Natã dos Santos Carvalho
DRE: 115027521

Relatório do Trabalho da Disciplina EEL891 - Aprendizado de Máquina

Rio de Janeiro

2019, v1.0

Natã dos Santos Carvalho
DRE: 115027521

Relatório do Trabalho da Disciplina EEL891 - Aprendizado de Máquina

Relatório sobre o trabalho de predição de dados imobiliários da disciplina de Aprendizado de Máquina no período de 2018.2 ministrada pelo professor Heraldo Almeida

Universidade Federal do Rio de Janeiro – UFRJ

Escola Politécnica

Engenharia Eletrônica e de Computação

Rio de Janeiro

2019, v1.0

Resumo

Neste relatório, será explicado o passo a passo para o desenvolvimento do código fonte apresentado no link do GitHub a seguir: https://github.com/natancarvalho/EEL891—Aprendizado-de-M-quina/blob/master/Untitled.ipynb

Palavras-chaves: Aprendizado de Máquina. Python.

Sumário

1 Análise e Tratamento de Dados

Para análise e tratamento de dados foram o utilizadas as seguintes bilbiotecas:

- **pandas** para geração de um modelo de dados estruturado e manipulação do mesmo;
- **numpy** e **scipy** para cálculos matemáticos;
- **seaborn** com ferramentas de estatística;
- **matplotlib** para visualização de dados.

1.1 Visualização dos Dados

Primeiramente, os dados foram separados conjuntos de treino e teste. Destinando 25% dos dados de treino para validação dos modelos.

Foi verificado a descrição do conjunto de treino fornecido pela biblioteca ***pandas*** e os primeiros dados do conjunto com as funções ".describe()"e ".head()"

1.2 Remoção de Outliers

A fim de remover possíveis outliers, foi exibido o "scatter plot"dos preços pra facilitar a visualização. Assim, são retiradas três amostras com valores excessivamente altos.

Posteriormente fora exibidos gráficos preço x atributos para verificar outros outliers. Feito isso, a remoção foi feita para continuar a análise de dados.

Foi feita uma Transformada de Box-Cox na coluna de preços para aprimorar a distribuição de preços, com lambda = 0.25.

1.3 Correlação de Dados

Antes de trabalhar com as variáveis numéricas, foi feita a normalização de todas.

Após a normalização, forão calculados os índices de correlação entre as variáveis numéricas e exibidas as que tiverem correlação positiva através da função ".corr()"e mostrado um mapa de calor das correlações.

2 Seleção de Atributos

Para a seleção de atributos foi escolhido o método Recursive Feature Extraction with Cross-Validation selection. Em seguida os dados de treino e teste foram manipulados de forma a manter somente as features selecionadas pelo método acima.

3 Modelos de Predição

Para realizar as predições foram escolhidos os modelos:

- **Regressão Linear**
- **Regressão Lasso**
- **Gradient Boosting Regression**

Foram mostrados os melhores e os piores resultados para o conjunto de treino completo e para o conjunto de treino com os atributos escolhidos pelo método citado anteriomente.

Por último, foi retornado os valores dos preços do imóveis e o data frame foi exportado para um arquivo csv.

Referências