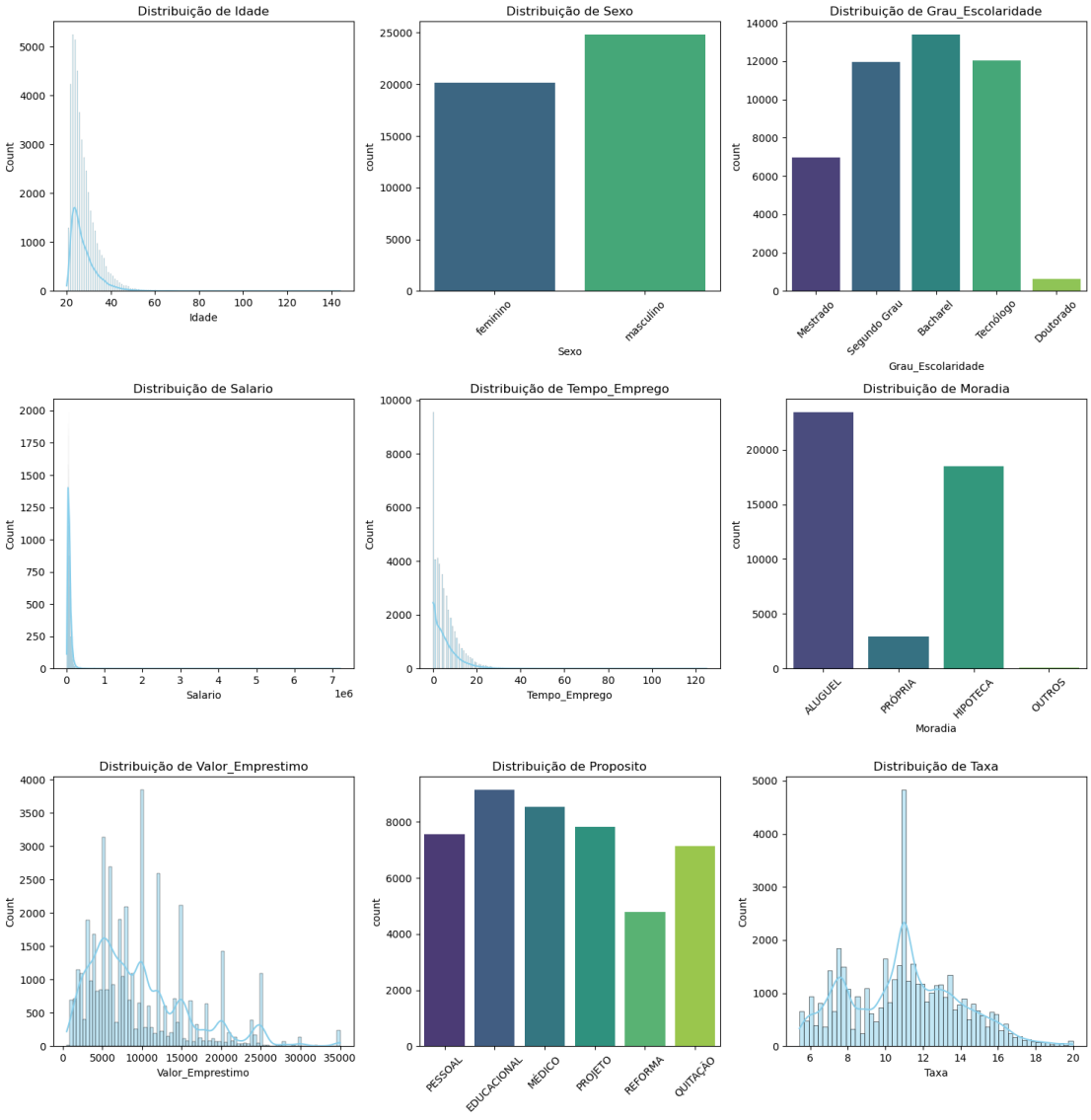
Relatório Técnico – Análise de Concessão de Crédito

