ORIE 4740 FINAL PROJECT REPORT

House Price Prediction in King County

Po-En Tsai <pt369>

Jiahui Xu <jx265>

Yuchen Zhu <yz2226>

May 17, 2017

ABSTRACT

A well-constructed housing price prediction tool would provide real estate guidance for people who either plan to buy a property or list their own for sale. The objective of this study is to develop a statistical model that predicts sales prices of the houses in King County based on 2014-2015 sales data and the external data from Zillow. The tools used included Python, R and Excel. The two statistical models used were linear regression and decision tree. The linear regression residual plot validated the linear assumption, but yielded relatively high MSE and low explained variance. Alternatively, the random forest tree model was used and outperformed the regression model. The model suggested strong significance of geographic house features, as well as space and quality of construction. The tuned model yielded lower mean squared error and higher explained variance despite some outliers. In general, the prediction gets more precise towards the low-priced houses. In fact, most outliers are high-priced houses that were difficult to predict with only 18 variables from the data set. 80% of the prediction errors were within 100K USD and 60% within 50K. In terms of percentage, the majority predictions had low errors around 15% of the actual prices. For real estate guidance purpose, the prediction is robust enough to provide valuable information for homebuyers and property sellers since the errors are relatively small in terms of the percentage of the predicted prices in that area. Furthermore, clustering approach was used to group the data that are prone to underestimated predictions. Any new test data that fall into such clusters would be adjusted in prediction in addition to the random forest model. The clustering approach would be more robust as the given data set get larger. Overall, the combination of the models can provide price prediction, the under-predicted groups, and the approach to compensate for the underestimation.

INTRODUCTION AND GOAL

This project mainly focused on supervised learning regarding prediction of house prices. There were in total 18 variables from the main data set used to make predictions, plus zip code-based Zillow Home Value Index. The prediction methods included feature engineering, linear regression and random forest decision tree. K-fold cross-validation was used to validate the test error and check for overfitting. The result evaluation approaches included outliers and prediction errors confidence interval analysis. The two models were compared in term of the mean squared errors and percentages of explained variance. The random forest model outperformed the linear regression model with 1.58E10 as the test MSE and 88.3% as R-square. The prediction errors were within about 30% and 15% of the actual prices under 95% and 75% confidence intervals respectively. The principal component analysis was conducted to define clusters with high likelihood of price underestimation. The data with high test errors fell into the predefined clusters. Those groups were underestimated by 10% in general. The predicted prices could be compensated with price adjustments by the same amount.

DATA SET

Main Data Set

The data set was acquired from Kaggle as a comma-separated spreadsheet. It includes 21613 observations and 18 features plus the price and the sales ID columns. The house features include sales date, built year, number of rooms, square footage, and location coordinates etc. Preliminary check for missing or error-like values was conducted in Excel. As a result, none was found.

External Data Set: Zillow

The zip codes usually correspond to house price level in those areas as shown below.

