ISSN: 2317-0840

Modelagem e análise da satisfação dos consumidores de uma instituição financeira utilizando aprendizado de máquina

Igor Caetano Silva[1†], Ricardo Menezes Salgado[2]

[1] *Programa de Pós-Graduação em Estatística Aplicada e Biometria, Unifal-MG.*
[2] *Departamento de Ciência da Computação, Universidade Federal de Alfenas (Unifal-MG).*

Resumo: *Com um mercado cada vez mais competitivo, se torna imprescindível ter um bom relacionamento com seus clientes, fator que pode gerar uma otimização de recursos e imensa vantagem competitiva. Dentre as áreas que mais dependem deste relacionamento e marketing individual para retenção de seus consumidores estão os bancos e empresas de seguro. Dessa forma, o presente artigo tem como objetivo a criação de um modelo de aprendizado de máquina que possa prever a satisfação dos consumidores de uma instituição financeira multinacional com base em suas características e comportamento descritos pela base de dados fornecida pela organização. Solução que permitirão o banco agir com contramedidas e reverter casos de insatisfação, identificar e corrigir suas causas. Como proposta de resolução foram apresentados três modelos de aprendizado de máquina (Árvore de Decisão, Floresta Aleatória e XGBoost) que tiveram uma performance máxima de 82,02%, 81,61% e 83,39%, performances consideráveis e eficientes ao se considerar a natureza do problema e do conjunto de dados.*

Palavras-chave: Aprendizado de Máquina; Árvore de Decisão; Floresta Aleatória; Satisfação do Consumidor; XGBoost.

Abstract: *With a more and more competitive market, it becomes necessary to maintain a good relationship with consumers, a fact that can generate optimization of resources and an immense competitive advantage. Among the markets that depend most on this relationship to keep their customers, make them generate value through word-of-mouth marketing and avoid evasion to competitors are banks and insurance companies. Thus, this article has the objective of creating a model that can predict the satisfaction of a financial institution's customers, through its characteristics and behavior described in a database provided by the bank. In a way that allows the company to act with countermeasures to reverse the cases of dissatisfaction. As a proposed solution, three machine learning models were used (Decision Tree, Random Forest and XGBoost) that had a maximum performance of 82,02%, 81,61% and 83,39%, a considerable and efficient performance due the nature of the problem and database.*

Keywords: *Customer Satisfaction*; *Decision Tree*; *Machine Learning*; *Random Forest*; XGBoost.

Introdução

Desenvolver uma relação de qualidade com os clientes é a chave para obter sucesso em qualquer ramo da indústria, e, com o elevado grau de competitividade, o monitoramento desse relacionamento se torna cada vez mais essencial. Desse modo, junto ao crescimento do número de alternativas para aquisição, os consumidores estão crescentemente mais orientados ao valor agregado do que consomem (WOODRUFF, 1997, p. 139). Assim, identificar e analisar causas de descontentamento são ações importantes para a retenção e fidelização de seus clientes.

Visto que há custos ligados à introdução e à atração de novos clientes para o negócio, quanto mais o cliente permanecer com uma organização mais valor ele agrega. O desperdício de valor agregado nasce justamente quando a empresa utiliza recursos em atividades, processos e melhorias que não estão alinhadas com as prioridades de seus clientes (MITTAL, 2020). A retenção de clientes otimiza a rentabilidade principalmente por reduzir custos relacionados à prospecção de novos clientes, sendo assim, o objetivo de estratégias de retenção deve ser minimizar a deserção de clientes rentáveis (REICHHELD, 1996, p. 56).

† Autor correspondente: igor.caetano@sou.unifal.edu.br.

Rosenberg et al. (2017) aponta que a lealdade e satisfação do consumidor são vitais por duas razões. Primeiramente são um recurso escasso, no qual é mais fácil manter um cliente antigo que conquistar um novo. E segundo, a lealdade e satisfação de um cliente tem um efeito positivo na rentabilidade da companhia.

Outro fator a ser considerado é o apontamento de Fornell (1996, p. 6) que mesmo sendo importante para todos os negócios, a satisfação do consumidor é especialmente importante para manter a lealdade de clientes em negócios como bancos e companhias de seguro. Neste mesmo pensamento, Ioanna (2002, p. 62) afirma que é quase impossível criar produtos e serviços diferenciados em um ambiente competitivo, uma vez que, os bancos oferecem os mesmos serviços com pequenas variações. Sendo assim, o diferencial fica a cargo do gerenciamento e qualidade do serviço prestado.

Dessa forma, o crescimento exponencial da evolução e uso da tecnologia e da internet possibilita oportunidades de desenvolvimento de soluções para diversas áreas e integração entre as próprias. Do mesmo modo, com o poder de armazenar históricos e monitorar o presente, tem-se a possibilidade de, através de análises, modelos e estratégias, criar oportunidades para prever possíveis cenários no futuro através de ferramentas de classificação e automações que auxiliem futuras tomadas de decisão. Assim, torna-se imprescindível por parte das companhias, a utilização dessas tecnologias para conhecer as demandas e expectativas de seus consumidores para garantir sua estadia.

