Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231088

Predicting and analyzing long-term deposit possibility leveraging logistic regression and visual analysis

Heying Xu

School of Social Science, University of California, Irvine, California, 92697, USA

heyingx@uci.edu

Abstract. Keeping balanced deposits in commercial banks is of great importance for the rational use of money. This article focuses on logistic regression to forecast the long-term deposit activity of banks. And the visual analysis is helpful to reveal a strong association between people's likelihood of long-term deposits and their age and happiness. In detail, people are less likely to make long-term deposits after the age of 65, i.e., after the age of retirement. In addition, people are more likely to make long-term savings when they have a high life expectancy index, such as when they are married. Finally, the whole logistic regression model is analyzed by using precision, recall, and f1-score. It is concluded that this logistic regression model is very accurate in predicting the long-term deposits of bank customers. However, the concluding section also explores the idea that the overall logistic regression model should use a broader data set for its effectiveness and high accuracy.

Keywords: machine learning, logistic regression, bank deposit.

1. First section

Over the past decade, time deposits in commercial banks have grown by nearly 170%, 10 times faster than demand deposits [1]. This is because the interest rates on time deposits are much larger than those on demand deposits. So, more people are willing to invest through time deposits, a relatively safer and more rewarding way. This time deposit pass-through means that demand deposits have a very significant impact on the internal growth of banks and on the society's economy as a whole. Because when people are very dependent on time deposits, the existing money flow in the society is greatly reduced. This raises a very serious problem: there is a great imbalance between the supply of available money and people's demand for it [2].

Therefore, banks need to control the amount of time deposits in banks by forecasting people's demand for time deposit items. Next, when the bank predicts the next tendency of people for time deposits, it can change the degree of people's tendency for time deposits by regulating the interest rate. This is because when the interest rate is small, people's demand for deposits becomes smaller because their opportunity cost increases. Similarly, when the interest rate on time deposits becomes larger, the demand for time deposits also becomes smaller. Of course, using machines for forecasting is more accurate because they can increase the accessibility of data and analyze big data quickly and comprehensively [3]. In addition, logistic regression is a very suitable model for predicting the term deposit items of banks. Because logistic regression can analyze the observed data very effectively and can be adjusted to reduce some potential bias [4].

DOI: 10.54254/2755-2721/20/20231088

2. Method

2.1. Dataset

Data from the UCI Machine Learning Repository is chosen in this study. This data is about a telemarketing campaign of a Portuguese bank. The goal is to predict whether a customer will choose to subscribe a time deposit or not by analyzing the variables in the data. In this data there is one numeric variable: age, and six categorical variables: Job (status), Marital (status), Education (status), Default (Yes or No), Housing (Yes or No), and Loan (Yes or No).

2.2. Logistic regression

Since it is necessary to analyze the probability of a customer's subscription to a time deposit (opt-in or opt-out), logistic regression is a good method to make predictions. Logistic regression is a widely used method for binary classification and can provide valid results to analyze the final impact of each variable on the two predicted outcomes. Logistic regression is a statistical method that can model the probability of binary variables of one or more predictor variables. As in this data whether the customer will subscribe to a time deposit: 1. the ratio of the probability of success: will subscribe to a time deposit with a binary variable of 1; 2. the probability of failure: will not subscribe to a time deposit binary variable of 0).

Also, the relationship between his predictor and response variables is linear on the logit scale. Thus, mathematically, the logistic regression model can be described as $\text{Logit}(p) = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp$ [5]. p is the probability of success, x1 to xp are the predictor variables, and β0 to βp are the coefficients or parameters of the model that need to be estimated from the data and this model is fitted by maximizing the likelihood function. The likelihood function is a function of the model parameters and the observed data, and it represents the probability of observing the data through the model. When the model is fitted, it can be used to predict the probability of success of a new observation based on the values of the predictor variables.

Second, logistic regression can also handle outliers in the data to help eliminate misinformation and extreme values that affect the prediction. Moreover, in a data set with seven different variables, logistic regression is computationally efficient and can analyze the relationship between the predictor and outcome variables to identify potential factors affecting the results.

2.3. Evaluation metrics

To evaluate this logistic regression method, the precision and recall and f1-score values from different situations (subscribe and do not subscribe to a time deposit) can help to determine the final result of this method [6].

Precision is a measure of the value of the proportion of positive situations correctly predicted out of all positive situations predicted. When the precision of the model is high, the false positive error is small. In addition, recall measures the proportion of correctly predicted positive situations to all actual positive situations [7]. A higher recall indicates that the model produces few false negative errors. Finally, the F1-score is the summed average of the accuracy and recall rates [8]. It ranges from zero to one, and a score closer to one shows better performance. Thus, the F1-score is a combined measure of accuracy and recall.

