**Predição do churn: Comparando técnicas de Machine Learning**

Carolaine Rangel Rodrigues¹\*; Emerson Aparecido Mouco Júnior2

1 Arquivista pela Universidade Federal do Espírito Santo. Rua Esther Oliveira Galveas, 185 – Jardim Camburi; 29092-260 Vitória, Espírito Santo, Brasil

2 Universidade de São Paulo. Escola Superior de Agricultura "Luiz de Queiroz” (ESALQ). Av. Pádua Dias, 11; 13418

- 900 Piracicaba, São Paulo, Brasil

\*autor correspondente: rodrigues.crangel@gmail.com

**Predição do churn: Comparando técnicas de Machine Learning**

**Resumo**

A retenção de clientes é crucial para o sucesso de uma empresa sendo mais econômica do que a aquisição de novos clientes. A taxa de rotatividade de clientes, conhecida como “churn”, é um indicador importante calculado pela proporção de clientes que cancelam o serviço em relação ao total de clientes em um período específico. Em mercados dinâmicos, o “churn” pode estar relacionado a insatisfação do cliente, estratégias competitivas, novos produtos ou regulamentações. Este trabalho propõe uma análise comparativa de técnicas de machine learning para predição de “churn” em uma empresa com modelo de assinatura mensal. Foram investigadas as técnicas de Aprendizagem de Árvores de Decisão, “Naïve Bayes” e “Random Forest”, a avaliação dos modelos incluiu métricas de acurácia, precisão, recall, matriz de confusão, curva ROC e Área da curva Área sob a curva [AUC] e para otimização dos hiperparâmetros da Árvore de Decisão foi utilizado a técnica de “Grid Search”. Este estudo visa contribuir para o avanço teórico e prático em análise preditiva de “churn”, oferecendo dados valiosos para gestores de empresas, melhorando a retenção de clientes e maximizando seu valor ao longo do tempo. As técnicas escolhidas são amplamente utilizadas por sua capacidade de lidar com grandes volumes de dados e identificar padrões complexos, sendo comparadas quanto as suas vantagens e limitações na predição de “churn”.

**Palavras-chave:** Retenção de clientes; Machine Learning; Aprendizado de Árvore, Random Forest, Naïve Bayes.

**Churn Prediction: Comparing Machine Learning Techniques**

**Abstract**

Customer retention is crucial for a company's success, being more cost-effective than acquiring new customers. The customer churn rate, known as "churn," is an important indicator calculated by the proportion of customers who cancel the service relative to the total number of customers in a specific period. In dynamic markets, churn may be related to customer dissatisfaction, competitive strategies, new products, or regulations. This paper proposes a comparative analysis of machine learning techniques for predicting churn in a company with a monthly subscription model. The techniques investigated included Decision Tree Learning, Naïve Bayes, and Random Forest. The model evaluation included metrics such as accuracy, precision, recall, confusion matrix, ROC curve, and Area Under the Curve (AUC). Grid Search was used to optimize the hyperparameters of the Decision Tree. This study aims to contribute to the theoretical and practical advancement of churn predictive analysis, offering valuable data to business managers, improving customer retention, and maximizing their lifetime value. The chosen techniques are widely used for their ability to handle large volumes of data and identify complex patterns, being compared in terms of their advantages and limitations in churn prediction.

**Keywords:** Customer Retention; Machine Learning; Decision Tree Learning; Random Forest; Naïve Bayes.

**Introdução**

A retenção de clientes é um dos pilares fundamentais para o sucesso de uma empresa e tem sido uma das principais preocupações, pois adquirir novos clientes podem custar cinco vezes mais caros do que reter os ativos. A taxa de rotatividade de clientes é conhecida como taxa de “churn” que, para uma empresa que possui um serviço mensal, é calculado pela divisão do número de clientes de cancelaram o serviço em determinado período pelo número total de cliente no mesmo período. (Kurtz e Clow, 1998).

