House Price Prediction Technical report

Stat 2360

Kuanyu Lai

4/13/2023

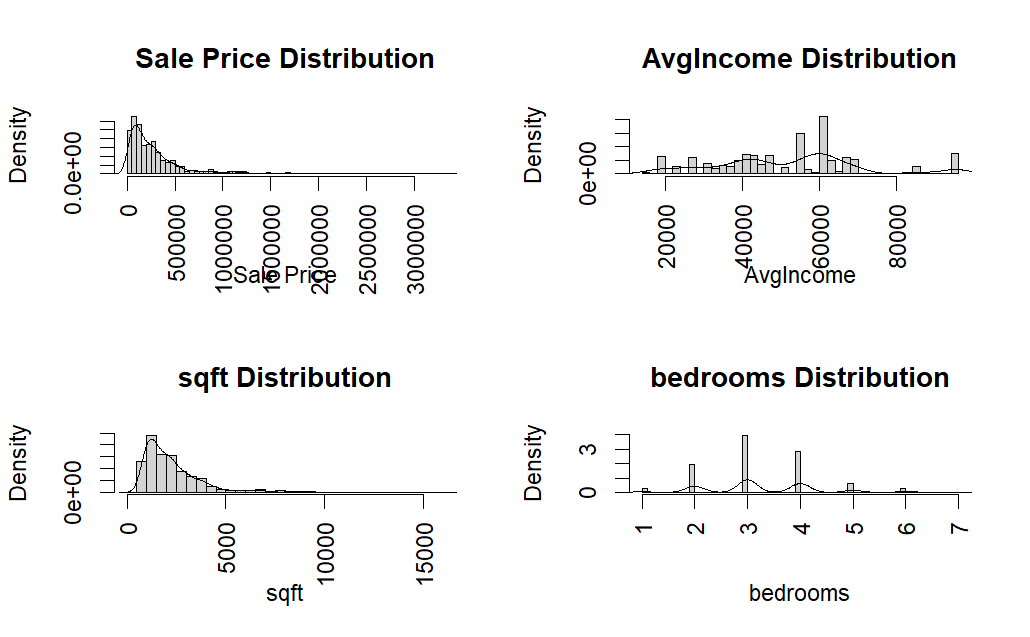
Abstract

For this project, we are trying to help PA Reality breaks into real estate as a data science consultant. Our job is to look into various properties and investigate which properties are most relative to house price so we make a better prediction in the future. In this specific task we are trying to develop a statistical model to predict house price so that in the future realtors can make judgment on whether the house is overprice or underprice by comparing listed price and prediction from the model.

Exploratory data analysis

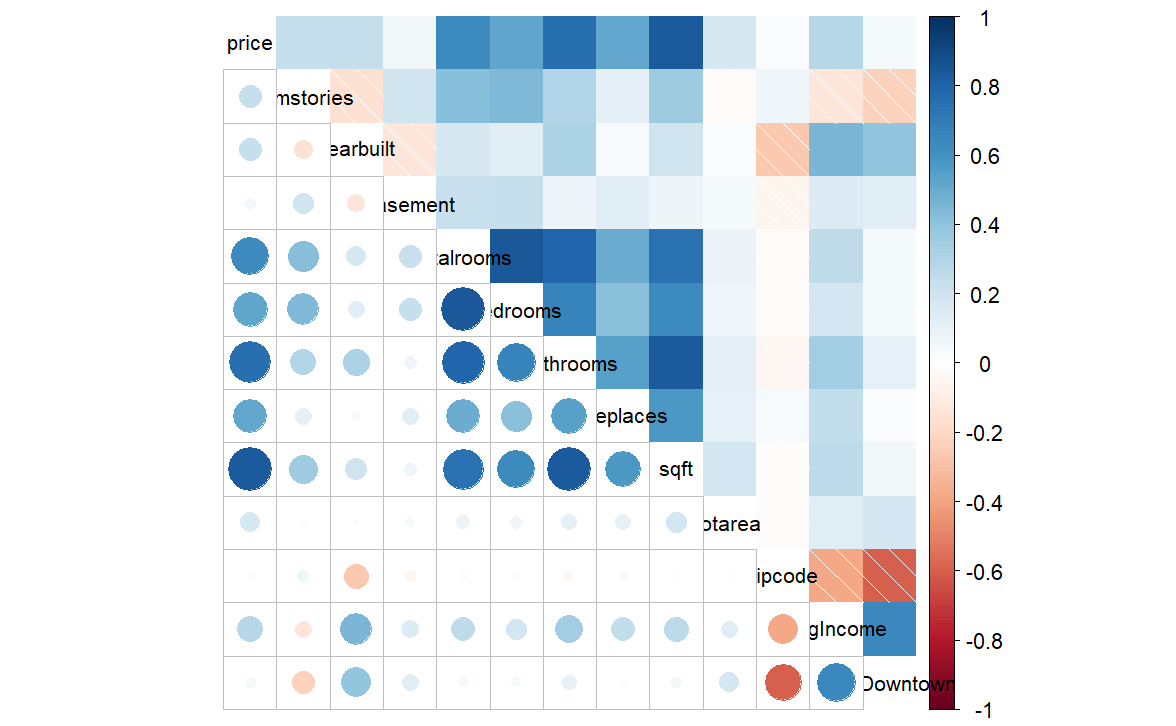
Upon looking at the data, by doing a quick summarize of the train data in r, there are several categorical variables including description of the family, rooftype, location along with more continuous variable like total rooms, avgincome, sqft etc. I noticed that we also have zip code including in the data which is usually very important in predicting house price, however in this case by plotting house price against zip code I found that most zip codes are clustered in two location and there is no clear indication whether it has a linear relationship with house price so in this project I decided to excluded it from making model.

