**Contexto e Objetivo**

O objetivo central deste projeto foi construir um modelo preditivo que pudesse identificar a probabilidade de inadimplência entre os clientes do A.Cash, um banco digital internacional. A análise se concentrou em permitir que o banco avaliasse melhor o risco de crédito, diminuindo perdas financeiras por inadimplência e otimizando a concessão de empréstimos. Para esse propósito, foi utilizado um conjunto de dados detalhado sobre as características dos clientes, suas intenções de empréstimo e histórico de crédito.

**Descrição dos Dados**

O dataset contém 34.501 registros com 12 colunas relacionadas a dados demográficos, informações de empréstimo e histórico de crédito. Observamos que algumas colunas possuem valores ausentes, especialmente em taxa\_juros\_emprestimo, tempo\_emprego\_pessoa e inadimplente\_pessoa. Esse tipo de lacuna requer estratégias específicas de imputação, dependendo da natureza das variáveis e da quantidade de valores ausentes.

A estrutura do notebook revela as seguintes etapas iniciais:

1. **Leitura e renomeação de colunas**: O notebook foi ajustado para nomes em português, facilitando o entendimento das variáveis para análises subsequentes.
2. **Exploração dos dados**: Inclui comandos para avaliar a quantidade de entradas, tipos de dados e a presença de valores ausentes.

**Modelagem de Regressão**

A regressão linear foi utilizada exclusivamente como uma etapa de pré-processamento para preencher valores ausentes na variável **taxa\_juros\_emprestimo**. No entanto, o foco da modelagem está nos modelos de classificação para prever a inadimplência, baseando-se em características financeiras e histórico de crédito dos clientes, sendo uma análise de inadimplência com amostras balanceadas (oversampling e undersampling) para o modelo de classificação da variável **inadimplente\_pessoa**. Vamos detalhar os principais passos seguidos:

1. **Pré-processamento:**

**Remoção de valores nulos** para variáveis preditoras específicas, focando inicialmente na coluna **grau\_emprestimo**.

**Codificação ordinal** aplicada à coluna **grau\_emprestimo**, atribuindo valores numéricos para facilitar o modelo.

**Normalização de Dados Numéricos:**As colunas numéricas foram escaladas para o intervalo [0, 1] utilizando o **MinMaxScaler**.

Este processo garantiu uniformidade na escala das variáveis, prevenindo que algumas características dominassem outras no modelo de regressão ou influenciassem negativamente os modelos de classificação. Além disso, a normalização reduziu a influência de valores extremos e preparou os dados para uso em modelos subsequentes que requerem entradas normalizadas.

1. **Modelagem de Regressão:**

**Modelo:** Uma regressão linear foi ajustada para prever a **taxa\_juros\_emprestimo** com base nas variáveis **valor\_emprestimo** e **grau\_emprestimo\_encoded**.

**Avaliação:** O modelo apresentou:

**Erro Médio Quadrático (MSE):** Aproximadamente **1.84**, indicando um erro médio relativamente baixo.

**Coeficiente de Determinação (R²):** Cerca de **0.82**, sugerindo que o modelo explica **82%** da variabilidade da taxa de juros.

**Limitações:** A robustez deste modelo pode ser limitada pela presença de outliers e pela distribuição assimétrica da variável dependente.

1. **Balanceamento de Classes:**

Para a variável de inadimplência (**inadimplente\_pessoa**), foi aplicado o **SMOTE** (oversampling) para gerar amostras adicionais de inadimplentes e o **ClusterCentroids** (undersampling) para reduzir o desequilíbrio entre as classes.

A inclusão da normalização no pipeline fortaleceu o pré-processamento ao criar condições favoráveis para a modelagem, contribuindo para a qualidade e desempenho tanto do modelo de regressão quanto dos modelos de classificação subsequentes.

**Avaliação Final dos Modelos de Classificação**

**Análise dos Resultados:**

No contexto de prever a inadimplência entre os clientes do A.Cash, é crucial equilibrar corretamente as métricas de **precisão** (quantos dos identificados como inadimplentes realmente o são) e **recall** (quantos dos inadimplentes foram identificados pelo modelo). Uma alta taxa de recall é importante para capturar a maioria dos clientes que realmente serão inadimplentes, reduzindo perdas financeiras. Entretanto, uma precisão baixa pode levar a muitos falsos positivos, afetando a relação com clientes bons pagadores.

