FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

***CUSTOMER EXPERIENCE* E *CUSTOMER HEALTH SCORE* DE CLIENTES SAAS:**

Uma análise da parametrização das métricas que influenciam o *Customer Churn*

ADRIANO AKISSUE MARANGONI CAREZZATO

SÃO PAULO
2019

ADRIANO AKISSUE MARANGONI CAREZZATO

CUSTOMER EXPERIENCE E *CUSTOMER HEALTH SCORE* DE CLIENTES SAAS:

Uma análise da parametrização das métricas que influenciam o *Customer Churn*

Trabalho Aplicado apresentado a Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas, como requisito para a obtenção do título de Mestre em Gestão para a Competitividade.
Linha de pesquisa: Tecnologia de Informação.
Orientador: Prof. Dr. Eduardo de Rezende Francisco

SÃO PAULO
2019

Carezzato, Adriano Akissue Marangoni.
Customer Experience e Customer Health Score de clientes SaaS : uma análise da parametrização das métricas que influenciam o Customer Churn / Adriano Akissue Marangoni Carezzato. - 2019.
145 f.

Orientador: Eduardo de Rezende Francisco.
Dissertação (mestrado profissional MPGC) – Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo.

1. Clientes - Fidelização. 2. Comportamento do consumidor. 3. Marketing de relacionamento. 4. Sistemas de informação gerencial. 5. Indústria de software. I. Francisco, Eduardo de Rezende. II. Dissertação (mestrado profissional MPGC) – Escola de Administração de Empresas de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 654

Ficha Catalográfica elaborada por: Raphael Figueiredo Xavier CRB SP-009987/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

ADRIANO AKISSUE MARANGONI CAREZZATO

CUSTOMER EXPERIENCE E *CUSTOMER HEALTH SCORE* DE CLIENTES SAAS:

Uma análise da parametrização das métricas que influenciam o *Customer Churn*

Trabalho Aplicado apresentado à Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas, como requisito para a obtenção do título de Mestre em Gestão para a Competitividade.

Linha de pesquisa: Tecnologia de Informação.

Orientador: Prof. Dr. Eduardo de Rezende Francisco

Data da Aprovação
____/____/_____

Banca Examinadora:

_____
Prof. Dr. Eduardo de Rezende Francisco
(Orientador) FGV – EAESP

_____
Prof. Dr. José Luiz Carlos Kugler
FGV – EAESP

_____
Prof. Dr. Nelson Lerner Barth
FGV – EAESP

_____
Prof. Dr. Eric David Cohen
FMU

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Eduardo de Rezende Francisco, pela orientação, paciência e pelo encorajamento nos momentos mais desafiadores, bem como pelos preciosos ensinamentos e serenidade em todos os nossos contatos durante esses anos.

A todos os professores do mestrado profissional, por toda experiência e conhecimento compartilhados. Em especial, o Prof. Dr. Alberto Luiz Albertin, coordenador do MPGC-TI.

Aos estimados membros da banca avaliadora, que se dispuseram a participar e a contribuir com meu aprendizado.

Agradeço à minha família, que ao longo da vida me deu todo apoio, carinho e orientação, me permitindo chegar onde estou. Em especial aos meus pais, Armando e Lúcia e minha irmã Fabíola, meus maiores exemplos em todos os sentidos. Além deles, agradeço: Wagner, Mara, Francisco, Renata, Gabriela, Leandro; aos pequenos Bernardo, João, Nina, Penélope e Margot.

Aos meus avós que sempre me deram amor, apoio e foram uma inspiração em diversos aspectos na minha vida.

Agradeço à amigos que direta ou indiretamente me auxiliaram nesse processo, em especial ao Renan, Cintia, Coiro, Julia, Cadu e Bruno, que além de me apoiarem me inspiraram a seguir estudando.

Aos meus estimados colegas de trabalho que me ajudaram a obter o acesso aos dados, me auxiliaram nas análises dos mesmos à luz de uma visão do negócio e me permitiram e incentivaram buscar meu objetivo de fazer uma pesquisa aplicada a uma empresa. Em especial, agradeço ao Silvio Menezes, Juliane Manarin Vasconcelos Lopes, Alisson Poli de Carvalho, Rafael Rui, Arlindo Correia Junior, Wellington Schiavon, André Mihalik e Humberto Aillon.

Aos meus amigos e colegas da primeira turma de Mestrado Profissional em Gestão para a Competitividade na linha de TI da FGV-EAESP, e demais linhas, que tornaram as inúmeras noites e dias de aulas mais enriquecedoras, tranquilas e divertidas.

Agradeço especialmente à minha esposa, melhor amiga, companheira de vida, inspiração para continuar sempre estudando e buscando ser uma pessoa melhor a cada dia, Mariana, incansável e agradável companhia e apoio durante toda essa jornada e nossa vida juntos.

“Courage is what it takes to stand up and speak.
Courage is also what it takes to sit down and listen.”
(Sir Winston Leonard Spencer-Churchill, 1874 – 1965)

RESUMO

Apesar da rotatividade de clientes, ou *Customer Churn*, nas indústrias de serviços ter sido estudada extensivamente, o *Software as a Service* ainda não representa um foco específico de pesquisas acadêmicas especialmente no que concerne as métricas que compõem o *Customer Health Score*. Buscou-se oferecer uma metodologia para criação de um modelo preditivo de *Customer Churn* a partir do cálculo de *Customer Health Score*. Levou-se em consideração seus diversos indicadores e variáveis que podem ser medidos em todas as etapas de interação dos clientes ao longo de sua jornada. Tomou-se como foco uma empresa líder na América Latina em *software*, cuja classe de produtos principal é o *ERP*, do inglês *Enterprise Resource Planning.* A análise foi realizada por meio de dois algoritmos preditivos – regressão logística, que é um modelo já bastante estudado e de uso bastante consolidado, e o *XGBoost*, que é um modelo mais novo e que tem apresentado bons resultados em diversas aplicações– utilizando-se principalmente dados coletados em interações do cliente com a companhia. Foram considerados dados de 40.635 clientes que compõem a carteira de clientes com receita recorrente da companhia. As contribuições de quatro categorias de variáveis que poderiam compor um *Customer Health Score* foram analisadas: informações de interação da companhia com o cliente, informações sobre o cliente, informações sobre os contratos existentes e informações comerciais. Os modelos testados nesse trabalho analisaram, ao todo, 396 variáveis candidatas coletadas num período de 19 meses – de setembro de 2017 a março de 2019 – representando um número de dados unitários superior a 220 milhões. O modelo de *XGBoost* apresentou bons resultados de indicadores quantitativos para o caso estudado tendo os melhores índices de *TPR* (0,6547), Precisão (0,7268) e *F-Score* (0,6889). A metodologia aplicada é descrita em nível operacional, permitindo assim que outros pesquisadores e analistas possam replicá-la em contextos diferentes e mais amplos.

Palavras-chave: *customer experience*, *customer churn*, métricas de *customer churn*, seleção de variáveis para predição de *churn*, calibragem de modelo preditor de *churn*.

ABSTRACT

Although Customer Churn in service industries has been studied extensively, Software as a Service does not yet represent a specific focus of academic research, especially on what concerns the metrics that make up the Customer Health Score. This research sought to offer a methodology for creating a predictive model of Customer Churn starting from the Customer Health Score calculation. It took into consideration the various indicators and variables that can be measured at all stages of the customer journey. The focus of this study was a leading software company in Latin America, whose main product class is the Enterprise Resource Planning (ERP). The analysis was performed using two predictive algorithms - logistic regression, which is a well-studied and widely used model, and XGBoost, which is a newer model that has shown good results in many applications - using mainly data collected during customer interactions with the company. Data from the 40,635 company's customers portfolio, which currently compose its recurring revenue, were considered. Contributions from four categories of variables that could make up a Customer Health Score were analyzed: company-customer interaction information, customer information, existing contract information, and sales information. The models tested in this study analyzed a total of 396 candidate variables collected over a 19-month period – from September 2017 to March 2019 - representing a number of unit data of over 220 million. The XGBoost model presented good quantitative indicators results for the case study having the best TPR (0.6547), Precision (0.7268) and F-Score (0.6889) indices. The applied methodology is described at the operational level, thus allowing other researchers and analysts to replicate it in different and broader contexts.

Keywords: customer experience, customer churn, customer churn metrics, churn prediction feature selection, churn predictor model calibration.

LISTA DE FIGURAS

Figura 1 – Gráfico da previsão de receita mundial de serviços de nuvens públicas até 2022 (em US$ bilhões). 16
Figura 2 - Desenvolvimento temporal e maturidade teórica do *Customer experience*..... 22
Figura 3 - Gráfico da relação entre probabilidade e odds..... 33
Figura 4 - Exemplo de curva ROC. 36
Figura 5 - Tela da carga dos dados originais na ferramenta Radiant para execução da regressão logística. 56
Figura 6 - Seleção do Modelo de regressão logística na ferramenta Radiant. 57
Figura 7 – 1ª Rodada de geração do modelo de regressão logística na ferramenta Radiant para as variáveis originais. 58
Figura 8 - 2ª Rodada de geração do modelo de regressão logística na ferramenta Radiant para as variáveis originais. 59
Figura 9 – Gráfico do número de variáveis versus Scores de cross-validation (primeira rodada). 61
Figura 10 - Número de variáveis versus Scores de cross-validation (segunda rodada)..... 61
Figura 11 - Geração do modelo de regressão logística na ferramenta Radiant para as variáveis selecionadas a partir de filtros e wrappers. 63
Figura 12 – Indicadores VIF e Odds Ratio das variáveis do modelo de regressão logística, selecionadas a partir de filtros e wrappers. 64
Figura 13 – Variáveis remanescentes após aplicação da regressão logística com eliminação de variáveis pelo modelo XGBoost. 66
Figura 14 – 1ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost. 67
Figura 15 – Indicadores da 1ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost. 68
Figura 16 - 3ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost. 69
Figura 17 – Indicadores da 3ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost. 70
Figura 18 – Etapa de avaliação do modelo. 71
Figura 19 – Indicadores de avaliação do modelo..... 72
Figura 20 - Carteira de investimentos em Tecnologia da Informação. 79

LISTA DE TABELAS

Tabela 1 - Previsão de receita mundial de serviços de nuvens públicas até 2022 (em US$ bilhões). 15
Tabela 2 - Relação entre probabilidade e chance. 32
Tabela 3 - Amostra dos dados originais dos indicadores que compunham o Health Score adotado pela empresa. 40
Tabela 4 - Amostra dos dados dos indicadores após tratamentos e criação de novas dimensões transformadas. 43
Tabela 5 -Variáveis selecionadas a partir da aplicação de filtros e wrappers. 47
Tabela 6 - Lista das 20 variáveis que têm maior importância segundo a análise realizada com o XGBoost. 49
Tabela 7 - Cálculo dos scores dos indicadores que compõem o Health Score original utilizado pela companhia. 52
Tabela 8 - Resumo da quantidade e percentual de clientes da amostra que preencheram o NPS e que têm pedidos de Churn Parcial e Churn Total. 54
Tabela 9 – Comparativo quantitativo dos modelos aplicados no estudo. 75
Tabela 10 – Categorias de Implicação Gerencial na comparação dos modelos aplicados no estudo. 75
Tabela 11 - Matriz de confusão da regressão logística com as variáveis originais em quantidades. 76
Tabela 12 - Matriz de confusão da regressão logística com as variáveis originais em percentuais do total do número total de dados utilizados. 76
Tabela 13 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação de filtro e do wrapper para redução do número de variáveis em quantidades. 76
Tabela 14 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação de filtro e do wrapper para redução do número de variáveis em percentuais do total do número total de dados utilizados. 76
Tabela 15 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação do XGBoost para redução do número de variáveis em quantidades. 77
Tabela 16 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação do XGBoost para redução do número de variáveis em percentuais do total do número total de dados utilizados. 77
Tabela 17 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação apenas do XGBoost em quantidades. 77
Tabela 18 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação apenas do XGBoost em percentuais do total do número total de dados utilizados. . 77
Tabela 19 - Lista das 20 variáveis que têm maior importância segundo a análise realizada com o XGBoost. 83

LISTA DE QUADROS

Quadro 1 - Cálculos de indicadores da Matriz de Confusão para duas classes. 35
Quadro 2 - Lógicas de criação e tipos de variáveis geradas a partir dos dados disponíveis. 44

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface* (Interface de Programação de Aplicativos)
ARR	*Annual Recurring Revenue* (Receita Recorrente Anual)
AUC	*Area Under the Curve* (Área Abaixo da Curva)
B2B	*Business-to-Business* (Empresa-para-Empresa)
B2C	*Business-to-Customer* (Empresa-para-Cliente)
BPaaS	*Business Process as a Service* (Processo de Negócio como Serviço)
CAC	*Customer Acquisition Cost* (Custo de Aquisição de Cliente)
CDR	*Churn Detection Rate* (Taxa de Detecção de *Churn*)
CEO	*Chief Executive Officer* (Diretor Executivo)
CHS	*Customer Health Score* (Pontuação de Saúde do Cliente)
CLC	*Customer Lifecycle* (Ciclo de Vida do Cliente)
CLTV	*Customer Lifetime Value* (Valor do Cliente em Todo Tempo de Vida)
CNPJ	Cadastro Nacional de Pessoa Jurídica
CRM	*Customer Relationship Management* (Gestão de Relacionamento com Clientes)
CSAT	*Customer Satisfaction Score* (Pontuação de Satisfação do Cliente)
CX	*Customer Experience* (Experiência do Cliente)
ERP	*Enterprise Resource Planning* (Planejamento de Recursos Empresariais)
Fn	*False negative* (Falso negativo)
Fp	*False positive* (Falso positivo)
IaaS	*Infrastructure as a Service* (Infraestrutura como Serviço)
KPI	*Key Performance Indicator* (Indicador-Chave de Performance)
NPS	*Net Promoter Score* (Pontuação Líquida de Promotores)
OKRs	*Objectives and Key Results* (Objetivos e Resultados-Chave)
OOT	*Out of Time* (Fora do Tempo)
PaaS	*Platform as a Service* (Plataforma como Serviço)
P&D	Pesquisa e Desenvolvimento
ROC	*Receiver Operating Characteristic* (Características de Operação do Receptor)
SaaS	*Software as a Service* (*Software* como Serviço)
SLA	*Service Level Agreement* (ANS - Acordo de Nível de Serviço)
SI	Sistemas de Informação
SOA	*Service Oriented Architecture* (Arquitetura Orientada a Serviço)
TI	Tecnologia da Informação
Tn	*True positive* (Verdadeiro negativo)
Tp	*True positive* (Verdadeiro positivo)
TPR	*True Positive Rate* (TVP - Taxa de Verdadeiro Positivo)

SUMÁRIO

1 INTRODUÇÃO 13

1.1 Contexto e justificativa 13

1.2 Perguntas de pesquisa 19

1.3 Objetivos 19

1.3.1 Objetivo 19

1.3.2 Etapas da Pesquisa 20

1.4 Estrutura do trabalho 20

2 REVISÃO DA LITERATURA / FUNDAMENTAÇÃO TEÓRICA 21

2.1 *Customer Experience, Lifeclycle, Lifetime Value* 21

2.2 *Customer Health Score* 24

2.3 *Customer Churn* 24

2.4 *SaaS* 25

3 METODOLOGIA 28

3.1 Aprendizado de máquina 28

3.1.1 Definição de machine learning (aprendizado de máquina) 28

3.1.2 Etapas da metodologia de seleção de algoritmos de machine learning 29

3.1 Métodos de classificação 30

3.1.1 Regressão logística 30

3.2 Avaliação dos modelos preditivos 34

3.2.1 Matriz de Confusão 34

3.2.2 Curva ROC (Receiver Operating Characteristic Curve) 36

3.3.1 Primeira fase da coleta de dados 37

3.3.2 Segunda fase da coleta de dados 39

3.4 Ferramentas de manipulação de dados 39

3.5 Análise preliminar, tratamento e limpeza de dados 39

3.6 Criação de variáveis candidatas / *Feature Engineering* 44

3.7 Processo de seleção de variáveis para inclusão nos modelos de regressão logística e *XGBoost* 46

3.8 Escolha do algoritmo de predição 50

4 RESULTADOS E DISCUSSÃO 51

4.1 Primeira etapa de análise 51

4.2 Segunda etapa da análise 60

4.3 Terceira etapa da análise 65
5 CONCLUSÕES 73
5.1 Implicações gerenciais 79
5.2 Sugestão de novas métricas 82
5.3 Limitações 84
5.4 Estudos futuros 85
REFERÊNCIAS BIBLIOGRÁFICAS 86
APÊNDICE A - VARIÁVEIS CANDIDATAS CRIADAS 94
APÊNDICE B - TRECHOS DE CÓDIGO EM PYTHON – ANÁLISE INICIAL DOS DADOS 107
APÊNDICE C - TRECHOS DE CÓDIGO EM PYTHON – SELEÇÃO DE VARIÁVEIS 112
APÊNDICE D - TRECHOS DE CÓDIGO EM PYTHON – APLICAÇÃO DO XGBOOST 127
APÊNDICE E - TRECHOS DE CÓDIGO EM PYTHON – SALVANDO AS VARIÁVEIS RESULTANTES DO XGBOOST 138

1 INTRODUÇÃO

Nas últimas décadas, as tecnologias de computação emergentes levaram a uma profunda evolução na capacidade de as empresas coletarem, armazenarem e analisarem grandes conjuntos de dados. Dessa forma, para cada cliente, milhares ou até milhões de objetos de dados são armazenados, permitindo a análise do histórico completo de compras dele. Outrossim, as mudanças no relacionamento entre empresas e clientes, devido às recentes mudanças econômicas e sociais, fizeram as empresas mudarem do *marketing* de transação para o *marketing* de relacionamento (MIGUÉIS *et al.*, 2012).

1.1 Contexto e justificativa

Quando pensamos em histórias de sucesso da comercialização de *software* como serviço − do termo em inglês *Software as Service* ou seu acrônimo *SaaS* − algumas das empresas que vêm a nossa mente são: Salesforce, Adobe, Microsoft, Shopify, Workday, Zendesk, LinkedIn. Antes de entender aonde esse mercado pode chegar, porém, surge uma questão: quais fatores podem levar uma empresa a se destacar em seu mercado de atuação? Ou, indo mais a fundo, que premissa pode manter uma determinada corporação como um dos motores de uma disrupção em um mercado e entre as líderes desse mercado ao longo de décadas?

Desde 1999, cerca de 20 anos atrás, Marc Benioff, cofundador e *CEO* da empresa Salesforce, respaldou-se no seguinte sonho: "Tornar o uso da tecnologia dentro das empresas tão simples quanto fazer uma compra pela *Amazon*" (MUCH, 2019).

Passadas duas décadas, a visão de Benioff é uma realidade e, tendo se tornado conhecida principalmente por ter trilhado um caminho de sucesso na "nuvem", a companhia pode ser considerada a pioneira e a responsável por criar o caminho pelo qual outras gigantes da tecnologia como Adobe e Microsoft – selecionando apenas algumas das maiores dentre as anteriormente citadas – passariam alguns anos mais tarde. Essas empresas têm alcançando resultados significativos que apontam que esse é mesmo um caminho de sucesso a ser seguido.

Desde seus primórdios, o que tornou a Salesforce um caso a ser estudado por outros fabricantes de *software* foi o modelo de cobrança por assinatura. Nesse modelo, os clientes, independentemente de seu porte ou segmento, teriam a possibilidade de ter o acesso ao *software* e expandi-lo de acordo com o crescimento do próprio negócio, sem ter a necessidade de se

preocupar com investimentos e/ou gestão de uma infraestrutura que garantissem seu funcionamento. Foi uma forma pioneira de se oferecer um serviço por meio de aplicativos *web*, algo que posteriormente passou a ser conhecido como "Computação em Nuvem" ou, em inglês, *Cloud Computing*. Esse termo, hoje em dia, é corriqueiro, mas, no final da década de 1990, mostrou-se algo inovador e revolucionário.

Youseff e Butrico Da Silva (2008) definiram a Computação em Nuvem como sendo o conjunto de serviços usados sob demanda por meio de redes. Esses autores dividem os serviços de nuvem em cinco camadas: aplicativos em nuvem, ambiente de *software* em nuvem, infraestrutura em nuvem, *kernel* de *software* e *firmware* / *hardware*. A forma mais conhecida de aplicativos em nuvem é a do *software* como serviço – ou em inglês *Software as a Service* (*SaaS*). Nessa modalidade, um aplicativo é usado via rede, sem ser instalado no dispositivo do usuário. Portanto, *SaaS* é um modelo de distribuição de *software* no qual os aplicativos são hospedados por um fornecedor ou provedor de serviços e disponibilizados aos clientes em uma rede, geralmente na internet.

O *SaaS* está se tornando um modelo de distribuição cada vez mais predominante, pois tecnologias subjacentes que oferecem suporte a serviços da *web* e arquitetura orientada a serviços – ou, em inglês, *Service Oriented Architecture* (*SOA*) – amadurecem e novas abordagens de desenvolvimento se tornam populares. O *SaaS* também é frequentemente associado a um modelo de licenciamento de assinatura pré-pago. Os aplicativos *SaaS* também devem interagir com outros dados e outros aplicativos em uma variedade igualmente ampla de ambientes e plataformas. Ele é implementado com mais frequência para fornecer aos clientes corporativos funcionalidade de *software* comercial a um baixo custo, permitindo que esses clientes obtenham os mesmos benefícios do *software* operado internamente e comercialmente licenciado sem a complexidade associada de instalação, gerenciamento, suporte, licenciamento e alto custo inicial.

De acordo com o Gartner (2019), o mercado de *SaaS* crescerá em 2019 cerca de 17,5%. O segmento de crescimento mais rápido será justamente o de serviços de infraestrutura de sistemas em nuvem, ou infraestrutura como serviço − em inglês *Infrastructure as a Service* (*IaaS*) −, que deve crescer 27,5% em 2019, como se pode observar na Tabela 1, na página seguinte.

Tabela 1 - Previsão de receita mundial de serviços de nuvens públicas até 2022 (em US$ bilhões).[1]

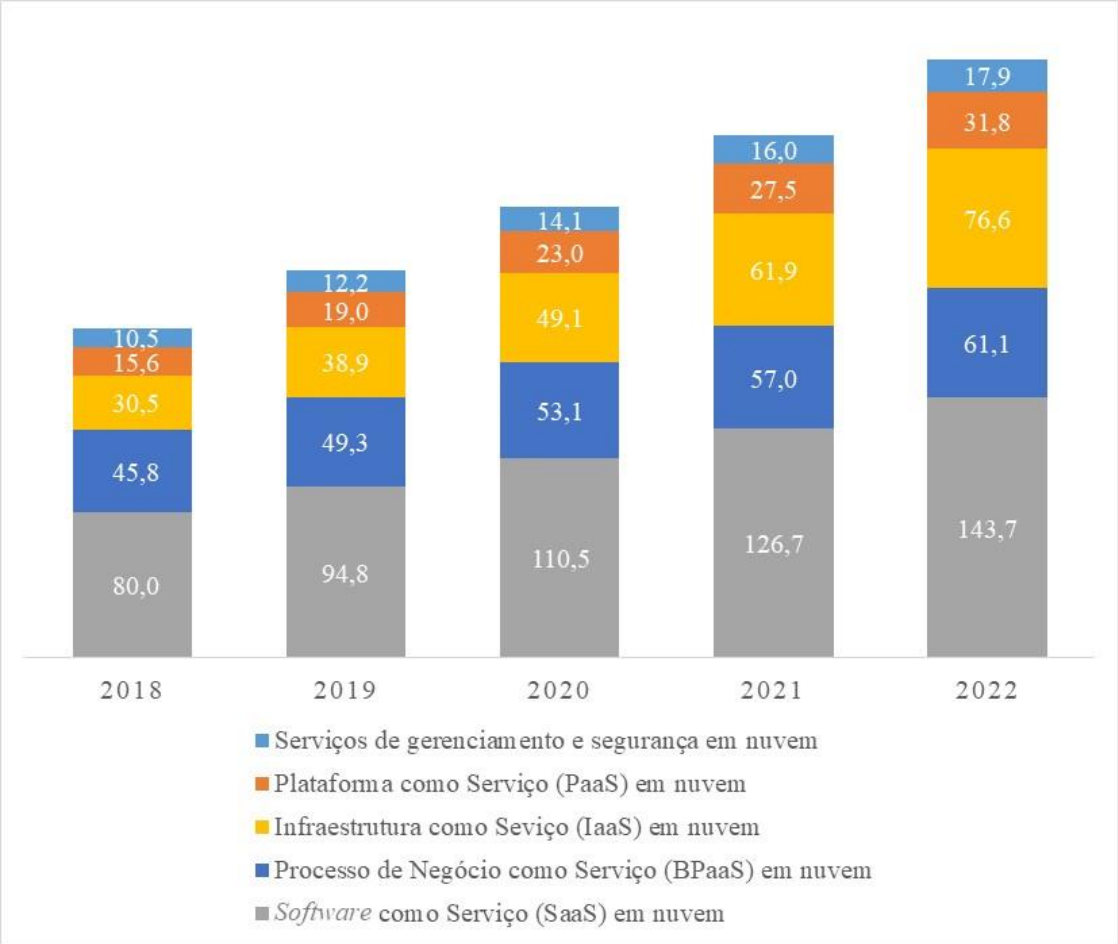
Tipo de serviço em nuvem	2018	2019	2020	2021	2022
Software como Serviço (*SaaS*) em nuvem	80,0	94,8	110,5	126,7	143,7
% Variação ano sobre ano		18,5%	16,6%	14,7%	13,4%
Processo de Negócio como Serviço (*BPaaS*) em nuvem	45,8	49,3	53,1	57,0	61,1
% Variação ano sobre ano		7,6%	7,7%	7,3%	7,2%
Infraestrutura como Seviço (IaaS) em nuvem	30,5	38,9	49,1	61,9	76,6
% Variação ano sobre ano		27,5%	26,2%	26,1%	23,7%
Plataforma como Serviço (*PaaS*) em nuvem	15,6	19,0	23,0	27,5	31,8
% Variação ano sobre ano		21,8%	21,1%	19,6%	15,6%
Serviços de gerenciamento e segurança em nuvem	10,5	12,2	14,1	16,0	17,9
% Variação ano sobre ano		16,2%	15,6%	13,5%	11,9%
Mercado Total	182,4	214,3	249,8	289,1	331,2
% Variação ano sobre ano		17,5%	16,6%	15,7%	14,6%

Fonte: elaboração própria, com base em Gartner (2019).

Ainda segundo o Gartner (2019), até 2022, aproximadamente 90% das organizações que compram *IaaS* em nuvem pública passarão a comprá-la de um provedor integrado de *IaaS* e plataforma como serviço – do inglês, *Platform as a Service* (*PaaS*) –, e usarão as variáveis de *IaaS* e *PaaS* desse provedor. Quando se trata da categoria de processos de negócios como um serviço *BPaaS* – do inglês, *Business Process as a Service* –, o Gartner (2019) espera ver a receita crescer 7,9%, para US$ 49,3 bilhões em 2019. Globalmente falando, como atesta a Figura 1, na página seguinte, o Gartner prevê que, até 2022, as receitas da nuvem totalizem US$ 278 bilhões.

[1] Os totais podem não bater com os valores parciais devido ao arredondamento.

Figura 1 – Gráfico da previsão de receita mundial de serviços de nuvens públicas até 2022 (em US$ bilhões).



Fonte: elaboração própria, com base em Gartner (2019).

Os serviços de aplicativos na nuvem − ou *SaaS* − continuam sendo o maior segmento do mercado de nuvem, com uma receita que deve crescer 17,8% e alcançar o marco de US$ 85,1 bilhões em 2019. Para empresas de *software*, isso significará novas oportunidades de negócios, bem como desafios na adaptação ao novo ambiente de negócios.

A transição de um provedor de negócios de *software* empacotado para um provedor de *SaaS* seguindo as melhores práticas desse modelo impactará não só o gerenciamento estratégico, mas também *marketing* e vendas. Até agora, a literatura acadêmica não parece ter abordado esse problema, embora muitos profissionais venham publicando um volume significativo de material relacionado ao tema na *web*. Kaplan (2006) e Mallya (2009) se concentraram nas diferenças entre *marketing* e venda de *software* tradicional e *SaaS*, enfatizando o fornecimento de uma solução para o cliente, em vez de vender *SaaS* como tal. Mallya (2009), por sua vez, concentrou-se em fornecer valor para o cliente ao vender *SaaS*. Em

geral, alguns dos estudos encontrados destacaram a facilidade de vender *SaaS*, enquanto outros se voltaram para a dificuldade nesse mesmo processo.

Ao mesmo tempo em que há um enorme potencial de crescimento do mercado *SaaS*, há que se considerar o quanto esse conceito ainda é recente. A Adobe, uma seguidora na estrada pavimentada pela Salesforce, começou sua migração para a nuvem apenas em 2009. Já a Microsoft, outra gigante desse mercado que também tem dirigido seus negócios por esse caminho, verificou em 2017, pela primeira vez, uma superação na ordem de 43% dos resultados financeiros gerados pela assinatura corporativa do *Office 365* em relação ao licenciamento tradicional. No mesmo ano, a empresa viu cair o faturamento das licenças permanentes em 17%. Com tantas evidências, ninguém duvida que o modelo *SaaS* veio para ficar e que sua sustentação se dá por um pilar central chamado "sucesso do cliente" − em inglês, *Customer Success*. Nessa estruturação, a Salesforce nasceu pronta e serviu como inspiração para toda uma nova economia que despontou na sequência.

Em 2000, o americano Jeremy Rifkin escreveu a obra **The Age of Access**[2], na qual previa um futuro em que a necessidade de aquisição de um bem seria suprida pelo acesso a uma experiência. Não é preciso dizer que essa tese passou a ser um divisor de águas dentro dos mais variados negócios da economia. A indústria de *software* respondeu à valorização do acesso, prevista por Rifkin (2000), transformando a compra de licença em um modelo de assinatura, mas esse não foi um movimento rápido. Na verdade, ele ainda está acontecendo na sociedade atual. A escala, nesse modelo, dá-se de forma mais ampla, pois usa a transformação digital e a fluidez da internet para apoiar essa que já é chamada de 4ª Revolução Industrial e que vem rompendo modelos tradicionais de negócios. Nada faz mais sentido hoje do que a integração das tecnologias digitais, físicas e biológicas. É a projeção de uma inteligência artificial que está mudando a forma como vivemos. No ambiente corporativo, as análises se tornaram preditivas.

E ainda há muito potencial de crescimento. O *2017 Subscription Economy® Index™* (*SEI*), estudo conduzido pela Zuora (2017), uma das mais representativas plataformas de assinaturas do mercado, apontou que, até 2020, 50% das maiores empresas verão grande parte dos seus negócios depender de sua capacidade do aprimoramento de produtos, serviços e experiências no âmbito digital. Além do mercado de *software*, o modelo de assinaturas já é visto em saúde, educação, academias de ginástica, agricultura, indústria automotiva, *marketing* etc. E essa é mesmo a ideia da economia da recorrência: conquistar e fidelizar clientes rumo a

[2] A obra foi traduzida para o português (**A Era do Acesso**) e publicada no Brasil no ano seguinte, pela Makron Books.

uma jornada de longa duração. As oportunidades oferecidas pelo modelo *SaaS* aos negócios são inúmeras e beneficiam estruturas de todos os portes.

Diversos fatores contribuem para isso:

- fácil instalação;
- acessibilidade;
- versões sempre atualizadas;
- suporte simplificado;
- possibilidade de integração com outros programas – APIs (Interfaces de Programação de Aplicativo, em inglês *Application Programming Interfaces*);
- desempenho escalável;
- geração de dados sobre o comportamento do usuário;
- acesso a dados preditivos.

Para os fornecedores, um dos maiores atrativos para a adoção do modelo *SaaS* é, sem dúvida, a receita recorrente trazida pelos seus clientes. No entanto, as renovações de assinatura do fornecedor tornam a empresa muito mais sensível à satisfação do cliente. Quando um cliente adquire uma licença de *software* tradicional, o produto é instalado nos servidores do cliente e o fornecedor não tem a visibilidade de como o cliente está usando o produto. Os fornecedores *SaaS* hospedam seus aplicativos executando o *software* em servidores gerenciados por eles, o que permite que coletem dados valiosos sobre como os clientes estão usando seus produtos, incluindo quem os está usando, quando, por quanto tempo e com que frequência. Muitas indústrias, como telecomunicações e bancos, já contam com padrões de uso para prever a rotatividade de clientes e a satisfação do cliente. Os fornecedores de *SaaS* têm a capacidade de prever características semelhantes em seus clientes.

A rotatividade de clientes, ou *Customer Churn*, nas indústrias de serviços tem sido estudada extensivamente. No entanto, o *Software as a Service* ainda não representa um foco específico de pesquisas acadêmicas. Embora a indústria de *SaaS* compartilhe muitas semelhanças com os setores de telecomunicações e outros serviços, há algumas diferenças que merecem ser exploradas, especialmente devido ao rápido crescimento do mercado de *SaaS* e à crescente disponibilidade e confiabilidade da computação em nuvem.

Esta pesquisa descreve um estudo sobre métricas que podem ser relacionadas ao *Customer Churn* na indústria de *SaaS* e que poderiam compor o *Customer Health Score*. O

estudo foi conduzido utilizando-se informações de uma empresa líder em *software* na América Latina, que tem como seu carro chefe a linha de produtos *ERP*, do inglês *Enterprise Resource Planning*. Este trabalho traz uma análise das diversas métricas que podem ser utilizadas para compor o *Customer Health Score*, descreve um experimento da própria empresa que aplica métodos existentes para prever o *Customer Churn*, trazendo uma sugestão de pesos a serem utilizados no caso especificamente de produtos *SaaS*. É importante se destacar que o trabalho possuí um caráter exploratório, sendo os dados utilizados os que permitiram concluir o comportamento do *Customer Churn*. Foram trabalhadas as variáveis que se encontravam disponíveis na empresa no momento da análise, portanto, elas não são fruto de referencial teórico levantado sobre o assunto. Ao final, serão apresentadas oportunidades de pesquisa futura para entender melhor a rotatividade de clientes na indústria de *SaaS*.

1.2 Perguntas de pesquisa

Tendo em vista o objetivo e as etapas metodológicas apresentadas, elencam-se as seguintes perguntas de pesquisa:

1. Dados de interação com os clientes, contratos existentes e informações comerciais apresentam influência no *Customer Health Score*?
2. Que ações gerenciais podem contribuir para a diminuição do *Customer Churn* a partir de análises de dados?

1.3 Objetivos

1.3.1 Objetivo

O objetivo geral desta pesquisa é calibrar o modelo de previsão de *Customer Churn*, ou rotatividade de clientes, a partir do cálculo de *Customer Health Score*. Esse indicador será composto de variáveis selecionadas por algoritmos de *Machine Learning* em dados de clientes obtidos de uma grande empresa nacional de *software*.

Busca-se aumentar o potencial de previsão do *Customer Churn*, de modo a prevenir perdas de receitas diretas e indiretas para a companhia provocadas pela saída desses clientes e, consequentemente, auxiliar empresas cujo modelo comercial se baseia em receitas recorrentes.

1.3.2 Etapas da Pesquisa

Para alcançar esse objetivo, as seguintes etapas metodoógicas foram estabelecidas. O caráter prescritivo deste projeto permitiu que essas etapas fossem elencadas junto aos objetivos e perguntas da pesquisa.

a) testar a significância das variáveis que compõem o modelo atual de cálculo de *Customer Health Score* utilizado pela companhia para previsão de *Customer Churn*;
b) reunir variáveis de interação com o cliente, contratos existentes e informações comerciais para ampliar rol de variáveis candidatas a comporem o *Customer Health Score*, visando o aumento da previsão de *Customer Churn*;
c) coletar dados temporais das variáveis selecionadas e restringir às mais significantes para previsão de *Customer Churn* através de aplicação de algoritmos de predição;
d) utilizar as análises e resultados para apontar desdobramentos gerenciais que possibilitem a redução de *Customer Churn*;

1.4 Estrutura do trabalho

Este trabalho está dividido em quatro seções.

Na *Revisão de literatura / Fundamentação teórica*, são apresentados os fundamentos que embasam o estudo realizado.

Na *Metodologia de pesquisa*, é explicitada a abordagem da pesquisa desenvolvida e são expostos os métodos adotados para coleta e análise de dados.

Na *Apresentação e discussão de resultados*, são descritos os resultados obtidos na pesquisa e analisados os resultados frente ao referencial teórico e ao contexto atual da empresa em foco.

Finalmente, nas *Conclusões*, são apresentadas as considerações decorrentes da análise, bem como as contribuições teóricas e práticas do estudo, as limitações da pesquisa e sugestões para estudos futuros.

