Análise dos Dados de Inadimplência: Perfil e Comportamento dos Devedores

Discentes: Alexander Lira e Diego Wilson
Docente: Prof. Jodavid Ferreira
EE3

Departamento de Estatística
Universidade Federal de Pernambuco - UFPE

23 de setembro de 2024

Sumário I

Introdução

Análise de Inadimplência: Perfil e Comportamento dos Devedores

Nesta análise, será realizado um estudo sobre inadimplência com o objetivo de identificar padrões e fatores que influenciam o comportamento dos clientes em relação ao pagamento de suas dívidas. Utilizando um conjunto de dados com variáveis socioeconômicas e comportamentais, vamos construir dois modelos de machine learning diferentes para prever a inadimplência. Os modelos serão comparados em termos de desempenho, permitindo avaliar qual deles oferece as melhores previsões e pode ser mais eficaz para a tomada de decisões no contexto da análise de crédito.

Objetivo Geral

Objetivo da Análise

O objetivo desta análise é conectar os padrões identificados nos dados de inadimplência com o comportamento de pagamento dos clientes, intervindo em soluções que possam impactar positivamente a experiência e previsibilidade no contexto do crédito. A análise permitirá um entendimento detalhado sobre os fatores que influenciam a inadimplência, usando técnicas de machine learning para suportar decisões mais assertivas no gerenciamento de crédito.

Metodologia e Modelagem

Pré-processamento e Modelagem dos Dados

O processo foi dividido em duas etapas principais: pré-processamento dos dados e construção dos modelos de machine learning:

- **Tratamento de dados faltantes:** Substituição dos valores ausentes pela mediana categorizada pela idade, para manter a coerência entre as faixas etárias e evitar viés.
- **Descrição das variáveis:** Exploração das variáveis socioeconômicas e comportamentais para compreender sua relação com a inadimplência.
- **Balanceamento da variável resposta:** Garantiu que o modelo não estivesse enviesado para a classe majoritária, melhorando a precisão das previsões.

Metodologia e Modelagem

Pré-processamento e Modelagem dos Dados

O processo foi dividido em duas etapas principais:
pré-processamento dos dados e construção dos modelos de machine learning:

Modelos utilizados:

- **Decision Tree:** Modelo interpretável que cria uma árvore de decisões baseada nas variáveis preditoras.
- **Random Forest:** Modelo mais robusto, composto por várias árvores de decisão, oferecendo maior precisão ao reduzir o risco de overfitting.

Os modelos foram comparados em termos de desempenho para identificar o mais eficiente na previsão de inadimplência.

Bibliotecas

Bibliotecas Utilizadas

As bibliotecas utilizadas para manipulação e análise dos dados:

- `pycaret.classification`: Para a construção e comparação de modelos de machine learning.
- `sklearn.tree`: Para visualização e exportação de árvores de decisão.
- `sklearn.metrics`: Para cálculo de métricas de desempenho como matriz de confusão, acurácia, e relatórios de classificação.
- `sklearn.model_selection`: Para divisão dos dados em treino e teste.
- `pandas` e `numpy`: Para manipulação e análise dos dados.
- `seaborn` e `matplotlib.pyplot`: Para visualização de gráficos.

Sobre os Dados

Fonte de Dados

Os dados foram obtidos do repositório Kaggle, uma fonte amplamente utilizada e confiável para análise de dados. Disponível em: `https://www.kaggle.com`.

Variáveis dos Dados

Principais Variáveis

As principais variáveis incluídas no conjunto de dados são:

- **inadimplente:** variável de resposta, indicando se o cliente está inadimplente (1) ou adimplente (0);
- **util_linhas_inseguras:** percentual de utilização das linhas de crédito inseguras;
- **idade:** idade do cliente;
- **vezes_passou_de_30_59_dias:** número de vezes que o cliente passou de 30 a 59 dias em atraso no pagamento.

Variáveis dos Dados (Cont.)

- **razao_debito:** razão entre o valor do débito e o total do crédito disponível;
- **salario_mensal:** salário mensal do cliente;
- **numero_linhas_crdto_aberto:** número de linhas de crédito abertas atualmente em nome do cliente;
- **numero_vezes_passou_90_dias:** número de vezes que o cliente passou 90 dias ou mais em atraso;
- **numero_emprestimos_imobiliarios:** número de empréstimos imobiliários que o cliente possui;
- **numero_de_vezes_que_passou_60_89_dias:** número de vezes que o cliente passou de 60 a 89 dias em atraso;
- **numero_de_dependentes:** número de dependentes do cliente.

