DCS 550 Data Mining (DSC550-T302 2227-1)

Bellevue University

Term Final Project: Prediction of House Prices

Author: Jake Meyer

Date: 8/13/2022

Introduction

The accurate prediction of a house price is beneficial to many different stakeholders. Homeowners, realtors, and real-estate companies are often interested in the important features that impact the price of a home. Realtors and real-estate companies need to ensure the listing value for a home is accurate prior to engaging with their clients. A homeowner needs to understand how a particular home improvement will enhance the value of their home. In addition, folks interested in purchasing or selling a home will want to understand whether the price matches the house value. The objective for this project is to develop a method for predicting house prices through various types of regression analysis. The aptitude to predict an accurate house price will be highly desirable for folks involved in the real-estate market. Stakeholders will benefit from this analysis by making informed data-driven decisions to maximize financial gains.

The “Housing Prices” dataset used in this study is from Kaggle (M Yasser H, 2022). The thirteen features include house price, area of home, number of bedrooms, number of bathrooms, number of stories, main road connection status, guest room status, basement included, hot water heater included, air conditioning included, number of parking spots, whether the house is in a preferred location, and the furnishing status. The dataset contains 545 observations for the given features. The intent for this project is to predict the price of a home based on the features included within this dataset. Other factors (such as location, home age, lot size, market conditions, home condition, landscaping status, etc.…) would be recommended to include in future analysis. A link to the dataset is provided below:

Kaggle “Housing Prices” Dataset – [Housing Prices Dataset | Kaggle](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/code).

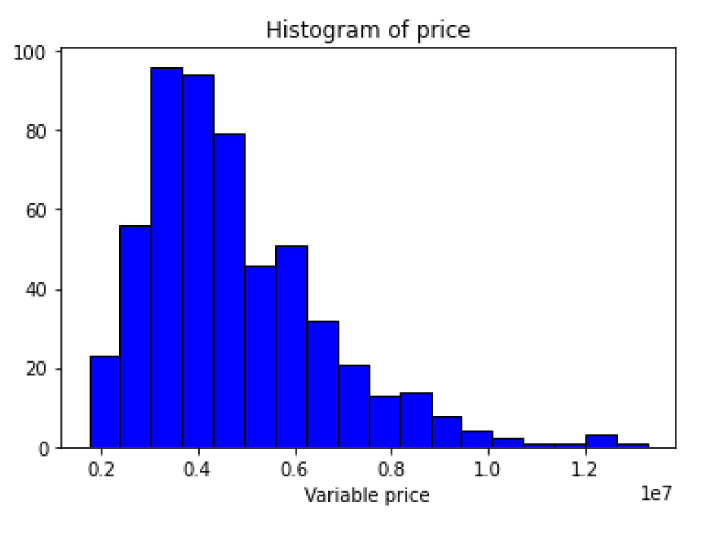
Milestones 1

The focus for Milestone 1 is on data selection and Exploratory Data Analysis (EDA). As described in the introduction, the dataset utilized for this analysis is from Kaggle. The questions and topics of interest considered during the EDA phase are shown below:

* Understand the distribution of the target feature “House Price”.
* Understand the count of values for each categorical feature.
* Do any of the categorical features appear to have a relationship with house pricing?
* Which features appear to be correlated with house price?
* Are there any missing values or outliers within the dataset?

The histogram for “House Price” helps illustrate the distribution. There is a positive skew with a Fisher\_Pearson skew value of 1.21. The histogram for this target feature is shown in Figure 1:

Figure 1: Histogram of House Price



The mean House Price is $4,766,729.00 and the median House Price is $4,340,000.00. This $426,729.00 difference suggests some of the data points may be influencing the mean value (on the high side). Outliers are identified and handled in the data preparation phase. The other feature identified with skew is the area of the home with a Fisher\_Pearson value of 1.32.

The Countplot for each categorical feature is shown below in Figure 2. These plots help visualize the balance for the categorical features present within the dataset. The tables in Appendix A also help quantify the count of unique values. The first set of tables show the ordered categorical data while the second set of data shows the categorical features without order.

