Trabalho 1 - Introdução ao Aprendizado de Máquina

Luís Rafael Sena

Dezembro 2024

Contents

1 Introdução

Este relatório apresenta o desenvolvimento do Trabalho 1 da disciplina EEL891 - Introdução ao Aprendizado de Máquina, ofertada no período 2024-2. O objetivo deste trabalho foi construir um sistema preditivo que auxilie na decisão de concessão de crédito, analisando a probabilidade de inadimplência de novos clientes com base em dados históricos.

Para isso, utilizamos um conjunto de dados contendo 20.000 solicitações de crédito, que incluíam variáveis relacionadas ao perfil financeiro dos solicitantes e os desfechos das solicitações (quitação ou inadimplência). A partir desse histórico, treinamos diferentes modelos de aprendizado de máquina, que foram avaliados em um conjunto de teste oculto contendo 5.000 registros, com as métricas calculadas automaticamente pelo Kaggle.

Este trabalho utilizou bibliotecas amplamente reconhecidas, como pandas, numpy e scikit-learn, que foram fundamentais para as etapas de pré-processamento, modelagem e análise dos dados. Este relatório detalha cada uma dessas etapas, discute os desafios enfrentados e apresenta os resultados obtidos.

2 Pré-Processamento dos Dados

O pré-processamento dos dados é essencial para garantir a qualidade e consistência dos dados de entrada, permitindo que os modelos de aprendizado de máquina sejam treinados de forma eficaz. Nesta seção, detalhamos as estratégias adotadas para tratar diferentes aspectos do conjunto de dados.

2.1 Tratamento de Dados Não Numéricos

As variáveis categóricas foram convertidas em representações numéricas por meio de *one-hot encoding*, uma técnica que cria novas colunas binárias para cada categoria presente. Essa abordagem foi escolhida para preservar as informações originais e evitar a introdução de hierarquias artificiais nos dados.

Para variáveis binárias, optou-se por manter uma única coluna com valores 0 e 1, reduzindo redundâncias e garantindo a simplicidade do conjunto de dados.

2.2 Tratamento de Dados Faltantes

Valores ausentes foram tratados com estratégias específicas para cada tipo de dado:

- **Dados categóricos:** Foram preenchidos com a mediana das categorias disponíveis, uma vez que a mediana é robusta a outliers e representa melhor a centralidade em variáveis categóricas.
- **Dados numéricos:** Optou-se pelo preenchimento com o valor zero, simplificando o processamento e evitando vieses na imputação.

Essa abordagem foi escolhida devido à simplicidade computacional e aos bons resultados observados em experimentos preliminares.

2.3 Normalização

Os dados numéricos foram normalizados utilizando o *RobustScaler*, que transforma as variáveis para uma escala reduzida e é robusto à presença de outliers. Essa técnica garantiu que variáveis com diferentes amplitudes tivessem o mesmo peso durante o treinamento dos modelos.

2.4 Seleção de Características

A matriz de correlação foi utilizada para identificar variáveis redundantes e pouco relevantes. A análise demonstrou que a coluna `grau_instrucao` apresentava uma correlação baixa com a variável alvo (`inadimplente`) e outras variáveis preditivas, sendo, portanto, excluída do conjunto final.

3 Modelos Testados

Diversos modelos foram explorados, com o objetivo de identificar a melhor abordagem para o problema em questão. Os algoritmos selecionados e seus principais parâmetros ajustados estão descritos a seguir:

3.1 Random Forest

O *Random Forest* destacou-se como o modelo mais robusto, devido à sua capacidade de lidar com dados heterogêneos e capturar interações complexas entre variáveis. Foram testadas diferentes configurações de profundidade máxima (até 16) e número de estimadores, com foco na otimização do desempenho.

3.2 K-Nearest Neighbors (KNN)

O método KNN foi testado utilizando valores ímpares de k (potências de 3), para minimizar empates durante as classificações. Adicionalmente, foram exploradas diferentes métricas de distância (euclidiana e Manhattan) e pesos (uniformes e baseados em distância).

3.3 Regressão Logística

A Regressão Logística foi implementada como um modelo base para comparação. Apesar de sua simplicidade, demonstrou desempenho competitivo em cenários menos complexos.

3.4 Gradiente Descendente Estocástico (SGD)

O SGD foi testado com várias funções de erro (logarítmico, quadrático e perceptron) e taxas de aprendizado. Essa abordagem permitiu avaliar a eficácia do método em um conjunto de dados com alta dimensionalidade.

4 Métricas de Avaliação

4.1 Divisão dos Dados

O conjunto de dados foi dividido da seguinte forma:

- **Treino:** 81% dos registros para treinamento inicial dos modelos.
- **Teste:** 9% dos registros para validação intermediária.
- **Validação:** 10% dos registros para ajustes finais e avaliação.

4.2 Acurácia

A acurácia foi a métrica principal utilizada para avaliar o desempenho dos modelos, sendo definida como a proporção de predições corretas no conjunto de validação.

4.3 Matriz de Correlação

A matriz de correlação foi essencial para entender as relações entre as variáveis e reduzir redundâncias no conjunto de dados, contribuindo para um modelo mais eficiente e interpretável.

5 Resultados e Discussão

5.1 Modelo Selecionado

O modelo *Random Forest* com profundidade máxima de 16 e 100 estimadores apresentou o melhor desempenho no conjunto de validação, destacando-se por sua robustez e capacidade de generalização.

5.2 Impacto do Pré-Processamento

O uso do *RobustScaler* e a imputação baseada em zero e mediana foram determinantes para a qualidade dos dados de entrada, mitigando problemas relacionados a outliers e valores ausentes.

5.3 Desempenho no Kaggle

O modelo final obteve uma acurácia de 59% no conjunto de teste do Kaggle, garantindo uma posição de destaque na competição e demonstrando a eficácia da abordagem adotada.

6 Conclusão

Este trabalho destacou a importância de um pipeline bem estruturado para problemas de classificação, desde o pré-processamento até a seleção e ajuste de modelos. Os resultados demonstraram que o uso de técnicas robustas, como *Random Forest* e *RobustScaler*, pode levar a soluções eficientes e competitivas.