Em um mercado dinâmico, a taxa de “churn” pode estar relacionada ao baixo nível de satisfação do cliente, estratégias competitivas agressivas, novos produtos ou regulamentações. Para visualizar essa taxa é necessária uma análise dos dados que podem ser apresentados em painel de visualização. (Seid e Woldeyohannis, 2022)

A análise preditiva de “churn” utilizando técnicas de machine learning tornou-se uma área de pesquisa e aplicação essencial para empresas. Essas técnicas permitem que as empresas identifiquem padrões nos dados dos clientes e prevejam sinais em clientes iniciais que terão a maior probabilidade de cancelar o seu serviço ou produto (Saias et al., 2018).

Diante desde contexto, este trabalho propõe realizar uma análise comparativa de técnicas de machine learning aplicadas à predição de “churn” de uma empresa que possui um modelo de assinatura mensal. Foram investigadas e comparadas abordagens como Aprendizagem de Árvores de Decisão, “Naïve Bayes” e “Random Forest”. O objetivo é avaliar a eficácia, acurácia e precisão das três técnicas em prever o “churn” considerando uma base de dados artificiais com base em afirmações semelhantes para o mundo real.

Para alcance do objetivo será explorado e comparado os modelos Aprendizagem de Árvores de Decisão, “Naïve Bayes” e “Random Forest”. Cada técnica será dividida para ser treinada e posteriormente testada. Para avaliação será apresentado métricas como acurácia, precisão, recall, além da apresentação da matriz de confusão, a curva de ROC e Área sob a curva [AUC].

Este estudo visa fornecer dados sobre as vantagens e limitações de cada técnica de machine learning para a predição de “churn” em uma empresa com modelo de assinatura mensal. Ao contribuir para o avanço do conhecimento teórico e prático em análise preditiva de “churn”, este trabalho oferece avaliações valiosas para os gestores de empresas, ajudando-os a melhorar a retenção de clientes e maximizar o valor dos clientes ao longo do tempo.

Além disso, as técnicas de machine learning escolhidas para este estudo são amplamente utilizadas em diversas indústrias devido à sua capacidade de lidar com grandes volumes de dados e identificar padrões complexos. Árvores de Decisão são conhecidas pela sua facilidade de interpretação e simplicidade, “Naïve Bayes” é eficiente para classificações probabilísticas em grandes conjuntos de dados, enquanto “Random Forest” combina múltiplas árvores de decisão para melhorar a precisão e a robustez das previsões. A aplicação dessas técnicas no contexto de predição de “churn” não só permite uma análise comparativa robusta mas também oferece uma visão prática sobre como essas metodologias podem ser implementadas em ambientes empresariais reais para melhorar a tomada de decisão estratégica.

O artigo está organizado da seguinte maneira: na seção 2, serão apresentadas as técnicas de aprendizagem e os critérios de avaliação; na seção 3, serão expostos os resultados da simulação; e na seção 4, serão discutidas as conclusões sobre os modelos apresentados.

**Metodologia ou Material e Métodos**

O conjunto de dados utilizado para este estudo foi obtido de uma empresa fictícia que presta serviços no modelo de assinatura mensal, representando uma amostra de clientes ao longo de um período de 6 meses. Os dados consistem em variáveis como localização geográfica, plano contratado, dados de utilização do sistema (vendas e pedidos dos últimos dias, módulos contratados), modelo de negócio, canal de aquisição, se a empresa pertence a uma franquia, forma de pagamento, se houve desconto no momento da contratação, quantidade de chamados ao time de suporte, e fase do ciclo de vida que indica se o cliente cancelou ou não o seu contrato durante o período de observação.

Os dados foram pré-processados para lidar com valores ausentes e normalizar variáveis numéricas. Além disso, foi realizado uma análise exploratória para entender a distribuição das variáveis e identificar possíveis relações.

Para análise foram selecionados três modelos de machine learning para comparação da predição do “churn”, sendo: Aprendizagem de Árvores de Decisão [AD], “Naïve Bayes” [NB] e “Random Forest” [RF]. Para aplicação de cada modelo será implementado pacotes específicos disponíveis no Software R como: “rpart” para AD, “naivebayes” para NB, “randomForest” para RF e “ggplot2” “rpart.plot” para criação dos gráficos.

