**Modelo para Concessão de Crédito da PoD Bank**

****

**1.Riscos na concessão de crédito**

A concessão de crédito envolve riscos significativos para as empresas fornecedoras, especialmente quando há inadimplência por parte dos clientes. Aqui estão alguns problemas e riscos associados à inadimplência na concessão de crédito:

* **Perda Financeira:**

A inadimplência pode resultar em perda financeira direta para a empresa, uma vez que os valores devidos podem não ser recuperados. Isso impacta diretamente a rentabilidade e a saúde financeira da empresa.

* **Redução da Liquidez:**

A inadimplência pode afetar a liquidez da empresa, pois os fundos que deveriam ser recebidos podem não estar disponíveis para cobrir despesas operacionais, pagamento de fornecedores ou investimentos.

* **Aumento de Custos de Recuperação:**

Recuperar valores de clientes inadimplentes pode exigir esforços significativos e custos adicionais. Isso inclui custos legais, custos de cobrança e tempo dedicado à gestão de recuperação.

* **Avaliação de Crédito Inadequada:**

Uma avaliação inadequada do crédito, incluindo a falta de análise detalhada do histórico financeiro dos clientes, capacidade de pagamento e comportamento de crédito passado, pode levar a concessões de crédito para clientes de alto risco.

* **Fraude:**

A concessão de crédito sem uma avaliação adequada pode abrir espaço para atividades fraudulentas. Os clientes podem fornecer informações falsas ou intencionalmente buscar crédito com a intenção de não honrar os compromissos.

**1.1 Principais KPIs (Indicadores Chave de Desempenho) de crédito**:

* Taxa de Inadimplência média da carteira atual: 8%
* Taxa de inadimplência média da carteira futura: 4,3%
* Taxa de Aprovação: 70% dos novos clientes
* Limite de Crédito Utilizado
* Prazo Médio de Pagamento
* Índice de Fraude na Concessão de Crédito

**2. Entendimento do Negócio**

A PoD Bank, uma startup no setor financeiro, conquistou um sólido posicionamento de mercado e maturidade ao atender a população com escassa informação de crédito. A empresa oferece crédito imobiliário, veicular e empréstimo pessoal. Diante desse crescimento, identificou a necessidade de aprimorar suas decisões de crédito, superando a dependência exclusiva do Business Intelligence (BI).

Assim, para fortalecer ainda mais sua abordagem, a PoD Bank optou por estabelecer uma área dedicada ao Planejamento e Modelagem de Crédito, centrada na implementação de modelos estatísticos. Essa estratégia visa aprimorar a eficácia na análise de riscos e garantir uma tomada de decisão mais fundamentada, refletindo o compromisso contínuo da empresa em oferecer soluções de crédito inovadoras e sustentáveis.

**3. Objetivo do Projeto**

Este projeto visa a criação de um modelo de risco de crédito, concebido como uma abordagem generalista que, além de atender às necessidades atuais, servirá como um sólido ponto de partida para futuras evoluções na definição do público-alvo.

O objetivo primordial é desenvolver um modelo preditivo destinado à avaliação do risco de inadimplência de potenciais clientes. Essa ferramenta tem a finalidade de embasar decisões estratégicas e fundamentadas no que diz respeito à concessão de crédito. O modelo será estruturado para fornecer insights valiosos aos analistas, contribuindo significativamente para a análise de riscos e, consequentemente, para a tomada de decisões mais informadas.

Além disso, o modelo proposto busca ser uma ferramenta de auxílio essencial para os analistas, oferecendo suporte na avaliação e decisão sobre a concessão ou não de crédito a novos clientes, aumentando a eficiência e a precisão na identificação de potenciais riscos de inadimplência.

**4. Sobre os Dados**

Os dados utilizados para a análise são provenientes de fontes internas e externas, conferindo assim uma perspectiva ampla e detalhada acerca do perfil dos solicitantes. Os dados internos abrangem informações cadastrais e transacionais.

Base de dados disponibilizados pela PoD Bank.

* **application\_train.csv/application\_test.csv**: Estes são os principais dados de treino e teste com informações sobre cada solicitação de empréstimo na PoD Bank. Cada empréstimo tem sua própria linha e é identificado pela variável SK\_ID\_CURR. Os dados de aplicação de treinamento vêm com a variável TARGET, indicando 0: o empréstimo foi pago ou 1: o empréstimo não foi pago (inadimplência).
* **previous\_application.csv**: Contém informações sobre aplicações de empréstimo anteriores de um cliente na PoD Bank.
* **installments\_payments.csv**: Detalha o histórico de pagamentos de empréstimos anteriores na PoD Bank.
* **bureau.csv**: Fornece dados de crédito de outras instituições financeiras.
* **POS\_CASH\_balance.csv**: Informações sobre o histórico de pagamentos de POS (Point of Sale) ou empréstimos em dinheiro na PoD Bank.
* **bureau\_balance.csv**: Informações mensais sobre créditos anteriores do cliente em outras instituições financeiras.
* **credit\_card\_balance.csv**: Informações mensais sobre saldos de cartões de crédito do cliente na PoD Bank.

