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

Joelle Wosiack

Comparação de estratégias de Machine Learning: um exemplo em análise de churn

Curitiba
2025

Joelle Wosiack

Comparação de estratégias de Machine Learning: um exemplo em análise de 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. Wagner Hugo Bonat

Curitiba
2025

Comparação de estratégias de Machine Learning: um exemplo em análise de churn

Comparison of Machine Learning Strategies: An Example in Churn Analysis

Joelle Wosiack[1], Wagner Hugo Bonat[2]

[1]Aluna do programa de Especialização em Data Science & Big Data, jojoelle.wosiack@gmail.com
[2]Professor do Departamento de Estatística - DEST/UFPR, wbonat@gmail.com

O presente trabalho tem como objetivo analisar a efetividade de diferentes arquiteturas de aprendizado de máquina na previsão do *churn* em uma base de dados fictícios provinda do *Kaggle*. O *churn*, ou cancelamento de serviços e/ou produto por parte dos clientes é um desafio comum em diversos setores e sua prevenção é crucial para a manutenção da competitividade das empresas. A partir dos dados disponíveis, é realizado a limpeza deles, a análise descritiva das variáveis e outras análises cabíveis, com os dados processados, é aplicado três modelagens diferentes, sendo a regressão logística como base, e dois modelos mais robustos, o *Random Forest* e redes neurais. As redes neurais são modelos sofisticados e populares atualmente, e apesar do grande potencial delas, para o contexto e a base de dados utilizadas, elas não possuem a melhor performance. O estudo destaca a importância de selecionar o algoritmo adequado para o contexto específico da análise, reforçando que a sofisticação do algoritmo não garante, necessariamente, a melhor performance.
Palavras-chave: *Churn*, análise de dados, *random forest*, redes neurais

This study aims to analyze the effectiveness of different machine learning architectures in predicting churn in a fictitious data set from Kaggle. Churn, or cancellation of services and/or products by customers, is a common challenge in several sectors and its prevention is crucial to maintaining the competitiveness of companies. Based on the available data, data are cleaned, descriptively analyzed, and other appropriate analyses are performed. With the processed data, three different models are applied, the basis being logistic regression, and two more robust models, Random Forest and neural networks. Neural networks are sophisticated and popular models today, and despite their great potential, for the context and database used, they do not have the best performance. The study highlights the importance of selecting the appropriate algorithm for the specific context of the analysis, reinforcing that the sophistication of the algorithm does not necessarily guarantee the best performance.
Keywords: Churn, data analysis, random forest, neural networks

1. Introdução

Churn refere-se ao abandono ou cancelamento de um serviço e/ou produto por parte dos clientes, é algo comum em diversos setores no qual o cliente mantém uma relação contínua com a empresa [1]. De modo geral, manter um cliente já existente é mais barato do que adquirir um novo. Segundo estudos de Philip Kotler os custos de atrair um novo cliente pode ser 5 vezes maior do que manter um atual [2]. Assim, para as empresas, conseguir reduzir as taxas de *churn* pode significar vantagens competitivas no mercado.

Identificar o *churn* e seus motivos pode ajudar empresas a melhorar seus produtos, serviços e relacionamento com os clientes. A taxa de *churn* alta pode significar problemas com a qualidade dos produtos ou serviços prestados, falhas no atendimento aos clientes, precificação errada, entre outros problemas potenciais [3].

Reduzir *churn* significa aumentar a retenção dos clientes, e isso está estritamente ligado à previsibilidade de receita e lucratividade da empresa. Clientes fiéis tendem a comprar mais, fazer propaganda orgânica da empresa e contribuem para o crescimento do negócio [4].

Prever o *churn* antes de que este aconteça, dá margem às empresas de agir proativamente, tendo estratégias mais eficientes de retenção dos clientes, aumentando a eficiência operacional e reduzindo desperdícios de recursos em ações desnecessárias.

Apesar dos benefícios conhecidos da prevenção do *churn*, a predição deste não é algo simples, possui desafios técnicos e estratégicos. Identificar precocemente o cliente que tem potencial de virar *churn* é complicado, pois às vezes é difícil entender os sinais de que isso irá acontecer, a coleta de dados confiáveis e suficientes para entender o cenário completo às vezes é muito delicada de ser feita, a ação rápida antes que o *churn* ocorra também é difícil, a personalização na abordagem aos clientes por vezes é necessárias, mas impossível, entre outros problemas.

