**ClassIFIcation on Bank Marketing Data Set to predict client’s intention of a term subscription**

Contributors:

* Tina Pai
* Yang zang
* Balaji Avvaru

# Introduction

Marketing campaigns develop strategies to enhance businesses. Companies use direct marketing by targeting segments of customers and contacting them to meet specific goals. Centralizing customer remote interactions in a contact center eases operational management of campaigns. Such centers allow communicating with customers through channels such as telephone (fixed-line or mobile).

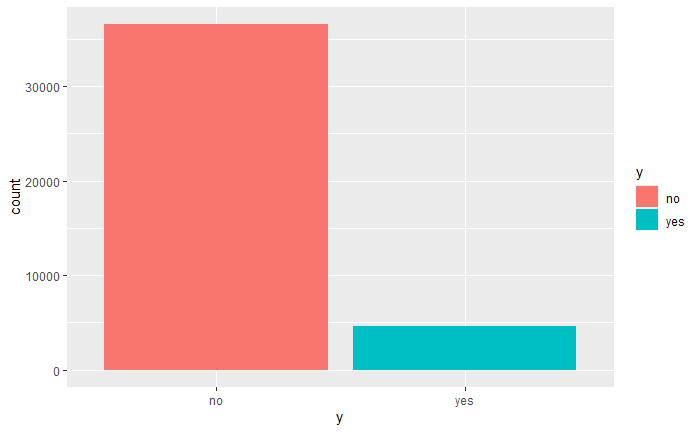
The success of bank marketing campaign is predicted with customer features, campaign information and economic attributes. Here we attempt to analyze the eﬀect of telemarketing on attracting new clients in a ﬁnance industry by looking at the success of telemarketing calls for selling bank long-term deposits recorded by a Portuguese retail bank.

# Data Description

We obtained dataset from UCI [<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>] containing data originally sourced from the direct marketing campaign performed by a Portuguese bank. The bank collected data from May 2008 to November 2010 and the data consist of 41188 observations and 21 variables. The target response is a binary, categorical variable indicating whether a client subscribed to a term deposit or not.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Description | Min | Max | Median | Mean |
| age | Age | 17 | 98 | 38 | 40.02 |
| job | Type of job:  'admin.',  'blue-collar',  'entrepreneur',  'housemaid',  'management',  'retired',  'self-employed',  'services',  'student',  'technician',  'unemployed',  'unknown' | N/A | N/A | N/A | N/A |
| marital | Marital status  'divorced',  'married',  'single',  'unknown'; | N/A | N/A | N/A | N/A |
| education | Education  'basic.4y',  'basic.6y',  'basic.9y',  'high.school',  'illiterate',  'professional.course',  'university.degree',  'unknown' | N/A | N/A | N/A | N/A |
| default | Has credit in default?  'no',  'yes',  'unknown' | N/A | N/A | N/A | N/A |
| housing | Has housing loan?  'no',  'yes',  'unknown' | N/A | N/A | N/A | N/A |
| loan | Has personal loan?  'no',  'yes',  'unknown' | N/A | N/A | N/A | N/A |
| contact | Contact communication type  'cellular'  'telephone' | N/A | N/A | N/A | N/A |
| month | Last contact month of year  'Jan',  'Feb',  'Mar', ...,  'Nov',  'Dec' | N/A | N/A | N/A | N/A |
| day\_of\_week | Last contact day of the week  'Mon',  'Tue',  'Wed',  'Thu',  'Fri' | N/A | N/A | N/A | N/A |
| duration | Last contact duration, in seconds | 0.00 | 4918.0 | 180.0 | 258.3 |
| campaign | Number of contacts performed during this campaign and for this client | 1.0 | 56.0 | 2.0 | 2.568 |
| pdays | Number of days that passed by after the client was last contacted from a previous campaign | 0.0 | 999.0 | 999.0 | 962.5 |
| previous | Number of contacts performed before this campaign and for this client | 0.0 | 7.0 | 0.0 | 0.173 |
| poutcome | Outcome of the previous marketing campaign  'failure',  'nonexistent',  'success' | N/A | N/A | N/A | N/A |
| emp.var.rate | Employment variation rate - quarterly indicator | -3.4 | 1.4 | 1.1 | 0.0819 |
| cons.price.idx | Consumer price index - monthly indicator | 92.20 | 94.77 | 93.75 | 93.58 |
| cons.conf.idx | Consumer confidence index - monthly indicator | -50.8 | -26.9 | -41.8 | -40.8 |
| euribor3m | Euribor 3-month rate - daily indicator | 0.634 | 5.045 | 4.857 | 3.621 |
| nr.employed | Number of employees - quarterly indicator | 4964 | 5228 | 5191 | 5167 |
| Y | has the client subscribed a term deposit?  'yes',  'no' | N/A | N/A | N/A | N/A |

According to the statistics summary, there seems to be an issue of imbalance in our data with 36548 cases in ’No’ and 4640 cases in ’Yes’. A dataset is imbalanced if the classiﬁcation categories are not approximately equally represented. Imbalance in data, especially in class, always contributes to worse prediction. We used “upsample” technique to construct balanced dataset. The general idea of this method is to randomly sample (with replacement) the minority class to be the same size as the majority class.



# Exploratory Analysis

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

Figure 1-10 shows the spine plots 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. We ran chi-square test on all categorical variables with response variable to see whether any of the predictors has dependency relationship with response variables. Reviewing the frequency tables and chi-square p-values for the remaining categorical variables , it appears that the variables “job”, “marital”, “education”, “contact”, “month”, “day\_of\_week” and “poutcome” could contribute, the proportions vary across the factor levels within each variable.

