Predicting Melbourne House Prices from 2016

MATH2319 Machine Learning Applied Project Phase 2

Sangyoon Lee (s3727080)

March 31, 2018

Contents

1 Introduction

2 Methodology

3 Feature Selection and Ranking

4 Hyperparameter Tune-Fining

    4.1 Naïve-Bayes

    4.2 Random Forest

    4.3 k-Nearest Neighbour

    4.4 Threshold

5 Evaluation

6 Discussion

7 Summary

1 Intoduction

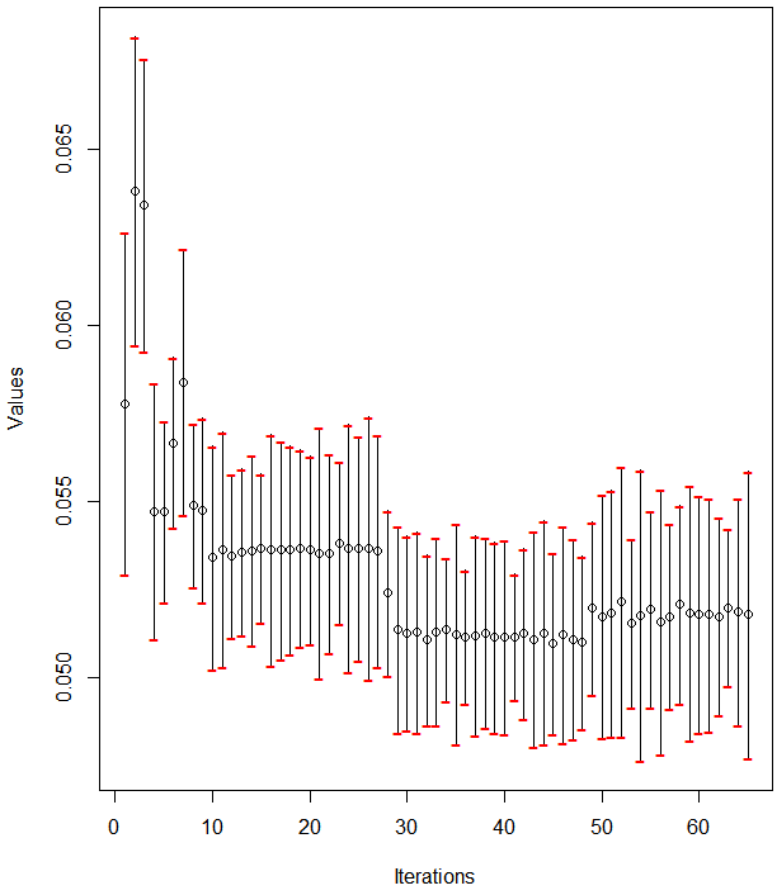
The objective of this project is to build classifiers to explore relationships of Melbourne house prices which are over 1,000,000 AUD (1 million AUD) or less in a year from January of 2016 Melbourne, Australia. House Prices in Melbourne have been greatly changing over past few years. The data sets were sourced publicly available results posted from Domain.com.au, collected by Tony Pino at kaggle.com. In phase 1, we preprocessed the data in ready to use to lessen data redundancy attributes and less granular. In phase 2, we built three binary classifiers on the data. Section 2 describes our methodology. Section 2 describes feature selection method. Section 4 discusses the three classifiers, fine-tuning process and detailed analysis. Section 5 compares performance of the classifiers using resampling method. Section 6 discusses limitation of our methodology through result in section 5 evaluations. In last section, summarizes the overview of our project.

2 Methodology

Firstly we performed feature selection and ranking to see importance of each descriptive features. Then we considered three different classifiers: Naive Bayes (NB), Random Forest (RF), and k-Nearest Neighbor (KNN). We split up a preprocessed dataset from previous phase into 2 sets, 70% data into training set and 30% into test set. These data are divided in random, and expected to behave similar proportion of over 1,000,000 AUD (1M AUD) houses and below 1M AUD houses. Approximately 60% of houses are below 1,000,000 AUD and 40% are above 1,000,000 AUD. Each sets have same proportion of house prices.