Estrutura dos Dados

Descrição do DataFrame

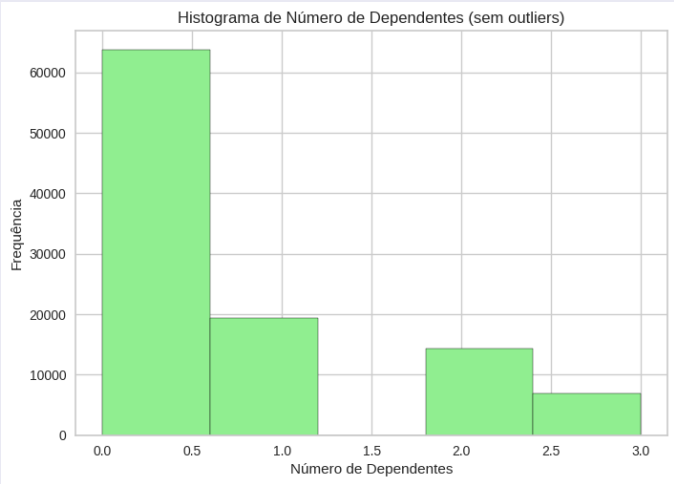
O conjunto de dados contém 110.000 entradas e 11 colunas.
Abaixo, estão descritas as principais características:

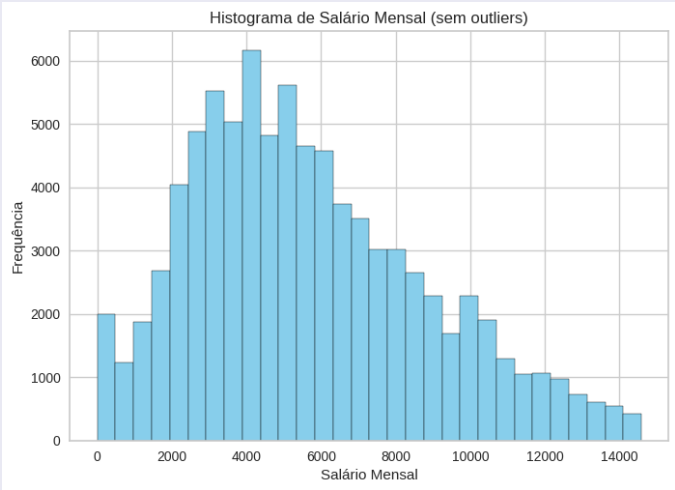
- O DataFrame inclui tanto variáveis do tipo `int64` (7 colunas) quanto `float64` (4 colunas).
- Algumas variáveis contêm valores ausentes, como:
  - **salário_mensal**: 88.237 entradas preenchidas (21.763 valores ausentes).
  - **número_de_dependentes**: 107.122 entradas preenchidas (2.878 valores ausentes).

