FIN 550: Big Data Project
EXECUTIVE SUMMARY

Your Group Number: __3___

Group member names: Joel Esparza, Gisselle Pena, Akshara Raghav, Aditi Tiwari, Aman Urfi , Alexandra Watson, Ivy Zhao, Devanand Lankapalli

FIN 550: Big Data Project

Executive Summary

Team 3: Joel Esparza, Akshara Raghav, Aditi Tiwari, Alexandra Nicole Watson, Ivy Zhao, Devanand Prasad Lankapalli, Gisselle Pena, Aman Ahmad Urfi

Case Overview

In this data science project, we focus on optimizing real estate value predictions in Cook County, the second most populous county in the United States, covering Chicago and over 130 neighboring suburbs. Fair market value assessments are critical to property tax calculations, but the Cook County Assessor's Office (CCAO) has historically been vague and inefficient. Our goal is to enhance this valuation process by leveraging predictive models. The dataset "historic_property_data.csv" serves as our training ground, allowing us to build models that relate property characteristics to their actual sales prices. The complexity of this task is compounded by the diversity of properties scattered across numerous cities. Our prediction model will be tested against the "predict_property_data.csv" dataset to accurately estimate home value. 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.

Methodology

Our methodology involves integrated steps to data cleaning, data preparation, variable selection, then fitting predictive algorithms. We will introduce our steps and methods in detail subsequently.

1) Data Cleaning and preparation:

1.1 Loading Libraries: R code is written by loading necessary libraries, including `tidyverse`, `glmnet`, and `  `, for subsequent data manipulation and modeling tasks.

1.2. Data Import and Exploration: The dataset is loaded from a CSV file, and its structure is examined using `str()` and `head()` functions to understand the data's characteristics. We have used the 'historic_property_values.csv' for testing predictive models. The predictions are made on a separate file named 'predict_property_data.csv'.

1.3. Feature Selection: Columns that are not predictors are removed, and the remaining columns are renamed for clarity. This step involves excluding specific columns deemed irrelevant for predicting property sale prices.

1.4. Handling Missing Values: A function is created to count the number of null values in each column. Columns exceeding a specified null threshold (20%) are then removed, and remaining null values are dropped.

1.5. Handling Unique Values: A function is implemented to identify columns with too few or too many unique values. Columns meeting these criteria are dropped to improve model generalization.

1.6. Conversion of Categorical Variables: Certain categorical variables are converted to factors using a 'for loop' over the specified columns.

1.7. Winsorization: A winsorization function is applied to numeric variables, capping extreme values to enhance model robustness. After winsorizing the data, we visualized the distribution of sale price. Ggplot is then used to display the distribution of sale prices.

1.8. Data Partitioning: The dataset is split into training (70%) and test sets (30%) using the specified seed for reproducibility.

2) Variable Selection:

2.1. Lasso Regression: we performed forward selection, backward selection, and step wise selection. However, they provided relatively high MSE. Finally, we performed a lasso regression model to the training data for variable selection by introducing a penalty term that encourages some coefficients to be exactly zero. It is used to explore the sequence of lambda values and examine the dimensions of the resulting coefficient matrix. Lasso regression gives us a relatively lower MSE. Lasso regression provides better prediction accuracy and model interpretability.

3) Modeling:

