EAS485/587 Project Phase 2

David Huang, Seungmin Lee

November 2022

For Phase 2, the objectives are applying five different significant and relevant algorithms whether it is ML, MR, or statistical models to the data and create visualizations the models. There should be at least two algorithms that are outside of class. The algorithms that are used are: Word-Count, KNN Clustering, Mean Clustering, and a variety of trees: Decision Tree, Random Forest, and Gradient Boosting. From the previous phase of exploratory data analysis, only five features were selected for the modeling phase: Remodel Repair SQFT, Total New Add SQFT, Total Valuation Remodel, Number Of Floors, and Total Lot SQFT with the target variable as Total Job Valuation.

1 K-Means Clustering

We wish to use k-Means as a way to categorize the groups that have a possible correlation to each other. Using the heat map from Phase 1 as reference, we decided to see the relationship between Total New SQFT and Total Job Valuation. Before starting with the k-Means clustering, we will first use an elbow curve to determine the optimal value for k.

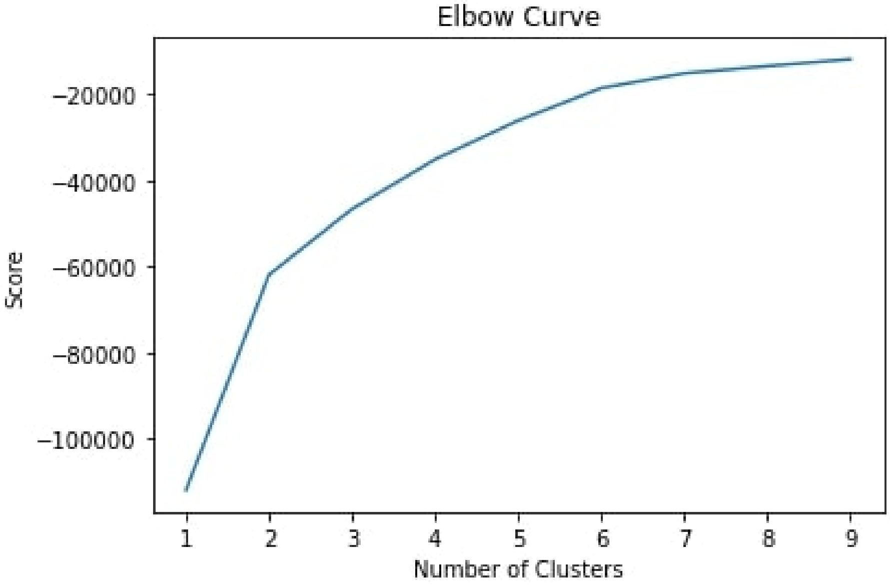


Figure 1: Elbow Curve for K-Means

The slope of the curve is the greatest for the first 2 values of k but significantly reduces as k increases. We will be using $k = 6$ as afterwards there are minimal increases in accuracy.

