1. Introdução

O objetivo deste trabalho é construir e avaliar um modelo preditivo de risco de crédito capaz de estimar a probabilidade de inadimplência de clientes de cartão de crédito. A partir dessas probabilidades, busca-se propor políticas de cessão de crédito, isto é, regras operacionais que orientem a decisão de manter um cliente na carteira da instituição ou ceder esse crédito a terceiros.

A base utilizada é o conjunto de dados “Default of Credit Card Clients”, amplamente conhecido na literatura e disponibilizado pela UCI. Ela reúne informações sociodemográficas, características de crédito, histórico de faturas e pagamentos, além do indicador de default no mês subsequente. O problema é formulado como uma tarefa de classificação binária, em que a classe “1” representa clientes inadimplentes e a classe “0” representa clientes adimplentes.

A proposta metodológica combina preparação cuidadosa dos dados, análise exploratória e redução de dimensionalidade, seleção de variáveis por algoritmo genético e modelagem via regressão logística com ponderação de classes. Adicionalmente, diferentes limiares de decisão são testados para representar políticas mais equilibradas ou mais agressivas em relação à detecção de inadimplentes, e uma validação cruzada é empregada para avaliar a robustez dos resultados.

2. Base de dados

Após a leitura do arquivo original em formato Excel e a remoção da coluna de identificação, o conjunto passa a contar com 30.000 observações e 24 variáveis explicativas, além da variável alvo, renomeada para `default` para simplificar o tratamento. As variáveis incluem características sociodemográficas, como sexo, nível educacional, estado civil e idade, informações de crédito, como o limite total de cartão de crédito (`LIMIT_BAL`), e um conjunto de indicadores que retratam o histórico de atraso de pagamento, valores de fatura e montantes efetivamente pagos ao longo de seis meses.

A variável resposta `default` é tratada como um fator binário, com “0” indicando ausência de inadimplência e “1” indicando ocorrência de default. A distribuição das classes revela um cenário típico de desbalanceamento: aproximadamente 22% das observações correspondem a inadimplentes, enquanto cerca de 78% são adimplentes. Esse desbalanceamento é um aspecto central do problema, pois impacta diretamente a forma como o modelo deve ser treinado e avaliado.

3. Preparação e pré-processamento dos dados

O processo de preparação dos dados começa pela definição dos tipos adequados das variáveis. As variáveis categóricas, como sexo, nível de escolaridade, estado civil e indicadores de atraso de pagamento, são convertidas para o tipo fator. Essa conversão é essencial para

que, em etapas posteriores, o R reconheça essas variáveis como categorias e gere corretamente as dummies necessárias à modelagem.

Em seguida, é realizado o tratamento de valores faltantes. Para variáveis numéricas, adota-se a imputação pela mediana, reduzindo a sensibilidade a valores extremos. Para variáveis categóricas, utiliza-se a imputação pela moda, isto é, pela categoria mais frequente. Esse procedimento permite manter todas as observações, evitando a perda de informação decorrente da exclusão de linhas com dados ausentes.

Um passo adicional é o tratamento de outliers por meio de winsorização. Em cada variável numérica, valores abaixo do primeiro percentil são elevados a esse valor, e valores acima do nono nono percentil são reduzidos a este. Dessa forma, atenua-se o impacto de observações extremas sem remover dados, o que é particularmente importante em variáveis financeiras, sujeitas a valores muito altos em poucos casos.

O conjunto de dados limpo é então dividido em subconjuntos de treino e teste, utilizando aproximadamente 70% das observações para o treino e 30% para o teste, com semente aleatória fixa para garantir reprodutibilidade. Na sequência, emprega-se a função `model.matrix` para transformar as variáveis originais em uma matriz numérica de características, com criação automática de variáveis dummies para os fatores. A coluna de intercepto gerada automaticamente é removida, e os dados são normalizados utilizando média e desvio padrão calculados apenas sobre o conjunto de treino. Essas estatísticas são então aplicadas ao conjunto de teste, evitando vazamento de informação.

