Uso de Algoritmos de Classificação no Contexto de Dados Financeiros

Daniel B. de Salles, Guilherme A. D. Maciel, Halliday G. C. dos Santos, Iago C. Andrade, Vinicius N. Targa

[1]Departamento de Computação
Universidade Federal de Ouro Preto (UFOP) – Ouro Preto, MG – Brazil

daniel.bortot@aluno.ufop.edu.br, guilherme.maciel@aluno.ufop.edu.br,

halliday.santos@aluno.ufop.edu.br, iago.andrade@aluno.ufop.edu.br,

vinicius.targa@aluno.ufop.edu.br

Abstract. *This report aims to introduce classification algorithms with and without attribute selection, carry out their application in the treatment of data from the German Credit Dataset (credit-g) and evaluate them in order to identify the best model.*

Resumo. *Este relatório tem como objetivo introduzir algoritmos de classificação com e sem a seleção de atributos, realizar sua aplicação no tratamento de dados do German Credit Dataset (credit-g) e avaliá-los de maneira a identificar o melhor modelo.*

1. Introdução

Este trabalho se baseia no uso de três algoritmos de classificação na análise de dados financeiros. Os algoritmos usados eram ou não capazes de selecionar atributos, sendo eles: regressão logística (*logistic regression*), árvore de decisão (*decision tree*) e *Random Forest*. A base dados utilizada foi o *dataset* conhecido como *German Credit Dataset* (*credit-g*)[Dua and Graff 2017], disponibilizado pelo professor Dr. Hans Hofmann.

A regressão logística é um modelo de classificação binária que nos permite estimar a probabilidade associada à ocorrência de determinado evento em função dos valores conhecidos de outras variáveis. Uma função sigmoide é utilizada nesse modelo, por conseguinte seus resultados se encontram contidos no intervalo [0,1].[Smola and Vishwanathan 2008]

Uma árvore de decisão é capaz de predizer o rótulo associado com um elemento instanciado ao percorrer uma árvore de um nó raiz até um nó folha. Em cada nó do caminho raiz-folha, o filho sucessor é escolhido com base em uma divisão do espaço de entrada. Normalmente, esta divisão é baseada em uma das características do elemento ou em um conjunto predefinido de regras de divisão.[Shalev-Shwartz and Ben-David 2021]

Random Forests por sua vez, são conjuntos de árvores criados com o objetivo de reduzir as chances de ocorrer *overfitting*. Este tipo de classificador consiste em uma coleção de árvores de decisão, onde cada árvore é construída aplicando um algoritmo a um conjunto de dados de treino e um vetor aleatório adicional, amostrado de variáveis

aleatórias independentes e identicamente distribuídas. A predição deste modelo é obtida através da ”maioria de votos” dentre as predições das árvores individuais.[Breiman 2001]

Por fim, o *dataset credit-g* classifica as pessoas descritas como risco de crédito bom ou ruim através de um conjunto de vinte atributos diversos (por exemplo, situação da conta corrente existente, histórico de crédito, finalidade de crédito, quantidade de crédito, situação de suas economias, entre outros).

2. Metodologia

O pré-processamento dos dados foi realizado em etapas. Primeiro, foi realizada uma limpeza dos dados nulos e dos dados duplicados. Em seguida, atributos nominais foram transformados em atributos numéricos, e aqueles que não foram considerados úteis para a análise foram removidos (mais especificamente, status pessoal, sexo, idade, número de pessoas pelas quais o indivíduo é responsável, telefone e se é ou não um trabalhador estrangeiro). Realizou-se a impressão da matriz de correlação dos dados para verificar se algum dos atributos possuía alguma correlação que justificasse sua remoção — tal fato não foi observado no *dataset* utilizado, apresentado na figura 1. Então, usou-se o *one-hot encoding* para transformar as variáveis categóricas em binárias para assegurar compatibilidade com os algoritmos. Por sinal, nenhuma instância de indivíduos de sexo ’Feminino’ e estado civil ’Solteiro’ foi encontrada, bem como propósito ’Férias’.