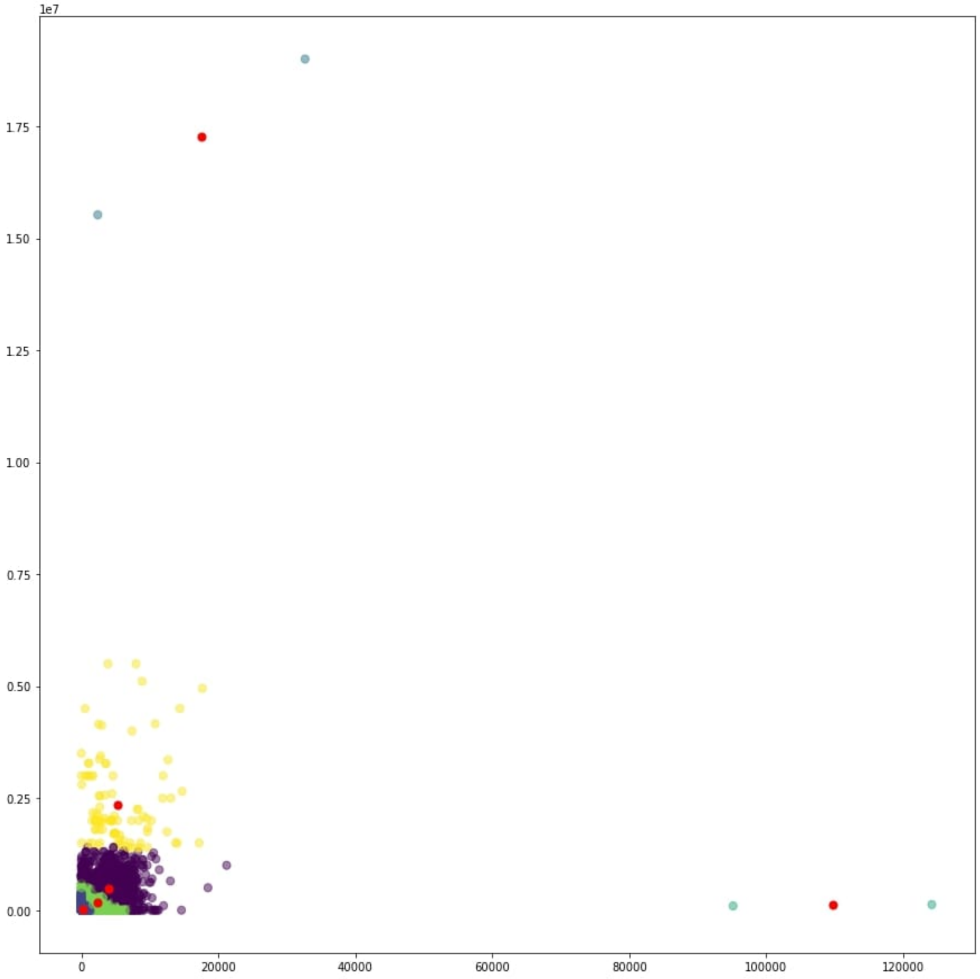


Figure 2: K-Means Clustering

As shown above, there are several outliers leaving inconsistencies. The cluster centroids for the yellow patch, along with the two on the top left and bottom right have too much of a gap between the clusters. To more accurately depict the results of k-Means, we will only work with data within 3 standard deviations.

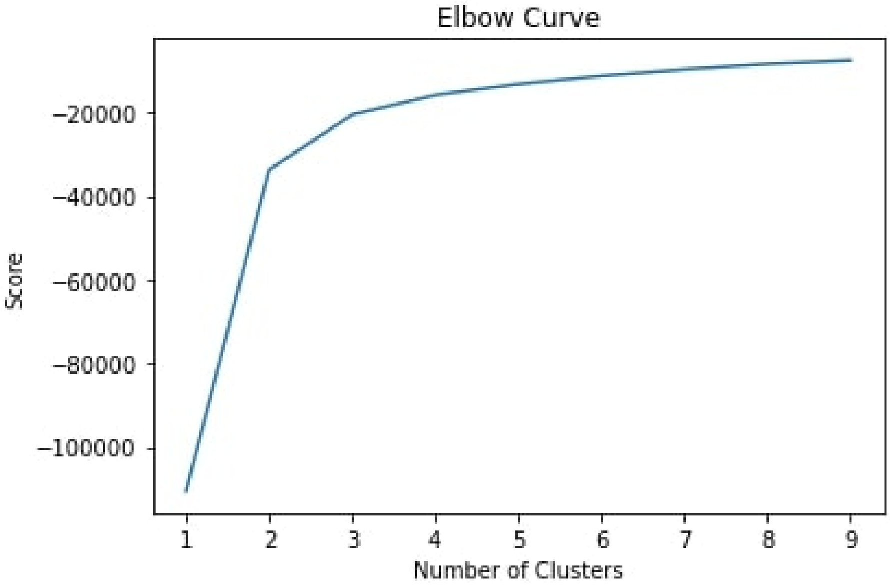


Figure 3: Elbow Curve for K-Means Reduced Features

We will assign out k value to be 4 as this is the last instance of any noticeable growth in the slope.