Pensando em tudo isso, a área da ciência de dados oferece várias técnicas que permitem transformar os dados disponíveis em *insights* úteis. Técnicas de aprendizado de máquina permitem a criação de modelos capazes de aprender padrões complexos e fazer previsões baseadas neles, a fim de identificar previamente clientes que podem virar *churn* e também para identificar o melhor tipo de campanhas de retenção destes.

Com isso, o presente trabalho busca analisar a efetividade de técnicas de aprendizado de máquina com a finalidade de indicar fatores associados ao *churn* para que ações futuras possam ser pensadas. Por se tratar de uma análise necessária para inúmeros setores do mercado, espera-se que o presente estudo possa servir de referência para empresas que buscam avaliar técnicas de prevenção de *churn*.

2. Materiais e métodos

2.1. Materiais

Após a Lei Geral de Proteção de Dados - LGPD (Lei n° 13.709) [5] ser sancionado em agosto de 2018, com o intuito de garantir a segurança dos dados, encontrar dados públicos ficou mais difícil, portanto para a elaboração do trabalho em questão, foi utilizado dados fictícios de uma base do *Kaggle* [6]. A base contém 10.000 registros, sendo cada um de um cliente diferente, dos quais 20,37% deles são usuários que cancelaram o serviço, ou seja, são *churn*. Além disso, a base conta com 14 variáveis, conforme consta na Tabela 1.

Variável	Descrição
RowNumber	Número da linha
CustomerId	Identificador do cliente
Surname	Sobrenome do cliente
CreditScore	Pontuação de crédito
Geography	País de origem do cliente
Gender	Gênero do cliente
Age	Idade do cliente
Tenure	Tempo de vínculo
Balance	Saldo na conta
NumOfProducts	Número de produtos
HasCrCard	Flag de cartão de crédito
IsActiveMember	Flag de cliente ativo
EstimatedSalary	Salário estimado
Exited	Flag de churn

Tabela 1: Variáveis disponíveis na base de dados.

Em posse da base de dados, foram efetuados alguns tratamentos nos dados para melhor entendimento deles e também com o intuito de prepará-los para a utilização nos algoritmos de aprendizado de máquina.

2.2. Métodos

O processo foi iniciado através da limpeza de dados, que é uma etapa fundamental para garantir a qualidade dos dados, isso influencia diretamente na precisão, utilidade e confiabilidade da análise e dos modelos posteriormente criados [7]. A limpeza dos dados evita distorções e vieses na análises realizadas, pode melhorar a performance dos modelos, garante padronização e coerência, reduz retrabalho e consequentemente custos, além de facilitar a visualização e interpretação dos dados.

Com a finalidade de analisar os dados disponíveis, foram utilizadas algumas bibliotecas padrões do *Python*, como *Pandas* [8], *Numpy* [9], *Seaborn* [10] e *Matplotlib* [11].

Após a limpeza, é realizada a análise dos dados, etapa esta que permite entender o comportamento das variáveis, detectar padrões claros, validar hipóteses iniciais, selecionar variáveis relevantes e auxiliar na escolha de algoritmos que façam sentido para os dados disponíveis.

É nesta etapa que há o entendimento da distribuição das variáveis numéricas, que ajuda a identificar dispersão, simetrias ou assimetrias nos dados. Vemos também a frequência de variáveis categóricas para identificar possíveis desbalanceamentos nas categorias, detectar erros de codificação e entender o perfil dos clientes. É analisada a correlação entre as variáveis, a comparação geral entre clientes que são *churn* e que não são e analisado a variável alvo.

Se durante a análise de dados é percebido alguma informação fora do comum, como uma idade de cliente muito superior a cem anos, por exemplo, é possível re-

alizar a limpeza de dados novamente, ou seja, não são etapas sequenciais fixas, mas apenas um fluxo geral e comum a se seguir.

Por fim é realizado a modelagem dos dados, é nessa parte que os dados são transformados em previsões acionáveis, ou seja, com o histórico de clientes, é possível prever se um novo cliente é propenso a se tornar *churn*. É através da modelagem que é possível identificar padrões mais complexos, que são difíceis ou invisíveis na análise dos dados.

