UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Maria Luiza Rabelo Serpa

RANDOM FOREST APLICADO NA ANÁLISE DE *CHURN*: COMPARAÇÃO DO AJUSTE COM DADOS COMPLETOS *VERSUS* AJUSTE EM ESTRATOS DEFINIDOS POR COVARIÁVEL CATEGÓRICA

Belo Horizonte
2023

Maria Luiza Rabelo Serpa

RANDOM FOREST APLICADO NA ANÁLISE DE *CHURN*: COMPARAÇÃO DO AJUSTE COM DADOS COMPLETOS *VERSUS* AJUSTE EM ESTRATOS DEFINIDOS POR COVARIÁVEL CATEGÓRICA

Monografia de Especialização apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Especialista em Estatística.

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte

2023

Serpa, Maria Luiza Rabelo

S486r Random forest aplicado na análise de Churn [manuscrito]: comparação do ajuste com dados completos versus ajuste em estratos definidos por covariável categórica / Maria Luiza Rabelo Serpa —2023.
32.f. il.

Orientador . Guilherme Lopes de Oliveira.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: 31-32

1. Estatística. 2. Churn de clientes.3. Processo estocástico. 4. Árvores de decisão. I. Oliveira, Guilherme Lopes de. II. Universidade Federal de Minas Gerais I. Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 283ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE MARIA LUIZA RABELO SERPA.

Aos trinta dias do mês de março de 2023, às 08:00 horas, com utilização de recursos de videoconferência à distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Maria Luiza Rabelo Serpa**, intitulado: "*Random Forest aplicado na análise de Churn: comparação do ajuste com dados completos versus ajuste em estratos definidos por covariável categórica*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 30 de março de 2023.

Guilherme Lopes de Oliveira

Prof. Guilherme Lopes de Oliveira (Orientador)
DECOM / CEFET-MG

Guilherme Augusto Veloso

Prof. Guilherme Augusto Veloso
UFF/RJ

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Maria Luiza Rabelo Serpa, número de registro 2020680020, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que entregou para seu orientador, o professor Guilherme Lopes de Oliveira, o trabalho, que aprovou a versão final. O trabalho foi apresentado no dia 10 de março de 2023 com o título "Random Forest aplicado na análise de Churn: comparação do ajuste com dados completos versus ajuste em estratos definidos por covariável categórica".

Belo Horizonte, 24 de abril de 2023

Roberto da Costa Quinino:80871291720
Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2023.04.24 10:50:57 -03'00'

Prof. Roberto da Costa Quinino
Coordenador do curso de Especialização em Estatística
Departamento de Estatística / UFMG

AGRADECIMENTOS

À minha mãe, minha gratidão eterna por todo amor, toda dedicação e por ser minha referência na vida.

Agradeço ao meu companheiro de vida, Raoni, pelo incentivo, apoio e por ter acreditado em mim mesmo em momentos que eu não acreditava. Sou eternamente grata pelo companheirismo e paciência nos momentos difíceis.

Ao meu orientador, obrigada por sua excelência, ensinamentos e por ter me incentivado a persistir. O mundo precisa de mais professores inspiradores e gentis como você. Com certeza esse trabalho não teria acontecido sem sua ajuda.

Aos professores da Especialização de Estatística da UFMG, minha admiração por toda disponibilidade, adaptabilidade e dedicação em compartilhar conhecimento, mesmo em momentos difíceis e incertos.

RESUMO

A perda de clientes para concorrentes, conhecida como *churn*, ocorre quando um cliente decide mudar de uma empresa para outra. Em busca de se manter no mercado, as empresas devem entender a motivação dos clientes a se afastarem se quiserem reduzir as taxas de *churn* e reter seus clientes atuais. Uma maneira de compreender melhor o *churn* de clientes, detectar padrões e prever o comportamento dos clientes é utilizando uma técnica de aprendizado de máquina conhecida como *Random Forest*. Esse é um método *ensemble* que combina múltiplas árvores de decisão e cada uma delas contribui na identificação da classe mais popular. Este estudo avalia o impacto de uma variável categórica geográfica ao utilizarmos o modelo *Random Forest* para prever o *churn* de clientes de uma instituição financeira na Alemanha, Espanha e França, com base em dados disponíveis em um repositório público. Para isso foram criados um modelo com a base de dados completa e modelos com a base de dados estratificada por país. Os dados dos modelos foram analisados utilizando algumas métricas sobre a qualidade de predição. A premissa de que a nacionalidade dos clientes seria capaz de impactar o modelo estatístico se mostrou verdadeira em certo sentido para o método e dados utilizados. As principais diferenças foram observadas na sensibilidade e no F_1 *score*, ambos ligados à qualidade da classificação dos *churns* verdadeiros na França, país que representa a maior parte da base de dados proporcionalmente, e Alemanha. Embora os achados deste estudos sejam limitados aos dados abordados e sujeitos às condições de análises aqui especificadas, os resultados mostraram que, no caso da classificação binária via *Random Forest*, a exploração sobre estratificação ou não dos dados pode gerar conclusões interessantes do ponto de vista prático.

Palavras-chave: *Churn* de Clientes. *Random Forest*. Árvores de decisão.

ABSTRACT

Customer churn, also known as customer turnover, is the loss of existing customers to competitors. It occurs when a customer decides to switch from one company's product or service to another. In order to thrive, companies must understand what drives their customers away if they want to reduce churn rates and retain their current customers. One way to better understand customer churn, detect patterns and predict customer behavior is by using a machine learning technique known as Random Forest. It is an ensemble method that combines multiple decision trees and outputs the most popular class. This study evaluates the impact of the geographic categorical variable when using the Random Forest model to predict customer churn of a financial institution in Germany, Spain and France, based on a public data repository. To do this, a model was created with the complete database, and models with the database stratified by country. The data from the models were analyzed using quality prediction metrics. The premise that the customer's nationality would be able to impact the statistical model proved to be true in a way for the method and data used. The main differences were observed in the sensitivity and F_1 score, both related to the classification quality of the true churns in France, the country that represents the largest part of the database proportionally, and Germany. Although the findings of this study are limited to the data addressed and subject to the conditions of analysis specified here, the results showed that, in the case of binary classification via Random Forest, the exploration to stratify or not of the data can generate interesting conclusions from a practical point of view.

Keywords: Customer Churn. Random Forest. Decision Trees

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia, surgimento das *fintechs* e maior acesso à informação, instituições financeiras têm visto o mercado se tornar mais dinâmico e altamente competitivo (ARAÚJO, 2022). Para lidar com esses fatores, muitas empresas têm investido em melhores métodos de retenção de clientes, utilizando muitas vezes modelos estatísticos associados a poder computacional para personalizar a experiência em suas plataformas, melhorando dessa forma o relacionamento com seus consumidores. Esses modelos utilizam dados que quantificam características comportamentais e assim podem indicar possíveis necessidades ou possíveis insatisfações.

Franceschi (2019) cita que empresas têm alterado o foco estratégico para atuar em retenção de clientes em vez de focar em aquisição de novos, pois clientes de longo prazo tendem a evitar quebra de fidelidade e geram maiores lucros para a empresa. Ainda, a definição de estratégias para evitar a perda de clientes é um dos fatores que permitem a longevidade do negócio (NESLIN *et al.*, 2006).