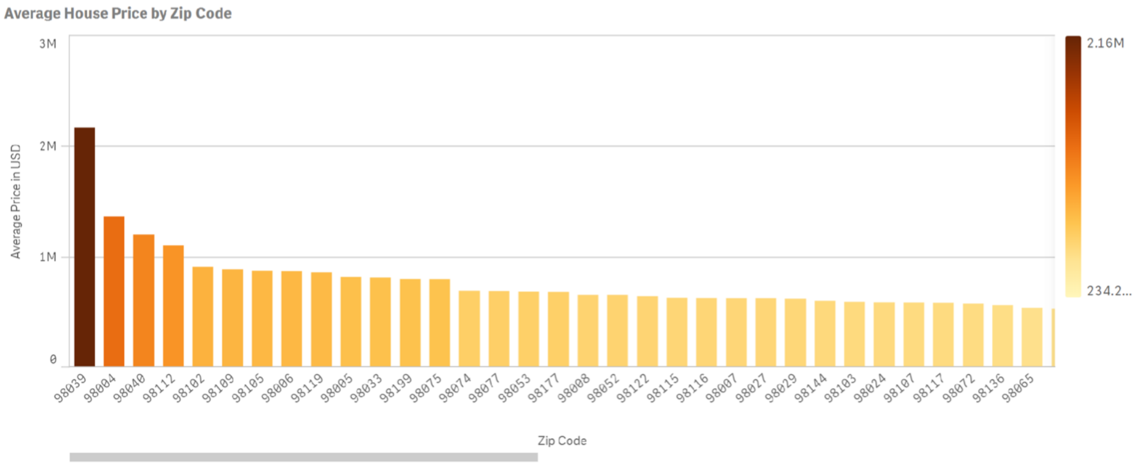


Figure 1: Zip Code Effects

Zillow Home Value Index provided an estimation of the house price level corresponding to the zip codes. The time-series Zillow data set generates the indices for a given neighborhood in a given month. The indices were integrated into the main data set based on listed year, month and zip code.

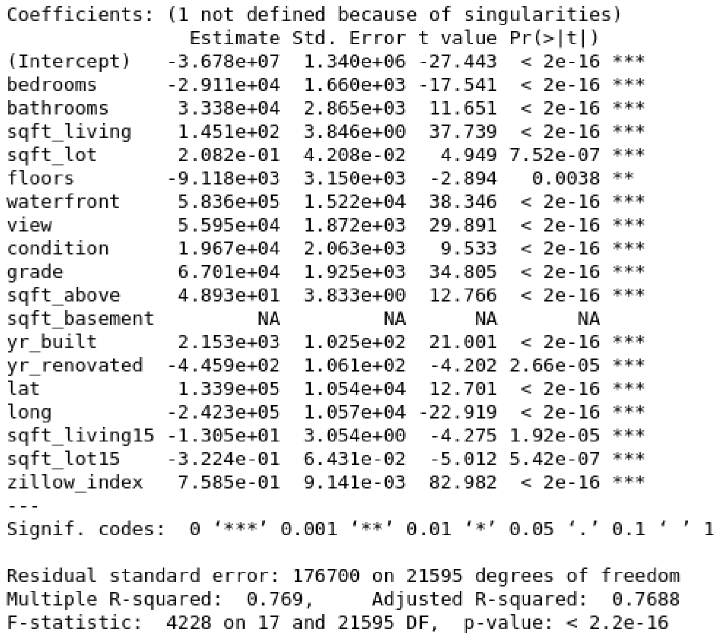
Feature Engineering

The “yr_built” column was transformed into the duration between the built year and the sold year. The “yr_renovated” column was transformed into the duration between the year of renovation and the year of sales. The zip code column was transformed into the Zillow Index as mentioned above. All three columns are numeric in terms of format.

LINEAR REGRESSION

The data was fitted into a linear model. The output from the model is shown below.

Table 1: Linear Regression Summary



```
Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.678e+07  1.340e+06 -27.443  < 2e-16 ***
bedrooms      -2.911e+04  1.660e+03 -17.541  < 2e-16 ***
bathrooms      3.338e+04  2.865e+03  11.651  < 2e-16 ***
sqft_living    1.451e+02  3.846e+00  37.739  < 2e-16 ***
sqft_lot       2.082e-01  4.208e-02   4.949 7.52e-07 ***
floors        -9.118e+03  3.150e+03  -2.894   0.0038 **
waterfront     5.836e+05  1.522e+04  38.346  < 2e-16 ***
view           5.595e+04  1.872e+03  29.891  < 2e-16 ***
condition      1.967e+04  2.063e+03   9.533  < 2e-16 ***
grade          6.701e+04  1.925e+03  34.805  < 2e-16 ***
sqft_above     4.893e+01  3.833e+00  12.766  < 2e-16 ***
sqft_basement         NA         NA      NA       NA
yr_built       2.153e+03  1.025e+02  21.001  < 2e-16 ***
yr_renovated  -4.459e+02  1.061e+02  -4.202 2.66e-05 ***
lat            1.339e+05  1.054e+04  12.701  < 2e-16 ***
long          -2.423e+05  1.057e+04 -22.919  < 2e-16 ***
sqft_living15 -1.305e+01  3.054e+00  -4.275 1.92e-05 ***
sqft_lot15    -3.224e-01  6.431e-02  -5.012 5.42e-07 ***
zillow_index   7.585e-01  9.141e-03  82.982  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 176700 on 21595 degrees of freedom
Multiple R-squared:  0.769,     Adjusted R-squared:  0.7688
F-statistic:  4228 on 17 and 21595 DF,  p-value: < 2.2e-16
```

The adjusted R-square is 0.7688. The training mean square error was calculated to be 3.11E10. The 10-fold cross validation error was 3.13E10. The MSEs from training data and CV data suggested low likelihood of overfitting. Most of the predictors were significant as they had low p-values. As the residual plot shown below, the variance was non-constant, including a few outliers with large residuals.