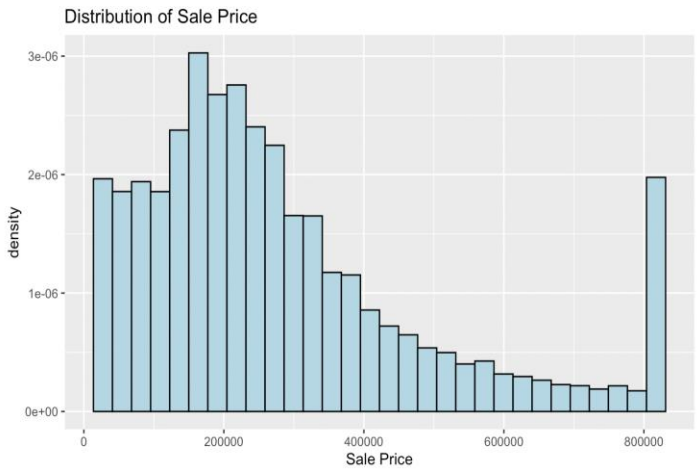
3.1. Cross-validated Lasso Regression Modeling: we performed cross-validation to find the appropriate lambda value. K-fold is set to 5 and is applied to the training set. K fold = 5 means that the dataset is split into 5 parts, and the model is trained and evaluated 5 times, each time using a different fold as the validation set and the remaining folds as the training set. Alpha is set to 1 to perform L1 Lasso Regularization. This model is used to extract the optimal lambda value and its corresponding coefficients. Predictions are then made using the best lambda value coefficient.

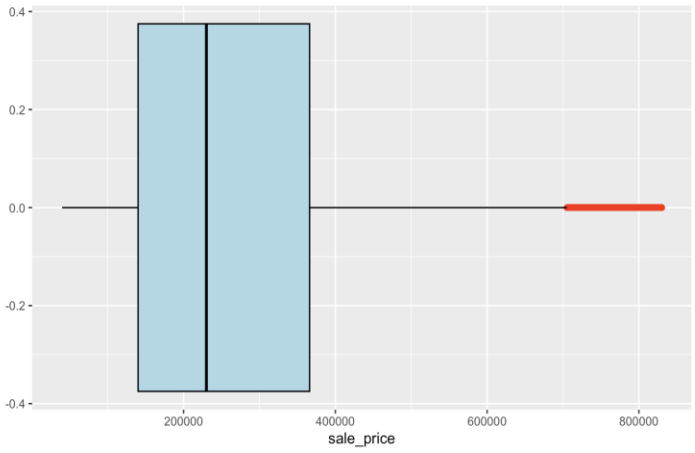
3.2. Prediction and MSE Calculation: The model is used to predict sale prices on the test set, and the Mean Squared Error (MSE) is calculated to evaluate prediction accuracy. The lowest MSE value obtained is 5,002,819,932. The range of absolute differences is displayed in the appendix. The mean absolute difference between actual and predicted prices was $53,642. In this context, this value is relatively low as compared to the mean housing price which is $248,444. 31.5% of the predicted prices were within $25,000 of the actual price.