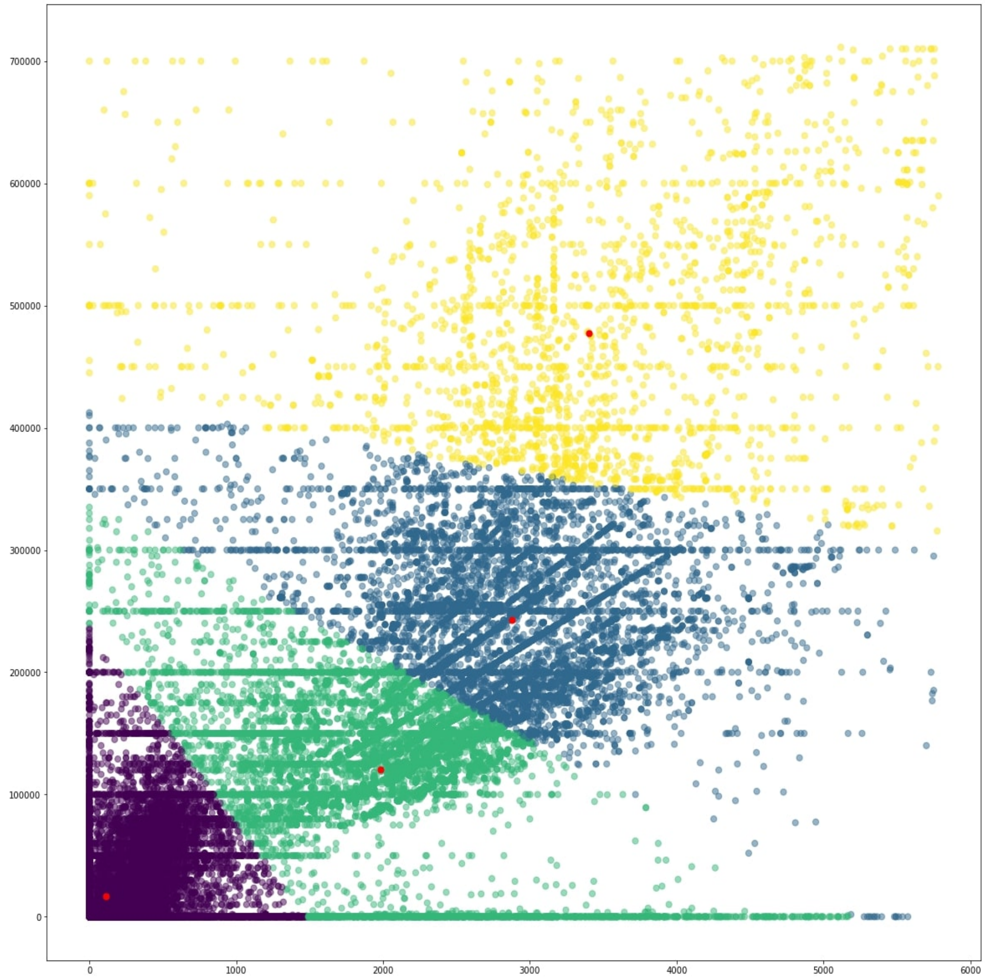


Figure 4: K-Means Clustering with 4 Clusters

Although not perfect, we have separated the data into 4 groups, corresponding to the SQFT and Total Job Valuation. This can be further expanded through the use of other models as we now have boundary conditions shown by the clustering groups. These conditions would represent the separate class used when classifying pay grade groups based on SQFT.

2 KNN Clustering

Similar to k-Means, we wish to see if we can classify the groups through a different approach by using k-NN classification. We will first see the accuracy of the classification model. The models accuracy came out to be 0.1556. The accuracy of this current model is inefficient. However, we wish to find the correlation between SQFT and Total Job Valuation from our finding from the heat map produced from Phase 1. What if we change our k value?

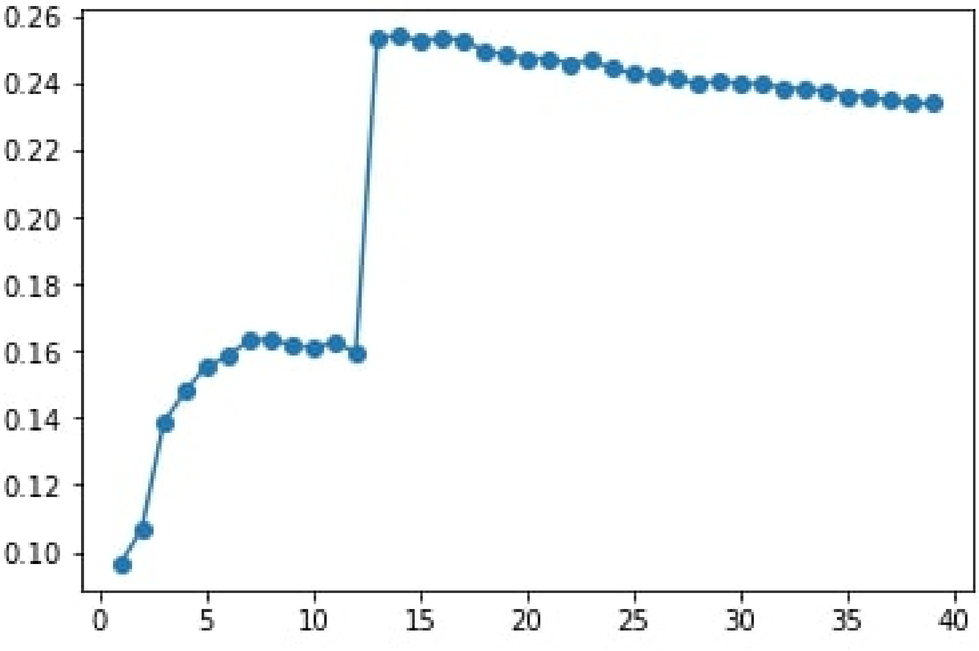


Figure 5: KNN Clustering

There is a substantial increase in accuracy between the k values [2,3] and [12,13]. Regardless, k-NN would not be the best model for our purpose as the highest score achieved is 0.25, which is still not enough.