**5. Configuração do ambiente**

A metodologia adotada para o desenvolvimento do modelo de risco de crédito envolveu a utilização de diversas bibliotecas, métodos e funções essenciais. Abaixo estão listadas as principais importações realizadas no ambiente Python:

Bibliotecas, Métodos e Funções utilizadas

* \*\*pandas as pd:\*\*

- Utilizada para manipulação e análise eficiente de dados tabulares.

* \*\*numpy as np:\*\*

- Essencial para operações matemáticas e manipulação eficiente de arrays multidimensionais.

* \*\*matplotlib.pyplot as plt:\*\*

- Ferramenta gráfica para visualização de dados, utilizada na criação de gráficos e plots.

* \*\*seaborn as sns:\*\*

- Facilita a criação de gráficos estatísticos atraentes e informativos.

* \*\*from sklearn.model\_selection import train\_test\_split:\*\*

- Método para dividir o conjunto de dados em conjuntos de treino e teste.

* \*\*import pickle:\*\*

- Usada para serializar e desserializar objetos Python, facilitando o armazenamento de modelos treinados.

* \*\*from sklearn.preprocessing import LabelEncoder:\*\*

- Função para codificar variáveis categóricas em formato numérico.

* \*\*from sklearn.preprocessing import OneHotEncoder:\*\*

- Função para realizar codificação one-hot em variáveis categóricas.

* \*\*from sklearn.tree import DecisionTreeClassifier:\*\*

- Implementação de um classificador baseado em árvore de decisão.

* \*\*from sklearn.ensemble import GradientBoostingClassifier:\*\*

- Implementação de um classificador baseado em Gradient Boosting.

* \*\*import statsmodels.api as sm:\*\*

- Fornece funcionalidades para modelagem estatística, especialmente útil em análises mais detalhadas.

* \*import warnings:\*\*

- Utilizado para gerenciar e filtrar avisos (warnings) no código.

* \*\*import xgboost as xgb:\*\*

- Biblioteca que oferece uma implementação eficiente do algoritmo XGBoost, utilizado para treinamento do modelo performático.

* \*\*from sklearn.metrics import accuracy\_score, recall\_score, precision\_score, f1\_score, confusion\_matrix, roc\_curve, auc, roc\_auc\_score:\*\*

**6. Etapa de Feature Engineering**

Feature Engineering refere-se ao processo de seleção, transformação e criação de variáveis a serem utilizadas pelo modelo preditivo. Este é um passo crítico para garantir que o modelo tenha acesso a informações relevantes e significativas para realizar previsões precisas.

No processo de Feature Engineering, focado na melhoria da representação do conjunto de dados da aplicação, foram criadas 11 novas variáveis a partir das características já presentes na base "application". Cada variável tem como objetivo enriquecer a compreensão do perfil do aplicante, proporcionando insights valiosos para o modelo de avaliação de risco de crédito. A seguir, apresentamos cada variável juntamente com sua fórmula de cálculo e explicação:

1. APP\_INCOME\_CREDIT\_RATIO (Razão de Renda para Crédito do Aplicante):

Fórmula: abt\_00['APP\_INCOME\_CREDIT\_RATIO'] = abt\_00['AMT\_INCOME\_TOTAL'] / abt\_00['AMT\_CREDIT']

Explicação: Indica a proporção entre a renda total do aplicante e o montante total de crédito solicitado, oferecendo insights sobre o comprometimento financeiro em relação ao valor do crédito.

1. APP\_ANNUITY\_INCOME\_RATIO (Razão de Anuidade para Renda do Aplicante):

Fórmula: abt\_00['APP\_ANNUITY\_INCOME\_RATIO'] = abt\_00['AMT\_ANNUITY'] / abt\_00['AMT\_INCOME\_TOTAL']

Explicação: Representa a proporção da renda total do aplicante destinada ao pagamento da anuidade, auxiliando na avaliação da capacidade do aplicante de arcar com o compromisso mensal em relação à sua renda.

1. APP\_AGE (Idade do Aplicante):

Fórmula: abt\_00['APP\_AGE'] = abt\_00['DAYS\_BIRTH'] / -365

Explicação: Calcula a idade do aplicante em anos, considerando o campo 'DAYS\_BIRTH', geralmente representando a idade em dias com valores negativos.