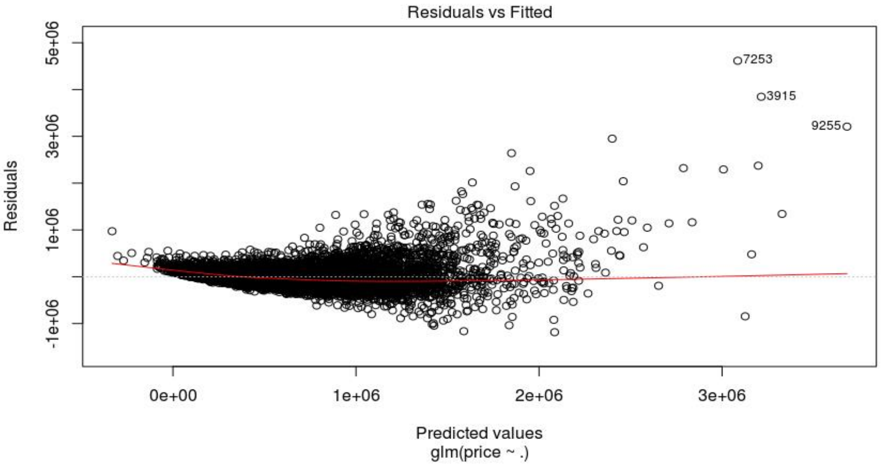


Figure 2: Linear Regression Residual Plot

RANDOM FOREST

Unlike the linear regression model, the random forest decision tree model de-correlates the trees through random selection and takes the average from many trees. Prior to tuning the model, an initial run was conducted to see the general performance compared to the linear regression model. The data set was randomly split into half as a training set and a test set. The MSE and R-square produced by the test set was 1.60E10 and 87% respectively, which outperformed the linear regression model by a large margin.

MODEL SELECTION

As summarized in the table below, the random forest model proved to be more promising in this prediction project. It was thus selected as the final model for development.

Table 2: Model Results Comparison

Model	**Linear Regression**	**Random Forest (Preliminary)**
Adjusted R-square	0.76	0.87
Test MSE	3.13E10	1.60E10

MODEL DEVELOPMENT AND IMPROVEMENT

Model Parameter Selection Analysis

The random forest model was tuned in terms of the number of random features and trees. The “mtry” selection was evaluated in terms of mean squared error (MSE) and explained variance (R^2) from the test set. The candidates included “m=18” which is the number of total predictors. The other three were “m=12”, “m=9”, and “m=5”. Even though the results did not show large separations between the selections, “m=9” showed a slight edge over the rest. The test error started to stabilize once the number of trees became larger than 100.