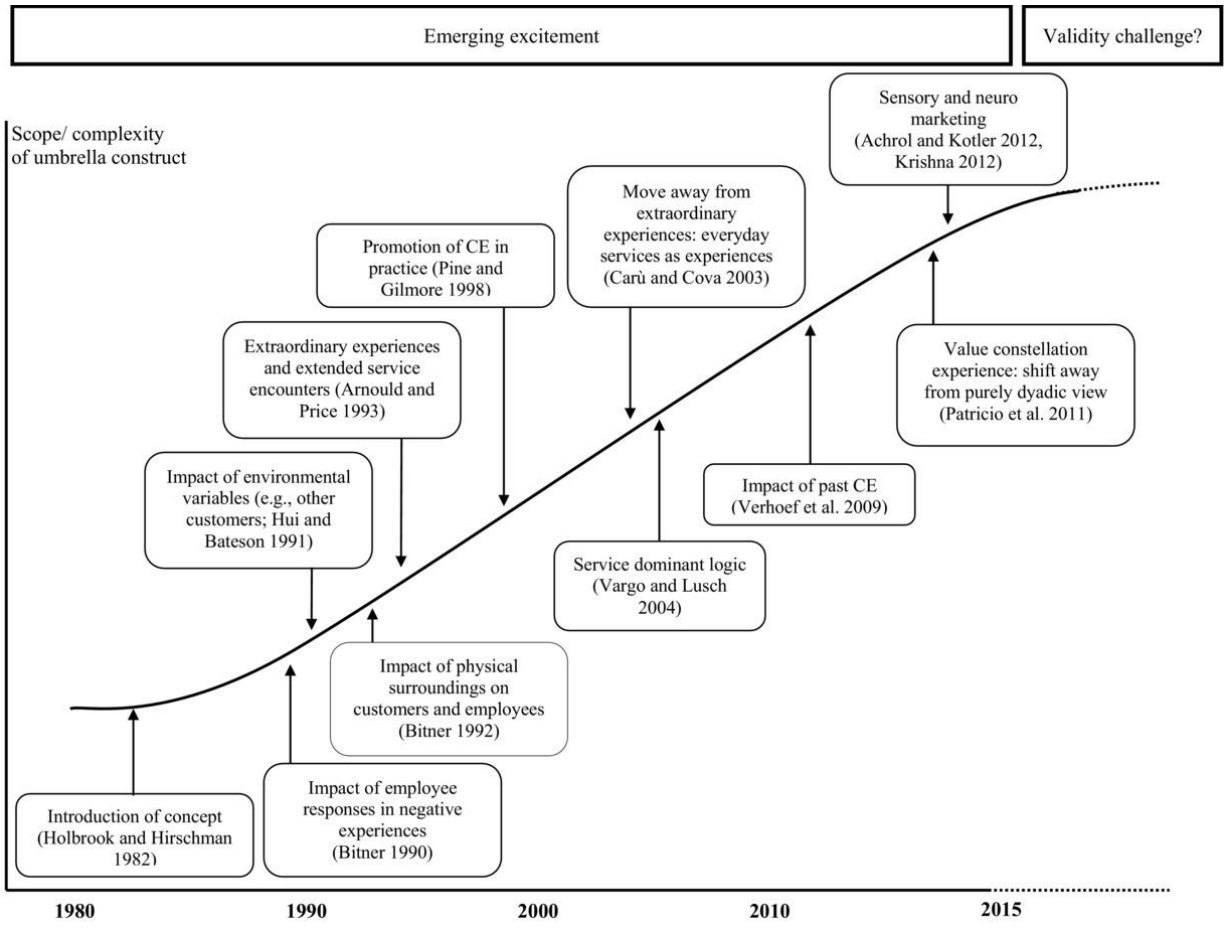
2 REVISÃO DA LITERATURA / FUNDAMENTAÇÃO TEÓRICA

2.1 *Customer Experience, Lifeclycle, Lifetime Value*

A dificuldade de se definir o exato conceito de *Customer Experience* (*CX*) − que, em tradução livre para português seria a Experiência do Consumidor −, pode ser explicada pelo fato de que, já há vários anos, muitos estudos se voltaram para tentar caracterizá-lo, motivo pelo qual muitas definições distintas estão disponíveis em fontes diversas. Segundo Kranzbühler *et al.* (2018), estudos iniciais abordaram os recursos gerenciáveis de encontros únicos de serviço: em particular, como mitigar os efeitos de experiências negativas de serviço, concentrando-se principalmente nos fatores sob o controle da empresa. Com o tempo, a pesquisa avançou desde o encontro de serviço em si até a consideração de fatores ambientais que também podem estar além do controle da empresa. Ainda segundo os autores, para esse fim, foi iniciado um extenso volume de pesquisas para desvendar o mundo dos serviços, abrangendo não só ambiente físico, atmosfera, *layout* espacial, funcionalidade de uma loja, mas também sinais, símbolos e artefatos. Mais recentemente, o campo emergente de *marketing* sensorial descreveu o uso de gatilhos subconscientes que atraem os sentidos humanos do olfato, paladar, tato, visão e audição para criar um *CX* sensorial. Consequentemente, o conceito tem sido maleável, respondendo a desenvolvimentos emergentes no campo e na prática do *CX*, expandindo-se em escopo e profundidade ao longo do tempo. A

Figura **2** mostra uma síntese da trajetória da definição do termo:

Figura 2 – Desenvolvimento temporal e maturidade teórica do *Customer experience*.



Fonte: Kranzbühler *et al.* (2018) (mantido original)

O conceito "Experiência do Consumidor" foi introduzido pela primeira vez por Holbrook e Hirschman (1982). Eles postularam que os consumidores não são solucionadores de problemas puramente racionais. Portanto, a ponderação cognitiva de valor e preço pode explicar apenas uma pequena fração do comportamento do consumidor. Os autores propuseram a substituição da teoria anterior do processamento de informações por uma abordagem experimental, enfatizando emoções, sentimentos e subconsciência. O objetivo era capturar holisticamente a experiência de consumo, que não é apenas cognitiva, mas também hedônica, simbólica e estética. Dessa forma, a CX pode ser caracterizada como um fenômeno subjetivo que não é totalmente controlável pela empresa (VERHOEF *et al.*, 2009).

Segundo Kranzbühler *et al.* (2018), além da noção holística de Holbrook e Hirschman (1982), os primeiros estudos abordaram os recursos gerenciáveis dos encontros únicos de serviço: em particular, como mitigar os efeitos de experiências negativas de serviço, concentrando-se principalmente nos fatores sob o controle da empresa. Com o tempo, a

pesquisa foi além do encontro de serviço em si para considerar fatores ambientais que também podem estar além do controle da empresa. Para esse fim, foi iniciado um extenso volume de pesquisas investigando outros aspectos dos serviços, como ambiente físico, atmosfera, layout espacial, funcionalidade de uma loja, mas também sinais, símbolos e artefatos. Mais recentemente, o campo emergente do marketing sensorial descreve o uso de gatilhos subconscientes que apelam aos sentidos humanos do olfato, paladar, tato, visão e audição para criar um Customer Experience sensorial. Conseqüentemente, o conceito de Customer Experience tem sido intocável, respondendo a desenvolvimentos emergentes em seu campo e em sua prática, expandindo-se em escopo e profundidade ao longo do tempo. Em contraste com as idéias obtidas em encontros únicos de serviço, surgiu uma literatura paralela examinando vários encontros de serviço, enriquecendo o conceito de Customer Experience considerando as influências temporais em uma base longitudinal. Reconhecendo que as experiências não ocorrem exclusivamente em nenhum momento, esses estudos enquadraram as experiências em uma série de encontros repetidos e seqüenciais. Isso complica ainda mais o entendimento de Customer Experience, pois esses encontros também incluem fatores que estão além do controle da empresa (KRANZBÜHLER *et al.*, 2018).

Bó (2012) define *Customer Experience* como a sensação ou percepção dos clientes sobre o produto / marca resultante de suas interações como um todo com a marca durante o ciclo de vida do cliente. O ciclo de vida do cliente − do inglês *Customer Lifeclycle* (*CLC*) −, por sua vez, baseia-se no conceito do ciclo de vida do produto, uma estrutura descritiva que classifica a evolução dos mercados de produtos em quatro estágios distintos: introdução, crescimento, maturidade e declínio (MOON, 2005; RINK; SWAN, 1979).

Outrossim, as interações com os clientes em cada estágio tendem a ter efeito no período em que os clientes permanecerão na empresa (FREEZE; KULKARNI, 2007). Portanto, são necessários diferentes tipos de ações e comunicações em cada um dos quatro estágios citados anteriormente. A perspectiva do ciclo de vida é significativa para o *marketing*, porque rastreia as atividades do cliente e as combina com ações adequadas à medida que os clientes passam pelos estágios. Isso garante que, em cada fase, a empresa priorize a maximização do valor do cliente em todo tempo de vida − em inglês, *Customer Lifetime Value* (*CLTV*), o que geralmente aumenta seu fluxo de caixa (HANSOTIA, 2002). O *CLTV* pode ser definido como a receita total que uma empresa pode esperar que um cliente gere ao longo do seu relacionamento comercial com ela (FONTANELLA, 2019).

2.2 *Customer Health Score*

As organizações de serviços precisam ser proativas na compreensão dos níveis de satisfação atuais dos clientes antes que esses clientes virem um *Churn* (BOLTON, 1998). O *Customer Health Score* (CHS) – que, em tradução livre para português seria a *Score* de Satisfação do Cliente – correlaciona-se com a propensão à perda ou crescimento da receita advinda de um determinado cliente. Uma boa pontuação de *Health Score*, por definição, é uma indicação de satisfação do cliente. Uma pontuação ruim é um sinal de alerta precoce para sua saída (FORRESTER CONSULTING, 2014). Outras definições mencionam o nível de lealdade do cliente e o valor agregado como partes da saúde do cliente (RABOY, 2013; ABBOTT, 2017), portanto, o conceito está vinculado ao bem-estar de um cliente de alguma forma.

2.3 *Customer Churn*

Muitas empresas sofrem com o problema substancial de saída de clientes, devido à concorrência acirrada resultante de mercados saturados, condições dinâmicas de mercado e introdução contínua de novas ofertas competitivas. Em resposta, muitas delas passaram de uma estratégia centrada em ofertas, projetada para vender o maior número possível de produtos para uma abordagem de retenção orientada para o cliente, o *Customer Experience*, que busca explicitamente reduzir a rotatividade de clientes – em inglês *Customer Churn* (BLATTBERG; KIM; NESLIN, 2010). *Customer Churn* pode ser definido como a taxa de clientes que saem (não renovam ou cancelam sua inscrição em um serviço) em um determinado período de tempo (JANZER, 2017). Assim também, pode ser definido como a mudança de lealdade do cliente de um provedor de serviços para outro (FARQUAD; RAVI; RAJU, 2014).

Na indústria de *software* em específico, devido à nova forma de licenciamento via *SaaS*, o cliente não possui mais um contrato permanente com o fabricante e pode cancelar a assinatura a cada período de licenciamento. Além disso, os clientes muitas vezes não têm mais um investimento inicial tão alto, o que diminui a barreira de saída deles, e ainda por cima podem escolher entre um número crescente de ofertas *SaaS* disponíveis no mercado. Como resultado dessa somatória de fatores, o relacionamento via *SaaS* é caracterizado por custos de troca relativamente baixos, a ponto de os clientes poderem trocar seus fornecedores de *software* com mais facilidade, se comparado ao modelo tradicional de venda de licenças (BENLIAN; KOUFARIS; HESS, 2011). Consequentemente, os fornecedores de *software* tornaram-se

dependentes da renovação contínua da assinatura de seus clientes, na medida em que a capacidade de recuperar o alto custo de aquisição do cliente − em inglês *Customer Acquisition Cost* (*CAC*) −, no que tange às vendas iniciais de licenças, característico do modelo perpétuo de licenciamento, desapareceu, senão totalmente, ao menos parcialmente, devido ao novo modelo de licenciamento via *SaaS* (DEMPSEY; KELLIHER, 2018). Uma vez que, em média, uma empresa *SaaS* leva cerca de doze meses para equilibrar as despesas com o *CAC* de cada consumidor, fica claro que o *Customer Churn* se tornou um risco significativo para o desempenho econômico das empresas de *SaaS* (GE *et al.*, 2017).

Dempsey e Kelliher (2018) introduziram a "Lei de Assinatura de *Churn*" − em inglês, *Subscription Law of Churn* −, que postula que uma empresa de *SaaS* não crescerá e nem ao menos sobreviverá caso a receita total com a assinatura não seja maior ou igual à receita perdida devido ao *Customer Churn*. Assim, um dos principais desafios do modelo de negócio *SaaS* implica garantir a renovação da assinatura, pois ela é crucial para a sobrevivência e sucesso do fornecedor desse serviço. Portanto, a fim de lidarem com essas mudanças, os fabricantes de *software* começaram a se concentrar cada vez mais em seu relacionamento com o cliente como um meio de garantir fluxos de receita recorrentes e sustentáveis (DEMPSEY; KELLIHER, 2018).

2.4 *SaaS*

Software as a Service ou *SaaS* – que, em tradução livre para português seria *Software* como Serviço – é a camada final do modelo de serviços em nuvem (NARANG; HUDIARA; KHURMI, 2014). É a forma mais conhecida de distribuição de aplicativos em nuvem, na qual um aplicativo é usado por meio de uma rede sem ser instalado no dispositivo do usuário (TYRVÄINEN; SELIN, 2014). Portanto, pode-se definir o *SaaS* como sendo um modelo de distribuição de *software* no qual os aplicativos são hospedados por um fornecedor ou provedor de serviços e disponibilizados aos clientes em uma rede, geralmente via internet (SATYANARAYANA, 2012). De uma maneira mais simples, pode-se afirmar que *SaaS* refere-se ao uso de serviços e aplicativos específicos pela internet (GOSCINSKI; BROCK, 2010; ZHANG; CHENG; BOUTABA, 2010).

O *SaaS* surge como uma abordagem inovadora para o fornecimento de aplicativos de *software* baseados em tecnologia de computação em nuvem (CHOU; CHOU, 2007). O modelo de cobrança do serviço *SaaS* é geralmente baseado em um pagamento calculado por usuário (JU *et al.*, 2010; LEE *et al.*, 2009; WALZ; GRIER, 2010), sendo que alguns desenvolvedores

cobram por uma quantidade mínima de usuários e armazenamento também de forma apartada (JU *et al.*, 2010). Nesse modelo de distribuição de aplicativos, uma forma alternativa para a cobrança por usuário é os provedores de *SaaS* permitirem aos usuários o consumo com base em suas necessidades e cobrarem pelos serviços de acordo com o uso real de cada um. Em outras palavras, permitirem a utilização de um modelo *pay-per-use* (ARMBRUST *et al.*, 2010). Essa abordagem de fornecimento de serviços "sob demanda" é semelhante ao modelo utilizado por serviços públicos fornecidos por governos: um usuário apenas assina um aplicativo sem a necessidade de comprar, instalar e de dispender esforços para manter as licenças de *software* atualizadas, assim como se consome energia da rede pública ao invés de um gerador particular por exemplo (GOSCINSKI; BROCK, 2010). Em outras palavras, o modelo de receita é uma assinatura por serviço, com pagamentos regulares e recorrentes em troca de acesso contínuo a um aplicativo (SUKOU; GRANT, 2013).

Outro fator que conta a favor do *SaaS* é a melhora na qualidade dos serviços de *software* por meio de atualizações automáticas de aplicativos e do backup de dados (XIN; LEVINA, 2008). Ainda do ponto de vista do cliente, a facilidade de implementação de uma opção de *SaaS* é importante, uma vez que os clientes podem usar o aplicativo *SaaS* de qualquer computador conectado à internet que tenha um navegador web. Isso simplifica bastante o *rollout* de serviços baseados em *SaaS* em grandes empresas com muitos tipos de *software* e *hardware* (SUKOU; GRANT, 2013).

Nesse contexto, como nenhum custo é necessário para comprar e manter / atualizar o aplicativo, e também porque a aquisição restringe-se apenas à quantidade de serviços de computação necessários, o *SaaS* reduz radicalmente os requisitos de capital para muitas empresas (FREITAS; FREITAS NETO, 2017).

Os usuários, por sua vez, têm a garantia de compatibilidade e colaboração mais fácil, porque têm a certeza de estarem utilizando o mesmo *software*. Empresas maiores são um bom exemplo para se observar nesse cenário. Grandes equipes precisam ser capazes de trabalhar de forma colaborativa para atingir as metas da sua empresa. Ao usar o *SaaS*, os usuários podem acessar o mesmo *software* a partir de vários dispositivos, no escritório ou em trânsito, o que facilita a colaboração entre as equipes (NARANG; HUDIARA; KHURMI, 2014).

Dessa forma, o *SaaS* permite que as organizações terceirizem muitos de seus aplicativos, incluindo ferramentas genéricas (por exemplo, *software* antivírus, *e-mail*, pacote de escritório) e aplicativos de negócios (por exemplo, contabilidade, gerenciamento de relacionamento com clientes − *CRM*, planejamento de recursos empresariais − *ERP*). Com base na computação em nuvem, as organizações também podem terceirizar suas infraestruturas de

TI (por exemplo, armazenamento, *backup* e computação) na forma de Infraestrutura como Serviço (*IaaS*) e plataformas de TI (por exemplo, banco de dados e *business intelligence*) na forma de plataforma como serviço (*PaaS*) (YANG *et al.*, 2015). Portanto, em uma perspectiva de negócios, o *SaaS* é um modelo emergente no setor de *software*, já que os fornecedores são responsáveis não apenas pelo desenvolvimento do aplicativo, mas por todo o conjunto de serviços para fornecer a experiência do cliente, incluindo implementação, teste, treinamento e solução de problemas, manutenção, hospedagem, upgrades e segurança (JU *et al.*, 2010).

3 METODOLOGIA

Nessa seção serão apresentados o percurso metodológico e as técnicas que nortearam este estudo. É importante reforçar que os dados utilizados se aplicam à empresa escolhida, e os resultados não podem ser generalizados, tendo essa pesquisa um caráter exploratório. Através da aplicação dos modelos de regressão logística e *XGBoost* de buscou-se responder

1. Dados de interação com os clientes, contratos existentes e informações comerciais apresentam influência no *Customer Health Score*?
2. Que ações gerenciais podem contribuir para a diminuição do *Customer Churn* a partir de análises de dados?

3.1 Aprendizado de máquina

3.1.1 Definição de machine learning (aprendizado de máquina)

Machine learning − que, em tradução livre para português seria aprendizado de máquina -, pode ser definido como fazer com que os computadores modifiquem ou adaptem suas ações para que estas se tornem mais precisas ao longo do tempo (MARSLAND, 2014). Pode-se ainda, entender a capacidade de aprendizado como sendo algo essencial para se caracterizar um comportamento inteligente. De forma geral, habilidades como memorizar, observar e explorar cenários para aprender estão relacionadas a aprendizagem (FACELI et al., 2011).

Em inúmeros autores do tema, afirmam que o aprendizado de máquina não tenta emular um comportamento inteligente, tal qual a inteligência artificial, mas sim elencar mecanismos pelos quais determinadas tarefas são aprendidas por máquinas tendo como foco a inferência indutiva e a capacidade de generalização (BATISTA, 2019).

Ainda segundo Batista (2019), os problemas de aprendizado geralmente são classificados como:

- Supervisionados: dado um volume de exemplos e suas respectivas respostas corretas, um algoritmo elencado faz uma generalição com o intuito de dar respostas corretas à maioria, senão todas, possíveis entradas futuras. Dá-se o nome de conjunto de treinamento a esse volume de dados, uma vez quem a partir deles o algoritmo é treinado

para fazer generalizações. Dentre os algoritmos clássicos do aprendizado supervisionado, pode-se elencar além da análise de regressão logística e o *boosting*, que serão abordados nesse estudo, a análise de regressão linear e o *support vector machine*.

- Não-supervisionados: esses algoritmos trabalham identificando similaridades entre os objetos de entrada, pois não são fornecidas nesse caso as respostas corretas para cada objeto de entrada. É possível, nesse caso, procurar explicar as relações entre as variáveis ou entre os objetos de entrada. Portanto, é chamado por esse nome pois não existe um vetor de respostas para supervisionar a análise que será feita nos dados de treinamento. A análise de cluster é uma ferramenta muito comum em aprendizado não-supervisionado.

O problema objeto desse estudo, que é a classificação de variáveis, é um dos mais estudados no caso de aprendizado supervisionado. Ele pode ser entendido como o ato de se receber dados de entrada e definir a qual das N classes cada um pertence tendo como base os exemplos utilizados na fase de treinamento (BATISTA, 2019).

3.1.2 Etapas da metodologia de seleção de algoritmos de machine learning

De forma geral, as etapas descritas por Marsland (2014) e resumidas por Batista (2019) descrevem como selecionar, aplicar e avaliar os algoritmos de *machine learning* a serem usados em uma determinada resolução de problema:

1. **coleta e preparação dos dados** − os dados normalmente não estão prontos para uso, principalmente se o problema é novo. Nesses casos, eles precisam ser coletados desde o início e preparados. O volume de dados é uma variável importante nessas situações, uma vez que os algoritmos de aprendizado de máquina requerem por padrão quantidades significativas de dados. Algumas exceções a esse padrão são novos algoritmos que foram desenvolvidos para minimizar essa necessidade de grande quantidade de dados.
2. **seleção de variáveis** − consiste na identificação de variáveis relevantes na solução do problema em análise. Nessa etapa, é necessário um conhecimento prévio do problema e dos dados e/ou emprego de técnicas de redução de dimensionalidade para se obter um conjunto menor de variáveis a serem consideradas no modelo.

3. **escolha do algoritmo** − uma vez definido o problema a ser tratado, deve-se selecionar um ou mais algoritmos de *machine learning* para teste.
4. **determinação dos parâmetros** − diversos algoritmos necessitam que alguns parâmetros sejam digitados manualmente ou então requerem testes prévios para identificação dos valores apropriados para esses parâmetros.
5. **treinamento** − tendo os dados e os algoritmos em mãos, o treinamento consiste no uso de recursos computacionais a fim de construir uma função F(x), para predizer as a saídas de novos dados. Essa função também pode ser definida como sendo um classificador.
6. **avaliação** − finalmente, é feita a seleção das métricas mais adequadas que serão utilizadas na avaliação do desempenho do classificador montado. Dados que foram previamente reservados e, portanto, não foram utilizados no treinamento, são cruciais nessa etapa final de checagem do classificador.

3.1 Métodos de classificação

Diversos métodos estatísticos podem ser utilizados para criar um modelo de previsão de *Customer Churn*. O desafio da previsão, não exclusivamente no campo da análise de *Churn*, encontra-se em grande parte focado na classificação dos clientes com relação à sua satisfação com a empresa contratada. Tendo em vista que se pode utilizar uma grande diversidade de metodologias com esse intuito, esta seção irá apresentar dois métodos de aprendizado de máquina que serão utilizados no problema de classificação de *Customer Churn* explorado nesse trabalho.

3.1.1 Regressão logística

Ao longo do tempo, métodos de regressão tornaram-se um componente integrante de qualquer análise de dados relacionada à descrição do relacionamento entre uma variável de resposta e uma ou mais variáveis explicativas, que normalmente serve como uma linha de base para comparar o desempenho de modelos mais sofisticados (HOSMER; LEMESHOW; STURDIVAN, 2000). No modelo de regressão logística, a sua variável resposta pode ser classificada em três categorias distintas: binomial, multinomial ou ordinal. Na primeira, a variável dependente pode assumir dois valores diferentes, não obrigatoriamente numéricos, mas usualmente codificados como 0 ou 1. A segunda refere-se a casos em que a variável dependente

pode assumir mais do que dois valores distintos, também opcionalmente numéricos. Finalmente, a última possui uma variável resposta ordinal e, portanto, a codificação para um valor numérico deve preservar a ordem das categorias. Em todas as opções, pode-se utilizar uma ou mais variáveis independentes categóricas ou contínuas. Em todos os casos, a regressão logística apresenta uma probabilidade de uma variável resposta pertencer a uma categoria específica com base em uma ou mais variáveis preditoras (BATISTA, 2019).

A regressão logística pode ser definida como uma técnica de mineração de dados que é utilizada, entre outras aplicações, para prever a probabilidade de ocorrência do *Customer Churn* (MIGUÉIS *et al.*, 2012; FARIS; AL-SHBOUL; GHATASHEH, 2014; FARQUAD; RAVI; RAJU, 2014). Ela é baseada em uma abordagem matematicamente orientada para analisar como variáveis afetam umas às outras. A previsão é feita pela formação de um conjunto de equações que conectam valores de entrada (por exemplo, que afetam o *Customer Churn*) ao campo de saída (probabilidade de ocorrer o *Customer Churn*). As equações (1), (2) e (3) fornecem as fórmulas matemáticas para um modelo de regressão logística (NISBET; ELDER; MINER, 2009):

$$p(y = 1 \mid x_1, \dots, x_n) = f(y) \tag{1}$$

$$\mathrm{f(y)} = \frac{1}{(1+ e^{-y})} \tag{2}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3}$$

Onde:

- y é a variável meta para cada indivíduo j (cliente na modelagem de *Churn*) e um rótulo de classe binária (0 ou 1);
- β_0 é uma constante;
- β_i é o peso dado para a variável específica x_i associada com cada consumidor j (j = 1, ... , m).

Os conjuntos de dados do cliente são analisados para formar as equações de regressão. Um processo de avaliação para cada cliente é executado no conjunto de dados e um valor de probabilidade de *Customer Churn* (p) é obtido (HASSOUNA *et al.*, 2015). Dessa forma, apesar de o modelo logístico não realizar classificações, ele pode ser usado como um classificador. Para isso, basta escolher um ponto de corte e atribuir uma das classes às entradas cuja probabilidade p estimada é maior do que esse ponto de corte. Analogamente, deve-se atribuir a outra classe às entradas cuja probabilidade p estimada é menor do que esse ponto de corte

(BATISTA, 2019). Portanto, um cliente pode estar em risco de *Churn* se sua probabilidade p calculada for superior a um valor predefinido como 0,5 ou 0,7, dependendo do valor de corte que se deseja adotar conforme a avaliação dos aspectos relevantes do negócio, podendo ser até menor que 0,5 em alguns casos (HASSOUNA *et al.*, 2015).

Diversos autores pontuam que é intuitivo considerar probabilidades como uma forma de quantificar as chances de um determinado evento ocorrer. No entanto, dentre diversas outras maneiras de se representarem as chances de um *Customer Churn*, existe uma amplamente utilizada, conhecida por *odds* − que, em tradução livre para português seria chance -. Conforme ilustra a equação (4), a chance de um evento é a razão entre o número esperado de vezes que um evento ocorrerá pelo número esperado de vezes que não ocorrerá (BATISTA, 2019):

$$C = \frac{p(x)}{1-p(x)} \tag{4}$$

Ainda segundo descrição de Batista (2019), é possível perceber que, para uma chance menor do que 1, a probabilidade é menor que 50%, ao passo que, para uma chance maior do que 1, a probabilidade será maior do que 50%. Ou seja, um coeficiente de chance de 1 corresponde a uma variável que tem pouca ou nenhuma influência na regressão. A equação (4), assim como as probabilidades, são limitadas inferiormente por 0. Porém, não há um limitante superior para a chance. Portanto, quando os valores de chance são próximos a zero, as probabilidades são também muito baixas e, quando são próximos de ∞, as probabilidades são igualmente muito altas, próximas a 100%, como demonstra a Tabela 2 Na , é possível observar que, ao plotarmos os dados contidos na Tabela 2 em um gráfico com o eixo da chance em escala logarítmica, a curva se comporta como um "S" característico do sigmoide, como mostra a

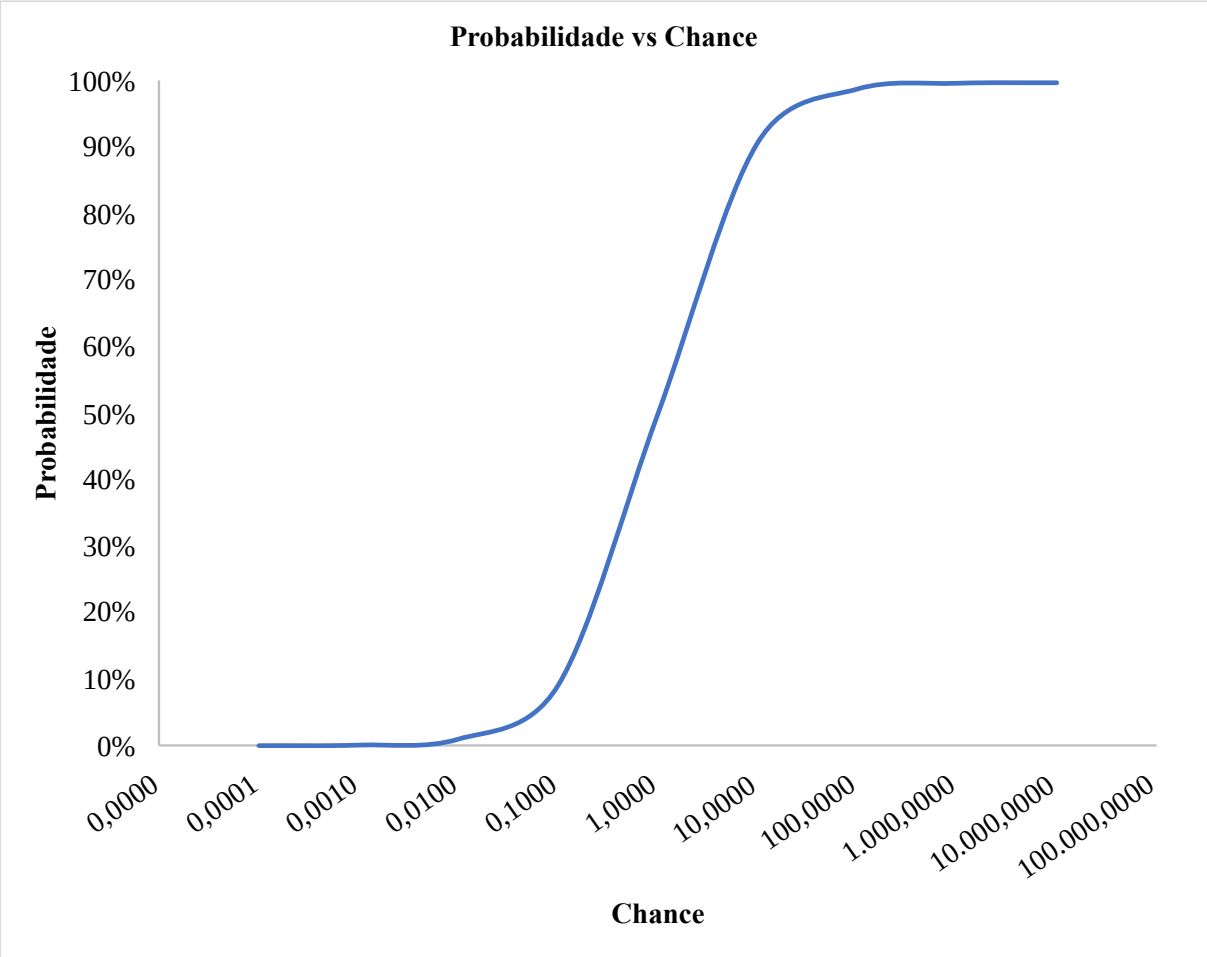
Figura 3.

Tabela 2 - Relação entre probabilidade e chance.

Chance	Probabilidade
0,0001	0,0100%
0,001	0,0999%
0,01	0,9901%
0,1	9,0909%
1	50,0000%
10	90,9091%
100	99,0099%
1.000	99,9001%
10.000	99,9900%

Fonte: elaboração própria, com base em Batista (2019).

Figura 3 - Gráfico da relação entre probabilidade e chance.



Fonte: elaboração própria, com base em Batista (2019).

Sobre o uso da análise de regressão, é preciso ser considerada com cautela, pois pode gerar resultados enganosos ao ser realizada a avaliação de causalidade e impacto. A multicolinearidade resultante de fortes correlações entre variáveis independentes também é um fator preocupante nos modelos de regressão logística (NISBET; ELDER; MINER, 2009). A existência de forte multicolinearidade leva a conclusões incorretas sobre as relações entre variáveis independentes e dependentes, pois inflaciona as variações das estimativas de parâmetros e indica a magnitude incorreta do coeficiente de regressão. Mais ainda, sob certas circunstâncias, a regressão logística pode ser usada para aproximar e representar sistemas não lineares, apesar de ser uma abordagem linear (HASSOUNA *et al.*, 2015).

Apesar de ser um método relativamente simples, a regressão logística pode produzir resultados satisfatórios e, por isso, é amplamente utilizada em diversas aplicações na indústria. Ela é afetada por variáveis redundantes, que podem ser removidas na etapa de seleção de variáveis. Outrossim, uma quantidade grande de categorias pode dificultar a interpretação de

atributos independentes nominais, uma vez que será estimado um coeficiente para cada um, e todos são usualmente interpretados (BATISTA, 2019).

3.1.2 XGBOOST

XGBoost é um método de aprendizado de máquina que tem sido utilizado por muitos cientistas de dados para resolução de problemas de aprendizado supervisionado. Ele é muito popular por ser aplicável tanto em problemas de classificação quanto de regressão (WANG *et. al.*, 2017). Seu nome é a abreviação de "*Extreme Gradient Boosting*", sendo que o termo "*Gradient Boosting*", referente ao modelo no qual foi baseado, foi proposto originalmente por Friedman (2001) em seu artigo intitulado: "*Greedy Function Approximation: A Gradient Boosting Machine*".

Ele utiliza uma estratégia de aumento de árvores, que é extremamente eficaz e, além disso, geralmente utilizada para aprendizado de máquina (GOHIYA; LOHIYA; PATIDAR, 2018). O fator mais importante por trás do sucesso do XGBoost é seu uso escalável em diversos cenários. Ele é cerca de dez vezes mais rápido do que outras soluções populares existentes, se processado em uma única máquina, e escala para bilhões de exemplos em configurações distribuídas ou de memória limitada tipo *docker*. A escalabilidade do XGBoost se deve a vários sistemas importantes e otimizações algorítmicas (CHEN; GUESTRIN, 2016).

3.2 Avaliação dos modelos preditivos

A questão crítica no uso de diferentes métodos de modelagem de *Customer Churn* se refere à avaliação eficiente do desempenho desses métodos e ao *benchmarking* e comparação do desempenho relativo entre modelos concorrentes (HASSOUNA *et al.*, 2015).

Esta seção discute os dois métodos que foram utilizados neste estudo à luz desses problemas.

3.2.1 Matriz de Confusão

Uma Matriz de Confusão contém informações sobre classificações reais e previstas feitas por um sistema de classificação (AKAY, 2009). Segundo Souza (2019), a pesquisadora mostra as seguintes frequências:

- **verdadeiro positivo** (*true positive - TP*) − ocorre quando, no conjunto real, a classe que estamos buscando foi prevista corretamente. Por exemplo, quando houve um *Customer Churn* e o modelo previu corretamente que haveria.
- **falso positivo** (*false positive - FP*) − ocorre quando no conjunto real, a classe que estamos buscando foi prevista incorretamente. Por exemplo, não houve um *Customer Chrun*, mas o modelo previu que haveria. É também conhecido como taxa de "alarme falso" (em inglês, *false alarm*, e corresponde ao Erro tipo I).
- **falso verdadeiro** (*true negative - TN*) − ocorre quando, no conjunto real, a classe que não estamos buscando foi prevista corretamente. Por exemplo, não houve um *Customer Chrun*, e o modelo previu corretamente que não haveria.
- **falso negativo** (*false negative - FN*) − ocorre quando, no conjunto real, a classe que não estamos buscando foi prevista incorretamente. Por exemplo, quando houve um *Customer Chrun* e o modelo previu incorretamente que não haveria. É também conhecido como taxa de "não notado" (em inglês, *miss rate*, e corresponde ao Erro tipo II).

O Quadro 1 mostra como são calculadas essas frequências:

Quadro 1 - Cálculos de indicadores da Matriz de Confusão para duas classes.

Classe	**predita C+**	**predita C-**	**Taxa de Erro da Classe**	**Taxa de Erro Total**
Verdadeira C+	*Tp*	*Fn*	$\frac{Fn}{Tp + Fn}$	$\frac{Fp + Fn}{n}$
Verdadeira C-	*Fp*	*Tn*	$\frac{Fp}{Tf + Tn}$	

Fonte: elaboração própria, com base em James *et al.* (2013).

Acurácia de classificação (5), sensibilidade (6), especificidade (7), valor preditivo positivo (8) e valor preditivo negativo (9) podem ser definidos usando os elementos da Matriz de Confusão como (AKAY, 2009):

$$\text{Acurácia de classificação: } (\%) = \frac{Tp+Tn}{Tp+Fp+Fn+Tn} \quad (5)$$

$$\text{Sensibilidade } (\%) = \frac{Tp}{Tp+Fn} X\ 100 \quad (6)$$

$$\text{Especificidade } (\%) = \frac{Tn}{Fp+Tn} X\ 100 \quad (7)$$

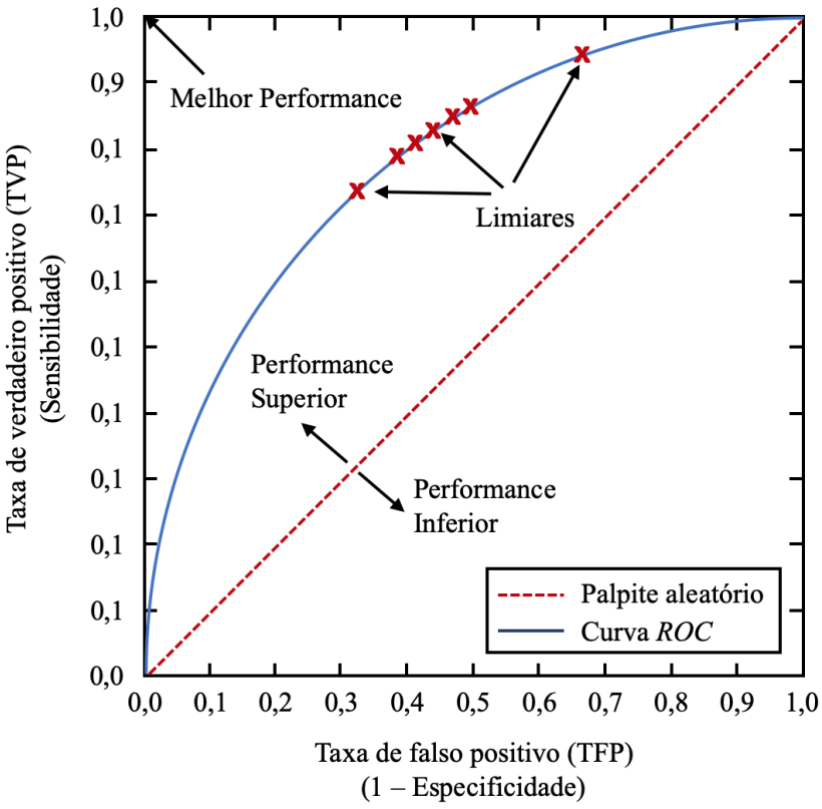
$$\text{Valor Preditivo Positivo} = \frac{Tp}{Tp+Fp}\ X\ 100 \quad (8)$$

$$\text{Valor Preditivo Negativo} = \frac{Tn}{Fn+Tn}\ X\ 100 \quad (9)$$

3.2.2 Curva ROC (Receiver Operating Characteristic Curve)

A curva ROC (*Receiver Operating Characteristic*, – que, em tradução livre para português seria a Características de Operação do Receptor -) é uma representação das relações entre a taxa de verdadeiros positivos (ou benefícios) e a taxa de falsos positivos (ou custos), desenhada nos eixos x e y em uma escala linear (KARAHOCA; KARAHOCA; AYDIN, 2007). A ROC (ou COR, em português) representa as relações entre a taxa de *Customer Churn* prevista corretamente como *Customer Churn* e a taxa de não-*churners* prevista incorretamente como *Customer Churn*. Em outras palavras, a ROC fornece implicações relativas entre benefícios e custos. A curva ROC consiste de pontos correspondentes aos resultados da previsão. A Figura 4 apresenta um exemplo de uma curva ROC (HASSOUNA *et al.*, 2015):

Figura 4 - Exemplo de curva ROC.