A análise de c*hurn* é responsável por identificar a propensão dos clientes a se desligarem da instituição, possibilitando a definição de estratégias que podem ser utilizadas para cada tipo de consumidor. Essa análise é um exemplo clássico de problemas de classificação (SCHNEIDER, P. H, 2016), pois tem como objetivo prever classes para os dados analisados, ou seja, temos como resposta uma variável categórica que, neste caso, representa duas possíveis situações: se o cliente tem a propensão de continuar a fidelidade (não *churn*) ou se deseja se desligar da instituição (*churn*). Neste problema de classificação binária, é comum que seja feita a substituição dos valores categóricos pelos valores 1, que representaria a ocorrência de *churn*, e 0, que representaria a não ocorrência de *churn*. Ao prever se um cliente ficará ou não com a empresa, as empresas podem tomar medidas para abordar os problemas que possam levá-los a sair.

Em problemas de modelagem e classificação de *churn*, assim como em outros contextos de análise estatística, é comum a utilização de diversos tipos de fatores/variáveis preditoras, como, por exemplo, variáveis quantitativas contínuas ou discretas e também variáveis qualitativas como categóricas, nominais ou ordinais. A abordagem de tratamento, escala e codificação dessas variáveis pode impactar no desempenho dos modelos criados e seus resultados.

Este trabalho dará atenção especial à questão de variáveis categóricas no contexto de modelos e métodos estatísticos para classificação. Em certos casos, os dados de diferentes regiões geográficas, empresas ou comunidades, por exemplo, são agregados e a indicação sobre o subgrupo (região, empresa ou comunidade) é incluída no modelo/método de classificação como variável de controle. Fazendo isso no caso de um modelo de regressão logística, por exemplo, ter-se-á a estimação de um intercepto comum a todos os subgrupos e os coeficientes individuais dos demais fatores presentes no modelo, bem como os seus erros-padrão, serão estimados de acordo com as informações/padrões observados nos dados agregados (dados de todos os subgrupos conjuntamente). Ainda, uma mesma distribuição é assumida para todas as observações e erros/resíduos associados. Em outros métodos de classificação, como árvores de decisão e *Random Forest*, as classificações também podem ser diferentes ao se fazer a análise separadamente por subgrupo.

A capacidade preditiva é uma das preocupações principais ao se aplicar algum modelo/método de classificação. Ela está relacionada ao potencial do método em identificar/classificar a correta condição de cada observação, por exemplo, se um determinado cliente se trata de um potencial *churn* ou não. Neste contexto, uma pergunta que surge na presença de uma variável categórica que segrega as observações em subgrupos é: existe diferença na capacidade preditiva dos modelos/métodos aplicados separadamente para os dados dentro de cada subgrupo com relação ao que seria obtido ao se analisar todos os dados conjuntamente e incluindo tal variável de controle?

A fim de investigar empiricamente a hipótese acima, neste trabalho será explorado um conjunto de dados para classificação de *churn* disponível na literatura, no qual existe uma variável que identifica as observações como provenientes de três diferentes países. Um método de classificação, *Random Forest*, será aplicado ao conjunto de dados agregado e aos três conjuntos de dados específicos de cada país separadamente, a fim de se comparar as medidas de capacidade preditiva nos dois contextos de análise para este conjunto de dados.

O texto está organizado como: o Capítulo 2 apresenta a base de dados selecionada; o Capítulo 3 apresenta uma breve definição de alguns métodos utilizados para classificações binárias, com atenção especial ao *Random Forest*; os resultados são apresentadas no Capítulo 4; e o Capítulo 5 fecha o trabalho com a discussão e considerações finais.

2 BASE DE DADOS

A base de dados selecionada para ser analisada neste trabalho foi obtida através do site Kaggle[1]. Esses dados trazem informações sobre consumidores do setor bancário e foi disponibilizada inicialmente com o objetivo de encontrar modelos capazes de prever a ocorrência de *churn* de acordo com as informações desses consumidores. A base de dados contém 10000 observações para cada uma das 14 variáveis disponíveis descritas no Quadro 1, sendo a variável *Exited* a variável resposta de interesse. Em resumo, as variáveis dessa base se dividem em três variáveis identificadoras, três variáveis intervalares, duas variáveis nominais, três variáveis proporcionais e três variáveis binárias.

Quadro 1 - Descrição e tipo das variáveis da base de dados

Variável	**Tipo**	**Descrição**
RowNumber	Identificador	Número da linha
CustomerId	Identificador	Identificador único
Surname	Identificador	Sobrenome
CreditScore	Intervalar	Classificador de crédito
Geography	Nominal	País
Gender	Nominal	Gênero
Age	Ratio	Idade (em anos)
Tenure	Intervalar	Quantidade de anos de fidelidade
Balance	Ratio	Saldo bancário
NumOfProducts	Intervalar	Quantidade de produtos bancários consumidos
HasCrCard	Binário	Se tem cartão de crédito
IsActiveMember	Binário	Se é cliente ativo
EstimatedSalary	Ratio	Salário estimado
Exited	Binário	Se não é mais cliente (*churn*)

Fonte: Elaborado pela autora.

[1] https://www.kaggle.com/

As variáveis *RowNumber*, *CustomerId* e *Surname* são identificadores de usuários e para fins de criar modelos de previsão são consideradas irrelevantes, portanto não foram consideradas neste estudo.

Foram então utilizadas as variáveis restantes, nas quais temos a variável dependente *Exited* e nas variáveis independentes, temos seis variáveis numéricas e temos quatro que são categóricas. A variável *Geography* contém três possíveis valores, França, Alemanha ou Espanha; a variável *Gender* contém as categorias Feminino e Masculino; a variável *HasCrCard* tem duas categorias, indicando se o cliente tem ou não o serviço de cartão de crédito; e por fim, a variável *IsActiveMember,* que também tem apenas duas possibilidades, informa se o cliente é ou não ativo da instituição financeira.

A variável *Geography* é de grande relevância para os objetivos deste trabalho. Partindo do pressuposto que a localização dos consumidores pode impactar a forma como os mesmos se relacionam com instituições bancárias, esta pode ser uma variável de segregação que impacta substancialmente a capacidade preditiva do modelo estatístico selecionado.

3 METODOLOGIA

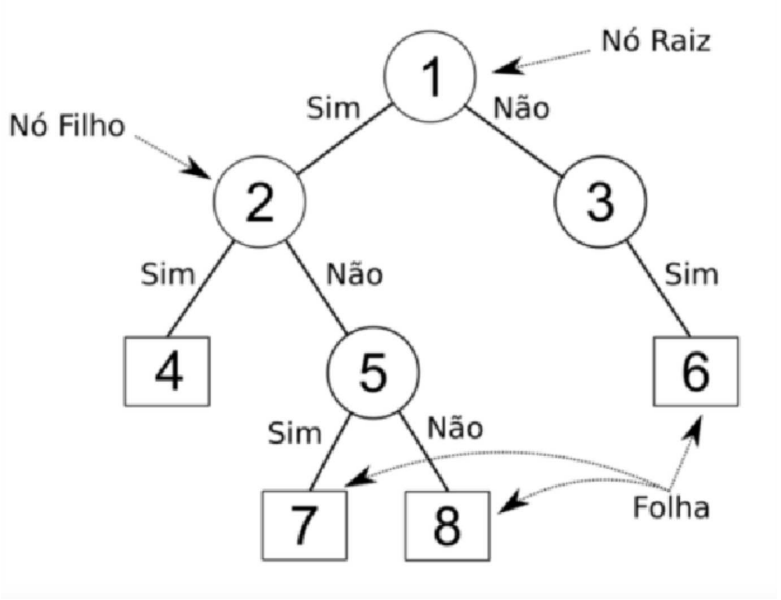
Um problema de classificação binária é a atribuição de elementos de um conjunto de dados a duas classes distintas, como, por exemplo, a classificação de usuários inadimplentes ou a identificação de possível *turnover* de clientes. Neste tipo de modelo, a resposta varia entre duas categorias geralmente representados pelos valores 1, associado à ocorrência do evento em estudo, e 0, associado à não ocorrência do evento. Regressão logística, árvores de decisão, florestas aleatórias (em inglês, *Random Forest*), XGBoost, KNN (K-*nearest neighbors*, ou "K-vizinhos mais próximos") e redes neurais estão entre os métodos estatísticos e de *machine learning* mais utilizados para predizer as classificações (WU *et al.*, 2008).

