Al Alipour
August 23, 2021

Predicting Housing Prices in Nashville

Problems Statement

There are various factors impacting housing prices such as housing type, living area, acreage, number of bedrooms, etc. Moreover, in a free market, prices can vary significantly depending on the buyers and sellers. It can be quite confusing for buyers to find the right price of a property and for sellers to price their property reasonably. The current project aimed at developing a model for predicting housing prices in Nashville metropolitan area based on different features of a property. The tool developed provides a means of acquiring a reasonable initial estimate of the price of a property for buyers and sellers in Nashville metropolitan area.

Data and Preparation

Main data for the project were acquired from Kaggle website (https://www.kaggle.com/tmthyjames/nashville-housing-data). The dataset included more than 56,000 records of property sales in Nashville metropolitan area including features such as property address, land use, sale date, etc. The sale records were all for the period between 2013-2016. About half the records, however, missed information on a large number of features. Besides sale records, data were acquired on US mortgage rate (data from Freddie Mac available on Fred Economic Data website, 2021), US average hourly wage (US Bureau of Labor Statistics website, 2021), US unemployment rate (US Bureau of Statistics website, 2021), and Nashville unemployment rate (US Bureau of Statistics website, 2021) for the period of analysis. However, data on US unemployment rate were not used given its high correlation with Nashville unemployment rate.

Data were explored, organized and cleaned through the data wrangling phase of the project. Some of the records (rows) and features (columns) were combined, dropped, and/or changed to prepare the data for further analysis. Missing data were explored to find out whether the data were missing randomly or systematically (Fig. 1), and whether missing values could be

imputed. While some of the data were imputed, most of the missing data were dropped during exploratory data analysis because there was no way to reliably impute them. Finally, data on mortgage rates, hourly wages, and unemployment rates were joined with the rest of the data.