Fonte: elaboração própria, adaptado de Hassouna *et al.* (2015).

Ainda segundo *Hassouna et al.* (2015), o melhor modelo de desempenho possível é quando a curva ROC passa por ou perto de (x, y) = (0, 1). Nesse caso, ambos − a sensibilidade e a especificidade do modelo − serão 100%, em outras palavras, não haveria falsos negativos nem falsos positivos, respectivamente. Determinados modelos, como por exemplo a regressão logística, que também foi utilizada neste estudo, resultam em uma pontuação, não se restringindo a produzir decisões de classe binária (como se há *Customer Churn* ou não). Para produzir um classificador binário, nesse caso, são usados limites, de forma que, se o resultado do classificador for maior que um limite, a classe de classificação será um *Churn*, caso contrário, não será um *Churn*. Na Figura 4, é possível ver a curva ROC para um preditor aleatório representado pela linha diagonal que divide esse espaço ROC em duas partes. As curvas ROC que passam perto dessa linha correspondem a classificadores de adivinhação aleatórios (como, por exemplo, a classificação a partir do lance de uma moeda para o alto).

De maneira geral, se os modelos com curvas *ROC* passarem pela parte superior esquerda da curva *ROC*, é porque apresentam melhor desempenho. A área sob a curva *ROC* (chamada *AUC*, do inglês *Area Under the Curve* − que, em tradução livre para português seria a Área Abaixo da Curva -) é usada como uma métrica de desempenho (KARAHOCA; KARAHOCA; AYDIN, 2007). O valor da *AUC* pode assumir valores entre 0 e 1. Normalmente, os modelos que apresentam melhor desempenho possuem uma *AUC* maior. Além disso, é fácil ver que modelos com valor maior do que 0,5 de *AUC* apresentam desempenho melhor do que modelos aleatórios, porque a área sob a linha diagonal do ROC é 0,5 (HASSOUNA *et al.*, 2015).

3.3 Coleta de dados

3.3.1 Primeira fase da coleta de dados

Inicialmente, foram utilizados diversos dados cadastrais, de venda e indicadores associados aos clientes, à sua experiência e às interações com a companhia durante sua jornada. Esses dados foram coletados em uma única data no mês de abril de 2019. O número total de empresas clientes consideradas inicialmente no estudo foi de cerca de 40.635. Os dados das empresas clientes utilizados na primeira etapa foram os que compunham o *Health Score* original, selecionados pela área de *Customer Experience* da empresa. Além desses dados, foram

utilizados os respectivos *scores* individuais de cada indicador e o código de cada cliente. Os indicadores são:

- quantidade total de *tickets*;
- *score* da quantidade total de *tickets*;
- quantidade de *tickets* de Ouvidoria;
- *score* da quantidade dos *tickets* de Ouvidoria;
- quantidade de *tickets* com SLA atrasado;
- *score* da quantidade de *tickets* com SLA atrasado;
- NPS;
- *score* do NPS;
- quantidade de solicitações de cancelamento total;
- *score* da quantidade de solicitações de cancelamento total;
- quantidade de solicitações de cancelamento parcial;
- *score* da quantidade de solicitações de cancelamento parcial;
- situação do cliente;
- *score* da situação do cliente;
- quantidade de títulos em atraso;
- *score* da quantidade de títulos em atraso.

3.3.2 Segunda fase da coleta de dados

Na segunda fase da coleta de dados, além de informações das empresas clientes, foram agregados dados coletados de outras fontes internas da empresa. São eles:

- tempo como cliente;
- localização do cliente;
- unidade de vendas (própria, franquia ou canal complementar);
- quantidade de produtos/serviços diferentes contratados;
- tipos de produtos contratados;
- informações de faturamento;
- macrossegmento e subsegmento dos clientes;
- assuntos dos *tickets* de atendimento;
- prioridade do *ticket*;
- dados de bonificação concedida ao cliente;
- valores de reajuste de contratos.

O período de tempo dos dados coletados nessa segunda etapa foi de setembro de 2017 a março de 2019.

3.4 Ferramentas de manipulação de dados

Na manipulação dos dados, foram utilizadas duas ferramentas distintas: a linguagem Python e a Radiant. Da linguagem Python, foram usadas as bibliotecas pandas, numpy, scipy, matpltlib e sklearn. A ferramenta Radiant é uma interface baseada em navegador *web*, independente de plataforma e de código aberto, para análise de negócios na linguagem R.

3.5 Análise preliminar, tratamento e limpeza de dados

A análise inicial verificou quais desses dados poderiam ser utilizados para se prever o *Customer Churn*. Primeiramente, eliminaram-se as colunas de *Score* do conjunto de dados, pois essas apresentavam os cálculos correntes dos pesos e, portanto, seriam eventualmente substituídas por novas. Uma amostra desses dados originais pode ser verificada na Tabela 3.

Tabela 3 - Amostra dos dados originais dos indicadores que compunham o *Health Score* adotado pela empresa.[3]

Código do cliente	Nome do Cliente	*Healthscore*	Quantidade total de tickets	*Score* do total de tickets	Quantidade de tickets de ouvidoria	*Score* da quantidade de tickets de ouvidoria	Quantidade de tickets com SLA atrasado	Score dos tickets com SLA atrasado	Quantidade de solicitações de cancelamento total
99000	Empresa_1	93,3	0	12,5	0	20	0	20	0
99014	Empresa_2	98,5	1	12,5	0	20	1	20	0
99025	Empresa_3	93,3	2	12,5	0	20	2	20	0
99095	Empresa_4	98,5	0	12,5	0	20	0	20	0
99699	Empresa_5	84,8	1	12,5	0	20	0	20	2
99910	Empresa_6	98,5	0	12,5	0	20	0	20	0
99920	Empresa_7	98,5	0	12,5	0	20	0	20	0
99955	Empresa_8	98,5	0	12,5	0	20	0	20	0
99958	Empresa_9	97,3	18	11,3	0	20	3	20	0
99959	Empresa_10	98,5	0	12,5	0	20	0	20	0
99960	Empresa_11	98,5	0	12,5	0	20	0	20	0
99969	Empresa_12	98,5	0	12,5	0	20	0	20	0
99971	Empresa_13	98,5	0	12,5	0	20	0	20	0
99973	Empresa_14	98,5	2	12,5	0	20	2	20	0
99980	Empresa_15	98,5	2	12,5	0	20	1	20	0
99987	Empresa_16	93,3	1	12,5	0	20	0	20	0
99990	Empresa_17	98,5	0	12,5	0	20	0	20	0
99991	Empresa_18	98,5	0	12,5	0	20	0	20	0

(Continua)

Código do cliente	*Score* da quantidade de solicitações de cancelamento total	Quantidade de solicitações de cancelamento parcial	*Score* da quantidade de solicitações de cancelamento parcial	Situação do cliente	*Score* da situação do cliente	Quantidade de títulos em atraso	*Score* da quantidade de títulos em atraso	NPS	*Score* NPS
99000	7,5	0	5	Nenhum	12,5	1	12,3	-	3,5
99014	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99025	7,5	0	5	Nenhum	12,5	1	12,3	-	3,5
99095	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99699	3	2	2	Sol. Canc. Parc.	6,3	0	17,5	-	3,5
99910	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99920	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99955	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99958	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99959	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99960	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99969	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99971	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99973	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99980	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99987	7,5	0	5	Nenhum	12,5	1	12,3	-	3,5
99990	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5
99991	7,5	0	5	Nenhum	12,5	0	17,5	-	3,5

Fonte: elaboração própria, a partir de dados compartilhados pela empresa estudada.

Em seguida, foi feita uma análise nas dimensões remanescentes. Nessa análise, constatou-se que, dentre os 40.635 clientes cujos dados haviam sido levantados, o valor do NPS estava preenchido em apenas 6.768, ou seja, em cerca de 16,7% dos clientes. Como esse percentual é muito baixo, foi necessário realizar uma transformação no campo do NPS e o resultado foi gravado em uma coluna nova chamada "NPS_CAT". A Lógica utilizada na transformação foi:

- NPS_CAT:
    - **se NPS >= 9**: P (Promotor);
    - **se NPS <=8**: ND (Neutro/Detrator);
    - **se não estiver preenchido**: NR (Não Respondeu).

Outros dois campos foram criados a fim de se realizarem transformações: "Parcial" e "Total", que são transformações dos campos "Quantidade de solicitações de cancelamento parcial" e "Quantidade de Solicitações de Cancelamento Total" respectivamente. Essas transformações foram feitas seguindo as seguintes lógicas:

- PARCIAL:
    - **se quantidade de solicitações de cancelamento parcial > 0: 1;**
    - **caso contrário: 0.**

- TOTAL:
    - **se quantidade de solicitações de cancelamento total > 0: 1;**
    - **caso contrário: 0**

Criar os campos "Parcial" e "Total" teve o objetivo de utilizar essas novas variáveis como variáveis dependentes ou, em outras palavras, como variáveis a serem previstas por um modelo de regressão logística. Uma amostra dos dados resultantes desses ajustes de 15 empresas pode ser vista com maiores detalhes na Tabela 4. Os trechos de código que correspondem a esse processo de ANÁLISE PRELIMINAR se encontram no APENDICE – B.

Tabela 4 - Amostra dos dados dos indicadores após tratamentos e criação de novas dimensões transformadas.[4]

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
99000	0	0	0	0	0	1	-	NR	0	0
99014	1	0	1	0	0	0	-	NR	0	0
99025	2	0	2	0	0	1	-	NR	0	0
99095	0	0	0	0	0	0	-	NR	0	0
99699	1	0	0	2	2	0	-	NR	1	1
99910	0	0	0	0	0	0	-	NR	0	0
99920	0	0	0	0	0	0	-	NR	0	0
99955	0	0	0	0	0	0	-	NR	0	0
99958	18	0	3	0	0	0	-	NR	0	0
99959	0	0	0	0	0	0	-	NR	0	0
99960	0	0	0	0	0	0	-	NR	0	0
99969	0	0	0	0	0	0	-	NR	0	0
99971	0	0	0	0	0	0	-	NR	0	0
99973	2	0	2	0	0	0	-	NR	0	0
99991	0	0	0	0	0	0	-	NR	0	0

Fonte: elaboração própria.

[4] **Nota:** C1 = Código do cliente, C2 = Quantidade total de *tickets*, C3 = Quantidade de *tickets* de ouvidoria, C4 = Quantidade de *tickets* com SLA atrasado, C5 = Quantidade de solicitações de cancelamento total, C6 = Quantidade de solicitações de cancelamento parcial, C7 = Quantidade de títulos em atraso, C8 = NPS, C9 = NPS categórico, C10 = *Churn* parcial, C11 = *Churn* total.

3.6 Criação de variáveis candidatas / *Feature Engineering*

Em seguida, foram projetados novas *features* − que, em tradução livre para português seriam as variáveis -, para criar pontuações que identificam clientes que possivelmente irão se tornar um *Churn*. Uma *feature* é uma propriedade ou característica mensurável individual de um fenômeno que está sendo observado (BISHOP, 2006). Foram projetados quatro grupos de variáveis por cálculos e transformações realizadas nos dados que serão descritos em maiores detalhes mais à frente.

Para tanto, foi utilizada uma extensa base contendo informações sobre vendas, contratos, produtos, unidades de venda, macrossegmento e subsegmento dos clientes, *tickets* de atendimento, dentre outras. O período dos dados coletados vai de setembro de 2017 a março de 2019.

É possível categorizar as variáveis em três grandes grupos: "Variáveis de quantidade", “Variáveis de frequência” e " Variáveis de dias desde/para". O Quadro 2 apresenta um resumo de algumas das lógicas que foram utilizadas para se calcularem as novas variáveis.

Quadro 2 - Lógicas de criação e tipos de variáveis geradas a partir dos dados disponíveis.

Variáveis de quantidade	Média Máximo Mediana Total Atual/Média Atual/Máximo Atual/Mediana Atual/Total	Quantidade por/ nesse (a)	Produto Unidade Macrossegmento Subsegmento Região	No decorrer do(s) último(s)	1 Mês 2 Meses 3 Meses 6 Meses 7 Meses 8 Meses 12 Meses
Variáveis de frequência	Número de *tickets* com esse(a)	Assunto Tipo Prioridade	No decorrer do(s) último(s)	1 Mês 2 Meses 3 Meses 6 Meses 7 Meses	
Dias para	Carência	Iniciar Terminar			

Dias desde	Início do Contrato

Fonte: elaboração própria.

No caso das "Variáveis de Quantidade", pode-se utilizar a segunda coluna para realizar um agrupamento (*groupby*) e a terceira coluna como uma janela de tempo para calcular cada tipo de medida de "Montante" da primeira coluna (média, máximo, mediana, total etc.). Em cada uma dessas colunas determinou-se um número de medidas, respectivamente, e assim suas combinações forneceram quantidades diferentes de variáveis.

Ao todo foram criadas 396 variáveis candidatas. Uma lista completa com todas elas e a indicação de quais foram utilizadas na etapa de seleção de variáveis encontra-se no APÊNDICE - A. Alguns exemplos de trechos dos nomes das principais *variáveis*, o que representam e suas quantidades totais são:

- **variáveis iniciadas em "*bow*"**: soma do total de *tickets* de chamado do cliente que possuem uma determinada palavra no assunto no período dos últimos 7 meses. Por exemplo, bow_esocial_last_7_months representa o número total de *tickets* abertos que contêm no assunto a palavra "esocial". Foram criadas 100 variáveis com essa característica;
- **variáveis iniciadas em "ticket_grp_"**: contagem do total de tickets de chamado do cliente que possuem um conteúdo direcionado a um determinado departamento no período dos últimos 7 meses. Por exemplo, ticket_grp_CST-Faturamento_last_7_months representa o número total de tickets abertos que foram direcionados para o time de faturamento da empresa nos últimos 7 meses. Foram criadas 101 variáveis com essa característica;
- **variáveis iniciadas em "*churn_rate_for_same_*"**: taxa de *Customer Churn* calculada que compartilham um determinada característica, essa podendo ser: a unidade que fez a venda para o cliente; macro-segmento, segmento ou sub-segmento do cliente; região do cliente; produto. Todas essas taxas foram calculadas para períodos anteriores específicos de 1, 2, 3 ou 6 meses. Foram criadas 24 variáveis com essa característica;
- **variáveis iniciadas em "*total_contracts_with_same _*"**: total de contratos que compartilhem uma determinada característica com o contrato observado, essa podendo ser: a unidade que fez a venda para o cliente; macro-segmento, segmento ou sub-segmento do cliente; região do cliente; produto. Todas essas taxas foram calculadas para períodos anteriores específicos de 1, 2, 3 ou 6 meses. Foram criadas 24 variáveis com essa característica;

- **variáveis iniciadas em "*tickets_of_priority_*"**: soma do total de *tickets* de chamado do cliente que possuem uma determinada prioridade de atendimento, essa podendo ser: baixa (*low*); normal (*normal*); alta (*high*); urgente (*urgent*). Todas essas somas foram calculadas para períodos anteriores específicos de 1, 2, 3 ou 6 meses. Foram criadas 24 variáveis com essa característica;

3.7 Processo de seleção de variáveis para inclusão nos modelos de regressão logística e *XGBoost*

Em preparação para uso no modelo supervisionado, normalizaram-se as variáveis consideradas e selecionaram-se as mais relevantes, diminuindo efeitos de escala. Em seguida, uma parte dos dados foi separada em um conjunto de treinamento / teste e uma outra parte em um conjunto "fora de tempo", ou *Out-of-Time* (*OOT*). Usando o conjunto de treinamento / teste, aplicou-se um teste KS (Kolmogorov-Smirnov) univariado, além de uma medida de taxa de detecção de *Churn* -, em inglês, *Churn Detection Rate* (*CDR*) -. A lista de variáveis gerada foi organizada em ordem decrescente de ranking, formado pela média dos valores de KS e *CDR*. Essa lista permitiu reduzir expressivamente o número total de variáveis candidatas, pois considerou-se apenas as variáveis com ranking maior que a mediana, ou em outra palavras, foram eliminadas cerca de metade das variáveis nessa etapa. Em seguida, aplicou-se um *wrapper* de estimador de regressão logística que utiliza a classificação de variáveis com a eliminação de variáveis recursiva sendo eliminadas assim praticamente metade dessas variáveis remanescentes. Esse *wrapper* foi aplicado uma segunda vez para reduzir ainda mais o número de variáveis sendo que o número de variáveis foi reduzido uma vez mais para cerca da metade novamente. Dessa forma, após a aplicação dessas metodologias de redução da dimensionalidade, as variáveis iniciais foram reduzidas em cerca de 8 vezes ($(\frac{1}{2})^3$).

Juntamente com os *wrappers* foi aplicada uma técnica de validação estratificada de dobra cruzada (*stratified cross-fold validation)* para selecionar as melhores variáveis.

Ao final dessa etapa de seleção, foram mantidas apenas as 20 variáveis mais impactantes no modelo, como mostra a Tabela 5 na próxima página.

Tabela 5 -Variáveis selecionadas a partir da aplicação de filtros e *wrappers*.

Nome da Variável
Número de *tickets* de atendimento com conteúdo direcionado a área de "Contratos de vendas" nos últimos 7 meses
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 6 meses
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 3 meses
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 2 meses
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 1 meses
Taxa de *Churn* para o mesmo produto do cliente nos últimos 6 meses
Taxa de *Churn* para o mesmo produto do cliente no último mês
Quantidade total de contratos com o mesmo produto do cliente nos últimos 2 messes
Valor total de novas vendas nos últimos 6 meses
Taxa de *Churn* para o mesmo produto do cliente nos últimos 3 meses
Quantidade total de contratos com o mesmo produto do cliente no último mês
Quantidade total de contratos com o mesmo produto do cliente nos últimos 6 meses
Quantidade total de contratos com o mesmo macrosegmento do cliente no último mês
Quantidade total de contratos com o mesmo macrosegmento do cliente nos últimos 3 meses
Quantidade total de contratos com o mesmo macrosegmento do cliente nos últimos 2 meses
Valor médio de novas vendas nos últimos 12 meses
Taxa de *Churn* para o mesmo macrosegmento do cliente nos últimos 6 meses
Taxa de *Churn* para o mesmo macrosegmento do cliente nos últimos 3 meses
Taxa de *Churn* para o mesmo macrosegmento do cliente nos últimos 2 meses
Taxa de *Churn* para o mesmo macrosegmento do cliente no último mês

Fonte: elaboração própria.

Os trechos de código que correspondem a esse processo de seleção de variáveis se encontram no APÊNDICE – C.

Posteriormente, foi aplicado um outro algoritmo para a redução do número de variáveis, compreendido no modelo *XGBoost*. O intuito foi de se obter resultados para comparação com o modelo descrito até o momento nessa seção. Também na aplicação do modelo *XGBoost* foram selecionadas apenas 20 variáveis mais relevantes para efeito de comparação, como mostra a Tabela 6 na próxima página. Os trechos de código que correspondem a esse segundo processo de seleção de variáveis se encontram no APENDICE – D.

Tabela 6 - Lista das 20 variáveis que têm maior importância segundo a análise realizada com o XGBoost.

Nome da Variável	Importância
Taxa de *Churn* para o mesmo produto consumido pelo cliente no último mês	13
Taxa de *Churn* para o mesmo produto consumido pelo cliente nos últimos 6 meses	12
Tempo como cliente da empresa (dias)	12
Número total de contrato com o mesmo segmento do cliente nos últimos 6 meses	10
Taxa de *Churn* para o mesmo produto consumido pelo cliente nos últimos 2 meses	8
Taxa de *Churn* para o mesmo subsegmento do cliente nos últimos 6 meses	8
Valor médio de novas vendas nos últimos 6 meses	7
Número de *tickets* de atendimento com a palavra "Contato" no assunto nos últimos 7 meses	7
Número de *tickets* de atendimento com conteúdo direcionado a área de "Contratos de vendas" nos últimos 7 meses	7
Número de *tickets* de atendimento com a prioridade “Urgente" nos últimos 7 meses	6
Taxa de *Churn* para o mesmo macrosegmento do cliente nos últimos 6 meses	6
Taxa de *Churn* para o mesmo segmento do cliente nos últimos 3 meses	6
Taxa de *Churn* para o mesmo produto do cliente nos últimos 3 meses	6
Número de *tickets* de atendimento com a palavra "Cancelamento" no assunto nos últimos 7 meses	5
Número de *tickets* de atendimento com conteúdo direcionado a "Outros" nos últimos 7 meses	5
Quantidade total de clientes que deram *Churn* nos últimos 6 meses	5
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 3 meses	5
Quantidade total de contratos com o mesmo segmento do cliente no último mês	5
Quantidade total de contratos com o mesmo produto do cliente no último mês	4
Dias até o início da carência do cliente	4

Fonte: elaboração própria, a partir da linguagem Python.

3.8 Escolha do algoritmo de predição

Com o número de variáveis projetadas diminuído, foram aplicados os modelos de regressão logística e *XGBoost* por se tratarem de modelos bem conhecidos e robustos de predição (NIE *et al.*, 2011, CHEN; GUESTRIN, 2016).

É importante lembrar que em ambos os modelos aplicados a variável dependente foi o *churn*, ou deserção do cliente.

3.9 Checagem do desempenho dos algoritmos

A checagem do desempenho de previsão de *Churn* dos modelos utilizados foi realizada através das análises matriz de confusão e da curva ROC.

4 RESULTADOS E DISCUSSÃO

Nas próximas seções, serão descritas as etapas de aplicação da metodologia, algumas das decisões que foram tomadas e os resultados obtidos ao longo do desenvolvimento deste trabalho.

4.1 Primeira etapa de análise

A primeira etapa consistiu em analisar as métricas e o cálculo do *Customer Health Score* praticados pela empresa originalmente. O cálculo dos *scores* dos indicadores e uma amostra dos dados que compõem o *Health Score* original estão descritos na Tabela 7 na próxima página.

Tabela 7 - Cálculo dos *scores* dos indicadores que compõem o *Health Score* original utilizado pela companhia.[5]

Peso	20,0%	20,0%	17,5%	12,5%	12,5%	12,5%	5,0%
Valor / Indicador	C1	C2	C3	C4	C5	C6	C7
100	0	<=1	0	0	<=3	Outro Status	Nota do NPS da última pesquisa respondida pelo cliente multiplicada por 10
90	1	2	Não há	Não há	4 ou 5	Não há	
80	2	3	Não há	Não há	6 ou 7	Não há	
70	3	4	1	1	8 ou 9	Não há	
60	4	5	Não há	Não há	10 ou 11	Não há	
50	5	6	Não há	Não há	12 ou 13	Não há	
40	6	7	Não há	2	14 ou 15	Não há	
30	7	8	>=2	3	16 ou 17	Não há	
20	8	9	Não há	Não há	18 ou 19	Não há	
10	9	10	Não há	4	20 ou 21	Não há	
0	>=10	>=11	Não há	>=5	>=22	Solicitação de Cancelamento	

Fonte: elaboração própria, a partir de documentos compartilhados pela empresa estudada.

[5] **Nota:** C1 = Tickets de Ouvidoria, C2 = Tickets com SLA atrasado, C3 = Títulos de Pagamento em Atraso, C4 = Solicitações de Cancelamento, C5 = Total de Tickets, C6 = Situação do Cliente, C7 = NPS.

A partir desses *scores* parciais é calculado o *Health Score* de cada cliente, sendo ele o resultado da somatória do *rating* de cada uma dessas métrica que é igual ao valor do *score* dividido por 100 e multiplicado por seu respectivo peso, como mostram as equações (10) e (11).

$$\mathrm{R_i} = \frac{\mathrm{SCM_i \ x \ PSM_i}}{100} \tag{10}$$

$$HS = \sum \mathrm{R_i} \tag{11}$$

Onde:

- $\mathrm{R_i}$ é o *Rating* de cada métrica.
- $\mathrm{SCM_i}$ é o *score* de cada métrica.
- $\mathrm{PSM_i}$ é o peso de cada métrica.
- *HS* é o *Health Score* de cada cliente.

Os pesos utilizados no cálculo do *Health Score* original foram atribuídos pela equipe de *Customer Experience* da companhia baseada em entrevistas realizadas com executivos da própria empresa. Dessa forma, não possuíam rigor estatístico, apresentando poder de previsão de *Churn* muito aquém do desejado.

Notou-se também que a quantidade de clientes com pedido de cancelamento no momento exato em que foram analisados os indicadores originais é muito pequena se comparada com o universo total de clientes na base. Isso mostra que a base original de dados considerada é insuficiente para se analisar o *Customer Churn*. A Tabela 8, na próxima página, mostra que apenas, no período considerado originalmente, que é de 1 mês, cerca de 0,18% dos clientes estavam em situação de pedido de *Churn* Parcial e 0,34% dos clientes em situação de pedido de *Churn* Total.

Tabela 8 - Resumo da quantidade e percentual de clientes da amostra que preencheram o NPS e que têm pedidos de Churn Parcial e Churn Total.[6]

Descrição	NPS Preenchido	*Churn* Parcial	*Churn* Total
Sim	6.768	72	140
# Total Clientes	40.635	40.635	40.635
% do Total	**16,66%**	**0,18%**	**0,34%**

Fonte: elaboração própria

Essa informação obtida na análise preliminar, que demonstra percentuais baixos de ambos os indicadores de *Churn* em relação ao montante total de clientes, por si só já denota que o volume da amostra era insuficiente para se chegar a um modelo satisfatório no cenário original considerado pela empresa. Por essa razão, posteriormente foi realizada uma análise que abrangia um período maior de meses, descrita na seção 4.1.2.

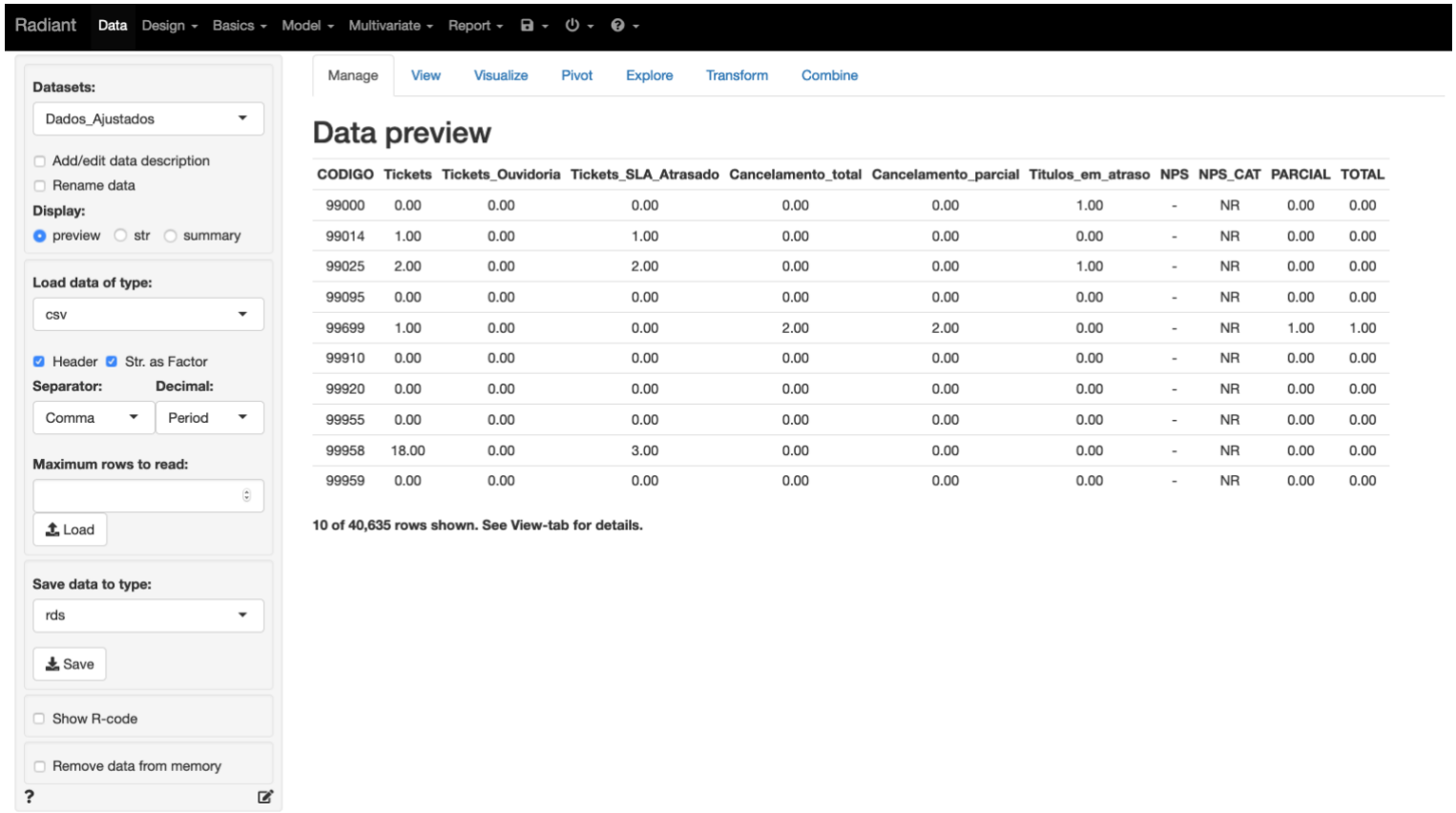
Após realizados os ajustes nos dados descritos na seção 3.5, foram carregados na ferramenta *Radiant*, como mostra a Figura 5, e foi gerado um modelo de regressão logística a partir deles, como mostra a Figura 6. Para gerar esse modelo inicial foram utilizadas as variáveis: Tickets; Tickets_Ouvidoria; Tickets_SLA_Atrasado; Titulos_em_atraso; NPS_CAT. A variável resposta adotada foi a TOTAL, que representa os casos de *Churn* Total. Foi adotada essa variável, pois além de representar os casos com maior potencial de prejuízo para a companhia, havia quase o dobro de casos de *Churn* Total em relação ao de *Churn* Parcial, o que poderia melhorar a acurácia do modelo final.

Após essa primeira rodada, pode-se observar que a variável Tickets_Ouvidoria apresentou um *p.value* muito elevado, próximo a 1, indicando baixa relevância dessa variável com o modelo, como pode ser observado na Figura 7. Além disso, outras três variáveis utilizadas também apresentaram um *p. value* elevado. Esses resultados preliminares se mostraram insatisfatórios, pois de um total de cinco variáveis selecionadas, apenas duas se mostraram significativas. Ao eliminar as três variáveis não significativas, verifica-se que houve um resultado contraditório, já que a quantidade total de tickets contribuiu para o aumento do

6 **Nota:** O NPS é respondido de forma opcional pelos clientes. A pesquisa é feita em uma frequência não determinada e por esses motivos há poucos dados preenchidos nesse campo, muitos dos quais já estão defasados.

Churn, enquanto a quantidade de tickets com *SLA* atrasado contribuiu com a redução do *Churn*, como pode ser visto na Figura 8.

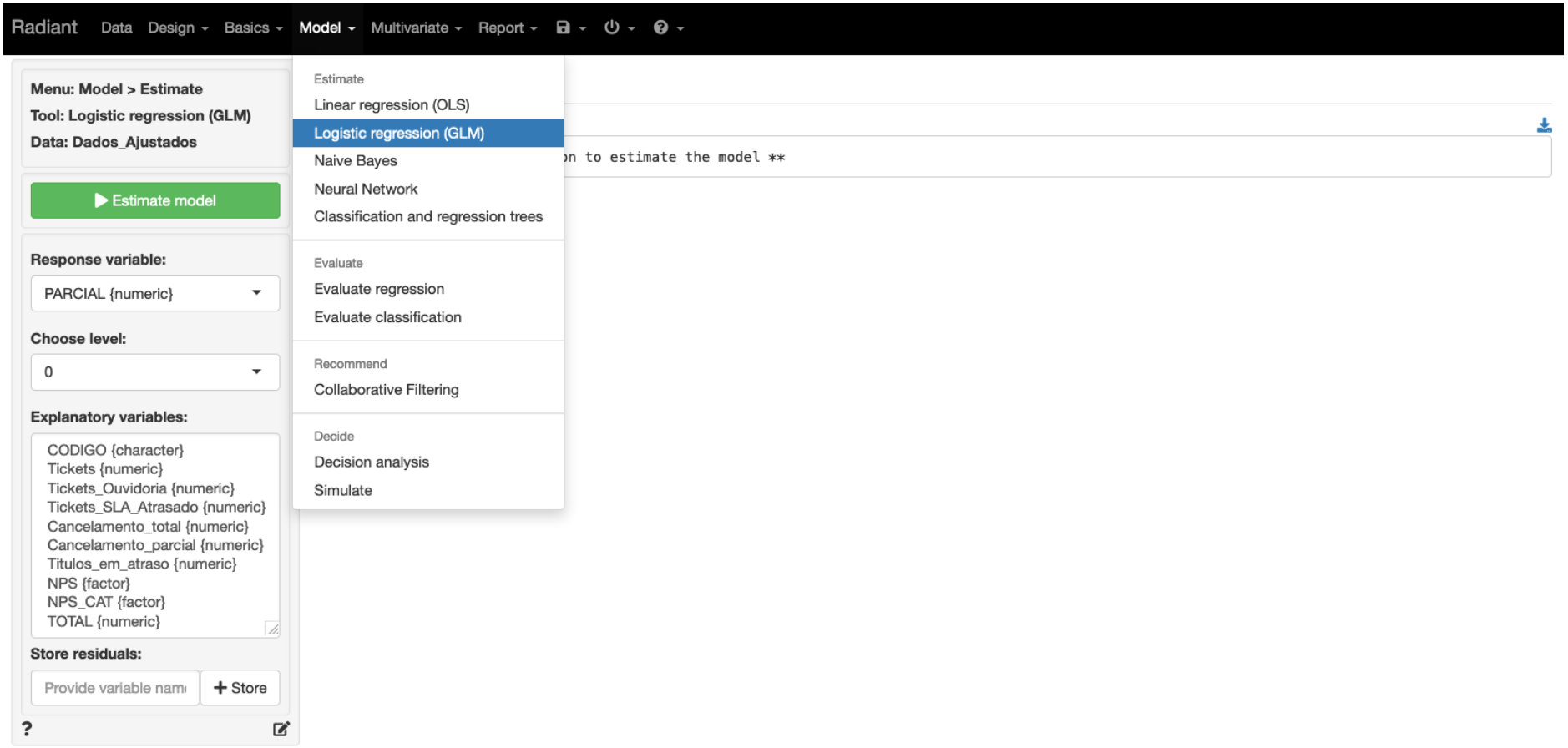
Figura 5 - Tela da carga dos dados originais na ferramenta *Radiant* para execução da regressão logística.



CODIGO	Tickets	Tickets_Ouvidoria	Tickets_SLA_Atrasado	Cancelamento_total	Cancelamento_parcial	Titulos_em_atraso	NPS	NPS_CAT	PARCIAL	TOTAL
99000	0.00	0.00	0.00	0.00	0.00	1.00	-	NR	0.00	0.00
99014	1.00	0.00	1.00	0.00	0.00	0.00	-	NR	0.00	0.00
99025	2.00	0.00	2.00	0.00	0.00	1.00	-	NR	0.00	0.00
99095	0.00	0.00	0.00	0.00	0.00	0.00	-	NR	0.00	0.00
99699	1.00	0.00	0.00	2.00	2.00	0.00	-	NR	1.00	1.00
99910	0.00	0.00	0.00	0.00	0.00	0.00	-	NR	0.00	0.00
99920	0.00	0.00	0.00	0.00	0.00	0.00	-	NR	0.00	0.00
99955	0.00	0.00	0.00	0.00	0.00	0.00	-	NR	0.00	0.00
99958	18.00	0.00	3.00	0.00	0.00	0.00	-	NR	0.00	0.00
99959	0.00	0.00	0.00	0.00	0.00	0.00	-	NR	0.00	0.00

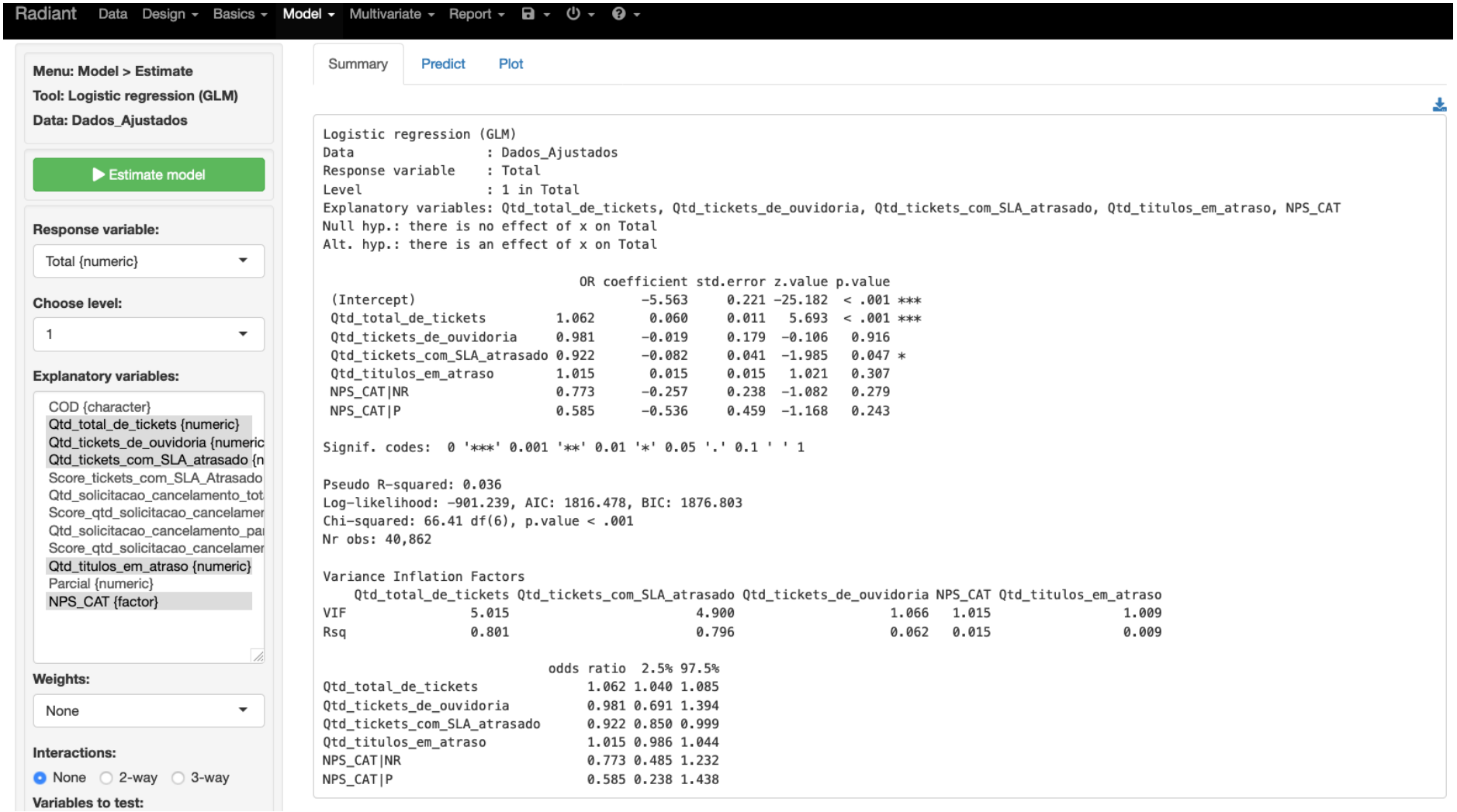
Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 6 - Seleção do Modelo de regressão logística na ferramenta Radiant.