3.1 Regressão logística, árvores de decisão e *Random Forest*

A regressão logística (COX, 1958) é um dos modelos estatísticos mais comuns para resolver problemas de classificação binária. Esse modelo permite a inclusão de covariáveis de forma simples, as quais tem seus efeitos estimados e avaliados quanto à significância estatística. A regressão logística muitas vezes é utilizada como modelo base para comparação da eficiência de modelos mais complexos.

Outro método utilizado para resolver problemas de classificação é o método de árvores de decisão (JAMES *et al.*, 2013). Esse método é baseado na utilização de regras hierárquicas de separação e que podem ser representadas através de um desenho em formato de árvore com nós principais e suas ramificações, conforme exemplificado na Figura 1. A separação e escolha dos nós e suas ramificações ocorre a partir de estratificação ou segmentação das variáveis presentes no conjunto de dados (nós), também chamado de espaço preditor, resultando em uma regra de particionamento que será utilizada para fazer a predição. Em um determinado problema, é natural que não exista uma única árvore de decisão e, além disso, pode-se chegar a um mesmo resultado com diferentes árvores de classificação.

Figura 1: Exemplo de árvore de decisão.



Fonte: SATO, *et al.* (2013)

Para a definição dos nós e ramos que irão compor uma árvore de decisão podem ser utilizados diferentes métodos, dentre os quais se destacam o cálculo da entropia, índice de Gini ou regressão (JAMES et al., 2013). Tais métodos referem-se à fórmulas e cálculos responsáveis pela definição da estrutura final da árvore, visando encontrar a estrutura mais otimizada para o problema em questão. Os métodos utilizados pelos algoritmos visam identificar as variáveis, dentre o conjunto de todas as variáveis preditoras, aquelas que possuem maior relação com a variável resposta, colocando-as no topo da árvore, em seus nós principais. Em todos os passos do algoritmo, a busca é feita no conjunto de todas as variáveis envolvidas na análise.

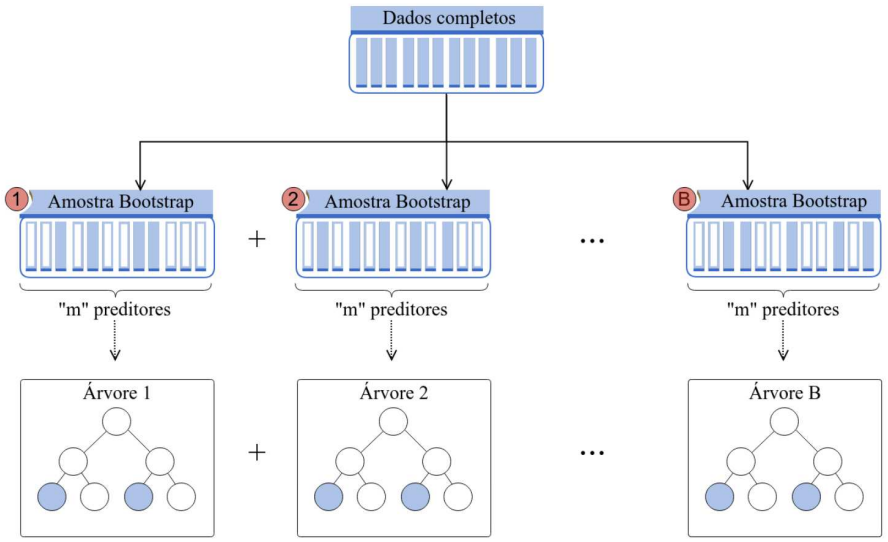
Conforme Kotsiantis (2013) menciona, métodos de árvore de decisão são modelos sequenciais que combinam uma sequência lógica de testes simples e por essa razão são mais interpretáveis que modelos mais complexos. No entanto, esses modelos normalmente não são competitivos, em termos de performance e predição, se comparados a outros modelos de aprendizado supervisionado. Métodos que combinam várias árvores de decisão, como, por exemplo, o *Random Forest* e *Bagging*, foram criados para contornar os problemas relacionados com o método

clássico, porém como efeito colateral a interpretabilidade dos modelos é prejudicada, já que são de maior complexidade.

A técnica escolhida para realização deste trabalho foi o método *Random Forest*, pois foi um dos métodos que obteve os melhores resultados nos artigos científicos que utilizaram o banco de dados descrito no Capítulo 2 (COŞER *et al.*, 2020; RAHMAN e KUMAR, 2020; TÉKOUABOU *et al*., 2022).

Breiman (2001) define o método *Random Forest* como um classificador que consiste em uma coleção grande de árvores de decisão estruturadas, onde os preditores são distribuídos de forma independente dentre as árvores e cada uma delas contribui na identificação da classe mais popular para cada observação. Esse método surgiu como forma de minimizar a correlação entre as árvores do método *Bagging*, que também está definido em Breiman (2001), através da definição de uma quantidade *B* de árvores de decisão em amostras *bootstrap* nos dados de treino (parcela dos dados selecionada para ajuste do modelo e definição da regra de classificação), sendo que em cada separação que ocorre dentro da árvore é considerada uma amostra randômica de *m* preditores do total de *p* preditores, conforme pode ser visto na Figura 2.

Figura 2 - Mecanismo de funcionamento do *Random Forest*.



Fonte: http://cursos.leg.ufpr.br/ML4all/

James *et al.* (2013) reforça que esse mecanismo de separação dentro de cada árvore não permite que sejam consideradas todas as opções disponíveis de preditores e isso é o que permite que esse método consiga minimizar os impactos da alta correlação entre as árvores. Portanto, o propósito do método é reduzir a variância enquanto mantém baixo viés. As principais vantagens da utilização do *Random Forest* são a relativa rapidez no treinamento do modelo e alta capacidade de predição. Ele é robusto quando existem *outliers*, pode ser usado tanto em problemas de regressão e classificação binária ou multiclasses, fornece medidas da importância das variáveis, possibilidade de modificação dos pesos das classes, entre outras (CUTLER, CUTLER e STEVENS, 2012).

3.2 Etapas da Análise e Métricas de Avaliação

Em estudos que envolvem a classificação das unidades amostrais entre categorias predefinidas, é comum que o conjunto de dados seja dividido aleatoriamente em dados de treino (parcela dos dados selecionada para ajuste do modelo e definição da regra de classificação) e dados de validação/teste (parcela dos dados selecionada para predição e avaliação da regra construída). Para todos os modelos ajustados neste trabalho, considerou-se a proporção 80/20, ou seja, 80% dos dados foram selecionados para treino e 20% dos dados para teste.