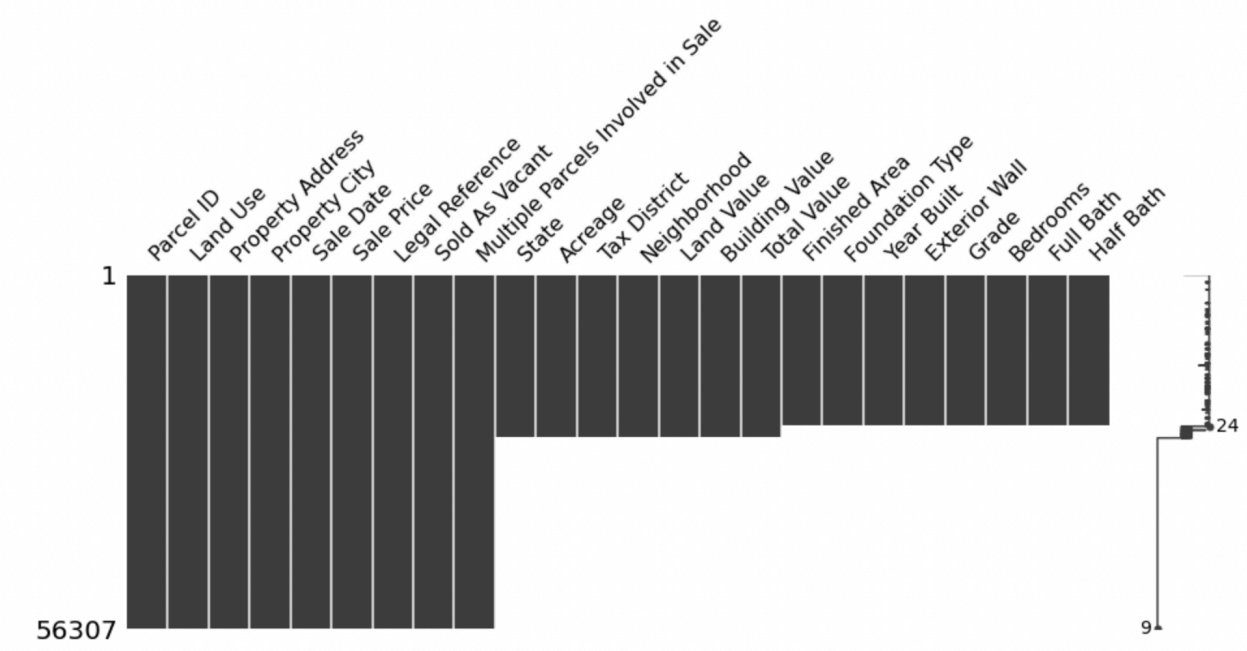

Figure 1. Distribution of missing values

Exploratory Data Analysis

Data were further explored using different visualizations to find potential patterns, identify required corrections, and discover associations. It was clarified in this step that sale price for a property was highly correlated with three other pricing features including land value, building value and total value. As the source of these values must have been from appraisals of the properties, it was argued that the goal of the analysis was to predict a home's sale price in Nashville based on a free market and with no need for an appraisal of each individual home. Thus, it would be possible to predict a property sale price not included in the dataset based on the model developed. As a result, while these values could have helped increase the accuracy of the model, it was decided to drop all these three features to expand the applicability of the final model to properties with no appraisal of their land, building, or total value.

Other issues such as presence of data on vacant land rather than actual homes and duplicate land use types were identified in this step as well and accordingly dealt with. Another problem identified in this step was the presence of outliers. For example, some sale prices were incredibly low, or the number of bedrooms and bathrooms were zero for some properties with large living areas. The outliers were identified in this step and appropriately dealt with during preprocessing. Finally, after dealing with the problems identified, correlations between some of the numerical features and sale price were further investigated (Fig. 2). It was made clear that sale price was highly correlated with factors such as living area, acreage, bedrooms, and bathrooms. It was also made clear that factors such as unemployment rate and mortgage rate were negatively correlated with sale price.

