MODELOS DE APRENDIZADO DE MÁQUINA: UMA APLICAÇÃO EM R[1]

Lucas Henrique Fleischmann[2], Elisa Henning[3], Raquel Sant'Ana[4]

[1] Vinculado ao projeto "Métodos estatísticos e de aprendizado de máquina para análise do consumo de água em edificações"
[2]Acadêmico (a) do Curso de Licenciatura em Matemática – CCT – Bolsista PIBIC/CNPQ
[3] Orientador, Departamento de Matemática – CCT – elisa.henning@udesc.br
[4] Acadêmico de Pós-Graduação em Engenharia Civil – CCT

A análise de dados se tornou parte crucial no mundo moderno. Com a diversidade, em formato e categoria, dos dados e o desenvolvimento da computação, modelos mais sofisticados do que uma simples regressão linear, por exemplo, começaram a ser explorados: modelos de aprendizado de máquina.

No processo de aprendizagem, trabalhei com alguns desses modelos, de modo a entender os métodos e, posteriormente, a forma de aplicá-los. Para isso, utilizei uma base de dados de algumas casas de Joinville, em que a informação de interesse era o valor do imóvel, ou seja, uma resposta quantitativa.

Detalhadamente, os dados utilizados são de 99 casas dos bairros Glória, Santo Antônio e Costa e Silva, da cidade de Joinville, localizada no estado de Santa Catarina, Brasil. Delas, são informações conhecidas: rendimento médio mensal em salários mínimos; valor do imóvel – que ocupou o papel de variável resposta; área da edificação; zoneamento; pavimentos (1 ou 2); número de dormitórios; número de banheiros; número de garagens cobertas; mobília; existência de piscina; idade da edificação; sua conservação; se é geminada; a área total do terreno; o posicionamento do imóvel; e ocorrência de inundação.

Essa aplicação dos modelos de aprendizado de máquina foi feita com o pacote H2O. Para a análise em si, a plataforma foi executada dentro do software R. Com o H2O é possível construir os modelos de aprendizado de máquina de forma fácil de manipular, com a vantagem de possuir várias explicações dentro do próprio pacote. Com poucos parâmetros já é possível treinar um modelo. Dentro das funções, podemos especificar e melhorar esses modelos.

Mais especificamente, nesse pacote, tem-se a função H2O AutoML: um algoritmo de aprendizado supervisionado que treina os dados, num tempo determinado pelo usuário, com vários modelos gerais (por exemplo, GBM e Florestas Aleatórias). E, além disso, com modelos *stacked ensembles*, que são modelos mais complexos e poderosos que usam como base os modelos previamente treinados pela função. Um desses modelos é o *All Models*, que inclui todos os modelos de base, e o outro é o *Best of Family*, que utiliza apenas o melhor modelo de cada grupo treinado.

Com isso, os procedimentos foram a preparação e limpeza dos dados; a separação dos dados em treino e teste; a aplicação dos modelos com os dados de treino e a amostra de teste para medir a acurácia dos modelos (momento em que estudei as métricas).

Para a separação, foi escolhida a proporção de 80% dos valores para treino e 20% dos valores para teste. Uma justificativa para essa escolha pode ser encontrada em Gholamy *et al.* (2018).

Um dos modelos individualmente treinados foi o de regressão linear, uma aproximação direta utilizada para dados quantitativos. Busca-se estimar os valores dos coeficientes da regressão com a minimização dos quadrados dos resíduos (James, 2021). Também foram treinados os modelos de Florestas Aleatórias e *Gradient Boosting Machine*, que são modelos baseados em árvores de decisão, que separam os dados observados em vários subconjuntos disjuntos e atribuem a média dos valores para cada subconjunto. As Florestas Aleatórias objetivam remover a correlação entre as árvores, enquanto que o *Boosting* atualiza o modelo gradativamente com base nos resíduos (James, 2021).

Para validar os modelos, pode-se utilizar diferentes métricas. Neste estudo, em que a variável resposta era quantitativa, as métricas utilizadas foram *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), *Median Absolute Percentage Error* (MedianAPE), Correlação, R^2, e *Deviance*.

Na Figura 1 está um mapa de calor que mostra a importância de cada variável (eixo y) em cada modelo (eixo x). Percebe-se que a área da edificação (m^2) é a variável de maior importância para os modelos.

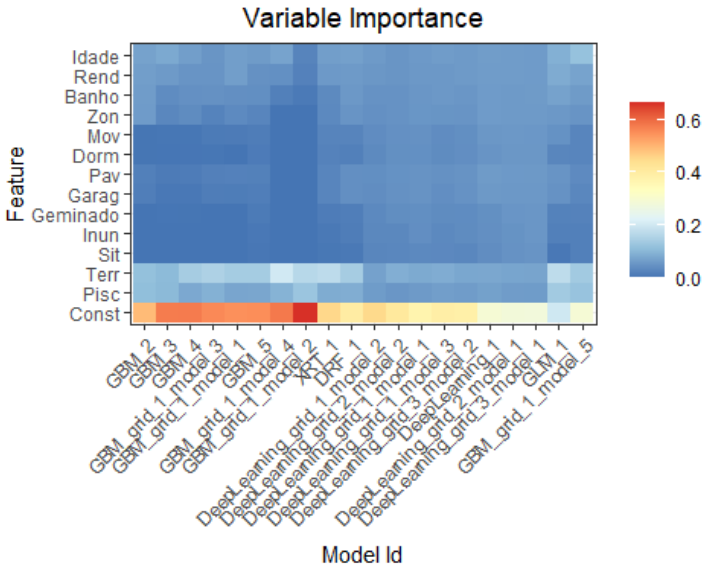

Figura 1. *Importância de cada variável considerada nos modelos treinados para a função AutoML.*

Palavras-chave: Aprendizado de máquina. Modelos. Métricas.

Referências:

JAMES, G. *et al*. **An Introduction to Statistical Learning**: with Applications in R. Second edition, 2021.
LEDELL, E.; POIRIER, S. **H2O AutoML**: Scalable Automatic Machine Learning. 2020.
GHOLAMY, A. *et al*. **Why 70/30 or 80/20 relation between training and testing sets**: a pedagogical explanation. University of Texas, El passo, 2018.