# Apply Logistic Model to Analyze the Odds of Default

Hairong Sun 1004116618

09/12/2020

**Abstract**

Default, failing the obligation to repay a loan is the major issue in the banking system. It indicates whether a bank could receive back the money on time and in full. This is may influence the regular operation of a bank and the financial system. Therefore, an indicator that predict a borrower whether would default is necessary. Bank and financial institutions can decide whether to loan based on this indicator to avoid future loss. This analysis used the survey data collected in 2008 to 2010 of a Portuguese bank (Citation 1). The prediction is associated with predictors such as age, job type, and loan etc. After selecting the better model, the result finds that default possibility is negatively related to age and bank balance.

**key words:** Default, Banking, Logistic Regression model, Stepwise AIC, Prediction

**Code and data supporting this analysis is available at: here (click "here")**

## Introduction

In the financial world, default is a failure to fulfill an obligation, especially to repay a loan. To be more specific, the borrower is unable to repay the debt on time or in full. Default is the risk that financial system trying to avoid and minimize. The goal of the analysis is to find the probability of default and testing how strong these variables are related to the testing goal. Investopedia states that the default rate is the highest based on S&P index which is about 3.28% (Citation 2). It is clear that the lower the risk of default is, the safer the financial environment is. Therefore, it is essential to introduce a tool to examine whether a person would default or not. In other word, whether the borrower would return the borrowed money on time and in full.

This analysis could provide financial sectors to strengthen the default assumption. In this case, the financial sector would lend or not lend to an applicant according to this indicator. Before building a model, this analysis is mainly using the dataset from a marketing campaign by a Portuguese banking institution from May 2008 to November 2010. The researching goal is finding the individual's probability of default, by using attributes such as age, job, marital, bank account balance, housing, and loan.

Before doing the analysis, assuming that the more stable marital, job, and housing situation is, the less likely a person would default. This analysis is going to use logistic regression to predict the log odds of default, and how strong each variable related to the interest of this report. Results are presented in the mathematical expression and graphs. In the discussion part, the next step and limitations indicates that expanding data collection is necessary.

## Data

The datasets are obtained from Kaggle website. The data is collected by phone calls based on a marketing campaign of a Portuguese banking institution from May 2008 to November 2010 (Citation 1). The data is helpful for this analysis because it contains related information such as age, job, education, loan, default, etc. The interest of this analysis is driven by the dataset because the initial assumption is that there is a strong relationship between probability of default and factors such as age, education, bank balance, etc. The benefit of collecting data via phone is that it saves money and time while collecting large amount of data. However, there are also drawbacks such as people not welling to provide some private information such as job type and education. Therefore, there are some "unknown" existing in datasets. By cleaning the data, observations with "unknown" input have been removed.

The datasets are "train.csv" and "test.csv" respectively. There are 45,211 observations and 17 variables in the train dataset, and 4521 observations and 17 variables in the test dataset. In order to prepare the data for the analysis, which is whether a person would default, attributes such as contact, day, month, etc. (9 in total) are removed due to irrelevance. These removed factors are more related to the marketing strategy. Moreover, the binary variable has been changed from yes/no format to 1/0. 1 stands for yes while 0 stands for no. This step is for creating the logistic regression in the modelling step.

Here is Table 1 that contains the basic information of the cleaned datasets. The attributes age (in years, min=18, max=95) and balance (yearly bank balance in Euro, min=-8020, max=102000) are numerical variables. Factors job (job types), marital (marital status), education (education level), default (whether has credit in default, yes/no), housing (whether has housing loan, yes/no), loan (whether has personal loan, yes/no) are categorical variables.

