Predicting Banking Marketing Data Analysis and Preparation

Akshay Kumar Reddy

April 2023

Contents

1 Data and goals **2**

2 Datasets **2**

3 Results **6**

4 Conclusion **6**

5 References **10**

1 Data and goals

In this project, we study different approaches to predict the success of bank telemarketing. As an instrument, we have a dataset related to direct marketing campaigns based on phone calls from a Portuguese banking institution. Often, more than one contact with the same client was required, in order to assess if the product (bank term deposit) would be (yes) or not (no) subscribed.

The data under study here is called Bank Marketing Dataset (BMD) and it was found in the Machine Learning Repository (UCI). The data is publicly available in the URL https://archive. ics.uci.edu/ml/datasets/Bank+Marketing. The dataset size is considerably large, especially if we consider its origin. Data from clients of financial institutions are usually difficult to find, and when found, are rarely available in this quantity. In the BMD data, we have 41188 observations, with eighteen features.

The eighteen features are briefly described in Table 1, where in the left column we have the original feature name in the dataset, and in the right column its description, also mentioning if the feature is numeric, categorial, and with how many levels (if categorical, of course). The first one called y is the response, the desired target. The other features are presented in the same order that they appear in the dataset.

To know better the data some descriptive analysis is performed, see Figure 1 and Figure 2.

Feature	Description
y	desired target. has the client subscribed a term deposit? (no, yes)
age	numeric
job	type of job, twelve categories
marital	marital status, four categories
eduacation	eight categories
housing	has housing loan? (no, yes, unknown)
loan	has personal loan? (no, yes, unknown)
contact	contact communication type (cellular, telephone)
month	last contact month of year (twelve levels, months)
day.of.week	last contact day of the week (five levels, days)
campaign	number of contacts performed during this campaign and for this client
previous	number of contacts performed before this campaign and for this client
poutcome	previous marketing campaign (failure, nonexistent, success)
emp.var.rate	numeric. employment variation rate - quarterly indicator
cons.price.idx	numeric. consumer price index - monthly indicator
cons.conf.idx	numeric. consumer confidence index - monthly indicator
euribor3m	numeric. euribor 3 month rate - daily indicator
nr.employed	numeric. number of employees - quarterly indicator

Figure 1: Table 1: Features description of the Bank Marketing Dataset (BMD).

2 Datasets

This introduction intends to provide an overview of the dataset under analysis. The dataset contains information about individuals and their respective financial status, employment status, and contact information. Specifically, it consists

Figure 2: Scatterplot lower triangular matrix and correlation upper triangular matrix for all the quantitative features presented in the Bank Marketing Dataset BMD

of 16 variables, with each variable being assigned to a particular category. These categories include age, job, marital status, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, and poutcome.

The age variable denotes the age of the individuals in the dataset. The job variable represents the job title of the individuals, i.e. whether they are unemployed, self-employed, technicians, entrepreneurs, services, blue-collar, or management. The marital status variable indicates whether the individual is married, single, divorced, or widowed. The education variable reflects the highest level of education the individual has achieved. It can be primary, secondary, tertiary, or unknown.

The default variable shows whether or not the individual has credit in default. It can be either yes or no. The balance variable shows the amount of money the individual has in their bank account. The housing variable indicates whether or not the individual has a housing loan. It can be either yes or no. The loan variable indicates whether or not the individual has a personal loan. It can be either yes or no. The contact variable denotes the method of contact between the individual and the financial institution. It can be either cellular or unknown. The day variable shows the day of the month the individual contacted the financial institution. The month variable indicates the month of the year the individual contacted the financial institution. The duration variable denotes the length (in seconds) of the contact between the individual and the financial institution.

The dataset is especially useful for marketers who need to understand the customer base in order to target potential customers and make better investment decisions. The age of the customers in the dataset ranges from 20 to 59, and the

vast majority of customers are between 30 and 40 years old. Most customers are married and are at least educated to a secondary level.

There is an even split between customers who have a default status and those who don't. The balance of the customers ranges from -88 to 9374. Most customers have a loan and housing, although there is a sizable minority who have neither. Of the customers, 86.2 per cent had cellular contact, 6.5 per cent had an unknown contact, and 7.3 percent had a telephone contact. Regarding the month, the majority of customers were contacted in May followed by April and August. The average duration of the contract was about 181.2 seconds. Most customers had a campaign of 1 or 2. The majority of customers had a pdays of -1, indicating that they were not previously contacted. About half of the customers had a previous outcome of unknown.

The rest of the customers had a previous outcome of failure or other. Finally, about 40 of the customers had a y value of yes, indicating that they are likely to subscribe to a term deposit. This indicates that financial institutions should focus their marketing efforts towards this group of customers in order to get the best return on their investment. In conclusion, this dataset provides a comprehensive overview of 2000 customers, including their financial and socio-demographic attributes. The dataset is useful for marketers who need to understand the customer base in order to target potential customers and make better investment decisions Mean of the integrated profile.

In Figure 1 we see the scatterplots and correlations, two-by-two, for all the eight numerical features in the BMD. In more than half of them, we see a random behaviour, that is also described by a correlation close to zero or between the interval -0.3 and 0.3. A (very) strong (and positive) correlation is seen in three cases. emp.var.rate vs. euribor3m (cor. 0.97), euribor3m vs. no. employed (cor. 0.95), and emp.var.rate vs. no. employed (cor. 0.91), i.e., involving only three features - employment variation rate, Euro Interbank Offered Rate (Euribor) and a number of employees. During the analysis, this point can be better studied.

Already in Figure 2, we have the frequencies for each level of the categorical features in the BMD. First, we see that the desired target is unbalanced, with more than 85 per cent of the observations corresponding to clients that didn't subscribe to a term deposit. An equilibrium between levels is only present in the day.of.week last contact feature. By this Figure, we can also see that the last contact of most of the clients was in may (month feature), that most of the clients have a nonexistent previous marketing campaign (poutcome feature), that they are married (marital feature) and that most have a job in the administrative sector.

Using these features, described in Table 1, the goal here is to test several algorithms to see how good they are to predict the desired target, i.e., predicting, given the seventeen features, if the bank term deposit would be or not subscribed. The algorithms used for this task are described in the next section, together with some extra information about the analysis procedure.

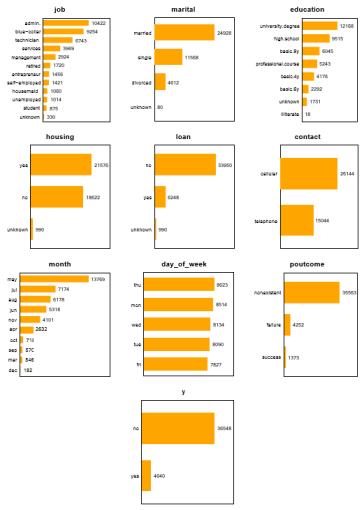
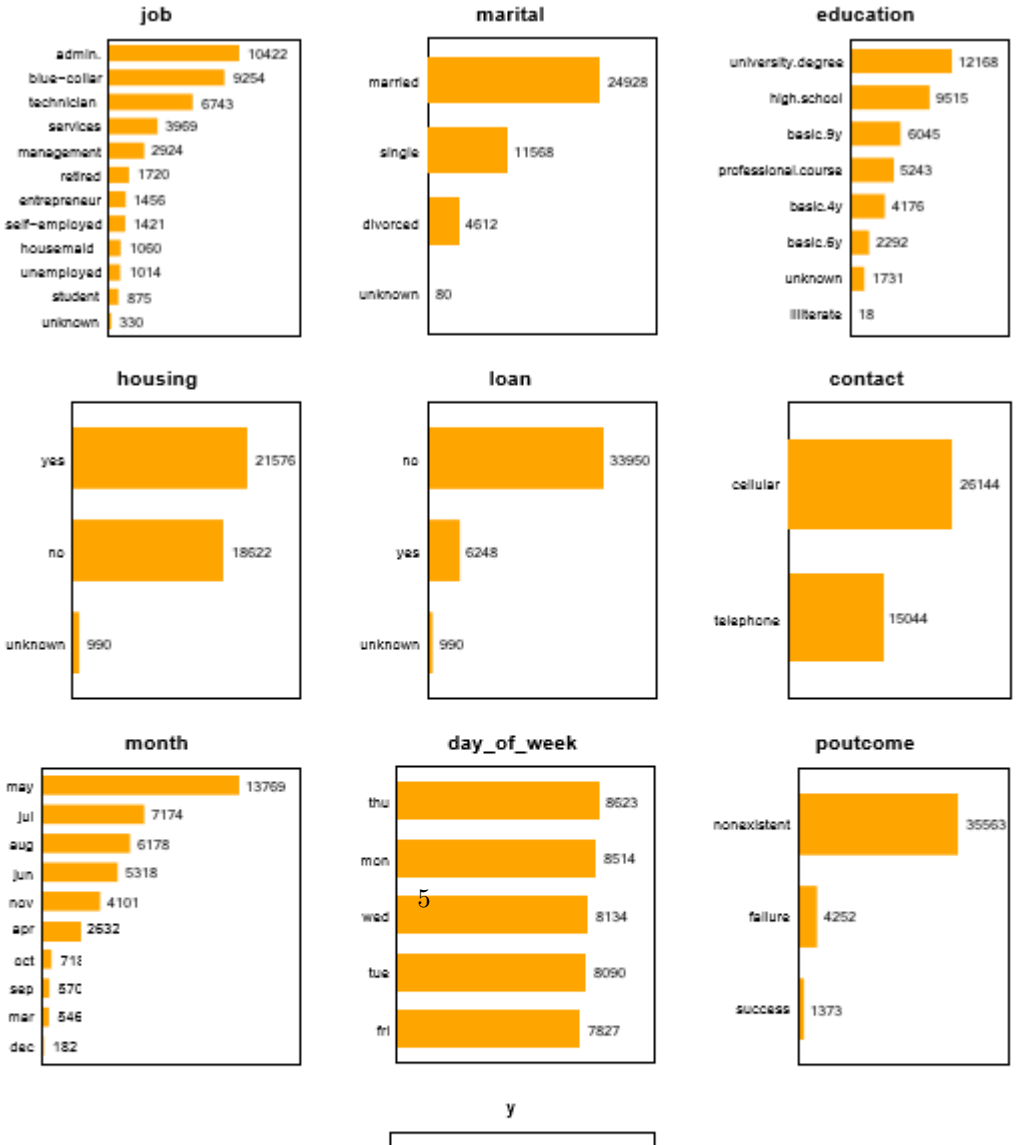


Figure 3: Figure 2: Bar plots for all the qualitative features presented in the Bank Marketing Dataset (BMD).



3 Results

With the GLMs and LM we are able to do feature selection. Here we do this via AIC. Which features are kept and which features are dropped can be seen in Table 2. $The main measure that can be used to compare all the fifteen algorithms is the Receiver Operating Character$ $//en.wikipedia.org/wiki/Receiver_operating_characteristic$.

When dealing with ROC curves the main measure returned is the Area Under the Curve (AUC), that is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The AUC for each model is presented in Figure 3. The highest is obtained with the probit link function in the GLM.

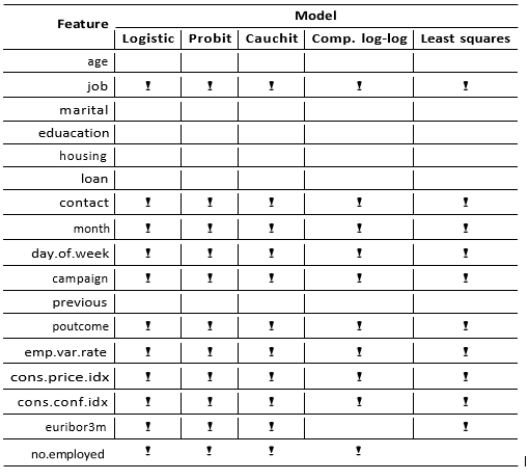
Other measures as specificity, sensitivity and risk, are presented in Table 3. The risk here is defined as the proportion of observations in the test dataset that are wrongly classified by the trained model.

The GLMs and the LM approaches return a probability for each observation, where close to zero means that the observation is more likely to be prominent from the no class - a client that did not subscribe to a term deposit. However, with the ROC curve, we obtain an optimized threshold for this decision. Thus, to compute the risk in this model we use this obtained threshold, instead the default value half - that in the first moment is the logical choice since the returned probability is between zero and one. This optimal threshold is defined as the cutting point that returns the best specificity and sensitivity - in general, we don't want a good specificity value but with a bad sensitivity, or vice-versa. We want the best possible value - combination - for both, at the same time. So the threshold of this scenario is the used value to compute the risk in this model. The other algorithms return directly the class label, not a probability. Again, these values - specificity, sensitivity and the classification risk/error - can be checked in Table 3.

4 Conclusion

Keep a feature means that the feature was significant, statistically significant, in describing the difference between the classes of the desired target - if the bank term deposit would be or not subscribed. In Table 2 we can see a very high concordance between the models, in a general form. Each model finished with eleven, from seventeen, features. This are the dropped, nonsignificant in describing the difference between classes, features in all models: age, marital status, education, housing loan, personal loan, and previous number of contacts performed before this campaign and for this client.

Looking by the AUC, Figure 3, the best model is the GLM with probit link function. However, very similar values are obtained with the others link functions and with the LM. With the other algorithms the AUC's are considerable smaller, but always above 0.55 (a not bad value, but also not so good). Looking to the other computed measures in Table 3, we see that for all the algorithms we obtain a very good specificity, true negative rate, and a bad or not so good sen-



Feature	Model				
	Logistic	Probit	Cauchit	Comp. log-log	Least squares
age					
job	!	!	!	!	!
marital					
eduacation					
housing					
loan					
contact	!	!	!	!	!
month	!	!	!	!	!
day.of.week	!	!	!	!	!
campaign	!	!	!	!	!
previous					
poutcome	!	!	!	!	!
emp.var.rate	!	!	!	!	!
cons.price.idx	!	!	!	!	!
cons.conf.idx	!	!	!	!	!
euribor3m	!	!	!		!
no.employed	!	!	!	!	

Figure 4: Table 2: Remaining features in each model after features selection by AIC

Feature	Model				
	Logistic	Probit	Cauchit	Comp. log-log	Least squares
age					
job	!	!	!	!	!
marital					
eduacation					
housing					
loan					
contact	!	!	!	!	!
month	!	!	!	!	!
day.of.week	!	!	!	!	!
campaign	!	!	!	!	!
previous					
poutcome	!	!	!	!	!
emp.var.rate	!	!	!	!	!
cons.price.idx	!	!	!	!	!
cons.conf.idx	!	!	!	!	!
euribor3m	!	!	!		!
no.employed	!	!	!	!	