Pré Processamento e Análise Exploratória

Histogramas das variáveis com valores nulos

Figura: Histograma de Número de Dependentes e Salário Mensal





Justificativa do Tratamento de Nulos

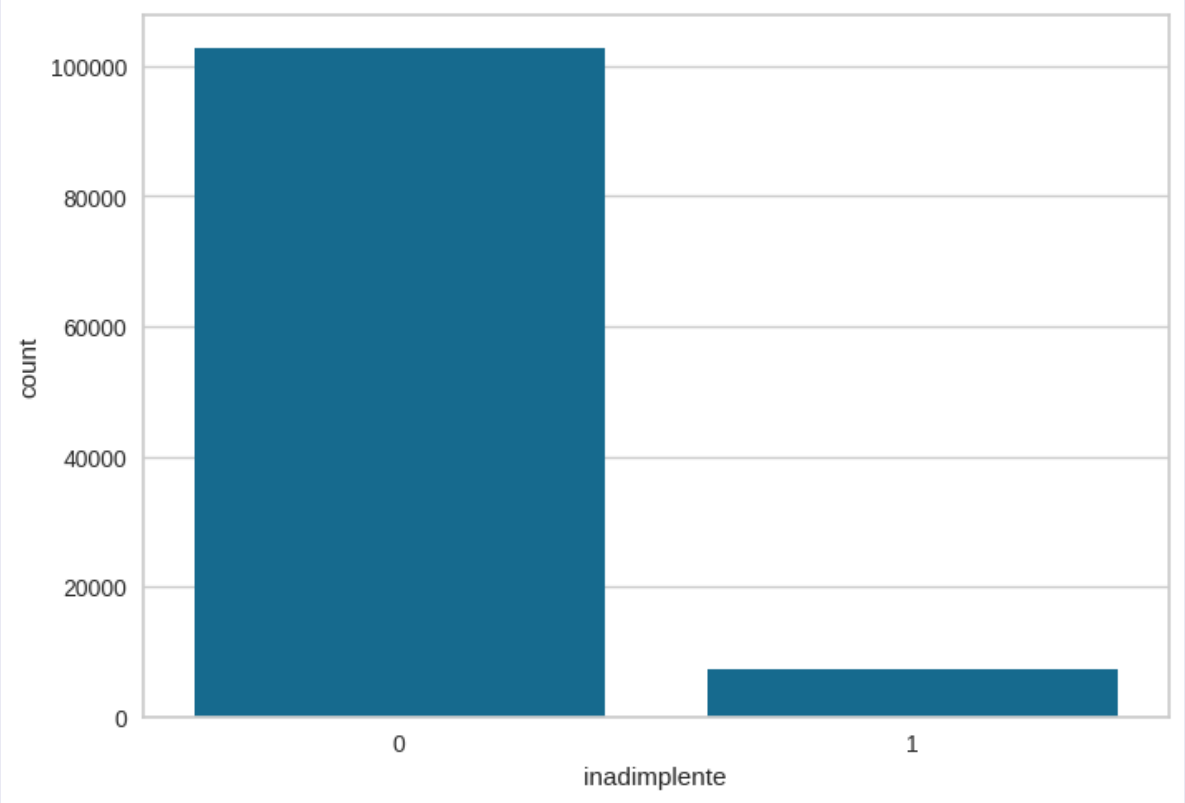
Justificativa do Tratamento de Nulos

A presença de valores nulos nas variáveis *salário_mensal* e *número_de_dependentes* pode impactar negativamente a precisão dos modelos de machine learning. Para evitar perda de dados, optamos por substituir os valores nulos pela mediana categorizada pela idade para o salário e apenas a mediana para numero de dependentes, garantindo a consistência dos dados sem introduzir viés ou distorções significativas. As ultimas observações nulas restantes foram excluidas por serem apenas 3 observações que eram outliers com relação a variável idade.

