Análise de Risco de Crédito com Aprendizado de Máquina

Apresentação sobre a análise de risco de crédito utilizando aprendizado de máquina, realizada por uma equipe de quatro participantes.

Introdução ao Trabalho

1

Definição do Problema

No setor financeiro, a análise de risco de crédito é essencial para as instituições, pois ajuda a prever o comportamento dos clientes em relação a empréstimos, evitando prejuízos e aumentando a segurança financeira, o que por sua vez pode ajudar com a redução de juros, visto que não será necessário "compensar" a perda.

2

Objetivo do Projeto

Neste trabalho, usamos algoritmos de aprendizado de máquina, como a Árvore de Decisão e o Random Forest, para analisar uma base de dados de crédito. O objetivo é identificar padrões e avaliar os fatores determinantes do risco de crédito, especialmente em uma base de dados desbalanceada.

Descrição dos Dados e Pré-processamento

Base de Dados

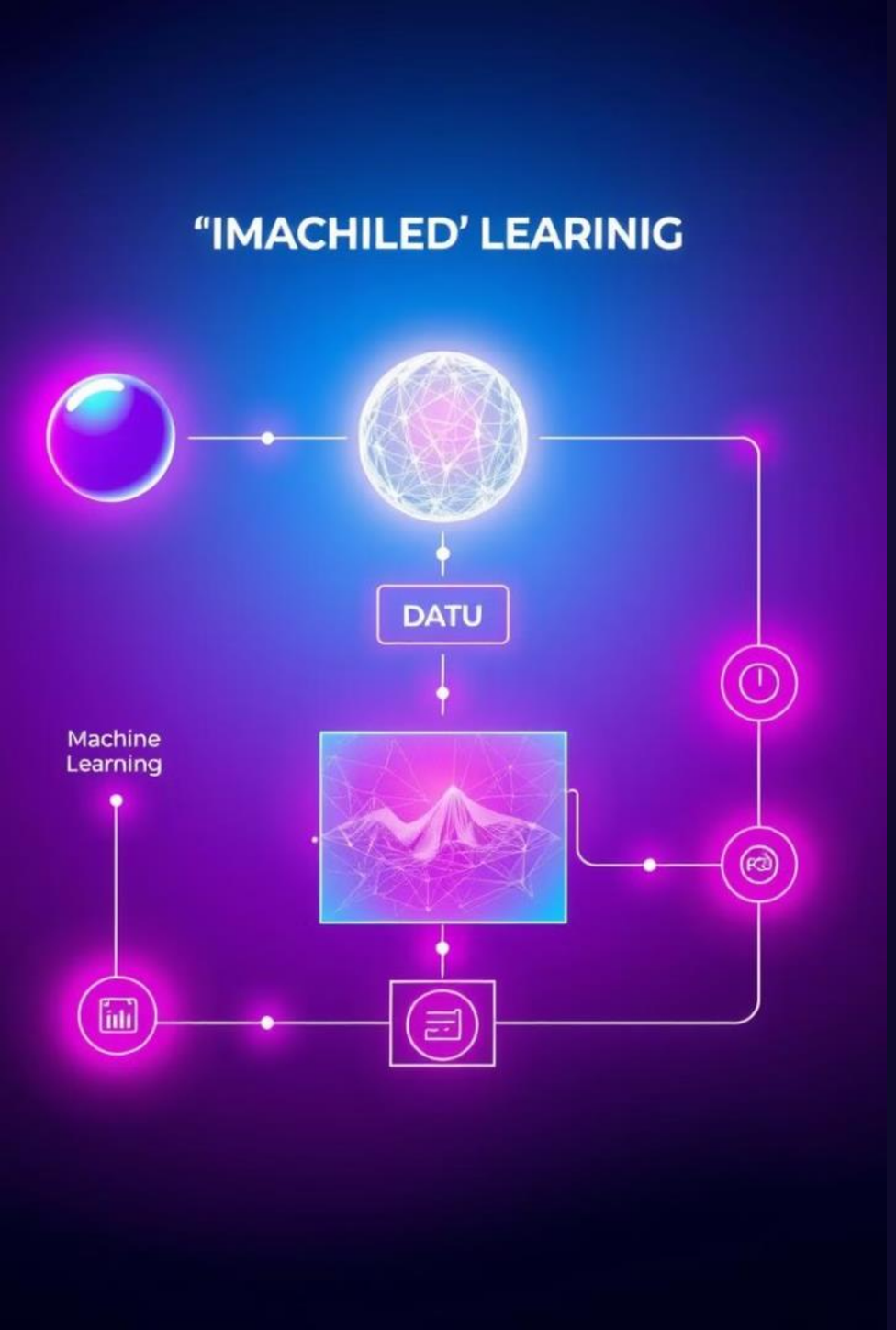
Usamos uma base de dados com variáveis demográficas (idade, sexo), econômicas e financeiras dos clientes, incluindo o valor e a duração do crédito, tipo de moradia e status das contas de poupança e corrente. Esses dados foram usados para classificar os clientes em duas categorias: Risco Alto e Risco Baixo, com o objetivo de identificar padrões de crédito que indicam a probabilidade de inadimplência.