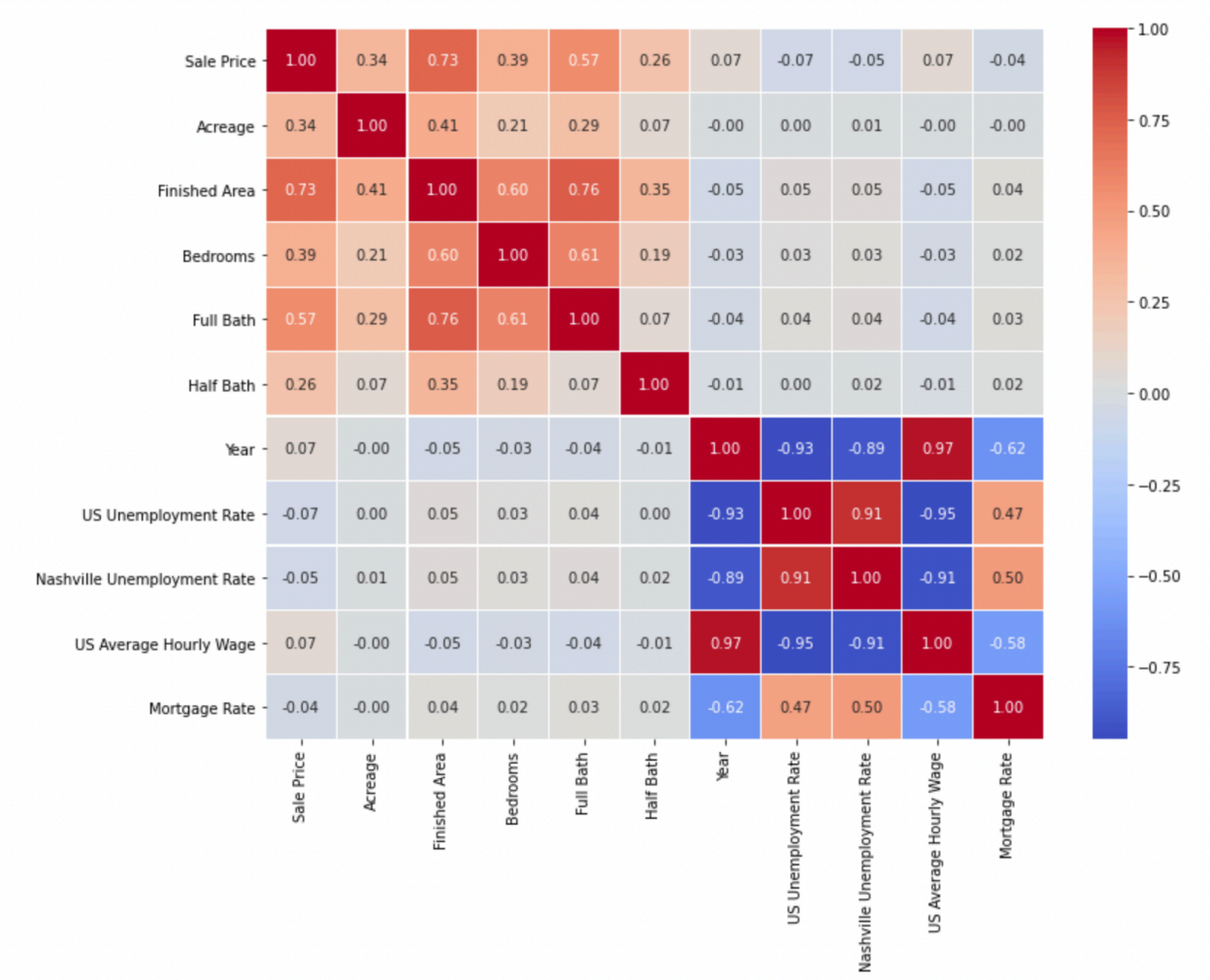

Figure 2. Correlations between sale price and numerical features

Preprocessing

Data were standardized and split into training (75%) and test (25%) sets in this step. However, besides these standard processes, some other actions were taken in this step as well. These included imputing some of the missing data, feature engineering, and dealing with outliers. Feature engineering composed of combining some of the categorical features (e.g., neighbourhood and city) and performing one-hot encoding as well as creating a new feature (age when sold) from year built of a property and its sale year. Dealing with outliers, in this step, was associated with properties with very low and very high sale prices per square foot of living area. It was observed that there were properties with a sale price of less than $50 per square foot of living area. It was argued that this value was very low and probably associated with errors. A comprehensive search on the internet showed that houses in Nashville metropolitan area currently sell for an average price of $165-$303 per square foot. Since the prices from the period of analysis (2013-2016) to present have almost doubled, it was plausible to assume that the median price per square foot for the period of analysis should have been around $100-$150. Thus, a conservative approach was taken and it was assumed that the prices outside of the range of $50-$350 per square foot of living area were associated with errors, and the related rows were dropped. Please note that this approach is particularly plausible for this project because the project goal is to estimate the price of ordinary properties in Nashville and there is less focus on extreme outliers.

Modelling

Modelling initially consisted of trying out seven different machine learning algorithms with their default options and hyperparameter values to identify the top three performing models for further analysis and hyperparameter tuning. The seven algorithms included k-nearest neighbours, linear regression, random forest, gradient boosting, support vector machine, gaussian naive bayes, and a neural network model. The initial neural network model consisted of three hidden layers each with 32 nodes and 25% dropouts. Among the seven initial models, the linear regression model showed an odd behaviour on the test set despite performing well on the training set. This was manifested in the form of a large negative value for R-squared and a very large value for

Root Mean Square Error (RMSE). This behaviour must be associated with a bug in scikit-learn linear regression algorithm. Since the performance of the linear regression model on the training set was not among the top three models, it was dropped from further analysis. Gaussian naive bayes and support vector machine models did not perform well on either the training or test sets (Fig. 3). Moreover, k-nearest neighbours model performed less satisfactory compared to the remaining three models (Fig. 3). Thus, random forest, gradient boosting and the neural network models were selected for further analysis and hyperparameter tuning.