Balanceamento da Variável Inadimplente

Gráfico de Barra

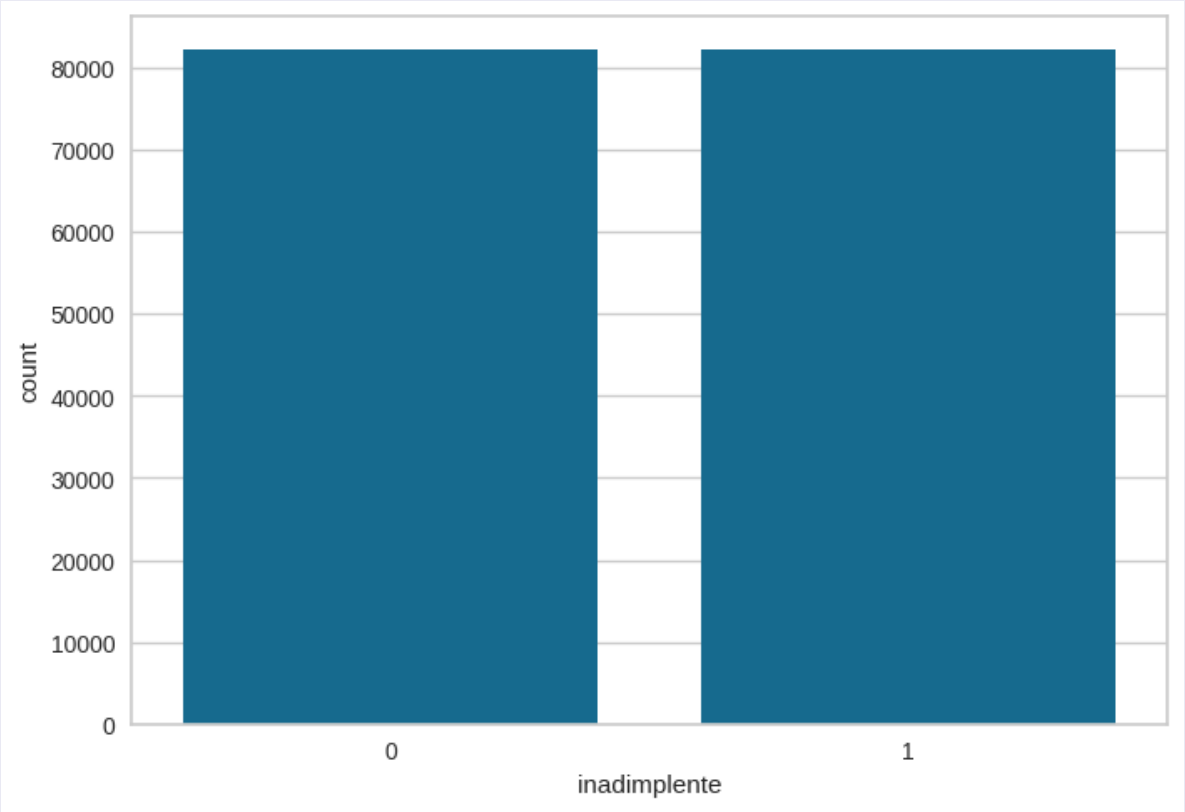
Figura: Variável Resposta



Resultado do Balanceamento

Aplicação do SMOTE da Biblioteca imblearn.over_sampling

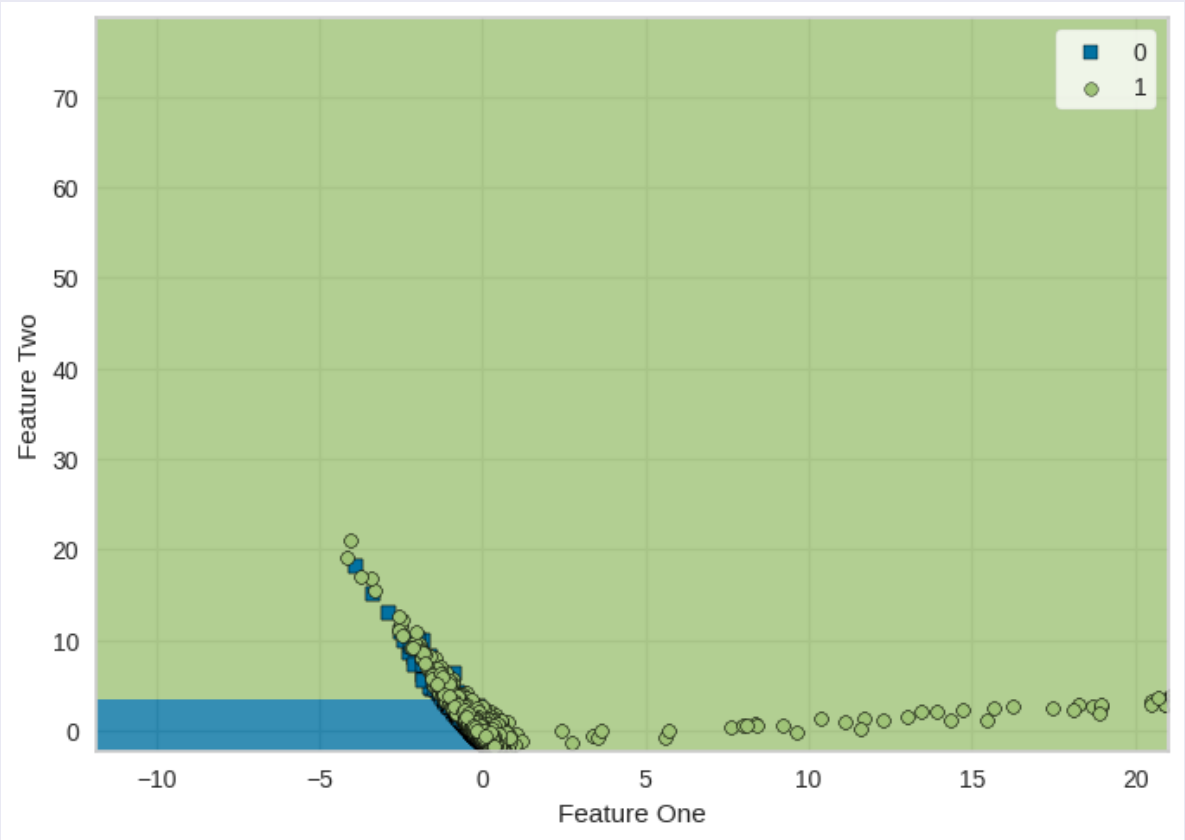
Figura: Variável Resposta Final



Decision Tree

Parte 1

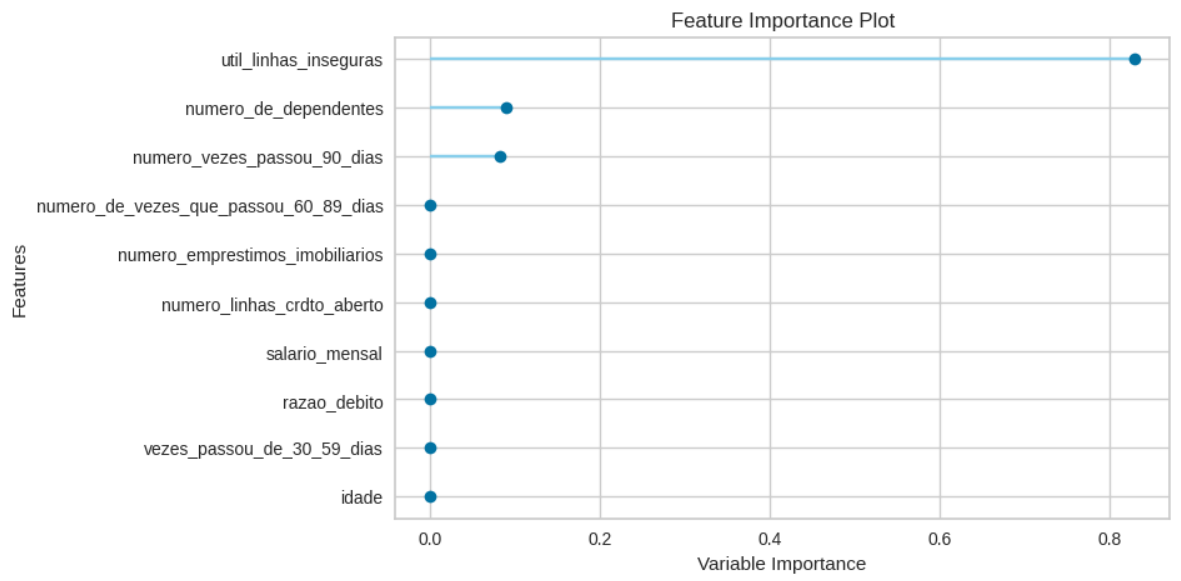
Figura: Modelo Decision Tree



Decision Tree

Parte 2

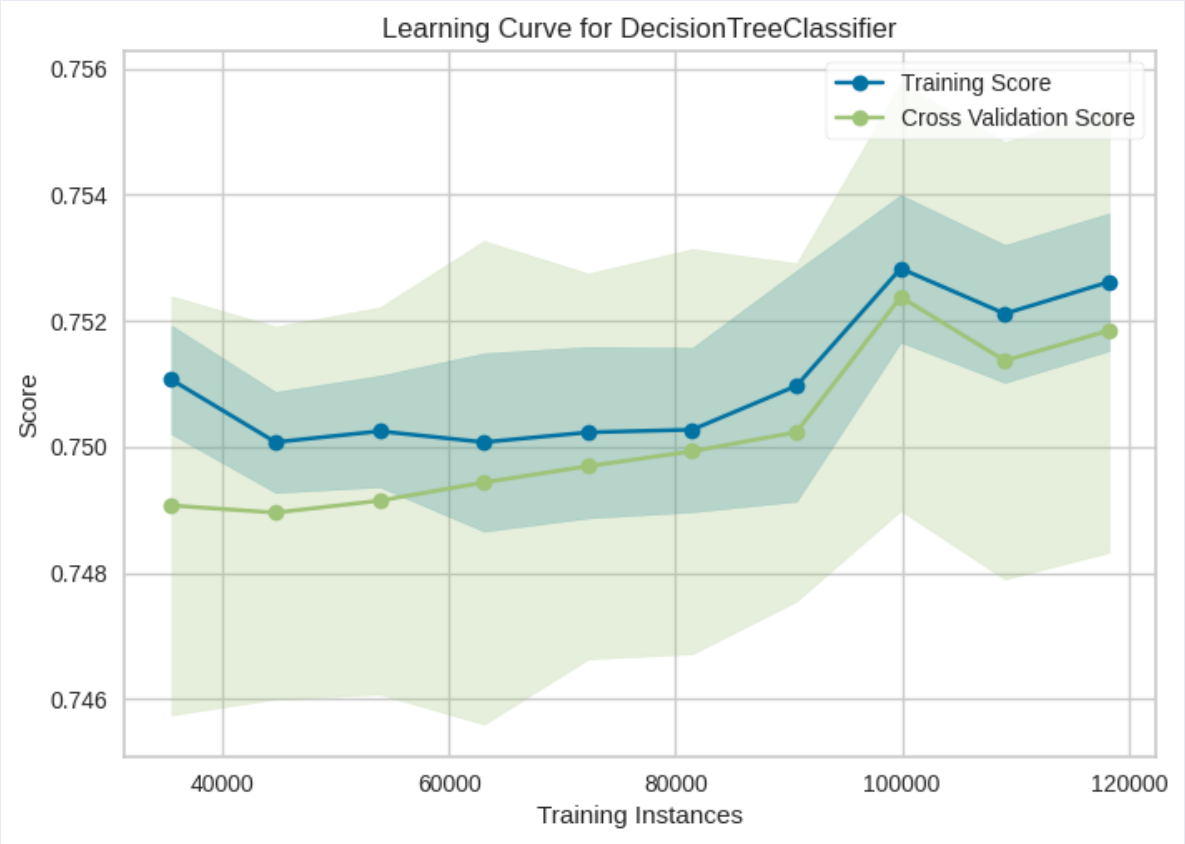
Figura: Modelo Decision Tree



