Logo

Description automatically generated

ALY6020 Module 4 Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 05/11/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

Nashville Area is growing and attracting many investing opportunities from the real estate agencies. Working for a real estate company, we were given the responsibility to build a model to accurately predict the best value deals for the company as they are interested in investing in the area. This will help them in acquiring the best trade during their next visit to site in coming weeks. We planned to strategically decide which are the most important factors to determine the Sale value of the property using several models and predict which factors or variables to focus upon.

In this assignment we have used, Linear Regression Model, Decision Tree Regression Model, Random Forest Regression Model and Gradient Boost Model to approach to this week’s problem solution. We noticed there were huge number of missing values and outliers in our dataset hence proper imputation techniques and outlier treatment procedure were implemented to resolve those issues. I have also clubbed few columns values where I felt the numbers of levels were unnecessary huge. All this is further discussed in detail in the Analysis section below.

# Analysis

## Data Profiling & Cleaning

Nashville Housing Dataset contains the information about the properties collected between the time 2013 and 2016. Initial dataset has 56636 number of observations with 31 columns. We can see from the screenshot below there are several numerical (integer and float) and categorical (object) variables. Also, the count of records in each column suggests about the missing data in those columns which we will further investigate.

Table

Description automatically generated

Fig 1: Dataset Summary

A lot of properties registered in the dataset are from Nashville City and majority property owners are from Nashville as well. Although the Property City count, and the City count doesn’t match up. This could possibly suggest two probabilities that one the data is certainly missing from the City column and second that people from other cities are also investing in properties from Nashville. The second point strengthens the need of our analysis into the price value factors investigation in Nashville area more as we are concerned about investing there.

Text

Description automatically generated

Fig 2a: Property City

Text

Description automatically generated

Fig 2b: Property Owner’s City

We saw that the average Sale price of the properties is about $327,211 when compared with the Total Property Value which on average is $232,397. There is a great difference of about $95,000. This means the investment is good for future and has good ROI. The properties are little older averaging around 60 years with most of them having 3 bedrooms 2 full baths and on few occasions 1 half bath as well.

Table

Description automatically generated

Fig 3: Statistical Data Analysis of Numeric Variables

We can see most of the time the property is not sold as vacant and multiple parcels are also not involved in the sale. Maximum property is from Urban Services Tax District so this could be the tax district we can focus upon while our next visit and could think of buying one there as this is a hot spot having hot properties listed. Also, there is a very good chance of buying a property available which has been used by single family over the time, thus less chances of going through a property dispute occurring over the time.

Graphical user interface

Description automatically generated

Fig 4: Statistical Data Analysis of Categorical Variables

We couldn’t find any duplicate values in our dataset.



Fig 5: Duplicate value count

I have then dropped few insignificant columns which I deemed won’t be helping in our further analysis before proceeding with our data cleansing process. In the below screenshot we can find the dataset after removed columns and having Null values in them which needs to be treated. Columns like Acreage, Tax District, Land Value, Building Value, Total Value, Finished Area, Foundation Type, Year Built, Exterior Wall, Grade, Bedrooms, Full Bath, and half Bath have huge amount of missing values.

Text

Description automatically generated

Fig 6: Null Values count in the dataset

Below screenshot reveals that maximum property has 3 Bedrooms, 2 Full Bath and mostly 0 Half bath.

Table

Description automatically generated

Fig 7: Unique Count of Bedroom, Full Bath, and Half Bath levels

Finally, we handled the numerical variables null values with their mean values and categorical variables with their mode value. Once we do not have any null values left, we can look for Outlier values in the variables to properly treat them. But before going there I converted float datatype variables and few object datatypes as integer datatypes as needed.

## Outlier Treatment

I focused upon variables Acreage, Land Value, Building Value, Finished Area, Year Built, Bedrooms, Full Bath, and Half Bath to check if they have any outliers and found most of them has extreme outliers. I have used Quantile based flooring and capping technique to handle outliers for Finished Area, Building Value and Land Value. Post treating these outliers we got a better skewed value for these variables, all of them left skewed.

A picture containing diagram

Description automatically generated

Fig 8a: Outlier Detection

  

Fig 8b: Skewness Value post Outlier Treatment

## Categorical Data Encoding and Correlation Measure

Before label encoding the categorical variable, we decided to calculate the age of the property and drop the Year Built column from the dataset. I also clubbed different levels of Land Use variables under fewer levels to customize the dataset. The column level was then set to the levels as follows ‘SINGLE FAMILY’, ‘CONDO’, ‘VACANT RESIDENTIAL LAND’, ‘DUPLEX’, ‘TRIPLEX ‘, ‘QUADPLEX ‘ and ‘Others’.