**Table 1: Baseline Characteristics of 2010 a Portuguese Bank Marketing Survey**

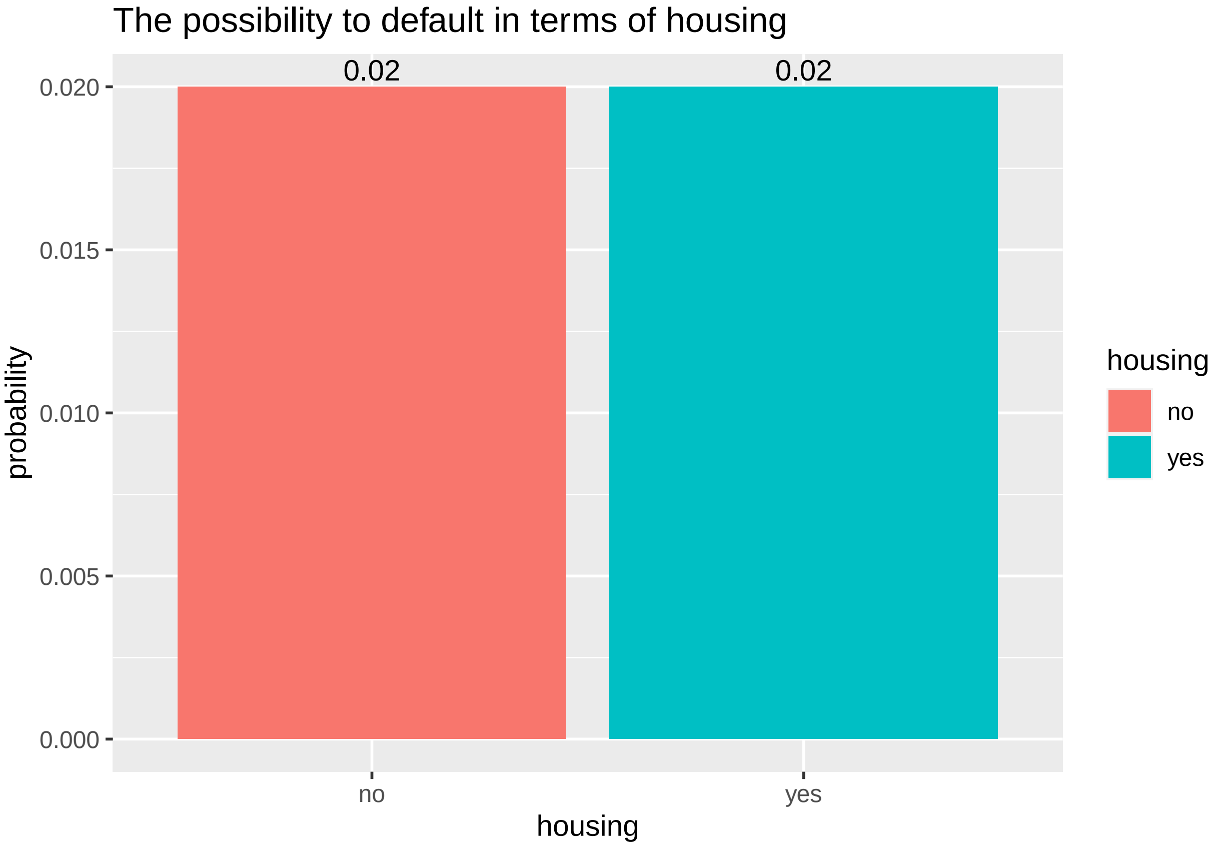
```
##                            Stratified by group
##                             Testing data        Training data       p      test
##   n                            4,311              43,193
##   age (mean (SD))              40.97 (10.50)       40.76 (10.51)     0.219
##   job (%)                                                            0.326
##      admin.                      461 (10.7)         5000 ( 11.6)
##      blue-collar                 905 (21.0)         9278 ( 21.5)
##      entrepreneur                157 ( 3.6)         1411 (  3.3)
##      housemaid                   107 ( 2.5)         1195 (  2.8)
##      management                  942 (21.9)         9216 ( 21.3)
##      retired                     216 ( 5.0)         2145 (  5.0)
##      self-employed               179 ( 4.2)         1540 (  3.6)
##      services                    404 ( 9.4)         4004 (  9.3)
##      student                      68 ( 1.6)          775 (  1.8)
##      technician                  746 (17.3)         7355 ( 17.0)
##      unemployed                  126 ( 2.9)         1274 (  2.9)
##   marital (%)                                                        0.045
##      divorced                    503 (11.7)         5028 ( 11.6)
##      married                    2664 (61.8)        25946 ( 60.1)
##      single                     1144 (26.5)        12219 ( 28.3)
##   education (%)                                                      0.837
##      primary                     671 (15.6)         6800 ( 15.7)
##      secondary                  2298 (53.3)        23131 ( 53.6)
##      tertiary                   1342 (31.1)        13262 ( 30.7)
##   balance (mean (SD))       1,410.66 (3,015.65) 1,354.03 (3,042.10)  0.243
##   housing = yes (%)             2475 (57.4)        24292 ( 56.2)     0.144
##   loan = yes (%)                 683 (15.8)         7107 ( 16.5)     0.312
##   group = Training data (%)        0 ( 0.0)        43193 (100.0)    <0.001
##   default (mean (SD))           0.02 (0.13)         0.02 (0.13)      0.581
```

