DS 6372 - Applied Statistics : Inference & Modeling

Dr. Turner - Final Project

Antonio Debouse

Paul Huggins

Lijju Mathew

Introduction

Predicting the success rate of telemarketing calls for a bank selling long-term deposits can be a useful metric when determining the overall health of a bank. Marketing efforts are a commonplace in today's business models to enhance profitability and customer base. This paper will focus on data from 2008 to 2013 supplied by a Portuguese retail bank in an attempt to predict the success rates of the telemarketing efforts. Noticeably, this data was gathered during the global financial downturn of 2008 and subsequent years so any application to the current market needs to be used carefully. Multiple prediction techniques such as logistic regression, random forests, linear discriminant analysis and others are utilized and then compared to evaluate which model(s) provide the highest degree of prediction accuracy.

Data Description

The data consists of 41,188 different subjects contacted and the associated attributes. There are 20 unique variables and 1 target prediction variable. There are a total of 12,718 missing values out of the 864,948 total values (1.4%) and they are labeled as 'unknown'. The predicted variable is labeled 'y' and represents whether or not the client purchased a long-term deposit (yes/no). Overall, there were 36,548 no's and 4,640 yes's, resulting in an 11.2% success rate (88.8% failure rate). Below is a data table with descriptions:

Variable	Type	Description	Missing Values
Age	Numeric	Age of client	0
Job	Categorical	Type of Job	330
Marital	Categorical	Marital Status	80
Education	Categorical	Education	1,731
Default	Categorical	Credit Default?	8,597
Housing	Categorical	Housing Loan?	990
Loan	Categorical	Personal Loan?	990
Contact	Categorical	Communication Type	0
Month	Categorical	Last Contact Month	0
Day_of_week	Categorical	Last Contact Day of Week	0
Duration	Numeric	Last Contact Duration	0
Campaign	Numeric	# of Contacts Performed	0
Pdays	Numeric	# of Days after Last Contact	0

Previous	Numeric	# of Contacts before Campaign	0
Poutcome	Categorical	Outcome of Prior Marketing Campaign	0
Emp.var.rate	Numeric	Employment Variation Rate	0
Cons.prics.idx	Numeric	Consumer Price Index	0
Cons.conf.idx	Numeric	Consumer Confidence Index	0
Euribor3m	Numeric	Euribor 3 Month Rate	0
Nr.Employed	Numeric	Number of Employees	0
y	**Categorical**	**Subscribed to Term Deposit?**	**0**

Exploratory Data Analysis

While exploring data we identified multiple trends and used this to shape our approach in modelling. A summary table focusing on the continuous variables is shown in Figure 1 along with histograms and a box plot plot of the variables is in Figure 2. Below are the observations that we identified from our exploratory data phase.

- Missing/Unknown Values Treatment
    - We Identified that there was a significant amount of null values in our data set. (Figure 3) After identifying the null values in the dataset we decided to use the Mice package to fill in the missing values with predicted values. This method runs a built-in function to predict the missing variables using the complete rows in the dataset as the training model. We took this approach so that there is consistency when addressing the correlations and deciding on what factors we could use in our models.

- Correlations between variables
    - From those factors we have found that pdays, previous, nr.employed, emp.var.rate, euribor3m, cons.price,id with a high amount of significance. That is shown in our correlation matrix and heat map below. See Appendix (Figure 4)
- Outliers
    - We ran the box plots for the variables and found that the variable pdays has a good number of outliers with value 999. On further investigation, the values of 999 means the client was not contacted. This value was sticking out as an outlier and hence we recoded the value to -1. See Appendix (Figure 5)
    - The accuracy metrics decreased when removing continuous variable observations that surpassed the 99th and 95th percentile; the dataset decreased dramatically along with important information the models needed to improve their predictions.

- PCA
    - Principal Component Analysis was done, and PCs were plotted (Figure 6) to get a sense of how good the prediction would be. This gave an indication that we would get marginally good predictions.
- Duration has a direct correlation with the predicted value (y) and is determined at the conclusion of the call. A value of 0 is automatically a 'no' for the y value and the duration value is not known before the call begins unlike the other 19 predictors to be used. This variable was dropped when the modelling procedures begin due to the direct correlation with the outcome

Objective One

The first objective was to perform a logistic regression technique along with similar regression models on the full dataset and then interpret the coefficients, confidence intervals, and perform hypothesis tests. Due to the simplistic nature of the models, no interaction terms were utilized. The goal of this objective was to let the data guide us into an easy to interpret model to explore the dataset and relationships.

Model Selection:

Below is the approach that we took for model selection.

1) Run a full model with all the predictors and do a forward, backward, stepwise feature selection.
2) Taking correlation into account by removing redundant variables and using significant predictors from above feature selection we built a very lean model.
3) Then we added significant predictors from feature selection iteratively and built 7 models with manual intuition.
4) We choose the top 3 models based on their ability to accurately predict yes (specificity metric) and then their overall accuracy. We placed the most significance on a model's "true positive" metric (specificity) because it would be valuable for a bank to be able to know which customer is more likely to buy a term deposit. An employee will have more success pitching a term deposit to a likely buyer, or spend that opportunity selling a different product to a customer not likely to buy a term deposit.

