Real Estate Pricing Valuation

Sept 20, 2018

Hardit Singh

AGENDA

- Problem Statement
- Data Overview
- Exploratory Data Analysis
  - Distribution of single variables using density plots
  - Effect of a variable conditioned on other variables using Correlation Matrix
  - Relationship among Crime Rate and House Value
- K-Fold Cross Validation Models
- Leave-One-Out Model Comparison
- Further Questions

PROBLEM STATEMENT

Scenario: The real estate agency offers an automated valuation service for prospective customers who consider selling their homes. The current valuation method is based on historical prices of nearby properties, which are often outdated and the process requires a substantial amount of work to manually adjust values.

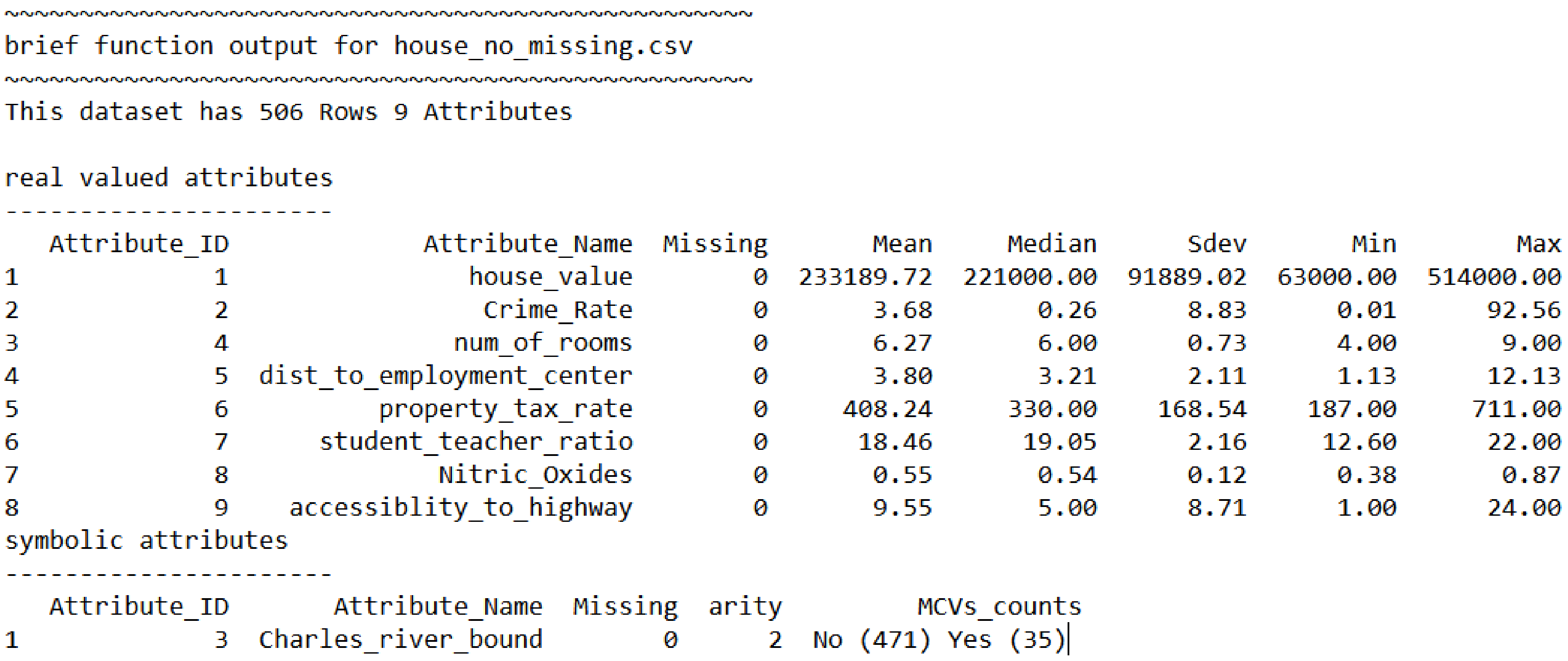
Objective: Improve the current valuation service to better estimate selling price by incorporating sophisticated modelling techniques using the available historical transaction data.