1. APP\_CHILDREN\_INCOME\_RATIO (Razão de Filhos para Renda do Aplicante):

Fórmula: abt\_00['APP\_CHILDREN\_INCOME\_RATIO'] = abt\_00['CNT\_CHILDREN'] / abt\_00['AMT\_INCOME\_TOTAL']

Explicação: Indica a proporção da renda do aplicante associada ao número de filhos, proporcionando insights sobre a carga financeira representada pelos filhos em relação à renda.

1. EXT\_SOURCE\_MEAN (Média das Fontes Externas):

Fórmula: abt\_00['EXT\_SOURCE\_MEAN'] = (abt\_00['EXT\_SOURCE\_1'] + abt\_00['EXT\_SOURCE\_2'] + abt\_00['EXT\_SOURCE\_3']) / 3

Explicação: Calcula a média das três fontes externas de pontuação de crédito (EXT\_SOURCE\_1, EXT\_SOURCE\_2, EXT\_SOURCE\_3), indicando a confiabilidade de crédito.

1. APP\_EMPLOYED\_LENGTH (Tempo de Emprego do Aplicante em Anos):

Fórmula: abt\_00['APP\_EMPLOYED\_LENGTH'] = abt\_00['DAYS\_EMPLOYED'] / -365

Explicação: Representa o tempo de emprego do aplicante em anos, calculado a partir do campo 'DAYS\_EMPLOYED'.

1. APP\_GOODS\_CREDIT\_RATIO (Razão de Bens para Crédito do Aplicante):

Fórmula: abt\_00['APP\_GOODS\_CREDIT\_RATIO'] = abt\_00['AMT\_GOODS\_PRICE'] / abt\_00['AMT\_CREDIT']

Explicação: Indica a razão entre o valor dos bens adquiridos pelo aplicante e o montante total do crédito, fornecendo insights sobre o tipo de compra em relação ao crédito.

1. APP\_CREDIT\_DOWN\_PAYMENT (Pagamento Inicial do Aplicante):

Fórmula: abt\_00['APP\_CREDIT\_DOWN\_PAYMENT'] = abt\_00['AMT\_GOODS\_PRICE'] - abt\_00['AMT\_CREDIT']

Explicação: Representa o valor do pagamento inicial feito pelo aplicante em relação ao preço total dos bens.

1. APP\_NEW\_DAYS\_EMPLOYED\_PERC (Percentual de Dias Empregado em Relação à Idade):

Fórmula: abt\_00['APP\_NEW\_DAYS\_EMPLOYED\_PERC'] = abt\_00['DAYS\_EMPLOYED'] / abt\_00['DAYS\_BIRTH']

Explicação: Indica a porcentagem de dias que o aplicante esteve empregado em relação à sua idade, oferecendo insights sobre a estabilidade no emprego ao longo do tempo.

1. APP\_NEW\_INCOME\_PER\_PERSON (Renda por Pessoa na Família):

Fórmula: abt\_00['APP\_NEW\_INCOME\_PER\_PERSON'] = abt\_00['AMT\_INCOME\_TOTAL'] / abt\_00['CNT\_FAM\_MEMBERS']

Explicação: Calcula a renda média por pessoa na família do aplicante, considerando o número total de membros da família.

1. APP\_NEW\_PAYMENT\_RATE (Taxa de Pagamento do Aplicante):

Fórmula: abt\_00['APP\_NEW\_PAYMENT\_RATE'] = abt\_00['AMT\_ANNUITY'] / abt\_00['AMT\_CREDIT']

Explicação: Representa a taxa de pagamento mensal em relação ao montante total do crédito, auxiliando na avaliação do compromisso financeiro mensal do aplicante em relação ao crédito concedido.

**7. Metodologia**

No contexto deste projeto serão implementados dois modelos:

* Um modelo performático, que busca o alcance de boas métricas de separação das classes, porém, não traz explicabilidade acerca do cálculo do score;
* Um modelo de Regressão Logística que busca a elaboração de uma equação simples para o cálculo dos scores, de forma que os coeficientes desta equação tragam explicabilidade do Score atribuído.

A elaboração de um "modelo performático" será fundamental como parâmetro comparativo em relação ao modelo de Regressão Logística. O termo "modelo performático" refere-se à busca por uma abordagem que demonstre eficácia e eficiência na consecução dos objetivos estabelecidos. Nesse sentido, este modelo será concebido com a finalidade de fornecer resultados sólidos e alinhados às metas predefinidas, estabelecendo-se como um padrão de referência para a avaliação do desempenho.

Por outro lado, a escolha da Regressão Logística como técnica principal destaca-se pela sua capacidade de analisar a relação entre variáveis e prever probabilidades em situações em que a variável dependente é binária. Assim, a Regressão Logística será empregada como o modelo preditivo principal para avaliar o risco de inadimplência dos potenciais clientes.

Em ambos os modelos, todavia, o quesito principal para avaliação é a ordenação dos Scores atribuídos, ou seja, o modelo deve ser capaz de separar faixas de Scores e calcular uma taxa de evento que tenha ordenação ao longo das faixas.