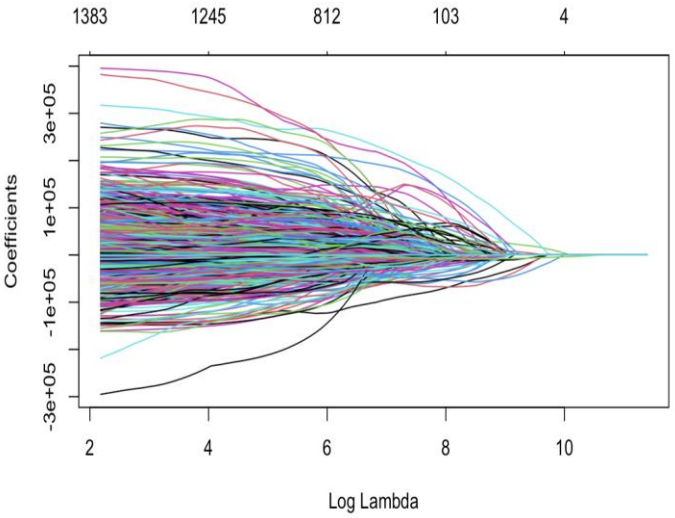
Conclusion

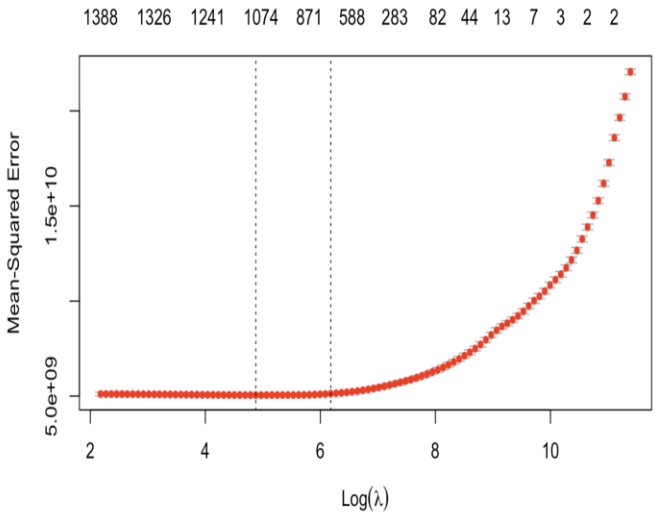
Cross-validated lasso regression performed the best with the testing set and was hence used to make predictions of housing prices for 10,000 households in the 'predict_property_data.csv' file. This model produced a low mean absolute difference, and a large range of predicted prices were predicted with negligible error. The difference between mean and median ($284,639 & $248,444) is relatively low. This difference means that the predictions are slightly skewed to the right implying that some households predicted to have higher prices are pulling the mean to the right. The first and third quartiles are $146,328 & $377,360 respectively representing the spread of the predictions. The minimum and maximum are extreme predictions and could be treated as outliers.

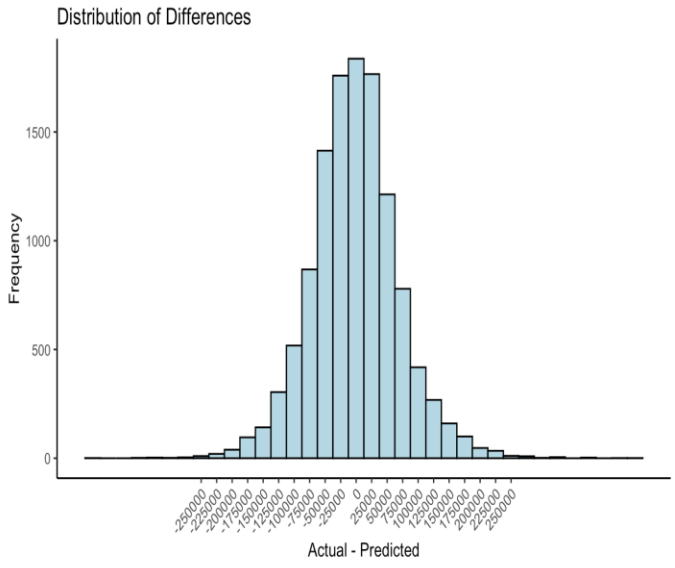
Appendix

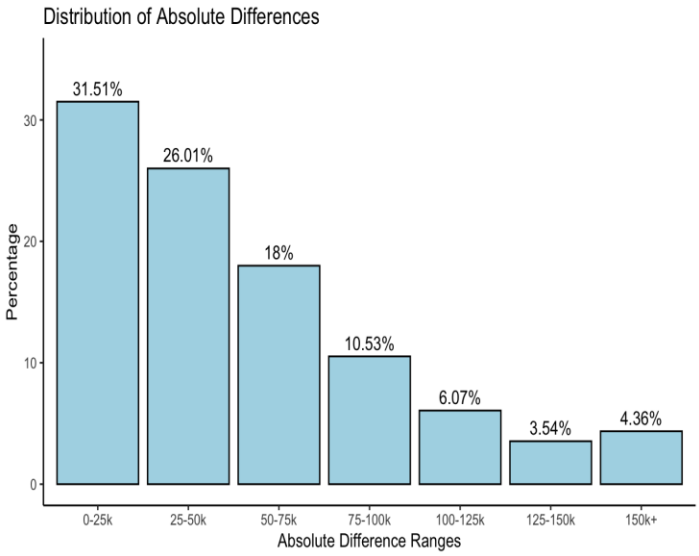

Distribution of Sale Price
density
Sale Price
3e-06
2e-06
1e-06
0e+00
0
200000
400000
600000
800000