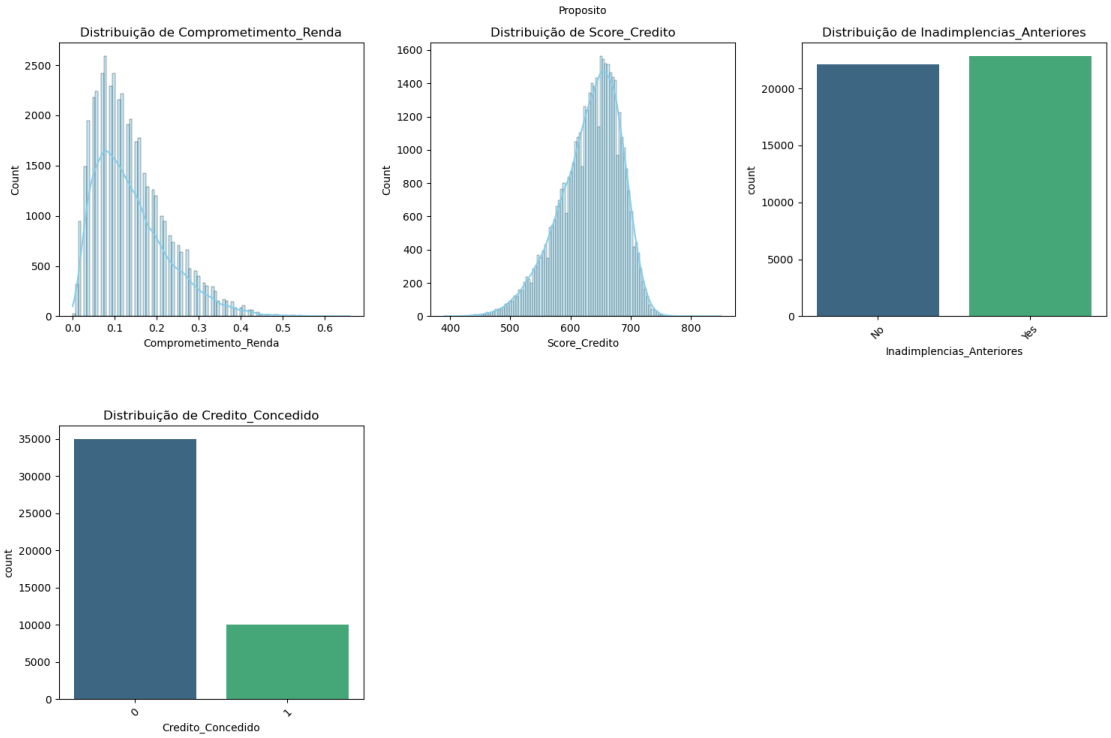
Beatriz Almeida Felício

Introdução

Esse projeto tem como objetivo modelar a concessão de crédito utilizando técnicas de ciência de dados. Para isso, foi utilizado um conjunto de dados contendo 45.000 registros e 14 variáveis, entre atributos financeiros e demográficos dos solicitantes de empréstimo. A variável alvo é Crédito_Concedido, que indica se o empréstimo foi aprovado (1) ou negado (0).

