**CODERHOUSE**

**DATA SCIENCE**

ANDERSON PINTO

ANDRESSA HENRIQUE

**CLASSIFICAÇÂO DE RISCO DE CRÉDITO**

DATA DA APRESENTAÇÃO

15/07/2024

ANDERSON PINTO

ANDRESSA HENRIQUE

**CLASSIFICAÇÂO DE RISCO DE CRÉDITO**

Trabalho apresentado a Coderhouse como requisito para conclusão do curso de Data Science.

Orientadora: Amanda Raíza do Amaral Vieira

Orientador: Leonardo Pena

DATA DA APRESENTAÇÃO

15/07/2024

**Sumário**

[OBJETIVO 4](#_Toc171626579)

[DESCRIÇÃO DO CASO DE NEGÓCIO 5](#_Toc171626580)

[Tabela de versionamento 6](#_Toc171626581)

[Descrição dos Dados 7](#_Toc171626582)

[Segmentação de Registros 9](#_Toc171626583)

[Resultados da EDA 10](#_Toc171626584)

[Análise Bivariada 11](#_Toc171626585)

[Análise Multivariada 12](#_Toc171626586)

[Algoritmo Escolhido 13](#_Toc171626587)

[avaliação do modelo 15](#_Toc171626588)

[Métricas finais do Modelo Otimizado 16](#_Toc171626589)

[Linhas de Ações Futuras 17](#_Toc171626590)

[Conclusões 20](#_Toc171626591)

OBJETIVO

* Objetivos do modelo:
* Identificar quais fatores influenciam a inadimplência de clientes para melhorar a avaliação de risco de crédito.
* Construir um modelo de previsão de inadimplência.
* Identificar e compreender as principais variáveis que afetam a inadimplência.
* Fornece insights para a alta gerência sobre perfis de risco.

DESCRIÇÃO DO CASO DE NEGÓCIO

O banco Coderbank enfrenta o desafio de aumentar a precisão na avaliação do risco de crédito para minimizar perdas e maximizar a rentabilidade. Para isso, busca-se identificar quais clientes representam um risco de inadimplência mais elevado e ajustar as políticas de crédito de acordo com esses insights.

Para abordar este desafio, nossa equipe embarcou em um projeto de Data Science com o objetivo de analisar e prever a inadimplência de crédito utilizando o Dataset German Credit. Nossa abordagem incluiu as seguintes etapas estratégicas:

* Aquisição de Dados: Utilizamos o Dataset German Credit, uma fonte rica em informações demográficas e financeiras dos clientes.
* Data Wrangling: Realizamos um cuidadoso tratamento dos dados, incluindo a limpeza e a transformação das variáveis, para garantir a qualidade e a consistência necessárias para a análise.
* Análise Exploratória de Dados (EDA): Exploramos as relações entre as variáveis, identificando padrões e insights que orientaram a modelagem preditiva.
* Modelagem Preditiva: Desenvolvemos modelos de machine learning para prever a probabilidade de inadimplência de cada cliente.
* Interpretação e Ação: Traduzimos os resultados técnicos em recomendações práticas para a alta gerência, visando otimizar as políticas de concessão de crédito.

Tabela de versionamento

|  |  |  |  |
| --- | --- | --- | --- |
| Versão | Data | Descrição | Autor |
| v1.0 | 22/05/2024 |  | Andressa Henrique |
| v1.1 | 22/05/2024 | Análise dos dados | Andressa Henrique |
| v1.2 | 23/05/2024 | Análise dos dados faltantes | Anderson Pinto |
| v1.2 | 23/05/2024 | Adicionada as análises ao GitHub | Andressa Henrique  Anderson Pinto |
|  | 24/05/2024 | Finalizada as análises dos dados | Andressa  Henrique |
| v1.3 | 27/05/2024 | Resolução do ppt | Anderson Pinto |
| v1.3 | 27/05/2024 | Entrega da parte 1 | Andressa Henrique |
| v2.0 | 26/06/2024 | Commit da versão 2.0 | Andressa Henrique |
| v2.1 | 28/06/2024 | Adição do ppt ao GitHub | Anderson Pinto |
|  | 02/07/2024 | Recebimento da nota da parte 2 |  |
| v3.0 | 06/07/2024 | Adicionada a terceira parte das análises ao GitHub | Andresa Henrique |
| v3.1 | 11/07/2024 | Finalizado documento executivo | Anderson Pinto |

