FIN 550: Big Data Project EXECUTIVE SUMMARY

Team Members:

1. **Doraiswamy, Pavithra**
2. **Iyer, Balakrishnan Mohan**
3. **Jillepalli, Mani Chandana**
4. **Mahmoudi Meymand**
5. **Nadeem, Nimra**

Case Overview:

In this specific final term project, we tackled the challenge of enhancing real estate value predictions in Cook County, Illinois. Our primary goal was to refine the property valuation process used by the Cook County Assessor's Office (CCAO), which has historically faced issues with transparency and efficiency. Utilizing the "historic_property_data.csv" as a training ground for the model development, our team built predictive models to accurately relate property characteristics to their sale prices. The project's complexity was heightened by the diverse range of properties in Cook County, which includes the bustling urban environment of Chicago and over 130 suburbs. Our methodologies were tested and validated against a new dataset, "predict_property_data.csv," ensuring our models' robustness and applicability in real-world scenarios. By leveraging advanced analytics and machine learning techniques in R, we aim to uncover patterns and relationships in real estate data, providing a solid foundation for accurate value predictions. This streamlined approach ensures transparency, efficiency, and accuracy of the Cook County property appraisal process.

1- Methodology:

The methodology adopted by our team was comprehensive, incorporating several critical stages from initial data handling to advanced predictive modeling. That involves data cleaning and preprocessing, variable selection, model formation, and cross-validation.

1) Data Cleaning and preparation:

1. Data Cleaning and Preparation:

    - **Loading Libraries:** Essential R packages such a*s 'tidyverse'* for data manipulation and ***'glmnet'*** for advanced modeling are loaded to facilitate the subsequent tasks.

    - **Data Import and Exploration:** We begin by importing the "historic_property_data.csv" and use str() and head() functions for an initial examination to understand the dataset's structure and main characteristics.

    - **Feature Selection:** Non-predictive columns are removed, and the remaining columns are renamed to enhance clarity and focus on relevant predictors for property sale prices.

    - **Handling Missing Values:** A function to count nulls in each column is implemented. Columns with more than 10% missing values are discarded, and other missing values are appropriately imputed or excluded to maintain data integrity.

- **Handling Unique Values:** We identify and remove columns with excessively high or low unique values to prevent model overfitting and enhance generalization.

- **Conversion of Categorical Variables:** Categorical variables are transformed into factors, facilitating more accurate analysis and model training.

- **Z-Value Method:** To address the outliers, we employ the Z-value method, calculating Z-scores to identify and handle outliers in numerical data. This method standardizes the treatment of outliers by capping extreme values, which helps in normalizing the data distribution and improves the robustness of our predictive models.

- **Data Partitioning:** The dataset is split into training (60%) and test sets (40%) using the specified seed for reproducibility.

2) Variable Selection:

In our project, we utilized several variable selection methods to identify the most predictive features for our models aimed at estimating real estate values in Cook County. These methods were essential in enhancing model accuracy and interpretability, ultimately helping us to achieve more reliable predictions. The following approaches were employed during the variable selection phase:

1. Forward Selection, Backward Elimination, and Stepwise Selection:

Initially, we employed traditional variable selection techniques, including forward selection, backward elimination, and stepwise selection:

- **Forward Selection** : starting with no variables in the model, adding one variable at a time that provides the most significant improvement to the model fit, and continuing this process until no further improvement is observed.

- **Backward Elimination** : begins with all potential predictors in the model, removing the least significant variable that offers the least contribution to the model fit, and repeating this process until only significant variables remain.

- **Stepwise Selection:** combines the principles of forward selection and backward elimination, allowing for variables to be added or removed in a stepwise manner based on their statistical significance and contribution to model performance.

While these methods are widely used for their simplicity and effectiveness in many scenarios, they led to relatively high Mean Squared Errors (MSE) in our initial tests. This indicated that while traditional methods could identify relevant predictors, they might not efficiently handle multicollinearity or interact effectively with the complex data structure present in our dataset.

2. Lasso Regression

Given the limitations observed with the initial methods, we advanced our approach by implementing Lasso Regression for variable selection. Lasso (Least Absolute Shrinkage and Selection Operator) Regression is particularly effective in reducing overfitting in models that

suffer from a high degree of multicollinearity or when the number of predictors is particularly high compared to the number of observations:

- **Penalty Application:** Lasso regression improves upon standard linear regression by introducing a regularization term (λ, lambda) with the L1 norm, which imposes a penalty on the absolute size of the regression coefficients. By doing this, Lasso effectively drives the coefficients of less important variables to zero, thus performing variable selection during the model fitting process.

- **Lambda Optimization:** We explored a sequence of lambda values to find the optimal balance between model complexity and performance. This involved examining the dimensions of the coefficient matrix across various lambda values, identifying the point where the increase in lambda leads to the most substantial reduction in model error.

- **Outcome:** The Lasso approach resulted in a lower MSE compared to previous methods, indicating better prediction accuracy and model interpretability. By reducing the number of features included in the final model only to those with significant predictive power, Lasso helped in simplifying the model while enhancing its predictive performance.

3) Modeling:

1- Cross-validated Lasso Regression Modeling:

In our modeling approach, we utilized a 10-fold cross-validation technique to ensure robust evaluation of the Lasso regression model. This method involves dividing the original dataset into

10 equal parts, using each in turn for validation while training on the remaining nine. Such a strategy enhances the model's generalizability and prevents overfitting to specific data subsets, thereby ensuring that our predictions remain consistent across different samples of data.
To optimize the model's performance, we focused on selecting the best lambda (λ) value, which determines the strength of the penalty applied to the coefficients in Lasso regression. By testing different lambda values across each fold of our cross-validation process, we were able to identify the value that minimizes the mean squared error (MSE), effectively balancing bias and variance in our model. This careful tuning helps in honing the model to achieve optimal predictive accuracy on unseen data.

Furthermore, we set the alpha parameter to 1, aligning our model strictly with L1 regularization typical of Lasso regression. This configuration is instrumental in enhancing the sparsity of the model — selectively shrinking coefficients of less critical variables to zero. Consequently, the model retains only those predictors that are most impactful, simplifying model interpretation and focusing on the most influential factors driving property valuations in Cook County. This strategic implementation of Lasso regression not only refines the model's predictive capabilities but also provides clear insights into the variables that significantly affect real estate prices.

2-Prediction and MSE Calculation

Upon finalizing our model, we proceeded to utilize it for making predictions on the test set, which was set aside during the model training phase. This step is crucial as it provides an unbiased evaluation of the model's effectiveness in predicting new, unseen data. The primary metric used to assess the accuracy of our predictions was the Mean Squared Error (MSE). The MSE quantifies the average squared difference between the estimated values and the actual value, offering a comprehensive measure of prediction accuracy.

The lowest MSE value obtained through our model testing was 15903578320, indicating the model's robustness in capturing the underlying data patterns without fitting to random noise. This phase of our analysis not only confirms the model’s precision but also highlights its applicability in real-world scenarios, providing valuable insights for stakeholders interested in accurate property valuation, such as real estate analysts and policymakers.

.

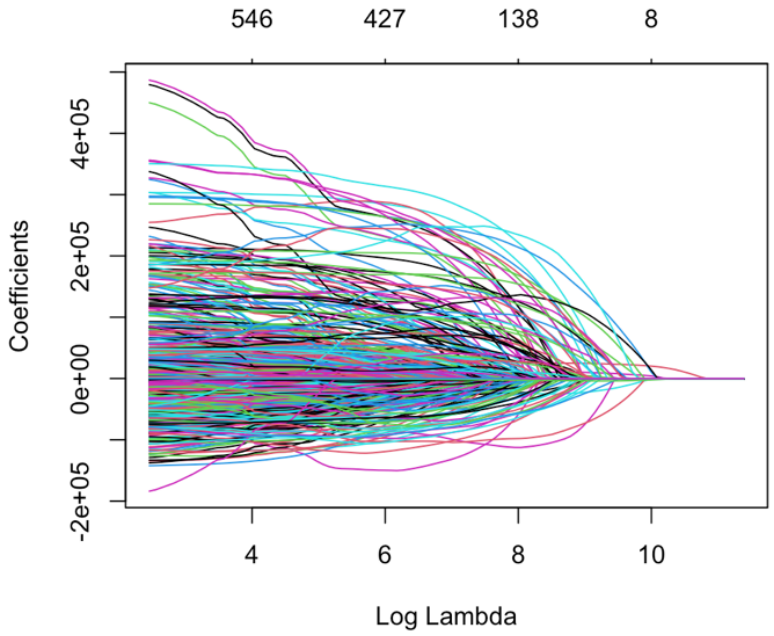
3- Conclusion:

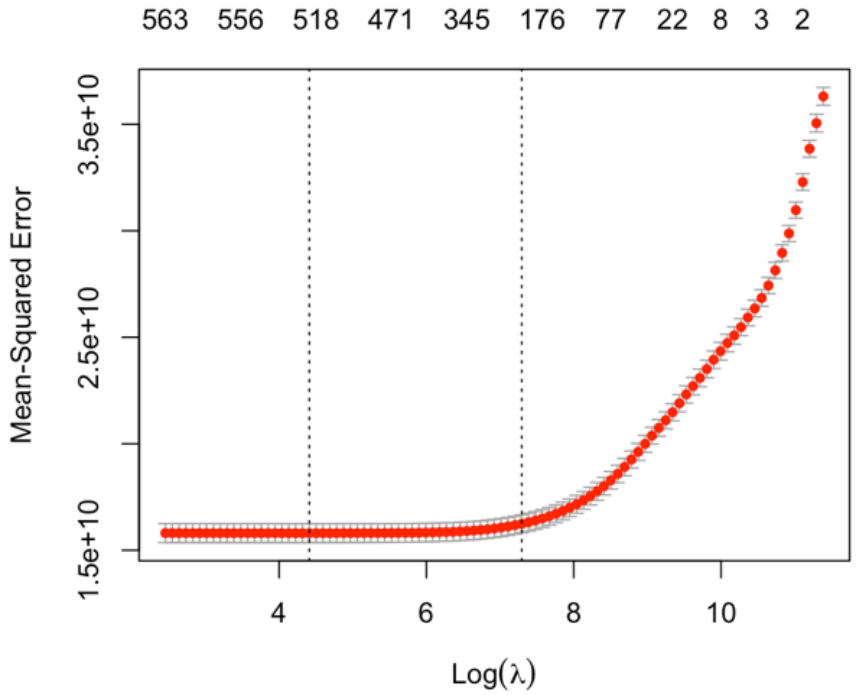
Our comprehensive modeling approach, centered on cross-validated Lasso regression, has proven highly effective in predicting property values within Cook County. By implementing rigorous variable selection and model optimization techniques, we achieved a model that not only offers high predictive accuracy but also enhances interpretability, a crucial factor in real estate valuation. The successful reduction in the Mean Squared Error (MSE) to15903578320 highlights the model's practical relevance and reliability. These results are particularly promising as they suggest that our model can significantly aid the Cook County Assessor's Office in refining their property assessment processes.

Moreover, the insights derived from our analysis provide a solid foundation for future enhancements. The model's ability to identify key value drivers through Lasso regression allows for targeted policy and decision-making processes, ensuring fair and equitable property tax assessments across the county.

In conclusion, the methodologies and findings from this project do not only advance the technical aspects of property valuation but also contribute to more transparent, efficient, and fair property tax assessments in Cook County. As we move forward, these advancements lay the groundwork for further research and refinement, potentially influencing broader applications within the field of real estate analytics.

Appendix:


546
427
138
8
Coefficients
4e+05
2e+05
0e+00
-2e+05
4
6
8
10
Log Lambda


563
556
518
471
345
176
77
22
8
3
2
Mean-Squared Error
3.5e+10
2.5e+10
1.5e+10
4
6
8
10
Log(λ)