3 Random Forest, Decision Tree, and Gradient Boosting

The algorithms of Random Forest[3] and Gradient Boosting[1] stem from Decision Trees. The reason why these three were chosen were because decision trees are able to handle regression models with categorical and continuous attributes for the data set. On top of that, the common analysis between the three decision trees algorithms is that random forest normally outperforms decision trees and gradient boosting sometimes improves the mean square error and R^2 values. The reason is because random forest introduces bagging and feature randomness which allows it to better perform compared to a decision tree and reduces over-fitting. Bagging or Bootstrap Aggregating is when multiple different samples of the training data is selected and then using those samples to fit the model. From there, the accuracy is calculated and the average of all the model's accuracy becomes the random forest's accuracy[2]. Feature randomness is when the features combination are selected randomly for each decision tree in a random forest[4].

Using these two methods, it will optimize the random forest to be a better model compared to decision trees. This was shown through using decision tree as a base model comparison with random forest and gradient boosting. For all the models, it was trained using the same set of data and tested using the same set of data.

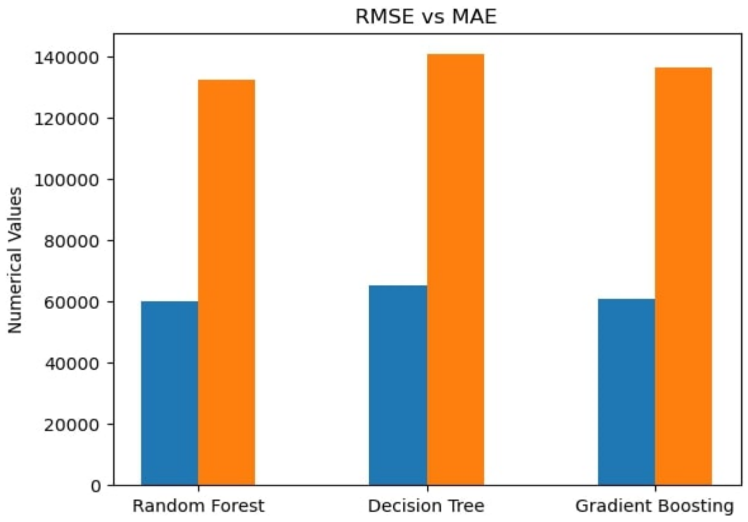


Figure 6: RMSE vs MAE

Figure 6 demonstrates a much higher RMSE value compared with the MAE values for all three models. The ideal is for a lower RMSE value and thus a lower MSE value because RMSE represents the difference between the predicted values and the actual values so the lower the RMSE value, the better the model is at predicting. Currently all the models' RMSE is very high with random forest being the lowest. Besides the RMSE value, the R^2 value is also very important because it represents the amount of variance covered within the model. The R^2 value for random forest is the highest but it is still relatively low which means all of these models are not good predictors of the total job valuation.

The reason why these models are not good predictors would be the large variance and the very large range of data points. There should be some removal of outliers as well as double checking the decision trees are not being overfitted. This can be done through more pruning and through being more specific with the parameters of the random forest and gradient boosting algorithms.

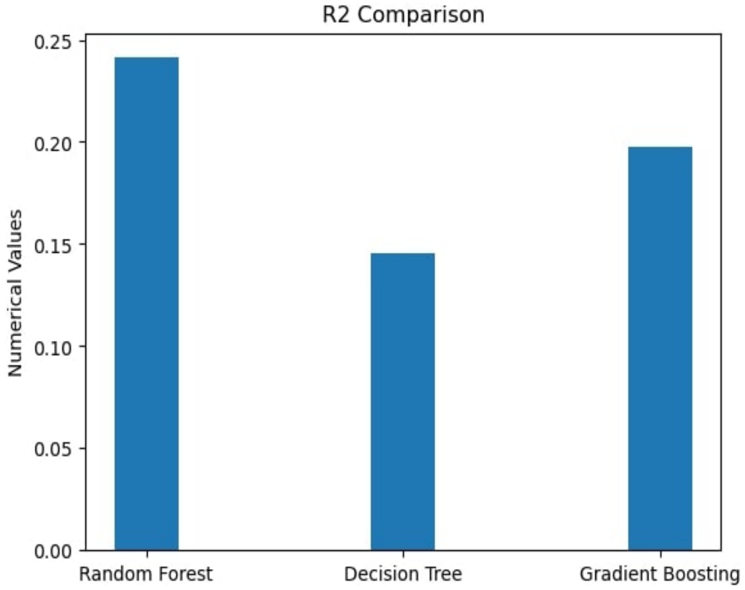
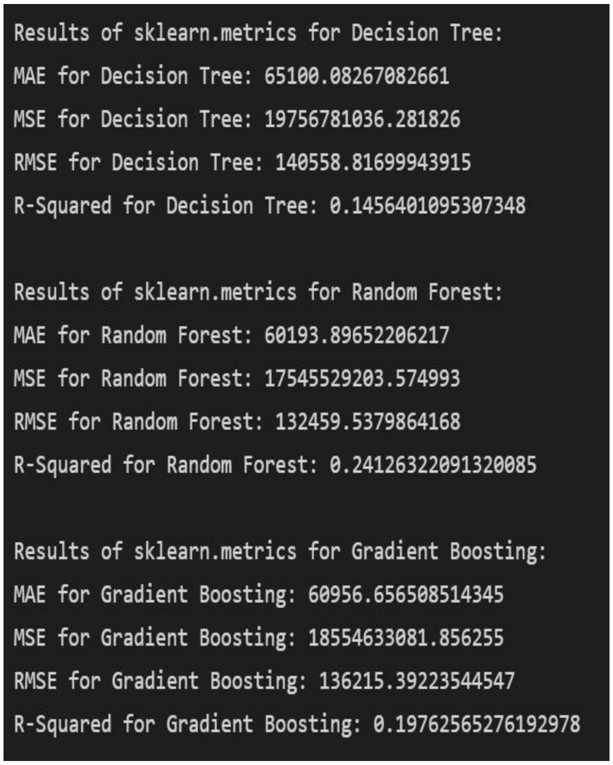