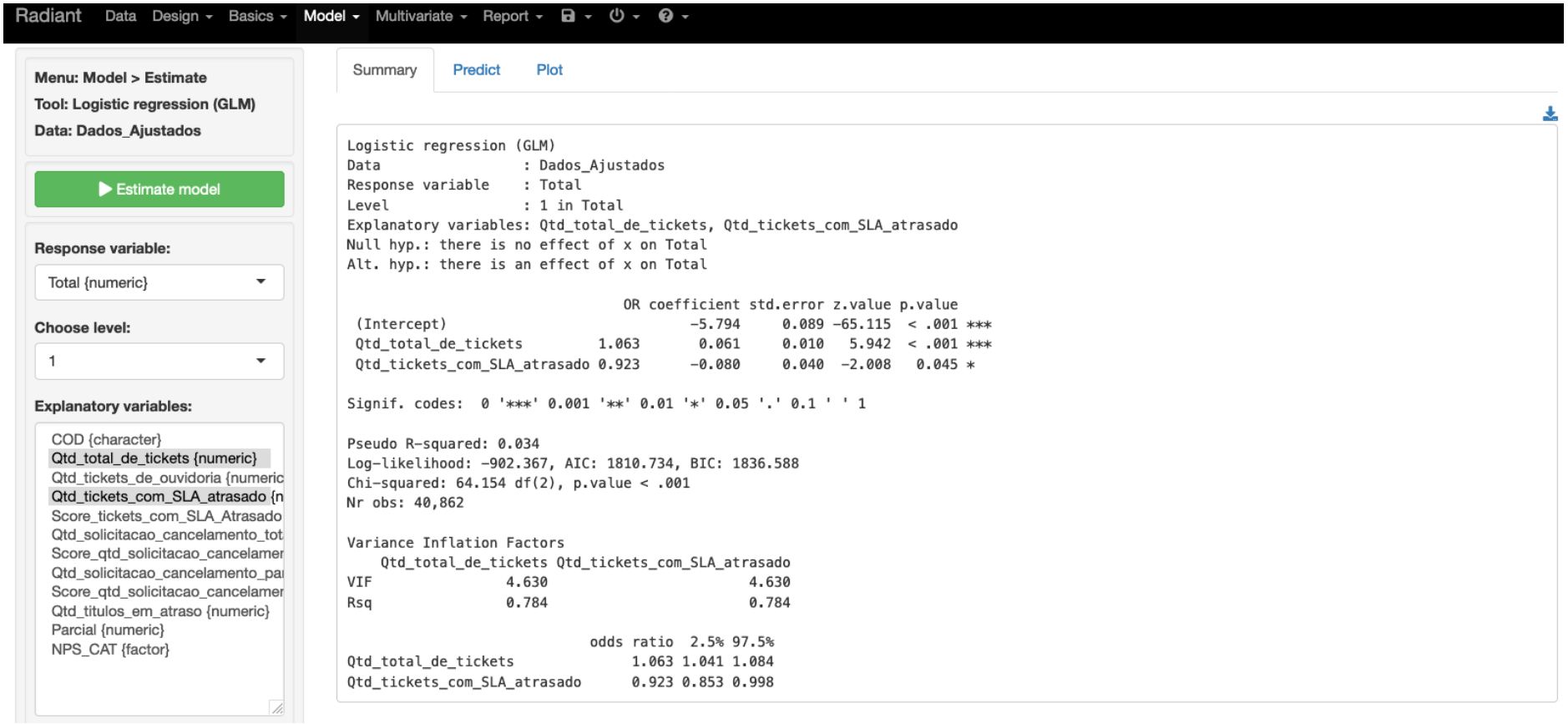
Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 7 – 1ª Rodada de geração do modelo de regressão logística na ferramenta Radiant para as variáveis originais.



Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 8 - 2ª Rodada de geração do modelo de regressão logística na ferramenta Radiant para as variáveis originais.



Fonte: elaboração própria, a partir da ferramenta Radiant.

Sendo assim, foi necessário um aprofundamento na análise. Como os dados iniciais representavam apenas uma fotografia "estática" desses clientes em um determinado mês, não havia volume de casos de *Customer Churn* suficiente para se criar um modelo confiável. Por isso, foram criadas novas variáveis, utilizando-se para isso os dados disponíveis ao longo de diversos meses.

4.2 Segunda etapa da análise

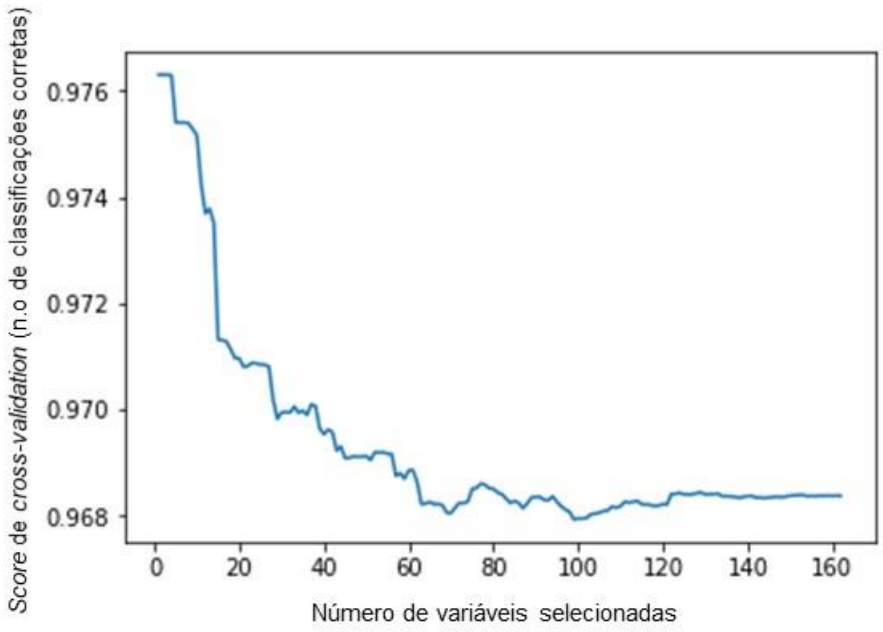
Após realizar a análise preliminar utilizando o conjunto de dados iniciais, foi constatado que seria mais apropriada a criação de novas variáveis. Essas variáveis foram calculadas tendo como base outros dados disponíveis, como explicado anteriormente na seção 3.3.2. O objetivo era avaliar seus efeitos no *Customer Churn* ao longo do tempo.

Em seguida, foi construído um modelo de previsão de *Customer Churn* usando dados de treinamento e prevendo em ambos um subconjunto de testes e dados *OOT*. Foram criadas algumas variações das variáveis para posteriormente selecionar o melhor conjunto de parâmetros a serem considerados na predição do *Customer Churn*, que comporiam uma nova equação do *Customer Health Score*.

Como descrito na seção 3.4, foram aplicadas as técnicas de filtro e *wrapper* com o intuito de se reduzir a dimensionalidade do problema. A aplicação do processo de seleção foi realizada em duas etapas, como mostrado nas Figuras

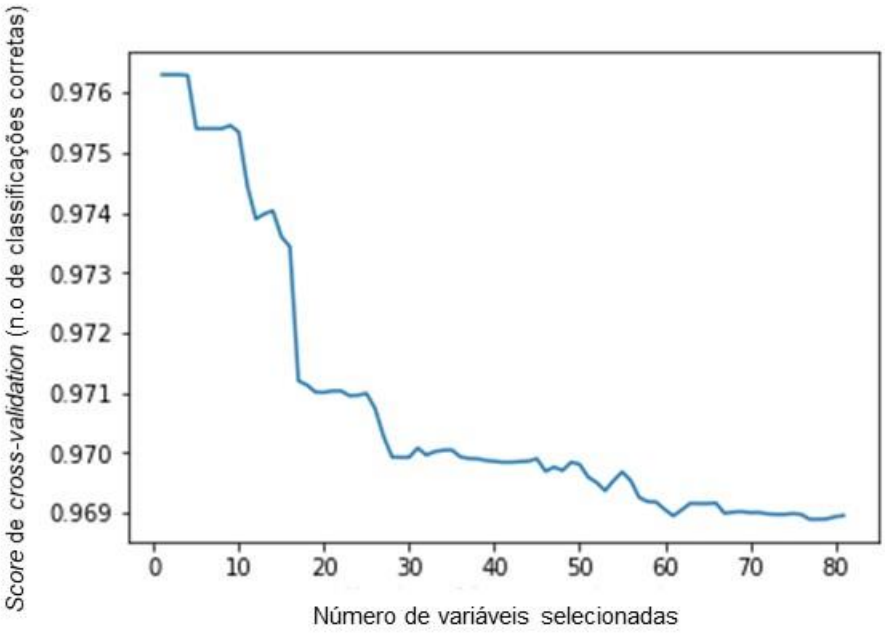
Figura *9* e Figura ***10*** na próxima página. Após essas etapas, chegou-se a uma lista de 20 variáveis de maior influência no *Customer Churn* segundo o método utilizado.

Figura 9 – Gráfico do número de variáveis versus *Scores* de *cross-validation* (primeira rodada).



Fonte: elaboração própria, a partir da linguagem Python.

Figura 10 - Número de variáveis versus *Scores* de *cross-validation* (segunda rodada).



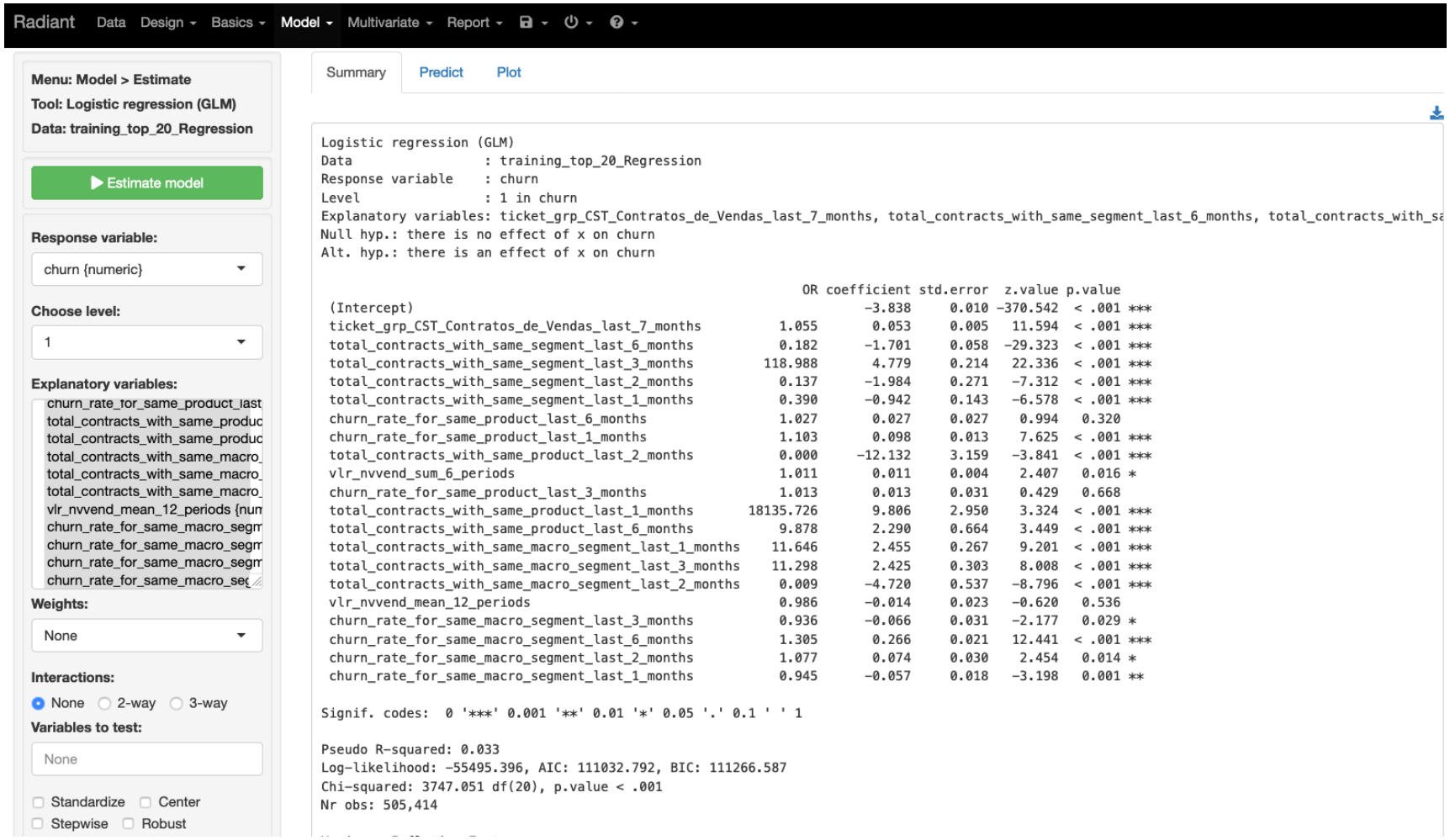
Fonte: elaboração própria, a partir da linguagem Python.

As 20 variáveis foram então carregadas na ferramenta Radiant e um modelo de regressão logística foi aplicado com a opção de uma seleção *stepwise* com eliminação reversa de variáveis. Para a checagem da significância das variáveis incluídas no modelo de regressão

logística, foram utilizados os indicadores *p.value*, *Variance Inflation Factor* (*VIF*) e o *Odds Ratio* (*OR*). Nas análises as variáveis que apresentaram o indicador de *p.value* maiores que 0,05 foram excluídas uma a uma e o modelo novamente calculado, sempre eliminando a variável de maior *p.value* em cada rodada. Terminado o processo de eliminação das variáveis passou-se a analisar o *VIF* e o *OR*. O primeiro, indica se existe algum problema de multicolinearidade no modelo. A multicolinearidade ocorre quando o modelo inclui vários fatores que são correlacionados entre si e apenas uma variável resposta. Em outras palavras, isso ocorre quando existem fatores ligeiramente redundantes. O segundo, indica se a variável tem uma influência positiva ou negativa na probabilidade de *Custumer Churn* no caso dos modelos analisados nesse estudo. Quando o *OR* for maior do que 1, quanto maior for o valor da variável maior a probabilidade de ocorrência do *Churn*. O inverso ocorre caso o *OR* seja menor do que 1, ou seja, quanto maior o valor da variável menor é a chance de ocorrer o *Churn*. Um *OR* igual ou muito próximo a 1 indica uma variável com pouca relevância no modelo criado.

No resultado da regressão apresentado na próxima página, na Figura 11, é possível notar que apenas as variáveis *"Taxa de Churn para o mesmo produto nos últimos 3 meses", "Taxa de Churn para o mesmo produto nos úlimos 6 meses"* e a *"Valor médio de novas vendas nos últimos 12 meses"* apresentaram um *p.value* acima de 0,05, portanto não são relevantes.

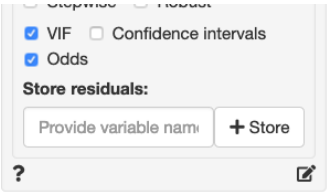
Figura 11 - Geração do modelo de regressão logística na ferramenta Radiant para as variáveis selecionadas a partir de filtros e *wrappers*.



Fonte: elaboração própria, a partir da ferramenta Radiant.

A partir da análise dos indicadores *VIF* e *Odds Ratio*, obtidos pela regressão logística das variáveis selecionadas a partir de filtros e *wrappers*, verifica-se que a maioria delas apresenta valores muito superiores ao considerado adequado para ambos indicadores, acima de 10 no caso do *VIF* e muito acima de 1 para o *Odds Ratio*. Sendo assim, a combinação do método de seleção de variáveis por filtros e wrappers e o modelo de regressão logística, não foi considerado satisfatório para esse conjunto de dados. Um outro fator determinante é o tempo envolvido no processo de seleção de variáveis, o que motivou o uso de uma ferramenta diferente, o *XGBoost,* tendo a sua agilidade apontada pela literatura como grande vantagem.

Figura 12 – Indicadores *VIF* e *Odds Ratio* das variáveis do modelo de regressão logística, selecionadas a partir de filtros e *wrappers*.



Variance Inflation Factors

	VIF	Rsq
total_contracts_with_same_product_last_2_months	105132.059	1.000
total_contracts_with_same_product_last_1_months	91677.398	1.000
total_contracts_with_same_product_last_6_months	4646.936	1.000
total_contracts_with_same_macro_segment_last_2_months	1952.925	0.999
total_contracts_with_same_segment_last_2_months	792.008	0.999
total_contracts_with_same_macro_segment_last_3_months	628.653	0.998
total_contracts_with_same_segment_last_3_months	500.366	0.998
total_contracts_with_same_macro_segment_last_1_months	475.818	0.998
total_contracts_with_same_segment_last_1_months	217.763	0.995
churn_rate_for_same_product_last_3_months	56.218	0.982
churn_rate_for_same_product_last_6_months	45.844	0.978
total_contracts_with_same_segment_last_6_months	35.190	0.972
churn_rate_for_same_product_last_1_months	9.801	0.898
churn_rate_for_same_macro_segment_last_3_months	9.727	0.897
churn_rate_for_same_macro_segment_last_2_months	9.678	0.897
churn_rate_for_same_macro_segment_last_6_months	4.292	0.767
churn_rate_for_same_macro_segment_last_1_months	3.816	0.738
vlr_nvvend_mean_12_periods	1.205	0.170
vlr_nvvend_sum_6_periods	1.200	0.166
ticket_grp_CST_Contratos_de_Vendas_last_7_months	1.016	0.016

	odds ratio	2.5%	97.5%
ticket_grp_CST_Contratos_de_Vendas_last_7_months	1.055	1.045	1.064
total_contracts_with_same_segment_last_6_months	0.182	0.163	0.204
total_contracts_with_same_segment_last_3_months	118.988	78.231	180.979
total_contracts_with_same_segment_last_2_months	0.137	0.081	0.234
total_contracts_with_same_segment_last_1_months	0.390	0.295	0.516
churn_rate_for_same_product_last_6_months	1.027	0.974	1.084
churn_rate_for_same_product_last_1_months	1.103	1.076	1.132
total_contracts_with_same_product_last_2_months	0.000	0.000	0.003
vlr_nvvend_sum_6_periods	1.011	1.002	1.020
churn_rate_for_same_product_last_3_months	1.013	0.954	1.076
total_contracts_with_same_product_last_1_months	18135.726	55.958	5877746.970
total_contracts_with_same_product_last_6_months	9.878	2.688	36.303
total_contracts_with_same_macro_segment_last_1_months	11.646	6.903	19.647
total_contracts_with_same_macro_segment_last_3_months	11.298	6.241	20.452
total_contracts_with_same_macro_segment_last_2_months	0.009	0.003	0.026
vlr_nvvend_mean_12_periods	0.986	0.944	1.031
churn_rate_for_same_macro_segment_last_3_months	0.936	0.881	0.993
churn_rate_for_same_macro_segment_last_6_months	1.305	1.251	1.361
churn_rate_for_same_macro_segment_last_2_months	1.077	1.015	1.142
churn_rate_for_same_macro_segment_last_1_months	0.945	0.912	0.978

Fonte: elaboração própria, a partir da ferramenta Radiant.

4.3 Terceira etapa da análise

Após realizada a seleção de variáveis através dos indicadores mencionados na seção 4.2, foi realizada um novo processo de seleção utilizando o modelo *XGBoost*. A partir das variáveis principais identificadas pelo modelo, foi aplicado novamente o modelo de regressão logística. Ao longo da próximas páginas, nas Figuras Figura ***14***, Figura ***15***, Figura ***16*** e Figura ***17*** é possível ver a evolução da aplicação do modelo de regressão logística com a eliminação das variáveis que possuíam *p.value* maior do que 0,05, como foi feito na aplicação anterior. Além disso, é possível observar que os valores de VIF estão muito altos, - excedem 5 ou 10, indicando uma problemática de multicolinearidade (JAMES *et al.*, 2013) - para algumas variáveis. A presença de colinearidade entre variáveis implica que a informação provida em relação à variável resposta pode ser redundante (JAMES *et al.*, 2013). Ainda no que concerne as variáveis, é possível notar que, em especial, a variável "*Quantidade total de contratos com o mesmo segmento do cliente nos últimos 3 meses*" apresentou um valor de *Odds Ratio* muito acima dos demais, o que pode ter distorcido o modelo, já que isso significa que a variável apresenta uma influência desproporcional na variável resposta (*Churn*).

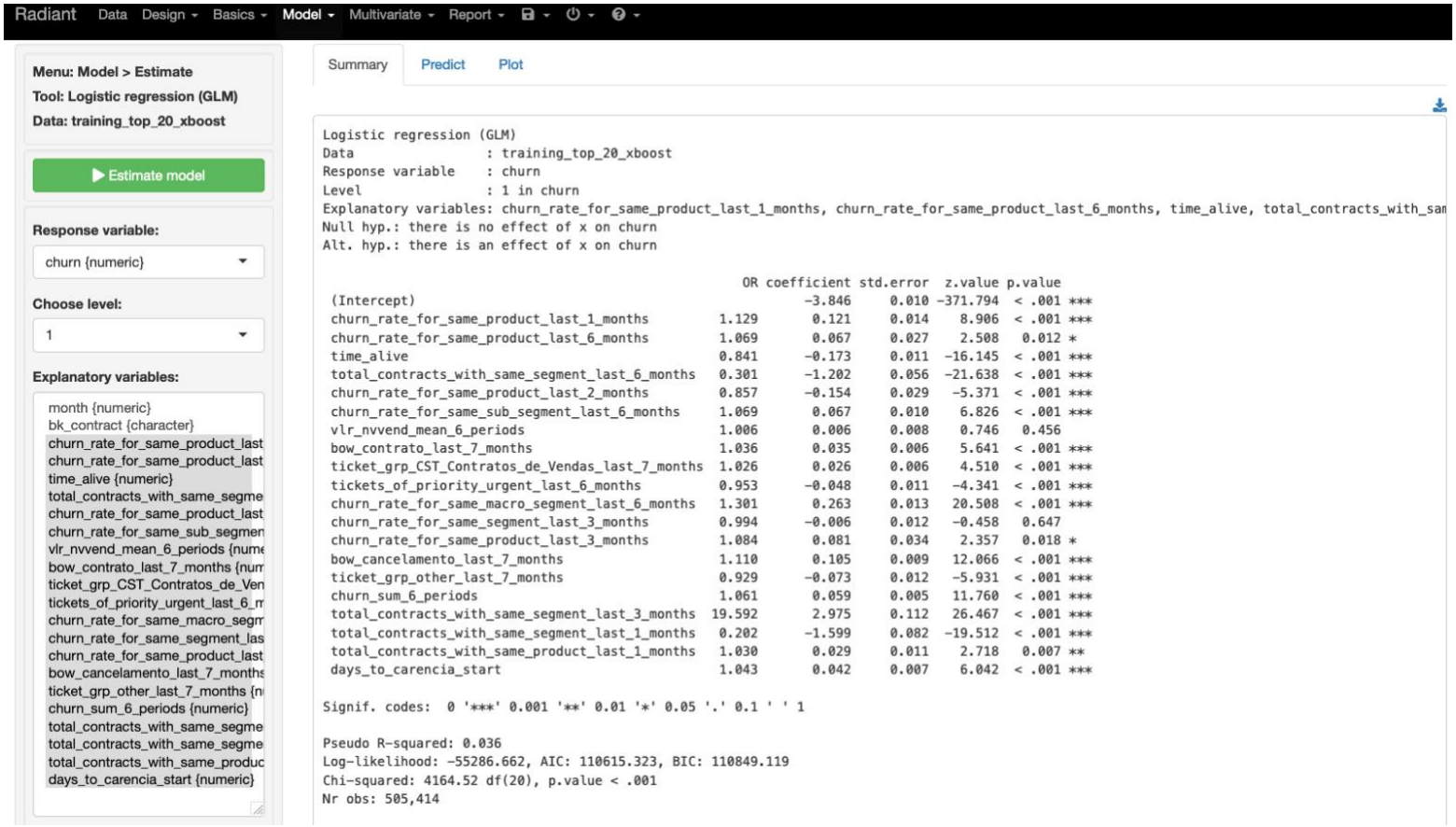
Foi feita uma avaliação final do modelo através de indicadores, como mostram as Figura 18 e Figura 19. Destaca-se o valor do indicador AUC (*Area Under the Curve)* de 0,65. O valor desse indicador representa a área abaixo da curva ROC. Um modelo cujas previsões estão 100% incorretas, apresenta AUC de 0, enquanto um modelo cujas previsões estão 100% corretas, apresenta AUC de 1 (HASSOUNA *et al.*, 2015). Logo, deseja-se que esse indicador se aproxime de 1, mas é esperado que um classificador tenha um desempenho melhor que o acaso quando esse valor supera 0,5.

Figura 13 – Variáveis remanescentes após aplicação da regressão logística com eliminação de variáveis pelo modelo *XGBoost*.

```
------------------------------------------------------
Backward stepwise selection of variables
------------------------------------------------------
Logistic regression (GLM)
Data                 : resultado_xboost
Response variable    : churn
Level                : 0 in churn
Explanatory variables: month, time_alive, churn_rate_for_same_product_last_6_months, total_contracts_with_same
Null hyp.: there is no effect of x on churn
Alt. hyp.: there is an effect of x on churn

                                                 OR coefficient std.error z.value p.value
 (Intercept)                                            503.295    38.171  13.185  < .001 ***
 month                                        0.998      -0.002     0.000 -13.074  < .001 ***
 time_alive                                   1.001       0.001     0.000  26.677  < .001 ***
 churn_rate_for_same_product_last_6_months    0.264      -1.332     0.482  -2.763   0.006 **
 total_contracts_with_same_segment_last_6_months  1.000  -0.000     0.000 -12.733  < .001 ***
 churn_rate_for_same_product_last_1_months    0.092      -2.384     0.346  -6.884  < .001 ***
 churn_rate_for_same_product_last_2_months    2.595       0.954     0.591   1.614   0.106
 vlr_nvvend_sum_12_periods                    1.000       0.000     0.000   1.869   0.062 .
 bow_cancelamento_last_7_months               0.996      -0.004     0.000 -13.902  < .001 ***
 churn_rate_for_same_segment_last_6_months    0.000     -12.661     1.258 -10.063  < .001 ***
 tickets_of_priority_low_last_3_months        1.000       0.000     0.000   6.541  < .001 ***
 churn_rate_for_same_sub_segment_last_6_months  0.000    -8.239     0.640 -12.873  < .001 ***
 vlr_nvvend_mean_6_periods                    1.000       0.000     0.000   2.131   0.033 *
 total_contracts_with_same_product_last_1_months  1.000  -0.000     0.000  -2.724   0.006 **
 bow_contrato_last_7_months                   0.999      -0.001     0.000  -5.236  < .001 ***
 ticket_grp_CST_Contratos_de_Vendas_last_7_months 0.999  -0.001     0.000  -5.619  < .001 ***
 vlr_nvvend_max_6_periods                     1.000      -0.000     0.000  -2.434   0.015 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.028
Log-likelihood: -60391.635, AIC: 120817.27, BIC: 121007.985
Chi-squared: 3480.069 df(16), p.value < .001
Nr obs: 550,466
```

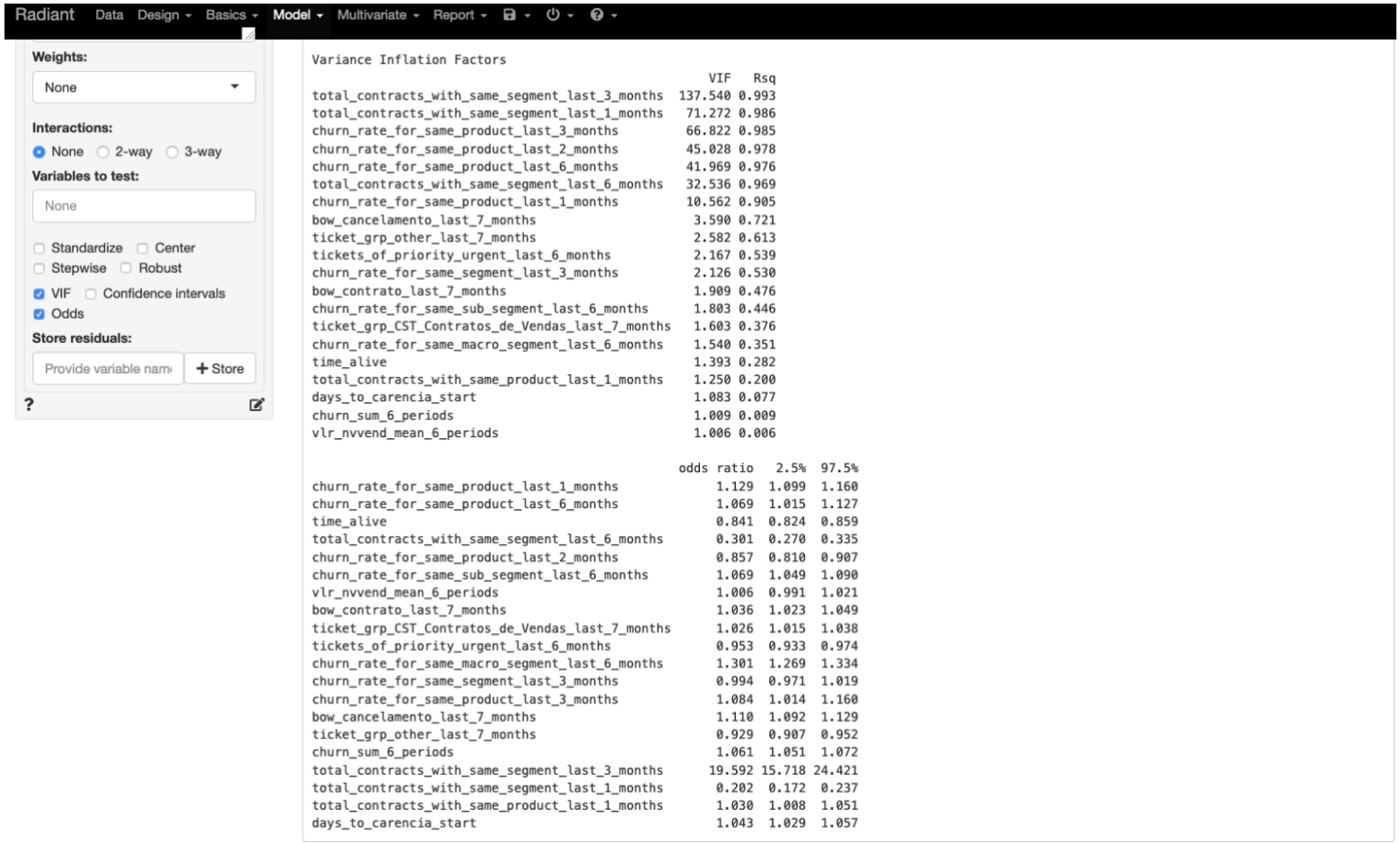
Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 14 – 1ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost.



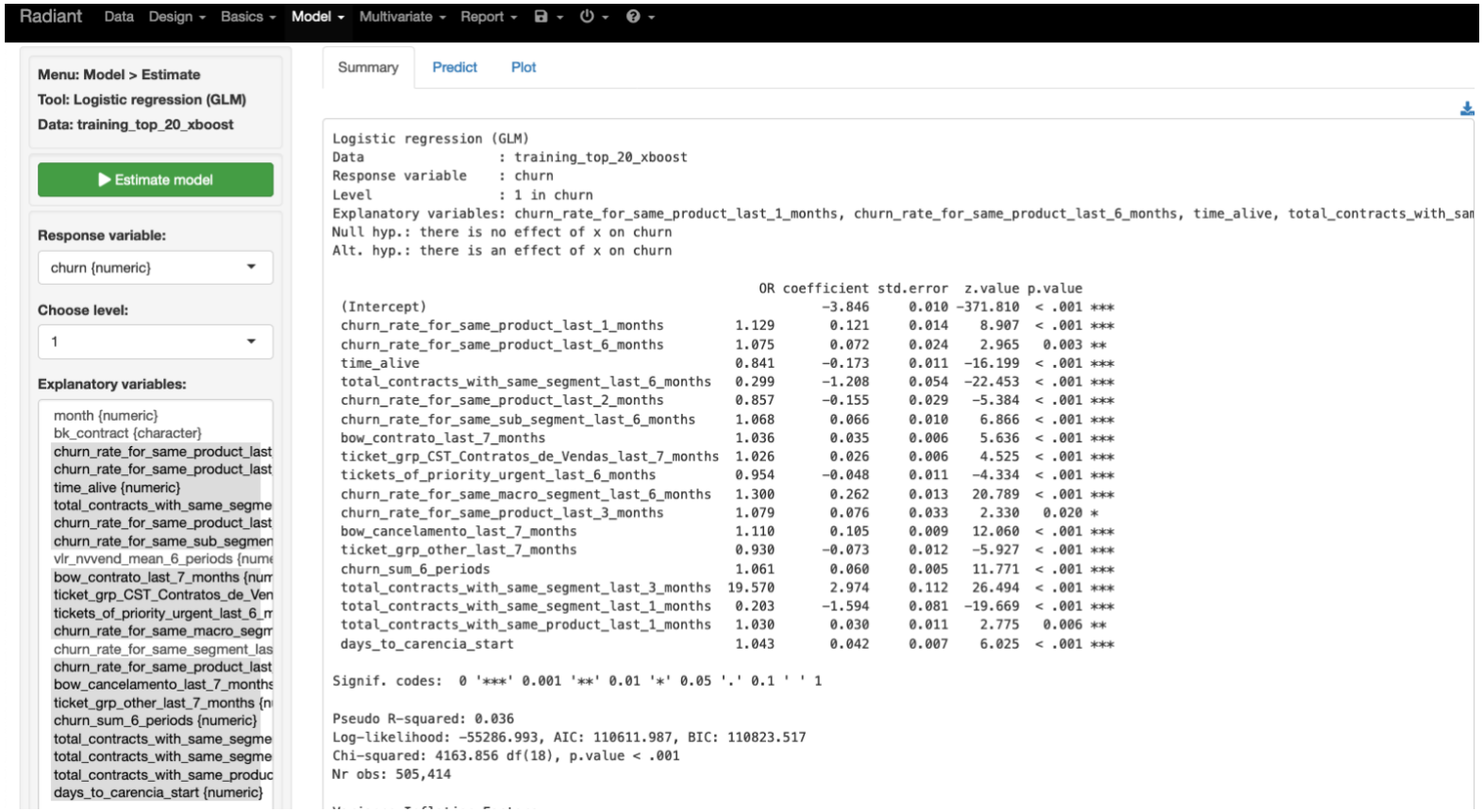
Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 15 – Indicadores da 1ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost.



