Are You Paying the Right Price for Your Property?

Mihir P. Gandhi

Northeastern University

Week 6: Project Presentation

03/27/2019

ALY6040- Data Mining Applications [CRN: 21288]

Instructor: Prof. Justin Grosz

Contents

Introduction

Buying a new house is a milestone in any person's life as years as one may end up investing a huge chuck of their savings or tend to get housing loans from financial institutions. Thus, knowing the right amount for the property is highly critical for any person. Let us assume you wish to buy a property in an area where you have no idea of the surroundings and the property prices and trends. Naturally, you would tend towards getting the deal brokered from a local realtor. There could be a possibility that, the realtor would try to dupe you buy inflating the property price which would be an unnecessary burden on your finances. Having an estimate beforehand would prove highly beneficial for such situations. Thus, with the help of this project, a potential buyer would be able to estimate the property price in the King County area of Seattle. Similar models could be constructed with the help of property data present in that area.

Creating a model which would predict the price of the houses involves initial cleaning of data, exploring the trends of the data, creating statistical models, eliminating unwanted variables and finally creating a model which is significantly high in accuracy.

Data Key:

Variable	Description	Variable	Description
id	Notation for the house	sqft_living sqft_lot	Square footage of living and lot area
date	Date the house was sold	waterfront	If house has a view to the waterfront
price	Price at which the house was sold	condition	How good is the condition of house
bedrooms bathrooms	Number of bedrooms and bathrooms in the house	grade	Overall grade given based on the King County grading scale
sqft_above sqft_basement	Square footage of house above and at basement	yr_built yr_renovated	Year the house was built or renovated
zipcode	Postal Zipcode of the property	lat long	Geographical location of the house
sqft_living15	Square footage of house after 2015	view	View of the house

Analysis

Dataset

The data set extracted from Kaggle contains the past property sale data in King County, Seattle. The data set includes 21.6k entries and 21 variables. Amongst those 21 variables, 19 variables describe the features of the house such as no. of bedrooms, no. of bathrooms, floors etc. and the other two variables are the price of the house and the transaction ID. Since, we are creating a model to predict the price of each house, we set “price” as the dependent variable for analysis.

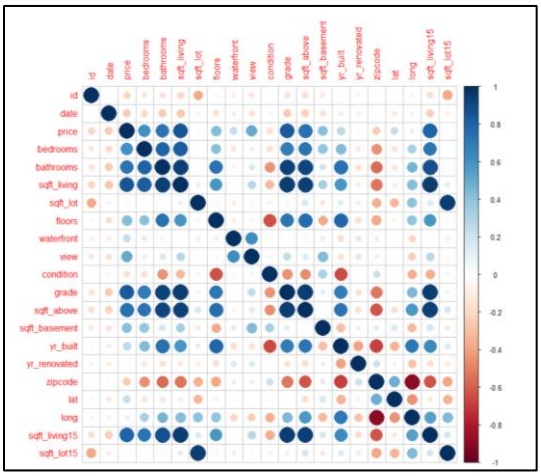
Data Cleaning

Data cleaning was performed in three major steps: NA or missing values, outliers and typos. The number of NA values for each column was checked and the results showed that none of the variables in the data set contained any NA values. The missing values were checked in Excel by searching for “blanks” and no missing values were found in the data set. The variable “date” contained junk values “T000000” in every entry and were subsequently removed from the variable. The outliers were checked using the boxplots and the variables were grouped by the range of the values as the outliers would be easily interpretable. On observing the plots, some outliers were observed but since they were not at a greater distance from the cluster of data points, they were not removed from the data set. One clear outlier was observed in the “bedrooms” boxplot. A data point was observed at a value of 33 which seemed irrelevant to the row corresponding to it when compared. The data frame containing the value of 33 bedrooms would have had more resemblance if the number of bedrooms were 3. Thus, taking the value as a typo, the no. of bedrooms was changed from 33 to 3 and was added to the data set. The variable “yr_renovated” contained more than 95% of value which were 0. Thus, due to this high

percentage, the variable was removed. The variable “bedrooms” was changed from character to integer as it only contained numeric values.

Correlation Matrix

To gain further insights for the variables present in the dataset, a correlation matrix was constructed which is shown below.