Figure 7: R^2 Comparison

RMSE and MAE both represents error; however, MAE measures the average magnitude of the errors in a set of predictions, without considering their direction of positive or negative. Both of these variables should be as low as they can. One main difference is that for RMSE, it gives a heavier weight for larger errors whereas, MAE is normally steady. Since for this project, there is a need to penalize the larger errors, the RMSE value is a great way to describe the models.



```
Results of sklearn.metrics for Decision Tree:
MAE for Decision Tree: 65100.08267082661
MSE for Decision Tree: 19756781036.281826
RMSE for Decision Tree: 140558.81699943915
R-Squared for Decision Tree: 0.1456401095307348

Results of sklearn.metrics for Random Forest:
MAE for Random Forest: 60193.89652206217
MSE for Random Forest: 17545529203.574993
RMSE for Random Forest: 132459.5379864168
R-Squared for Random Forest: 0.24126322091320085

Results of sklearn.metrics for Gradient Boosting:
MAE for Gradient Boosting: 60956.656508514345
MSE for Gradient Boosting: 18554633081.856255
RMSE for Gradient Boosting: 136215.39223544547
R-Squared for Gradient Boosting: 0.19762565276192978
```

Figure 8: RMSE, MSE, and R^2 Values Comparison

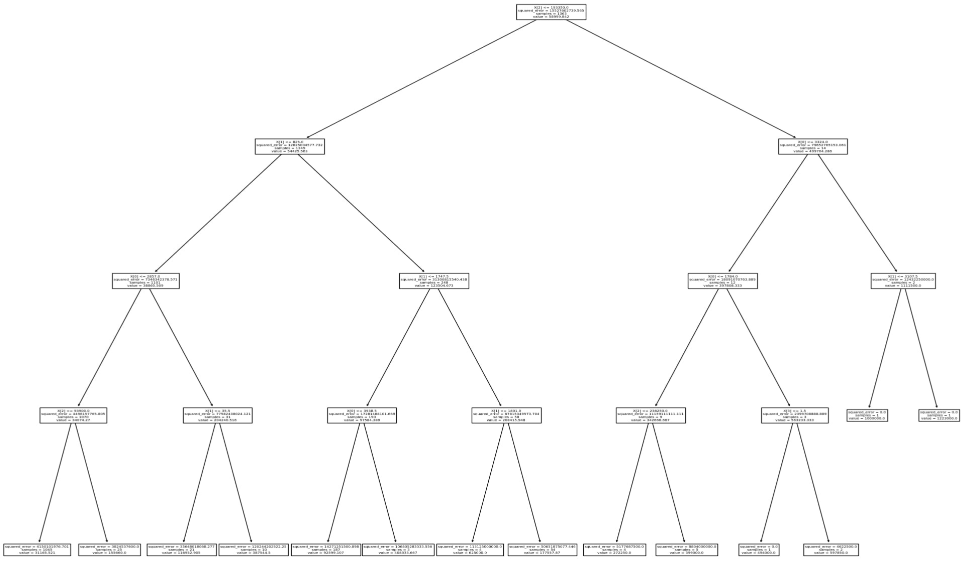


Figure 9: Decision Tree

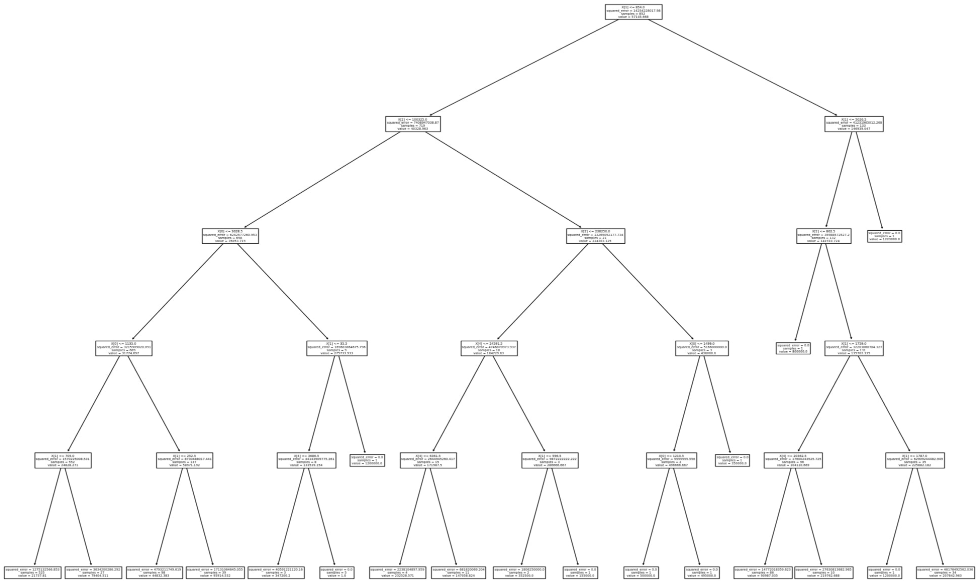

X[1] <= 854.0 squared_error = 14254228017.98 samples = 852 value = 57145.668
X[2] <= 100325.0 squared_error = 7408947038.87 samples = 719 value = 40328.963
X[1] <= 5026.5 squared_error = 41231985012.268 samples = 133 value = 146939.047
X[0] <= 3628.5 squared_error = 6242577260.953 samples = 698 value = 35053.719
X[2] <= 238250.0 squared_error = 13269092177.734 samples = 21 value = 224303.125
X[1] <= 862.5 squared_error = 35988572527.2 samples = 132 value = 141910.724
squared_error = 0.0 samples = 1 value = 1223000.0
X[0] <= 1135.0 squared_error = 3215909020.091 samples = 689 value = 31774.697
