ANÁLISE COMPARATIVA DE MODELOS TRADICIONAIS E DE APRENDIZADO DE MÁQUINA PARA PROBLEMAS DE CREDIT SCORE

Palavras-Chave: Aprendizado de máquina, Credit score, Classificação de padrões

Autores(as):

Gabriel Henrique Cerqueira, UNICAMP

Prof.a Dra. Ivette Luna, UNICAMP

INTRODUÇÃO

A pontuação de crédito é essencial para a avaliação do risco financeiro dos solicitantes de empréstimos, influenciando a decisão de concessão e as taxas de juros aplicadas. Tradicionalmente, sistemas de pontuação de crédito utilizam métodos manuais, avaliando fatores como histórico de pagamentos e uso de crédito, com técnicas como Regressão Logística e Análise de Sobrevivência. No entanto, os avanços em aprendizado de máquina têm permitido a automação e a consideração de variáveis mais sofisticadas, como padrões de gastos e renda. Neste contexto, esta pesquisa visa comparar os sistemas tradicionais de pontuação de crédito com os métodos modernos baseados em aprendizado de máquina, focando em identificar as diferenças na eficácia e eficiência desses métodos. A análise de crédito é crucial para a economia moderna, facilitando o acesso a empréstimos para diversos fins, como aquisição de imóveis e investimentos. O uso de técnicas de aprendizado de máquina pode aprimorar o processo, oferecendo maior precisão e identificação de padrões que podem não ser visíveis para os analistas humanos.

Estudos como o de TUKSON et al. (2016), mostram que modelos de aprendizado de máquina superam os métodos tradicionais de pontuação de crédito, com uma confiabilidade aproximada de 80%. Além disso, DASTILE et al. (2020) sugere a inclusão de variáveis macroeconômicas para uma avaliação mais completa.

Assim, este estudo busca aprofundar a compreensão sobre como o aprendizado de máquina pode aprimorar os sistemas de pontuação de crédito, avaliando sua viabilidade e potencial para melhorar a precisão e eficiência desses sistemas em instituições financeiras.

METODOLOGIA

A pesquisa utiliza dados públicos do banco de dados de Taiwan (I-Cheng, Y. 2016), composta por 23 variáveis e 30.000 observações, sendo 17 variáveis numéricas e 6 variáveis categóricas, além da variável resposta binária que assume o valor 1 se o cliente não realizar o pagamento pendente (default) e 0, caso contrário.

A Tabela 1 descreve a estrutura da base utilizada e a natureza de cada variável considerada na pesquisa.

NOME	Descrição	Tipo
ID	Identificador	Inteiro
LIMIT_BAL	Limite de Crédito	Inteiro
Sex	Sexo do usuário	1 = Homem; 2 = Mulher
EDUCATION	Nível de Escolaridade	Qualitativa (1 = pós-graduação; 2 = universidade; 3 = ensino médio; 4 = outros)
MARRIAGE	Estado Civil	Qualitativa (1 = casado; 2 = solteiro; 3 = outros)
AGE	Idade em anos	Inteiro
PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6	Status de Pagamento (abril a setembro de 2005)	Qualitativa (-1 = pagamento pontual; 1 = atraso de 1 mês; 2 = atraso de 2 meses; ...; 8 = atraso de 8 meses; 9 = atraso de 9 meses ou mais)
BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6	Valor da Fatura (abril a setembro de 2005)	Inteiro
PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	Valor Pago (abril a setembro de 2005)	Inteiro
default_payment_next month	Pagamento Pendente no Próximo Mês	Binário. 1 = Pagamento Pendente; 0 =Pagamento não Pendente

Tabela 1: descrição da estrutura da base de dados

Foram utilizadas quatro técnicas de modelagem preditiva para avaliar o risco de crédito e a elegibilidade de clientes: Árvores de Decisão, Florestas Aleatórias, Regressão Logística e KNN (K-Nearest Neighbors).

A Regressão Logística é amplamente utilizada para problemas de classificação binária e é definida pela equação (GUJARATI, 2011).:

$$P_i = \frac{1}{1 + e^{(-Z_i)}} \quad (1)$$

onde

$$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Sendo $X_{ji}, j = 1, \ldots, k$ as variáveis explanatórias ou atributos para cada observação. O "thresholding" é utilizado para determinar a classificação final com base nas probabilidades estimadas (Huang, X. 2023). A técnica de Eliminação Recursiva de Atributos (RFE) é empregada para selecionar as características mais importantes e reduzir a dimensionalidade, melhorando a precisão e a capacidade de generalização do modelo (Guyon, I. Elisseeff, A. 2003). Para a estimação dos parâmetros, o Método de Newton-Cholesky é usado para otimizar a função de verossimilhança, ajustando os parâmetros até a convergência (Tanveer, A. Huaxin, C. 2020).

