Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Bianca Cavalcante Bileski

Análise da viabilidade da aplicação de técnicas para evitar churn

Curitiba
2023

Bianca Cavalcante Bileski

Análise da viabilidade da aplicação de técnicas para evitar churn

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. André Grégio

Curitiba
2023

Análise da Viabilidade da Aplicação de Técnicas para Evitar Churn

Bianca C. Bileski[1]

[1]Departamento de Estatística, Universidade Federal do Paraná Av. Cel. Francisco H. dos Santos 100, Jardim das Américas, 81530-000, Curitiba, PR, Brasil*

A perda de clientes para uma empresa pode ser uma vitória para a concorrência. Este prejuízo não reflete somente na carteira de clientes, mas na totalidade de sua receita. Afinal, o custo para conquistar uma nova clientela costuma ser maior do que o investimento necessário para mantê-la fidelizada. Com o Open Finance, os bancos terão a possibilidade de desenvolver mais análises de rotatividade de clientes - o *churn*. Deste modo, os clientes permitirão compartilhamento de suas informações entre diferentes instituições autorizadas pelo Banco Central, possibilitando que elas tenham acesso a uma vasta gama de dados do consumidor. Com a adoção desse programa, os bancos entram em uma nova fase, a qual será crucial detectar os padrões de comportamento do cliente e definir abordagens precisas para prever potenciais *churn*, essa prevenção do churn permite que as empresas desenvolvam programas de fidelidade e campanhas de retenção de clientes resultando em redução de custos e consequentemente aumento da receita. O objetivo do presente trabalho é analisar a viabilidade e técnicas para evitar *churn* em uma base de dados bancária fictícia. O estudo foi feito com conceitos estatísticos, contendo fase de análise descritiva estatística, análise de correlação e aplicação de modelos de predição. Essas análises, podem servir de referência para gestores que desejam avaliar técnicas para evitar *churn* e a adoção de ferramentas de *machine learning* para predições de rotatividade de clientes.
Palavras-chave: *Churn*, Banco, Predição, Clientes

For a company, the loss of customers can mean a victory for the competition. This loss does not reflect only on the client portfolio, but on its entire revenue. After all, studies show that the cost of winning over a new clientele is usually higher than the investment required to keep them loyal. With Open Finance, banks will be able to develop more customer turnover analyzes - churn. In this way, customers will allow the sharing of their information between different institutions authorized by the Central Bank, allowing them to have access to a wide range of consumer data. With the adoption of this program, banks enter a new phase, where it will be crucial to detect customer behavior patterns and define accurate approaches to predict potential churn, this churn prevention allows companies to develop loyalty programs and customer retention campaigns. customer retention resulted in cost reduction and consequently increased revenue. The objective of this work is to analyze the feasibility and techniques to avoid churn in a fictitious bank database. The study was carried out with statistical concepts, containing a phase of descriptive statistical analysis, dynamic analysis and application of prediction models. These analyzes can serve as a reference for managers who want to evaluate techniques to avoid churn and the adoption of machine learning tools for predicting customer turnover.
Keywords: *Churn, Bank, Prediction, Clients*

1. Introdução

Em diversos setores do mercado a perda de um cliente pode significar uma vitória para a concorrência. Este prejuízo não reflete somente na carteira de clientes mas na totalidade da receita da empresa. Afinal, o custo para conquistar novos clientes costuma ser maior do que o investimento necessário para mantê-los fidelizados.

De acordo com Kumar e Shah [1] um fator crucial para o crescimento e prosperidade das empresas do século XXI é a proatividade na interação com os clientes. É inegável que a retenção de clientes é um dos principais pilares para garantir a sustentabilidade e saúde financeira de uma organização de sucesso. Há diversos estudos [3] [2] que abordam a retenção de clientes e oferecem soluções diversas para o problema, no entanto, é necessário identificar a questão real: o *churn*,

*databileski@gmail.com

que corresponde à saída de um cliente da empresa, para o fim de abordá-la adequadamente.

As novas oportunidades vêm surgindo com a rápida evolução das tecnologia de big data e pela enorme disponibilidade de dados que as empresas podem capturar por diversas fontes [5] e essas informações se tornam cada vez mais as principais aliadas no acompanhamento de métricas empresariais. Uma aplicação de previsão particularmente proeminente da gestão de relacionamento com o cliente (*CRM - Customer Relationship Management*) é a previsão de rotatividade de clientes (*CCP - Customer Churn Prediction*), que é definida como um método de identificação de clientes que mostram uma alta tendência para abandonar a empresa [6].

Os dados são importantes aliados no acompanhamento do *churn*. Diversas técnicas relacionadas aos dados podem ser aplicadas para evitá-lo, incluindo: análise preditiva, segmentação de clientes, monitoramento de indicadores, pesquisas de satisfação, automação, acompanhamento pós-atendimento, campanhas de reengajamento e acompanhamento da concorrência.

Em 30 de março de 2022 foi lançado no Brasil o Open Finance: um novo sistema financeiro.