Outcome: A predictive model that estimates selling price using known characteristics of properties.

Data available: CSV file containing house value of historical transactions with the following attributes:

Crime Rate, Charles River Bound (Yes/No), Crime Rate, Num Of Rooms, Distance To Employment Center, Property Tax Rate, Student Teacher Ratio, Nitric Oxides, Accessibility To Highway, House Value

DATA OVERVIEW



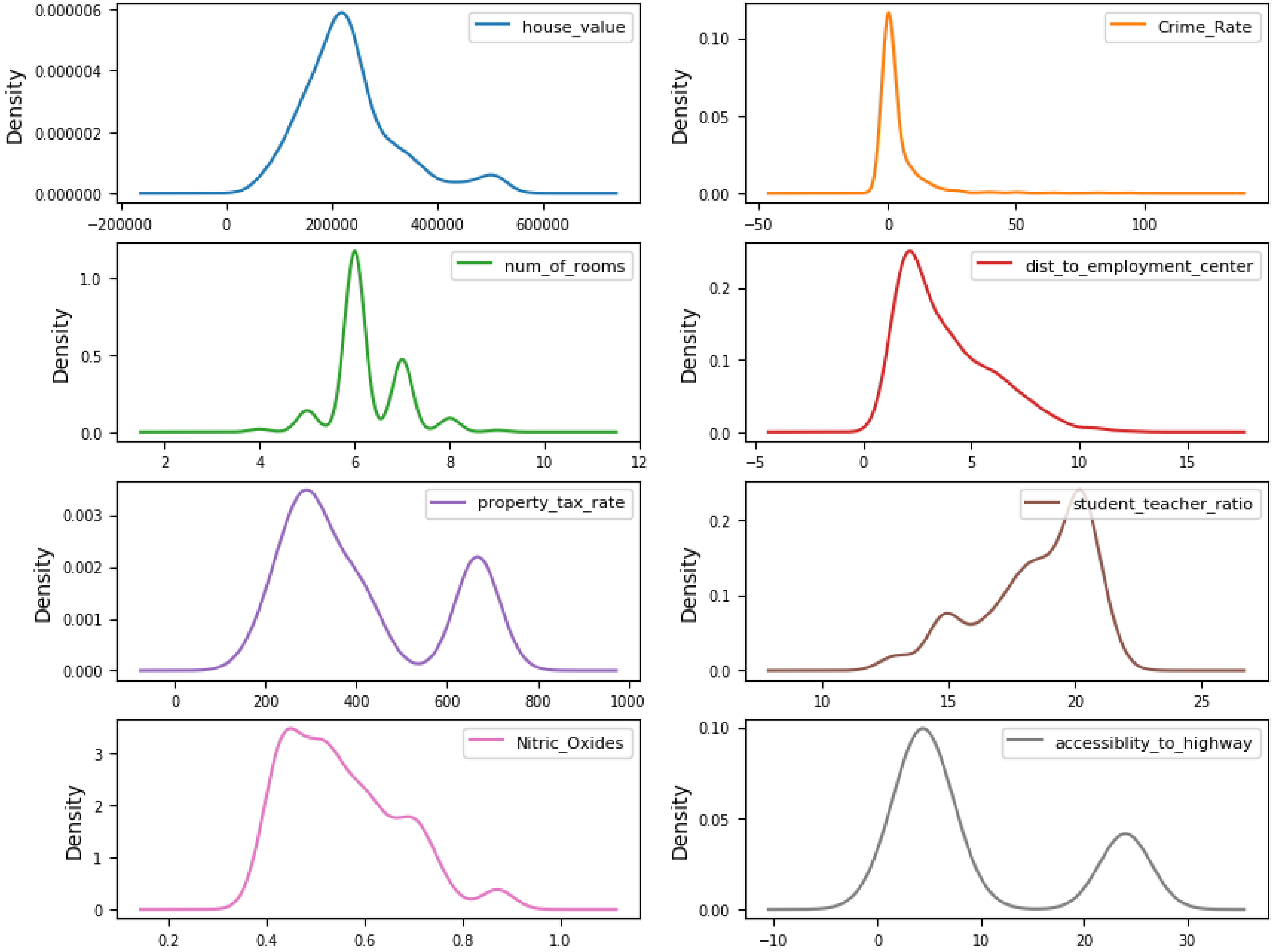
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
brief function output for house_no_missing.csv
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
This dataset has 506 Rows 9 Attributes