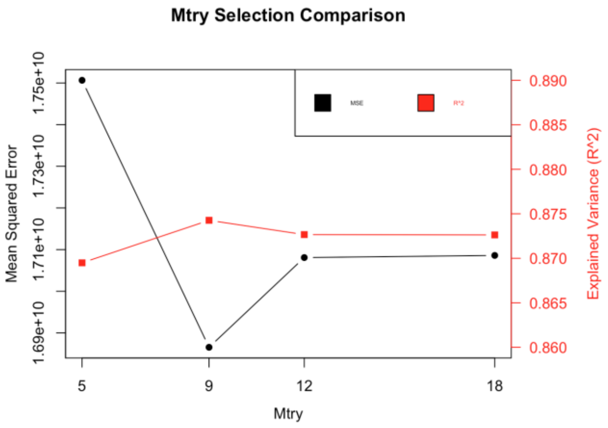


Figure 3: Number of Features Comparison

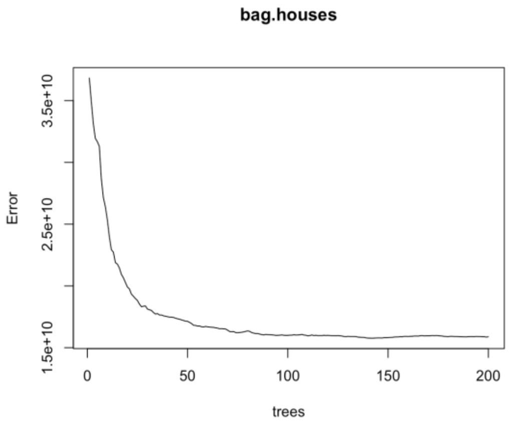


Figure 4: Number of Trees Comparison

Test Errors and Outliers Analysis

The model was then tuned with number of random features equal to 9 and number of trees equal to 200. K-fold method was used to perform cross-validation on the whole data set with K equal to 10. With these set parameters, the model was run again and produced the test error plot as shown below.

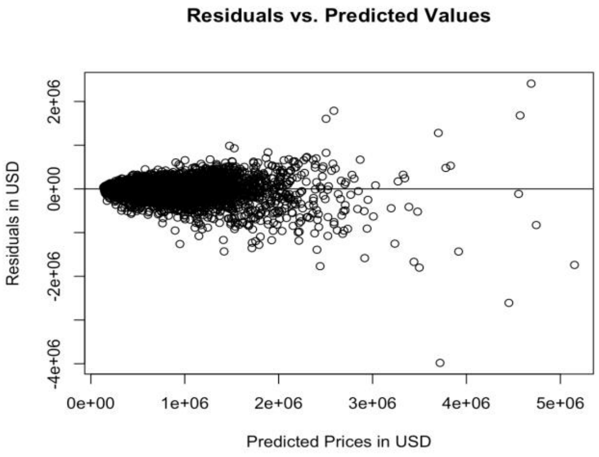


Figure 5: Random Forest Model Residual Plot

In general, the model had better performance on regular-priced houses. As the house prices became higher, the model tended to underestimate the values. Graphically speaking, the far-off-line observations were more likely to be below the solid line. The prediction with the largest test error (at the bottom of the plot) was listed as 7.7 million and was underestimated by almost 4 million. A further investigation suggested that the feature descriptions match the actual house features. It is in an expensive area with nice views. This outlier was concluded to be realistic rather than an error-like observation. The test error was likely caused by the limitation of the data set itself. The test mean squared error from the test set was calculated to be 1.583E10. And the average OOB error from the training sets was 1.581E10 and the training set had 2.7E9 MSE. Despite the outliers, the model still performed well with 88.3% explained variance.

Feature Importance Analysis

As expected, the Zillow external data set contributed to the prediction significantly. Like the Zillow index, the longitude and latitude that correspond to geographic features were also significant in prediction improvement. “Sqft_living” and grade also proved to be important features which correspond to living area and quality of house construction respectively. Below are the rankings of features in terms of effects of MSE and node purity (training RSS).

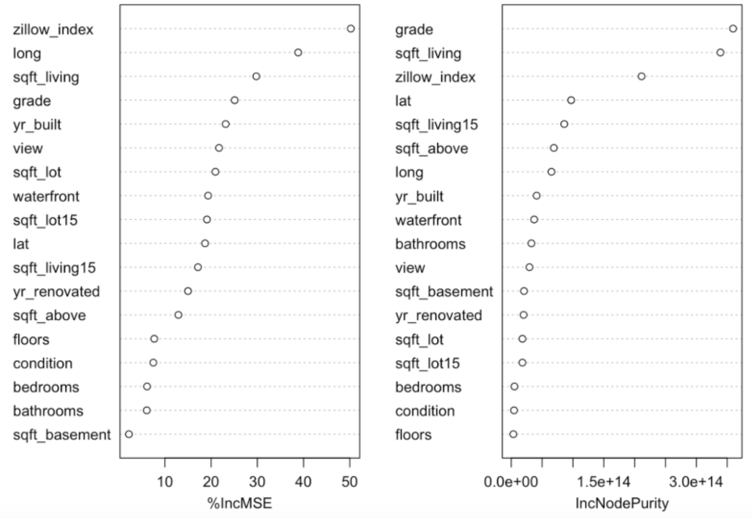


Figure 6: Features Ranking

Model Result Evaluation

The test set residuals were plotted onto a histogram as shown below. The distribution has much more density near zero. Most residuals are within +/-100K USD.

