PREDICTION OF REAL ESTATE PRICES

Elena Gillis

DS6003

Homework 1 – Pyspark

January 30, 2019

MOTIVATION

- Dataset:
  - Real estate prices in Taipei City, Taiwan
  - Dimensions: 414 x 7
  - Features:
    - Date of transaction
    - Age of the house
    - Distance from nearest metro station
    - Number of convenience stores in area
    - Geographic coordinates

CODE

- Random Forest regression in pyspark:

```
# Train a Random Forest model
rf = RandomForestRegressor()
rfModel = rf.fit(trainingDF)

# transform test set to make predictions (predict on test set)
predictionsAndLabelsDF = rfModel.transform(testDF)
```

- Evaluation

```
Root Mean Squared Error (RMSE) on test data = 6.21435
R-squared = 0.74881
```

VISUALIZATION