Athens University of Economics and Business

MSc in Business Analytics

Statistics For Business Analytics II

Project I 2021-2022

Bank telemarketing phone calls dataset

Name: *Panagiota Gkourioti*

Student ID: *P2822109*

Professor: *Karlis Dimitrios*

Contents

1. **Introduction – description of the problem**......................................3
2. **Exploratory Data Analysis**..6
3. **Model building & Evaluation**..12
4. **Conclusion**..20
5. **References**..21
6. **Appendix**...22

1. Introduction – description of the problem

Telemarketing is a widely adopted direct advertising approach towards potential customers through telephone communication. Customer targeting consists of identifying prospective buyers of a product or subscribers of a service offered within the context of a marketing campaign. It is beneficial for businesses to constrict the range of potential customers and to explore their characteristics, thus increasing the success rate as well as efficiently reducing the marketing costs.

This marketing technique also enables banks and other financial organizations to focus on those customers who present the largest likelihood of subscribing to their products, offers, and other packages. Most often than not, identifying these group of customers poses a challenge to financial institutions.

The current study focuses on a retail bank that conducted a telemarketing campaign, aiming to persuade customers into subscribing for long-term deposits. Within the campaign, the agents make phone calls to a list of clients to sell the product (outbound) or, if meanwhile the client calls the contact-center for any other reason, he is asked to subscribe the product (inbound). The dataset contains information collected from a retail bank between May 2008 to June 2010 for all the customers who were contacted during a particular year to open term deposit accounts in a total of near 40 thousand phone contacts. CI). The dataset has 20 input variables and 1 output variable as a target, which are described in Table 1.1. They include both numerical and categorical variables, regarding the socio-demographics of the clients, the client-bank relationship, the contact context and the socio-economic macro context.

	Features	**Description**
dependent variable		
	SUBSCRIBED	whether the client subscribed a long-term deposit (yes, no)
personal information		
	age	age in years of the costumer
	job	type of job (admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)
	marital	marital status (divorced, married, single, unknown)
	education	completed education (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
	default	whether the customer has loans in delay (yes, no, unknown)
	housing	whether the customer has a mortgage account (yes, no, unknown)
	loan	whether the customer has a personal credit (yes, no, unknown)
contact information		
	contact	contact communication type (cellular, telephone)
	month	last contact month of year (jan, feb, mar, ..., nov, dec)
	day_of_week	last contact day of the week (mon,tue,wed,thu,fri)
	duration	last contact duration, in seconds
historic information		
	pdays	number of days since the last call for a previous campaign
	previous	total number of past calls
	poutcome	outcome of the previous marketing campaign (failure, nonexistent, success)
socio-economic information		
	emp.var.rate	quarterly data of the employment variation rate
	cons.price.idx	monthly data of the consumer price index
	cons.conf.idx	monthly data of the consumer confidence index
	euribor3m	daily data of the Euribor 3 month rate
	nr.employed	quarterly number of employees in thousands
hindsight information		
	campaign	number of contacts performed during this campaign and for this client

Table 1.1.: Features description

The business objective of the current project is to determine which factors influence the purchase of term deposits and construct a model that can explain the decision of a client to subscribe or not to a term deposit.

The first crucial part of the analysis is transforming the data. This included removing duplicate rows, determining the correct data types for each variable, removing variable pdays, as it provided similar information with poutcome variable regarding people who have never participated in a previous marketing campaign, and grouping education variable into three levels (basic, intermediate, advanced) for better interpretation of the model.

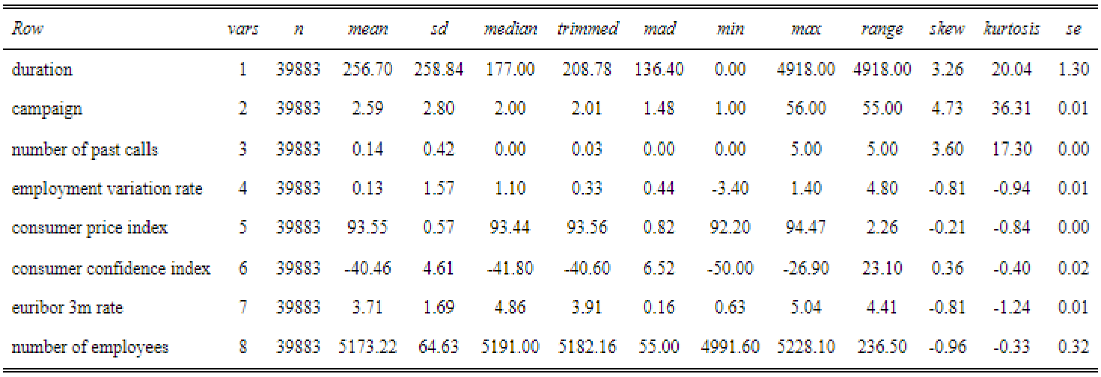
Before we proceed further into variable selection and modelling, the second step is the exploratory data analysis, which aims to find patterns in data regarding the behavior of the client and what influences their decision to subscribe to a bank deposit.

The following step is to implement variable selection algorithms in order to remove any unnecessary information and keep only the most important factors that will help in constructing an accurate model.

Finally, after the model building procedure is described, the goodness of fit of the model will be tested through some measures and the results will be presented as well as the interpretation of coefficients.

2. Exploratory Data Analysis

A brief description of the data is presented below in Table 2.1., to get an initial picture of the distribution of quantitative variables, by observing at the descriptive statistics. It appears that each client was contacted by the bank about two times on average, but, for some cases, up to 56 times. The duration of the calls also ranges from zero seconds- which implies that the call was declined- up to one and a half hours- which increases the chances of convincing customer to subscribe to long-term deposit. With regard to socio-economic variables, it should be noted that there is a decline in consumer confidence, which could lead to consumers decreasing their spending.



Row	*vars*	*n*	*mean*	*sd*	*median*	*trimmed*	*mad*	*min*	*max*	*range*	*skew*	*kurtosis*	*se*
duration	1	39883	256.70	258.84	177.00	208.78	136.40	0.00	4918.00	4918.00	3.26	20.04	1.30
campaign	2	39883	2.59	2.80	2.00	2.01	1.48	1.00	56.00	55.00	4.73	36.31	0.01
number of past calls	3	39883	0.14	0.42	0.00	0.03	0.00	0.00	5.00	5.00	3.60	17.30	0.00
employment variation rate	4	39883	0.13	1.57	1.10	0.33	0.44	-3.40	1.40	4.80	-0.81	-0.94	0.01
consumer price index	5	39883	93.55	0.57	93.44	93.56	0.82	92.20	94.47	2.26	-0.21	-0.84	0.00
consumer confidence index	6	39883	-40.46	4.61	-41.80	-40.60	6.52	-50.00	-26.90	23.10	0.36	-0.40	0.02
euribor 3m rate	7	39883	3.71	1.69	4.86	3.91	0.16	0.63	5.04	4.41	-0.81	-1.24	0.01
number of employees	8	39883	5173.22	64.63	5191.00	5182.16	55.00	4991.60	5228.10	236.50	-0.96	-0.33	0.32

Table 2.1.: Descriptive statistics for quantitative variables

Moreover, for better understanding of the data, the distributions of quantitative variables are visualized with histograms[1] and vioplots[2], while frequencies of occurrences for each level of categorical variables are visualized with bar plots[3].

Subsequently, there will be a more detailed examination of relationships between variables and, most importantly, how each attribute influences the decision of the potential customer. Thus, pairwise comparisons will be used in conjunction with univariate analysis to give an indication of which are the statistically important attributes and should be included in the final model.

As observed by the bar plots of the response variable, the marketing campaign did not have a successful outcome for the majority of the contacts. Specifically, by exploring

[1] See Appendix, Figure 2.1.
[2] See Appendix, Figure 2.2.
[3] See Appendix, Figure 2.3.

the associations with each of the qualitative variables individually, we come to the following conclusions.

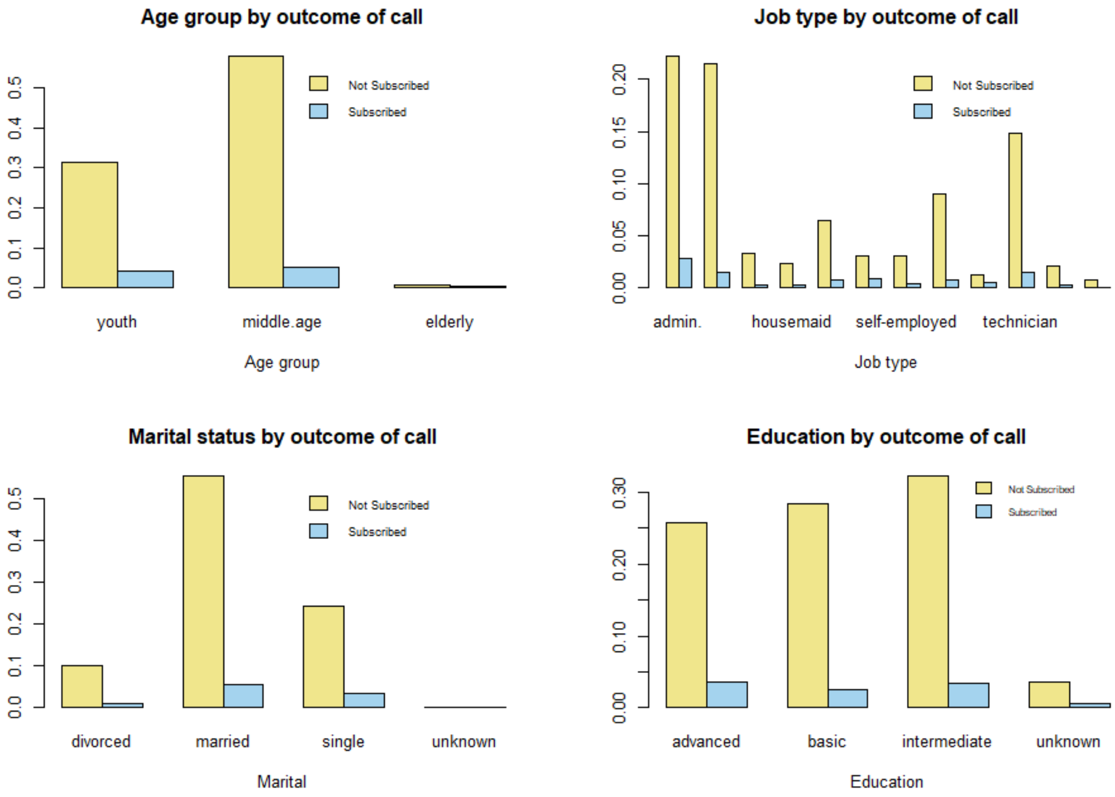


Figure 2.4.: Bar plots of association between qualitative variables and response variable

In Figure 2.4., it appears that, in general, people at middle age (between 35 and 65 years old) are more likely to subscribe to a term deposit. Moreover, customers who have a job of admin have the highest rate of subscribing a term deposit, but they are also the highest when it comes to not subscribing. This is simply because there are more customers working as admin than any other profession. Majority of the customers are married. In absolute terms, their subscription is high. But in relative terms, the clients who are single are subscribed higher. Similar applies for education level, the majority has intermediate education level, but in relative terms, the clients who have advanced education level are subscribed higher.