X[1] <= 35.5 squared_error = 169683864675.796 samples = 9 value = 275733.933
X[4] <= 24591.5 squared_error = 4746870973.937 samples = 18 value = 184729.63
X[0] <= 1499.0 squared_error = 5166000000.0 samples = 3 value = 438000.0
squared_error = 0.0 samples = 1 value = 800000.0
X[1] <= 1759.0 squared_error = 32203868784.327 samples = 131 value = 135702.335
X[1] <= 705.0 squared_error = 1570225008.531 samples = 552 value = 24828.271
X[1] <= 252.5 squared_error = 8730488017.441 samples = 137 value = 58971.192
X[4] <= 3686.5 squared_error = 44143909775.361 samples = 8 value = 133539.154
squared_error = 0.0 samples = 1 value = 1200000.0
X[4] <= 6361.5 squared_error = 2644945260.417 samples = 15 value = 171987.5
X[1] <= 596.5 squared_error = 9872222222.222 samples = 3 value = 286666.667
X[0] <= 1210.5 squared_error = 5555555.556 samples = 2 value = 496666.667
squared_error = 0.0 samples = 1 value = 350000.0
X[4] <= 20362.5 squared_error = 17600243525.725 samples = 96 value = 104110.669
X[1] <= 1787.0 squared_error = 62909244482.949 samples = 35 value = 225882.182
squared_error = 1275132566.853 samples = 525 value = 21737.81
squared_error = 3634200286.292 samples = 27 value = 79404.511
squared_error = 4793211749.619 samples = 98 value = 44832.383
squared_error = 17131064645.055 samples = 39 value = 95914.532
squared_error = 40591221120.16 samples = 3 value = 347200.2
squared_error = 0.0 samples = 5 value = 1.0
squared_error = 2238104897.959 samples = 4 value = 232528.571
squared_error = 681820069.204 samples = 11 value = 147058.824
squared_error = 1806250000.0 samples = 2 value = 352500.0
squared_error = 0.0 samples = 1 value = 155000.0
squared_error = 0.0 samples = 1 value = 500000.0
squared_error = 0.0 samples = 1 value = 495000.0
squared_error = 14772018359.623 samples = 86 value = 90987.035
squared_error = 27630813682.965 samples = 10 value = 219762.688
squared_error = 0.0 samples = 1 value = 1200000.0
squared_error = 46176492562.036 samples = 34 value = 207842.963