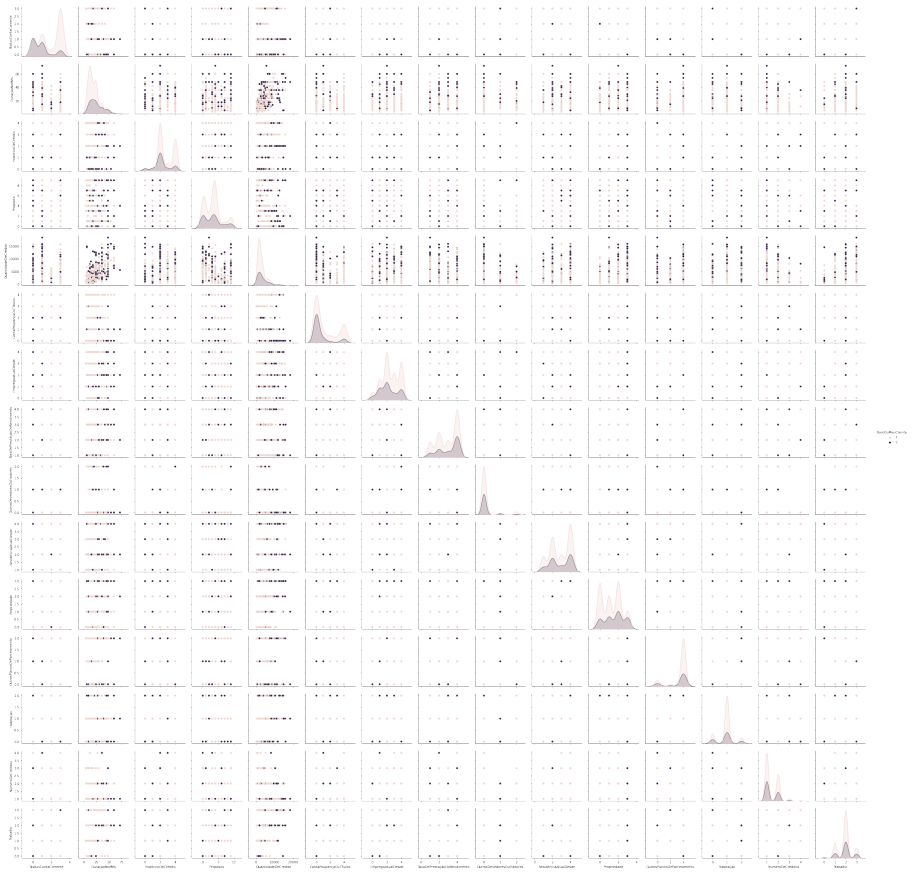

Figure 1. Matriz de correlação dos dados do dataset credit-g.

A normalização dos atributos numéricos foi então realizada, agilizando a execução e permitindo maior integridade dos dados, eliminando possíveis redundâncias. A seguir,

verificou-se a distribuição dos dados com base no rótulo (cliente bom ou ruim), e foi possível observar, por exemplo:

1. Se o status da conta corrente possuísse um valor menor ou igual a 200 Deutsche Marks, há uma tendência a ser categorizado cliente ruim, mostrado na figura 2;
2. Em contra partida, não possuir uma conta corrente esteve mais associado com o rótulo de cliente bom, também demonstrado pela figura 2;

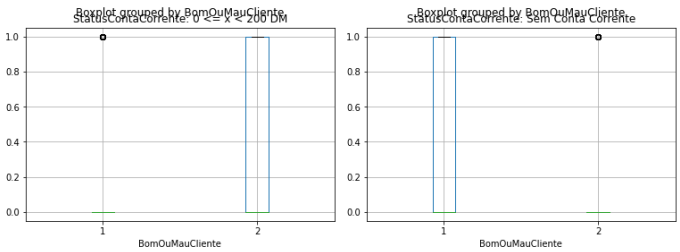

Figure 2. Box plot de distribuição das variáveis referentes ao status da conta corrente.

3. Estar no mesmo emprego por sete ou mais anos foi uma tendência dos clientes categorizados como bons, bem como possuir um imóvel como propriedade, descrito pela figura 3;

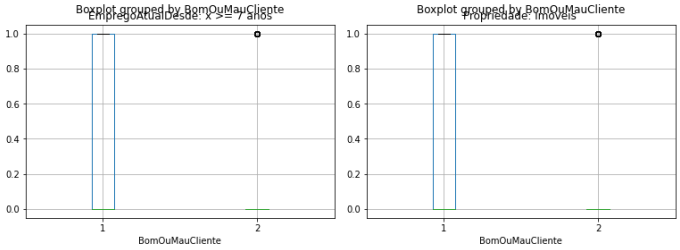

Figure 3. Box plot de distribuição das variáveis referentes à situação empregatícia e propriedade de imóveis.