As “price” was our dependent variable, we checked the correlation of “price” with the other variables. The variables which were not significantly correlated with “price” were removed from the data set. Thus, the variables “id”, “date”, “sqft_lot” and “sqft_basement” were removed from the data set. Moreover, since the data from 2015 was not relevant for our analysis as there was not much renovation done in one year, we removed the variables “sqft_living15” and “sqft_lot15” from the dataset. Thus, 7 of the 21 irrelevant variables were removed from the data set by just observational inferences.

Linear Modelling

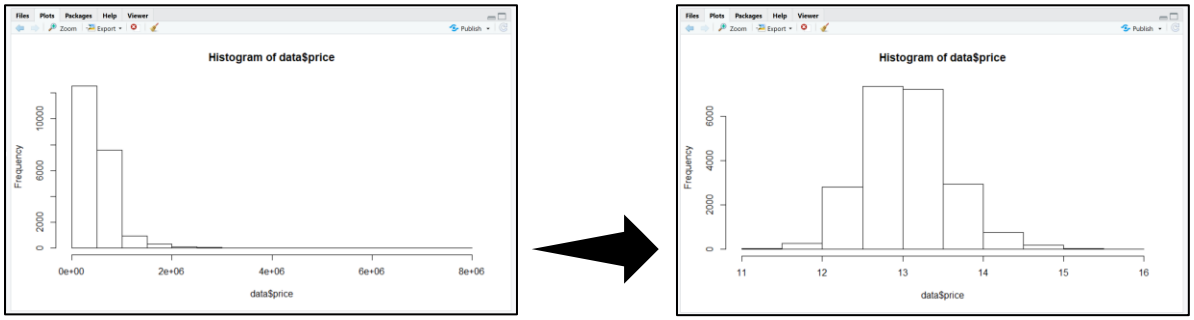
As the data was now clean and free of redundant variables, we performed the linear modelling techniques on the data set. The data set was split into train and test models under the ratio of 80:20. The linear model was created for the training model which yielded a R2 value of 70.12. R2 is a measure of the accuracy of the model which meant that the initial linear model for

the training data set was 70.12%. The variable “floors” was seen to have low statistical significance to the model due to a higher p value. The overall p-value of < 2.2e-16 proved that the dependent and the independent variables were statistically significant. With the significant nature of the model, the test data set was linearly regressed which yielded a R2 value of 69.21 which meant that the model was 69.21% accurate. Like in the training model, “floors” was not significant due to the higher p-value. Also, “sqft_above” was seen to be insignificant to the model. To check the multicollinearity of the variables, the VIF (variance inflation factor) values were computed and since none of the values were seen to be above 10, no independent variable was overfitting to the model. As the accuracy of the test model was considerable, a predictive model was created which gave the value of price of $478,366.2 for the input values of independent variables.

```
> ## creating predictive model
> house1 <- data.frame(bedrooms = 3, bathrooms = 2, sqft_living = 1800, floors = 1, waterfront =
 0, view = 1, grade = 8, condition = 4, sqft_above = 1800, yr_built = 1981, zipcode = 98178, lat
 = 47.5112, long = -122.257)
> price <- predict(lm_modeltest, house1)
> price
       1
478366.2
>
```

Optimized Linear Modelling

To further optimize the model, the values of the dependent variable “price” were converted to their logarithmic equivalent values as the distribution of the values was right skewed. With the help of logarithmic transformation, the data followed close to a normal distribution.



Also, the variables “floors” was removed from the data set as it did not provide any statistical significance to the linear model created initially. Thus, the with the same ratio of train and test data sets, a linear model was initially created for the training data set. The results from the training data set yielded a R2 value of 76.27 which was 6.15% higher than the initial training model. Applying the log transformation reduced the range of the data points which made the dependent variable normally distributed, leading to a higher accuracy. All the variables in the model had high statistical significance based on the p-values and the overall model was also significant as the p-value was less than 2.2e-16. Applying the modelling technique to the test data set, we obtained a R2 value of 75.72 which was 6.51% higher than the original test linear model. Also, none of the values in the VIF analysis were over 10, which meant that no one of the variables were overfitting to the data. Thus, we created a refreshed predictive model with a higher accuracy.

```
> ## creating predictive model for new linear model
> house1 <- data.frame(bedrooms = 3, bathrooms = 2, sqft_living = 1800, waterfront = 0, view = 1
, grade = 8, condition = 4, sqft_above = 1800, yr_built = 1981, zipcode = 98178, lat = 47.5112,
long = -122.257)
> price <- predict(lm2_test, house1)
> price
       1
12.97786
>
```

The value 12.97786 was the logarithmic value of the original price and thus, by taking the antilog of the value, we arrived at a value of $432,726. The model with a higher accuracy gave the price of the house $45,640.2 less than the original predicted value. There was a significant change in the predicted value when the accuracy of the model was increased by 6.5%.