> "O Open Finance, ou sistema financeiro aberto, é a possibilidade de clientes de produtos e serviços financeiros permitirem o compartilhamento de suas informações entre diferentes instituições autorizadas pelo Banco Central e a movimentação de suas contas bancárias a partir de diferentes plataformas e não apenas pelo aplicativo ou site do banco, de forma segura, ágil e conveniente."[8]

A partir desse lançamento, os bancos começaram a ter acesso a uma gama de dados maior possibilitando novas análises e abrindo oportunidade para melhorias na retenção de clientes e consequentemente na receita dessas instituições.

Com isso, o presente trabalho busca analisar a viabilidade e técnicas para evitar *churn* em uma base de dados bancária fictícia e foram executadas as seguintes etapas:

- Limpeza dos dados;
- Análise descritiva estatística;
- Análise de correlação;
- Análise chi-quadrado;
- Aplicação de modelos de predição (*SVM - Support Vector Machine, Logistic Regression, Random Forests* e *Decision Tree*).

Por se tratar de uma análise essencial para diversos setores do mercado, espera-se que esse estudo sirva de referência para gestores que desejam avaliar técnicas para evitar *churn* e a adoção de ferramentas de *machine learning* para predições de rotatividade de clientes. Diante disso, os principais resultados esperados são:

- Identificar principais fatores de rotatividade de produtos bancários;
- Identificar características de clientes que tendem a sair;
- Identificar características de clientes fiéis;
- Identificar a viabilidade de técnicas para evitar o *churn*.

2. Material e Métodos

2.1. Material

Em 14 de agosto de 2018 foi sancionada, no Brasil, a Lei Geral de Proteção de Dados - LGPD (Lei nº 13.709) [7] que tem como objetivo garantir a segurança de dados pessoais. Essa lei promoveu importantes alterações no Marco Civil da internet em 2014 e com ela surgiu uma imensa dificuldade para encontrar dados de clientes para estudos.

Com a implementação do Open Finance, os bancos estão ampliando o acesso a uma vasta quantidade de dados, o que tem impulsionado a importância da análise de *churn* neste mercado. Nesse contexto, os dados escolhidos para dar continuidade ao trabalho foram os bancários.

Devido a essa escassez de dados públicos ocasionada pela LGPD, para a elaboração do trabalho em questão, foram utilizados dados fictícios de clientes bancários em uma base de dados do Kaggle [9]. Ela contém 10.000 registros, sendo 20,37% deles rotulados como *churn*, e 14 variáveis diferentes que as definições constam na **Tabela 1** a seguir:

Descrição de Variáveis		
Nº	Variável	Descrição
01	*RowNumber*	Número da linha
02	*CustomerId*	Identificador do cliente
03	*Surname*	Apelido do cliente
04	*CreditScore*	Pontuação de crédito
05	*Geography*	País de origem
06	*Gender*	Gênero do cliente
07	*Age*	Idade do cliente
08	*Tenure*	Quanto tempo é cliente?
09	*Balance*	Saldo na conta
10	*HasCrCard*	Possui cartão de crédito?
11	*NumOfProducts*	Número de produtos
12	*IsActiveMember*	É um cliente ativo?
13	*EstimatedSalary*	Estimativa Salarial
14	*Churn*	Deixou de ser cliente?

Tabela 1: Tabela contendo as definições das 14 variáveis contidas na base de dados extraída do Kaggle [9]

Baseando-se nos dados mencionados, foram realizados alguns processos de tratamento para viabilizar a utilização dessas informações.

2.2. Métodos

Para as análises do presente trabalho, foi utilizada a linguagem Python, com as seguintes bibliotecas: Pandas, Numpy, Seaborn, Matplotlib, Sklearn, Plotly. Além disso, foram aplicadas as seguintes metodologias:

- **Limpeza e manipulação dos dados**

Com a quantidade de dados aumentando diariamente no mundo, novos desafios são encontrados, entre eles a coleta dessas informações. A coleta de dados e aquisição muitas vezes introduzem erros em dados, por exemplo, valores ausentes, erros de digitação, formatos mistos, entradas replicadas para a mesma entidade do mundo real, outliers e violação das regras de negócios [10].

Por esses motivos, a limpeza de dados é um processo fundamental na ciência de dados e possui um papel crucial para a qualidade e confiabilidade dos resultados obtidos. Além disso, ela garante eficiência computacional, descoberta de *insights* relevantes e a tomada de decisão se torna mais fácil e com um impacto maior. Por exemplo, um apelido de cliente não irá impactar nas decisões dele, mas a idade dele pode influenciar muito em seu comportamento.

- **Análise descritiva estatística**

Após a limpeza e manipulação dos dados, uma das técnicas mais utilizadas para análise de dados é a estatística descritiva que tem como função descrever os dados examinados. Com ela, pode-se resumir, sumarizar e explorar o comportamento dos dados.

A análise descritiva estatística foi aplicada através de gráficos, tabelas e medidas de síntese como porcentagens e médias para a melhor compreensão dos dados.

