PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

Mharcos Vinicius Gonçalves de Hungria

USO DE INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DE CRÉDITO

São Paulo
Agosto de 2023

Mharcos Vinicius Gonçalves de Hungria

USO DE INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

São Paulo

Agosto de 2023

SUMÁRIO

1. Introdução

A Inteligência Artificial (IA) e o Aprendizado de Máquina (AM) têm revolucionado a maneira como abordamos uma variedade de problemas complexos em diferentes setores, e o campo da análise de crédito não é exceção. Nos últimos anos, a interseção entre IA, AM e análise de crédito tem ganhado destaque como uma abordagem inovadora e eficaz para avaliar o risco financeiro de indivíduos e empresas. A capacidade da IA em lidar com grandes volumes de dados de maneira rápida e precisa, juntamente com os avanços contínuos em algoritmos de aprendizado de máquina, tem proporcionado um novo conjunto de ferramentas e técnicas que têm o potencial de aprimorar significativamente a precisão das decisões de crédito.

Nesta era digital, onde dados são gerados em uma escala exponencial, as instituições financeiras enfrentam o desafio de processar informações detalhadas sobre os históricos financeiros e de pagamento de seus clientes em tempo hábil. Aqui é onde a Inteligência Artificial entra em jogo, permitindo a análise automatizada desses dados para identificar padrões, tendências e correlações que poderiam passar despercebidos em abordagens tradicionais. Além disso, a IA também tem a capacidade de incorporar uma ampla gama de variáveis, incluindo fontes não convencionais de dados, como redes sociais e comportamentos online, enriquecendo assim a avaliação de crédito.

Ao explorar o uso da IA na análise de crédito, surgem questões cruciais sobre a interpretação das decisões tomadas pelos algoritmos. A transparência e a justiça desses modelos são tópicos de discussão importantes, uma vez que decisões baseadas apenas em dados históricos podem perpetuar vieses existentes. Portanto, a busca por modelos de IA e AM que sejam éticos, imparciais e compreensíveis continua sendo um foco essencial nesse campo.

Esta introdução tem como objetivo destacar a importância crescente da Inteligência Artificial e do Aprendizado de Máquina na análise de crédito. Nos próximos segmentos, serão explorados os métodos específicos, desafios e impactos desse uso inovador de tecnologias emergentes no cenário da análise de crédito moderna.

2. Descrição do Problema e da Solução Proposta

O problema central abordado neste trabalho é aprimorar a precisão e eficiência na avaliação de crédito, enfrentado por instituições financeiras ao analisar a viabilidade de concessão de empréstimos e linhas de crédito. A justificativa para essa pesquisa reside na necessidade de superar as limitações das abordagens tradicionais, que muitas vezes se baseiam em análises subjetivas ou em um número limitado de variáveis, resultando em decisões de crédito potencialmente imprecisas e arriscadas.

A motivação por trás deste estudo é aproveitar as capacidades da Inteligência Artificial e do Aprendizado de Máquina para criar um modelo preditivo robusto e confiável, capaz de avaliar o risco de crédito de maneira objetiva e precisa. O objetivo principal é desenvolver um sistema automatizado que possa analisar uma ampla gama de variáveis financeiras, históricos de pagamento, comportamento de gastos e até mesmo dados não financeiros relevantes, como interações online e perfis de redes sociais, a fim de tomar decisões informadas sobre a concessão de crédito.

Para alcançar esse objetivo, a solução proposta envolve a implementação de algoritmos de Aprendizado de Máquina, especificamente os modelos de classificação. Algoritmos como Regressão Logística, Árvores de Decisão, Máquinas de Vetores de Suporte (SVM) e Redes Neurais Artificiais podem ser explorados para criar um modelo preditivo que atribui uma probabilidade de risco a cada candidato a crédito. Essa probabilidade é calculada com base nas características individuais do solicitante, o histórico de crédito, os dados financeiros disponíveis e outros fatores relevantes.

O Aprendizado de Máquina desempenha um papel fundamental na construção e treinamento desses modelos. A partir de um conjunto de dados de treinamento que inclui exemplos de concessões de crédito bem-sucedidas e malsucedidas, os algoritmos aprendem a identificar padrões e correlações que podem indicar os fatores mais influentes na decisão de crédito. O processo de treinamento envolve ajustar os parâmetros do modelo para otimizar sua capacidade de generalização, ou seja, a capacidade de fazer previsões precisas em novos casos não vistos anteriormente.