Decision Tree

Parte 3

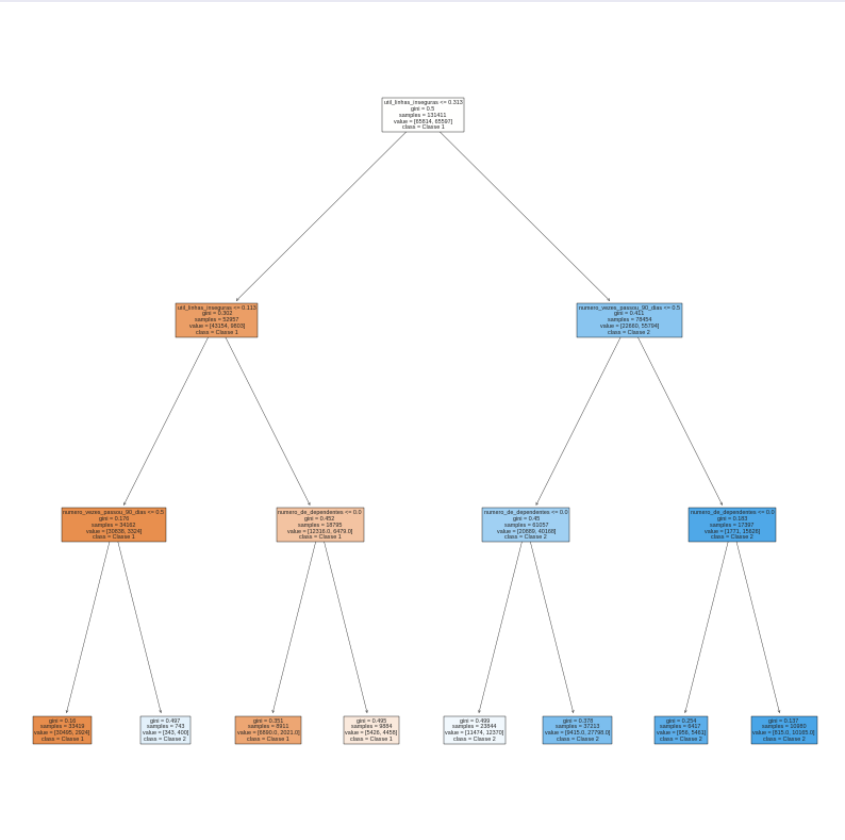
Figura: Modelo Decision Tree



Decision Tree

Parte 4

Figura: Modelo Decision Tree



Desempenho do Modelo: Decision Tree Classifier

Métricas de Desempenho

- **Accuracy:** 75.64%
  - Aproximadamente 76% das previsões foram corretas.
- **AUC:** 0.8318
  - Indica boa capacidade do modelo em distinguir entre as classes.

Desempenho do Modelo: Decision Tree Classifier

Métricas de Desempenho

- **Recall:** 85.51%
  - O modelo conseguiu identificar 85% dos clientes inadimplentes.
- **Precisão:** 71.60%
  - 72% das previsões de inadimplência foram corretas.
- **F1-Score:** 77.94%
  - Um bom equilíbrio entre precisão e recall.

Desempenho do Modelo: Decision Tree Classifier

Métricas de Desempenho

- **Kappa:** 0.5121
  - Indica uma concordância moderada entre as previsões e a realidade.
- **MCC:** 0.5222
  - Reflete uma correlação positiva entre as classes previstas e reais.