- **Análise de correlação**

A correlação é a dependência recíproca entre duas ou mais variáveis. Existem diversas técnicas para analisar a correlação em uma base de dados. O método utilizado no trabalho foi a Correlação de Pearson que pode ser detectada com o auxílio da Matriz de Correlação que faz uma distribuição dos Coeficientes de Correlação de Pearson em uma matriz de temperatura, onde as cores mais claras representam valores maiores, positivos ou negativos.

O cálculo do coeficiente de correlação é feito da seguinte forma:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Figura 1: Fórmula do coeficiente de correlação de Pearson: Sejam x_i e y_i os valores das variáveis X e Y. $\overline{x}$ e $\overline{y}$ são respectivamente as médias dos valores x_i e y_i. [18]

Com esse cálculo, é possível discernir quais variáveis possuem correlação, fazendo com que essa etapa torne-se parte essencial para o desenvolvimento das análises dos dados.

Entretanto, para uma correlação assertiva é necessário entender os tipos dos dados, nesse caso foram identificadas quatro principais variáveis categóricas que não entrariam na análise de correlação de Pearson.

Diante desse cenário, tornou-se indispensável aplicar análises distintas.

- **Análise chi-quadrado**

O teste estatístico chi-quadrado é aplicado determinar se existe relação ou associação entre duas variáveis categóricas. Diante disso, para as variáveis *IsActiveMember, Geography, Gender* e *HasCrCard* foi aplicada a análise chi-quadrado relacionando-as com a variável de *Churn*.

Há três maneiras principais para interpretar os valores:

- Valor p: quando for obtido um valor p menor que 0,05 (5%) é considerado significativo, ou seja, as variáveis possuem relação.
- Estatística chi-quadrado: Quanto maior a estatística chi-quadrado, maior a evidência de que as variáveis estão relacionadas. Ele precisa ser combinado com os valores críticos da tabela chi-quadrado para uma maior significância.
- Graus de liberdade: Os graus de liberdade são importantes para que seja possível identificar o valor crítico da tabela chi-quadrado, que é utilizado para determinar a significância do resultado.

- **Aplicação de modelos estatísticos**

Uma das formas de analisar a viabilidade e técnicas para evitar *churn* é a aplicação de modelos estatísticos preditivos. Para o presente trabalho, foram aplicados quatro modelos de preditivos, eles são capazes de indicar se um cliente é um potencial *churn*.

Os modelos foram selecionados por serem indicados para problemas de classificação, sendo eles: *SVM - Support Vector Machine, Logistic Regression, Random Forests e Decision Tree*. Ambos são recomendados para decisões de classificação binárias, como a previsão de *churn* (clientes que saem ou ficam).

O modelo *SVM - Support Vector Machine* é um algoritmo de aprendizado de máquina que tem como principal foco o treinamento e classificação de dados. Nesse algoritmo, são definidas fronteiras entre os dados, definindo o hiperplano que melhor diferencia as duas classes (clientes que sairam ou ficaram no banco). [13]

A *Logistic Regression* é um modelo de classificação binário simples e linear, ele possui uma fácil interpretação e funciona bem para relações entre variáveis preditoras aproximadamente lineares. [11]

A *Decision Tree* é um modelo de *machine learning* que divide o conjunto de dados em decisões hierárquicas, relacionando as variáveis afim de definir variáveis decisivas para a previsão de *churn*. É facilmente interpretável e permite visualizar a lógica em formato de árvore. [12]

Ainda no contexto de árvores, o modelo *Random Forest* é uma extensão da *Decision Tree*, sendo a combinação de várias árvores de decisão, transformando-o em uma solução mais robusta, reduzindo assim o overfitting. [14]

Além de serem modelos indicados para o problema, eles tem ampla utilização na literatura, o que evidencia seu excelente desempenho em contextos relacionados à rotatividade de clientes.

3. Resultados e Discussões

3.1. Limpeza dos dados

Por se tratar de uma base de dados pública do Kaggle [9] não foram encontradas grandes dificuldades para a manipulação e limpeza dos dados. Entretanto, três variáveis foram descartadas por não possuírem reflexos na saída desses clientes e apresentando irrelevância para a aplicação de testes: *RowNumber, CustomerId* e *Surname*.

Sendo assim, para a análise de dados restaram onze variáveis de maior relevância: *CreditScore, Geography, Tenure, Age, Gender, Balance, HasCrCard, NumOfProducts, IsActiveMember, EstimatedSalary* e *Churn*.

3.2. Análise descritiva estatística

Para iniciar as análises descritivas estatísticas, os dados foram resumidos em tabelas com medidas de síntese, conforme **Tabela 2**:

Estatísticas	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
Count	10	10	10	10	10	10
Mean	650.53	38.92	5.01	76485.89	1.53	100090.24
Std	96.65	10.48	2.89	62397.4	0.58	57510.49
Min	350	18	0	0	1	11.58
25%	584	32	3	0	1	51002.11
50%	652	37	5	97.198.54	1	100193.91
75%	718	44	7	127644.24	2	149388.25
Max	850	92	10	250898.09	4	199992.48

Tabela 2: Tabela de sínteses estatísticas das variáveis: CreditScore, Age, Tenure, Balance, NumOfProducts e EstimatedSalary

Por meio dela, é viável identificar algumas características nos dados que possibilitou análises mais profundas para possíveis clusterizações de clientes.