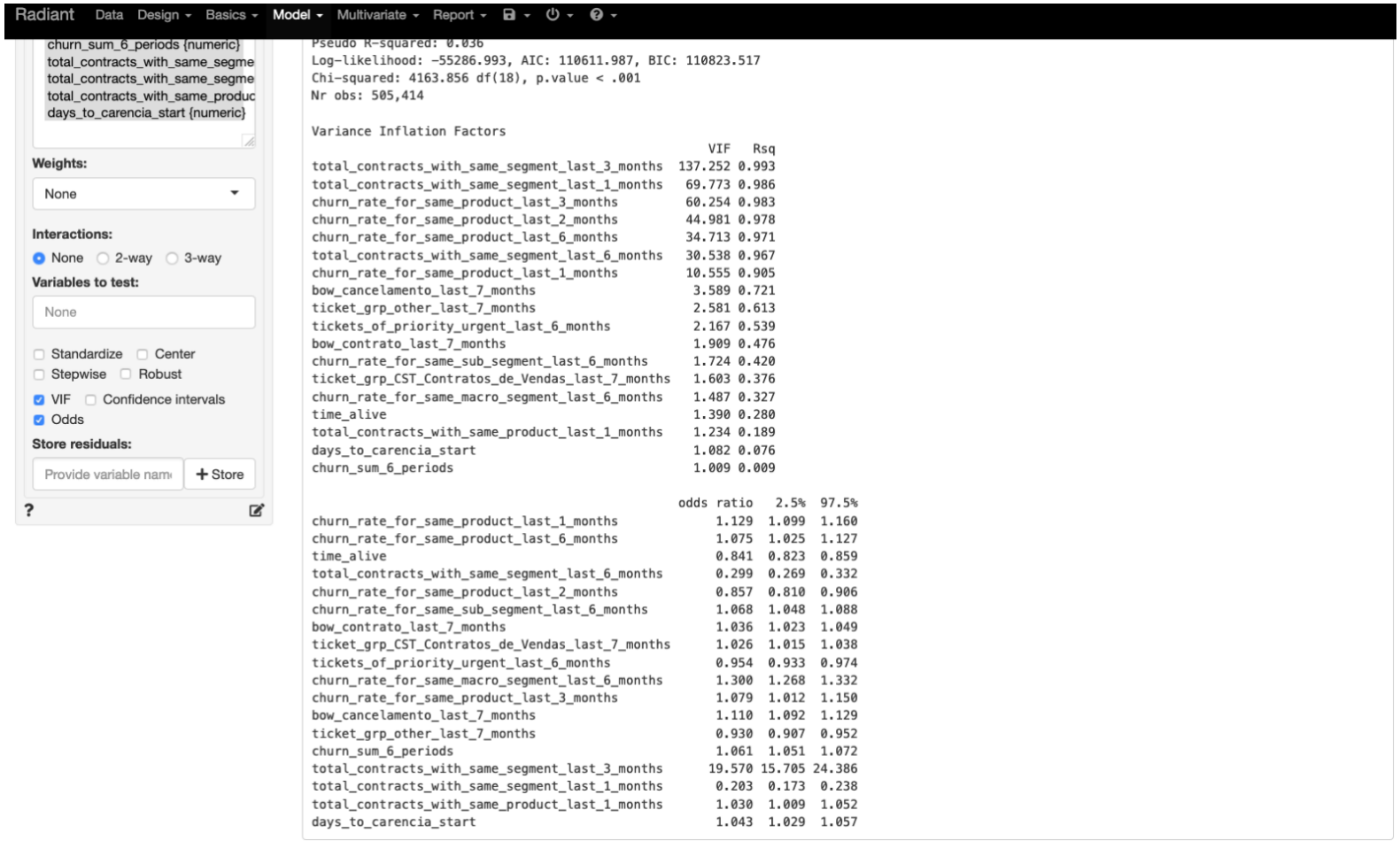
Fonte: elaboração própria, a partir da ferrameta Radiant.

Figura 16 - 3ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost.



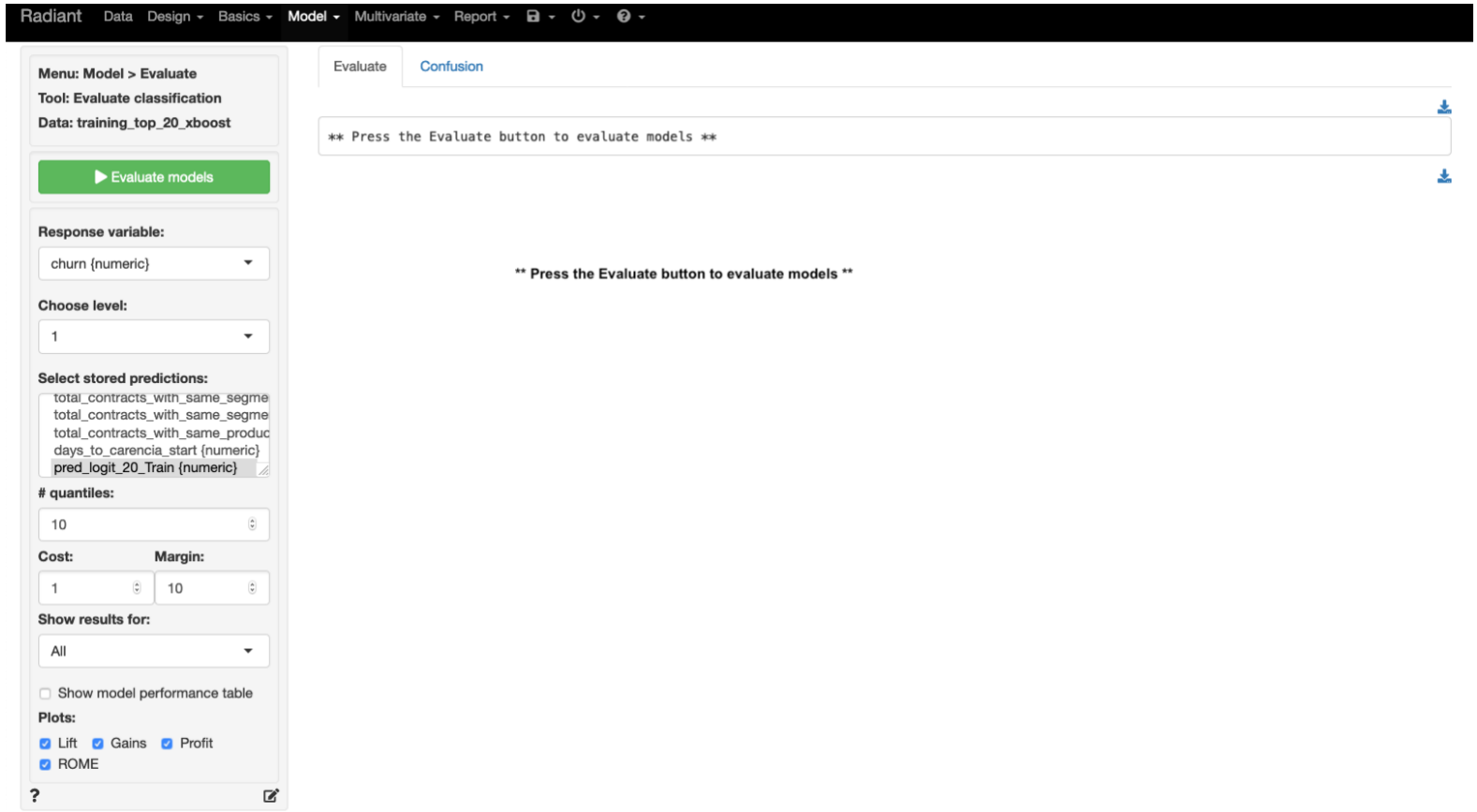
Fonte: elaboração própria, a partir da ferrameta Radiant.

Figura 17 – Indicadores da 3ª Aplicação do modelo de regressão logística nas variáveis selecionada pelo XGBoost.



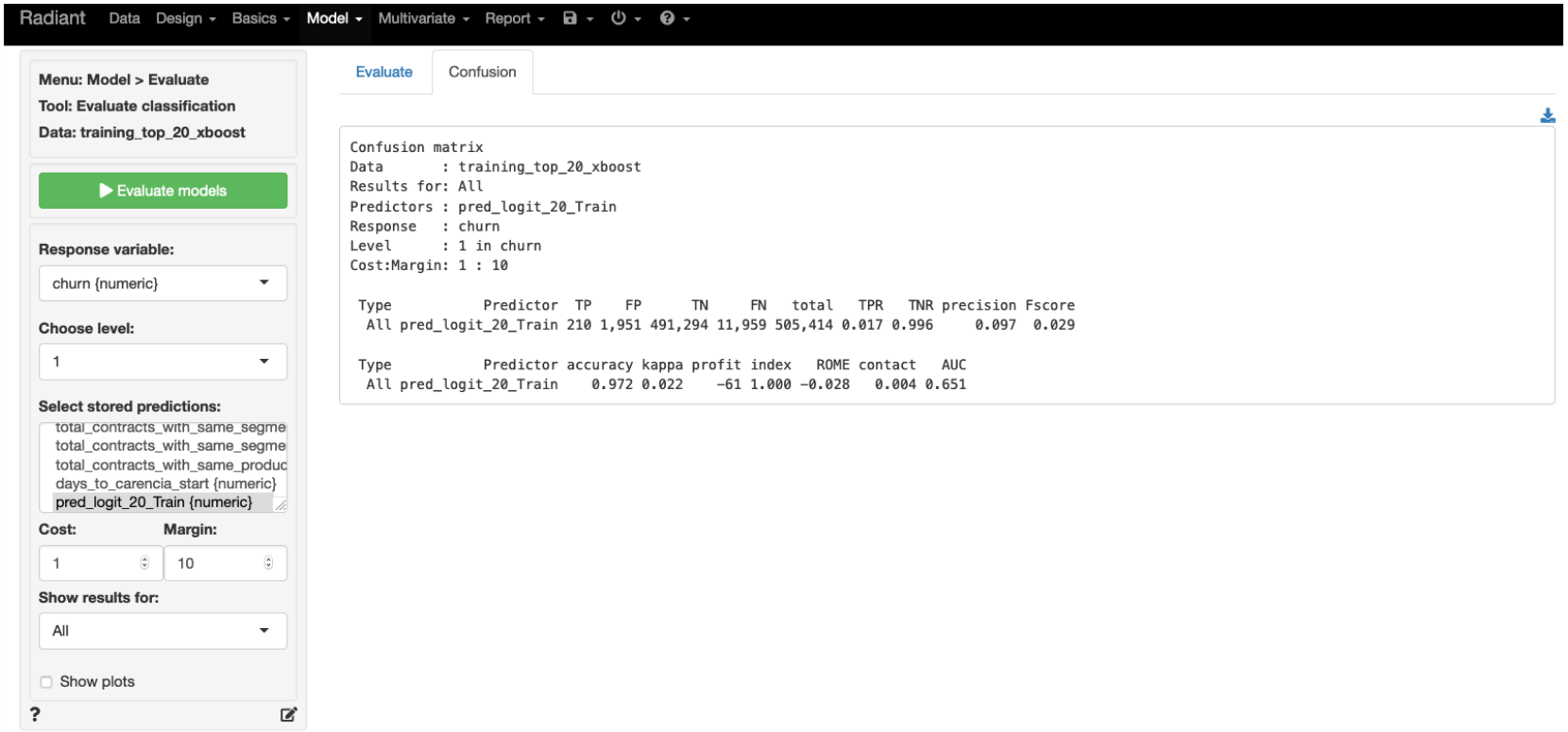
Fonte: elaboração própria, a partir da ferrameta Radiant.

Figura 18 – Etapa de avaliação do modelo.



Fonte: elaboração própria, a partir da ferramenta Radiant.

Figura 19 – Indicadores de avaliação do modelo.



Fonte: elaboração própria, a partir da ferramenta Radiant.

5 CONCLUSÕES

O objetivo geral desta pesquisa foi calibrar o modelo de previsão de *Customer Churn*, ou rotatividade de clientes, a partir do cálculo de *Customer Health Score*. Para isso, foram utilizadas variáveis selecionadas por algoritmos de *Machine Learning* em dados de clientes obtidos de uma grande empresa nacional de *software*. Buscou-se aumentar o potencial de previsão do *Customer Churn*, de modo a prevenir perdas de receitas diretas e indiretas para a companhia provocadas pela saída desses clientes e, consequentemente, auxiliar empresas cujo modelo comercial se baseia em receitas recorrentes.

Para alcançar esse objetivo geral, foram estabelecidos alguns objetivos específicos, de caráter metodológico. Primeiramente, com o intuito de avaliar o modelo atual de cálculo do *Customer Health Score*, foi aplicada uma regressão logística nos dados dos indicadores que o compunham a fim de se testar a significância dessas variáveis. Logo no início da etapa de preparação e análise dos dados, foi possível notar que os resultados foram ruins no que concerne a aderência dessas variáveis ao modelo, uma vez que, a quantidade de clientes com pedido de cancelamento no momento exato em que foram analisados os indicadores originais é muito pequena se comparada com o universo total de clientes na base. Além disso, os pesos utilizados no cálculo do *Health Score* original foram atribuídos pela equipe de *Customer Experience* da companhia baseada em entrevistas realizadas com executivos da própria empresa, não possuindo um rigor estatístico.

Em segundo lugar, reuniram-se variáveis de interação com o cliente, contratos existentes e informações comerciais para ampliar rol de variáveis candidatas a comporem o *Customer Health Score*, visando o aumento da previsão de *Customer Churn*. Para realizar essa tarefa as variáveis foram calculadas tendo como base outros dados disponíveis na companhia. O valor que foi agregado nessa análise foi o de se avaliar além de um escopo maior de variáveis, essas foram medidas ao longo do tempo cumprindo assim dois objetivos específicos de uma vez. Foi aplicado um modelo de regressão logística novamente, no entanto, foi reduzida a dimensionalidade do problema a priori através do uso de técnicas de filtro e wrapper em cima desse volume maior de variáveis.

Em terceiro lugar, nas mesmas variáveis candidatas criadas foi aplicado o modelo XGBoost para redução da dimensionalidade no lugar das técnicas de filtro e *wrapper*. Essa

abordagem se mostrou mais eficiente e eficaz na medida em que dispendeu menos tempo e trouxe resultados melhores se comparados ao experimento anterior.

Por fim, as 20 variáveis remanescentes do processo de redução de dimensionalidade executado através do *XGBoost*, por si só poderiam ser utilizadas como um direcionamento de quais variáveis seriam as mais importantes dentre todas as geradas nesse estudo.

Nas Tabela 9 e Tabela 10 é possível ver uma comparação do desempenho dos modelos utilizados durante as etapas desse estudo, quantitativamente e qualitativamente na ordem. Dentro os indicadores quantitativos apresentados, destaca-se que tanto a Acurácia quanto o *AUC* são indicadores que, embora normalmente muito utilizados para avaliação de modelos, no caso estudado, variaram pouco entre os modelos. Isso ocorreu, pois o número de casos de *Customer Churn* versus o universo total da amostra era pequeno, e assim, não sendo indicados para a comparação de desempenho entre os modelos nesse caso estudado. Os indicadores que representam melhor os aspectos observados nos modelos são a Precisão, *F-Score* e o *TPR*. Os três indicadores demonstram que o modelo de *XGBoost* apresentou os melhores resultados para o caso estudado. Nas Tabelas Tabela ***11***, Tabela ***12***, Tabela ***13***, Tabela ***14***, Tabela ***15***, Tabela ***16***, Tabela ***17***Tabela 18 são apresentadas as matrizes de confusão dos quatro métodos testados, tanto em quantidades quanto em percentuais do total do número total de dados utilizados em cada modelo.

Tabela 9 – Comparativo quantitativo dos modelos aplicados no estudo.

Preditor	*TP*	*FP*	*TN*	*FN*	Total	*TPR*	*TNR*	Precisão	*F-Score*	Acurácia	*AUC*
Reg. Log. c/ vars. originais	0	1	8.144	28	8.173	0,0000	0,9999	0,0000	0,0000	0,99645	0,79129
Reg. Log. c/ vars. pós filtro e wrapper	11	18	58.348	1.535	59.912	0,0071	0,9997	0,3793	0,0140	0,97408	0,66752
Reg. Log. c/ vars. pós XGBoost	14	52	58.314	1.532	59.912	0,0091	0,9991	0,2121	0,0174	0,97356	0,70428
XGBoost apenas	1.788	672	2.083	943	5.486	0,6547	0,7561	0,7268	0,6889	0,70561	0,70530

Fonte: elaboração própria, a partir da ferramenta Radiant, códigos em Python e cálculos em planilha.

Tabela 10 – Categorias de Implicação Gerencial na comparação dos modelos aplicados no estudo.

Preditor	Tempo demandado	Coerência dos Dados	Possibilidade de Encaminhamento
Reg. Log. c/ vars. originais	Baixo	Incoerente	Muito Baixa
Reg. Log. c/ vars. pós filtro e wrapper	Muito Alto	Pouco Coerente	Baixa
Reg. Log. c/ vars. pós XGBoost	Médio	Pouco Coerente	Média
XGBoost apenas	Baixo	Pouco Coerente	Baixa

Fonte: elaboração própria, a partir da ferramenta Radiant, códigos em Python e cálculos em planilha.

Tabela 11 - Matriz de confusão da regressão logística com as variáveis originais em quantidades.

Classe	predita C+	predita C-	Totais
Verdadeira C+	0	28	**28**
Verdadeira C-	1	8.144	**8.145**
Totais	**1**	**8.172**	**8.173**

Fonte: elaboração própria.

Tabela 12 - Matriz de confusão da regressão logística com as variáveis originais em percentuais do total do número total de dados utilizados.

Classe	predita C+	predita C-	Totais
Verdadeira C+	0,0%	0,3%	0,3%
Verdadeira C-	0,0%	99,6%	99,7%
Totais	0,0%	100,0%	100,0%

Fonte: elaboração própria.

Tabela 13 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação de filtro e do *wrapper* para redução do número de variáveis em quantidades.

Classe	predita C+	predita C-	Totais
Verdadeira C+	11	1.535	**1.546**
Verdadeira C-	18	58.348	**58.366**
Totais	**29**	**59.883**	**59.912**

Fonte: elaboração própria.

Tabela 14 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação de filtro e do wrapper para redução do número de variáveis em percentuais do total do número total de dados utilizados.

Classe	predita C+	predita C-	Totais
Verdadeira C+	0,0%	2,6%	2,6%
Verdadeira C-	0,0%	97,4%	97,4%
Totais	0,0%	100,0%	100,0%

Fonte: elaboração própria.

Tabela 15 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação do XGBoost para redução do número de variáveis em quantidades.

Classe	predita C+	predita C-	Totais
Verdadeira C+	14	1.532	**1.546**
Verdadeira C-	52	58.314	**58.366**
Totais	**66**	**59.846**	**59.912**

Fonte: elaboração própria.

Tabela 16 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação do XGBoost para redução do número de variáveis em percentuais do total do número total de dados utilizados.

Classe	predita C+	predita C-	Totais
Verdadeira C+	0,0%	2,6%	2,6%
Verdadeira C-	0,1%	97,3%	97,4%
Totais	0,1%	99,9%	100,0%

Fonte: elaboração própria.

Tabela 17 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação apenas do XGBoost em quantidades.

Classe	predita C+	predita C-	Totais
Verdadeira C+	1.788	943	**2.731**
Verdadeira C-	672	2.083	**2.755**
Totais	**2.460**	**3.026**	**5.486**

Fonte: elaboração própria.

Tabela 18 - Matriz de confusão da regressão logística com as variáveis remanescente após aplicação apenas do XGBoost em percentuais do total do número total de dados utilizados.

Classe	predita C+	predita C-	Totais
Verdadeira C+	32,6%	17,2%	49,8%
Verdadeira C-	12,2%	38,0%	50,2%
Totais	44,8%	55,2%	100,0%

Fonte: elaboração própria.

Em termos práticos, em outras palavras, para a companhia é menos prejudicial contatar um cliente, ou até dezenas deles, que não irá provavelmente se tornar um *Churn* de forma errônea do que não classificar um cliente com alta probabilidade de *Churn* como sendo um cliente que a companhia deva contatar e tomar ações preventivas de *Churn*. No limite, um

modelo que classifique 100% dos clientes como não *Churn* irá errar menos de 1% dos Tn (verdadeiros negativos), o que a priori parece ser um bom número. No entanto, pode-se observar na Tabela 9 que o modelo de regressão logística aplicado às variáveis originais que compunham o *Customer Health Score*, se comportou exatamente dessa forma, com TNR de quase 100%, mas em compensação foi o pior modelo de todos para prever o *Customer Churn*, uma vez que seu TPR foi de 0. Nesse caso avaliado, foram perdidos 28 clientes, que potencialmente no caso da empresa estudada poderiam ter na média cerca de R$70.000 de *LTCV* cada, ou seja, fazendo com que a companhia perdesse cerca de quase R$2 milhões e o modelo não acusasse nada para os gestores ou pessoas que poderiam atuar para reverter essas perdas de receita.

Além disso, é importante enfatizar que, se por um lado o modelo de regressão logística da forma que foi utilizado com uma equação linear, não consegue captar de maneira eficiente os efeitos de não-linearidade que o problema apresenta, o modelo *XGBoost* por outro foi capaz de absorver melhor esses efeitos não-lineares. Isso ocorre, pois em seu caso, árvores mais profundas também ajudam a capturar a interação das variáveis. Portanto, não há necessidade de procurar transformações apropriadas. Quando se depara com um problema de alta dimensão, *XGBoost* "supera" a maldição da dimensionalidade por não depender de nenhuma métrica de distância. Além disso, a semelhança entre os pontos de dados é aprendida com os dados através do ajuste adaptativo das vizinhanças. Isso torna o modelo imune à maldição da dimensionalidade.

No entanto, devido à dificuldade de interpretação de seus resultados em relação ao efeito promotor ou preventor do *Churn*, o *XGBoost* carece de análises complementares que dêem suporte à interpretabilidade da influência das variáveis.

O último objetivo específico que era de se utilizar as análises e resultados para apontar desdobramentos gerenciais que possibilitem a redução de *Customer Churn*, serão tratados no próximo tópico dessa conclusão.

Vale ressaltar que a validade preditiva dos modelos apresentados se estende por um determinado tempo, sendo que os resultados devem ser revalidados e possivelmente os modelos recalibrados a cada ano, semestre ou até trimestre dependendo do tipo de negócio que for analisado. Uma atualização anual ao mínimo seria desejável pois possivelmente abarcaria as sazonalidades anuais das variáveis e os reflexos nas mesmas das mudanças de estratégias macros da companhia como um todo.

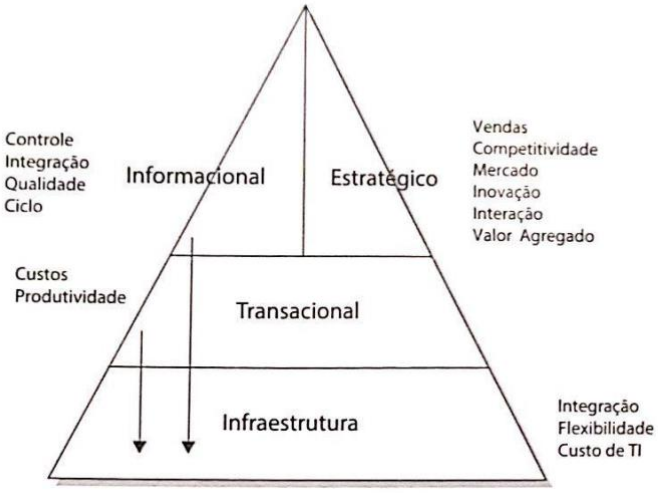
5.1 Implicações gerenciais

5.1.1 Condições de contorno para o desenvolvimento de iniciativas de Customer Analytics

Com o objetivo de manter um bom relacionamento com seus clientes e garantir um crescimento sustentável de suas carteiras, empresas de diversos setores da economia têm investido em entender as necessidades de seus clientes e os fatores que mais podem afetar o *Customer Experience* ao longo de toda a sua jornada com o negócio.

No entanto, nem todas as empresas possuem uma cultura de dados apropriada para que essa compreensão do *Customer Experience* seja efetiva e baseada em análises estatísticas robustas. Uma cultura adequada de dados, está ligada a investimentos constantes tanto em uma boa infraestrutura de TI, quanto de coleta e manutenção de dados. Pode-se pensar nos investimentos em TI como uma carteira de investimentos, para a qual existem riscos e retornos associados, como demonstra a Figura 20 (ALBERTIN & ALBERTIN, 2016)

Figura 20 - Carteira de investimentos em Tecnologia da Informação.



Fonte: Albertin e Albertin (2016) *apud* Weill e Broadlent (1998).

Nesse sentido, não é possível tomar boas decisões estratégicas desassociadas de uma base de infraestrutura, que contempla desde servidores até boas redes de computadores, uma base transacional, que abarca sistemas e ferramentas de *software* que integrem as áreas e uma base informacional robusta, que inclui boa engenharia de dados e análises que permeiam diversas etapas do *Customer Journey*. Todas essas áreas envolvem também investimentos em

pessoas qualificadas para operacionalizarem e garantirem a manutenção dessa cultura. Portanto requerem esforços de captação e seleção de profissionais do mercado, passando por capacitação dos profissionais que já compõem o quadro de funcionários da empresa e envolvimento das lideranças na manutenção dessa cultura.

A visão de que essas questões levantadas, além dos investimentos necessários, é responsabilidade de apenas uma área ou departamento de TI do negócio é ultrapassada e impede a aplicação de uma verdadeira cultura de dados, que é imprescindível para desenvolvimentos de iniciativas estratégicas como a proposta nesse trabalho.

Nesse sentido, não à toa, em empresas prestadoras de serviço (MEIRELLES, 2019), onde o investimento em TI é mais elevado em comparação com as demais indústrias, se vê uma cultura de dados mais madura. Isso se traduz em um volume maior de inovações em mercados de serviço. Um mercado onde isso é uma realidade atualmente, no Brasil, é o financeiro, em que se observa empresas incumbentes investindo em toda a cadeia mostrada na Figura 20. Essas empresas estão reagindo a um movimento recente do crescimento do número de *start-ups*, chamadas também de *fintechs*, que surgiram nos últimos anos e tomaram parte do *marketshare* das grandes instituições financeiras. Segundo a 8ª Edição do Radar Fintechlab (2019), o número de fintechs e iniciativas de eficiência financeira atuantes no Brasil passou de 453 empresas em agosto de 2018 para 604 em junho de 2019, representando um crescimento percentual de 33 pontos. Consequentemente, houve uma pressão na margem desses negócios devido ao aumento súbito da concorrência.

Com o cenário do mercado de serviços se acirrando devido ao rápido desenvolvimento de novas empresas e iniciativas, é importante evitar que clientes já captados pelas empresas incumbentes sejam perdidos para novos entrantes. Sendo assim, o presente trabalho apresenta uma alternativa para atuar preventivamente à rotatividade de clientes, diminuindo esforços de captação de novos clientes, o custo de aquisição (*CAC*) com novos clientes e a perda de receitas recorrentes. No caso específico da indústria de *SaaS*, os fabricantes de *software* tornaram-se dependentes da renovação contínua da assinatura de seus clientes, na medida em que a capacidade de recuperar o alto *CAC* −, no que concerne às vendas iniciais de licenças, característico do modelo perpétuo de licenciamento, diminuiu substancialmente (DEMPSEY; KELLIHER, 2018)

5.1.2 Dados necessários para implantação da metodologia proposta

Para as empresas, este trabalho fornece uma metodologia para criação de modelos preditivos de *Customer Churn* e indiretamente o *Customer Health Score* a partir de diversos indicadores que podem ser medidos em todas as etapas de interação desse cliente com o negócio em sua jornada. Por meio dos resultados obtidos, é possível medir em diversas etapas do *Customer Journey* a influência de várias métricas que afetam esses indicadores principais e, assim, embasar tomadas de decisão que visem a melhoria dos serviços e do relacionamento com o cliente. Como consequência, é possível desenvolver, a partir desse trabalho, um plano de ações que vise a redução do *Customer Churn* e também minimizar a perda de receita associada a essa quebra dos contratos por parte dos clientes.

No entanto, as interações com os clientes em cada estágio tendem a ter efeito no período em que os clientes permanecem efetivamente na empresa (FREEZE; KULKARNI, 2007). Portanto, para que a utilização do método proposto seja possível, deve-se coletar dados de interações com os clientes, que envolvem todas as ocasiões em que há contato entre ele e a empresa, incluindo: solicitações, alterações, acessos, dúvidas, reclamações, etc. Portanto, são necessários diferentes tipos de ações e comunicações em cada um desses estágios. A perspectiva do ciclo de vida é importante para o *marketing*, uma vez que acompanha essas interações do cliente com a empresa e as associa a atividades adequadas à medida que os clientes passam por essas etapas. Isso faz com que, em cada momento, a empresa seja capaz de priorizar a maximização do *Customer Lifetime Value* (*CLTV*), normalmente aumentando seu fluxo de caixa (HANSOTIA, 2002).

Ao mesmo tempo, deve-se coletar dados relacionados aos contratos existentes, que envolvem: tempo de vida do cliente, condições de contrato, produto fornecido, condições comerciais, etc. Finalmente, é necessário coletar informações comerciais relevantes, que incluem: valores de bonificação, tempo de carência, reajustes, segmento do cliente, canal de atendimento, região em que se encontra, etc.

Como é possível perceber, a reunião desses dados exige da empresa competências variadas que permitam coletá-los e armazená-los de forma organizada e eficiente, para que estejam disponíveis e, principalmente, sejam confiáveis. Para empresas que não possuem um quadro de funcionários preparados para compor essas competências e/ou que não possuem engenharia e tratamento de dados confiáveis para armazenamento, é necessário que se adequem, já que concentrar-se em seu relacionamento com o cliente é um bom meio de garantir fluxos de receita recorrente e sustentável (DEMPSEY; KELLIHER, 2018).

Além disso, a manutenção periódica do próprio modelo permitirá a inclusão de novas métricas ao longo do tempo, e uma gestão mais voltada para dados, o que também possibilitará aos gestores a tomada de decisões mais embasadas. Isso pode permitir que a companhia evite perdas e faça mudanças estratégicas de maneira mais rápida e segura, na medida em que esse modelo atingir uma maturidade maior.

5.1.3 Aspectos técnicos referentes à metodologia proposta

Neste trabalho, optou-se por utilizar um algoritmo de regressão logística para que o resultado pudesse ser mais facilmente traduzível em pesos para as variáveis que seriam sugeridas para compor o cálculo do novo *Customer Health Score*. Além disso, outro motivo para se ter optado pela regressão logística é que o desempenho desse algoritmo pode servir como uma base de comparação para outros algoritmos mais complexos que podem ser testados em seguida. O modelo preditivo *XGBoost* pode ser usado com boa eficiência para seleção das variáveis mais relevantes para análise de *Churn*. No entanto, para se obter um entendimento completo do impacto de cada variável, seja ela promotora ou preventora de *Churn* e em qual medida, é necessário utilizar-se o modelo de regressão logística. Sabe-se que o modelo *XGBoost* também tem a vantagem de além de ser robusto, assim como o modelo de regressão logística (NIE *et al.*, 2011), traz bons resultados exigindo um tempo menor para seu cálculo (NISBET; ELDER; MINER, 2009, WANG *et. al.*, 2017).

5.2 Sugestão de novas métricas

Após rodar o modelo de *XGBoost*, chegou-se a uma lista de novas 20 variáveis que mais influenciam o *Customer Churn*. A Tabela 19 contém as variáveis selecionadas e seu respectivo grau de importância atribuído pelo modelo *XGBoost*.

Tabela 19 - Lista das 20 variáveis que têm maior importância segundo a análise realizada com o XGBoost.

Nome da Variável	Importância
Taxa de *Churn* para o mesmo produto consumido pelo cliente no último mês	13
Taxa de *Churn* para o mesmo produto consumido pelo cliente nos últimos 6 meses	12
Tempo como cliente da empresa (dias)	12
Número total de contrato com o mesmo segmento do cliente nos últimos 6 meses	10
Taxa de *Churn* para o mesmo produto consumido pelo cliente nos últimos 2 meses	8
Taxa de *Churn* para o mesmo subsegmento do cliente nos últimos 6 meses	8
Valor médio de novas vendas nos últimos 6 meses	7
Número de *tickets* de atendimento com a palavra "Contato" no assunto nos últimos 7 meses	7
Número de *tickets* de atendimento com conteúdo direcionado a área de "Contratos de vendas" nos últimos 7 meses	7
Número de *tickets* de atendimento com a prioridade “Urgente" nos últimos 7 meses	6
Taxa de *Churn* para o mesmo macrosegmento do cliente nos últimos 6 meses	6
Taxa de *Churn* para o mesmo segmento do cliente nos últimos 3 meses	6
Taxa de *Churn* para o mesmo produto do cliente nos últimos 3 meses	6
Número de *tickets* de atendimento com a palavra "Cancelamento" no assunto nos últimos 7 meses	5
Número de *tickets* de atendimento com conteúdo direcionado a "Outros" nos últimos 7 meses	5
Quantidade total de clientes que deram *Churn* nos últimos 6 meses	5
Quantidade total de contratos com o mesmo segmento do cliente nos últimos 3 meses	5
Quantidade total de contratos com o mesmo segmento do cliente no último mês	5
Quantidade total de contratos com o mesmo produto do cliente no último mês	4
Dias até o início da carência do cliente	4

Fonte: elaboração própria, a partir da linguagem Python.

5.3 Limitações

Por se tratar de um estudo de caso exploratório realizado em uma única empresa, os resultados aqui obtidos não são estatisticamente generalizáveis. Tendo sido realizado o estudo em uma empresa brasileira de grande porte do mercado de *software*, majoritariamente fornecedora de *software* do tipo *ERP*, os resultados obtidos podem não ser aplicáveis a empresas de *software* de outra natureza, que não possuam as mesmas características, forças de barganha com o cliente e barreiras de saída equivalentes às da empresa estudada.

Outras limitações que merecem ser observadas relacionam-se aos dados utilizados no estudo. Primeiramente, havia uma restrição na disponibilidade de dados históricos confiáveis. Quanto mais recentes, mais especificamente de janeiro de 2018 em diante, mais confiáveis se tornaram os dados. Isso ocorreu, pois o processo de armazenamento desses dados se tornou mais estruturado dentro da empresa desde o mês de janeiro de 2018. Dessa forma, não foi possível realizar uma análise tão extensa nos dados em relação ao histórico das informações anteriores a esse mês.

Outra questão que concerne aos dados, é que alguns deles, embora já em parte disponíveis na empresa, estavam desorganizados, ou seja, não otimizados para uso em análises como a realizada nesse trabalho. Nesse sentido, por exemplo, existem dados de uso das ferramentas e serviços de *software* por parte dos usuários que a companhia ainda não organizou de forma a serem analisados facilmente. Devido ao volume muito grande desses dados de telemetria, a solução encontrada pela empresa foi armazená-los em uma nuvem da Microsoft Azure. Existem duas questões em relação a esses dados de uso dos *software* que dificultam sua utilização: primeiramente, é necessário organizá-los e eventualmente aplicar transformações para que seja possível interpretá-los de maneira que possam ser aplicados em decisões no negócio; em segundo lugar, muitos dos produtos e serviços *SaaS* da companhia possuem dados de uso que ainda não foram mapeados e, portanto, ficariam de fora da análise. A indisponibilidade de dados de uso de alguns dos produtos e serviços *SaaS*, diminuem a relevância dessa informação no modelo e por isso optou-se por não utilizar esses dados neste estudo.

Finalmente, o estudo apresenta limitações em relação aos modelos aplicados. O modelo de regressão logística possui restrições no que tange o número máximo de variáveis a serem consideradas e também no que se refere aos efeitos combinados, ou multicolinearidade, das variáveis na função objetivo. Poderia ter sido feita uma outra análise utilizando a regressão

logística e eliminando-se as variáveis com maior colinearidade a cada iteração da aplicação do modelo. Em uma segunda etapa da análise, foi aplicado o modelo *XGBoost* com o intuito se minimizar ambos esses efeitos. No entanto, com isso, perdeu-se parte da interpretabilidade das variáveis na medida em que não se podia mais inferir se essas variáveis tinham efeitos positivo ou negativo no *Customer Churn*. Só se podia afirmar que elas tinham maior influência no modelo em relação às outras variáveis avaliadas e não se seu efeito era positivo ou negativo. Para minimizar esse efeito indesejado, esse novo conjunto de variáveis selecionadas foi novamente submetida a um modelo de regressão logística, reintroduzindo os problemas citados incialmente nesse parágrafo.

5.4 Estudos futuros

Em primeiro lugar, propõe-se a aplicação de estudo de caso exploratório semelhante ao feito neste trabalho em diversas organizações, inclusive em algumas com características distintas da empresa analisada nesse estudo. Poderia ser interessante, tanto do ponto de vista acadêmico quanto do prático, a aplicação de estudo semelhante em empresas de menor porte, mais jovens e que possuam produtos com barreiras de saída menores – ou em outras palavras, produtos que são mais facilmente substituíveis - e um menor tempo médio de relacionamento com seus clientes por exemplo.

Em segundo lugar, poderiam ser incluídas outras métricas, que ainda não foram consideradas nos modelos testados aqui. Por exemplo, um dos indicadores que poderia ser acrescentado na confecção das variáveis seria o dado de uso de *software* e serviços *SaaS* por parte dos clientes. Essa informação, *a priori*, aparentemente apresenta grande potencial de trazer mais idéias alternativas de como melhorar o modelo de previsão do *Customer Churn*.

Em terceiro lugar, além desse indicador de uso, outros indicadores que medem a satisfação dos clientes durante sua jornada na empresa poderiam ser incluídos. Por exemplo, já estão sendo implementadas pesquisas de satisfação dos clientes em que é obtido um *Customer Satisfaction Score* (*CSAT*) durante os projetos de consultoria para a implantação dos *ERPs*. Em cada *milestone* (marco), em que é entregue uma etapa do projeto de implantação, é medido o *CSAT* dos clientes. O *CSAT* de cada uma dessas etapas poderia ser incluso na lista de insumos que serão utilizados na confecção das variáveis que serão utilizadas nos modelos.

Finalmente, outros modelos preditivos poderiam ser testados na companhia com o intuito de melhorar a eficiência das previsões de *Customer Churn*.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBOTT, L. Customer Health Score: advice from three customer success experts. **Wootric**, 26 mar. 2017. Disponível em: https://www.wootric.com/blog/customer-health-score-advice-from-three-customer-success-experts/. Acesso em: 15 set. 2019.

AKAY, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, i. 2. 2009.

ALBERTIN, A. L.; ALBERTIN, R. M. M., **Tecnologia de informação e desempenho empresarial: as dimensões de seu uso e sua relação com os benefícios do negócio**.p.52, São Paulo: Editora Atlas S.A., 2016.

ARMBRUST, M.; STOICA, I.; ZAHARIA, M.; FOX, A.; GRIFFITH, R.; JOSEPH, A. D., RABKIN, A. A view of cloud computing. **Communications of the ACM**, n. 53, i. 4, p. 50-58, 2010. Disponível em: http://delivery.acm.org/10.1145/1730000/1721672/p50-armbrust.pdf?ip=179.93.191.50&id=1721672&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1568767482_41826f3f68701386f871081069c327e9. Acesso em: 15 set. 2019.

BATISTA, M. R. S. **A utilização de algoritmos de aprendizado de máquina em problemas de classificação**. 2019. Dissertação (Mestrado Profissional em Matemática, Estatística e Ciências da Computação Aplicadas à Indústria) − Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, *Campus* São Carlos, São Carlos, 2019.

BENLIAN, A.; KOUFARIS, M.; HESS, T. Service quality in Software-as-a-Service: developing the SaaS-Qual measure and examining its role in usage continuance. **Journal of Management Information Systems**, v. 28, i. 3, p. 85−126, 2011. Disponível em: https://pdfs.semanticscholar.org/ae88/a9ef29279761401ff644e6d01553d3ff54b2.pdf. Acesso em: 15 set. 2019.