Inicialmente o *Random Forest* foi ajustado considerando todos os dados disponíveis (Alemanha, Espanha e França) e, para a separação dos dados de treino e teste, foi considerada a proporção dos três países contidos na variável *Geography*. Isso foi feito com o intuito de garantir que cada país contribuiria igualmente para construção da regra de classificação, bem como na avaliação da assertividade da regra construída. Em um segundo momento, o modelo *Random Forest* foi ajustado separadamente para cada país.

Com o objetivo de verificar e comparar a eficiência de modelos estatísticos, é importante a utilização de métricas de avaliação da capacidade preditiva. Para a avaliação e comparação dos modelos criados neste trabalho, foram utilizadas 5 dessas métricas: acurácia, precisão, sensibilidade, especificidade e F_1 *score*. Vale mencionar que nos artigos científicos pesquisados que analisaram esta mesma base de dados, não houve consenso também sobre a melhor métrica de avaliação.

As métricas definidas abaixo foram escolhidas por serem mais comumente utilizadas para avaliar modelos de classificação.

Sensibilidade

A sensibilidade mede a porcentagem de verdadeiros positivos (VP) que o modelo classifica corretamente. Ou seja, no contexto deste trabalho, a sensibilidade representa o percentual de acerto na classificação dos indivíduos que de fato representam um *churn* na base de dados. Essa é uma métrica que dá mais ênfase aos erros chamados de falsos negativos (FN) que ocorrem quando o modelo classifica um *churn* observado como sendo não *churn*. Essa medida está demonstrada na Equação (1):

$$Sensibilidade = \frac{VP}{VP + FN} \quad (1)$$

Especificidade

Ao contrário da sensibilidade, a métrica de especificidade mede a porcentagem de verdadeiros negativos (VN) que o modelo classifica corretamente. Ou seja, no contexto deste trabalho, a especificidade representa o percentual de acerto na classificação dos indivíduos que de fato não representam um *churn* na base de dados. Essa é uma métrica que dá mais ênfase aos erros chamados de falsos positivos (FP) que ocorrem quando o modelo classifica um não *churn* observado como sendo *churn*. Ela está definida conforme Equação (2):

$$Especificidade = \frac{VN}{VN + FN} \quad (2)$$

Acurácia

A métrica de acurácia é definida como a distância total entre os valores estimados e os valores reais, conforme demonstrado na Equação (3). Ela é dada pela porcentagem de classificações corretas em relação ao total de predições. Utilizando a matriz de confusão, classificações corretas são tanto os verdadeiros positivos, quanto os verdadeiros negativos.

$$Acurácia = \frac{VP + VN}{total\ de\ predições} \quad (3)$$

Precisão

Precisão é uma métrica muito utilizada em problemas de classificação e é definida conforme mostrado na Equação (4). Ela avalia a relação entre o total de classificações corretas de *churns* (verdadeiros positivos) dentro do total de classificações positivas (*churns* identificados) feitas pelo modelo:

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

F_1 *score*

A métrica *F_1 score* é usada para avaliar a acurácia das predições de classificações binárias e seu cálculo é apresentado na Equação (5). Ela corresponde à média harmônica da precisão e da sensibilidade (CHICCO e JURMAN, 2020). Quando calculada para o conjunto de classificações negativas (*não churn*), esta leva em consideração a precisão e a especificidade.

$$F_1 score = 2 \times \frac{VP}{VP + FP + FN} \tag{5}$$

A Equação (5) considera um fator multiplicativo igual a dois, pois de acordo com a documentação da biblioteca do Scikit Learn (PEDREGOSA, F. et al., 2011), isto faz com que a sensibilidade e a precisão contribuam igualmente para o resultado do F_1 *score*.

3.3 *Software* Utilizado

O desenvolvimento deste trabalho foi feito através da linguagem Python (VAN ROSSUM & DRAKE Jr., 1995) e foi utilizado a biblioteca Scikit Learn (PEDREGOSA, F. et al., 2011) para o ajuste do modelo, para a separação das bases de treino e teste, e para a avaliação das métricas de desempenho do modelo preditivo.

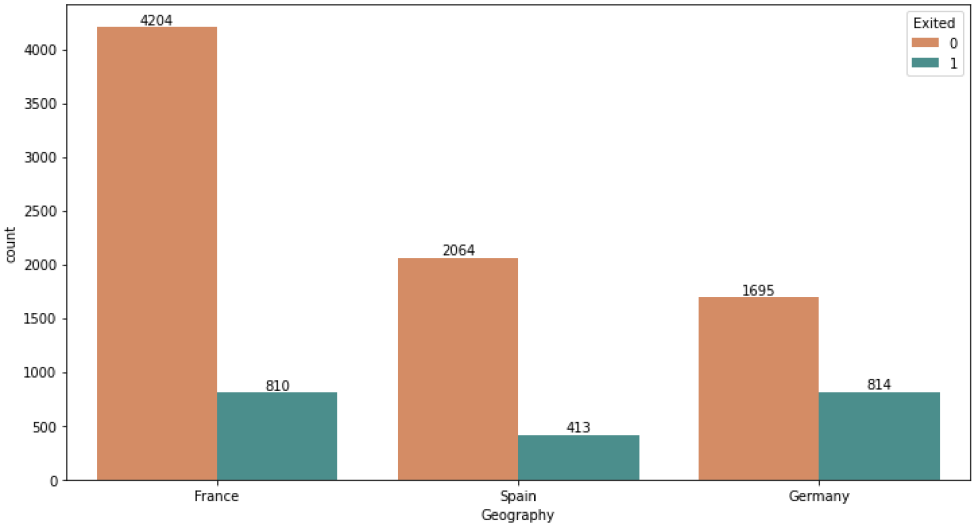
4 RESULTADOS

4.1 Análise descritiva

O principal objetivo de modelos de predição é prever os valores corretos da variável dependente, ou também chamada de variável *target*, portanto é importante entender suas características e como ela se comporta com relação a cada variável independente. Sendo assim, a variável dependente *Exited* é uma variável binária que contém 7963 (79,63%) registros com o valor igual a 0, ou seja, de consumidores que foram retidos (“não *churns*”) e 2037 (20,37%) registros com o valor igual a 1, ou seja, de consumidores que fecharam suas respectivas contas bancárias (*churns*). De acordo com a variável *Geography*, 50,14% dos dados são de indivíduos da França; 25,39% são da Alemanha; e 24,47% são da Espanha.

A relação das variáveis *Exited* e *Geography* pode ser vista na Figura 3 e demonstra que as taxas de *churn* (variável *Exited*=1) nos três países são diferentes, sendo 16,15% na França, 32,44% na Alemanha e 16,67% na Espanha. Isto reforça o objetivo deste trabalho em observar o impacto da segregação dos dados conforme essa variável no poder preditivo do modelo. Dentre os indivíduos observados como *churn*, 39,96% são da Alemanha; 39,76% são da França; e 20,28% são da Espanha.