- **Análise dos dados**

Levando em conta que a base em questão é fictícia, alguns dados estão organizados de maneira pouco realista. A fim de facilitar a compreensão da distribuição desses dados, foram criados quatro gráficos: *Age, Tenure, Balance* e *EstimatedSalary*, como ilustrado na **Figura 2**. Esses gráficos permitem uma interpretação mais clara dos dados, tornando o processo mais acessível.

Figura 2: Gráficos de distribuição das variáveis: Age, Tenure, Balance e EstimatedSalary.

Ao examinar a distribuição dos dados, presente no gráfico apresentado na **Figura 2**, constatou-se que a idade e o saldo da conta bancária se comportam de uma forma praticamente normal.

Os dados relacionados ao saldo bancário (*Balance*), exibem um pico de valores zerados. Essa ocorrência é considerada comum no mundo real, já que muitas pessoas podem solicitar portabilidade de salário ou optar por não movimentar suas contas. É importante destacar que essas informações zeradas não influenciaram as análises realizadas.

Enquanto os dados de *Tenure* e *EstimatedSalary* apresentam uma distribuição equilibrada, com uma média de 5 anos de *Tenure* (tempo que é cliente no banco) e salário mínimo e máximo de 11,58 e 199.992,48, respectivamente.

Os dados relacionados a Estimativa Salarial fogem da realidade, visto que o salário médio anual para os países Alemanha, França e Espanha, segundo a Organização para Cooperação e Desenvolvimento Econômico (OECD), são respectivamente US$30.721 [15], US$34.375 [17] e US$27.155 [16]. Enquanto as médias da base são, respectivamente, US$119.730, US$62.093 e US$61.818.

- **Relacionando variáveis com *churn***

Para um melhor entendimento das variáveis relevantes, foram realizadas várias análises para identificar aquelas que mostram um comportamento distinto com base na rotulação de *Churn*. Essas análises podem ser visualizadas nos gráficos gerados, como mostrado na **Figura 3**.

Mediante a avaliação dos dados apresentados no Gráfico das Variáveis Categóricas (**Figura 3**), constata-se que as mulheres demonstram uma maior propensão a deixar de ser clientes. Além disso, é notório que ter um cartão de crédito não exerce influência sobre a decisão de sair do banco.

Outro aspecto relevante a ser destacado é que os clientes inativos apresentam uma maior tendência a deixar o banco. Tal fenômeno pode ser atribuído a diferentes motivos, como a ausência de suporte adequado,

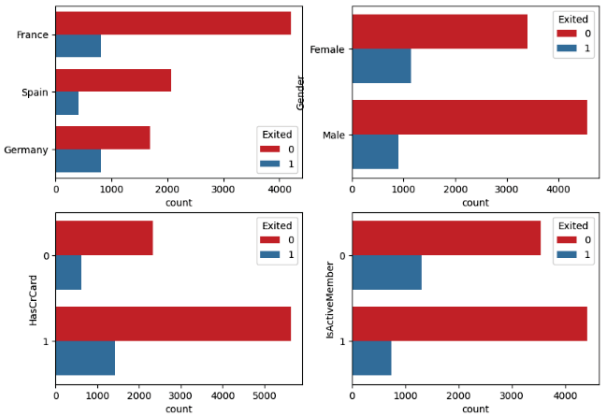

Figura 3: Gráfico de Variáveis Categóricas - Distribuição entre clientes que resultaram em Churn nas variáveis: Geography, Gender, HasCrCard e IsActiveMember

falta de identificação com os produtos oferecidos e insatisfação geral.

Ao avaliar as categorias presentes na **Figura 3**, repara-se que, embora a França tenha o maior número de clientes na base de dados, não é o país com a maior taxa de *churn*. Nesse contexto, verifica-se que a Alemanha apresenta uma porcentagem de *Churn* mais elevada, totalizando 32,44%.

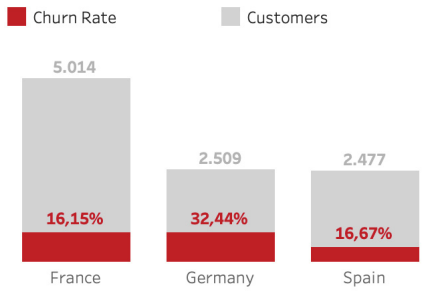

Figura 4: Churn - Países: Distribuição de Churn Rate por País

Adicionalmente, foi realizada uma análise utilizando gráficos boxplot **Figura 4**, permitindo uma visualização mais clara dos padrões e comportamentos dos clientes em relação às variáveis em estudo.

Essa abordagem proporcionou *insights* valiosos sobre a distribuição dos dados, identificação de valores discrepantes e comparação das distribuições entre diferentes grupos ou categorias.