**7.1 Modelo Performático**

O Modelo Performático desenvolvido adota a utilização do algoritmo XGBoost (eXtreme Gradient Boosting). Este algoritmo é reconhecido por sua capacidade de fornecer resultados precisos e rápidos, tornando-se uma escolha ideal para lidar com conjuntos de dados complexos.

Foi implementada a criação de novas features em cada conjunto de dados. Essa estratégia visa capturar nuances e relações latentes, contribuindo para uma compreensão mais abrangente do perfil de crédito dos indivíduos.

A fase de Feature Selection é conduzida utilizando o próprio algoritmo XGBoost, que avalia a importância de cada variável. Esse processo permite a identificação e destaque das variáveis mais essenciais, contribuindo para um modelo mais eficiente e interpretável.

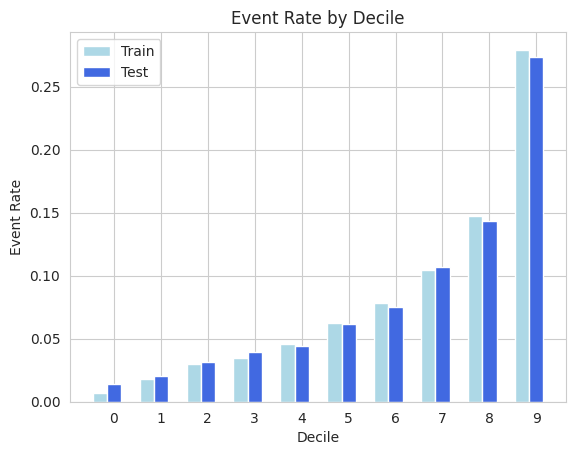
A preparação dos dados é realizada filtrando a base apenas com as features consideradas relevantes. Além disso, variáveis categóricas são codificadas para garantir uma representação numérica adequada, mantendo a integridade das informações.

O modelo performático é treinado e otimizado utilizando a base de dados **fe\_application\_01**, identificada como proporcionando os melhores resultados de desempenho durante o processo de validação.

A seguir, apresentamos as features selecionadas após a etapa de Feature Selection, destacando as variáveis que desempenham um papel crucial na avaliação de risco de crédito:

1. APP\_NEW\_PAYMENT\_RATE
2. EXT\_SOURCE\_3
3. REG\_CITY\_NOT\_LIVE\_CITY
4. OCCUPATION\_TYPE
5. FLOORSMAX\_MEDI
6. NAME\_EDUCATION\_TYPE
7. EXT\_SOURCE\_2
8. EXT\_SOURCE\_1
9. LIVINGAREA\_MEDI
10. NAME\_INCOME\_TYPE

O gráfico a seguir ilustra a ordenação das variáveis no modelo performático desenvolvido utilizando o algoritmo XGBoost. Essa representação visual destaca a importância relativa de cada variável na predição de risco de crédito.



**7.2 Regressão Logística**

Após a importação do conjunto de dados, conduzimos uma abrangente exploração inicial, analisando a volumetria, métricas descritivas como média, moda, mínimo e máximo das variáveis, bem como a distribuição do Target, que revelou uma proporção de aproximadamente 8% para o evento de inadimplência (domínio 1).

Realizamos o split dos dados pelo modo Out-of-Sample, adotando uma divisão de 70/30, onde 70% dos dados foram alocados para o treinamento do modelo, enquanto 30% foram reservados para testes e validação.

Para compreender a estrutura dos dados, empregamos uma função para examinar os metadados da base, fornecendo informações úteis como a quantidade e a porcentagem de

Optamos por separar os dados em dois DataFrames, um contendo exclusivamente as variáveis numéricas, sujeitas à seleção de variáveis, e outro contendo apenas as variáveis categóricas. Esta abordagem permitiu uma análise mais aprofundada das variáveis categóricas, considerando sua ordenação para futuras etapas do modelo.

Para a seleção de variáveis, adotamos o método de Feature Importance, utilizando o algoritmo GradientBoostingClassifier. Esse processo identificou as variáveis mais relevantes com um corte de 10% em relação à variável de maior importância. As variáveis selecionadas, como ['APP\_GOODS\_CREDIT\_RATIO', 'APP\_NEW\_PAYMENT\_RATE', 'EXT\_SOURCE\_1', 'EXT\_SOURCE\_2', 'EXT\_SOURCE\_3'], foram mantidas nos DataFrames de treino e teste.

Realizamos a junção das variáveis selecionadas com as variáveis categóricas, mantendo essa abordagem também para a base de teste.