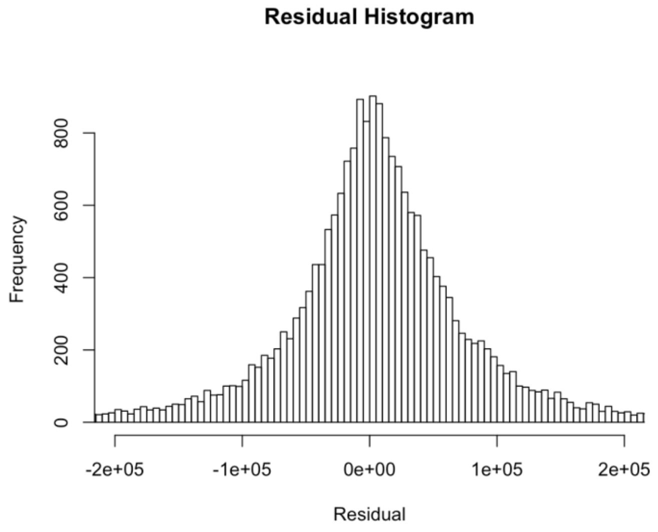


Figure 7: Residual Histogram

For each prediction, the error percentage metric was defined as the equation shown below.

error percentage = (test error /predicted value) * 100%

Even though the test errors on the histogram did show a clear trend, it was not a traditional normal distribution. Instead, quantiles of the test error distribution were used to estimate the confidence interval. 10.9% of the observations had extremely low or high predicted values, among which there were not enough data to clearly define a distribution.