Tendo como base as figuras **Figura 5** e **Figura 6**, destaca-se que em relação às idades dos clientes, nota-se que aqueles que permanecem no banco estão concentrados principalmente nas faixas etárias entre 35 e 40 anos, enquanto os que saem estão predominantemente na faixa entre 40 e 60 anos.

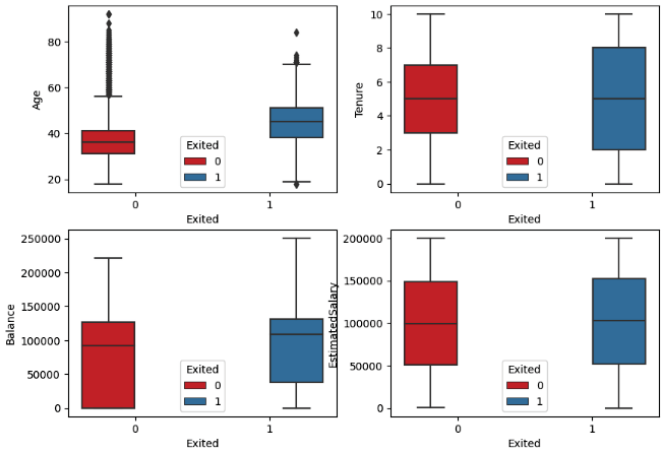

Figura 5: BoxPlot - *Age, Tenure, Balance* e *EstimatedSalary*

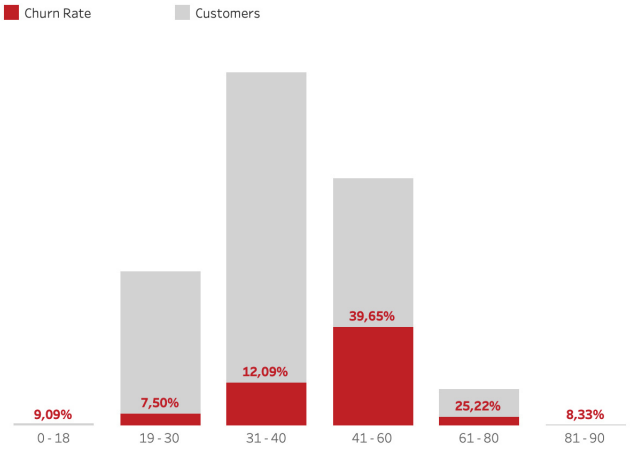

Figura 6: *Churn* - Age: Distribuição de Churn Rate por Age

Considerando a segmentação das idades apresentada na **Figura 6**, percebe-se que o banco está perdendo clientes de seu segundo maior grupo etário. Isso pode sugerir que a empresa não esteja focando corretamente no público-alvo, pois está está perdendo clientes em uma de suas faixas etárias mais representativas.

Outro aspecto importante é que os clientes com saldos mais altos em suas contas (*Balance*) são mais propensos a *Churn* em comparação com aqueles que possuem saldos mais baixos. Essa situação pode se tornar um problema significativo para os bancos, uma vez que clientes com saldos mais altos são fundamentais para a distribuição de caixa da empresa, dificultando assim a disponibilidade de recursos para empréstimos e outras atividades.

Além disso, notou-se que clientes que têm permanecido no banco por períodos de 2 a 3 anos e de 7 a 8 anos (*Tenure*) apresentam maior probabilidade de deixar a instituição, como pode ser visto na **Figura 7** apresentada a seguir. Isso pode indicar que o suporte oferecido pelo banco pode não estar atendendo adequadamente às necessidades desses clientes, o que influencia negativamente sua decisão de permanecerem.

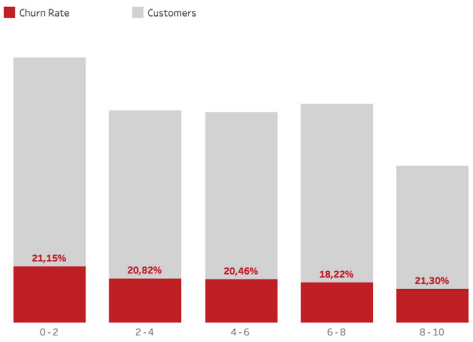

Figura 7: *Churn* - Tenure: Distribuição de Churn Rate por Tenure

Ao analisar a **Figura 5**, observa-se que não há grandes variações nas faixas em relação à métrica *EstimatedSalary*. No entanto, é preciso mencionar que esses dados podem não refletir a realidade, o que nos faz considerá-los com cautela em nossas análises.

- **Relacionando variáveis**

No contexto das ponderações previamente apresentadas, emerge a oportunidade de relacionar as variáveis que analisamos com uma das mais representativas, a idade do cliente.

