ISyE 6740 – Fall 2020
Final Report

Team 82: Yang_Hsuan-Han, 903565940

Project Title: Study of housing valuation by investigating correlation between housing price and housing features

Problem Statement

Purchasing a property can be a very important decision people have to make in their lives whether for living or investing purposes, and they need to take a considerable amount of time and efforts to gather enough relevant information to come up with the best bargain given the property's valuation and features. Instead of being lured by real estate agents into purchasing properties at a high cost, it is important for people to gather enough housing information and data from credible sources, and perform data analysis to keep track of the latest regional housing price and have a fair price range in mind while making offers based on the features of properties.

In this project, the goal is to provide insights to housing valuation by regressing on the history housing data and investigating the correlations between housing features and price. Therefore, this project plans to tackle the formulated hypothesis which is whether there are significant correlations between certain housing features and housing price using various modelling techniques. By employing and comparing multiple models, we may be able to observe the pattern of dataset and the prominent housing features that may have high influence to housing price, and discuss the underlying reasons for variation in terms of model performance.

Data Source

Data is collected from a previous Kaggle[1] competition hosted by Zillow. Data consists of information on housing features such as bathroom count, bedroom count, size of property in square ft, garage square feet, lot size in square feet, year built, number of stories and etc. Data includes housing information in the US from 2016 to 2016. Data source is collected in the form of csv file that contains a total of 58 columns and 5,970,432 records.

Since the purpose of this project is to perform exploratory data analysis using various modelling techniques, in order to avoid any potential issues during data collection such as insufficient computational power, and to narrow down the scope of project, the data to work with is preprocessed to drop unused columns, remove records with missing information (price). The final data to work with consists of 11 columns and 90,275 records, with 10

[1] Zillow. (2017). *Kaggle.* https://www.kaggle.com/c/zillow-prize-1/data

predictors including number of stories, lot size, property size, property tax, land tax, total tax, bedroom count, bathroom count, garage size and year built, with response variable being sale price.

Methodology

The project consists of three parts, namely data collection, data processing and data analysis. Data was collected from Kaggle which contains raw csv file. Preliminary data cleaning, casting and sorting were done to ensure the data consists of relevant housing information, and there are no missing or dirty data that may lead to error during analysis or misleading conclusions. Furthermore, since dataset were collected from the same source with multiple tables, data was concatenated by joining tables containing price information and property information using property id.

Prior to data analysis, data was randomly split into training and testing samples with 80-to-20 ratio in order to evaluate model performance in the latter part. For data analysis, statistical techniques including linear regression, K-nearest-neighbor regression (KNN), decision tree regression, random forest regression and neural network were performed to investigate the correlation between predicting variables including housing features and response variable which is price of property. Reason for choosing these techniques is to compare various regression models' performances and understanding the underlying reasons for the variation in performances.

By performing linear regression on training samples, we obtained the strength of correlation between housing features and price to observe which housing feature has the most prominent impact on property price. Afterwards, we used the trained model to predict future value and used the test samples to validate the accuracy of the trained model using mean absolute error (MAE). Similar steps were taken for other types of regression to see which model best fit the data and investigate the underlying reasons (i.e. concepts of the chosen model) for such result.

In addition, to investigate the difference in terms of performance between linear and non-linear classifiers for fitting the given dataset, techniques including random forest, decision tree, KNN and neural network were employed using training samples, and resulting models were used to predict future values as well. Validation were done using test samples to obtain a MAE score for the model.

By comparing the results between regression and random forest, we can observe the variation of performance of models given the dataset, investigate whether dataset fits better against linear or non-linear classifiers, and discuss the underlying concepts of models and the potential reasons that may lead to certain results. Finally, exploratory data visualization is done by plotting results for each model to observe and compare model behavior given dataset, and discuss the conclusion as to whether there exists correlation between housing features and housing price, and what are the most suitable model to utilize given the dataset.

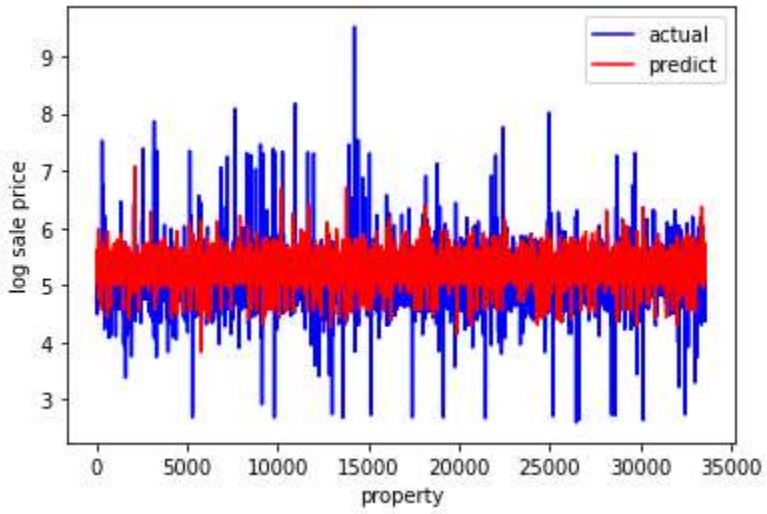
Evaluation and Final Results

Following is a preview of processed dataset to work with:

ofstories	lotsizesquarefeet	calculatedfinishedsquarefeet	structuretaxvaluedollarcnt	landtaxvaluedollarcnt	taxvaluedollarcnt	bedroomcnt	bathroomcnt	garagetotals
1.437764	7528.0	1684.0	122754.0	237416.0	360170.0	3.0	2.0	347.6039
1.437764	3643.0	2263.0	346458.0	239071.0	585529.0	4.0	3.5	468.0000
1.437764	11423.0	2217.0	61994.0	57912.0	119906.0	2.0	3.0	347.6039
1.437764	70859.0	839.0	171518.0	73362.0	244880.0	2.0	2.0	347.6039
2.000000	6000.0	2283.0	169574.0	264977.0	434551.0	4.0	2.5	598.0000

A total of five models were computed using the training samples. The first linear regression model is compiled using all 10 housing features as predictors and sale price as response variable. The results and test statistics are as follows:

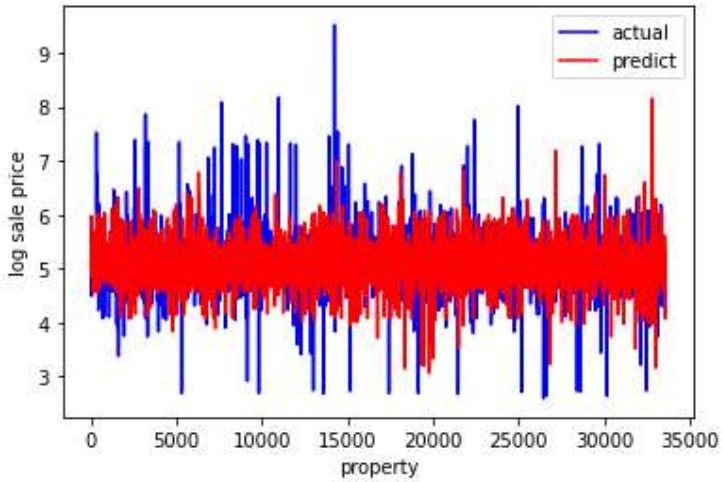
```
Mean Abs Err:  0.13611350600977665
Mean Sq Err:  0.04066478105133046
```



After fitting the data with training samples and using the testing samples to evaluate the accuracy of predicted values, the mean absolute error is around 0.1361 with the actual vs. predicted value shown in the graph above. Sale price is scaled so the plot is legible.

The second K-nearest-neighbor regression model is compiled using 5 predictors (number of stories, garage size, bedroom count, bathroom count and year built) which yield the lowest MAE after iterations of trial. The results and test statistics are as follows:

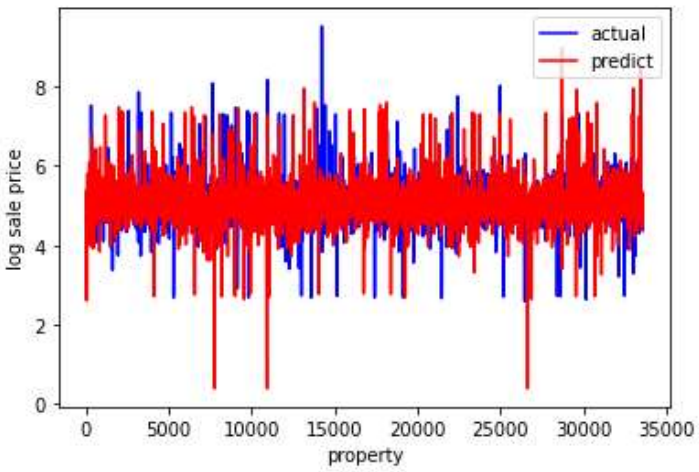
```
Mean Abs Err:  0.11868855861576613
Mean Sq Err:  0.04746701867693105
```



Using similar way to fit the data with training samples with n=5 and using the testing samples to evaluate the accuracy of predicted values, the mean absolute error is around 0.1187 with the actual vs. predicted value shown in the graph above. Notice that the MAE of kNN with non-linear classifier is lower than linear regression, suggesting that the data may work better with non-linear classifier given the selected 5 predictors.