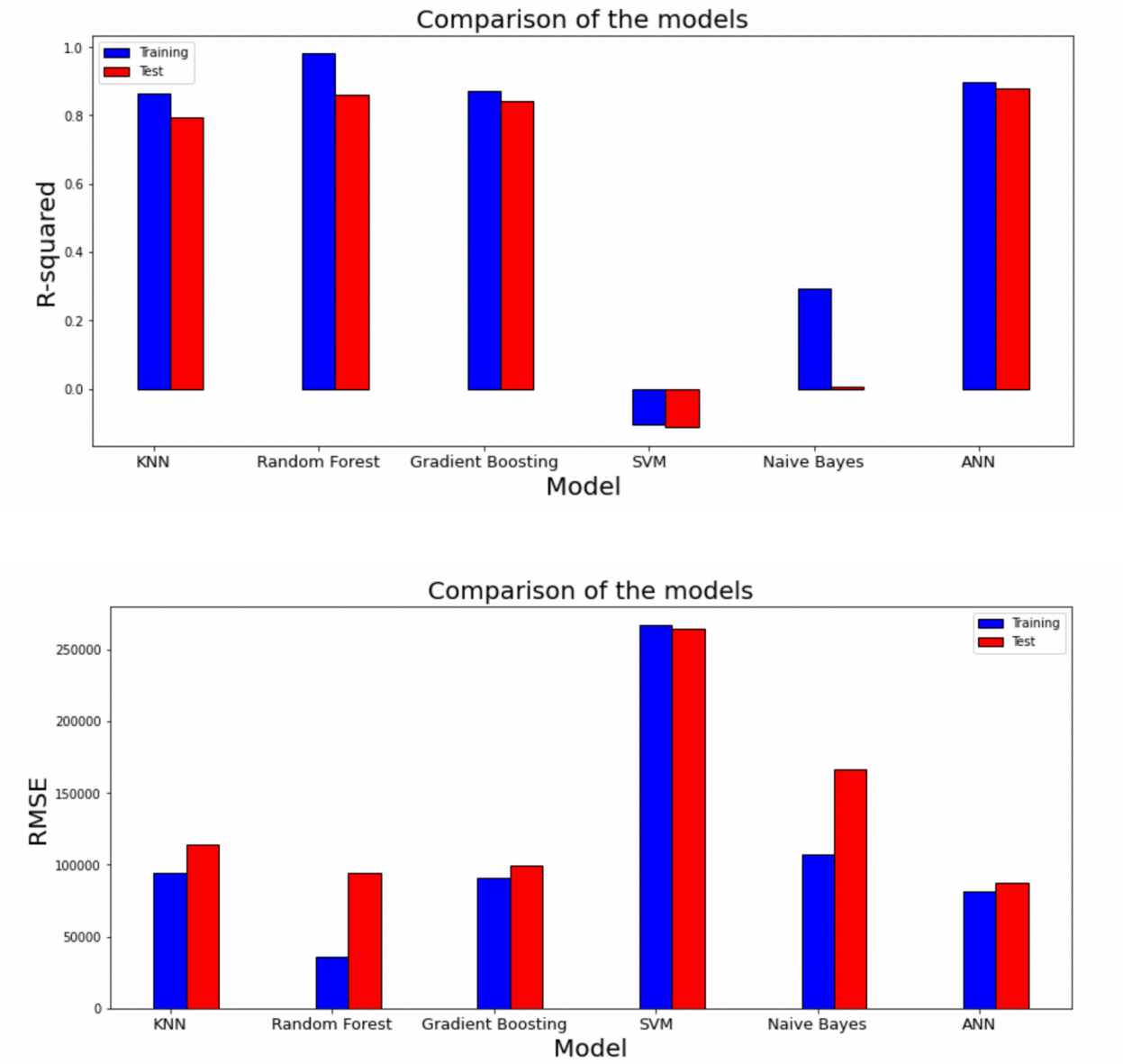

Figure 3. Modelling performance of different algorithms based on R-squared and RMSE

