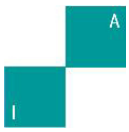
Inteligencia Artificial, 20(59) (2020), 123-137
doi: 10.4114/intartif.vol20iss59pp123-137



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Comparação de algoritmos de aprendizado de máqina na previsão do churn em uma empresa de tecnologia na área de educação continuada a distância

Gabriel Dillenburg Martins, Me. Diego Silva
Programa de Graduação em Engenharia de Produção
Centro Universitário Metodista (IPA).
Porto Alegre, RS, Brasil

Abstract The definition and modeling of consumer loyalty and satisfaction has been a central problem in managing customer relationships. Churn prediction models, in addition to working on solutions on this topic, aim to indicate users who have a high propensity of attrite, allowing for improvement of the efficiency of customer retention campaigns and reduction of costs associated with churn. Although cost reduction and retention is their main objective, churn prediction models are usually evaluated using statistical performance measures with computational tools, such as machine learning. The objective of the article was to develop and validate predictive churn models with data from more than 150,000 customers from a technology company in the field of distance continuing learning and to compare the performance of the different machine learning algorithms implemented for the task. After selecting 13 variables from the literature, the models were developed based on 4 steps: (I) training of balanced and unbalanced data sets; (II) generalization / testing on an independent data set; (III) statistical comparison of the algorithms; and (IV) evaluation of the models with higher accuracy. The models that presented the best performance were xgboost, for unbalanced classes, with an average accuracy of 86 % and 84 % of average area under the curve in the test step, and bagging classifier for balanced classes, with an average accuracy of 76 % and 81 % of average area under the curve in the test and generalization phase.

Resumo A definição e modelagem da lealdade e satisfação do consumidor tem sido um problema central na gestão do relacionamento com o cliente. Modelos de predição de churn, além de atuar em soluções neste tema, têm o objetivo de indicar os usuários que possuem uma alta propensão de atrito, permitindo melhoria da eficiência de campanhas de retenção de clientes e redução de custos associados ao churn. Embora a redução de custos e retenção seja seu objetivo principal, os modelos de previsão de churn são normalmente avaliados usando medidas de desempenho estatísticas com ferramentas computacionais, como aprendizado de máquinda. O objetivo do artigo foi desenvolver e validar modelos preditivos de churn com dados de mais de 150.000 clientes de uma empresa de tecnologia na área de educação continuada a distância e comparar a performance dos diferentes algoritmos de aprendizado de máquina implementados para a tarefa. Após a seleção de 13 variáveis a partir da literatura, os modelos foram desenvolvidos a partir de 4 etapas: (I) treinamento dos conjuntos de dados balanceados e desbalanceados; (II) generalização/teste em um conjunto de dados independente; (III) comparação estatística dos algoritmos; e (IV) avaliação dos modelos com maior acurácia. Os modelos que apresentaram as melhores performance foram o *xgboost*, para classes desbalanceadas, com acurária média de 86% e 84% de média de área sob a curva na etapa de teste, e o *bagging classifier*, para classes balanceadas, com acuráricia média de 76% e 81% de média de área sob a curva na fase de teste e generalização.

Keywords: Machine Learning, Churn prediction, Customer Satisfaction.
Palabras Clave: Aprendizado de Máquina, Predição de Churn, Satisfação do cliente.

ISSN: 1137-3601 (print), 1988-3064 (on-line)

1 Introdução

Estudiosos de marketing de relacionamento enfatizam que geralmente é mais caro conquistar novos clientes do que manter os já existentes [4]. Nessa perspectiva, a compreensão do fenômeno Churn, isto é, o abandono do cliente aos serviços prestados pela empresa, permite que a organização atue nos fatores principais que o influenciam, como satisfação do cliente, por exemplo, a fim de minimizá-lo [25].

No contexto geral, *Churn* tem relação com a retenção de clientes e pode ser definido como o número de indivíduos que deixam de se relacionar com a empresa em um determinado período. Embora a definição ampla e que abrange diferentes contextos, *Churn*, dependendo da ótica, pode ter outras definições. *Churn* (ou attrition) significa a interrupção do contrato de um cliente com uma empresa, em geral com troca para uma empresa concorrente. Assim, o gerenciamento de churn consiste na evolução de técnicas que permitam à empresa manter seus clientes mais lucrativos [3]. *Churn* (taxa de) pode ser uma medida que contribua com as organizações a identificarem o nível de insatisfação das suas bases de clientes em relação ao produto ou serviço prestado, preços melhores praticados pela concorrência, melhores performances comerciais dos competidores e outras razões que tenham associação com o ciclo de vida do cliente [25].

Assim, o primeiro passo para a análise de *Churn* é a classificação dos clientes e, uma vez identificados os clientes que se desligaram, a análise dos dados de seu relacionamento com a empresa pode permitir o desenvolvimento de modelos que determinem padrões de comportamento. Investigar grande volume de dados e construir modelos que visam a predizer o comportamento dos clientes, a mineração de dados utilizando modelos como redes neurais, modelos de regressão, árvores de decisão e outras técnicas estatísticas multivariadas vem a ser uma abordagem confiável. Todavia, é indispensável identificar previamente as variáveis de entrada (independentes) que podem ser úteis para prever o comportamento da variável de saída (dependente) em modelo preditivo de *Churn* [3].

Dessa forma, os autores relatam que geralmente as variáveis de entrada são:

1- Dados sobre o comportamento e perfil do cliente (provenientes de dados históricos);
2- Dados demográficos do cliente (idade, renda);
3- Dados sobre faturamento;
4- Dados de serviços solicitados (tipos de serviços usados, tempo de uso);
5- Informações e registros do call center (tempo de ligação, conteúdo da ligação);
6- Informações sobre o mercado (despesas com propaganda realizadas pela concorrência, por exemplo);

De maneira geral, grande parte da literatura ao longo do tempo destaca a influência da satisfação do cliente em sua lealdade com a empresa. Para Fornell et. al [9], a satisfação do cliente influencia diretamente na lealdade dos consumidores para com as organizações. Embora amplamente divulgada e aplicada como uma estratégia de negócio eficaz das empresas, gestão de relacionamento com o cliente (CRM) não possui uma definição formal universalmente aceita [25]. Para Gummesson [11], CRM são os valores e estratégias do marketing de relacionamento (MR), transformados em aplicação prática dependentes da ação humana e da tecnologia da informação. Neste contexto, a necessidade de atrelar a tecnologia no gerenciamento do relacionamento com o cliente surge como indispensável no objetivo de reter o consumidor. A lealdade do cliente é frequentemente interpretada como retenção, que é um dos pilares dos conceitos aplicados em CRM [12].