Correlation Heat map below is not very clear; however, it was observed that maximum positive relation Sales Price has is with whether Multiple parcels were Involved in sales or not and highest negative relation with Land Use. In the entire dataset the strongest positive correlation is between variables Finished Area Building Value. Building Value also has a very strong positive relation with Land Value. Exterior wall and Land Value has the strongest negative relation in the entire set. Following this Tax district also has quite a evident negative relation with Full Bath and Total Value.

Treemap chart

Description automatically generated

Fig 9: Correlation Heat Map

Upon checking the multicollinearity, we didn’t find any present for our dataset.

Text

Description automatically generated

Fig 10: Multicollinearity Output

## Linear Regression Model

We created a Train Test split as 70:30 ratio and set the random seed as 42. For linear regression model we set the random state as 30. We fit the model with all the variables in the beginning and got the variables which were insignificant to the model. We verified the same with finding the most significant variables using forward selection method. We got almost the same set of variables just one additional and we decided to keep it in the model. Set of most significant variables are as follows 'Multiple Parcels Involved in Sale', 'Total Value', 'Sold As Vacant', 'Land Use', 'Sale Year', 'Sale Day', 'Land Value' and 'Tax District'. Upon dropping insignificant variables, we re-ran our linear regression model and found the accuracy over the training set remains the same to be 11.4%. Upon fitting the model and predicting on the test dataset we got a slightly improved accuracy of 15%. This means our model is better in predicting Test dataset. Most significant variables comes out for this model to be ‘Sold As Vacant’, 'Multiple Parcels Involved in Sale' and ‘Sale Year’. Fit of the model is 773733.49.

Text

Description automatically generated

Fig 11: Linear Regression Model Summary

## Decision Tree Regression Model

The model accuracy for Decision Tree Regression Model comes out to be -33.87 which is not making sense to me. The Fit comes out to be 4965564.08. The most significant variables for this model are ‘Total Value’, ‘Land Use’ and ‘Sale Year’.

A picture containing qr code

Description automatically generated

Fig 12: Decision Tree

Chart, histogram

Description automatically generated

Fig 13: Decision Tree Significant Variables

## Random Forest Regression Model

The model accuracy for Random Forest Regression Model comes out to be 83%. The Fit comes out to be 346586.35. The most significant variables for this model are ‘Sale Year’, ‘Sale Day’ and ‘Land Use’.



Fig 14: Random Forest Regression Model Summary

Chart, bar chart

Description automatically generated

Fig 15: Random Forest Significant Variables

## Gradient Boost Model

The model accuracy for Gradient Boost Model comes out to be 46%. The Fit comes out to be 617730.32. The most significant variables for this model are ‘Multiple Parcels Involved in Sale’, ‘Land Use’ and ‘Sale Day’.



Fig 14: Gradient Boost Model Summary

Chart, bar chart, histogram

Description automatically generated

Fig 15: Gradient Boost Significant Variables

# Conclusion

1. All the models have their own set of weakness and strength.
2. Will not recommend Decision Tree because of very poor accuracy.
3. Random Forest has the best Accuracy.
4. Gradient Boost has better fit than Random Forest.
5. Land use seems to be the most common significant variable.
6. When was the last sale made seems to be an appropriate variable to predict the future sales price.

# Reference

1. BigLuluu. (2021, July 28). Nashville-real-estate-investment---linear-regression-decision-tree-random-forest-and-gradient-boos/Nashville real estate investment - linear regression, decision tree, random forest and gradient boost modeling.ipynb at main · BigLuluu/Nashville-real-estate-investment---linear-regression-decision-tree-random-forest-and-gradient-boos. GitHub. Retrieved May 11, 2022, from <https://github.com/BigLuluu/Nashville-Real-Estate-Investment---Linear-Regression-Decision-Tree-Random-Forest-and-Gradient-Boos/blob/main/Nashville%20Real%20Estate%20Investment%20-%20Linear%20Regression%2C%20Decision%20Tree%2C%20Random%20Forest%20and%20Gradient%20Boost%20Modeling.ipynb>
2. dhanashreechavan8. (n.d.). Investing-in-Nashville-logistic-regression-decision-tree-random-forest-gradient-boost/ALY 6020 - assignment 4.ipynb at main · dhanashreechavan8/investing-in-Nashville-logistic-regression-decision-tree-random-forest-gradient-boost. GitHub. Retrieved May 11, 2022, from <https://github.com/dhanashreechavan8/Investing-in-Nashville-Logistic-Regression-Decision-Tree-Random-Forest-Gradient-Boost/blob/main/ALY%206020%20-%20Assignment%204.ipynb>
3. Singh, D. (2019, October 22). Deepika Singh. Pluralsight. Retrieved May 11, 2022, from <https://www.pluralsight.com/guides/cleaning-up-data-from-outliers>

# Appendix

Note: Code is attached separately.