Existem várias técnicas de modelagem que podem ser utilizadas para a predição do *churn* [12], aqui foram utilizados três modelos, sendo o mais simples deles a regressão logística disponível no *Scikit-learn* [13], que é um modelo estatístico muito utilizado em problemas de classificação binária, ele modela a probabilidade de ocorrência de um evento (neste caso o *churn*) como uma função logística (*sigmóide*) dos atributos de entrada, produzindo coeficientes que indicam a direção e força da influência de cada variável, possuindo uma boa interpretabilidade e baixa complexidade computacional. Foi utilizado o *Random Forest* também disponível no *Scikit-learn* [14], que é um modelo de aprendizado em conjunto, ou seja, ele combina várias árvores de decisão, no qual cada árvore é treinada com uma amostra aleatória dos dados e a decisão final é feita por votação. E por fim criado alguns modelos neurais com o *TensorFlow* [15], redes neurais artificiais são modelos inspirados no cérebro humano, compostos por camadas de nós (também chamados de neurônios, de onde vem a nomeação) que aprendem representações mais complexas dos dados. É um tipo de modelagem que permite boa performance em problemas complexos, com muitos dados e muitas *features*, porém eles possuem uma baixa interpretabilidade, por vezes até mesmo considerados "caixas-pretas".

Como forma de analisar os modelos, foram utilizadas métricas disponíveis no *Scikit-learn* [16] que servem para avaliar modelos de classificação binária. Foi utilizado a matriz de confusão que é uma tabela que resume o desempenho do modelo mostrando o número de previsões corretas e incorretas, separados pelas classes disponíveis, através dele é possível entender onde o modelo está errando. E também utilizado o relatório de classificação, no qual é possível ver a precisão do modelo, que é entre os clientes que o modelo previu como *churn*, quantos realmente são *churn*, a sensibilidade do modelo (*recall*), entre os clientes que realmente são *churn*, quantos o modelo conseguiu identificar, e a média harmônica entre precisão e sensibilidade, conhecido como *F1-Score*. As métricas fornecem uma visão mais completa do desempenho dos modelos para cada classe, permitindo avaliar a performance de cada modelo.

Estatística	CreditScore	Balance	EstimatedSalary
Média	650,58	76.485,89	100.090,24
Desvio pad.	96,60	62.397,41	57.510,49
Mínimo	350	0	11,58
25%	584	0	51.002,11
50%	652	97.198,54	100.193,92
75%	718	127.644,24	149.388,25
Máximo	850	250.898,09	199.992,48

Tabela 2: Análise estatística de algumas variáveis.

Para finalizar, foi visto como os modelos trabalham e quais as variáveis que foram mais importantes para a predição deles.

3. Resultados

O processo foi iniciado por meio da limpeza dos dados, mas devido a procedência da base ser do *Kaggle* [3], não houve dificuldades nessa etapa, mas ela não foi descartada, pois foram desconsideradas três variáveis, são elas: *RowNumber*, *CustomerId* e *Surname*, visto que elas não permitem a generalização dos dados e são insignificantes para a análise e para a modelagem.

Após isso, foram realizadas diversas análises nos dados, desde algumas simples, como entender a distribuição de quantos usuários possuem crédito ativo, que representam 70,55%, e quantos não possuem crédito ativo, que são 29,45%, ou então entender que 51,51% dos usuários são membros ativos, enquanto 48,49% não são ativos, e outras análises mais avançadas.

Foi realizada a análise descritiva das variáveis, no qual é possível perceber na Tabela 2 que o saldo na conta, em muitos casos, está zerado, representando pouco mais de 36% da base total. Além disso, a idade média dos clientes é de 39 anos, sendo que o mais novo tem 18 anos e o mais idoso possui 92 anos.

Durante a análise foi possível constatar que a idade e a pontuação da conta se comportam de forma normal, que o tempo de vínculo e o salário estimado possuem uma distribuição uniforme, como pode ser visualizado na Figura 1, o que reflete o fato da base ser criada a partir de dados fictícios, portanto reproduzem pouco a realidade. A análise das mesmas variáveis em contraste a coluna *Exited*, coluna alvo também seguem de maneira muito uniforme.

Modelo	F1-Score	F1-Score (Não Churn)	Precisão	Recall
Regressão Logística	0.34	0.89	0.40	0.30
Random Forest	0.62	0.92	0.65	0.59
Rede Neural	0.60	0.90	0.63	0.58

Tabela 3: Métricas do modelos.