3. Coleta de Dados

Os dados foram obtidos a partir do dataset "Credit Risk Dataset", disponível na plataforma Kaggle. Esse conjunto de dados é composto por informações relacionadas à análise de crédito de indivíduos, incluindo características pessoais, financeiras e de histórico de crédito. O dataset está estruturado em formato tabular, com as instâncias organizadas em linhas e as características em colunas. Cada linha corresponde a um indivíduo solicitante de empréstimo, enquanto as colunas representam as diferentes características que descrevem esses solicitantes.

Nome do dataset: Credit Risk Dataset **Descrição:** Este conjunto de dados contém colunas que simulam dados de crédito **Link:** https://www.kaggle.com/datasets/laotse/credit-risk-dataset		
Nome do Atributo	Descrição	Tipo
person_age	Idade do indivíduo solicitante.	int
person_income	Renda anual do solicitante.	int
person_home_ownership	Situação de posse da moradia do solicitante. Valores possíveis: 'MORTGAGE', 'RENT', 'OWN', 'OTHER'.	string
person_emp_length	Tempo de emprego em anos.	float
loan_intent	Propósito da aplicação do empréstimo.	string
loan_grade	Classificação atribuída ao empréstimo.	string
loan_amnt	Valor do empréstimo solicitado.	int
loan_int_rate	Taxa de juros do empréstimo.	float
loan_status	Status do empréstimo, onde 0 representa não inadimplente e 1 representa inadimplente.	int
loan_percent_income	Proporção entre o valor do empréstimo e a renda anual.	float
cb_person_default_on_file	Histórico de inadimplência.	string
cb_person_cred_hist_length	Extensão do histórico de crédito do solicitante.	int

4. Processamento/Tratamento de Dados

Nesta etapa, executei o tratamento dos dados usando os seguintes métodos:

- Limpeza de Dados: Preenchi valores ausentes usando a média para características numéricas e moda para categóricas.

- Codificação de Variáveis Categóricas: Converti características categóricas em números usando one-hot encoding.

- Normalização/Padronização: Padronizei as características para que todas tivessem a mesma escala.

- Engenharia de Recursos: Criei uma característica chamada "loan_income_ratio" para mostrar a relação entre o valor do empréstimo e a renda anual.

- Seleção de Características: Usei a Importância de Características para escolher as mais influentes.

- Tratamento de Desbalanceamento: Usei oversampling da classe inadimplente para equilibrar as classes.

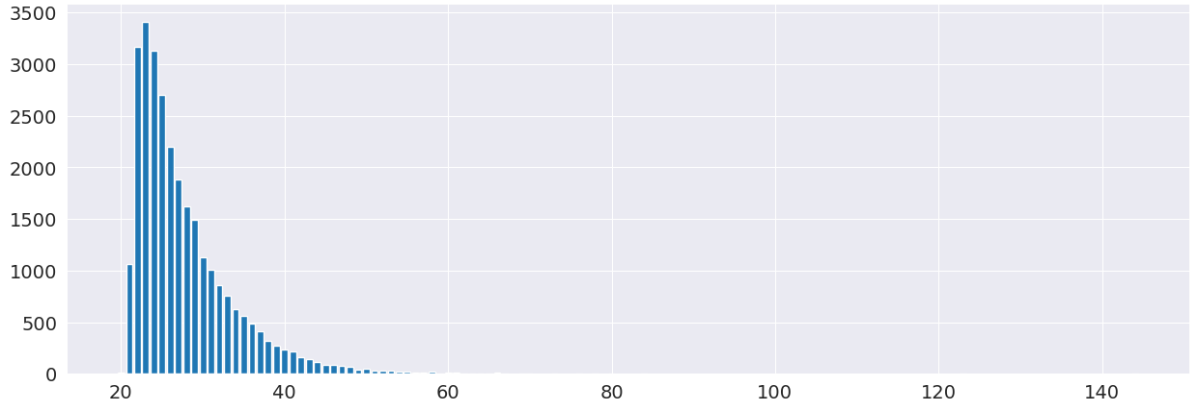
Na etapa de desenvolvimento, optei por utilizar os algoritmos de Regressão Logística, Árvore de Decisão e Random Forest para construir os modelos de predição. A Regressão Logística foi escolhida devido à sua interpretabilidade, enquanto a Árvore de Decisão e o Random Forest podem lidar com interações complexas entre características.