**The dataset is acquired from Kaggle (click "Kaggle")**

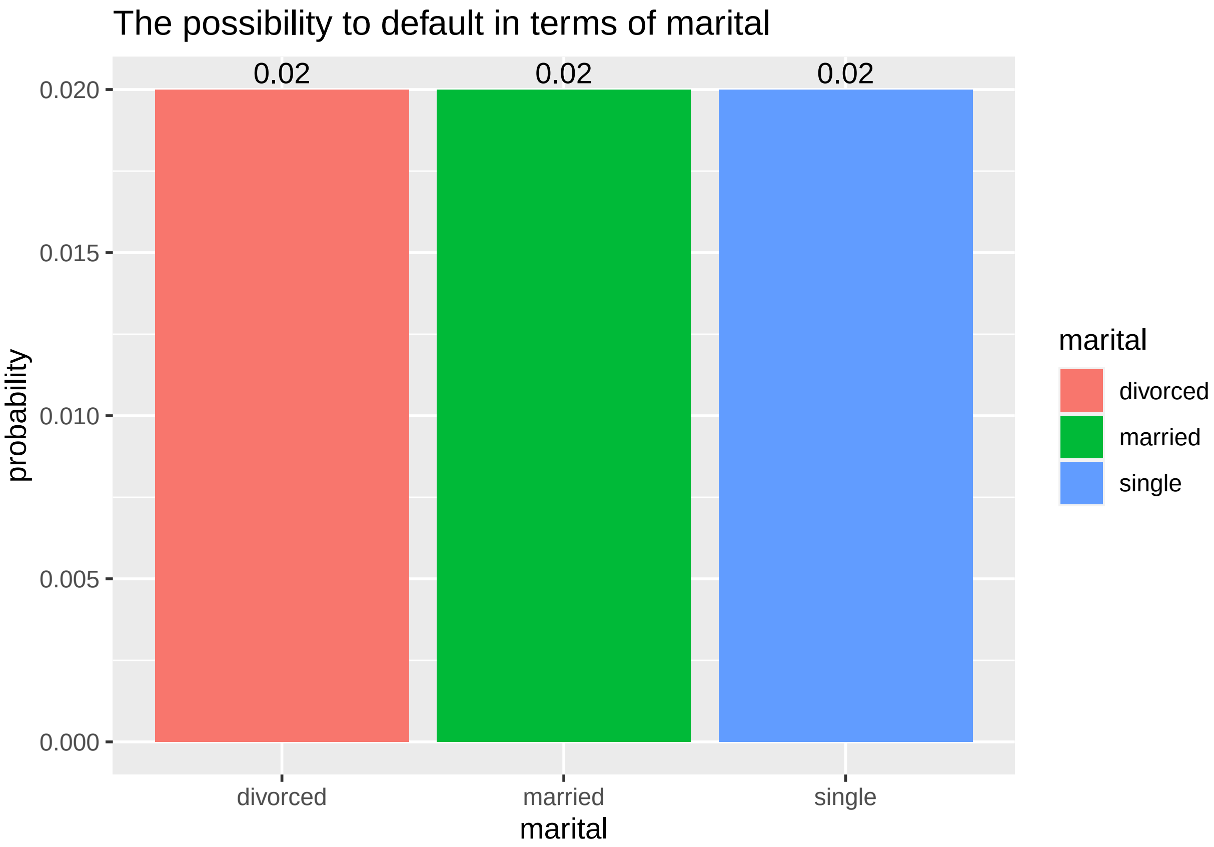
### Model

The logistic regression model is built by using R Studio. In the modelling process, there are three models created in total and one only chosen after comparing the AICs. The models other than the full model are created based on the probability plot.

This analysis used all of the attributes in the cleaned training dataset because these factors all contribute to the probability of default. The first step is to create a full model that contains all the attributes (age, job, marital, education, balance, housing, loan). Before creating more models, two figures of possibility to default have been drawn. The first one is in terms of housing factor. This graph indicates that the probability to default for people are about the same (2%) regardless of whether the borrower is holding a housing loan. Therefore, the second model contains factors age, job, marital, education, balance, and loan (full model removed housing factor).



The second one is in terms of marital factor. Similarly, the graph shows that the possibility to default is the same (2%) for people regardless of their marital status (divorced, married, single). Therefore, the third model is built by factors age, job, education, balance, housing, and loan. (full model removed marital factor).



**Stepwise AIC:** Comparing the AIC of these three models which are full model, full model removed housing factor, and full model removed marital factor. The AIC of the full model is 6439.98. The AIC of the model that removed housing factor is 6480.96. The AIC of the model that removed marital factor is 6450.09. 6439.98 < 6450.09 < 6480.96. The smaller the AIC is, the better the model is. Therefore, the full model is chosen for this analysis and been applied to later testing process.

Based on the selected logistic regression model. The mathematical expression is shown below:

## Table 2: Statistics of the full model

| Coefficients | Estimated coefficient | P-value |
|---|---|---|
| (Intercept) | -2.856 | < 2e-16 |
| age | -0.01407 | 0.003105 |
| job blue-collar | 0.5028 | 0.000865 |
| job entrepreneur | 1.232 | 6.14e-10 |
| job housemaid | 0.4568 | 0.085472 |
| job management | 0.6783 | 7.22e-05 |
| job retired | 0.3658 | 0.153793 |
| job self-employed | 0.7901 | 0.000574 |
| job services | 0.1819 | 0.296499 |
| job student | -1.063 | 0.075058 |
| job technician | 0.2499 | 0.107438 |
| job unemployed | 0.8958 | 9.11e-05 |
| marital married | -0.3645 | 0.000862 |
| marital single | -0.1378 | 0.280646 |
| education secondary | 0.002047 | 0.985812 |

| Coefficients | Estimated coefficient | P-value |
|---|---|---|
| education tertiary | -0.4344 | 0.004259 |
| balance | -0.002222 | < 2e-16 |
| housing yes | -0.5001 | 2.70e-10 |
| loan yes | 0.6791 | < 2e-16 |

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 age + \beta_2 job\ bluecollar + ... + \beta_{12} marital\ married + ... + \beta_{14} education\ secondary + ... + \beta_{18} loan\ yes$$

P is the probability to default, and (p/1-p) is the odds to default for the clients in the Portuguese banking institution in 2008-2010. Factors job blue collar, job entrepreneur, . . . , job unemployed, marital married, marital single, education secondary, and education tertiary are dummy variables. $\beta_0$ to $\beta_{18}$ are coefficients and associate with each variable. There are dummy variables for job types, marital status, and education level. For example, the value of $\beta_1$4 (0.002047) represents the change in the log odds of default when the education level for an observation is secondary vs primary and tertiary. The dummy variables could only be 1 or 0. For example, if the education of an observation is secondary, then education secondary in this case should be 1. Other education types should be 0 (Citation 3).

For the intercept, it means that when job = admin., marital = divorced, education =primary, housing = no, loan = no, the log odds of default is -2.856. Remain other conditions the same, when the job changed from admin. to entrepreneur, the log odds of default would increase by 1.232. It is the same rule with other dummy variables.