O modelo de AD constrói uma árvore composta de nós de decisão e folhas. O modelo NB é utilizado para classificar grandes conjuntos de dados, calculando a probabilidade de uma instância pertencer a cada classe com base nas probabilidades condicionais dos preditores. Já o modelo RF combina técnicas de ensemble com árvores de decisão para fazer previsões em conjuntos de dados complexos e de alta dimensionalidade, criando várias árvores de decisão durante o treinamento e fazendo previsões agregadas (Schneider, 2016).

Para a avaliação do desempenho dos modelos, foram utilizadas métricas como acurácia, que apresenta a proporção de previsões corretas em relação ao total de previsões; precisão, que apresenta a proporção de verdadeiros positivos em relação ao total de positivos previstos; recall, que é a proporção de verdadeiros positivos em relação ao total de verdadeiros positivos na amostra; além da matriz de confusão, curva ROC e Área sob a Curva [AUC]. Também foi aplicado o procedimento de Grid Search para estimar os melhores hiperparâmetros, garantindo um desempenho otimizado do modelo de AD (Strauss et al., 2022).

O software R, versão 4.4.1, foi utilizado para a validação dos modelos em um ambiente de desenvolvimento integrado no RStudio. A comparação dos modelos foi realizada através da análise da matriz de confusão, precisão e recall, com o objetivo de avaliar o desempenho de cada um (Oliveira, 2008).

Este trabalho foi conduzido de acordo com os princípios éticos da pesquisa científica, no qual garante a confidencialidade e anonimato dos dados dos clientes, não utilizando informações sensíveis e anonimizando todos os dados

**Resultados e Discussão**

A base de dados contém 2.717 clientes que adquiriram os serviços de uma empresa nos primeiros seis meses de 2024. O percentual de "churn" em relação ao total da amostra é de 43%. As variáveis incluídas na amostra estão descritas na Tabela 1.

Tabela 1. Variáveis da amostra

|  |  |  |
| --- | --- | --- |
| Tipo | Nome | Descrição |
| Factor | ativo | Cliente ativou ou cancelado (1 ou 0) |
| Numérico | total\_tickets\_suporte | Total de chamados do cliente no suporte |
| Numérico | vendas\_30\_dias | Faturamento dos clientes nos últimos 30 dias |
| Numérico | pedidos\_30\_dias | Número de venda dos clientes nos últimos 30 dias |
| Factor | modelo\_negocio | Modelo de atuação do cliente |
| Factor | forma\_pagamento | Forma de pagamento (boleto ou cartão) |
| Factor | plano\_escolhido | Plano adquirido pelo cliente |
| Factor | desconto | Desconto aplicado na aquisição do cliente |
| Factor | franquia | Cliente participante de uma franquia |
| Numérico | mensalidade | Valor que o cliente paga para utilizar a plataforma |
| Factor | UF | Estado do cliente |
| Factor | canal | Canal de aquisição de captura do cliente |
| Factor | plano\_pagamento | Pagamento mensal ou anual |
| Numérico | pedidos\_delivery | Faturamento dos clientes nos últimos 30 dias no delivery |
| Numérico | vendas\_delivery | Número de venda dos clientes nos últimos 30 dias no delivery |
| Numérico | dias\_boleto\_vencido | Quantidade de dias que o cliente não realizou o pagamento do sistema |

Fonte: Resultados originados da pesquisa

Durante a etapa de pré-processamento, foram verificados os tipos de cada variável e realizados os ajustes necessários. Além disso, foi conferida a existência de valores "null" e feita uma análise de balanceamento dos dados para identificar possíveis tendências nos resultados. Não foram encontrados valores faltantes, e a base de dados apresentou-se balanceada. Em seguida, os dados foram divididos em conjuntos de treino e teste, sendo 70% dos dados alocados para o treino.

Inicialmente, foi construído um modelo de AD com um parâmetro de complexidade “cp = 0.0028”. Ao analisar o custo de complexidade da árvore de treino, foi possível identificar um erro de 43%, uma divisão de 13 nós, um erro relativo de 0,16 e um erro de validação cruzada [cross-validation error - xerror] de 0,40. Na Figura 1, além de visualizar o erro relativo e o tamanho da árvore, é possível observar o grau de importância de cada variável.