At the beginning I decide to look at some distributions of some continuous variable, the plot is below:



We can see that the distribution of house price and square feet are skewed to the right which indicates that if we try to fit a linear model, we might have to do some variable transformation. As for average income and bedroom number, both seems fine and don’t need further transformation.

Next, we continue to investigate relationships between price and other continuous variable by using correlation plot, below is the correlation graph for each predictor. The darker the blue color indicates stronger the relationship between two variables, if the color is orange or dark red, it indicates that two variables have no or negative relationship.



From this correlation plot we can clearly see with covariates correlations and each predictor’s correlation with house price. We can see the sqft and bathrooms are two variables that are most related house price, we also discover that there are some correlations within predictors such as total rooms, bathrooms and bedrooms which make sense because we can say total rooms is 100 percent correlated so when we try to fit a model, we might want to drop total rooms.

Model Consideration

When considering what statistical model to fit, since we only have training set data and limited amount of data point, I prefer a more conservative method like ridge/lasso. Another method that comes to my mind when choosing model is random forest because it is very effective on data set that contains both continuous variables and categorical variables. RF is also prone to outliers which from the diagnostic plot of linear models there seems to be some. In addition, RF is also very good at choosing important variables by using permutation which shuffle the value of each feature then measure the decrease in accuracy of the prediction, and the feature which has the largest accuracy drop would be consider to be the most important feature.

Model fitting

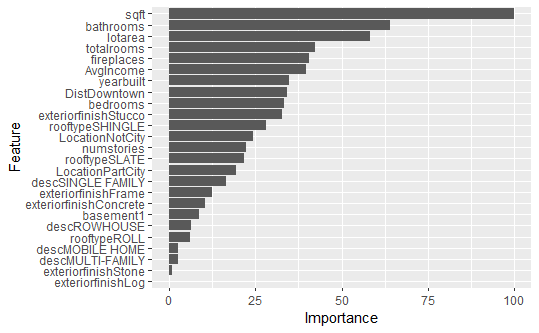
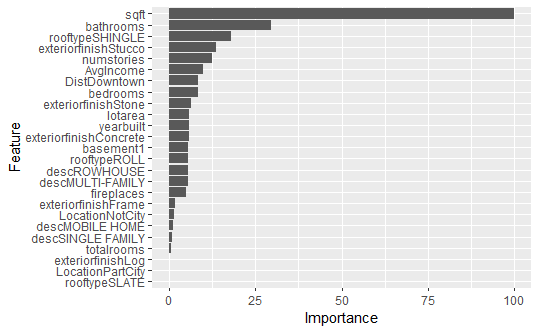
In this project I fitted three kinds of model – Linear, Lasso, and random forest. For linear model I first try out to fit the full model (exclude zip code) which got a adjusted r square of 75.18 percent, after fitting different model by dropping some variables, I got to increase adjusted r square to 76 percent. As for prediction, it has a test correlation of 62 percent and RMSE of 25268. As for lasso regression there is not much difference

|  |  |  |  |
| --- | --- | --- | --- |
|  | lasso | linear | RF |
| R2 | 0.63 | 0.62 | -- |
| RMSE | 251690 | 252679 | 184885 |

There is not much difference between lasso and regression, I think this is because we have 600 observations and 12 variables, in this case n >> p where OLS is more ideal so that might be why they have similar performance. As for random forest, it performs far better than other two.

feature Importance

When building model, we want to decide which predictors are most important in predicting response variables. From graph below we can see feature importance frow lasso and RF.



In both graph sqft is the most important feature followed by number of bathrooms, so base on the feature importance from both models. I would conclude that the result is pretty robust.

Conclusion

Personally, I am fairly confident in my approach and the model I choose. One downside I think my model is that because data is kind of noisy so when generating decision trees in random forest there will be many noisy trees which may affect classification accuracy. The most challenging part of this project is that