Para Akhila et al. [1], um dos principais objetivos de um típico sistema de gerenciamento de relacionamento, é classificar e prever um grupo de potenciais *Churners* de um grande conjunto de clientes para conceber campanhas de retenção rentáveis e direcionadas, na tentativa de manter um relacionamento de longo prazo com clientes valiosos. Perante o exposto, é possível dizer que objetivo de entender os padrões de comportamento que levam o cliente a deixar de se relacionar com a empresa *(Churn)*, são importantes para estratégia do modelo de negócio. Métodos estatísticos utilizados com técnicas da ciência da computação que auxiliam em análises preditivas, como aprendizado de máquina contribuem como parte fundamental nesse processo analítico.

Com a atual situação do mercado da educação no Brasil [8], vêm aumentando a necessidade de as empresas gerirem o relacionamento com o cliente de forma mais eficiente, visto que a concorrência neste setor vem crescendo e, assim, dificultando a retenção de clientes. Neste caso, conseguir identificar se o cliente está propenso a deixar de consumir o produto ou serviço da empresa e prever quando isso pode ocorrer pode ser de grande valor para organização, pois possibilita tomadas de decisão antecipadas

e até mesmo reconhecer problemas de produto, operacionais e inclusive os estratégicos que impactam diretamente no cliente.

Desse modo, diversos estudos atuais estão enfatizando técnicas de aprendizado de máquina com foco em problemas e aplicações reais, auxiliando no desenvolvimento de modelos preditivos que contribuem com as organizações em seus negócios.

O Grupo X, empresa da qual os dados foram utizados no presente trabalho, possui mais de 15 anos de atuação somente na área de educação continuada a distância, focando em diferentes campos de atuação de ensino, mas principalmente na área profissional da saúde. No momento da elaboração da pesquisa, a empresa já tinho tido mais de 250.000 clientes que em algum momento compraram o seu produto. Considerando que o autor atua profissionalmente na empresa e devido a lógica de assinatura de serviço do modelo de negócio da organização, promovendo uma alta rotatividade de clientes, o problema de pesquisa abordado foi entender quais as diferenças entre distintos modelos preditivos de aprendizado de máquina para prever o *Churn*.

2 Objetivo

Este artigo apresenta o desenvolvimento e a comparação de modelos preditivos criados a partir de diferentes técnicas de aprendizado de máquina, com a finalidade de prever o *Churn*, utilizando os dados de comportamento e compra de clientes de uma empresa de tecnologia na área de educação continuada a distância (Grupoa X Educação). O propósito das comparações estatísticas entre os diferentes algoritmos, além de analisar a viabilidade de ser implementado eventualmente na prática pelas organizações que promovem serviços educacionais, é permitir que estudos mais profundos sejam inciados através do presente estudo.

Entretanto, o objetivo geral do trabalho é analisar diferentes modelos preditivos de aprendizado de máquina para prever o *Churn*. Para atender o objetivo geral e o problema de pesquisa, foram determinados os objetivos específicos: identificar dados históricos de compra e comportamento de clientes relevantes para a construção dos modelos preditivos; identificar as previsões executadas pelos diferentes modelos preditivos; comparar os resultados dos modelos preditivos.

Não obstante, a compreensão e introdução de novos aspectos relevantes em adotar técnicas de aprendizado de máquina no setor educanional proporciona diferentes abordagens para soluções atreladas ao problema de previsão do *Churn* e retenção de clientes.

3 Revisão de Literatura

3.1 Gestão do relacionamento com o cliente

A gestão do relacionamento com o cliente (CRM) é uma prática importante para as organizações e é o ponto de partida para a fundamentação teórica deste presente trabalho. Conforme exposto na parte introdutória desta pesquisa, existem diversas definições que são abordadas ao explorar o conceito de CRM. Entretanto, de maneira geral, na literatura existe uma concordância coletiva que traduz estes conceitos não como uma tecnologia em si, mas como um movimento estratégico da organização.

Para Buttle [6], CRM é uma estratégia de negócios que maximiza a lucratividade, a receita e a satisfação do cliente com implementação de processos focados no consumidor, os quais possuem efeitos tanto nos âmbitos tático e estratégico como no operacional das empresas. Dentro deste contexto, CRM é uma estratégia de negócio que visa entender, antecipar e gerenciar as necessidades dos clientes atuais e potenciais de uma organização para melhorar sua retenção, lealdade e lucratividade [26].

3.2 Satisfação e Retenção de Clientes

A satisfação do cliente tem sido um considerável objeto de estudo e suas definições e importâncias são abrangentes. Conforme Oliveira [19], as organizações devem permanentemente avaliar se o desempenho da empresa está conciliado com as premissas estratégicas estabelecidas, as quais consideram que a satisfação do cliente é um dos grandes desafios para as organizações, pois é dessa satisfação que se traduz o primeiro passo para o processo de fidelização.

A criação de relacionamentos duradouros se sustenta na premissa de que conquistar novos clientes é muito mais oneroso que manter e estimular a base de clientes existentes [3].

3.3 Churn

Churn tem relação com a retenção de clientes e pode ser definido como o número de indivíduos que deixam de se relacionar com a empresa em um determinado período [25]. Segundo Buttle [6], aumentar a taxa de retenção ou diminuir a taxa de *Churn* aumenta consideravelmente a base de cliente ativos. Para o autor, as organizações alcançam melhores resultados quando gerenciam a base de clientes a fim de identificar, adquirir, satisfazer e reter os clientes mais lucrativos.

Conforme Kotler e Keller [14], para as organizações reduzirem as taxas de *Churn*, ao menos três ações devem ser tomadas: definir e medir a sua taxa de retenção (a); distinguir as causas de desgaste na relação com clientes identificando aquelas que podem ser mais bem administradas (b); e comparar o valor vitalício do cliente com os custos de redução da taxa de *Churn* (c). Para os autores, o item (a) depende do modelo de negócio e a área de atuação da empresa. Já para os item (b) e (c), o objetivo é focar nos motivos possíveis de serem tratados e naqules que o custo seja menor do que o lucro perdido, respectivamente.

3.4 Aprendizado de Máquina e Algoritmos

Aprendizado de máquina é uma área da inteligência artificial que visa desenvolver técnicas computacionais sobre o aprendizado, bem como fazer com que essas técnicas construam sistemas capazes de aprender e adquirir conhecimento de forma automática. Podemos dizer que aprendizado de máquina utiliza técnicas dos ramos da estatística e computação que compreendem dois objetivos principais: o primeiro é o desempenho preditivo de modelos e o segundo é automatizar o processo de modelagem das bases de dados observados ou aprendizado com os dados observados [28].