Por outro lado, o algoritmo KNN classifica novos casos com base na votação majoritária dos seus K vizinhos mais próximos, utilizando uma função de distância como a distância Minkowski, definida por:

$$d_m(X, Y) = (\frac{1}{n}\sum_{i=1}^{n} |x_i - y_i|^{\lambda})^{\frac{1}{\lambda}}$$

onde $\lambda > 1$ é a ordem da distância, utilizando neste caso λ=2, que representa a distância Euclidiana (MERIGÓ J. CASANOVAS, M., 2010). O algoritmo KNN é simples e eficaz para amostras pequenas e grandes, mas o processamento pode ser intensivo para grandes conjuntos de dados, e a sensibilidade a valores atípicos é uma limitação importante (Turkson, R. et al., 2016).

As Árvores de Decisão são modelos que segmentam os dados em subconjuntos com base em critérios de divisão, como o ganho de informação ou o índice de Gini (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O objetivo é reduzir a heterogeneidade dos subconjuntos em relação à variável alvo, minimizando a entropia ou o erro quadrático. Essas árvores geram regras do tipo SE-ENTÃO, onde cada nó representa uma decisão com base em um atributo, e cada folha representa um resultado de classificação ou previsão (ZHANG; HU; PATTNAIK, 1999). Apesar de serem eficazes e interpretáveis, as árvores de decisão podem sofrer com sobreajuste (QUINLAN, 1993). Técnicas como poda e uso de comitês de máquinas (ensembles) como florestas aleatórias são usadas para melhorar a robustez e precisão do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O modelo de Florestas Aleatórias é um comitê de máquinas composto por múltiplas árvores de decisão independentes, treinadas em subconjuntos aleatórios dos dados usados para o treinamento do comitê (BREIMAN, 2001). Os subconjuntos aleatórios são gerados via a técnica de bootstrap aggregating (bagging). A predição final é obtida pela agregação das predições das árvores, seja pela média (regressão) ou pela moda (classificação) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). As Florestas Aleatórias são robustas ao sobreajuste, conseguem lidar com muitas variáveis e detectar interações complexas, fornecendo uma medida da importância de cada variável. Contudo, o treinamento e a predição podem ser computacionalmente intensivos, e a combinação de muitas árvores reduz a interpretabilidade do modelo (ZHENG; PADMANABHAN; MUKHERJEE, 2007).

RESULTADOS E DISCUSSÃO

Para o ajuste dos modelos e a posterior validação dos desempenhos, o conjunto de dados foi dividido em conjunto de treino (70%) e de teste (30%), usando o método train_test_split no Python.

A Regressão Logística foi identificada usando a técnica de eliminação recursiva de características (RFE) com o solver 'newton-cholesky' para selecionar as cinco principais variáveis preditoras. A Tabela 2 mostra que o modelo alcançou uma acurácia de 0,81, precisão de 0,79, Log Loss de 6,93, recall de 0,81, AUC-ROC de 0,59 e F1-Score de 0,76.

Para o modelo KNN, o valor de K foi otimizado com a técnica de grid-search, encontrando um K ótimo igual a 11. Com essa configuração, o algoritmo KNN obteve uma acurácia de 0,77, precisão de 0,71, Log Loss de 8,31, recall de 0,77, AUC-ROC de 0,54 e F1-Score de 0,72.

A Árvore de Decisão foi ajustada para otimizar a sua estrutura em combinação com técnicas de poda, considerando profundidade máxima de 3, número mínimo de amostras por folha de 1 e número mínimo de amostras para dividir um nó de 2. Com esses parâmetros, o modelo resultante alcançou uma acurácia de 0,82, precisão de 0,77, Log Loss de 6,46, recall de 0,82, AUC-ROC de 0,64 e F1-Score de 0,66.

Para a Floresta Aleatória, também se usou a técnica de grid-search para determinar os melhores hiperparâmetros: profundidade máxima de 10, número mínimo de amostras por folha de 1, número mínimo de amostras para dividir um nó de 10 e número de árvores igual a 50. O modelo resultante obteve uma acurácia de 0,82, precisão de 0,80, Log Loss de 6,50, recall de 0,82, AUC-ROC de 0,65 e F1-Score de 0,80.

As matrizes de confusão indicam que tanto a Floresta Aleatória quanto a Árvore de Decisão têm uma alta taxa de acertos, especialmente para a classe 0 (não-default), mas a classe 1 (default) apresenta uma taxa de erro maior, característica comum em problemas de classificação com classes desbalanceadas. As curvas ROC na Figura 1 mostram que a Floresta Aleatória (AUC-ROC de 0,65) e a Árvore de Decisão (AUC-ROC de 0,64) têm maior capacidade de separabilidade entre as classes positiva e negativa em comparação com os outros modelos.