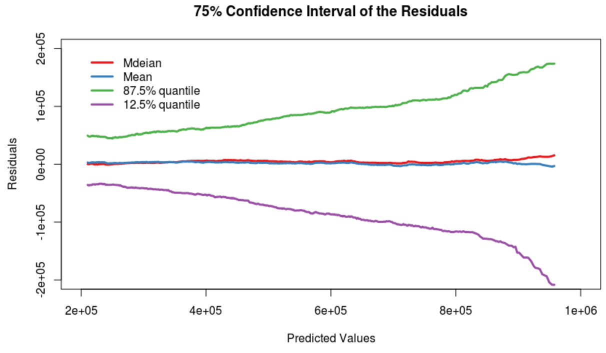


Figure 9: 75% CI of Residuals

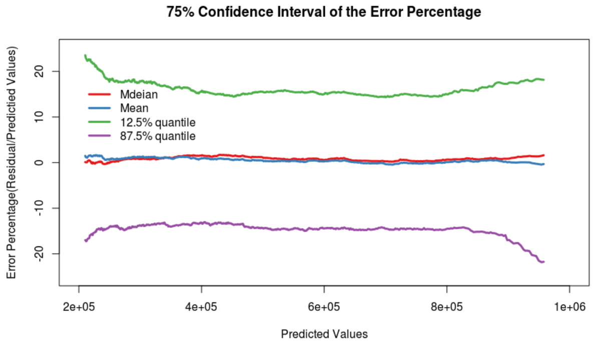


Figure 10: 75% CI of Residual Percentages

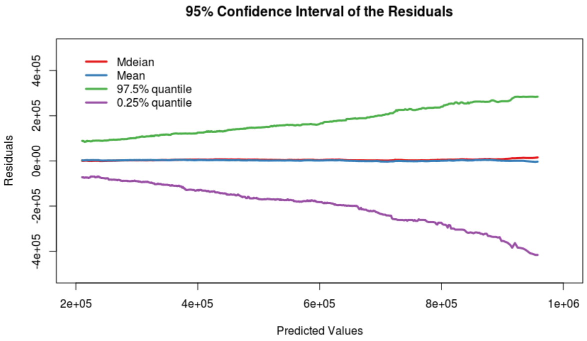


Figure 11: 75% CI of Residuals

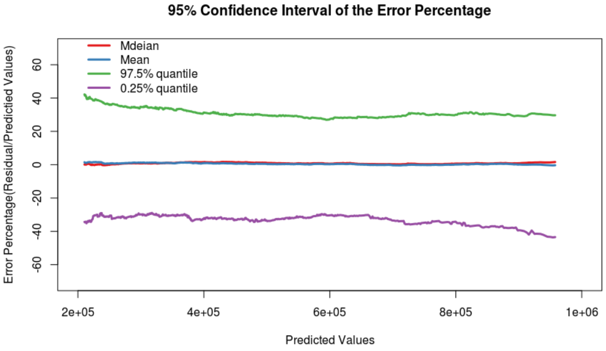


Figure 12: 75% CI of Residual Percentages

Higher predicted values had higher test errors. But the errors in terms of percentage, were almost constant throughout the all the test data. The performances on different sections of the test data were recorded in the table below.

Table 3: Confidence Intervals of Different Price Ranges

Range of Predicted Values	256,681(+- 10%)	354,465(+- 10%)	461,126(+- 10%)	591,774(+- 10%)	878,880(+- 10%)
Number of Observations	2078	3069	3488	3206	1382
95% CI lower	-32.216996%	-30.8529960%	-32.446686%	-29.6251785%	-37.827359%
95% CI upper	36.44726633%	32.67755036%	30.26179526%	27.32607786%	29.87274133%
75% CI lower	-13.9530253%	-13.9036520%	-13.555905%	-14.614453%	-15.5171236%
75% CI upper	17.93017664%	16.1194738%	14.79991402%	15.04731104%	17.3533267%

CLUSTERING: AN ATTEMPTION TO IDENTIFYING HIGH ERROR PRONE TEST DATA

The purpose of clustering approach is to provide additional information of the expected test errors in addition to the random forest model. The principal component analysis (PCA) was conducted to capture more characteristics of the observations. Then K-means was used to cluster data with similar characteristics. The findings suggested，by using the 9th and 11th PCs and Choosing K=10, the model could form two clusters with test errors higher than 10% of the predicted prices. They are clusters 3 and 9 as shown in figure 13.

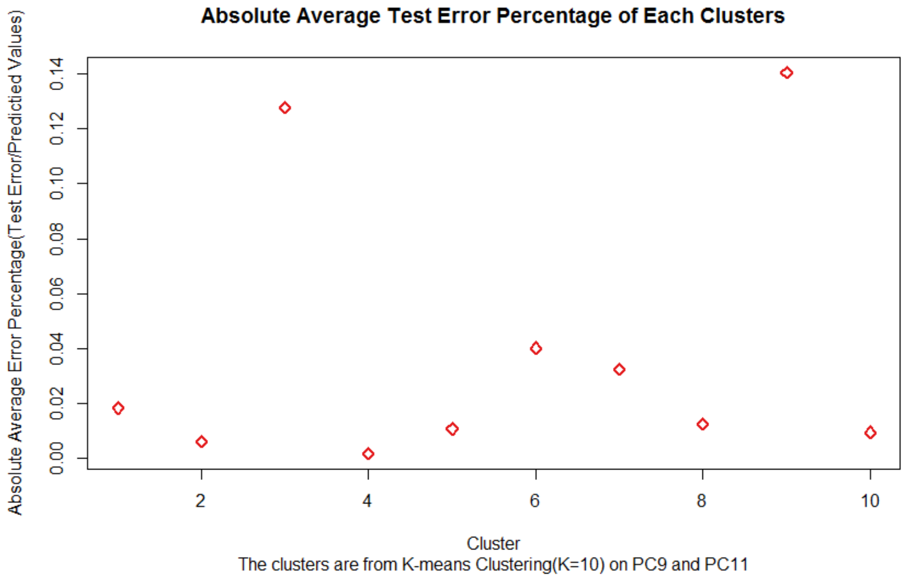


Figure13: Average Test Errors of Different Clusters

Figure 14 was plotted to show where the high-test-error data were distributed.

