Previsão de Inadimplência Utilizando Técnicas de Credit Scoring e Machine Learning

Resumo

Este projeto é fruto da aplicação de conhecimento e pesquisa em Ciência de Dados, abordando o tema Gestão de Risco de Crédito e Credit Scoring.

O objetivo geral era construir um modelo de machine learning capaz de classificar clientes de forma binária, baseando-se em treinamento com dados de cartão de crédito.

Utilizamos diversas técnicas de preparação e limpeza da base de dados, análises univariáveis e bivariáveis, assim como aplicações de testes de hipótese para validar suposições pertinentes.

Aplicamos diversas técnicas de pré-processamento dos dados, com desafios referentes a balanceamento de classes, tratamento de variáveis categóricas, redução de dimensionalidade e programação de computadores.

Por fim, construímos modelos com dois algoritmos comumente usados em mercado financeiro, sendo eles Random Forest e Regressão Logística. Avaliamos as métricas e definimos pelo melhor desempenho.

Finalizamos sugerindo algumas oportunidades de continuação do estudo e aplicação de outras técnicas possíveis para aumento de desempenho.

Introdução

A análise de riscos em cartões de crédito é um campo fundamental dentro do setor financeiro, que busca garantir a segurança das transações e a proteção tanto dos consumidores quanto das instituições financeiras.

A crescente popularidade do uso de cartões de crédito como meio de pagamento tem gerado um ambiente complexo e desafiador, no qual a detecção precoce de possíveis ameaças é essencial para evitar prejuízos financeiros significativos.

Um dos principais pilares da análise de riscos em cartões de crédito é a avaliação do histórico de pagamento do titular do cartão. Esse aspecto inclui a análise de como o cliente tem gerenciado suas dívidas e honrado seus compromissos financeiros no passado. A análise do histórico de pagamento é frequentemente realizada por meio de sistemas de pontuação de crédito, como o FICO Score nos EUA, que avaliam o risco de inadimplência com base em vários fatores, incluindo histórico de pagamento, dívidas atuais, duração do histórico de crédito e outros.

Outro ponto crítico na análise de riscos em cartões de crédito é a detecção de atividades suspeitas e a prevenção de fraudes. Isso envolve o monitoramento constante de transações em busca de padrões incomuns que possam indicar uma fraude em andamento. Técnicas avançadas, como aprendizado de máquina e anáise de Big Data, são frequentemente utilizadas para identificar comportamentos fraudulentos e tomar medidas rápidas para proteger os consumidores e as instituições financeiras.

A análise de riscos em cartões de crédito é uma discplina dinâmica e em constante evolução, essencial para garantir a estabilidade e a segurança do sistema financeiro. Ela desempenha um papel crucial na concessão responsável de crédito e na prevenção de fraudes, contribuindo para a confiança dos consumidores e a saúde financeira das instituições.

Nosso objetivo específico é entregar um modelo classificador com acurácia mínima de 75% e Recall mínimo de 70%.

Metodologia

Coleta de Dados

Os dados utilizados neste projeto são públicos, utilizados em competições de Data Science e disponibilizados na plataforma Kaggle.

Link de acesso: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

Os dados estão distribuídos em 2 datasets, sendo um relacionado aos dados pessoais dos clientes e o segundo ao eventos relacionados a inadimplência.

Encontramos aqui o primeiro desafio, construir nossa variável resposta e unificar os dados em um único dataset. Para este processo, utilizamos uma técnica conhecida como Vintage Analysis, que utiliza as ocorrências de inadimplência em janelas de tempo para determinar qual a janela é mais representativa, baseando-se na definição de negócio da instituição financeira para o produto em análise.