Fundamentalmente, o campo de aprendizado de máquina pode ser dividido em três principais categorias: aprendizagem não-supervisionada, supervisionada e por reforço.

A primeira categoria, intitulada como aprendizagem não-supervisionada, é guiada pelos dados fornecidos, uma vez que não é necessário conhecimento prévio sobre as classes existentes. O aprendizado supervisionado tem o objetivo de deduzir conceitos sobre os dados, com base nas instâncias rotuladas apresentadas (exemplos) e, por fim, a aprendizagem por reforço é a mais geral das três categorias que, em vez de ser informado sobre o que fazer por um instrutor, um agente de aprendizagem por reforço deve justamente aprender a partir do reforço, onde o agente de aprendizagem experimentando movimentos aleatórios, poderá eventualmente construir um modelo de previsão de seu ambiente [24].

De modo geral, os algoritmos de aprendizagem podem ser classificados de duas maneiras, segundo o modo em que os exemplos são apresentados [22]:

1- Não-incremental: necessita que todos os modelos estejam disponíveis de maneira simultânea para que o modelo de aprendizagem induza um conceito. Em problemas de aprendizagem em que todos os exemplos estão disponíveis e que provavelmente não irão ocorrer mudanças, é vantajoso usar esses algoritmos.

2- Incremental: aqui o indutor não necessita construir a hipótese a partir do início, no qual é possível adicionar novos exemplos ao conjunto de treinamento. Dessa maneira, no modo incremental o indutor apenas tenta atualizar a hipótese antiga sempre que novos exemplos são adicionados ao conjunto de treinamento.

No mundo real, é comum trabalharmos com dados imperfeitos. Quando esses dados imperfeitos (exemplos com os mesmos valores de atributos, mas com classes diferentes) são derivados do próprio processo de aquisição que os dados são gerados ou até mesmo quando o problema ocorre na etapa de transformação, se diz que existem ruídos nos dados [20].

Os algoritmos utilizados serão descritos brevemente a seguir:

3.4.1 Logistc Regression

Técnica de modelagem estatística de classificação utilizada para análises preditivas de churn [27]. Geralmente é usada como referência em comparação com outras técnicas para análises de dados no departamento de inteligência de negócios das empresas.

3.4.2 Decision Tree

Técnica bem conhecida e teve muitas aplicações bem-sucedidas para problemas do mundo real. Árvore de decisão é uma técnica simbólica de aprendizado que organiza as informações extraídas de um conjunto de dados de treinamento em uma estrutura hierárquica composta por nós e ramificações. A saídas de uma árvore de decisão podem ser organizadas na forma de uma árvore, sendo assim considerada um dos *algoritmos de machine learnig* com uma maior facilidade de interpretabilidade. Além disso, uma árvore de decisão tem a capacidade de criar modelos usando dados numéricos ou categóricos [15].

3.4.3 Random Forest

É um algoritmo de aprendizado de máquina da família de algoritmos *"ensemble"* (em conjunto). Um método de conjunto é uma técnica que combina as previsões de aprendizado de máquina com múltiplos algoritmos juntos para fazer previsões mais precisas do que qualquer modelo trabalhando individualmente. A ideia principal do Random Forest é construir uma "floresta" de "árvores" de decisão aleatória e usar para novas classificações [20]

3.4.4 Bagging Classifier

A Agregação de Bootstrap (ou *Bagging* para abreviar) é um método em conjunto simples e muito poderoso. A agregação de Bootstrap é um procedimento que pode ser usado para reduzir a variação de algoritmos que têm alta variação. Neste método, são gerados conjuntos sucessivos e independentes de amostras a partir do conjunto de treinamento. *Bagging* com árvores de decisão são menos propensos de sobreajuste (*overfitting*) aos conjuntos de treinamento [29].

3.4.5 Adaboost Classifier

Um classificador AdaBoost (*Adaptive Boosting*) trabalha com a combinação de classificadores modelados por um mesmo algoritmo (conhecido como algoritmo fraco) e sua execução é ajustado de acordo com os erros cometidos pelo classificador anterior [2].

Em geral, o diferencial de algoritmos *boosting* é a busca por gerar novos classificadores melhores e mais adequados para o problema, principalmente por, de forma isolada, corrigirem e aumentarem a eficiência dos classificadores gerados. A principal diferença entre o *Bagging* e *Boosting*, é que este último gera conjuntos de treinos e classificadores de forma sequencial, baseado nos resultados das interações anteriores. Já o *Bagging* gera esses conjuntos de forma aleatória e pode gerar os classificadores de forma paralela [7].

3.4.6 XGBoosting Classifier

Xgboost (*Extreme Gradient Boosting*) é uma extensão regularizada das tradicional técnicas "*ensemble*"que pertencem à família CART (*classification and regression tree*). Como um método *boosting* de árvores, sua função fundamental prevê uma nova associação de classificação após cada iteração. Isso é feito de maneira aditiva, o que significa que as previsões são feitos de classificadores fracos que melhoram constantemente o erro dos classificadores anteriores. Amostras classificadas incorretamente receba pesos mais altos na próxima etapa, forçando o classificador para se concentrar em seu desempenho nas seguintes iterações [23].

4 Métodos

Caracterizada como descritiva, para Gil (2008) [10], a pesquisa descritiva descreve as características de determinadas populações ou fenômenos, uma de suas peculiaridades está na utilização de técnicas padronizadas de coleta de dados, tais como o questionário e a observação sistemática. O presente trabalho também é caracterizado como uma pesquisa de classificação quantitativa, empregando técnica de coleta documental e de análises estatísticas, dispondo de dados tão somente da empresa de educação continuada (Grupo X Educação), com informações de compra e comportamento de clientes, originalmente emitidas a partir de 2013 e armazenadas em banco de dados pela própria organização.

Após a coleta de dados, as etapas de seleção, codificação, tabulação e análise serão elaboradas através da linguagem de programação Python e suas principais bibliotecas disponíveis para análise de dados, análise exploratória, visualização de dados e aprendizado de máquina, como Pandas, Numpy, Matplotlib e Scikit-learn, respectivamente.

Para a implementação dos algoritmos do modelo preditivo, será utilizado a plataforma de ambiente computacional web Jupyter Notebook. Neste ambiente é possível implementar os códigos das bibliotecas e visualizar dados e resultados de análises.