3. Result

3.1. Visual analysis

By using this logistic regression model to predict the likelihood of bank customers' subscription of a term deposit, the visual analysis graph in the model shows that there is a strong correlation between age group and the likelihood of subscribing to a long-term deposit. As this histogram distribution in Figure 1 shows, the age group of sixty and above is more reluctant to make a long-term deposit of their money.

DOI: 10.54254/2755-2721/20/20231088

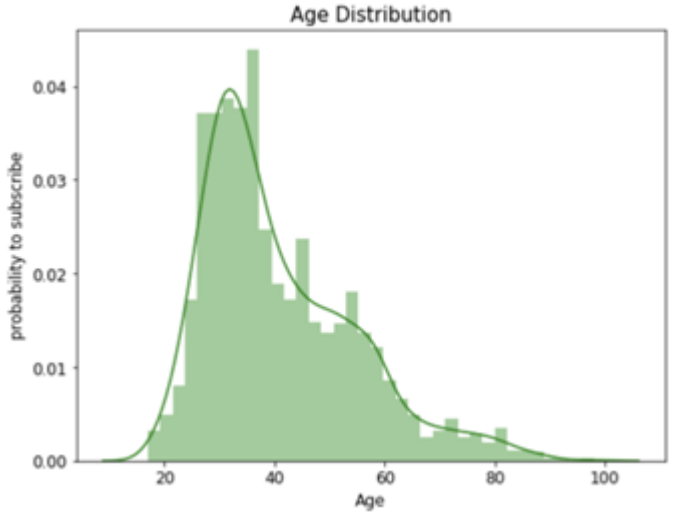

Figure 1. Relationship between age and subscribe probability.

Also, because the United States Social Security normal retirement age is between 65 and 67 [9]. it is possible to analyze that customers in the retired group are more reluctant to make long-term deposits. In addition, the image also indicates that the working population group which aged between 20 and 40 is more likely to make time deposits. Thus, it can be found that the group with high liquidity is more willing to make long-term deposits. The reverse is also true, as the older retired group with low liquidity is less likely to engage in deposits compared to the working group. Another reason is that older people will be more reluctant to engage in such online operations. For example, they prefer to use cash rather than bank cards. Therefore, from the age of 40 onwards, when people are gradually working less and getting older, they are less likely to subscribe to long term savings as a negative slope and therefore less likely to participate in savings every year.

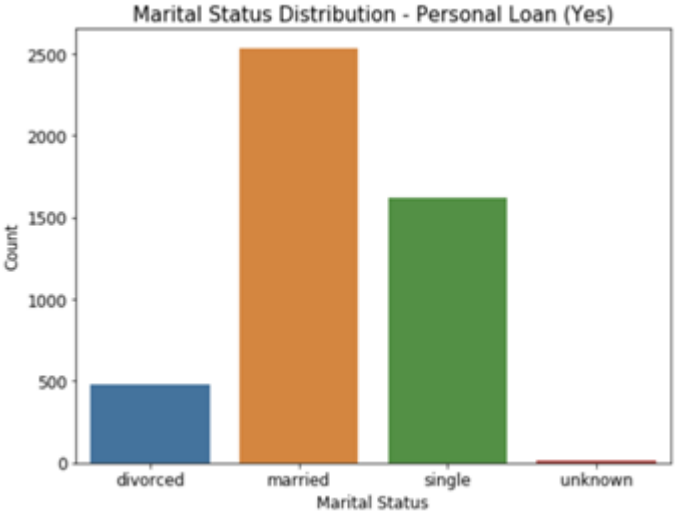

Figure 2. Relationship between marital status and subscribe probability.

Secondly, the relationship between the likelihood of subscribing to long-term deposits and family status is also shown in Figure 2 generated by using logistic regression. According to the images, people are more willing to make long-term deposits when they are married. In addition, people are less likely to make long-term deposits when they are single. In particular, divorced people are the least likely to make long-term deposits. Also, according to the study, people are happier and healthier after marriage [10]. This means that people are more likely to make long-term deposits when they are happier. Conversely, people are not as happy when they are single as they are when they are married, so they are

DOI: 10.54254/2755-2721/20/20231088

less likely to choose long-term savings among single people. Finally, people are least likely to make long-term bank deposits in the case of a sad divorce. This also accurately reflects the data in the image. Married people are more likely to choose long-term deposits than single people than divorced people. Therefore, this image can be derived from the relationship between people's daily happiness and the likelihood of making long-term deposits. The higher the daily happiness people obtain, the more likely they are to participate in long-term deposits with banks.

3.2. Evaluation results

After using the logistic regression to prediction, its accuracy also can be determined by the result of precision, recall, and f1-score. Results in Table 1 and Figure 3 are the numbers that average precision number is equal to 0.89; average recall number is equal to 0.90; average f1-score is equal to 0.89.

Table 1. Evaluation results.

