A STUDY OF THE PREDICTION OF THE MODEL LOGISTIC REGRESSION IN MACHINE LEARNING: FOCUSING ON A SURVEY

Namkil Kang
Far East University, South Korea

The ultimate goal of this paper is to make the logistic regression model learn train data so that it can predict our subjects' vote in the last presidential election. When it comes to categorical data such as gender and regional places, it is necessary to readjust them. By using one hot encoding, we readjusted them. A point to note is that cross validation without random is better than cross validation with random. More specifically, the logistic regression model works well with cross validation without random, whereas it does not work well with cross validation with random. A further point to note is that when C (a parameter) was 100, the logistic regression model obtained the best score (100%) in train data and test data by using grid search. On the other hand, when C was 23, the logistic regression model obtained the best score (100%) in train data and test data by using random search. A major point of this paper is that when C is 10, the accuracy rate of the model logistic regression is 100% in both of train data and test data. Finally, the Receiver Operating Characteristic (ROC) analysis tells us that the space between true positive and false positive is big. This in turn implies that the accuracy rate of the classification model logistic regression is high. It is clear from our findings that the logistic regression model is good enough and that it works well for train data and test data.
Keywords: machine learning, python, prediction, learning, ROC, logistic regression

1. Introduction

The main purpose of this paper is to train the logistic regression model so that it can predict our subjects' vote in the last presidential election. Put differently, we train the logistic regression model to predict whether or not the subjects took a vote in the presidential election. This research was carried out by python. We let the company survey-top conduct a survey with respect to political interest, the subjects' support of the Korea government, and their vote in the last presidential election. Additionally, we obtained information on gender and regional places where 107 people live in. 107 participated in our survey and 107 sets of information were obtained. First, this research aims to provide information on data and preprocesses them in terms of one hot encoding. The term one hot encoding refers to readjusting categorical data. We attempt to make categorical data such as gender and regional places true or false. Second, we train the logistic regression model to predict whether the subjects took a vote in the presidential election by using cross validation without random, cross validation with random (K-fold cross validation), and shuffle-split cross validation. When it comes to the cross validation without random, this method consists of 25 folds. Among 25 folds, 20 folds are used for train data, whereas 5 folds are used for test data. This method is the most widely used one.

Talking about the cross validation with random, it usually uses 5-fold cross validation. That is to say, data are divided into 5. 4 are used for train data and 1 is used for test data. As for the shuffle-split cross validation, it refers to mixing data. The weak point of this method is that some data cannot belong to any of train data or test data. Third, we probe into grid search and random search. More specifically, to make the logistic regression model predict whether or not 107 people took a vote in the last presidential election, we use grid search and random search. By doing so, we obtain the best parameter and the best accuracy of the model logistic regression. When it comes to grid search, it indicates that a researcher sets a parameter or parameters. On the other hand, talking about random search, it indicates that a researcher sets the scope of a parameter so that the model logistic regression can obtain the best score. Finally, we make the model logistic regression train and predict 107 people's responses. Again, we use grid search and random search in terms of cross validation. Also, we attempt to evaluate the model logistic regression in terms of the ROC (Receiver Operating Characteristic) analysis. An ROC curve is a graph showing the performance of the classification model (the logistic regression model). Two parameters, namely true positive and false positive consist of the graph. The term true positive is a test result that correctly indicates the presence of a condition or characteristic. On the other hand, the term false positive is a test result which wrongly indicates that a particular condition or attribute is present. The more the space between true positive and false positive is big, the more the logistic regression model is good.