Entretanto, dados considerados para análise efetiva da satisfação do consumidor não tem apenas grande variedade (diferentes tipos de dados, como variáveis de caráter quantitativos e qualitativos), mas também volume (quantidades exorbitantes de instâncias), tornando difícil a avaliação manual. Essas características tornam as técnicas de aprendizado de máquina bastante convenientes, pois essas focam na obtenção e entendimento de relação úteis dentro de grandes bases de dados, visto que segundo Jordan e Mitchell (2015), estas se otimizam automaticamente através da experiencia. Neste contexto, o presente estudo tem como foco avaliar o nível de satisfação de clientes de uma instituição financeira multinacional com o objetivo de criar um modelo capaz de calcular a probabilidade de um cliente estar insatisfeito futuramente com seus produtos e serviços baseado em certas características e comportamentos históricos.

O presente trabalho encontra-se organizado da seguinte forma: a seguir apresenta a fundamentação de conceitos básicos para contextualização do estudo. Depois são apresentados os métodos utilizados e suas respectivas etapas, descrevendo a base de dados, preparação, métricas utilizadas, os modelos propostos e suas características. Por fim, são descritos a discussão dos resultados obtidos, a conclusão do trabalho e sugestões para futuros estudos.

Algoritmos de classificação

O aprendizado de máquina tem como seu principal foco duas questões correlatas: Como um sistema computacional pode ser aprimorado automaticamente através de experiências e quais são as leis estatísticas e computacionais que governam os conhecimentos de aprendizado, incluindo computadores, humanos e organizações (JORDAN; MITCHELL, 2015).

Dada a amplitude das duas questões, existem diversas técnicas e métodos que podem ser utilizados para cada tipo de problema e suas respectivas especificidades, como Naive-Bayes, Aprendizado Profundo, Máquina de Vetores de Suporte, entre outros. A seguir, são apresentados os três algoritmos de classificação utilizados nesta pesquisa.

Árvore de Decisão

Árvore de Decisão é um algoritmo de aprendizado de máquina supervisionado que divide base de dados em pequenos grupos de dados, baseando-se em medidas descritivas, até o ponto em que possam ser descritos com um rótulo. Conforme pode ser visualizado na Figura 1.

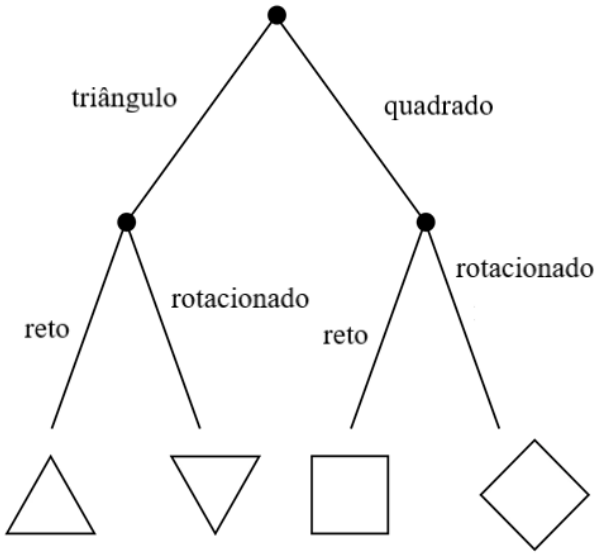

Figura 1. Exemplo de Árvore de Decisão, demonstrando as etapas do processo de classificação de um modelo de árvore de decisão que distingue formas geométricas.
Fonte: Adaptado Plenio e Vitellilipartz (2001),

Segundo Myles et al. (2004), este algoritmo é composto por funções discriminantes ou regras de decisão, estruturadas de forma hierárquica, que são aplicadas recursivamente para segregar as informações do banco de dados em classes. Esta segregação é feita nos nós estabelecidos pela árvore, o nó raiz é uma das variáveis do conjunto de dados utilizado, já o nó-folha é o rótulo ou valor que será gerado como resposta de uma regra de decisão.

Floresta Aleatória

A grande maioria dos problemas relacionados a Aprendizado de Máquina podem ser divididos em dois grupos, o de classificação (no qual se prediz a que classe certa observação pertence) e regressão (no qual se prediz o valor de certa instância), e um dos algoritmos mais utilizados em ambos os casos é o Floresta Aleatória, em diversas aplicações.

O Floresta Aleatória é um método *ensemble* que consiste na criação de árvores de decisão individuais, no qual cada uma é cultivada de acordo com um parâmetro aleatório (BHARATHIDASON; VENKATAESWARAN, 2014).

Neste modelo, um subconjunto de recursos preditivos é considerado durante cada divisão selecionada aleatoriamente, assim, a decisão das árvores leva a uma única previsão conjunta, calculando a média de todas as previsões individuais (BREIMAN, 2001, p. 5). Assim, o resultado mais frequente se torna a previsão do modelo, conforme demonstrado na Figura 2.

A maior vantagem do Floresta Aleatória é o fato das árvores protegerem-se de seus erros individuais. Por cada árvore ser um modelo individual, algumas podem estar certas e outras incorretas em suas previsões, porém como um modelo *ensemble* e prevendo como um grupo, é possível obter um rendimento melhor que com suas previsões individuais.

Dentre alguns dos diversos meios de aplicação deste método de aprendizagem de máquina, temos: astronomia (GAO et al., 2009), ecologia (CUTLER et al., 2007), agricultura (LOW et al.; 2012) e biotecnologia (BOULESTREIX et al., 2012).