Para facilitar a compreensão das variáveis da base de dados que utilizamos, segue um dicionário de dados:

Dataset: application_record.csv

- ID = Identificador do Cliente
- CODE_GENDER = Gênero do Cliente
- FLAG_OWN_CAR = Refere-se a propriedade de carro pelo cliente
- FLAG_OWN_REALTY = Refere-se a propriedade residencial pelo cliente
- CNT_CHILDREN = Número de Filhos do Cliente
- AMT_INCOME_TOTAL = Receita Anual do Cliente
- NAME_INCOME_TYPE = Origem da Receita do Cliente
- NAME_EDUCATION_TYPE = Grau de Formação do Cliente
- NAME_FAMILY_STATUS = Estado Civil do Cliente
- NAME_HOUSING_TYPE = Tipo de Moradia
- DAYS_BIRTH = Data do Aniversário relativa ao dia atual (hoje = 0)
- DAYS_EMPLOYED = Total de dias Desempregado (valor positivo) e Empregado (Valor Negativo)
- FLAG_MOBILE = Refere-se a propriedade de Celular do Cliente
- FLAG_WORK_PHONE = Refere-se a propriedade de telefone corporativo do Cliente
- FLAG_PHONE = Refere-se a propriedade de telefone fixo do cliente
- FLAG_EMAIL = Refere-se a propriedade de Email do Cliente
- OCCUPATION_TYPE = Tipo de Ocupação do Cliente
- CNT_FAM_MEMBERS = Quantidade de pessoas na família do Cliente

Dataset: credit_record.csv

- ID = Identificador do Cliente
- MONTHS_BALANCE = Medida relativa do mês devedor, sendo 0 o mês atual
- STATUS = Tempo de Atraso nos Pagamentos DPD (Days Past Due):
  - 0 - 1-29 dias,
  - 1 - 30-59 dias,
  - 2 - 60-89 dias,
  - 3 - 90-119 dias,
  - 4 - 120-149 dias,
  - 5 - dividas baixadas com mais de 150 dias,
  - C - Quitado no mês,
  - X - Nenhum empréstimo no mês

Análise Exploratória de Dados (EDA)

Nossos dados estavam divididos em 19 variáveis, sendo 14 variáveis categóricas e 5 variáveis numéricas.

Analisamos cada variável categórica individualmente, observando a frequência e proporção em relação as classes de inadimplência. Algumas observações e suposições ficaram evidentes e foram passíveis de confirmação estatística através de testes de hipótese Chi², Proporção e ANOVA.

Para as variáveis numéricas, tratamos outliers com a técnica IQR (Interquartile Range) e analisamos as estatísticas de posição e dispersão. Devido a natureza das mesmas, aplicamos escalas Likert para transformá-las em variáveis categóricas e então analisamos frequência e proporção em relação as classes de inadimplência. De forma análoga as variáveis categóricas analisadas anteriormente, algumas suposições surgiram e então foram testadas estatísticamente com os mesmos testes de hipóteses já comentados.

Por fim, utilizando do Coeficiente de Cramer, fizemos uma tabela de associação para identificar possível multicolinearidade entre as variáveis, com o objetivo de reduzir este efeito em nosso modelo de machine learning, por consequência aumentar a generalização e interpretabilidade, reduzindo a dimensionalidade.

Insights de interesse:

1. Clientes do Gênero Masculino são mais inadimplentes que Clientes do Gênero Feminino.
2. Clientes sem Veículo Próprio são mais inadimplentes que Clientes com Veículo Próprio.
3. Clientes com Residência Própria são menos inadimplentes que Clientes que não possuem o ativo.
4. Clientes com Filhos são menos inadimplentes que Clientes Sem Filhos.
5. Estudantes são 100% adimplentes.
6. Pensionistas são mais inadimplentes do que as demais formas de origem de receita.
7. Clientes do funcionalismo público são mais inadimplentes que clientes do setor privado.
8. Clientes com formação Acadêmica Completa são 100% adimplentes.

