Chicago House Price Prediction and Classification

Motivation and Main Business Idea

In a metropolitan city like Chicago, housing prices can vary widely. This study aims to create models that predict housing prices and classifications as accurately as possible. Our selected models will enable people to either get an expectation of whether the house they're looking for falls into the expensive or affordable category or gauge the house price.

This study uses different variables to predict/classify house prices and assist potential home buyers/sellers in making decisions. To enhance versatility, we divided Chicago into five areas. For instance, users can choose the appropriate models to determine if the house they plan to purchase is priced appropriately or not.

Users can consider multiple factors to determine whether a house might be affordable. How can our models help someone decide the area they want to move into or better understand their budget given the conditions they set? We'll attempt to answer this question by experimenting with various models in R using the 2021 Chicago Housing data. We'll eventually implement different combinations of variables for each model and analyze summary statistics and graphs to determine the optimal model.

Benefits from Customer Perspective

Clients can obtain a price range for future housing values, allowing them to make profitable investments at the appropriate time.

Benefits from a Real Estate Company Perspective

It is feasible to evaluate what factors can increase the value of a property and construct more beneficial properties.

Data and Empirical Methodology

Data description: The 2021 Chicago Housing Data consists of 12 columns. Some variable types need to be modified for the desired output.

Variables: ZIP, HOUSEID, HPRICE, LOG_PRICE, SQFT, LOG_SQFT, BEDROOM, BATHROOM, GARAGE, AGEBLD, FIREPLACE, SOLD_30DAY

```
'data.frame':   5300 obs. of  12 variables:
 $ ZIP       : int  60002 60002 60002 60002 60002 60002 60002 60002 60002 60002 ...
 $ HOUSEID   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ HPRICE    : int  141000 250000 81000 180000 240000 180000 272500 65000 313000 159900 ...
 $ LOG_PRICE : num  11.9 12.4 11.3 12.1 12.4 ...
 $ SQFT      : int  936 2716 900 2774 2556 2244 2333 1660 3092 1168 ...
 $ LOG_SQFT  : num  6.84 7.91 6.8 7.93 7.85 ...
 $ BEDROOM   : int  2 4 2 4 4 4 5 5 4 3 ...
 $ BATHROOM  : num  1 2.5 1 3 2.5 2.5 4 2 3.5 1 ...
 $ GARAGE    : num  1 3 2.5 2 3 2 2 2 4 2 ...
 $ AGEBLD    : int  43 6 48 23 23 15 11 1 11 25 ...
 $ FIREPLACE : int  0 1 0 1 1 0 1 0 1 0 ...
 $ SOLD_30DAY: int  0 0 1 0 0 0 0 0 0 0 ...
```

To predict both the price and the binary classification, we created two sets of data, keeping only the variables we'll use. The first set consists of HPRICE, SQFT, BEDROOM, BATHROOM, GARAGE, AGEBLD, and FIREPLACE. This set is used for linear regression and multiple regression models. The second set consists of SQFT, BEDROOM, BATHROOM, GARAGE, AGEBLD, FIREPLACE, and PRICE_LEVEL. This set is used for LPM (Linear Probability Model), Logistic Regression, and Random Forest models. In other words, we used linear and multiple regression models to predict HPRICE and LPM, logistic regression, and random forest to predict price classification.

For HPRICE, the mean is 366803, the median is 324000, and the mode is 350000. We classified houses with price above 324000 as expensive and other as affordable. We divided HPRICE into dummy variables 0 and 1 by the median price (324000).

Dummy Variable: PRICE_LEVEL (0 means affordable and 1 means expensive)

We can look at the summary statistics such as P-value and R-squared to determine which variable(s) plays a greater role in predicting the house price for linear and multiple regression. We can look at ROC/AUC for the remaining models to determine the classification method.

H0: We can predict house price/classification based on given variables
H1: We cannot predict house price/classification based on given variables