Gráfico

Descrição gerada automaticamente com confiança baixa

Figura 1: Análise de “cp” e grau de importância das variáveis no modelo AD

Fonte: Resultados originados da pesquisa

A precisão de 95% observada no modelo de treino e de 84% no modelo de teste indica que há espaço para melhorias por meio da aplicação de técnicas de otimização. A Tabela 2 permite a comparação das métricas, enquanto a Figura 2 apresenta as curvas ROC e as áreas sob a curva (AUC) dos modelos de treino e teste, revelando desempenhos ligeiramente distintos.

Tabela 2. Avaliação do modelo inicial proposto, com “cp=0. 0028”, modelo AD

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1902 | 0.95 | | 0.93 | 0.92 | 0.93 | 0.96 |
| Teste | 815 | 0.84 | | 0.82 | 0.85 | 0.78 | 0.85 |

Fonte: Resultados originados da pesquisa

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figura 2: Curvas ROC do modelo proposto nas bases de treino e teste modelo AD

Fonte: Resultados originados da pesquisa

Diante dos resultados insatisfatórios do modelo inicial, decidiu-se aplicar o método de "Grid Search" para otimizar os parâmetros "minsplit", "maxdepth", "minbucket" e "cp", com o objetivo de minimizar o erro. A Tabela 3 mostra uma leve melhoria nos resultados obtidos.

Tabela 3. Avaliação do modelo final após utilização da técnica de "Grid Search"

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1902 | 0.97 | | 0.94 | 0.93 | 0.96 | 0.97 |
| Teste | 815 | 0.84 | | 0.83 | 0.85 | 0.80 | 0.86 |

Fonte: Resultados originados da pesquisa

A Figura 3 apresenta a matriz de confusão após o ajuste dos parâmetros, evidenciando um número relativamente alto de verdadeiros positivos (TP) e verdadeiros negativos (TN), juntamente com uma quantidade reduzida de falsos positivos (FP) e falsos negativos (FN), especialmente na base de treino. Isso sugere que o modelo está se comportando de maneira eficaz após a otimização.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figura 3: Matriz de confusão nas bases de treino e teste modelo AD

Fonte: Resultados originados da pesquisa



Em seguida, foi testado o modelo Naïve Bayes (NB), escolhido pela sua simplicidade e eficiência em problemas de classificação com grandes volumes de dados. Utilizou-se o parâmetro "Laplace" igual a 1 para evitar probabilidades nulas. Os resultados das métricas avaliadas entre os conjuntos de treino e teste estão apresentados na Tabela 4. Observa-se uma precisão bastante próxima entre os conjuntos, com uma taxa de acerto acima de 90% em ambos os casos. No entanto, nota-se uma capacidade limitada em identificar apenas 57% dos casos verdadeiros negativos [TN] e falsos negativos [FN] na base de teste, sendo inferior ao modelo AD.

Tabela 4. Avaliação do modelo proposto, com “laplace=1 no modelo NB

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1902 | 0.95 | | 0.75 | 0.59 | 0.96 | 0.78 |
| Teste | 815 | 0.94 | | 0.74 | 0.57 | 0.95 | 0.76 |

Fonte: Resultados originados da pesquisa

Ao analisar as curvas ROC e as áreas sob a curva (AUC), conforme apresentado na Figura 4, observa-se que tanto a base de treino quanto a de teste apresentam áreas com valores semelhantes. Isso indica que o modelo possui uma capacidade consistente de discriminação entre as classes em ambas as bases. A proximidade dos valores sugere que o modelo não está super ajustado, mantendo uma performance equilibrada ao generalizar para novos dados.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figura 4: Curvas ROC do modelo proposto nas bases de treino e teste no modelo NB

Fonte: Resultados originados da pesquisa

Como mencionado anteriormente, o modelo demonstrou uma sensibilidade de 58% na base de teste. Esta observação é reforçada ao examinar a matriz de confusão na Figura 5, onde se nota um número significativamente maior de falsos negativos (FN) em ambas bases.