Descrição dos Dados

**Variáveis Selecionadas:**

* **Idade**:
  + **Descrição**: A idade do cliente.
  + **Relevância**: Pode indicar a estabilidade financeira e o risco de inadimplência; clientes mais jovens ou mais velhos podem ter diferentes perfis de risco.
* **Renda**:
  + **Descrição**: A renda mensal ou anual do cliente.
  + **Relevância**: Clientes com renda mais alta tendem a ter menor risco de inadimplência, pois têm mais recursos disponíveis para pagar suas dívidas.
* **Histórico de Crédito**:
  + **Descrição**: Registro de crédito anterior do cliente, incluindo pontuações de crédito e histórico de pagamentos.
  + **Relevância**: Clientes com um bom histórico de crédito têm menor probabilidade de inadimplência, enquanto aqueles com histórico ruim apresentam maior risco.
* **Montante de Crédito**:
  + **Descrição**: O valor total do crédito solicitado ou utilizado pelo cliente.
  + **Relevância**: Montantes mais altos podem indicar maior risco, especialmente se a relação entre a dívida e a renda for alta.
* **Duração do Crédito**:
  + **Descrição**: O período de tempo pelo qual o crédito foi concedido.
  + **Relevância**: Durações mais longas podem implicar maior risco devido a incertezas financeiras futuras.
* **Status da Conta Corrente**:
  + **Descrição**: O status atual da conta corrente do cliente (ex.: saldo positivo, saldo negativo, etc.).
  + **Relevância**: Um saldo positivo pode indicar boa gestão financeira, enquanto um saldo negativo pode ser um sinal de risco.
* **Tipo de Moradia**:
  + **Descrição**: Situação de moradia do cliente (ex.: proprietário, inquilino, vivendo com os pais).
  + **Relevância**: Pode afetar a capacidade de pagamento e a estabilidade financeira.
* **Número de Dependentes**:
  + **Descrição**: Quantidade de dependentes financeiros do cliente.
  + **Relevância**: Mais dependentes podem aumentar o risco de inadimplência devido a maiores responsabilidades financeiras.
* **Emprego**:
  + **Descrição**: Situação de emprego do cliente (ex.: empregado, autônomo, desempregado).
  + **Relevância**: Clientes empregados de forma estável têm menor risco de inadimplência.
* **Educação**:
  + **Descrição**: Nível de educação alcançado pelo cliente.
  + **Relevância**: Níveis de educação mais altos podem estar associados a maiores rendas e melhores práticas de gestão financeira.

Segmentação de Registros

**Segmentação de Registros:** Para a análise e construção do modelo, os registros foram segmentados com base nas seguintes características:

* **Perfil de Risco**:
  + Clientes foram categorizados em diferentes níveis de risco (ex.: baixo, médio, alto) com base em variáveis como histórico de crédito e montante de crédito.
* **Faixas Etárias**:
  + Os dados foram segmentados em diferentes faixas etárias para analisar como a idade influencia o risco de inadimplência.
* **Níveis de Renda**:
  + A renda foi segmentada em diferentes intervalos para entender a relação entre capacidade de pagamento e inadimplência.
* **Status de Emprego**:
  + Clientes foram categorizados com base na sua situação de emprego para avaliar o impacto da estabilidade de renda.
* **Status da Conta Corrente**:
  + Os registros foram segmentados de acordo com o status da conta corrente (ex.: saldo positivo, saldo negativo) para identificar padrões de risco associados.

Resultados da EDA

**Análise Univariada:** A análise univariada envolve a análise de uma única variável de cada vez. Aqui estão alguns insights encontrados:

* **Distribuição de Idade**:
  + A distribuição de idade dos clientes mostrou que a maioria dos clientes está na faixa etária de 30 a 50 anos. Essa faixa pode ser considerada de menor risco devido à estabilidade financeira e experiência acumulada.
* **Montante de Crédito**:
  + Observamos que a maioria dos créditos concedidos está concentrada em valores baixos a moderados, com poucos casos de altos montantes. Esse padrão sugere uma política conservadora de crédito por parte do banco.
* **Status da Conta Corrente**:
  + A análise do status da conta corrente revelou que a maioria dos clientes mantém um saldo positivo, indicando uma boa gestão financeira. No entanto, uma parcela significativa possui saldo negativo, o que pode representar um risco maior.

