**Anirudh Vunnam**

**06/26/2022**

**Introduction**

In this assignment, Housing dataset containing records of house prices in different areas in the United States between the years 2013 to 2016. This dataset contains variables that could influence the pricing of houses, these include the variable 'sale price', variable 'acreage', variable 'city', and so on. By analyzing the dataset, the real estate company can question what causes people to buy properties based on the output of the analysis. Conclusions on the manner in which houses are going over their asking price have to be achieved regarding the overpricing of the properties. To achieve this, the report focuses on interpreting and analyzing the relationship between sale prices and other property records related to it. By doing this, the data analyst can use the conclusions obtained from the analysis, as a reference for accurately pricing properties to attract prospective buyers.

**Data cleaning**

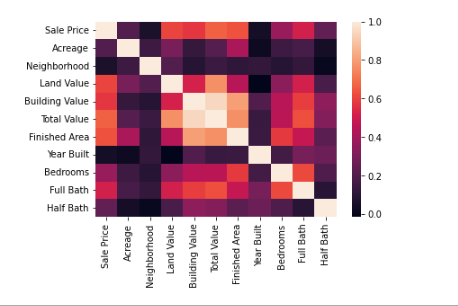
The housing dataset analyzed contains 56636 records and 31 variables. These variables have to be tweaked to design the prediction model using clean data. By checking for missing values ,we have found out that the data has multiple variables missing values, The most being the variable ‘Suite/ Condo #’ which is missing about 95% of the required values, while variables 'Owner Name’, variable 'Address', variable 'City', variable 'Acreage', variable 'Tax District', variable 'Neighborhood', variable 'image', variable 'Land Value', variable 'Building Value', 'Total Value', 'Finished Area', 'Foundation Type', variable 'Year Built', variable 'Exterior Wall', variable 'Grade', variable 'Bedrooms', variable 'Full Bath’, and variable 'Half Bath' missing almost half the values. Outliers only exist in the variable named “State”. We then explore the features with categorical classes to check for imbalanced data, and by doing this. It was found that the variable 'Sold as Vacant' and variable 'Multiple Parcels Involved in Sale" are binary categorical features with imbalanced data and to prevent false outlier detection of the minority classes in the above-mentioned features, we will be excluding them from the outlier detection and removal step. After correcting the outliers and dropping variables that could significantly cause inaccuracies the data shape changed to have 56636 records and 24 variables that are significant to create an accurate model.

**EDA**

In this section, we will explore the different variables that may drive house prices to a certain degree. From the data, we identify the data types in place, and that is float64 has 10 variables, int64 has 1 variable, which is Sale Price, and objects are 13. The exploration is divided into two parts: numerical variables and categorical variables.

**Understand numerical variables**

To understand numerical variables, correlations between all numerical features that is variable 'sale price', variable 'acreage', variable 'Neighborhood', variable 'Land Value',variable 'Building value',variable 'total value',variable 'finished area',variable 'year built',variable 'bedrooms', variable 'full bath',and variable 'half bath'. Based on the heatmap of correlations, some preliminary conclusions related to the problem can be obtained. We can already see that the building value of houses correlates with the house's total value. Most numerical features have a weak correlation with the variable “sale price”, and according to the color depth, it can be judged that variable “land value” and variable “total value” have a high degree of correlation. However, the strong correlation



*F2. Heat map of correlations*

Between variable “Finished Area” and variable “Total value” requires further analysis. Based on the above preliminary conclusions, we can think that among the numerical variables, variable “Land Value”, variable “Finished Area”, and variable “Full Bath” have a significant impact on the pricing of a particular house. These three variables need to be focused on in the next analysis part to understand the data accurately.

**Understand categorical variables**

To understand the impacts of categorical features, count plots were drawn to judge whether each category variable has a significant impact on the overpricing and underpricing of houses. After observing the table in the appendix, we can get a preliminary conclusion: the variable “Property City" shows that the prices in Nashville supersede any other city. We can also conclude that most of the prices hiked in 2015 and that most purchases were done in the middle months of the year. Among the remaining variables, the difference in the price status of different values of each variable is relatively small. To correctly analyze all variables further analysis needs to be done to understand the relationship between each variable. Based on the above preliminary conclusions, we can think that among the categorical variables, the variable "Property City”, variable “log\_sale\_prices”, and variable “land use” have a significant impact on the pricing of houses in the cities. These three variables need to be focused on in the next analysis part.

