FIN 550: Final Project

EXECUTIVE SUMMARY

Your Team Name (be creative): Team No Team – Seunggyun Shin, ss6

□ Individual Submission

Case Overview

This project aims to estimate fair market value of properties in Cook County which is the second most populous county in the United States. Accurate estimation in market price is crucial as it is related to budgeting property taxes, one of the largest portions of revenue for local government. Throughout this project, the data provided by Cook County Assessor's Office (CCAO) that contains information about 50,000 properties will be leveraged in predictive analysis for fair market value of properties.

Methodology

1. **EDA & Data Manipulation**

a) NA values

- Columns with NA values were detected for historic and predict data sets. For 10 columns *(Feature 1)* that have more than 1,000 NA values in the predict data set was eliminated as those columns will not be able to be used in prediction since there is no value provided.
- For columns that have less than 1,000 NA values, data loss was calculated in case dropping all the rows with NA values. Dropping rows with NA values, 2.58 percent of observations was eliminated. As the data loss was less than the pre-defined threshold 5 percent, simply dropping rows was the best and most effective method for following reasons.

• Filling NA with representative values such as mean or median without solid background knowledge can cause increase bias of the data and distort correlations between variables.

• Replacing NA with values based on predictive modeling was disregarded due to timely manner as the data has the large number of columns containing NA values.

b) Data type

- Data types of columns were investigated comparing those to descriptions provided by the code book. Multiple incorrect data types were observed, and redefined based on the code book.

• ex) char_bsmt (Basement), char_heat (Central heating), char_ext_wall (Wall material), and many other columns were in numeric type, which had to be changed to characters

c) Column modification

- There were some columns that had to be modified for a better analysis.

- The column “char_cnst_qlty” which describes a construction quality of the property has 3 categories (1 – Deluxe, 2 – Average, 3 – Poor). Commonly deluxe is a better assessment than poor that the level of categories was reordered.
- The column “char_gar1_size” which describes a size of garage has 8 categories from 1 to 8 depending on the number of cars can fit the garage. However, categories are in an ambiguous order. For example, “3” means that 2 cars can fit the garage while “7” means that 0 cars can fit the garage. So, all categories were transformed into numeric value depending on the number of cars fitting the garage.
- The column “char_type_resd” which describes the type of residence has 9 categories. Categories from “6” to “9” has small proportions in the data. Moreover, there is no significant difference in meaning with “5”. So, they were all combined to the category “5” to balance the proportion.

2. Response Distribution & Variable Selection

- Before modeling, the distribution of the response variable “sale_price” was observed *(Feature 2)*. The distribution is right skewed that log transformation on response variable was considered to normalize the distribution. However, by choosing modeling methods that do not require normality of response variable, the log transformation provided worse prediction results.
- The data has 23 categorical variables and some of those variables contain excessive number of categories *(Feature 3)*. These variables with more than 50 categories were dropped from the data as excessive number of categories will generate corresponding number of dummy variables, which will massively expand the dimension of the data during modeling process (2036 columns including dummies if columns are not dropped), and cause errors and inaccurate predictions if the categories do not match between training and testing data. High dimension can trigger dimension curse and unnecessarily increase the running time of models. Dropping these columns does not have negative impact on predictive modeling as excessive number of categories do not give significant information, and those can be explained by other columns as well.

  • ex) Town codes contain information of school districts.

3. Modeling

- The response variable “sale_price” is a continuous variable and the key purpose of this project is to minimize MSE in prediction.
- Multiple linear regression with stepwise variable selection was used as a baseline model. As the final data contains 45 prediction variables, there is high possibility of multicollinearity (There were variables causing multicollinearity depending on Variance Inflation Factor Analysis) so that modeling methods which are effective for data with multicollinearity was selected. Lasso regression and ridge regression were used due to their effectiveness in variable shrinkage, and Random Forest and Xgboost were utilized as well since they are comparatively free from multicollinearity and expected to have high performance on data with the large number of variables.
- The data was divided in train and test set (8:2) to validate performance of modeling methods.
- Among MLR models, backward elimination applying AIC had the best prediction result. However, the result was not considered reliable that it is hard to conclude that variables will have exact linear relationships with the response.

- Lasso and ridge regression utilizing cross validation might be more reliable than the MLR model, but focusing on MSE, they had worse results than MLR.
- Xgboost had significantly less running time than Random Forest and better prediction results than multiple linear, lasso, and ridge regression.
- Even without detailed parameter tuning, Random Forest had the best model performance. However, it took lots of time to run each model.
- As Random Forest had the lowest MSE value in predicting test set (Feature 4), the model was selected as a final prediction method. Tuning on hyperparameters (mtry, ntree) was conducted using 3-fold cross validation, and found out mtry = 10, ntree = 500 has the best performance in prediction.
- The final model was trained using the hyperparameters above with the whole data set rather than applying hyperparameter tuning using cross validation for the whole data set due to the concerns about overfitting.
- Before making the final predictions, NA values in the predict property data was detected. Since there were only single digit numbers of NA values in 16 columns, those values were replaced with the majority value of each column.