BENLIAN, A.; KOUFARIS, M.; HESS, T. The role of Saas service quality for continued Saas use: empirical insights from Saas using firms. In: THIRTY FIRST INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, St. Louis, 2010. **[Proceedings]**. Disponível em: https://www.researchgate.net/publication/221599026_The_role_of_SAAS_service_quality_for_continued_SAAS_use_Empirical_insights_from_SAAS_using_firms. Acesso em: 15 set. 2019.

BERNAZZANI, S. **The 6 Stages of the SaaS Customer Lifecycle.** Disponível em: https://blog.hubspot.com/service/saas-customer-lifecycle. Acesso em: 15 set. 2019.

BISHOP, C. **Pattern recognition and machine learning.** Berlin: Springer, 2006.

BLATTBERG, R. C.; KIM, B. D.; NESLIN, S. A. **Database marketing:** analyzing and managing customers. New York: Springer, 2010.

BOLTON, R. N. A dynamic model of the duration of the customer's relationship with a continuous service provider: The Role of Satisfaction. **Marketing Science**, n.17, i. 1, p. 45-65, 1998.

CHEN, T. AND GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. InProceedings of the 22Nd ACM SIGKDD International Conference on Knowl- edge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

CHOU, D.; CHOU, A. Analysis of a new information systems outsourcing practice: *Software*-as-a-service business model. **International Journal of Information Systems and Change Management**, n. 2, i. 4, p. 392-405, 2007.

CHURAKOVA, I.; MIKHRAMOVA, R. *Software* as a Service: Study and Analysis of SaaS Business Model and Innovation Ecosystems. **Faculteit Economie En Bedrijfskunde**, p. 1–130, 2010.

DEMPSEY, D.; KELLIHER, F. **Industry trends in cloud computing: alternative Business-to-Business revenue models**. Cham: Springer International Publishing, 2018.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina. 1st. ed. [S.l.]: LTC, 2011. ISBN 8521618808,9788521618805.

FALVEY, P. M. *et al. Software* as a Service : A Primer on SaaS Metrics & Benchmarks Shea & Company Overview People ▪ Industry Expertise ▪ Process Excellence. n. December 2017, 2018.

FARIS, H. & AL-SHBOUL, B. & GHATASHEH, N. (2014). A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry. Computational Collective Intelligence. Technologies and Applications: 6th International Conference, ICCCI 2014, Seoul, Korea, September 24-26, 2014. Proceedings (pp.353-362). DOI. 10.1007/978-3-319-11289-3_36.

FARQUAD, M.; RAVI, V.; RAJU, S. B. Churn prediction using comprehensible support vector machine: an analytical CRM application. **Applied Soft Computing**, v. 19, i. 1, p. 31-40, 2014. Disponível em: http://dx.doi.org/10.1016/j.asoc.2014.01.031. Acesso em: 15 set. 2019.

FINTECHLAB. 8ª edição do Radar Fintechlab registra mais de 600 iniciativas. **Fintechlab**. Disponível em: https://fintechlab.com.br/index.php/2019/06/12/8a-edicao-do-radar-fintechlab-registra-mais-de-600-iniciativas/. Acesso em 19 set 2019.

FONTANELLA, C. How to calculate customer lifetime value. **HubSpot**. Disponível em: https://blog.hubspot.com/service/how-to-calculate-customer-lifetime-value. Acesso em: 15 set. 2019.

MORGAN, B., **What Is Customer Experience?** Disponível em: https://www.forbes.com/sites/blakemorgan/2017/04/20/what-is-customer-experience-2/#6f84cf6970c2. Acesso em: 15 set. 2019.

FORRESTER CONSULTING, **Measuring Customer HealthTo Drive The Right Conversations**. Disponível em: http://www.gainsight.com/wpcms/wp-content/uploads/2014/07/WP-Forrester-3-Measuring-Customer-Health.pdf. Acesso em: 15 set. 2019.

FOSTER, I.; VASILIADIS, V.; TUECKE, S. *Software*-as-a-Service as a path to *software* sustainability. **Go.Sandbox.Globuscs.Info**, [s.d.].

FRANK, B. (RADFORD U.; PITTGES, J. (RADFORD U. Analyzing *Customer Churn* in the *Software* as a Service (SaaS) Industry. p. 1–8, 2009.

FREEZE, R. D.; KULKARNI, U. Knowledge management capability: defining knowledge assets. **Journal of Knowledge Management**, v. 11, i. 6, p. 94−109, 2007. Disponível em: https://doi.org/10.1108/13673270710832190. Acesso em: 15 set. 2019.

FREITAS, A. L. P.; FREITAS NETO, M. M. Assessing the service quality in *Software*-as-a-Service from the customers' perspective: a methodological approach and case of use. **Production**, v. 27, p. 1–13, 2017. Disponível em: http://www.scielo.br/pdf/prod/v27/0103-6513-prod-27-e20170020.pdf. Acesso em: 15 set. 2019.

FRIEDMAN, J.H. Greedy function approximation: a gradient boosting machine. Annals of statistics, (2001) 1189-1232.

GAO, Y. *et al.* A bi-level decision model for *Customer Churn* analysis. **Computational Intelligence**, v. 30, n. 3, p. 583–599, 2014.

GARTNER. Gartner forecasts worldwide public cloud revenue to grow 17.5 Percent in 2019. B.A. Katie Costello. **Gartner**, 2 abr. 2019. Disponível em: https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g. Acesso em: 15 set. 2019.

GE, Y.; HE, S.; XIONG, J.; BROWN, D. *Customer Churn* analysis for a Software-as-a-service company. In: SYSTEMS AND INFORMATION ENGINEERING DESIGN SYMPOSIUM (SIEDS),Virginia, USA, 28 April 2017. **[Proceedings]**, p. 106−111.

https://repositorio.ucs.br/xmlui/bitstream/handle/11338/1146/Tese%20Giancarlo%20Dal%20B%C3%B3.pdf?sequence=1&isAllowed=y GIANFORTE, G. (2012)

GOSCINSKI, A.; BROCK, M. Toward dynamic and attribute based publication, discovery and selection for cloud computing. **Future Generation Computer Systems**, n. 26, i. 7, p. 947-970, 2010. Disponível em: http://dx.doi.org/10.1016/j.future.2010.03.009. Acesso em: 15 set. 2019.

HADDEN, J. *et al.* Computer assisted *Customer Churn* management: State-of-the-art and future trends. **Computers and Operations Research**, v. 34, n. 10, p. 2902–2917, 2007.

HALVORSRUD, R.; KVALE, K.; FØLSTAD, A. Improving service quality through customer journey analysis. **Journal of Service Theory and Practice**, v. 26, n. 6, p. 840–867, 2016.

HANSOTIA, B. Gearing up for CRM: antecedents to successful implementation. **Journal of Database Marketing**, v. 10, i. 2, p. 121–132, 2002. Disponível em: https://doi.org/10.1057/palgrave.jdm.3240103. Acesso em: 15 set. 2019.

HASSOUNA, M.; TARHINI, A.; ELYAS, T.; ABOUTRAB, M. S. *Customer Churn* in mobile markets: a comparison of techniques. **International Business Research**, v. 8, n. 6, p. 224-237, 2015. Disponível em: https://pdfs.semanticscholar.org/dfcb/012b63e3549888a784c1ce13275df7666afa.pdf?_ga=2.191498986.1521246979.1568646812-178287196.1568646812. Acesso em: 15 set. 2019.

HIMALAYA GOHIYA, HARSH LOHIYA, KAILASH PATIDAR. 2018 A SURVEY OF XGBOOST SYSTEM. International Journal of Advanced Technology & Engineering Research (IJATER)

HOLBROOK, M. B.; HIRSCHMAN, E. C. The experiential aspects of consumption − consumer fantasies, feelings, and fun. **Journal of Consumer Research**, v. 9, p. 132-140, 1982. Disponível em: https://www.jstor.org/stable/2489122?seq=1#page_scan_tab_contents. Acesso em: 15 set. 2019.

HOSMER, D. W.; LEMESHOW, S.; STURDIVAN, R. X. **Applied logistic regression**. 2. ed. Nova Jersey: Wiley, 2000.

JAHROMI, A. T.; EWING, M. Maximizing the Profit of a Churn Management Campaign by Offering Customer-Specific Incentives. p. 2–4, 2014.

JAHROMI, A. T.; STAKHOVYCH, S.; EWING, M. Managing B2B *Customer Churn*, retention and profitability. **Industrial Marketing Management**, v. 43, n. 7, p. 1258–1268, 2014.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning with applications in R**. Berlim: Springer 2013

JANZER, A. **Subscription marketing: strategies for nurturing customers in a world of Churn**. 2.ed. Cuesta Park Consulting, 2017.

JU, J.; WANG, Y.; FU, J.; WU, J.; LIN, Z. Research on key technology in SaaS. In: 2010 INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND COGNITIVE INFORMATICS 7, Kuala Lumpur, Malaysia, September 2010. **[Proceedings]**.

KALTENCKER, N. *et al.* The disruptive potential of *software* as a service: Validation and application of an ex-ante methodology. In: THE 24TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS (ICIS). **[Proceedings]**, v. 4, p. 3808–3826, 2013.

KAPLAN, J. M. Software-as-a-Service myths. **Business Week on-line**, 17 abr. 2006. Disponível em: https://www.augusoft.net/news/*software*-as-service.htm. 2006. Acesso em: 15 set. 2019.

KARAHOCA, A.; KARAHOCA, D.; AYDIN, N. GSM churn management using an adaptive neuro-fuzzy inference system. In: THE 2007 INTERNATIONAL CONFERENCE ON INTELLIGENT PERVASIVE COMPUTING, Valencia, Spain, 2007. **[Proceedings]**, p. 323-326.

KRANZBÜHLER, A.; KLEIJNEN, M. H.; MORGAN, R. E.; TEERLING, M. The multilevel nature of *Customer Experience* research: an integrative review and research agenda. **International Journal of Management Reviews**, Bath, v. 20, i. 2, p. 433-456, 2018. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/ijmr.12140. Acesso em: 15 set. 2019.

Lee, J. Y., Lee, J. W., Cheun, D. W., & Kim, S. D. (2009, December 2-4). A quality model for evaluating Software-as-a-service in cloud computing. In Proceedings of the 2009 Seventh ACIS International Conference on Software Engineering Research, Management and Applications (pp. 261-266). Haikou, China: IEE.MALLYA, S. SaaS sales strategy. **Prudent Cloud**, jun. 2009. Disponível em: http://www.prudentcloud.com/saas/saas-sales-strategy-25062009. Acesso em: 15 jun. 2019.

MARSLAND, S. **Machine learning: an algorithmic perspective**. 2.ed. Londres: Chapman & Hall, 2014.

MEIRELLES, F. **30ª Pesquisa Anual do Uso de TI nas Empresas, 2019**. Disponível em: https://eaesp.fgv.br/sites/eaesp.fgv.br/files/pesti2019fgvciappt_2019.pdf. Acesso em: 15 jun. 2019.

MIGUÉIS, V. L.; van den POEL, D.; CAMANHO, A.S.; FALCÃO-E-CUNHA, J. Modeling partial *Customer Churn*: on the value of first product-category purchase sequences. 2012. **Expert Systems with Applications**, v. 39, i. 12, p. 11250-11256, 15 set. 2012. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417412005969. Acesso em: 15 set. 2019.

MOON, Y. Break free from the product life cycle. **Harvard Business Review**, s.v., p. 87-94, May 2005. Disponível em: https://hbr.org/2005/05/break-free-from-the-product-life-cycle. Acesso em: 15 set. 2019.

MUCH, M. How salesforce's Marc Benioff revolutionized the *software* industry. **Investor's Business Daily**, 2 nov. 2019. Disponível em: https://www.investors.com/news/management/leaders-and-success/how-salesforces-marc-benioff-revolutionized-the-*software*-industry/. Acesso em: 15 set. 2019.

NARANG, P.; HUDIARA, I. S.; KHURMI, S. Various software metrics aid for measuring cloud SAAS (Software-as-a-Service). **International Journal of Computer Science and Technology**, India, v. 5, i. 4, spl. 1, p. 61-63, 2014. Disponível em: http://ijcst.com/vol54/Spl1/13-Pallavi-Narang.pdf. Acesso em: 15 set. 2019.

NENONEN, S.; RASILA, H.; JUNNONEN, J. M. Customer Journey – a method to investigate user experience. **W111 Research Report Usability of Workplaces Phase 2**, n. Schmitt 1999, p. 54–63, 2008.

NIE, G.; ROWE, W.; ZHANG, L.; TIAN, Y.; SHI, Y. Credit card churn forecasting by logistic regression and decision tree. **Expert Systems with Applications**, v. 38, i. 12, p.15273-15285, 2011.

NISBET, R.; ELDER, J.; MINER, G. **Handbook of statistical analysis and data mining applications**. Cambridge, Massachusetts: Academic Press, 2009.

OPANASENKO, M. Customer journey in B2B SaaS business models. 2017. GORDINI, N.; VEGLIO, V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. **Industrial Marketing Management**, v. 62, p. 100–107, 2017. **5 Experience-Maps-Using-Customer-Journey-Maps-to-Improve-Customer-Experience -- important.pdf**. , [s.d.]. SUKOW, A. E. R.;

ORACLE, **Seven Power Lessons for Customer Experience Leaders**. Disponível em: http://www.oracle.com/us/corporate/acquisitions/rightnow/seven-power-lessons-wp-1502937.pdf. Acesso em: 15 set. 2019.

PASI, T.; JOONA, S. How to Sell SaaS: A Model for Main Factors of Marketing and Selling *Software*-as-a-Service. **Lecture Notes in Business Information Processing**, v. 80, n. October, p. 125–139, 2011.

RABOY, O. What is *Customer Health Score* and how do I build one for my business? **Totango**, 26 mar. 2013. Disponível em: https://www.totango.com/blog/2013/03/what-iscustomer-health-score-and-how-do-i-build-one-for-my-business/. Acesso em: 15 set. 2019.

RIFKIN, J. **The Age of Access:** The new culture of hypercapitalism, where all of life is a paid-for experience. New York: Putnam's Sons, 2000.

RINK, D. R.; SWAN, J. E. Product life cycle research: a literature review. **Journal of Business Research**, v. 7, i. 3, p. 219–242, 1979.

SATYANARAYANA, S. Cloud computing : SaaS. **GESJ:** Computer Science and Telecommunications, n. 4, i. 36, p. 76-79, 2012. Disponível em: https://fenix.tecnico.ulisboa.pt/downloadFile/1126518382178096/1986.pdf. Acesso em: 15 set. 2019.

SOUZA, E. G. de. Entendendo o que é Matriz de Confusão com Python. **Medium**, 27 mar. 2019. Disponível em: https://medium.com/data-hackers/entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509. Acesso em: 15 set. 2019.

SPRINGER, T. *et al.* Breaking the back of *Customer Churn*. **Bain & Company**, p. 01–08, 2014. BANDULET, F. *Software*-as-a-Service as Disruptive Innovation in the Enterprise Application Market: An Empirical Analysis of Revenue Growth and Profitability among SaaS Providers Working Paper. n. Setembro 2015.

SUKOW, A. E. R.; GRANT, R. Forecasting and the role of churn in *Software*-as-a-Service. **IBusiness**, n. 5, p. 49-57, 2013. Disponível em: https://www.researchgate.net/publication/301902794_Forecasting_and_the_Role_of_Churn_in_Software-as-a-Service_Business_Models. Acesso em: 15 set. 2019.

TYRVÄINEN, P.; SELIN, J. A model for main factors of marketing and selling Software-as-a-Service department of computer science and information systems. In: Regnell B.; van de WEERD, I., DE TROYER, O. (eds). **Software Business. Lecture Notes in Business Information Processing**, v. 80. Springer, Berlin, Heidelberg. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-21544-5_2. Acesso em: 15 set. 2019.

VERHOEF, P. C.; LEMON, K. N., PARASURAMAN, A.; ROGGEVEEN, A.; TSIROS, M.; SCHLESINGER, L. A. Customer experience creation: determinants, dynamics, and management strategies. **Journal of Retailing**, v. 85, p. 31– 41, 2009. Disponível em: https://www.rug.nl/staff/p.c.verhoef/jr_customer_experience.pdf. Acesso em: 15 set. 2019.

WALZ, J.; GRIER, D. A. Time to push the cloud. **IT Professional**, n. 12, i. 5, p. 14-16, 2010.

WHEELER, R.; ADVISOR, R. Cloud versus Traditional *Software* – From a business *software* developer's perspective. p. 1–8, 2019.

XIN, M.; LEVINA, N. Software-as-a service model: elaborating client-side adoption factors. In: THE 29TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, Paris, France, 2008. BOLAND, R.; LIMAYEM, M., PENTLAND, B. (eds.) **[Proceedings]**, p. 14-17, 2008.

YANG, J.; SUN, j.; ZHANG, Y.; WANG, W. Understanding SaaS adoption from the perspective of organizational users: A tripod readiness model. **Computers in Human Behavior**, n. 45, p. 254-264, 2015.

YOUSSEF, L.; BUTRICO, M.; DA SILVA, D. Toward a unified ontology of cloud computing. In: 2008 GRID COMPUTING ENVIRONMENTS WORKSHOP, Austin, TX, USA, 12-16 nov. 2008. **[Greetings]**, p. 1-10.

ZHANG, Q.; CHENG, L.; BOUTABA, R. Cloud computing: state-of-the-art and research challenges. **Journal of Internet Services and Applications**, n. 1, i. 1, p. 7-18, 2010. Disponível em: http://dx.doi.org/10.1007/s13174-010-0007-6. Acesso em: 15 set. 2019.

ZUORA. 2017 Subscription Economy ® Index™. **Zuora**, 2017. Disponível em: https://www.zuora.com/wp-content/uploads/2017/06/Zuora-2017-Subscription-Economy-Index.pdf. Acesso em: 15 set. 2017.

WEIZENG WANG, YULIANG SHI, GAOFAN LYU and WANGHUA DENG 2017 Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform 2nd International Conference on Artificial Intelligence and Engineering Applications (AIEA 2017) ISBN: 978-1-60595-485-1

APÊNDICE A - VARIÁVEIS CANDIDATAS CRIADAS[7]

# Variável	Nome da Variável	Utilizada
1	bow_acesso_last_7_months	Sim
2	bow_alteracao_last_7_months	Sim
3	bow_alterar_last_7_months	Sim
4	bow_ambiente_last_7_months	Sim
5	bow_arquivo_last_7_months	Sim
6	bow_atendimento_last_7_months	Sim
7	bow_atualizacao_last_7_months	Sim
8	bow_baixa_last_7_months	Sim
9	bow_banco_last_7_months	Sim
10	bow_base_last_7_months	Sim
11	bow_cadastro_last_7_months	Sim
12	bow_calculo_last_7_months	Sim
13	bow_campo_last_7_months	Sim
14	bow_cancelamento_last_7_months	Sim
15	bow_chamado_last_7_months	Sim
16	bow_chat_last_7_months	Sim
17	bow_cliente_last_7_months	Sim
18	bow_codigo_last_7_months	Sim
19	bow_consulta_last_7_months	Sim
20	bow_conta_last_7_months	Sim
21	bow_contabil_last_7_months	Sim
22	bow_contas_last_7_months	Sim
23	bow_contrato_last_7_months	Sim
24	bow_criacao_last_7_months	Sim
25	bow_dados_last_7_months	Sim
26	bow_data_last_7_months	Sim
27	bow_documento_last_7_months	Sim
28	bow_duvida_last_7_months	Sim
29	bow_entrada_last_7_months	Sim
30	bow_envio_last_7_months	Sim

(continua)

[7] **Nota:** A tabela indica quais *variáveis* foram utilizadas e quais foram desconsideradas nas análises. Alguns nomes das *variáveis* apresentam aparente erro no nome devido ao uso de dados descritos com "ç" ou letras acentuadas que não compões o alfabeto em inglês.

# Variável	Nome da Variável	Utilizada
31	bow_erro_last_7_months	Sim
32	bow_error_last_7_months	Sim
33	bow_esocial_last_7_months	Sim
34	bow_evento_last_7_months	Sim
35	bow_exclusao_last_7_months	Sim
36	bow_falha_last_7_months	Sim
37	bow_faturamento_last_7_months	Sim
38	bow_ferias_last_7_months	Sim
39	bow_financeiro_last_7_months	Sim
40	bow_fiscal_last_7_months	Sim
41	bow_fluig_last_7_months	Sim
42	bow_folha_last_7_months	Sim
43	bow_geracao_last_7_months	Sim
44	bow_gerar_last_7_months	Sim
45	bow_icms_last_7_months	Sim
46	bow_importacao_last_7_months	Sim
47	bow_inclusao_last_7_months	Sim
48	bow_integracao_last_7_months	Sim
49	bow_lancamento_last_7_months	Sim
50	bow_liberacao_last_7_months	Sim
51	bow_log_last_7_months	Sim
52	bow_ltda_last_7_months	Sim
53	bow_mail_last_7_months	Sim
54	bow_mata_last_7_months	Sim
55	bow_mensagem_last_7_months	Sim
56	bow_nf_last_7_months	Sim
57	bow_nfe_last_7_months	Sim
58	bow_nota_last_7_months	Sim
59	bow_notas_last_7_months	Sim
60	bow_nova_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
61	bow_novo_last_7_months	Sim
62	bow_pagamento_last_7_months	Sim
63	bow_pedido_last_7_months	Sim
64	bow_ponto_last_7_months	Sim
65	bow_portal_last_7_months	Sim
66	bow_problema_last_7_months	Sim
67	bow_problemas_last_7_months	Sim
68	bow_processo_last_7_months	Sim
69	bow_producao_last_7_months	Sim
70	bow_produto_last_7_months	Sim
71	bow_protheus_last_7_months	Sim
72	bow_re_last_7_months	Sim
73	bow_registro_last_7_months	Sim
74	bow_reinf_last_7_months	Sim
75	bow_rejeicao_last_7_months	Sim
76	bow_relatorio_last_7_months	Sim
77	bow_res_last_7_months	Sim
78	bow_rescisao_last_7_months	Sim
79	bow_retorno_last_7_months	Sim
80	bow_rm_last_7_months	Sim
81	bow_rotina_last_7_months	Sim
82	bow_saldo_last_7_months	Sim
83	bow_servico_last_7_months	Sim
84	bow_sistema_last_7_months	Sim
85	bow_social_last_7_months	Sim
86	bow_solicitacao_last_7_months	Sim
87	bow_sped_last_7_months	Sim
88	bow_tabela_last_7_months	Sim
89	bow_taf_last_7_months	Sim
90	bow_teste_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
91	bow_ticket_last_7_months	Sim
92	bow_totvs_last_7_months	Sim
93	bow_transferencia_last_7_months	Sim
94	bow_tss_last_7_months	Sim
95	bow_urgente_last_7_months	Sim
96	bow_usuario_last_7_months	Sim
97	bow_valor_last_7_months	Sim
98	bow_venda_last_7_months	Sim
99	bow_versao_last_7_months	Sim
100	bow_xml_last_7_months	Sim
101	carencia_span	Sim
102	cf_score_last_2_mean	Não
103	cf_score_last_3_mean	Não
104	cf_score_last_4_mean	Não
105	cf_score_last_5_mean	Não
106	cf_score_last_6_mean	Não
107	cf_score_last_7_mean	Não
108	cf_score_last_8_mean	Não
109	cf_score_minus_1	Não
110	cf_score_minus_2	Não
111	cf_score_minus_3	Não
112	cf_score_minus_4	Não
113	cf_score_minus_5	Não
114	cf_score_minus_6	Não
115	cf_score_minus_7	Não
116	cf_score_minus_8	Não
117	churn_rate_for_same_franquia_last_1_months	Sim
118	churn_rate_for_same_franquia_last_2_months	Sim
119	churn_rate_for_same_franquia_last_3_months	Sim
120	churn_rate_for_same_franquia_last_6_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
121	churn_rate_for_same_macro_segment_last_1_months	Sim
122	churn_rate_for_same_macro_segment_last_2_months	Sim
123	churn_rate_for_same_macro_segment_last_3_months	Sim
124	churn_rate_for_same_macro_segment_last_6_months	Sim
125	churn_rate_for_same_product_last_1_months	Sim
126	churn_rate_for_same_product_last_2_months	Sim
127	churn_rate_for_same_product_last_3_months	Sim
128	churn_rate_for_same_product_last_6_months	Sim
129	churn_rate_for_same_region_last_1_months	Sim
130	churn_rate_for_same_region_last_2_months	Sim
131	churn_rate_for_same_region_last_3_months	Sim
132	churn_rate_for_same_region_last_6_months	Sim
133	churn_rate_for_same_segment_last_1_months	Sim
134	churn_rate_for_same_segment_last_2_months	Sim
135	churn_rate_for_same_segment_last_3_months	Sim
136	churn_rate_for_same_segment_last_6_months	Sim
137	churn_rate_for_same_sub_segment_last_1_months	Sim
138	churn_rate_for_same_sub_segment_last_2_months	Sim
139	churn_rate_for_same_sub_segment_last_3_months	Sim
140	churn_rate_for_same_sub_segment_last_6_months	Sim
141	churn_sum_12_periods	Sim
142	churn_sum_6_periods	Sim
143	client_embed_0	Não
144	client_embed_1	Não
145	client_embed_10	Não
146	client_embed_11	Não
147	client_embed_12	Não
148	client_embed_13	Não
149	client_embed_14	Não
150	client_embed_15	Não

(continua)

# Variável	Nome da Variável	Utilizada
151	client_embed_16	Não
152	client_embed_17	Não
153	client_embed_18	Não
154	client_embed_19	Não
155	client_embed_2	Não
156	client_embed_20	Não
157	client_embed_21	Não
158	client_embed_22	Não
159	client_embed_23	Não
160	client_embed_24	Não
161	client_embed_3	Não
162	client_embed_4	Não
163	client_embed_5	Não
164	client_embed_6	Não
165	client_embed_7	Não
166	client_embed_8	Não
167	client_embed_9	Não
168	days_to_carencia_end	Sim
169	days_to_carencia_start	Sim
170	month	Sim
171	product_embed_0	Não
172	product_embed_1	Não
173	product_embed_10	Não
174	product_embed_11	Não
175	product_embed_12	Não
176	product_embed_13	Não
177	product_embed_14	Não
178	product_embed_15	Não
179	product_embed_16	Não
180	product_embed_17	Não

(continua)

# Variável	Nome da Variável	Utilizada
181	product_embed_18	Não
182	product_embed_19	Não
183	product_embed_2	Não
184	product_embed_20	Não
185	product_embed_21	Não
186	product_embed_22	Não
187	product_embed_23	Não
188	product_embed_24	Não
189	product_embed_3	Não
190	product_embed_4	Não
191	product_embed_5	Não
192	product_embed_6	Não
193	product_embed_7	Não
194	product_embed_8	Não
195	product_embed_9	Não
196	sk_cliente_old_nunique_12_periods	Sim
197	sk_cliente_old_nunique_6_periods	Sim
198	sk_prod_nunique_12_periods	Sim
199	sk_prod_nunique_6_periods	Sim
200	ticket_grp_Bemacash_last_7_months	Sim
201	ticket_grp_ChefCorporativoN1_last_7_months	Sim
202	ticket_grp_CloudAtendimentoSPÂ\xa0_last_7_months	Sim
203	ticket_grp_CloudBancodeDadosÂ\xa0_last_7_months	Sim
204	ticket_grp_CloudPrimeClientesJV_last_7_months	Sim
205	ticket_grp_CloudPrimeProtheusSP_last_7_months	Sim
206	ticket_grp_CloudSustentaÃ§Ã£oJVÂ\xa0_last_7_months	Sim
207	ticket_grp_CloudTSS_last_7_months	Sim
208	ticket_grp_CMNETAtlÃ¢ntica_last_7_months	Sim
209	ticket_grp_CMNETBackOffice_last_7_months	Sim
210	ticket_grp_CMNETFrontOffice_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
211	ticket_grp_CMNETLATAMFrontOffice_last_7_months	Sim
212	ticket_grp_CMNETPDV_last_7_months	Sim
213	ticket_grp_ConsultoriaDatasul_last_7_months	Sim
214	ticket_grp_CST-Atendimento_last_7_months	Sim
215	ticket_grp_CST-CadastroeAssessoriaPortal_last_7_months	Sim
216	ticket_grp_CST-CobranÃ§aFLY01_last_7_months	Sim
217	ticket_grp_CST-CobranÃ§a_last_7_months	Sim
218	ticket_grp_CST-ContratosdeVendas_last_7_months	Sim
219	ticket_grp_CST-ContratosMPN_last_7_months	Sim
220	ticket_grp_CST-Faturamento_last_7_months	Sim
221	ticket_grp_CST-PMOLocal_last_7_months	Sim
222	ticket_grp_CST-Senhas_last_7_months	Sim
223	ticket_grp_CST-Viagens_last_7_months	Sim
224	ticket_grp_CSTâ€“ContratosMPN_last_7_months	Sim
225	ticket_grp_DatasulControladoriaÂ\xa0_last_7_months	Sim
226	ticket_grp_DatasulEstoqueeCustos_last_7_months	Sim
227	ticket_grp_DatasulFaturamento_last_7_months	Sim
228	ticket_grp_DatasulFinanÃ§asÂ\xa0_last_7_months	Sim
229	ticket_grp_DatasulFiscalÂ\xa0_last_7_months	Sim
230	ticket_grp_DatasulFrameworkÂ\xa0_last_7_months	Sim
231	ticket_grp_DatasulManufaturaÂ\xa0_last_7_months	Sim
232	ticket_grp_DatasulMateriaisÂ\xa0_last_7_months	Sim
233	ticket_grp_DatasulNotaFiscalEletrÃ´nica_last_7_months	Sim
234	ticket_grp_DatasulRecursosHumanosÂ\xa0_last_7_months	Sim
235	ticket_grp_DatasulSaÃºdeÂ\xa0_last_7_months	Sim
236	ticket_grp_DatasulTecnologia_last_7_months	Sim
237	ticket_grp_EMAILNOTIFICAÃ‡Ãƒ*O_last_7_months	Sim
238	ticket_grp_ESocialTotvs_last_7_months	Sim
239	ticket_grp_Fly01Manufatura_last_7_months	Sim
240	ticket_grp_Fly01SaÃºde-PersonalMedOnline_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
241	ticket_grp_FLY01SAÃšDE_last_7_months	Sim
242	ticket_grp_Fly01VarejoÂ\xa0_last_7_months	Sim
243	ticket_grp_GCADSP_last_7_months	Sim
244	ticket_grp_LogixFinanÃ§as_last_7_months	Sim
245	ticket_grp_LogixVendaseNFe_last_7_months	Sim
246	ticket_grp_other_last_7_months	Sim
247	ticket_grp_OuvidoriaTotvs_last_7_months	Sim
248	ticket_grp_PIMSAgroindÃºstriaÂ\xa0_last_7_months	Sim
249	ticket_grp_PrimeDatasul_last_7_months	Sim
250	ticket_grp_PrimeProtheus_last_7_months	Sim
251	ticket_grp_PrimeRM_last_7_months	Sim
252	ticket_grp_ProtheusADVPL_last_7_months	Sim
253	ticket_grp_ProtheusAtivoeContÃ¡bil_last_7_months	Sim
254	ticket_grp_ProtheusAutomaÃ§Ã£oComercial_last_7_months	Sim
255	ticket_grp_ProtheusAutomaÃ§Ã£oFiscalÂ\xa0_last_7_months	Sim
256	ticket_grp_ProtheusCompras/GCP/GCT_last_7_months	Sim
257	ticket_grp_ProtheusEstoque/Custos/ACD/DCL_last_7_months	Sim
258	ticket_grp_ProtheusFaturamentoeCRM_last_7_months	Sim
259	ticket_grp_ProtheusFinanceiroePlanejamento_last_7_months	Sim
260	ticket_grp_ProtheusFiscalArquivoMagnÃ©tico_last_7_months	Sim
261	ticket_grp_ProtheusFiscalÂ\xa0_last_7_months	Sim
262	ticket_grp_ProtheusFramework_last_7_months	Sim
263	ticket_grp_ProtheusManufatura_last_7_months	Sim
264	ticket_grp_ProtheusMateriais_last_7_months	Sim
265	ticket_grp_ProtheusMercadoInternacionalÂ\xa0_last_7_months	Sim
266	ticket_grp_ProtheusRecursosHumanosJOIÂ\xa0_last_7_months	Sim
267	ticket_grp_ProtheusRecursosHumanosÂ\xa0_last_7_months	Sim
268	ticket_grp_ProtheusReinf_last_7_months	Sim
269	ticket_grp_ProtheusServiÃ§os_last_7_months	Sim
270	ticket_grp_ProtheusSPEDÂ\xa0_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
271	ticket_grp_ProtheusTMS_last_7_months	Sim
272	ticket_grp_ProtheusTradePadrÃ£o_last_7_months	Sim
273	ticket_grp_RMBI_last_7_months	Sim
274	ticket_grp_RMC&P_last_7_months	Sim
275	ticket_grp_RMControladoria_last_7_months	Sim
276	ticket_grp_RMEducacional_last_7_months	Sim
277	ticket_grp_RMeSocial_last_7_months	Sim
278	ticket_grp_RMFinanÃ§asÂ\xa0_last_7_months	Sim
279	ticket_grp_RMFramework_last_7_months	Sim
280	ticket_grp_RMGestÃ£oRH_last_7_months	Sim
281	ticket_grp_RMLabore_last_7_months	Sim
282	ticket_grp_RMMateriaisÂ\xa0_last_7_months	Sim
283	ticket_grp_RMTecnologia_last_7_months	Sim
284	ticket_grp_RMSSuporte_last_7_months	Sim
285	ticket_grp_ServiÃ§osCorporativos-Terceiros_last_7_months	Sim
286	ticket_grp_SQUAD1CLOUDPROTHEUS_last_7_months	Sim
287	ticket_grp_SQUAD2CLOUDPROTHEUS_last_7_months	Sim
288	ticket_grp_SQUAD3CLOUDPROTHEUS_last_7_months	Sim
289	ticket_grp_SQUADCLOUDRM_last_7_months	Sim
290	ticket_grp_TAFe-SocialProtheus_last_7_months	Sim
291	ticket_grp_TAFe-SocialRM_last_7_months	Sim
292	ticket_grp_TAFeSOCIAL/REINF/TSS_last_7_months	Sim
293	ticket_grp_TIADMVENDAS(CRM)_last_7_months	Sim
294	ticket_grp_TIINFRAFIELDSERVICE_last_7_months	Sim
295	ticket_grp_TIINFRASERVICEDESK_last_7_months	Sim
296	ticket_grp_TIPlataformasDigitais_last_7_months	Sim
297	ticket_grp_VirtualAgeFaturamento_last_7_months	Sim
298	ticket_grp_VirtualAgeFinanceiro_last_7_months	Sim
299	ticket_grp_VirtualAgeFiscal_last_7_months	Sim
300	ticket_grp_VirtualAgeVarejo_last_7_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
301	tickets_of_priority_high_last_1_months	Sim
302	tickets_of_priority_high_last_2_months	Sim
303	tickets_of_priority_high_last_3_months	Sim
304	tickets_of_priority_high_last_6_months	Sim
305	tickets_of_priority_low_last_1_months	Sim
306	tickets_of_priority_low_last_2_months	Sim
307	tickets_of_priority_low_last_3_months	Sim
308	tickets_of_priority_low_last_6_months	Sim
309	tickets_of_priority_normal_last_1_months	Sim
310	tickets_of_priority_normal_last_2_months	Sim
311	tickets_of_priority_normal_last_3_months	Sim
312	tickets_of_priority_normal_last_6_months	Sim
313	tickets_of_priority_urgent_last_1_months	Sim
314	tickets_of_priority_urgent_last_2_months	Sim
315	tickets_of_priority_urgent_last_3_months	Sim
316	tickets_of_priority_urgent_last_6_months	Sim
317	tickets_of_type_incident_last_1_months	Sim
318	tickets_of_type_incident_last_2_months	Sim
319	tickets_of_type_incident_last_3_months	Sim
320	tickets_of_type_incident_last_6_months	Sim
321	tickets_of_type_problem_last_1_months	Sim
322	tickets_of_type_problem_last_2_months	Sim
323	tickets_of_type_problem_last_3_months	Sim
324	tickets_of_type_problem_last_6_months	Sim
325	tickets_of_type_question_last_1_months	Sim
326	tickets_of_type_question_last_2_months	Sim
327	tickets_of_type_question_last_3_months	Sim
328	tickets_of_type_question_last_6_months	Sim
329	tickets_of_type_task_last_1_months	Sim
330	tickets_of_type_task_last_2_months	Sim

(continua)

# Variável	Nome da Variável	Utilizada
331	tickets_of_type_task_last_3_months	Sim
332	tickets_of_type_task_last_6_months	Sim
333	time_alive	Sim
334	total_contracts_with_same_franquia_last_1_months	Sim
335	total_contracts_with_same_franquia_last_2_months	Sim
336	total_contracts_with_same_franquia_last_3_months	Sim
337	total_contracts_with_same_franquia_last_6_months	Sim
338	total_contracts_with_same_macro_segment_last_1_months	Sim
339	total_contracts_with_same_macro_segment_last_2_months	Sim
340	total_contracts_with_same_macro_segment_last_3_months	Sim
341	total_contracts_with_same_macro_segment_last_6_months	Sim
342	total_contracts_with_same_product_last_1_months	Sim
343	total_contracts_with_same_product_last_2_months	Sim
344	total_contracts_with_same_product_last_3_months	Sim
345	total_contracts_with_same_product_last_6_months	Sim
346	total_contracts_with_same_region_last_1_months	Sim
347	total_contracts_with_same_region_last_2_months	Sim
348	total_contracts_with_same_region_last_3_months	Sim
349	total_contracts_with_same_region_last_6_months	Sim
350	total_contracts_with_same_segment_last_1_months	Sim
351	total_contracts_with_same_segment_last_2_months	Sim
352	total_contracts_with_same_segment_last_3_months	Sim
353	total_contracts_with_same_segment_last_6_months	Sim
354	total_contracts_with_same_sub_segment_last_1_months	Sim
355	total_contracts_with_same_sub_segment_last_2_months	Sim
356	total_contracts_with_same_sub_segment_last_3_months	Sim
357	total_contracts_with_same_sub_segment_last_6_months	Sim
358	vlr_bonifi_max_12_periods	Sim
359	vlr_bonifi_max_6_periods	Sim
360	vlr_bonifi_mean_12_periods	Sim

(continua)

# Variável	Nome da Variável	Utilizada
361	vlr_bonifi_mean_6_periods	Sim
362	vlr_bonifi_std_12_periods	Sim
363	vlr_bonifi_std_6_periods	Sim
364	vlr_bonifi_sum_12_periods	Sim
365	vlr_bonifi_sum_6_periods	Sim
366	vlr_cancel_max_12_periods	Sim
367	vlr_cancel_max_6_periods	Sim
368	vlr_cancel_mean_12_periods	Sim
369	vlr_cancel_mean_6_periods	Sim
370	vlr_cancel_std_12_periods	Sim
371	vlr_cancel_std_6_periods	Sim
372	vlr_cancel_sum_12_periods	Sim
373	vlr_cancel_sum_6_periods	Sim
374	vlr_nvvend_max_12_periods	Sim
375	vlr_nvvend_max_6_periods	Sim
376	vlr_nvvend_mean_12_periods	Sim
377	vlr_nvvend_mean_6_periods	Sim
378	vlr_nvvend_std_12_periods	Sim
379	vlr_nvvend_std_6_periods	Sim
380	vlr_nvvend_sum_12_periods	Sim
381	vlr_nvvend_sum_6_periods	Sim
382	vlr_reajuste_max_12_periods	Sim
383	vlr_reajuste_max_6_periods	Sim
384	vlr_reajuste_mean_12_periods	Sim
385	vlr_reajuste_mean_6_periods	Sim
386	vlr_reajuste_std_12_periods	Sim
387	vlr_reajuste_std_6_periods	Sim
388	vlr_reajuste_sum_12_periods	Sim
389	vlr_reajuste_sum_6_periods	Sim

Fonte: elaboração própria.