**Analysis**

**Modeling**

After completing the preliminary analysis of the data set by using different techniques to manage and shape the data, this report will complete further analysis of the factors that influence the house prices by establishing models based on the preliminary conclusions drawn in the exploratory data analysis. Here, the logistic regression model, decision tree model, Gradient Boost, Artificial neural network model, and Random Forest regression will be used to predict future trends. After the model is established, to accurately find a means to uniformly investigate the different variables involved. The model will be optimized to provide the best model features that are significant through feature engineering techniques to improve the accuracy of the model and to determine the categorization of house prices.

**Logistic regression model**

The first model we will investigate is the logistic regression model. Logistic regression has its own merits. Firstly, is that this model works pretty well with binary classification problems and the problem being studied involves a dependent variable. Secondly, the model is great at interpreting data involving multiple variables. The influence of different features on the final result and these variables are significant in generating decisions for the company when seeking profitability in attracting prospective customers. Divide the data set into a training data set and a testing data set at a ratio of 3 to 1, and the variables that have NaN values are replaced by a constant zero. At the same time, to prevent errors one of the values of each categorical variable is deleted. Finally, we design a logistic regression model that is established for the prediction of variable relationships in determining house prices. We can get some conclusions. Among the numerical variables, variable “acreage”, variable “building value”, and variable “land value” have significant effects on the pricing of houses, and the variable's total value' whose coefficients are positive, in turn, means that the higher the total value of land, the more the sale price. At the same time, the influence of all the zero values under the values is significant and positively correlated. Concurrently, the better the bathroom space, the greater the house value

**Decision tree model**

The second model we analyze is the decision tree model. To use it in the analysis we exploit the advantages that it offers. The decision tree structure is not only easy to explain but can handle both numerical variables and categorical variables at the same time and have a good fit. Based on the result of the decision tree model in the appendix, from the results the top variables in importance are variable “property city” variable “acreage”, variable “land value”, variable “finished area”, and variable “building value”. When we combine the original data set, we can just say that dependent variables, “land use” should be used as an indicator to refer to how to price houses, this is because the above variables have a significant influence on the pricing of houses.

**Random Forest Regression**

The third model that we will analyze is the Random forest, this is an estimator that fits a couple of classifying decision trees handling different samples of the housing dataset and employs the use of averages to increase the prediction accuracy and manage the overfitting of data.to actualize a fully grown tree that can potentially be excess on some dataset, the default values for controlling the size of the tree have to be determined. This is done to reduce memory consumption. From the dataset. This model uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by the committee is more accurate than that of any individual tree. the single linear regression models of these variables, a linear regression model including all variables, a gradient boost model using the same inputs, and a random forest model using the same inputs. The results of this comparison were that the random forest model outperformed all the others, generating the best results and the smallest amount of error of any of the models. All the other models performed about as expected compared to the single linear model, indicating that the single linear model was, in fact, a good baseline for comparison.

**Gradient Boost**

In the fourth model, the gradient boost model is a technique used to classify and regress tasks, it works by giving a prediction model that is in the form of an ensemble of weaker models used in predicting occurrences. These are the decision tree that results from gradient boosted trees. It works by outperforming the random forest by regularization. This is the training set fitting done too closely to degradation of the model’s ability to general. A crucial use of gradient boosting is regularization since it allows shrinking the number of observations in the terminal nodes of the tree. This is used to ignore any splits that lead to nodes having a few training set instances. This model works efficiently with the data variables from the dataset to generate the predictive outcome of price evaluation. However, this model sacrifices interpretability and intelligibility. To achieve both interpretability and performance. The models' implementation is also difficult because of the higher computational demand.

**Neural Network**

In the last model, we will analyze the neural network model to predict the accuracies of the model data. This involves a series of algorithms that are used to identify the relationships between the different variables in the data set, this is done through a process that imitates the logic executed by the human brain. They do refer to systems of artificial neurons that manipulate the data to obtain predictive conclusions. This model works by adapting to input changes, this network works by generating the best possible results of an expression without the need to redefine the output criteria. This model is used on the data to predict the house price by using the significant variable “land value”, and variable “total value “.To determine the influence these variables have on the final sale price of the house.