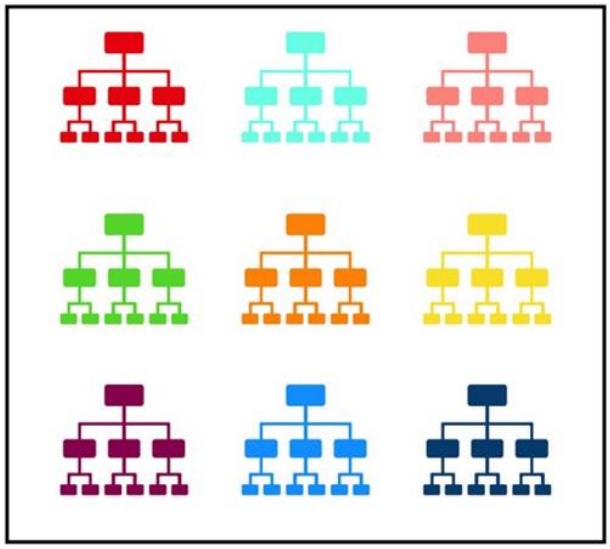

Figura 2. Visualização do processo de decisão de um modelo Floresta Aleatória, na qual, em um grupo de 9 árvores, 6 previram o resultado 1 e 3 previram o resultado 0. Com isso, a predição final do modelo é 1.
Fonte: Adaptado de Yiu (2019).

XGBoost

Métodos *ensemble* muitas vezes são usados para reduzir viés nas predições do algortimo, porém há também métodos *ensemble* como *boosting*, que cria membros *ensemble* sequencialmente ao invés de paralelamente, como realizado no *Floresta Aleatória*. Dessa forma, o elemento criado mais recente aprende com os erros de previsão dos elementos anteriores.

Já o *Gradient Boosting* é uma abordagem na qual novos modelos são treinados para prever os resíduos dos modelos anteriores. Mais especificamente o algoritmo de *boosting* executa iterações n vezes para aprender a função que faz predições minimizando a função de perda, a cada iteração é adicionado um novo estimador para tentar corrigir a predição de cada instância (MITCHELL; FRANK, 2017), conforme pode ser visualizado na Figura 3.

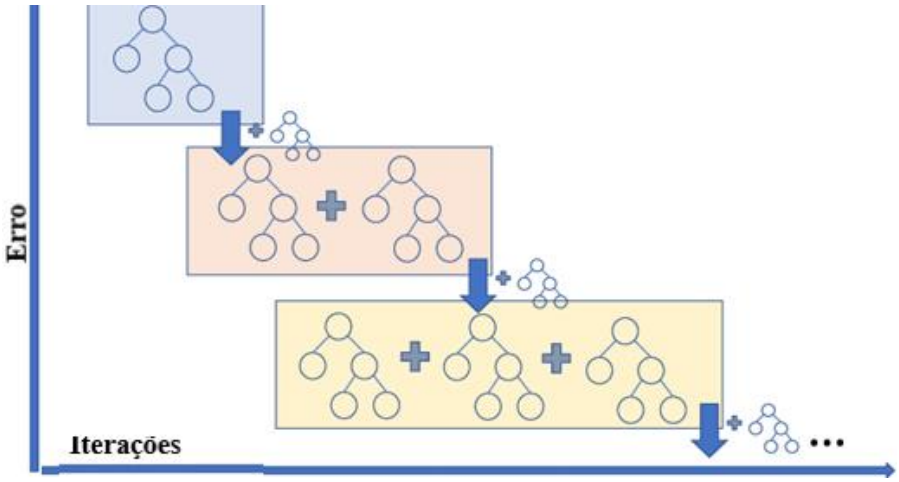

Figura 3. Representação do processo de iteração dos algoritmos de *Gradient Boosting.*
Fonte: Adaptado de Baturynska (2021).

O XGBoost dá também nome a uma biblioteca de código aberto que permite combinar o *Gradient Boosting* e o Floresta Aleatória, sendo assim, superior a diversos modelos de classificação. O modelo foi utilizado em diversas soluções vencedoras em portais de competições de Ciência de Dados, como Kaggle (17 de 29 soluções vencedoras em 2015) e KDDCup, as 10 melhores soluções em 2015 usaram XGBoost (BEKKERMAN, 2015).

Apesar das similaridades com o método *Floresta Aleatória*, suas diferenças distinguem muito o processo de classificação principalmente em como as árvores são construídas e como os resultados são combinados.

Conforme Glen (2019), o fato de o XGBoost construir as árvores de forma sequenciada e não paralela como o *Floresta Aleatória* permite que a adição da nova árvore melhore as deficiências presentes nas já existentes. Assim, os resultados são combinados durante o processo de criação das árvores e não no final do processo (por "regra da maioria"). Entretanto, apesar de todas as vantagens, o método é mais complexo e pode causar *overfitting* (sobreajuste aos dados de treino) em bancos de dados com muitos "ruídos".

Material e Métodos

Na Figura 5 é possível observar o fluxo do código do modelo criado para este trabalho e cada uma de suas etapas, as quais serão descritas com detalhes nas subseções seguintes.