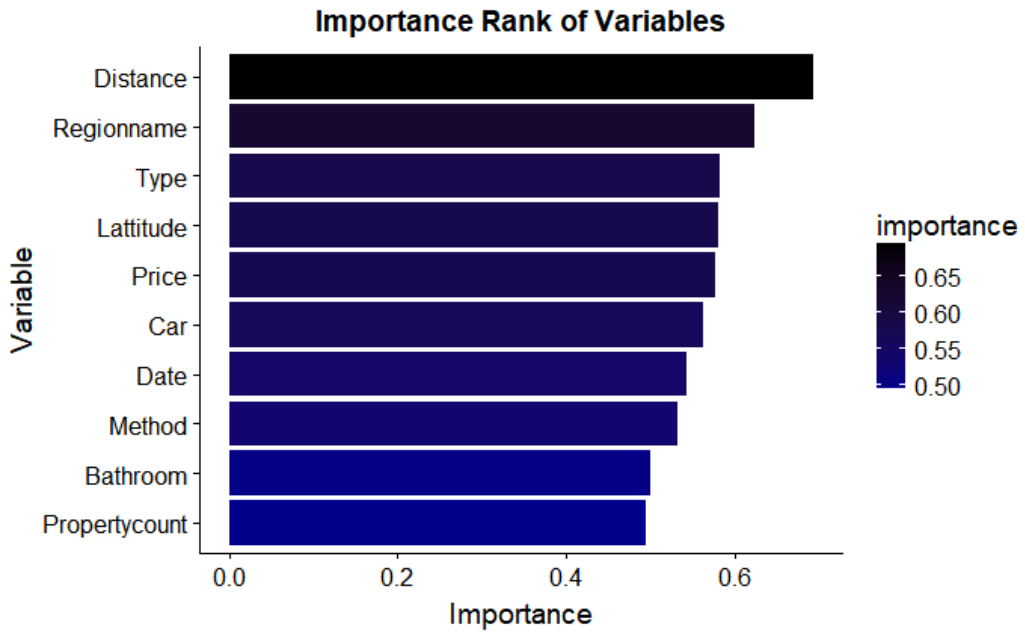
3 Feature Selection and Ranking

We used spFSR package for feature selection and ranking. With the stochastic nature of the SPSA-FSP algorithm, 65 iterations were made to find the most optimal number of iterations for feature selection. Naïve-Bayes model was used.




```
## Best iteration = 45
## Number of selected features = 10
## Best measure value = 0.05097
## Std. dev. of best measure = 0.00257
```

From the graph, bars indicates the mean accuracy rate. The optimal iteration is 45 times with 10 selected features.



Referring to the graph, distance is the most important variable related to determine the housing price. And general regions (West, North West, North, and Metropolitan) follows the next importance rank. Property counts in the suburb and number of bathrooms are the least important variables.

```
## Error % with 10 best MLR features =   5.6
## Error % with full set of features =   7.3
## Error % with 10 best spFSR features = 5.1
```

Error rate is lower with spFSR, compare to 10 best MLR features and full set of features.

4 Hyperparameter Tune-Fining

4.1 Naïve-Bayes

We tested mean errors with Laplacian parameter in range of 0 to 30 with step side 1, the grid search to determine the optimal value of Laplacian smoothing parameter. The optimal Laplacian parameter we obtained was 0 with a mean error of 0.1135147