Conclusion

According to the variance importance from the final random forest model (Feature 5), variables "meta_certified_est_bldg" and "meta_certified_est_land" had outstanding influences in predicting sale prices of properties. This makes sense as those two variables describes estimated values of building and lands by assessors. Besides those two variables, median income, building square feet, white percentage, and the number of bathrooms had significant impact on the prediction. This proves that the model is reasonable as incomes, sizes of buildings, and the number of bathrooms might be crucial to estimate the prices of properties. However, the variable white percentage should be deeply investigated through further analysis. For instance, analysis on difference among races in the country can be a good comparison with the finding.

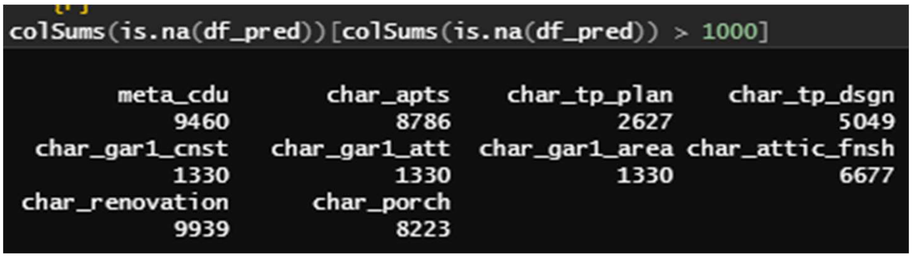
The distribution of final predictions (Feature 6) is right skewed, which is similar to the distribution of sale price in the trained data set (historic property). This confirms that the prediction well-reflected the information provided from the training data.

Summary Statistics

	Historic property data	**Final Prediction**
Minimum	100	6,173
1st Quantile	130,000	158,122
Median	221,000	247,365
Mean	297,124	324,965
3rd Quantile	355,000	377,120
Maximum	9,400,000	5,006,132

Appendix

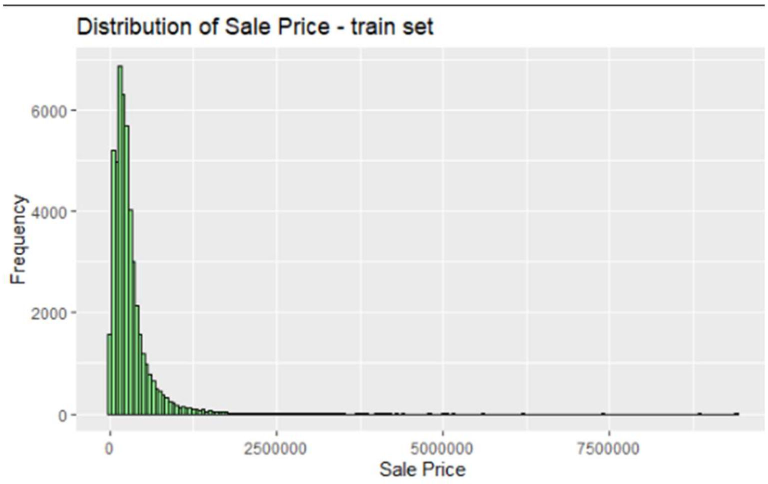
Feature 1. Columns with more than 1000 NA values

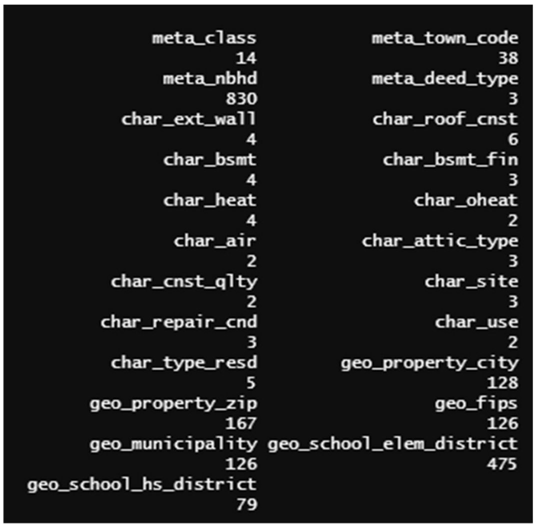


```
colSums(is.na(df_pred))[colSums(is.na(df_pred)) > 1000]

       meta_cdu        char_apts      char_tp_plan    char_tp_dsgn
           9460             8786              2627            5049
 char_gar1_cnst    char_gar1_att   char_gar1_area char_attic_fnsh
           1330             1330              1330            6677
char_renovation       char_porch
           9939             8223
```

Feature 2. Distribution of Sale Price in Training data / Feature 3. Distinct values in each category