4. Análise exploratória e PCA

A análise exploratória de dados fornece uma visão inicial sobre a estrutura do conjunto. São examinadas as dimensões do dataset, as primeiras linhas e um resumo estatístico das variáveis numéricas, permitindo identificar faixas de valores típicas, assimetrias e possíveis problemas de escala. A distribuição da variável `default` é analisada em termos de frequências absolutas e relativas, o que confirma o desbalanceamento entre adimplentes e inadimplentes.

A partir dessa visão geral, alguns gráficos são construídos para melhor compreender o comportamento das variáveis. Um histograma da idade, por exemplo, mostra em que faixas etárias se concentra a maior parte dos clientes. Um boxplot do limite de crédito segmentado por status de default permite comparar a distribuição do limite entre clientes que se tornaram inadimplentes e aqueles que permanecem adimplentes, o que auxilia na identificação de padrões de risco associados ao tamanho do limite.

Também é aplicada uma Análise de Componentes Principais (PCA) à matriz de características normalizadas. O objetivo é investigar a correlação entre as variáveis e visualizar a distribuição das observações em um espaço de menor dimensão. As primeiras componentes principais explicam uma fração relevante, mas não dominante, da variância total, sinalizando que o problema é de natureza genuinamente multidimensional. O gráfico de dispersão das duas

primeiras componentes, colorido de acordo com a classe de `default`, evidencia sobreposição considerável entre as classes, embora se observe alguma tendência de separação em determinadas regiões do espaço.

5. Seleção de variáveis por algoritmo genético

Dada a quantidade de variáveis geradas após a codificação em dummies e a possibilidade de multicolinearidade, opta-se por utilizar um algoritmo genético para seleção de variáveis. Nesse contexto, cada indivíduo na população do GA representa um subconjunto de variáveis, codificado como um vetor binário em que o valor “1” indica que uma variável é selecionada e o valor “0” indica que é excluída.

O cálculo do fitness é feito sobre uma subamostra fixa do conjunto de treino, o que reduz o custo computacional. Para cada indivíduo, ajusta-se um modelo de regressão logística utilizando apenas as variáveis selecionadas, com pesos de classe definidos para lidar com o desbalanceamento. A partir das probabilidades previstas, calcula-se uma medida de desempenho, como a área sob a curva ROC (AUC) ou uma combinação de sensibilidade, especificidade e um termo de penalização proporcional ao número de variáveis incluídas. Dessa forma, o GA procura subconjuntos que conciliam bom poder discriminatório e parcimônia.

O algoritmo genético utiliza operadores padrão de seleção, cruzamento e mutação, bem como elitismo, de modo a preservar os melhores indivíduos ao longo das gerações. Um critério de parada antecipada é adotado quando não há melhoria significativa do fitness por um número pré-determinado de iterações, o que torna o processo mais eficiente. Ao final, obtém-se o subconjunto de variáveis selecionadas e um registro da evolução do fitness, permitindo compreender como o GA convergiu para uma solução considerada adequada.

6. Regressão logística com ponderação de classes

Com o subconjunto de variáveis selecionadas pelo algoritmo genético, ajusta-se o modelo principal de regressão logística binária. Essa escolha é justificada pela ampla utilização desse tipo de modelo em risco de crédito, pela interpretação transparente dos coeficientes em termos de odds de inadimplência e pela capacidade de trabalhar naturalmente com probabilidades.

O desbalanceamento de classes é tratado por meio de pesos na função de verossimilhança. Observações da classe de inadimplentes recebem maior peso relativo, de modo que erros nessa classe tenham impacto mais forte na estimação dos parâmetros. A ideia é aproximar a contribuição das classes na calibração do modelo, sem alterar artificialmente a proporção de observações.