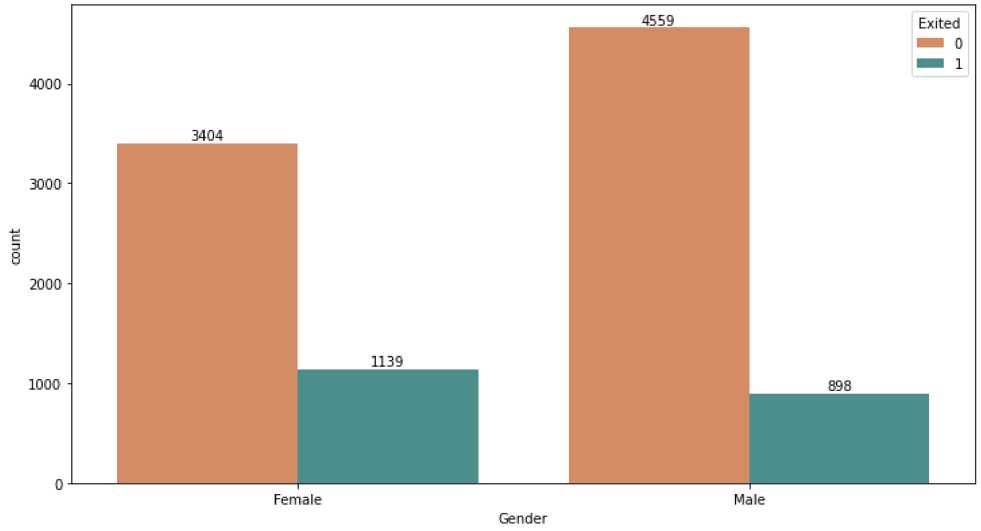
Figura 3 - Distribuição da variável *Exited* em relação a variável *Geography*



Fonte: Elaborado pela autora.

A relação das variáveis *Exited* e *Gender* pode ser vista na Figura 4 e demonstra como as taxas de *churn* se diferenciam entre gêneros, sendo 16,45% para homens e 25,07% para mulheres. Vale ressaltar que neste banco de dados a contribuição de observações entre os sexos é semelhante, sendo 54,57% para observações masculinas e 45,43% para observações femininas.

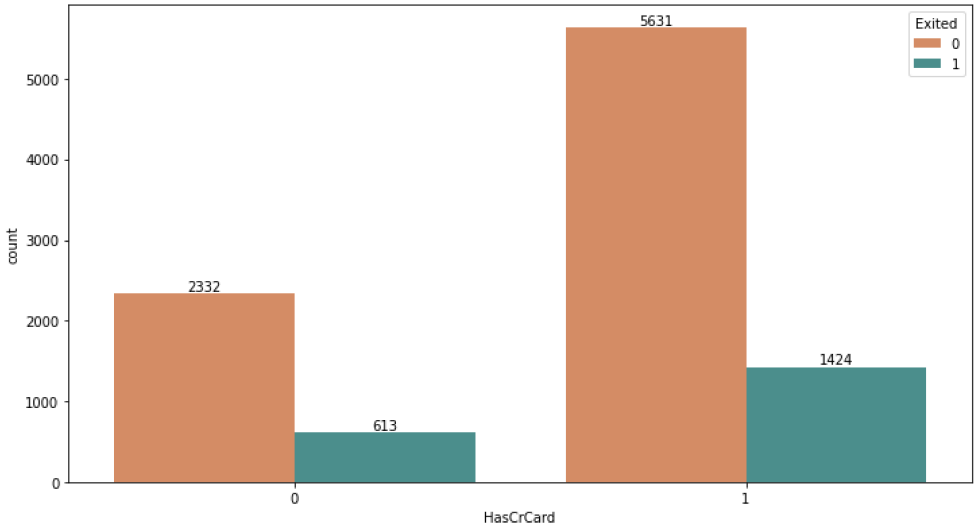
Figura 4 - Distribuição da variável *Exited* em relação a variável *Gender*



Fonte: Elaborado pela autora.

A variável *HasCrCard* informa se o cliente usufrui do serviço de cartão de crédito ou não, ou seja, caso o cliente tenha esse serviço contém o valor 1 ou contém 0, caso contrário. A relação das variáveis *Exited* e *HasCrCard* pode ser vista na Figura 5 e mostra que existe um desbalanceamento entre classes, sendo 70,55% a contribuição da classe 1 e 29,45% da classe 0. No entanto, as taxas de *churn* entre categorias não se diferem muito, sendo 20,18% para quem tem cartão de crédito e 20,81% para a outra categoria.

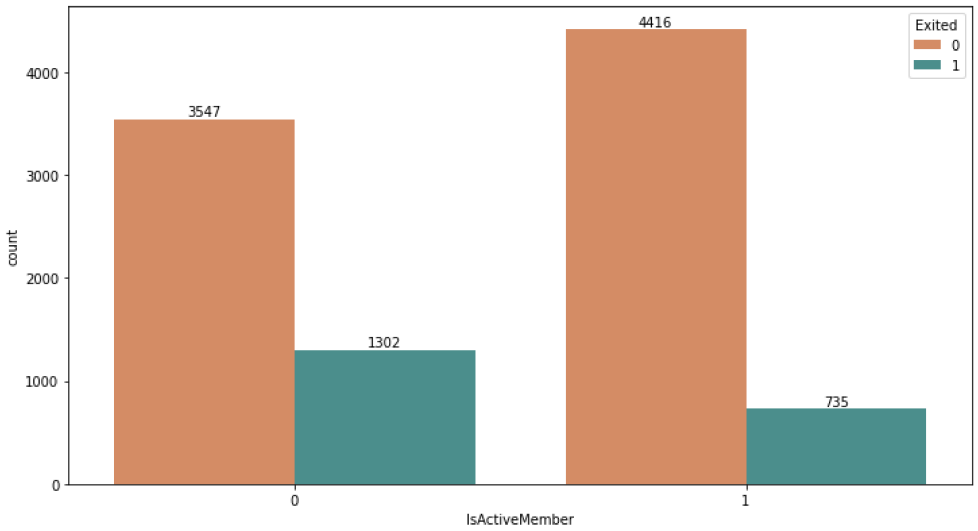
Figura 5 - Distribuição da variável *Exited* em relação a variável *HasCrCard*



Fonte: Elaborado pela autora.

A variável *IsActiveMember* informa se o consumidor é um cliente ativo ou não, ou seja, caso o cliente seja ativo contém o valor 1 ou contém 0, caso contrário. A relação das variáveis *Exited* e *IsActiveMember* pode ser vista na Figura 6 e mostra que as taxas de *churn* entre categorias se diferem, sendo 14,27% para os clientes ativos e 26,85% para a outra categoria. A contribuição das classes nessa variável é semelhante, sendo 51,51% para clientes ativos e 48,49% para clientes não ativos.