2. Methods

We let the company survey-top conduct a survey. 107 people participated in the survey. The survey provides information such as gender, regional places where 107 people live in, the interest of politics, the subjects' support of the Korea government, a question of having a progressive idea, the subjects' support of a party, and a vote in the last presidential election. When it comes to gender, we assigned 1 to a male, whereas we assigned 2 to a female. Talking about the regional places, we assigned 1 to the metropolitan area, we assigned 2 to the Chungcheong area, we assigned 3 to the Honam area, we assigned 4 to the Youngnam area, and we assigned 5 to others. As for the interest of politics, the subjects' support of the Korea government, and a question of having a progressive idea, we assigned 1 to the positive response, whereas we assigned 0 to the negative response. With respect to the subjects' support of a party, 1 is assigned to the ruling party, 2 are to the opposition party, and 3 are to others. When it comes to a vote in the last presidential election, we assigned 1 to the positive response, whereas we assigned 0 to the negative response. Finally, as a research tool, python was used. We used the classification model logistic regression to predict whether or not 107 people took a vote in the last presidential election. We obtained feedback through grid search and random search and evaluated the classification model logistic regression.

3. A Vote in the Last Presidential Election

3.1. Data Preprocessing and One Hot Encoding

In what follows, we aim at preprocessing raw data and readjusting them. When it comes to categorical data such as gender and regional places, it is necessary to readjust categorical data. The following table shows low data about gender and regions:

Table 1 Categorical data

Number	Gender	Region
1	1	4
2	1	1
3	2	5
4	1	1
102	2	1
103	1	1

As can be seen from table 1, we assigned 1 or 2 to a male or a female, respectively. As illustrated in Table 1, we assigned 1, 2, 3, 4, 5 to the metropolitan area, the Chungcheong area, the Honam area, the Youngnam area, and others, respectively. However, this figure itself does not mean a scale of values. This figure is for a convenient classification. If we do not readjust these categorical data, they will count as y-label. Simply put, these data are different from the subjects' support of the Korean government, the interest (intention) of politics, and a question of asking a progressive idea. Thus, the best way of solving this problem is to make the categorical data true or false, as exemplified in Table 2 and Table 3:

Table 2 One Hot Encoding

Number	Gender	Region
1	male	Youngnam
2	male	Sudo
3	female	Others
4	male	Sudo
5	female	Youngnam
102	female	Sudo
103	male	Sudo
104	female	Sudo
105	female	Sudo
106	male	Sudo

As exemplified in Table 3, categorical data was readjusted in terms of one hot encoding:

Table 3 One Hot Encoding

Number	Gender	Chungchong	Honam	Sudo	Youngnam	Others
1	False	True	False	False	False	True
2	False	True	False	False	True	False
3	True	False	False	False	False	False
4	False	True	False	False	True	False
5	True	False	False	False	False	True
102	True	False	False	False	True	False
103	False	True	False	False	True	False
104	True	False	False	False	True	False

105	True	False	False	False	True	False
106	False	True	False	False	True	False

It is worthwhile pointing out that the classification model logistic regression does not work for categorical data. The model logistic regression works for the interest of politics, the subjects' support of the Korean government, and the question of having a progressive idea. This model learns these data and predict whether 107 people took a vote in the presidential election or not.

3.2. Cross Validation without Random, Cross Validation with Random, and Shuffle-Split Cross Validation

In the following, we train the classification model logistic regression to predict whether or not 107 people took a vote in the presidential election. For this, we will use cross validation without random, cross validation with random, and shuffle-split cross validation.

To begin with, we will use cross validation without random. The so-called cross validation without random is made up of 25 folds. 80% (20 folds) of data are used for train data, whereas 20% (5 folds) are used for test data. First, we imported LogisticRegression and cross val score and we divided train data into 5 sets. The following table shows the accuracy rate of the logistic regression model in 5 sets, respectively:

Table 4 Cross Validation without Random

Set	The Accuracy Rate of the Logistic Regression Model
Set 1	93.75
Set 2	100
Set 3	93.75
Set 4	93.75
Set 5	93.75
Average	95

As indicated in Table 4, train data are divided into 5 sets. The accuracy rate of set 1 is 93.75%. Exactly the same can be said of set 3, set 4, and set 5. Their accuracy rate is 93.75%, respectively. It is worthwhile noting that in set 2, the accuracy rate of the logistic regression model is 100%. Note that the average of the accuracy rate is 95%. This in turn indicates that through cross validation without random, the accuracy rate of the classification model logistic regression is good enough.