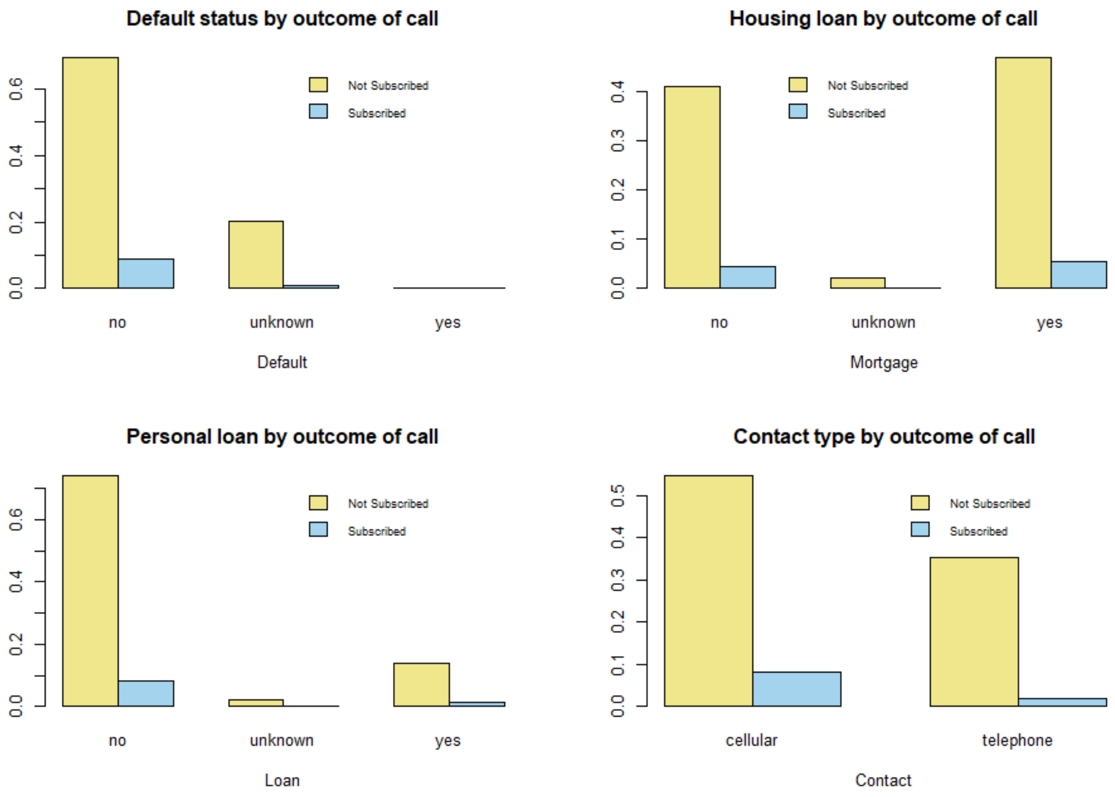


Figure 2.5.: Bar plots of association between qualitative variables and response variable

In Figure 2.5., it is important to note that the majority of the clients do not have loans in delay or a personal loan and, therefore, they are more likely to subscribe to term deposit. On the other hand, there is no significant difference between those who have housing loan and those who haven't, in terms of subscription. The contact with clients occurs in most cases via cellphone and the subscriptions are also more in those cases.

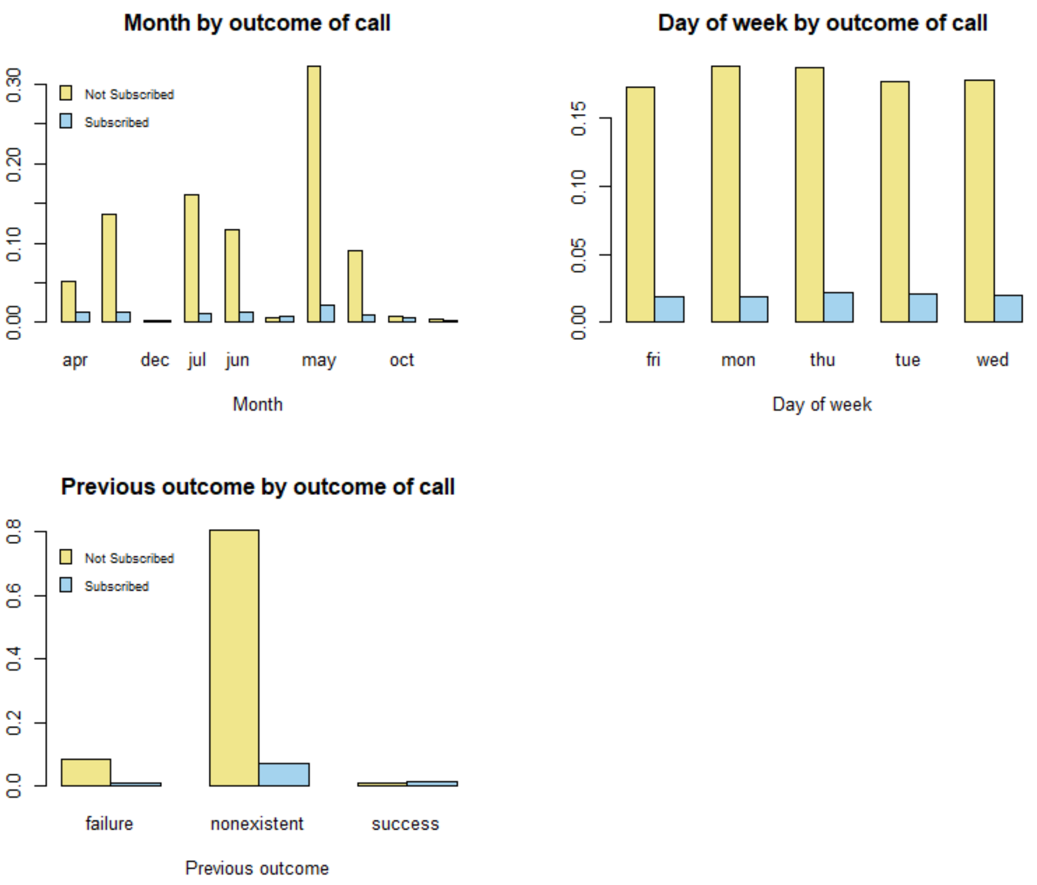
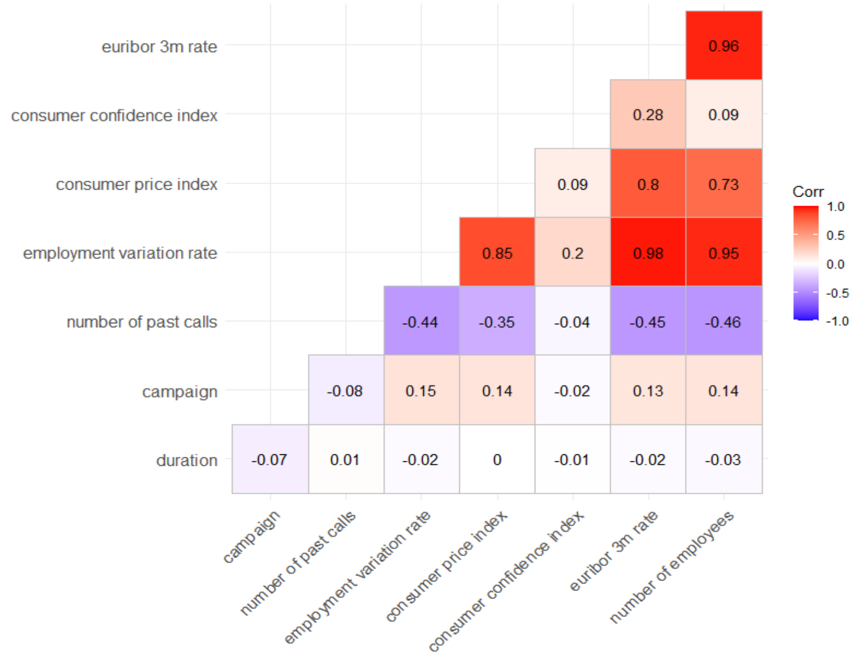


Figure 2.6.: Bar plots of association between qualitative variables and response variable

In Figure 2.6., it appears that most contacts took place in May as well as most successful subscriptions, while summer months and April also had high success rates based on the frequency of the calls, meaning that month variable influences the subscriptions. Regarding weekdays, the frequency of calls and the success rate is almost the same for all weekdays, therefore the decision of the potential customer does not change for each day of the week and it is assumed that this variable will not be included in the model. For the majority of people contacted, the previous marketing campaign outcome does not exist, meaning that they are new customers who have not been contacted earlier. Moreover, for the customers that had a successful outcome from the previous campaign, majority of those customers did subscribe for a term deposit. Therefore, it can be assumed that this feature may hold some value in predicting the target variable.



	campaign	number of past calls	employment variation rate	consumer price index	consumer confidence index	euribor 3m rate	number of employees
euribor 3m rate							0.96
consumer confidence index						0.28	0.09
consumer price index					0.09	0.8	0.73
employment variation rate				0.85	0.2	0.98	0.95
number of past calls			-0.44	-0.35	-0.04	-0.45	-0.46
campaign		-0.08	0.15	0.14	-0.02	0.13	0.14
duration	-0.07	0.01	-0.02	0	-0.01	-0.02	-0.03

Table 2.2.: Correlation Matrix

The bivariate visualizations observed in correlation matrix (Table 2.2.) depict some strong linear relationships between socio-economic variables. Specifically, between Euribor 3m rate, number of employees, employment variation rate and consumer price index. This shows a multicollinearity effect between these predictors, therefore only one of them should be included in the final model.