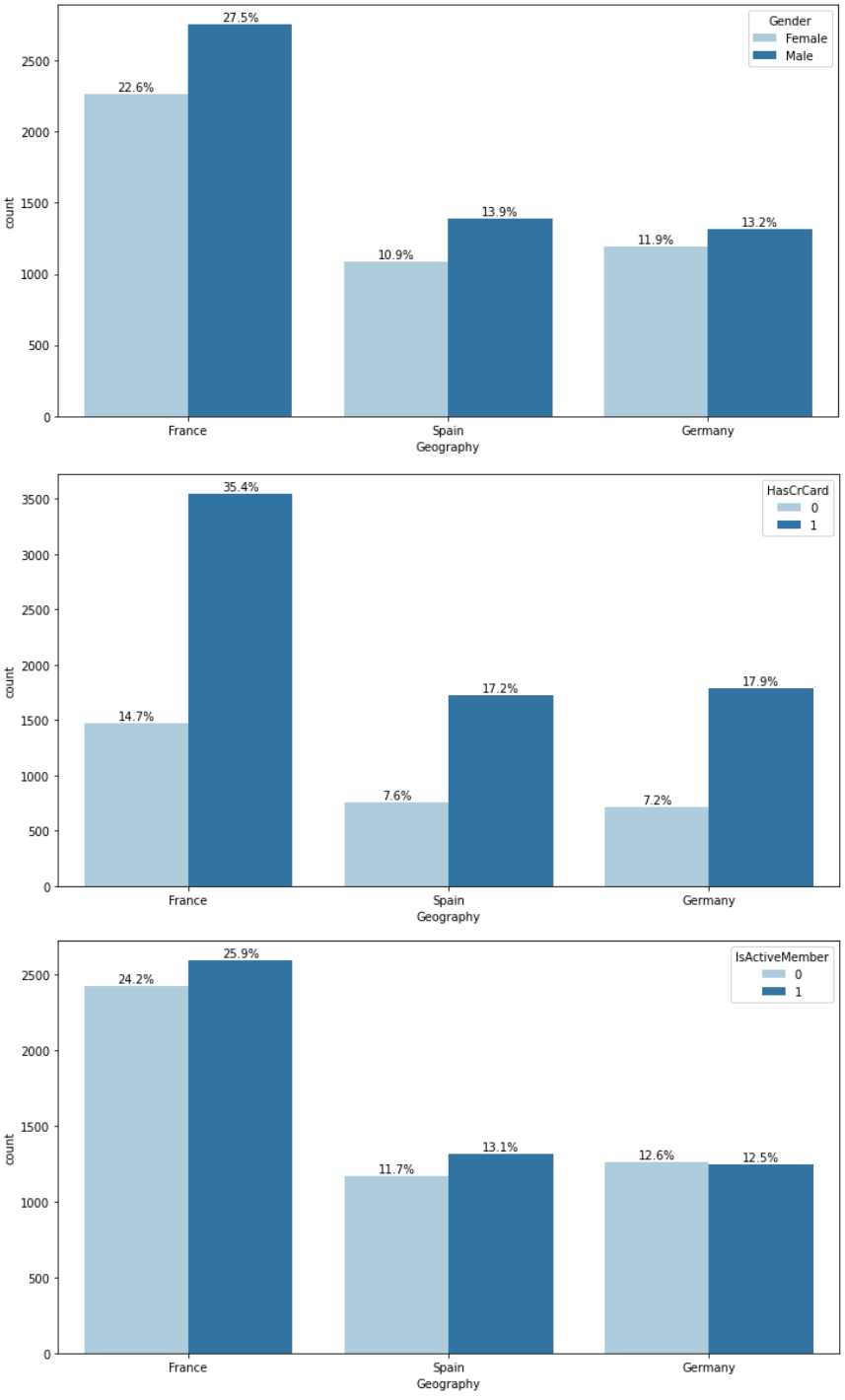
Figura 6 - Distribuição da variável *Exited* em relação a variável *IsActiveMember*



Fonte: Elaborado pela autora.

A Figura 7 apresenta a distribuição das variáveis *Gender*, *HasCrCard* e *IsActiveMember* separadamente para cada país. No geral, parece não haver grandes diferenças entre os países quanto à distribuição dessas variáveis.

Figura 7 - Distribuição das variáveis *Gender*, *HasCrCard* e *IsActiveMember* para cada país (França, Espanha e Alemanha), respectivamente



Fonte: Elaborado pela autora.

Na Tabela 1 é possível observar a relação da variável Exited em relação às variáveis quantitativas por meio de estatísticas descritivas, como a média, mediana, desvio-padrão e valores mínimos e máximos. Observa-se também que estas variáveis têm unidades de medida diferentes e grandezas escalares diferentes. No entanto, não foi realizado para a elaboração deste trabalho nenhum tratamento em relação a padronização de escala dessas variáveis.

A Tabela 1 se refere às estatísticas das variáveis quantitativas tanto para quando houve *churn*, ou seja, quando a variável dependente *Exited* assume o valor igual a 1, quanto para clientes que continuam ligados a instituição financeira, ou seja, não houve *churn* e a variável dependente assume o valor igual a 0. No geral, os consumidores que cancelaram suas contas bancárias (*churn*) são em média um pouco mais velhos, têm balanço bancário maior e salário estimado maior quando comparados com o grupo que não encerraram suas contas. Outro ponto observado foi uma pequena queda na média do *score* de crédito do grupo em que houve *churn* e na média de produtos consumidos por esses clientes.

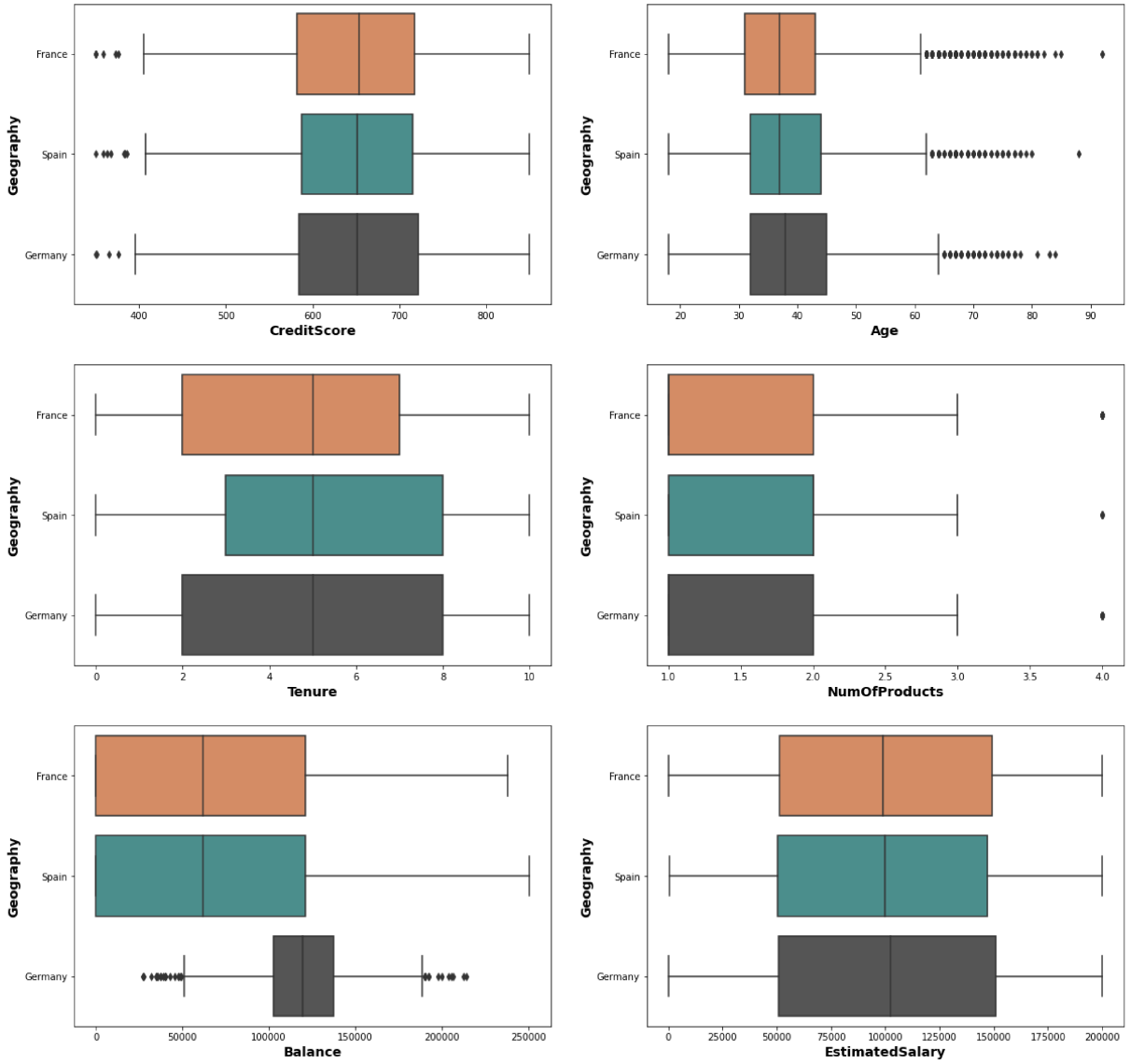
Tabela 1 - Medidas descritivas das variáveis quantitativas

