A red and white logo

Description automatically generated

**College of Professional Studies, Northeastern University**  
**ALY6020**

**Predictive Analytics**  
**Professor Ajit Appari**  
**October 17th, 2024**

**INTRODUCTION:**

This project aims to analyze a marketing campaign dataset to derive insights that will guide strategic decision-making for optimizing campaign performance. The analysis is divided into five parts:

1. **Data Cleansing**: The dataset will undergo extensive cleaning, including renaming columns, handling missing values, and removing irrelevant data. This ensures accuracy and consistency, setting the foundation for effective analysis.
2. **Decision Tree Classifier**: A Decision Tree model will be built to classify campaign performance. This involves generating performance metrics, feature importance tables, and plotting a ROC curve to visualize model accuracy. The model's insights will help identify key factors driving campaign success.
3. **Random Forest Classifier**: The project will also implement a Random Forest Classifier, which is a more robust model leveraging multiple decision trees. Using the same metrics, the model’s performance will be compared against the Decision Tree.
4. **Gradient Boost Classifier**: Gradient Boosting, known for its accuracy, will be used to refine predictions. Its performance will be evaluated similarly to previous models.
5. **Model Comparison and Recommendation**: The final part consolidates results from all models, comparing metrics, feature importance, and ROC curves to determine the most suitable model for improving campaign strategy.

**DATA ANALYSIS:**

**Part 1: Data Cleansing Process**

The data cleansing process is critical for ensuring high-quality data to support accurate model predictions. The following steps were taken:

**Missing Values:** Identified and handled missing values across columns using median imputation for numerical columns such as 'Creative Width' and 'Creative Height.' Mean imputation was used for budget-related columns like 'Approved Budget.'

**Renaming Columns:** Columns were renamed for clarity, such as changing 'no\_of\_days' to 'Days Campaign Running.' This made the dataset more readable and easier to interpret for the analysis.

**Dropping Irrelevant Columns:** To streamline the model-building process, columns with little impact on the target variable or redundant information (e.g., 'Position in Content,' 'Unique Reach') were dropped.

**Duplicate Records:** Checked for duplicate records to ensure the dataset was free of duplicates that could bias the model.

**Outliers:** Identified and addressed any extreme outliers to prevent distortion of the model's predictions. We used IQR-based filtering for numeric variables.

**Summary stats for the data:**

A screenshot of a computer

Description automatically generated\

### PART 2: Decision Tree Classifier

### The **Decision Tree Classifier** model achieved an accuracy of 70%, with a precision of 77% for classifying undervalued properties (class 1) and 34% for overvalued properties (class 0). However, the recall for class 0 was lower, indicating that the model struggles to identify overvalued properties accurately. The F1-score for undervalued properties was 0.81, reflecting better overall performance for class 1. Feature importance analysis revealed that 'Building Value,' 'Finished Area,' and 'Acreage' were the most significant factors influencing the classification, helping to identify the key drivers of property valuation.

**PLOT 1: ROC CURVE FOR DECISION TREE:**

A line graph with a blue line

Description automatically generated

**PART 3: RANDOM FOREST MODEL:**

The **Random Forest Classifier** achieved an accuracy of 75%, with a precision of 75% for identifying undervalued properties (class 1). However, it struggled to identify overvalued properties (class 0) with 0% precision and recall. The model performed well in classifying undervalued properties with a recall of 100% and an F1-score of 0.86. Feature importance analysis showed that 'Parcel ID,' 'Legal Reference,' and 'Sold As Vacant' were the most significant factors influencing the model's predictions. Despite the high performance for undervalued properties, the model's inability to classify overvalued properties effectively is a limitation.

PLOT 2: ROC CURVE FOR RANDOM FOREST MODEL:

A graph of a line

Description automatically generated

**PART 4: GRADIENT BOOSTING MODEL:**

The **Gradient Boosting Classifier** achieved an accuracy of 76% on a small test set, with a precision of 76% and a recall of 99% for undervalued properties (class 1). However, it struggled significantly with overvalued properties (class 0), with a precision of 33% and a recall of just 1%, resulting in a low F1-score of 0.02 for class 0. The most important features identified by the model were 'Finished Area', 'Neighborhood,' and 'Building Value.' These variables played a significant role in predicting property valuation, though the model's inability to classify overvalued properties remains a limitation.

PLOT 3: GRADIENT BOOSTING MODEL:

A graph of a curve

Description automatically generated

PART 5:

In **Part 5**, the performance of the three models—Decision Tree, Random Forest, and Gradient Boosting—was consolidated for comparison. The table below summarizes the key performance metrics. The Random Forest model achieved the highest accuracy (74.87%) and perfect recall for classifying undervalued properties, while Gradient Boosting provided a balance between precision (74.91%) and recall (98.97%) with a strong F1-score (0.85). The Decision Tree model had the lowest accuracy (69.73%) but still performed reasonably well in terms of precision and recall.

The **Consolidated Feature Importance Table** highlights key factors such as 'Finished Area', 'Building Value', and 'Neighborhood' as influential across models, with 'Finished Area' showing the highest importance in Gradient Boosting. 'Acreage' and 'Building Value' were also consistently important across models, reflecting their significance in property valuation.

Based on the comparison, **Gradient Boosting** offers the best overall performance, balancing both precision and recall and is recommended for use in identifying undervalued properties.

PLOT 4: COMPARISON FOR MODELS:

A graph of a tree

Description automatically generated

CONCLUSION:

In conclusion, the analysis leveraged three machine learning models—Decision Tree, Random Forest, and Gradient Boosting—to classify properties in the Nashville area as over- or undervalued based on various features. While all models performed reasonably well, the Gradient Boosting model provided the most balanced performance, with high precision (74.91%) and recall (98.97%) for undervalued properties. It identified key features such as 'Finished Area', 'Building Value', and 'Neighborhood' as the most important factors influencing property valuation.

The Random Forest model achieved the highest recall but struggled with precision for overvalued properties, making Gradient Boosting the more consistent performer. The Decision Tree, while effective, had lower accuracy overall compared to the other models.

Given its superior balance between accuracy and feature importance, Gradient Boosting is the recommended model to guide the real estate company's investment strategy, helping them to identify the best value deals in the Nashville area.

REFERENCES:

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Wang, Y., & Li, J. (2016). Machine learning in real estate valuation: A review and further research directions. Journal of Real Estate Literature, 24(2), 271-298. https://doi.org/10.5555/jrel.24.2.271