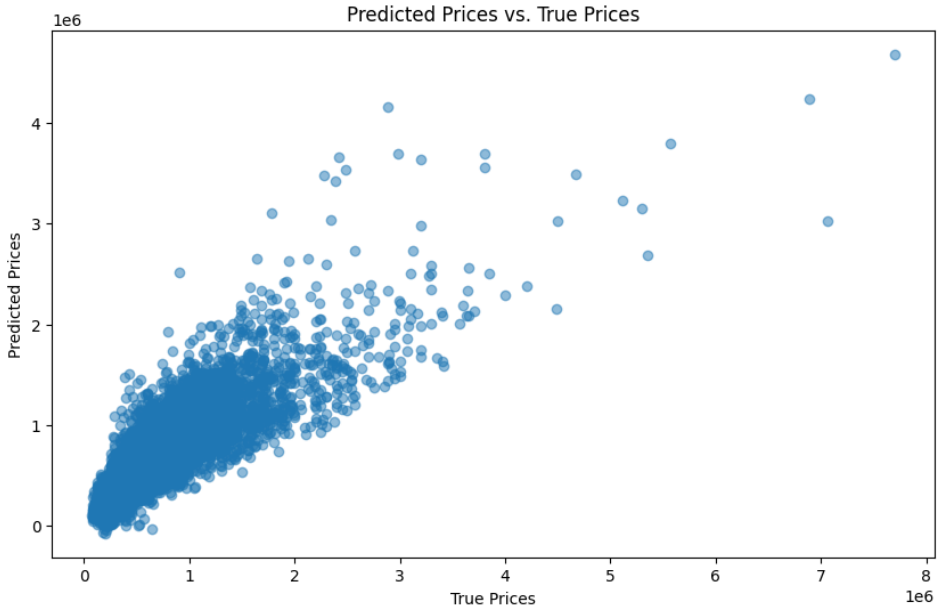
Tyler Graham
CS5610
November 27, 2023

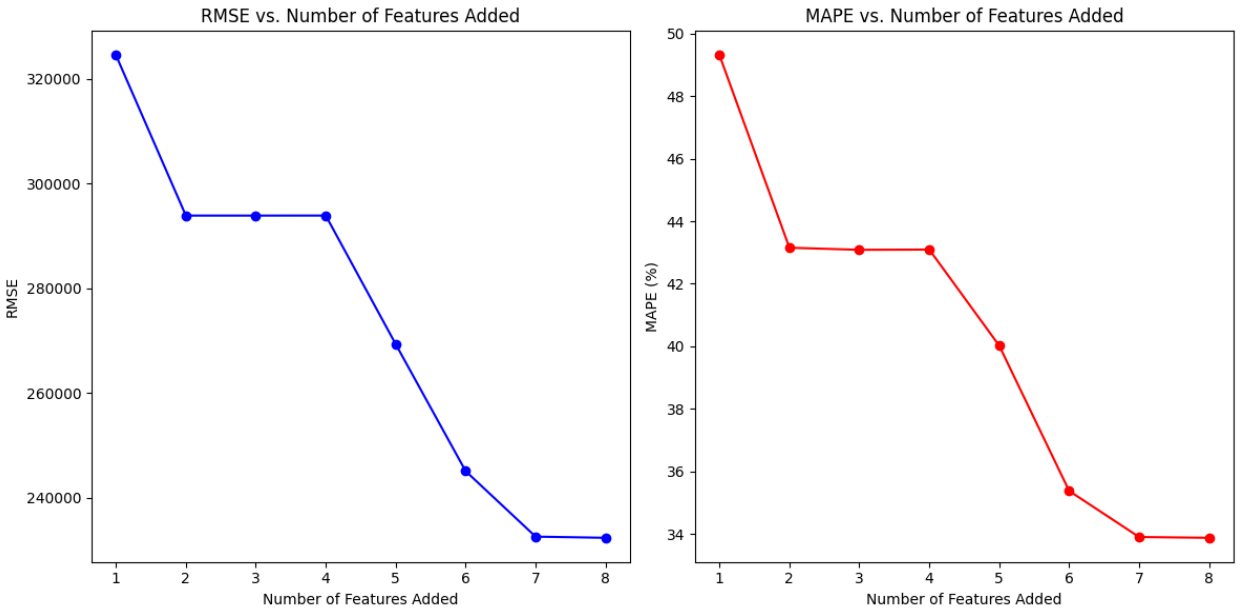
Project 4

A home is often an individual's most expensive purchase in their lifetimes. The ability to predict the value of a home is an invaluable tool both to investors and home owners as they purchase or sell their homes. This analysis aims to evaluate different statistical models to predict house prices. The initial data set contains features related to each home sold's location, their physical characteristics (bedrooms and bathrooms, etc) and the condition of the home (conditions, year renovated, etc).