A base de dados disponibilizada para este estudo é uma base real de uma instituição financeira multinacional de grande porte. Os dados que permitem a identificação da companhia foram omitidos por questão de estratégia comercial. O conjunto utilizado é composto por 370 *atributos* e a variável *alvo* com um total de 151,838 observações, as quais 76,020 compondo o conjunto de treino e 75,818 compondo o conjunto de teste (divisão proposta pela própria instituição que disponibilizou os dados).

Dentre os atributos, todos são numéricos e não foram disponibilizadas descrições individuais, apenas uma coluna nomeada ID com um valor individual para a identificação dos clientes/instâncias. Entretanto, sabe-se que são características dos clientes (como idade, localização, dentre outras) e dados comportamentais (como informações de movimentação de conta, saldo bancário, entre outras).

Nas análises iniciais do conjunto de treino foi observado um desbalanceamento no mesmo em relação à variável *alvo*, tendo 96,04% das instâncias com valor 0 (que indica cliente satisfeito) e apenas 3,96% com valor 1 (que indica cliente insatisfeito).

Apesar do banco de dados estar com a proporção condizente com de um negócio de sucesso, a mesma pode dificultar o treinamento do modelo, uma vez que, a grande parte dos dados representam apenas um dos resultados a serem previstos e a minoria as instâncias que realmente queremos detectar que são os clientes insatisfeitos. É válida também a menção que há uma presença muito grande do valor zero no banco de dados, sendo maioria em muitas variáveis, fator que também pode afetar no aprendizado do modelo.

Por se tratar de um banco de dados com um grande número de *atributos* e sem descrição detalhada para identificar correlações iniciais lógicas ou em grupos, os demais atributos foram analisados individualmente, nas quais nenhuma instância apresentou ausência de algum valor ou valor inválido.

Preparação dos Dados

No processo de limpeza dos dados, *outliers* foram observados apenas em uma das variáveis (var_3) que apresentava o valor -9999, que foi substituído pelo valor mais comum na variável (presente em 97,56% das instâncias): 2.

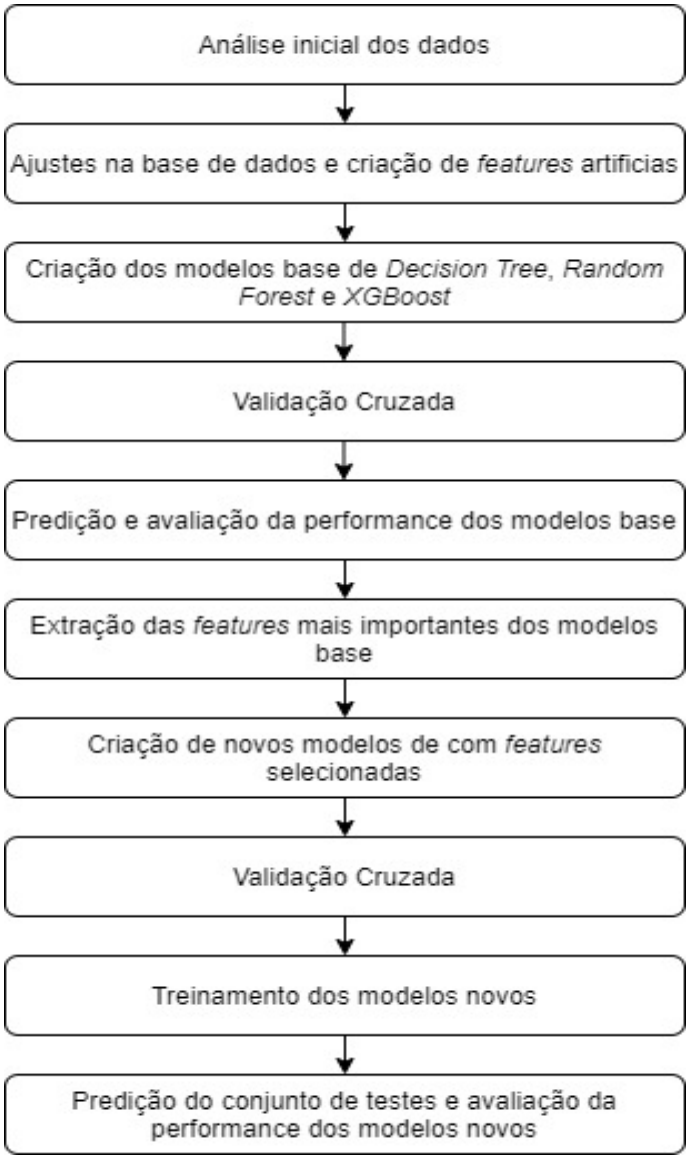

Figura 5. Fluxograma do código desenvolvido para o trabalho.
Fonte: Autores.

Depois, foram removidos os atributos que são redundantes e que não contribuem para o processo de aprendizagem do modelo: 34 *atributos* com valores duplicados, 29 *atributos* com valores constantes e 165 *atributos* com valores escassos (neste caso, variáveis que apresentam o valor 0 para mais de 99,00% das instâncias).

Em seguida, foram criadas novas *atributos*, atributos artificiais que possam prover informações significativas para o aprendizado do modelo. Foram criados 9 atributos, em sua maioria medidas descritivas, sendo elas: média aritmética, média geométrica, variância, coeficiente de variação, desvio padrão, curtose, assimetria e contagem de zeros que aparecem em cada instância (excluindo destes cálculos, obviamente, a variável *alvo*).

