Project Report

Introduction:

The question to be studied in this project is how do the variables “year of construction”, “number of floors”, “number of underground floors” and “floor to ceiling height” influence “whether there are elevators or not in the building” according to the 2012 US Commercial Building Energy Consumption Survey (CBECS). If they are related, this project will predict “whether there are elevators or not in the building” based on the analysis.

This question is interesting because the factors seem to be relevant to elevators, but if there is no data to analyze, it is difficult for us to judge whether there are elevators in a building based on these simple values. However, through the analysis of this project, after mastering these pieces of information, it is possible to make predictions with certain accuracy. Logistic regression will be used to analyze relationships and forecasts and cross-validation could also be significant to improve the accuracy in this project.

Data and Methodology:

The data come from the 2012 US CBECS. The dependent variable is “whether there are elevators or not” named ELEVTR in the data set. The independent variables are Year of construction (YRCON), Number of floors (NFLOOR), Number of underground floors (BASEMNT) and Floor to ceiling height (FLCEILHT). R language is applied to complete all the analysis.

Since “Whether there are elevators” is a typical binomial distribution variable, it is appropriate to use logistic regression to analyze its factors. The specific steps of analysis are shown below.

(1) Firstly, data are filtered and cleaned, and only valid information is retained. In order to facilitate subsequent prediction and testing, the overall data is divided into training set and test set and the proportion is 8:2.

(2) Secondly, perform logistic regression and judge whether each independent variable has sufficient influence on the dependent variable based on the p-value. Also, eliminate insignificant factors to improve the prediction accuracy.

(3) Thirdly, predict “whether there are elevators based on the logistic regression results and test the accuracy of the prediction.

(4) Fourthly, in order to make the logistic regression more reasonable and the prediction more reliable, cross-validation is used to compare the models and data are divided into 10 folds.

(5) Fifthly, in order to show the relationship between the variables, choose the best performing one out of 10 folds in the cross-validation, that is, the one with the highest accuracy.

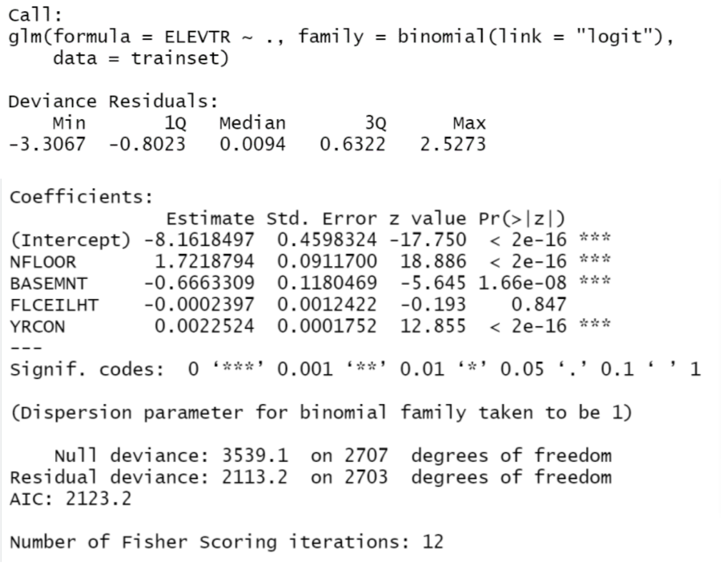
(6) Lastly, draw ROC diagram and calculate AUC value to test its performance.

Please refer to the link for specific R codes.

https://github.com/Lehao25/Stats506_public/tree/master/final_project/R+code.R

Results:

Firstly, in step 2, the result of the logistic regression that shows the relationship between elevators and the factors could be seen in Table 1 below.

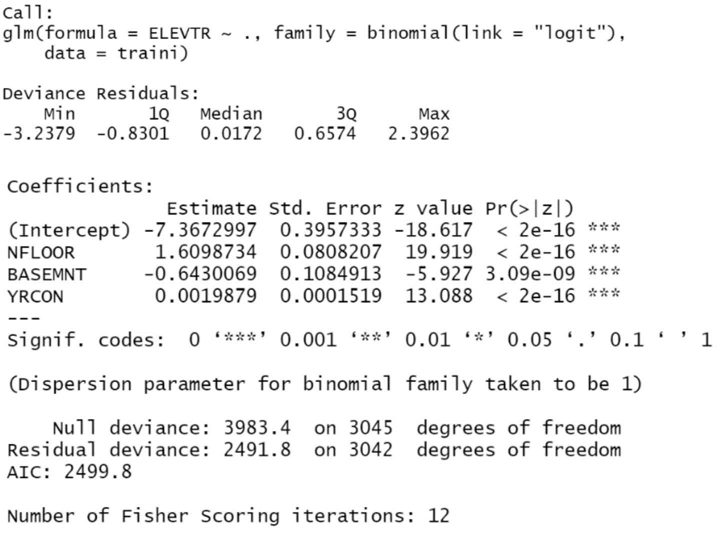


```
Call:
glm(formula = ELEVTR ~ ., family = binomial(link = "logit"),
    data = trainset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3067  -0.8023   0.0094   0.6322   2.5273

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.1618497  0.4598324 -17.750  < 2e-16 ***
NFLOOR       1.7218794  0.0911700  18.886  < 2e-16 ***
BASEMNT     -0.6663309  0.1180469  -5.645 1.66e-08 ***
FLCEILHT    -0.0002397  0.0012422  -0.193    0.847
YRCON        0.0022524  0.0001752  12.855  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3539.1  on 2707  degrees of freedom
Residual deviance: 2113.2  on 2703  degrees of freedom
AIC: 2123.2

Number of Fisher Scoring iterations: 12
```

