SPARK HW

DS 6003

Name: Mengyao Zhang (mz6jv)

Jan. 30, 2019

Motivation

- Data
  - *The HousePrices dataset is a simulated dataset from Kaggle. It contains 500K observations and 16 variables.*
  - *The dependent variable is Prices.*
- Goals
  - *Use pyspark to read in data as a spark dataframe directly from S3.*
  - *Prepare data in the correct form for analysis.*
  - *Use MLlib to build a linear regression model to predict Prices.*
  - *Visualize the data.*

Code Snippet

- The following code shows how we can create multiple-element vector as a feature column using *VectorAssembler*

```
1 from pyspark.ml.feature import VectorAssembler
2
3 # vectorize the df
4 feature_names = df2.schema.names[0:15]
5 vectorAssembler = VectorAssembler(inputCols = feature_names, outputCol = 'features')
6 new_df = vectorAssembler.transform(df2)
7 new_df = new_df.select(['features', 'Prices'])
8 new_df.show(3)
```

```
+--------------------+------+
|            features|Prices|
+--------------------+------+
|[30.0,2.0,1.0,4.0...| 33000|
|[31.0,1.0,4.0,4.0...| 38775|
|[4.0,1.0,4.0,3.0,...| 14350|
+--------------------+------+
only showing top 3 rows
```

Visualization

- Linear regression model with *Area* as a feature.
  - *R2 = 0.021*
- Linear regression model with *Floors* as a feature.
  - *R2 = 0.38*
- Linear regression with all features.
  - *R2 = 0.99*
- It may be helpful to visualize the correlation matrix to see which features are more important.