AUC (*Area Under the Curve*)

A métrica escolhida para avaliação da performance dos modelos foi a AUC, também conhecida como Área Sob a Curva, que fornece uma medida agregada de desempenho em todos os limites de classificação possíveis.

O AUC é como a probabilidade de o modelo classificar um exemplo positivo aleatório mais alto que um exemplo negativo aleatório de acordo com *Google Developers* (2020). Permitindo assim avaliar o quão eficiente o modelo é em distinguir as classes num intervalo de 0 a 1.

A curva utilizada para o cálculo do AUC é a ROC (*Receiver Operating Characteristic*) que é formada pelos valores do percentual de Positivos (sensibilidade do modelo) e a percentual de Falsos Positivos em diferentes limiares de classificação, ilustrado na Figura 6.

Figura 6. Taxa de Positivos classificados como positivos (Verdadeiros Positivos) vs. Taxa de Negativos classificados como negativos (Falsos Negativos) em diferentes limiares de classificação, formando a AUC em cinza.
Fonte. Adaptado de Google Developers (2020).

Como a maioria das métricas de avaliação de performance, quanto maior o AUC, melhor o desempenho do modelo em prever negativos e positivos corretamente. Para referência, um AUC com valor 0 significa que o modelo está prevendo negativos como positivos e vice-versa, já com um valor de 0.5 significa que aleatoriamente o modelo consegue prever negativos e positivos corretamente, porém sem consistência, e um modelo com AUC de valor 1 implica um modelo que consegue prever corretamente todos os negativos como negativos e todos os positivos como positivos.

Modelos

Foram utilizados três modelos de classificação na solução proposta, Árvore de Decisão e Floresta Aleatória da biblioteca *Scikit Learn* e o XGBoost, ambos da linguagem de programação Python.

Inicialmente, o conjunto de treino foi dividido em dois subconjuntos de treino e teste (serão referenciados como subset_1 e subset_2) na proporção de 80/20. Prontamente, construiu-se um modelo simples de cada um dos modelos de classificação escolhidos, todos treinados com o subset_1 e com o hiperparâmetro *max_depth* igual a 15 respectivamente. Este que define o número máximo de divisões que podem ser feitas por cada uma das árvores do modelo, rigorosamente falando, limitam a complexidade e profundidade das árvores.

Para definir qual a melhor quantidade de árvores a serem criadas por cada modelo de múltiplas

árvores, foi feito um processo de validação cruzada para os modelos de *Floresta Aleatória* e XGBoost. A validação cruzada foi feita com 15 grupos, ou seja, o modelo dividirá o subconjunto subset_1 em 15 grupos e repetirá a rotina 15 vezes, testando cada combinação de treino e teste destes grupos. O hiperparâmetro *n_estimators* dos modelos, que delimita o número de árvores criadas, foi ajustado de acordo com o processo.

Tabela 1. Performance dos modelos iniciais

Modelo	subset_2	Conjunto de Teste
Árvore de Decisão	0,750880	0,73973
Floresta Aleatória	0,829280	0,81618
XGBoost	0,830945	0,82270

Depois, foram treinados os modelos e feitas as previsões do subset_2 e do conjunto de testes principal. A avaliação da performance de cada um dos modelos pode ser observada na Tabela I. Em seguida, foram extraídos os *atributos* que apresentaram maior importância para aprendizado de cada um dos três modelos. A partir disso, foram criados mais 3 subconjuntos derivados do subset_1: subset_DT, que contém apenas os 65 atributos mais importantes para o modelo de Árvore de Decisão, subset_RF, que contém apenas os 65 atributos mais relevantes para o modelo Floresta Aleatória e subset_XGB que contém os 65 atributos que mais contribuíram para o modelo XGBoost.

Logo, foram elaborados três novos modelos de Árvore de Decisão, Floresta Aleatória e XGBoost para serem treinados com cada subconjunto criado, com objetivo de analisar como os modelos respondem a um conjunto de dados menor e mais concisos. Foram utilizados o hiperparâmetros *max_depth* igual a 7, repetidos os processos de validação cruzada, e finalmente, treinados os modelos. Em sequência foram realizadas as previsões de cada um dos modelos para o subset_2 e o conjunto de testes, cujas performances podem ser observadas na Tabela II.

Tabela 2. Performance dos Modelos.

Conjunto de Treino	Modelo	subset_2	Conjunto de Teste
subset_DT	Árvore de Decisão	0,808070	0,814030
subset_DT	Floresta Aleatória	0,808070	0,804550
subset_DT	XGBoost	0,845672	0,833690
subset_RF	Árvore de Decisão	0,820014	0,813350
subset_RF	Floresta Aleatória	0,820014	0,810290
subset_RF	XGBoost	0,842233	0,832690
subset_XGB	Árvore de Decisão	0,808760	0,820230
subset_XGB	Floresta Aleatória	0,808760	0,805110
subset_XGB	XGBoost	0,846498	0,833980

Resultados

Nas performances dos primeiros modelos, a pequena diferença no percentual de performance dos modelos nos dois conjuntos de dados subset_2 e o conjunto de teste (0,01115, 0,01310 e 0,008245 para *Árvore de Decisão*, *Floresta Aleatória* e XGBoost respectivamente) mostra que todos os 3 foram capazes de generalizar de forma eficiente o comportamento dos dados e que de fato aprenderam com o conjunto de treino e não apenas decoraram seu comportamento de suas instâncias.