According to the above results, it is not difficult to find that the p-value of FLCEILHT is significantly greater than 0.05, indicating that it has no significant effect on the elevator. Therefore, the factor “Floor to ceiling height” is eliminated in the subsequent analysis.

Secondly, in step 3, after predicting the probability based on the new logistic regression result, the accuracy which is the ratio of the correct number to the total number is calculated. Here, the accuracy is 0.7622 that is not bad. However, cross-validation might make it better.

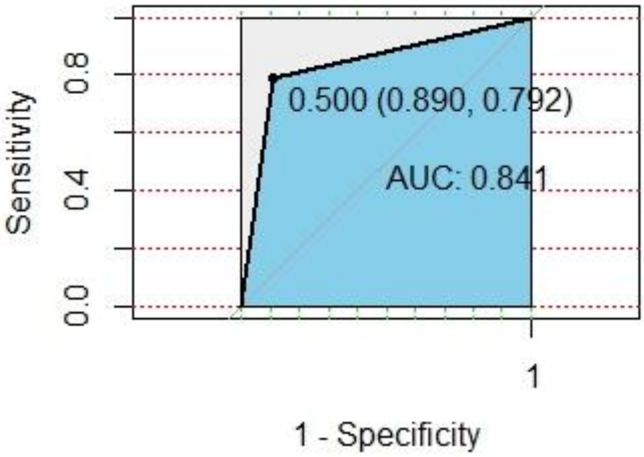
Thirdly, in step 4 and 5, data are divided into 10 folds and the best performing fold with the highest accuracy is chosen. Finally, the 10th group is chosen, and the test set accuracy is 0.8260 while the training set accuracy is 0.7876. These are indeed more accurate than the previous conclusions given by logistic regression analysis without cross-validation. The result of the logistic regression through 10-fold cross-validation that shows the relationship between elevators and the factors could be seen in Table 2 below.



```
Call:
glm(formula = ELEVTR ~ ., family = binomial(link = "logit"),
    data = traini)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2379  -0.8301   0.0172   0.6574   2.3962

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.3672997  0.3957333 -18.617  < 2e-16 ***
NFLOOR       1.6098734  0.0808207  19.919  < 2e-16 ***
BASEMNT     -0.6430069  0.1084913  -5.927 3.09e-09 ***
YRCON        0.0019879  0.0001519  13.088  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3983.4  on 3045  degrees of freedom
Residual deviance: 2491.8  on 3042  degrees of freedom
AIC: 2499.8

Number of Fisher Scoring iterations: 12
```

Still, these three factors are very significant according to the very small p-value. It should be noted that even if the AIC has risen compared to before, the current results are still more accurate. Therefore, cross-validation is important in this case.

Finally, in step 6, ROC diagram is drawn, and AUC value is calculated to test the performance of the model. The figure below shows the details.



It can be seen from the figure that the ROC curve is close to the upper left corner, which shows again that the prediction is accurate. Also, the AUC value which indicates the Area Under Curve is 0.841, close to 1, showing that the forecast is quite accurate.

Conclusion and Discussion:

Through the analysis of this project, it is not difficult to find that, among the four variables, Year of construction, Number of floors and Number of underground floors are very related to elevators, while the Floor to ceiling height has no influences on elevators. It is also worth studying that the logistic regression and cross-validation play very important roles in the prediction and finally give a good prediction with an accuracy of about 80%.

The strength of this project is choosing a suitable logistic regression model, and comprehensively consider using the cross-validation method to process the data and thus improve a certain accuracy. The limitation of this project is that it does not consider the interaction between the variables, in particular, “Number of floors” and “Number of underground floors” are likely to be related rather than independent. If the interaction relationship can be further considered, the prediction may be more precise, which is also the goal of further research.