**Optimize**

Based on the conclusions of the above models, the model will be optimized next to identify the best model to use. Because there is a high correlation between variable “total value” and variable “sale price” we have to preserve the characteristics of the two variables at the same time without interfering with the accuracy of the model affected by the correlation between the dependent variables. As a criterion used to judge, Property city is more operative than a state criterion

After optimizing the data set the accuracy of the decision tree model has been improved through various. Based on the visualization of model results we can find that the variable “land value” is still the most influential factor, so it should be used as the first level indicator to judge whether the price of the house will be overpriced or underpriced. We can conclude that the five variables, variable “total value”, variable “finished area”, variable “building value”, variable “acreage”, and variable “land use”, should be used as the main judgment indicators to predict whether the pricing of a house will be increased or decreased.

Finally, we can compare the F1 score which is the weighted average of precision and recall, and accuracy in the table below to get the conclusion that the optimization was a success.

|  |  |  |
| --- | --- | --- |
| ***model*** | ***accuracy*** | ***F1 score*** |
| Logistic regression | 0.79 | 0.27 |
| Decision tree | 0.81 | 0.46 |
| Neural Network | 0.89 | 0.36 |
| Random Tree Regression | 0.91 | 0.54 |
| Gradient Boost | 0.76 | 0.42 |

**Conclusion**

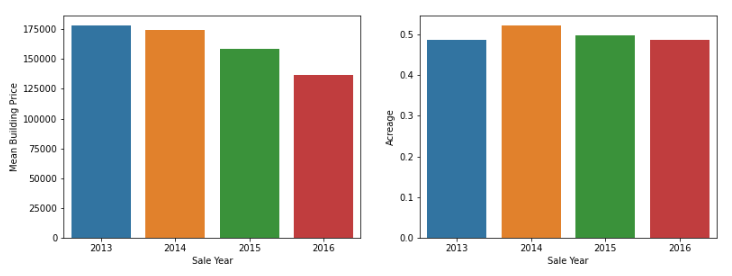
Based on the above analysis, we can find that it is not simple to accurately predict the factors that influence the setting of house prices. However, through the above series of analyses, we can still make a certain degree of judgment on whether a house is likely to be overpriced or underpriced. According to the conclusion of the model, we can find that the value placed on the land and building of the house has a significant impact on whether a house will be overpriced or not and this is of importance. The intention of overpricing houses is related to the area these houses are located. In addition, when putting a price on a house, it should be noted that according to the conclusions drawn by our model, the better the land value and building materials used or the structure of the houses the more price it holds.

**Reference**

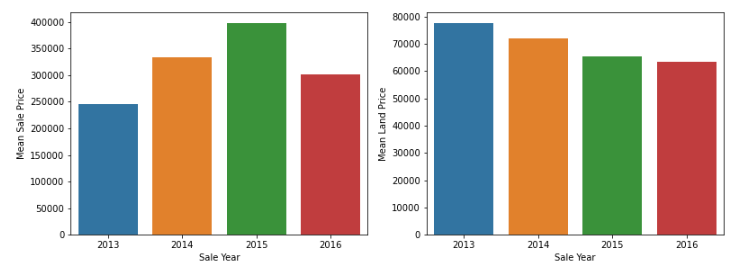
House price growth. (2019). https://doi.org/10.1787/fe060d6b-en

**Appendix 1**

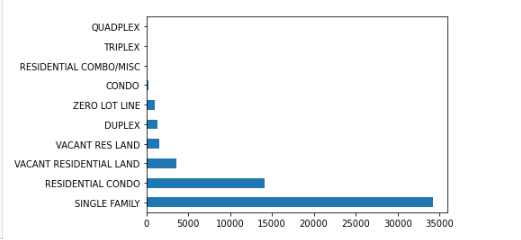
Visualization of the Mean of Building Price and Acreage against Sale Year



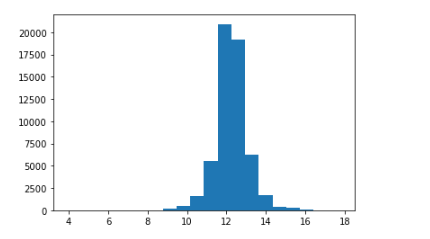
Visualization of the Mean of Sale Price and Land Price against Sale Year