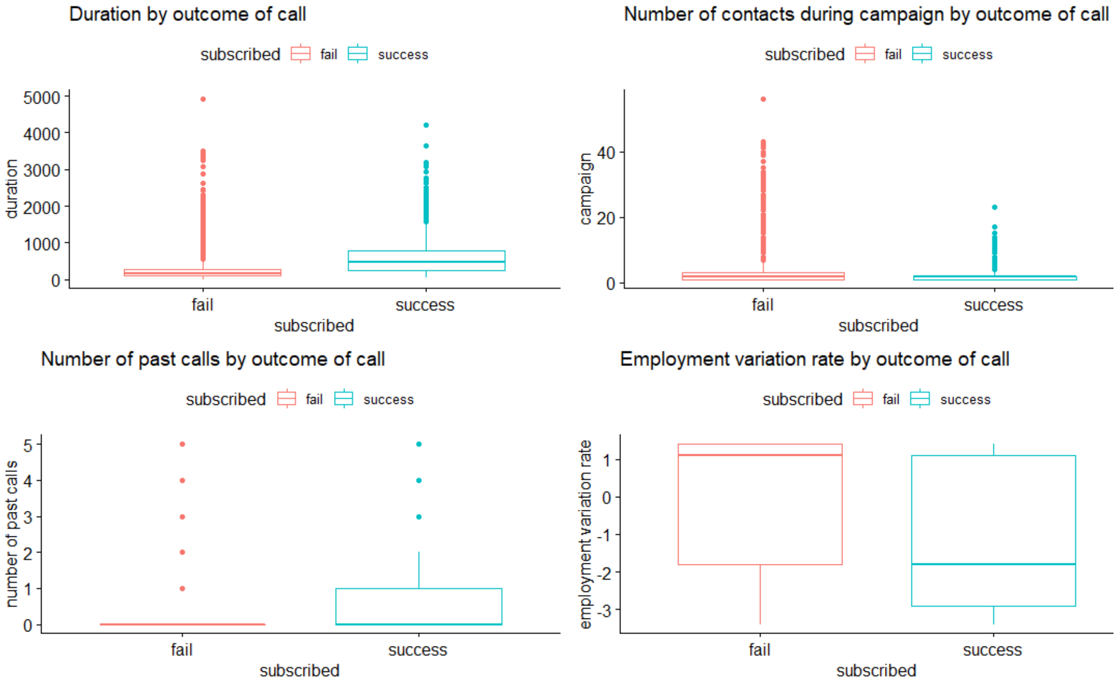


Figure 2.7.: Boxplots of association between quantitative variables and response variable

In Figures 2.7. and 2.8., the association between quantitative variables and response variable is visualized using boxplots. It is important to note that the duration of calls highly affects the output target, because, as previously mentioned, if duration is zero, then the response of the customer is negative. As observed also in Figure 2.7., the duration of calls for successful outcomes is higher than in calls that were not successful, which makes sense because those who are interested in subscribing to term deposit, stay on the call longer. Number of contacts during campaign and number of past calls do not seem to have a strong influence on the success rate, while employment variation rate is higher when the outcome of the call is negative. This can be justified by the uncertainty that clients might have when the employment variation is high.

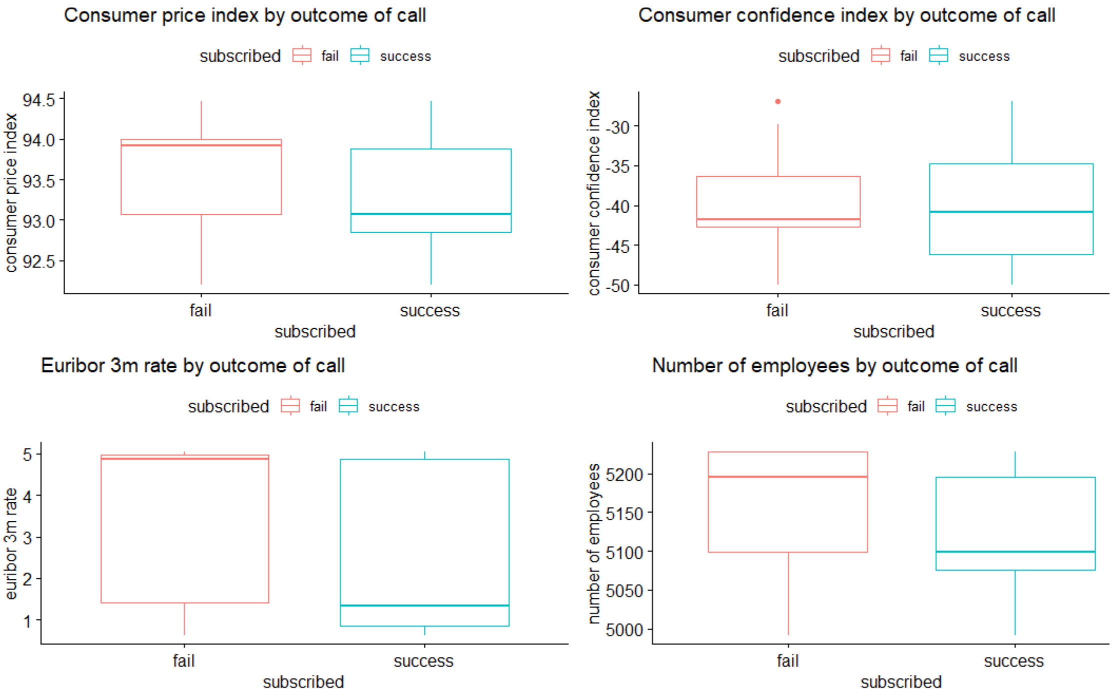


Figure 2.8.: Boxplots of association between quantitative variables and response variable

Similarly, in Figure 2.8., when the consumer price index is high, the outcome of the call is most likely to be failure, which makes sense because when prices are high, people tend to be more conservative. Euribor 3m rate also affects the decision of a potential customer, because when it is higher, the interest on the loan rises, so people are not so willing to subscribe to term deposits. Number of employees affects the outcome of the call as well, because when number of employees is larger, the call is more likely to fail.

3. Model building & Evaluation

This section will focus on the model building procedure and evaluation. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary), specifically in the current study, whether or not the client subscribed a long-term deposit (success/failure) and the predicted values are probabilities and not values. Its formula is $\log(p/1\text{-}p)=\beta_0+\beta_1x_1$.

The large number of potential useful features for a model demands a stricter choice of relevant attributes. Feature selection is a useful step to discard irrelevant inputs, leading to simpler data-driven models that are easier to interpret and that tend to provide better predictive performances. To achieve that, we implement variable selection algorithms (LASSO - AIC/BIC).

Initially, LASSO method filters the full model and selects the most appropriate covariates. LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the accuracy and interpretability of the model. Specifically, it forces the sum of the absolute value of the coefficients to be less than a fixed value, which sets certain coefficients to zero, thereby removing them from the model. In logistic regression, Lasso regularization works by adding a penalty term to the log likelihood function. The regularization parameter is “lambda” and measures the degree to which the coefficients are penalized.

By choosing lambda within one standard error from the minimum, LASSO shrinks more parameters towards zero, as observed also in Figures 3.1., 3.2. and the model becomes significantly simpler, by selecting twenty-nine parameters and dealing with the multicollinearity issue. Specifically, it removes variables housing and loan, then variable previous, that provides similar information with poutcome, and cons.price.idx, euribor3m, which are highly correlated as previously noted.

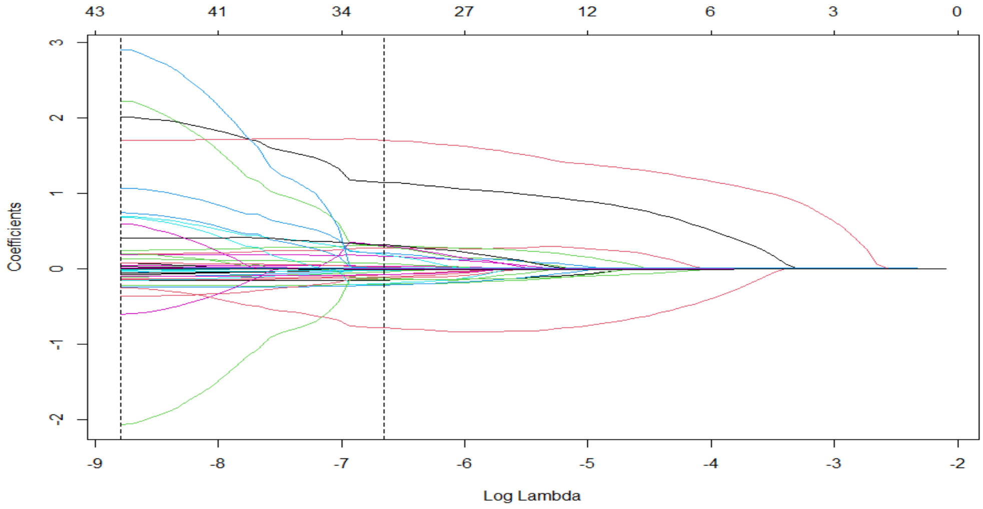


Figure 3.1.: Variable coefficients per lambda

As displayed in Figure 3.2, the lambda with one standard error from the minimum is chosen instead of the minimum lambda, because the GLM deviance remains the same for less parameters.

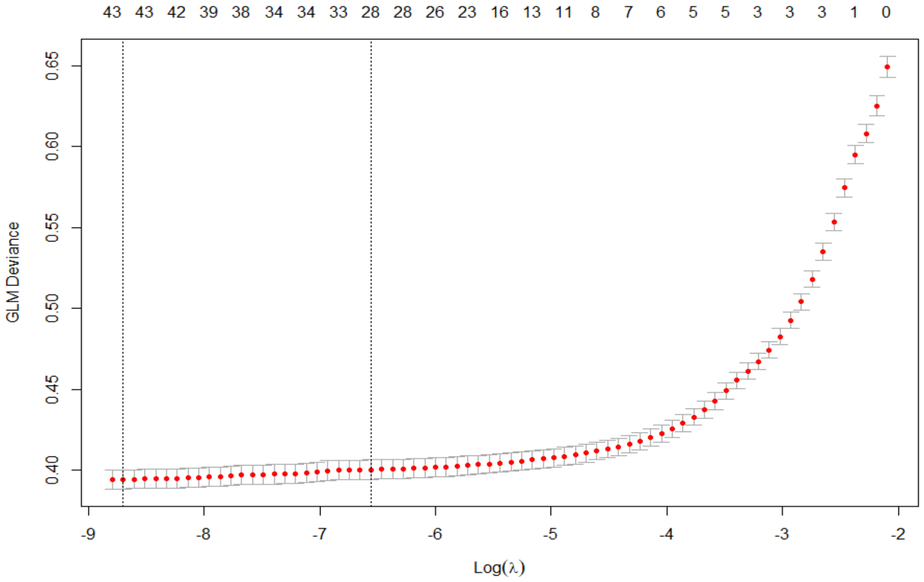


Figure 3.2.: Misclassification error per lambda

By checking the summary of the lasso model, it is found that some variables are not statistically significant since their p-values are over 0.05 and can be considered as zero. Therefore, we remove those variables (marital, contact, cons.conf.idx) and fit again the general linear model. The summary output of the second lasso model is presented in Table 3.1., Appendix.

Subsequently, the Bayesian Information Criterion (BIC) method is performed, using the stepwise selection with penalty = log(n). It is a method for scoring and selecting a model and is appropriate for models' fit under the maximum likelihood estimation framework. BIC is generally preferred instead of AIC (Akaike Information Criterion) when the focus is interpretation rather than prediction. Nevertheless, AIC method will also be performed at a next stage and the models will be compared in order to choose the best fit.

By comparing the AIC improvements from dropping and adding each candidate variable from the current model, we end up with a model having the smallest AIC. The model ends up with eight variables: age, default, month, duration, campaign, poutcome, emp.var.rate and nr.employed.

Furthermore, multicollinearity between the covariates should be checked using the Vif test.[4] This shows that emp.var.rate and nr.employed are correlated, as previously observed in the correlation table of quantitative variables. Therefore, we decide to remove the emp.var.rate variable to fix the multicollinearity issue and fit again the general linear model into object step.model1. The model from the BIC method now consists of seven variables: age, default, month, duration, campaign, poutcome and nr.employed. Summary output of this model[5] indicates statistically significant p-values and the residual deviance is 15946 on 39852 degrees of freedom.

After fitting a model to the observed data, one of the next essential steps is to investigate how well the proposed model fits the observed data. Two of the most commonly used goodness of fit measures are the deviance and Pearson's chi-squared (X^2) goodness of fit test statistics.