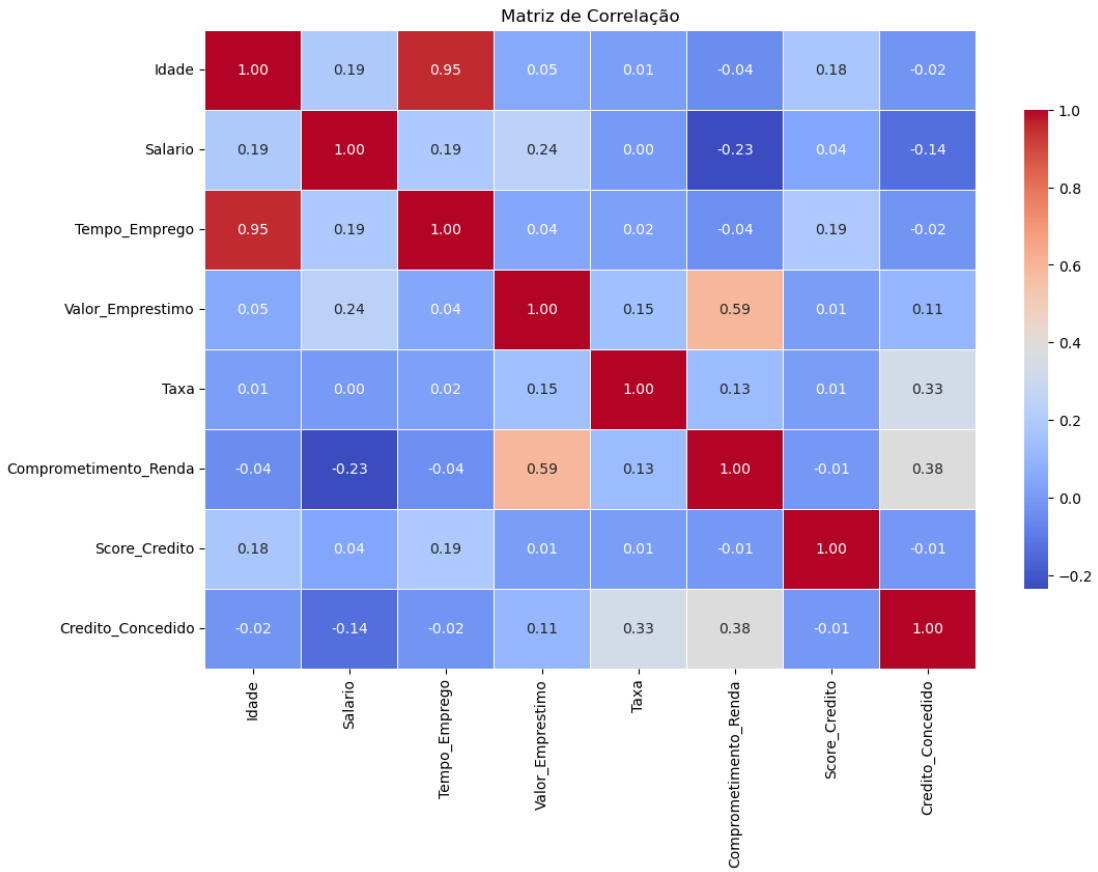
O conjunto de dados inclui variáveis como idade, sexo, salário, tempo de emprego, valor do empréstimo, taxa aplicada, comprometimento da renda, entre outras. Foram analisadas as distribuições de cada variável, correlações entre elas e possíveis desequilíbrios na variável alvo. As visualizações correspondentes acompanham esse relatório e evidenciam os principais padrões e relações encontradas.





Principais Interpretações

1. **Desequilíbrio de Classes**: A variável Crédito_Concedido apresentou desequilíbrio, com predominância de uma das classes. Esse fator foi considerado durante o pipeline de modelagem, utilizando técnicas de **undersampling** para balancear os dados.
2. **Outliers**: Foram identificados outliers em variáveis como Idade e Tempo_Emprego. Esses valores foram tratados para minimizar seu impacto negativo nos modelos, utilizando técnicas robustas.
3. **Correlação entre Variáveis**:
    - Correlação esperada entre **Idade** e **Tempo de Emprego**.
    - Relação inversa entre **Salário** e **Comprometimento de Renda**, sugerindo que pessoas com maior renda comprometem menos sua renda com dívidas.
    - Relação direta entre **Salário** e **Crédito Concedido**, indicando influência do poder aquisitivo na decisão de crédito.
    - Aumento do **Valor do Empréstimo** tende a elevar o **Comprometimento de Renda**.
    - Taxas menores de juros estão associadas a maior número de **Créditos Concedidos**.



Pré-Processamento

O pipeline de preparação dos dados incluiu:

- Remoção de outliers;
- Normalização com RobustScaler, para maior robustez frente a valores extremos;
- Codificação de variáveis categóricas:
    - Sexo, Moradia e Proposito: codificados com One Hot Encoding;
    - Inadimplencias_Anteriores: transformada com Label Binarizer;
- Divisão estratificada dos dados, respeitando o desequilíbrio das classes;
- Balanceamento utilizando undersampling da classe majoritária.

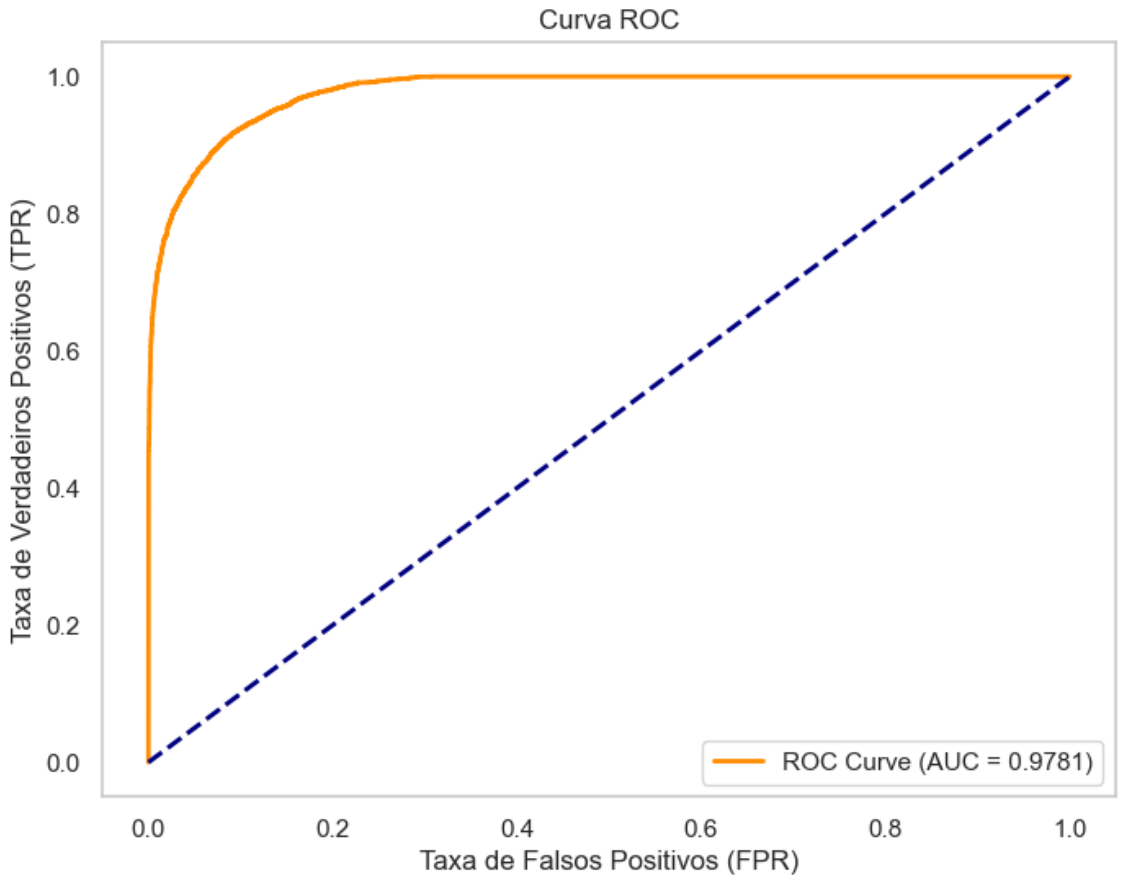
Modelagem

Foram treinados e avaliados os seguintes modelos:

- Regressão Logística
- Random Forest
- XGBoost
- LightGBM

As métricas utilizadas para avaliação foram Acurácia, Precisão, Revocação e F1 Score. O melhor desempenho foi alcançado com o **modelo LightGBM**, que, após ajuste de hiperparâmetros, apresentou os seguintes resultados:

- Accuracy: 0.9348
- Precision: 0.8999
- Recall: 0.7952
- F1 Score: 0.8443
- AUC Score (ROC Curve): 0.9781



Conclusão

O modelo final demonstrou excelente desempenho na predição da concessão de crédito, especialmente no equilíbrio entre precisão e recall. A análise detalhada das variáveis permitiu entender os fatores que mais influenciam a decisão de crédito, e o uso de técnicas apropriadas de pré-processamento e balanceamento foi fundamental para garantir a robustez do modelo. Esse pipeline pode ser aplicado a cenários reais, auxiliando instituições financeiras na tomada de decisão de forma eficiente e baseada em dados.