Results

Descriptive Analytics

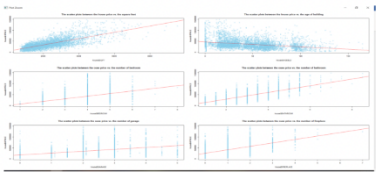
```
      ZIP          HOUSEID         HPRICE          LOG_PRICE          SQFT         LOG_SQFT    
 60002  : 100   Min.   :   1   Min.   :  18075   Min.   : 9.802   Min.   : 700   Min.   :6.551  
 60004  : 100   1st Qu.:1326   1st Qu.: 237975   1st Qu.:12.380   1st Qu.:1445   1st Qu.:7.276  
 60010  : 100   Median :2650   Median : 324000   Median :12.688   Median :2050   Median :7.626  
 60013  : 100   Mean   :2650   Mean   : 366803   Mean   :12.662   Mean   :2248   Mean   :7.616  
 60014  : 100   3rd Qu.:3975   3rd Qu.: 440000   3rd Qu.:12.995   3rd Qu.:2800   3rd Qu.:7.937  
 60025  : 100   Max.   :5300   Max.   :1485000   Max.   :14.211   Max.   :9546   Max.   :9.164  
 (other):4700                                                                                  
    BEDROOM        BATHROOM        GARAGE          AGEBLD          FIREPLACE       SOLD_30DAY    
 Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :  0.00   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 23.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :3.000   Median :2.500   Median :2.000   Median : 44.00   Median :1.0000   Median :0.0000  
 Mean   :3.515   Mean   :2.502   Mean   :2.166   Mean   : 46.35   Mean   :0.6758   Mean   :0.2619  
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.500   3rd Qu.: 65.00   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :8.000   Max.   :8.500   Max.   :5.000   Max.   :165.00   Max.   :7.0000   Max.   :1.0000  
```

- Each zip area has 100 observations
- The mean house price is $366,803, the min price is $18,075, the max price is $1,485,000
- The mean of square feet is 2248, the min SQFT is 700, the max SQFT is 9546
- The mean of the age of the building is 46.35, the min age is 0, the max age is 165
- The mean of the number of bedrooms is 3.5, the min number is 1, the max number is 8
- The mean of the number of bathrooms is 2.5, the min number is 1, the max number 8.5
- The mean of the number of garages is 2.2, the min number is 0, the max number is 5
- The mean of the number of fireplaces is 0.68, the min number is 0, the max number is 7

Correlation, ANOVA and Scatter plot between house price and each independent variable

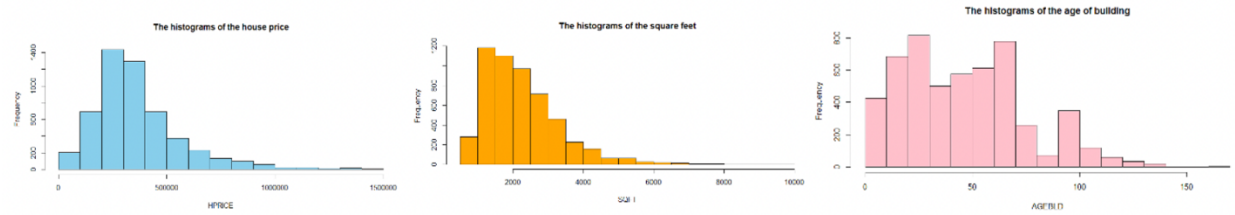
```
> house %>% summarise(correlation = cor(HPRICE, SQFT))
  correlation
1   0.7377394
> house %>% summarise(correlation = cor(HPRICE, AGEBLD))
  correlation
1  -0.3099846
> house %>% summarise(correlation = cor(HPRICE, BEDROOM))
  correlation
1   0.4716259
> house %>% summarise(correlation = cor(HPRICE, BATHROOM))
  correlation
1   0.7347457
> house %>% summarise(correlation = cor(HPRICE, GARAGE))
  correlation
1   0.3250635
> house %>% summarise(correlation = cor(HPRICE, FIREPLACE))
  correlation
1   0.5686125
```

```
> res_aov <- aov(HPRICE~SQFT+BEDROOM+BATHROOM+GARAGE+FIREPLACE+AGEBLD,
+               data = house)
> summary(res_aov)
              Df    Sum Sq   Mean Sq  F value  Pr(>F)
SQFT           1 1.241e+14 1.241e+14 7491.887 < 2e-16 ***
BEDROOM        1 1.253e+11 1.253e+11    7.560 0.00599 **
BATHROOM       1 1.266e+13 1.266e+13  763.980 < 2e-16 ***
GARAGE         1 1.313e+10 1.313e+10    0.792 0.37348
FIREPLACE      1 3.389e+12 3.389e+12  204.521 < 2e-16 ***
AGEBLD         1 5.876e+10 5.876e+10    3.546 0.05975 .
Residuals   5293 8.771e+13 1.657e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Depending on the correlation value and scatter plot results, we can see all variables have a positive relationship with the house price, except the age of the building, which has a negative relationship with the house price. Meanwhile, the square feet and the number of bathrooms look like they have a strong relationship with house prices. The number of fireplaces has a medium-strong relationship with the house price. Furthermore, in the Analysis of variance, according to the f-values and p-values, we can identify that almost all variables have a significant impact on the house price, except the number of bedrooms.

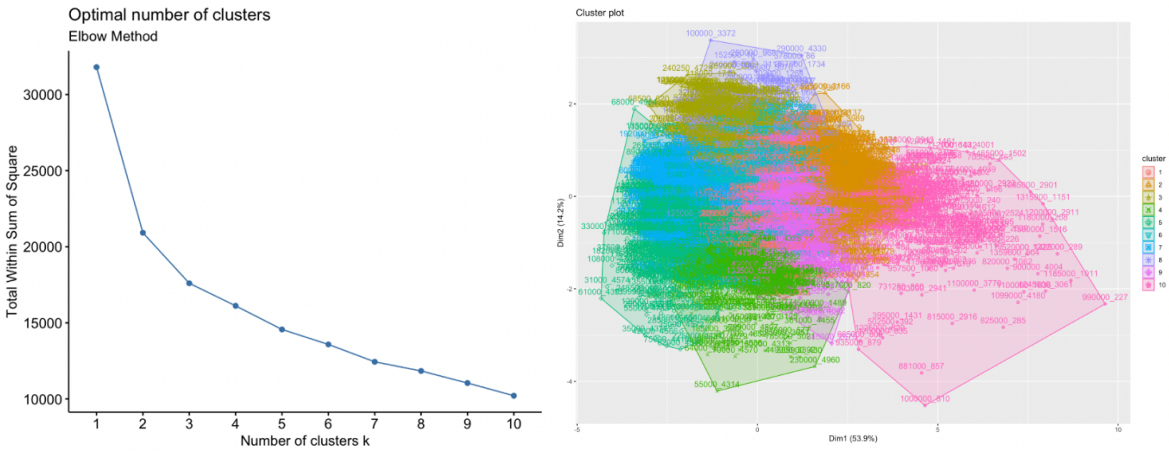
Histograms to Visualize Distribution of Variables



From these grams, we can see that houses under 0.5 million dollars, less than 4,000 square feet, and less than 100 years old are more popular and marketable. Also, people prefer a house with 2 to 4 bedrooms, bathrooms, garages, and one fireplace.

K-Means Clustering

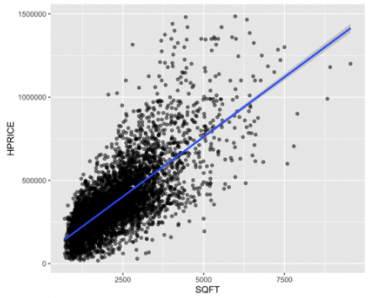
We started with K-means clustering to determine what type of groups exist within the data set. We assigned HPRICE as labels and the rest as data. After scaling the data, we calculated the distance metrics between our observations using the distance function. We used the elbow plot by using the within-sum squares method and chose ten clusters, and found that it is roughly 68%, which means more than half are correctly identified.



Linear Regression

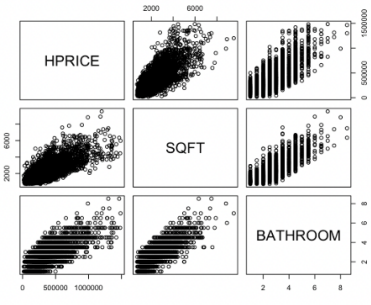
We performed linear regression in R using the LM function by setting total square feet as independent variable and pricing levels as dependent variables. Then we calculated the least squares estimate for the y-intercept and the slope. There is a positive correlation by looking at the plot. The coefficients section tells us the least squares estimate for the fitted line. The p-value for square feet is statistically significant, which is less than 0.05. A significant p-value for SQFT means that it will give us a reliable guess of the house price. By looking at the r-squared, SQFT can explain 54.43% of the variation in house price.