By comparing the fitted model with the null model (a model with just the intercept)[6], we conclude that step.model1 (model after Lasso and BIC) fits significantly better than the null model. Specifically, the difference in deviance for the two models is 9972, the degrees of freedom for the difference between the two models is 18 and the p-value is less than 0.001, which tells us that our model fits significantly better than null model.

[4] See Appendix, Table 3.2.
[5] See Appendix, Table 3.3.
[6] See Appendix, Table 3.4.

Another goodness of fit measure that examines how well the model explains the data is the pseudo R^2 metric. Most notable is McFadden's R^2 which is defined as $1-[\ln(L_M)/\ln(L_0)]$ where $\ln(L_M)$ is the log likelihood value for the fitted model and $\ln(L_0)$ is the log likelihood for the null model with only an intercept as a predictor. It ranges from 0 to 1 with values closer to zero indicate that the model has no predictive power. McFadden's pseudo R^2 ranging from 0.2 to 0.4 indicates very good model fit, therefore step.model1 that has a pseudo R^2 of around 0.385, shows that the model does a very good fit of the data [7].

The next step of the analysis is to implement AIC method (Akaike Information Criterion) for variable selection and compare the results with the BIC method. The BIC method penalizes free parameters more strongly, so it is expected that the model from AIC will be larger. Based on the p-values of Wald Chi-square tests for lasso.model2[8], we conclude that education and days.of.week variables have low statistical significance for the model. After fixing multicollinearity between the covariates by removing emp.var.rate and also removing days.of.week and education variables, we end up with a model with eight variables: age, job, default, month, duration, campaign, poutcome and nr.employed. Summary output of this model[9] indicates that residual deviance is lower, at 15899 on 39841 degrees of freedom. Moreover, by comparing the fitted model with the null model (a model with just the intercept)[10], we conclude that step.model2 fits significantly better than the null model. Specifically, the difference in deviance for the two models is 10019, the degrees of freedom for the difference between the two models is 29 and the p-value is less than 0.001, which tells us that our model fits significantly better than null model. Moreover, the median deviance is very close to zero, so model is unbiased towards both directions of the outcomes. Lastly, the McFadden's R^2 is now 0.3866, slightly improved than the previous model[11].

In order to compare the two models, step.model1 and step.model2, we implement the likelihood ratio test, to further test the goodness of fit of the two competing statistical models based on the ratio of their likelihoods. The best model is the one that makes the data most likely or maximizes the likelihood function. Based on the Table

[7] See Appendix, Table 3.5.
[8] See Appendix, Table 3.6.
[9] See Table 3.12.
[10] See Appendix, Table 3.7.
[11] See Appendix, Table 3.8.

3.9.(Appendix), the p-value less than 0.05, so the null hypothesis is rejected and it can be assumed that the larger model has a statistically significant improvement in terms of goodness of fit.

The final step of the analysis is to check if the model satisfies the assumptions of logistic regression.

The logistic regression method assumes that:

- ✓ The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.

This assumption is satisfied, since subscribed variable has two levels (0,1).

- ✓ There is a linear relationship between the logit of the outcome and each predictor variables.

Figure 3.3. shows that duration variable is perfectly linearly associated with the subscribed outcome in logit scale.

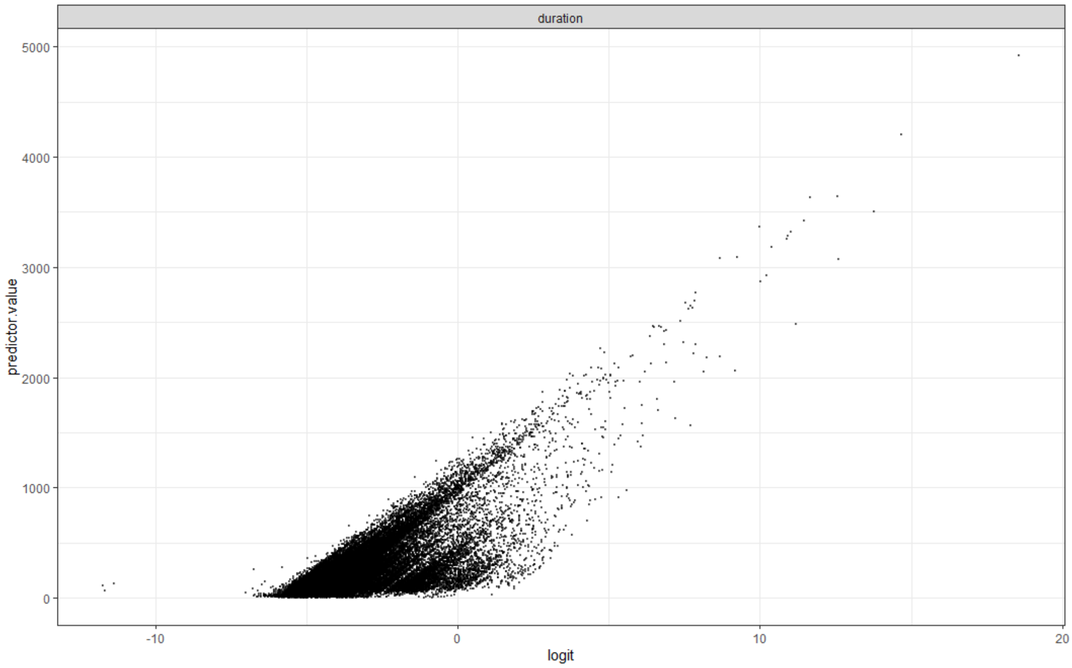


Figure 3.3.: Linearity between duration and logit of subscribed

- ✓ There are no influential values (extreme values or outliers) in the continuous predictors.

In Figure 3.4., we notice that there are not many strongly influential outliers.

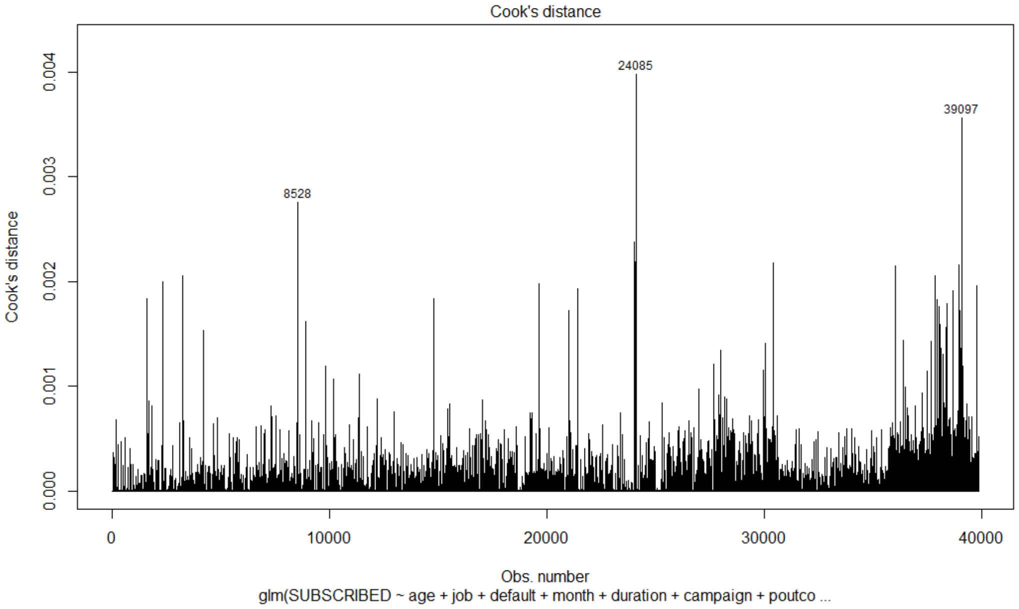


Figure 3.4.: Influential observations

- ✓ There are no high intercorrelations (i.e., multicollinearity) among the predictors.

Table 3.10. in Appendix shows no problem of multicollinearity in data.

- ✓ Indepedence of errors (Figure 3.5.)

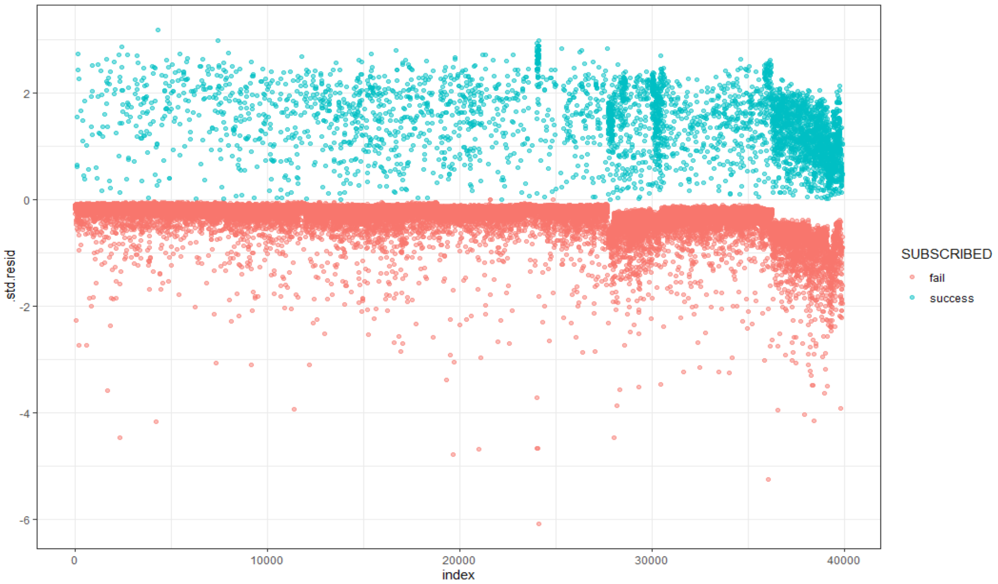
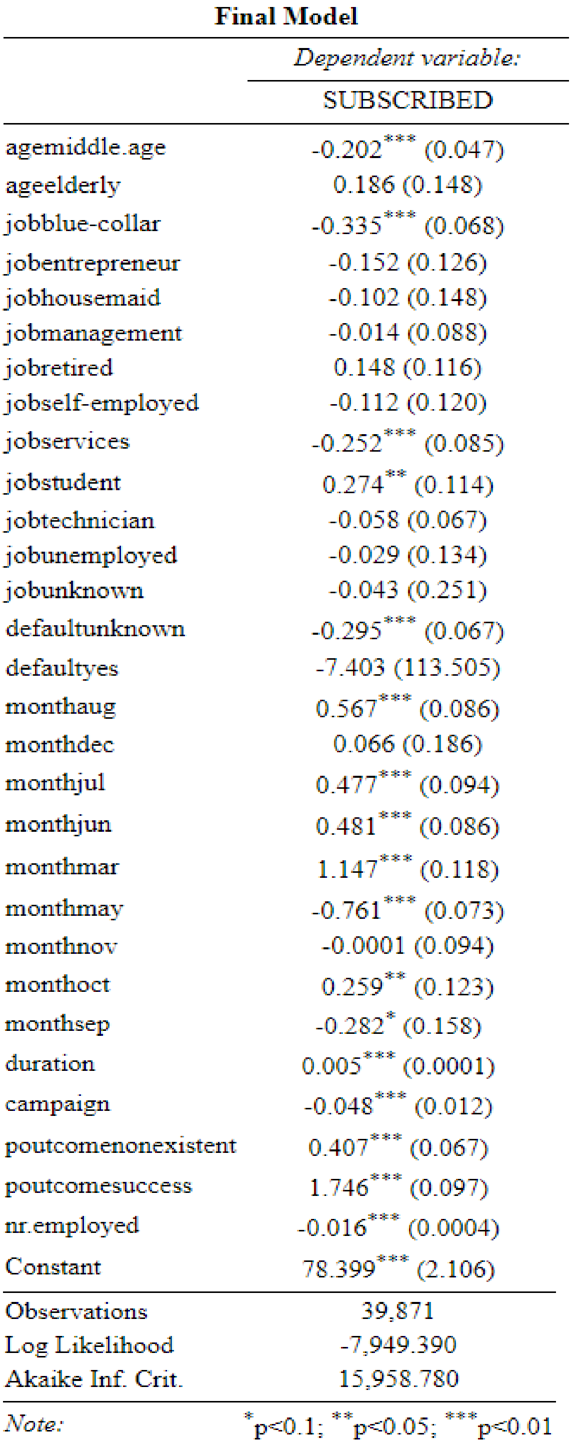


Figure 3.5.: Residuals vs fitted values

Consequently, we conclude to the final model, step.model2, which is presented below along with the interpretation of coefficients. This can also be derived from Table 3.11. in Appendix, along with confidence intervals.