The accuracy of the testing process is about 0.9835. The confusion matrix below summarized and evaluated the predicted results on the effectiveness and accuracy of the operation. Also, it helps to check the true positive and true negative rate of the model. In this matrix, 0 represents no (would not default) and 1 represents yes (would default).

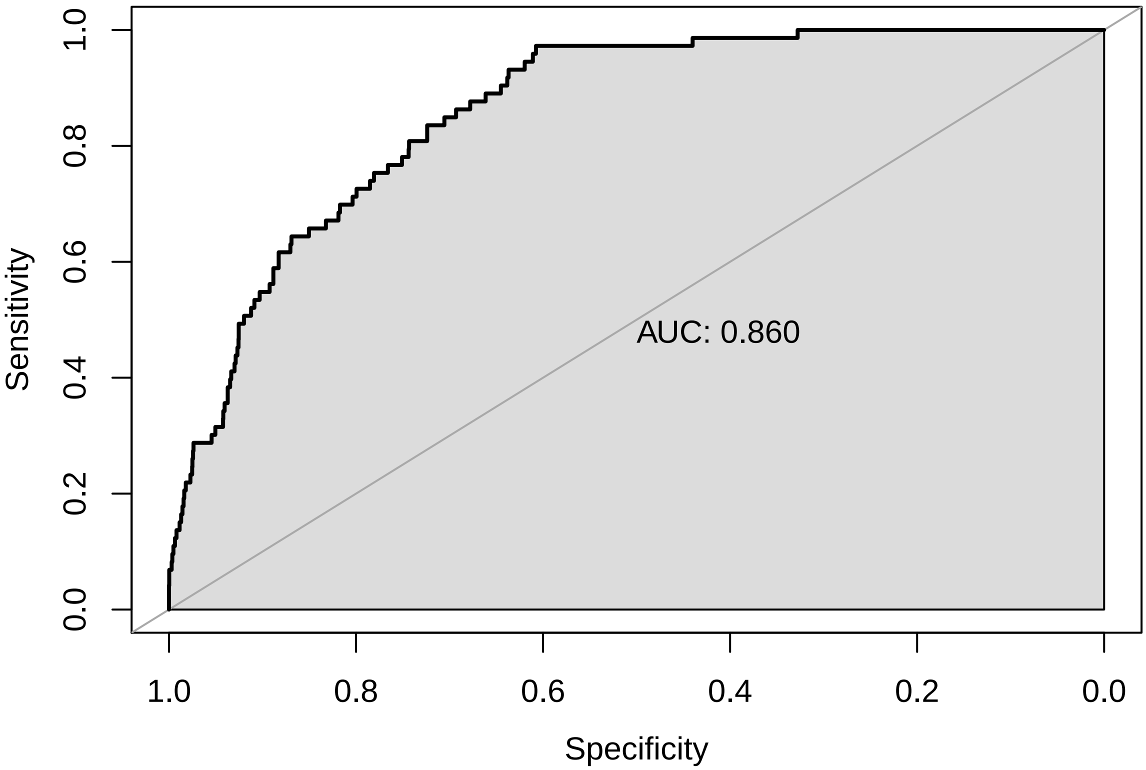
There are 4237 true positives (predicted 4237 cases of not default correctly), 70 false positives (predicted 70 default case wrongly to not default), 1 false negative (predicted 1 not default case wrongly to default), and 3 true negatives (predicted 3 default cases correctly).

The true positive rate is: 4237/4307 (0.9837), the true negative rate is: 3/4 (0.75). Therefore, the model performs well when predicting the non-default cases while perform not that well when predicting default cases.

## Table 3: Confusion Matrix of testing result

| | Reference | |
|---|---|---|
| Prediction | 0 | 1 |
| 0 | 4237 | 70 |
| 1 | 1 | 3 |