Compare Competing Models:

Model	Accuracy	Sensitivity	Specificity	VIF	MisCalc Rate	ROC/AUC
Model 1	81.6%	85.8	48.9	< 5	0.183	0.738
Model 2	81.1%	84.9	50.9	< 5	0.189	0.721
Model 3	80.9%	84.4	53.2	< 5	0.193	0.738

From the above three competing models we chose Model 3 which has high sensitivity and highest specificity sacrificing little accuracy. Model 3 is a simple model that contains the least amount of variables as well. The variables used in each model were:

- Model 3 : job, education, campaign, poutcome, nr.employed
- Model 2 : job, month, campaign, pdays, poutcome, emp.var.rate, cons.price.idx
- Model 1 : age, job, contact, day_of_week, campaign, pdays, poutcome, emp.var.rate, nr.employed

Assumption:

- After assessing the information and going through the data, the distribution of the residuals is fairly normal as shown on the residual plots. Additionally, the data is also normally distributed as per the QQ plot. The assumptions have been met for us to proceed. Independence is also assumed. As far as the outliers and influential point analysis goes, the Cook's D plot has few outliers, but they seem that there are not significant outliers. Additionally, the leverage plots seem to be fairly uniform. See Appendix (Figure 7) for the plots.
- We ran a goodness of fit test with the Hosmer-Lemeshow test and Chi Square test. Both gave a p value closer to 1 indicating no evidence of poor fit. See Appendix (Figure 8)

Metrics:

Below are the metrics collected for the final model (Model 3 Above) for the objective. See Appendix (Figure 9).

- Model Summary
- Confidence Intervals
- VIF
- ROC Curve
- Confusion Matrix (Accuracy, Sensitivity, Specificity)
- Miscalculation Rate

Regression Equation:

log(y) = 45.9 + (-0.028) job blue-collar + (-0.298) job entrepreneur + (1.779) job housemaid +
(0.732) job retired + (-0.279) job self-employed + (-0.255) job services +
(0.590) job student + (0.829) job technician + (-0.270) job unemployed +
(0.168) education basic.6y + (1.460) education basic.9y +
(0.349) educationhigh.school + (0.377) education illiterate +
(0.711) education professional.course + (0.690) education university.degree +
(-0.220) campaign +
(0.805) poutcome nonexistent + (2.177) poutcome success +
(-0.009) nr.employed

Parameter Interpretation:

- The log odds of clients subscribing to term deposit among clients with blue collared jobs is 0.97 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with an entrepreneur job is 0.74 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with a housemaid job is 5.9 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with a management job is 2.67 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients who are retired is 2.07 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with a self-employed job is 0.75 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with a services job is 0.77 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients being a student is 1.80 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with a technician job is 2.2 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients unemployed is 0.76 times more compared to clients with admin jobs, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with basic 6y education is 0.76 times more compared to clients with basic 4y education, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with basic 9y education is 4.3 times more compared to clients with basic 4y education, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with high school education is 1.41 times more compared to clients with basic 4y education, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with no education is 1.45 times more compared to clients with basic 4y education, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with professional education is 2.03 times more compared to clients with basic 4y education, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with university education is 1.99 times more compared to clients with basic 4y education, having other predictors constant.

- The log odds of clients subscribing to term deposit with every increase in contact in the current campaign is 0.82 times the odds of subscribing, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with the previous outcome of the campaign being successful is 8.82 times more compared to clients with previous outcome as failure, having other predictors constant.
- The log odds of clients subscribing to term deposit among clients with the previous outcome of the campaign nonexistent is 2.23 times more compared to clients with previous outcome as failure, having other predictors constant.
- The log odds of clients subscribing to term deposit with every increase in number of employees is 0.99 times the odds of subscribing, having other predictors constant.

Conclusion:

The best model for objective 1 is a simple model with higher specificity that sacrifices a little sensitivity and accuracy. A few predictors (job, education, campaign, poutcome, nr.employed) were used in the model to avoid overfitting the model and still have a high level of accuracy. Adding predictors resulted in a decrease in specificity but an increase in accuracy and sensitivity. The low miscalculation rate and high ROC/AUC value also indicate that this is a better model.

Objective Two

The next step was to build a more complex model from objective 1 using the same set of predictors. First, an ANOVA table was built to determine which variables were significant with each other. The ANOVA table can be seen in Figure 10. Secondly, the model was built using the following predictors: job, education, campaign, outcome, job*education, job*campaign, education*campaign, campaign*nr.employed, education*nr.employed, education*poutcome and outcome*nr.employed. The variables were converted to numeric before the interaction terms were applied. The same stepwise procedure was run on the data. Below is a table comparing the simple and complex models. The complex model had a higher specificity and AUC value compared to the simple model.