O melhor resultado no conjunto de teste foi obtido pelo modelo XGBoost, que ficou a uma diferença pequena da performance do modelo Floresta Aleatória (0,00652), já em relação ao modelo Árvore de Decisão teve uma performance muito superior (0,08297). Resultados que, de forma geral, apresentam performance de classificação satisfatória na métrica de avaliação escolhida.

Dentre as 65 variáveis mais relevantes para os modelos selecionadas para formar os subconjuntos subset_DT, subset_RF e subset_XGB, 31 estavam presentes em todos os 3 conjuntos. Em meio dessas, destaca-se a var_15, atributo de maior relevância nos modelos de Árvore de Decisão e Floresta Aleatória e 7ª mais relevante no modelo XGBoost.

Ao analisá-la individualmente, verifica-se que a variável apresenta o valor mínimo de 5 e máximo de 105, além de uma distribuição que se concentra no intervalo de 23 e 27 (48% das instâncias em ambos os conjuntos de treino e teste se encontram neste intervalo), conforme representado na Figura 7 e na Figura 8.

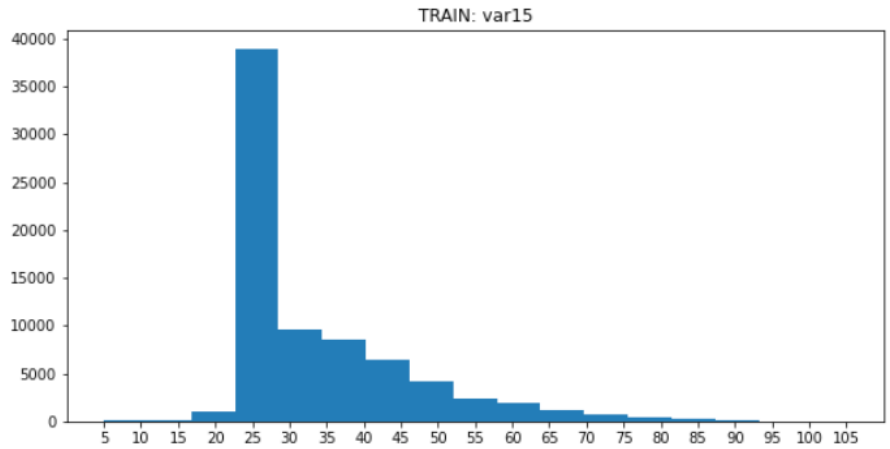

Figura 7. Histograma demonstrando a distribuição dos valores da feature_15 no conjunto de treino.
Fonte: Autores.

Figura 8. Histograma demonstrando a distribuição dos valores da feature_15 no conjunto de teste.
Fonte: Autores.

Em ambos os conjuntos a variável apresenta distribuição semelhante, e pelos valores da variável pode-se presumir que a variável se refere a idade dos clientes, o que demonstraria uma predominância de consumidores mais jovens. Foi observado também que nenhum dos clientes insatisfeitos estava abaixo do valor 23, uma variável para apontar este fato foi criada durante as primeiras análises, mas não resultou em nenhum acréscimo considerável no aprendizado do modelo e foi descartada.

Nota-se que 7 dos 9 atributos artificiais criados durante a preparação dos dados ficaram entre os 20 atributos mais importantes e foram inclusas nos conjuntos subset_DT e subset_RF. Já dentre os *atributos* do subset_XGB tem-se apenas o atributo n0. Os modelos criados a partir dos novos subconjuntos, tiveram diferença pequena de acurácia no desempenho de predições do subset_2 e do conjunto de testes. Assim, como os primeiros modelos demonstram terem sido treinados de forma eficiente, conseguindo generalizar o aprendizado.

Entretanto as performances melhoraram em alguns casos, o modelo de Árvore de Decisão treinado com o subset_XGB se comparado ao inicial, aumentou em 8,05%. Já o modelo Floresta Aleatória diminuiu -1,11%, se comparado o melhor desempenho (treinado com o subset_RF) com o modelo inicial.

O modelo XGBoost apresentou a melhor avaliação dentre os modelos em todas as modalidades, chegando a atingir 83,4% de desempenho. Curiosamente, os modelos XGBoost apresentaram performance similar em todos os 3 modelos criados, tendo a maior diferença de apenas 0,13%. Sendo assim, o modelo mais indicado para o problema proposto e tem-se o potencial de ser uma ferramenta poderosa para a instituição em questão.

Discussão

Na literatura, foram encontradas pesquisas relevantes para o tema e que estão relacionadas ao presente estudo. Kumar e Zymbler (2019) propuseram modelos de Máquina de Vetores de Suporte, Redes Neurais Artificiais e Redes Neurais Convulsionais para analisar *feedback* de clientes de diversas companhias aéreas na rede social Twitter com o objetivo de aprimorar a experiência do consumidor. As variáveis utilizadas foram extraídas dos textos dos consumidores e classificados pelo modelo. Os resultados mostraram que uma análise de associação aprofundada com as mensagens poderia oferecer insights poderosos para melhorar a experiência do consumidor.