	Precision	Recall	F1-score	Support
0	0.92	0.98	0.95	7073
1	0.68	0.34	0.45	927
Average/ total	0.89	0.90	0.89	8000

Firstly, when the value of precision is high, it means the false positive error is small. In this model, the precision is 0.89 which is a high value reflecting this logistic regression model has a high accuracy. Secondly, when the value of recall is high, it means less false negative errors. In this model, the recall is 0.90 which is a high value showing that this logistic model has less errors. Thirdly, when the value of f1-score is close to 1, it means that the overall performance is good. In this model, the f1-score is 0.89 which is very close to 1 meaning that this logistic model has a very good performance which also means a high accuracy. Therefore, this logistic regression model can be said to be more accurate in predicting the long-term deposits of bank customers.

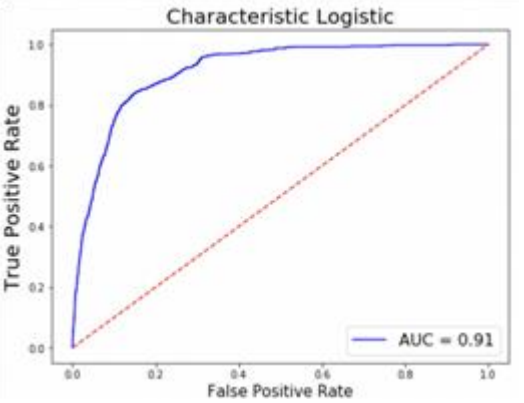

Figure 3. AUC value of prediction.

4. Discussion

After analyzing the accuracy of the results using precision, recall, and f1-score, these three numbers clearly reflect the success of this prediction. All values of precision, recall, and f1-score demonstrate the high accuracy for this logistic regression model. Therefore, this shows that the logistic regression model used in this case is very useful and accurate in predicting the probability of long-term deposit customers of the bank. Also, an inaccuracy appears here, that is, in the case of 0. This is because in case 0, the whole model has very low precision, recall, and f1-score, which means that the accuracy of case 0 among the two cases is low despite the overall high accuracy of the model. Therefore, the prediction of this model needs a larger database to support and remove outliers. If the long-term deposit probability is expected for prediction for the whole country or the whole world, the participation data should be

DOI: 10.54254/2755-2721/20/20231088

collected from more banks or even from more banks in different countries. This would not lead to the overall forecast being too limited. Therefore, ensuring the diversity of data is very helpful for the accuracy of the forecast results.

5. Conclusion

Based on the resulting values of precision, recall, and f1-score, the logistic regression model is very suitable for predicting the long-term deposits of banks. This is because the logistic regression model is well suited to analyze the variables (age, occupation, marital status, loans, etc.) for two different scenarios (customers opting to participate in time deposits and customers not participating in time deposits). Thus, it can be concluded from the analysis that age has a positive relationship with the probability of participating in time deposits until the age of 45 and a negative relationship after the age of 45. Alternatively, the daily happiness index of an individual is positively related to the probability of participation in fixed deposits. These are the key variables that predict customer participation in time deposits. Of course, data from just one bank is not convincing. Therefore, in future logistic regression model predictions, the likelihood of long-term deposits in global banks can be predicted by collecting different banks from more countries. Because such huge data can better reduce the influence of outlier on the data results as well as expand the variables to be more diversified. This way the forecasts based on big data will be more representative. It also provides a more comprehensive analysis of banks and the global economic system.

References

[1] Gramley, L. E., & Chase Jr, S. B.: Time deposits in monetary analysis. Fed. Res. Bull., 51, 1380 (1965).

[2] Artavanis, N., Paravisini, D., Robles-Garcia, C., Seru, A., & Tsoutsoura, M.: Deposit withdrawals. Unpublished working paper.1-61 (2019).

[3] Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., & Livingood, W.: A review of machine learning in building load prediction. Applied Energy, 285, 116452 (2021).

[4] Connelly, L.: Logistic regression. Medsurg Nursing, 29(5), 353-354 (2020).

[5] Azen, R., & Traxel, N.: Using dominance analysis to determine predictor importance in logistic regression. Journal of Educational and Behavioral Statistics, 34(3), 319-347 (2009).

[6] Yacouby, R., & Axman, D.: Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems, 79-91 (2020).

[7] Chicco, D., & Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13 (2020).

[8] Derczynski, L.: Complementarity, F-score, and NLP Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, 261-266 (2016).

[9] Deshpande, M., Fadlon, I., & Gray, C.: How sticky is retirement behavior in the US? Responses to changes in the full retirement age. National Bureau of Economic Research, (No. w27190 (2020).

[10] Rosen - Grandon, J. R., Myers, J. E., & Hattie, J. A.: The relationship between marital characteristics, marital interaction processes, and marital satisfaction. Journal of counseling & Development, 82(1), 58-68 (2004).