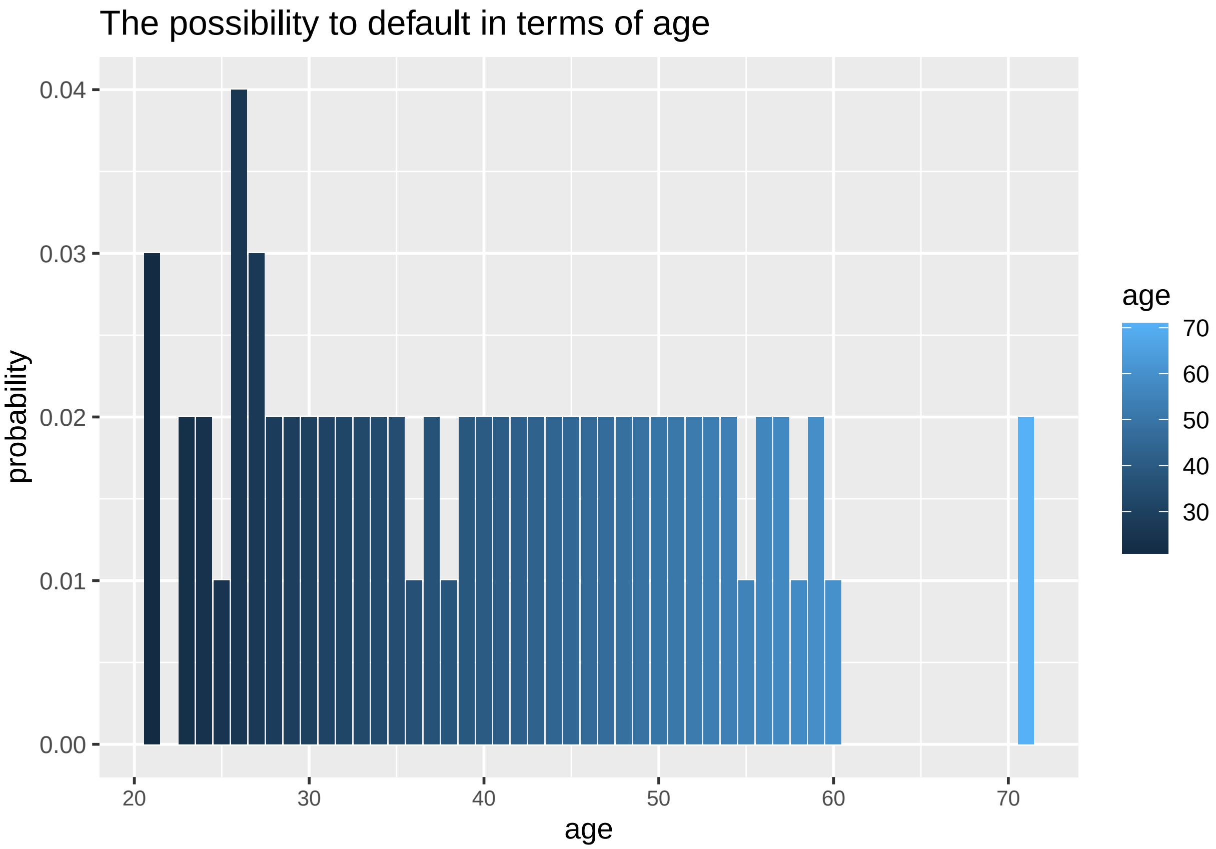
```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.8605
```

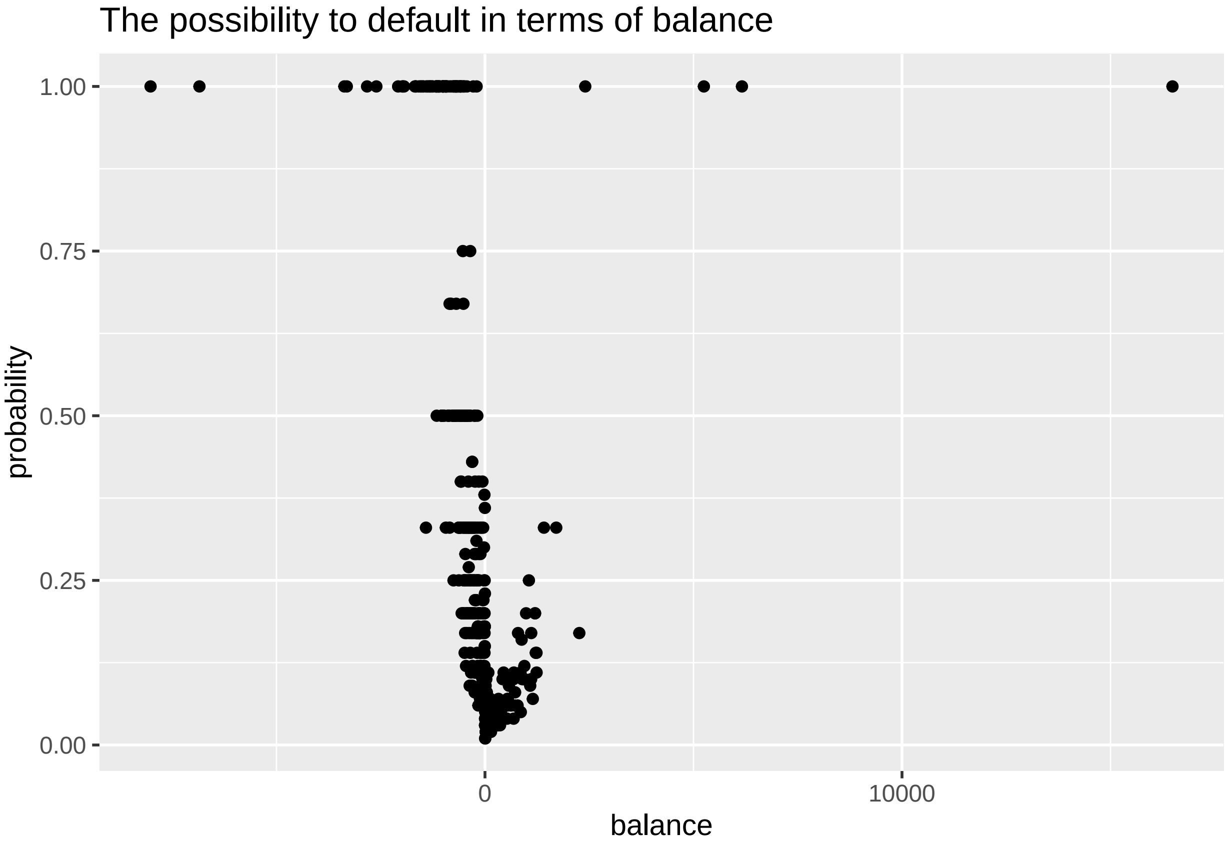


**Explanation of the ROC curve** ROC (Receiver Operating Characteristics) curve showing the sensitivity and Specificity (1 on the left and 0 on the right). It means that 1 subtracting the false positive rate. The AUC is 0.86 (max = 1), it means that the result is 86% credible and it is more accurate than the random guessing. Therefore, this model is performing well.

