Dear CEO and Realtors of PA Reality,

This report focused on predicting housing prices for homes in the Pittsburgh area using various predictive models. I needed to start by analyzing the dataset, which included exploring the data and checking for missing values, outliers, and inconsistencies in categorical variables, where the data had variables such as price, square foot, location, and more. I used different strategies such as imputation, outlier detection, and data cleaning techniques to address these issues.

Next, I considered several predictive models such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Ridge Regression. I evaluated the performance in these models using the Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error through 5-cross-validation. In these results, I identified that the Random Forest model was the best-performing model with the lowest MSE, RMSE, and MAE compared to the other models.

During the identification process, we saw several data oddities in the dataset, including missing values in some variables, outliers in certain numeric variables, and inconsistencies in categorical variables. For example, I have to use log transformations to combat skewness in variables like prices and fireplaces. Afterward, I utilized strategies like imputation for missing values, outlier detection to handle outliers, log transformation techniques for extreme values, and data techniques for inconsistencies in categorical variables.

With the dataset, I considered several potential predictive models, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Ridge Regression. These models were based on their suitability for the dataset and provided accurate predictions for the target variable

price. By measuring the predictive accuracy of each model, I needed to perform various concepts such as Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error. With this, I did cross-validation of a 5-fold to be sure I could obtain strong results. With these results, I identified the model of Random Forest with the lowest MSE, RMSE, and MAE as the best predictive accuracy-performing model.

To determine the important variables for the predictive models, we used measures such as a Decision Tree, Random Forest, and Gradient Boosting. Taking advantage of these variables is found by displaying the most important plots that identify influential features in the prediction. However, for Ridge Regression, we use the coefficients of the variables to assess their importance.

The most challenging aspects of the dataset were the missing data and outliers I needed to identify. I had to use imputation techniques to handle missing values and outlier detection and treatment to handle outliers. But there were limitations in the effectiveness of the techniques at hand, where missing values got imputed inaccurately, and some outliers were difficult to identify and treat. Moreover, the categorical variables and inconsistencies required data-cleaning efforts to allow consistency and accuracy in the model predictions. So, I tried to mitigate these issues by using techniques of treatment like log transformations for skewness. The log transformations gave us a performance better than it was, but there is always room for more validation in the data.

Based on the results, the best-performing model for predictive accuracy was the Random Forest since it had the lowest MSE, RMSE, and MAE compared to the other models. The performance

of this model may have been affected by factors like data quality, model assumptions, and changes in the data distribution. Thus, while the Random Forest model performs much better, I can acknowledge that there may be room for improvement and validation before fully trusting its predictions. I would recommend more improvements in sections of data cleaning and data inconsistencies in preprocessing to gain reliability and accuracy in the model.

In conclusion, we have the process of developing a predictive model with the Pittsburgh dataset while performing EDA, data oddities, and evaluating multiple models for predictive accuracy. With this, the results showed that the Random Forest model was the best-performing model.

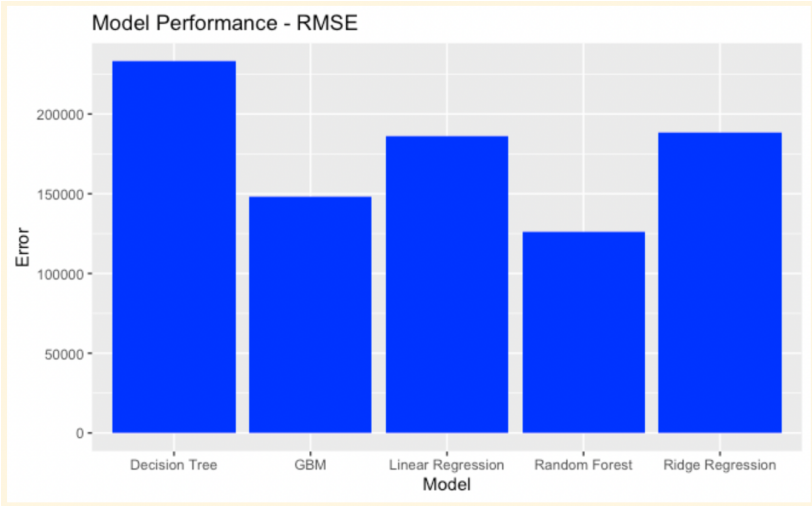
Sincerely,

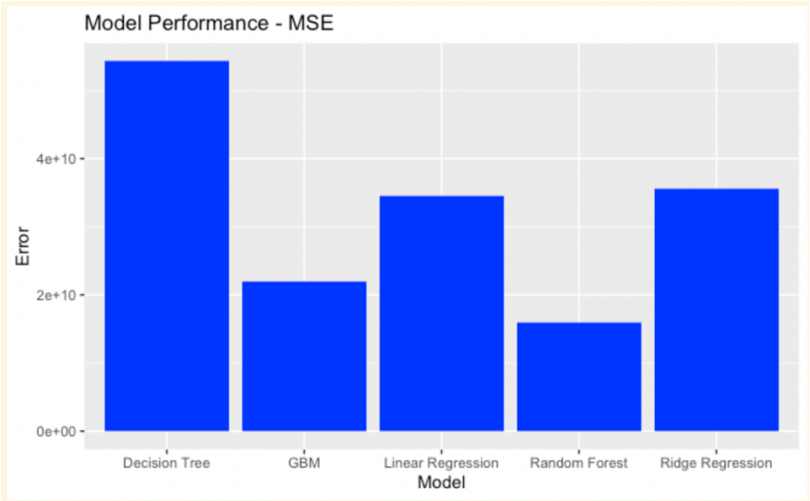
Stephen Skasko

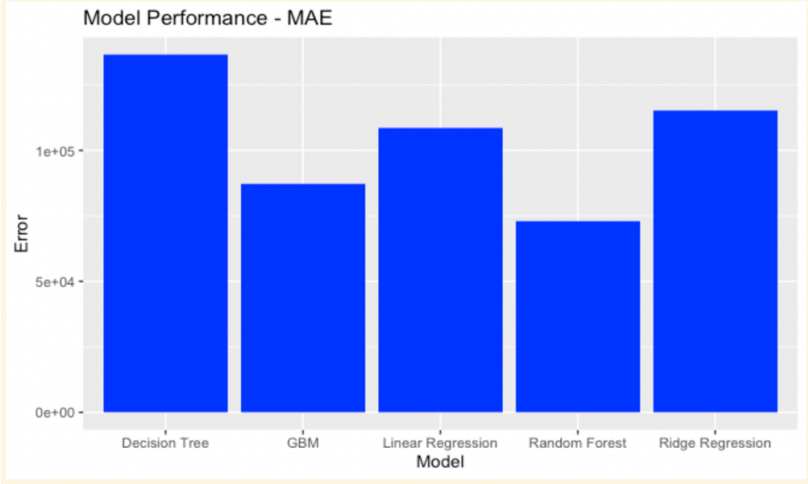
Data Science Consultant

Predictive Housing Prices

Model Type Best (1) - Worst (5)	MSE	RMSE	MAE
3. Linear Regression	34531538892	185826.6	108528.2
5. Decision Tree	54289986539	233002.1	136477
1. Random Forest	15888289863	126048.8	72887.65
2. Gradient Boosting Machine	22000077297	148324.2	87432.39
4. Ridge Regression	35511984281	188446.2	115216.5






Model Performance - MAE
Error
1e+05
5e+04
0e+00
Decision Tree
GBM
Linear Regression
Random Forest
Ridge Regression
Model