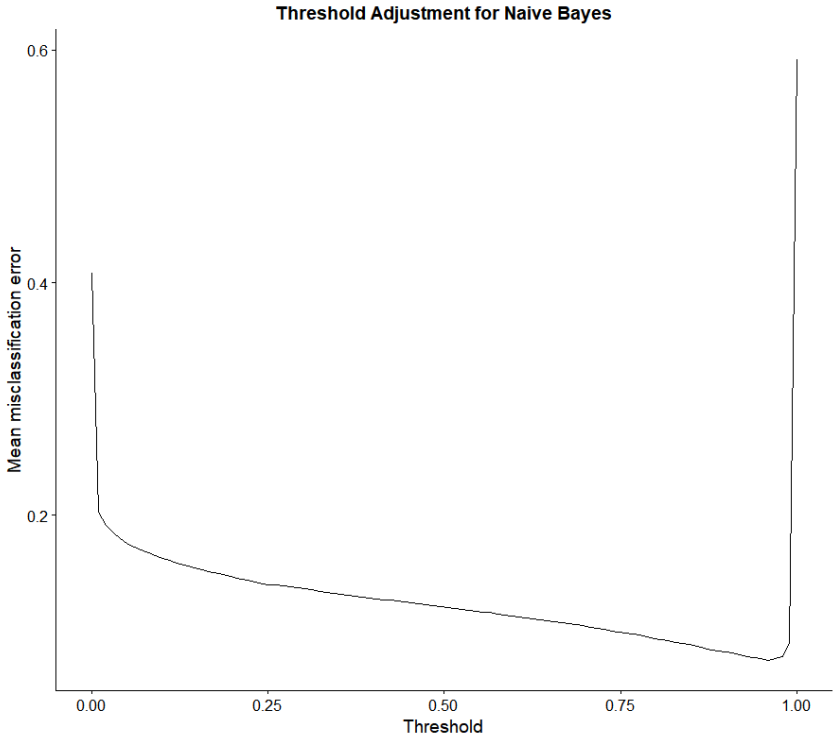
4.2 Random Forest

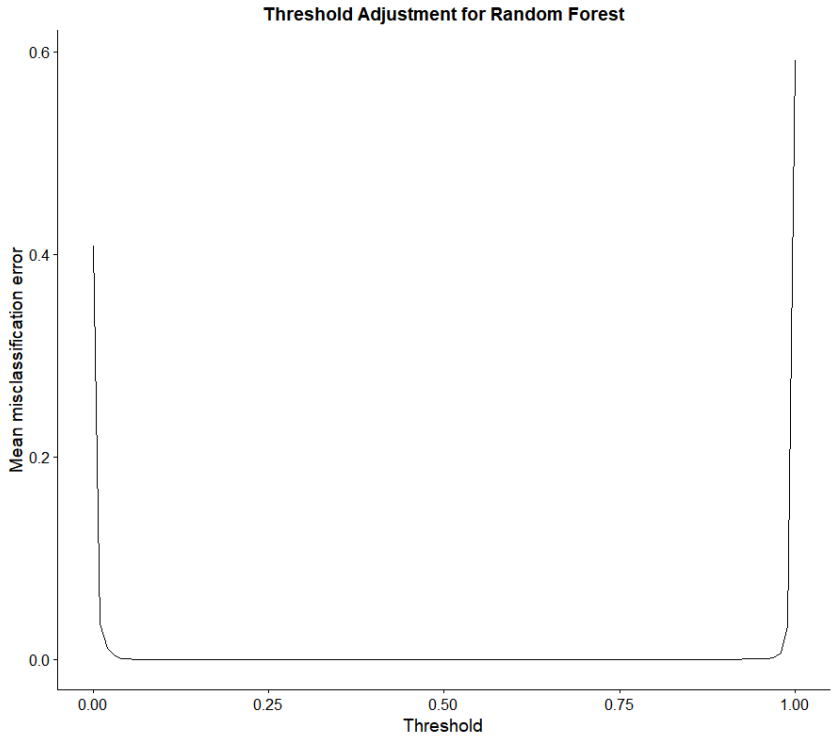
Breiman (2001) suggests that, $\text{mtry} = \sqrt{p}$, where p is the number of descriptive features. We have 10 descriptive features and our suggested mtry is $\sqrt{\text{p}} = \sqrt{10} = 3.16228$. Therefore we decided to test mtry with 2, 3, 4 and 5. The test provided mtry = 3 with mean test error 0.0000524.

