Univ.AI

Deep Dive of Housing in India

Sahil Joshi Chetan Jagadeesh Pulkit Pradeep Gupta Kshitiz Pradeep Gupta

Motivation and About the Project

- To gather insights of the current housing conditions in India
- To use these insights and predict the Housing Price in major metropolitan cities
- Calculation of Housing Quality of Living Index.
- Use 40 different amenities to predict the housing prices.
- Multiple models like Decision Tree, Random Forest, XGBoost where used and evaluated

Data and Labels

- Total 640 csv files were scraped from the website using Beautiful Soup.

- These datasets were then merged into a master dataset which was used to gather insights from.

- The datasets for the training and predictions was gathered from the Kaggle. All the different metropolitan cities data was merged before splitting into training and testing data.

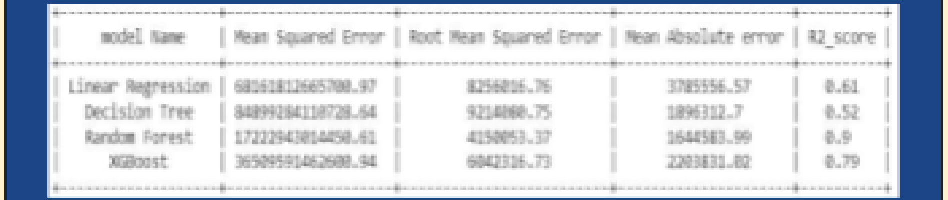
References

The dataset used for the EDA was gathered from the 2011 Census data from https://censusindia.gov.in.

[1] Das, Bhaswati & Mistri, Avijit. (2013). Household Quality of Living in Indian States: Analysis of 2011 Census. Environment and Urbanization Asia. 4. 151-171. 10.1177/0975425313477755 paper was referred for feature creation

Model and Results

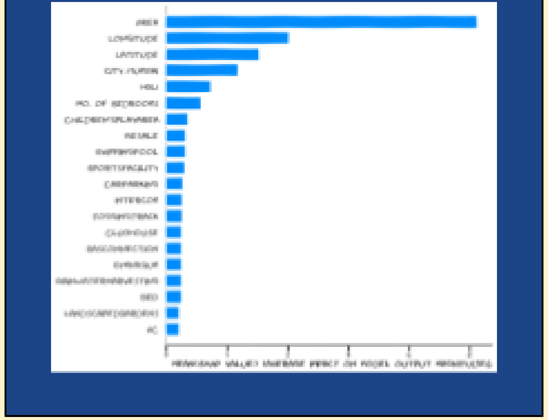
- Four Models : Linear Regression, Decision Tree, Random Forest and XGBoost where constructed to predict the housing prices.
- Random forest and XGBoost models where optimized to achieve good results.
- We have compared the models over different regression metrics like MSE,RMSE,MAE and R2Score.



```
+-------------------+----------------------+-------------------------+---------------------+----------+
|     model Name    |  Mean Squared Error  | Root Mean Squared Error | Mean Absolute error | R2_score |
+-------------------+----------------------+-------------------------+---------------------+----------+
| Linear Regression | 68161812665700.97    |        8256016.76       |      3785556.57     |   0.61   |
|   Decision Tree   | 84899284110720.64    |        9214080.75       |      1896312.7      |   0.52   |
|   Random Forest   | 17222943014450.61    |        4150053.37       |      1644583.99     |   0.9    |
|      XGBoost      | 36509591462600.94    |        6042316.73       |      2203831.02     |   0.79   |
+-------------------+----------------------+-------------------------+---------------------+----------+
```

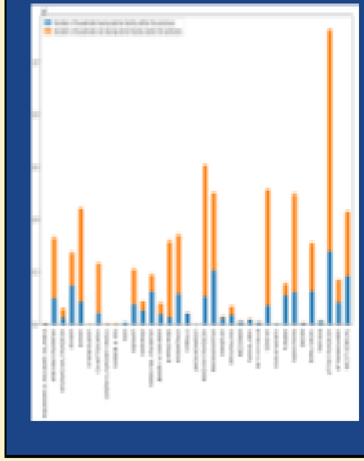
Observations

- Top features that models like Random Forest and XGBoost used for prediction are:

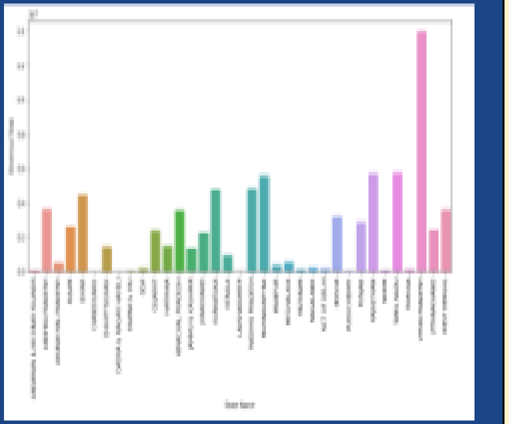


Explanatory Data Analysis

Latrine facility across state



Electrical household items usage



Conclusion and Future Work

- We would like to explore more models like lightgbm and Neural Networks.
- We would try Stacking or cascading techniques and evaluate the performance of such methods.
- Probably spend more time creating more features for better prediction.