real valued attributes
----------------------
   Attribute_ID             Attribute_Name  Missing       Mean     Median      Sdev       Min        Max
1             1                house_value        0  233189.72  221000.00  91889.02  63000.00  514000.00
2             2                 Crime_Rate        0       3.68       0.26      8.83      0.01      92.56
3             4               num_of_rooms        0       6.27       6.00      0.73      4.00       9.00
4             5  dist_to_employment_center        0       3.80       3.21      2.11      1.13      12.13
5             6          property_tax_rate        0     408.24     330.00    168.54    187.00     711.00
6             7      student_teacher_ratio        0      18.46      19.05      2.16     12.60      22.00
7             8              Nitric_Oxides        0       0.55       0.54      0.12      0.38       0.87
8             9    accessiblity_to_highway        0       9.55       5.00      8.71      1.00      24.00
symbolic attributes
----------------------
   Attribute_ID       Attribute_Name  Missing  arity         MCVs_counts
1             3  Charles_river_bound        0      2   No (471) Yes (35)
```

Findings

- There are no missing values in the data
- High standard deviation in house values. The house values range from $63,000 to $514,000

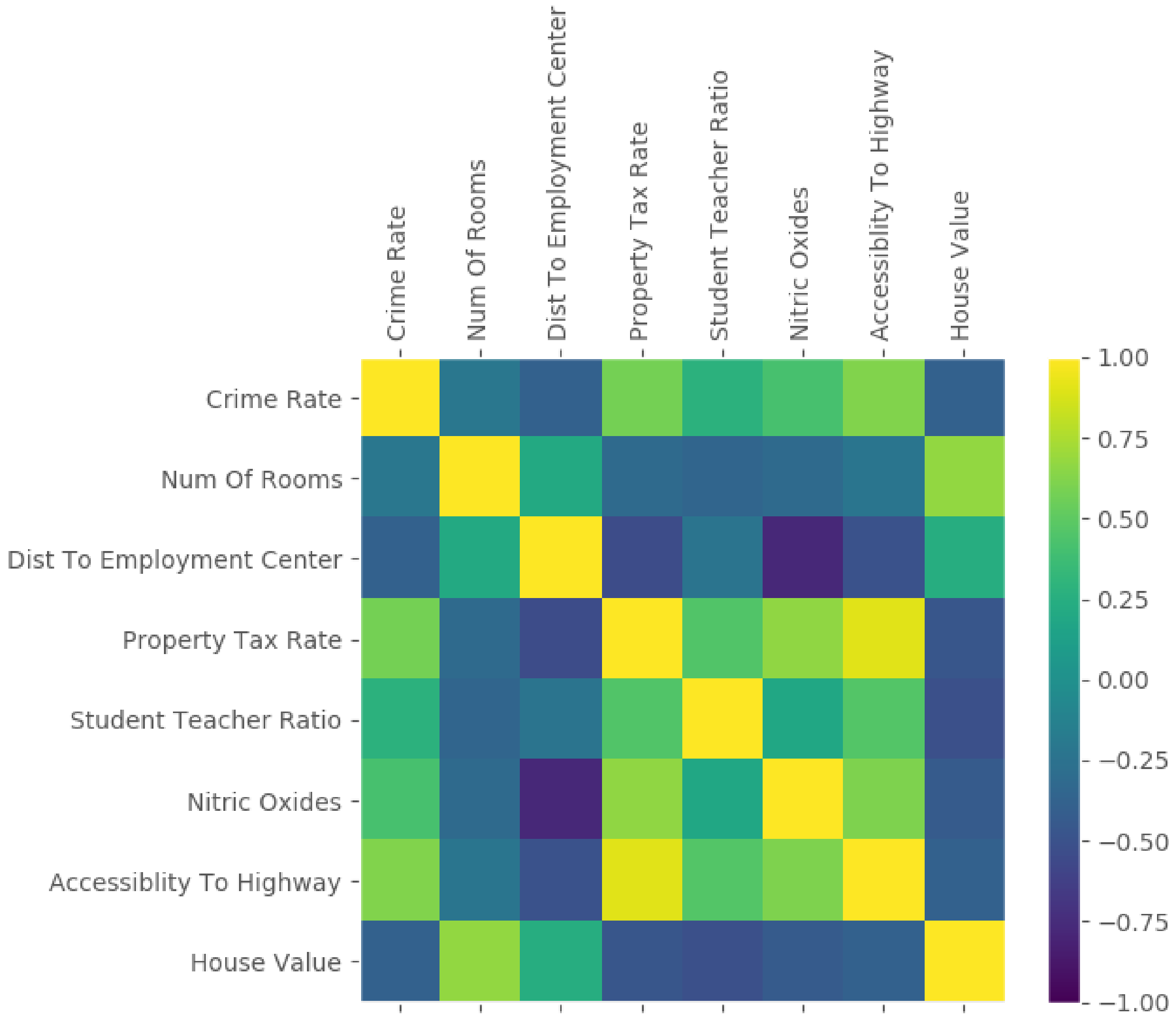
DENSITY PLOTS



Findings

- House Value is approximately normally distributed across the dataset with mean at $233,189
- Crime Rate has a left skewed distribution, suggesting that most houses lie in regions with low crime rate
- On average houses have 6 rooms
- Student Teacher Ratio is right skewed suggesting more number of houses lie in regions where a teacher teaches many students
- Two distinct peaks observed in distribution of accessibility to highways meaning houses lie either close to highway or away

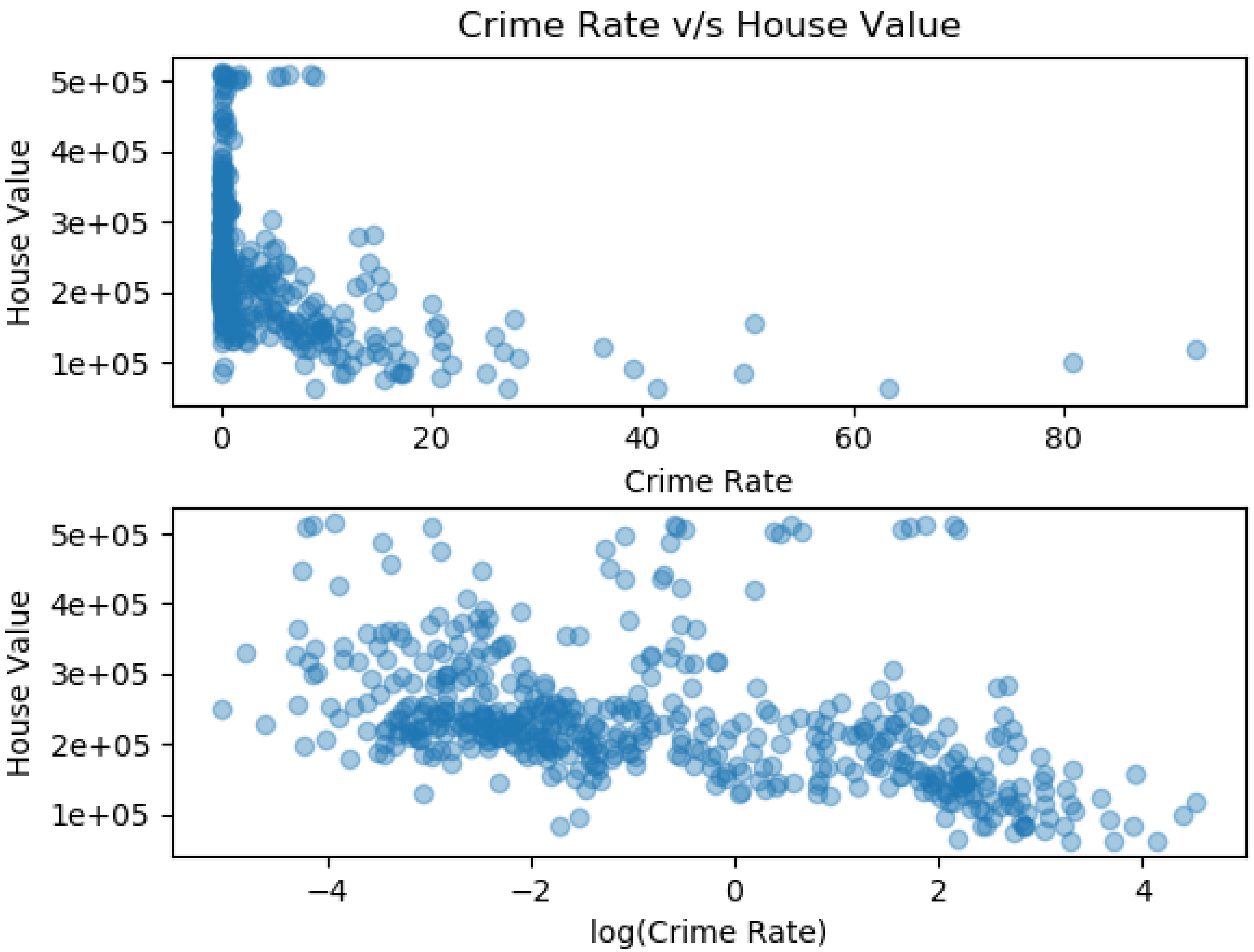
CORRELATION MATRIX



Findings

- House Value is correlated to Crime Rate in the negative direction suggesting that the price of a house decreases with an increase in the crime rate
- House Value is correlated to Number of Rooms in the positive direction suggesting that the price of a house increases with an increase in the number of rooms