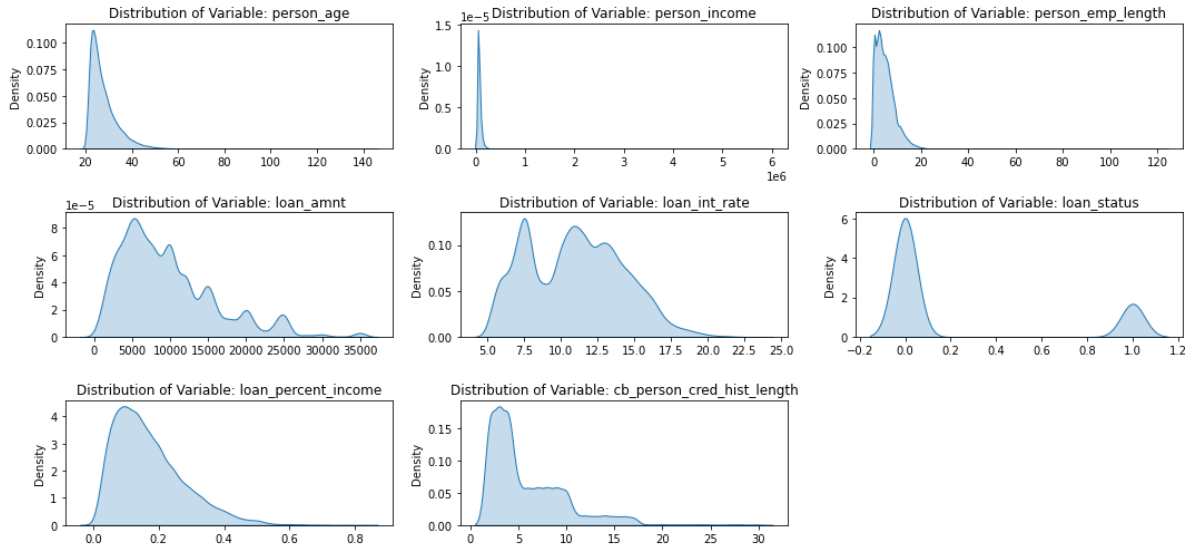
Cada escolha foi fundamentada na minha compreensão dos dados subjacentes e nos objetivos específicos do projeto, garantindo a pertinência das decisões tomadas. Essas etapas estruturais são essenciais para a construção de modelos preditivos que se mostrem eficazes na análise de crédito. Além disso, para assegurar a replicabilidade do processo, todos os códigos relacionados a cada uma dessas etapas foram cuidadosamente documentados.

5. Análise e Exploração dos Dados

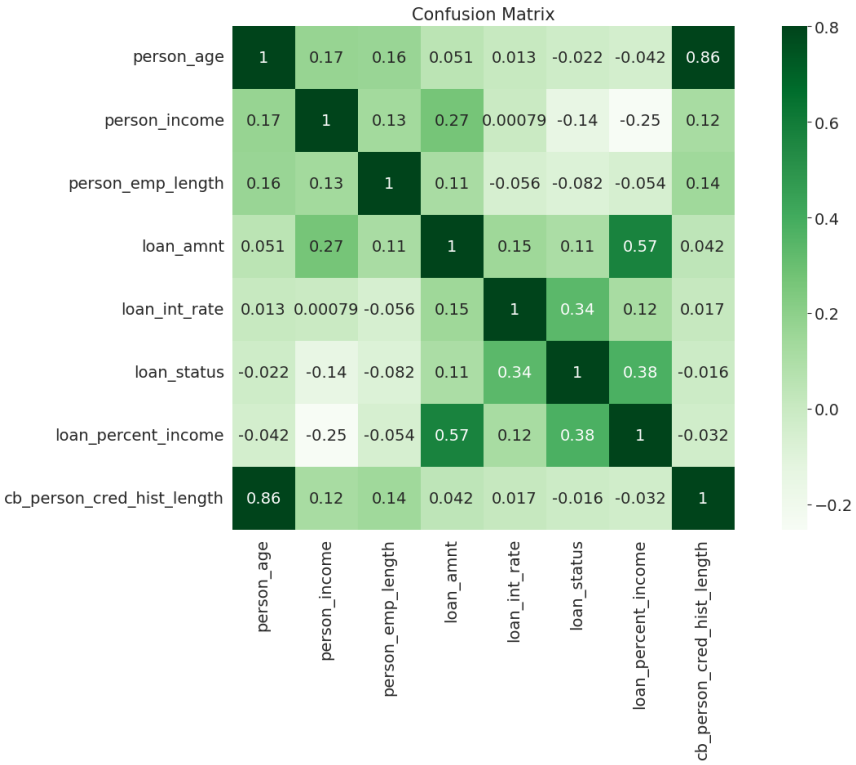
A exploração dos dados é conduzida de maneira analítica, com o objetivo de identificar padrões, levantar hipóteses e ganhar insights que possam fornecer uma compreensão mais profunda do problema em questão. O uso de ferramentas estatísticas é essencial, e a análise pode envolver voltar a passos anteriores para obter mais dados ou refinar o tratamento realizado. Para isso, utilizo as bibliotecas gráficas Matplotlib e Seaborn em Python, permitindo a criação de visualizações informativas.

- Análise Descritiva: Começo com uma análise descritiva das características principais do conjunto de dados. Ploto histogramas e gráficos de barra para visualizar a distribuição das idades, valores de empréstimos, taxas de juros e outras variáveis relevantes. Isso ajuda a identificar tendências iniciais e possíveis outliers.

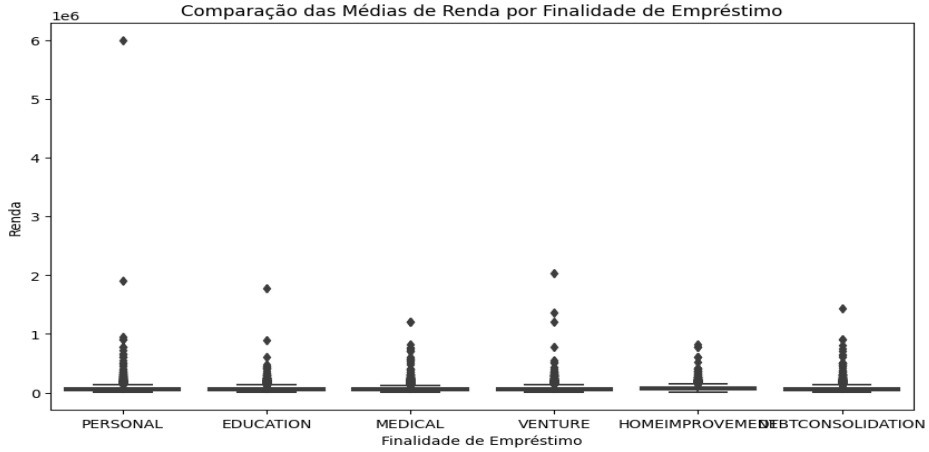




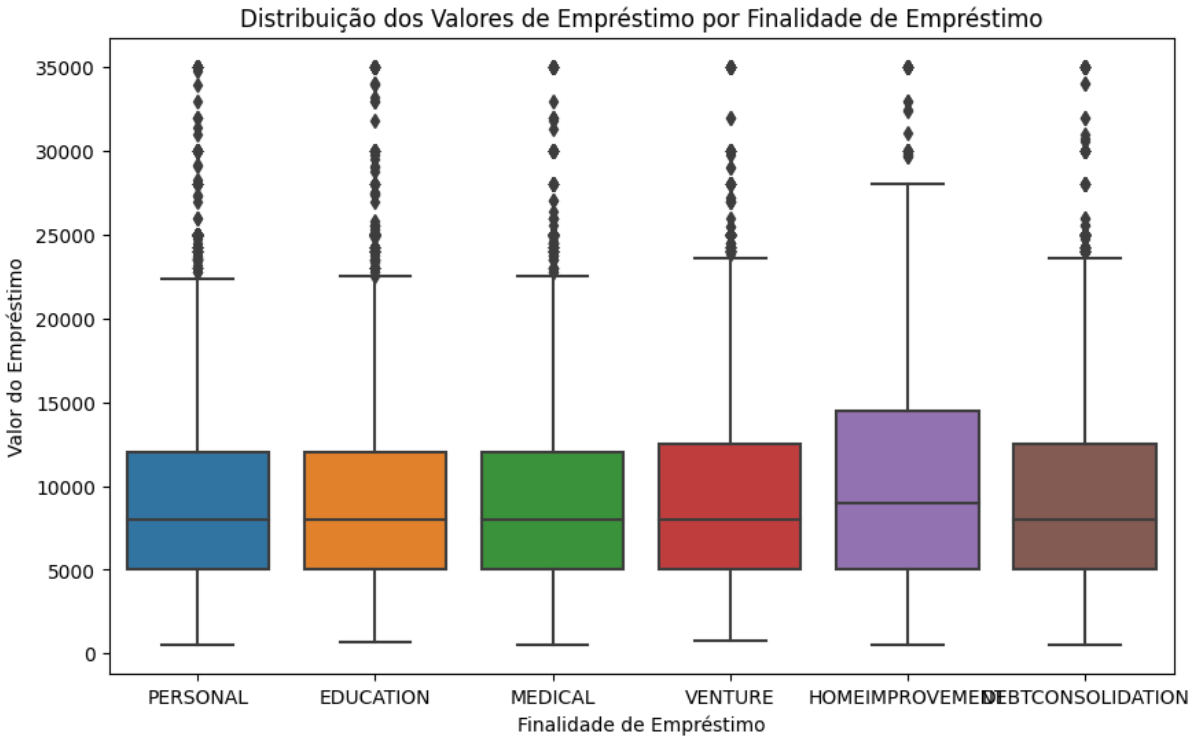
- Exploração de Relações: Investigo relações entre características usando gráficos de dispersão e mapas de calor de correlação. Procuro entender como a idade, renda, quantidade do empréstimo e outras variáveis se relacionam. Por exemplo, traço gráficos de dispersão entre idade e renda para visualizar possíveis correlações.

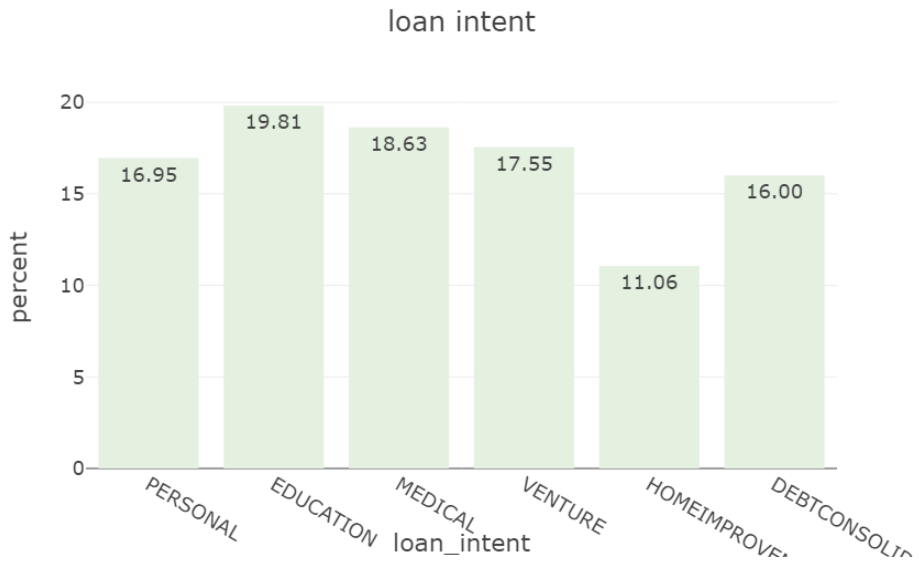


- Testes Estatísticos: Para avaliar hipóteses e padrões, aplico testes de hipóteses e calculo intervalos de confiança. Por exemplo, posso testar se há diferenças significativas na média da renda entre diferentes categorias de finalidade de empréstimo usando testes t. Esses testes me ajudam a validar as hipóteses e obter insights estatisticamente sólidos.



- Visualizações Avançadas: Além de gráficos básicos, crio visualizações mais sofisticadas. Um exemplo é o boxplot para avaliar a distribuição dos valores de empréstimo por finalidade. Também posso gerar mapas de calor para mostrar correlações entre características numéricas.





6. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Nesta etapa, concentro-me na meticulosa preparação dos dados para atender às exigências específicas dos modelos de Aprendizado de Máquina escolhidos. São executadas diversas etapas essenciais, incluindo a criação de atributos relevantes, o balanceamento da base de dados quando necessário e, por fim, a divisão estratégica dos dados em conjuntos de treino, validação e teste.

Com o intuito de otimizar o desempenho dos modelos de Aprendizado de Máquina, desenvolvo atributos adicionais que se originam da análise exploratória realizada anteriormente. Além do "loan_income_ratio" previamente criado, introduzo o atributo "loan_percent_income", que reflete a proporção da renda anual representada pelo valor do empréstimo. Esses atributos adicionais, bem concebidos, têm o potencial de aprimorar a capacidade dos modelos em capturar informações cruciais e nuances nos dados.

Diante da identificação de desequilíbrio entre as classes de status de empréstimo, recorro à técnica de oversampling por meio da biblioteca SMOTE. Esta abordagem viabiliza a geração de novas instâncias sintéticas para a classe minoritária (inadimplentes), alcançando, assim, um equilíbrio proporcional nos dados. Este passo é de suma importância, uma vez que assegura que os modelos não sejam tendenciosos em relação às classes de interesse.

Dividir a base de dados em conjuntos distintos — treino, validação e teste — é uma etapa vital para a avaliação precisa dos modelos. Adoto uma abordagem estratificada para preservar a representatividade das classes em cada conjunto. A alocação percentual é realizada de maneira criteriosa: 70% dos dados são alocados para treino, enquanto 15% são destinados tanto à validação quanto ao teste. O conjunto de validação desempenha um papel crucial no ajuste dos hiperparâmetros dos modelos, contribuindo assim para o refinamento das previsões.

Exemplo de Código (Criação de Atributos e Divisão da Base):

```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE

# Criação de atributos
df['loan_percent_income'] = df['loan_amnt'] / df['person_income']

# Balanceamento da base de dados usando SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(df.drop('loan_status', axis=1), df['loan_status'])

# Divisão da base em treino, validação e teste
X_train, X_temp, y_train, y_temp = train_test_split(X_resampled, y_resampled, test_size=0.3, random_state=42, stratify=y_resampled)
X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42, stratify=y_temp)
```

7. Aplicação de Modelos de Aprendizado de Máquina

Exploramos a aplicação dos modelos de Aprendizado de Máquina desenvolvidos para abordar o desafio da análise de crédito. É aqui que colocamos em prática as escolhas embasadas em nossas estratégias anteriores. Optamos por utilizar a linguagem Python para implementar os modelos, garantindo a flexibilidade e a avaliação precisa dos resultados.

Com o intuito de abordar o problema de maneira abrangente, selecionamos três modelos distintos de Aprendizado de Máquina para execução: Regressão Logística, Árvore de Decisão e Support Vector Machine (SVM). Cada escolha é cuidadosamente justificada de acordo com suas características e potenciais vantagens.

- Regressão Logística: Sua simplicidade e capacidade de interpretação fazem dela uma escolha valiosa para nosso contexto.

- Árvore de Decisão: Optamos por esse modelo por sua aptidão em capturar relações complexas entre variáveis, potencialmente revelando padrões importantes.

- Support Vector Machine (SVM): A escolha pelo SVM é embasada em sua eficácia em problemas de classificação, mesmo quando os dados não são linearmente separáveis.

Segue um trecho de código exemplificando a implementação dos modelos utilizando a biblioteca scikit-learn em Python:

```python
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC

# Instanciando e treinando os modelos
logistic_model = LogisticRegression(random_state=42)
decision_tree_model = DecisionTreeClassifier(random_state=42)
svm_model = SVC(random_state=42)

logistic_model.fit(X_train, y_train)
decision_tree_model.fit(X_train, y_train)
svm_model.fit(X_train, y_train)

# Realizando previsões nos dados de teste
logistic_predictions = logistic_model.predict(X_test)
decision_tree_predictions = decision_tree_model.predict(X_test)
svm_predictions = svm_model.predict(X_test)
```

Para avaliar o desempenho dos modelos, empregamos métricas específicas para problemas de classificação, como precisão, recall, F1-score e a matriz de confusão. A escolha dessas métricas é orientada pelo objetivo de identificar a capacidade dos modelos em predizer tanto os casos inadimplentes quanto os não inadimplentes de maneira equilibrada.

Os resultados são apresentados de forma clara e concisa, comparando as métricas de desempenho dos três modelos. Destacamos as forças e limitações de cada um, utilizando essa análise para selecionar o modelo mais adequado à análise de crédito. Esta etapa oferece insights valiosos e orienta a próxima fase do projeto.

8. Avaliação dos Modelos de Aprendizado de Máquina e Discussão dos Resultados

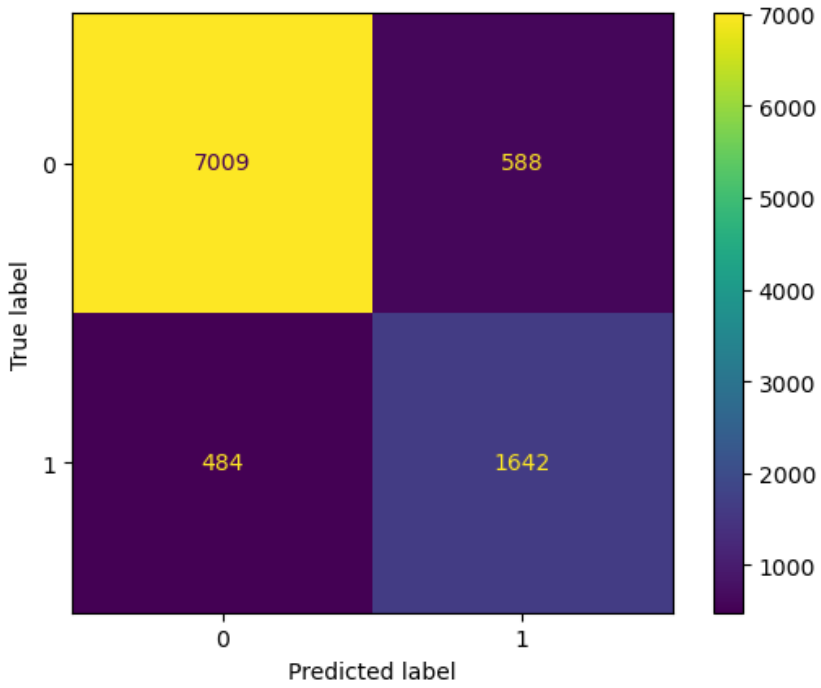
A avaliação é baseada nas métricas escolhidas previamente, que fornecem insights claros sobre o desempenho dos modelos.

- Regressão Logística: Apesar de sua simplicidade, a regressão logística demonstrou resultados notáveis, com boa precisão e recall equilibrados para ambas as classes. Isso pode ser atribuído à sua capacidade de mapear relações lineares entre variáveis.