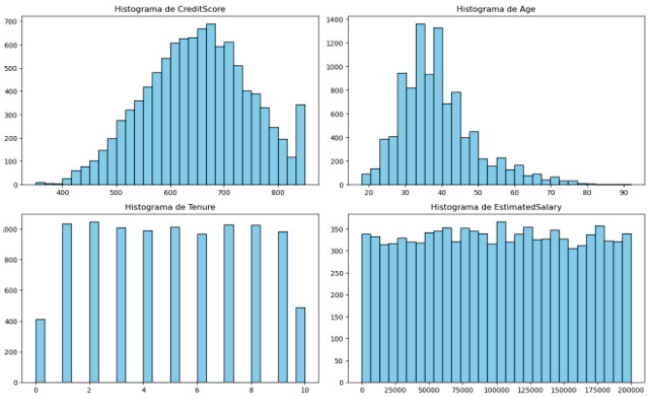

Figura 1: Histograma das variáveis CreditScore, Age, Tenure e EstimatedSalary.

Após as análises, mas antes da modelagem, as variáveis categóricas de localização e gênero foram convertidas em variáveis binárias (0 ou 1), criando uma nova coluna para cada categoria distinta.

Então foi realizada a modelagem, com três tipos de modelos sendo utilizados: Regressão Logística, *Random Forest* e Redes Neurais, todos avaliados por meio de métricas extraídas com *Scikit-learn*. A Tabela 3 resume as principais métricas obtidas para cada modelo.

A regressão logística obteve desempenho razoável, porém mostrou dificuldades em prever corretamente os casos de *churn*. O modelo de *Random Forest* se destacou em todas as métricas e obteve a melhor performance geral.

Foram testadas diferentes arquiteturas de redes neurais, a mais robusta utilizada foi composta por três camadas ocultas (64, 32 e 16 neurônios), com funções de ativação *ReLU*, além de *BatchNormalization* e *Dropout*, que servem para treinar melhor e evitar o *overfitting*. A camada de saída possuía um neurônio com ativação *sigmoide*. Apesar dos ajustes, os modelos neurais não superaram o desempenho do *Random Forest*.

O modelo de *Random Forest* também permitiu extrair a importância relativa das variáveis para a predição de *churn*, conforme Figura 2:

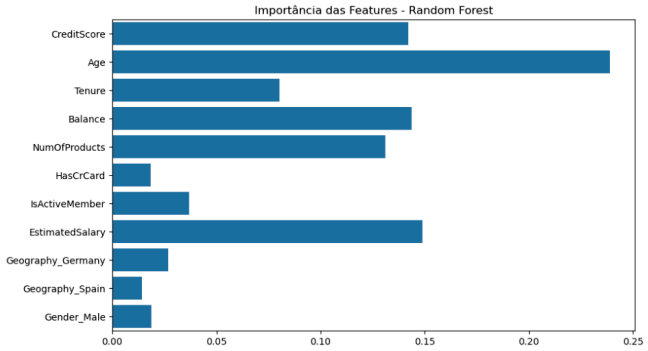

Figura 2: Importância das variáveis do *Random Forest*.

Esses dados são especialmente úteis para orientar ações estratégicas de retenção, permitindo focar esforços em clientes mais propensos ao *churn* com base em seu perfil comportamental e financeiro.

4. Discussão

Ao longo do estudo realizado, foram testados diversos modelos de aprendizado de máquina sobre uma base de dados fictícia, com o intuito de avaliar a capacidade preditiva dos algoritmos e entender suas diferenças de desempenho. O fato dos dados serem sintéticos não traduzem a realidade, tanto a realidade dos dados no mundo real em si, como também do desempenho dos modelos, visto que se os dados fossem reais, a performance dos modelos possivelmente seriam diferente, talvez até obtendo um resultado final diferente de qual modelo foi melhor. Ainda assim, os testes realizados fornecem uma base útil para compreender o funcionamento das técnicas aplicadas e seus possíveis usos em contextos práticos.

A expectativa inicial era de que as redes neurais apresentassem o melhor desempenho entre as três abordagens, especialmente considerando o destaque que esse tipo de arquitetura tem recebido nos últimos anos. As redes neurais são de fato muito poderosas e versáteis, podendo lidar bem com problemas complexos e não lineares. No entanto, os resultados obtidos no presente trabalho mostram que nem sempre complexidade e robustez geram o melhor resultado. Para a base de dados disponível e o problema em questão, o modelo *Random Forest* demonstrou ter a melhor performance. Isso reforça a importância de avaliar diferentes algoritmos para cada contexto específico, pois nem sempre o algoritmo mais popular e sofisticado será o mais adequado.