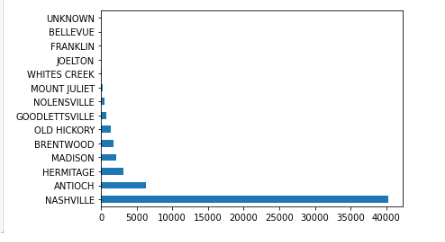
Visualization of the Land Use against the Sale prices



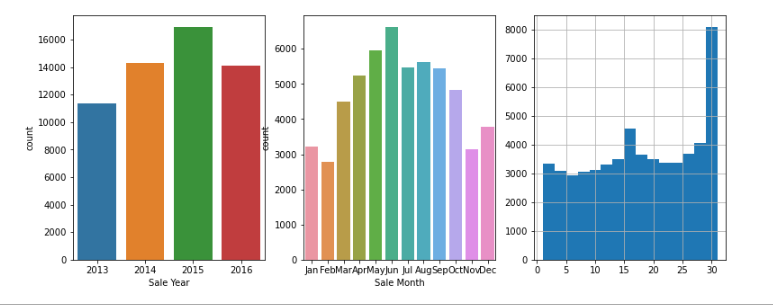
Visualization of the Log sale Price



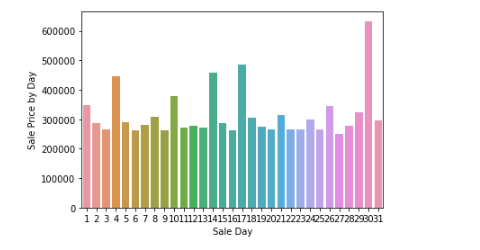
Visualization of Property City against Sale Price



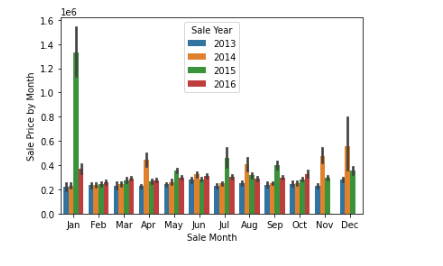
Visualization of Count Plot on Sale Year, Sale Month, and Sale Date



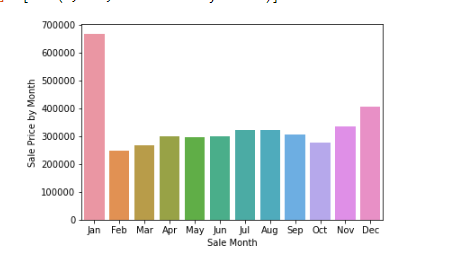
Visualization of Sale Price by Day against Sale Day



Visualization of the Sale Price by Month against Sale Month

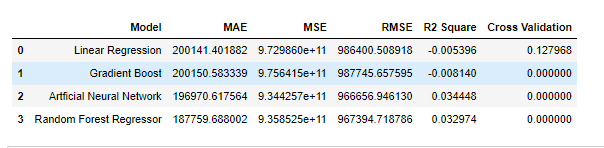


Visualization of the Sale Price by Month against Sale Month

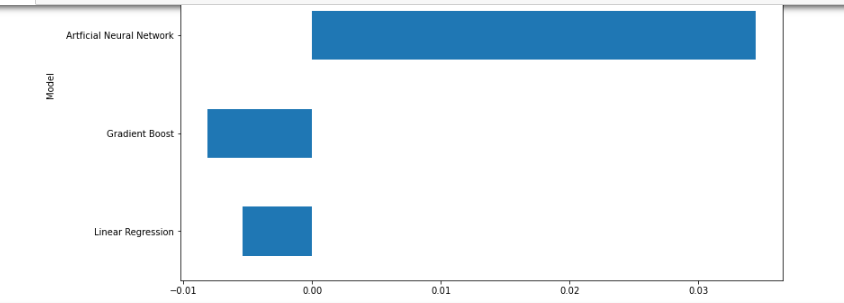


**Appendix 2**

The summary of model comparison

****

Visual representation of the Model Performance

****

**Appendix 3**

The summary of Neural Network performance