Figure 10: Random Forest Tree

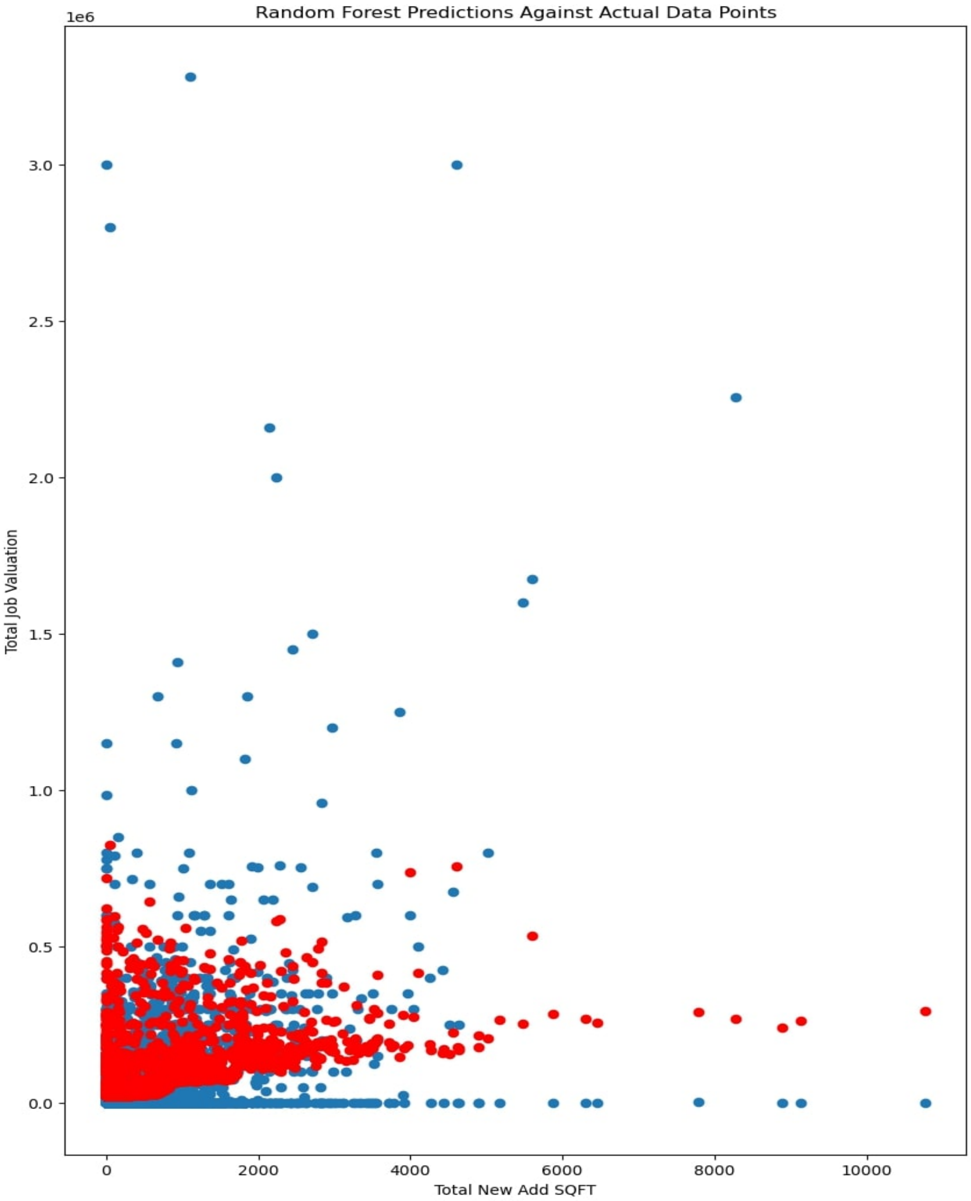


Figure 11: Random Forest Data Points with Actual Points

There are some interesting portions in regards to the scatter plot because the total job valuation for the total new added SQFT seems to be increasing at a faster rate as the total new added SQFT increases. However, there isn't a huge range of the total job valuation compared to the actual data points. There are almost no predicted values above $1.0 * 10^6$ portion but there are actual data points.