4.1 Conjunto de dados e seleção preliminar de variáveis

Os dados dos clientes do Grupoa X Educação foram usados para criar os modelos preditivos de *Churn*, assegurando-se a privacidade dos usuários que consomem ou consumiram os produtos da empresa. A base de dados contempla 407.441 observações de compra/assinatura (linhas) de 157.336 clientes, sendo essas então informações que caracterizam o momento do usuário no período de cada compra efetuada.

Atributos como canal de compra, produto, área profissional de atuação, renda, tempo de permanência, acesso aos produtos digitais, forma e condição de pagamento, região de residência, entre outros, foram considerados na coleta como relevantes para a construção dos modelos preditivos.

Das 407.441 assinaturas, 90.658 (22,2%) foram canceladas. Os cancelamentos ocorrem quando solicitado diretamente pelo cliente através dos canais de atendimento, ou de maneira motivada pela própria empresa, quando há inadimplência por parte do usuário do produto adquirido.

Através da aquisição de variáveis no conjunto de dados do Grupo X, que foram semelhantes às indicadas em modelos preditivos publicados na literatura anteriormente para prever o churn ou situações de alto risco de cancelamento [3], foram selecionadas 13 fatores de risco para análise.

A coleta e seleção manual das variáveis foram executadas priorzando a privacidade dos usuários, assim quaisquer informações coletadas do banco de dados da empresa se limitaram a ser sobre comportamento do usuário em relação aos produtos e serviços fornecidos pela empresa, bem como de informações tão somente operacionais.

Quaisquer variáveis que implicavam em alto custo de processamento devido a alta cardinalidade foram excluídas para que os modelos obtidos pudessem ser aplicados com méticras simples e com menor tempo na fase de treinamento. Esse processo ocorreu no momento de coleta e análise exploratória dos conjuntos de dados.

O subconjunto de dados final foi então validado e, com isso, resultou na base de dados final das 13 variáveis candidatas conforme mensionado anteriormente. Mais detalhes das variáveis estão descritos na Tabela 1 e na Tabela 2. A Tabela 2 também apresenta a variável alvo da análise de previsão do *churn*, *"Churn"*.

A base de dados original foi dividida randomicamente em duas partes com o rateio de 75:25. A primeira parte (75%) da base de dados foi utilizada para o trinamento e a segunda (25%) para os testes do modelo na fase de generalização.

O processo de treinamento ocorreu também com o conjunto de dados balanceado após o pré-processamento da base original. Os modelos foram comparados e avalidados sobre a média de acertos na previsão correta do *churn*. Os algoritmos de *machine learning*, *logistic regression, decision tree, random forest, bagging (bootstrap aggregation), adaboost e xgboost* foram utilizados com os parâmetros-padrão de suas bibliotecas [16] [21].

Tabela 1: Variáveis Numéricas.

Variável	Mínimo	Mediana	Média	Máximo	Desvio Padrão	Descrição
Renda	2300.0	8900.0	6348.4	8900.0	2991.3	Renda do cliente baseada na profissão
Freq acum	1.0	3.0	5.24	127.0	6.62	Frequência acumulada de compras
Temp prim compra	0.0	2.0	3.33	15.3	3.82	Tempo desde a primeira compra
Temp ult compra	0.0	0.9	0.78	15.1	1.25	Tempo desde a última compra

Tabela 2: Variáveis Categóricas.

Variável	Número de níveis	Frequência em cada nível	Descrição	Valores possíveis
Novo Cliente	2	0:277.577 1: 129.864	Indica se o cliente é novo ou antigo	0: Não 1: Sim
Região	5	1: 36.508 2: 97.579 3: 39.555 4: 161.294 5: 72.505	Regiões que contemplam as unidades federativas do cliente	1: Centro-oeste 2: Nordeste 3: Norte 4: Sudeste 5: Sul
Pagou prim	2	0: 59.205 1: 348.236	Pagou a primeira parcela da assinatura	0: Não 1: Sim
Área	7	1: 86.730 2: 2.016 3: 55.536 4: 235.506 5: 9.948 6: 7.161 7: 10.544	Área profissional em que o produto é focado	1: Enfermagem 2: Farmácia 3: Fisioterapia 4: Medicina 5: Nutrição 6: Psicologia 7: Veterinária
Canal	5	1: 28.074 2: 18.685 3: 20.888 4: 217.995 5: 121.799	Canal de entrada para aquisição de um novo ciclo do produto	1: Call Center 2: E-commerce 3: Receptivo 4: Renovação 5: Representantes
Forma pgmt	3	1: 49.974 2: 354.841 3: 2.626	Forma de pagamento adotada pelo cliente	1: Cartão de crédito 2: Débito em conta 3: Outros
Cond pgmt	3	1: 12.380 2: 35.365 3: 359.696	Condição de pagamento adotada pelo cliente	1: À vista 2: Parcelado 6x 3: Parcelado 12x
Acesso portal	2	0: 253.902 1: 153.539	Indica se o cliente acessou o portal online com conteúdo digital	0: Não 1: Sim
Churn	2	0: 316.783 1: 90.658	Indica se o cliente cancelou a assinatura	0: Não 1: Sim

4.2 Preparação dos dados

4.2.1 Trasformação das variáveis categóricas

Quando necessário, a transformação dos tipos de dados categóricos foi realizada por meio de binarização. Nessa fase de pré-processamento, as variáveis categóricas com n níveis são tranformadas em n-1 *dummy* variáveis, as quais possuem os valores igual a "1"quando o novo ponto pertence ao nível representado pela variável *dummy*, caso contrário, igual a "0" [17].

4.2.2 Balanceamento das classes

Como descrito no início da metodologia, as classes (variável alvo) da base de dados original não estão balanceadas, com apenas 22,2% das observações sendo de clientes que cancelaram o produto. Contudo, econtram-se diversas soluções para o problema de balanceamento na literatura, de modo que a maioria dos algoritmos tradicionais assumem erros supondo que as distribuições das classes são relativamente equilibradas [13]. Com isso, o algoritmo SMOTE (*Synthetic Minority Over-sampling Technique*) [18] foi utilizado no conjunto de treinamento para igualar a distruibuição das classes, aumentando o volume da classe minoritária, criando um novo conjunto de treino.