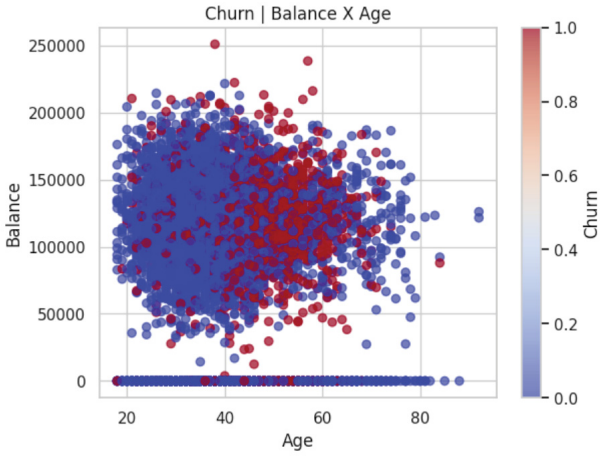

Figura 8: *Gráfico de Dispersão | Churn - Balance X Age*

Ao estabelecer uma comparação entre os dados de saldo de conta (Balance) e a idade dos clientes (*Age*), na **Figura 8**, nota-se que os saldos zerados das conta não representam, de forma expressiva, uma tendência ao *Churn*.

Contudo, um achado notável é a presença de uma probabilidade mais eminente para o *Churn* em todas

as faixas de saldo de conta quando associadas às idades situadas entre 40 e 60 anos.

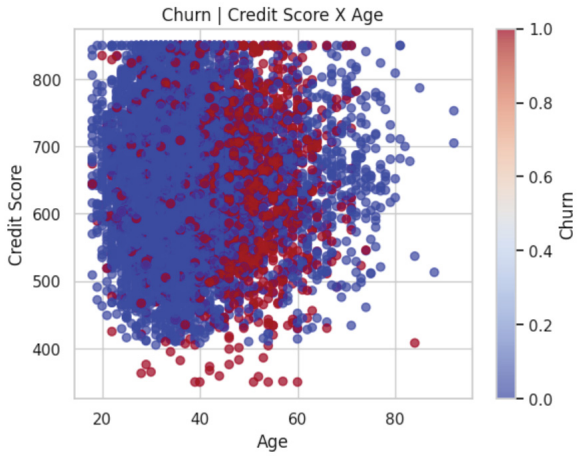

Figura 9: *Churn - Credit Score X Age*

Adicionalmente, foi feita uma análise comparativa entre idade dos clientes (*Age*) e o *Score* de Crédito (*CreditScore*), apresentada na **Figura 9**, que revelou que os clientes com pontuações de crédito mais baixas têm maior propensão a deixar o banco. Além disso, mantém-se evidente que a faixa etária compreendida entre 40 e 60 anos continua a apresentar a maior inclinação à evasão.

Por fim, realizamos uma análise detalhada das duas variáveis mais significativas (*Age* e *IsActiveMember*) na **Figura 10**, revelando que indivíduos acima de 40 anos que se encontram como clientes inativos apresentam uma propensão significativamente maior de abandonar a instituição bancária.

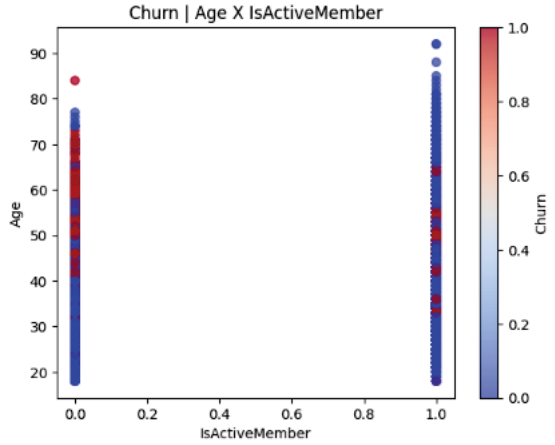

Figura 10: *churn - Age X IsActiveMember*

Esta constatação destaca a importância dessas variáveis como potenciais indicadores-chave para a previsão de *Churn* no contexto do banco em questão.

3.3. Análise de correlação

A matriz de correlação foi aplicada, conforme **Figura 11** abaixo, e nela, é possível observar algumas relações mais fortes com *churn*:

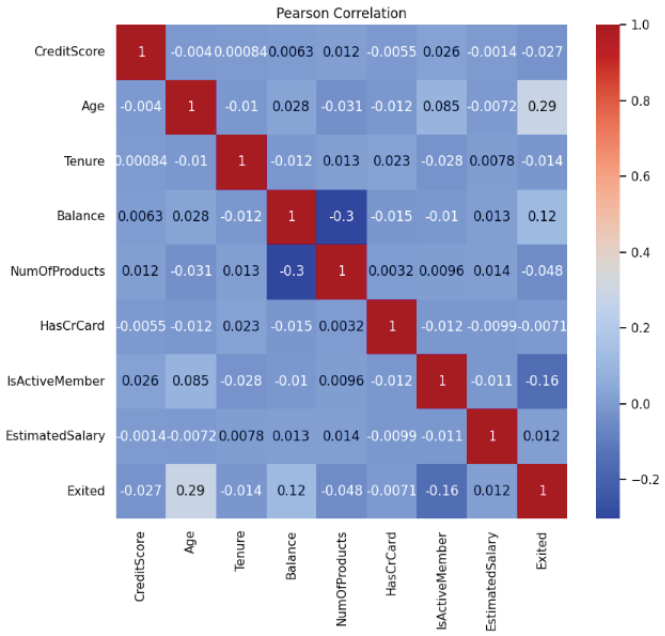

Figura 11: Matriz de Correlação de Pearson

Considerando que valores positivos na matriz de correlação apresentada acima, indicam uma correlação positiva, enquanto valores negativos indicam uma correlação negativa. Diante disso, é perceptível que as métricas *Age, Balance* e *EstimatedSalary* possuem uma correlação positiva mais representativa, comparando as outras métricas, com o *Churn*. Isso significa que, quando uma dessas variáveis aumenta, a probabilidade de o cliente sair (*Churn*) também aumenta.

Por outro lado, destaca-se que existe um valor negativo que se sobressai em relação aos outros resultados, sendo a variável IsActiveMember a mais relevante nesse contexto. Isso indica que clientes inativos têm maior tendência a deixar o banco (*Churn*).

3.4. Análise chi-quadrado

A análise chi-quadrado foi aplicada, tal como representado na **Figura 12** abaixo, para as variáveis descritivas da base (*IsActiveMember, Geography, Gender* e *HasCrCard*), relacionando-as com a variável de *Churn*.

A tabela apresentada na **Figura 12**, reforça as análises anteriores. Nela, verifica-se que é de grande relevância avaliar as métricas de ser um membro ativo, morar em determinado país e o gênero do cliente.

Variable	Chi_Statistic	P_value
IsActiveMember	242.98	8.78
Geography	301.25	3.83
Gender	112.91	2.25
HasCrCard	0.47	0.49

Figura 12: Tabela de métricas chi-quadrado

3.5. Aplicação de modelos de predição

Como visto anteriormente, um dos métodos para evitar o *Churn* é prever comportamentos dos clientes e algoritmos de *machine learning* focados em modelos preditivos são essenciais para isso.

O trabalho em questão utilizou quatro algoritmos diferentes: *Decision Tree, Random Forest, SVM e Logistic Regression*. Para ilustrar o teste dos modelos, foi elaborada um gráfico para de comparativo de acurácia entre os algoritmos, conforme **Figura 13**:

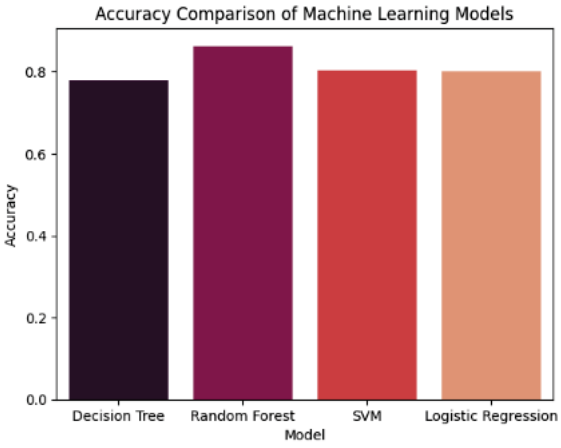

Figura 13: Comparativo de Acurácia - Modelos Preditivos

Devido à maior robustês do modelo de Floresta Aleatória (*Random Forest*), ele apresentou um desempenho superior comparando aos outros. Ainda assim, todos os modelos exibiram uma acurácia significativa, evidenciando que, ao analisar as variáveis corretas, é viável alcançar resultados satisfatórios.

Além do estudo de acurácia, foi desenvolvida uma análise dos escores das características do modelo. Ao avaliar o resultado da **Figura 14** que é o gráfico de Estudo de Scores do Modelo de Melhor Acurácia (*Random Forest*), pôde-se identificar que as principais métricas analisadas foram a idade, estimativa salarial, pontuação de crédito e saldo.

O resultado desse estudo indica que todas as análises anteriores estão de acordo com a aplicação dos

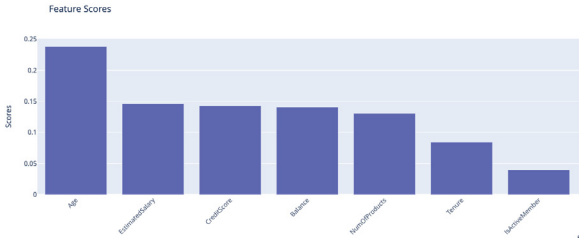

Figura 14: Estudo de Scores do Modelo de Melhor Acurácia (*Random Forest*)

modelos. Resultando assim, em um resultado satisfatório de um modelo que apresenta aproximadamente 87% de acurácia.

4. Conclusão

A Análise da viabilidade da aplicação de técnicas para evitar *churn* revela-se essencial para as empresas que buscam manter uma empresa saudável e com clientes fidelizados. Neste artigo foram abordadas diferentes técnicas estatísticas para compreender e prever a evasão de clientes (*Churn*).