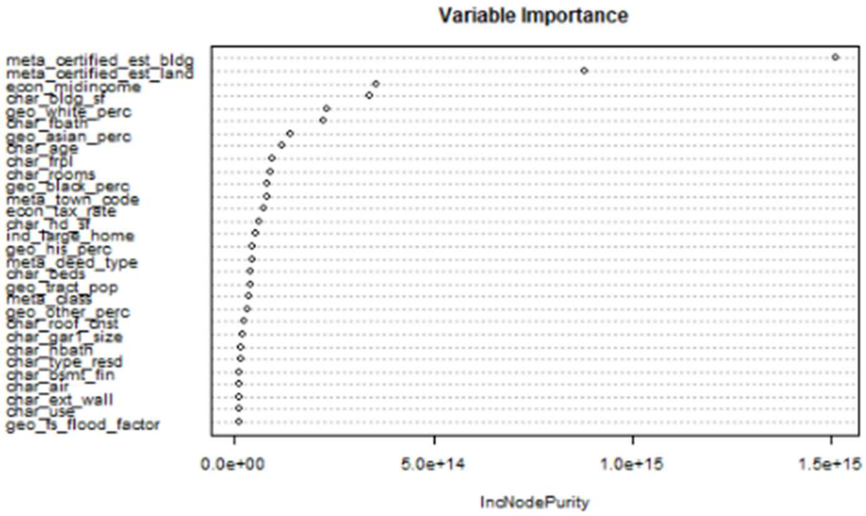


```
            meta_class          meta_town_code
                    14                      38
             meta_nbhd          meta_deed_type
                   830                       3
         char_ext_wall          char_roof_cnst
                     4                       6
             char_bsmt           char_bsmt_fin
                     4                       3
             char_heat              char_oheat
                     4                       2
              char_air         char_attic_type
                     2                       3
        char_cnst_qlty               char_site
                     2                       3
       char_repair_cnd                char_use
                     3                       2
        char_type_resd       geo_property_city
                     5                     128
      geo_property_zip                geo_fips
                   167                     126
      geo_municipality geo_school_elem_district
                   126                     475
geo_school_hs_district
                    79
```

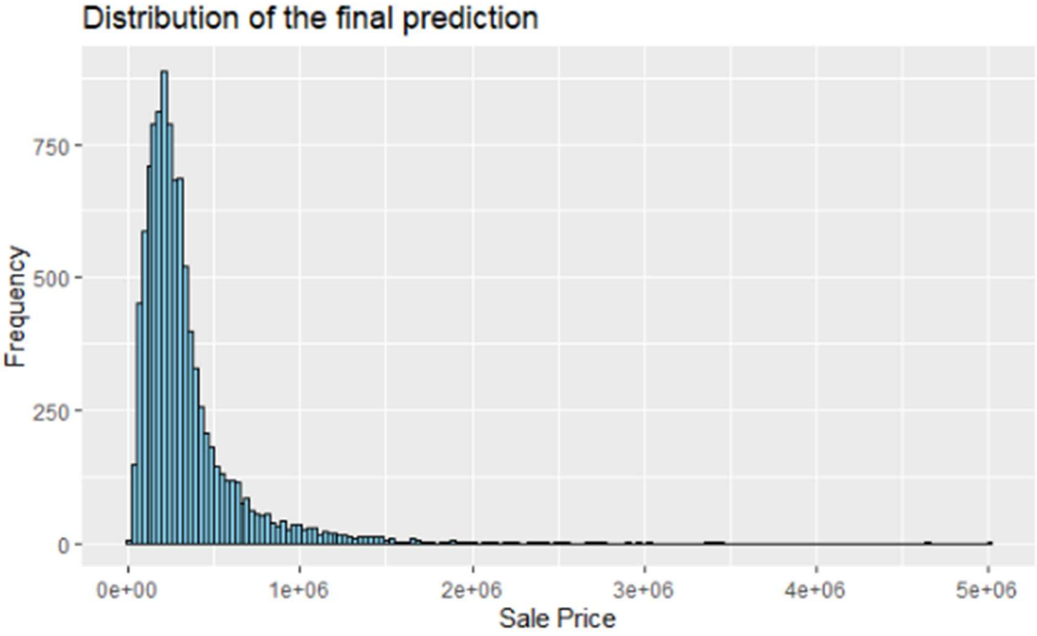
Feature 4. MSE values by modeling methods

Method	MSE
Lasso Regression	20,539,953,279
Ridge Regression	20,878,107,768
Linear Regression (Full model)	20,508,491,523
Linear Regression (Backward elimination)	20,494,297,897
Random Forest (mtry = 10, ntree = 500)	17,881,301,719
Random Forest (mtry = 15, ntree = 500)	18,277,462,537
Random Forest (mtry = 15, ntree = 100)	18,847,826,593
Xgboost (nrounds = 300)	19,178,502,942

Feature 5. Variance Importance in the final model



Feature 6. Distribution of final predictions


Distribution of the final prediction
Frequency
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Sale Price