**EXECUTIVE SUMMARY**

**Final Project: Predicting House Prices**

Ziwan Peng

## State the problem

I have been hired by the tax authority of the City of Boston to perform a re-assessment of Taxes of residential single-family homes in the greater Boston area. Based on the information in the date set, I developed a linear regression model and a random forest model to evaluate properties in the Boston area.

## KEY FINDINGS

* Instinctively, we would think the age of a house has a strong relationship with the house pricing. But from the density histogram and correlation matrix, we can see that the relationship is weak as a -0.1787 correlation coefficient.
* The average assessed value between residential interior conditions ranges a lot, from a high of $907K in Excellent to $289K in Poor.
* Generally, house pricing would increase as the population increases because there are more demands. However, Dorchester Center, the city with the largest population of 477k has the second-lowest average assessed value of $401k. Jamaica Plain, the city with only 354k population has the highest average assessed value of $670k.

## performance

RSQ is the most common metric to evaluate a model’s performance. It ranges from 0 to 1. The closer to 1 the RSQ is, the more variance explained by the model and the better performance the model has.

In the linear regression model, both test data and train data have a very high RSQ value (the RSQ is 0.8319 for test data and 0.8244 for train data). Over 82% variance of values of properties in the Boston area can be explained by the linear regression model. In addition, only 2 predictors have a P-value above 5% so that most of the variables in the model are statistically significant and useful to explain assessed housing value. We can conclude that this linear regression model performs well to predict the values of properties in the Boston area.

In the random forest model, test data and train data have a very high RSQ value (the RSQ is 0.8119 for test data and 0.8681 for train data). But there’s an overfitting concern about the random forest model that the R-square in test data is obviously higher than that in train data. This model performs much better in train data than test data. Therefore, even though this model can be a useful model to predict the values of properties, we still need to do further optimizations to develop a better one.

## RECOMMANDATIONS

* This linear regression model has a great performance. But we didn’t check the assumptions under this model. One of the main issues would be multicollinearity, which means it’s possible that many variables have strong correlations so that it will affect the model in our case, such as building style with kitchen style and interior condition with overall condition.
* As mentioned before, we should deal with the overfitting issue that happens in random forest prediction.
* In this case, we did variable selections “by hand” but there’s lots of techniques helpful to remove unsignificant variables like stepwise and VIF.
* When dealing with null values, we use mode for categorical variables and mean for numerical variables. But during real feature engineering process, we should be more careful with what value is the most appropriate one to impute.

## KAGGLE