APÊNDICE B - TRECHOS DE CÓDIGO EM PYTHON – ANÁLISE INICIAL DOS DADOS

In [13]:

```
import pandas as pd

import pandas as pd

% matplotlib inline


df = pd.read_csv('Relatorio Empodera Clientes - 08_05_2019 - 08_44_40.csv', sep = "
;", low_memory = False)

pd.set_option('display.max_columns', 5)

df.head()
```

Out[13]:

	CodT	Nome	...	_Valor_bruto_recorrente_mantido	_Valor_revertido
0	99000	TOTVS S.A.	...	NaN	NaN
1	99014	TOTVS S.A.	...	NaN	NaN
2	99025	TOTVS S.A.	...	NaN	NaN
3	99095	EURO TOTVS	...	NaN	NaN
4	99699	CST SENHAS	...	NaN	NaN

5 rows × 35 columns

In [14]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34945 entries, 0 to 34944
Data columns (total 35 columns):
CodT                                34945 non-null object
Nome                                34945 non-null object
CNPJ/CPF                            33250 non-null object
Status                              34945 non-null object
Healthscore                         34945 non-null float64
NPS                                 34945 non-null float64
Situacao                            149 non-null object
Segmento                            34863 non-null object
Unidade                             33179 non-null object
Momento                             0 non-null float64
Ambiente                            0 non-null float64
SaaS_contratado                     33323 non-null object
SaaS_pagante                        33323 non-null object
CDU                                 33323 non-null object
Qtd_licencas                        33323 non-null object
Contratos_ativos                    33323 non-null float64
Ultima_nota_NPS                     0 non-null float64
Campanha_NPS                        0 non-null float64
Assinatura_Contrato                 33176 non-null object
Inserção_Empodera                   34945 non-null object
Data_Cancelamento                   17 non-null object
Agentes                             2 non-null object
_Participantes                      0 non-null float64
_Sponsor                            0 non-null float64
```

```
Sponsor_E-mail                      0 non-null float64
Sponsor_Telefone                    0 non-null float64
ESN                                 33385 non-null object
ESN_E-mail                          28411 non-null object
ESN_Telefone                        21287 non-null object
GSN                                 32305 non-null object
GSN_E-mail                          31206 non-null object
GSN_Telefone                        20372 non-null object
_Outro                              0 non-null float64
_Valor_bruto_recorrente_mantido     0 non-null float64
_Valor_revertido                    0 non-null float64
dtypes: float64(14), object(21)
memory usage: 9.3+ MB
```

In [15]:

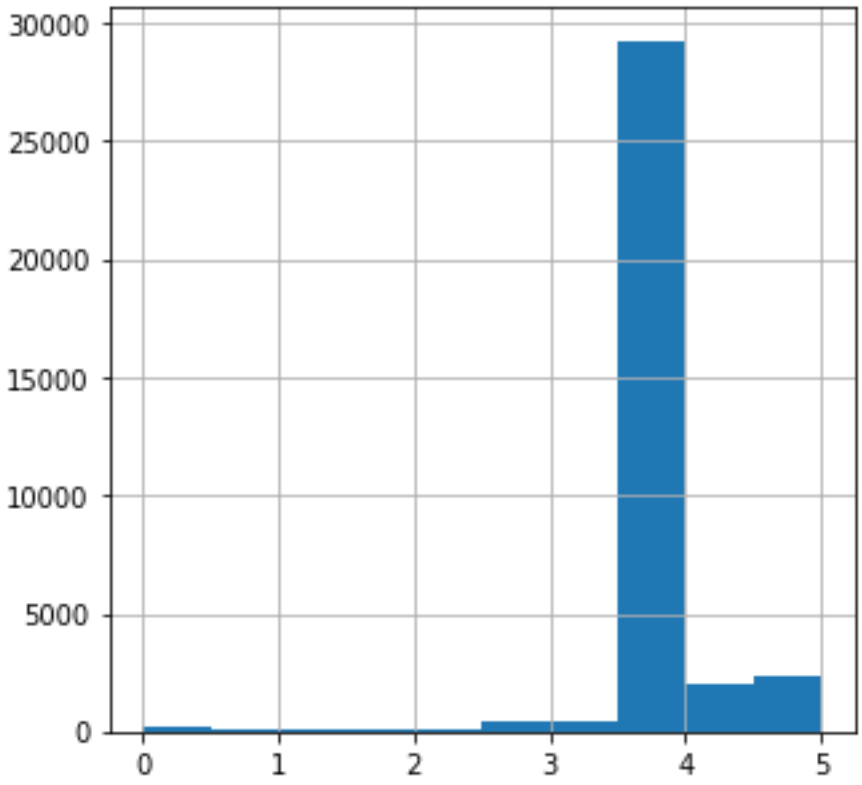
```
sum(df.duplicated())
```

Out[15]:

```
0
```

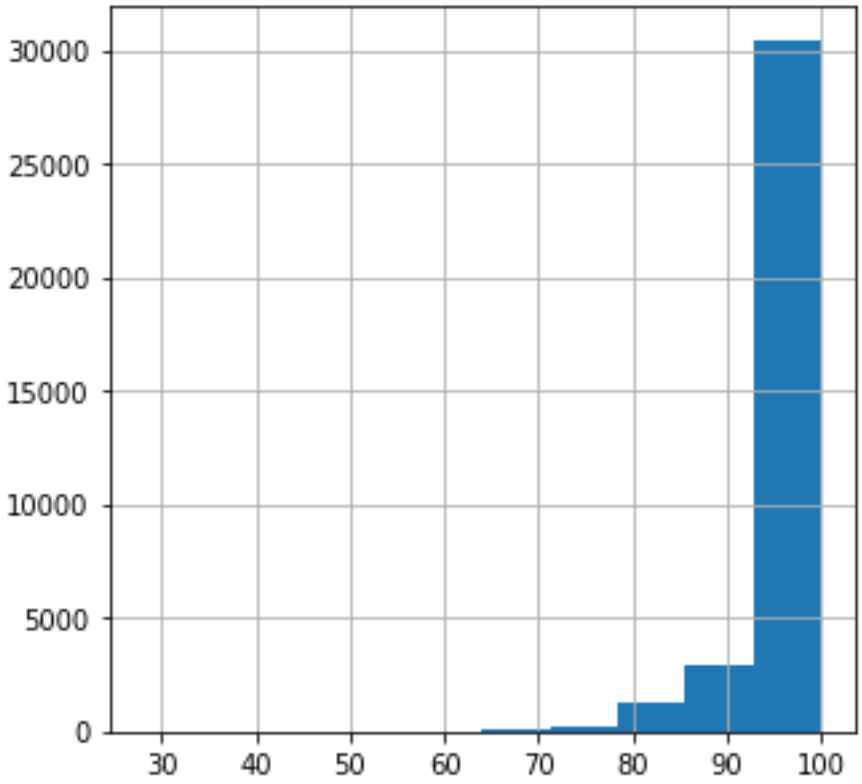
In [24]:

```
df['NPS'].hist(figsize=(5,5));
```



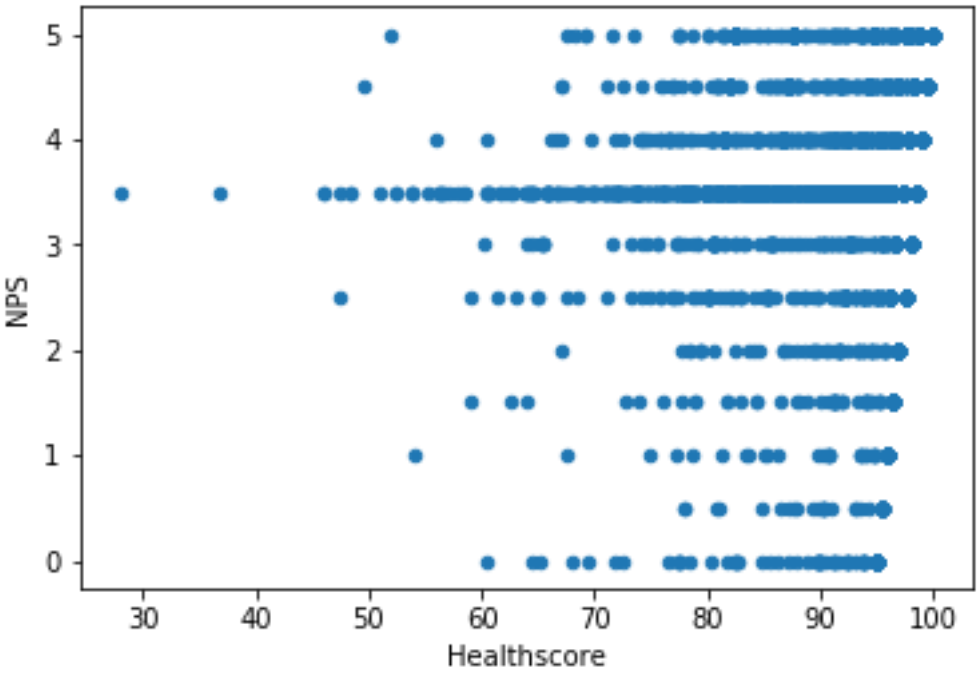
In [25]:

```
df['Healthscore'].hist(figsize=(5,5));
```



In [18]:

```
df.plot(x='Healthscore', y= 'NPS', kind= 'scatter');
```



In [19]:

```
df.describe()
```

Out[19]:

	Healthscore	NPS	...	_Valor_bruto_recorrente_mantido	_Valor_revertido
count	34945.000000	34945.000000	...	0.0	0.0
mean	96.227808	3.556160	...	NaN	NaN
std	5.127429	0.487687	...	NaN	NaN
min	28.000000	0.000000	...	NaN	NaN
25%	97.250000	3.500000	...	NaN	NaN
50%	98.500000	3.500000	...	NaN	NaN
75%	98.500000	3.500000	...	NaN	NaN
max	100.000000	5.000000	...	NaN	NaN

8 rows × 14 columns

In [20]:

```
df.groupby('Situacao').mean().NPS
```

Out[20]:

```
Situacao
Solicitação de Cancelamento            3.389610
Solicitação de Cancelamento Parcial    3.500000
Solicitação de Cancelamento Total      3.595745
Name: NPS, dtype: float64
```

In [21]:

```
df.groupby('Situacao').mean().Healthscore
```

Out[21]:

```
Situacao
Solicitação de Cancelamento            80.470779
Solicitação de Cancelamento Parcial    84.470000
Solicitação de Cancelamento Total      73.542553
Name: Healthscore, dtype: float64
```

In [22]:

```
df.groupby('Situacao').mean().Contratos_ativos
```

Out[22]:

```
Situacao
Solicitação de Cancelamento            16.434211
Solicitação de Cancelamento Parcial    17.040000
Solicitação de Cancelamento Total      24.425532
Name: Contratos_ativos, dtype: float64
```

In [23]:

```
df.groupby('Situacao').mean()
```

Out[23]:

	Healthscore	NPS	...	_Valor_bruto_recorrente_mantido	_Valor_revertido
Situacao					
Solicitação de Cancelamento	80.470779	3.389610	...	NaN	NaN
Solicitação de Cancelamento Parcial	84.470000	3.500000	...	NaN	NaN

	Healthscore	NPS	...	_Valor_bruto_recorrente_mantido	_Valor_revertido
Situacao					
Solicitação de Cancelamento Total	73.542553	3.595745	...	NaN	NaN

3 rows × 14 columns

APÊNDICE C - TRECHOS DE CÓDIGO EM PYTHON – SELEÇÃO DE VARIÁVEIS

In [9]:

```
import pandas as pd
from scipy.stats import zscore, ks_2samp
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV, RFE
from sklearn import linear_model
from sklearn.model_selection import StratifiedKFold
import pickle
```

Read in the dataset containing the variables.

In [10]:

```
df = pd.read_parquet("Tabela_Referencias@20190401.parquet")
numerics = ['int32', 'int64', 'float64', 'bool']
expert_variables = df
#df.select_dtypes(include=numerics)
remove_list=
['month','churn','cf_score_minus_8','cf_score_minus_7','cf_score_minus_6','cf_score_minus_5
','cf_score_minus_4','cf_score_minus_3','cf_score_minus_2','cf_score_minus_1','cf_score_last
_8_mean','cf_score_last_7_mean','cf_score_last_6_mean','cf_score_last_5_mean','cf_score_la
st_4_mean','cf_score_last_3_mean','cf_score_last_2_mean','product_embed_0','product_embe
d_1','product_embed_2','product_embed_3','product_embed_4','product_embed_5','product_e
mbed_6','product_embed_7','product_embed_8','product_embed_9','product_embed_10','prod
uct_embed_11','product_embed_12','product_embed_13','product_embed_14','product_embed
_15','product_embed_16','product_embed_17','product_embed_18','product_embed_19','produ
ct_embed_20','product_embed_21','product_embed_22','product_embed_23','product_embed_
24','client_embed_0','client_embed_1','client_embed_2','client_embed_3','client_embed_4','cli
ent_embed_5','client_embed_6','client_embed_7','client_embed_8','client_embed_9','client_e
mbed_10','client_embed_11','client_embed_12','client_embed_13','client_embed_14','client_e
mbed_15','client_embed_16','client_embed_17','client_embed_18','client_embed_19','client_e
mbed_20','client_embed_21','client_embed_22','client_embed_23','client_embed_24']
selected_columns = [i for i in expert_variables.columns if i not in remove_list]
```

In [11]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 565326 entries, 0 to 565325
Columns: 397 entries, bk_contract to churn
dtypes: bool(1), float64(155), int32(233), int64(1), object(7)
memory usage: 1.2+ GB
```

In [12]:

```
pd.set_option('display.max_columns', 5)
```

In [13]:

```
sum(df.duplicated())
```

Out[13]:
0

In [14]:
```
expert_variables_scale = expert_variables.copy()
cols = list(expert_variables_scale.drop(remove_list, axis =
1).select_dtypes(include=numerics).columns)
expert_variables_scale['time_alive'] =
expert_variables_scale['time_alive'].fillna(expert_variables_scale['time_alive'].mean())
expert_variables_scale['time_alive'].head()
```

Out[14]:
0 0.0
1 212.0
2 212.0
3 212.0
4 212.0
Name: time_alive, dtype: float64

Filter for training and out-of-time data.

In [15]:
```
#in_train_valid = (expert_variables_scale['Date'] > '2010-01-14') &
(expert_variables_scale['Date'] < '2010-11-01')
#in_oot = expert_variables_scale['Date'] >= '2010-11-01'
in_train_valid = (expert_variables_scale['month'] < 201902) #&
(expert_variables_scale['Date'] < '2010-11-01')
in_oot = expert_variables_scale['month'] >= 201902
train_valid = expert_variables_scale[in_train_valid].sort_index() #.drop('Date', axis = 1)
oot = expert_variables_scale[in_oot].sort_index() #.drop('Date', axis = 1)
```

Z-scale all of the variables.

In [16]:
```
train_valid[cols] = train_valid[cols].apply(zscore)
oot[cols] = oot[cols].apply(zscore)
```

/anaconda3/lib/python3.7/site-packages/scipy/stats/stats.py:2253: RuntimeWarning: invalid value encountered in true_divide
return (a - mns) / sstd

In [18]:
```
train_valid = expert_variables_scale
train_valid_filter=train_valid
```

Get the shape of the training and out of time data.

Add RANDOM variable to the dataframe and remove the Recnum column.

In [19]:
```
## Add random variable for test and remove the `Recnum` column.
#train_valid_filter = train_valid.drop('Recnum', axis = 1)
train_valid_filter['RANDOM'] = np.random.rand(len(train_valid_filter))
```

In [20]:

```
%%time

#target_months_val = [201806, 201807, 201808, 201809, 201810, 201811, 201812,
201901,201902, 201903]  #Todos os meses de validacao.
target_months_val = [201806]
months = np.concatenate((np.arange(201701, 201713), np.arange(201801, 201813),
np.arange(201901, 201904)))  #todos os meses no DW
month_map = dict(zip(months, np.arange(len(months)).astype(int)))
months
num_training_months = 7  # Número de meses "look back". Ou seja, usamos os ultimos 7
Meses pra prever o mes seguinte.
for i, month_val in enumerate(target_months_val):
    print (i, month_val)
```

```
0 201806
CPU times: user 127 µs, sys: 72 µs, total: 199 µs
Wall time: 158 µs
```

In [21]:

```
features=train_valid_filter

#cat_cols = ['ds_seg_cnta', 'regi_cnta', 'ds_macro_seg_cnta', 'frqd_cnta', 'ds_sub']  # Variaveis
categoricas.

all_cols = [i for i in features.columns] #if i not in ['month', 'churn']]  # Todas as colunas que
vao ser utilizadas no training.
                                                                    # Excluimos aqui essas 4
colunas ['bk_contract','month', 'sk_cliente', 'churn']

# for each validation month...
for i, month_val in enumerate(target_months_val):
    print(f'month_val = {month_val}')
    # get validation data

    mask_val = (features['month'] == month_val).values
    month_val_ind = month_map[month_val]
    months_tr = months[month_val_ind-num_training_months:month_val_ind]
    mask_tr = (features['month'].isin(months_tr)).values

    print(months_tr)

    X_tr = features.loc[mask_tr, all_cols]
    y_tr = features.loc[mask_tr, 'churn']

    X_val = features.loc[mask_val, all_cols]
    y_val = features.loc[mask_val, 'churn']
```

```
month_val = 201806
[201711 201712 201801 201802 201803 201804 201805]
```

In [22]:

```
train_valid_filter_backup=train_valid_filter

train_valid_filter=X_tr

train_valid_filter.head()
```

Out[22]:

	bk_contract	month	...	churn	RANDOM
49440	CON99991	201711	...	False	0.563268
49441	CONT00010	201711	...	False	0.557440
49442	CONT00015	201711	...	False	0.718091
49443	CONT00018	201711	...	False	0.201300
49444	CONT00019	201711	...	False	0.425192

5 rows × 398 columns

Compute univariate KS and CDR for each column.

In [23]:

```
ks_cdr = pd.DataFrame(index = cols).reindex(columns = ['KS', 'CDR'])

churn_yes = train_valid_filter[train_valid_filter.churn == 1]
churn_no  = train_valid_filter[train_valid_filter.churn == 0]

total_churn = train_valid_filter.churn.sum()

for col in cols:
    ## Compute the KS score for the column.
    ks_cdr.loc[col, 'KS'] = ks_2samp(churn_no[col].values, churn_yes[col].values)[0]

    ## Compute the CDR for the column.
    ## Start by computing the number of churn in the 3rd percentile from both the high end and
low end.
    churn_from_lo = train_valid_filter.churn[train_valid_filter[col] <=
train_valid_filter[col].quantile(0.03)].sum()
    churn_from_hi = train_valid_filter.churn[train_valid_filter[col] >=
train_valid_filter[col].quantile(0.97)].sum()
    ## Compute and store the churn rate.
    ks_cdr.loc[col, 'CDR'] = max(churn_from_lo, churn_from_hi) / total_churn
```

In [26]:

```
## Compute the rank for both KS and CDR
for var in ['KS', 'CDR']: ks_cdr[var + '_rank'] = ks_cdr[var].rank()

## Compute the average rank.
ks_cdr['avg_rank'] = (ks_cdr['KS_rank'] + ks_cdr['CDR_rank']) / 2

## Sort by the average rank.
ks_cdr = ks_cdr.sort_values(['avg_rank', 'CDR_rank'], ascending = False)

## Save the ks_cdr table.
#ks_fdr.to_csv('../data/ks_cdr_table.csv')
```

Collect the best 50% of variables to use in the wrapper function and create the training and test datasets.

In [27]:

```
ks_cdr
```

Out[27]:

	KS	CDR	KS_rank	CDR_rank	avg_rank
bow_ltda_last_7_months	0.015135	1.000000	192.0	265.0	228.50
ticket_grp_CST - CobranÃ§a FLY01_last_7_months	0.015015	1.000000	190.0	265.0	227.50
ticket_grp_RM MateriaisÂ _last_7_months	0.014169	1.000000	184.0	265.0	224.50
ticket_grp_RM GestÃ£o RH_last_7_months	0.012753	1.000000	177.0	265.0	221.00
ticket_grp_RM Framework_last_7_months	0.012231	1.000000	173.0	265.0	219.00
ticket_grp_RM Controladoria_last_7_months	0.011596	1.000000	168.0	265.0	216.50
ticket_grp_RM Tecnologia_last_7_months	0.010692	1.000000	161.0	265.0	213.00
ticket_grp_Protheus AutomaÃ§Ã£o FiscalÂ _last_7_months	0.010182	1.000000	156.0	265.0	210.50
ticket_grp_RM FinanÃ§asÂ _last_7_months	0.009493	1.000000	152.0	265.0	208.50
bow_rm_last_7_months	0.008721	1.000000	143.0	265.0	204.00
bow_rejeicao_last_7_months	0.008380	1.000000	137.0	265.0	201.00
ticket_grp_Protheus Manufatura_last_7_months	0.008376	1.000000	136.0	265.0	200.50
ticket_grp_Protheus FiscalÂ _last_7_months	0.018698	0.972962	209.0	190.0	199.50
ticket_grp_Protheus ServiÃ§os_last_7_months	0.007704	1.000000	132.0	265.0	198.50
ticket_grp_RM Labore_last_7_months	0.020817	0.971760	213.0	184.0	198.50
ticket_grp_CST - Contratos MPN_last_7_months	0.007087	1.000000	128.5	265.0	196.75
ticket_grp_CST â€“ Contratos MPN_last_7_months	0.007087	1.000000	128.5	265.0	196.75
ticket_grp_Protheus ADVPL_last_7_months	0.006996	1.000000	126.0	265.0	195.50
bow_ferias_last_7_months	0.016384	0.972962	201.0	190.0	195.50
bow_alterar_last_7_months	0.006900	1.000000	125.0	265.0	195.00
bow_cancelamento_last_7_months	0.096083	0.858402	298.0	87.5	192.75
ticket_grp_RM BI_last_7_months	0.006567	1.000000	120.0	265.0	192.50
ticket_grp_CST - Contratos de Vendas_last_7_months	0.069668	0.888644	284.0	101.0	192.50
vlr_bonifi_sum_12_periods	0.006556	1.000000	119.0	265.0	192.00
bow_baixa_last_7_months	0.015537	0.972962	194.0	190.0	192.00
vlr_bonifi_max_12_periods	0.006325	1.000000	116.0	265.0	190.50
ticket_grp_Protheus Fiscal Arquivo MagnÃ©tico_last_7_months	0.016523	0.969557	203.0	178.0	190.50
bow_fluig_last_7_months	0.006250	1.000000	115.0	265.0	190.00
ticket_grp_Protheus Faturamento e CRM_last_7_months	0.013481	0.976167	179.0	200.0	189.50
bow_evento_last_7_months	0.006091	1.000000	113.0	265.0	189.00
...	...	...	...	...	...
ticket_grp_TAF e-Social RM_last_7_months	0.000224	1.000000	7.0	265.0	136.00
bow_ticket_last_7_months	0.005785	0.965952	105.0	167.0	136.00

	KS	CDR	KS_rank	CDR_rank	avg_rank
bow_protheus_last_7_months	0.007032	0.957941	127.0	145.0	136.00
tickets_of_priority_normal_last_6_months	0.020512	0.726818	212.0	60.0	136.00
ticket_grp_CMNET AtlÃ¢ntica_last_7_months	0.000200	1.000000	6.0	265.0	135.50
total_contracts_with_same_franquia_last_6_months	0.043008	0.039656	251.0	19.5	135.25
ticket_grp_TAF e-Social Protheus_last_7_months	0.000170	1.000000	5.0	265.0	135.00
ticket_grp_Virtual Age Faturamento_last_7_months	0.000134	1.000000	4.0	265.0	134.50
ticket_grp_CMNET PDV_last_7_months	0.000107	1.000000	3.0	265.0	134.00
ticket_grp_CMNET LATAM Front Office_last_7_months	0.000000	1.000000	1.5	265.0	133.25
ticket_grp_Chef Corporativo N1_last_7_months	0.000000	1.000000	1.5	265.0	133.25
bow_nota_last_7_months	0.009131	0.931104	147.0	119.0	133.00
churn_rate_for_same_franquia_last_3_months	0.049181	0.031845	261.0	5.0	133.00
bow_problemas_last_7_months	0.004307	0.971160	81.0	182.5	131.75
tickets_of_priority_normal_last_1_months	0.015832	0.790306	197.0	65.0	131.00
churn_rate_for_same_region_last_2_months	0.048471	0.029641	260.0	2.0	131.00
bow_solicitacao_last_7_months	0.008248	0.937913	135.0	124.0	129.50
churn_rate_for_same_franquia_last_1_months	0.041131	0.034448	249.0	9.0	129.00
bow_chat_last_7_months	0.004948	0.965752	91.0	166.0	128.50
bow_portal_last_7_months	0.004483	0.967755	84.0	172.5	128.25
sk_cliente_old_nunique_12_periods	0.021901	0.066693	216.0	35.5	125.75
bow_dados_last_7_months	0.006088	0.948328	112.0	135.0	123.50
churn_rate_for_same_franquia_last_2_months	0.035488	0.030443	244.0	3.0	123.50
bow_transferencia_last_7_months	0.004086	0.967555	76.0	170.5	123.25
bow_alteracao_last_7_months	0.006346	0.941118	117.0	127.0	122.00
tickets_of_priority_normal_last_3_months	0.014132	0.737232	183.0	61.0	122.00
tickets_of_priority_normal_last_2_months	0.012277	0.752453	174.0	63.0	118.50
bow_totvs_last_7_months	0.005697	0.946926	103.0	132.0	117.50
bow_sistema_last_7_months	0.004831	0.949930	88.0	137.0	112.50
ticket_grp_Cloud Atendimento SPÂ _last_7_months	0.003445	0.961947	66.0	154.5	110.25

323 rows × 5 columns

In [28]:

```
ks_cdr
```

Out[28]:

	KS	CDR	KS_rank	CDR_rank	avg_rank
bow_ltda_last_7_months	0.015135	1.00000	192.0	265.0	228.50
ticket_grp_CST - CobranÃ§a FLY01_last_7_months	0.015015	1.00000	190.0	265.0	227.50
ticket_grp_RM MateriaisÂ _last_7_months	0.014169	1.00000	184.0	265.0	224.50

	KS	CDR	KS_rank	CDR_rank	avg_rank
ticket_grp_RM GestÃ£o RH_last_7_months	0.012753	1.00000	177.0	265.0	221.00
ticket_grp_RM Framework_last_7_months	0.012231	1.000000	173.0	265.0	219.00
ticket_grp_RM Controladoria_last_7_months	0.011596	1.000000	168.0	265.0	216.50
ticket_grp_RM Tecnologia_last_7_months	0.010692	1.000000	161.0	265.0	213.00
ticket_grp_Protheus AutomaÃ§Ã£o FiscalÂ _last_7_months	0.010182	1.000000	156.0	265.0	210.50
ticket_grp_RM FinanÃ§asÂ _last_7_months	0.009493	1.000000	152.0	265.0	208.50
bow_rm_last_7_months	0.008721	1.000000	143.0	265.0	204.00
bow_rejeicao_last_7_months	0.008380	1.000000	137.0	265.0	201.00
ticket_grp_Protheus Manufatura_last_7_months	0.008376	1.000000	136.0	265.0	200.50
ticket_grp_Protheus FiscalÂ _last_7_months	0.018698	0.972962	209.0	190.0	199.50
ticket_grp_Protheus ServiÃ§os_last_7_months	0.007704	1.000000	132.0	265.0	198.50
ticket_grp_RM Labore_last_7_months	0.020817	0.971760	213.0	184.0	198.50
ticket_grp_CST - Contratos MPN_last_7_months	0.007087	1.000000	128.5	265.0	196.75
ticket_grp_CST â€“ Contratos MPN_last_7_months	0.007087	1.000000	128.5	265.0	196.75
ticket_grp_Protheus ADVPL_last_7_months	0.006996	1.000000	126.0	265.0	195.50
bow_ferias_last_7_months	0.016384	0.972962	201.0	190.0	195.50
bow_alterar_last_7_months	0.006900	1.000000	125.0	265.0	195.00
bow_cancelamento_last_7_months	0.096083	0.858402	298.0	87.5	192.75
ticket_grp_RM BI_last_7_months	0.006567	1.000000	120.0	265.0	192.50
ticket_grp_CST - Contratos de Vendas_last_7_months	0.069668	0.888644	284.0	101.0	192.50
vlr_bonifi_sum_12_periods	0.006556	1.000000	119.0	265.0	192.00
bow_baixa_last_7_months	0.015537	0.972962	194.0	190.0	192.00
vlr_bonifi_max_12_periods	0.006325	1.000000	116.0	265.0	190.50
ticket_grp_Protheus Fiscal Arquivo MagnÃ©tico_last_7_months	0.016523	0.969557	203.0	178.0	190.50
bow_fluig_last_7_months	0.006250	1.000000	115.0	265.0	190.00
ticket_grp_Protheus Faturamento e CRM_last_7_months	0.013481	0.976167	179.0	200.0	189.50
bow_evento_last_7_months	0.006091	1.000000	113.0	265.0	189.00
...	...	...	...	...	...
ticket_grp_TAF e-Social RM_last_7_months	0.000224	1.000000	7.0	265.0	136.00
bow_ticket_last_7_months	0.005785	0.965952	105.0	167.0	136.00
bow_protheus_last_7_months	0.007032	0.957941	127.0	145.0	136.00
tickets_of_priority_normal_last_6_months	0.020512	0.726818	212.0	60.0	136.00
ticket_grp_CMNET AtlÃ¢ntica_last_7_months	0.000200	1.000000	6.0	265.0	135.50
total_contracts_with_same_franquia_last_6_months	0.043008	0.039656	251.0	19.5	135.25
ticket_grp_TAF e-Social Protheus_last_7_months	0.00017	1.00000	5.0	265.0	135.00
ticket_grp_Virtual Age Faturamento_last_7_months	0.000134	1.000000	4.0	265.0	134.50
ticket_grp_CMNET PDV_last_7_months	0.000107	1.000000	3.0	265.0	134.00

	KS	CDR	KS_rank	CDR_rank	avg_rank
ticket_grp_CMNET LATAM Front Office_last_7_months	0.000000	1.000000	1.5	265.0	133.25
ticket_grp_Chef Corporativo N1_last_7_months	0.000000	1.000000	1.5	265.0	133.25
bow_nota_last_7_months	0.009131	0.931104	147.0	119.0	133.00
churn_rate_for_same_franquia_last_3_months	0.049181	0.031845	261.0	5.0	133.00
bow_problemas_last_7_months	0.004307	0.971160	81.0	182.5	131.75
tickets_of_priority_normal_last_1_months	0.015832	0.790306	197.0	65.0	131.00
churn_rate_for_same_region_last_2_months	0.048471	0.029641	260.0	2.0	131.00
bow_solicitacao_last_7_months	0.008248	0.937913	135.0	124.0	129.50
churn_rate_for_same_franquia_last_1_months	0.041131	0.034448	249.0	9.0	129.00
bow_chat_last_7_months	0.004948	0.965752	91.0	166.0	128.50
bow_portal_last_7_months	0.004483	0.967755	84.0	172.5	128.25
sk_cliente_old_nunique_12_periods	0.021901	0.066693	216.0	35.5	125.75
bow_dados_last_7_months	0.006088	0.948328	112.0	135.0	123.50
churn_rate_for_same_franquia_last_2_months	0.035488	0.030443	244.0	3.0	123.50
bow_transferencia_last_7_months	0.004086	0.967555	76.0	170.5	123.25
bow_alteracao_last_7_months	0.006346	0.941118	117.0	127.0	122.00
tickets_of_priority_normal_last_3_months	0.014132	0.737232	183.0	61.0	122.00
tickets_of_priority_normal_last_2_months	0.012277	0.752453	174.0	63.0	118.50
bow_totvs_last_7_months	0.005697	0.946926	103.0	132.0	117.50
bow_sistema_last_7_months	0.004831	0.949930	88.0	137.0	112.50
ticket_grp_Cloud Atendimento SPÂ _last_7_months	0.003445	0.961947	66.0	154.5	110.25

323 rows × 5 columns

In [29]:

```
filter_vars = list(ks_cdr[ks_cdr.avg_rank >= np.median(ks_cdr.avg_rank)].index)
filter_vars
```

Out[29]:

```
['bow_ltda_last_7_months',
 'ticket_grp_CST - CobranÃ§a FLY01_last_7_months',
 'ticket_grp_RM MateriaisÂ\xa0_last_7_months',
 'ticket_grp_RM GestÃ£o RH_last_7_months',
 'ticket_grp_RM Framework_last_7_months',
 'ticket_grp_RM Controladoria_last_7_months',
 'ticket_grp_RM Tecnologia_last_7_months',
 'ticket_grp_Protheus AutomaÃ§Ã£o FiscalÂ\xa0_last_7_months',
 'ticket_grp_RM FinanÃ§asÂ\xa0_last_7_months',
 'bow_rm_last_7_months',
 'bow_rejeicao_last_7_months',
 'ticket_grp_Protheus Manufatura_last_7_months',
 'ticket_grp_Protheus FiscalÂ\xa0_last_7_months',
 'ticket_grp_Protheus ServiÃ§os_last_7_months',
 'ticket_grp_RM Labore_last_7_months',
 'ticket_grp_CST - Contratos MPN_last_7_months',
 'ticket_grp_CST â€“ Contratos MPN_last_7_months',
 'ticket_grp_Protheus ADVPL_last_7_months',
```

'bow_ferias_last_7_months',
'bow_alterar_last_7_months',
'bow_cancelamento_last_7_months',
'ticket_grp_RM BI_last_7_months',
'ticket_grp_CST - Contratos de Vendas_last_7_months',
'vlr_bonifi_sum_12_periods',
'bow_baixa_last_7_months',
'vlr_bonifi_max_12_periods',
'ticket_grp_Protheus Fiscal Arquivo MagnÃ©tico_last_7_months',
'bow_fluig_last_7_months',
'ticket_grp_Protheus Faturamento e CRM_last_7_months',
'bow_evento_last_7_months',
'bow_contrato_last_7_months',
'ticket_grp_FLY01 SAÃšDE_last_7_months',
'ticket_grp_Fly01 SaÃºde - Personal Med Online_last_7_months',
'bow_reinf_last_7_months',
'bow_social_last_7_months',
'bow_retorno_last_7_months',
'vlr_bonifi_mean_12_periods',
'ticket_grp_Protheus SPEDÂ\xa0_last_7_months',
'bow_processo_last_7_months',
'ticket_grp_Protheus Recursos HumanosÂ\xa0_last_7_months',
'churn_sum_6_periods',
'bow_notas_last_7_months',
'total_contracts_with_same_segment_last_6_months',
'total_contracts_with_same_segment_last_3_months',
'total_contracts_with_same_segment_last_2_months',
'total_contracts_with_same_segment_last_1_months',
'bow_xml_last_7_months',
'churn_rate_for_same_product_last_6_months',
'ticket_grp_ESocial Totvs_last_7_months',
'ticket_grp_Protheus Materiais_last_7_months',
'days_to_carencia_end',
'bow_falha_last_7_months',
'ticket_grp_Protheus Financeiro e Planejamento_last_7_months',
'vlr_nvvend_std_6_periods',
'carencia_span',
'bow_erro_last_7_months',
'days_to_carencia_start',
'bow_teste_last_7_months',
'ticket_grp_Protheus Ativo e ContÃ¡bil_last_7_months',
'bow_rescisao_last_7_months',
'churn_rate_for_same_product_last_1_months',
'bow_log_last_7_months',
'vlr_nvvend_mean_6_periods',
'vlr_nvvend_std_12_periods',
'bow_taf_last_7_months',
'bow_envio_last_7_months',
'tickets_of_type_incident_last_3_months',
'total_contracts_with_same_product_last_2_months',

'vlr_nvvend_sum_6_periods',
'vlr_nvvend_max_6_periods',
'bow_venda_last_7_months',
'tickets_of_priority_urgent_last_2_months',
'churn_sum_12_periods',
'churn_rate_for_same_product_last_3_months',
'vlr_cancel_mean_6_periods',
'tickets_of_priority_urgent_last_6_months',
'vlr_cancel_std_6_periods',
'time_alive',
'bow_registro_last_7_months',
'vlr_cancel_sum_6_periods',
'vlr_cancel_max_6_periods',
'tickets_of_priority_urgent_last_1_months',
'bow_conta_last_7_months',
'total_contracts_with_same_product_last_1_months',
'total_contracts_with_same_product_last_3_months',
'total_contracts_with_same_product_last_6_months',
'bow_producao_last_7_months',
'tickets_of_type_incident_last_6_months',
'tickets_of_priority_urgent_last_3_months',
'bow_usuario_last_7_months',
'bow_icms_last_7_months',
'bow_cadastro_last_7_months',
'tickets_of_priority_low_last_3_months',
'total_contracts_with_same_macro_segment_last_1_months',
'bow_exclusao_last_7_months',
'bow_ponto_last_7_months',
'vlr_nvvend_sum_12_periods',
'churn_rate_for_same_product_last_2_months',
'bow_geracao_last_7_months',
'bow_servico_last_7_months',
'tickets_of_type_incident_last_2_months',
'vlr_nvvend_max_12_periods',
'tickets_of_priority_low_last_6_months',
'total_contracts_with_same_macro_segment_last_3_months',
'total_contracts_with_same_macro_segment_last_2_months',
'vlr_nvvend_mean_12_periods',
'ticket_grp_Datasul Recursos HumanosÂ\xa0_last_7_months',
'bow_mata_last_7_months',
'bow_base_last_7_months',
'bow_gerar_last_7_months',
'tickets_of_priority_low_last_2_months',
'bow_res_last_7_months',
'bow_folha_last_7_months',
'bow_relatorio_last_7_months',
'churn_rate_for_same_macro_segment_last_3_months',
'vlr_bonifi_max_6_periods',
'tickets_of_type_incident_last_1_months',
'bow_atendimento_last_7_months',

'vlr_bonifi_sum_6_periods',
'tickets_of_priority_low_last_1_months',
'vlr_bonifi_mean_6_periods',
'bow_error_last_7_months',
'ticket_grp_Protheus Framework_last_7_months',
'bow_banco_last_7_months',
'ticket_grp_Datasul FinanÃ§asÂ\xa0_last_7_months',
'ticket_grp_RM Educacional_last_7_months',
'bow_arquivo_last_7_months',
'bow_pedido_last_7_months',
'ticket_grp_Datasul ManufaturaÂ\xa0_last_7_months',
'bow_inclusao_last_7_months',
'ticket_grp_Protheus Trade PadrÃ£o_last_7_months',
'bow_esocial_last_7_months',
'bow_integracao_last_7_months',
'bow_urgente_last_7_months',
'bow_sped_last_7_months',
'bow_calculo_last_7_months',
'bow_financeiro_last_7_months',
'bow_atualizacao_last_7_months',
'churn_rate_for_same_macro_segment_last_2_months',
'ticket_grp_SQUAD 2 CLOUD PROTHEUS_last_7_months',
'tickets_of_type_task_last_1_months',
'vlr_reajuste_std_6_periods',
'vlr_reajuste_mean_6_periods',
'bow_chamado_last_7_months',
'ticket_grp_Ouvidoria Totvs_last_7_months',
'vlr_reajuste_sum_6_periods',
'bow_codigo_last_7_months',
'bow_importacao_last_7_months',
'churn_rate_for_same_macro_segment_last_1_months',
'ticket_grp_Cloud TSS_last_7_months',
'vlr_reajuste_max_6_periods',
'ticket_grp_Consultoria Datasul_last_7_months',
'ticket_grp_EMAIL NOTIFICAÃ‡Ãƒ*f*O_last_7_months',
'bow_nf_last_7_months',
'ticket_grp_Bemacash_last_7_months',
'tickets_of_type_problem_last_6_months',
'ticket_grp_Datasul Faturamento_last_7_months',
'churn_rate_for_same_segment_last_3_months',
'churn_rate_for_same_sub_segment_last_6_months',
'churn_rate_for_same_sub_segment_last_3_months',
'bow_entrada_last_7_months',
'ticket_grp_Datasul ControladoriaÂ\xa0_last_7_months',
'bow_nfe_last_7_months']

In [30]:

```
X = train_valid_filter.loc[:, filter_vars]
y = train_valid_filter.loc[:, 'churn']
```

Run the wrapper.

