Análise Preditiva de Inadimplência com Dados Financeiros: Um Estudo com o Conjunto HMEQ

André Matteucci[1], Enzo Casagrande[1], Enzo Koji[1], Felipe Ribeiro da Silva[1]

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

{10403403,10400726,10403411,10400831}@mackenzista.com.br

Abstract. *This article describes the development of a predictive model based on a dataset containing data from Home Equity Line of Credit (HELOC) applications, a credit modality that allows property owners to use the value of their real estate as collateral to obtain loans at reduced interest rates. The study addresses banking automation in decision-making from a macro perspective, employing exploratory data analysis techniques, data preparation, and machine learning applications.*

Resumo. *Este artigo descreve a construção de um modelo preditivo com base em um dataset contendo dados de solicitações de Linha de Crédito com Garantia de Imóvel (HELOC), modalidade que permite ao proprietário utilizar o valor de seu imóvel como garantia para obtenção de empréstimos com taxas de juros reduzidas. O estudo aborda a automação bancária na tomada de decisão a partir de uma perspectiva macro, utilizando técnicas de análise exploratória, preparação e uso de machine learning.*

1. Introdução

1.1. Contextualização

A Linha de Crédito com Garantia de Imóvel (HELOC - Home Equity Line of Credit) representa uma importante modalidade de crédito no mercado financeiro, permitindo que proprietários de imóveis utilizem o patrimônio acumulado em suas propriedades como garantia para obtenção de empréstimos com condições mais favoráveis, muito usada em outros países, como nos Estados Unidos e no Brasil é conhecido por hipoteca. Esta modalidade se destaca por oferecer taxas de juros significativamente mais baixas quando comparadas a empréstimos sem garantia.

1.2. Justificativa

Com o avanço da tecnologia e a crescente digitalização do setor bancário, as instituições financeiras têm implementado sistemas de análise de crédito para otimizar o processo decisório na concessão destes empréstimos. Estes sistemas utilizam técnicas avançadas de análise de dados e aprendizado de máquina para avaliar o risco de inadimplência dos solicitantes, permitindo decisões mais rápidas, consistentes e baseadas em evidências, garantindo uma melhor tomada de decisão.

1.3. Objetivo

O objetivo deste projeto é desenvolver modelos preditivos baseados em técnicas de aprendizado de máquina para análise de risco de crédito em Linha de Crédito com Garantia de Imóvel, visando aumentar a acurácia na concessão de crédito e reduzir a inadimplência fazendo uso do dataset público HMEQ (Home Equity Dataset).

1.4. Opção do projeto

Para este projeto, optou-se pela **Opção Framework**, com o uso de Python e bibliotecas relevantes no contexto de machine learning a fim de criar um modelo preditivo de regressão logística.

2. Descrição do Problema

Como qualquer processo de análise de concessão de crédito, a Linha de Crédito com Garantia de Imóvel envolve diversas decisões críticas que impactam tanto a sustentabilidade e viabilidade do ponto de vista da instituição financeira quanto o acesso por parte dos clientes.

Com o avanço da Ciência de Dados, ainda que haja regulações que limitam o escopo de aplicabilidade na área, é possível empregá-la de modo a automatizar parte destes processos. Nesse cenário, o problema abordado por este projeto consiste em aplicar técnicas de machine learning para construir um modelo preditivo capaz de estimar a probabilidade de inadimplência de um cliente com base em dados históricos de solicitações de crédito, dados demográficos e de relacionamento do cliente com a instituição.

Sendo assim, o objetivo é fornecer uma ferramenta de apoio para a tomada de decisão bancária na concessão de crédito, objetivando minimizar o risco de inadimplência.

3. Ética e Responsabilidade de IA

Em primeira instância, é importante relembrar que, embora a Inteligência Artificial proporcione auxílios notáveis na análise de risco de crédito e inadimplência, é de extrema importância satisfazer as questões éticas e regulatórias presentes no setor bancário perante o desenvolvimento de tecnologias que manipulam dados pessoais.

Nessa perspectiva, a LGPD estabelece normas rigorosas para a coleta, tratamento e armazenamento de dados sensíveis. Dessa forma, a base de dados utilizada para a aprendizagem de nosso Machine Learning não possui informações pessoais dos clientes (nome, CPF, endereço etc.) assegurando que eles não terão seus dados expostos, promovendo a privacidade, segurança e ética.

Além disso, adotar práticas responsáveis é crucial para estar em conformidade com os padrões regulatórios, garantindo que o uso da Inteligência Artificial respeite a privacidade e os direitos dos clientes. Embora os modelos de IA utilizados no processo de análise de inadimplência operem de forma autônoma, todos os cuidados serão tomados para proteger os dados pessoais, assegurando que nenhuma informação sensível seja exposta. Com isso, buscamos inovar na análise de risco de crédito enquanto mantemos o compromisso com a privacidade, segurança e ética.

