Trabalho Individual - Analise e Seleção de Modelos

Data de Entrega: 28/11/2024

Descrição:

Neste trabalho, você utilizará a base de dados *sao-paulo-properties.csv* com informações de preços do aluguel de apartamentos e diversas características desses imóveis em São Paulo. O objetivo é comparar modelos e escolher aquele de maior potencial preditivo do preço. Após a escolha do melhor modelo você deve realizar a predição para o valor dos imóveis da base de dados *sao-paulo-properties-new.csv*.

Instruções:

- Treine pelo menos 4 modelos apropriados para a variável "*Price"* sendo que entre eles deve existir, necessariamente, um **Modelo de Regressão Linear** (regularizado ou não), um **Modelo baseado em árvores** e um **Modelo não paramétrico** (ex.: k vizinhos);

- Utilize algum método de validação para avaliar o desempenho dos modelos. (Deixe explícita a estratégia escolhida). Escolha o melhor modelo sob uma perspectiva preditiva.

- Realize a predição de preço para os imóveis contemplados na base de dados *sao-paulo-properties-new* e salve em uma tabela com apenas duas colunas, são elas, "X" e "*pred"* que representam, respectivamente, o código único do imóvel presente na base de dados disponibilizada e o valor predito do modelo construído por você.

    - Note que, essa base de dados possui todas as variáveis da *sao-paulo-properties* com exceção do preço. Após o envio do trabalho será calculado e divulgado o erro das predições obtidas por cada um dos trabalhos entregues.

    - A métrica de erro considerada será o Erro Quadrático Médio (EQM)

$$EQM = 1/n \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Em que,

- n é o número de observações,
- y_i é o valor real da variável resposta,
- $\widehat{y_i}$ é o valor predito pelo modelo.

Formato de Entrega:

Serão entregues **3 arquivos** (se atente aos formatos, pois entregas fora do padrão não serão considerada), são eles:

1. Arquivo *.pdf* com o relatório dos resultados, gerado pelo **R Markdown** ou **Jupyter Notebook,** contendo descrição dos modelos testados e a estratégia de seleção do modelo final.

2. Arquivo *.Rmd* ou *.ipynb* que gerou o arquivo em **1;**

3. Arquivo em formato *.csv* com tabela contendo apenas duas colunas. Uma das colunas denominada ***"X"*** que contempla o código único do imóvel presente na base de dados *sao-paulo-properties-new.csv* e a outra coluna denominada **"*pred*"** com as predições realizadas pelo modelo escolhido para cada um dos imóveis (identificados por "X") da base *sao-paulo-properties-new.csv* . A seguir um exemplo de como devem aparecer as colunas (os valores são apenas ilustrativos):

X	**pred**
3	2000
35	1500
26	850
…	…

Para esse exemplo teríamos o valor de aluguel de R$2.000,00 predito para o imóvel 3.