	Credit-Score	**Age**	**Tenure**	**Balance**	**NumOf-Products**	**Estimated-Salary**
média (*Exited* = 1)	645,35	44,83	4,93	91108,54	1,48	101465,68
média (*Exited* = 0)	651,85	37,41	5,03	72745,30	1,54	99738,39
mediana (*Exited* = 1)	646,00	45	5	109349,29	1	102460,84
mediana (*Exited* = 0)	653,00	36	5	92072,68	2	99645,04
desvio-padrão (*Exited* = 1)	100,32	9,76	2,94	58360,79	0,80	57912,42
desvio-padrão (*Exited* = 0)	95,65	10,13	2,88	62848,04	0,51	57405,59
min (*Exited* = 1)	350,00	18	0	0,00	1	11,58
min (*Exited* = 0)	405,00	18	0	0,00	1	90,07
max (*Exited* = 1)	850	84	10	250898,09	4	199808,10
max (*Exited* = 0)	850,00	92	10	221532,80	3	199992,48
total (*Exited* = 1)	2037	2037	2037	2037	2037	2037
total (*Exited* = 0)	7963	7963	7963	7963	7963	7963

Fonte: Elaborado pela autora.

A Figura 8 apresenta a distribuição das variáveis *Credit-Score, Age, Tenure, Balance, NumOf-Products* e *Estimated-Salary* separadamente para cada país. Percebe-se uma maior discrepância apenas na distribuição das variáveis *Tenure* e *Balance*, sendo que nesta última os dados da Alemanha têm comportamento muito diferente dos da França e Espanha.

Figura 8 - Distribuição das variáveis, respectivamente por linha, *Credit-Score, Age, Tenure, NumOfProducts, Balance* e *EstimatedSalary* para cada país (França, Espanha e Alemanha)



Fonte: Elaborado pela autora.

4.2 Resultados do Modelo *Random Forest*

Conforme mencionado na Seção 3.2, inicialmente foi criado um modelo *Random Forest* a partir da base de dados completa, indicado por M. Completo, e três outros modelos criados a partir da base separada por país, referenciados como M. Separado, sendo que cada modelo contém um país diferente. As estatísticas de

avaliação da capacidade preditiva de todos os modelos estão apresentadas na Tabela 2. Elas se referem aos resultados obtidos para tais métricas no conjunto de dados de validação. Para fins de comparação, no caso do M. Completo, além de apresentar as estatísticas para a base completa, elas também foram calculadas separadamente para os dados de cada país. Isto é, após obter as previsões com base no ajuste do modelo com os dados completos, os dados de cada país foram separados a fim de se obter as estatísticas de previsão individualizadas em cada um deles.

Tabela 2 - Estatísticas de avaliação da capacidade preditiva dos modelos *Random Forest* ajustados com base nos dados completos e nos dados separados por país

Modelo	Acurácia	Precisão (Classe 0)	Precisão (Classe 1)	Especif.	Sensib.	F_1 *score* (Classe 0)	F_1 *score* (Classe 1)
M. Completo	0,87	0,87	0,80	0,97	0,43	0,92	0,56
França - M. Completo	0,88	0,89	**0,81**	0,98	**0,34**	0,93	**0,48**
França - M. Separado	0,88	0,89	**0,76**	0,97	**0,43**	0,93	**0,55**
Alemanha - M. Completo	**0,82**	**0,83**	0,80	**0,94**	0,56	0,88	0,66
Alemanha - M. Separado	**0,80**	**0,80**	0,80	**0,92**	0,57	0,86	0,66
Espanha - M. Completo	0,88	0,88	0,80	0,98	**0,38**	0,93	0,51
Espanha - M. Separado	0,88	0,88	0,80	0,98	**0,38**	0,93	0,51

Fonte: Elaborado pela autora.

Para França e a Alemanha, foram identificadas algumas diferenças relevantes ao se comparar as estatísticas de predição para o modelo com todos os dados com as estatísticas obtidas com os dados separados por país.

No caso da França, país que corresponde à maioria das observações da base de dados, houve um aumento na sensibilidade e no F_1 *score*. Isto indica que, ao se separar a base de dados exclusivos deste país, tem-se um maior acerto na classificação dos clientes que representam um *churn* (aumento de 9% no acerto

desse grupo, segundo a alteração da sensibilidade). Vale notar que esse comportamento veio acompanhado de uma diminuição na precisão da classe 1 (clientes que representam um *churn*). Isto indica que em geral houve mais predições da classe 1 no modelo separado, o que inclui possivelmente um número mais elevado de falsos positivos (FN) aumento o denominador da métrica de precisão. Contudo, este efeito não gerou grandes alterações nas métricas de predição associadas aos clientes que não representam um *churn* (classe 0).

O oposto aconteceu no modelo separado da Alemanha, que quando comparado com o modelo completo, observa-se uma diminuição nas estatísticas de acurácia, precisão da classe 0 e especificidade. Ou seja, houve leve piora na classificação de clientes que verdadeiramente não são *churns*. Tanto no modelo completo quanto no modelo separado, a Alemanha é o país com maior sensibilidade e F_1 *score*, isto é, neste país a classificação correta de *churn* tende a ser maior e não foi afetada pela separação ou não da base de dados.

A Espanha apresentou métricas idênticas em ambos os casos, tendo uma sensibilidade menor quando analisada individualmente do que a obtida no modelo geral (considerando todos os dados conjuntamente).

5 DISCUSSÃO E CONSIDERAÇÕES FINAIS

O presente trabalho abordou o problema de classificação de *churn* (quando um cliente se desliga de uma instituição) através do método *Random Forest*. O modelo foi ajustado em diferentes contextos, com o objetivo de avaliar o impacto da estratificação baseada em uma variável categórica quanto ao seguinte questionamento: há diferença na capacidade preditiva do modelo ao se considerar os dados conjunta ou separadamente de acordo com as categorias desta variável?