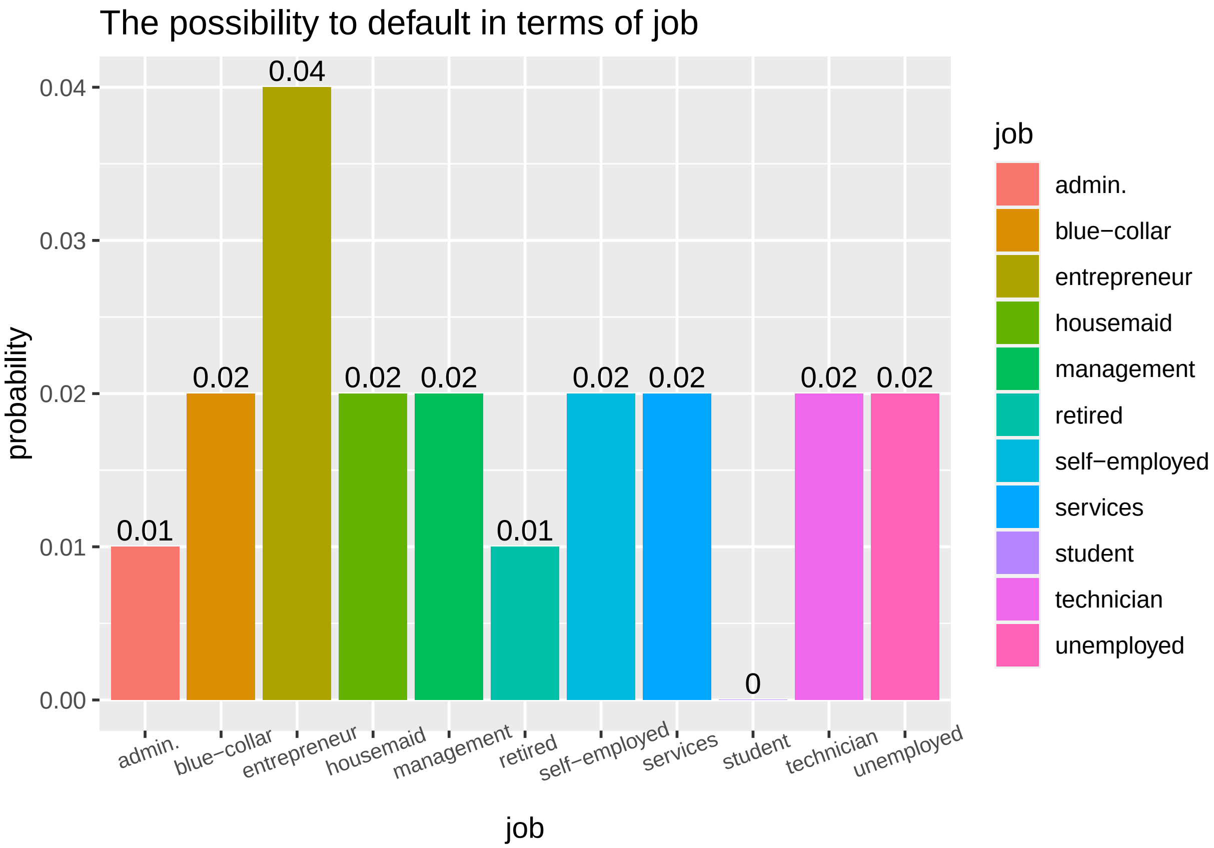
## Results

Here are the relationships between the probability to default and each variable in the model respectively. In this graph, it shows the relationship between possibility to default and age. Clearly, when the age is between 20 to 30, there are more possibilities to default. However, the possibility of default is low after age 55 compared to other age ranges. The reason might be that when a person is young and just begin to accumulate wealth. Once they failed to return the credit, they have no spare money for it. The only result is default. When the age is higher, the wealth is accumulated to an extent. Some people would use other wealth to repay the borrowed credit, therefore, the possibility to default is low for senior people.