Now attention is paid to cross validation with random (K-fold cross validation). The so-called cross validation with random uses 5-fold cross validation. Data are divided into five sets. We imported LogisticRegression and KFold. Thus, data are divided into five sets (5 splits). When it comes to the accuracy rate of 5 test sets, that of the classification model logistic regression is as follows:

Table 5 Cross Validation with Random

Set	The Accuracy Rate of the Logistic Regression Model

Set 1	100
Set 2	100
Set 3	87.5
Set 4	68.75
Set 5	100
Average	91.25

It is probably worthwhile pointing out that in set 1, set 2, and set 5, the accuracy rate of the classification model logistic regression is 100%, respectively. By using cross validation with random, the logistic regression model seems to work well for 5 test sets, but if we consider set 3 and set 4, things are different. More specifically, the accuracy rate of the logistic regression model in set 4 is the lowest (68.75%). Notice that the average of the accuracy rate of the logistic regression model is 91.25%. This amounts to saying that cross validation without random is better than cross validation with random. Simply put, the classification model works well for train data with cross validation without random.

Finally, attention is paid to shuffle-split cross validation. It refers to mixing data. We imported LogisticRegression and ShuffleSplit. Through the ShuffleSplit, data were divided into test data and train data. Test size is 0.5 (50%) and train size is 0.5 (50%). The accuracy rate of the logistic regression model through shuffle-split cross validation is as follows:

Table 6 Shuffle-Split Cross Validation

Number	**The Accuracy Rate of the Model Logistic Regression**
1	100
2	92.5
3	95
4	92.5
5	95
6	85
7	87.5
8	95
9	82.5
10	92.5
Average	91.75

As indicated in Table 6, shuffle-split cross validation is better than cross validation with random. However, cross validation without random is the best among 3 methods. It therefore seems safe to contend that the logistic regression model works well with cross validation without random, whereas it does not work well with cross validation with random.

3.3. Grid Search and Random Search

This section aims to probe into grid search and random search. We use grid search and random

search and make the logistic regression model train and predict whether or not 107 people took a vote in the election. The term grid search indicates that a researcher sets parameters. On the other hand, the term random search indicates that a researcher sets the scope of a parameter and the classification model (the logistic regression model) obtains the best parameter and its best accuracy automatically.

To begin with, let us turn our attention to grid search. We imported LogisticRegression and GridSearchCV. The parameters that we set are 0.001, 0.01, 0,1, 10, and 100. If a parameter is 0, to divide the positive answer and the negative one into two will be a straight line in the graph. However, if a parameter is bigger than 10 (100), to divide them into two will be a bent S shape. Through grid search, we obtained the best parameter and the best score (the best accuracy rate). When C (a parameter) was 100, we obtained the best score (100%). Quite interestingly, when it comes to test data, the accuracy rate of the logistic regression model is also 100%.

The following graph shows the accuracy rate per five parameters. It is worth noting that whenever C (a parameter) is bigger, the accuracy rate of the model logistic regression increases. When C (a parameter) is 100 in the case of train data and test data, the accuracy rate of the logistic regression model is 100%. This in turn implies that when we set a parameter as 100, the logistic model works well for both of train data and test data:

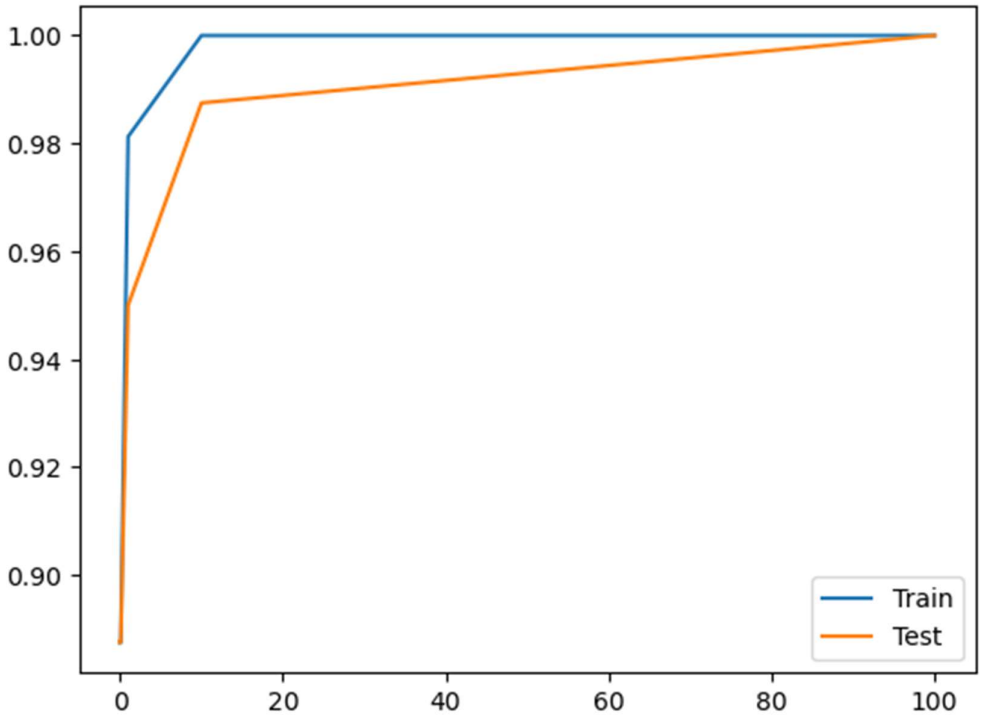


Figure 1 The Accuracy Rate of the Model Logistic Regression (Grid Search)

Now let us turn our attention to random search. We imported LogisticRegression and RandomizedSearchCV. We trained the logistic regression model and made it predict 107 people's responses. We set the scope of parameters from 1 to 100. When C was 23, the logistic regression model obtained the best score (100%). More importantly, in the case of test data, the accuracy rate of the logistic regression model is also 100%. The following graph shows the

accuracy rate of the logistic regression model. Talking about train data, when C is around 23, the accuracy rate of the logistic regression model is the highest, whereas C is bigger than 23, that of the logistic regression model is low. It must be stressed that when C is around 100, the accuracy rate is 100% again. Talking about test data, when C is more than 10, the accuracy rate of the logistic regression model is 100%.