Verificamos a correlação de Pearson entre as variáveis numéricas, não identificando alta correlação, evitando assim multicolinearidade e possíveis problemas de overfitting. Além disso, analisamos o Information Value (IV) das variáveis, excluindo aquelas consideradas inúteis para a predição, bem como as de baixo IV, indicando fraco poder preditivo. A variável 'CODE\_GENDER' foi removida devido à sua sensibilidade no contexto de modelagem de crédito.

A partir da análise anterior, e baseados no IV, decidimos excluir variáveis que são inúteis para predição e algumas com IV baixo que possuem fraco poder preditivo. Optamos também por remover a variável 'CODE\_GENDER' por se tratar de uma variável sensível para esse problema de modelagem de crédito. **As variáveis excluídas foram:** ['CODE\_GENDER', 'FLAG\_OWN\_REALTY', 'NAME\_TYPE\_SUITE', 'FLAG\_OWN\_CAR', 'WEEKDAY\_APPR\_PROCESS\_START', 'NAME\_HOUSING\_TYPE', 'WALLSMATERIAL\_MODE']. Excluímos as mesmas variáveis da base de teste também.

Para codificar as variáveis categóricas com cardinalidade superior a 20, empregamos a técnica de Label Encoding com o algoritmo LabelEncoder, transformando categorias em números e posteriormente em bins. Essa transformação foi aplicada também à base de teste, garantindo consistência.

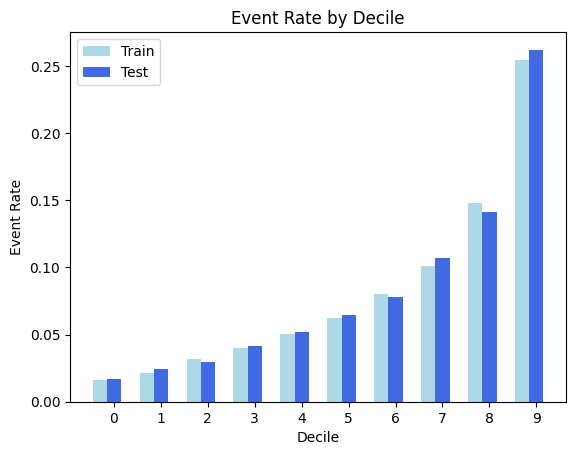
Dividimos as variáveis com cardinalidade menor ou igual a 20, avaliando a taxa de evento de cada uma para verificar a existência de ordenação. Realizamos agrupamentos das categorias quando necessário. As variáveis com cardinalidade maior que 20 foram divididas em 5 bins inicialmente. Essa transformação foi aplicada tanto na base de treino quanto na de teste.

Após garantir a ordenação das variáveis, procedemos com a dummificação, gerando variáveis dummy para cada categoria. Essa transformação foi salva em um arquivo .pkl para produção e aplicada na base de teste.

Verificamos novamente a correlação entre as variáveis após o processo de dummy, identificando algumas com alta correlação. Removemos as variáveis ['TFB\_EMERGENCYSTATE\_MODE\_Yes', 'TFB\_NAME\_EDUCATION\_TYPE\_Higher education', 'TFB\_NAME\_INCOME\_TYPE\_Student'] devido à correlação superior a 0.85 e inferior a -0.85. A transformação foi salva e aplicada na base de teste para manter a consistência entre os conjuntos de dados.

Essa metodologia detalhada visa assegurar a qualidade e robustez do modelo de Regressão Logística, garantindo a eficácia na avaliação de risco de crédito.

Para o modelo baseado em Regressão Logística, a figura abaixo apresenta a ordenação das variáveis conforme sua influência na modelagem de risco de crédito.



**8.Desenvolvimento do Modelo**

A etapa de desenvolvimento de modelos desempenha um papel central na eficácia da concessão de crédito, especialmente em cenários diversos, como financiamento imobiliário, veículos e empréstimo pessoal. Essa fase crítica envolve a criação de algoritmos preditivos capazes de avaliar o risco associado a cada transação, proporcionando uma base sólida para a tomada de decisões informadas. Neste processo, são adotadas diversas etapas, desde a definição de objetivos específicos até a implementação e monitoramento contínuo do modelo.

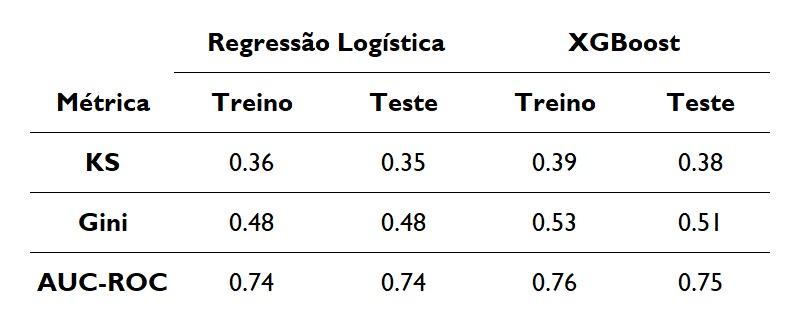
**8.1 Escolha do modelo**

A seleção dos modelos para o projeto de avaliação de risco de crédito foi cuidadosamente orientada pela natureza específica da tarefa e pelos objetivos almejados. Cada modelo escolhido oferece vantagens distintas, contribuindo de maneiras únicas para a eficácia do sistema. Abaixo, apresentamos um resumo das razões por trás da escolha de cada modelo:

O modelo XGBoost (eXtreme Gradient Boosting) foi escolhido devido à sua robustez e eficiência comprovadas em problemas de classificação complexos, como é o caso da avaliação de risco de crédito.