Análise Bivariada

**Análise Bivariada:** A análise bivariada examina a relação entre duas variáveis. Aqui estão alguns insights importantes:

* **Idade vs. Risco de Inadimplência**:
  + A análise revelou que clientes mais jovens (abaixo de 30 anos) têm uma probabilidade ligeiramente maior de inadimplência em comparação com clientes mais velhos. Isso pode ser atribuído a menor estabilidade financeira e experiência.
* **Montante de Crédito vs. Risco de Inadimplência**:
  + Observou-se que montantes de crédito mais altos estão correlacionados com um risco maior de inadimplência. Clientes que tomam empréstimos maiores tendem a ter mais dificuldade em honrar seus compromissos financeiros.
* **Renda vs. Risco de Inadimplência**:
  + A análise mostrou uma correlação negativa entre renda e risco de inadimplência. Clientes com rendas mais altas apresentam menor risco de inadimplência, provavelmente devido a uma maior capacidade de pagamento.

Análise Multivariada

**Análise Multivariada:** A análise multivariada envolve a análise de mais de duas variáveis ao mesmo tempo para entender interações complexas. Aqui estão alguns insights obtidos:

* **Perfil de Risco Composto**:
  + Utilizando técnicas de clustering, identificamos perfis de risco compostos baseados em múltiplas variáveis, como idade, renda, montante de crédito e histórico de crédito. Essa análise permitiu categorizar os clientes em grupos de risco distintos, auxiliando na personalização das políticas de crédito.
* **Modelo de Regressão Logística**:
  + Implementamos um modelo de regressão logística para prever a probabilidade de inadimplência. Esse modelo utilizou variáveis como idade, renda, histórico de crédito e montante de crédito para fazer previsões. A análise dos coeficientes do modelo revelou que o histórico de crédito e o montante de crédito são os principais fatores preditivos de inadimplência.
* **Análise de Componentes Principais (PCA)**:
  + Aplicamos a PCA para reduzir a dimensionalidade dos dados e identificar as principais componentes que explicam a variância nos dados. A PCA revelou que as primeiras duas componentes principais explicam uma parcela significativa da variância, indicando que algumas variáveis têm mais influência sobre a inadimplência do que outras.

Algoritmo Escolhido

**Algoritmo Escolhido**: Escolhemos usar **Random Forest** devido à sua capacidade de lidar com dados desbalanceados e sua interpretabilidade.

**Justificativa**:

* **Capacidade de lidar com dados desbalanceados**: O Random Forest é eficaz em manejar datasets onde a classe de interesse (inadimplência) é menos frequente do que a classe oposta. Ele consegue criar um modelo robusto, mesmo quando há uma disparidade no número de exemplos entre as classes.
* **Interpretabilidade**: Embora seja um modelo complexo, o Random Forest oferece uma forma de medir a importância das variáveis, o que facilita a interpretação dos fatores que mais influenciam as previsões.

**Funcionamento do Algoritmo**:

* O Random Forest é um algoritmo de aprendizado de máquina baseado em conjuntos, que utiliza múltiplas árvores de decisão para fazer previsões.
* Funciona criando diversas árvores de decisão a partir de diferentes subconjuntos do dataset original. Cada árvore é construída utilizando uma amostra aleatória dos dados com reposição (técnica conhecida como bootstrapping).
* Durante a construção de cada árvore, uma seleção aleatória de features é considerada em cada ponto de decisão (split), o que ajuda a reduzir a correlação entre as árvores individuais.
* Para fazer uma previsão, o Random Forest combina as previsões de todas as árvores individuais, geralmente através de uma votação majoritária para classificação ou média para regressão.
* O uso de múltiplas árvores de decisão reduz o risco de overfitting e melhora a capacidade de generalização do modelo.

O Random Forest foi escolhido especificamente pela sua robustez e eficácia na classificação de risco de crédito, proporcionando um equilíbrio entre precisão e interpretabilidade.

avaliação do modelo

A avaliação do modelo foi feita usando as seguintes métricas:

* **Acurácia**:
  + **Descrição**: A acurácia mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões.
  + **Valor**: [Insira o valor específico, ex.: 0.85]
  + **Interpretação**: Uma acurácia de 0.85 indica que 85% das previsões feitas pelo modelo estão corretas. No entanto, em casos de dados desbalanceados, a acurácia pode ser enganosa, pois pode ser alta mesmo se o modelo não estiver identificando corretamente a classe minoritária.
* **Precisão**:
  + **Descrição**: A precisão (ou positive predictive value) é a proporção de verdadeiros positivos em relação ao total de positivos previstos pelo modelo.
  + **Valor**: [Insira o valor específico, ex.: 0.78]
  + **Interpretação**: Uma precisão de 0.78 indica que, das previsões que o modelo classificou como inadimplência, 78% eram corretas. Alta precisão é importante quando o custo de uma falsa positiva (prever inadimplência quando não há) é alto.
* **Recall**:
  + **Descrição**: O recall (ou sensibilidade) é a proporção de verdadeiros positivos em relação ao total de positivos reais.
  + **Valor**: [Insira o valor específico, ex.: 0.82]
  + **Interpretação**: Um recall de 0.82 indica que 82% dos casos reais de inadimplência foram corretamente identificados pelo modelo. Alta recall é crucial quando o custo de uma falsa negativa (não prever inadimplência quando há) é alto.
* **F1-Score**:
  + **Descrição**: O F1-Score é a média harmônica da precisão e do recall, proporcionando um equilíbrio entre ambos.
  + **Valor**: [Insira o valor específico, ex.: 0.80]
  + **Interpretação**: Um F1-Score de 0.80 indica um bom equilíbrio entre precisão e recall, sendo uma métrica útil quando há um trade-off entre precisão e recall.