9. Clientes com Ensino Médio Completo são melhores pagadores que Clientes com demais níveis educacionais.
10. Cliente Casados possuem menor inadimplência que os Clientes com demais Estados Civis.
11. Clientes Solteiros possuem maior inadimplência em relação aos demais Estados Civis.
12. Clientes que moram em Cooperação (dividem apartamento) são menos inadimplentes que Clientes com as demais formas de moradia.
13. Clientes que moram de Aluguel ou em Residência Funcional (Custeada pela Empresa ou Governo) são mais inadimplentes que as demais formas de moradia.
14. Os Clientes mais inadimplentes estão localizados na faixa etária de 50-60 anos, enquanto os menos inadimplentes na faixa de 61-70 anos.
15. Os Clientes que possuem email ativo são menos inadimplentes que Clientes que não possuem o endereço eletrônico.
16. Clientes da área de Tecnologia são mais inadimplentes que as demais profissões. Clientes autônomos são os menos inadimplentes.
17. Clientes com receitas acima de 20k US$/ano são mais inadimplentes que Clientes com faixas de receita anual inferiores.
18. Clientes com receitas entre 50-100k US$/ano são os menos inadimplentes em nossa base de dados.
19. A média de receita de nossa carteira de clientes é de 181k US$/ano, sendo a mediana 171k US$/ano com um desvio padrão de 71K US$/ano.
20. O Tamanho médio das Famílias é de 2 pessoas com desvio padrão de 1 pessoa. Podemos afirmar que temos em nossa carteira de Clientes uma grande maioria de solteiros e casais com mais um depedente.
21. Em média, nossos Clientes tem 6.7 anos de tempo de serviço, porém com desvio padrão de 4.5 anos. Uma distribuição bastante desequilibrada apesar de uma mediana muito próxima da média (6 anos).
22. Média de Idade dos Clientes da nossa carteira é de 41 anos com desvio de 9 anos. Mediana = Média.
23. 75% de nossos Clientes tem até 1 Filho!

Pré-Processamento dos Dados e Seleção de Variáveis

Para esta etapa, fizemos uma cópia do dataset unificado e eliminamos os outliers de todas as variáveis numéricas com a técnica de IQR. Posteriormente, utilizamos a escala Likert para transformar estas mesmas variáveis em categóricas.

Como sabemos, modelos matemáticos trabalham apenas com números e não com textos, portanto aplicamos uma técnica de Label Enconding em todas as variáveis, transformando os textos em números inteiros representativos, conforme exemplo abaixo:

```
Variável Genero:

Número de Categorias: 2

Mapeamento: {'F': 0, 'M': 1}


Variável Prop:Carro:

Número de Categorias: 2

Mapeamento: {'N': 0, 'Y': 1}


Variável Prop:Casa:

Número de Categorias: 2

Mapeamento: {'N': 0, 'Y': 1}


Variável Orig_Receita:

Número de Categorias: 5

Mapeamento: {'Commercial associate': 0, 'Pensioner': 1, 'State servant': 2, 'Student': 3, 'Working': 4}
```

Após esta etapa, aplicamos outra técnica, denominada One Hot Enconding, criando variáveis dummies para cada classe numérica representada anteriormente. Nosso dataset então ficou com 75 variáveis e 296.731 observações.

Com os processos anteriores, preparamos nossos dados para serem aplicados a modelos matemáticos, e neste momento aplicamos a um algoritmo de Random Forest para obtermos uma reudção de dimensionalidade, ou seja, reduzir a quantidade de variáveis. O algoritmo possui um retorno sobre a entropia das variáveis, identificando quais variáveis possuem informação mais pura em relação a variável resposta. Utilizamos um threshold (valor de corte) de 0.02 e definimos 18 variáveis mais importantes.

Filtramos então, nosso dataset final e aplicamos a divisão em treino e teste, com a proporção de 70/30 respectivamente.

Modelagem e Validação

Construimos 4 versões de modelos usando os algoritmos de Regressão Logística e Random Forest.

A escolha destes algoritmos se deve a suas capacidades de adequação ao problema de classificação binária, entre outras vantagens:

Regressão Logística

- Interpretabilidade: oferece uma saída que pode ser interpretada em termos de probabilidade de um evento ocorrer, o que é intuitivo para a previsão de inadimplência.
- Eficiência: é um modelo computacionalmente eficiente para o tamanho de dados comumente usados no mercado financeiro.
- Modelagem de Relação Não Linear: embora seja um modelo linear, a RL tem capacidade de modelar relações não lineares através do logit (função logística), o que é muito útil para a previsão de eventos binários.