Figure 2: Countplots of Categorical Features

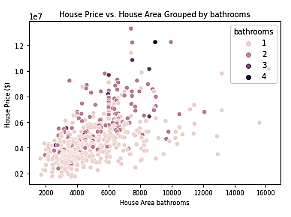
Square

Description automatically generated

The categorical features are plotted with boxplots in Appendix B. There is not a great balance for the categorical features within this dataset. Majority of homes had three bedrooms, one bathroom, one or two stories, connected to a main road, did not include a guest bedroom, did not include a basement, did not have hot water heating, did not have a parking spot, were not located in a preferred location, and were either fully furnished or semi-furnished.

There is a direct relationship between House Price and House Area. The scatterplots shown in Appendix C help illustrate the relationship between these two variables. As Home Area increased, the House Price tended to increase. Logically, the visualizations from the box plots make sense to explain the relationship between the categorical variables and House Price. Homes with more bedrooms, bathrooms, stories, connections to mainroads, guestrooms, basements, hot water heaters, air conditioning units, parking spots, in a preferred area, and furnished (to some degree) tended to have higher House Prices compared to homes that did not. The degree at which the House Prices increase did vary for each feature suggesting the relationship was stronger for some features compared to others. The Correlation Heatmap in Appendix C helps show the strength of the relationship between House Price and other numeric features. The features with the strongest correlation with House Price are Home Area, Number of Bathrooms, and Number of Stories respectively. All the features with numeric data currently indicate a positive correlation with House Price, however the correlation strength appears to be moderate to low. There are no missing values in this dataset. During the data preparation step (next phase), dummy variables are created for the remaining categorical variables. An example of a House Price vs. House Area Scatterplot by Number of Bathrooms is shown below:

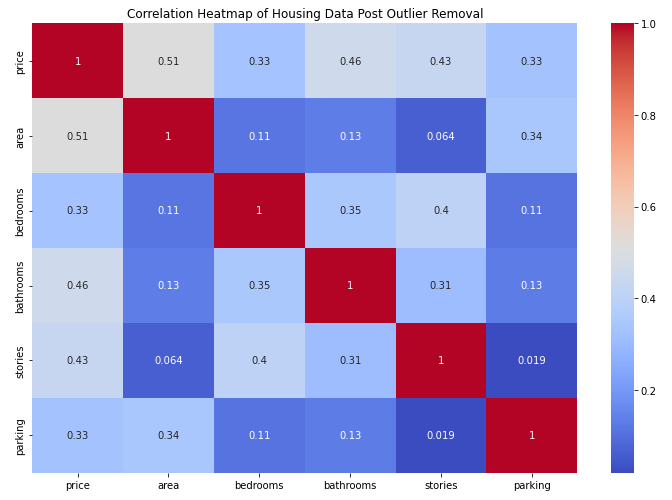
Figure 3: House Price vs. House Area by # of Bathrooms



Milestone 2

Data preparation is essential prior to the modeling phase. The steps taken for data preparation are handling outliers, eliminating nonessential features, splitting the data into training/test datasets, converting categorical features to dummy variables, and investigating feature selection opportunities. Fifteen outliers were removed using the interquartile range method. All fifteen of the outliers were on the upper side of the interquartile range which matches the direction of bias seen between the house price mean and median values. The removal of the outliers reduced the skew present within the dataset for house price. The revised Fisher\_Pearson skew value is 0.69 and the Correlation Heatmap is shown below:

Figure 4: Correlation Heatmap Post Outlier Removal

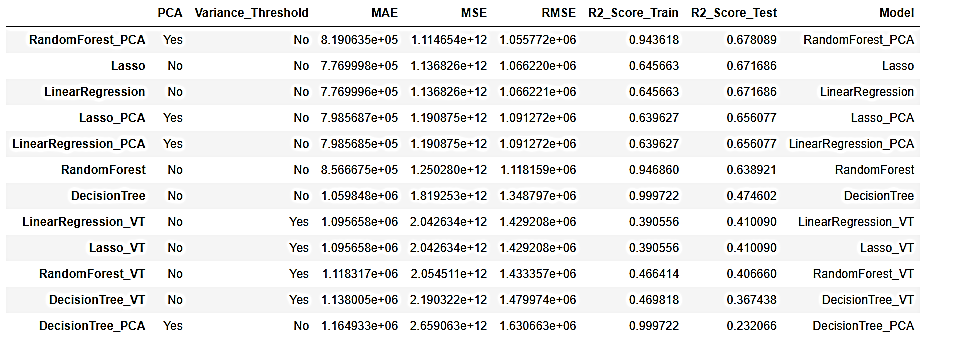


The Hot Water Heater feature was dropped due to the major imbalance. Only 4% of homes in this data set included a hot water heater. All other features are kept for the model building and evaluation phase. The dataset was split into Train (80%) and Test (20%) subsets. As previously discussed, the target feature was specified as the House Price. Dummy variables were created for the categorical columns without any order. The first dummy variable that was created was dropped to eliminate collinearity between dummy variables. There were three versions of training and test datasets prepared for the model building and evaluation phase. The first training and test data set include all the features after the dummy variables were created. The second training and test dataset had PCA applied to reduce the features down to 10 based on 90% of variance being retained. The third training and test data set included a Variance Threshold to reduce the features down to 7 based on a threshold set to 0.1. These dataset variations allow for several options to be performed during the model building and evaluation phase.

Milestones 3

The four machine learning model types evaluated in this section were Linear Regression, Decision Tree, Random Forest, and Lasso Regression. Each model was trained on the three different types of training data. The models were then evaluated based on R-Squared for test data set, R-Squared for training data set, Mean Absolute Error, Mean Square Error, and Root Mean Square Error. The summary of the metrics are shown in the table below:

Table 1: Model Evaluation Table



The best performance on the test dataset is from the Random Forest Model with PCA performed. The R-Squared Value for this model turned out as 0.6781 when evaluated on the test data. Although this performed the best based on the R-Squared Value, this is still not an ideal model to use for predicting home prices. The R-Squared value indicates there may be some additional features influencing the home price that are not included in this dataset. In addition, this model shows signs of being overfit to the training data due to the high R-Squared Value on the trained dataset. The most influential feature for this model is home area. Other important features for this model are number of bedrooms, number of parking spots, and number of bathrooms. The next two models that performed the best are the Lasso Regression and Linear Regression models trained without any PCA or High Variance Threshold. Both models performed similarly with R-Squared Values of 0.6712. The R-Squared Values were similar on the trained data set for both models as well. These models do not have the same overfit problem as the Random Forest Model.

Conclusion

The analysis/model building provides several summarizing arguments. The models that performed the best without any PCA or Variance Threshold were Lasso and Linear Regression. When PCA was applied, the Random Forest model performed the best. Lasso and Linear Regression performed slightly worse compared to their counterpart models trained without PCA or Variance Threshold. The Decision Tree model performed the worst with PCA. When Variance Threshold was applied, the performance for each model was worse compared to the initial models trained without any PCA or Variance Threshold. Caution with overfitting needs to be applied when considering the Decision Tree and Random Forest models.

Although the Random Forest model with PCA performed the best, the model is not recommended to be deployed. Recommendations for future improvements for this project are to perform hyperparameter tuning on the models to see if any additional improvements can be achieved for the model performance. In addition, the code can be consolidated using Pipelines and/or functions to improve readability. Cross-validation may be a good idea to include in the analysis to further evaluate the selected model's performance. Lastly, additional data such as more samples or new features may help understand some of the unexplained variance left out this dataset.