```
Call:
lm(formula = HPRICE ~ SQFT, data = house1)

Residuals:
    Min      1Q  Median      3Q     Max
-570563  -82057  -10825   66856  868991

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 44409.104   4486.619   9.898   <2e-16 ***
SQFT          143.428      1.803  79.543   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140100 on 5298 degrees of freedom
Multiple R-squared:  0.5443,    Adjusted R-squared:  0.5442
F-statistic:  6327 on 1 and 5298 DF,  p-value: < 2.2e-16
```



Multiple Regression

We specified house price as y value and used SQFT and bathroom to predict price. We plotted the data as a first step because it allows us to evaluate whether doing a linear regression, to begin with, is a good idea. We were able to see a relationship between SQFT and price, and BATHROOM with a price. We know that using SQFT and bathroom is better than using SQFT alone because the p-values are significant and the adjusted R2 rises to 59.97. It’s even better to use all variables to predict price with the R2 of 61.5.

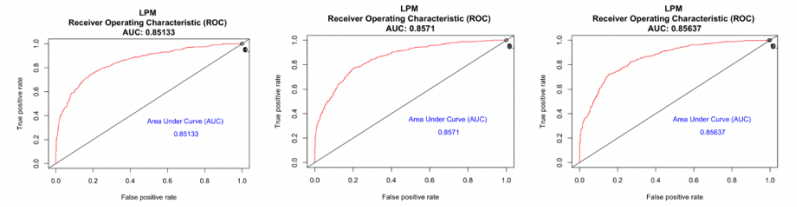



```
Call:
lm(formula = HPRICE ~ SQFT + BATHROOM, data = house_mr)

Residuals:
    Min      1Q  Median      3Q     Max
-423824  -81514   -8487   68671  820306

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23519.537   4893.826  -4.806 1.58e-06 ***
SQFT           80.706      2.864  28.180  < 2e-16 ***
BATHROOM    83495.922   3078.204  27.125  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131300 on 5297 degrees of freedom
Multiple R-squared:  0.5998,    Adjusted R-squared:  0.5997
F-statistic:  3970 on 2 and 5297 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = HPRICE ~ ., data = house1)

Residuals:
    Min      1Q  Median      3Q     Max
-419822  -77861   -8285   66973  779825

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3004.388  10061.211   0.299   0.7652
SQFT           71.247      3.184  22.378   <2e-16 ***
BEDROOM     -6440.498   2812.851  -2.290   0.0221 *
BATHROOM    79375.559   3152.727  25.177   <2e-16 ***
GARAGE      -3160.612   2480.600  -1.274   0.2027
AGEBLD        127.920     67.933   1.883   0.0598 .
FIREPLACE   42322.560   2976.520  14.219   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 128700 on 5293 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.615
F-statistic:  1412 on 6 and 5293 DF,  p-value: < 2.2e-16
```

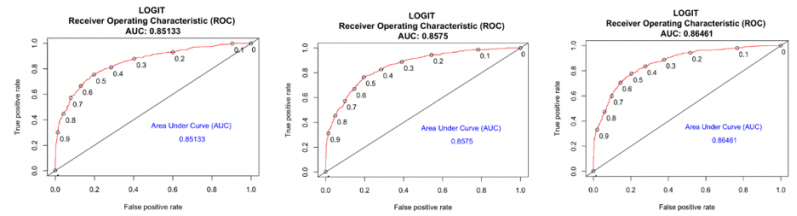
Linear Probability Model

For this model, we split the data into 80% training and 20% testing, using the PRICE_LEVEL to be the predictor. ROC/AUC plots suggest predicting using SQFT and BATHROOM.



Logistic Model

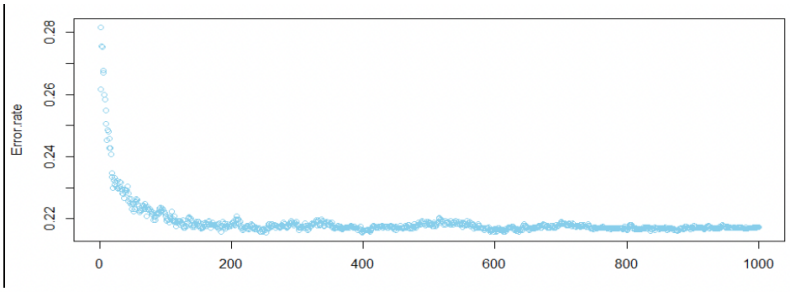
Then we did the same with the logistic model, the AUC is slightly higher than the LPM at 0.86461, still, using all the variables win. SQFT, SQFT+BATHROOM, ALL.



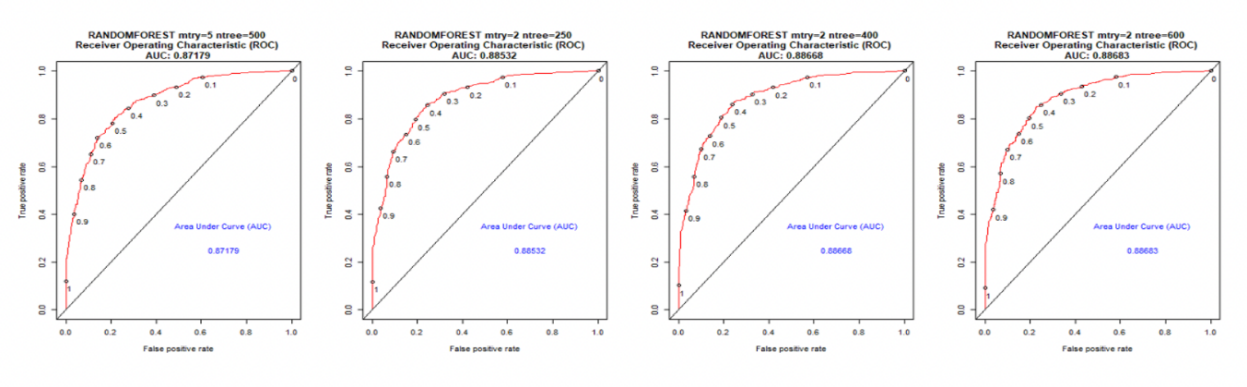
Random Forest

For random forest, we split the data into 80% training and 20% testing. We need to find the optimal number of trees and mtry for our model, so we used rf to perform classification.