- House Value is negatively correlated to Student Teacher Ratio suggesting that houses where students per teacher is low are priced higher
- Property Tax Rate is highly correlated to Accessibility To Highway in the positive direction
- Nitric Oxides is highly correlated to Distance to Employment Center in the negative direction

RELATIONSHIP BETWEEN CRIME RATE AND HOUSE VALUE



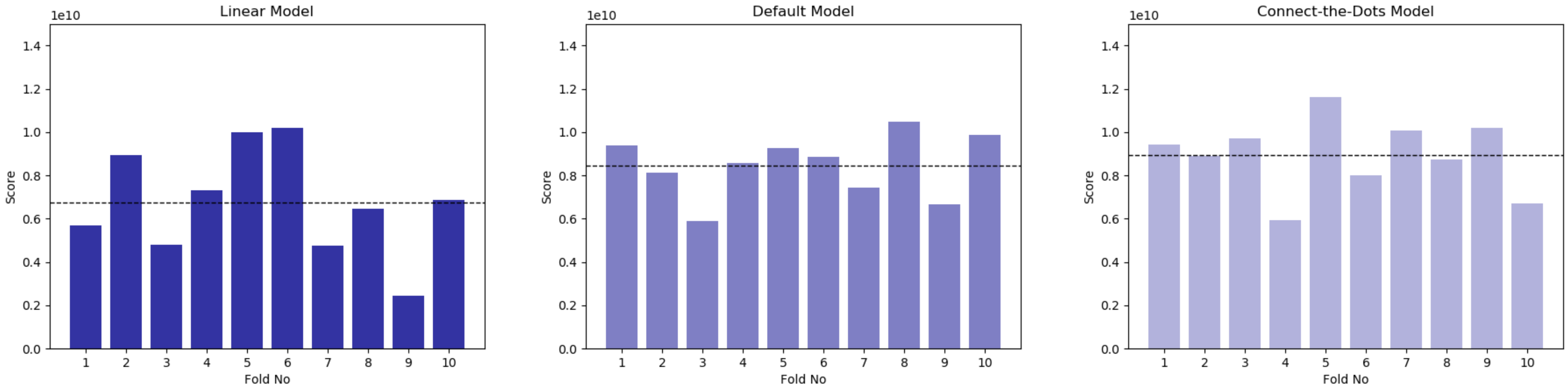
Findings

- The scatter plot of Crime Rate and House Value shows that Crime Rate is highly skewed and therefore the points are densely located in region
- On transforming Crime Rate data using log transformation, we notice an almost linear relationship between log(Crime Rate) and House Value with a negative slope
- It shows that houses located in regions with high crime rate are priced lower than houses located in regions with low crime rate

FURTHER ANALYTICAL QUESTIONS

- Are there any outliers in the data?
- Does Charles River Bound influence the house value? If so, how?
- Is living close to work costlier than living away from work?
- What are the long term effects of Property Tax Rate on house value?