Stepwise Regression and Gradient Boosting Model

To gain further insights on the significance of the variables and improve the overall accuracy of the model, the model was regressed stepwise in the backward direction to gain insights on the insignificant variables still present in the data set. The stepwise regression

technique proved that there are no more insignificant variables in the data set and none of the variables must be dropped.

Now, to further improve the accuracy of the model, gradient boosting technique was used with an initial iteration level of 10,000 for the training data set. We observed that, there were no variables with zero influence on the model "lat" variable had the maximum influence in predicting the price of the house. Thus, the location in which the house lies contributes highly affects the value of the price of the house in the given area. Further, we observed that, for the training data set, to predict the values with the minimum value of squared error loss, 2,744 iterations were required for the analysis. These number of iterations would also prevent the model from being overfitted. Thus, the values were predicted with the optimal number of iterations and comparing the predicted values to the original values, we found that, the predicted values were almost equal to the original values. Similar techniques were applied to the test data set and the observations concerning the variable were similar. Since, the gradient boosting model is a self-learning model, the number of iterations required to correctly predict the value of house price were 1,182 for minimum value of squared error loss.

Thus, with the help of this data, a final predictive model was created with the same values as seen in the previous models to compare the change in price. The data frame was readjusted based on the relative influence of the variables based on the gradient boosting model.

```
> ## creating predictive model for gradiant boosting model
> house1 <- data.frame(lat = 47.5112, sqft_living = 1800, grade = 8, long = -122.257, view = 1,
zipcode = 98178, yr_built = 1981, sqft_above = 1800, condition = 4, waterfront = 0, bathrooms =
2, bedrooms = 3)
> price <- predict(data.boost_test, n.trees = 1182, house1)
> price
[1] 12.87244
>
```

The value obtained was the logarithmic value of the price. Thus, taking the antilog value, we obtained a value of 388,481.2. Comparing the predicted value of the gradient boosted model

and the original linearly regressed model we see a difference of 89,885. Thus, there is a significant drop in the value of the given house as the accuracy of the model is increased.

Conclusion

Thus, we can see the change of predicted values based on the changes in the accuracy of the model. Reducing the range of the dependent variable by using logarithmic transformation proved significant in improving the accuracy of the model. We saw a difference of 89,885 dollars between our original linearly regressed model and the final gradient boosted model. One of the major parameters which affects the price of the house in King County is the location of the house which was seen from the high influence of the latitude and longitude in the model. Number of bedrooms and bathrooms do not contribute highly to the predictive model as a 1BHK apartment in the downtown area might cost higher than a 3BHK apartment in the suburbs. This model can help future home buyers to get a clear estimate of the value of the property in the area that they are targeting. One of the biggest advantages of utilizing this model is not falling prey to the inflated prices set by the local realtors in the area. Due to the high accuracy of the model, a buyer could challenge the price quoted by the realtor if there exists a significant difference between the quoted and the predicted value. Since, the parameters used for predicting the value of the house were quite general, predicting the house prices for different areas is quite possible provided the data is available for the given area.

Gradient boosting technique is quite useful when the nature of the predicted value is highly sensitive. Parameters like money, health, economy etc. would require a highly accurate model as small changes in these values could prove significant in real-life applications.

Reference

Walia, A. (2018, February 16). Gradient Boosting in R. Retrieved from: https://datascienceplus.com/gradient-boosting-in-r/

Harlfoxem. (2016). House Sales in King County, USA. Retrieved from: https://www.kaggle.com/harlfoxem/housesalesprediction

Ridgeway, G. (2006, April 2016). Generalized Boosted Models: A guide to the gbm package. Retrieved from: http://ftp.auckland.ac.nz/software/CRAN/doc/vignettes/gbm/gbm.pdf

R Bloggers. (2013, February 5). Collinearity and stepwise VIF selection. Retrieved from: https://www.r-bloggers.com/collinearity-and-stepwise-vif-selection/