Gráfico, Gráfico de barras, Gráfico de cascata

Descrição gerada automaticamente

Figura 5: Matriz de confusão nas bases de treino e teste modelo NB

Fonte: Resultados originados da pesquisa

Por fim, avaliou-se o modelo Random Forest (RF), reconhecido por sua robustez e capacidade de generalização. Para este estudo, optou-se por utilizar "ntree=200" para assegurar um equilíbrio adequado entre viés e variância. Os resultados estão apresentados na Tabela 5, onde observa-se uma precisão superior a 90% tanto na base de treino quanto na base de teste, alcançando taxas de acerto de 98% e 85%, respectivamente. No entanto, nota-se que a especificidade na base de teste, ou seja, a capacidade de identificar corretamente os casos negativos, foi significativamente inferior à da base de treino.

Tabela 5. Avaliação do modelo proposto, com “ntree=200”, modelo RF

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1847 | 0.99 | | 0.98 | 0.96 | 0.99 | 0.99 |
| Teste | 790 | 0.88 | | 0.87 | 0.90 | 0.82 | 0.93 |

Fonte: Resultados originados da pesquisa

O desempenho da curva ROC e AUC mostrou uma performance quase perfeita, chegando a quase 100%, na base de treino. No entanto, ao analisar a base de teste, observa-se uma pequena queda conforme ilustrado na Figura 6.

Gráfico

Descrição gerada automaticamente

Figura 6: Curvas ROC do modelo proposto nas bases de treino e teste modelo RF

Fonte: Resultados originados da pesquisa

O modelo revelou um menor valor na avaliação de verdadeiros negativos (TN) no modelo de treino comparado a base de teste, conforme evidenciado na Figura 7. Isso indica que a especificidade na base de teste é menor em comparação com a base de treino.

Gráfico, Gráfico de cascata

Descrição gerada automaticamente

Figura 7: Matriz de confusão nas bases de treino e teste modelo RF

Fonte: Resultados originados da pesquisa

Ao comparar os modelos utilizando a de precisão e recall (sensibilidade), conforme Figura 8, observamos que o modelo RF se destaca significativamente. Ele demonstra uma alta precisão e AUC tanto no conjunto de treino quanto no conjunto de teste, acompanhada por uma sensibilidade robusta, indicando sua capacidade de identificar corretamente a maioria dos casos positivos sem comprometer a precisão global. Em contraste, os modelos AD e NB exibem desempenhos variados. O AD mostra uma boa precisão e AUC no treino, mas uma sensibilidade mais baixa no teste, sugerindo dificuldade em capturar todos os casos positivos. O NB, embora mantenha alta precisão, apresenta uma AUC menor que demais modelos e uma sensibilidade limitada em ambos os conjuntos de dados, indicando sua dificuldade em identificar corretamente os casos positivos.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figura 8: Comparativo da precisão e “recall” dos modelos

Fonte: Resultados originados da pesquisa

Esses resultados ressaltam a importância de considerar não apenas a precisão, mas também o recall e AUC ao avaliar modelos de classificação, especialmente em contextos em que a identificação de todos os casos positivos é crucial.

**Considerações Finais**

Os resultados da análise indicam que o modelo Random Forest se destacou em termos de precisão e acurácia, tanto na base de treino quanto na de teste, sendo seguido pelo modelo de Árvores de Decisão e, por último, o Naïve Bayes. A utilização da curva ROC e da Área sob a Curva (AUC) permitiu avaliar a performance dos modelos, revelando que o Random Forest teve um desempenho significativamente superior em comparação com os demais. Este resultado não apenas valida a escolha do Random Forest para a predição de “churn”, mas também destaca sua capacidade de lidar com a complexidade dos dados, proporcionando previsões mais confiáveis.

A análise das matrizes de confusão confirmou que o modelo Random Forest apresentou uma notável capacidade de identificar corretamente os casos de “churn”, reduzindo a incidência de falsos negativos e falsos positivos. Essa eficácia é especialmente crítica em ambientes de negócios, onde a retenção de clientes se traduz diretamente em sustentabilidade e crescimento. A implementação de Grid Search para ajuste de hiperparâmetros foi fundamental, permitindo otimizar os modelos e extrair o máximo de seu potencial preditivo.

