Relátorio: Trabalho 2 - Introdução a aprendizado de máquina 2020-1

1. Objetivo
   O objetivo deste trabalho é o desenvolvimento de um regressor que prediga o preço de um imóvel

2. Os dados

   O conjunto de dados utilizado apresenta informações sobre imóveis do estado do Pernambucano, incluindo seus respectivos preços e atrativos.

3. Tratamento de dados

Colunas:

'Id', 'tipo', 'bairro', 'tipo_vendedor', 'quartos', 'suites', 'vagas', 'area_util', 'area_extra', 'diferenciais', 'churrasqueira', 'estacionamento', 'piscina', 'playground', 'quadra', 's_festas', 's_jogos', 's_ginastica', 'sauna', 'vista_mar', 'preco'.

Inicialmente, o conjunto de dados foi analisado usando a ferramenta “profile_report()” da biblioteca pandas, com isso foi possível descobrir várias informações importantes a respeito do conjunto de dados.

- Por apresentar valores nulos, alta cardinalidade, alta correlação com outra variável ou por agregar pouca informação ao modelo, as seguintes variáveis foram retiradas do “dataframe” inicial, utilizando o método “drop”:

'tipo', "diferenciais", ‘'tipo_vendedor', 'bairro', ‘'area_extra', 'estacionamento', 'churrasqueira', 'piscina', 'playground', 'quadra', 's_festas','s_jogos','s_ginastica', 'sauna','vista_mar'

- Foi aplicada uma escala nas variáveis do conjunto utilizando o método “StandardScaler”

- Os outliers, valores de y (preço) que se distanciam muito da média foram removidos

4. Modelos

Os seguindos modelos foram testados no conjunto de treinamento:

- Regressor polinomial
- Regressor Ridge
- RegressorLasso
- Regressor KNN
- Regressor ElasticNet
- Regressor SGD

Para os modelos de regressão polinomial os valores k do grau do polinômio foram variados entre 1 e 9.
Para o modelo de regressão KNN o valor de k foi variado entre 1 e 50.
Então, GridSearchCV foi utilizado para encontrar as melhores combinações de parâmetros para um dado modelo.

Os seguintes resultados foram obtidos:

REGRESSOR KNN:

| K | DENTRO da amostra | FORA da amostra |
|---|---|---|
| 25 | 0.0108 | 0.8194 |
| 26 | 0.0108 | 0.8194 |
| 27 | 0.0108 | 0.8194 |
| 28 | 0.0108 | 0.8194 |
| 29 | 0.0108 | 0.8194 |
| 30 | 0.0108 | 0.8194 |
| 31 | 0.0108 | 0.8194 |
| 32 | 0.0108 | 0.8194 |
| 33 | 0.0108 | 0.8194 |
| 34 | 0.0108 | 0.8194 |

REGRESSOR POLINOMIAL DE GRAU K:

| K | NA | DENTRO da amostra | FORA da amostra |
|---|---|---|---|
| 1 | 5 | 0.0347 | 0.8197 |
| 2 | 15 | 0.0310 | 0.8196 |
| 3 | 35 | 0.0275 | 0.8228 |
| 4 | 70 | 0.0260 | 1.2639 |
| 5 | 126 | 0.0242 | 15.1679 |
| 6 | 210 | 0.0225 | 136.4276 |
| 7 | 330 | 0.0208 | 33552.3751 |

REGRESSOR POLINOMIAL DE GRAU K COM REGULARIZACAO RIDGE (L2):

| K | NA | DENTRO da amostra | FORA da amostra |
|---|---|---|---|
| 1 | 5 | 0.0535 | 0.8198 |
| 2 | 15 | 0.0495 | 0.8197 |
| 3 | 35 | 0.0426 | 0.8194 |
| 4 | 70 | 0.0371 | 0.8288 |
| 5 | 126 | 0.0346 | 0.8971 |
| 6 | 210 | 0.0331 | 5.1842 |
| 7 | 330 | 0.0312 | 52.7763 |

REGRESSOR POLINOMIAL DE GRAU K COM REGULARIZACAO LASSO (L1):

| K | NA | DENTRO da amostra | FORA da amostra |
|---|---|---|---|
| 1 | 5 | 0.0560 | 0.8199 |
| 2 | 15 | 0.0560 | 0.8199 |
| 3 | 35 | 0.0560 | 0.8199 |
| 4 | 70 | 0.0560 | 0.8199 |
| 5 | 126 | 0.0560 | 0.8199 |
| 6 | 210 | 0.0560 | 0.8199 |
| 7 | 330 | 0.0560 | 0.8199 |

REGRESSOR POLINOMIAL DE GRAU K COM ElasticNet:

| K | NA | DENTRO da amostra | FORA da amostra |
|---|---|---|---|
| 1 | 5 | 0.0560 | 0.8199 |
| 2 | 15 | 0.0560 | 0.8199 |
| 3 | 35 | 0.0560 | 0.8199 |
| 4 | 70 | 0.0560 | 0.8199 |
| 5 | 126 | 0.0560 | 0.8199 |
| 6 | 210 | 0.0560 | 0.8199 |
| 7 | 330 | 0.0560 | 0.8199 |

REGRESSOR SGD:

| Métrica | DENTRO da amostra | FORA da amostra |
|---|---|---|
| mse | 0.0012 | 0.6720 |
| rmse | 0.0349 | 0.8197 |
| r2 | 0.6123 | -0.0003 |

Com isso, o modelo Regressor KNN  foi escolhido como definitivo para o projeto, com os parâmetros:

```
regressor_knn = KNeighborsRegressor(
        n_neighbors = 31,
        weights     = 'distance',
        p=1,
        n_jobs=-1
        )
```

E obtendo o rmse final de 0,34501.