O potencial de uso do modelo ajustado, *Random Forest*, está na sua capacidade de antecipar o compor-

tamento do *churn*, para a base analisada, de forma prática e interpretável. Empresas que operam com grandes volumes de clientes podem utilizar esse tipo de modelo como ferramenta de apoio à decisão, automatizando a identificação de clientes com maior risco de abandono. A partir disso, é possível realizar intervenções direcionadas, como ofertas personalizadas, contato proativo com o cliente ou reavaliação de condições contratuais.

Além disso, a análise das covariáveis mais importantes revela informações estratégicas para a prevenção do *churn*. Por exemplo, variáveis como idade, saldo em conta e salário estimado podem indicar que clientes mais idosos, com maior saldo e maior engajamento, tendem a permanecer mais tempo, enquanto clientes jovens, com pouco saldo e inatividade na conta, apresentam maior risco de saída.

Com base nisso, a empresa pode desenvolver estratégias segmentadas:

- Para clientes com baixo saldo, pode-se oferecer produtos ou serviços de engajamento, como programas de cashback, programas de milhas ou consultorias financeiras.
- Para clientes com salário estimado mais baixo, que podem ter menor poder de compra, podem ser abordados pacotes mais acessíveis, que promovam valor sem comprometer seu orçamento.

Por fim, os resultados não apenas evidenciam o desempenho superior do modelo *Random Forest* no cenário estudado, mas também mostram como o uso inteligente das variáveis preditoras pode subsidiar ações concretas de retenção de clientes, contribuindo para a redução de *churn* de maneira direcionada e eficiente.

5. Agradecimentos

Gostaria de agradecer a todos os docentes do curso de especialização em Data Science e Big Data que contribuíram para o meu aprendizado, e para a realização deste trabalho, em especial ao meu orientador Wagner Bonat. Agradecer também à minha família que me incentivou a dar continuidade aos meus estudos.

Referências

[1] O que é churn?. *Zendesk*, 2025. Disponível em https://www.zendesk.com.br/blog/o-que-e-churn/#. Acesso em: 06/05/ 2025.

[2] A segunda compra do cliente é mais importante do que a primeira. *Exame*, 2025. Disponível em: https://exame.com/colunistas/relacionamento-antes-do-marketing/a-segunda-compra-do-cliente-e-mais-importante-do-que-a-primeira/. Acesso em: 06/05/ 2025.

[3] Análise de churn: guia prático para empresas. *Stripe*, 2024. Disponível em: https://stripe.com/br/resources/more/churn-analysis-101-a-how-to-guide-for-businesses. Acesso em: 06/05/2025.

[4] Entenda porque investir na prevenção do churn e como direcionar estratégias personalizadas. *Cinnecta*, 2022. Disponível em: https://cinnecta.com/conteudos/prevencao-de-churn/. Disponível em: 06/05/2025.

[5] 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: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 07 maio 2025.

[6] Churn Modelling. *Kaggle*, 2025. Disponível em: https://www.kaggle.com/datasets/shubh0799/churn-modelling/data. Acesso em: 06/05/2025.

[7] Entenda por que a limpeza de dados é tão importante quanto a coleta. *Cortex*, 2020. Disponível em: https://www.cortex-intelligence.com/blog/inteligencia-de-mercado/limpeza-de-dados. Acesso em: 07/05/2025.

[8] pandas documentation. *Stripe*, 2025. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 07/05/2025.

[9] NumPy Documentation. *Stripe*, 2025. Disponível em: https://numpy.org/doc/. Acesso em: 07/05/2025.

[10] An introduction to seaborn. *Stripe*, 2024. Disponível em: https://seaborn.pydata.org/tutorial/introduction. Acesso em: 07/05/2025.

[11] Matplotlib: Visualization with Python. *Stripe*, 2024. Disponível em: https://matplotlib.org/. Acesso em: 07/05/2025.

[12] Previsão de churn: como escolher o melhor modelo de previsão para sua empresa. *Stripe*, 2024. Disponível em: https://stripe.com/br/resources/more/churn-prediction-101-how-to-choose-the-best-prediction-model-for-your-business. Acesso em: 07/05/2025.

[13] LogisticRegression. *Scikit-learn*. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>. Acesso em: 07/05/2025.

[14] RandomForestClassifier. *Scikit-learn*. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.

RandomForestClassifier.html. Acesso em: 07/05/2025.

[15] Uma plataforma completa de machine learning. *TensorFlow*. Disponível em: https://www.tensorflow.org/?hl=pt-br. Acesso em: 07/05/2025.

[16] Metrics and scoring: quantifying the quality of predictions. *Scikit-learn*. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Acesso em: 07/05/2025.