The first model implemented was a baseline linear regression, that aimed to predict the house priced based on the combined scaled input of the numerical variables and dummy variables created from the categorical variables. The performance of this model was initially evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), yielding scores of 190322.68 and 23.32%, respectively. The dataset is continued to be refined with different models and different feature engineering to continue to drive the RMSE and MAPE down.



Following the completion of the linear regression model, a separate tanganiel analysis was ran by adding variables one at a time into the model to attempt to find the any outliers to eliminate data or create new features bases on the variables. The below chart shows both the RMSE and MAPE as each variable is added into the model. The biggest jumps were adding in bedrooms (0), bathrooms (1), square feet(2), year_built (6). If computing resources were an issue, these five could be completed to run a model with an RMSE of 341412.262 and an MAPE of 53.61 which was approx. half of the previous model.



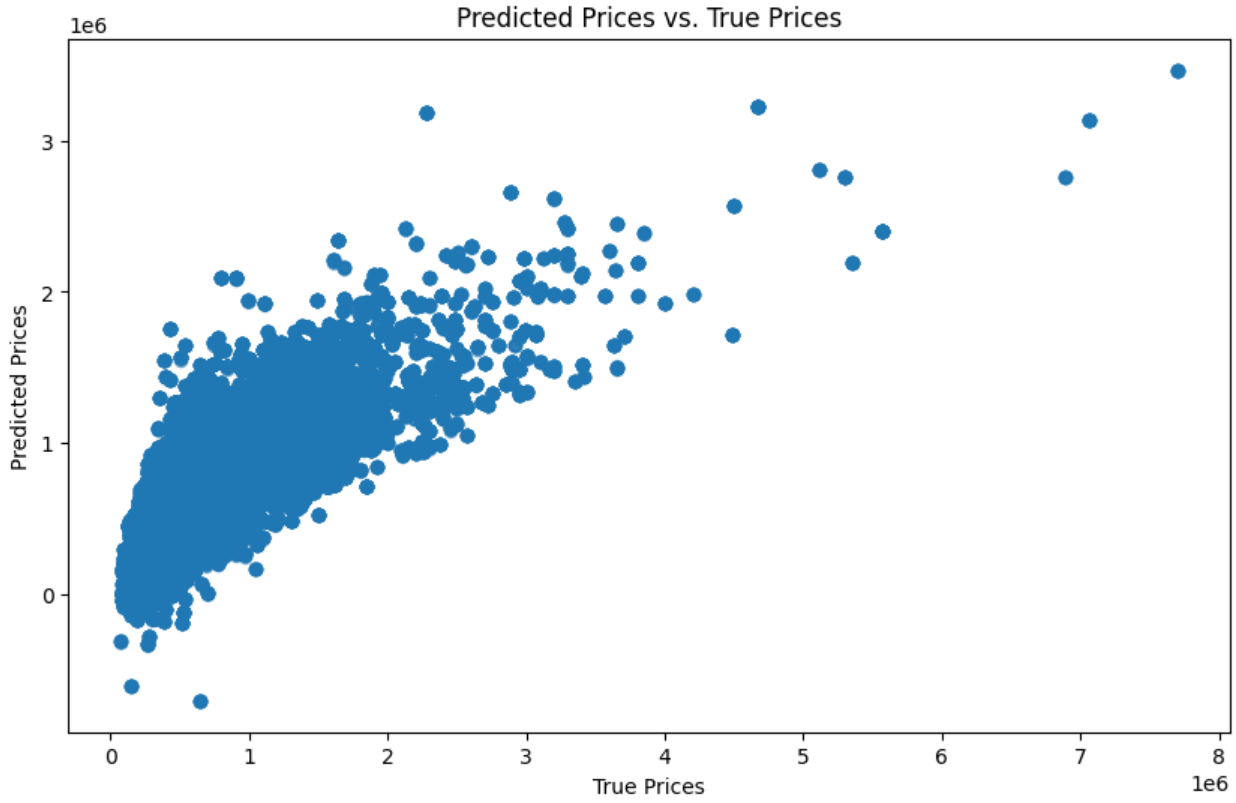
The second approach employed was the Random Forest Regressor, a more sophisticated machine learning technique known for its robustness and ability to handle non-linear relationships. The model was initially assessed with standard parameters, achieving a significantly lower RMSE of 49464.675. To further enhance its performance, a hybrid approach using cross-validation was employed to tune the below hyperparameters. This resulted in a slightly higher RMSE of 50951.3911 but an improved MAPE of 5.596%.

