Investing in Nashville

Jyothi Chandrakanth

College of Professional Studies, Northeastern University

ALY6020: Predictive Analytics

Course CRN: 21499

Professor Justin Grosz

Introduction

A real estate agency intending to invest significantly in the rapidly booming Nashville region. The dataset we were given pertains to real estate. The dataset we were given pertains to real estate which consists of various attributes that has an impact on the pricing of the land. They've gathered data on previous deals and would like us to create a model to assist them in precisely identifying the best value offers. There is concern that properties may sell for more than their listing price, and this dataset can assist us in observing this. The objective is to determine which qualities have the greatest impact on land pricing.

Analysis

Exploratory Data Analysis

The dataset consists of 56636 rows and 31 rows in total. In the process of the data cleaning procedure, we examined for missing value to ensure that we had clean and unbiased data for the modeling process. In order to identify any anomalous data in the dataset, the unique function was used to evaluate each of the variables for the presence of any unique values. The variables for which there was no data dictionary were eliminated from the analysis. The rows containing less than 1 % of missing data were deleted. The outliers that were present were not removed as it could be beneficial for the analysis. After cleaning the data, we have 24013 records and 18 columns to work with for the modeling. A new variable, *PriceType,* was created by calculating the difference between the Total Value and the Sale Value to identify whether the land was undervalued or overvalued, and then categorizing the results based on the outcomes of the calculations. The results with zero and less were labeled as Underpriced, whereas the results with more than zero were labeled as Overpriced. Furthermore, the values were converted into binary form as 0's (Underpriced) and 1's (Overpriced) to make things easier to compute the Logistic regression.

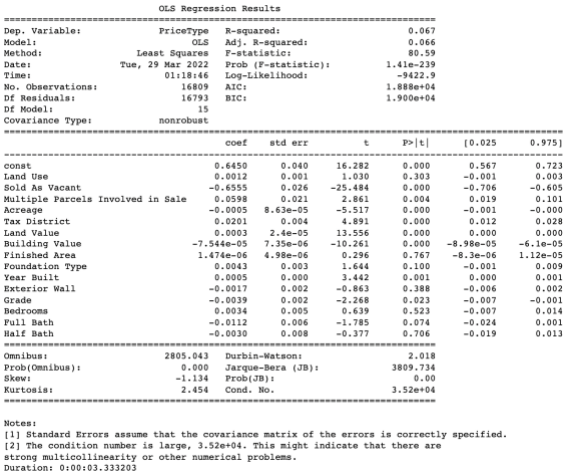
Modeling

1.Logistic Regression

It is used to statistically analyze and evaluate the relationship between the dependent variable and one or more independent variables. Ordinary Least Squares (OLS) regression analysis is a

statistical technique for detecting unknown parameters that comprises developing a strategy to decrease the sum of squared errors between observed and predicted values.

Because logistic regression evaluates data in numeric form, the dataset was transformed to numeric form using label encoder. *PriceType* is the dependent variable, whereas the other variables are the independent variables. The dataset was divided into train and test data in a 70:30 ratio. The model was then fitted, and the results are shown below.



OLS Regression Results

Dep. Variable:	PriceType	R-squared:	0.067
Model:	OLS	Adj. R-squared:	0.066
Method:	Least Squares	F-statistic:	80.59
Date:	Tue, 29 Mar 2022	Prob (F-statistic):	1.41e-239
Time:	01:18:46	Log-Likelihood:	-9422.9
No. Observations:	16809	AIC:	1.888e+04
Df Residuals:	16793	BIC:	1.900e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6450	0.040	16.282	0.000	0.567	0.723
Land Use	0.0012	0.001	1.030	0.303	-0.001	0.003
Sold As Vacant	-0.6555	0.026	-25.484	0.000	-0.706	-0.605
Multiple Parcels Involved in Sale	0.0598	0.021	2.861	0.004	0.019	0.101
Acreage	-0.0005	8.63e-05	-5.517	0.000	-0.001	-0.000
Tax District	0.0201	0.004	4.891	0.000	0.012	0.028
Land Value	0.0003	2.4e-05	13.556	0.000	0.000	0.000
Building Value	-7.544e-05	7.35e-06	-10.261	0.000	-8.98e-05	-6.1e-05
Finished Area	1.474e-06	4.98e-06	0.296	0.767	-8.3e-06	1.12e-05
Foundation Type	0.0043	0.003	1.644	0.100	-0.001	0.009
Year Built	0.0005	0.000	3.442	0.001	0.000	0.001
Exterior Wall	-0.0017	0.002	-0.863	0.388	-0.006	0.002
Grade	-0.0039	0.002	-2.268	0.023	-0.007	-0.001
Bedrooms	0.0034	0.005	0.639	0.523	-0.007	0.014
Full Bath	-0.0112	0.006	-1.785	0.074	-0.024	0.001
Half Bath	-0.0030	0.008	-0.377	0.706	-0.019	0.013

Omnibus:	2805.043	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3809.734
Skew:	-1.134	Prob(JB):	0.00
Kurtosis:	2.454	Cond. No.	3.52e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.52e+04. This might indicate that there are strong multicollinearity or other numerical problems.
Duration: 0:00:03.333203

Figure 1. Logistic Regression results

The observed R-Squared value is 0.46. We can eliminate variables with a p-value of less than 0.05, as they are not significant. *Multiple Parcels Involved in Sale* and *Sold as Vacant* with the highest coefficient values of 0.0598 and -0.6555 respectively, are the two most significant factors among the other variables with values less than 0.05. It took 03.333203 seconds for the model to execute. The Accuracy of the model is 0.74 while the mean squared error is 0.25

Variable	p-value	Co-eff
Sold As Vacant	0.000	-0.6555
Multiple Parcels Involved in Sale	0.004	0.0598

2. Decision Tree

It is a supervised non-parametric learning approach for classification and regression. The objective is to learn simple decision rules using data attributes to build a model that can predict the value of a target variable. By separating the source set into groups depending on a feature values testing, a tree may be "trained." Recursive partitioning is the process of repeating this method on every resulting subset. When all of the subsets at a node have the same value, the iteration is completed.

The previous model's data, which was partitioned into train and test, was fit. The MSE score was 0.23, while the accuracy was 0.76. The model took 00.115636 seconds to complete.
Based on the decision tree model the two most significant elements that determine the pricing of the land are *Sold as Vacant* and *Year Built*

3. Random Forest

Random forest is a Supervised Machine Learning Algorithm that is employed extensively in Classification and Regression applications. It constructs decision trees on distinct samples and uses the majority of votes for classification and average in case of regression. One of most essential properties of the Random Forest Algorithm is that it can deal the data set comprising continuous variables as in the regression problems and categorical variables in case of classification. It produces superior outcomes for classification issues.

The MSE value of the model was 0.26, whereas the accuracy obtained was 0.73. The model took 01.937160 seconds to run. Based on the Random Forest model the most significant factors that impact the pricing of the land are *Acreage, Land Value, Building Value, Finished Area and Year Built.*

4. Gradient Boosting

Gradient boosting is a strategy that stands out for its predictability and efficiency, especially when dealing with big and complicated datasets. The fundamental idea behind this technique is to develop models in a sequential manner, with each model attempting to decrease the mistakes of the preceding model. It's a strategy for combining predictions from multiple models into one by evaluating each prediction one at a time and modeling it depending on its predecessor's error

The MSE value of the model was 0.24, and the accuracy gained was 0.75. It took 18.152734 seconds for the model to execute. Based on the Gradient Boosting model the two most significant elements that determine the pricing of the land are *Sold as Vacant* and *Year Built*

Comparison of all the models

Model	Accuracy	Precision	Recall	MSE	Duration (In seconds)
Logistic Regression	0.74	0.74	0.99	0.25	03.333203
Decision Tree	0.76	0.76	0.99	0.23	00.115636
Random Forest	0.73	0.76	0.90	0.26	01.937160
Gradient Boosting	0.75	0.77	0.95	0.24	18.152734

Conclusion

We may conclude from the above findings that the Dssecision Tree model performed well, with the highest accuracy and lowest mean squared error and the duration to execute the model. According to the model the factors that have most influence on the overpricing and underpricing of the property are the *Sold as Vacant* and *Year Built*. To maximize the profit, the reality agency should concentrate whether the anyone living in the house when the house was sold and the year which the house was built.

References

Science, O. D. S. C.- O. D. (2019, May 30). *The Complete Guide to Decision Trees (part 1)*. Medium. Retrieved March 29, 2022, from https://odsc.medium.com/the-complete-guide-to-decision-trees-part-1-aa68b34f476d

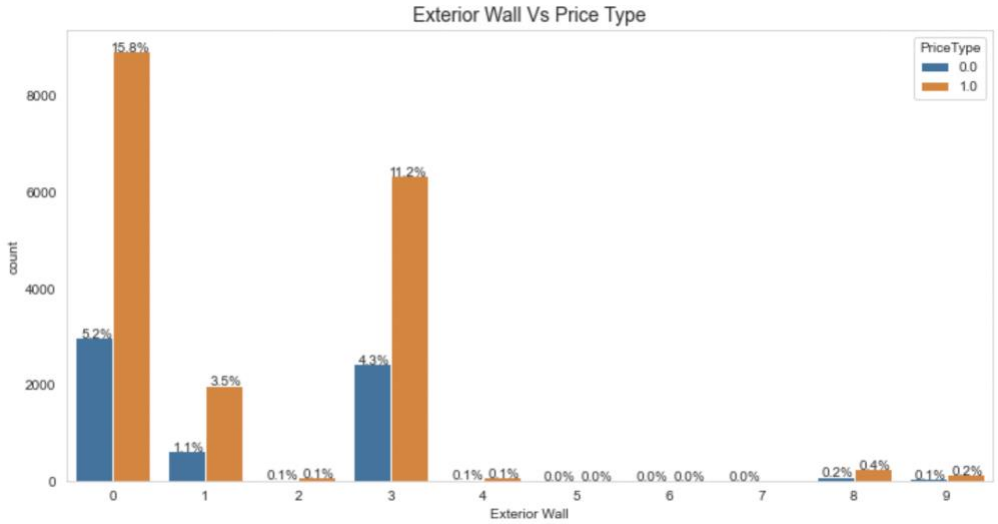
Sklearn.ensemble.randomforestclassifier. scikit. (n.d.). Retrieved March 29, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

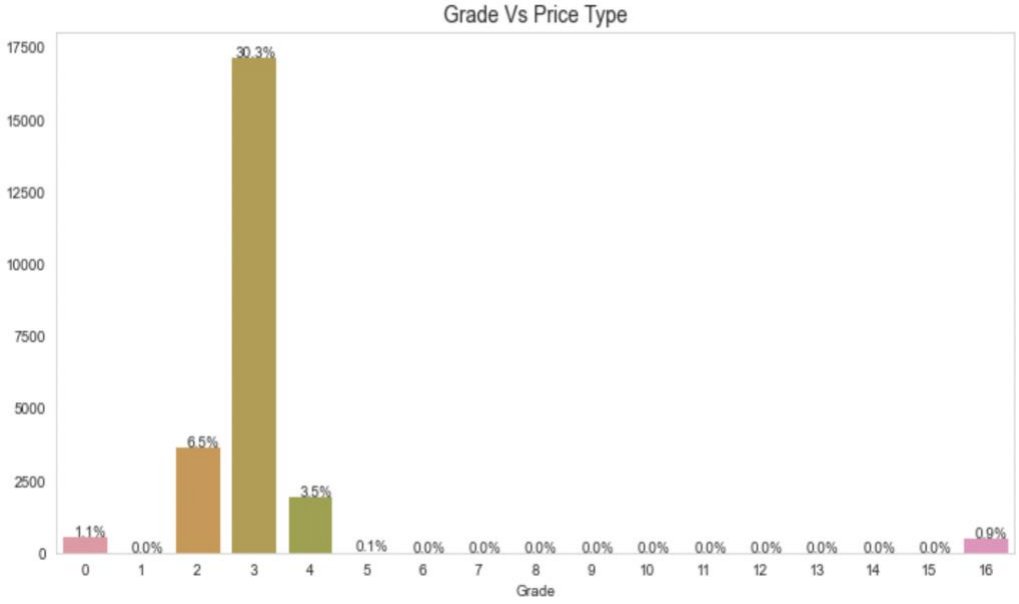
What is logistic regression? IBM. (n.d.). Retrieved March 29, 2022, from https://www.ibm.com/topics/logistic-regression