We decided to check whether there exists multicollinearity among continuous variables. Multicollinearity refers to the correlation between variables. To check the multicollinearity, we developed a correlation matrix on the variables. Looking at correlation matrix shows strong positive correlation between euribor3m and nr.employed (0.95), emp.var.rate and euribor3m (0.97) and emp.var.rate and nr.employed (0.91). There are negative correlations between previous and pdays (-0.59), previous and nr.employed (-0.50)

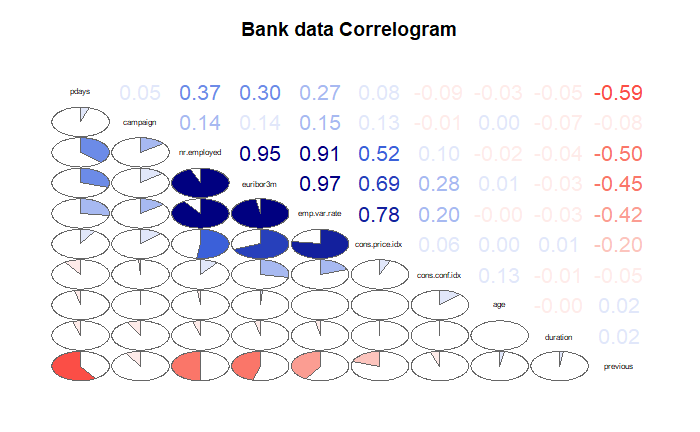


Figure: Correlation plot

## Missing Data

Data has some unknown data, which is not treated as missing, So, considering unknown as one the factor in those variables.

# Objective 1 - 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.

We ran a manual selection by starting with all the explanatory variables (excluding three variables which are independent of response variable based on Chi-square test) and having “y”(customer subscribed for term deposit or not) as the outcome in a logistic regression model, we took off the nonstatistical significant variables and then we adjusted for multicollinearity.

Key Assumptions:

Before running logistic regression we proceed making sure the following three key assumptions are met.

* Logistic regression requires the observations to be independent of each other. For this data set we do not have information if any of the observations recorded belonged to members from the same family. And hence we assume that all the observations are independent of each other.
* Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. During the EDA we identified highly correlated variables and we make sure to remove them when building the models.
* Logistic regression requires linearity of independent variables and log odds.

Lack of fit test

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, which is statistically signiﬁcant and implies that the null hypothesis should be rejected.

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

Hosmer and Lemeshow goodness of fit (GOF) test

data: model.main1$y, fitted(model.main1)

X-squared = 3501.9, df = 8, p-value < 2.2e-16

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

Parameter Interpretation

Figure 12 displays the coeﬃcient estimates for each factor level and Figure 13 displays the odd ratio estimates and the conﬁdence intervals for each level. Here is our interpretation of a subset of most interesting estimates.

# age: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 1.0024 times higher than a client 1 year younger. The 95% confidence interval is [0.9994, 1.00549].

# job (admin vs blue-collar): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for clients with admin job title relative to clients who are blue collar job tittle is 0.662. The 95% confidence interval is [0.601, 0.731].

# Marital (divorced vs single): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for divorced clients relative to single clients is 1.355. The 95% confidence interval is [1.231, 1.491].

# Education (basic 4-year vs University degree): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for clients with basic 4-year education relative to clients with university degree is 1.475. The 95% confidence interval is [1.312, 1.658].

# Contact (Cellular vs Telephone): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for clients using cellular phone relative to clients using telephone is 0.183. The 95% confidence interval is [0.169, 0.197].

# Month (April vs August): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for clients who last contacted in April relative to clients who last contacted in August is 0.130. The 95% confidence interval is [0.115, 0.148].

# Duration: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 1.0062 times higher than a client whose last contact duration is 1 second less. The 95% confidence interval is [1.0061, 1.0063].

# Campaign: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 0.9066 times higher than a client who contacted 1 time less during current campaign. The 95% confidence interval is [0.891, 0.922].

# pdays: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 0.99841 times higher than a client whose previous contact 1 day less. The 95% confidence interval is [0.9980, 0.9987].

# previous: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 1.408 times higher than a client who contacted 1 time less during previous campaign. The 95% confidence interval is [1.266, 1.567].

# poutcome (Failure vs Success): Holding all other explanatory variables fixed, the odds ratio of subscribing to a term-deposit for clients who did not subscribe in the previous campaign relative to clients who subscribed in the previous campaign is 2.053. The 95% confidence interval is [1.425, 2.957].

# Consumer Confidence index: Holding all other explanatory variables fixed, odds of a client subscribing a term-deposit is 1.0936 times higher than a client whose confidence index is 1 less. The 95% confidence interval is [1.086, 1.101].

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 below shows the ROC curve of the training dataset (orange) and the ROC curve on the test dataset (green). 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 0.913 and 0.905 for the test data, which indicates that we did not overﬁt the model and the predictability power of the model is quite high.

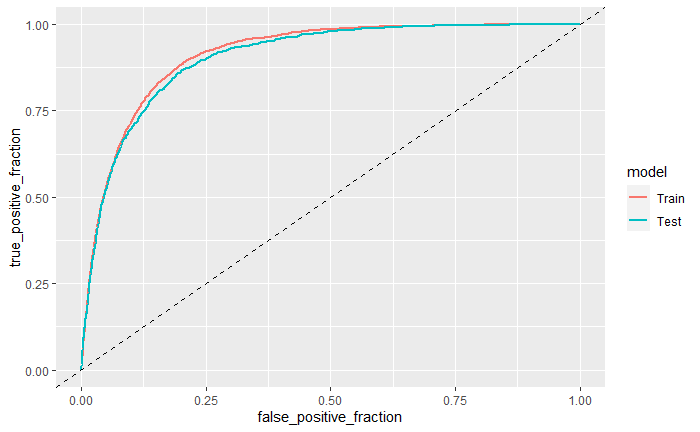


Figure: ROC curves for the balanced training dataset and the test dataset

The classiﬁcation tables (Table 1 and Table 2) 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 85.66% of the event (not subscribed for term deposit) and 81.25% of the non-event (subscribed for term deposit), with an overall rate of 83.45% on the training data. For the test data, the sensitivity is 85.293%, the speciﬁcity is 79.34% and the overall accuracy increase to 84.61.