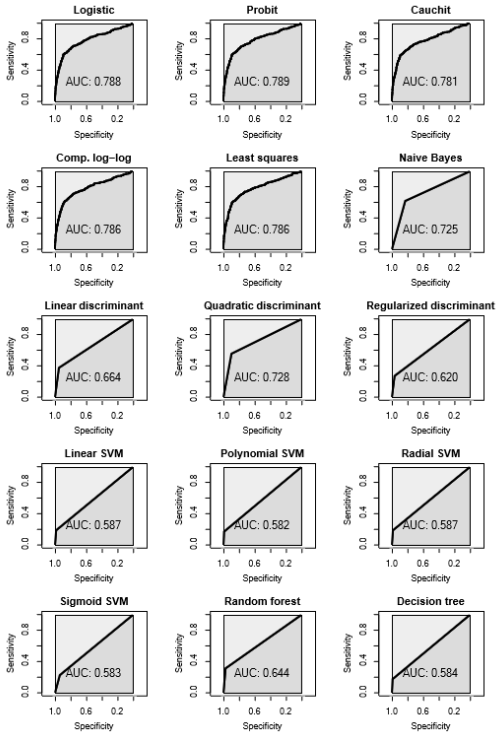
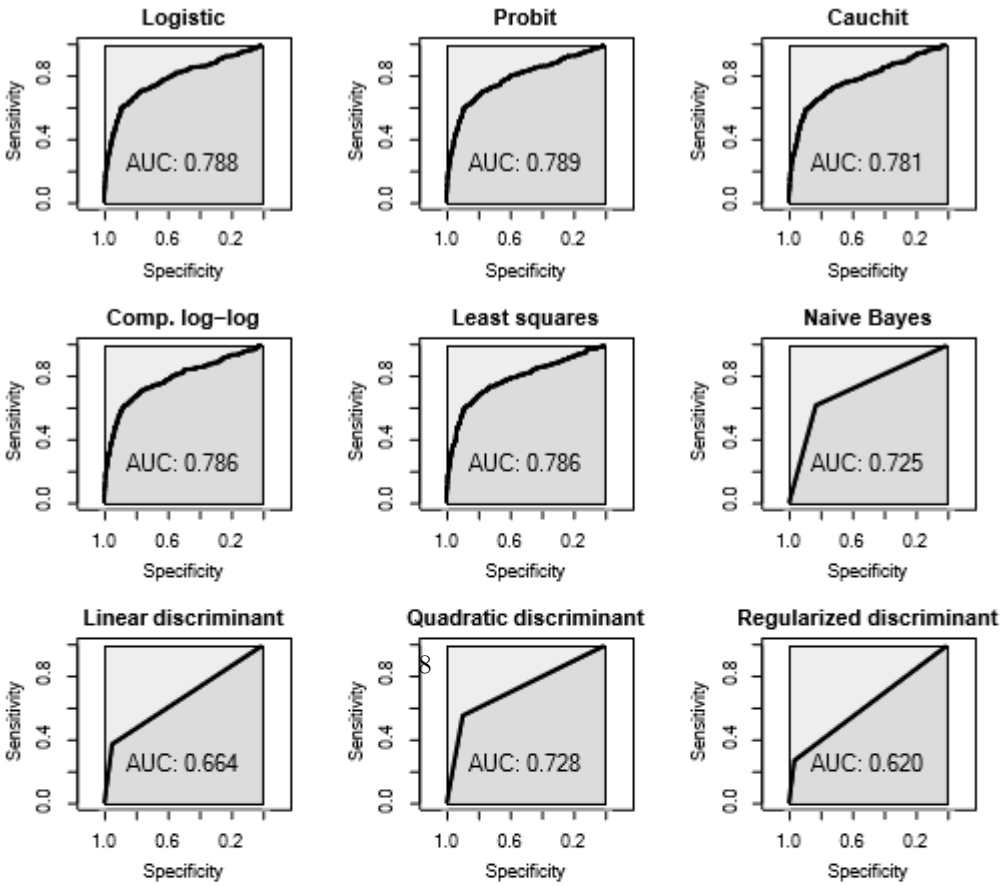


Figure 5: Figure 3: ROC curve for each model (in the test) with respective AUC and thresholds.



Model	Specificity	Sensitivity	Risk
Random baseline classifier	0.891	0.609	0.212
Probit regression (GLM with probit link)	0.888	0.609	0.197
Logistic regression model)	0.893	0.594	0.261
Neural network model (64-32-16-8-1	0.878	**0.614**	0.265
Neural network model (32-16-8-1)	0.886	0.609	0.16
Neural network model (16-8-1)	0.829	0.621	0.192
Neural network model (8-1)	0.950	0.377	0.107
Neural network model (4-1)	0.897	0.558	0.137
neural network model (2-1)	0.966	0.278	0.103

Figure 6: Table 3: Specificity, sensitivity and risk for each fitted model in the test Bank Marketing Dataset (BMD), in bold we have the best performances. The models in bold are the models with the best AUC.

Model	Specificity	Sensitivity	Risk
Random baseline classifier	0.891	0.609	0.212
Probit regression (GLM with probit link)	0.888	0.609	0.197
Logistic regression model)	0.893	0.594	0.261
Neural network model (64-32-16-8-1	0.878	**0.614**	0.265
Neural network model (32-16-8-1)	0.886	0.609	0.16
Neural network model (16-8-1)	0.829	0.621	0.192
Neural network model (8-1)	0.950	0.377	0.107
Neural network model (4-1)	0.897	0.558	0.137
neural network model (2-1)	0.966	0.278	0.103

sitivity, true positive rate. For all the algorithms we have a good risk value, less than 0.30, but a very good risk value is obtained only with the non-GLM/LM techniques.

5 References

[1] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[2] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071.

[3] Venables, W. N. Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

[4] Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.

[5] Liaw, A. Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

[6] Therneau, T., Atkinson, B. and Ripley, B. (2017). rpart: Recursive Partitioning and Re- gression Trees. R package version 4.1-11. https://CRAN.R-project.org/package=rpart.

[7] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, JC. and Mu¨ller, M. (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/