Desempenho do Modelo: Decision Tree Classifier

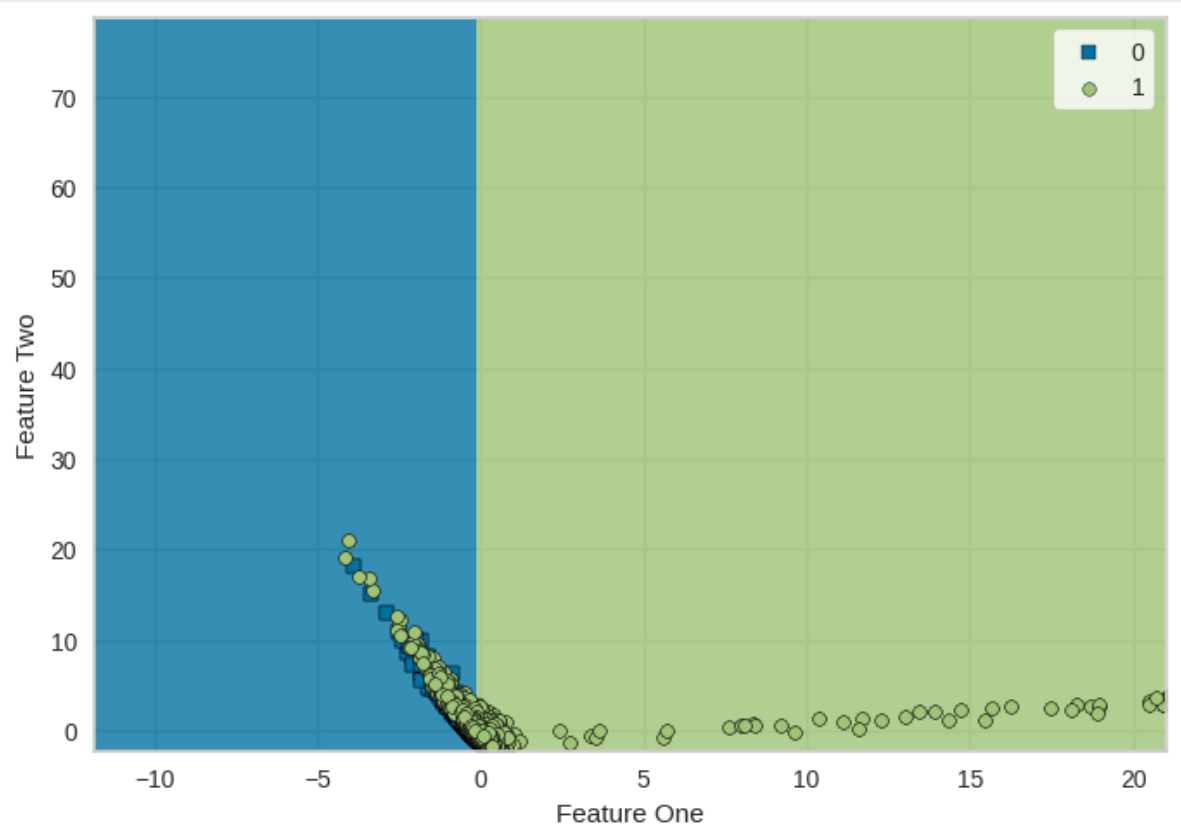
Classificação

- **Classe 0 (Adimplente):**
  - Precisão: 0.82
  - Recall: 0.66
  - F1-Score: 0.73
- **Classe 1 (Inadimplente):**
  - Precisão: 0.72
  - Recall: 0.86
  - F1-Score: 0.78