Regressão Logística com SMOTE

F1-score médio: 0.9040505909391824

Precisão média: 0.8255135786169981

Recall: 0.9991178628524647

Acurácia média: 0.8939515163806361

Regressão Logística com NearMiss

F1-score médio: 0.8489812322344822

Precisão média: 0.7559165129659371

Recall médio: 0.9682189626178681

Acurácia média: 0.8277697626768019

Random Forest com SMOTE

F1-score médio: 0.9099152116515423

Precisão média: 0.8465842102675423

Recall médio: 0.9835077770189485

Acurácia média: 0.9026195006971764

Random Forest com NearMiss

F1-score médio: 0.8538504151408031

Precisão média: 0.7748255747439741

Recall médio: 0.9509937730594593

Acurácia média: 0.8371797540126092

XGBoost com SMOTE

F1-score médio : 0.9007935891387391

Precisão média: 0.8696013016180133

Recall médio: 0.9343381728004955

Acurácia média: 0.8970966146743808

XGBoost com NearMiss

F1-score médio : 0.8995715992380735

Precisão média: 0.8560773976085352

Recall médio: 0.9477996094568804

Acurácia média: 0.8941760305383231

**Recomendação:**

* **Modelo Sugerido:** **Random Forest com SMOTE**

**Justificativa:**

* **Equilíbrio Ideal:** O Random Forest com SMOTE oferece um excelente equilíbrio entre precisão (0,8465) e recall (0,9835), garantindo que a maioria dos inadimplentes seja identificada sem gerar um número excessivo de falsos positivos.
* **Desempenho Superior:** Apresenta o maior F1-score médio (0,9099) e a maior acurácia média (0,9026), indicando um desempenho geral robusto.
* **Redução de Riscos:** Com um recall alto, o modelo minimiza o risco de perdas financeiras ao identificar a maioria dos clientes que potencialmente seriam inadimplentes.
* **Manutenção de Relacionamentos:** A precisão relativamente alta ajuda a evitar a recusa de crédito a clientes que são bons pagadores, mantendo um relacionamento positivo com a base de clientes.

**Considerações Finais:**

* **Importância do SMOTE:** Os modelos que utilizaram SMOTE superaram consistentemente aqueles que utilizaram NearMiss, indicando que o balanceamento da classe minoritária através de oversampling foi mais eficaz neste caso.
* **Avaliação Contínua:** É recomendado monitorar continuamente o desempenho do modelo após a implementação, ajustando conforme necessário para responder a mudanças nos dados ou no comportamento dos clientes.
* **Integração com Políticas Internas:** O modelo deve ser integrado com as políticas de risco e compliance do banco, garantindo que as decisões automatizadas estejam alinhadas com os regulamentos vigentes.

#### **Por que realizar Train-Test Split inicialmente?**

O **Train-Test Split** foi utilizado para dividir o conjunto de dados em dois subconjuntos:

**Conjunto de Treino:** Usado para ajustar (treinar) o modelo.

**Conjunto de Teste:** Reservado para avaliar o desempenho do modelo em dados que ele nunca viu antes.

Essa separação inicial é essencial para simular como o modelo se comportará em um cenário real, permitindo uma avaliação de generalização antes de qualquer ajuste mais refinado.

#### **Por que complementar com Validação Cruzada?**

Após realizar o Train-Test Split, foi empregada a **validação cruzada** para garantir maior confiança no desempenho do modelo:

Isso ajuda a reduzir o risco de que os resultados sejam influenciados por uma única divisão dos dados.

Fornece métricas mais robustas e estáveis, especialmente em conjuntos de dados com distribuição complexa ou limitada.

**Conclusão:**

Implementar o **Random Forest com SMOTE** permitirá ao A.Cash identificar eficazmente clientes com alto risco de inadimplência, reduzindo perdas financeiras e otimizando a concessão de empréstimos. Este modelo equilibra a necessidade de minimizar riscos com a manutenção de relacionamentos positivos com clientes confiáveis, alinhando-se perfeitamente ao objetivo central do projeto.

**Avaliação do Relatório com Base em Critérios Específicos**

A análise deste relatório se concentrará em cinco pontos principais para garantir que ele cumpra com os requisitos estratégicos:

1. **Descrição dos Dados**: O relatório inclui uma seção que detalha claramente os dados e suas características, assim como as etapas iniciais de tratamento de valores ausentes e preparação das variáveis.
2. **Objetivo da Análise**: A seção inicial apresenta o objetivo principal de maneira direta, enfatizando a necessidade de reduzir inadimplência e melhorar o risco de crédito.
3. **Variação de Modelos**: O relatório compara variações de classificadores, como SMOTE e ClusterCentroids, e especifica que o modelo Random Forest com SMOTE é o que melhor se adapta ao objetivo de identificar inadimplentes.
4. **Descobertas Principais**: As métricas do modelo, incluindo recall e precisão, são destacadas com explicações sobre o impacto desses resultados para o objetivo principal.
5. **Falhas do Modelo e Plano de Ação**: Possíveis limitações são identificadas, como a precisão moderada e falsos positivos. Ações sugeridas incluem ajustes de threshold, regras de negócio adicionais e exploração de dados complementares.

Este relatório atende aos requisitos principais e fornece uma base sólida para tomada de decisão estratégica no contexto de modelagem de crédito e mitigação de riscos no Alura Cash.