Final Model

	Dependent variable:
	SUBSCRIBED
agemiddle.age	-0.202*** (0.047)
ageelderly	0.186 (0.148)
jobblue-collar	-0.335*** (0.068)
jobentrepreneur	-0.152 (0.126)
jobhousemaid	-0.102 (0.148)
jobmanagement	-0.014 (0.088)
jobretired	0.148 (0.116)
jobself-employed	-0.112 (0.120)
jobservices	-0.252*** (0.085)
jobstudent	0.274** (0.114)
jobtechnician	-0.058 (0.067)
jobunemployed	-0.029 (0.134)
jobunknown	-0.043 (0.251)
defaultunknown	-0.295*** (0.067)
defaultyes	-7.403 (113.505)
monthaug	0.567*** (0.086)
monthdec	0.066 (0.186)
monthjul	0.477*** (0.094)
monthjun	0.481*** (0.086)
monthmar	1.147*** (0.118)
monthmay	-0.761*** (0.073)
monthnov	-0.0001 (0.094)
monthoct	0.259** (0.123)
monthsep	-0.282* (0.158)
duration	0.005*** (0.0001)
campaign	-0.048*** (0.012)
poutcomenonexistent	0.407*** (0.067)
poutcomesuccess	1.746*** (0.097)
nr.employed	-0.016*** (0.0004)
Constant	78.399*** (2.106)
Observations	39,871
Log Likelihood	-7,949.390
Akaike Inf. Crit.	15,958.780
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3.12.: Summary of step.model2

Given that the logit function is logit(p) = log(p/(1-p)), the mathematical formula of the model is the following:

$$\mathit{logit(SUBSCRIBED)} = 78.4 - 0.2*\mathit{AgeMiddle.age} - 0.335*\mathit{JobBlue\text{-}collar} - 0.252*\mathit{JobServices} + 0.274*\mathit{JobStudent} - 0.295*\mathit{DefaultUnknown} + 0.567*\mathit{MonthAugust} + 0.477*\mathit{MonthJuly} + 0.481*\mathit{MonthJune} + 1.147*\mathit{MonthMarch} - 0.761*\mathit{MonthMay} + 0.259*\mathit{MonthOctober} - 0.282*\mathit{MonthSeptember} + 0.005*\mathit{Duration} - 0.048*\mathit{Campaign} + 0.407*\mathit{PoutcomeNonexistent} + 1.746*\text{PoutcomeSuccess} - 0.016*\mathit{Nr.Employed}$$

- Agemiddle.age: The middle aged group has a $e^{-0.202} = 0.81$ times the odds of the youth group of subscribing.
- Jobblue-collar: The blue-collar job group has a $e^{-0.335} = 0.71$ times the odds of the admin job group of subscribing.
- Jobservices: The services job group has a $e^{-0.252} = 0.77$ times the odds of the admin job group of subscribing.
- Jobstudent: The student group has a $e^{0.274} = 1.31$ times the odds of the admin job group of subscribing.
- The default unknown group has a $e^{-0.295} = 0.74$ times the odds of the default no group of subscribing.
- Monthaug: The month August group has a $e^{0.567} = 1.76$ times the odds of the month April group of subscribing.
- Monthjul: The month July group has a $e^{0.477} = 1.61$ times the odds of the month April group of subscribing.
- Monthjun: The month June group has a $e^{0.481} = 1.61$ times the odds of the month April group of subscribing.
- Monthmar: The month March group has a $e^{1.147} = 3.14$ times the odds of the month April group of subscribing.
- Monthmay: The month Mary group has a $e^{-0.761} = 0.46$ times the odds of the month April group of subscribing.
- Monthoct: The month October group has a $e^{0.259}= 1.29$ times the odds of the month April group of subscribing.
- Monthsept: The month September group has a $e^{-0.282}= 0.75$ times the odds of the month April group of subscribing.
- Duration: An increase of 1 unit in duration multiplies the odds of subscribing by $e^{0.005} = 1.005$.
- Campaign: An increase of 1 unit in campaign multiplies the odds of subscribing by $e^{0.005} = 1.005$.
- Poutcomenonexistent: The non existent previous outcome group has a $e^{0.407} = 1.5$ the odds of the failure previous outcome group of subscribing.
- Poutcomesuccess: The success previous outcome group has a $e^{1.746} = 5.73$ the odds of the failure previous outcome group of subscribing.
- Nr.employed: An increase of 1 unit in number of employed multiplies the odds of subscribing by $e^{-0.016} = 0.98$.

4. Conclusion

Within the banking industry, optimizing targeting for telemarketing is a key issue, under a growing pressure to increase profits and reduce costs. In this study, we attempted to identify the features which contribute to a successful contact, namely the client subscribes to the term deposit.

These factors turned out to be age group, job type, default status of client, month in which the call was conducted, duration of call, number of contacts performed during this campaign for this client, previous outcome of call (or whether the client has not been contacted before) and number of employees.

The models tested were similar to each other, but we selected the one that fitted the data better and satisfied all of the assumptions.

One limitation of the analysis is that the duration of the call can be a confounding variable. From the exploratory data analysis, duration appeared to be a very important variable. However, it is possible that the interests of the consumers influenced the length of the call, and henceforth, caused the duration to be longer for those who were interested in the product and were willing to purchase term deposit. A future study could explore this relationship more.

One other limitation of the analysis was that there were high correlations between some variables, which could cause problems and, finally, that the number of days since the last call for any other campaign is indicated with ‘999’ days if there was no previous call. In further research it could be corrected.

Furthermore, the dataset could be enriched with more information regarding telemarketing calls and possibly provide new valuable knowledge to improve model accuracy.

5. References

- Statistics for Business Analytics II – course slides.
- Jiang, Y. (2018). “Using logistic regression model to predict the success of bank telemarketing.” International Journal on Data Science and Technology.
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems.
- Hosmer, D.W., T. Hosmer, S. Le Cessie and S. Lemeshow (1997). “A comparison of goodness-of-fit tests for the logistic regression model.” Statistics in Medicine.

6. Appendix

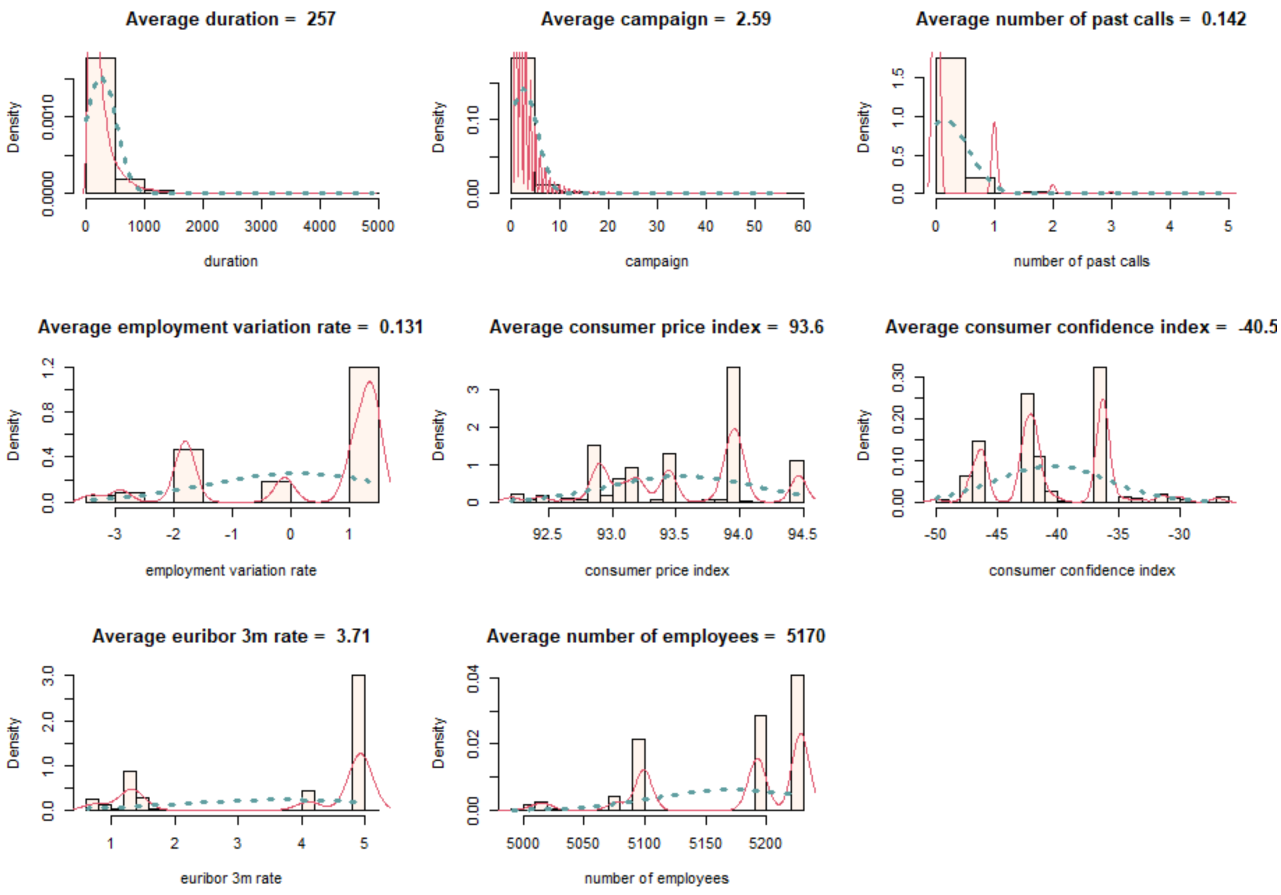


Figure 2.1: Histograms of quantitative variables

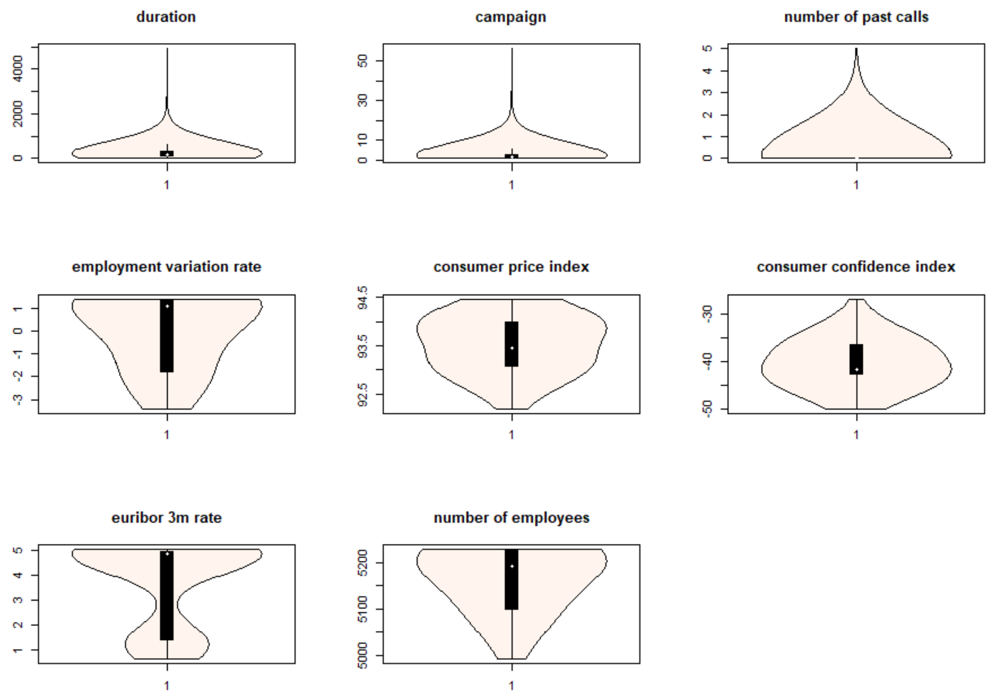


Figure 2.2: Vioplots of quantitative variables

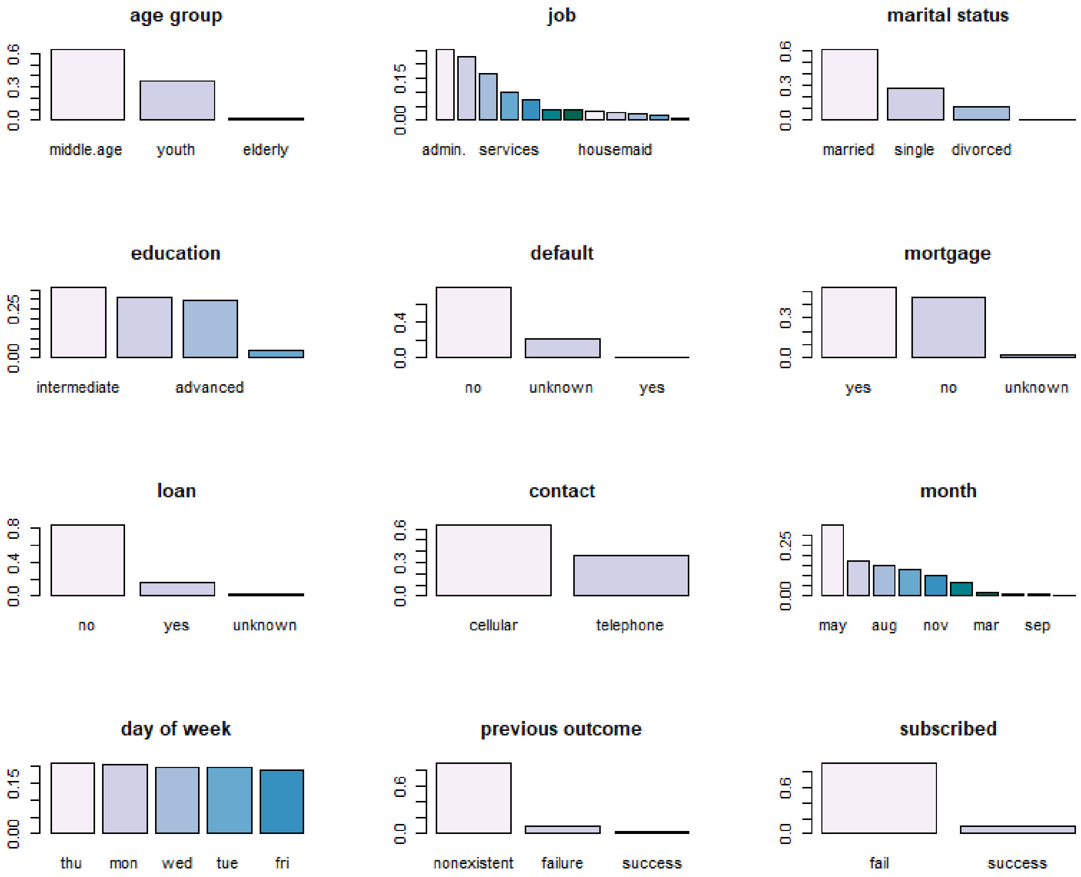
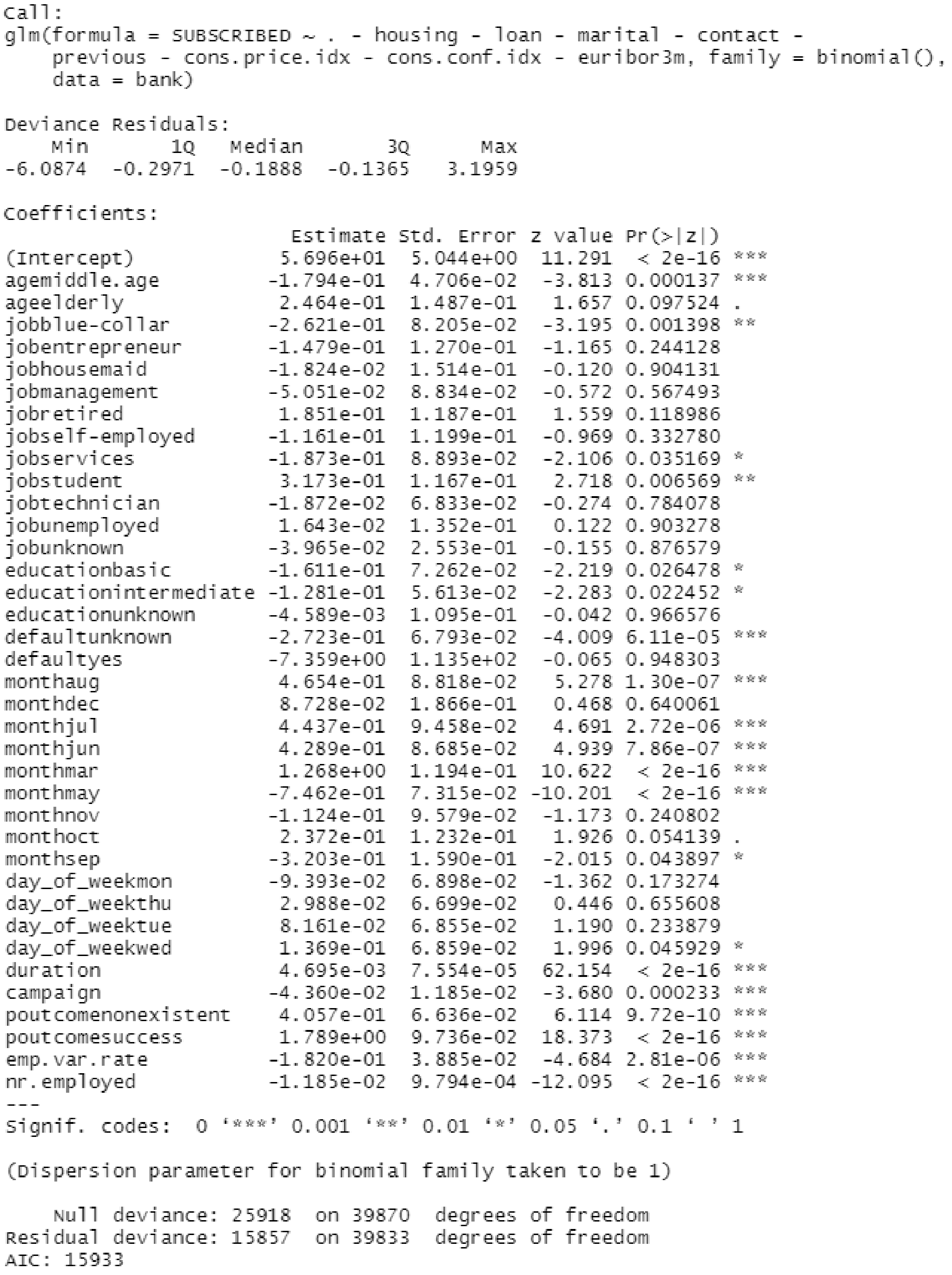
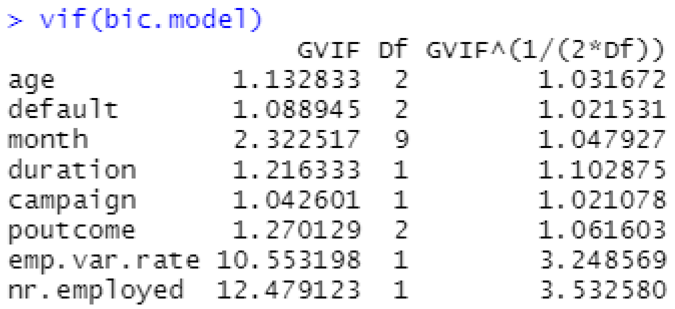


Figure 2.3: Bar plots of qualitative variables



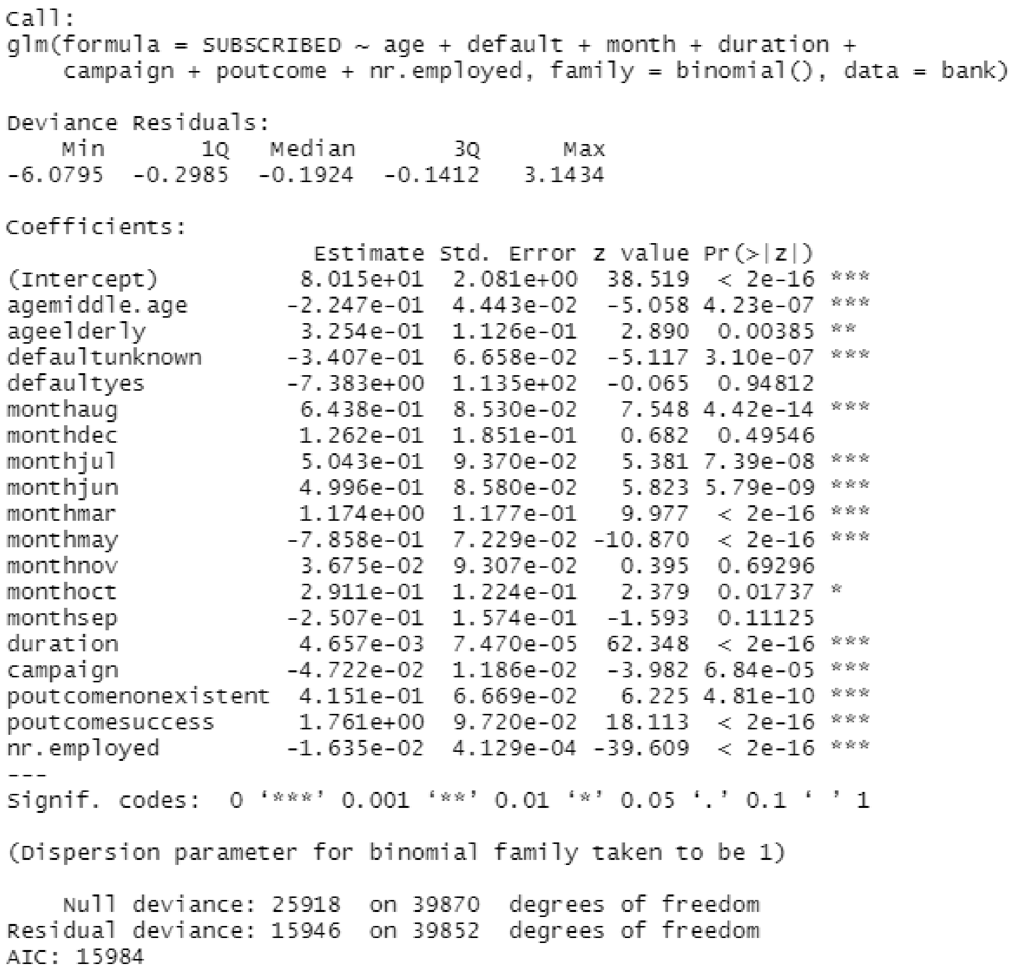
```
Call:
glm(formula = SUBSCRIBED ~ . - housing - loan - marital - contact - 
    previous - cons.price.idx - cons.conf.idx - euribor3m, family = binomial(), 
    data = bank)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.0874  -0.2971  -0.1888  -0.1365   3.1959  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            5.696e+01  5.044e+00  11.291  < 2e-16 ***
agemiddle.age         -1.794e-01  4.706e-02  -3.813 0.000137 ***
ageelderly             2.464e-01  1.487e-01   1.657 0.097524 .  
jobblue-collar        -2.621e-01  8.205e-02  -3.195 0.001398 ** 
jobentrepreneur       -1.479e-01  1.270e-01  -1.165 0.244128    
jobhousemaid          -1.824e-02  1.514e-01  -0.120 0.904131    
jobmanagement         -5.051e-02  8.834e-02  -0.572 0.567493    
jobretired             1.851e-01  1.187e-01   1.559 0.118986    
jobself-employed      -1.161e-01  1.199e-01  -0.969 0.332780    
jobservices           -1.873e-01  8.893e-02  -2.106 0.035169 *  
jobstudent             3.173e-01  1.167e-01   2.718 0.006569 ** 
jobtechnician         -1.872e-02  6.833e-02  -0.274 0.784078    
jobunemployed          1.643e-02  1.352e-01   0.122 0.903278    
jobunknown            -3.965e-02  2.553e-01  -0.155 0.876579    
educationbasic        -1.611e-01  7.262e-02  -2.219 0.026478 *  
educationintermediate -1.281e-01  5.613e-02  -2.283 0.022452 *  
educationunknown      -4.589e-03  1.095e-01  -0.042 0.966576    
defaultunknown        -2.723e-01  6.793e-02  -4.009 6.11e-05 ***
defaultyes            -7.359e+00  1.135e+02  -0.065 0.948303    
monthaug               4.654e-01  8.818e-02   5.278 1.30e-07 ***
monthdec               8.728e-02  1.866e-01   0.468 0.640061    
monthjul               4.437e-01  9.458e-02   4.691 2.72e-06 ***
monthjun               4.289e-01  8.685e-02   4.939 7.86e-07 ***
monthmar               1.268e+00  1.194e-01  10.622  < 2e-16 ***
monthmay              -7.462e-01  7.315e-02 -10.201  < 2e-16 ***
monthnov              -1.124e-01  9.579e-02  -1.173 0.240802    
monthoct               2.372e-01  1.232e-01   1.926 0.054139 .  
monthsep              -3.203e-01  1.590e-01  -2.015 0.043897 *  
day_of_weekmon        -9.393e-02  6.898e-02  -1.362 0.173274    
day_of_weekthu         2.988e-02  6.699e-02   0.446 0.655608    
day_of_weektue         8.161e-02  6.855e-02   1.190 0.233879    
day_of_weekwed         1.369e-01  6.859e-02   1.996 0.045929 *  
duration               4.695e-03  7.554e-05  62.154  < 2e-16 ***
campaign              -4.360e-02  1.185e-02  -3.680 0.000233 ***
poutcomenonexistent    4.057e-01  6.636e-02   6.114 9.72e-10 ***
poutcomesuccess        1.789e+00  9.736e-02  18.373  < 2e-16 ***
emp.var.rate          -1.820e-01  3.885e-02  -4.684 2.81e-06 ***
nr.employed           -1.185e-02  9.794e-04 -12.095  < 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: 25918  on 39870  degrees of freedom
Residual deviance: 15857  on 39833  degrees of freedom
AIC: 15933
```

Table 3.1.: Summary of lasso model



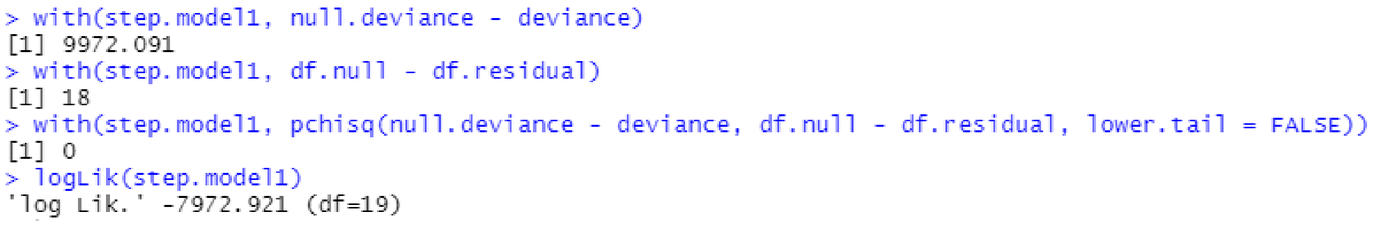
```
> vif(bic.model)
                  GVIF Df GVIF^(1/(2*Df))
age           1.132833  2        1.031672
default       1.088945  2        1.021531
month         2.322517  9        1.047927
duration      1.216333  1        1.102875
campaign      1.042601  1        1.021078
poutcome      1.270129  2        1.061603
emp.var.rate 10.553198  1        3.248569
nr.employed  12.479123  1        3.532580
```

Table 3.2.: Multicollinearity test for bic.model



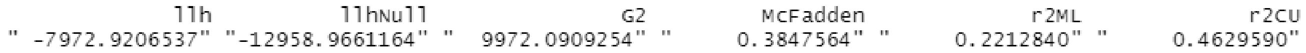
```
Call:
glm(formula = SUBSCRIBED ~ age + default + month + duration + 
    campaign + poutcome + nr.employed, family = binomial(), data = bank)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.0795  -0.2985  -0.1924  -0.1412   3.1434  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)          8.015e+01  2.081e+00  38.519  < 2e-16 ***
agemiddle.age       -2.247e-01  4.443e-02  -5.058 4.23e-07 ***
ageelderly           3.254e-01  1.126e-01   2.890  0.00385 ** 
defaultunknown      -3.407e-01  6.658e-02  -5.117 3.10e-07 ***
defaultyes          -7.383e+00  1.135e+02  -0.065  0.94812    
monthaug             6.438e-01  8.530e-02   7.548 4.42e-14 ***
monthdec             1.262e-01  1.851e-01   0.682  0.49546    
monthjul             5.043e-01  9.370e-02   5.381 7.39e-08 ***
monthjun             4.996e-01  8.580e-02   5.823 5.79e-09 ***
monthmar             1.174e+00  1.177e-01   9.977  < 2e-16 ***
monthmay            -7.858e-01  7.229e-02 -10.870  < 2e-16 ***
monthnov             3.675e-02  9.307e-02   0.395  0.69296    
monthoct             2.911e-01  1.224e-01   2.379  0.01737 *  
monthsep            -2.507e-01  1.574e-01  -1.593  0.11125    
duration             4.657e-03  7.470e-05  62.348  < 2e-16 ***
campaign            -4.722e-02  1.186e-02  -3.982 6.84e-05 ***
poutcomenonexistent  4.151e-01  6.669e-02   6.225 4.81e-10 ***
poutcomesuccess      1.761e+00  9.720e-02  18.113  < 2e-16 ***
nr.employed         -1.635e-02  4.129e-04 -39.609  < 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: 25918  on 39870  degrees of freedom
Residual deviance: 15946  on 39852  degrees of freedom
AIC: 15984
```

Table 3.3.: Summary of step.model1



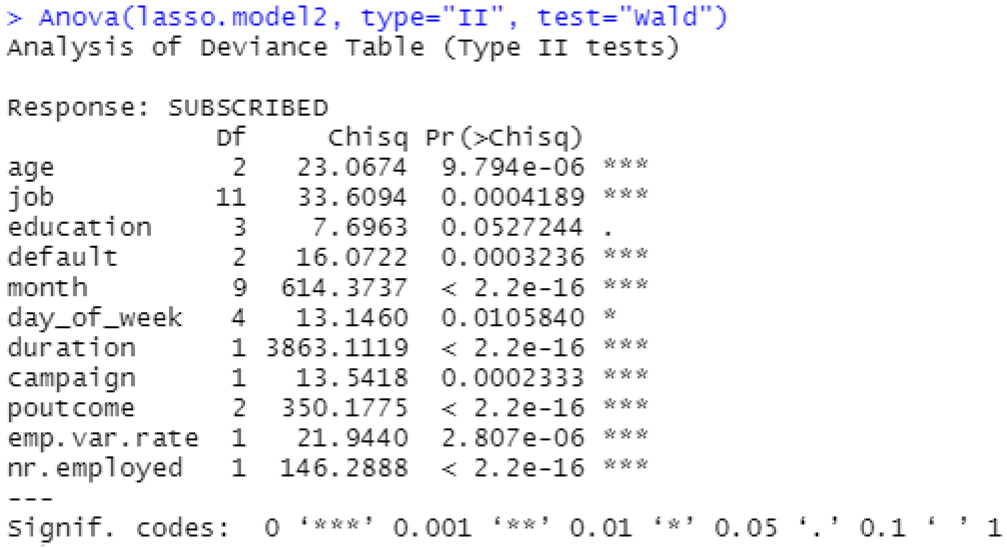
```
> with(step.model1, null.deviance - deviance)
[1] 9972.091
> with(step.model1, df.null - df.residual)
[1] 18
> with(step.model1, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 0
> logLik(step.model1)
'log Lik.' -7972.921 (df=19)
```

Table 3.4.: Goodness of fit measures for step.model1



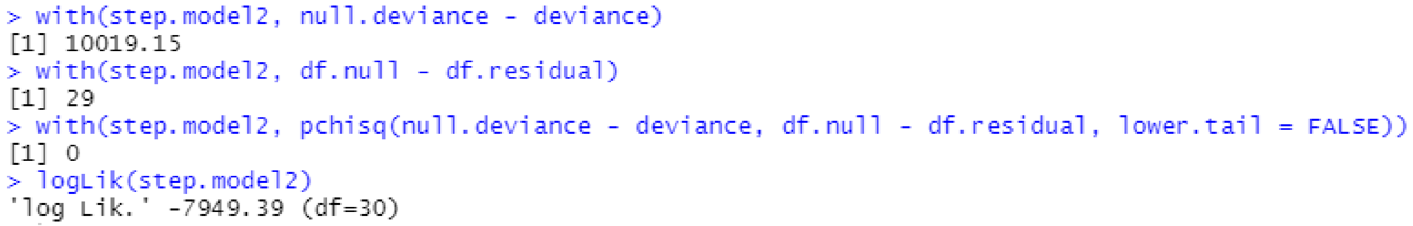
```
             llh          llhNull               G2         McFadden             r2ML             r2CU 
" -7972.9206537" "-12958.9661164" "  9972.0909254" "     0.3847564" "     0.2212840" "     0.4629590" 
```

Table 3.5.: Pdeudo R^2 metric for step.model1



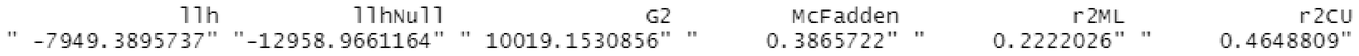
```
> Anova(lasso.model2, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)

Response: SUBSCRIBED
             Df     Chisq Pr(>Chisq)    
age           2   23.0674  9.794e-06 ***
job          11   33.6094  0.0004189 ***
education     3    7.6963  0.0527244 .  
default       2   16.0722  0.0003236 ***
month         9  614.3737  < 2.2e-16 ***
day_of_week   4   13.1460  0.0105840 *  
duration      1 3863.1119  < 2.2e-16 ***
campaign      1   13.5418  0.0002333 ***
poutcome      2  350.1775  < 2.2e-16 ***
emp.var.rate  1   21.9440  2.807e-06 ***
nr.employed   1  146.2888  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3.6.: Anova Type II tests for lasso.model2



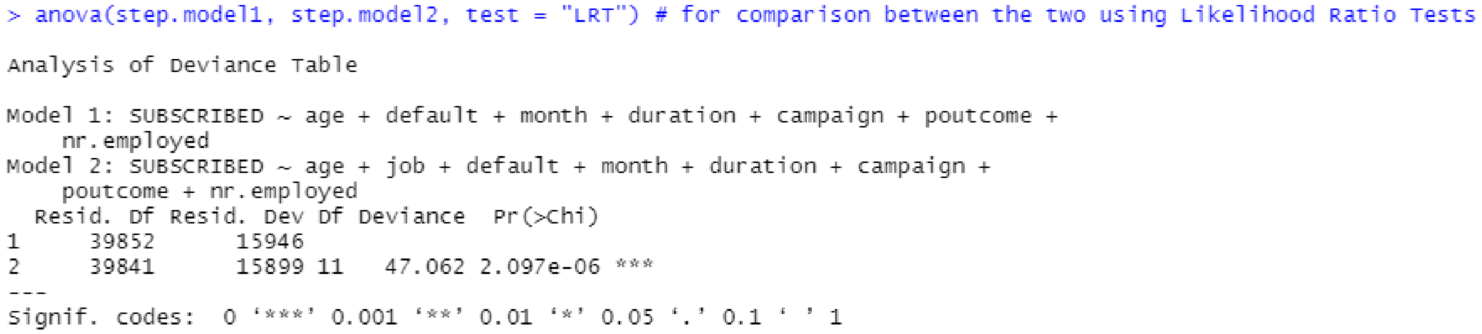
```
> with(step.model2, null.deviance - deviance)
[1] 10019.15
> with(step.model2, df.null - df.residual)
[1] 29
> with(step.model2, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 0
> logLik(step.model2)
'log Lik.' -7949.39 (df=30)
```

Table 3.7.: Goodness of fit measures for step.model2



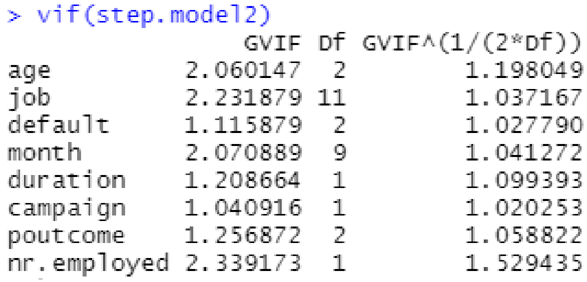
```
             llh          llhNull               G2         McFadden             r2ML             r2CU 
" -7949.3895737" "-12958.9661164" " 10019.1530856" "     0.3865722" "     0.2222026" "     0.4648809" 
```

Table 3.8.: Pdeudo R^2 metric for step.model2



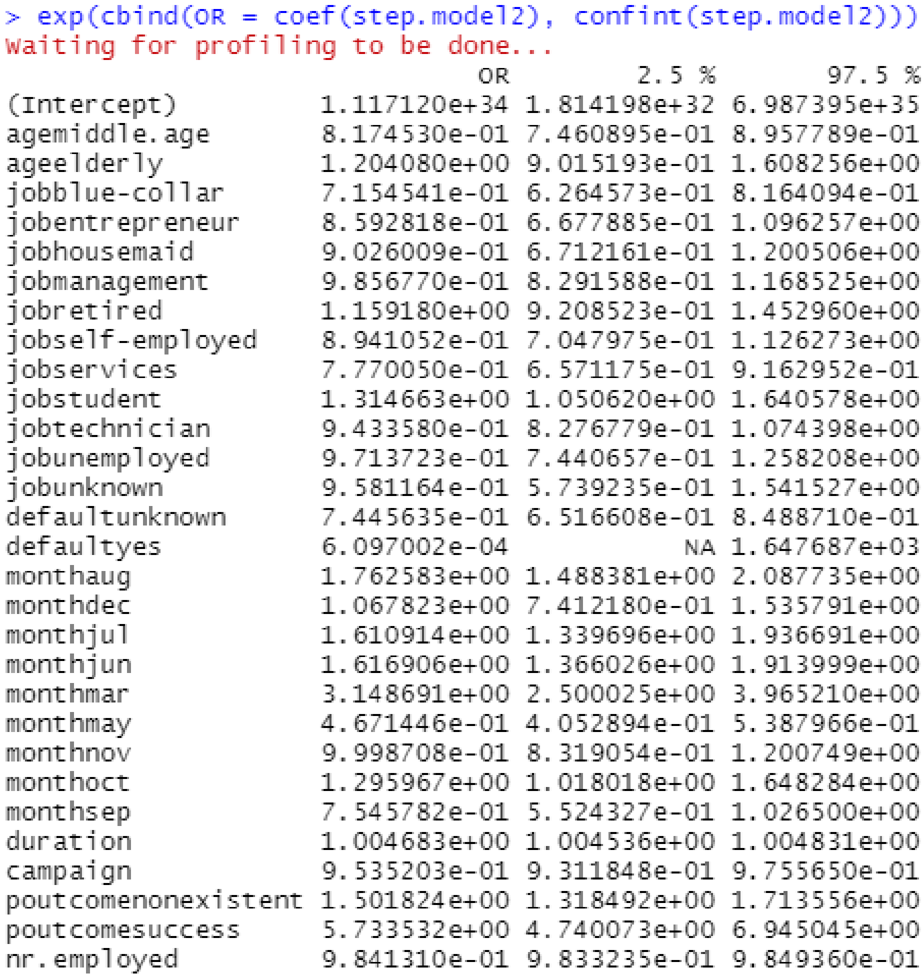
```
> anova(step.model1, step.model2, test = "LRT") # for comparison between the two using Likelihood Ratio Tests

Analysis of Deviance Table

Model 1: SUBSCRIBED ~ age + default + month + duration + campaign + poutcome + 
    nr.employed
Model 2: SUBSCRIBED ~ age + job + default + month + duration + campaign + 
    poutcome + nr.employed
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1     39852      15946                          
2     39841      15899 11   47.062 2.097e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3.9.: Anova Likelihood ratio test between step.model1 and step.model2



```
> vif(step.model2)
                GVIF Df GVIF^(1/(2*Df))
age         2.060147  2        1.198049
job         2.231879 11        1.037167
default     1.115879  2        1.027790
month       2.070889  9        1.041272
duration    1.208664  1        1.099393
campaign    1.040916  1        1.020253
poutcome    1.256872  2        1.058822
nr.employed 2.339173  1        1.529435
```

Table 3.10.: Multicollinearity test for step.model2



```
> exp(cbind(OR = coef(step.model2), confint(step.model2)))
Waiting for profiling to be done...
                              OR         2.5 %       97.5 %
(Intercept)         1.117120e+34 1.814198e+32 6.987395e+35
agemiddle.age       8.174530e-01 7.460895e-01 8.957789e-01
ageelderly          1.204080e+00 9.015193e-01 1.608256e+00
jobblue-collar      7.154541e-01 6.264573e-01 8.164094e-01
jobentrepreneur     8.592818e-01 6.677885e-01 1.096257e+00
jobhousemaid        9.026009e-01 6.712161e-01 1.200506e+00
jobmanagement       9.856770e-01 8.291588e-01 1.168525e+00
jobretired          1.159180e+00 9.208523e-01 1.452960e+00
jobself-employed    8.941052e-01 7.047975e-01 1.126273e+00
jobservices         7.770050e-01 6.571175e-01 9.162952e-01
jobstudent          1.314663e+00 1.050620e+00 1.640578e+00
jobtechnician       9.433580e-01 8.276779e-01 1.074398e+00
jobunemployed       9.713723e-01 7.440657e-01 1.258208e+00
jobunknown          9.581164e-01 5.739235e-01 1.541527e+00
defaultunknown      7.445635e-01 6.516608e-01 8.488710e-01
defaultyes          6.097002e-04            NA 1.647687e+03
monthaug            1.762583e+00 1.488381e+00 2.087735e+00
monthdec            1.067823e+00 7.412180e-01 1.535791e+00
monthjul            1.610914e+00 1.339696e+00 1.936691e+00
monthjun            1.616906e+00 1.366026e+00 1.913999e+00
monthmar            3.148691e+00 2.500025e+00 3.965210e+00
monthmay            4.671446e-01 4.052894e-01 5.387966e-01
monthnov            9.998708e-01 8.319054e-01 1.200749e+00
monthoct            1.295967e+00 1.018018e+00 1.648284e+00
monthsep            7.545782e-01 5.524327e-01 1.026500e+00
duration            1.004683e+00 1.004536e+00 1.004831e+00
campaign            9.535203e-01 9.311848e-01 9.755650e-01
poutcomenonexistent 1.501824e+00 1.318492e+00 1.713556e+00
poutcomesuccess     5.733532e+00 4.740073e+00 6.945045e+00
nr.employed         9.841310e-01 9.833235e-01 9.849360e-01
```

Table 3.11.: Interpretation of coefficients as odds-ratio along with confidence intervals