4.2.3 Processo geral de construção e avaliação do modelo

Os modelos foram construídos, avaliados e comparados considerando os 4 passos na sequência:

1- Treinamento dos conjuntos balanceados e desbalanceados;
2- Generalização/teste em um conjunto de dados independente;
3- Comparação estatística dos algoritmos;
4- Avaliação dos modelos com maior acurácia;

O processo completo é demonstrado na Figura 1. Primeiramente, a seleção manual de variáveis foi realizada conforme descrito na parte introdutória da metodologia (caixa "Seleção manual das variáveis"na Figura 1). Após a seleção, 25% do conjunto de dados (base "Teste"na Figura 1), contendo 101.860 observações, foi separado para generalização e teste, enquanto a outra parte (base "Treino"na Figura X), contendo 305.580 observações, para treinamento.

A base de treinamento gerada inicialmente pelo conjunto de dados bruto manteve a distribuição original das classes de 78:22, sendo 237.562 observações da classe "0"e 68.018 da classe "1". De acordo com o exposto na subseção "Balenceamento de classe", grande parte dos algoritmos tendem a trabalhar melhor com conjunto de dados balanceados. Com isso, o conjunto de trainamento original foi balanceado, mantendo uma distruibição entre as classes de 50:50, sendo 237.562 observações para cada classe. Assim é possível comparar a performance dos algoritmos de aprendizado trabalhando com conjuntos de treinamento balanceados e não-balanceados.

No segundo momento, foi realizado a generalização dos algoritmos com base nas bases de dados de treino. Nessa etapa, para todos os algoritmos foi realizado o processo através da técnica "validação cruzada", com um número total de 10 *folds* e configurados com a métrica *"accuracy"* para o parâmetro *scoring*, assim possibitando a comparação da acurácia dos modelos preditivos [5].

A terceira etapa consiste na comparação dos resultados obtidos dos modelos, avaliando a média de acertos do algoritmo, quando no momento de generalizar a novos dados e observações, estes predizem corretamente as classes observadas. Também é avaliado o desvio padrão dos resultados para cada modelo, comparando-os entre si independentemente da base de treino em que o algoritmo obteve o aprendizado dos dados.

Finalmente, foi efetuada uma avaliação mais profunda dos modelos que tiveram a acurácia mais alta, tanto para aquele que foi treinado em um conjunto de treino balanceado ou não.

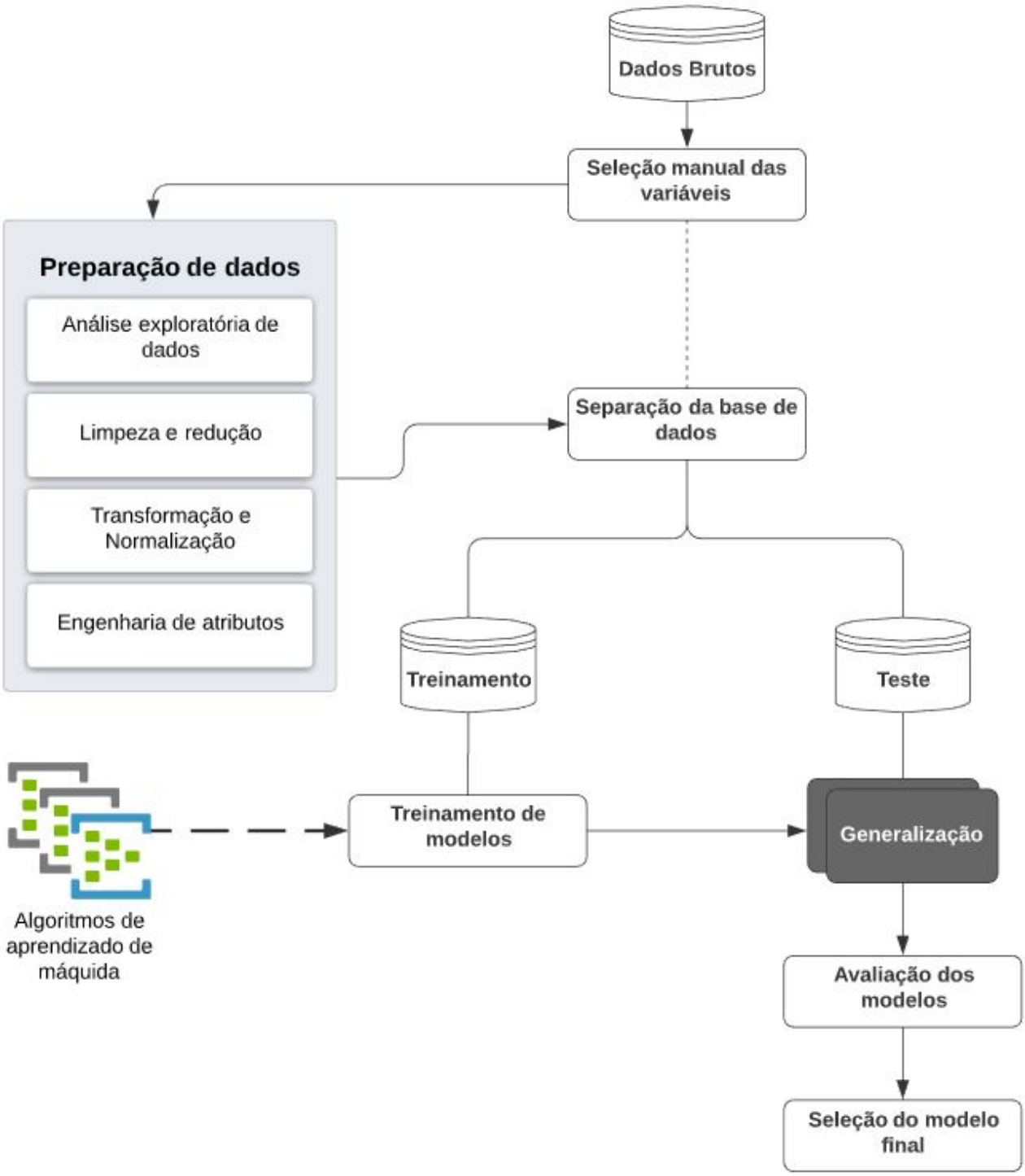


Figura 1: Processo de desenvolvimento e avaliação dos modelos de aprendizado de máquina.

5 Resultados

5.1 Amostra de estudo

Entre as 407.441 observações da base de dados utilizada na presente pesquisa, econtram-se 157.336 usuários dos serviços da empresa Grupo X Educação com compras e cancelamentos a partir de Janeiro de 2013 até Abril de 2020. A renovação automática da assinatura (encontrado no atributo "canal"da base de dados) é a maior proporção das observações de compra dos clientes, visto que é um processo característico e relevante para entender o contexto e modelo de négócio da organização, concentrando 53% das inscrições nos serviços da companhia. Outros detalhes sobre a amostra é possível ser encontrado na Tabela 1 e Tabela 2.

5.2 Resultados da gerenalização dos modelos

Os resultados desta pequisa estão organizados na Tabela 3 e seguiram as etapas descritas na seção metodológica. Em todos os modelos, sem exceção, foram utilizados a mesma proporção e o mesmo estado aleatório de saparação dos conjuntos de dados, possibilitando todos os modelos trabalharem exatamente com as mesmas informações.

A métrica utilizada para seleção de modelos durante o aprendizado e generalização foi a acurácia, portanto, o modelo com maior acurácia dentre aqueles que foram treinados com um conjunto de treinamento desbalanceado, foi selecionado para uma comparação mais detalhada com aquele que teve um melhor desempenho com os conjuntos de treino balanceados. Em suma, a avaliação de performance dos modelos selecionados evidenciou desempenho satisfatório, com uma acurácia média geral maior que 79% dos modelos implementados.

Para os modelos que foram treinados com as classes desbalanceadas, a média da acurácia foi de 85.7%, já para os que treinaram com as classes balanceadas tiveream um desempenho médio inferior, de 73.5%. Os modelos selecionados para uma comparação mais detalhada foram os que tiveram a maior acurácia dentre aqueles que foram treinados com o mesmo conjunto de dados, sendo eles: o algoritmo XGboost, com uma acurácia média de 86.61%, treinado com classes desbalanceadas, e o algoritmo Bagging Classifier com média de 76.84% de acurácia, treinado com classes balanceadas.

Nas Figuras 4 e 5, é possível obsevar a comparação e distribuição das médias de acertos das predições de cada algoritmo utilizado no presente artigo. Para facilitar a visualização gráfica, os algoritmos foram apresentados com abreviações, as quais são relacionadas a seguir: (LR) Logistic Regression, (CART) Decision Tree Classifier, (RFC) Random Forest Classifier, (BAGG) Bagging Classifier, (ADA) AdaBoost Classifier e (XGB) XGBoost Classifier.

Tabela 3: Resultados dos modelos.

Algoritmo	Balanceamento	Média	Desvio padrão
XGB	Não	86.61%	0.014
XGB	Sim	71.92%	0.098
LR	Não	86.57%	0.013
LR	Sim	69.88%	0.100
ADA	Não	86.53%	0.013
ADA	Sim	69.86%	0.099
RFC	Não	85.44%	0.016
RFC	Sim	76.38%	0.029
BAGG	Não	85.25%	0.016
BAGG	Sim	76.84%	0.024
CART	Não	84.26%	0.017
CART	Sim	76.14%	0.023

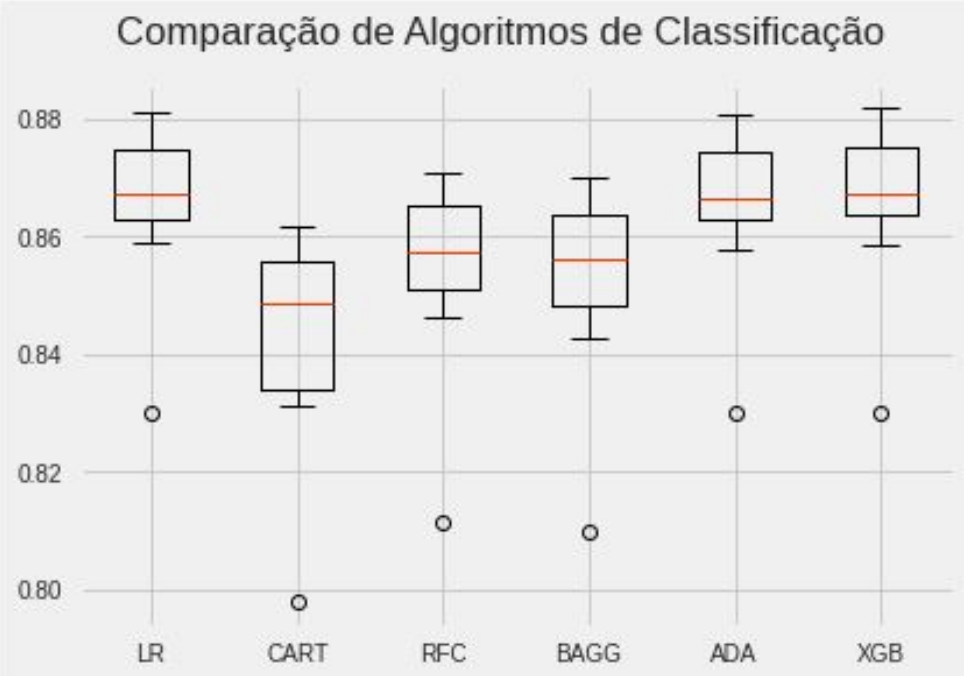


Figura 2: Comparação de modelos com base de treino desbalanceada.

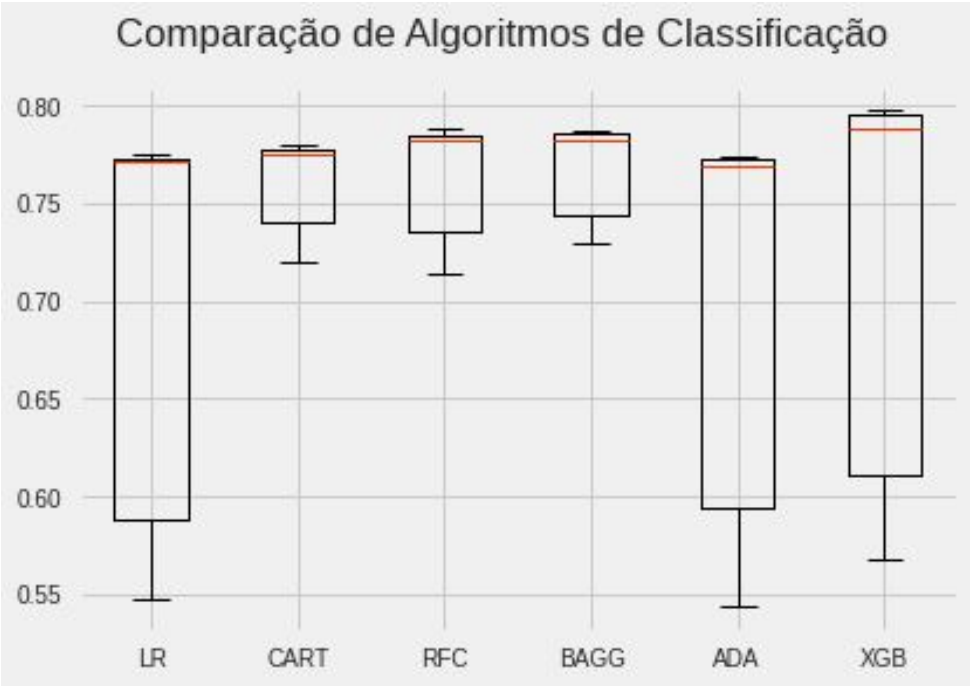


Figura 3: Comparação de modelos com base de treino balanceada.