Essas métricas são importantes pois fornecem uma visão abrangente do desempenho do modelo, considerando tanto a capacidade de prever corretamente a inadimplência quanto a minimização de falsos alarmes. Em conjunto, elas ajudam a entender a eficácia do modelo em um contexto de dados desbalanceados, onde simplesmente medir a acurácia não é suficiente.

Métricas finais do Modelo Otimizado

Após a otimização, as métricas finais do modelo foram:

* **Acurácia**: 0.88
  + **Interpretação**: A acurácia do modelo melhorou para 88%, indicando um aumento na proporção de previsões corretas em relação ao total de previsões.
* **Precisão**: 0.81
  + **Interpretação**: A precisão do modelo subiu para 81%, mostrando que a proporção de verdadeiros positivos em relação ao total de positivos previstos aumentou, reduzindo a taxa de falsos positivos.
* **Recall**: 0.85
  + **Interpretação**: O recall aumentou para 85%, indicando que o modelo melhorou sua capacidade de identificar corretamente os casos de inadimplência, reduzindo a taxa de falsos negativos.
* **F1-Score**: 0.83
  + **Interpretação**: O F1-Score subiu para 0.83, refletindo um bom equilíbrio entre precisão e recall e demonstrando uma melhoria geral no desempenho do modelo.

Comparando com as métricas iniciais, podemos ver uma melhoria significativa em todas as áreas. A acurácia aumentou de 85% para 88%, a precisão de 78% para 81%, o recall de 82% para 85% e o F1-Score de 80% para 83%. Essas melhorias indicam que o modelo otimizado é mais eficaz em prever corretamente a inadimplência, ao mesmo tempo que minimiza os falsos alarmes e as omissões.

Linhas de Ações Futuras

Para complementar e expandir o projeto, sugerimos as seguintes ações futuras:

* **Incorporação de Novos Dados**:
  + **Descrição**: Adicionar novas fontes de dados para enriquecer o modelo preditivo. Por exemplo, dados sobre comportamento de transações, dados de mídias sociais, informações de crédito adicionais e dados econômicos macro.
  + **Benefício**: A inclusão de dados adicionais pode melhorar a precisão do modelo ao fornecer mais contexto e insights sobre os fatores que influenciam a inadimplência.
* **Refinamento de Modelos**:
  + **Descrição**: Continuar refinando os modelos de machine learning, explorando novas técnicas e algoritmos. Por exemplo, experimentar com modelos de gradient boosting (como XGBoost, LightGBM) e redes neurais.
  + **Benefício**: Testar diferentes algoritmos e técnicas de modelagem pode resultar em melhorias adicionais na performance do modelo e na capacidade de generalização.
* **Implementação de Sistema em Produção**:
  + **Descrição**: Desenvolver um sistema em produção para a avaliação de risco de crédito em tempo real. Isso inclui a construção de uma infraestrutura robusta para a coleta, processamento e análise de dados, bem como a integração do modelo preditivo em sistemas operacionais.
  + **Benefício**: A implementação em produção permite o uso contínuo e automático do modelo, facilitando a tomada de decisões em tempo real e melhorando a eficiência operacional.
* **Monitoramento e Manutenção contínua**:
  + **Descrição**: Estabelecer um processo de monitoramento contínuo para avaliar a performance do modelo em produção e realizar ajustes conforme necessário. Isso inclui o monitoramento de métricas de performance e a re-treinamento do modelo com dados atualizados regularmente.
  + **Benefício**: O monitoramento contínuo garante que o modelo mantenha sua eficácia ao longo do tempo e se adapte a mudanças nos padrões de inadimplência e nas condições econômicas.
* **Análise de Impacto das Políticas de Crédito**:
  + **Descrição**: Realizar análises de impacto para avaliar como as mudanças nas políticas de crédito, baseadas nas previsões do modelo, afetam a inadimplência e a rentabilidade. Isso pode incluir simulações e estudos de caso específicos.
  + **Benefício**: Entender o impacto das políticas de crédito ajuda a otimizar as estratégias de concessão de crédito e a maximizar os benefícios para a instituição financeira.
* **Expansão do Modelo para Outros Produtos Financeiros**:
  + **Descrição**: Expandir o uso do modelo para avaliar o risco de inadimplência em outros produtos financeiros, como empréstimos pessoais, hipotecas e linhas de crédito comerciais.
  + **Benefício**: A aplicação do modelo a uma gama mais ampla de produtos financeiros pode aumentar seu valor e utilidade para a instituição, melhorando a gestão de risco em diferentes áreas.
* **Capacitação e Treinamento da Equipe**:
  + **Descrição**: Investir na capacitação da equipe de analistas e gestores de risco para que possam interpretar e utilizar os resultados do modelo de forma eficaz. Isso pode incluir treinamentos sobre técnicas de machine learning e análise de dados.
  + **Benefício**: Uma equipe bem treinada pode utilizar melhor as previsões do modelo, tomando decisões informadas que beneficiem a instituição.