4. Dataset

O dataset utilizado é o HMEQ_Data, obtido pelo Kaggle com licença CC0 1.0 Universal (Domínio Público), contendo 5960 linhas com os seguintes dados:

Coluna	Descrição
BAD	1 = cliente inadimplente; 0 = empréstimo quitado
LOAN	Valor do empréstimo solicitado
MORTDUE	Valor devido da hipoteca existente
VALUE	Valor estimado da propriedade atual
REASON	DebtCon = consolidação de dívidas; HomeImp = reforma residencial
JOB	Seis categorias ocupacionais
YOJ	Anos no emprego atual
DEROG	Número de registros negativos relevantes

DELINQ	Número de linhas de crédito em atraso
CLAGE	Idade da linha de crédito mais antiga (em meses)
NINQ	Número de consultas recentes ao crédito
CLNO	Número total de linhas de crédito
DEBTINC	Relação dívida/renda

A variável-alvo é a BAD, que indica se o solicitante é ou não inadimplente. A análise exploratória com mais detalhes pode ser consultada no arquivo .ipynb que faz parte do projeto.

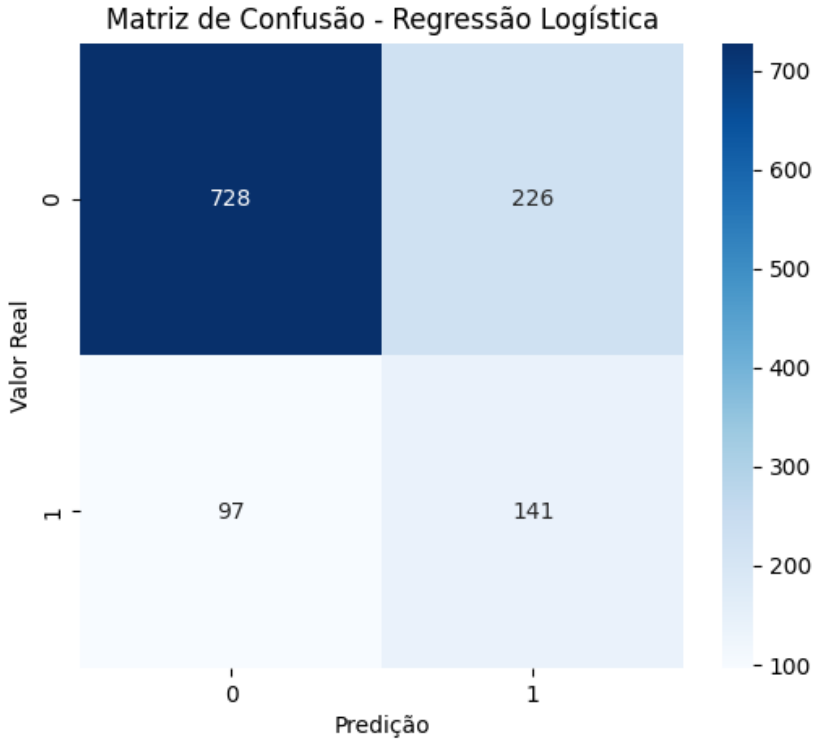
5. Metodologia

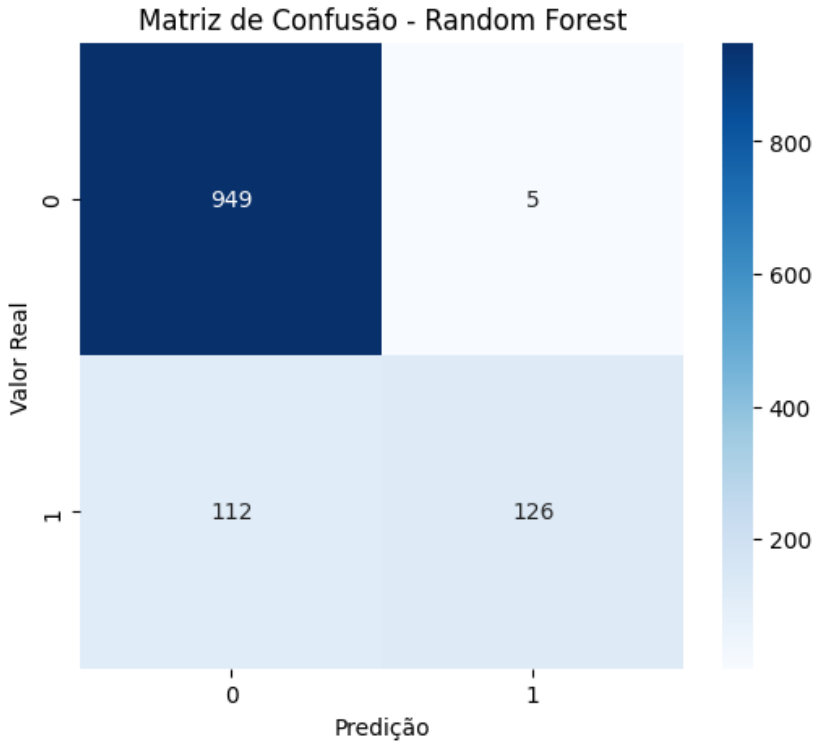
O presente projeto tem como objetivo a construção de um modelo preditivo capaz de identificar clientes com uma maior probabilidade de inadimplência. A metodologia é composta pelas seguintes etapas:

1) **Análise Exploratória de Dados**, na qual é realizada uma análise detalhada no dataset, buscando compreender a estrutura dos dados e identificar valores vazios e duplicados, outliers e verificar o balanceamento da variável-alvo.
2) **Tratamento dos dados**, na qual, foram aplicadas técnicas relevantes para ajuste do dataset de modo a eliminar os problemas.
3) **Construção do modelo preditivo**, na qual foram aplicados algoritmos de aprendizado de máquina com o objetivo de classificar corretamente os clientes quanto ao risco de inadimplência. Essa etapa envolve a divisão dos dados em conjuntos de treino e teste, além do ajuste de hiperparâmetros para otimização do desempenho.
4) **Validação**, na qual o modelo foi avaliado por meio de métricas de Acurácia, Precisão, Recall e F1-Score, com o intuito de garantir a robustez e a capacidade preditiva do modelo em diferentes cenários.

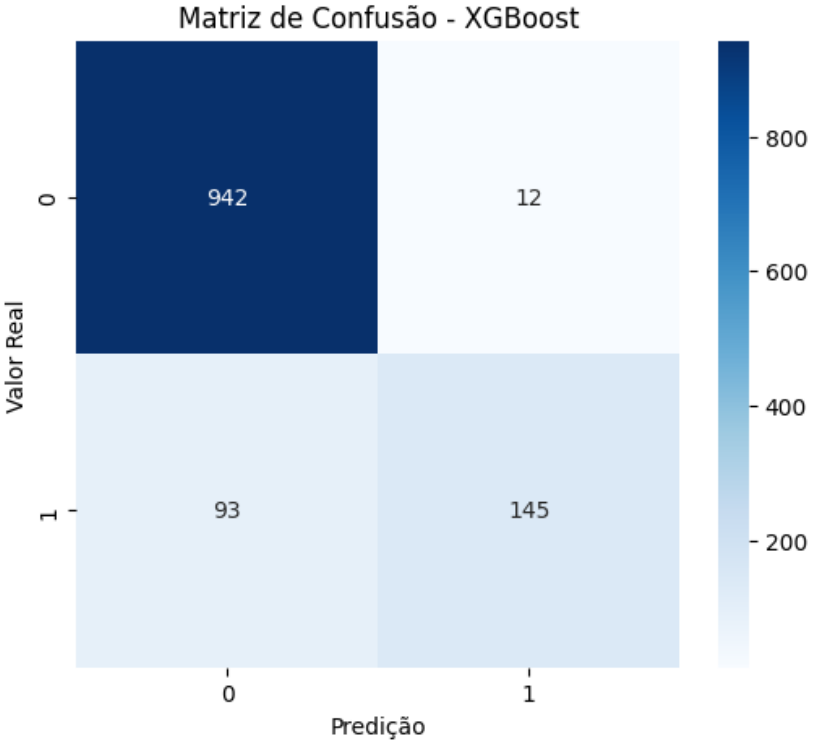
6. Resultados

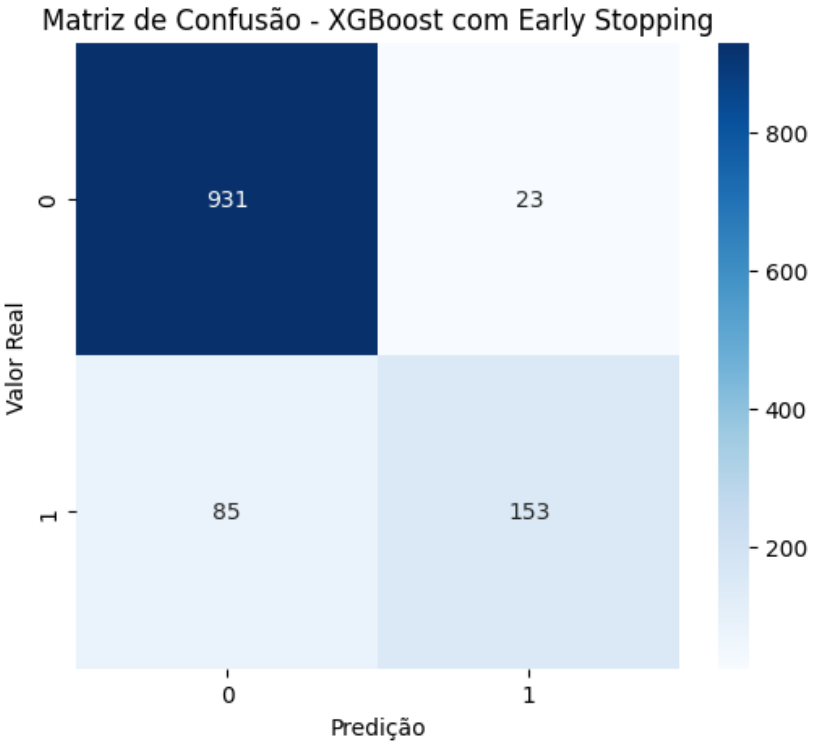
Inicialmente, foi feita a opção do treinamento de modelos de Regressão Logística e Random Forest. O modelo de Regressão Logística apresentou resultados mistos na classe 0 e insatisfatórios na classe 1. No Random Forest, os resultados na classe 0 melhoraram significativamente. Ainda que tenha havido melhora na classe 1, os resultados ainda não foram satisfatórios. Essas conclusões podem ser observadas por meio das matrizes de confusão:





Deste modo, foi explorado também o algoritmo XGBoost, com e sem Early Stopping. Os resultados foram similares em ambas as estratégias, mas a versão com Early Stopping resultou em um desempenho melhor na classe 1:





A tabela a seguir resume as principais métricas obtidas para cada um dos algoritmos desenvolvidos.

Métrica \ Algoritmo	Regressão Logística	Random Forest	XGBoost	XGBoost com Early Stopping
Acurácia	0.7290	0.9018	0.9119	0.9094
Precisão	0.3842	0.9618	0.9236	0.8693
Recall	0.5924	0.5294	0.6092	0.6429
F1-score	0.4661	0.6829	0.7342	0.7391

7. Conclusões