Hyper Parameters Variables for Random Forest Regressor

Hyperparameter	Explanation	Search Space
n_estimators	Number of decision trees in the random forest ensemble	Random integer between 100 and 500
max_depth	Maximum depth of each decision tree in the random forest	[None, 10, 20, 30, 40, 50]
min_samples_split	Minimum number of samples required to split an internal node in a decision tree	Random integer between 2 and 10
min_samples_leaf	Minimum number of samples required to be in a leaf node of a decision tree	Random integer between 1 and 4

Finally, the initial linear regression model was revisited, this time incorporating KBinDiscretizer for specific features such as square footage longitude and latitude. This method aimed to discretize continuous variables into bins, potentially capturing non-linear relationships more effectively. However, this approach did not significantly alter the initial linear model's performance, with RMSE and MAPE remaining at 190322.68 and 23.32%, respectively.

Next, one of the best ways to add additional gains to the accuracy of your model is to add datasets that contain additional variables. There was an included dataset containing a subset of population by zip code in the year 2010. This include the population, the minimum age, the maximum age, the gender and the GEO ID. Since, we have zip code in our original dataset this was used as the variable to merge the population data on to see if running a linear regression of having the initial dataset would increase the accuracy of the model. However, this approach did not improve significantly alter the initial linear model's performance, with RMSE and MAPE remaining at 214761.93 and 29.17%, respectively with the following graph of the predicted prices versus the true prices.



Comparing these approaches reveals that the Random Forest Regressor emerged as the most accurate model, significantly lowering the RMSE and improving the MAPE. Its ability to capture complex patterns in the data made it superior to the linear models. The simple linear regression, was less effective due to its inherent limitations in handling non-linear relationships and interactions between variables. The Linear Regression with KBinDiscretizer did not offer any significant advantages over the simple linear regression, indicating that merely discretizing certain features is not enough to capture the complexities of house pricing and attempting to reduce the initial dataset to eliminate additional outliers did not provide much decrease of computa-

tional resources at almost half the accuracy of the baseline model. The following table contains the final RMSE and MAPE values from this model.

RMSE & MAPE for all 5 Models

	RMSE	MAPE
Baseline Linear Regression	190,322.68	23.32%
Reducing Features to 5 Most Impactful	341,412.26	53.62%
Random Tree Regression	50,591.40	5.59%
Create categorical variables from continuous variables	190,322.68	23.32%
Additional Dataset	214,761.93	29.17%

One of the primary concerns in this dataset is the potential presence of missing data such as incomplete records or data entry errors. After more careful inspect of the additional zip code dataset it was found to include large amounts of incomplete records. It can significantly skew the results of a model, leading to biased or inaccurate predictions.

Additionally, The dataset's usage of historical data from 2015 and with 'yr_built' and 'yr_renovated' does not reflect the current real estate environment. Trends in housing change rapidly in price and desirable feature set, since the original data is from 2015 it may not capture recent developments or changes in neighborhood desirability.

Finally, the dataset is from a single geographic location of a suburb of Seattle, Washington. There are large amounts of homes in other states in the US and even more outside of the US. This dataset will not be accurate for other countries as their pricing, interest rates, cultural expectations of a home differ drastically as you venture outside of the US.

While the Random Forest Regressor stands out for the lowest error capabilities, to enhance the model's utility, future iterations should integrate more up to data, cleaner population and socioeconomic data from external sources. This approach will lead to more accurate, reliable, and contextually relevant predictions, better aligning the model to real estate valuation and decision-making processes.

It would be reasonable for the user to trust the ~5% error from the Random Forest Regressor if the data used was relatively (with the last 6 months) of data and trained on nearby geographic location. I would not trust the model the further the training and testing data is from the present and the further away the user is from the initial location.

In conclusion, Project 4 did find a suitable model to get the error within ~5%, however, there is some key issues with the data to be used in a modern predictor of home prices for everyone. The model is a great first step and can continue to be modified, cleaned and tuned to become accurate for a much larger subset of the population to predict home prices.