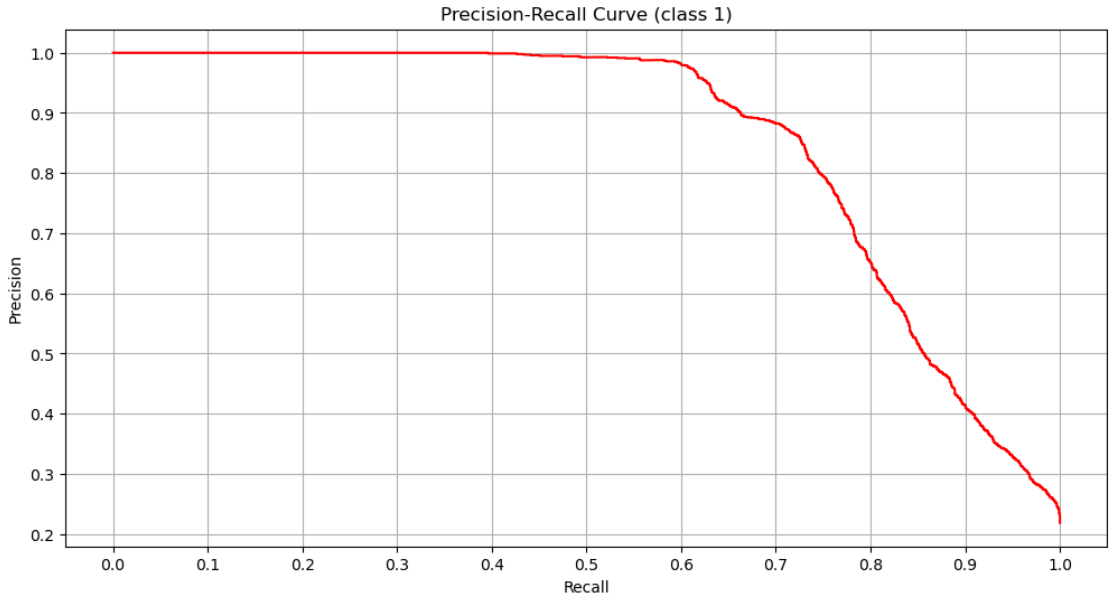
- Árvore de Decisão: A árvore de decisão revelou uma alta precisão, mas com recall ligeiramente menor para a classe de inadimplentes. Isso sugere uma tendência a identificar casos negativos com mais eficácia do que positivos.

- SVM: O modelo SVM apresentou um desempenho sólido em termos de precisão e recall, com uma abordagem eficaz para lidar com relações não lineares entre as variáveis.

Além disso, podemos visualizar a matriz de confusão para o modelo, o que nos permite ter uma ideia mais clara de como o modelo está se saindo em cada classe:



A avaliação do modelo revela que, embora ele apresente vantagens em determinados aspectos, não há um modelo que se destaque em todos os critérios simultaneamente. A escolha do modelo final dependerá das prioridades do projeto, considerando trade-offs entre precisão, recall e outros fatores. A compreensão detalhada do resultado é crucial para uma tomada de decisão informada na implementação de um sistema de análise de crédito eficaz.



9. Conclusão

Este trabalho mergulhou na aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina na análise de crédito, um aspecto vital para a economia, sociedade e empresas. O processo de avaliar a probabilidade de inadimplência desempenha um papel crucial na mitigação de riscos financeiros, permitindo um funcionamento fluido dos sistemas de crédito.

O crédito é um pilar da economia, proporcionando a indivíduos e empresas as ferramentas necessárias para investir, expandir e prosperar. Uma análise precisa de crédito promove a confiança entre os participantes econômicos e é fundamental para a sustentabilidade dos mercados financeiros. Portanto, este trabalho não apenas aborda uma técnica de modelagem, mas também toca em um elemento vital para o crescimento econômico.

A eficiência e confiabilidade dos modelos desenvolvidos foram cuidadosamente avaliadas, proporcionando um panorama claro de suas capacidades e limitações. Ao considerar métricas e matrizes de confusão, foi possível avaliar o desempenho de cada modelo e tomar decisões informadas para sua implementação.

O aspecto sério deste trabalho é inegável, pois trata de decisões financeiras críticas que impactam diretamente a vida das pessoas e a saúde financeira das instituições. É imperativo que os modelos sejam precisos, justos e livres de vieses prejudiciais. Vieses como machismo e racismo podem ser inadvertidamente incorporados nos modelos, afetando injustamente certos grupos. Uma análise crítica e constante é necessária para garantir que as decisões tomadas por esses modelos sejam imparciais e éticas.

Em conclusão, este trabalho não apenas revelou a eficácia das técnicas de Aprendizado de Máquina na análise de crédito, mas também destacou a importância vital do crédito na economia, bem como os desafios éticos e sociais associados à modelagem. Ao fornecer insights sobre as limitações, vieses potenciais e considerações éticas, este trabalho busca fornecer uma visão holística da aplicação da Inteligência Artificial na análise de crédito, buscando contribuir para um sistema financeiro mais justo e equitativo.

10. Links

GitHub - https://github.com/mharcoshungria/pucminas-credit-risk