Ao longo deste trabalho, foram abordadas as etapas principais do ciclo da ciência de dados: limpeza de dados e manipulação dos dados, análise descritiva estatística, análise de correlação, análise chi-quadrado e aplicação de modelos de predição.

Com tais etapas, esse artigo teve como objetivo identificar principais fatores de rotatividade dos produtos bancários, identificar características de clientes que tendem a sair, identificar características de clientes fiéis e identificar a viabilidade de técnicas para evitar *churn*.

Dentro desse contexto, tornou-se evidente que a faixa etária de 40 a 60 anos representa o grupo de clientes mais propenso a deixar o banco, especialmente aqueles que se mostram inativos. Diante dessa constatação, é possível melhorar essa situação por meio de campanhas de engajamento direcionadas a esse público específico, visando reter esses clientes e, ao mesmo tempo, focar em estratégias que atraiam a faixa etária correta do público-alvo.

Além de compreender os motivos pelos quais os clientes deixam o banco, é de extrema importância ressaltar que a análise de *churn* é um processo contínuo e dinâmico, pois os comportamentos dos clientes e do mercado estão em constante evolução. A volatilidade do cenário competitivo exige uma abordagem adaptativa e proativa para a retenção de clientes. Por esse motivo, a análise de *churn* se torna um compo-

nente fundamental para ser incorporado à cultura das instituições bancárias

Por fim, conclui-se que o fortalecimento e a retenção de clientes não é uma tarefa fácil, é um trabalho contínuo entre pessoas, máquinas e processos gerenciados e acima de tudo o comprometimento com a satisfação dos clientes. O presente trabalho identificou as principais ferramentas capazes de manter o cliente satisfeito e consequentemente, reter o mesmo na organização.

Agradecimentos

B.C.B. gostaria de agradecer ao meu orientador André Grégio e a todos os professores do curso DSBD - UFPR.

Referências

[1] Sharma, A., Panigrahi, D., & Kumar, P. (2013). A neural network based approach for predicting customer churn in cellular network services. *arXiv preprint arXiv:1309.3945*.

[2] Xiao, J., Xiao, Y., Huang, A., Liu, D., & Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems*, 43(1), 29–51.

[3] Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. *Journal of Retailing*, 80, 317-329.

[4] Wen, Z., Yan, J., Zhou, L., Liu, Y., Zhu, K., Guo, Z., Li, Y., & Zhang, F. (2018). Customer churn warning with machine learning. In *The Euro-China Conference on Intelligent Data Analysis and Applications*, pp. 343–350. Springer.

[5] Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, 38, 187–195.

[6] GANESH, J.; ARNOLD, M. J.; REYNOLDS, K. E. Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers. *Journal of Marketing*, v. 64, n. 3, p. 65-87, 2000. DOI: 10.1509/jmkg.64.3.65.18028.

[7] BRASIL. Lei No. 13.709, de 14 de agosto de 2018. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 15 ago. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 30 jun. 2023.

[8] BANCO CENTRAL DO BRASIL. Open Finance. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/openfinance. Acesso em: 30 jun. 2023.

[9] Kaggle. Predicting Churn for Bank Customers. Disponível em: https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers. Acesso em: 30 jun. 2023.

[10] ILYAS, Ihab F.; CHU, Xu. *Data Cleaning*. Morgan & Claypool, 2019. 282 páginas.

[11] Scikit-learn Developers. "Sklearn.linear_model.LogisticRegression - Scikit-Learn 0.21.2 Documentation." *Scikit-Learn.org*, 2014, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[12] Scikit-learn Developers. "1.10. Decision Trees — Scikit-Learn 0.22 Documentation." *Scikit-Learn.org*, 2009, https://scikit-learn.org/stable/modules/tree.html.

[13] Scikit-learn Developers. "1.4. Support Vector Machines — Scikit-Learn 0.20.3 Documentation." *Scikit-Learn.org*, 2018, https://scikit-learn.org/stable/modules/svm.html.

[14] Scikit-learn Developers. "3.2.4.3.1. Sklearn.ensemble.RandomForestClassifier — Scikit-Learn 0.20.3 Documentation." *Scikit-Learn.org*, 2018, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[15] OECD Better Life Index. Germany - PT. Disponível em: https://www.oecdbetterlifeindex.org/pt/paises/germany-pt/. Acesso em: 27 de julho de 2023.

[16] OECD Better Life Index. Título do país: Espanha - PT. Disponível em: <https://www.oecdbetterlifeindex.org/pt/paises/spain-pt. Acesso em: 27 de julho de 2023.

[17] OECD Better Life Index. Título do país: França - PT. Disponível em: https://www.oecdbetterlifeindex.org/pt/paises/france-pt/. Acesso em: 27 de julho de 2023.

[18] Silvia Shimakura, LEG/UFPR - Universidade Federal do Paraná. *Coeficiente de Pearson 2005-11-08*. Disponível em: http://www.leg.ufpr.br/~silvia/CE701/node79.html. Acesso em: 27 de julho de 2023.