The random forest model was the first model to go under comprehensive hyperparameter tuning by randomized search cross validation. One hundred random combinations of the hyperparameters were tested through four-fold cross validation. The gradient boosting and the neural network models were similarly tested by 100 combinations of the hyperparameters through four-fold cross validation. While the random forest model had the best performance on the training set, the optimal gradient boosting and the neural network models outperformed the optimal random forest model with respect to both R-squared and RMSE on the test set (Fig. 4). Moreover, the gradient boosting model had a slightly better performance than the neural network model on both metrics for both the training and test sets (Fig. 4).

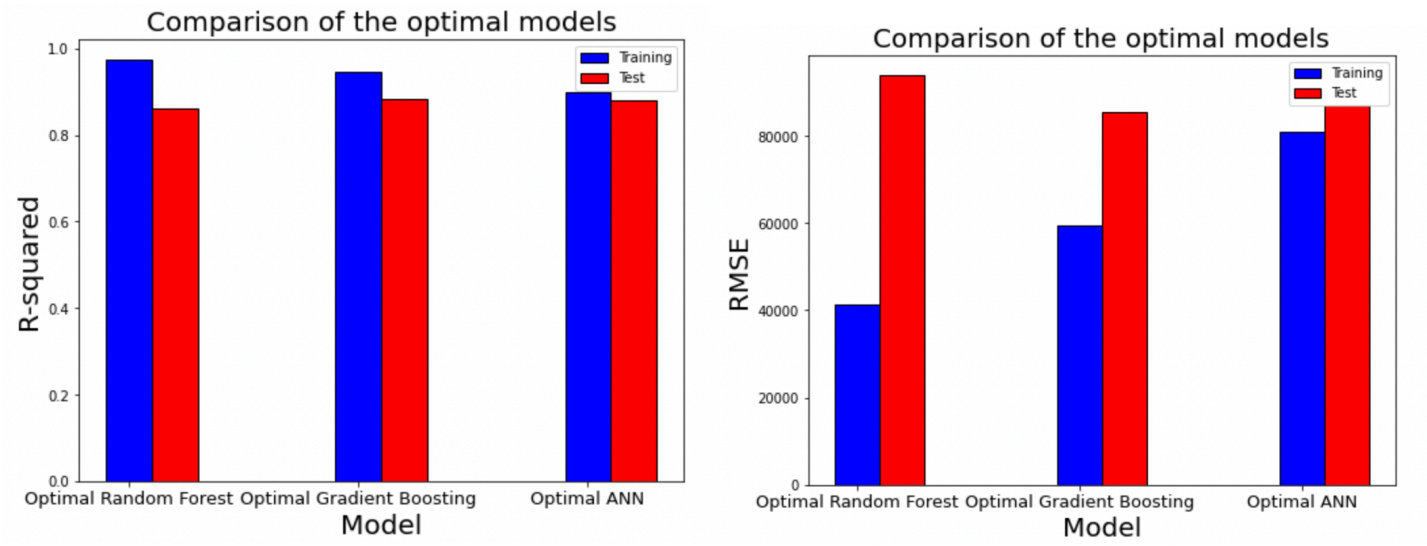

Figure 4. Modelling performance of the three optimal algorithms

Thus, the gradient boosting model with the following hyperparameters was selected as the optimal model for predicting housing prices in Nashville metropolitan area:

Loss function to be optimized: Huber
Learning rate: 0.01
Number of boosting stages to perform: 1200
Minimum number of samples required to split an internal node: 5
Minimum number of samples required to be at a leaf node: 10
Number of features to consider when looking for the best split: square-root of number of features
Maximum depth of the individual regression estimators: 50

The other hyperparameters were not changed from their scikit learn default values.

Relative importance of the features for the optimal gradient boosting model were also analyzed (Fig. 5). As expected, living area had the highest importance, while factors such as bedrooms and bathrooms, age, location, etc. played an important role as well.