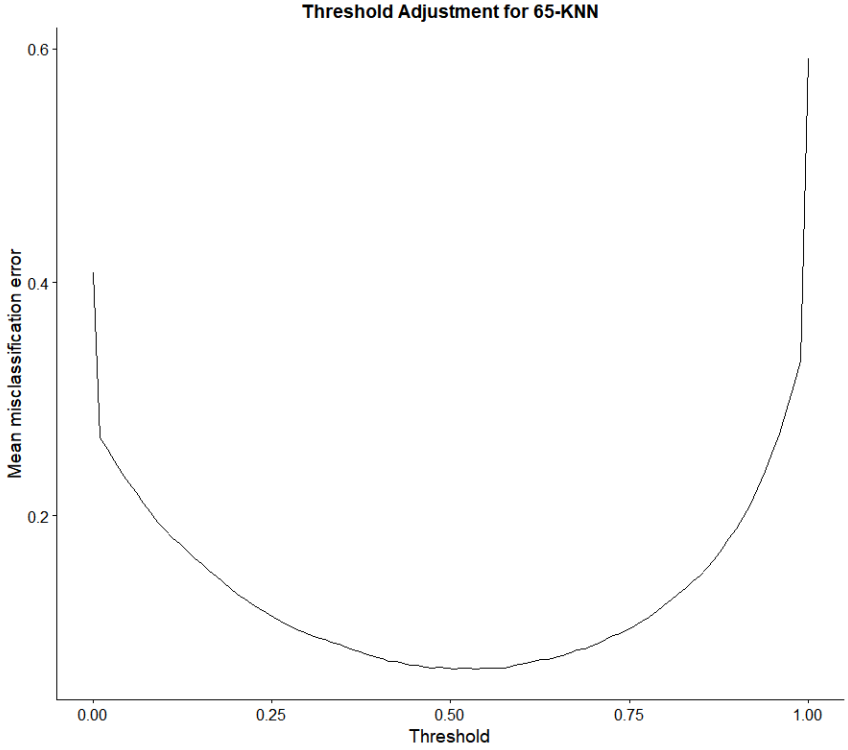
4.3 K-Nearest Neighbour

By using the optimal kernel, we ran a grid search on k = 2 to 65 with step side 1. The outcome was k = 31 with a mean test error of 0.0950082

4.4 Threshold Adjustment




Threshold Adjustment for Random Forest
Mean misclassification error
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Threshold


Threshold Adjustment for 65-KNN
Mean misclassification error
0.6
0.4
0.2
0.00
0.25
0.50
0.75
1.00
Threshold

5 Evaluation

The confusion matrix of Naïve-Bayes classifier

```
##Relative confusion matrix (normalized by row/column):
##        predicted
##true     < 1M      >= 1M     -err.-
##  < 1M   0.05/0.78 0.95/0.65 0.95
##  >= 1M  0.01/0.22 0.99/0.35 0.01
##  -err.-      0.22      0.61 0.62


##Absolute confusion matrix:
##        predicted
##true     < 1M >= 1M -err.-
##  < 1M    421  4985   4985
##  >= 1M   121  2648    121
##  -err.-  121  4985   5106
## -err.-  121 4985  5106
```

The confusion matrix of Random Forest

```
##Relative confusion matrix (normalized by row/column):
##        predicted
##true     < 1M      >= 1M     -err.-
##  < 1M   0.06/0.71 0.94/0.55 0.94
##  >= 1M  0.03/0.29 0.97/0.45 0.03
##  -err.-      0.29      0.55 0.53


##Absolute confusion matrix:
##        predicted
##true     < 1M >= 1M -err.-
##  < 1M    513  4129   4129
##  >= 1M   207  3326    207
##  -err.-  207  4129   4336
## -err.-  207 4129 4336
```

The confusion matrix of k-NN classifier.

```
##Relative confusion matrix (normalized by row/column):
##        predicted
##true     < 1M      >= 1M     -err.-
##  < 1M   0.04/0.71 0.96/0.61 0.96
##  >= 1M  0.03/0.29 0.97/0.39 0.03
##  -err.-      0.29      0.61 0.60


##Absolute confusion matrix:
```

```
##        predicted
##true     < 1M >= 1M -err.-
##  < 1M    220  4788   4788
##  >= 1M    92  3075     92
##  -err.-   92  4788   4880
## -err.- 92 4788 4880
```

Our classifiers generally shows the critically row TPR and high FNR. This means that all classifiers accurately predicted individual earning higher than 1M houses, but predicted inaccurately for below 1M houses. Based on our evaluation, we concluded that RF classifier which was the best, showed the least error rate in our classifiers.

6 Discussion

Referring to the previous section, all classifiers did not performed accurately with low priced houses. The main issue causing inaccurate result in our sample data, is the data set contains NA values and numerical features are replaced to mean value in our preprocessing. Applying better replacing method would improve biases in predicting below 1M houses.

Our prediction is biased as predicting price over 1M besides of the actual price. To mention our limitation in methodology, we used stratified sampling which caused imbalance class problem.

RF classifier is advantaged classifier compare to NB and KNN classifiers and showed the best performance. NB classifier assumes the distribution of descriptive features is normal nonetheless the distribution is not. Since latitude is not appropriate for neither categorical nor numerical data, that may cause KNN classifier underperforming. However, RF classifier performed comparably well because this classifier had advantage in having more samples to improve the accuracy of prediction.

7 Conclusion

In our methodology: Naïve-Bayes, Random Forest and K-NN Neighbor, Random Forest provided the best prediction. Firstly we divided data set into training set and test set via stratified sampling. Then we performed three different methods to calculate each error rates. Overall misclassification was significantly

high in classifying below 1M houses. There is possibility for reinforcing our methodology by having better preprocessing and sampling method.

References

Breiman, L. 2001. “Random Forests” Machine Learning.