É possível observar que existe uma maior variação das médias de acertos nas predições nos modelos com classes balanceadas, principalmente nos algoritmos *Logistic Regression, Adaboost e XGBoost.* Todavia, é necessário avaliar mais profundamente com outras métricas, além da acurácia, para entender a

performance do algoritmo e a sua capacidade de generalizar para novas observações. Na próxima seção será a comparação entre os algoritmos que tiveram um melhor resultado na primera verificação.

5.3 Avaliando os modelos com melhor performance

A área sob a curva ROC (ou AUC para abreviar) é uma métrica de desempenho para problemas de classificação binária. A AUC representa a capacidade de um modelo de discriminar entre classes positivas e negativas. Uma área de 1,0 representa um modelo que fez todas as previsões perfeitamente. Uma área de 0,5 representa um modelo tão bom quanto aleatório. ROC pode ser dividido em sensibilidade e especificidade. Um problema de classificação binária é realmente uma relação entre sensibilidade e especificidade [5]. Abaixo descrito com mais detalhes:

1- Sensibilidade: é a taxa de verdadeiros positivos (*True positive rate*), também chamada de *recall*. É o número de instâncias da classe positiva (primeira) que realmente previu corretamente.

2- Especificidade: também é chamada de taxa de falsos positivos. É o número de instâncias da classe (segunda) negativa que foi realmente prevista corretamente.

Analisando o gráfico da Figura 4, que representa o modelo aplicado com o algoritmo *Xgboost*, que foi treinado com classes desbalanceadas, e o gráfico da Figura 5, que representa o algoritmo *Bagging* treinado com conjunto de dados balanceados, é possível perceber que o modelo com *Xboost* possui um desempenho superior, pois a área sob a curva é maior. Em resumo, área sob a curva do *Xboost* para as classes "0"e "1"foi de 0.84, enquanto para o modelo *Bagging* foi de 0.81.

De modo geral, os dois algoritmos apresentaram uma performance ralativamente parecida. Entretanto, o *Xgboost* se demonstrou mais eficiente para a previsão do *Churn*, visto que este apresentou a maior acurácia entre todos os modelos analisados na pesquisa, a maior área sob a curva e, acima de tudo, apresentou-se mais flexível para trabalhar com classes desbalanceadas, que é uma realidade na prática dos conjuntos de dados reais.

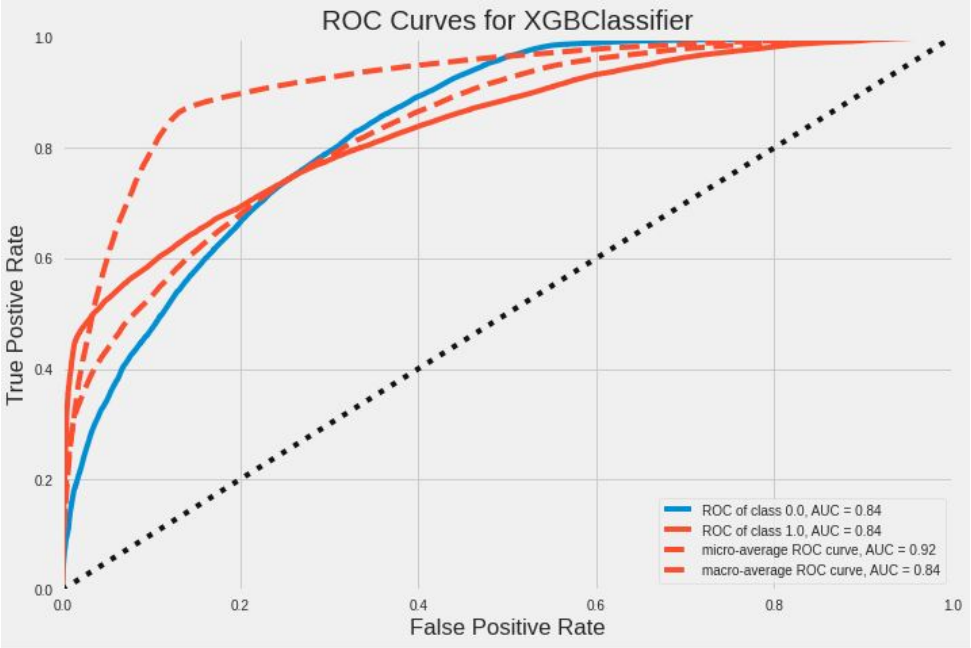


Figura 4: Comparação de modelos com base de treino balanceada.

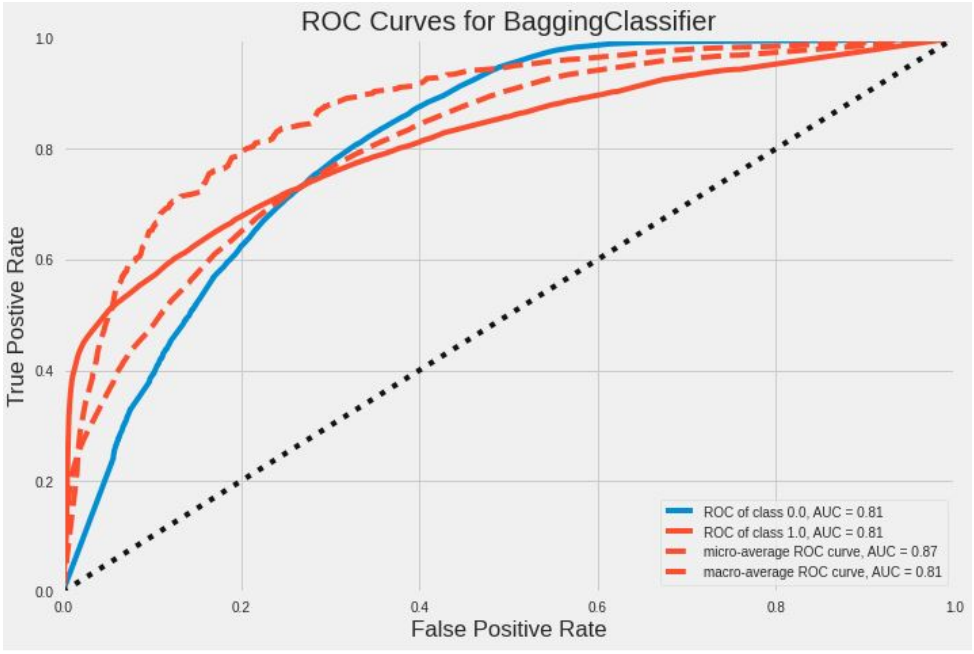


Figura 5: Comparação de modelos com base de treino balanceada.