Desbalanceamento de Classes

Um dos desafios foi o desbalanceamento dos dados. Cerca de 70% dos clientes estavam na categoria de Risco Baixo, o que influencia a performance dos modelos, tornando mais difícil a identificação dos clientes de Risco Alto.

Pré-processamento

No pré-processamento, removemos colunas desnecessárias, tratamos valores ausentes e transformamos variáveis categóricas em numéricas. Em seguida, dividimos os dados entre variáveis preditoras e a variável alvo, ‘Risco’.



Modelagem e Algoritmos Utilizados

Modelos Escolhidos

Para a modelagem, utilizamos dois algoritmos supervisionados: Árvore de Decisão e Random Forest. A Árvore de Decisão é fácil de interpretar e fornece um bom ponto de partida. Já o Random Forest oferece maior precisão, especialmente em conjuntos de dados complexos e desbalanceados, como o nosso.

Treinamento e Divisão de Dados

Dividimos os dados em conjuntos de treino e teste, para validar a performance dos modelos de forma realista. Após o treinamento, fizemos previsões no conjunto de teste para avaliar o desempenho de cada modelo.

Resultados e Avaliação dos Modelos

1 Principais Métricas

Para avaliar o desempenho, utilizamos métricas como acurácia, matriz de confusão e AUC (Área Sob a Curva ROC).

No modelo de Random Forest, conseguimos uma acurácia de 77% e uma AUC de 0,78, o que indica um bom poder de separação entre clientes de Risco Alto e Risco Baixo.

2 Matriz de Confusão

A matriz de confusão revelou que o modelo tem alta precisão para a classe de Risco Baixo, mas encontra dificuldades na classe de Risco Alto. Este é um efeito do desbalanceamento, que faz o modelo favorecer a classe majoritária.

3 Curva ROC e Importância das Features

Na Curva ROC, vemos que o modelo Random Forest supera a linha de classificação aleatória, demonstrando seu valor preditivo. A análise de importância das variáveis mostrou que as mais influentes são o valor do crédito, a idade do cliente e a duração do crédito.

Conclusão e Recomendações

1

Conclusão do Projeto

Este estudo destacou a importância de variáveis financeiras, como o valor e a duração do crédito, na classificação de risco de clientes. No entanto, o desbalanceamento do conjunto de dados afeta a precisão, especialmente para a classe de Risco Alto.

2

Recomendações para Melhorias

Para aprimorar o modelo, recomendamos técnicas de balanceamento de dados, como oversampling ou ajuste de pesos, para melhorar a identificação da classe de Risco Alto. Além disso, a exploração de algoritmos avançados, como o XGBoost, pode capturar padrões mais complexos.

3

Fechamento

A análise de risco de crédito é um desafio complexo, mas com a aplicação de técnicas de aprendizado de máquina, conseguimos avançar na direção de um modelo mais seguro e preciso. Entendemos que a importância de um modelo desse é conseguir que os bancos sejam mais rápidos na analise de crédito, sem a necessidade de um analista específico para isso. Acreditamos que a função desse colaborador será apenas de gerenciar as analises e com base nela tomar as decisões, sem precisar fazer isso desde o começo.

Referências Bibliográficas

Introdução ao Aprendizado de Máquina:

- Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006.
- Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.

Algoritmos e Modelos:

- Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.
- James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning: With Applications in R. Springer, 2013.

Banco de dados extraído de:

https://github.com/joaosoliveira0907/ML_7_aplicando_conhecimento