Uma vez ajustada a regressão logística, obtêm-se probabilidades estimadas de default para cada cliente nos conjuntos de treino e teste. Essas probabilidades permitem traçar curvas ROC, a partir das quais se calcula a área sob a curva (AUC). Os valores de AUC encontrados situam-se em torno de 0,76, tanto em treino quanto em teste, o que é compatível com um

modelo de capacidade discriminatória razoável no contexto de crédito e em linha com resultados reportados para esse conjunto de dados na literatura.

Além do AUC, as probabilidades são usadas para gerar tabelas de desempenho em escala de limiares, mostrando, para diferentes valores de corte, os valores de acurácia, sensibilidade, especificidade, precisão, F1-score e índice de Youden. Essas tabelas constituem a base para a definição de políticas de decisão.

7. Limiar de decisão e políticas de cessão de crédito

A regressão logística fornece probabilidades de default, mas a operacionalização do modelo depende da escolha de um limiar que transforme essas probabilidades em decisões binárias. No contexto deste trabalho, adota-se a interpretação de que clientes com probabilidade de default acima do limiar são considerados de alto risco e, portanto, candidatos à cessão de crédito, enquanto clientes com probabilidade abaixo ou igual ao limiar são considerados de risco mais baixo e permanecem na carteira.

Dois cenários básicos são analisados. No primeiro, busca-se um limiar que maximize o F1-score, métrica que equilibra sensibilidade e precisão. Esse limiar, situado aproximadamente em 0,65, produz uma política mais equilibrada, com acurácia elevada, boa especificidade e uma sensibilidade moderada. Em termos de negócio, essa abordagem tende a minimizar tanto falsos positivos quanto falsos negativos de maneira relativamente simétrica.

No segundo cenário, adota-se uma postura mais agressiva em relação à detecção de inadimplentes. Impõe-se uma sensibilidade mínima desejada, por exemplo em torno de 0,75, e, entre os limiares que satisfazem essa condição, escolhe-se aquele que apresenta maior F1-score. Essa escolha desloca o limiar para um valor mais baixo, em torno de 0,48, de forma a aumentar a taxa de detecção de inadimplentes às custas de maior número de falsos positivos e de redução da precisão. Em situações em que o custo de não identificar um inadimplente é muito superior ao de classificar incorretamente um adimplente como arriscado, essa política agressiva pode ser mais interessante para a instituição.

As métricas associadas a cada cenário, incluindo as respectivas matrizes de confusão nos conjuntos de treino e teste, são apresentadas e discutidas, evidenciando o compromisso entre captura de inadimplentes e manutenção de uma taxa aceitável de falsos positivos.

8. Validação cruzada

Para verificar a robustez do modelo e dos limiares de decisão, é aplicada validação cruzada do tipo K-fold sobre o conjunto de treino. Nesse procedimento, os dados são divididos em K subconjuntos de tamanho semelhante. Em cada iteração, um desses subconjuntos é utilizado como conjunto de validação e os demais compõem o conjunto de treino. O modelo de regressão logística é ajustado novamente em cada combinação, usando as mesmas variáveis selecionadas pelo algoritmo genético, e o limiar operacional é aplicado ao fold de validação.

Em cada fold são calculadas as principais métricas de desempenho, tais como acurácia, sensibilidade, especificidade, precisão, F1-score e AUC. Ao final, obtém-se a média e o desvio padrão dessas métricas ao longo dos K folds, possibilitando avaliar a estabilidade do modelo em diferentes partições dos dados.

Os resultados da validação cruzada indicam que as métricas variam dentro de faixas relativamente estreitas, sem degradação acentuada em nenhum fold específico. Isso sugere que o desempenho observado no split inicial de treino e teste não é fruto do acaso e que o modelo apresenta capacidade consistente de generalização para novos dados.

9. Interpretação econômica dos coeficientes

Um dos pontos centrais da escolha da regressão logística é a possibilidade de interpretar os coeficientes estimados em termos de odds ratios. Ao tomar a exponencial de cada coeficiente, obtém-se quanto as odds de default se multiplicam quando a variável em questão aumenta uma unidade, no caso de variáveis contínuas, ou quando se compara uma categoria dummy com a categoria de referência, no caso de variáveis categóricas.

