

**College of Professional Studies**

Name: Deepthi Jakki

Class Number and Name: ALY6020 Predictive Analytics

Assignment Name: Module4 Housing Dataset: Nashville

Instructor: Justin Grosz

**Introduction**

The real estate market in Nashville presents a dynamic and lucrative investment landscape. As the market grows, understanding the factors that drive property values is essential for informed investment decisions. Accurate valuation of properties is a crucial yet challenging aspect of real estate investment. The disparity between the sale price and the intrinsic value of properties can significantly affect investment outcomes. This complexity is driven by a myriad of factors, ranging from property-specific attributes to broader market trends.

The objective of this study is to leverage advanced data analytics and machine learning techniques to develop a series of predictive models. These models aim to accurately assess housing prices in the Nashville market and uncover the primary drivers of these valuations. The study involves:

* Establishing a foundational understanding of housing prices through a Linear Regression model, identifying key factors that, influence these prices.
* Enhancing the analysis with a Decision Tree model, examining its performance in comparison to Linear Regression.
* Expanding to more complex models like Random Forest and Gradient Boost, each adding a layer of sophistication and potentially improved accuracy.
* Conducting a comprehensive comparison of these models using multiple benchmarking metrics to ascertain the most effective approach for the real estate company’s investment decisions.

Through this analysis, the study will illuminate the key determinants of property values in Nashville, providing a nuanced understanding of what drives the real estate market in this region.

**Methodology**

This study adopts a methodical approach, encompassing data cleansing, exploratory analysis, model development, and rigorous evaluation. By comparing the strengths and weaknesses of each model, the analysis aims to provide insights into the most reliable methods for predicting housing prices and understanding the dynamics of the Nashville real estate market.

The findings and recommendations from this analysis will guide strategic investment decisions, ensuring that the real estate company capitalizes on the most lucrative opportunities in the Nashville market.

**Data Cleaning and Preparation**

The analysis is based on a comprehensive dataset detailing recent real estate transactions in the Nashville area. This dataset includes a wide range of variables such as property address, sale date, building value, finished area, and the key variable of interest – 'Sale Price Compared To Value'. An initial assessment of the dataset was conducted to understand its structure, completeness, and quality.

The data cleaning process involved several essential steps:

Removal of Irrelevant Features

Columns that were not relevant to the analysis, such as 'Suite/Condo #' and 'Unnamed: 0', were identified and removed. This step streamlined the dataset, focusing the analysis on impactful variables.

Handling Missing Values and outliers

Missing data was addressed by removing rows with incomplete information. This decision was made to preserve the quality of the analysis, given the ample size of the refined dataset. Considering the nature of the real estate industry the outliers are included in the study.

Data Type Conversion

The dataset was transformed to ensure data types were aligned with the analysis requirements. For example, the 'Sale Date' was converted to a DateTime format, facilitating temporal analysis.

Feature Engineering

Extraction of New Features

New features relevant to the Nashville market were derived from the existing data. For instance, extracting the year from the 'Sale Date' allowed for an analysis of price trends over time.

Encoding Categorical Variables

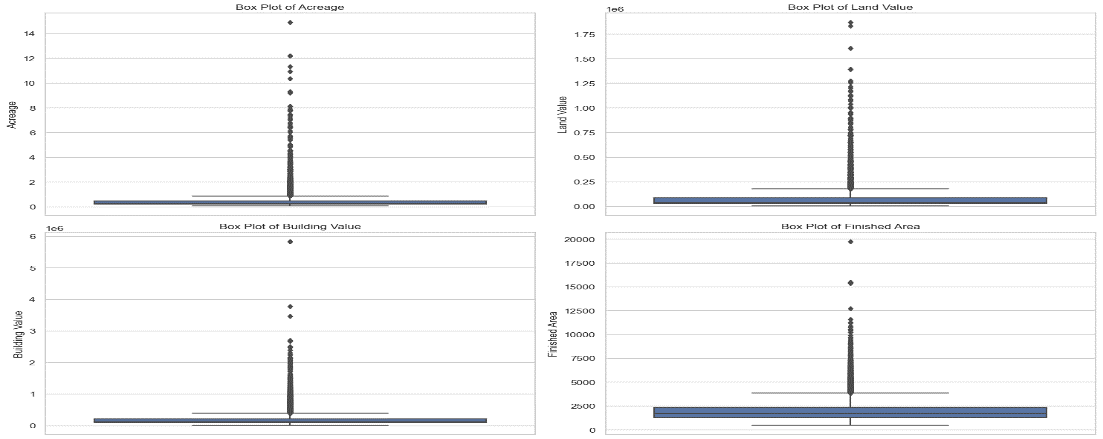
Categorical variables were transformed using techniques like one-hot encoding. This conversion of categorical data into a numerical format was crucial for incorporating these variables into the machine learning models.

Preparation

To align with the study's objectives, the dataset was filtered to include only those records pertaining to properties in Nashville. This refinement ensured that the subsequent analysis was concentrated on the market of interest.

**Exploratory Data Analysis**

The Nashville real estate dataset reveals a market characterized by diversity, offering a range of investment opportunities. The variance in acreage, from smaller plots to substantial landholdings, indicates a spectrum of property types, from compact urban residences to expansive rural estates. This diversity is crucial for a real estate portfolio that aims to cater to different market segments. The analysis revealed a significant variation in property sizes and values, indicating a dynamic and diverse Nashville real estate market.



The land and building values provide insights into the Nashville market's valuation trends. With average land values around $80,000 and building values at approximately $187,731, there's a notable range, highlighting the market's varied investment scales.

The predominance of 'Single Family' land use underscores a strong residential market focus. This is key for investors targeting residential property investments, offering insights into the demand patterns and potential growth areas in Nashville's residential sector.

The majority of properties not being sold as vacant indicates a market with a high turnover of established structures, a significant consideration for investors focusing on ready-to-occupy properties. Additionally, the trend of properties being classified as 'Over' in sale price compared to value suggests a competitive market with potential for high-value transactions.

|  |  |
| --- | --- |
| C:\Users\Admin\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\A6882D93.tmp | C:\Users\Admin\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B057428F.tmp |

**Preparing for Predictive Modeling: Addressing Multicollinearity through VIF Analysis**

Before delving into the construction of predictive models, a crucial preparatory step was undertaken: the Variance Inflation Factor (VIF) analysis. Multicollinearity arises when independent variables in a regression model are highly correlated. This correlation can lead to inflated standard errors and unreliable estimates of regression coefficients, making it challenging to discern the individual effect of each predictor. To mitigate the risk and enhance the reliability of the subsequent predictive models, variables with high VIF scores are eliminated from the analysis.

**Modeling**

Post-VIF analysis, the logistic regression model was built with a refined set of predictors. This ensured that the model's outcomes were based on variables with distinct and independent influences on the target variable, 'Sale Price Compared To Value'.

The Logistic Regression model was developed to predict the 'Sale Price Compared To Value' - whether properties in the Nashville real estate market are overvalued or undervalued. The model's coefficients offer insights into the factors influencing these valuations.

Interpretation of Model Results

Model Convergence: The model successfully converged after 6 iterations, indicating a stable solution.

Model Fit and Effectiveness:

The Pseudo R-squared value of 0.025, while modest, indicates that the model captures some of the variability in the data, but there's room for improvement.

Coefficient Analysis- Significant Predictors:

Sold As Vacant: A significant positive coefficient and an odd ratio above 29 indicates that properties sold as vacant are more likely to be valued differently from their sale price, possibly reflecting market perceptions of potential.

Building Value: Despite its relatively small coefficient, 'Building Value' has a profound statistical significance (p-value: 5.05e-24). This implies that even minor variations in building value can influence the likelihood of a property being considered overvalued or undervalued. It highlights the sensitivity of property valuation to the perceived value of the building.

Multiple Parcels Involved in Sale: The negative coefficient suggests that sales involving multiple parcels are less likely to be overvalued or undervalued.

Decision Tree Model:

This model is particularly good at capturing complex patterns that may not be apparent or easily modeled by a logistic regression. The decision tree identified 'Building Value', 'Sold As Vacant', and 'Land Value' as the top three features impacting predictions, with the importance of 0.377472, 0.353467, and 0.241714, respectively.

The decision tree has identified similar key features as the logistic regression model and has higher accuracy, its practical effectiveness in the context of the Nashville real estate market may be limited due to its low recall and ROC\_AUC scores.

Random Forest Method:

The Random Forest model provides a nuanced view of the Nashville real estate market, with its ensemble approach capturing complex interactions between features. The model identified 'Building Value', 'Sold As Vacant', and 'Land Value' as the top three features by importance, consistent with the previous models. This consistency underscores the strong influence these variables have on property valuation. The performance metrics show model is not effective in predicting true positive cases.

Comparison with Decision Tree and Logistic Regression Models

Improved Accuracy: The Random Forest model shows a marginal improvement in accuracy over the Decision Tree model, which may be due to the ensemble nature of Random Forests, where the aggregation of predictions from many trees tends to increase overall accuracy.

Precision and Recall: Unlike the Decision Tree model, the Random Forest model has a significantly higher precision but much lower recall. This implies that while it is more confident in its positive predictions, it fails to capture a large proportion of the actual positive cases.

Feature Importance Alignment: The Random Forest and the Decision Tree models align in identifying the most important features, suggesting these factors are robust predictors across different model types.

Gradient Boosting:

The Gradient Boost model was implemented to predict the binary classification of Nashville real estate as overvalued or undervalued. Gradient Boosting is known for its high performance, particularly in handling complex datasets with non-linear relationships. The model assigned the greatest importance to 'Building Value', 'Land Value', and 'Sold As Vacant', indicating these as the most influential predictors for the valuation status of properties. The model achieves an accuracy of 75.7% recall rate of 7.11%, precision of 50.4% and ROC\_AUC socre of 52.43%. The MSE model is 24.3%.

Comparison with Other Models

Accuracy and Recall: The Gradient Boost model has a similar accuracy to the Random Forest model, but like the Decision Tree and Random Forest, it struggles with recall.

Precision: The precision is slightly better than that of the Decision Tree model but lower than the Random Forest model, showing a moderate level of reliability in its positive predictions.

ROC\_AUC: The ROC\_AUC score is comparable to the Decision Tree and Random Forest models, suggesting similar abilities across these models.

The Gradient Boost model demonstrates a balance between accuracy and precision in predicting property valuation statuses in Nashville. However, the low recall indicates a potential issue in identifying all relevant cases, which could be critical depending on the business context. The consistency in the most important features across different models suggests these variables are key drivers of property valuation and should be focal points for real estate analysis and investment decision-making.

**Conclusion**

Given the company's objective to find the best value deals, which likely means prioritizing the identification of truly overvalued or undervalued properties (high recall), the Decision Tree model might be the initial choice due to its highest recall. However, its lower precision compared to the Random Forest model means that it would generate more false positives, which could increase the cost and time spent on investigating properties that are not mispriced.

If the company wants to minimize the risk of investing time and resources into properties that are not the best deals (high precision), then the Random Forest model would be the preferred choice. It offers the best balance of accuracy, precision, and reliability of predictions, although it may miss some true positives.

A potential strategy could be to use the Gradient Boost model to strike a balance between identifying true deals and ensuring the identified deals are likely to be correct. It offers an intermediate level of recall and precision, which might suit the company's need to balance the quantity and quality of investment opportunities.

The decision should also consider the operational costs associated with each model's false positives and false negatives.

**Bibliography:**

Canvas Resources, Aly6020 Predictive Analytics: Northeastern University (2023)

Annexures:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Model | Coefficient | Odds Ratio | Percent Change | P-Value |
| const | -1.128900 | 0.323389 | -67.661136 | 8.295121e-31 |
| Building Value | 0.000001 | 1.000001 | 0.000142 | 5.049526e-24 |
| Sold As Vacant | 3.376952 | 29.281394 | 2828.139448 | 1.370836e-19 |
| Land Value | -0.000001 | 0.999999 | -0.000121 | 2.795701e-06 |
| Multiple Parcels Involved in Sale | -0.557364 | 0.572717 | -42.728345 | 6.789425e-04 |
| Acreage | 0.104383 | 1.110026 | 11.002588 | 1.030117e-02 |
| Land Use | -0.072353 | 0.930203 | -6.979728 | 1.123900e-02 |
| Foundation Type | -0.030339 | 0.970117 | -2.988307 | 5.320118e-02 |
| Grade | 0.023905 | 1.024193 | 2.419340 | 2.162737e-01 |
| Neighborhood | -0.000011 | 0.999989 | -0.001072 | 3.298339e-01 |
| Exterior Wall | -0.003705 | 0.996302 | -0.369778 | 7.593821e-01 |
| Half Bath | 0.012155 | 1.012229 | 1.222933 | 7.814607e-01 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Decision Tree Classifier | | Random Forest | | Gradient Boost | |
| Feature Names | Importance | Feature Names | Importance | Feature Names | Importance |
| Building Value | 0.377472 | Building Value | 0.279993 | Building Value | 0.311901 |
| Sold As Vacant | 0.353467 | Sold As Vacant | 0.260598 | Land Value | 0.235552 |
| Land Value | 0.241714 | Land Value | 0.179600 | Sold As Vacant | 0.138341 |
| Exterior Wall | 0.027347 | Grade | 0.090420 | Acreage | 0.111451 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | True Negatives | False Positives | False Negatives | True Positives |
| Logistic Regression | 2678 | 35 | 855 | 17 |
| Decision Tree | 2640 | 73 | 807 | 65 |
| Random Forest | 2704 | 9 | 847 | 25 |
| Gradient Boost | 2652 | 61 | 810 | 62 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Logistic Regression | Decision Tree | Random Forest | Gradient Boost |
| Accuracy | 75.17% | 75.45% | 76.12% | 75.7% |
| Recall | 1.95% | 7.45% | 2.87% | 7.11% |
| Precision | 32.69% | 47.1% | 73.53% | 50.41% |
| ROC\_AUC | 50.33% | 52.38% | 51.27% | 52.43% |
| MSE | 24.83% | 24.55% | 23.88% | 24.3% |