References

Albon, Chris Albon. Machine Learning with Python Cookbook, O-Reilly, 2018

Yasser, M. (2022). *Housing Prices Dataset.* Kaggle. [Housing Prices Dataset | Kaggle](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/code)

Appendix A – Tables for Count of Categorical Features (Ordered and Non-Ordered)

Table 2: Ordered Categorical Feature Counts

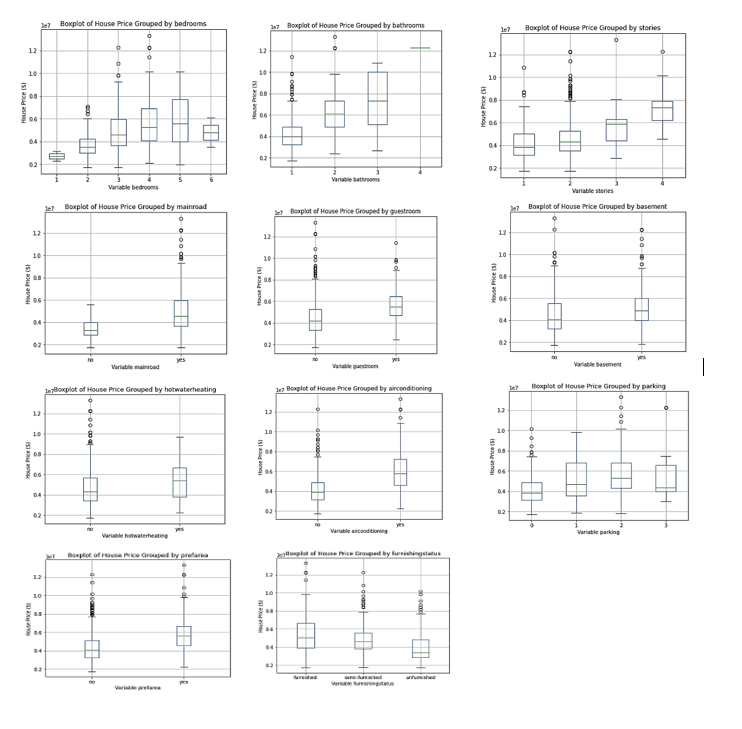


Table 3: Unordered Categorical Feature Counts



Appendix B – Boxplot of Categorical Features

Figure 5: Boxplots of Categorical Features



Appendix C – Relationship Plots- Scatterplot, Correlation Heatmap, and Pairwise Plot

Figure 6: Scatterplots of House Price vs. House Area by Categorical FeaturesCalendar

Description automatically generated

Figure 7: Correlation Heatmap

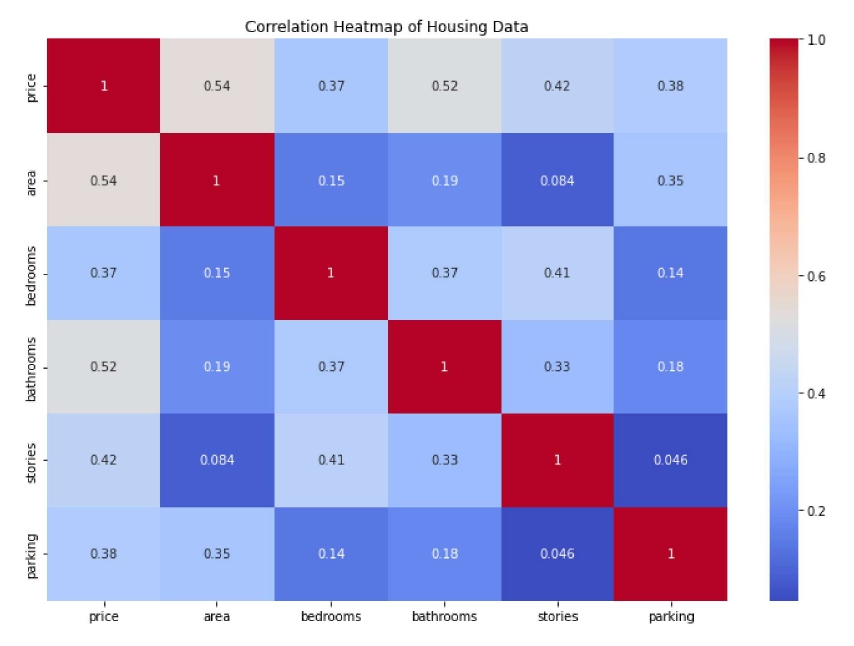


Figure 8: Pairwsie Plot for Housing Data Features