Model	Accuracy	Sensitivity	Specificity	MisCalc Rate	AUC/ROC
Simple	80.9%	84.4	53.2	0.193	0.739
Complex	80.0%	82.9	57.3	0.202	0.753

In addition to the simple and complex logistic regression models, additional models were run using a plethora of methods in an effort to generate an overall better result. Duration was dropped as a predictor just like in the first objective.

The first set of models utilized Linear and Quadratic Discriminant analysis using only the continuous variables. The variables included in the model were: age, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed. Each of these methods were run 3 times: once on the unbalanced original data, secondly on a balanced

training dataset with an 50/50 ratio between yes and no predictors, and thirdly on a Rose balanced training set. The test data consists of a random 30% split from the original data. The results of the models are below.

Method	Accuracy	Sensitivity	Specificity	Error Rate
LDA Rose	0.711	0.711	0.714	0.289
LDA Balanced	0.727	0.729	0.707	0.273
QDA Balanced	0.857	0.900	0.524	0.143
QDA Original	0.875	0.922	0.491	0.125
QDA Rose	0.873	0.927	0.447	0.127
LDA Original	0.891	0.960	0.328	0.109

The LDA Balanced and LDA Rose resulted in the best specificity values among all the LDA and QDA models. They did suffer from lower overall accuracies but still are in an acceptable range. Looking at the PCA plot in the appendix there does appear to be a more linear differentiator between yes and no leading our team to believe that is the reason LDA outperformed QDA.

The second set of models consisted of MDA, FDA, Ridge, Lasso, Elastic Net and Random Forest techniques on the original data, balanced data and Rose data sets. Under the MDA (Mixture Discriminant Analysis), each class is assumed to be a Gaussian mixture of subclasses. The FDA (Fisher's Discriminant Analysis) a non-linear combination of predictors is used such as splines. The models used variables deemed significant by the original stepwise procedures. The variables used were: age, job, marital, education, contact, month, day_of_week, campaign, outcome, cons.price.idx, cons.conf.idx and y. Below are the top 5 by each metric. The results for all of the models can be found in Figure 11 in the appendix.

Focusing on specificity, the best specificity came from the LDA Balanced and LDA Rose, mirroring the results above. Noticeably, the balanced datasets held four out of the top five spots coming from four different regression techniques. The balanced data consisted of a balanced training set (50/50 yes/no) and a random 30% sample of the original data acted as the test set. The results show that using this method resulted in the best models for determining positive 'yes' classes which we believe is a more important metric for this dataset.

Method	Accuracy	Sensitivity	Specificity	ErrorRate
LDA Rose	0.7114996	0.7111900	0.7139738	0.28850045
LDA Balanced	0.7266106	0.7291135	0.7068966	0.27338945
Lasso Rose	0.6511289	0.6503630	0.6574420	0.34887109
Net Rose	0.6609209	0.6617060	0.6544503	0.33907906
FDA Rose	0.6627013	0.6639746	0.6522064	0.33729870

Focusing on sensitivity, the best sensitivity came from the random forest on the original data. The top 5 models all were from the original data set. This is not surprising given that sensitivity is considering the correct guesses of 'no's in the dataset and the data consisted of over 84% of them. These results were expected and further enforced our belief that specificity was a better metric to use for this scenario. If the goal was to correctly predict the no's, then a random forest model on the original dataset would be the best model. The random forest models were built on a 10-fold cv grid search using both the categorical and numerical variables. Tests were done using different numbers of trees. The optimal number of trees for using the original dataset was 10. The balanced data was optimized using 24 trees and the rose data used 20 trees.

Method	Accuracy	Sensitivity	Specificity	ErrorRate
RF Original	0.9003804	0.9930940	0.1457101	0.09961965
Lasso Original	0.9002185	0.9907315	0.1634615	0.09978150
Net Original	0.8999757	0.9904589	0.1634615	0.10002428
Ridge Original	0.9008659	0.9898228	0.1767751	0.09913409
FDA Original	0.9019179	0.9870968	0.2085799	0.09808206

Focusing on accuracy and error rates, the random forest method on the original data set achieved the highest accuracies among the group at 90.2%. As seen below, the specificities for these models were greatly diminished compared to the other models. Similar to the previous table looking at sensitivity, accuracy is more heavily weighted by sensitivity than specificity given the unbalanced predictors in the original data set.

Method	Accuracy	Sensitivity	Specificity	ErrorRate
FDA Original	0.9019179	0.9870968	0.2085799	0.09808206
Ridge Original	0.9008659	0.9898228	0.1767751	0.09913409
RF Original	0.9003804	0.9930940	0.1457101	0.09961965
Lasso Original	0.9002185	0.9907315	0.1634615	0.09978150
Net Original	0.8999757	0.9904589	0.1634615	0.10002428

Conclusion

The best models for objective 2 utilized the balanced data set and achieved a specificity of 71.4% from the LDA technique. As mentioned previously, accuracy and sensitivity were not the primary metrics used in analyzing the results. Specificity is the accuracy of predicting when a customer does sign up for a long term deposit.