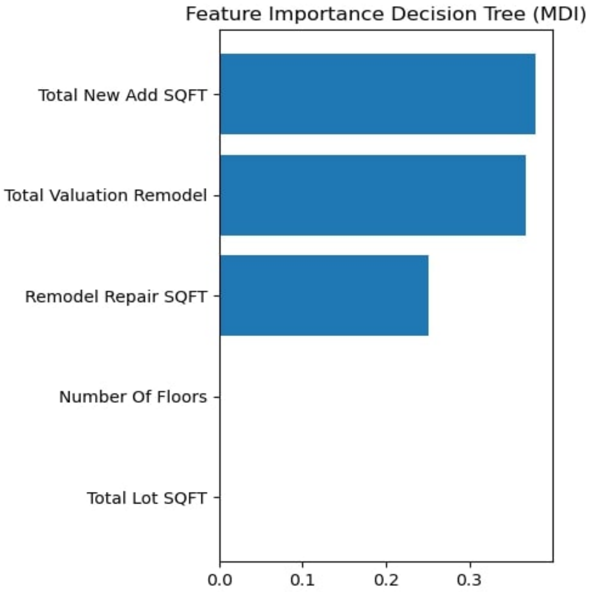


Figure 12: Feature Importance for Decision Tree

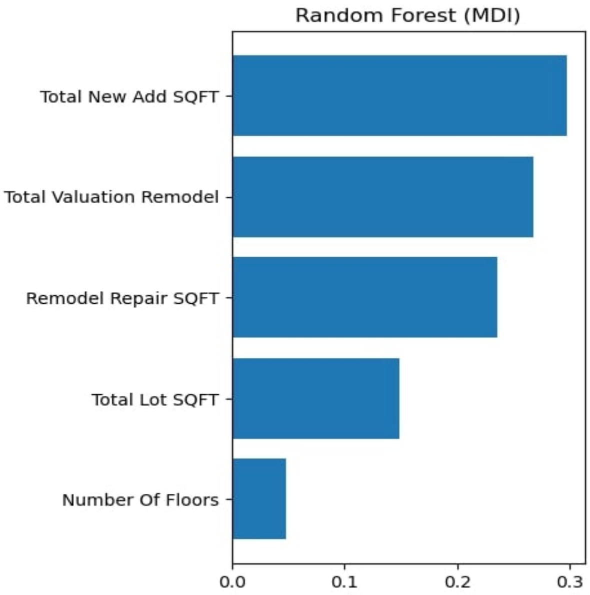


Figure 13: Feature Importance for Random Forest

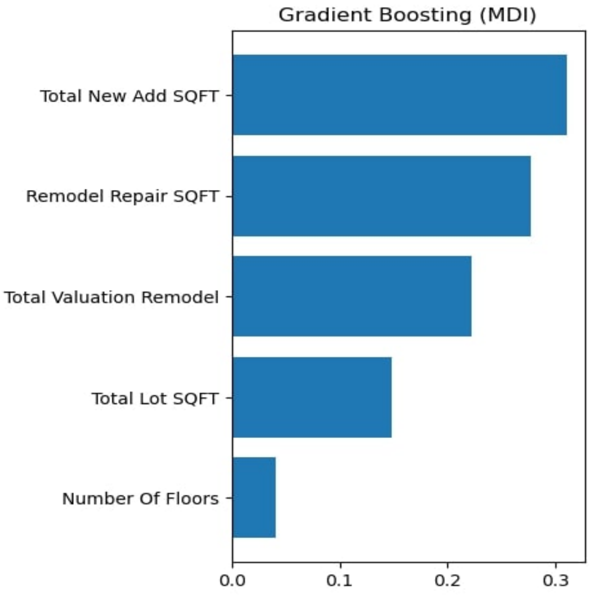


Figure 14: Feature Importance for Gradient Boosting

An interesting finding for the feature importance: figure 10, 12, and 14, between the three models is that, only 5 of the initial columns were significant which agrees with the EDA that was conducted using PCA on the columns. A common trend that is found throughout the all the feature importance is that the "Total New Add SQFT", "Remodel Repair SQFT", and "Total Valuation Remodel" all have the highest influence on the total job valuation. These are all numerical values. A future model that could be used to see if there is a linear relationship between the 3 features and the target variable would be multi-linear regression. The reason why currently Decision Trees related algorithms were used were because there were initially categorical features that were later removed due to their lack of influence. Decision Trees related algorithms also focuses on non-linear relationships which could be removed because currently there is no analysis on assuming that there is a linear relationship.

In regards to the question for the problem statement, there is no relationship between the amount of job valuation to the size of the lot; however, there is a relationship between the job valuation and the total new added SQFT and the remodel repair SQFT which would be indirectly related to the total lot SQFT because with a larger total lot SQFT, there would be a larger SQFT to be added and a larger remodel repair SQFT. There would be some linearity between the 3 SQFT features through that indirect relationship.

References

[1] https://blog.paperspace.com/implementing-gradient-boosting-regression-python/.

[2] https://medium.com/nerd-for-tech/bootstrap-aggregating-and-random-forest-model-9460e235537.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.randomforestregressor.html.

[4] https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd.