**Introduction**

The goal of this project is to help the real estate company understand the trends in Nashville. We are given a recent sales dataset. We try to find which of these houses are going above their asking price. The company wants to understand whether these houses are selling at a higher price or not. We build a model that can help them make their decision as to whether or not they want to invest in Nashville.

**Data Cleansing**

Data cleansing is the process of identifying corrupt, inaccurate, or irrelevant data. This is the most critical or important step in data science. Not cleaning the data properly could have dire consequences on the end results and the accuracy of the model. In our dataset, we observe that many of the columns have missing data. Refer appendix 1 table 1, Acreage, Tax district, land and building value, total value, finished area, foundation type, year built, exterior wall, grade, full and half bath columns have more than 50% of the data missing. Algorithms do not work with missing data. The simplest way to deal with missing data would be to remove them. On further analysis, when we filter out rows with missing Acreage values it was found that all the other column data was also missing. If we use imputation or mean, mode or median to replace the data we would be replacing more than 50% of the data which could impact the model in a big way. We will be using the total value column to determine our dependent variable which is also missing in these rows. The best way to deal with this data would be to remove the rows with missing data.

Sold as Vacant and Multiple parcels involved in sale columns have yes or no values which can be converted to 1 or 0 because our algorithms expect all independent variables to be numerical. We replace yes with 1 and no with 0 for these columns. We now convert the object columns to numerical by one hot encoding them. We use one hot encoding on Land use, Tax district, Foundation type, Exterior wall, and grade. Now, the data doesn’t contain any object/string values, it doesn’t contain any missing data or outliers.

**Dependent and Independent Variables**

We now divide the columns into dependent and independent variables. Our data doesn’t contain a dependent variable, so we create one based on what answers we want or what we want to investigate. Since we are looking at the price of the houses and trying to determine whether our real estate company should invest in Nashville or not, we are creating a dependent variable called Pricing. Pricing can have 2 values, true (1) or false (0) and we determine this data using sale price and total value of the house. If the total value of the house is less than selling price, we know the house was overpriced and if the total value was more than the selling price, the house was underpriced. If overpriced, the value is 1 and if its underpriced, we assign the value 0. All the other variables are taken as independent variables.

**Analysis of the Algorithms**

1. **Logistic Regression**

It is a supervised machine learning algorithm. It is used to predict the probability of the dependant/ target variable. Binomial logistic regression model is one of the most popularly used regression models for Logistic Regression. It helps categorize data into two classes and predict the value of a Chart, scatter chart

Description automatically generatednew input as belonging to either of the two classes. It works by using a Sigmoid function to map the output probabilities. We put 0 and 1 on x and y axis and the graph looks as follows. We then create a confusion matrix for the results to check the

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | | | |
| **Actual** |  | **Positive** | **Negative** |
| **Positive** | 36 | 1198 |
| **Negative** | 34 | 3535 |

accuracy and results. We use sklearn library to perform these functions. Due to the inconsistency in the data after one hot encoding, i.e., a lot of 1s or 0s, the model creates a singular matrix and fails to give the result. So, we have removed variables with high variance.

We can observe that the number of false positives are greater than the false negatives. Refer Appendix A Result 1 to see the classification report. We can see the most significant variables based on p values are sold as vacant, multiple parcels involved in sale, land value, building value, land use vacant residential land. We now check the coefficient values to see in what way these features are impacting our results. Sold As Vacant, if the land was used as vacant residential land and if the building value was high, these are negatively impacting which means if the these conditions are true, the probability of the houses being overpriced is less. While the value of the land and if multiple properties are involved in the sale have positive impact which means the probability of the houses being overpriced is more.

1. **Decision trees**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | | | |
| **Actual** |  | **Positive** | **Negative** |
| **Positive** | 113 | 1091 |
| **Negative** | 25 | 3574 |

Diagram

Description automatically generatedDecision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the

features of a dataset, branches represent the decision rules and each leaf node represents the outcome. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. We are using the Entropy criteria to do this classification. The splits aim to maximize the decrease in entropy i.e., data impurity. We see that same as logistic regression due to the imbalance in the data, we have good negative predictions but our model is not good at predicting positive values. We now plot a feature importance graph, which is derived from which feature was repeated the most in the tree. Refer appendix A Graph 1 for the results. We can observe that the four most important features are sold as vacant, building value, land value and year it was built.

1. Diagram

   Description automatically generated**Random Forest Algorithm**

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. We can give the maximum depth of the tree and how many trees the algorithm needs to create. We have checked for default 100, and for 1000 the confusion matrix for them are below. We can see that there is very little difference between the 2 and their accuracies. With 100 trees, it was 76.7% accurate and with 1000 trees it was 76.68% accurate.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | | | |
| **Actual** |  | **Positive** | **Negative** |
| **Positive** | 123 | 1111 |
| **Negative** | 8 | 3561 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | | | |
| **Actual** |  | **Positive** | **Negative** |
| **Positive** | 123 | 1111 |
| **Negative** | 9 | 3560 |

We also plot a feature importance graph as seen in appendix A graph 2. The 4 most important features are sold as vacant, year it was built, Land usage i.e., whether it was a residential vacant land and building value.

1. **Gradient Boosting**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | | | |
| **Actual** |  | **Positive** | **Negative** |
| **Positive** | 186 | 1035 |
| **Negative** | 25 | 3557 |

Gradient Boosting is a special type of Ensemble Learning technique that works by combining several weak learners(predictors with poor accuracy) into a strong learner(a model with strong accuracy). This works by each model paying attention to its predecessor’s mistakes and improves by reducing its errors. We can see like the other models, the false positives are high in number. The accuracy of this model is 77.9% which is greater than all the other models. We do not use this model to find the feature importance because of the lower accuracy of the trees they create in the beginning of learning.

**Comparison between the modules**

We now pick various benchmarks that are important to us to create a table for comparison, we are comparing the false negatives i.e. where we are saying it is underpriced but is actually overpriced, accuracy of the mode, significant/important features we got, precision i.e, ability of a classifier not to label positive to the negatives and speed in seconds since the epoch.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | False negatives | Accuracy  (%) | Significant/Important  Features | Precision for false | Speed |
| Logistic Regression | 34 | 74.5 | Sold as vacant,building value, land value | 0.51 | 0.2117 |
| Decision trees | 25 | 76.76 | Sold as vacant, building value,land value,year built | 0.82 | 0.1239 |
| Random forest n=100 | 8 | 76.7 | Sold as vacant,year built, land use if it was vacant residential land, building value | 0.94 | 1.0029 |
| Random Forest n=1000 | 9 | 76.68 | Sold as vacant,year built, land use if it was vacant residential land, building value | 0.93 | 10.1181 |
| Gradient Boosting | 25 | 77.93 |  | 0.88 | 7.7845 |

**Conclusion**

From the above analysis we can conclude the following

1. Our model is not good at predicting overpriced houses but very good at predicting underpriced houses which is good for the real estate company to know before they start investing.
2. Referring to the table above, we can see that random forest algorithm where the number of trees=100 has the lowest number of false negatives i.e., lowest number of wrongly labelled underpriced houses. Hence, it has the highest precision and a very good accuracy so the company should use this model.
3. From the above models, we can see that the most significant/important feature is sold as vacant i.e., was anybody living in the house (If the land use was vacant residential area also show up as important feature in the model). If a house is sold as vacant, it is mostly underpriced. Vacant house insurances generally cost a lot higher than occupied houses and have a slower emergency response time and the increased probability of a break-in occurring. The real estate company can look to invest in these houses. The company can upgrade security for these houses and increase the selling price when they provide their listings.
4. We can also see that the year it was built is another important feature that can be considered. Newer houses can be overpriced compared to older houses. The company can look at investing in houses that are older and then renovate them a little and sell at a higher price.

**Appendix A**

Table

Description automatically generatedTable 1: Missing Data

Table

Description automatically generatedResult 1: Logistic Regression ResultsTable

Description automatically generated

Graph 1: Decision tree feature importance

Chart

Description automatically generated with medium confidence

Graph 2: Random forest feature importance

Chart, bar chart

Description automatically generated

**References**

1. Decision Tree Classification Algorithm [Machine Learning Decision Tree Classification Algorithm - Javatpoint](https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)
2. Random Forest Algorithm [Machine Learning Random Forest Algorithm - Javatpoint](https://www.javatpoint.com/machine-learning-random-forest-algorithm)
3. Vacant Home Insurance [Vacant Home Insurance - Farmer Brown Insurance](https://farmerbrown.com/insurance/vacant-home-insurance/)