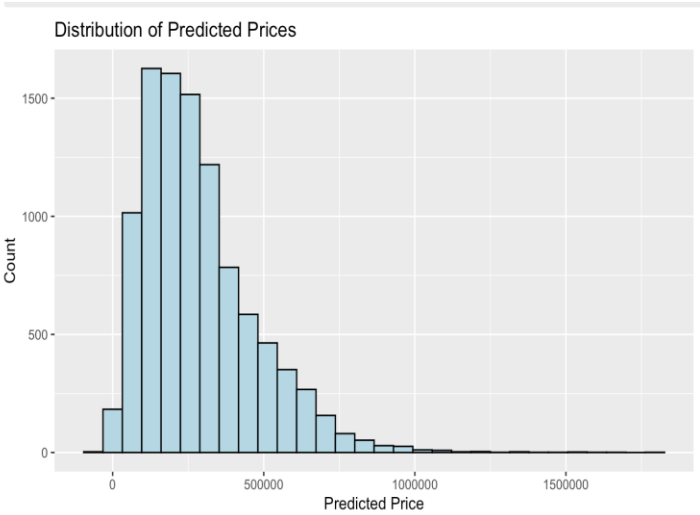

sale_price
0.4
0.2
0.0
-0.2
-0.4
200000
400000
600000
800000

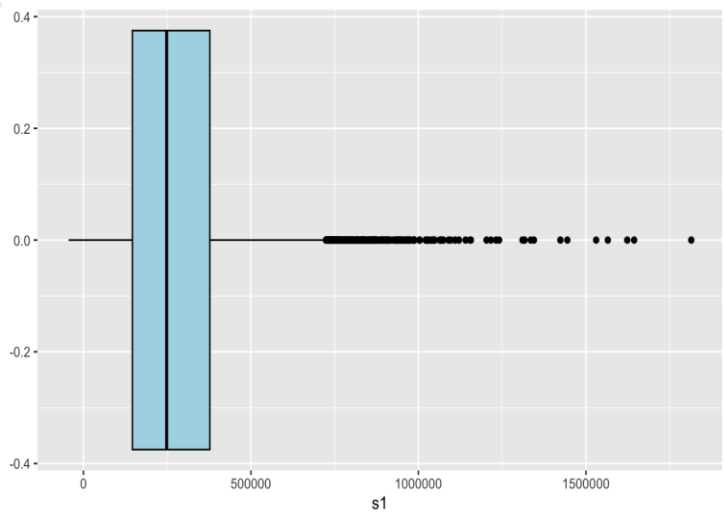

1383
1245
812
103
4
Coefficients
Log Lambda
3e+05
1e+05
-1e+05
-3e+05
2
4
6
8
10


1388 1326 1241 1074 871 588 283 82 44 13 7 3 2 2
Mean-Squared Error
Log(λ)
1.5e+10
5.0e+09
2
4
6
8
10


Distribution of Differences
Frequency
Actual - Predicted
1500
1000
500
0


Distribution of Absolute Differences
Percentage
Absolute Difference Ranges
31.51%
26.01%
18%
10.53%
6.07%
3.54%
4.36%
30
20
10
0
0-25k
25-50k
50-75k
75-100k
100-125k
125-150k
150k+


Distribution of Predicted Prices
Count
1500
1000
500
0
0
500000
1000000
1500000
Predicted Price


0.4
0.2
0.0
-0.2
-0.4
0
500000
1000000
1500000
s1