4. Curiosamente, ter o interesse de usar o crédito na compra de eletrodomésticos inclinou-se à categoria de cliente bom, enquanto a compra de um carro novo inclinou-se a cliente ruim, apresentados pela figura 4.

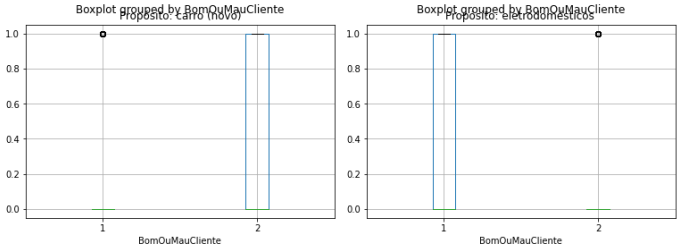

Figure 4. Box plot de distribuição das variáveis referentes ao interesse no uso do crédito.

Posteriormente, dividiu-se os conjuntos de treinamento e de teste. Dentre os 1000 indivíduos presentes no *dataset credit-g*, 70% foram utilizados no treinamento dos algoritmos, enquanto os 30% restantes foram utilizados para teste. Na etapa de treinamento dos modelos, a regressão logística recebeu os pesos de classe 49.9 para clientes bons e 50.1 para clientes ruins, definiu-se o número mínimo de amostras necessárias para estar em um nó folha como 30 para o algoritmo de árvore de decisão, e o número de árvores de decisão como 150 para o algoritmo de *Random Forests*.

3. Resultados

Foi feita uma validação cruzada com k-fold após treinamento dos modelos, e o resultados obtidos estão descritos pela tabela 1.

	Acurácia (desvio padrão)	F-measure (desvio padrão)
Regressão Logística	0.740 (0.047)	0.823 (0.040)
Árvore de Decisão	0.703 (0.051)	0.791 (0.043)
Random Forest	0.759 (0.053)	0.848 (0.037)

Table 1. Resultados da validação cruzada para os três algoritmos.

Após isso, foi realizada a predição no conjunto de teste, e as três matrizes geradas estão dispostas a seguir.

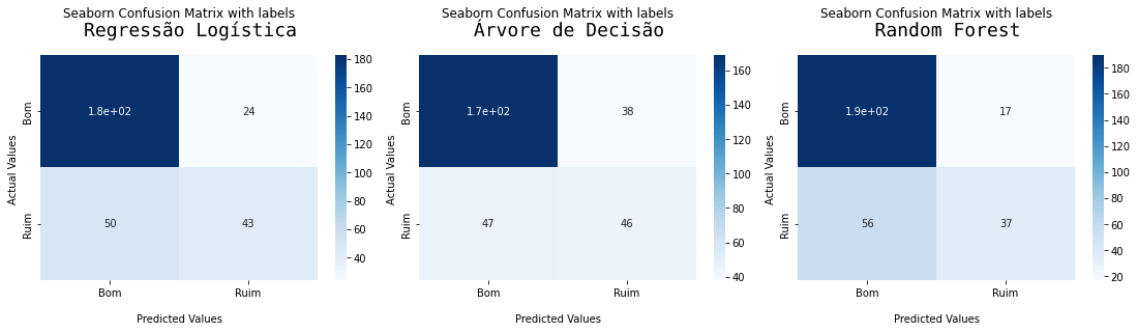

Figure 5. Matrizes de confusão dos algoritmos usados.

4. Conclusões

Dentre os algoritmos utilizados, a *Random Forest* possui a maior acurácia e a maior F-measure. Além disso, pelo conjunto de teste também foi possível perceber que a *Random Forest* possui melhor resultado em acertar clientes bons, enquanto a árvore de decisão é preferível para acertar clientes ruins, evidenciado pela tabela 2.

	Acertos para clientes bons	Acertos para clientes ruins
Regressão Logística	0.884	0.462
Árvore de Decisão	0.816	0.495
Random Forest	0.918	0.398

Table 2. Análise das matrizes de confusão.

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Shalev-Shwartz, S. and Ben-David, S. (2021). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Smola, A. and Vishwanathan, S. (2008). *Introduction to Machine Learning*. Cambridge University Press.