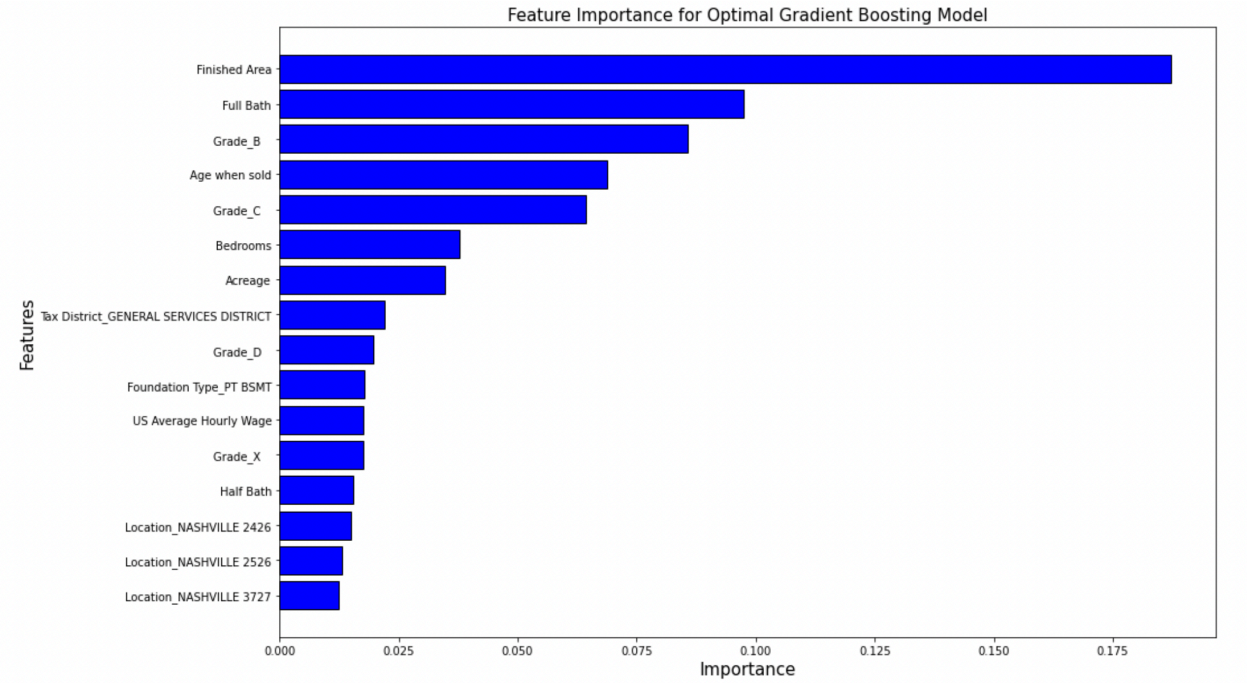

Figure 5. Relative importance of the features for the optimal gradient boosting model

Recommendations

1. The selected model should be utilized with care because on average the predicted price may still be around $85,000 different from the actual price of a property. It, however, should be noted that the estimates should be more accurate for properties with average prices and less accurate for extremely expensive properties.
2. The selected model can be used to identify when a property is overpriced or underpriced. This provides good information for whether or not to buy a property at the proposed price.
3. The selected model can help sellers to get a reasonable estimate of the value of their property, and provide them with information on potential improvements to increase their property's value.

4. To use the selected model for a different timeframe than the period of analysis (2013-2016), it is necessary to analyze how prices vary with time. This may be achieved by using a simple average factor or through more sophisticated time series analysis.

Future Research

It would be interesting to analyze how the selected model could be combined with a time series model for predicting housing prices and come up with predictions for present time. It would also be interesting then, to test the proposed model in a different city and analyze whether its predictions remain robust and accurate. It is however clear that some local considerations would need to be incorporated first so the model would be transferable to a new location. For instance, neighbourhoods would need to be analyzed for the new location. Another interesting research direction for the future could be to put focus on the most expensive properties and see how different models would perform on estimating these extremes.