Our team felt that specificity is an important metric for measuring the success of the models. Key information that could be beneficial in improving the models would be demographic information and a yes/no data set that is more equally balanced. A more balanced data set from the beginning would negate the need to generate a balanced set during the analysis that

drastically reduced the amount of training data used for the models. The dimensionality reduction provided by the LDA method proved to provide the best model. As seen in the PCA plot in the appendix, the split between yes and no appears to be more linear compared to quadratic.

Appendix

Figure 1 - Summary Statistics on Continuous Variables

Summary Statistics Table for Continuious Variables Grouped by Yes/No Responses	y: no (N = 36,548)	y: yes (N = 4,640)
Age		
mean (sd)	39.91 ± 9.90	40.91 ± 13.84
median (Q1, Q3)	38.00 (32.00, 47.00)	37.00 (31.00, 50.00)
min	17	17
max	95	98
Duration		
mean (sd)	220.84 ± 207.10	553.19 ± 401.17
median (Q1, Q3)	163.50 (95.00, 279.00)	449.00 (253.00, 741.25)
min	0	37
max	4918	4199
Campaign		
mean (sd)	2.63 ± 2.87	2.05 ± 1.67
median (Q1, Q3)	2.00 (1.00, 3.00)	2.00 (1.00, 2.00)
min	1	1
max	56	23
Previous Days Since Contacted		
mean (sd)	984.11 ± 120.66	792.04 ± 403.41
median (Q1, Q3)	999.00 (999.00, 999.00)	999.00 (999.00, 999.00)
min	0	0
max	999	999
Previous # of contacts performed campaign		
mean (sd)	0.13 ± 0.41	0.49 ± 0.86
median (Q1, Q3)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)
min	0	0
max	7	6
Employment Variation Rate		
mean (sd)	0.25 ± 1.48	-1.23 ± 1.62
median (Q1, Q3)	1.10 (-1.80, 1.40)	-1.80 (-1.80, -0.10)
min	-3.4	-3.4
max	1.4	1.4
Consumer Price Index		
mean (sd)	93.60 ± 0.56	93.35 ± 0.68
median (Q1, Q3)	93.92 (93.08, 93.99)	93.20 (92.89, 93.92)
min	92.201	92.201
max	94.767	94.767
Consumer Confidence Index		
mean (sd)	-40.59 ± 4.39	-39.79 ± 6.14
median (Q1, Q3)	-41.80 (-42.70, -36.40)	-40.40 (-46.20, -36.10)
min	-50.8	-50.8
max	-26.9	-26.9
Euribor 3-Month Rate		
mean (sd)	3.81 ± 1.64	2.12 ± 1.74
median (Q1, Q3)	4.86 (1.40, 4.96)	1.27 (0.85, 4.41)
min	0.634	0.634
max	5.045	5.045
Number of Employees		
mean (sd)	5,176.17 ± 64.57	5,095.12 ± 87.57
median (Q1, Q3)	5,195.80 (5,099.10, 5,228.10)	5,099.10 (5,017.50, 5,191.00)
min	4963.6	4963.6
max	5228.1	5228.1

Figure 2 - Histogram of all variables

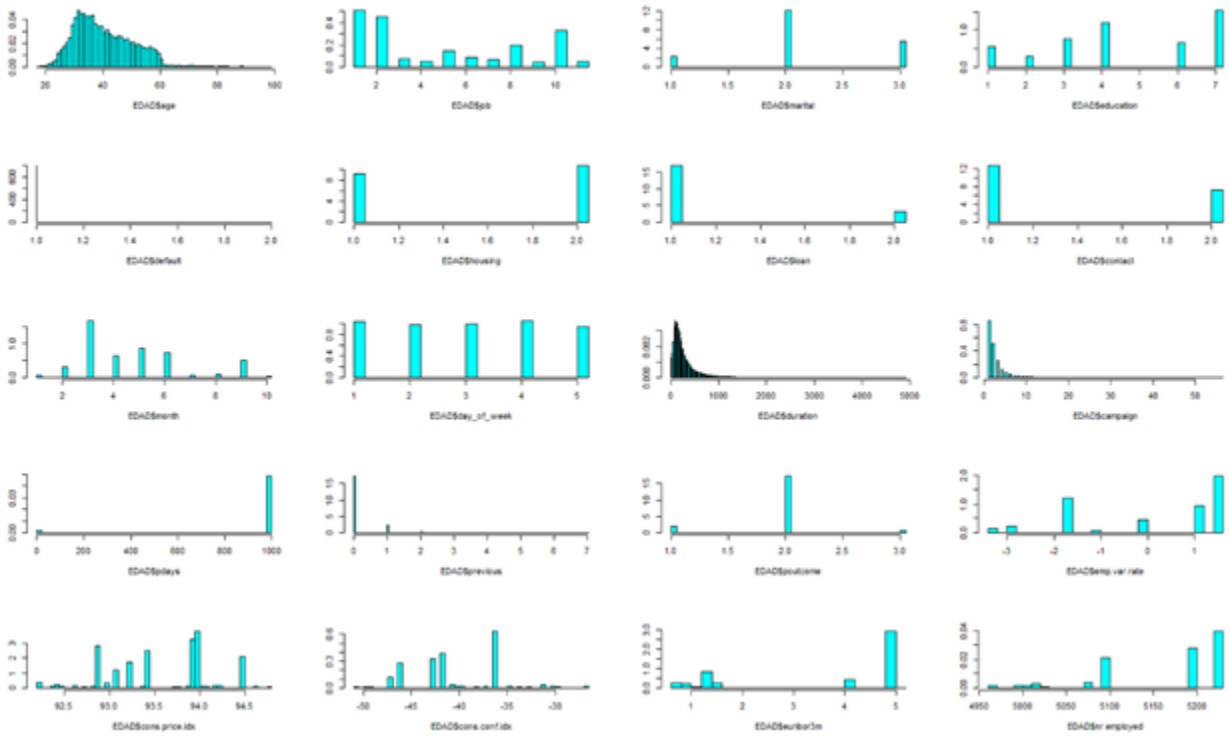

Figure 3 - Missing Values

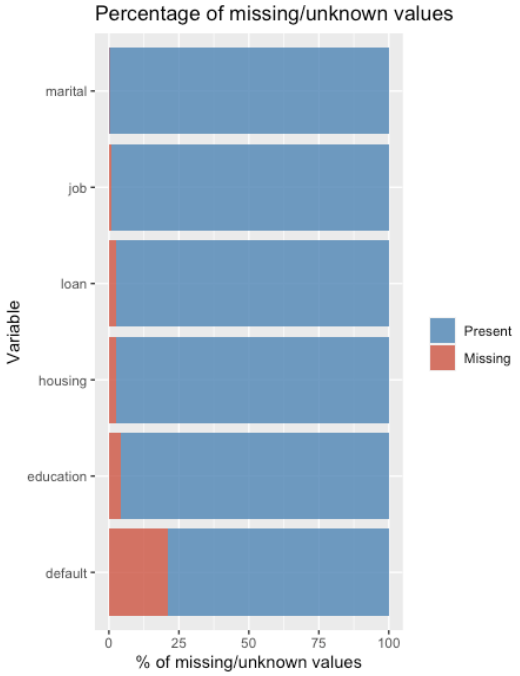

Figure 4 - Correlation Matrix

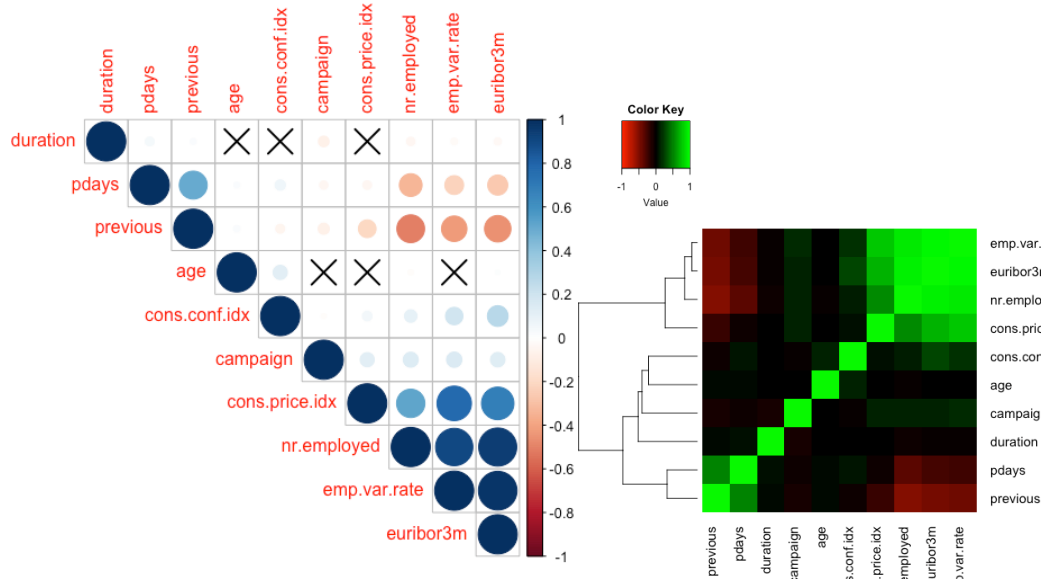

Figure 5 - Box plots

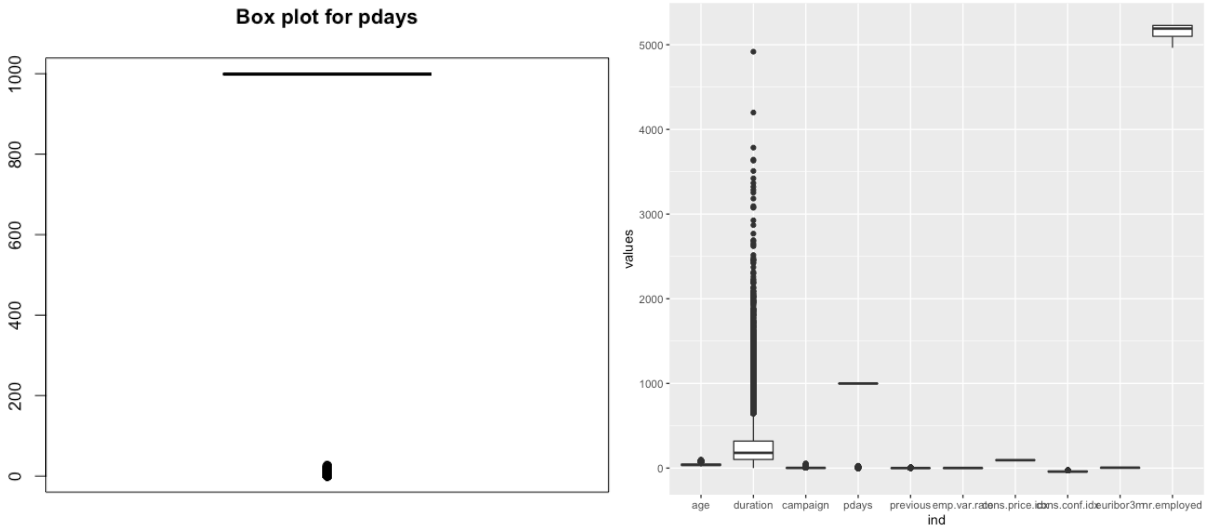

Figure 6 - PCA

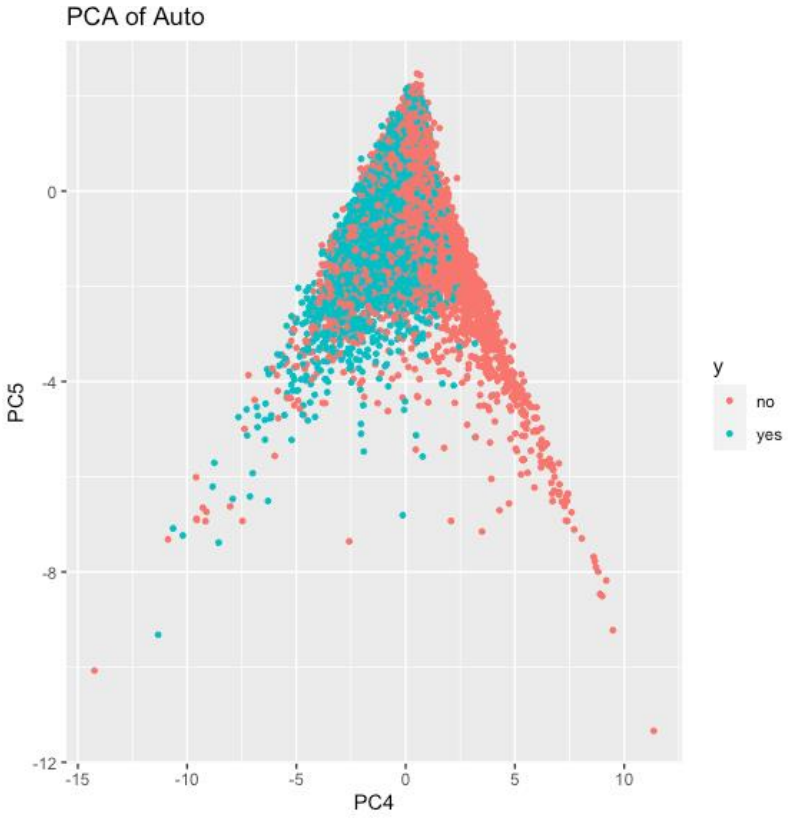

Figure 7 - QQ Plots, Cooks D, Leverage

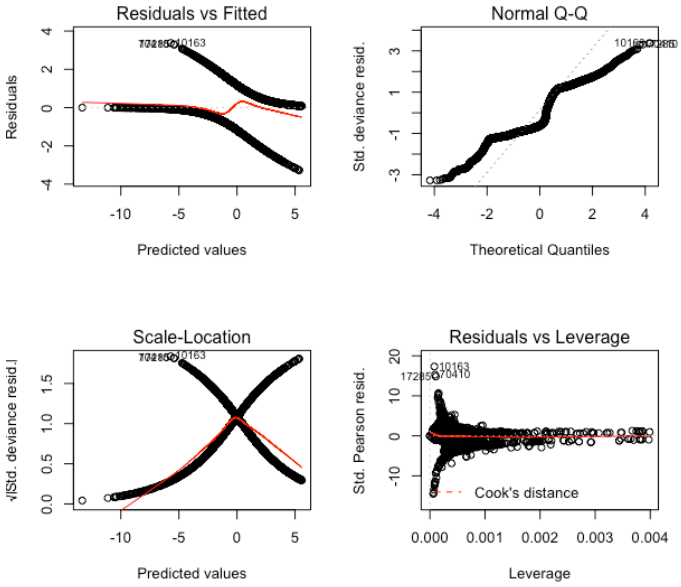

Figure 8 - Goodness of Fit Test

```
hl <- hoslem.test(mod$y, fitted(mod), g=10)
hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  mod$y, fitted(mod)
X-squared = 7.4866, df = 8, p-value = 0.4851
```

Figure 9 - Metrics for Objective 1 Model

```
> summary(step.os6.log)

Call:
glm(formula = y ~ job + education + campaign + poutcome + nr.employed,
    family = "binomial", data = logit.os.Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7474  -0.8773  -0.5221   0.9610   3.1609

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 45.9220468  1.0743593  42.744  < 2e-16 ***
jobblue-collar              -0.0280559  0.0458450  -0.612 0.540556
jobentrepreneur             -0.2987488  0.0896210  -3.333 0.000858 ***
jobhousemaid                 1.7790580  0.0666596  26.689  < 2e-16 ***
jobmanagement                0.9822990  0.0496651  19.778  < 2e-16 ***
jobretired                   0.7321423  0.0680274  10.762  < 2e-16 ***
jobself-employed            -0.2796216  0.0854764  -3.271 0.001070 **
jobservices                 -0.2556544  0.0608761  -4.200 2.67e-05 ***
jobstudent                   0.5901579  0.0860455   6.859 6.95e-12 ***
jobtechnician                0.8291080  0.0428630  19.343  < 2e-16 ***
jobunemployed               -0.2706723  0.0962600  -2.812 0.004925 **
educationbasic.6y            0.1680397  0.0808710   2.078 0.037721 *
educationbasic.9y            1.4606270  0.0556986  26.224  < 2e-16 ***
educationhigh.school         0.3494047  0.0606463   5.761 8.34e-09 ***
educationilliterate          0.3774656  0.4264585   0.885 0.376094
educationprofessional.course 0.7114359  0.0626799  11.350  < 2e-16 ***
educationuniversity.degree   0.6909084  0.0580222  11.908  < 2e-16 ***
campaign                    -0.2200794  0.0092820 -23.710  < 2e-16 ***
poutcomenonexistent          0.8059258  0.0458891  17.562  < 2e-16 ***
poutcomesuccess              2.1772180  0.0806538  26.995  < 2e-16 ***
nr.employed                 -0.0092636  0.0002112 -43.864  < 2e-16 ***
---
```

```
> step.os6.aic
[1] 37744.8
>
> # Confidence Intervals
> exp(cbind("Odds ratio" = coef(step.os6.log), confint.default(step.os6.log, level = 0.95)))
                              Odds ratio        2.5 %       97.5 %
(Intercept)                 8.783984e+19 1.069536e+19 7.214194e+20
jobblue-collar              9.723340e-01 8.887757e-01 1.063748e+00
jobentrepreneur             7.417457e-01 6.222566e-01 8.841798e-01
jobhousemaid                5.924273e+00 5.198695e+00 6.751119e+00
jobmanagement               2.670589e+00 2.422881e+00 2.943622e+00
jobretired                  2.079531e+00 1.819954e+00 2.376131e+00
jobself-employed            7.560698e-01 6.394465e-01 8.939630e-01
jobservices                 7.744095e-01 6.873103e-01 8.725464e-01
jobstudent                  1.804273e+00 1.524264e+00 2.135720e+00
jobtechnician               2.291274e+00 2.106648e+00 2.492081e+00
jobunemployed               7.628665e-01 6.317014e-01 9.212663e-01
educationbasic.6y           1.182984e+00 1.009581e+00 1.386169e+00
educationbasic.9y           4.308660e+00 3.863061e+00 4.805659e+00
educationhigh.school        1.418223e+00 1.259280e+00 1.597227e+00
educationilliterate         1.458583e+00 6.323086e-01 3.364600e+00
educationprofessional.course 2.036914e+00 1.801439e+00 2.303170e+00
educationuniversity.degree  1.995528e+00 1.781022e+00 2.235869e+00
campaign                    8.024551e-01 7.879885e-01 8.171872e-01
poutcomenonexistent         2.238768e+00 2.046201e+00 2.449458e+00
poutcomesuccess             8.821730e+00 7.531840e+00 1.033252e+01
nr.employed                 9.907792e-01 9.903691e-01 9.911893e-01
```

```
> vif.6
                GVIF Df GVIF^(1/(2*Df))
job         2.160664 10        1.039272
education   2.059768  6        1.062066
campaign    1.013858  1        1.006905
poutcome    1.180333  2        1.042320
nr.employed 1.230243  1        1.109163
```

```
> print("Miscalculation Rate
[1] "Miscalculation Rate"
> misClasificError
[1] 0.1933474
```

```
> obj1.mod6.cm
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  9254  651
       yes 1710  741

               Accuracy : 0.809
                 95% CI : (0.802, 0.816)
    No Information Rate : 0.887
    P-Value [Acc > NIR] : 1

                  Kappa : 0.283

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.844
            Specificity : 0.532
         Pos Pred Value : 0.934
         Neg Pred Value : 0.302
             Prevalence : 0.887
         Detection Rate : 0.749
   Detection Prevalence : 0.802
      Balanced Accuracy : 0.688

       'Positive' Class : no
```

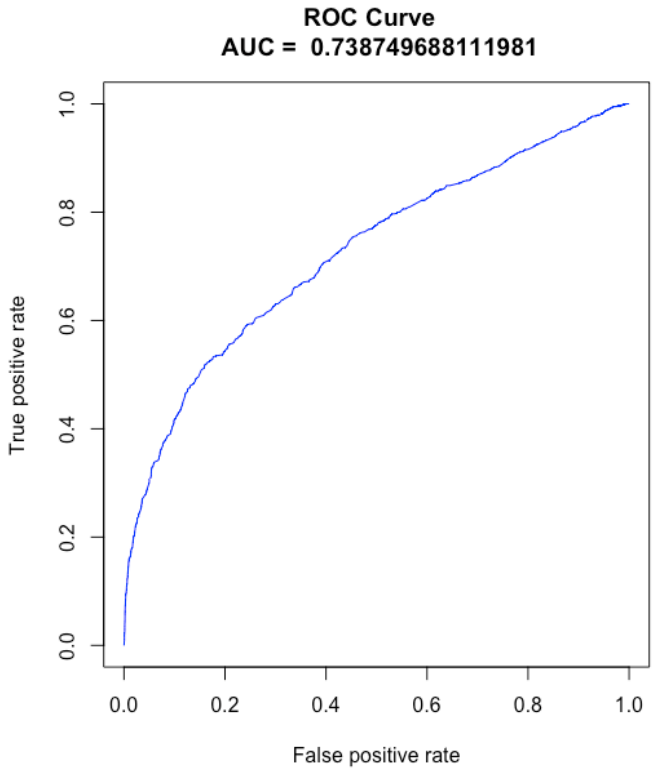

Figure 10 - ANOVA for Objective 2 Complex LR model

```
Analysis of Variance Table

Response: as.numeric(y)
                        Df Sum Sq Mean Sq F value  Pr(>F)
job                      1     50      50  247.98 < 2e-16 ***
education                1     13      13   64.49 1.0e-15 ***
campaign                 1    182     182  894.56 < 2e-16 ***
poutcome                 1    223     223 1099.20 < 2e-16 ***
nr.employed              1    676     676 3328.11 < 2e-16 ***
job:education            1     21      21  101.40 < 2e-16 ***
job:campaign             1      7       7   34.62 4.0e-09 ***
job:poutcome             1      2       2   11.40 0.00073 ***
job:nr.employed          1      0       0    0.02 0.89773
education:campaign       1      2       2    7.90 0.00495 **
education:poutcome       1      1       1    4.21 0.04026 *
education:nr.employed    1     10      10   48.37 3.6e-12 ***
campaign:poutcome        1      0       0    0.96 0.32629
campaign:nr.employed     1     16      16   76.99 < 2e-16 ***
poutcome:nr.employed     1      1       1    4.24 0.03940 *
Residuals            32464   6592       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11 - Results from all the objective 2 models

Method	Accuracy	Sensitivity	Specificity	ErrorRate
LDA Original	0.891	0.960	0.328	0.1093
LDA Balanced	0.727	0.729	0.707	0.2734
LDA Rose	0.711	0.711	0.714	0.2885
QDA Original	0.875	0.922	0.491	0.1253
QDA Balanced	0.857	0.900	0.524	0.1427
QDA Rose	0.873	0.927	0.447	0.1268
MDA Original	0.900	0.984	0.215	0.1003
MDA Balanced	0.735	0.754	0.583	0.2651
MDA Rose	0.762	0.789	0.539	0.2382
FDA Original	0.902	0.987	0.209	0.0981
FDA Balanced	0.678	0.683	0.640	0.3217
FDA Rose	0.663	0.664	0.652	0.3373
Ridge Original	0.901	0.990	0.177	0.0991
Ridge Balanced	0.679	0.684	0.640	0.3212
Ridge Rose	0.662	0.663	0.651	0.3379
Lasso Original	0.900	0.991	0.163	0.0998
Lasso Balanced	0.661	0.665	0.632	0.3388
Lasso Rose	0.651	0.650	0.657	0.3489
Net Original	0.900	0.990	0.163	0.1000
Net Balanced	0.688	0.695	0.634	0.3122
Net Rose	0.661	0.662	0.654	0.3391
RF Original	0.900	0.993	0.146	0.0996
RF Balanced	0.843	0.879	0.567	0.1565
RF Rose	0.892	0.979	0.227	0.1077