- Is the crime rate higher in toxic/polluted neighborhoods?
- What is more influential in determining house value: student-teacher ratio or distance to work?

K-FOLD CROSS VALIDATED MODELS for K=10

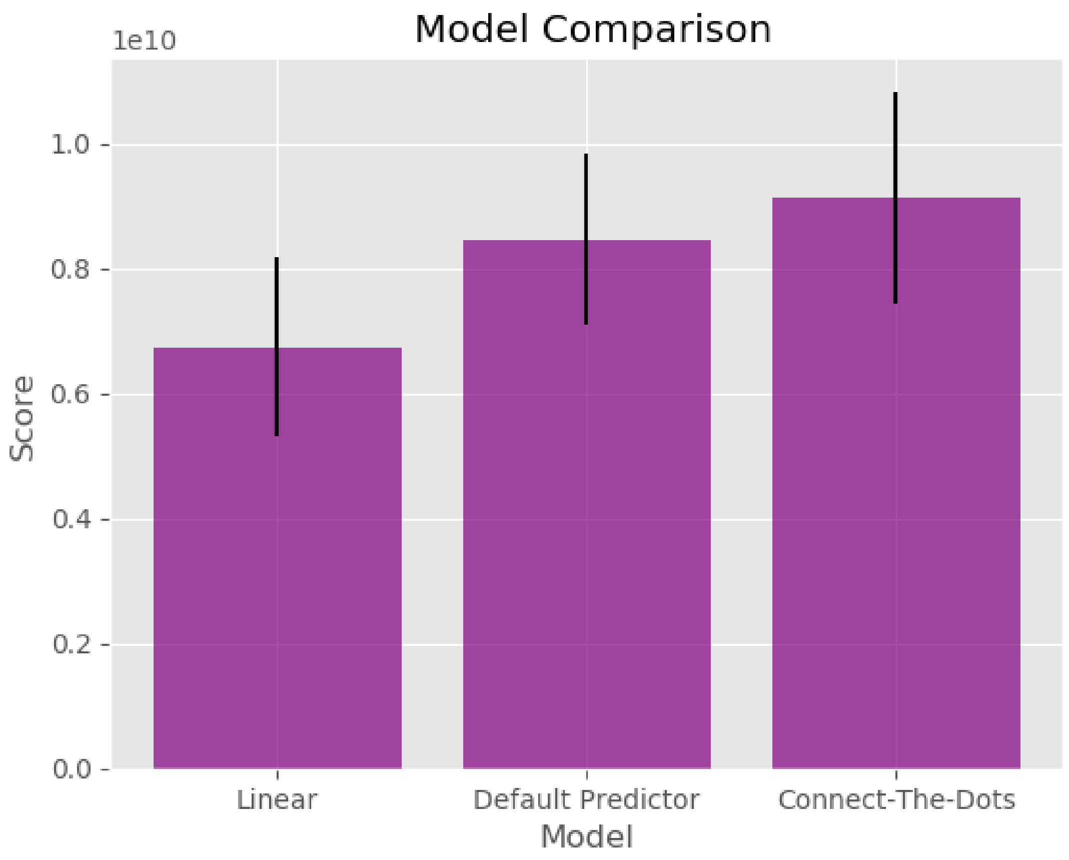


- The bar charts represent the Mean Squared Error (MSE) for each K in K-Fold Cross Validated Models for three modelling techniques
- The dashed line in each chart represents the mean of all the MSE bars

Findings

- Linear Model has the least error, followed by Default Predictor Model and the existing Connect-the-Dots Model

LEAVE-ONE-OUT MODEL COMPARISON



Findings

- On implementing the Leave-one-Out cross validation technique on the three models, we observe that Linear Model consistently returns lower mean of MSE
- The lines represent confidence intervals i.e a range of values which contains the population mean in 95% of instances.

Recommendation:

As the linear model results in the least error amongst the three models, we can say that it predicts the house value closest to the training data. Therefore it is recommended that it be used in the valuation system.