**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 AUC. Foi utilizado “Grid Search” para otimização dos hiperparâmetros. 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.

**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 será 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.

**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 serão pré-processados para lidar com valores ausentes e normalizar variáveis numéricas. Além disso, será realizada 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 e “randomForest” para RF.

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, serão 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 será aplicado o procedimento de Grid Search para estimar os melhores hiperparâmetros, garantindo um desempenho otimizado dos modelos (Strauss et al., 2022).

Será utilizado o software R versão 4.3.2 para validação dos modelos, que será executado em ambiente de desenvolvimento integrado RStudio. A comparação dos modelos será realizada por meio da comparação da matriz de confusão, precisão e recall para identificar o desempenho dos modelos (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 Preliminares**

A base de dados possui 2637 clientes que adquiriram o serviço de uma empresa durante o ano de 2023. O percentual de “churn” sobre o total da amostra é de 28.5%, e para evitar um desbalanceamento das classes, foi utilizado o método “oversampling” a fim de obter os melhores resultados com o modelo (Viloria et al. 2020). As variáveis contidas na amostra são apresentadas 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 |
| 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 |

Fonte: Resultados originados da pesquisa

Durante a etapa de pré-processamento, foram excluídas as colunas desnecessárias, ajustando o tipo das variáveis e transformando as variáveis com valores faltantes em “0”. Posteriormente, foi realizada a separação da base em treino e teste, com 70% dos dados concentrados na base de treino.

Inicialmente, foi construído um modelo de AD com um parâmetro de complexidade “cp = 0.0028”. Ao verificar o custo de complexidade da árvore de treino, foi possível identificar um erro de 49.61% e o grau de importância de cada variável, conforme Figura 1.

Uma imagem contendo Gráfico

Descrição gerada automaticamente

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, com 2663 nós, sugere um possível "overfitting". No entanto, ao aplicar as mesmas regras na base de teste, os resultados divergiram significativamente, indicando um claro "overfitting", conforme mostrado na Tabela 2. A Figura 2 compara as curvas ROC e as AUC dos modelos de treino e teste, revelando desempenhos bastante distintos.

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1847 | 0.95 | | 0.77 | 0.72 | 0.90 | 0.83 |
| Teste | 790 | 0.47 | | 0.26 | 0.28 | 0.19 | 0.79 |