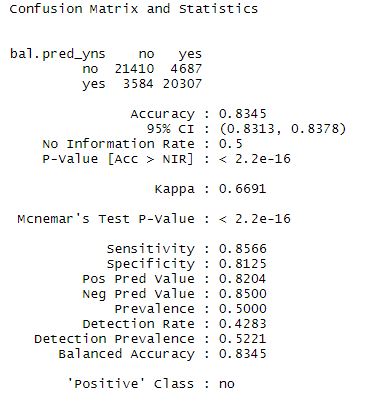
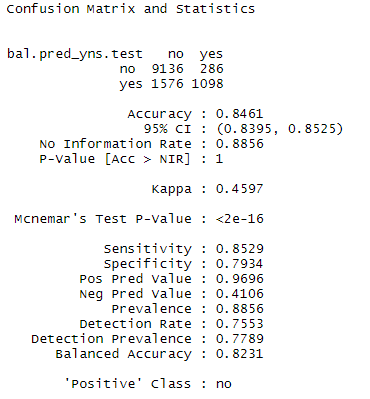
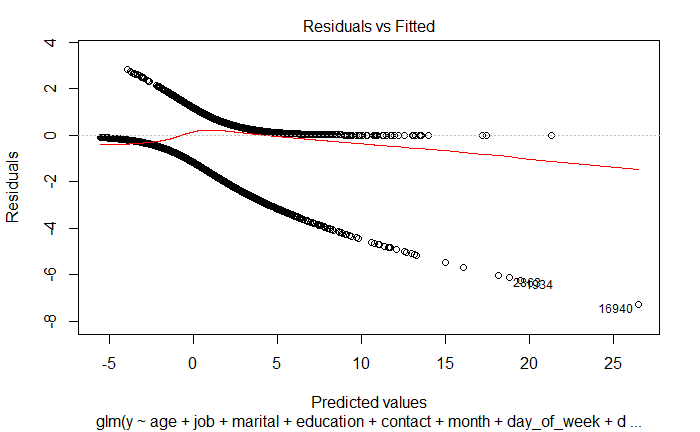
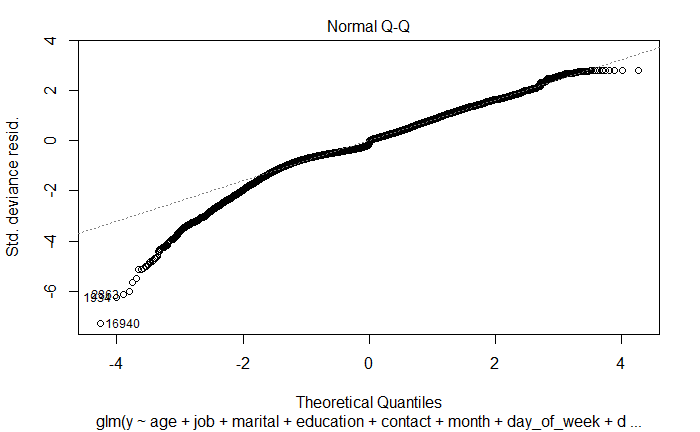


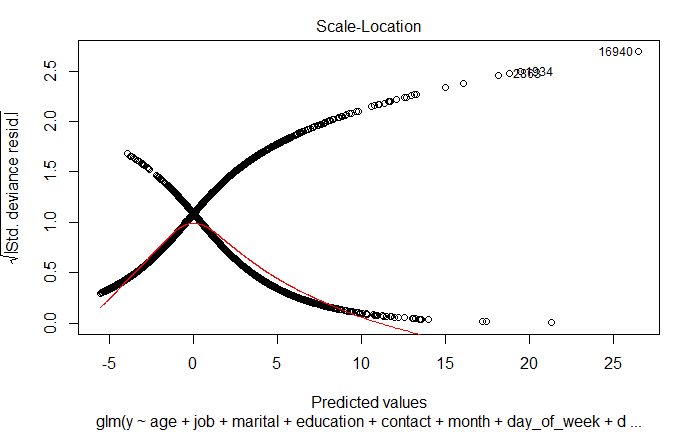
Table 1: Confusion matrix with test dataset Table 2: Confusion matrix with training dataset

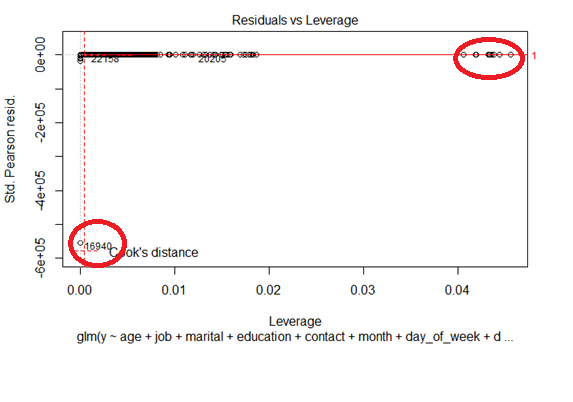
Residual diagnostics

Plots here help us examine Cook’s D graph:









When checking Cook’s D plot, if observations are outside the Cook’s distance (meaning they have a high Cook’s distance score) the observations are influential to the regression results. In this case, from the Cook’s D plot above, we see

* There are a few high leverage points, but the Cook’s distance for these points is not high. These points are likely not cause for concern.
* Observation 16940 (observation with max duration of 4918 seconds) looks like an outlier, but it is a low leverage point. This point should not cause influence on the fit. Refit the model without this observation which slightly increases the accuracy of the model.

Using Unblanced Dataset

The analyses we have done so far is 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.

At the signiﬁcant level of 0.05, age and day\_of\_week are non-signiﬁcant. Variables education and marital are statistically non-signiﬁcant here, whereas both of them were shown signiﬁcant in the prior model under the balanced dataset. We then removed the non-signiﬁcant predictors and reﬁt the model, the output.

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 below illustrates the ROC curve on the training dataset and test dataset. The AUC is 0.9175 for the model based on the training data and 0.9112 for the test data. The values are slighly higher than those that are obtained from the balanced model respectively.

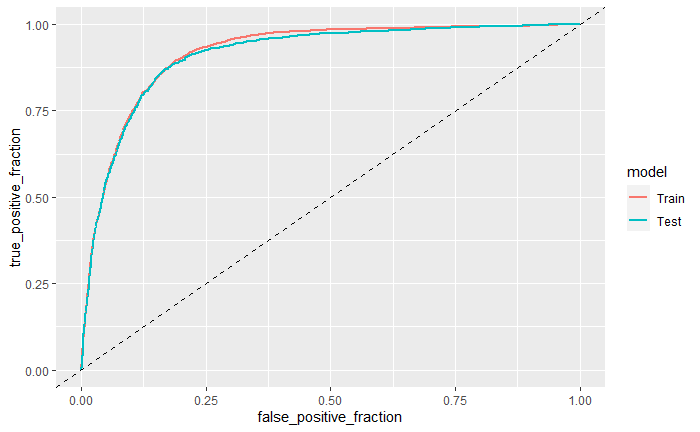


Figure: ROC curves for the unbalanced training dataset and the test dataset

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 97.56% of the event (not subscribed for term deposit) and 37.29% of the non-event (subscribed for term deposit), with an overall rate of 90.66% on the training data. For the test data, the sensitivity is 97.53%, the speciﬁcity is 37.37% and the overall accuracy increase to 90.76.

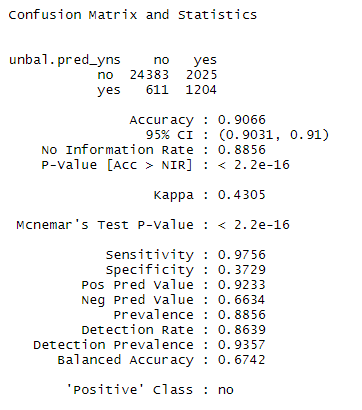
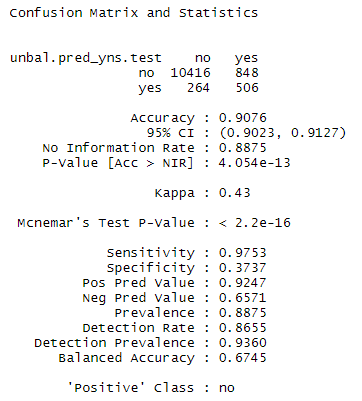


Table 1: Confusion matrix with test dataset Table 2: Confusion matrix with training dataset

Compared to the prior model with the balanced training data, the sensitivity is almost higher and the speciﬁcity is lower, 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 events resulting in higher sensitivity. On the other hand, the speciﬁcity 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.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training Dataset | | | | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC | Accuracy | Sensitivity | Specificity | AUC |
| LR (balanced dataset) | 83.45 | 85.66 | 81.25 | 0.913 | 84.61 | 85.29 | 79.34 | 0.905 |
| LR (unbalanced dataset) | 90.66 | 97.56 | 37.29 | 0.9175 | 90.76 | 97.53 | 37.37 | 0.9175 |

# Objective 2 - Additional Models

Logistic Regression model (LRM) with interaction

To add complexity to our model, we investigated what interactions may be significant. To do this, we created a list of all the possible pair combinations of the variables and used the Wald’s Z test on whether the interaction term’s coefficient was significantly different from 0.

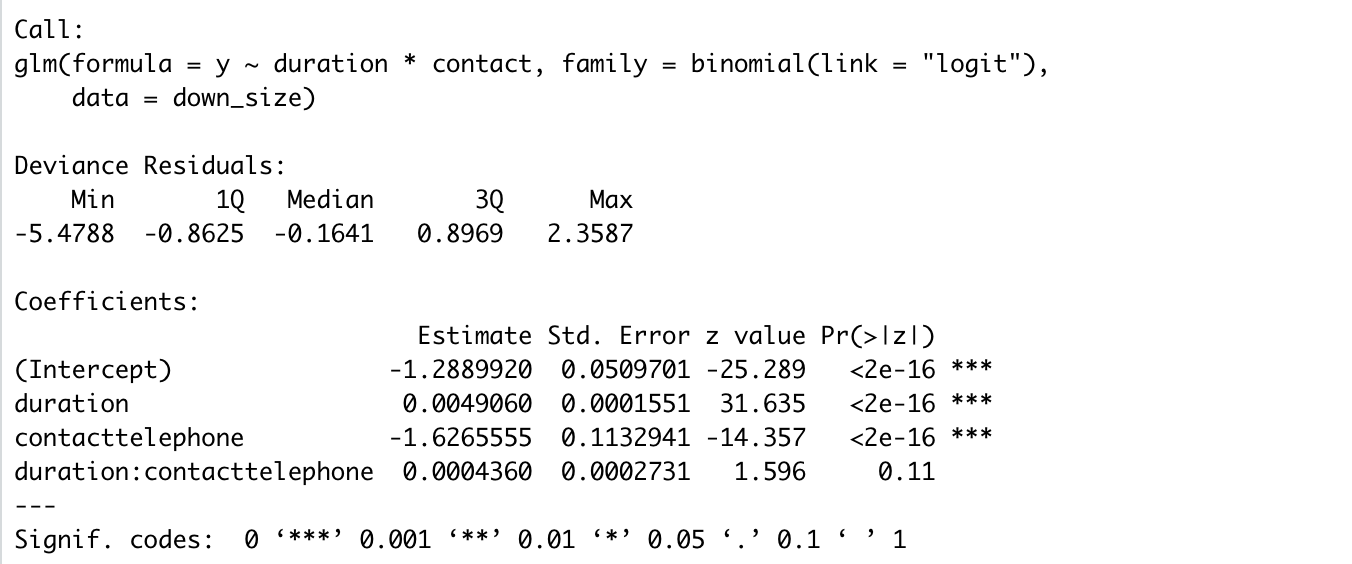
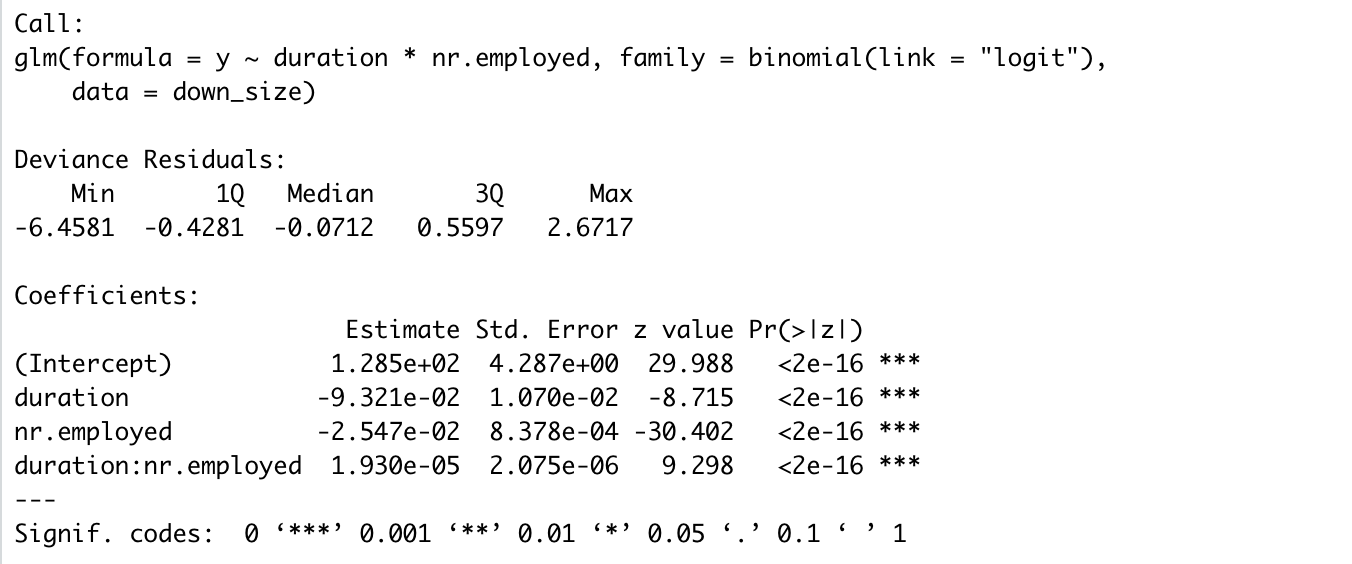


Figure : The interaction between duration and nr.employed is significant (p < 2e-16). The interaction between duration and contact is not significant (p = .11)

Taking into consideration the multiple tests, we chose .0001 as the p value threshold for significance in the interaction. We kept the interactions with p<.0001 and included them into the model.



Figure : the list of significant interaction terms

The complex model has some manually selected variables plus these selected interactions. The model is:

model.interaction1 <- glm(y ~ duration \* nr.employed + month + poutcome + emp.var.rate + cons.price.idx + job + contact + euribor3m + default + day\_of\_week + pdays + campaign + cons.conf.idx + duration\*nr.employed + duration\*poutcome + duration \* emp.var.rate + duration \* cons.price.idx + duration \* job + duration \* euribor3m + duration \* cons.conf.idx + nr.employed \* emp.var.rate + nr.employed \* euribor3m + nr.employed \* campaign + nr.employed \* cons.conf.idx + month \* cons.price.idx + month \* job + month \* contact + month \* default + month \* campaign + poutcome \* emp.var.rate + poutcome \* job + poutcome \* euribor3m + poutcome \* pdays + poutcome \* cons.conf.idx + emp.var.rate \* euribor3m + emp.var.rate \* campaign + emp.var.rate \* cons.conf.idx + cons.price.idx \* contact + cons.price.idx \* pdays + cons.price.idx \* cons.conf.idx + euribor3m \* campaign + euribor3m \* cons.conf.idx + default \* pdays + default \* campaign + default \* cons.conf.idx, data=bank, family="binomial")

The AUC of this model is .937. We picked the threshold of .5 to get a sensitivity of 85.7%, specificity of 89.7% and overall accuracy of 86.2%.

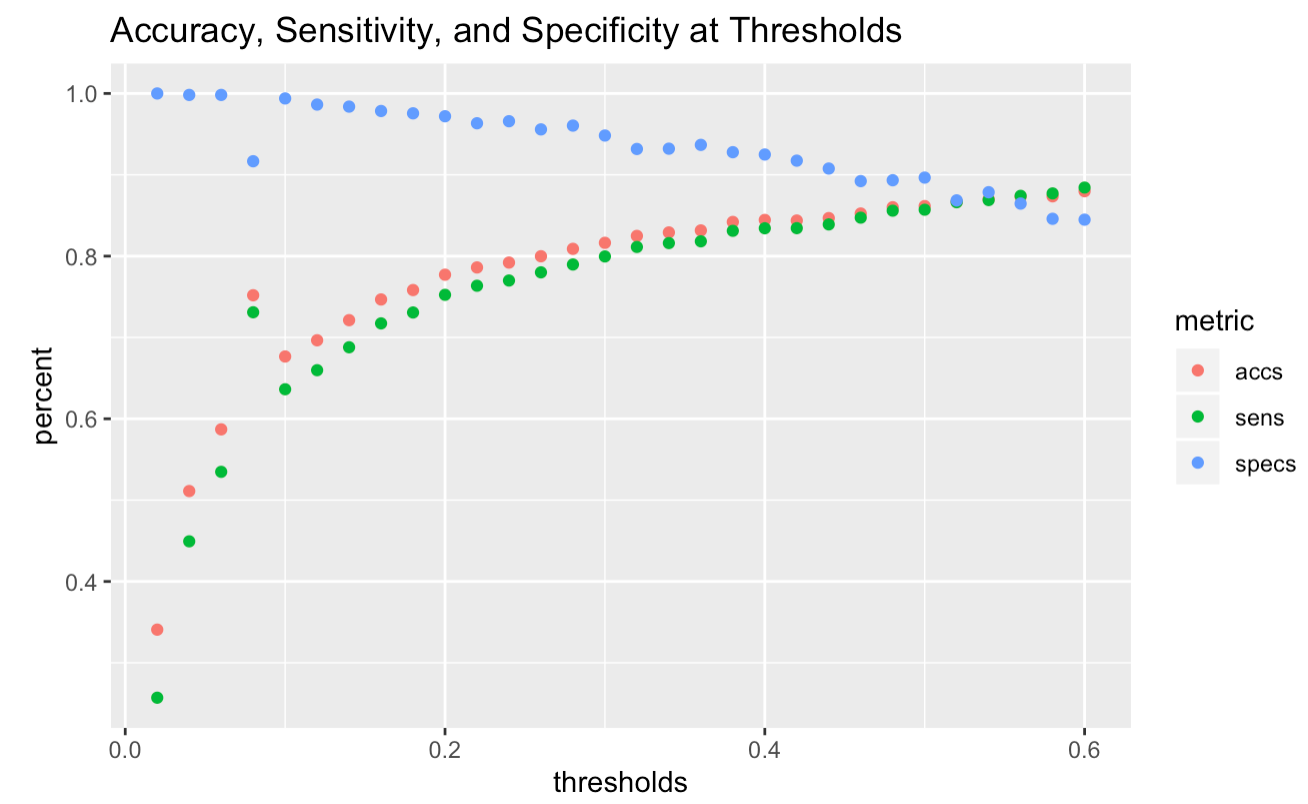


Figure : accuracy, sensitivity, and specificity across various thresholds

Linear Discriminant Analysis model (LDA)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| LDA | 85.6% | 86.9% | 84.1% | 0.912 |
| LDA log transformed | 81% | 80% | 89% | 0.914 |
| QDA | 82% | 88.3% | 75.6% | 0.899 |

|  |  |
| --- | --- |
|  |  |

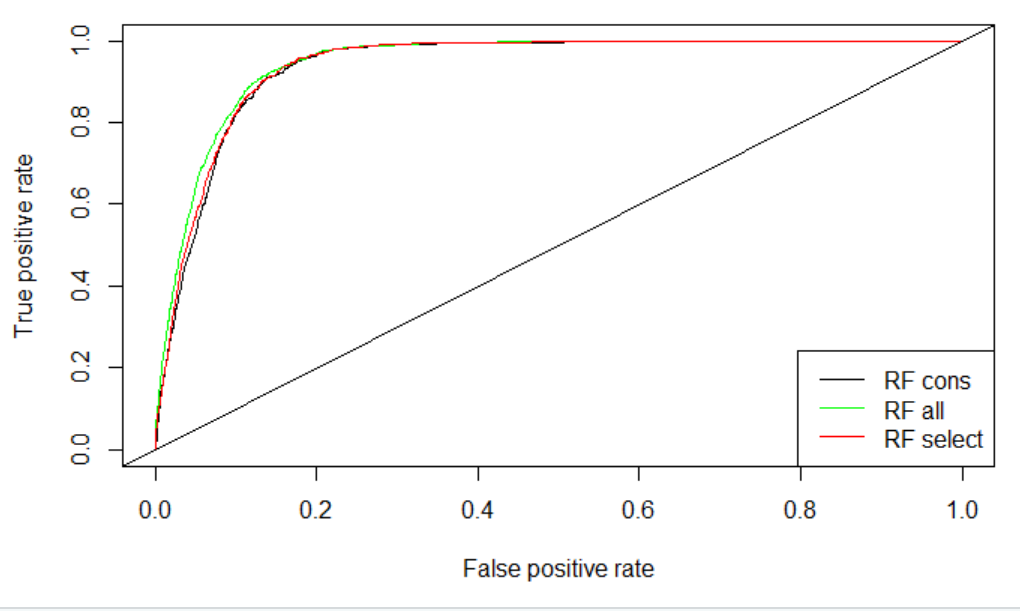
Non-parametric model. Knn (K nearest Neighbor)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| KNN (k=3) | 81.5% | 86.7% | 76.3% | 0.616 |
| KNN (k=5) | 83.6% | 83.6% | 82.9% | 0.647 |

Non-parametric model. Random Forest (RF)

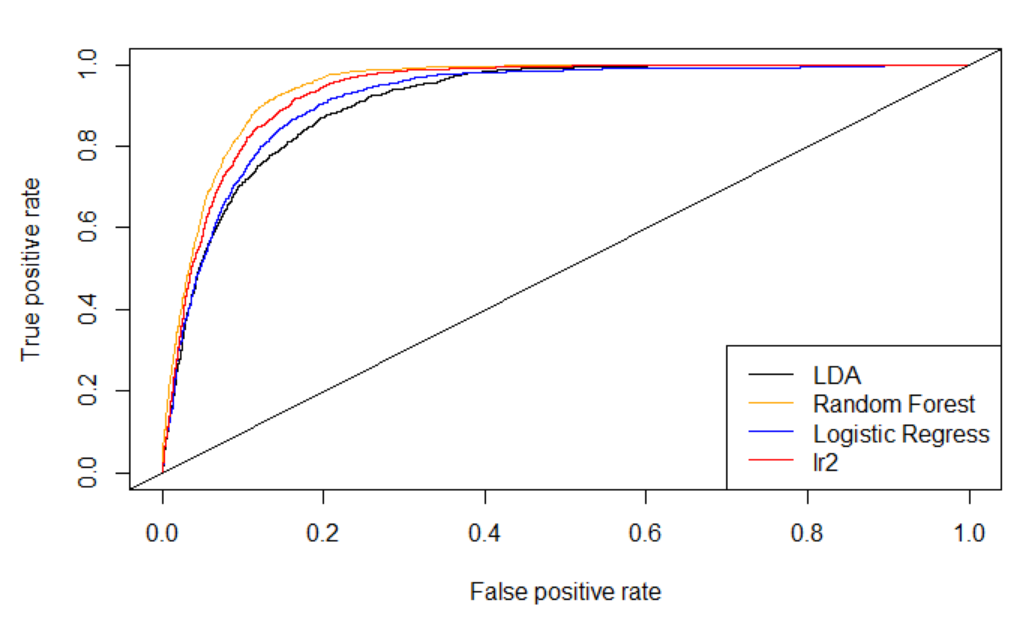
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| RF (balanced dataset)  All cons | 89.2% | 90.2% | 81.5% | 0.939 |
| RF (unbalanced dataset) | 91.2% | 96% | 52% | 0.94 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| RF  All cons | 89.2% | 90.2% | 81.5% | 0.939 |
| RF  All | 91% | 93% | 72% | 0.947 |
| RF  Select | 89.1% | 88.1% | 86% | 0.941 |



Comparison of all models and Conclusion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test Dataset | | | |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| LDA | 85.6% | 86.9% | 84.1% | 0.912 |
| KNN | 83.6% | 83.6% | 82.9% | 0.647 |
| RF | 87.8% | 88.1% | 86% | 0.941 |
| Logistic Regression  (Final) | 86.3% | 86.2% | 87.1% | 0.937 |



**APPENDIX**

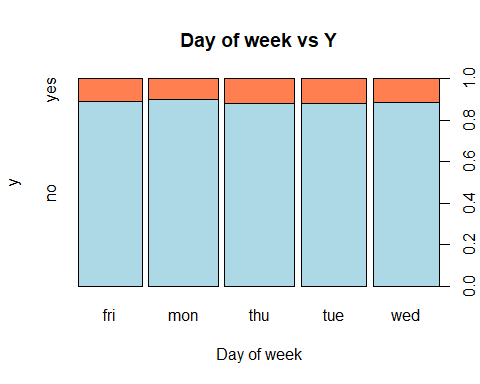


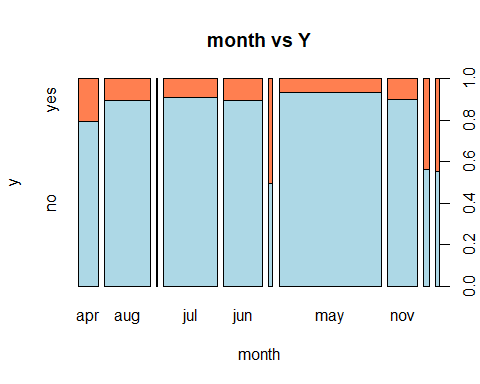
Figure 1: Spine plot of day of Week with response variable

Figure 2: Spine plot of month with response variable

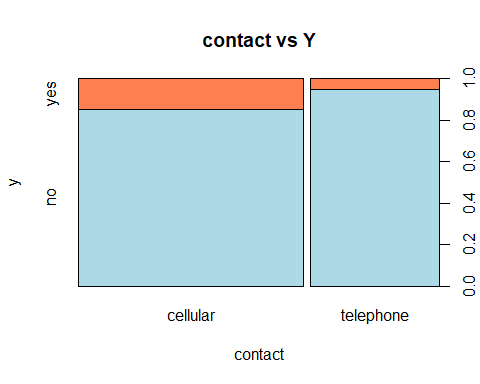


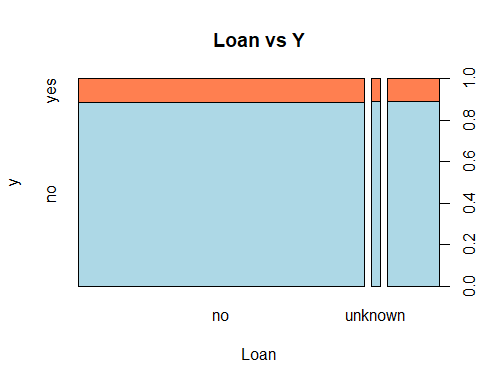
Figure 3: Spine plot of contact with response variable

Figure 4: Spine plot of Loan with response variable

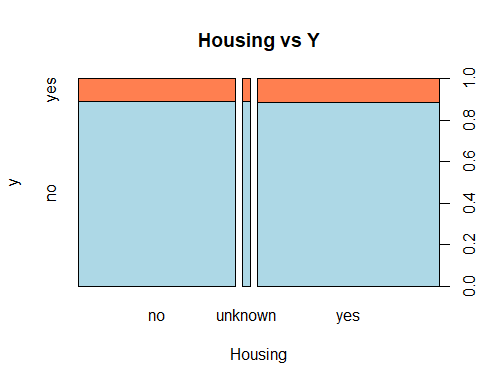


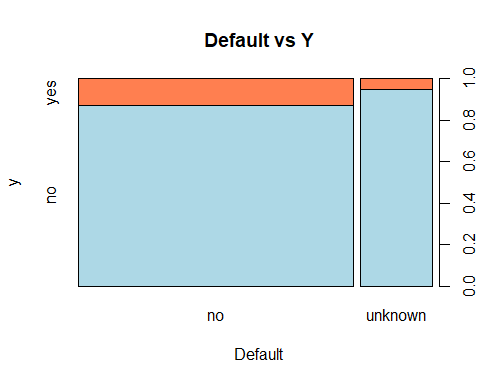
Figure 5: Spine plot of Housing with response variable

Figure 6: Spine plot of Default with response variable

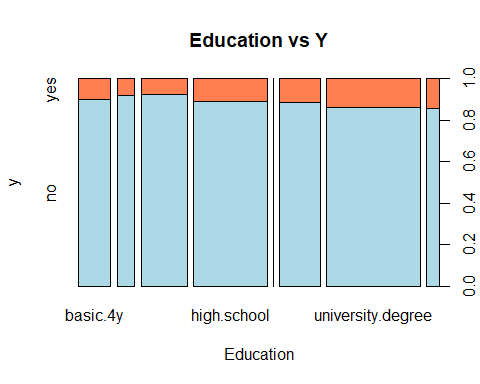


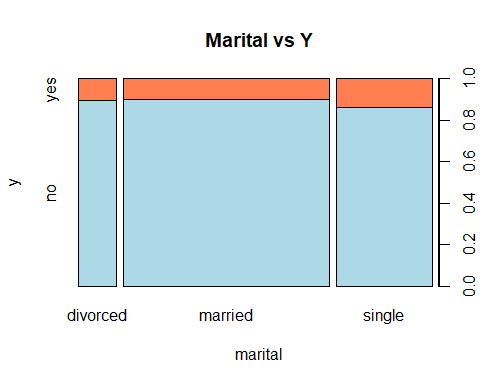
Figure 7: Spine plot of Education with response variable

Figure 8: Spine plot of Marital with response variable

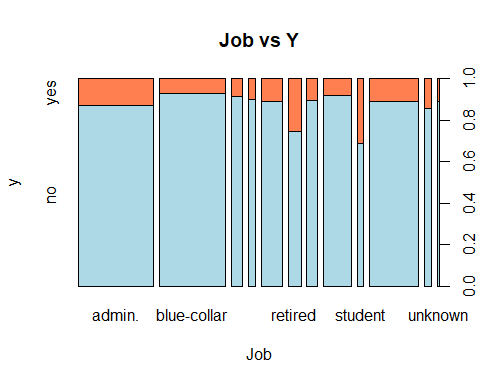


Figure 9: Spine plot of job with response variable

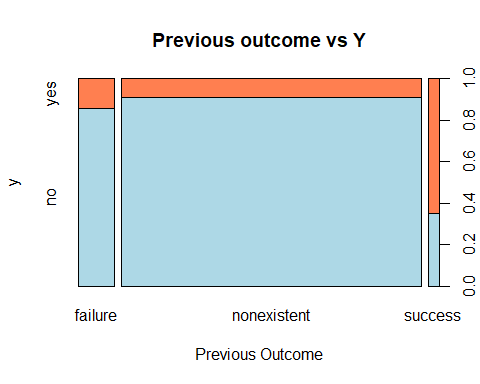


Figure 10: Spine plot of previous with response variable

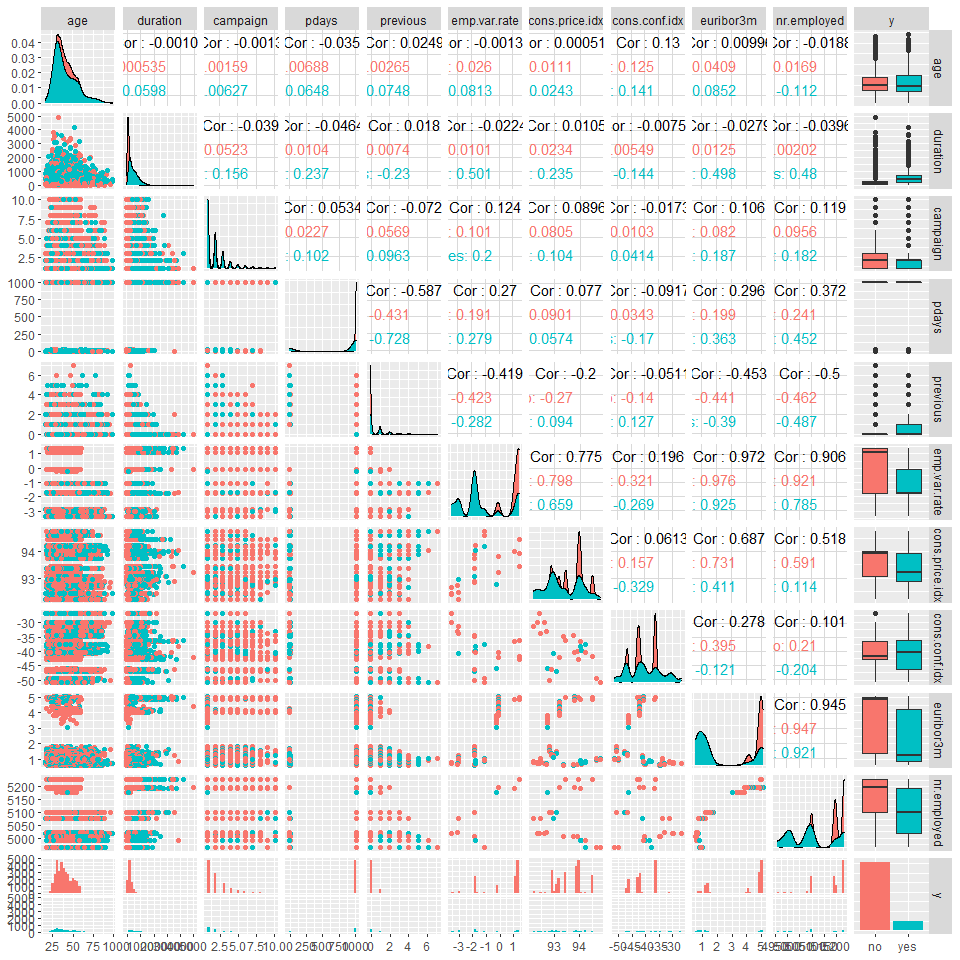


Figure 11: Pair wise plots for continuous variables with response variable

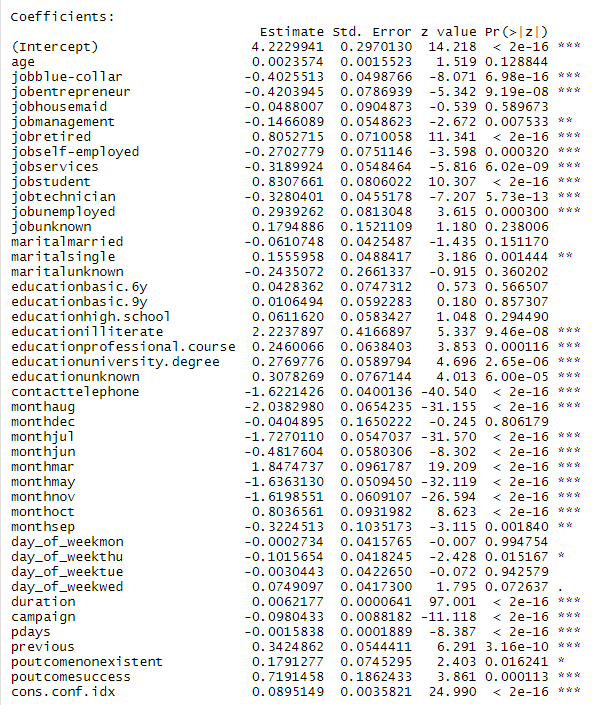


Figure 12: Coeﬃcient estimates

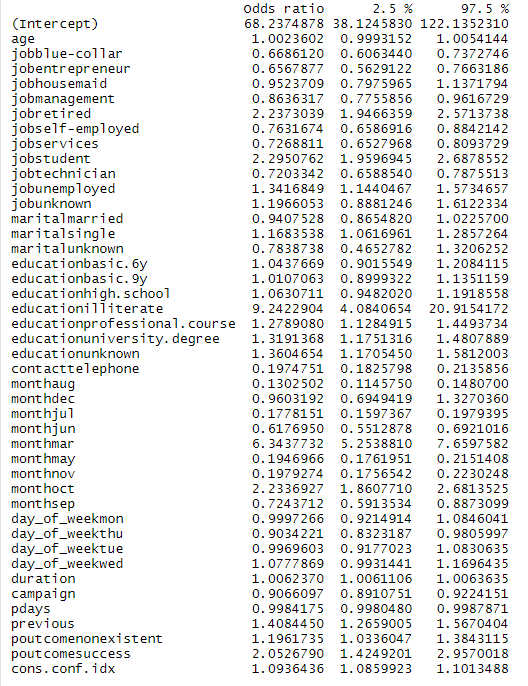


Figure 13: Odds Ratio estimates and conﬁdence intervals