A Regressão Logística foi escolhida por sua interpretabilidade, simplicidade e capacidade de fornecer insights diretos sobre a relação entre variáveis independentes e a variável dependente.

Ambos os modelos foram treinados e avaliados de maneira rigorosa durante o desenvolvimento, e a escolha final foi baseada em uma análise comparativa não só do desempenho, considerando métricas relevantes para o contexto de avaliação de risco de crédito, mas também pela questão da explicabilidade do modelo e das variáveis usadas. A combinação desses dois modelos distintos busca explorar as forças complementares de cada um, proporcionando uma abordagem abrangente e confiável para a tarefa em questão.



**8.2 Treinamento do Modelo**

**8.2.1 Treinamento do Modelo Performático (XGBoost)**

Durante a etapa de treinamento do modelo performático, empregamos o algoritmo XGBoost, conhecido por sua eficácia em tarefas complexas de classificação, como a avaliação de risco de crédito. A abordagem incluiu a seleção de features relevantes e a otimização de hiperparâmetros para melhorar o desempenho do modelo.

**Metodologia de Treinamento**

O modelo foi treinado utilizando as features selecionadas, proporcionando uma entrada refinada e relevante para a tarefa de previsão de risco de crédito.

**Otimização de Hiperparâmetros**

Para encontrar a combinação ideal de hiperparâmetros, geramos um grid de parâmetros que abrange diferentes configurações. Utilizamos a função GridSearchCV do módulo sklearn para realizar uma busca automática e identificar a combinação que resulta nas melhores métricas de desempenho.

**Grid de Parâmetros Utilizado**

grid\_param = {

'objective': ['binary:logistic'],

'learning\_rate': [0.1, 0.00001],

'max\_depth': [3],

'min\_child\_weight': [4],

'subsample': [0.5, 1.0],

'colsample\_bytree': [0.5, 1],

'n\_estimators': [200],

'scale\_pos\_weight': [11.5],

'reg\_lambda': [0.5, 1.5, 3],

'reg\_alpha': [0.5, 1.5, 3]

}

**Métricas do Modelo Tunado (Dados de Treino / Dados de Teste)**

* Acurácia: 0.70 / 0.70
* Precision: 0.17 / 0.17
* Recall: 0.70 / 0.68
* AUC-ROC: 0.76 / 0.75
* Gini: 0.53 / 0.51
* KS: 0.39 / 0.38

**8.2.2 Treinamento do Modelo de Regressão Logística**

Durante a fase de treinamento do modelo de regressão logística, utilizamos o algoritmo Logit da biblioteca Statsmodels para criar um modelo robusto capaz de fornecer previsões significativas no contexto da avaliação de risco de crédito.

**Metodologia de Treinamento**

O treinamento foi conduzido utilizando um conjunto de dados específico, onde aplicamos o algoritmo Logit para ajustar os parâmetros do modelo às características do problema em questão.

**Geração do Scorecard**

A partir do modelo treinado, geramos um Scorecard que contém informações cruciais para a interpretação do impacto de cada variável na previsão de risco. Este Scorecard inclui o nome da variável, o Coeficiente Beta associado, o P-Valor e a Estatística Wald de cada variável.

**Significância Estatística**

Todas as variáveis incluídas no treinamento do modelo apresentam P-Valor inferior a 0.05, indicando que cada variável é estatisticamente significativa para a previsão do risco de crédito. Essa observação ressalta a confiabilidade estatística do modelo treinado.

**Arquivo .pkl para Replicação**

Para garantir a replicação consistente do modelo em ambientes de produção, salvamos um arquivo .pkl contendo o modelo treinado. Isso permite que o modelo seja facilmente reproduzido, mantendo a consistência dos resultados.

