Faculdade de Computação e Informática

Modelo Preditivo de Inadimplência em Empréstimos P2P utilizando Algoritmos de Classificação Supervisionada

Theodoro Priosti de Almeida[1]

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

theodoro.almeida@mackenzie.br

Resumo. *Este projeto visa desenvolver um modelo de análise preditiva para avaliação de risco de crédito utilizando o dataset Lending Club, uma das maiores plataformas de empréstimos peer-to-peer dos Estados Unidos.*

Através de machine learning, pretende-se criar um sistema capaz de classificar solicitações de empréstimo quanto ao risco de inadimplência, contribuindo para decisões mais assertivas no mercado financeiro.

1. Introdução

O mercado de crédito desempenha um papel fundamental na economia global, facilitando o acesso ao capital para indivíduos e empresas. Com o advento das fintechs e plataformas de empréstimos, surgiram novas oportunidades e desafios na avaliação de risco de crédito. A Lending Club, fundada em 2006, tornou-se pioneira nesse segmento, conectando investidores a tomadores de empréstimo.

A avaliação tradicional de risco de crédito baseava-se principalmente em scores como o FICO e informações básicas dos solicitantes. No entanto, a disponibilidade de grandes volumes de dados e o desenvolvimento de algoritmos proporcionaram novas possibilidades para análises mais sofisticadas e precisas. A necessidade de modelos preditivos mais eficazes para avaliação de risco de crédito é evidente considerando:

- Crescimento do mercado: O setor de empréstimos digitais movimenta bilhões de dólares anualmente, demandando ferramentas de análise cada vez mais precisas.
- Redução de perdas: Modelos preditivos podem significativamente reduzir as taxas de inadimplência, protegendo investidores e instituições financeiras.
- Democratização do crédito: Algoritmos sofisticados podem identificar bons pagadores que seriam rejeitados por métodos tradicionais.

O objetivo final é desenvolver um modelo de machine learning para predição de risco de inadimplência em empréstimos, utilizando dados históricos da Lending Club, visando melhorar a precisão da análise de crédito e reduzir perdas financeiras.

O projeto se enquadra na opção “Framework”, onde será usado a linguagem Python com bibliotecas de Machine Learning para chegar no objetivo final.

2. Descrição do problema

O problema central abordado neste projeto é a predição de inadimplência em empréstimos pessoais, um desafio crítico no setor financeiro que envolve múltiplas dimensões. Naturalmente, a maioria dos empréstimos são pagos, criando um desafio de classificação com classes desbalanceadas

- Existem centenas de constantes que podem influenciar o risco de crédito
- Aprovar um empréstimo que entrará em default resulta em perda direta
- Rejeitar um bom pagador representa perda de oportunidade de lucro

3. Aspectos Éticos do Uso da IA

O desenvolvimento de sistemas de IA para análise de crédito levanta importantes questões éticas e de responsabilidade que devem ser cuidadosamente consideradas.

Será necessário garantir que o algoritmo não discrimine contra nenhum grupo e será necessário atualização e monitoramento constante para garantir futura precisão.

Dados pessoais dos usuários dever ser corretamente anonimizados, e os usuários cientes do uso.

4. Dataset

O dataset utilizado provém da Lending Club, uma das maiores plataformas de empréstimos dos Estados Unidos. Os dados foram disponibilizados publicamente pela empresa e cobrem o período de 2007 a 2018, representando um dos mais abrangentes conjuntos de dados sobre empréstimos disponíveis.

O dataset está disponível em Kaggle [wordsforthewise/lending-club], e foi pré processado pelo autor. O dataset está licenciado sob a licensa CC0: Domínio Público.

A análise exploratória, disponível no Github, revelou que a maioria dos empréstimos são pagos integralmente, com uma distribuição específica entre diferentes status de inadimplência, confirmando o desafio de classes desbalanceadas. O volume de empréstimos aumentou significativamente nos anos analisados, sendo a finalidade principal a consolidação de dívidas.

5. Metodologia

Preparação dos Dados:

O dataset foi processado resultando em 21.734 empréstimos (1% do dataset) com 80 features. Aplicou-se feature engineering, SMOTE para balanceamento, e divisão 70/15/15 (treino/validação/teste).

Modelos Implementados:

Foram implementados 4 algoritmos otimizados via GridSearchCV: Regressão Logística (AUC-ROC: 0.7050), Random Forest (0.7103), XGBoost (0.6924), e Rede Neural MLP (0.7147 - melhor modelo).

Interpretabilidade:

Utilizou-se SHAP values para identificar features mais importantes: issue_year (0.1085), grade_D (0.0287), grade_C (0.0252), verification_status (0.0231).

6. Resultados

A Rede Neural MLP apresentou melhor performance dentre os testados

• AUC-ROC (Teste): 0.7361

• Precisão: 0.2276 | Recall: 0.7223 | F1-Score: 0.3461

• Matriz de Confusão: TN=1.656, FP=1.160, FN=123, TP=321

Impacto de Negócio:

O modelo permite economia estimada de $3.945.000 (59,3% de redução vs baseline), possibilitando automação de 60-70% das decisões de baixo risco com foco de analistas em casos críticos.

7. Conclusão

O projeto alcançou com sucesso o objetivo de desenvolver um modelo preditivo de inadimplência. A Rede Neural MLP atingiu AUC-ROC de 0.7361 (próximo da meta de 0.75) e recall de 72,23% (superando meta de 65%).

O impacto de negócio superou expectativas: redução de 59,3% nas perdas vs 20% esperado. A análise SHAP revelou que ano de emissão, grades de risco e status de verificação são os principais preditores.

Limitações:

Os dados são pré-pandemia e podem não representar critérios atuais.

8. Entereços Web

Github: github.com/ovosimpatico/predictive-credit
Dataset: kaggle.com/datasets/wordsforthewise/lending-club
YouTube**:** https://youtu.be/rE8uOA6yZIQ

9. Bibliografia

Khandani, Amir & Kim, Adlar & Lo, Andrew. (2010). Consumer Credit-Risk Models Via Machine-Learning Algorithms. Journal of Banking & Finance. 34. 2767–2787. 10.1016/j.jbankfin.2010.06.001.

Lessmann, Stefan & Baesens, Bart & Seow, Hsin-Vonn & Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. (doi:10.1016/j.ejor.2015.05.030). 10.1016/j.ejor.2015.05.030.