

COMPARAÇÃO DE ACURÁCIAS EM MODELOS DE APRENDIZADO DE MÁQUINA

Gabriel Penha*

06/07/2022

1 INTRODUÇÃO

O estudo tem como objetivo tratar e aplicar modelos de aprendizado de máquina para problemas de classificação, o conjunto de dados utilizados foi sobre pagamentos inadimplentes, dados de crédito, histórico de pagamentos e extratos de contas de clientes de cartão de crédito entre abril e setembro. O banco apresenta 30000 observações e inicialmente 25 variáveis, sendo essas 14 numéricas e 10 categóricas. A variável de interesse é o pagamento no próximo mês.

2 METODOLOGIA

Primeiramente o estudo iniciou com uma análise descritiva procurando enxergar como as variáveis preditoras do banco se relacionavam com a variável resposta. Uma das ferramentas utilizadas para isso foi a matriz de correlação com a medida de associação V de Cramer para as variáveis categóricas.

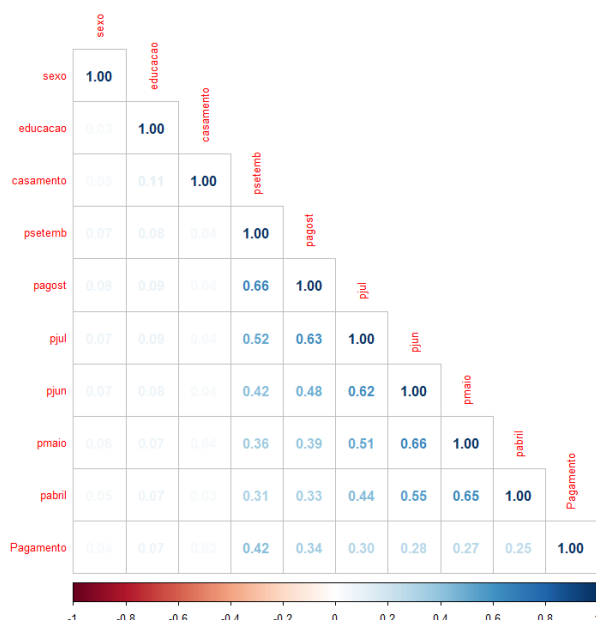



Figura 1 – Matriz de correlação das variáveis categóricas

Fonte: Autor (2022)

Pode-se visualizar que o pagamento está fortemente relacionado com os pagamentos anteriores e conforme vão se distanciando no tempo essa correlação vai diminuindo. Além disso, as variáveis sexo, educação e casamento quase não apresentaram correlação.

*  Departamento de Estatística, UFBA, Bacharelado em Estatística; gabriel.rodrigues4210@gmail.com.

Quanto as variáveis numéricas, ao fazer os bloxplots nenhuma apresentou uma diferença visível em relação a variável resposta e apresentavam valores muito altos, o máximo de alguns chegava a ser 100 vezes o valor do terceiro quartil da distribuição. Então os valores foram filtrados apenas para os menores que 400000.

Tentando retirar mais informações para o modelo foram criadas variáveis dummy para os meses onde eram 0 se o no mês pagou em dia, ou atrasou um mês e 1 se tiver atrasado mais tempo. Pois, estranhamente a chance de pagar no proximo mês era maior para os que apresentavam atraso de 3 meses ou mais.

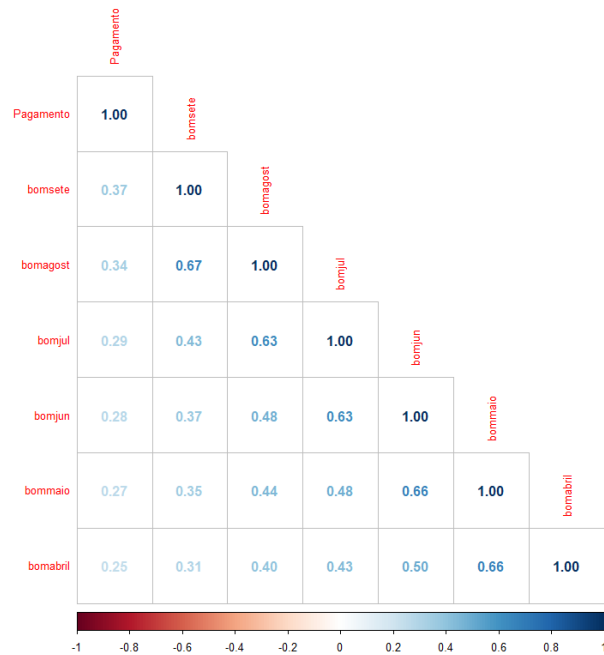


Figura 2 – Matriz de correlação das variáveis categóricas
Fonte: Autor (2022)

Além disso, a variável resposta era desbalanceada, apresentando 23364 observações para uma categoria e 6636 observações para a outra, então para que os modelos conseguissem apresentar uma capacidade preditiva próxima para as duas (uma qualidade razoável tanto para especificidade quanto para sensibilidade) foi realizada uma subamostragem, onde a variável resposta apresentou porcentagem semelhante para as duas categorias e a base final teve 12756 observações.

Os dados foram divididos com 75 por cento para treino e 25 por cento para teste, com controle de treino usando validação cruzada em 10 partes e 5 repetições.

Os modelos utilizados foram: GLM, KNN, RPART, LDA, GLMBoost, RF, TREEBAG, GBM

.	Modelo	Accuracy	Sensitivity	Specificity
GLM		0.7015	0.8107	0.5950
KNN		0.5945	0.5724	0.6161
RPART		0.6977	0.7954	0.6025
LDA		0.6968	0.8126	0.5839
GLMBoost		0.6983	0.8247	0.5752
RF		0.7034	0.7677	0.6399
TREEBAG		0.687	0.7235	0.6511
GBM (best - 100 trees)		0.7108	-	-

Tabela 1 – Métricas preditivas para cada modelo

3 CONCLUSÕES

Finalmente, o melhor modelo em termos de acuracia foi o GBM com 100 árvores, entretanto não soube retornar as medidas de sensibilidade e especificidade dele. Além disso, o melhor modelo em termo de sensibilidade foi o GLMBoost e

para especificidade foi o TreeBag. Finalmente, com o objetivo de um modelo que tivesse um melhor trade-off entre as duas medidas o modelo sugerido para esse problema foi o RF.

ANEXO A – ANEXO

No seguinte endereço é possível obter a base de dados e visualizar os procedimentos e os códigos utilizados para sua análise: <https://github.com/Gabriel4210/Data-Mining-Atv>