FIN 550: Final Project

EXECUTIVE SUMMARY

Your Team Name (be creative):____________________FJ Alliances____

Select whether this is an individual or group submission. **No more than 3 members per group.** Beyond the fact that all group members may submit the same answers, each submission must be separate work.

□ Individual Submission

■ Group Submission. Group member names: _____Yunzhe Yu, _ Ives He, Hanbin Yan____________________________________

Case Overview

Our group attempted to build a model for predicting and estimating the market value of homes based on residential properties for Cook County, Illinois. Our main goal was to minimize the mean square error of the predictions from historical property sales data and provide valuable insights to the Cook County Assessor's Office.

Methodology

Before modeling, we first need to ensure the consistency and reliability of the data, so our first step is to process the provided data. Our group processed the data in 3 steps: the first step was to process the missing values in which we filled in the numerical variables with median values to minimize the possible effect of extreme values on the results. The second step is to process the outliers, where we use the Quartile Distance method to detect outliers and set them as missing values to avoid these outliers from contaminating the model. The third step is to ensure the completeness and reasonableness of the data, and we exclude attributes that have missing or non-positive values in the sales price.

Next, to enhance the model performance, we used Random Forest, which is very stable and capable of handling complex nonlinear relationships without extensive preprocessing. We calculated the importance of the features and selected the top 30 based on their scores to balance the complexity and accuracy of the model. To optimize the performance of the model, we performed grid search hyperparameter tuning by adjusting the number of features (mtry) and the number of trees (ntree) considered in each segmentation.

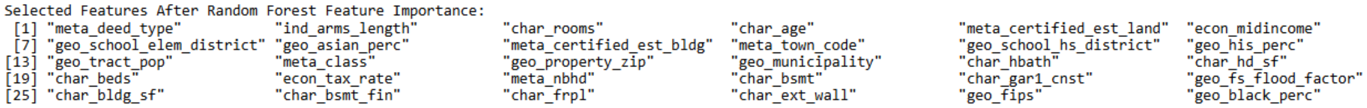
In training the model, we used 80% of the data for training and the remaining 20% for testing to verify the generalization ability of the model. Finally, we use the optimized model to make predictions on the new property dataset. And the results are exported in the form of CVS files

Conclusion

Describe your results, including summary statistics (e.g., min, max, mean, and quartiles) of the distribution of assessed property values. Describe your data file which reports the assessed property values you have generated.

The results of our model show significant progress in assessing the fair market value using the random forest model to minimize the mean square error. By analyzing the dataset, we achieved an optimal MSE of 0.1528612 , which means that our model is able to accurately find the relationship between property characteristics and its market value, providing a reliable basis for the assessment of housing plus resettlement, and we included the corresponding 10,000 records in the assessed_value.csv file.

Appendix



```
Selected Features After Random Forest Feature Importance:
 [1] "meta_deed_type"           "ind_arms_length"          "char_rooms"               "char_age"                 "meta_certified_est_land"  "econ_midincome"
 [7] "geo_school_elem_district" "geo_asian_perc"           "meta_certified_est_bldg"  "meta_town_code"           "geo_school_hs_district"   "geo_his_perc"
[13] "geo_tract_pop"            "meta_class"               "geo_property_zip"         "geo_municipality"         "char_hbath"               "char_hd_sf"
[19] "char_beds"                "econ_tax_rate"            "meta_nbhd"                "char_bsmt"                "char_gar1_cnst"           "geo_fs_flood_factor"
[25] "char_bldg_sf"             "char_bsmt_fin"            "char_frpl"                "char_ext_wall"            "geo_fips"                 "geo_black_perc"
```