O estudo teve como principal objetivo desenvolver um modelo preditivo capaz de estimar a probabilidade de inadimplência em solicitações de linha de crédito com garantia de imóvel, utilizando técnicas de ciência de dados e machine learning. A partir de da análise exploratória, preparação do conjunto de dados e aplicação de algoritmos de classificação, foi possível construir modelos que apoiam a tomada de decisão no setor bancário, promovendo uma redução de riscos significativa.

Os resultados encontrados mostram que, embora a regressão logística tenha apresentado desempenho limitado, principalmente na identificação de inadimplentes, modelos mais avançados como Random Forest e XGBoost demonstraram uma maior capacidade preditiva. O XGBoost com Early Stopping foi a abordagem com mais equilíbrio entre precisão e sensibilidade, alcançando o melhor F1-score dentre os modelos treinados.

Portanto, em última instância, conclui-se que o uso de técnicas de Machine Learning aplicadas a dados históricos de crédito podem ser uma ferramenta eficaz no apoio às decisões bancárias, desde que acompanhada de práticas responsáveis e éticas no tratamento de dados.

8. Vídeo da Apresentação

A apresentação está disponível no YouTube, por meio do seguinte link: https://youtu.be/zzY4w5ZFhKE

9. Referências

ITAÚ**.** Empréstimo com garantia de imóvel: como funciona? *Blog Itaú*, 16 ago. 2022. Disponível em: https://blog.itau.com.br/credito-garantia-imovel/emprestimo-com-garantia-de-imovel. Acesso em: 5 abr. 2025.

JUNIOR, R. Os Contratos Bancários e a Jurisprudência do Superior Tribunal de Justiça. Informativo Jurídico da Biblioteca Minstro Oscar Saraiva, Camaquã, 2003. Disponível em: https://www.stj.jus.br/publicacaoinstitucional/index.php/informativo/article/view/428/386. Acesso em: 7 abr. 2025.

LOPES, R.; COLOMBI, L.; MUTZ, F. Comparação de algoritmos de aprendizado de máquina para predição de pontuação de crédito. Anais do Computer on the Beach, v. 14, p. 424–431, 05 2023. Disponível em: <https://periodicos.univali.br/index.php/acotb/article/view/19479/11288>. Acesso em: 27 maio. 2025.

SUHADOLNIK, N.; UEYAMA, J.; DA SILVA, S. Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. *Journal of Risk and Financial Management*, Basel, 2023. Disponível em: https://doi.org/10.3390/jrfm16120496. Acesso em: 4 abr. 2025.