Franco, Garcia e Cataluña (2019) propuseram um modelo Naive Bayes para estudo da base de dados da *9th Yelp Dataset Challenge*, com o objetivo de identificar os termos mais relevantes nas resenhas dos hotéis da cidade de Las Vegas, bem como a sua influência na avaliação dos hotéis. O modelo teve uma performance bastante positiva, atingindo aproximadamente 86,00% na métrica AUC (*Area Under the Curve*) e 84,00% na métrica F1 Score, comprovando-se uma abordagem válida e apropriada para análise da satisfação dos hóspedes.

Outros trabalhos semelhantes e com resultados promissores foram também observados, como o estudo de Farhadloo, Patterson e Rolland (2016). Os autores utilizaram Análise Bayesiana para modelagem visando descobrir quais aspectos afetam a percepção e avaliação de um cliente. O conjunto de dados utilizado foi o do portal *Trip Advisor*. Os resultados demonstraram uma performance de 82,90% na métrica R2.

Siebert et al. (2019) utilizaram aprendizado de máquina para prever a satisfação de um cliente através dos dados históricos de uma companhia elétrica brasileira, modelo que teve performance de 89,05%, apresentando um Erro Percentual Médio Absoluto de apenas 1,36% comparado a uma pesquisa realizada com os clientes da empresa no ano anterior.

Embora, levem em consideração empresas diferentes e objetivos distintos, todos os estudos citados têm o mesmo escopo: predizer a satisfação de seus clientes através da análise de dados históricos para auxílio de gestão. Todos utilizam abordagens e modelos diferentes para soluções, e atingem resultados satisfatórios para suas problemáticas.

O estudo de Sabbeth (2019), utilizou uma base de dados de uma empresa de telecomunicações para comparar a performance de 9 modelos. Os resultados mostraram que o desempenho de um modelo pode

variar com uso de diferentes técnicas, como modelos híbridos, métricas de avaliação ou com um banco de dados diferente. Essa característica, torna o aprendizado de máquina uma área bastante propícia para criação e combinação de soluções de forma criativa e sistemática para obtenção de aprimoramento.

Dessa maneira, os resultados obtidos por este estudo podem ser considerados promissores ao que tange a proposta do trabalho, não apenas ao comparar com os resultados dos trabalhos similares, mas ao considerar também a complexidade da base de dados, com *atributos* semianônimas, variável *alvo* desbalanceada e subjetividade do tema.

Conclusão

Este artigo apresentou uma proposta utilizando metodologia baseada em *Data Science* e em modelos de Aprendizado de Máquina, buscando estabelecer relações entre comportamento e características de clientes de uma instituição financeira com seu grau de satisfação com a mesma.

O resultado do modelo de melhor performance foi de 83,40% e apesar de não ser plenamente acurado, é bastante eficiente e pode ser uma ferramenta poderosa para a companhia permitindo-a compreender mais a fundo a satisfação de seus clientes.

Por se tratar de uma base de dados sem descrição clara e específica para cada variável, é difícil ser assertivo em quais dados são de fatos relevantes e correlatos à insatisfação de um cliente, em especial por se tratar de um banco de dados tão extenso. Acredita-se que uma base de dados com atributos não-anônimos poderia agregar bastante para a análise do problema e consequentemente na performance do modelo, além de permitir estudos de possíveis causas da insatisfação e propostas de planos de ação.

Trabalhos futuros e similares, poderiam testar novas combinações de variáveis, outras combinações de modelos e hiperparâmetros para soluções alternativas, ou mesmo dessa mesma solução com outros períodos ou grupos diferentes de clientes. Espera-se que este trabalho ofereça *insights* valiosos na resolução de problemas similares e encoraje trabalhos semelhantes no que tange uso de dados para compreensão da satisfação do consumidor.

Agradecimentos

Este trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

BATURYNSKA, I. Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. *Research GateJournal of Intelligent Manufacturing*, 32(3), January, 2021. [Online]. Available: https://www.researchgate.net/publication/340524896_Prediction_of_geometry_deviations_in_additive_manufactured_parts_comparison_of_linear_regression_with_machine_learning_algorithms.

BEKKERMAN, R. The present and the future of the kdd cup competition: An outsider's perspective. *Linkedin*, August, 2015. [Online]. Available: https://www.linkedin.com/pulse/present-future-kdd-cup-competition-outsiders-ron-bekkerman.

BHARATHIDASON, S.; VENKATAESWARAN, C. J. Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees. *International Journal of Computer*

Applications, 101(13), 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.1781&rep=rep1&type=pdf.

BOULESTEIX, A. L.; et al. Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), October, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1072.

BREIMAN L. Random Forest, *Machine Learning*, 45(1), pp. 5–32, 2001.

BREIMAN, L. Random Forest. *Machine Learning*, 45(1), pp. 5-32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, United States of America, August, 2016. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2939672.2939785.

COHEN, D.; et al. Customer Satisfaction: A Study of Bank Customer Retention in New Zealand. *Lincoln University*, 109, New Zeland, 2006. [Online]. Available: http://researcharchive.lincoln.ac.nz/handle/10182/324.