Fonte: Resultados originados da pesquisa

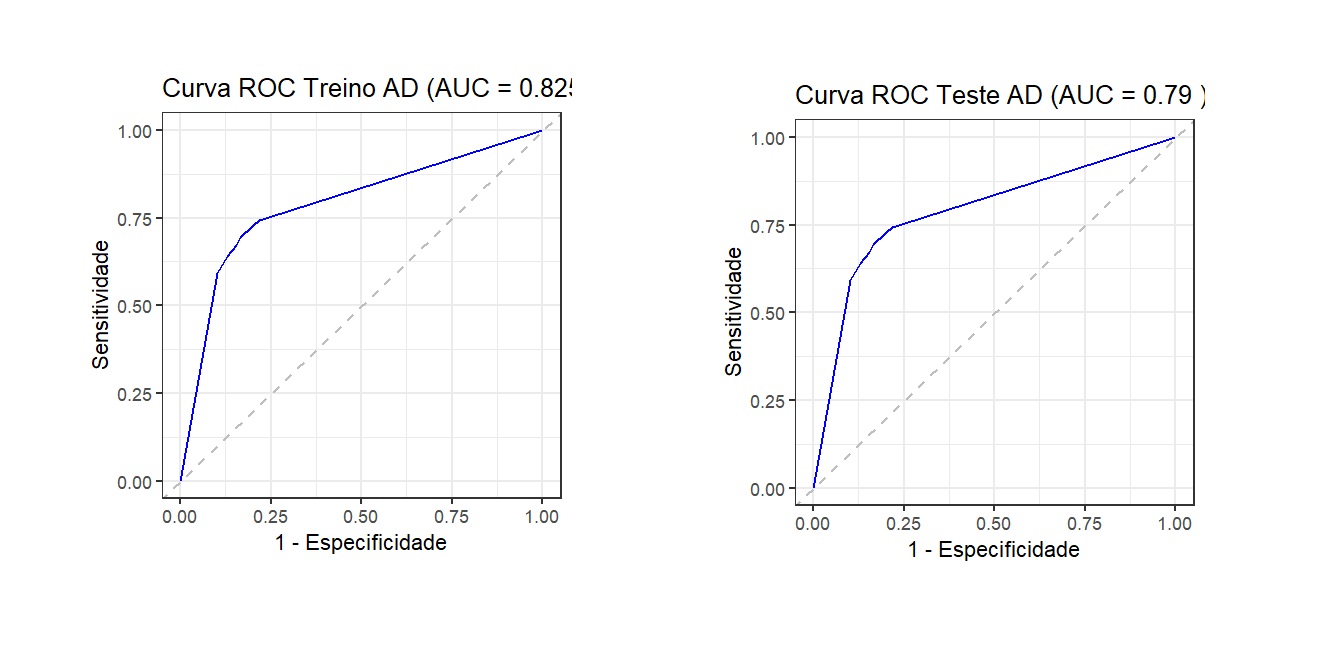


Figura 2: Curvas ROC do modelo proposto nas bases de treino e teste modelo AD

Fonte: Resultados originados da pesquisa

Diante dos resultados insatisfatórios do modelo inicial, optou-se por utilizar o método de "Grid Search" para otimizar os parâmetros como "minsplit", "maxdepth", "minbucket" e "cp", visando minimizar o erro entre eles. No entanto, ao validar o modelo final, observamos uma precisão de apenas 19% na base de treino e 46.5% na base de teste, revelando um desempenho ainda abaixo do esperado. A Figura 3 mostra a matriz de confusão após o ajuste dos parâmetros, destacando um número relativamente alto de verdadeiros positivos [TP] e verdadeiros negativos [TN], mas também uma quantidade significativa de falsos positivos [FP] e falsos negativos [FN], especialmente na base de teste.

Gráfico, Gráfico de cascata

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 3. Observa-se uma precisão bastante próxima entre os conjuntos, com uma taxa de acerto de 55% em ambos os casos. No entanto, nota-se uma capacidade limitada em identificar apenas 38% dos casos verdadeiros negativos [TN] e falsos negativos [FN].

Tabela 3. Avaliação do modelo proposto, com “laplace=1 no modelo NB

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1847 | 0.98 | | 0.55 | 0.38 | 0.98 | 0.32 |
| Teste | 790 | 0.96 | | 0.55 | 0.38 | 0.96 | 0.68 |

Fonte: Resultados originados da pesquisa

Ao analisar as curvas ROC e AUC, conforme apresentadas na Figura 4, nota-se que a base de treino teve um desempenho substancialmente inferior, enquanto a base de teste apresentou um desempenho moderado.

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 38%. Esta observação é reforçada ao examinar a matriz de confusão na Figura 5, onde se nota um número significativamente maior de falsos negativos (FN) e falsos positivos (FP) em comparação aos verdadeiros positivos (TP) e verdadeiros negativos (TN).

Gráfico, 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 4, 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 4. Avaliação do modelo proposto, com “ntree=200”, modelo RF

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Observações | | Precisão | Acurácia | Sensitivi-dade | Especificida-de | AUC |
| Treino | 1847 | 0.95 | | 0.98 | 0.98 | 0.96 | 0.99 |
| Teste | 790 | 0.94 | | 0.78 | 0.85 | 0.60 | 0.84 |

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 queda de quase 20%, conforme ilustrado na Figura 6.

Gráfico, Gráfico de dispersão

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 uma queda significativa na avaliação de verdadeiros negativos (TN) no modelo 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

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 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 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 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 ao avaliar modelos de classificação, especialmente em contextos em que a identificação de todos os casos positivos é crucial.

**Considerações Finais**

Os resultados preliminares indicam que o modelo Random Forest apresentou o melhor desempenho em termos de precisão e acurácia, tanto na base de treino quanto na base de teste, seguido pelo modelo Naïve Bayes e, por último, o modelo de Árvores de Decisão. A curva ROC e a AUC foram usadas para avaliar a performance dos modelos, mostrando que o modelo RF teve um desempenho significativamente superior em comparação com os outros dois modelos.

A análise das matrizes de confusão mostrou que o modelo RF teve uma maior capacidade de identificar corretamente os casos de “churn”, reduzindo a incidência de falsos negativos e falsos positivos. A utilização de técnicas de “oversampling” foi eficaz para balancear as classes e melhorar a performance dos modelos preditivos.

A utilização de Grid Search para ajuste de hiperparâmetros foi essencial para otimizar os modelos e obter os melhores resultados possíveis. No entanto, o modelo de Árvores de Decisão ainda apresentou problemas de “overfitting”, mesmo após os ajustes, o que sugere a necessidade de explorar outras técnicas de regularização ou modelos mais robustos.

Os próximos passos incluem a aplicação de técnicas de ensemble e a avaliação de outros algoritmos de machine learning para continuar melhorando a predição de “churn” e fornecer “insights” mais precisos para a retenção de clientes.

**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.

Viloria, A; Lezamab,O.B.P; Caruzo, N.M. (2020) Unbalanced data processing using oversampling: Machine Learning. Procedia Computer Science, 175:108-113. Disponível em: https://doi.org/10.1016/j.procs.2020.07.018>. Acesso em 25 abr. 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.