O banco de dados foi selecionado de um repositório público e já é bastante difundido por ser utilizado em competições que visam identificar um melhor método/modelo preditivo. O *Random Forest* foi selecionado por ter sido o método que, em geral, outros autores apontaram como sendo o mais eficiente para realizar predições nesses dados. A variável *Geography*, que estratifica os dados como sendo provenientes de três países (Alemanha, Espanha e França) foi considerada para os objetivos deste estudo.

A premissa de que a nacionalidade dos clientes seria capaz de impactar a capacidade preditiva do modelo estatístico utilizado neste trabalho se mostrou verdadeira em certo sentido. As principais diferenças foram observadas na sensibilidade e no F_1 *score*, ambos ligados à qualidade da classificação dos *churns* verdadeiros. Para França e Alemanha, o desempenho na classificação desse tipo de cliente foi melhorada ao se analisar individualmente os dados dos países. O oposto ocorreu para a Espanha.

Há um desbalanceamento na base de dados, com relação à proporcionalidade de cada país. A França corresponde a cerca de metade da base de dados e tem desempenho pior em classificar os churns ao ser analisada conjuntamente com os demais países. Uma possível justificativa para esse fato pode ser a melhor identificação de padrões nos dados franceses quando o efeito de confundimento dos dados dos demais países é removido. Do ponto de vista das instituições da França, apesar dos valores baixos de sensibilidade, mostra-se interessante a análise dos seus dados sem a inclusão de outros países, a fim de viabilizar uma melhor identificação dos clientes suscetíveis a *churn*. Para a Alemanha, segundo país mais frequente na base de dados, a estratificação não indicou melhora na capacidade preditiva dos *churns* (objetivo principal na prática),

tendo piorado um pouco a identificação dos não *churns*. Para este país, há uma maior assertividade na identificação de churns, independentemente se a análise é feita conjunta ou separadamente. Não houve diferença na capacidade preditiva para a Espanha ao se comparar os dois contextos analisados.

Em suma, para os dados analisados neste estudo, conclui-se que a separabilidade da base pode trazer vantagens na identificação de possíveis *churns* em certos casos. Percebeu-se que ao trabalhar os dados da França separadamente, a capacidade preditiva foi melhorada, possivelmente devido a comportamentos discrepantes das covariáveis para clientes desse país quando comparado aos demais. Embora tais discrepâncias não tenham sido observadas na análise descritiva dos dados, as especificações internas do método *Random Forest* podem tê-las identificado e gerado o efeito observado.

Alguns trabalhos e fóruns que abordam problemas de predição via modelos de regressão, apontam que a estratificação da base pode ser importante para permitir que a distribuição dos erros aleatórios sejam diferentes dentro de cada estrato[1]. A verificação de tal condição pode ser feita através da análise descritiva dos resíduos do modelo ajustado com os dados completos[2,3]. Embora os achados deste estudos sejam limitados aos dados abordados e sujeitos às condições de análises aqui especificadas, os resultados mostraram que, no caso da classificação binária via *Random Forest*, a exploração sobre estratificação ou não dos dados pode gerar conclusões interessantes do ponto de vista prático.

Por fim, uma possível limitação deste estudo diz respeito ao fato de que apenas um método de análise foi considerado, o *Random Forest*, mas outros podem ser explorados em análises futuras. Com a regressão logística, por exemplo, pode-se analisar mais diretamente possíveis mudanças nos efeitos das covariáveis sobre a chance de ocorrência ou não de *churn*, o que fica limitado no *Random Forest* por ser um método que não permite muito controle sobre os efeitos individuais das covariáveis.

[1]-https://stats.stackexchange.com/questions/468663/what-is-the-difference-between-a-linear-regression-with-a-dummy-variable-and-two

[2]-https://stats.stackexchange.com/questions/17110/should-i-run-separate-regressions-for-every-community-or-can-community-simply-b

[3]-https://stats.stackexchange.com/questions/12797/how-to-test-whether-a-regression-coefficient-is-moderated-by-a-grouping-variable

Outro ponto importante a ser comentado são as porcentagens das classes da variável dependente *Exited* nas bases de treino e teste, sendo que para a base de treino obteve-se aleatoriamente 79,49% de *não-churns* e 20,51% de *churns* enquanto na base de teste obteve-se 80,20% de *não-churns* e 19,80% de *churns*. Não foi realizada validação cruzada nem balanceamento do percentual de churn dentro da base de cada país com relação à base total, o que possivelmente pode impactar os resultados obtidos.

REFERÊNCIAS

ARAÚJO, J. M. A.. **Análise de sobrevivência e previsão de Churn de clientes de seguros de vida do banco do Brasil.** Dissertação (Mestrado Profissional em Computação Aplicada), Universidade de Brasília, Brasília, 2022.

BREIMAN, L.. **Random forests.** Machine Learning, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/a:1010933404324>.

CHICCO, D., JURMAN, G.. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.** BMC genomics, v. 21, p. 1-13, 2020. Disponível em: <https://doi.org/10.1186/s12864-019-6413-7>

COŞER, A. et al. **Propensity to churn in banking: what makes customers close the relationship with a bank?**. Economic Computation & Economic Cybernetics Studies & Research, v. 54, n. 2, 2020.

COX, D. R.. **The regression analysis of binary sequences**. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232, 1958.

CUTLER, A., CUTLER, D.R., STEVENS, J.R.. **Random Forests**. Em: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer, Boston, MA, 2012. Disponível em: <https://doi.org/10.1007/978-1-4419-9326-7_5>

FRANCESCHI, P. R.. **Modelagens preditivas de Churn: o caso do Banco do Brasil**. Dissertação (Mestrado) - Universidade do Vale do Rio dos Sinos, Porto Alegre, 2019.

JAMES, G. et al.. **An Introduction to Statistical Learning**. Springer New York, 2013. Disponível em: <https://doi.org/10.1007/978-1-4614-7138-7>

KAI, M. T.. **Precision and recall**. In: Encyclopedia of machine learning. Springer, 2011. p. 781-781. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_651>

KOTSIANTIS, S. B.. **Decision trees:** a recent overview. Artificial Intelligence Review, v. 39, p. 261-283, 2013. Disponível em: <https://doi.org/10.1007/s10462-011-9272-4>

RAHMAN, M., KUMAR, V.. **Machine Learning Based Customer Churn Prediction In Banking.** 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1196-1201, Coimbatore, India, 2020.

NESLIN, S. A. et al.. **Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models**. Journal Of Marketing Research (JMR), v. 43, n. 2, p. 204-211, 2006.

PEDREGOSA, F. et al.. **Scikit-learn: Machine learning in Python**. Journal of machine learning research, pp.2825–2830, 2011.

SATO, L. Y. et al.. **Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação do uso e cobertura da terra.** XVI Simpósio Brasileiro de Sensoriamento Remoto, Foz do Iguaçu, 2013.

SCHNEIDER, P. H.. **Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão.** Tese de Doutorado, 2016.

TÉKOUABOU, S.C.K. et al.. **Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods**. Mathematics, v. 10, n. 14, p. 2379, 2022

VAN ROSSUM, G., DRAKE Jr, F. L.. **Python reference manual.** Centrum voor Wiskunde en Informatica Amsterdam, 1995.

WU, X. et al.. **Top 10 algorithms in data mining**. Knowledge and Information Systems, 14(1), 1–37, 2008.