FIN 550: Big Data Project
EXECUTIVE SUMMARY

Your Group Number:_______________________5____________________________________

Group member names: __Courteney Chan, Xinyuan Chen, Ishika Gupta, Kaitlyn Ho, Chih-Yu Hsu, Rutuja Lohakare__
__________________________Raj Mehta, Yifeng Zhang____

Case Overview

Cook County is the second most populous county in the United States, with an estimated population of 5.1 million residents, and 2.27 million housing units as of the July 2022 US Census Bureau. The Cook County Assessor Office (CCAO), tasked with estimating market values, faced challenges during Joseph Berrios's 2010-2018 term. Around 70% of the market value estimates for commercial and industrial buildings remained unchanged for over two assessment periods (spanning six years). This anomaly affected residential properties, resulting in overestimated home values, leading to higher property taxes and other financial stressors for homeowners in lower socioeconomic brackets.

Our objective is to address this issue by developing an algorithm capable of predicting the fair market price for homes in Cook County. The CCAO's existing computer programs rely on square footage and bedroom count to estimate residential property value. Using a similar methodology, our algorithms were formed by identifying patterns among the houses to enhance the accuracy of the predictions.

Methodology

The data preprocessing used three datasets: the "historic_property_data" (encompassing the 50,000 most recent property sales), "predict_property_data" (for algorithm development), and "codebook" (defining dataset columns). Initial steps included removing non-predictive attributes based on the codebook and eliminating duplicate rows.

We then accounted for missing data values by imputing the neighborhood codes (meta_nbhd) and town codes (meta_town_code) using the Last Observation Carried Forward method, useful for time series datasets. Numeric missing values were imputed by the mean of their respective columns. Negative sale prices were adjusted to 0, and values were winsorized to address outliers. Any resulting rows with null values were removed.

Crucial predictors were identified, focusing on location, dimensions, and flood risk. Since the flood risk and direction were converted to a numerical scale, there was no need to change it to a factor variable. A correlation matrix (Image 1) and Variance Inflation Factor (VIF) analyses were performed to assess multicollinearity, ensuring distinct variable effects across models due to the extensive set of predictors. These results resulted in removing the township codes due to high correlation. Lastly, we performed linear regression, random Forest, and lasso modeling to design our algorithm, utilizing cross-validation.

Conclusion

The assessed property values have a minimum value of 42012, a maximum of 2180952, with an average of 308566. The first quartile is 184985, the median is 255128, and the third quartile is 372765.

Image 2 illustrates that most of the properties have lower market values, suggesting a right-skewed distribution. The highest frequency of properties is in the first few bins, with a sharp decline as market value increases, which is typical in real estate markets where there are lower-valued properties than high-valued ones. The market values shown range from close to $0 up to over $2,000,000, with very few properties in the highest value range.

However, we note that the linear regression has an MSE of 43,220,479,448, the random forest model's MSE is 15,276,646,289, and the lambda model has an MSE of 43,217,434,819. Combining these models, our final model has an MSE of 28,847,341,624. Thus, future steps may involve determining other predictors, specifically ones that determine the construction quality of homes, to improve the algorithm's accuracy.

Appendix

1.

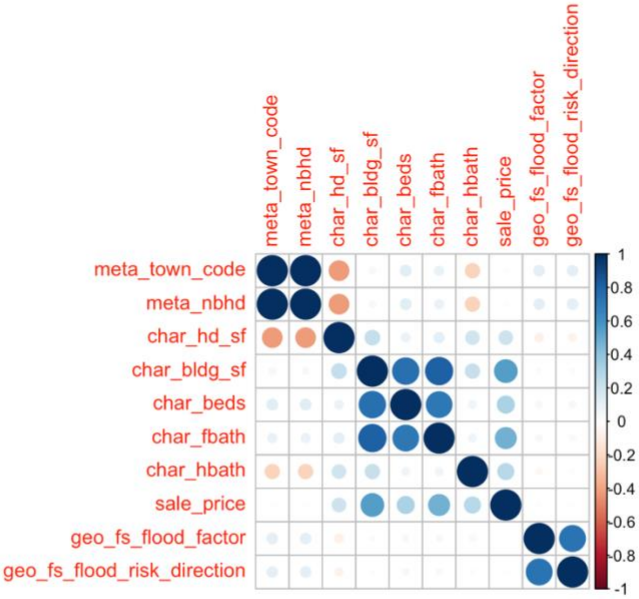

Image 1: Correlation Matrix of Numerical Variables

2.

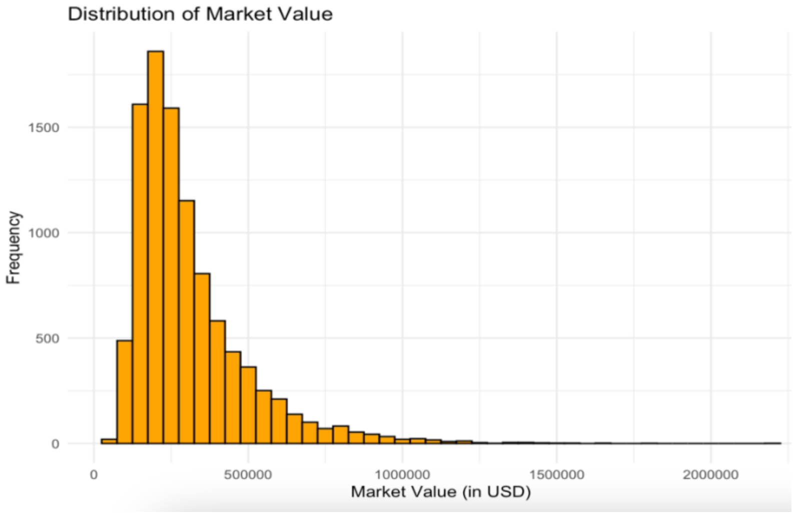

Image 2: Distribution of Market Values