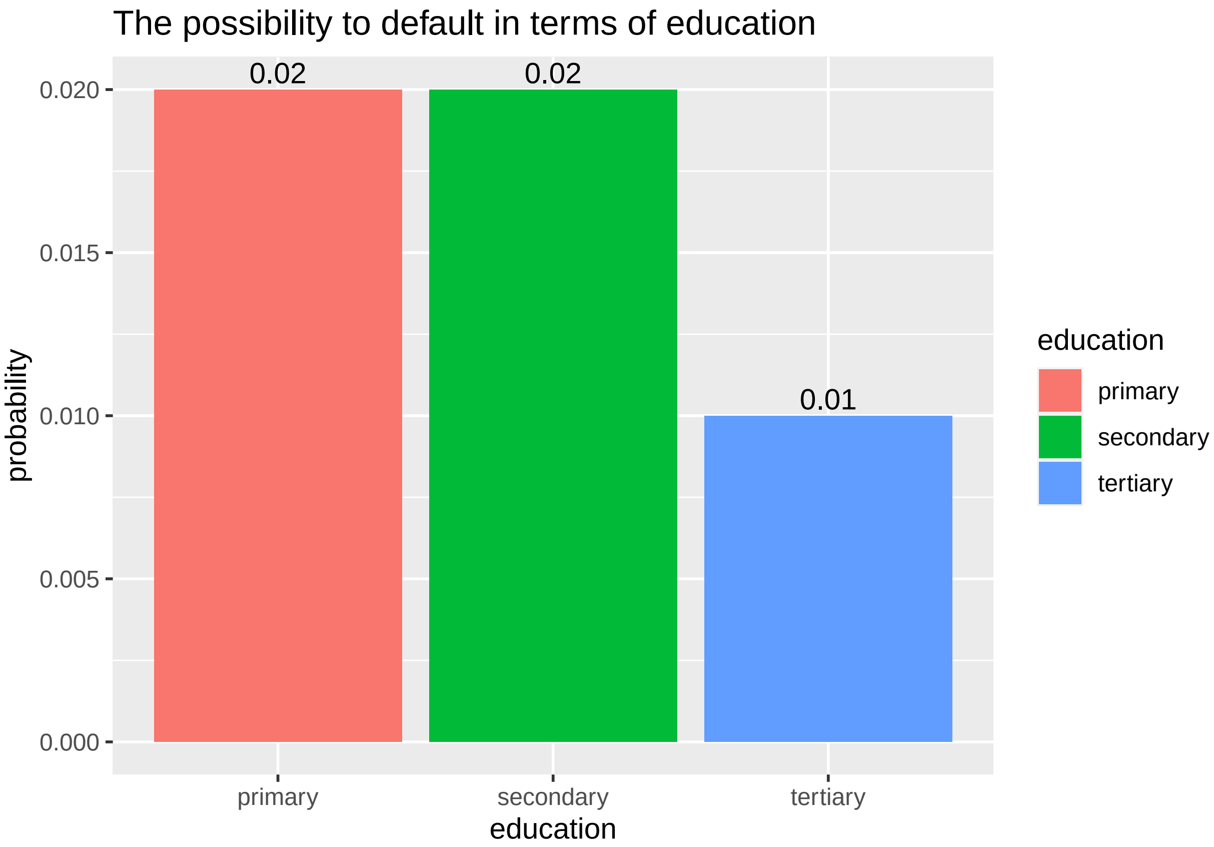

The possibility to default in terms of age


The possibility to default in terms of balance

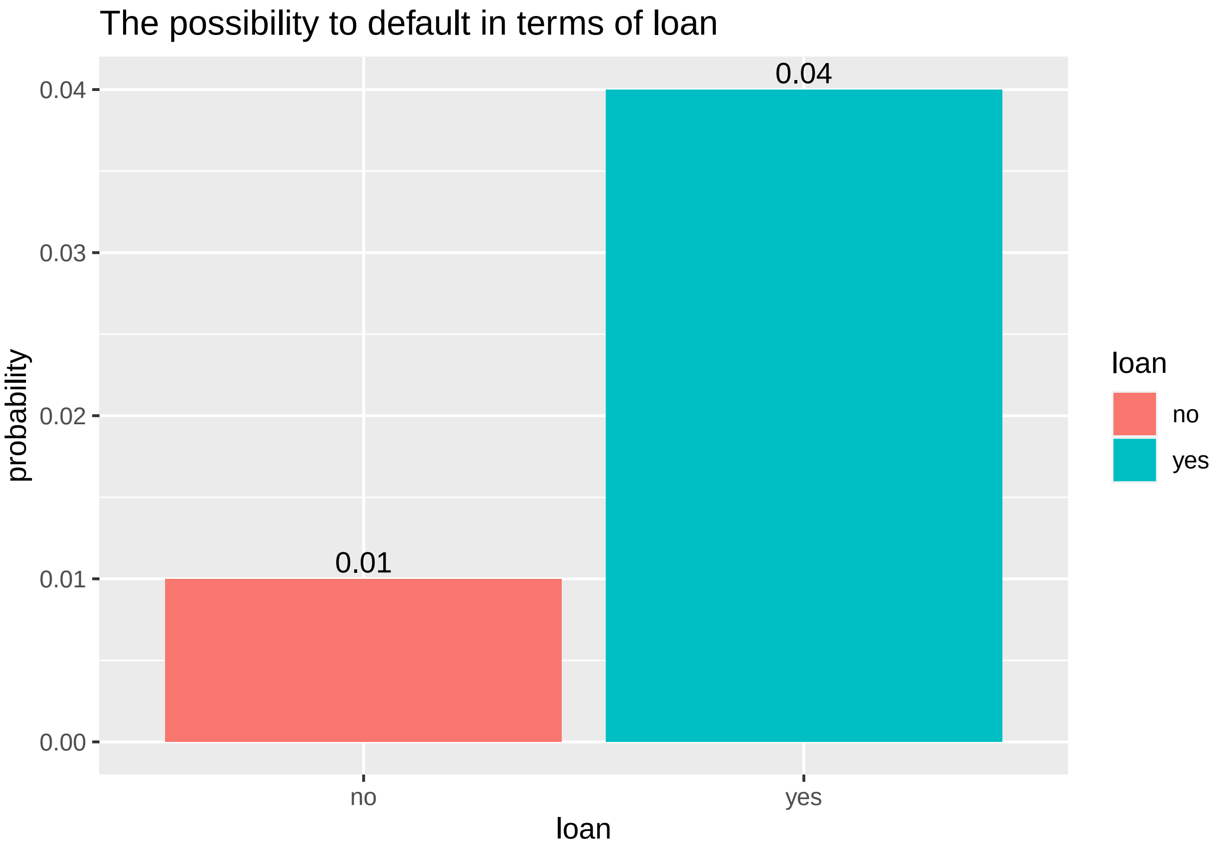
In this scatter plot, it shows the relationship between possibility to default and remaining balance. It is not hard to see that when the balance is negative, the possibility of default is high. When the account balance is greater than 0, the less possibility of default is comparing to those observations that have negative balance.