Além disso, a comparação dos modelos oferece insights valiosos sobre suas aplicabilidades. Enquanto o Random Forest se mostrou superior em termos de precisão e sensibilidade, o modelo de Árvores de Decisão é uma alternativa viável quando a interpretabilidade é necessária. Por outro lado, o Naïve Bayes, embora eficiente, apresentou limitações na identificação de casos positivos, o que sugere que sua aplicação deve ser cuidadosamente considerada em contextos críticos. Assim, a escolha do modelo deve ser baseada não apenas em métricas de desempenho, mas também nas especificidades e demandas do ambiente de negócios.

Por fim, esta análise reforça a importância de uma abordagem abrangente ao escolher e avaliar modelos de machine learning para a predição de “churn”. A combinação de técnicas de otimização e uma avaliação meticulosa das métricas permite não apenas melhorar a precisão preditiva, mas também contribuir para uma tomada de decisão mais informada. O estudo não só avança o conhecimento teórico na área de análise preditiva, mas também fornece recomendações práticas que podem ser implementadas por gestores para melhorar a retenção de clientes e maximizar seu valor ao longo do tempo.

**Referências**

Albrecht, T; Baier,D. (2020) Churn analysis using deep learning: Customer Classification from a Practical Point of View. Archives of Data Science, Series 6. Disponível em: <https://publikationen.bibliothek.kit.edu/1000134950/119767270>. Acesso em 23 mar.2024.

Kurtz, D. L; Clow, K. E. (1998). Services marketing. New York: John Wiley & Sons.

Oliveira. A.F.G. (2008) Testes estatísticos para comparação de médias. Revista Eletrônica Nutritime, 5: 777-778. Disponível em: < https://www.nutritime.com.br/wp-content/uploads/2020/02/Artigo-076.pdf>. Acesso em 22 mar. 2024.

Saias. J; Rato, L; Gonçalves, T. (2018) Estudo sobre a predição de churn. Relatório Técnico. Projeot APRA-CP: 2. Disponível em: <http://hdl.handle.net/10174/30175>. Acesso em: 17 mar. 2024.

Schneider. P. H. (2016) Análise preditiva de churn com ênfase em técnicas de machine learning: uma revisão. Dissertação de Mestrado. Fundação Getúlio Vargas, Escola de Matemática Aplicada. Rio de Janeiro, RJ. Brasil. Disponível em: <https://repositorio.fgv.br/server/api/core/bitstreams/51f11184-37a8-4cd6-94fd-07a330203506/content>. Acesso em 21 mar. 2024.

Seid, M.H; Woldeyohannis, M.M. (2022). Customer churn prediction using machine learning: commercial bank of ethiopia. In Proceedings of the 2022 International Conference on Information and Communication Technology for Development in Africa (ICT4DA). Disponível em: < https://ieeexplore.ieee.org/document/9971224>. Acesso em: 17 mar. 2024.

Seymen, O.F; Dogan, O; Hiziroglu.A.(2021) Customer churn prediction using deep learning. Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition: 520-529. Disponível em: <https://www.researchgate.net/publication/350906568>. Acesso em: 17 mar. 2024.

Strauss, E; Júnior, M.V; Ferreira, W.L.L. (2022) A importância de utilizar métricas adequadas de avaliação de performance em modelos preditivos de machine learning. Projectus 7: 52-62. Disponível em: <https://doi.org/10.15202/25254146.2022v7n2p52>. Acesso em 23 mar. 2024.

Tsai, C.F; Lu, Y.H. (2009) Customer churn prediction by hybrid neural networks. 36:12547-12553. Disponível, em:<https://www.sciencedirect.com/science/article/abs/pii/S0957417409004758>. Acesso em 19 mar. 2024.

Vafeiadis, T; Diamantaras, K.I; Sarigiannidis, G; Chatzisavvas, K.C. (2015) A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55: 1-9. Disponível em: <<https://doi.org/10.1016/j.simpat.2015.03.003>>. Acesso em 23 mar. 2024.

Xie, Y; Li, X; Ngai, E.W.T; Ying, W. (2009) Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36: 5445-5449. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0957417408004326>. Acesso em 17 mar. 2024.

Zhang, Y; Luo, S. (2018). Customer churn prediction using deep learning. Expert Systems with Applications: 105, 92-101.