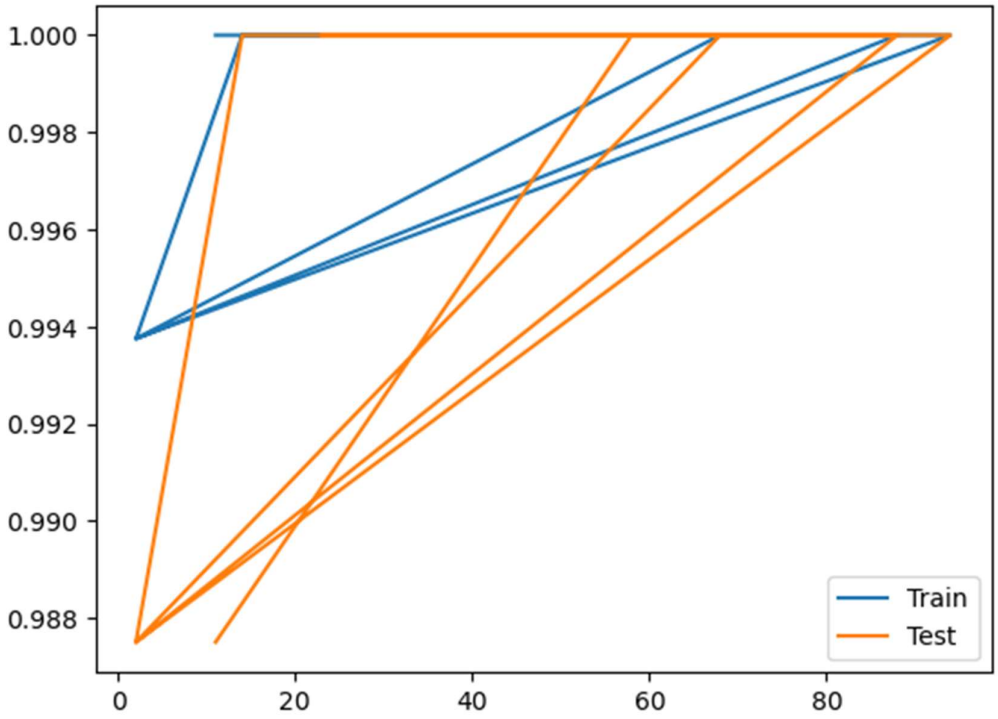


Figure 2 The Accuracy Rate of the Model Logistic Regression (Random Search)

3.4. Grid Search, Random Search, and Model Evaluation

The goal of this section is concerned with using grid search and random search and evaluating the model logistic regression in terms of the ROC analysis.

Now let us turn our attention to grid search. We imported LogisticRegression and GridSearchCV. We set cross validation. That is to say, data were divided into five sets. Most importantly, the parameter of grid is 0.001, 0.01, 0.1, 1, 10, and 100. We applied grid search to the logistic regression model and made it learn train data. Quite interestingly, the best parameter is 100 and the best cross-validity score (the best accuracy) is 100%.

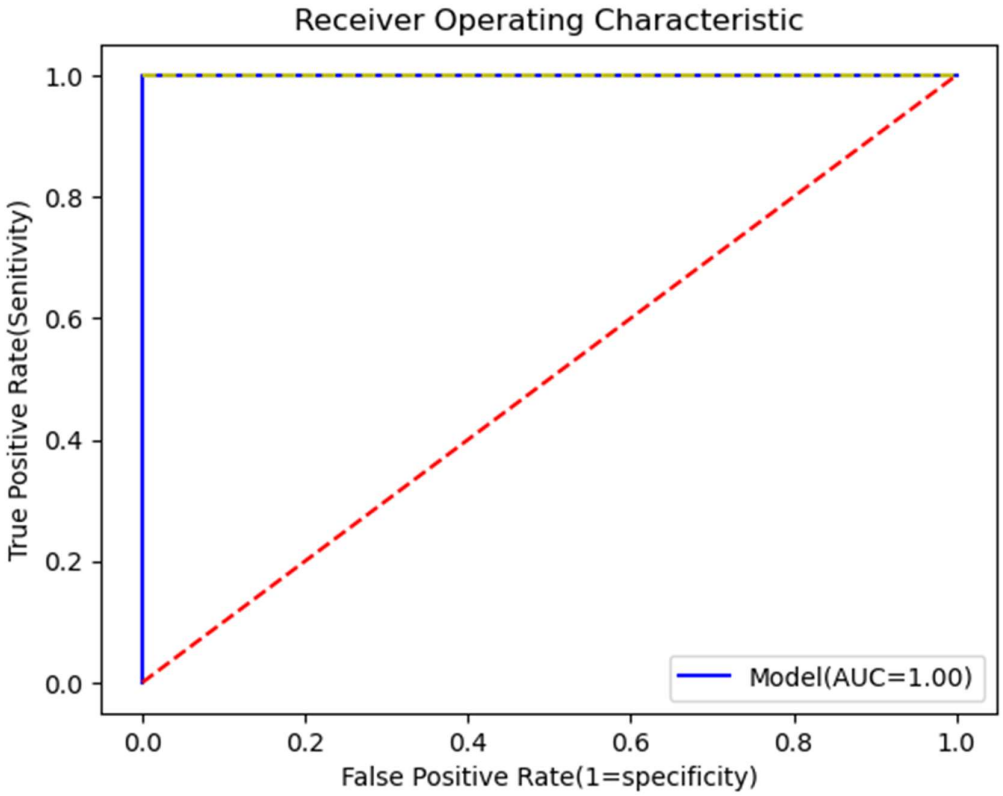
Now attention is paid to random search. We imported LogisticRegression and RandomizedSearchCV. We set cross validation. Simply put, data were divided into five sets. The parameter of random search is from 1 to 100. More importantly, the best parameter is 73, whereas the best cross-validity score is 100%. It seems thus reasonable to contend that when in the case of grid search, C is 100, the accuracy of the model logistic regression is the highest (100%), whereas when in the case of random search, C is 73, that of the model logistic regression is the highest (100%).