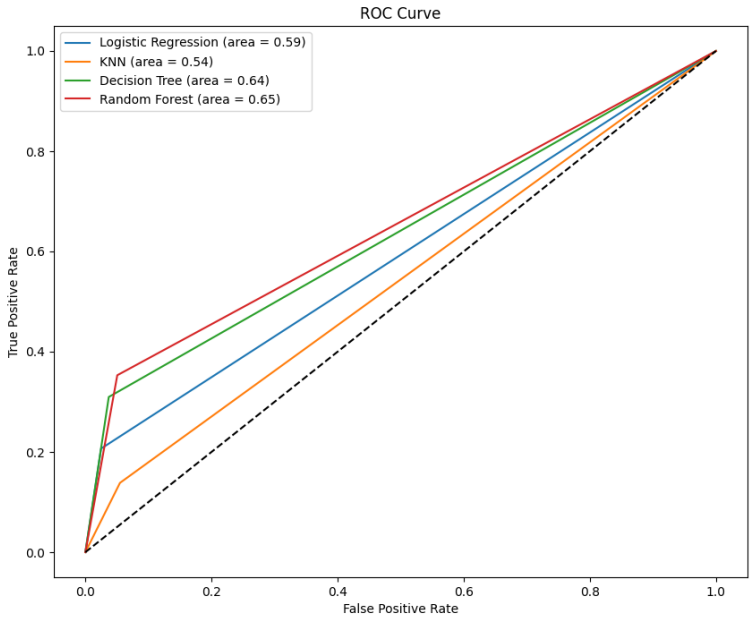

Figura 1- Curva ROC

	Acurácia	Precisão	Log Loss	Recall	AUC-ROC	f1-score
Logit	0,81	0,79	6,93	0,81	0,59	0,76
KNN	0,77	0,71	8,31	0,77	0,54	0,72
Árvore de Decisão	0,82	0,77	6,46	0,82	0,64	0,66
Floresta Aleatória	0,82	0,80	6,50	0,82	0,65	0,80

Tabela 2- Valores Finais

CONCLUSÕES

Neste estudo, avaliamos o desempenho de diferentes modelos de aprendizado de máquina na classificação de default de crédito, utilizando a Regressão Logística como referência e comparando-a com modelos de Floresta Aleatória, Árvore de Decisão e KNN. Analisamos as métricas de acurácia, precisão, Log Loss, recall, AUC-ROC e F1-Score para uma visão completa do desempenho de cada modelo.

Os resultados mostraram que tanto a Floresta Aleatória quanto a Árvore de Decisão superaram a Regressão Logística e o KNN em várias métricas. A Floresta Aleatória destacou-se com a melhor combinação de precisão (0,80), F1-Score (0,80) e AUC-ROC (0,65), evidenciando sua robustez na classificação e separação entre classes de default e não-default. A Árvore de Decisão, quando ajustada com parâmetros otimizados, também apresentou bom desempenho, especialmente na métrica Log Loss (6,46), mostrando sua confiabilidade na previsão de probabilidades.

Embora a Regressão Logística tenha apresentado alta acurácia (0,81) e precisão (0,79), seu desempenho em AUC-ROC (0,59) foi inferior ao dos modelos baseados em árvores, indicando uma menor capacidade de distinguir entre classes desbalanceadas. O KNN, com o valor otimizado de K igual a 11, teve um desempenho inferior em precisão e AUC-ROC, refletindo uma maior tendência a falsos positivos e uma menor capacidade de separação de classes.
Concluímos que os modelos baseados em árvores, especialmente a Floresta Aleatória, são mais eficazes para a pontuação de crédito. Estes modelos oferecem melhor acurácia, precisão e capacidade de generalização, tornando-os mais robustos na previsão de probabilidades. Recomendamos, portanto, o uso de técnicas de ensemble, como a Floresta Aleatória, e a exploração de ajustes finos de hiperparâmetros para otimizar a eficiência e a precisão das previsões em futuras pesquisas e aplicações práticas na área de crédito.

BIBLIOGRAFIA

BREIMAN, Leo. "Statistical modeling: The two cultures." Statistical Science, v. 16, n. 3, p. 199-231, 2001.

BREIMAN, Leo; FREIDMAN, Jerome; OLSHEN, Richard A.; STONE, Charles J. Classification and regression trees. Belmont: Wadsworth Publishing Company, 1986.

DASTILE, X. et. al. Statistical and machine learning models in credit scoring: A systematic literature survey, 2020.

GUJARATI, P. Econometria Básica, 2011.

Guyon, I. Elisseeff, A. An introduction to variable and feature selection, 2003.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009.

I-Cheng, Y. Default of Credit Card Clients, 2016.

MERIGÓ J., CASANOVAS, M. Decision Making with Distance Measures and Linguistic Aggregation Operators, 2010.

QUINLAN, J. R. C4*.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann, 1993.*

Tanveer, A., Huaxin, C. A review on machine learning forecasting growth trends and their real-time applications in different energy system, 2020.

TUKSON, R. et al. A machine learning approach for predicting bank credit worthiness, 2016.

ZHANG, G.; HU, M. Y.; PATTNAIK, D. P. Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis. *European Journal of Operational Research*, v. 116, n. 1, p. 16-32, 1999.