```
> cbind(1:12, oob.values)
      oob.values
 [1,]  1  0.2099057
 [2,]  2  0.2049528
 [3,]  3  0.2120283
 [4,]  4  0.2106132
 [5,]  5  0.2186321
 [6,]  6  0.2193396
 [7,]  7  0.2186321
 [8,]  8  0.2174528
 [9,]  9  0.2224057
[10,] 10  0.2183962
[11,] 11  0.2176887
[12,] 12  0.2195755
```



From the graphs above, we can identify the best number of mtry is 2 and the best number of trees might be 250, 400, or 600. Therefore, we conducted separate simulations for these three values and summarized all the AUC result graphs. (0.8718, 0.8853, 0.8867, 0.8868)



Consequently, the best model for classifying, in this case, is the random forest model with the number of mtry is 2 and trees is 600.

Lastly, we repeat the process for each area just like how we did for the entire Chicago area, and selected the best model for each area, just in case someone already has a place in mind, and he/she can use the best model specifically for that area.

Summary Table of Models for Each Area

%	North	West	Southwest	South	Other
Linear Regression (R^2)	54.44	59.86	66.26	36.05	55.01
Multiple Regression (Adj. R^2)	61.42	66.73	71.37	52.36	72.37
LPM (AUC)	87.02	92.55	90.35	81.24	94.30
Logistic Regression (AUC)	87.61	92.88	90.45	81.44	94.67
Random Forest (AUC)	88.32	92.11	91.66	76.52	94.33

Summary

We observed that the linear regression model demonstrates the correlation between variables while providing summary statistics that are less complex than those of other models. The Multiple R-squared and the P-value offered valuable insights into the data. On the other hand, the random forest model performed well in classifying the data with a relatively low OOB error rate and higher AUC, surpassing k-means clustering.

For Chicago, the best model to predict house prices is the multiple regression model, with independent variables including SQFT, the age of the building, and the number of bedrooms, bathrooms, garages, and fireplaces. Meanwhile, the best model to classify the price level is the random forest model with mtry set to 2 and the number of trees set to 600.

For the North area, the best model to predict house prices is the multiple regression model. The best model to classify the price level is the random forest model with mtry set to 1 and the number of trees set to 100.

For the West area, the best model to predict house prices is the multiple regression model, and the best model to classify the price level is the logistic regression model.

For the Southwest area, the best model to predict house prices is the multiple regression model, and the best model to classify the price level is the random forest model with mtry set to 2 and the number of trees set to 570.

For the South area, the best model to predict house prices is the multiple regression model, and the best model to classify the price level is the logistic regression model.

For the faraway areas, the best model to predict house prices is the multiple regression model, and the best model to classify the price level is the logistic regression model.

The project aimed to understand the effect of various factors on house prices and price classification and to design the best model for predicting the future housing market. Other factors like surroundings, crime rates, or interactions between independent variables may also influence house prices. Collecting more factor data and comparing it to previous price trends might benefit the model's development.