MSDS 6372 Project 2 - Using classification methods to determine an outcome of a telemarketing campaign.

Swee K Chew, Rene Pineda, Volodymyr Orlov

Introduction

While telemarketing might be considered as a cornerstone of modern advertising strategies by some companies, its role is highly questionable and sometimes it is viewed as a total waste of resources by others[1]. Here we attempt to analyze the effect of telemarketing on attracting new clients in a finance industry by looking at the success of telemarketing calls for selling bank long-term deposits recorded by a Portuguese retail bank. We apply multiple statistical methods and analyze outcomes of various models:

1. Logistic Regression models. We've built two models which can be used to predict our binary outcome and to interpret the effects of predictors.
2. We found that linear Discriminant Analysis model performed very poorly, in part due to the fact that only a handful of variables in the dataset are continuous.
3. Non-parametric model. This model has a slightly better performance than the Linear Regession models but is not easily interpretable.

Data Description

Our group focused on the Portuguese Bank Marketing data set[2]. The data is a result of a direct marketing campaign performed by a Portuguese bank. The bank collected data from May 2008 to November 2010 and the data consist of 45,211 observations and 17 variables. The target response is a binary, categorical variable indicating whether a client subscribed to a term deposit or not. For a complete list of variables please refer to Table 1.

The count plot of the binary response variable of the original dataset in Figure 1 suggests that the data is unbalanced. The number of 'no' responses is disproportionately higher than the 'yes' responses.

We decided to train all our models on balanced and unbalanced samples taken from the original data to find out whether the results would be different.
For the balanced sample, we took a random sample of 2500 'yes' and 2500 'no' responses. For the unbalanced sample, we simply randomly chose 5000 data points from our original dataset.

For our test sample, we selected 1000 data points which do not overlap with either the balanced, or unbalanced samples.

All dataset turned out to be clean and no imputation was nesessary.

Exploratory Data Analysis

For our analysis, we have used all variables. We separate variables into categorical and continuous and examine each group separately.

First, we explore the original full dataset. By looking at the histograms and boxplots of the continuous variables in Figure 2 and Figure 3, we find that *duration*, *balance*, *campaign* , *pdays* and *previous* variables might have an impact on our binary response, while *age* has no apparant effect on it. Also, the count plots of

[1] https://www.prospectresearch.co.uk/blog/telemarketing-still-effective/
[2] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

the categorical variables in Figure 4 indicate that *day* does not seem to have an impact on our dependent variable but the rest of the variables do.

After we obtain the balanced training sample, we examine the frequency tables of the categorical variables and the summary statistics of the continuous variables for further exploratory analysis.

Figure 5 shows the counts and percentage frequencies of the categorical variables for each factor level by the response variable. It lets us see if a specific level or group of a factor has a higher or lower count than its counterparts that might contribute to the likelihood of subscribing a term deposit.

The proportion of clients who subscribed to a term deposit seems to vary by job categories even for those with roughly the same sample size. For example, the proportion of subscribing to a term deposit is higher for clients who hold an administrative position and the proportion is lower for individuals who are self-employed. Thus, *job* possibly has an effect on the likelihood of a client subscribing to a term deposit.

Reviewing the frequency tables for the remaining categorical variables (Figure 6-13), it appears that all the variables could contribute, the proportions vary across the factor levels within each variable.

Figure 14 displays the summary statistics for each continuous variables by the response variable y, which allows us to see if there are any differences in characteristics between clients who subscribe a term deposit and who do not.

Except the *age* and *day* variables, the mean of the remaining continuous variables varies between two response groups. We decided not to transform any of these variables in order to build a simplier model that is easy to interpret.

Baseline model. Logistic Regression.

For our first two models, we fit logistic regression to balanced and unbalanced datasets. We estimate the performance of both models on the same test dataset to see if one model has a better predictive power than the other.

Model Assumptions

In this section, we assess whether the model's assumptions required for logistic regression analysis are met.

We use the Hosmer and Lemeshow Goodness-of-Fit test with the null hypothesis that the fitted model is correct. The output p-value is a number between 0 and 1 with higher values indicating a better fit. The p-value we obtain from the test is <0.0001 (Figure 15), which is statistically significant and implies that the null hypothesis should be rejected. Paul D. Allison, however, shows in his paper that the Hosmer and Lemeshow test is not accurate enough to evaluate model's fit[3]. Moreover, since our goal is to measure the predictive power of a model and not the goodness of fit, we will proceed despite not meeting the assumption.

We also look at the residual diagnostics for any potential leverage points. Figure 16 displays some of the residual and influential plots from the SAS output. When we review all the influential plots, there seems to be no leverage points.

Logistic regression also requires that there is little or no multicollinearity among the explanatory variables. The matrix scatter plot in Figure 17 and the correlation matrix in Figure 18 indicate that the continuous variables are not highly correlated with each other.

We assume that observations are independent of one another. Since the required assumptions have been addressed, we will proceed with model fitting.

[3]https://support.sas.com/resources/papers/proceedings14/1485-2014.pdf

Model Fit

First, the overall test is performed to test the null hypothesis that at least one coefficient is different from 0. Using the Likeliness Ratio test, we reject the null hypothesis at the significant level of 0.05 and conclude that the overall model is significan with the p-value <0.0001 (Figure 19).

We then include all the main effects, both categorical and continuous variables, to see which predictors are significant. Figure 20 (left) shows the output with all the main effects and their respective p-values. Based on the results, *education*, *default*, *age*, *balance*, *pdays*, and *previous* are non-significant at the alpha level of 0.05. Thus, we remove these predictors and refit the model. The new output is shown in Figure 20 (right).

Parameter Interpretation

Figure 21 displays the coefficient estimates for each factor level and Figure 22 displays the odd ratio estimates and the confidence intervals for each level. Here is our interpretation of a subset of most interesting estimates.

Job [categorical]:

The odds ratio of subscribing to a term-deposit for clients with unknown job title relative to clients who are entrepreneurs is 0.684 after accounting for other variables. The 95% confidence interval is [0.203,2.302]. In other words, the odds for someone with unknown job title to subscribe a term-deposit is 31.6% less than the odds for an entrepreneur.

Marital [categorical]:

The odds ratio for a single client subscribing a term-deposit relative to a married client is 0.727 after accounting for other variables. The 95% confidence interval is [0.607,0.870]. In other words, the odds for a single client to subscribe a term-deposit is 27.3% less than the odds for a married client.

Housing [categorical]:

The odds ratio of subscribing a term-deposit for clients with a housing loan relative to clients without a housing loan is 2.047 after accounting for other variables. The 95% confidence interval is [1.710,2.451]. In other words, the odds for someone with a housing loan to subscribe a term-deposit is 104.7% higher than the odds for someone without a housing loan.

Loan [categorical]:

The odds ratio of subscribing a term-deposit for clients with a personal loan relative to clients without a housing loan is 1.581 after accounting for other variables. The 95% confidence interval is [1.239,2.019]. In other words, the odds for someone with a personal loan to subscribe a term-deposit is 58.1% higher than the odds for someone without a personal loan.

Contact [categorical]:

The odds ratio of subscribing a term-deposit for clients whose contact communication type are unknown relative to clients who are communicated via cellular phone is 4.478 after accounting for other variables. The 95% confidence interval is [3.358,5.971]. In other words, the odds for someone who is contacted via an unknown method to subscribe a term-deposit is 347.8% higher than the odds for someone who is contacted via cellular.