**Variáveis Utilizadas no Treinamento:** As variáveis selecionadas para o treinamento do modelo foram cuidadosamente escolhidas com base em sua relevância e impacto percebido na avaliação de risco de crédito. A lista completa das variáveis utilizadas é apresentada a seguir:

['NAME\_CONTRACT\_TYPE\_Revolving loans', 'TFB\_FONDKAPREMONT\_MODE\_org spec account', 'TFB\_HOUSETYPE\_MODE\_block of flats', 'TFB\_HOUSETYPE\_MODE\_terraced house', 'TFB\_NAME\_EDUCATION\_TYPE\_Lower secondary', 'TFB\_NAME\_FAMILY\_STATUS\_Married', 'TFB\_NAME\_FAMILY\_STATUS\_Widow', 'TFB\_NAME\_INCOME\_TYPE\_Unemployed', 'TFB\_OCCUPATION\_TYPE\_Low-skill Laborers', 'TFB\_OCCUPATION\_TYPE\_Waiters/barmen staff', 'TFT\_EXT\_SOURCE\_2\_1', 'TFT\_EXT\_SOURCE\_2\_2', 'TFT\_EXT\_SOURCE\_2\_3', 'TFT\_EXT\_SOURCE\_2\_4', 'TFT\_EXT\_SOURCE\_3\_1', 'TFT\_EXT\_SOURCE\_3\_2', 'TFT\_EXT\_SOURCE\_3\_3', 'TFT\_EXT\_SOURCE\_3\_4', 'TFB\_TFT\_ORGANIZATION\_TYPE\_3', 'TFB\_TFT\_ORGANIZATION\_TYPE\_4', 'TFT\_APP\_NEW\_PAYMENT\_RATE\_1', 'TFT\_APP\_NEW\_PAYMENT\_RATE\_2', 'TFT\_APP\_NEW\_PAYMENT\_RATE\_3', 'TFT\_APP\_NEW\_PAYMENT\_RATE\_4', 'TFT\_EXT\_SOURCE\_1\_1', 'TFT\_EXT\_SOURCE\_1\_2', 'TFT\_EXT\_SOURCE\_1\_3', 'TFT\_EXT\_SOURCE\_1\_4', 'TFB\_TFT\_APP\_GOODS\_CREDIT\_RATIO\_1', 'TFB\_TFT\_APP\_GOODS\_CREDIT\_RATIO\_3', 'TFB\_TFT\_APP\_GOODS\_CREDIT\_RATIO\_4']

**8.3 Teste e Implementação**

Durante a fase de avaliação do modelo, conduzimos testes utilizando nossa base de teste derivada da principal. Esses testes foram projetados para avaliar a performance e as métricas do modelo de risco de crédito, fornecendo insights sobre a capacidade preditiva em um cenário mais próximo do ambiente de produção.

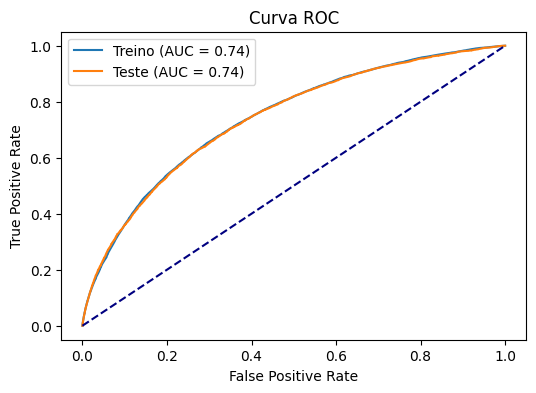
Metodologia de Testes:

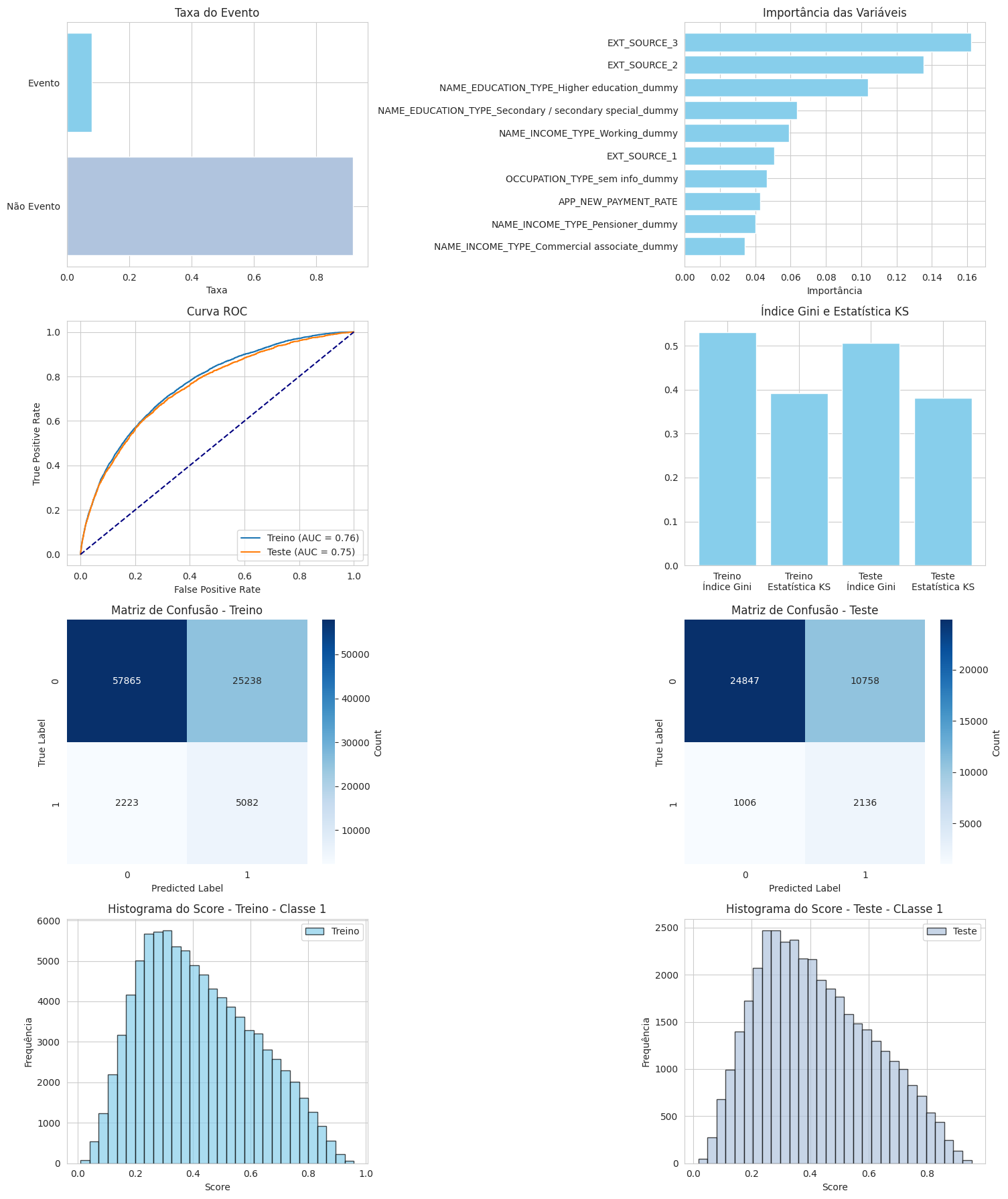
* Utilizamos uma base de teste derivada da base principal para simular condições do mundo real.
* A avaliação da performance foi realizada utilizando métricas específicas para problemas de classificação, como precisão, revocação, F1-score e a área sob a curva ROC

Resultados dos Testes:

* Os resultados obtidos durante os testes proporcionaram uma compreensão abrangente da eficácia do modelo na previsão de risco de crédito.
* Métricas como acurácia, precisão, revocação e a área sob a curva ROC foram examinadas para garantir uma visão completa do desempenho.

Os gráficos a seguir apresentam métricas cruciais de desempenho.





Dashboard com as métricas do modelo XGBoost.

**9. Conclusão**

A implementação do modelo de Regressão Logística para concessão de crédito na PoD Bank representa um avanço significativo na abordagem da empresa em relação à análise de riscos. Ao escolher estrategicamente a Regressão Logística, buscamos não apenas prever o risco de inadimplência, mas também garantir interpretabilidade e transparência nas decisões. O modelo se destaca por sua capacidade de analisar a relação entre variáveis e prever probabilidades em cenários binários.

O processo de desenvolvimento incluiu uma análise abrangente dos dados, Feature Engineering e uma metodologia rigorosa de treinamento. Complementamos esse modelo com o XGBoost, aproveitando a robustez desse algoritmo em problemas complexos de classificação. Durante o treinamento, métricas como Gini, KS e a área sob a curva ROC foram empregadas, evidenciando a eficácia do modelo de Regressão Logística na previsão de risco de crédito.

Além de proporcionar previsões precisas, o modelo oferece um Scorecard detalhado, contribuindo para a interpretação do impacto de cada variável. Essa transparência reforça a confiabilidade do modelo. Em síntese, o modelo de Regressão Logística é uma ferramenta valiosa, fundamentada e interpretável, aprimorando a eficácia na concessão de crédito e fortalecendo a posição da PoD Bank no mercado financeiro.

**10. Equipe de Desenvolvimento do Modelo**

Alexandre Alvarenga

[alexandresalvarenga@hotmail.com](mailto:alexandresalvarenga@hotmail.com)

Jackson Correa

[jackson.szcorrea@gmail.com](mailto:jackson.szcorrea@gmail.com)

José Santos

[santos.jcp@gmail.com](mailto:santos.jcp@gmail.com)

Suely Messias

[suelymesssias@gmail.com](mailto:suelymesssias@gmail.com)

Wagner Moraes

[wagnerjunior1@hotmail.com](mailto:wagnerjunior1@hotmail.com)

**Alexandre Alvarenga**

alexandresalvarenga@hotmail.c