FIN 550: Final Project
EXECUTIVE SUMMARY

Your Team Name (be creative): A+ team
■ Group Submission. Group member names: Chu-Chun Ku, Yi-Ting Lee, Hsiang Lee

Case Overview

The challenge we faced was that Cook County, Illinois, previously used outdated SPSS technology to assess property values, which lacked transparency and accuracy. Our goal is to transition to using R for property value predictions and to develop an effective predictive model that minimizes Mean Squared Error (MSE). This process is designed to be open and transparent, fostering public trust. Using historical property data, we aim to build the model to predict residential property values in Cook County and estimate market values for 10,000 properties, ensuring fair and accurate assessments.

Methodology

To predict property values for Cook County, the following data-driven approach was employed:

1. **Data Preparation**
   - **Dataset Loading:** Imported the historical property dataset (historic_property_data.csv) and its corresponding codebook (codebook.csv) to understand the variables.
   - **Variable Selection:** Identified relevant columns present in both the dataset and the codebook, focusing on predictors (var_is_predictor == TRUE) and the target variable (sale_price, renamed to assessed_value).
   - **Dataset Cleaning:** Created a clean dataset (m_data) containing only the relevant predictors and the target variable. Before model training ,missing values were handled by removing rows to ensure data integrity.
2. **Exploratory Data Analysis (EDA)**

- **Summary Statistics:** Calculated the mean and standard deviation for numeric predictors to understand their distributions.
- **Correlation Analysis:** Computed a correlation matrix to assess the relationships between numeric predictors and the target variable.

3. **Model Selection and Training:** The random forest algorithm was chosen for its ability to capture complex, non-linear relationships. Correlation analysis showed weak linear relationships for some variables, making random forest a better fit than linear models like Lasso regression.
    - **Initial Model:** Built an initial random forest model (mtry = 6, ntree = 100) to estimate variable importance.
    - **Feature Selection:** Based on the mean decrease in accuracy, selected the top 20 most important features and updated the dataset.
    - **Parameter Optimization:** Used 5-fold cross-validation to optimize the mtry parameter, ensuring a balance between bias and variance.
    - **Final Model:** Trained a refined random forest model (ntree = 1000) with optimized parameters to enhance predictive performance.
4. **Prediction and Evaluation**
    - **Prediction Dataset Cleaning:** Prepared the new dataset (predict_property_data.csv) by selecting the most important features and imputing missing values (median for numeric variables, mode for categorical variables).
    - **Prediction Generation:** Applied the trained random forest model to predict property values. Combined predictions with property IDs and exported the results in a structured format.
5. **Visualization and Analysis**

- **Distribution Assessment:** Visualized the distribution of predicted values using histograms and calculated summary statistics to evaluate their consistency.

Conclusion

1. **Data File Description:** The generated data file, predicted_assessed_values.csv, contains two columns:

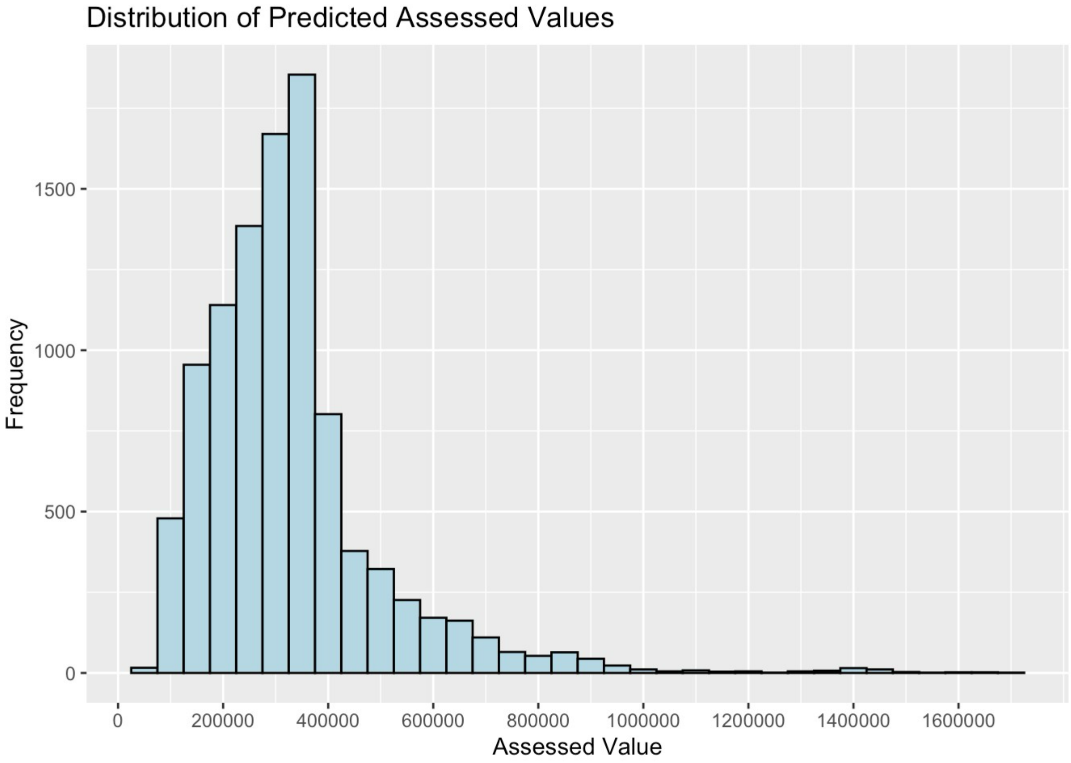
- **pid:** A unique identifier for each property, consistent with the input data.
- **assessed_value:** The property assessment value predicted using the random forest model, rounded to two decimal places.

2. **Summary Statistics of Predicted Property Assessment Values:** The distribution of the predicted property assessment values is summarized as follows:

- **Minimum (Min):** $47,914
- **First Quartile (1st Qu.):** $221,049
- **Median:** $308,657
- **Mean:** $328,825
- **Third Quartile (3rd Qu.):** $375,068
- **Maximum (Max):** $1,679,583
- **Standard Deviation:** $173,193.2

The distribution of the assessment values exhibits right skewness (positive skewness), with the majority of property values ranging between $200,000 and $400,000. A small number of high-value properties have assessment values exceeding $1,000,000.

Appendix


Distribution of Predicted Assessed Values
Frequency
1500
1000
500
0
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
Assessed Value