The third decision tree regression model is compiled using 6 predictors (lot size, property size, total tax value, bedroom count, garage size and year built) with the lowest MAE. The results and test statistics are as follows:

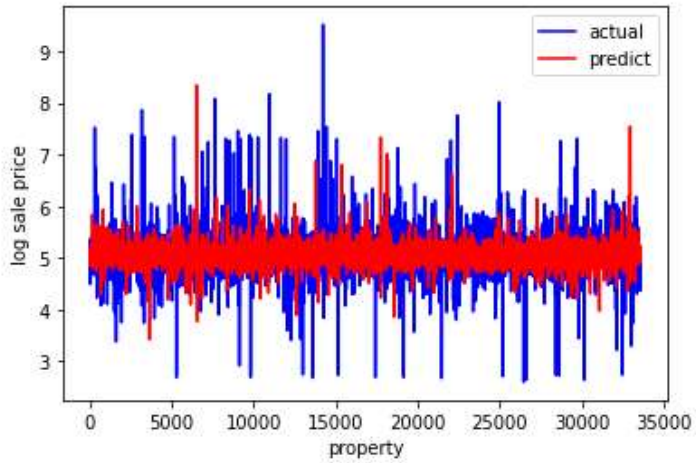
```
Mean Abs Err:  0.11705014384862579
Mean Sq Err:  0.059939241558329
```



The result suggests a MAE of around 0.1171 which is lower than linear regression and kNN. Similarly, non-linear classifier seems to outperform linear classifier, suggesting a non-linearity pattern within the data with the 6 predictors.

The fourth random forest regression model is compiled using the same 6 predictors (lot size, property size, total tax value, bedroom count, garage size and year built) with the lowest MAE. The results and test statistics are as follows:

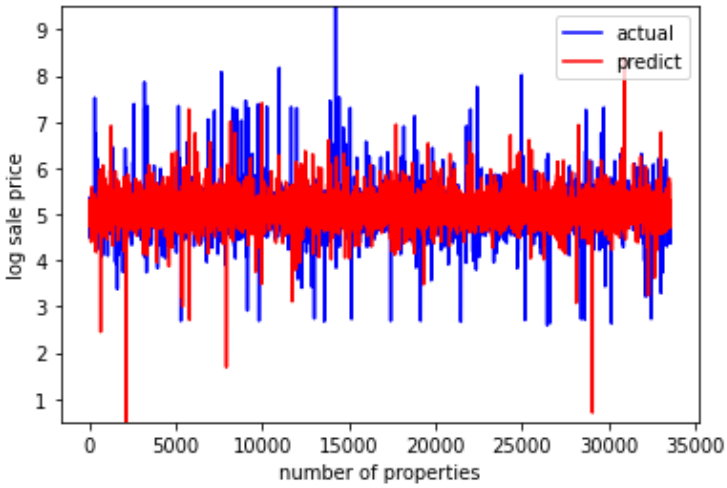
```
Mean Abs Err:  0.09403684783270932
Mean Sq Err:  0.03396017574968688
```



Notice the random forest model yields a MAE of around 0.0940 which is lower than decision tree, kNN and linear regression, and thus suggests a better accuracy and possibly better fit for the data. This is possibly due to the grouped decision trees that better captured the non-linearity of the data comparing to other non-linear classifiers.

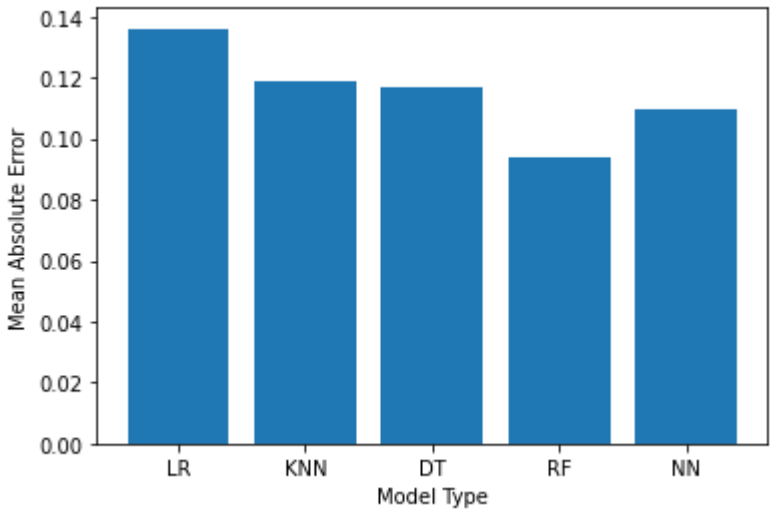
The fifth neural network model is compiled using the 7 housing features as predictors (lot size, property size, number of stories, property tax value, total tax value, garage size, year built) with the lowest MAE. Data is scaled prior to model implementation to ensure computing efficiency. The results and test statistics are as follows:

```
Mean Abs Err:  0.1093337732815457
Mean Sq Err:  0.048380066992038485
```



The MAE for neural network is around 0.1093 which is between random forest and decision tree. While the results are relatively close, random forest outperforms neural network in this case. With more test samples and tuning/scaling of the model, neural network should be able to yield better results.

The overall performance comparison is plotted in the following bar chart.



Overall, random forest with predictors of lot size, property size, total tax value, bedroom count, garage size and year built and response being sale price has the smallest mean absolute error, suggesting a relatively better fit for the dataset. Models with non-linear classifier generally performs better than linear classifier, suggesting the non-linearity pattern of the data which is expected given the relatively large dimension of dataset. For future studies, a more sophisticated approach to data processing and scaling such as PCA or SVD may be performed to process higher dimensional dataset, ensure better application and make further inferences to more types of model.