In [ ]:

```
run_rfecv_1 = True

if run_rfecv_1:
    #  Create a logistic regression estimator
    logreg = linear_model.LogisticRegression(solver='liblinear')
    # Use RFECV to pick best features, using Stratified Kfold
    rfecv1 =   RFECV(estimator=logreg, cv=StratifiedKFold(2), scoring='accuracy',
verbose=1)
    # Fit the features to the response variable
    rfecv1.fit(X, y)
    # Save the file using the pickle module
    #with open('rfe_cv_1.p', 'wb') as file: pickle.dump(rfecv1, file)
else:
    with open('rfe_cv_1.p', 'rb') as file: rfecv1 = pickle.load(file)
```

In [ ]:

```
print("Optimal number of features : %d" % rfecv1.n_features_)

# Plot number of features VS. cross-validation scores
plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score (nb of correct classifications)")
plt.plot(range(1, len(rfecv1.grid_scores_) + 1), rfecv1.grid_scores_)
plt.show()
```

In [74]:

```
wrapper_first_pass_keep = X.columns[rfecv1.ranking_ < np.median(rfecv1.ranking_)]

X_first_pass = X.loc[:, wrapper_first_pass_keep]
```

In [75]:

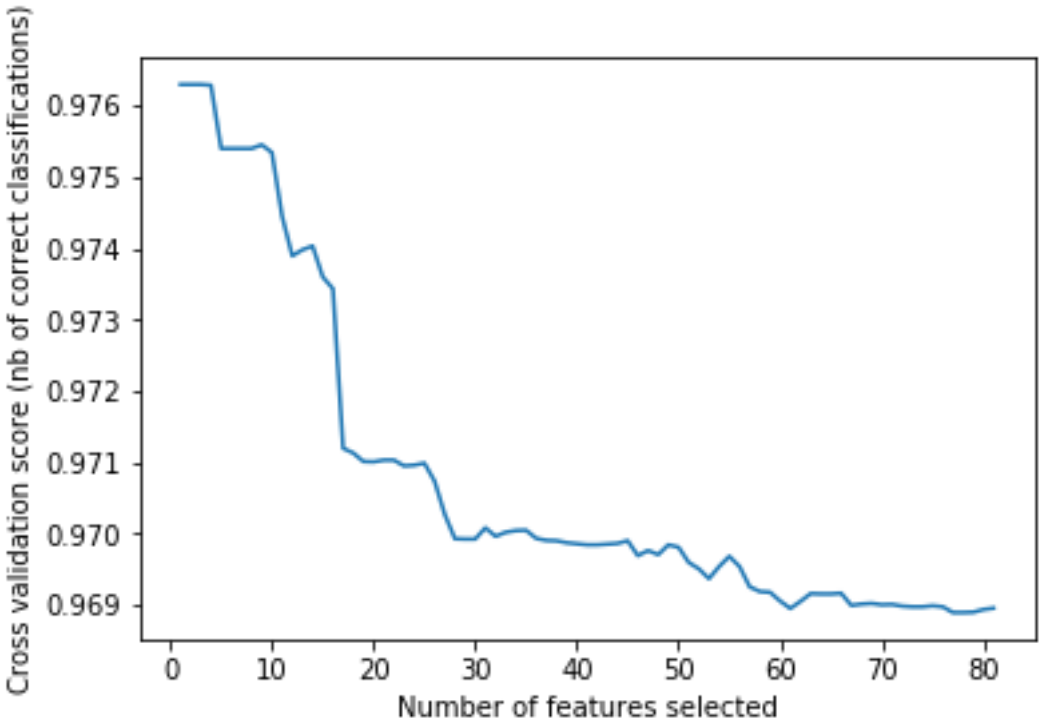
```
run_rfecv_2 = True

if run_rfecv_2:
    #  Create a logistic regression estimator
    logreg = linear_model.LogisticRegression(solver='liblinear')
    # Use RFECV to pick best features, using Stratified Kfold
    rfecv2 =   RFECV(estimator=logreg, cv=StratifiedKFold(2), scoring='accuracy',
verbose=1)
    # Fit the features to the response variable
    rfecv2.fit(X_first_pass, y)
    # Save the file using the pickle module
    with open('rfe_cv_2.p', 'wb') as file: pickle.dump(rfecv2, file)
else:
    with open('rfe_cv_2.p', 'rb') as file: rfecv2 = pickle.load(file)
```

In [76]:

```
print("Optimal number of features : %d" % rfecv2.n_features_)

# Plot number of features VS. cross-validation scores
```

```
plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score (nb of correct classifications)")
plt.plot(range(1, len(rfecv2.grid_scores_) + 1), rfecv2.grid_scores_)
plt.show()
```

Optimal number of features : 3



In [36]:

```
run_rfe_2 = True

if run_rfe_2:
    #  Create a logistic regression estimator
    logreg = linear_model.LogisticRegression(solver='liblinear')
    # Use RFE to pick best 15 features
    rfe2 = RFE(estimator=logreg, n_features_to_select=15, verbose=1)
    # Fit the features to the response variable
    rfe2.fit(X_first_pass, y)
    # Save the file using the pickle module
    with open('rfe_2.p', 'wb') as file: pickle.dump(rfe2, file)
else:
    with open('rfe_2.p', 'rb') as file: rfe2 = pickle.load(file)
```

```
NameError                                 Traceback (most recent call last)
<ipython-input-36-cf023e66b383> in <module>()
      7     rfe2 = RFE(estimator=logreg, n_features_to_select=15, verbose=1)
      8     # Fit the features to the response variable
----> 9     rfe2.fit(X_first_pass, y)
     10     # Save the file using the pickle module
     11     with open('rfe_2.p', 'wb') as file: pickle.dump(rfe2, file)
```

NameError: name 'X_first_pass' is not defined

In [80]:
```
print(col_to_keep)
```
```
Index(['ticket_grp_CST - Contratos de Vendas_last_7_months',
       'total_contracts_with_same_segment_last_6_months',
       'total_contracts_with_same_segment_last_3_months',
       'total_contracts_with_same_segment_last_2_months',
       'total_contracts_with_same_segment_last_1_months',
       'churn_rate_for_same_product_last_6_months',
       'churn_rate_for_same_product_last_1_months',
       'total_contracts_with_same_product_last_2_months',
       'vlr_nvvend_sum_6_periods', 'churn_rate_for_same_product_last_3_months',
       'total_contracts_with_same_product_last_1_months',
       'total_contracts_with_same_product_last_6_months',
       'total_contracts_with_same_macro_segment_last_1_months',
       'total_contracts_with_same_macro_segment_last_3_months',
       'total_contracts_with_same_macro_segment_last_2_months',
       'vlr_nvvend_mean_12_periods',
       'churn_rate_for_same_macro_segment_last_3_months',
       'churn_rate_for_same_macro_segment_last_6_months',
       'churn_rate_for_same_macro_segment_last_2_months',
       'churn_rate_for_same_macro_segment_last_1_months'],
      dtype='object')
```

In [14]:
```
columns_to_save=['churn','month','ticket_grp_CST - Contratos de Vendas_last_7_months',
       'total_contracts_with_same_segment_last_6_months',
       'total_contracts_with_same_segment_last_3_months',
       'total_contracts_with_same_segment_last_2_months',
       'total_contracts_with_same_segment_last_1_months',
       'churn_rate_for_same_product_last_6_months',
       'churn_rate_for_same_product_last_1_months',
       'total_contracts_with_same_product_last_2_months',
       'vlr_nvvend_sum_6_periods',
       'churn_rate_for_same_product_last_3_months',
       'total_contracts_with_same_product_last_1_months',
       'total_contracts_with_same_product_last_6_months',
       'total_contracts_with_same_macro_segment_last_1_months',
       'total_contracts_with_same_macro_segment_last_3_months',
       'total_contracts_with_same_macro_segment_last_2_months',
       'vlr_nvvend_mean_12_periods',
       'churn_rate_for_same_macro_segment_last_3_months',
       'churn_rate_for_same_macro_segment_last_6_months',
       'churn_rate_for_same_macro_segment_last_2_months',
       'churn_rate_for_same_macro_segment_last_1_months']
new_df_2 = expert_variables_scale.loc[mask_tr, columns_to_save]
new_df_2.shape
```
Out[14]:
(210652, 22)

In [15]:

```
new_df_2['churn'] = (new_df_2['churn'] == True ).astype(int)
new_df_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 210652 entries, 49440 to 260091
Data columns (total 22 columns):
churn                                                210652 non-null int64
month                                                210652 non-null int64
ticket_grp_CST - Contratos de Vendas_last_7_months   210652 non-null int32
total_contracts_with_same_segment_last_6_months      210652 non-null float64
total_contracts_with_same_segment_last_3_months      210652 non-null float64
total_contracts_with_same_segment_last_2_months      210652 non-null float64
total_contracts_with_same_segment_last_1_months      210652 non-null float64
churn_rate_for_same_product_last_6_months            210652 non-null float64
churn_rate_for_same_product_last_1_months            210652 non-null float64
total_contracts_with_same_product_last_2_months      210652 non-null float64
vlr_nvvend_sum_6_periods                             210652 non-null float64
churn_rate_for_same_product_last_3_months            210652 non-null float64
total_contracts_with_same_product_last_1_months      210652 non-null float64
total_contracts_with_same_product_last_6_months      210652 non-null float64
total_contracts_with_same_macro_segment_last_1_months    210652 non-null float64
total_contracts_with_same_macro_segment_last_3_months    210652 non-null float64
total_contracts_with_same_macro_segment_last_2_months    210652 non-null float64
vlr_nvvend_mean_12_periods                           210652 non-null float64
churn_rate_for_same_macro_segment_last_3_months      210652 non-null float64
churn_rate_for_same_macro_segment_last_6_months      210652 non-null float64
churn_rate_for_same_macro_segment_last_2_months      210652 non-null float64
churn_rate_for_same_macro_segment_last_1_months      210652 non-null float64
dtypes: float64(19), int32(1), int64(2)
memory usage: 36.2 MB
```

In [16]:

```
new_df_2.head()
new_df_2.to_csv("originais_selecionado.csv", index=False)
```

APÊNDICE D - TRECHOS DE CÓDIGO EM PYTHON – APLICAÇÃO DO XGBOOST

In [14]:

```
import pandas as pd
from scipy.stats import zscore, ks_2samp
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV, RFE
from sklearn import linear_model
from sklearn.model_selection import StratifiedKFold
import pickle
import xgboost as xgb
from sklearn.metrics import mean_squared_error
```

Lê o conjunto de dados contendo as fatures criadas anteriormente.

In [15]:

```
df = pd.read_parquet("Tabela_Referencias@20190401.parquet")
numerics = ['int32', 'int64', 'float64', 'bool']
expert_variables = df
#df.select_dtypes(include=numerics)
```

In [16]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 565326 entries, 0 to 565325
Columns: 397 entries, bk_contract to churn
dtypes: bool(1), float64(155), int32(233), int64(1), object(7)
memory usage: 1.2+ GB
```

In [17]:

```
df.month.unique()
```

Out[17]:

```
array([201709, 201710, 201711, 201712, 201801, 201802, 201803, 201804,
       201805, 201806, 201807, 201808, 201809, 201810, 201811, 201812,
       201901, 201902, 201903])
```

In [18]:

```
sum(df.duplicated())
```

Out[18]:

```
0
```

In [19]:

```
df.shape
```

Out[19]:

```
(565326, 397)
```

Remove unwanted features.

In [20]:

```
remove_list=
['cf_score_minus_8','cf_score_minus_7','cf_score_minus_6','cf_score_minus_5','cf_score_min
us_4','cf_score_minus_3','cf_score_minus_2','cf_score_minus_1','cf_score_last_8_mean','cf_s
core_last_7_mean','cf_score_last_6_mean','cf_score_last_5_mean','cf_score_last_4_mean','cf
_score_last_3_mean','cf_score_last_2_mean','product_embed_0','product_embed_1','product_
embed_2','product_embed_3','product_embed_4','product_embed_5','product_embed_6','prod
uct_embed_7','product_embed_8','product_embed_9','product_embed_10','product_embed_11
','product_embed_12','product_embed_13','product_embed_14','product_embed_15','product_
embed_16','product_embed_17','product_embed_18','product_embed_19','product_embed_20'
,'product_embed_21','product_embed_22','product_embed_23','product_embed_24','client_em
bed_0','client_embed_1','client_embed_2','client_embed_3','client_embed_4','client_embed_5'
,'client_embed_6','client_embed_7','client_embed_8','client_embed_9','client_embed_10','clie
nt_embed_11','client_embed_12','client_embed_13','client_embed_14','client_embed_15','clie
nt_embed_16','client_embed_17','client_embed_18','client_embed_19','client_embed_20','clie
nt_embed_21','client_embed_22','client_embed_23','client_embed_24']
selected_columns = [i for i in expert_variables.columns if i not in remove_list]
expert_variables_scale = expert_variables.copy()
cols = list(expert_variables_scale.drop(remove_list, axis =
1).select_dtypes(include=numerics).columns)
pd.set_option('display.max_columns', 5)
df_2 = expert_variables_scale[cols]
df_2.head()
```

Out[20]:

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
0	201709	1.0	...	3074.0	False
1	201709	1.0	...	698.0	False
2	201709	1.0	...	33.0	False
3	201709	1.0	...	206.0	False
4	201709	1.0	...	33.0	False

5 rows × 325 columns

In [21]:

```
#df_2.churn.astype(int)
df_2['churn'] = (df_2['churn'] == True ).astype(int)
df_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 565326 entries, 0 to 565325
Columns: 325 entries, month to churn
dtypes: float64(90), int32(233), int64(2)
memory usage: 903.6 MB
/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy

Separate training data and OOT data.

In [22]:
```
in_train_valid = (df_2['month'] < 201902)
in_oot = expert_variables_scale['month'] >= 201902
train_valid = df_2[in_train_valid].sort_index()
oot = df_2[in_oot].sort_index()
```

In [23]:
```
train_valid.head()
```
Out[23]:

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
0	201709	1.0	...	3074.0	0
1	201709	1.0	...	698.0	0
2	201709	1.0	...	33.0	0
3	201709	1.0	...	206.0	0
4	201709	1.0	...	33.0	0

5 rows × 325 columns

In [24]:
```
oot.head()
```
Out[24]:

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
505414	201902	0.0	...	2139.0	0
505415	201902	1.0	...	3168.0	0
505416	201902	1.0	...	46.0	0
505417	201902	1.0	...	1885.0	0
505418	201902	1.0	...	46.0	0

5 rows × 325 columns

In [12]:
```
train_valid.shape
```
Out[12]:
(505414, 325)

In [13]:

```
oot.shape
```

Out[13]:

(59912, 325)

In [26]:

```
X, y = train_valid.iloc[:,1:-1],train_valid.iloc[:,-1]
```

In [27]:

```
X.head()
```

Out[27]:

	sk_cliente_old_nunique_12_periods	sk_prod_nunique_12_periods	...	churn_rate_for_same_sub_segment_last_1_months	total_contracts_with_same_sub_segment_last_1_months
0	1.0	1.0	...	0.032206	3074.0
1	1.0	1.0	...	0.018625	698.0
2	1.0	1.0	...	0.000000	33.0
3	1.0	1.0	...	0.029126	206.0
4	1.0	1.0	...	0.000000	33.0

5 rows × 323 columns

In [28]:

```
X.shape
```

Out[28]:

(505414, 323)

Filtro para dados de treinamento e out-of-time.

Convert the dataset into an optimized data structure called Dmatrix that XGBoost supports and gives it acclaimed performance and efficiency gains.

In [29]:

```
data_dmatrix = xgb.DMatrix(data=X,label=y)
```

```
/anaconda3/lib/python3.7/site-packages/xgboost/core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version
  if getattr(data, 'base', None) is not None and \
/anaconda3/lib/python3.7/site-packages/xgboost/core.py:588: FutureWarning: Series.base is deprecated and will be removed in a future version
  data.base is not None and isinstance(data, np.ndarray) \
```

In [30]:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=123)
```

In [31]:

```
xg_reg = xgb.XGBClassifier(objective ='binary:logistic', colsample_bytree = 0.3,
learning_rate = 0.1,
            max_depth = 5, alpha = 10, n_estimators = 10)
```

In [32]:

```
xg_reg.fit(X_train,y_train)

preds = xg_reg.predict(X_test)
```

```
/anaconda3/lib/python3.7/site-packages/sklearn/preprocessing/label.py:151:
DeprecationWarning: The truth value of an empty array is ambiguous. Returning False, but in
future this will result in an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
```

In [33]:

```
preds
```

Out[33]:

```
array([0, 0, 0, ..., 0, 0, 0])
```

In [34]:

```
rmse = np.sqrt(mean_squared_error(y_test, preds))
print("RMSE: %f" % (rmse))
```

```
RMSE: 0.157642
```

In [35]:

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,preds)
```

Out[35]:

```
array([[98571,     0],
       [ 2512,     0]])
```

In [36]:

```
print((2510+2)/110274)
```

```
0.022779621669659213
```

In [37]:

```
df_2[df_2['churn'] == 1].head()
```

Out[37]:

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
91	201709	2.0	...	1037.0	1
208	201709	1.0	...	1037.0	1
238	201709	1.0	...	1146.0	1

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
481	201709	1.0	...	1477.0	1
502	201709	1.0	...	1037.0	1

5 rows × 325 columns

In [38]:

```
df_3 = df_2[df_2['churn'] == 0].sample(13715).append(df_2[df_2['churn'] == 1])
df_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 27430 entries, 379090 to 564941
Columns: 325 entries, month to churn
dtypes: float64(90), int32(233), int64(2)
memory usage: 43.8 MB
```

In [39]:

```
df_3.head()
```

Out[39]:

	month	sk_cliente_old_nunique_12_periods	...	total_contracts_with_same_sub_segment_last_1_months	churn
379090	201809	1.0	...	1635.0	0
107802	201801	1.0	...	2088.0	0
109380	201801	1.0	...	177.0	0
133163	201801	0.0	...	17.0	0
152761	201802	1.0	...	2115.0	0

5 rows × 325 columns

In [40]:

```
X, y = df_3.iloc[:,:-1],df_3.iloc[:,-1]
```

In [41]:

```
y.head()
```

Out[41]:

```
379090    0
107802    0
109380    0
133163    0
152761    0
Name: churn, dtype: int64
```

In [42]:

```
data_dmatrix = xgb.DMatrix(data=X,label=y)
```

```
/anaconda3/lib/python3.7/site-packages/xgboost/core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version
```

```
  if getattr(data, 'base', None) is not None and \
/anaconda3/lib/python3.7/site-packages/xgboost/core.py:588: FutureWarning: Series.base is deprecated and will be removed in a future version
  data.base is not None and isinstance(data, np.ndarray) \
```

In [43]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=123)
```

In [44]:

```
xg_reg = xgb.XGBClassifier(objective ='binary:logistic', colsample_bytree = 0.3,
learning_rate = 0.1,
            max_depth = 5, alpha = 10, n_estimators = 10)
```

In [45]:

```
xg_reg.fit(X_train,y_train)

preds = xg_reg.predict(X_test)
```

```
/anaconda3/lib/python3.7/site-packages/sklearn/preprocessing/label.py:151:
DeprecationWarning: The truth value of an empty array is ambiguous. Returning False, but in future this will result in an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
```

In [46]:

```
confusion_matrix(y_test,preds)
```

Out[46]:

```
array([[2111,  644],
       [ 925, 1806]])
```

In [47]:

```
tn, fp, fn, tp = confusion_matrix(y_test,preds).ravel()
(tn, fp, fn, tp)
```

Out[47]:

```
(2111, 644, 925, 1806)
```

In [48]:

```
(tp+tn)/(tn+fp+fn+tp)
```

Out[48]:

```
0.7139992708713088
```

In [49]:

```
preds
```

Out[49]:

```
array([1, 0, 0, ..., 1, 1, 0])
```

In [50]:

```
y_train
```

Out[50]:

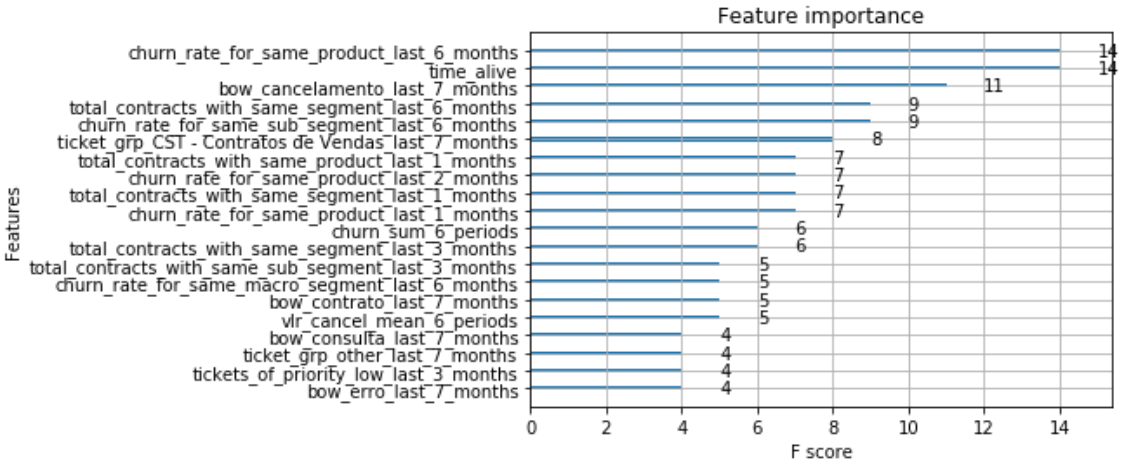
```
243235    0
455307    0
529234    0
524415    0
275328    0
```

411203 1
434855 0
79961 1
181615 0
215216 0
553156 1
507965 0
167057 0
415175 0
468585 0
302064 0
75250 1
348096 0
185687 0
482888 0
511947 0
564355 0
106978 1
258622 1
126984 0
518783 0
228153 0
231834 1
321650 0
452061 0
..
527190 0
531182 1
42955 0
224041 1
306091 0
183272 1
63165 1
509552 0
260110 0
194649 1
502466 0
466978 0
377452 1
452166 0
46373 0
295659 1
293788 1
397086 1
372947 1
405924 0
358236 1
424411 1
520276 1
385587 0

```
440202    0
74760     1
336333    1
194122    1
94924     1
274882    1
Name: churn, Length: 21944, dtype: int64
```

In [51]:

```
xgb.plot_importance(xg_reg, max_num_features=20)
plt.rcParams['figure.figsize'] = [15, 15]
plt.show()
```



In [52]:

```
X_oot, y_oot = oot.iloc[:,:-1],oot.iloc[:,-1]
```

In [53]:

```
xg_reg.fit(X_train,y_train)

preds_train = xg_reg.predict_proba(X_train)
preds_test = xg_reg.predict_proba(X_test)
preds_oot = xg_reg.predict_proba(X_oot)
```

In [54]:

```
preds_test
```

Out[54]:

```
array([[0.26777387, 0.73222613],
       [0.6352314 , 0.36476862],
       [0.5928117 , 0.4071883 ],
       ...,
       [0.4783404 , 0.5216596 ],
       [0.45394325, 0.54605675],
       [0.55960846, 0.44039154]], dtype=float32)
```

In [55]:

```
preds_test.shape
```

Out[55]:

```
(5486, 2)
```

In [56]:

```
y_train.shape
```

Out[56]:

(21944,)

In [57]:

```
X_test.shape
```

Out[57]:

(5486, 324)

In [58]:

```
dataset = pd.DataFrame({'Non-Churn': preds_test[:, 0], 'Churn': preds_test[:, 1]})
>>> print(dataset)
```

	Non-Churn	Churn
0	0.267774	0.732226
1	0.635231	0.364769
2	0.592812	0.407188
3	0.344525	0.655475
4	0.662676	0.337324
5	0.623026	0.376974
6	0.456233	0.543767
7	0.316960	0.683040
8	0.528947	0.471053
9	0.349213	0.650787
10	0.363332	0.636668
11	0.317604	0.682396
12	0.602878	0.397122
13	0.436395	0.563605
14	0.286423	0.713577
15	0.516227	0.483773
16	0.448312	0.551688
17	0.486245	0.513755
18	0.697321	0.302679
19	0.546978	0.453022
20	0.602878	0.397122
21	0.362796	0.637204
22	0.431533	0.568467
23	0.537327	0.462673
24	0.611351	0.388649
25	0.338418	0.661582
26	0.574727	0.425273
27	0.579207	0.420793
28	0.559608	0.440392
29	0.588176	0.411824
...	...	...
5456	0.592812	0.407188
5457	0.390658	0.609342
5458	0.646416	0.353584
5459	0.534386	0.465614
5460	0.425137	0.574863

```
5461   0.559608  0.440392
5462   0.496781  0.503219
5463   0.535039  0.464961
5464   0.576556  0.423444
5465   0.660363  0.339637
5466   0.496781  0.503219
5467   0.524625  0.475375
5468   0.537616  0.462384
5469   0.559608  0.440392
5470   0.456479  0.543521
5471   0.386400  0.613600
5472   0.443547  0.556453
5473   0.559608  0.440392
5474   0.406337  0.593663
5475   0.794996  0.205004
5476   0.474252  0.525748
5477   0.509051  0.490949
5478   0.279709  0.720291
5479   0.499943  0.500057
5480   0.559608  0.440392
5481   0.338740  0.661260
5482   0.372301  0.627699
5483   0.478340  0.521660
5484   0.453943  0.546057
5485   0.559608  0.440392

[5486 rows x 2 columns]
```

In [ ]:

```
preds_train[:, 1]
preds_test[:, 1]
preds_oot[:, 1]
```

APÊNDICE E - TRECHOS DE CÓDIGO EM PYTHON – SALVANDO AS VARIÁVEIS RESULTANTES DO XGBOOST

In [1]:

```
import pandas as pd
from scipy.stats import zscore, ks_2samp
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV, RFE
from sklearn import linear_model
from sklearn.model_selection import StratifiedKFold
import pickle
```

Read in the dataset containing the expert variables.

In [2]:

```
df = pd.read_parquet("Tabela_Referencias@20190401.parquet")
numerics = ['int32', 'int64', 'float64', 'bool']
expert_variables = df
#df.select_dtypes(include=numerics)
remove_list= ['month','churn', 'cf_score_minus_8','cf_score_minus_7','cf_score_minus_6','cf_score_minus_5','cf_score_minus_4','cf_score_minus_3','cf_score_minus_2','cf_score_minus_1','cf_score_last_8_mean','cf_score_last_7_mean','cf_score_last_6_mean','cf_score_last_5_mean','cf_score_last_4_mean','cf_score_last_3_mean','cf_score_last_2_mean','product_embed_0','product_embed_1','product_embed_2','product_embed_3','product_embed_4','product_embed_5','product_embed_6','product_embed_7','product_embed_8','product_embed_9','product_embed_10','product_embed_11','product_embed_12','product_embed_13','product_embed_14','product_embed_15','product_embed_16','product_embed_17','product_embed_18','product_embed_19','product_embed_20','product_embed_21','product_embed_22','product_embed_23','product_embed_24','client_embed_0','client_embed_1','client_embed_2','client_embed_3','client_embed_4','client_embed_5','client_embed_6','client_embed_7','client_embed_8','client_embed_9','client_embed_10','client_embed_11','client_embed_12','client_embed_13','client_embed_14','client_embed_15','client_embed_16','client_embed_17','client_embed_18','client_embed_19','client_embed_20','client_embed_21','client_embed_22','client_embed_23','client_embed_24']
selected_columns = [i for i in expert_variables.columns if i not in remove_list]
```

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 565326 entries, 0 to 565325
Columns: 397 entries, bk_contract to churn
dtypes: bool(1), float64(155), int32(233), int64(1), object(7)
memory usage: 1.2+ GB
```

In [4]:

```
pd.set_option('display.max_columns', 5)
```

In [5]:

```
df.describe()
```

Out[5]:

	month	sk_cliente_old_nunique_12_periods	...	churn_rate_for_same_sub_segment_last_1_months	total_contracts_with_same_sub_segment_last_1_months
count	565326.000000	565326.000000	...	565326.000000	565326.000000
mean	201804.118280	1.035311	...	0.023774	1371.328816
std	56.084484	0.639773	...	0.020059	931.498319
min	201709.000000	0.000000	...	0.000000	0.000000
25%	201802.000000	1.000000	...	0.015838	635.000000
50%	201806.000000	1.000000	...	0.020339	1209.000000
75%	201811.000000	1.000000	...	0.027404	2117.000000
max	201903.000000	60.000000	...	0.336767	3174.000000

8 rows × 389 columns

In [6]:

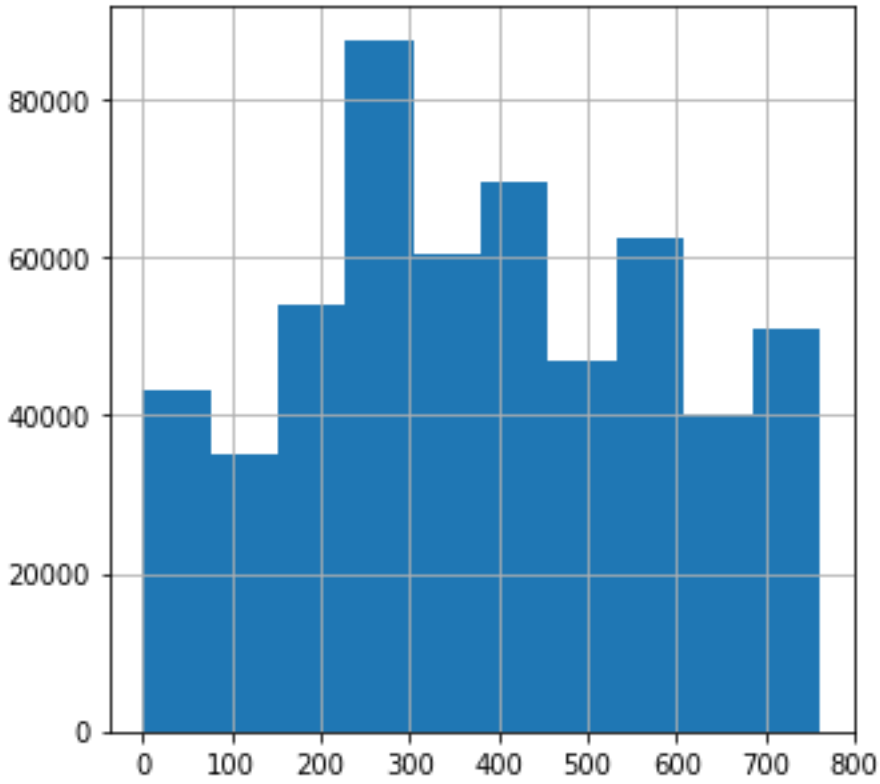
```
sum(df.duplicated())
```

Out[6]:

```
0
```

In [7]:

```
df['time_alive'].hist(figsize=(5,5));
```



In [8]:

```
expert_variables_scale = expert_variables.copy()

cols = list(expert_variables_scale.drop(remove_list, axis = 1).select_dtypes(includ
e=numerics).columns)

expert_variables_scale['time_alive'] = expert_variables_scale['time_alive'].fillna(
expert_variables_scale['time_alive'].mean())

expert_variables_scale['time_alive'].head()
```

Out[8]:

```
0      0.0
1    212.0
2    212.0
3    212.0
4    212.0
Name: time_alive, dtype: float64
```

Filter for training and out-of-time data.

In [9]:

```
in_train_valid = (expert_variables_scale['month'] < 201902) #& (expert_variables_sc
ale['Date'] < '2010-11-01')

in_oot = expert_variables_scale['month'] >= 201902

train_valid = expert_variables_scale[in_train_valid].sort_index() #.drop('Date', ax
is = 1)

oot = expert_variables_scale[in_oot].sort_index() #.drop('Date', axis = 1)
```

In [10]:

```
train_valid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 505414 entries, 0 to 505413
Columns: 397 entries, bk_contract to churn
dtypes: bool(1), float64(155), int32(233), int64(1), object(7)
memory usage: 1.1+ GB
```

In [11]:

```
oot.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 59912 entries, 505414 to 565325
Columns: 397 entries, bk_contract to churn
dtypes: bool(1), float64(155), int32(233), int64(1), object(7)
memory usage: 128.3+ MB
```

In [12]:

```
train_valid['time_alive'].head()
```

Out[12]:

```
0      0.0
1    212.0
2    212.0
3    212.0
4    212.0
Name: time_alive, dtype: float64
```

Z-scale all of the variables.

In [ ]:

```
train_valid[cols] = train_valid[cols].apply(zscore)

oot[cols] = oot[cols].apply(zscore)
```

```
/anaconda3/lib/python3.7/site-packages/scipy/stats/stats.py:2253: RuntimeWarning: i
nvalid value encountered in true_divide
  return (a - mns) / sstd
```

Get the shape of the training and out of time data.

In [ ]:

```
train_valid['time_alive'].head()
```

In [ ]:

```
print('Training Shape:', train_valid.shape)
```

In [ ]:

```
print('OOT Shape:', oot.shape)
```

In [ ]:

```
columns_to_save=['churn','month', 'bk_contract', 'churn_rate_for_same_product_last_
1_months','churn_rate_for_same_product_last_6_months','time_alive','total_contracts
_with_same_segment_last_6_months','churn_rate_for_same_product_last_2_months','chur
n_rate_for_same_sub_segment_last_6_months','vlr_nvvend_mean_6_periods','bow_contrat
o_last_7_months','ticket_grp_CST - Contratos de Vendas_last_7_months','tickets_of_p
riority_urgent_last_6_months','churn_rate_for_same_macro_segment_last_6_months','ch
urn_rate_for_same_segment_last_3_months','churn_rate_for_same_product_last_3_months
','bow_cancelamento_last_7_months','ticket_grp_other_last_7_months','churn_sum_6_pe
riods','total_contracts_with_same_segment_last_3_months','total_contracts_with_same
_segment_last_1_months','total_contracts_with_same_product_last_1_months','days_to_
carencia_start']
```

In [ ]:

```
new_df_1 = train_valid[columns_to_save]
new_df_1.shape
```

In [ ]:

```
new_df_2 = oot[columns_to_save]
new_df_2.shape
```

In [ ]:

```
new_df_1['churn'] = (new_df_1['churn'] == True ).astype(int)
new_df_1.info()
```

In [ ]:

```
new_df_2['churn'] = (new_df_2['churn'] == True ).astype(int)
new_df_2.info()
```

In [39]:

```
new_df_1.to_csv("training_top_20_xboost.csv", index=False)
```

In [40]:

```
new_df_2.to_csv("oot_top_20_xboost.csv", index=False)
```