Now attention is paid to a final model. When C is 10, let us consider the accuracy of the logistic regression model. Most importantly, in the case of train data, the accuracy rate of the model logistic regression is 100% and in the case of test data, that of the model logistic

regression is also 100%. This in turn suggests that the classification model logistic regression works well for both of train data and test data.

Now we attempt to evaluate the model logistic regression in terms of the ROC (Receiver Operating Characteristic) analysis:

Figure 3 Receiver Operating Characteristic



An ROC curve is a graph showing the performance of the classification model. Two parameters, true positive and false positive are made up of the graph. The so-called true positive is a test result that correctly indicates the presence of a condition or characteristic. On the other hand, the so-called false positive is a test result which wrongly indicates that a condition or attribute is present. To go into detail, the Receiver Operating Characteristic (ROC) analysis is used to determine whether the relevant model performs better than random guessing. This graph forms a triangle, which in turn indicates that the space between true positive and false positive is big. The more the space between true positive and false positive is big, the more the logistic regression model works well for train data and test data. It can thus be concluded that the classification model logistic regression works well for both of train data and test data that we obtained in the survey. For the computational analyses of linguistics, see Kang (2023a, 2023b, 2023c, 2023d, 2023e, 2023f).

4. Conclusion

To sum up, we have made the logistic regression model learn train data so that it can predict our subjects' vote in the last presidential election. In section 3.1, we have preprocessed raw data and readjusted them. When it comes to categorical data such as gender and regional places,

it is necessary to readjust them. In section 3.2, we have argued that cross validation without random is better than cross validation with random. More specifically, the logistic regression model works well with cross validation without random, whereas it does not work well with cross validation with random. In section 3.3, we have maintained that when C (a parameter) was 100, the logistic regression model obtained the best score (100%) in train data and test data by using grid search. We have contended that when C was 23, the logistic regression model obtained the best score (100%) in train data and test data by using random search. In section 3.4, we have argued that when C is 10, the accuracy rate of the model logistic regression is 100% in both of train data and test data. Finally, the Receiver Operating Characteristic (ROC) analysis tells us that the space between true positive and false positive is big. This in turn indicates that the classification model logistic regression is good enough and that it works well for train data and test data.

References

[1] Kang, N. (2023a). K-Pop in BBC News: A Big Data Analysis. *Advances in Social Sciences Research Journal 10*(2), 156-169.

[2] Kang, N. (2023b). K-Dramas in Google: A NetMiner Analysis. *Transaction on Engineering and Computing Sciences 11*(1), 193-216.

[3] Kang, N. (2023c). A Comparative Analysis of Tolerate and Put up with in the COCA. *Semiconductor and optoelectronics* 42(1): 1468-1476.

[4] Kang, N. (2023d). Sure of and Sure about in Corpora and ChatGPT. *Journal of Harbin Engineering University* 44(7): 1347-1351.

[5] Kang, N. (2023e). Turn out adj and Turn out to be adj in the Now Corpus and ChatGPT. *Journal of Harbin Engineering University* 44(8): 825-831.

[6] Kang, N. (2023f). Care for and Like in Corpora and ChatGPT. *Semiconductor and optoelectronics* 42(2): 188-198.

A Survey

Please read the following questions and answer our questions.

1. Where do you live?

1) the Metropolitan area 2) the Chungcheong area 3) the Honam area 4) the Youngnam area 5) others

2. Are you a male or a female?

1) male 2) female

3) Are you interested in politics?

1) Yes 2) No

4) Do you have a progressive idea?

1) Yes 2) No

5) Did you support Moon’s government?

1) Yes 2) No

6) Did you take a vote in the last presidential election?

1) Yes 2) No

7) Which party do you support?

1) the ruling party 2) the opposition party 3) others