No modelo ajustado, o coeficiente associado ao limite de crédito (`LIMIT_BAL`) é negativo, e seu odds ratio é inferior a um, o que indica que clientes com limites mais altos tendem, ceteris paribus, a apresentar menor probabilidade de inadimplência. Por outro lado, variáveis relacionadas a atrasos de pagamento recentes, como determinadas dummies derivadas das variáveis `PAY_0`, `PAY_2` e subsequentes, apresentam coeficientes positivos e odds ratios superiores a um, sugerindo que atrasos passados aumentam sensivelmente o risco futuro de default.

As variáveis associadas aos valores pagos, como `PAY_AMT1` e `PAY_AMT2`, em geral exibem coeficientes negativos, indicando que pagamentos mais elevados reduzem a probabilidade de inadimplência. Variáveis sociodemográficas, como escolaridade e estado civil, em muitos casos apresentam efeitos mais discretos ou estatisticamente menos significativos, o que é coerente com a expectativa de que o comportamento de pagamento e o uso do crédito sejam as fontes mais relevantes de informação para o risco.

Essa interpretação detalhada dos coeficientes e dos odds ratios permite traduzir o modelo em termos de regras de negócio, facilitando sua comunicação para áreas não técnicas da instituição, como equipes de crédito, risco e governança.

10. Gráficos e visualizações produzidas

Ao longo do pipeline, são gerados diversos gráficos que reforçam a compreensão do comportamento dos dados e do desempenho do modelo. Entre as principais visualizações, incluem-se gráficos de análise exploratória, como o histograma da distribuição de idade e o boxplot de limite de crédito segmentado por status de default, que ajudam a contextualizar o perfil da carteira.

Do ponto de vista da modelagem, são elaborados gráficos de PCA que exibem as observações projetadas nas duas primeiras componentes principais, coloridas conforme a classe de `default`, bem como as curvas ROC para os conjuntos de treino e teste, com o valor de AUC indicado em destaque. Além disso, é gerado um gráfico da evolução do fitness ao longo das gerações do algoritmo genético, permitindo visualizar como a busca por subconjuntos de variáveis mais adequados se estabiliza ao longo do tempo.

Também são produzidas visualizações que relacionam as diferentes métricas de desempenho com o limiar de decisão, destacando graficamente os limiares escolhidos para as políticas de equilíbrio e de sensibilidade elevada. Por fim, um gráfico de barras com os principais coeficientes da regressão logística, ordenados por magnitude, sintetiza as variáveis de maior impacto na probabilidade de inadimplência, contribuindo para a interpretação do modelo.

11. Conclusões e perspectivas

O trabalho descreve um pipeline completo para modelagem de risco de crédito em uma base real de clientes de cartão de crédito, desde o pré-processamento dos dados até a definição de políticas de decisão fundamentadas em probabilidades de default. A combinação de algoritmo genético para seleção de variáveis, regressão logística ponderada para modelagem e análise de limiares para calibrar políticas de cessão mostrou-se viável e coerente com as práticas de risco de crédito.

Os resultados obtidos, com valores de AUC em torno de 0,76 e desempenho estável em validação cruzada, indicam que o modelo é capaz de discriminar de forma razoável entre clientes adimplentes e inadimplentes. A possibilidade de ajustar o limiar de decisão permite adaptar a política de cessão aos objetivos estratégicos da instituição, seja privilegiando o equilíbrio entre erros, seja enfatizando a redução de falsos negativos em cenários de maior aversão ao risco.

A interpretação dos coeficientes e odds ratios fornece uma visão econômica clara de como diferentes fatores, em especial o histórico de atraso e o comportamento de pagamento, influenciam o risco de default. Essa característica é particularmente importante para fins de transparência, governança e comunicação com áreas de negócio.