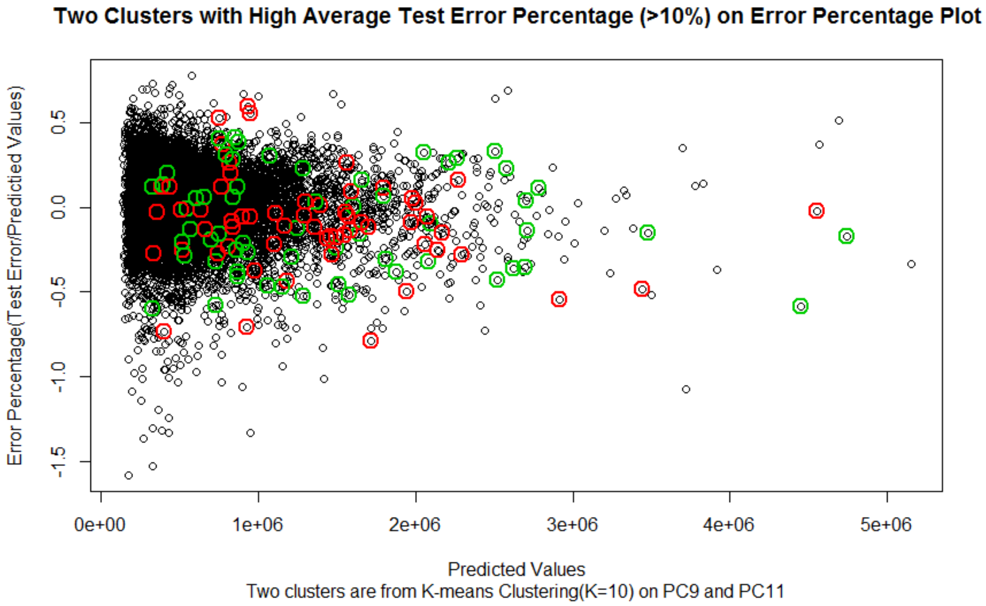


Figure 14: High Test Error Clusters Distribution on the Residual Plot

Based on the findings mentioned above, the hypotheses could be made that K-means clustering on the training set can help find high test error data from the test set. The framework is shown in figure 16.

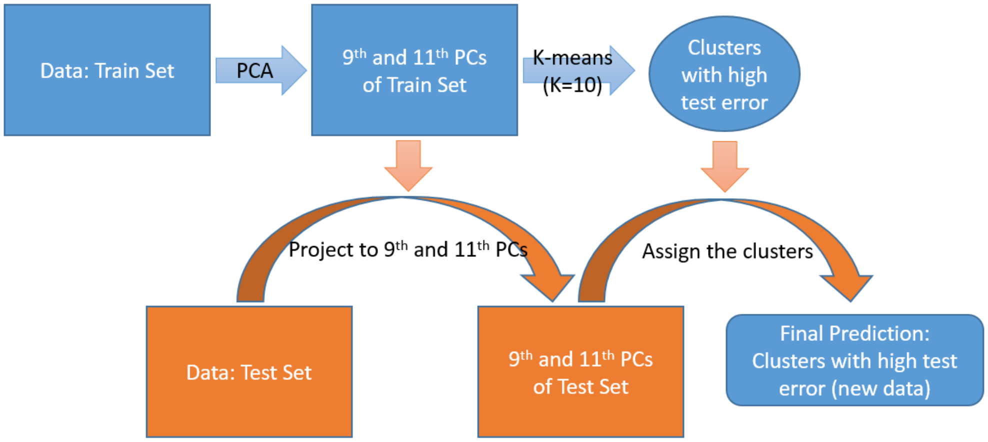

Data: Train Set
PCA
9th and 11th PCs of Train Set
K-means (K=10)
Clusters with high test error
Project to 9th and 11th PCs
Assign the clusters
Data: Test Set
9th and 11th PCs of Test Set
Final Prediction: Clusters with high test error (new data)

Figure 17: Framework of the hypotheses

10-fold cross validation was conducted using the same process. In each iteration, there was a cluster with high average test errors from the training set and a predicted clusters from the test set. As the result, the average test errors of the clusters from test set were slightly lower than the test error percentage form the train set (>-5%), as shown in figure 17. This result then was compared with the distribution of the test error percentage from all the observations in figure 16, which has mean value 0.0044 and median value 0.0086.

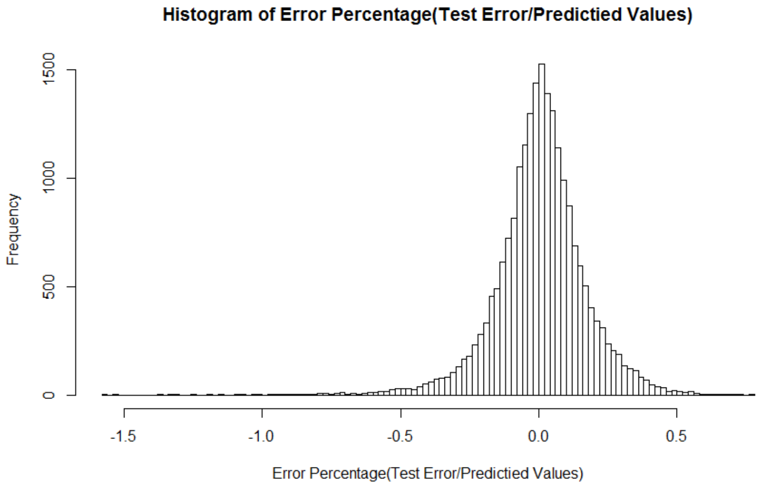


Figure 16: Histogram of Test Error Percentage

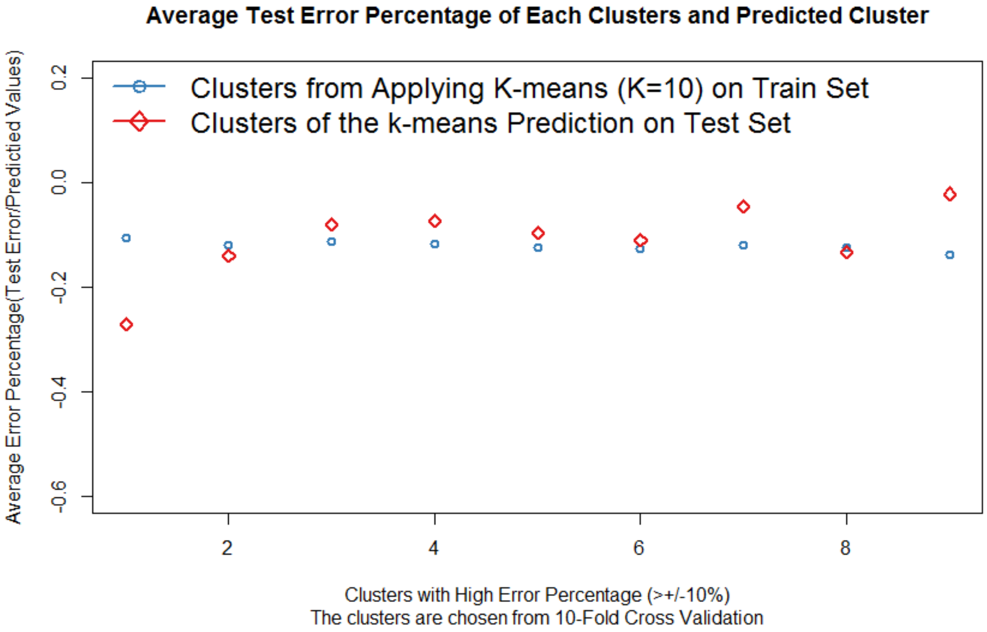


Figure 17: Average Test Error Percentage Each Cluster vs. Predicted Cluster

It was concluded that PCA and K-means clustering may help improve the confidence interval of the prediction model. However, it is difficult to quantitatively show the impact on the prediction improvement due to lack of observations. Each of the four clusters only contained about 15 to 30observations. More data would help illustrate the impact of this method.

SUMMARY

In our analysis, we first used linear regression model to predict the housing price, and then using the random forest model with clustering approach to adjust the underestimated predictions. With 18 predictors in the dataset, the linear regression model, compared to the random forest model, produces a larger MSE for both the training data and the test data and a smaller adjusted R-square. Thus, we picked the random forest model as our final model. To further improve the precision of our tree model, we conducted model parameter selection analysis, residuals and outlier analysis, feature importance analysis, and model result evaluation. We also performed PCA and K-means clustering to analyze test error data distribution. As a conclusion, the random forest model in our case outperforms the linear regression with a much lower test error and variance.