FIN 510: Final Project
EXECUTIVE SUMMARY

Your Team Name (be creative): AppleLTK

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: Ting-An Kuo(takuo2), Yi Liang(yil19)

Case Overview

Concisely describe the problem and your objectives as a data scientist on the team.

Being the members of CCAO data science team, our objective is to build the machine learning models to predict and calculate the fair market value of the residential properties in Cook County. This is a challenging task. To maintain CCAO's professional reputation, we aimed to predict the market values as accurate as possible.

Methodology

Describe the data approach and methodology you use. Justify your choice of methodology. You should be precise but avoid jargon. This is a document that will potentially be shared with the data science team at CCAO, so make the language is clear professional and appropriate for them.

Dataset Usage

We want to predict accurate value of the properties, so we refer to the data of recently sold properties nearby. We used the following three files for modeling:

1. "**predict_property_data.csv**" contains data on 10,000 properties whose fair market values we're going to assess.
2. "**historic_property_data.csv**" contains data on 50,000 other properties that have recently sold, including their recent sales price (variable "sale_price").
3. "**codebook.csv**" describes the variables included in each property file.

Data Cleansing

Firstly, we loaded the packages: dplyr, rpart and glmnet and then read in the three files mentioned above. We created 3 dataframes named codebook, df and pred_df to store the data respectively. In order to build a model precisely, we wrote a function to remove the outliers. (Our outlier means that the range below the 5th percentage and above the 95th percentage.) Next, we used is.na() to assign 0 to empty logistic variables. After cleaning all the data, we randomly set the training and testing set for further model testing.

Variable Selection

To choose which variables we should maintain and which variables we should abandon, we took the following few steps. According to Lecture 10, we want to imitate the lecture example to predict prices of houses based on their specification information. We fit all the predictors in a linear regression model.

Model Selection

Since our outcome is the predicted fair market price of the residential property, that would be numeric variable not categorical variable. Therefore, we firstly gave up the following models: Classification Tree (Used with a categorical outcome variable), Regression Tree (Suffer from high variance), Logistic Regression (used to model the relationship between a binary response and a set of predictor variables) and Discriminant Analysis (A classification method: classify a record to the closest class). At first step, we fit all the specification variables into a linear regression. Use lm() and specify the "sale_price" as dependent variable and specify "char_beds", "char_hbath", "char_hd_sf", "geo_fs_flood_factor", "char_rooms", "char_frpl", "char_bldg_sf", "econ_tax_rate", "char_age", "char_fbath", "econ_midincome", "geo_fs_flood_risk_direction", "char_heat", "char_gar1_att", "char_bsmt", "char_attic_type", "char_tp_plan", "char_ext_wall", "char_type_resd", "char_roof_cnst", "char_oheat", "char_gar1_size", "char_gar1_area", "char_repair_cnd", "char_bsmt_fin", "char_air", "geo_floodplain", "geo_ohare_noise", "geo_withinmr100", "geo_withinmr101300" and "ind_garage" as independent variables. In this step, we used training data to train and build the linear regression model. Next, we used testing data to predict a set of predicted sales prices. Further, we calculated the mean sum of square is 10,360,797,510, assigned it to a variable named mse1. In order to build a more efficient and more accurate linear regression, we want to reduce useless and redundant predictors. There are couple ways to reduce, we chose stepwise regression instead of exhaustive search, because exhaustive search would spend lots of time to get the results. We decided to ran stepwise in "both" way. That is, calculated the AIC each time and add or drop the variables each time to obtain a set of predictors with the lowest AIC. By stepwise reduction, we got the mean sum of square is 10,366,058,281, assigned it to a variable named mse2.

Comparing mean sum of square, we can discover that stepwise regression slightly decreased the MSE by reducing some redundant variables. Shown in the following table:

```
##   c..LM....Stepwise.Regression.. c.mse1..mse2.
## 1                             LM   10360797510
## 2            Stepwise Regression   10366058281
```

As a result, we decided to the selected independent variables and run a linear regression to predict sales prices.

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.

Summary of the statistics of our predicted sales prices are shown in the following table:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-129362  169244  242393  268320  346951 1122637
```

We can conclude that the model is fairly suitable, because the mse is not very high. But, the model still has some drawbacks. For example, we predicted some properties as “negative” values, which is impossible in the real world. Probably because some properties’ specifications are regarded as disadvantages based on our model. Therefore, when we obtained an accurate number from the model, we still need to recheck whether the value is reasonable or not.

Appendix

Include tables and plots here. Enumerate each table and plot, give each a descriptive title, and make sure elements are labeled clearly. Tables and plots should correspond to output from your code.