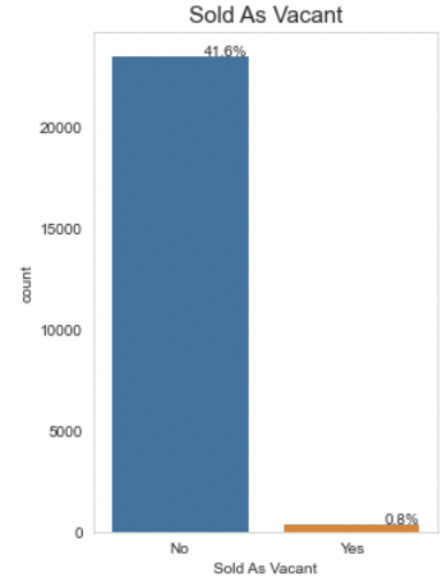
Kumar, A. (2020, June 30). *Introduction to the gradient boosting algorithm*. Medium. Retrieved March 29, 2022, from https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b

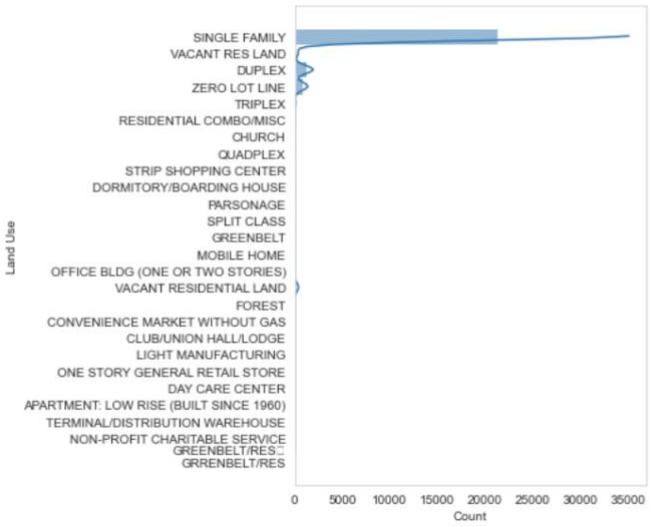
Appendix

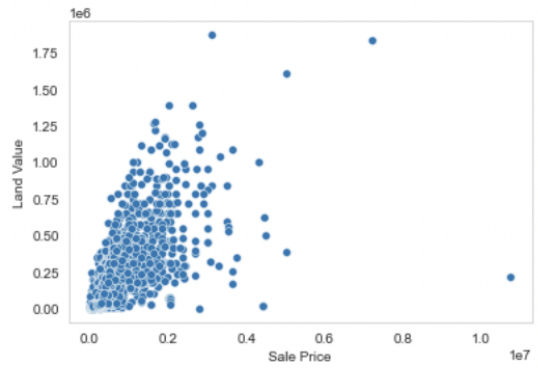
EDA

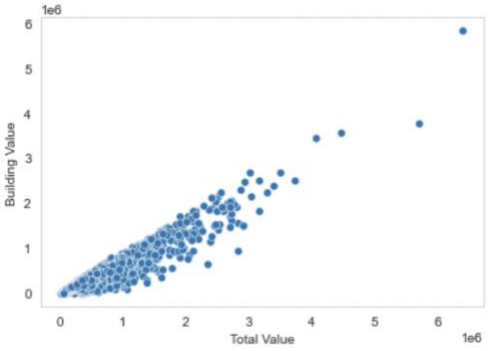












Decision Tree