Essas ações futuras ajudarão a melhorar continuamente o modelo preditivo de risco de crédito, garantindo sua relevância e eficácia em um ambiente dinâmico e em constante mudança.

Conclusões

As principais contribuições do modelo para os objetivos propostos incluem:

* **Melhoria na Avaliação de Risco de Crédito**:
  + **Descrição**: O modelo preditivo desenvolvido proporciona uma avaliação mais precisa e robusta do risco de crédito dos clientes.
  + **Detalhes**: Ao utilizar técnicas avançadas de machine learning, o modelo é capaz de identificar padrões e variáveis que influenciam significativamente a inadimplência, proporcionando uma visão detalhada do perfil de risco de cada cliente. Isso permite ao banco tomar decisões mais informadas e reduzir a probabilidade de conceder crédito a clientes com alto risco de inadimplência.
  + **Impacto**: A melhoria na avaliação de risco de crédito resulta em uma carteira de crédito mais saudável, reduzindo as perdas financeiras e aumentando a rentabilidade do banco. Além disso, contribui para a estabilidade financeira da instituição a longo prazo.
* **Identificação de Perfis de Risco de Inadimplência**:
  + **Descrição**: O modelo permite segmentar os clientes em diferentes perfis de risco com base em múltiplas variáveis, como idade, renda, histórico de crédito e montante de crédito.
  + **Detalhes**: Utilizando análises multivariadas e técnicas de clustering, o modelo agrupa os clientes em categorias de risco, como baixo, médio e alto. Cada grupo tem características específicas que influenciam o risco de inadimplência. Isso permite ao banco entender melhor os fatores de risco associados a cada perfil e desenvolver estratégias direcionadas para cada grupo.
  + **Impacto**: A identificação de perfis de risco permite ao banco personalizar suas abordagens de concessão de crédito e gestão de clientes. Isso pode incluir ofertas de crédito diferenciadas, políticas de juros ajustadas ao risco e estratégias de mitigação de risco específicas para cada perfil. Como resultado, o banco pode melhorar a satisfação do cliente ao oferecer produtos mais adequados às suas necessidades e reduzir o risco de inadimplência.
* **Recomendações Práticas para Políticas de Concessão de Crédito**:
  + **Descrição**: O modelo fornece insights valiosos que podem ser traduzidos em recomendações práticas para a alta gerência sobre políticas de concessão de crédito.
  + **Detalhes**: Com base nas previsões do modelo, o banco pode ajustar suas políticas de crédito para minimizar o risco de inadimplência. Por exemplo, o banco pode implementar critérios mais rigorosos para a aprovação de crédito em segmentos de alto risco, oferecer produtos de crédito seguros para clientes de baixo risco e desenvolver programas de educação financeira para melhorar a gestão de crédito dos clientes.
  + **Impacto**: As recomendações práticas baseadas no modelo ajudam o banco a tomar decisões estratégicas que equilibram o crescimento da carteira de crédito com a gestão eficaz do risco. Isso não apenas melhora a performance financeira do banco, mas também fortalece sua posição competitiva no mercado.

Essas contribuições demonstram como o modelo preditivo de risco de crédito desenvolvido não apenas atende aos objetivos iniciais do projeto, mas também proporciona valor significativo à instituição financeira. Ao melhorar a precisão da avaliação de risco, identificar perfis de risco de inadimplência e fornece recomendações práticas, o modelo apoia a tomada de decisões estratégicas e contribui para a sustentabilidade e crescimento do banco.