Month [categorical]:

The odds ratio of subscribing a term-deposit for clients who are last contacted in September relative to those who are last contacted in November is 0.080 after accounting for other variables. The 95% confidence interval is [0.042,0.156]. In other words, the odds for a client who is last contacted in September to subscribe a term-deposit is 92% less than the odds for a client who is last contacted in November.

Poutcome [categorical]:

The odds ratio of subscribing a term-deposit for clients with the unknown previous marketing campaign outcome relative to clients with the failure previous marketing campaign outcome is 1.566 after accounting for other variables. The 95% confidence interval is [1.233,1.989]. In other words, the odds for a client with the unknown previous marketing campaign outcome to subscribe a term-deposit is 56.6% higher than the odds for a client with the failure previous marketing campaign outcome.

Day [Continuous]:

For every 1 unit increases in last contact day of the month, the odds of a client subscribing a term-deposit will increase by a multiplicative factor of 1.013 holding all other variables constant. The odds ratio (for a clients with the last contact day on the 15th compared to the 14th) is 1.013. The 95% confidence interval is [1.002,1.023].

Duration [Continuous]:

The odds of a client subscribing a term-deposit for a client is 1.006 times higher than a client whose last contact duration is 1 second less after accounting for other variables. The 95% confidence interval is [1.005,1.006]. In other words, for every minute increase in the duration of last contact, the odds of a client subscribing a term-deposit will increase by a multiplicative factor of 1.409 (exp[60*0.00572]) holding all other variables constant.

Campaign [Continuous]:

For every 1 unit increases in number of contacts performed during the campaign, the odds of a client subscribing a term-deposit will decrease by a multiplicative factor of 0.0894 holding all other variables constant. The odds ratio (10 contacts made compared to 11 contacts) is 0.0894. The 95% confidence interval is [0.859,0.929].

Prediction Performance

Using the resulting model from the logistic regression, we examine the ROC curve on the balanced training dataset and also on the test dataset for the predictability power of the model.

Figure 23 shows the ROC curve of the training dataset (top) and the ROC curve on the test dataset (bottom). The area under the curve (AUC) is commonly used to assess the prediction performance of the logistics model, the closer it's to 1, the better the prediction is. The AUC based on the training data is 0.9096 and 0.9124 for the test data, which indicates that we did not overfit the model and the predicitibility power of the model is quite high.

The classification tables in Figure 24 can also be used to assess how well the model performs in classifying the dichotomous response variable. The accuracy is measured by its sensitivity (the ability to predict an event correctly) and specificity (the ability to predict a nonevent correctly). At the probability level of 0.5, the model can correctly classify 81.1% of the event and 84.2% of the non-event, with an overall rate of 82.7% on the training data. For the test data, the sensitivity drops to 29.3%, with more false positive

predictions of 54.1% of the event. However, the specificity and the overall accuracy increase to 97.9% and 93.9% respectively.

It could be the results of having very low counts of 'yes' responses in the test dataset and setting the probability cutpoint to 0.5. The test data contains only 58 'yes' records out of 1000 observations. We could adjust the cutpoint to predict more events correctly but at the expense of more false predictions.

Using Unblanced Training Dataset

The analyses we have done so far are based on the balanced training dataset. We would like to find out if we will get a different logistic regression model if the training dataset is unbalanced, thus we repeat the analyses using the unbalanced training dataset.

Due to the disproportionate sample size ratio of approximately 1:7 (yes:no), it's difficult to determine whether any of the variables have an influence on the likelihood of a client subscribing a term deposit just by looking at the frequency tables and the summary statistics table. Thus, we simply include all the variables in the model and let it decide which predictors are significant.

At the significant level of 0.05, *default*, *age*, *balance*, *pdays*, and *previous* are non-significant (Figure 25 (left)). The *education* variable is statistically significant here, whereas it was shown non-significant in the prior model under the balanced dataset. We then remove the non-significant predictors and refit the model, the output is shown in Figure 25 (right).

Using the resulting model that is built with the unbalanced dataset, we examine the ROC curve of the training dataset and also on the same test dataset to determine the predictability power of the model.

Figure 26 (top) illustrates the ROC curve on the traning dataset and Figure 26 (bottom) displays the ROC curve on the test dataset. The AUC is 0.9012 for the model based on the training data and 0.9054 for the test data. The values are slighly lower than those that are obtained from the balanced model respectively.

The classification table in Figure 27 (top) displays the sensitivity and the specificity of the model. At the probability level of 0.5, the model can correctly classify 31.9% of the event and 97.2% of the non-event, with an overall rate of 89.2% on the training data. For the test data, the sensitivity drops to 27.6%, with more false positive predictions of 54.3% of the event. However, the specificity and the overall accuracy increase to 93.9% and 98.0% respectively.

Compared to the prior model with the balanced training data, the sensitivity is much lower and the specificity is higher, which makes sense since the latter model is built based on the disproportionate ratio of 'no' and 'yes' responses, having a much higher observations of 'no' than 'yes'. Thus, the model can more accurately classify the nonevents resulting in higher specificity. On the other hand, the sensitivity is low due to the small number of 'yes' records in the training dataset. Thus, there is not enough information for the model to correctly classify the event.

Since the prediction accuracy is better with the balanced training data, we will only use the balanced data in fitting additional models and for further analyses.

Additional Models

Logistic Regression model (LRM) with transformed variables

Motivation:

The transformation of varaibles for this objective is focused on creating categories for the continuous variables. This responds to two reasons:

- A logistic regression model will typically assign a weight to a continuous feature, and always think that every feature is either positively or negatively related to the outcome variable. However, for some

variables (for example *balance*) the feature might be positively related with the outcome for one range of values, and negatively related for other ranges. Discretization of continuous features is simple but is a useful way to include additional information that might solve this problem, and we will pay special attention to those variables that were deemed as non-significant by the baseline model.

- In other instances, the creation of categorical variables responds to the need of highlighting information that is hidden or not explicit in the continuous variable. For example, the variable *pdays* contains information about whether a client was previously contacted or not, but because this is indicated by assigning the variable a value of -1, this information can't be picked up by a regression model that uses the continuous variable.

The upper and lower limits for the categories are based on our analysis of the distribution of the features.

Age:

Create three categories: Adult (up to 35 yo), Middle aged (36 to 60 yo), and Elderly (65 yo and more)

Balance:

Create categories for negative balance, zero balance, and 5 levels for positive balance: $0 to $100, $101 to $500, $501 to $2,000, $2,0001 to $10,000, and more than $10,000.

Campaign:

Create categories for clients that were contacted only once or twice and for those who were contacted more than two times during the campaign.

pdays:

Add variable to indicate whether a client was previously contacted or not. Additionally, convert days to months and create three categories depending on how much time had passed since the client was last contacted.

previous:

Create a category for clients that were previously contacted only once or twice and another category for those who were contacted more times.

Model Building

To create the logistic regression model, we use the Glmnet package in R, which fits a generalized linear model via penalized maximum likelihood. We perform a cross-validation fit, which is shown in Figure 28. The potential model that minimizes the misclassification error includes between 23 to 42 features.

We can generate a list of the coefficients for the value of lambda that gives minimum mean cross-validated error, which is part of the output of the Glmnet package. By examining these coefficients, we can tell that the new categories we created for the *balance*, *campaign*, and *pdays* variables are not selected by the model, similar to the results of the first model we produced. However, it is interesting to notice that the selection process picks up the negative balance, zero balance, balance between $2,000 and $10,000, and balance greater than $10,000 as important, indicating that splitting *balance* into categorical variables is useful.

Prediction Performance

Regarding predictive accuracy, the model with the new categorical variables shows a similiar performance compared to the first model developed on the balanced dataset. The model achieves an accuracy of 82.4% for the training set and 85.9% for the test set. This new model shows the same performance when we measure the AUC indicator, which is the same than the first logistic regression model (0.912). The ROC curve of the new model is shown in Figure 29.

Linear Discriminant Analysis model (LDA)

Next, we develop a LDA model using only the continuous predictors. This poses a serious challenge because only 6 variables in the dataset are continuous. Additionally, the logistic regression models we have developed indicate that out of these 6 variables, four are non-significant (*age*, *balance*, *pdays* and *previous*) and their predictive power is low, although one continuous variable (*duration*) is perhaps the strongest predictor of all.

Likely due to these limitations, the LDA model performs poorly. Examining the confusion matrix, we conclude that the model has an accuracy of only 74.1% on the training set and 64.7% on the testing set, much lower than the logistic regression models. Based on the ROC curve in Figure 30, the AUC score of 0.805 is also much lower compared to the logistic regression models.

The conclusion from the LDA model is that the two categories in the outcome binomial variable are not clearly separable on the continuous features. Thus, the LDA model is not an appropriate method for this specific dataset.

Non-parametric model. Random Forest (RF)

The third additional model we developed is based on the Random Forest package in R. To make this model works smoothly, we decided to modify the original dataset as follows: i) Create dummy variables for all the categorical predictors, ii) Create an additional dummy variable that indicates whether the client was previously contacted or not.

Reviewing the prediction accuracy and the AUC of the random forest model in ROC curve (Figure 31), it performs significantly better than LDA model, and marginally better than the logistic regression models. The accuracy of the model is 85.4% on the training set and 82.6% on the testing set. This model is especially good at predicting the 'yes' cases (clients who will actually sign up for the term deposit), with a sensitivity of around 88% on both the training and testing sets. However, this model performs relatively poorly on the overall accuracy on the test set, which is brought down by a poor sensitivity rate. This might be due to the fact that the proportion of 'yes' in the test set was very low.

Other disadvantages of this model are as follows:

- The model is not easily interpretable: we can have an idea about which factors impact the outcome by using the "importance" function, which displays the mean Gini gain produced by the X's over all trees, and the mean decrease in classification accuracy after permuting X's over all trees. Based on this, we can assess which variables are more important on a relative scale, but there is no absolute measure of this and the model is not interpretable.
- May require some work to tune the model to the data: the Random Forest has two main tuning parameters: the number of trees created (ntrees, default 500), and the number of features that are randomly selected at each split (mtry, default = the sq root of the number of features). The model we ran has the default parameters. We attempt to tune the model by increasing and decreasing the parameters, however, we do not obtain a better overall performance.

Comparison of all models and Conclusion

Table 2 shows a comparison of the performance of the four models that are fitted on the balanced data, along with four metrics.

Based on this information, we can conclude that the random forest model is suitable if the goal is to obtain the model with high predictive power. The limitation of low interpretability of the results can be overcome by understanding how different factors affect the outcome, as explained in the Exploratory Data Analysis section. However, if the interpretability is crucial, one could use the logistic regression models to better understand how individual factor levels can influence the likelihood a client subscribing to a term deposit.

Code

All codes used to generate models, plots and report related to this work can be found in https://github.com/VolodymyrOrlov/MSDS6372_Project2

Tables and Figures

Variable Name	Variable Type	Description
job	categorical	type of job ('admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
marital	categorical	marital status ('divorced', 'married', 'single', 'unknown')
education	categorical	'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
default	categorical	has credit in default? ('no', 'yes', 'unknown')
housing	categorical	has housing loan? ('no', 'yes', 'unknown')
loan	categorical	has personal loan? ('no', 'yes', 'unknown')
contact	categorical	contact communication type ('cellular', 'telephone')
month	categorical	last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec')
poutcome	categorical	outcome of the previous marketing campaign ('failure', 'nonexistent', 'success')
age	continuous	age of the contact
balance	continuous	average yearly balance, in euros
day	continuous	last contact day
duration	continuous	last contact duration, in seconds
campaign	continuous	number of contacts performed during this campaign and for this client
pdays	continuous	number of days that passed by after the client was last contacted from a previous campaign
previous	continuous	number of contacts performed before this campaign and for this client

Table 1: List of variables.

Summary	Training Set Statistics			Test Set Statistics			
Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
LR model 1	0.827	0.811	0.842	0.939	0.293	0.979	0.912
LR model 2	0.824	0.796	0.852	0.859	0.863	0.793	0.912
LDA	0.741	0.658	0.824	0.647	0.827	0.636	0.805
Random Forest	0.854	0.884	0.825	0.826	0.879	0.825	0.923

Table 2: Performance characteristics of all models.

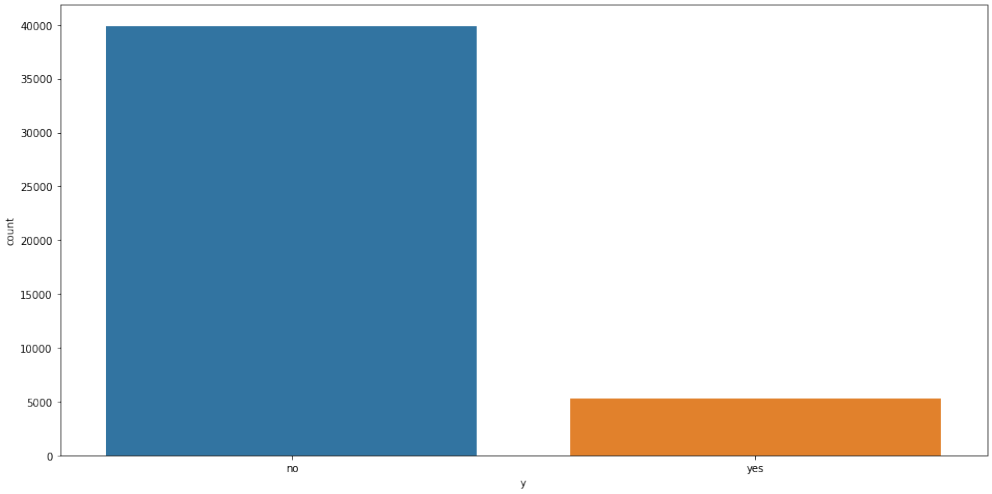


Figure 1: Count plot of the response variable.

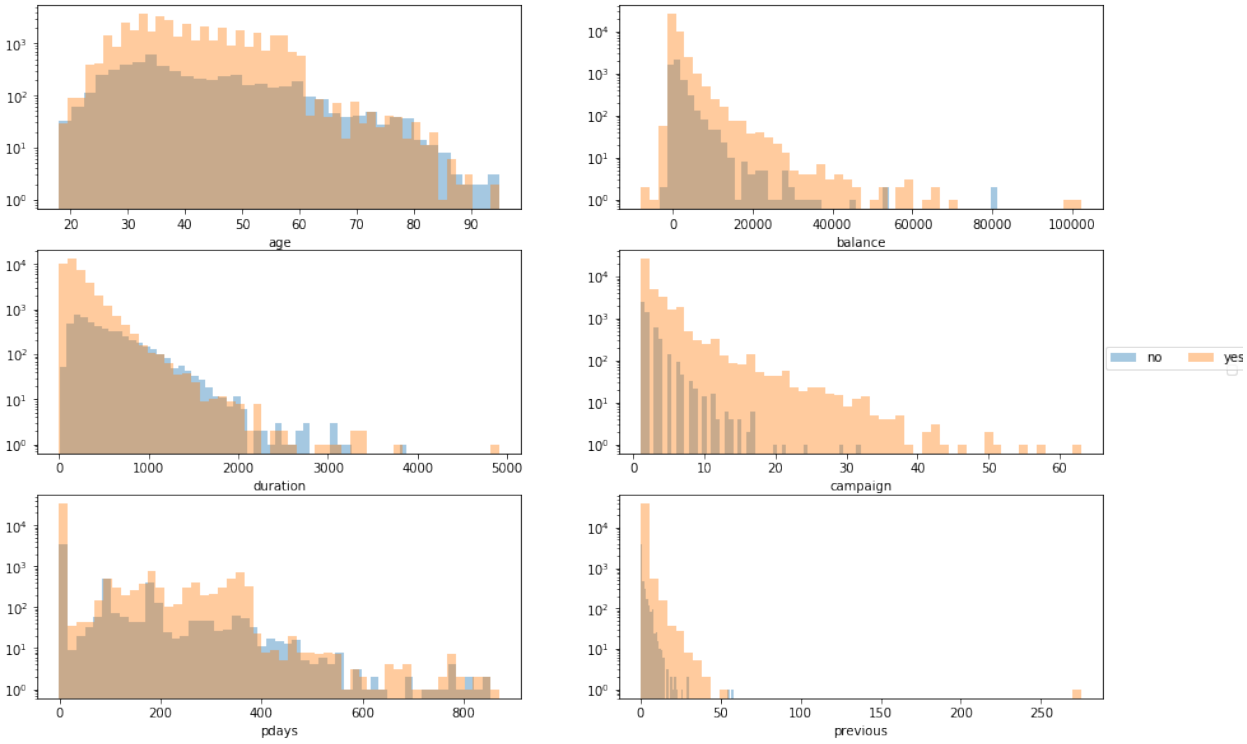


Figure 2: Histograms of continuous variables.

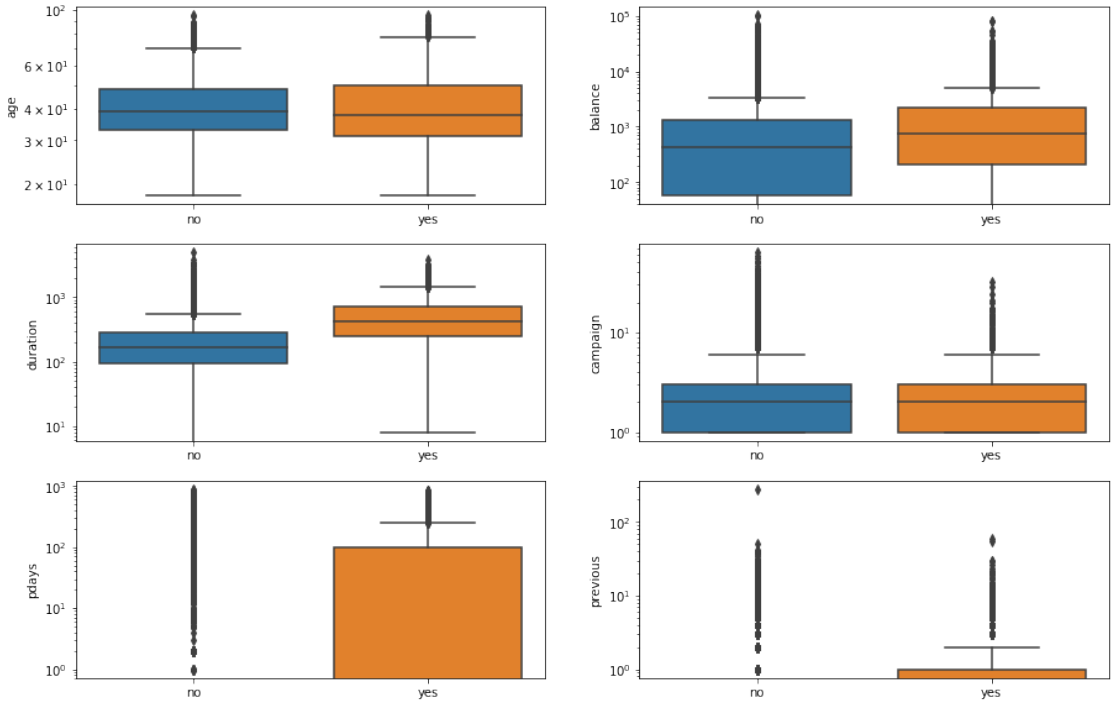


Figure 3: Boxplots of continuous variables.

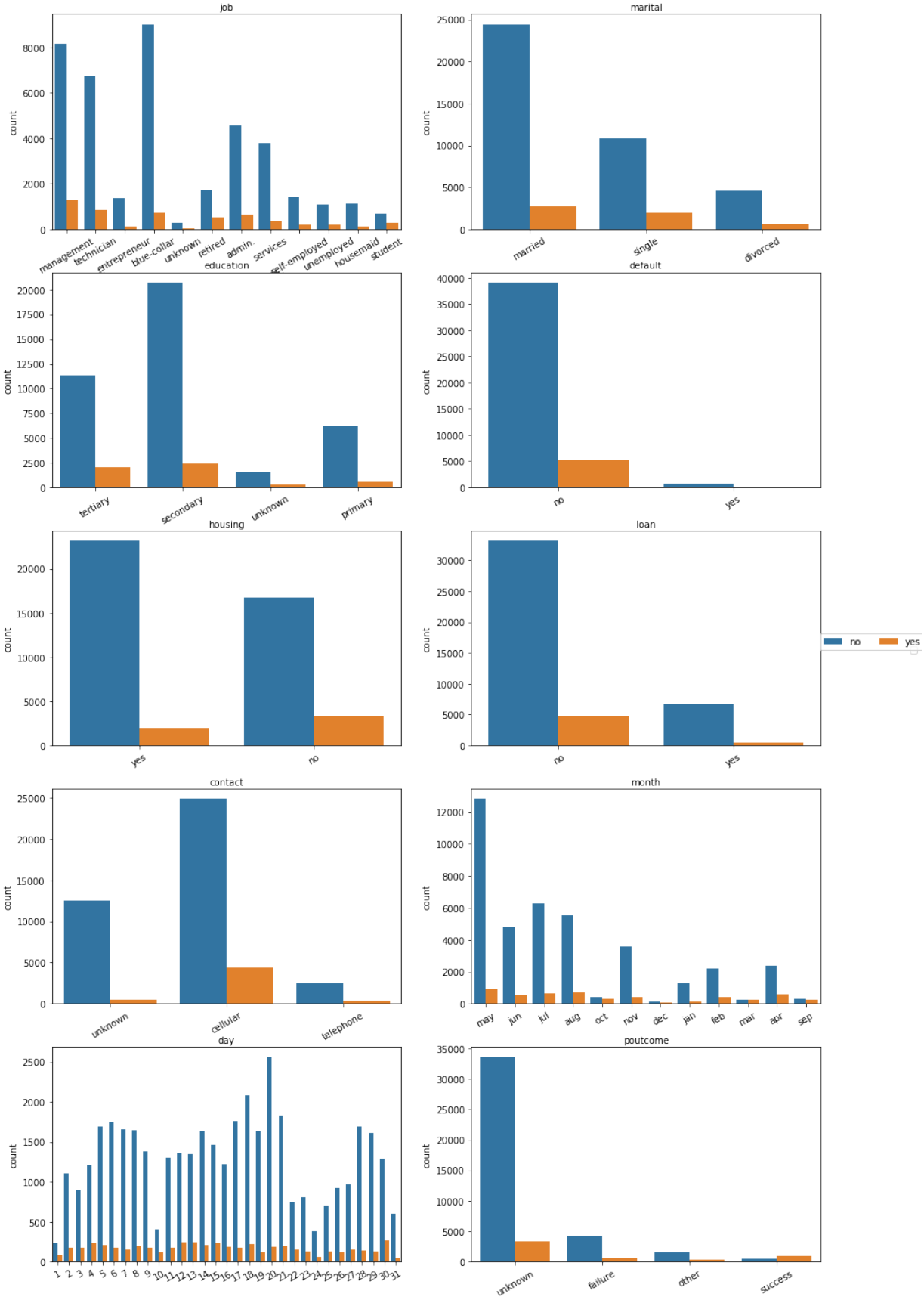


Figure 4: Count plots of categorical variables.

Frequency
Percent
Row Pct
Col Pct

Table of job by y			
job	y		
	no	yes	Total
admin.	274 5.48 47.74 10.96	300 6.00 52.26 12.00	574 11.48
blue-collar	569 11.38 63.50 22.76	327 6.54 36.50 13.08	896 17.92
entrepreneur	84 1.68 60.00 3.36	56 1.12 40.00 2.24	140 2.80
housemaid	70 1.40 61.95 2.80	43 0.86 38.05 1.72	113 2.26
management	534 10.68 46.19 21.36	622 12.44 53.81 24.88	1156 23.12
retired	106 2.12 29.53 4.24	253 5.06 70.47 10.12	359 7.18
self-employed	92 1.84 53.18 3.68	81 1.62 46.82 3.24	173 3.46
services	224 4.48 55.17 8.96	182 3.64 44.83 7.28	406 8.12
student	44 0.88 26.04 1.76	125 2.50 73.96 5.00	169 3.38
technician	425 8.50 51.02 17.00	408 8.16 48.98 16.32	833 16.66
unemployed	68 1.36 42.77 2.72	91 1.82 57.23 3.64	159 3.18
unknown	10 0.20 45.45 0.40	12 0.24 54.55 0.48	22 0.44
Total	2500 50.00	2500 50.00	5000 100.00

Figure 5: Frequency table of job type by the response variable.

Table of marital by y			
	y		
marital	no	yes	Total
divorced	263 5.26 46.30 10.52	305 6.10 53.70 12.20	568 11.36
married	1497 29.94 53.66 59.88	1293 25.86 46.34 51.72	2790 55.80
single	740 14.80 45.07 29.60	902 18.04 54.93 36.08	1642 32.84
Total	2500 50.00	2500 50.00	5000 100.00

Figure 6: Frequency table of marital status by the response variable.

Table of education by y			
	y		
education	no	yes	Total
primary	390 7.80 57.18 15.60	292 5.84 42.82 11.68	682 13.64
secondary	1276 25.52 52.51 51.04	1154 23.08 47.49 46.16	2430 48.60
tertiary	753 15.06 44.53 30.12	938 18.76 55.47 37.52	1691 33.82
unknown	81 1.62 41.12 3.24	116 2.32 58.88 4.64	197 3.94
Total	2500 50.00	2500 50.00	5000 100.00

Figure 7: Frequency table of education level by the response variable.

Table of default by y			
	y		
default	no	yes	Total
no	2443 48.86 49.65 97.72	2477 49.54 50.35 99.08	4920 98.40
yes	57 1.14 71.25 2.28	23 0.46 28.75 0.92	80 1.60
Total	2500 50.00	2500 50.00	5000 100.00

Figure 8: Frequency table of default (has credit or not) by the response variable.

Table of housing by y			
	y		
housing	no	yes	Total
no	1056 21.12 40.31 42.24	1564 31.28 59.69 62.56	2620 52.40
yes	1444 28.88 60.67 57.76	936 18.72 39.33 37.44	2380 47.60
Total	2500 50.00	2500 50.00	5000 100.00

Figure 9: Frequency table of housing loan by the response variable.

Table of loan by y			
	y		
loan	no	yes	Total
no	2077 41.54 47.87 83.08	2262 45.24 52.13 90.48	4339 86.78
yes	423 8.46 63.99 16.92	238 4.76 36.01 9.52	661 13.22
Total	2500 50.00	2500 50.00	5000 100.00

Figure 10: Frequency table of personal loan by the response variable.

Table of contact by y			
	y		
contact	no	yes	Total
cellular	1568 31.36 43.16 62.72	2065 41.30 56.84 82.60	3633 72.66
telephone	165 3.30 46.61 6.60	189 3.78 53.39 7.56	354 7.08
unknown	767 15.34 75.72 30.68	246 4.92 24.28 9.84	1013 20.26
Total	2500 50.00	2500 50.00	5000 100.00

Figure 11: Frequency table of contact type by the response variable.

Table of month by y			
	y		
month	no	yes	Total
apr	133 2.66 33.67 5.32	262 5.24 66.33 10.48	395 7.90
aug	353 7.06 52.22 14.12	323 6.46 47.78 12.92	676 13.52
dec	9 0.18 17.31 0.36	43 0.86 82.69 1.72	52 1.04
feb	136 2.72 39.88 5.44	205 4.10 60.12 8.20	341 6.82
jan	77 1.54 53.47 3.08	67 1.34 46.53 2.68	144 2.88
jul	411 8.22 58.13 16.44	296 5.92 41.87 11.84	707 14.14
jun	318 6.36 54.08 12.72	270 5.40 45.92 10.80	588 11.76
mar	14 0.28 11.02 0.56	113 2.26 88.98 4.52	127 2.54
may	793 15.86 65.11 31.72	425 8.50 34.89 17.00	1218 24.36
nov	217 4.34 51.79 8.68	202 4.04 48.21 8.08	419 8.38
oct	25 0.50 14.04 1.00	153 3.06 85.96 6.12	178 3.56
sep	14 0.28 9.03 0.56	141 2.82 90.97 5.64	155 3.10
Total	2500 50.00	2500 50.00	5000 100.00

Figure 12: Frequency table of last contact month by the response variable.

Table of poutcome by y			
	y		
poutcome	no	yes	Total
failure	264 5.28 47.23 10.56	295 5.90 52.77 11.80	559 11.18
other	98 1.96 39.04 3.92	153 3.06 60.96 6.12	251 5.02
success	32 0.64 6.24 1.28	481 9.62 93.76 19.24	513 10.26
unknown	2106 42.12 57.27 84.24	1571 31.42 42.73 62.84	3677 73.54
Total	2500 50.00	2500 50.00	5000 100.00

Figure 13: Frequency table of the outcome of the previous marketing campaign by the response variable.

y	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
no	2500	age	2500	40.9488000	10.1197995	19.0000000	82.0000000
		balance	2500	1340.18	3515.47	-4057.00	102127.00
		day	2500	15.7592000	8.3140883	1.0000000	31.0000000
		duration	2500	223.5128000	204.1116716	5.0000000	2055.00
		campaign	2500	2.8908000	3.3139807	1.0000000	32.0000000
		pdays	2500	37.2648000	97.7754708	-1.0000000	791.0000000
		previous	2500	0.4876000	1.6345381	0	23.0000000
yes	2500	age	2500	41.7436000	13.6081421	18.0000000	95.0000000
		balance	2500	1806.64	3553.44	-1944.00	81204.00
		day	2500	15.1104000	8.4987001	1.0000000	31.0000000
		duration	2500	541.2540000	397.7814977	8.0000000	3881.00
		campaign	2500	2.1344000	1.9123283	1.0000000	24.0000000
		pdays	2500	71.0044000	120.7742346	-1.0000000	828.0000000
		previous	2500	1.1768000	2.3401992	0	26.0000000

Figure 14: Summary statistics of the continuous variables by the response variable.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
248.4426	8	<.0001

Figure 15: Hosmer and Lemeshow Goodness-of-Fit Test Result.

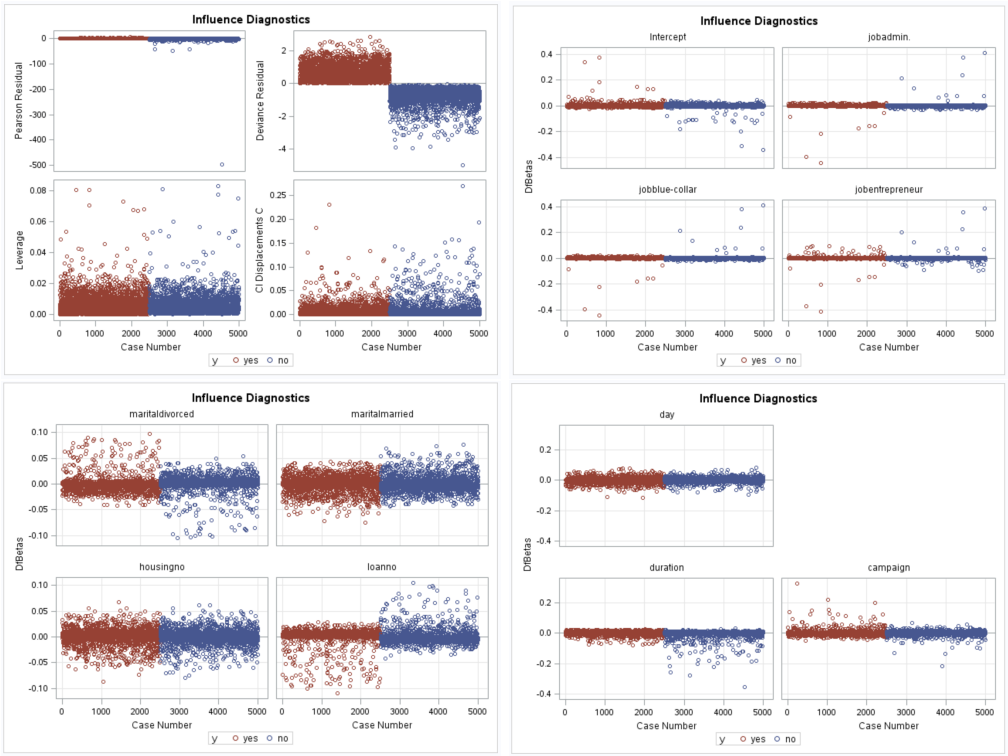


Figure 16: Residual and influential diagnostics plots.

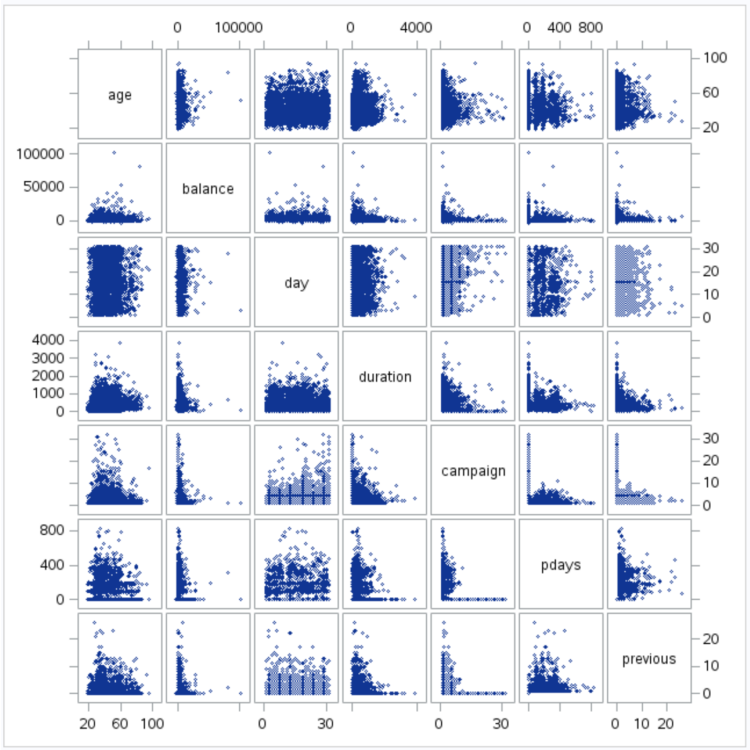


Figure 17: Matrix scatterplot of the continuous explanatory variables.

Pearson Correlation Coefficients, N = 5000 Prob > |r| under H0: Rho=0							
	age	balance	day	duration	campaign	pdays	previous
age	1.00000	0.11421 <.0001	0.00750 0.5962	0.00851 0.5473	0.01525 0.2811	0.00428 0.7622	0.03865 0.0063
balance	0.11421 <.0001	1.00000	0.00269 0.8493	0.00705 0.6183	-0.02037 0.1499	0.02621 0.0639	0.03650 0.0098
day	0.00750 0.5962	0.00269 0.8493	1.00000	-0.01949 0.1682	0.14773 <.0001	-0.05929 <.0001	-0.06073 <.0001
duration	0.00851 0.5473	0.00705 0.6183	-0.01949 0.1682	1.00000	-0.02972 0.0356	-0.04892 0.0005	-0.04073 0.0040
campaign	0.01525 0.2811	-0.02037 0.1499	0.14773 <.0001	-0.02972 0.0356	1.00000	-0.11413 <.0001	-0.07357 <.0001
pdays	0.00428 0.7622	0.02621 0.0639	-0.05929 <.0001	-0.04892 0.0005	-0.11413 <.0001	1.00000	0.51174 <.0001
previous	0.03865 0.0063	0.03650 0.0098	-0.06073 <.0001	-0.04073 0.0040	-0.07357 <.0001	0.51174 <.0001	1.00000

Figure 18: Correlation matrix of the continuous explanatory variables.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2977.1649	34	<.0001
Score	2166.9099	34	<.0001
Wald	1294.5344	34	<.0001

Figure 19: Overall test of Logistic Regression.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
job	11	35.2071	0.0002
marital	2	7.9349	0.0189
education	3	7.2142	0.0654
default	1	2.2222	0.1360
housing	1	59.3520	<.0001
loan	1	11.7028	0.0006
contact	2	99.0057	<.0001
month	11	275.6875	<.0001
poutcome	3	158.6775	<.0001
age	1	0.6323	0.4265
balance	1	1.2500	0.2636
day	1	5.3142	0.0212
duration	1	883.7383	<.0001
campaign	1	30.8742	<.0001
pdays	1	2.1870	0.1392
previous	1	1.3714	0.2416

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
job	11	39.5350	<.0001
marital	2	12.3050	0.0021
housing	1	60.8293	<.0001
loan	1	13.5356	0.0002
contact	2	104.6192	<.0001
month	11	280.7542	<.0001
poutcome	3	209.1446	<.0001
day	1	5.4485	0.0196
duration	1	885.8576	<.0001
campaign	1	31.8568	<.0001

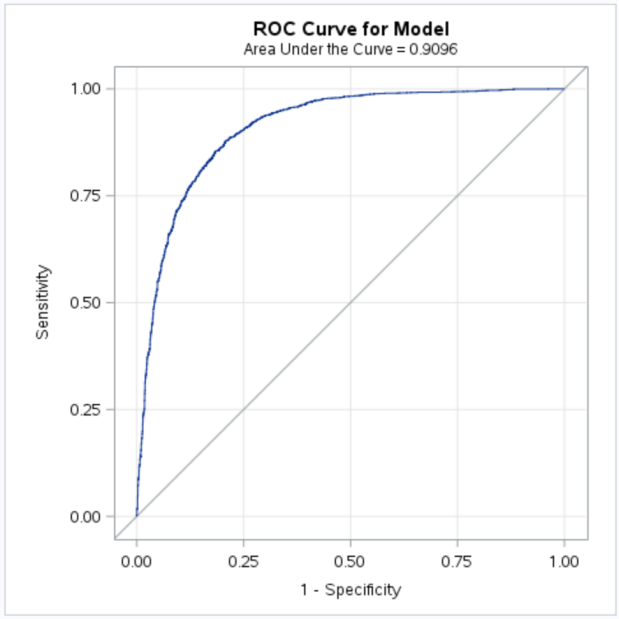
Figure 20: Type 3 analysis of effects with all the predictors (left) and with only the predictors that are significant (right) using the balanced dataset.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1555	0.6899	9.7622	0.0018
job	admin.	1	0.2225	0.5819	0.1463	0.7021
job	blue-collar	1	-0.3132	0.5791	0.2925	0.5886
job	entrepreneur	1	-0.3805	0.6195	0.3772	0.5391
job	housemaid	1	-0.7390	0.6258	1.3945	0.2376
job	management	1	0.1115	0.5745	0.0377	0.8461
job	retired	1	0.4006	0.5885	0.4632	0.4961
job	self-employed	1	-0.1287	0.6087	0.0447	0.8325
job	services	1	-0.0728	0.5876	0.0154	0.9013
job	student	1	0.6388	0.6154	1.0778	0.2992
job	technician	1	0.0348	0.5772	0.0036	0.9519
job	unemployed	1	0.1883	0.6105	0.0952	0.7577
marital	divorced	1	-0.1395	0.1413	0.9750	0.3234
marital	married	1	-0.3187	0.0918	12.0552	0.0005
housing	no	1	0.7163	0.0918	60.8293	<.0001
loan	no	1	0.4583	0.1246	13.5356	0.0002
contact	cellular	1	1.4992	0.1468	104.3097	<.0001
contact	telephone	1	1.4695	0.2098	49.0614	<.0001
month	apr	1	-1.3355	0.3362	15.7828	<.0001
month	aug	1	-2.2444	0.3275	46.9703	<.0001
month	dec	1	-0.9769	0.5351	3.3327	0.0679
month	feb	1	-1.5135	0.3368	20.1879	<.0001
month	jan	1	-2.8792	0.3862	55.5859	<.0001
month	jul	1	-2.5386	0.3302	59.1221	<.0001
month	jun	1	-1.3321	0.3378	15.5508	<.0001
month	mar	1	0.5403	0.4314	1.5686	0.2104
month	may	1	-2.3057	0.3249	50.3638	<.0001
month	nov	1	-2.5200	0.3375	55.7433	<.0001
month	oct	1	-0.3193	0.4007	0.6350	0.4255
poutcome	failure	1	0.4486	0.1221	13.5022	0.0002
poutcome	other	1	0.6441	0.1756	13.4551	0.0002
poutcome	success	1	2.9173	0.2057	201.1034	<.0001
day		1	0.0125	0.00537	5.4485	0.0196
duration		1	0.00572	0.000192	885.8576	<.0001
campaign		1	-0.1124	0.0199	31.8568	<.0001

Figure 21: Tables of Coefficient estimates.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
job admin. vs unknown	1.249	0.399	3.908
job blue-collar vs unknown	0.731	0.235	2.275
job entrepreneur vs unknown	0.684	0.203	2.302
job housemaid vs unknown	0.478	0.140	1.628
job management vs unknown	1.118	0.363	3.447
job retired vs unknown	1.493	0.471	4.730
job self-employed vs unknown	0.879	0.267	2.899
job services vs unknown	0.930	0.294	2.941
job student vs unknown	1.894	0.567	6.328
job technician vs unknown	1.035	0.334	3.209
job unemployed vs unknown	1.207	0.365	3.995
marital divorced vs single	0.870	0.659	1.147
marital married vs single	0.727	0.607	0.870
housing no vs yes	2.047	1.710	2.451
loan no vs yes	1.581	1.239	2.019
contact cellular vs unknown	4.478	3.358	5.971
contact telephone vs unknown	4.347	2.881	6.558
month apr vs sep	0.263	0.136	0.508
month aug vs sep	0.106	0.056	0.201
month dec vs sep	0.376	0.132	1.075
month feb vs sep	0.220	0.114	0.426
month jan vs sep	0.056	0.026	0.120
month jul vs sep	0.079	0.041	0.151
month jun vs sep	0.264	0.136	0.512
month mar vs sep	1.716	0.737	3.998
month may vs sep	0.100	0.053	0.188
month nov vs sep	0.080	0.042	0.156
month oct vs sep	0.727	0.331	1.594
poutcome failure vs unknown	1.566	1.233	1.989
poutcome other vs unknown	1.904	1.350	2.686
poutcome success vs unknown	18.491	12.356	27.674
day	1.013	1.002	1.023
duration	1.006	1.005	1.006
campaign	0.894	0.859	0.929

Figure 22: Tables of Odds Ratio estimates and confidence intervals.



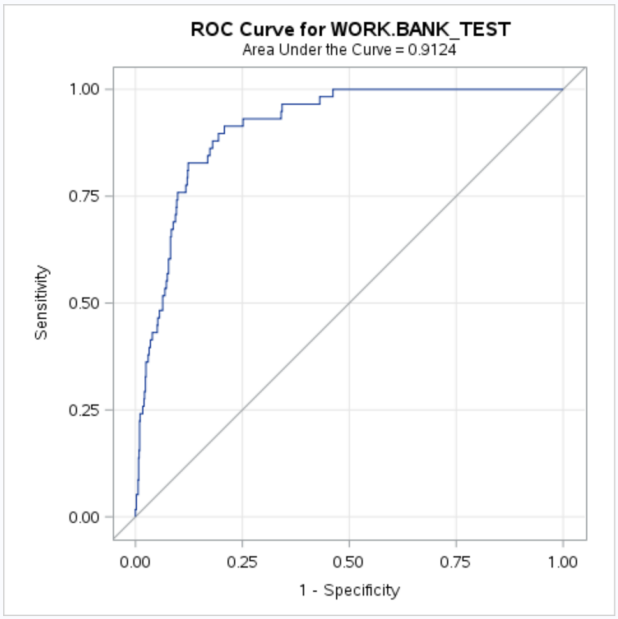


Figure 23: ROC curves for the balanced training dataset and the test dataset.

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	2028	2106	394	472	82.7	81.1	84.2	16.3	18.3

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	17	922	20	41	93.9	29.3	97.9	54.1	4.3

Figure 24: The classification table based on the balanced training (top) and test (bottom) datasets.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
job	11	27.1108	0.0044
marital	2	11.1156	0.0039
education	3	9.5613	0.0227
default	1	0.1290	0.7195
housing	1	16.2540	<.0001
loan	1	4.4419	0.0351
contact	2	66.7767	<.0001
month	11	118.8510	<.0001
poutcome	3	112.1621	<.0001
age	1	1.3277	0.2492
balance	1	1.2985	0.2545
day	1	5.4582	0.0195
duration	1	441.0185	<.0001
campaign	1	10.3696	0.0013
pdays	1	0.3409	0.5593
previous	1	0.3747	0.5404

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
job	11	31.1484	0.0010
marital	2	9.4299	0.0090
education	3	9.4554	0.0238
housing	1	18.4157	<.0001
loan	1	5.2017	0.0226
contact	2	66.2303	<.0001
month	11	119.1495	<.0001
poutcome	3	135.7497	<.0001
day	1	5.3815	0.0204
duration	1	444.9350	<.0001
campaign	1	10.5051	0.0012

Figure 25: Type 3 analysis of effects with all the predictors (left) and with only the predictors that are significant (right) using the unbalanced training dataset.

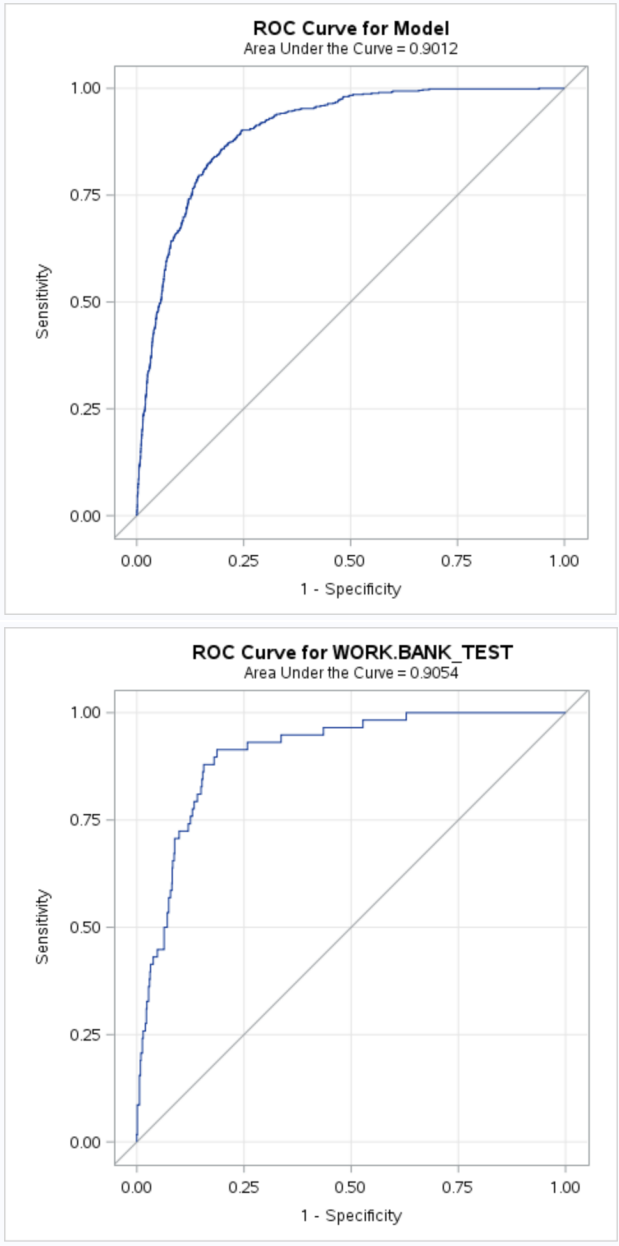


Figure 26: ROC curves for the unbalanced training dataset and the test dataset.

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	196	4264	122	418	89.2	31.9	97.2	38.4	8.9

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	16	923	19	42	93.9	27.6	98.0	54.3	4.4

Figure 27: The classification table based on the unbalanced training (top) and test (bottom) datasets.

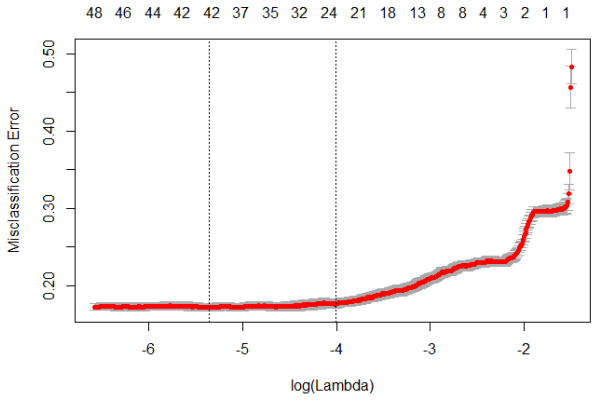


Figure 28: Misclassification error for the second LR model.

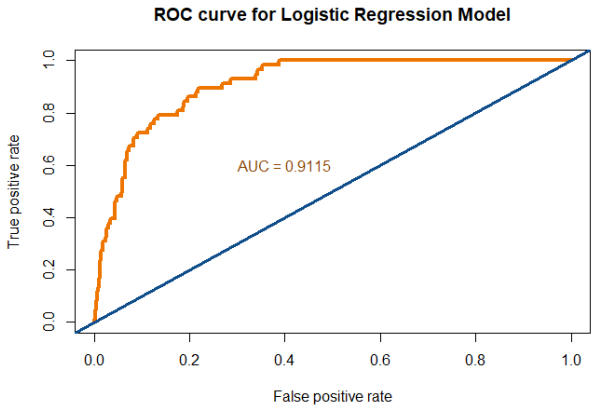


Figure 29: ROC curve for the second LR model.

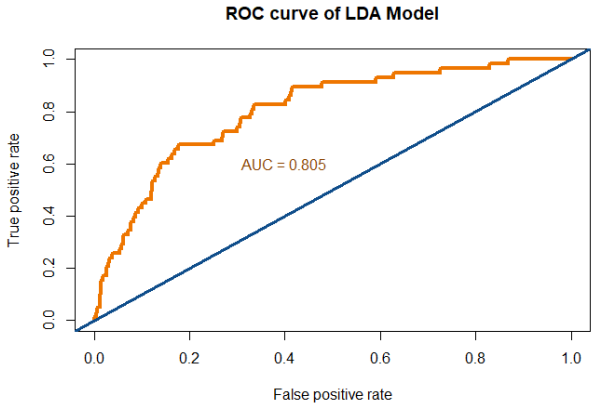


Figure 30: ROC curve for the LDA model.

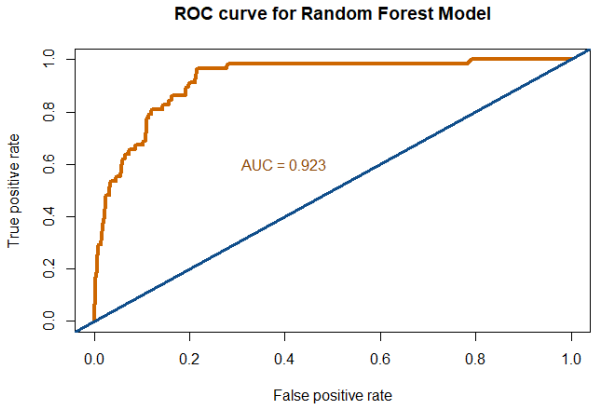


Figure 31: ROC curve for the Random Forest model.