- Bom Desempenho com Variáveis Categóricas: lida bem com esse tipo de variável, sendo as mesmas muito comum neste tipo de problema de negócio.

Árvores de Decisão e Random Forest

- Interpretabilidade: Uma única árvore de decisão é fácil de visualizar e entender, o que pode ser valioso para a tomada de decisões de crédito, onde é importante explicar o motivo por trás de uma decisão de inadimplência.
- Manuseio de Dados Desbalanceados: O Random Forest pode ser ajustado para lidar melhor com conjuntos de dados desbalanceados ajustando os pesos das classes ou por meio de técnicas de amostragem.
- Redução de Overfitting: Embora uma única árvore de decisão possa ser propensa ao overfitting, um Random Forest reduz esse risco ao construir várias árvores e decidir com base na média ou maioria dos votos.
- Manuseio Automático de Variáveis Categóricas: As árvores de decisão podem processar variáveis categóricas sem a necessidade de pré-processamento como a codificação one-hot.

Para avaliar a performance dos modelos, utilizamos as métricas de Acurácia, Precisão, Recall, F1-Score, AUC-ROC e AUC-PR. Fizemos ainda uma reavaliação da Acurácia, utilizando uma técnica chamada CrossValidation K Fold, que de forma resumida, cria diversos subconjuntos randomizados de dados de teste e calcula a média dos resultados. Isso proporciona uma avaliação mais robusta do modelo, já que o mesmo está sendo testado com diversas partes diferentes do conjunto de dados.

Outro ponto importante relacionado aos modelos, foram as técnicas de Balanceamento de Classe utilizada. Fizemos 2 modelos de Regressão Logística e 2 modelos de Random Forest. Para cada algoritmo, utilizamos uma técnica de Class Weight, balanceamento por pesos de frequência das classes, e a técnica de SMOTE, balanceamento pela adição de dados sintéticos ou oversampling.