CUTLER, D. R.; et al. Random Forest for classification in ecology. *Ecology*, 88(11), 2007. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0539.1.

FARHADLOO, M.; PATTERSON, R. A.; ROLLAND, E. Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, 90, October, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167923616301026.

FORNELL, C. A National Customer Satisfaction Barometer: The Swedish Experience. *Journal of Marketing*, 56, January, pp. 6-21, 1996.

FORNELL, C.; et al. The American Customer Satisfction Index: Nature, purpose, and findings, *Journal of Marketing*, 1996. [Online]. Available: http://triton.nfh.uit.no/dok/fornell-1996.pdf.

FRANCO, M. J. S.; GARCIA, A. N.; CATALÑA, F. J. R. A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, 101, August, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0148296318306672.

GAJARE, S. Data Science Methodology and Approach, *Geeks for Geeks*. [Online]. Available: https://www.geeksforgeeks.org/data-science-methodology-and-approach/.

GAO, D.; ZHANG, X.; ZHAO, H. Random Forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 9(2), 2009. [Online]. Available: https://iopscience.iop.org/article/10.1088/1674-4527/9/2/011/meta.

GLEN, S. Decision Tree vs. Random Forest vs. Gradient Boosting. *Data Science Central*, July, 2019. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained.

GOOGLE DEVELOPERS; *Classification: ROC Curve and AUC, Machine Learning Crash Course*. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

HILL, N.; ROCHE, G.; ALLEN, R. Customer Satisfaction: The customer experience through the customer's eyes. *Cogent Publishing Ltd*, London, 2007.

IOANNA, P. D. The Role of Employee Development in Customer Relations: The Case of UK Retail Banks. *Corporate Communication*, 7(1), pp. 62-77, 2002.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects, *Science*, July, 2015. [Online]. Available: https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf.

KAYNAK, E.; KUCUKEMIROGLU, O. Bank and Product Selection: Hong Kong. *The International Journal of Bank Marketing*, 10(1), pp. 3-17, 1992.

KUMAR, S.; ZYMBLER, M. A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, 62, 2019. [Online]. Available: https://link.springer.com/article/10.1186/s40537-019-0224-1.

LEEDS, B. 'Mystery Shopping' Offers Clues to Quality Service. *Bank Marketing*, 24(11), November, 2020, pp. 24-27, 1992.

LOW, F.; et al. Per-field crop classification in irrigated agricultural regions in Middle Asia using random orest and support vector machine ensemble. *International Society for Optics and Photonics*, United Kingdom, September, 2012. [Online]. Available: https://spie.org/Publications/Proceedings/Paper/10.1117/12.974588?SSO=1.

MITCHELL, R.; FRANK, E. Accelerating the XGBoost algorithm using GPU computing, *Peerj Computer Science*, July, 2017. [Online]. Available: https://peerj.com/articles/cs-127/.

MITTAL, V. To Cut Costs, Know Your Customer. *MIT Sloan*, 2021 (Winter), 2020. [Online]. Available: https://sloanreview.mit.edu/article/to-cut-costs-know-your-customer/.

MOTROC, G. Interview with Reynold Xin, co-founder and Chief Architect at Databricks. *Jaxenter*. [Online]. Available: https://jaxenter.com/apache-spark-machine-learning-interview-143122.html.

MYLES, A. J. et al. *An introduction to decision tree modeling. Journal of Chemmometrics*, vol. 18, pp. 275-285, 2004.

PLENIO, M. B.; VITELLILIPARTZ, V. The physics of forgetting: Landauer's erasure principle and information theory. *Imperial College*, London, UK, 2001. [Online]. Available: https://arxiv.org/pdf/quant-ph/0103108.pdf.

PROVOST, F.; FAWCETT, T. *Data Science for Business: What You Need to Know about Data Mining and Data*. O'Reilly Media Inc., 2013.

REICHHELD, F. F. Learning from Customer Defections. *Harvard Business Review*, March/April, pp. 56-69, 1996.

ROSENBERG, J. L.; CZEPIEL, A. J. Journal of Consumer: A marketing approach customer retention. *MCB Up Limited*, United Kingdom, 2017.

SABBETH, S. F. Machine-Learning Techniques for Customer Retention: A Comparative Study. *International Journal of Advanced Computer Science and Applications*, 9(2), 2018. [Online]. Available: https://thesai.org/Downloads/Volume9No2/Paper_38-Machine_Learning_Techniques_for_Customer_Retention.pdf.

SCIKIT LEARN. *sklearn.ensemble.RandomForestClassifier*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

SIEBERT, L. C. Predicting customer satisfaction for distribution companies using machine learning. *International Journal of Energy Sector Management*, December, 2019. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/IJESM-10-2018-0007/full/html.

SLATER, S. F. Developing a Customer Value-Based Theory of the Firm. *Journal of the Academy of Marketing Science*, 25, Spring, pp. 162-167, 1997.

WOODRUFF, R. B. Customer Value: The Next Source of Competitive Advantage. *Journal of Academy of Marketing Science*, 25(2), pp. 139-153, 1997.

XGBOOST. *XGBoost Documentation*. [Online]. Available: https://xgboost.readthedocs.io/en/latest/.

YIU, T. *Understanding Random Forest*. July, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.