6 Conclusão

A comparação entre diferentes famílias de algoritmos mostrou que todos eles produzem uma relativa similaridade em seus resultados utilizando os mesmos conjuntos de dados, possibilitando assim a viabilidade para a previsão do *Churn*. O modelo preditivo final pra prever o *Churn* foi o *Xgboost*, apesar de performar melhor sem o balanceamento das classes.

Contudo, todos os algoritmos apresentaram uma acurácia média acima de 84% partindo ao menos de um conjunto de treinamento, sugerindo que, com outras técnicas mais avançadas de pré-processamento e de ajustes finos em parâmetros, é possível implementar um modelo com alto desempenho para prever o *Churn* em uma empresa de educação continuada a distância.

Por fim, com os dados apresentados e analisados foi possível alcançar os objetivos propostos da pesquisa. De modo que, em função da limitação de tempo, não foi possível utilizar-se de todos os recursos disponíveis para a otimização de cada modelo aplicado. Desse modo, como sugestão de futuras pesquisas, se propõe os ajustes de parâmetros de cada algoritmo, bem como uma transformaçao e tratamento mais robusto dos dados para aperfeiçoamento e otimização das predições.

Referências

[1] Akhila, G Adhipathy.K., and J. Pamina. Analysing the behaviour of customers to predict churn in telecom sector. *International Journal of Emerging Technology and Innovative*, 2019.

[2] T. An and M. Kim. A new diverse adaboost classifier. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 359–363, 2010.

[3] Edson. BARRETO, Iná Futino; CRESCITELLI. *Marketing de relacionamento: como implantar e avaliar resultado.* Pearson Education do Brasil, 1 edition, 2013. ISBN 8581431844.

[4] Delane Botelho and Frederico Damian Tostes. Modelagem de probabilidade de churn. *Revista de Administração de Empresas*, 50:396 – 410, 12 2010.

[5] J. Brownlee. *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-end*. 2016.

[6] Francis Buttle and Stan Maklan. *Customer Relationship Management: Concepts and Technologies*. 01 2015.

[7] Bruno Butilhão. Chaves. Estudo do algoritmo adaboost de aprendizagem de máquina aplicado a sensores e sistemas embarcados. *Dissertação Mestrado. - São Paulo. Universidade de São Paulo. Escola Politécnica*, 2011.

[8] Silveira Daniel. Em meio à crise, mercado de educação é o que mais cresce em número de empresas no brasil, diz ibge. *Jornal G1, Globo*, 9 2019. <https://g1.globo.com/economia/noticia/2019/06/26/em-meio-a-crise-mercado-de-educacao-e-o-que-mais-cresce-em-numero-de-empresas-no-brasil-diz-ibge.ghtml>.

[9] Claes Fornell, Michael D. Johnson, Eugene W. Anderson, Jaesung Cha, and Barbara Everitt Bryant. The american customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing*, 60(4):7–18, 1996.

[10] Antônio Carlos. Gil. *Métodos e técnicas de pesquisa social*. Atlas, 2008.

[11] Evert GUMMESSON. *Marketing de Relacionamento Total*. Bookman, 3 edition, 2010. ISBN-13 978-8577806270.

[12] Anders Gustafsson, Michael Johnson, and Inger Roos. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing - J MARKETING*, 69:210–218, 10 2005.

[13] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[14] Keller Kevin Lane . Kotler, Philip. *Administração de Marketing*. Editora Pearson, 2018.

[15] Maria Carolina Monard and José Augusto Baranauskas. Conceitos sobre aprendizado de máquina. In *Sistemas Inteligentes Fundamentos e Aplicações*, pages 89–114. Manole Ltda, Barueri-SP, 1 edition, 2003.

[16] Website Oficial. Machine learning in python. *Scikit-learn*, 2020. <https://scikit-learn.org/stable/>.

[17] Website Oficial. Machine learning in python - one hot encoder. *Scikit-learn*, 2020. <https://bit.ly/2Y8F0LH>.

[18] Website Oficial. Synthetic minority over-sampling technique (smote). *Inbalanced Learn*, 2020. <https://bit.ly/3dHNmAr>.

[19] Marcos Antonio. Oliveira. Monitoramento da satisfação de cliente em contexto business-to-business: um survey em empresas com certificado ISO 9001-2000 no Estado de São Paulo. *Dissertação de Mestrado - Escola Politécnica da Universidade de São Paulo. São Paulo*, 2007.

[20] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 154–168, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[21] Website Oficial XGBoost Python Package. *XGBoost Documentation*, 2020. <https://xgboost.readthedocs.io/en/latest/index.html>.

[22] Ronaldo Cristiano Prati. Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos. 2006.

[23] Butch Quinto. *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*. 01 2020.

[24] S. Russell and P. Norvig. *Inteligência artificial*. CAMPUS - RJ, 2004.

[25] Pedro Henrique. SCHNEIDER. Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão. *Dissertação (Mestrado em Matemática Aplicada) - Escola de Matemática Aplicada, Fundação Getúlio Vargas - FGV, Rio de Janeiro*, 2016.

[26] R. Swift. *CRM: customer relationship management: o revolucionário marketing de relacionamento com o cliente*. Campus, 2001.

[27] Thanasis Vafeiadis, Kostas Diamantaras, G. Sarigiannidis, and Konstantinos Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 06 2015.

[28] Bruno. Vasconcelos. Poder preditivo de métodos de Machine Learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países . *Dissertação de Mestrado - Universidade de Brasília, Brasília.*, 2017.

[29] Masoumeh Zareapoor and Pourya Shamsolmoali. Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48:679–686, 12 2015.