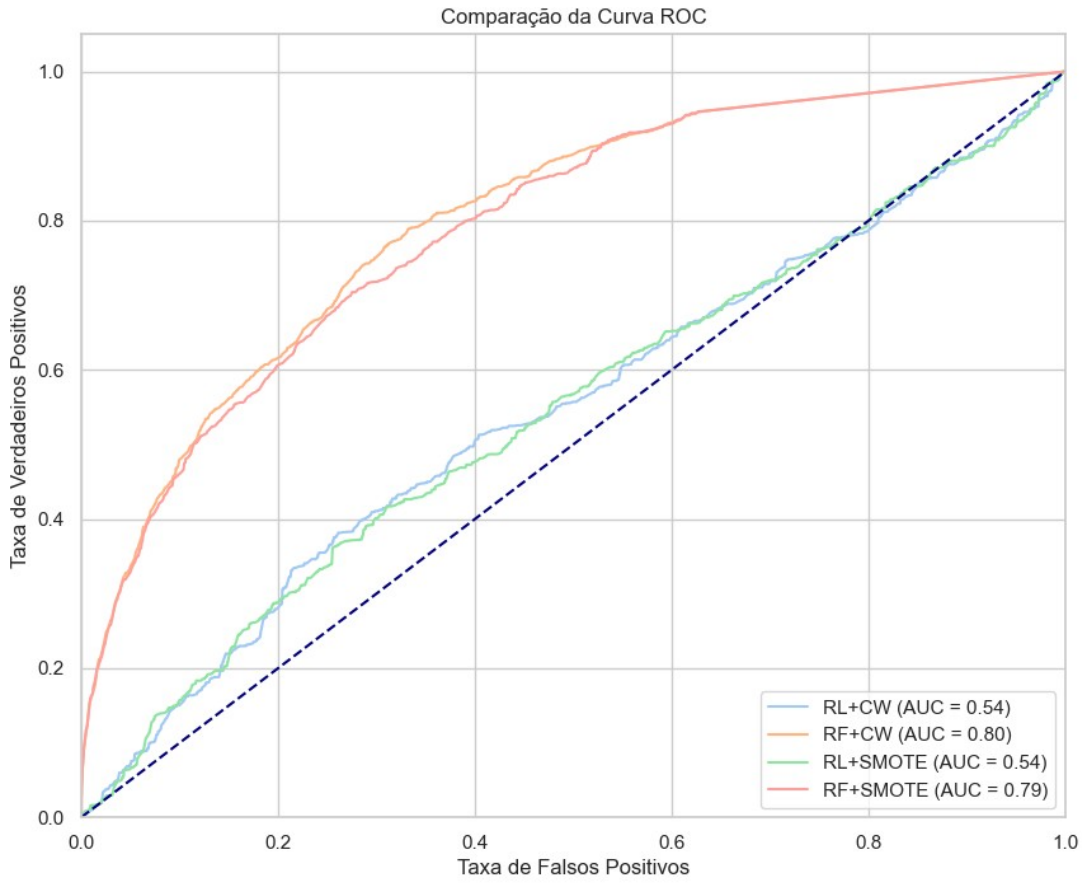
Resultados

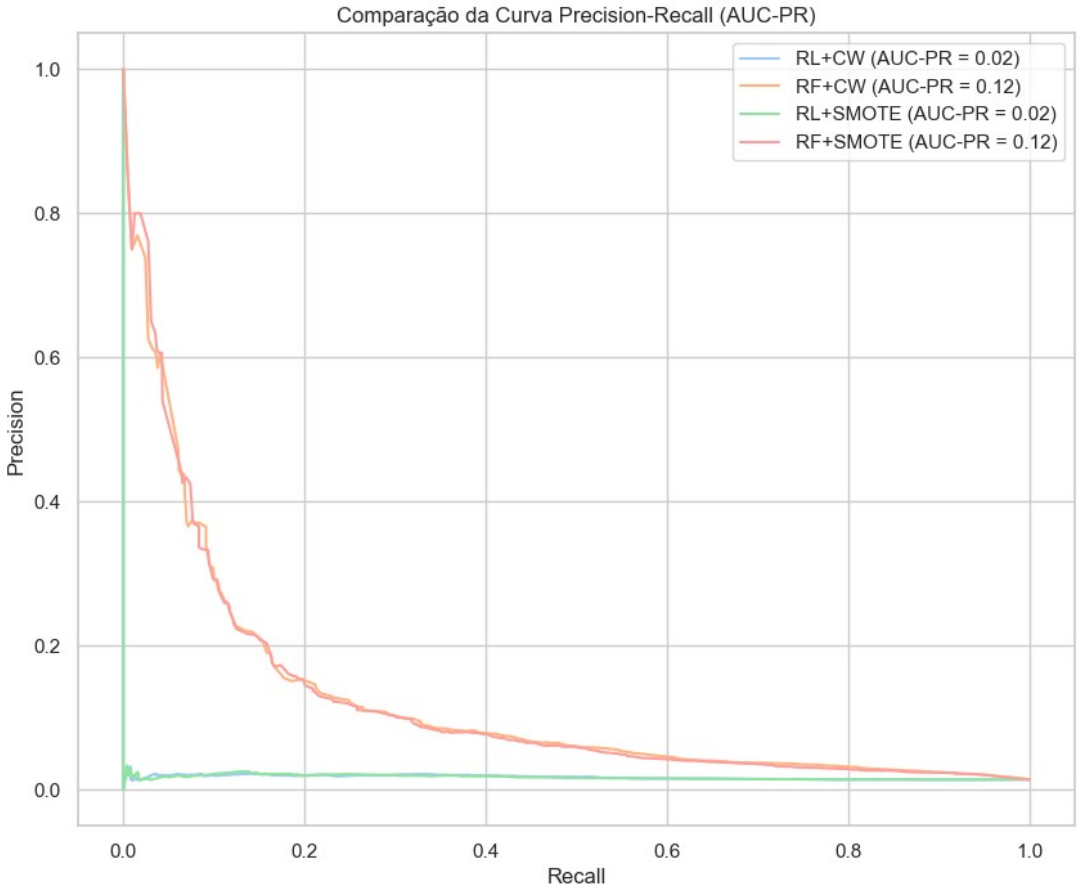
Segue tabela comparativa para as métricas dos modelos testados:

	Acuracia	Media_Acur_KFold	Recall_Macro_Avg	Precision_Macro_Avg	F1-Score_Macro_Avg	AUC_ROC	AUC_PR
RL+CW	0.556493	0.534902	0.541256	0.502352	0.372509	0.544982	0.017586
RF+CW	0.749944	0.790474	0.715370	0.515947	0.463770	0.801434	0.118060
RL+SMOTE	0.542047	0.985722	0.535091	0.501989	0.365980	0.543609	0.017565
RF+SMOTE	0.784071	0.985868	0.703216	0.516702	0.476766	0.791931	0.117327

Como podemos observar, o Modelo com Random Forest com Class Weight obteve acurácia mais realista, mesmo após o cross validation (0.79), assim como melhor Recall e AUC-ROC. É importante aqui ressaltar que o nível de erro é mais importante que a acurácia, visto que temos dados desbalanceados no teste, desta forma o Recall se torna imprescindível para a decisão do modelo.

Por fim, apresentamos as curvas ROC, comparando a taxa de Verdadeiros Positivos com Falsos Positivos e Precisão em relação ao Recall


Comparação da Curva ROC
Taxa de Verdadeiros Positivos
Taxa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
RL+CW (AUC = 0.54)
RF+CW (AUC = 0.80)
RL+SMOTE (AUC = 0.54)
RF+SMOTE (AUC = 0.79)


Comparação da Curva Precision-Recall (AUC-PR)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
RL+CW (AUC-PR = 0.02)
RF+CW (AUC-PR = 0.12)
RL+SMOTE (AUC-PR = 0.02)
RF+SMOTE (AUC-PR = 0.12)

Conclusões e Recomendações

Observamos pelos resultados que obtivemos, que os modelos Random Forest foram mais eficientes nas classificações. Referente a técnica de balanceamento de classes, é possivel notar uma melhora mínima na acuracidade com dados sintéticos, porém percebemos que ao aplicar o Cross Validation K Fold, tivemos overfitting. Isso nos mostra que o modelo não aprendeu os padrões dos dados, o que nos remete a utilizar como modelo final e mais generalizado, o modelo Random Forest + Class Weight.

Para finalizarmos nosso Projeto, podemos citar alguns pontos de melhoria e novos estudos:

- Utilizar mais variáveis, reduzindo o fator de corte para o Coeficiente Cramer, buscando aumentar as métricas do modelo.
- Aplicar outras técnicas para Eliminar Outliers das variáveis numéricas:
  - Z-Score
  - Isolation Forest
  - DBSCAN
  - EllipticEnvelope
  - Análise de Influência (Distância de Cook ou DFFITS)
- Aplicar outras técnicas para definir variáveis importantes para o modelo:
  - Outros Modelos baseados em Árvores de Decisão
  - Modelos Lineares Reguladores
  - Métodos de Embalagem: RFE (Recursive Feature Elimination) ou Métodos de Seleção Sequencial
  - Métodos baseados em Permutação
  - Método SHAP (Shapley Additive Explanations)
- Aplicar outros Algoritmos na construção do modelo:
  - SVM (Suport Vector Machine)
  - GBM (Gradient Boost Machine)
  - Redes Neurais Artificiais.
- Aplicar ajustes de Hiperparâmetros para afinar o modelo e melhorar as métricas, tomando cuidado com excessos e overfitting.

Finalizamos este projeto com o intuito de viver aprendizados adquiridos em livros e teses estudadas sobre o tema Gestão de Risco de Crédito e Credit Scoring, aplicando técnicas e habilidades de programação Python e Ciência de Dados.

Referências

- Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, Second Edition By Naeem Siddiqi Copyright © 2017 by SAS Institute, Inc.
- Artigo - Complete Guide to Credit Risk Modelling - A Complete Guide to Credit Risk Modelling (listendata.com)

- Artigo - Credit Risk: Vintage Analysis - Credit Risk : Vintage Analysis (listendata.com)
- Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions Swati Tyagi - American International Journal of Business Management (AIJBM)