Random Forest

Parte 1

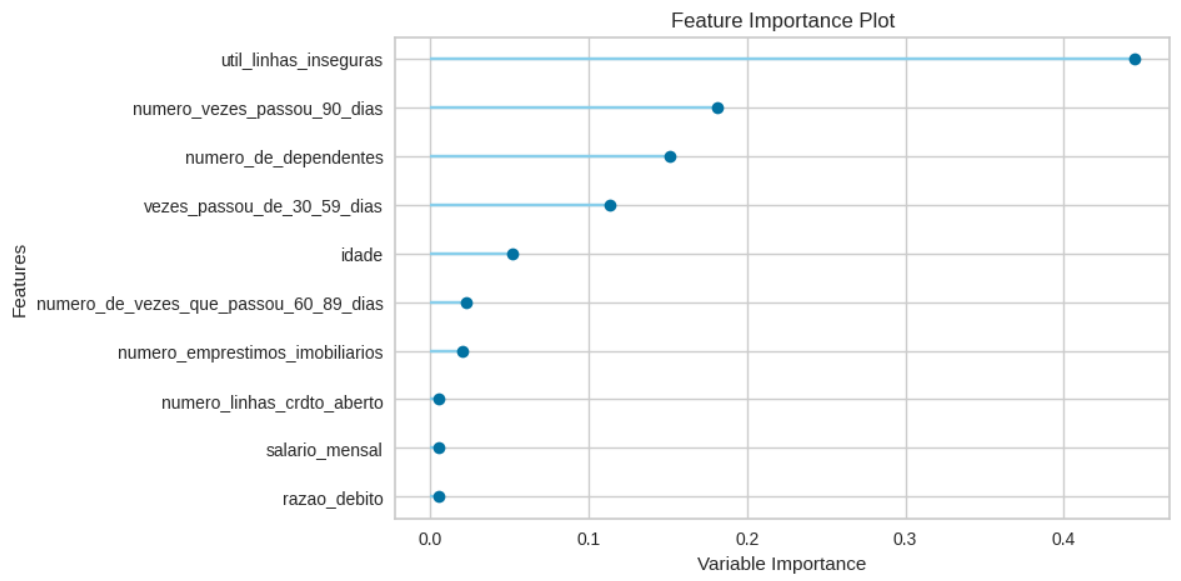
Figura: Modelo Random Forest



Random Forest

Parte 2

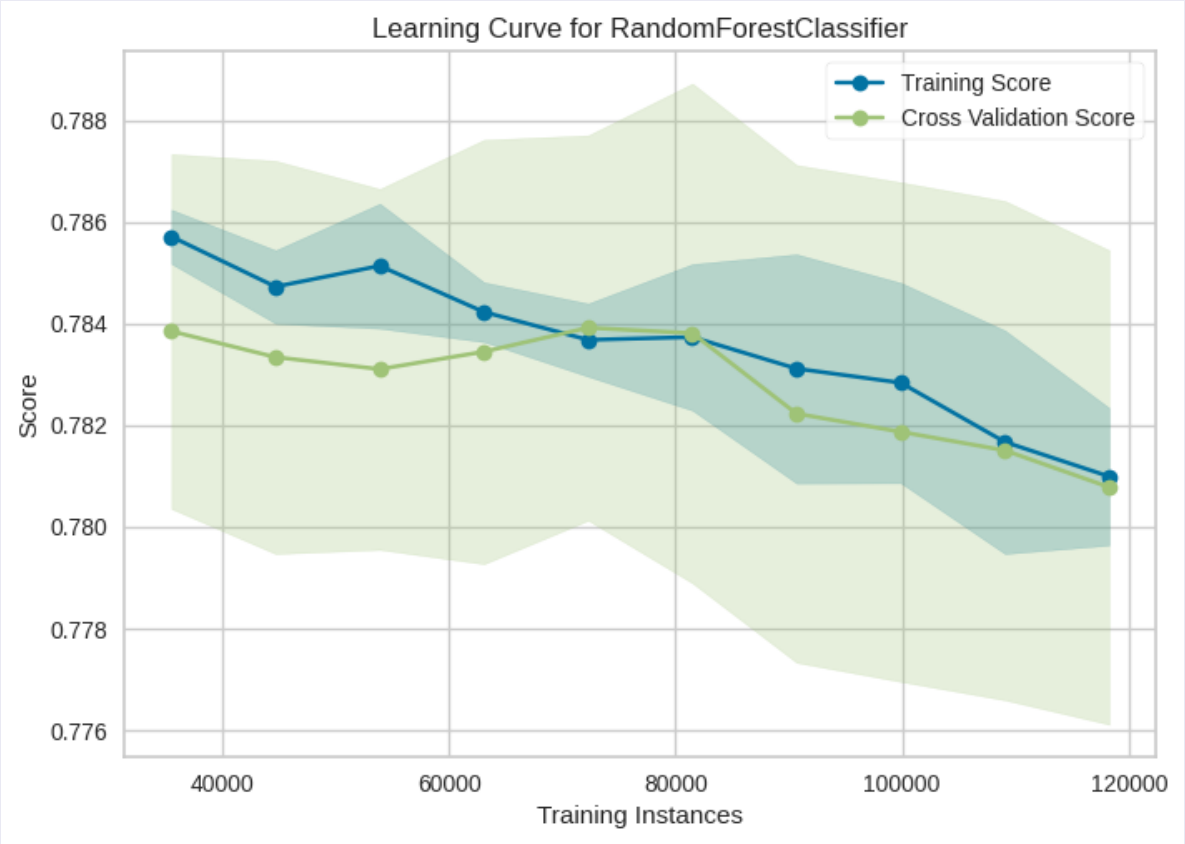
Figura: Modelo Random Forest



Random Forest

Parte 3

Figura: Modelo Random Forest



Desempenho do Modelo: Random Forest Classifier

Métricas de Desempenho

- **Accuracy:** 78.32%
  - Aproximadamente 78% das previsões foram corretas, melhor que o modelo Decision Tree.
- **AUC:** 0.8682
  - Excelente capacidade de distinção entre classes.

Desempenho do Modelo: Random Forest Classifier

Métricas de Desempenho

- **Recall:** 79.42%
  - O modelo identificou 79% dos inadimplentes.
- **Precisão:** 77.93%
  - Alta taxa de acerto nas previsões de inadimplência.

Desempenho do Modelo: Random Forest Classifier

Métricas de Desempenho

- **F1-Score:** 78.67%
  - Bom equilíbrio entre precisão e recall, superior ao modelo anterior.
- **Kappa:** 0.5663
  - Concordância moderada a boa entre previsões e realidade.
- **MCC:** 0.5664
  - Boa correlação entre as classes previstas e reais.

Desempenho do Modelo: Random Forest Classifier

Classificação

- **Classe 0 (Adimplente):**
  - Precisão: 0.79
  - Recall: 0.77
  - F1-Score: 0.78
- **Classe 1 (Inadimplente):**
  - Precisão: 0.78
  - Recall: 0.79
  - F1-Score: 0.79

Conclusão

Análise dos Modelos

- O modelo **Random Forest** superou o **Decision Tree** em diversas métricas:
  - **Accuracy** superior (78.32% vs 75.64%).
  - Melhor **AUC** (0.8682 vs 0.8318), indicando maior capacidade preditiva.
- Ambos os modelos demonstraram resultados promissores:
  - O **Recall** para a classe inadimplente é crucial, pois indica a capacidade de identificar clientes que podem representar risco, ou seja, estão inadimplentes.
- Para um contexto de análise de crédito, o modelo Decision Tree, embora tenha um recall maior, não é preferível devido ao seu desempenho geral mais robusto no random forest como por exemplo a precisão.

Obrigado!