In this graph, it shows the relationship between possibility to default and job. The highest possibility of default is entrepreneur (0.04) while the lowest is student (0). It is reasonable because the job type of entrepreneur bears more risks naturally. There are more gain and loss in this job type. However, for students, their costs are mainly come from their parents. Therefore, students would not lend much, which caused low possibility of default.



In this graph, it shows the relationship between possibility to default and education. It shows that people who received tertiary education have lower possibility of default (0.01) compare to the possibility of observations that have primary and secondary education, which is 0.02.



In this graph, it shows the relationship between possibility to default and loan. It shows that people who have personal loan usually have higher default possibility (0.04) compare to the default possibility of people that have no personal loan (0.01). It implies that the more a person loan, the higher the default possibility is.

## Summary

This analysis trying to build a logistic regression model to predict the log odds of default. To build a final model, there are other 2 models built based on the plots that describe the relationship between possibility of default and housing factor and marital factor respectively to run a stepwise AIC. Comparing the AIC of three models, it shows that the AIC of the full model which contains all the variables in the data is the lowest (6439.98). This implies the full model is the best choice of predicting the log odds to default. Then applying the model concluded from the training data to the testing data. The result is shown by the confusion matrix, the ROC curve, and AUC. According to the confusion matrix, both true positive rate (0.9837) and true negative rate (0.75) are high enough. Moreover, the AUC is 0.86, which proofs that the model predicted most of the default conditions in the test dataset correctly. In the result part, there are more analysis according to the relationship between possibilities of default and each attribute respectively.

## Conclusion

To conclude, the accuracy of the model that predict the log odds of default is 0.9835, which can predict most of the observations correctly based on information such as age, job type, marital status, education, etc. It means that this model could be used in the real-life cases when individuals applying for loan. It can help to determine the possibility of default of a person. According to Investopedia's explanation, the bank credit card default rate is the highest based on S&P index which is about 3.28% (Citation 1). It is a damage to both lenders and borrowers. The lenders would suffer the loss in money while the borrower suffer the loss of credibility. By using the model, about 98% of the money would be loan to the credible person based on the 2008 -2010 Portuguese banking data.

There are also some interesting findings. People who are in the age range of $20 - 30$, bank account balance $< 0$, entrepreneurs have higher possibility of default. People who received tertiary education have lower possibility to default.

## Discussion (Weaknesses & Next steps)

### Weaknesses

Based on the data itself and the former analysis, there are three weaknesses should be drawn. The first one is that the data obtained from the Kaggle website is out of date. It documented the survey data of a Portuguese banking institution from May 2008 to November 2010 and released in 2012. However, now it is 10 years after the data collection. After the recession caused by the US financial crisis in 2008, world financial market has suffered from it. This is the time when the data was collected (Citation 5). In contrast, in recent years, the world economy witnessed a rise and also keep meeting new challenges such as the shock brought by the COVID-19. Therefore, the financial world has changed, the habit of lending money of people would also change. Thus, due to the different financial context of different years, the model predicted based on the data collected in 2013 is not sufficient to predict nowadays investors borrowing habit and default behavior.

Secondly, the dataset collected data only from one bank in one country. Therefore, the model can only interpret the default habit of the client there. Doubts remain as to whether the analytical model can be applied to other countries. The default habit would be different when changing a country or even a bank. Therefore, the geographical limitation caused the concluded model not inclusive.

The third weakness is based on the model. Even though the accuracy of this model is about 0.9835, which seems like close to a perfect model. However, take a close look of the original training data, there are 98.2% of the sample that would not default and there are 1.8% of the sample would default. The amount of data would default is small compare to the data that would not default. Moreover, after applying the model to the test data, by looking at the simulated outcome, it is obvious that the model predicted most of the people who would not default. The model does not perform well in terms of predict the people who would default. The key of ensuring a safe financial environment is to distinguish those borrowers who would default and stop lending them money. In this case, this model failed to achieve this goal.

### Next steps

To solve the problem mentioned in the weakness section, expanding the data collection is the first step to make. To be more specific, collecting data from more recent years and more banks in different countries could make the model be more persuasive. Secondly, collecting large amount of data that people would default is essential to train the default model. When the accuracy of predicting the people who would default increases, then the model could help to minimize the loss of a bank or any financial institutions.

## References

1. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
2. Kagan, J. (2020, December 14). Default Rate Definition. Retrieved December 23, 2020, from https://www.investopedia.com/terms/d/defaultrate.asp
3. Sun, H., & Zhang, J. (2020). Apply logistic regression Model to analyze the number of children someone has (pp. 1-14, Rep.).
4. Ggplot2 barplots : Quick start guide - R software and data visualization. (n.d.). Retrieved December 23, 2020, from http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization
5. Financial crisis of 2007–2008. (2020, December 22). Retrieved December 23, 2020, from https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008

6. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

7. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

8. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

9. Kazuki Yoshida and Alexander Bartel (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone