Direct Marketing Campaign Modeling and Prediction

Sowmya Mani, Nicole Norelli, and Nnenna Okpara

Introduction

Direct marketing campaigns require communication between a customer and the marketing organization. The goal of direct marketing is to obtain favorable responses from customers for different promotions. This analysis will explore a data set consisting of direct marketing campaign information to better understand and predict if a customer will subscribe to a term deposit. We will use logistic regression to build an interpretable model, and we will compare it with a more complex logistic regression model, a linear discriminant analysis (LDA/QDA) model, and a random forest model to obtain the most accurate subscription predictions.

Data Description

The data set contains 17 variables describing 45,211 customers. It was obtained from the University of California Irvine Machine Learning Repository. The data comes from a Portuguese banking institution's direct marketing campaign. The marketing campaign was based on phone calls, with the goal of persuading a client to subscribe to a term deposit. The data was collected from May 2008 to November 2010. Original variable names, types, and descriptions are available in Appendix Table1.

Exploratory Data Analysis

Missing Data & Preliminary EDA

The data set had no missing data (Figure 1) but the variables with unknown data were: job (288), education (1,857), contact (13,020), and poutcome (36,959). An examination of the unknown data showed that it was not unknown at random (Figure 2). The observations were entered sequentially, and the blocks of unknown values from contact and poutcome were contained within the 2008 customers. An exploration of term deposit subscription proportion by year (Figure 3) showed that the proportions were quite different in each year. 2008 in particular had relatively few term deposit subscriptions. Of course, the 2008 global financial crisis most likely contributed to this. Because the models we are building are more likely to be used during typical financial times, and not during a financial crisis, we decided to limit our analysis to the data from 2009 and 2010. This left us with a data set of 17,481 customers. We felt the data from 2009 and 2010 was more representative of typical consumer behavior and would better explain and predict term deposit subscriptions. The exclusion of 2008 data also eliminated the blocks of unknown values from contact and poutcome.

We chose to include the remaining unknown values in our analysis. Each variable containing unknown values was categorical, and we kept “unknown” as a factor level for job, education, contact, and poutcome. We also added a categorical variable to indicate year.

Next, we modified the pdays variable. The original data coded pdays as “-1” to indicate that a customer had not been previously contacted. The other pdays values (number of days since a client was last contacted from a previous campaign) ranged from 1 to 871. We were able to apply the pdays coefficient in the models to just those previously contacted by changing the “-1” values to “0”. We explored adding a dummy variable to represent the clients who had not been previously contacted. Ultimately we did not keep the dummy variable (temporarily named pcontact) because it was almost identical to the “unknown” level in the poutcome variable (Figure 4).

Finally, we eliminated one outlier in the variable “previous”. It had a value of 275, while the rest of the variable “previous” ranged from 0 to 58. With no other values near this outlier, we decided to restrict the range of our model to clients contacted less than 100 previous times before the current campaign. More data would be required to accurately model clients contacted more than 100 times.

Data Sets: Training and Test

We randomly split our data into an 80% training and 20% test set, with 13,985 observations in the training and 3,496 in the test set. See Appendix Table 2 for the training set summary statistics.

EDA Highlights (Training Data)

Scatterplots and distributions of the continuous variables in the training set can be seen in Figure 5. An examination of the correlations between the continuous variables in our training data (Figure 6) shows one moderate correlation of 0.467 between the variables pdays and previous. Because pdays refers to the days passed since a client was last contacted from a previous campaign, and previous is the number of contacts before this campaign, correlation between the two variables makes sense, particularly because clients who had never been contacted previously would have a value of 0 in both variables.

Most continuous variables did not show a strong relationship with our outcome variable. The most promising was duration (Figure 7), with longer phone calls showing an association with term deposit subscriptions. Many of the categorical variables in our data set did show relationships with our outcome variable. For example, the “success” level of poutcome showed a much higher proportion of term deposit subscriptions than the other levels (Figure 8). This indicates that clients with successful previous campaign outcomes are more likely to subscribe to a term deposit.

We also searched for potential interactions between variables. The clearest visual evidence of an interaction was between year and month. Figure 9 shows that the proportion of yes and no outcomes in many months was different between 2009 and 2010.

Objective 1: Interpretable Logistic Regression Model

Our goal for objective one was to build an interpretable logistic regression model that identified important relationships between predictors and the client's decision to purchase a term deposit.

Model Selection

We attempted a variety of methods to build our interpretable model. These included custom builds guided by intuition and EDA, as well as forward, backward, stepwise, and LASSO variable selection. We compared all models using AIC, accuracy metrics, and AUC. Ultimately we prioritized AUC and parsimony. Our final choice for Model 1 was a customized version of our stepwise model, with a statistically insignificant predictor (age) removed. This model had the highest AUC (Figure 10) of the interpretable models we built (AUC = 0.875), and it had the smallest number of predictors when compared to any other models feature selection suggested.

Checking Assumptions

Logistic regression assumptions require observations to be independent. Each observation was a different client. Details of the source of the client list are unavailable. The data was collected over a period of two years, so there is some danger of serial effects, but we will proceed with caution.

A Hosmer and Lemeshow goodness of fit test rejects the hypothesis that these data fit a logistic regression model (p-value < .0001); however, the sample size of the training set is 13,985. This test is known to always reject if a data set is too large, so this is of little concern.

A look at the Cook's distance and leverage plot (Figure 11) shows a few points to examine, but there do not appear to be any overly influential points. The common theme for the points with higher residuals is that they had the highest values of duration (between 2,389 and 3,102 seconds). While there were not a lot of clients with phone calls this long, they did not appear to be errors and were all under one hour. Additionally, we examined variance inflation factors (VIFs) and did not find any high values.

Parameter Interpretation

Model 1:
Predicted probability of term deposit = Job + Education + Housing + Loan + Contact + Day + Month + Duration + Campaign + Poutcome + Year
The interpretation of parameters in Model 1 matched up nicely with our expectations from the EDA. See Appendix Table 3 for Model 1 coefficients, z-values, p-values, and 95% confidence intervals for coefficients.

Interpretation Example: Duration
The duration variable is statistically significant in Model 1 (p-value < .0001). For every 10 second increase in phone call duration, the odds of subscribing to a term deposit increase by a multiplicative factor of $e^{0.0036539 * 10} = 1.037$, holding other explanatory variables fixed. A 95%

confidence interval for this multiplicative factor is ($e^{0.00344488 * 10}$, $e^{0.00386289 * 10}$) = [1.035, 1.039]. The association between longer calls and term deposit subscriptions could be useful in determining sales training and tactics.

Interpretation Example: Poutcome
The odds ratio of subscribing to a term deposit for clients who had a successful outcome in the previous marketing campaign relative to clients who had a failed outcome in the previous marketing campaign is $e^{1.8137813}$ = 6.1336, after accounting for other explanatory variables. A 95% confidence interval for this odds ratio is [5.1192, 7.3490]. This has practical significance, indicating that clients with previous successful outcomes are much better candidates for the current campaign than those with previous failed outcomes.

Conclusions

Overall, the trends we observed in our EDA were reflected in our simple Model 1. Variables we thought to be important, such as poutcome or duration, show both statistical and practical significance. All of the relationships in Model 1 between the included variables and the outcome were perceptible during our EDA.

Objective 2: Complex Logistic Regression, QDA, Random Forest

Our goal for Objective 2 was to build additional competing models that improved the accuracy of client term deposit purchase predictions. We constructed complex logistic regression, quadratic discriminant analysis, and random tree models to compare with our simple, interpretable model from Objective 1.

Complex Logistic Regression Model (Model 2)

The first step in building our complex logistic regression model was to explore interactions and quadratic terms. Through intuition and EDA, we realized the variable year likely interacted with both month and day. We also added squared terms for each of the continuous variables, letting model selection determine which interaction and squared terms to include. Results indicated that the interactions of year and day, year and month, as well as the squared terms $duration^2$ and $pdays^2$ were statistically significant.

Our year variables posed two problems. First, there was no data collected in December 2010, so the interaction became an issue for some predictions. More importantly, the primary objective of this model is to predict. In reality, it is unlikely that a predictive model would be built in 2021 to predict outcomes in 2009 and 2010. Therefore, we omitted the year variable and its two interactions from our predictive models. We compared models built with forward, stepwise, and LASSO feature selection on AIC, AUC, and accuracy metrics. Our best complex logistic model had an AUC of 0.871 (Figure 12) and came from stepwise feature selection that used AIC as the selection criterion.

Model 2:

Predicted probability of term deposit = Age + Job + Marital + Education + Housing + Loan + Contact + Day + Month + Duration + Campaign + Pdays + Previous + Poutcome + Duration^2 + Pdays^2

LDA/QDA (Model 3)

Linear discriminant analysis (LDA) models assume the explanatory variables have multivariate normality, equal covariance matrices, and independence. Most of our continuous variables did not have normal distributions (Figure 13), and our attempts to correct the distributions with various log and square strategies were unsuccessful. Scatterplots of pairs of the continuous variables color coded by response (Figure 14) show the equal covariance matrices assumption to also be unmet. This led us to choose a quadratic discriminant analysis (QDA) model, which does not require equal covariance matrices. Of course the normality assumption was still problematic, but we decided to proceed with caution to see how the model's predictions performed.

Before we built the model, we conducted principle component analysis (PCA) on the continuous variables. We knew from our EDA that the majority of our most promising predictors were categorical, so we did not expect high performance from a model built with only continuous predictors. A scatterplot of our first two principal components color-coded by response (Figure 15) shows some separation, but not enough to expect good performance out of QDA. PCA analysis shows that we would require 7 PCs to explain 94% of the variance (Figure 16).

We compared a variety of QDA models, attempting to remove any variables that could be unimportant to improve performance. None of our attempts significantly improved our accuracy metrics, so we decided to keep all continuous predictors in our model. Our final QDA model had an AUC of 0.731 (Figure 17).

Random Forest (Model 4)

To build a realistic predictive model, we trained and tested our random forest model without the year variable as well. We used 500 trees and a mtry value of 4. This number of randomly selected variables was the optimal mtry value with respect to out-of-bag error estimates. Variable importance plots (Figure 18) indicate that the duration and month variables are some of the more important predictors in this model. The random forest model had an AUC of 0.890 (Figure 19).

Model Comparisons

A comparison of the performance of the four models is in Table 4. The random forest model (Model 4) outperforms the other models we built, with the highest AUC (Figure 20) and highest sensitivity. Model 4 was also second in overall accuracy, only 0.0046 behind Model 1 (simple logistic). Model 4's superior performance could be due to the nonparametric random forest modeling some complexity we were unable to see in our EDA or add to our complex logistic regression model. It is notable that our simple logistic (Model 1) performed slightly better than our complex logistic (Model 2). This is likely due to the presence of the year variable in the

simple model. We were not surprised to find the QDA (Model 3) performed the worst, as the continuous variables in our model did not appear to be great predictors of term deposit subscriptions.

Table 4
Model Comparisons

	AUC (test)	Accuracy (test)	Sensitivity (test)	Specificity (test)	Cut-off
Logistic Model 1 (simple)	0.875	0.7875	0.8128	0.7769	0.2
Logistic Model 2 (complex)	0.871	0.7775	0.8385	0.7599	0.2
QDA Model 3	0.731	0.6899	0.6462	0.7025	0.15
Random Forest Model 4	0.890	0.7829	0.9115	0.7459	0.2

As the goal of these campaigns is to obtain subscriptions to term deposits, identifying clients who are likely to subscribe is the most important objective. Sensitivity (correctly identifying clients who say yes to the term deposit) is essential for a predictive model that is useful and practical. To that end, we adjusted the prediction cut-off values for each of our models. Our goal was to maximize sensitivity while still keeping specificity and overall accuracy to an acceptable level. Our random forest (Model 4) correctly identified 711 of the 780 yes outcomes in our test set. Of course the specificity decreased to 0.7459 and overall accuracy decreased to 0.7829, but our priority is predicting purchases, and we are willing to accept a higher number of false positives and lower overall accuracy to achieve this goal.

Conclusion

Our best predictive model was Model 4 (Random Forest Model). Although we adjusted the cut-off and worsened the overall accuracy, it was still within acceptable limits and Model 4 was by far the best in terms of sensitivity. As sensitivity was a crucial aspect of our predictive model, this makes it the clear choice. It is likely that the nonparametric random forest modeled some complexity the other models missed. Our logistic models performed well, but our simple Model 1 was not built for predictions due to the inclusion of the year variable. Of course, Model 3 (QDA) was the poorest performer, but this was expected. Our continuous variables were not our strongest predictors, and the variables we used did not meet the normality assumptions for QDA.

Limitations

One limitation to our predictive models is the duration variable, as the duration of the phone call is not known ahead of time. It could be useful in a real-time update scenario, where all the other variables are known and the caller can see the probability change as time increases. If the model is meant to identify potential clients from a database, however, a new model without a duration variable might be required.

Another limitation is the time frame of this data. Some of the variables in the data set could have different relationships with the outcome than they did more than ten years ago. For example, contact indicates a communication type of unknown, telephone, or cellular. Cell phones are much more prevalent today, and people who continue to use landlines today may differ in meaningful ways from people who kept landlines in 2009.

This data set is from a Portuguese banking institution, so we should use caution when extending the scope of inference for these models. There may be cultural differences for which we cannot account, and it is unclear how the banking institution obtained the client list used for this campaign. Of course, no causal inferences can be made from our explanatory model as this is an observational study.

Future Directions

Access to data from years subsequent to 2010 could be useful to improve our predictive model as 2008 was a financial crisis period and year 2009 was not as good as the industries were getting back from the crisis. Our current predictions are based on purchasing decisions from a combination of 2009 and 2010 data, but the proportions of subscription purchases were still different between these two years. While we are fairly confident these data are more representative of typical consumer behavior than the 2008 data, it is possible that ongoing economic recovery resulted in lower than normal subscription rates in 2009 as well. If this was the case, then 2010 could be a better data set for a predictive model. It would be helpful to compare overall subscription rates during a few of the subsequent years to further explore this issue. Of course, more up-to-date information would be extremely helpful.

References

For original data set, please see:
https://github.com/NicoleNorelli/6372Project2/blob/main/bank-full.csv

For principle code:
https://github.com/NicoleNorelli/6372Project2/blob/main/Project2Models6372NN.RMD

https://github.com/NicoleNorelli/6372Project2/blob/main/Project2ComplexModels6372NN.Rmd

https://github.com/NicoleNorelli/6372Project2/blob/main/Bank_Sow_v2.Rmd

For final models:
https://github.com/NicoleNorelli/6372Project2/blob/main/Project2FinalModels6372.Rmd

For additional EDA, visualizations, and data exploration:
https://github.com/NicoleNorelli/6372Project2/blob/main/Project2EDA6372NN.Rmd

https://github.com/NicoleNorelli/6372Project2/blob/main/Bank_Sow_v2.docx

Appendix

Table 1

Original variable descriptions

Variable Name	Data Type	Description
age	Numeric	Age
job	Factor	Admin., unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services
marital	Factor	Married, divorced, single (divorced means divorced or widowed)
education	Factor	Unknown, secondary, primary, tertiary
default	Factor	Has credit in default? (yes/no)
balance	Numeric	Average yearly balance (in Euros)
housing	Factor	Has housing loan? (yes/no)
loan	Factor	Has personal loan? (yes/no)
contact	Factor	Contact communication type: unknown, telephone, cellular
day	Numeric	Last contact day of the month
month	Factor	Last contact month of the year
duration	Numeric	Last contact duration (in seconds)
campaign	Numeric	Number of contacts performed during this campaign and for this client (includes last contact)
pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
previous	Numeric	Number of contacts performed before this campaign and for this client
poutcome	Factor	Outcome of previous marketing campaign: unknown, other, failure, success
y	Factor	Response variable: has the client subscribed a term deposit? (yes/no)

Table 2

Training set summary statistics

Characteristic	N = 13,985
age, Mean (SD) Min Max	40 (12) 18 95
job, n (%)	
admin.	1,843 (13%)
blue-collar	2,685 (19%)
entrepreneur	376 (2.7%)
housemaid	280 (2.0%)
management	2,897 (21%)
retired	963 (6.9%)
self-employed	437 (3.1%)
services	1,237 (8.8%)
student	624 (4.5%)
technician	2,040 (15%)
unemployed	537 (3.8%)
unknown	66 (0.5%)
marital, n (%)	
divorced	1,488 (11%)
married	7,550 (54%)
single	4,947 (35%)
education, n (%)	
primary	1,872 (13%)
secondary	7,149 (51%)
tertiary	4,362 (31%)
unknown	602 (4.3%)
default, n (%)	121 (0.9%)
balance, Mean (SD) Min Max	1,442 (3,048) -4,057 102,127
housing, n (%)	7,756 (55%)
loan, n (%)	1,631 (12%)
contact, n (%)	
cellular	12,558 (90%)
telephone	1,238 (8.9%)
unknown	189 (1.4%)
day, Mean (SD) Min Max	14 (8) 1 31
month, n (%)	
jan	1,114 (8.0%)
feb	2,122 (15%)
mar	388 (2.8%)
apr	2,395 (17%)
may	4,617 (33%)
jun	675 (4.8%)
jul	397 (2.8%)
aug	845 (6.0%)
sep	451 (3.2%)
oct	537 (3.8%)
nov	291 (2.1%)
dec	153 (1.1%)
duration, Mean (SD) Min Max	267 (251) 0 3,102
campaign, Mean (SD) Min Max	2 (2) 1 23
pdays, Mean (SD) Min Max	97 (138) 0 871
previous, Mean (SD) Min Max	1 (3) 0 55
poutcome, n (%)	
failure	3,347 (24%)
other	1,322 (9.5%)
success	1,168 (8.4%)
unknown	8,148 (58%)
y, n (%)	3,108 (22%)
id, Mean (SD) Min Max	36,436 (5,035) 27,730 45,211
year, n (%)	
2009	11,923 (85%)
2010	2,062 (15%)

Table 3

Model 1 coefficients

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.6590281	0.2032132	-22.927	< 2e-16	***
jobblue-collar	-0.3306582	0.1037829	-3.186	0.00144	**
jobentrepreneur	-0.5635533	0.1955900	-2.881	0.00396	**
jobhousemaid	-0.1300422	0.1889792	-0.688	0.49137	
jobmanagement	0.0061455	0.0976748	0.063	0.94983	
jobretired	0.0483610	0.1142460	0.423	0.67207	
jobself-employed	-0.0775205	0.1585802	-0.489	0.62495	
jobservices	-0.1425890	0.1177365	-1.211	0.22586	
jobstudent	0.0601339	0.1245440	0.483	0.62921	
jobtechnician	-0.0729717	0.0951414	-0.767	0.44309	
jobunemployed	-0.1898567	0.1461852	-1.299	0.19403	
jobunknown	-0.1901727	0.3193244	-0.596	0.55148	
educationsecondary	0.2112835	0.0921315	2.293	0.02183	*
educationtertiary	0.4110908	0.1037754	3.961	7.45e-05	***
educationunknown	0.2481380	0.1420650	1.747	0.08070	.
housingyes	-0.7052064	0.0620790	-11.360	< 2e-16	***
loanyes	-0.2858664	0.0955147	-2.993	0.00276	**
contacttelephone	-0.5003948	0.0996739	-5.020	5.16e-07	***
contactunknown	-1.6713815	0.2525049	-6.619	3.61e-11	***
day	0.0267328	0.0033858	7.896	2.89e-15	***
monthfeb	1.3990019	0.1477082	9.471	< 2e-16	***
monthmar	2.7525298	0.1693802	16.251	< 2e-16	***
monthapr	1.4995269	0.1325282	11.315	< 2e-16	***
monthmay	1.1455802	0.1360434	8.421	< 2e-16	***
monthjun	2.5268316	0.1575433	16.039	< 2e-16	***
monthjul	2.2350589	0.1751459	12.761	< 2e-16	***
monthaug	2.1875072	0.1474579	14.835	< 2e-16	***
monthsep	2.2330499	0.1689553	13.217	< 2e-16	***
monthoct	2.0179585	0.1572513	12.833	< 2e-16	***
monthnov	2.6279951	0.1843426	14.256	< 2e-16	***
monthdec	2.6366144	0.2240066	11.770	< 2e-16	***
duration	0.0036539	0.0001066	34.265	< 2e-16	***
campaign	-0.0819894	0.0186665	-4.392	1.12e-05	***
poutcomeother	0.1689044	0.1056138	1.599	0.10976	
poutcomesuccess	1.8137813	0.0922363	19.665	< 2e-16	***
poutcomeunknown	0.4775334	0.0694347	6.877	6.09e-12	***
year2010	1.0684874	0.0695511	15.363	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	Odds ratio	2.5 %	97.5 %
(Intercept)	0.009475667	0.006362598	0.01411189
jobblue-collar	0.718450688	0.586214808	0.88051578
jobentrepreneur	0.569183006	0.387940765	0.83509990
jobhousemaid	0.878058341	0.606267125	1.27169431
jobmanagement	1.006164450	0.830860393	1.21845608
jobretired	1.049549452	0.838989531	1.31295328
jobself-employed	0.925408064	0.678187180	1.26274885
jobservices	0.867110388	0.688425398	1.09217415
jobstudent	1.061978697	0.831962463	1.35558851
jobtechnician	0.929627112	0.771479346	1.12019404
jobunemployed	0.827077618	0.621030904	1.10148687
jobunknown	0.826816362	0.442180577	1.54603194
educationsecondary	1.235262452	1.031185416	1.47972741
educationtertiary	1.508462245	1.230837239	1.84870775
educationunknown	1.281636725	0.970150185	1.69313238
housingyes	0.494006612	0.437412354	0.55792327
loanyes	0.751363023	0.623085420	0.90604976
contacttelephone	0.606291278	0.498699308	0.73709570
contactunknown	0.187987189	0.114602927	0.30836196
day	1.027093284	1.020300034	1.03393176
monthfeb	4.051154428	3.032839146	5.41138234
monthmar	15.682255487	11.252053726	21.85673328
monthapr	4.479569499	3.454840993	5.80823920
monthmay	3.144264992	2.408345452	4.10505990
monthjun	12.513794144	9.189416738	17.04080339
monthjul	9.347032263	6.631156034	13.17523094
monthaug	8.912967036	6.675839948	11.89977321
monthsep	9.328272638	6.698634025	12.99021115
monthoct	7.522950856	5.527587787	10.23860529
monthnov	13.845981601	9.647417509	19.87176426
monthdec	13.965840050	9.003104312	21.66415955
duration	1.003660572	1.003450827	1.00387036
campaign	0.921281706	0.888185051	0.95561165
poutcomeother	1.184007001	0.962621700	1.45630685
poutcomesuccess	6.133596733	5.119216975	7.34897721
poutcomeunknown	1.612093029	1.406977619	1.84711107
year2010	2.910973115	2.540014290	3.33610898

Figure 1

Original Full Data Set: Null data summary

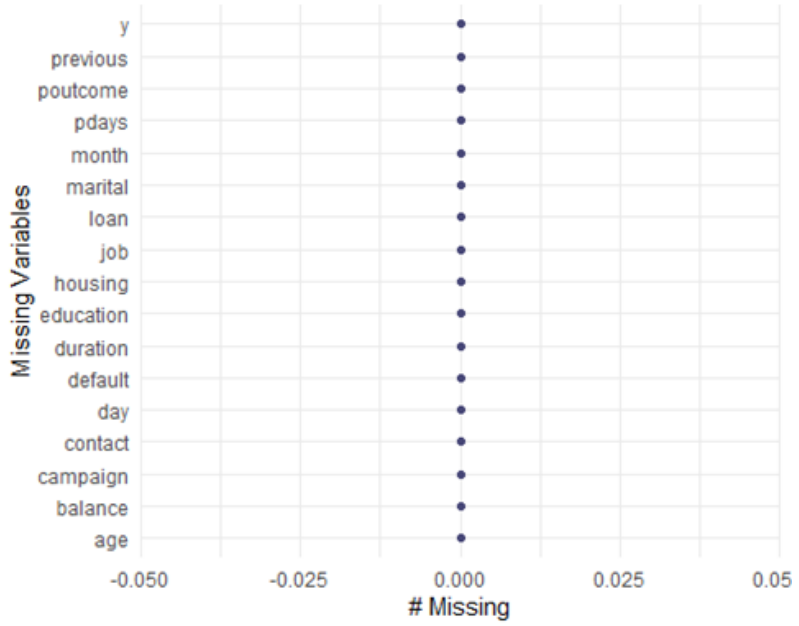

Figure 2

Original Full Data Set: Unknown Data

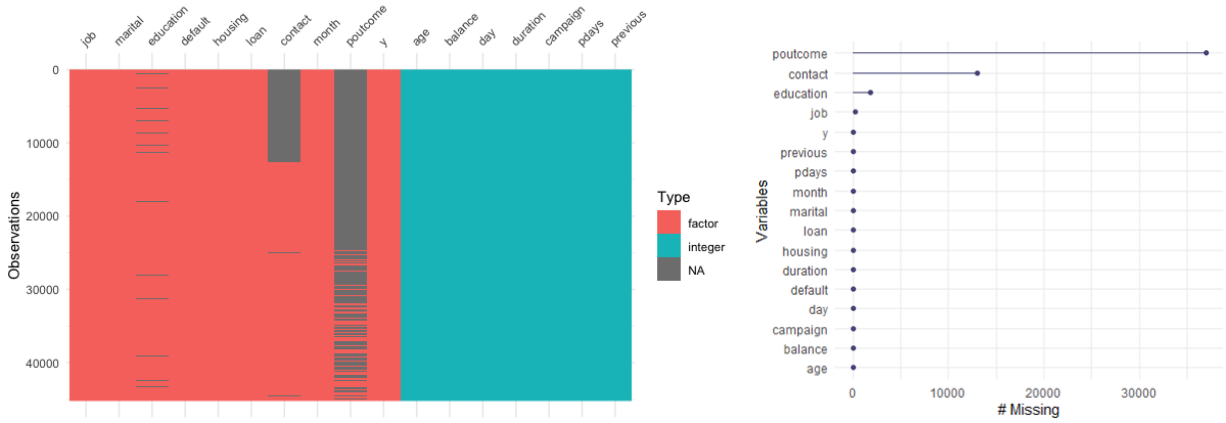

Figure 3

Original Full Data Set: Term Deposits by Year

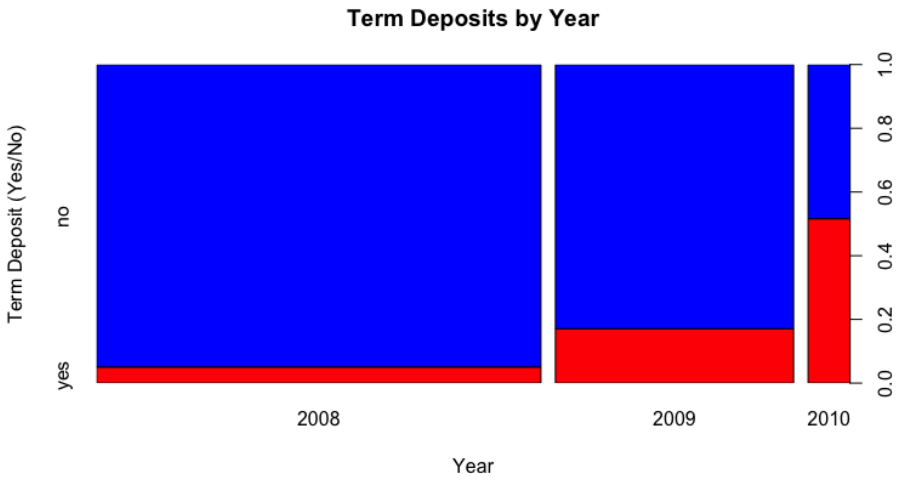

Figure 4

Proportion of temporary dummy variable (pcontact) in each level of poutcome

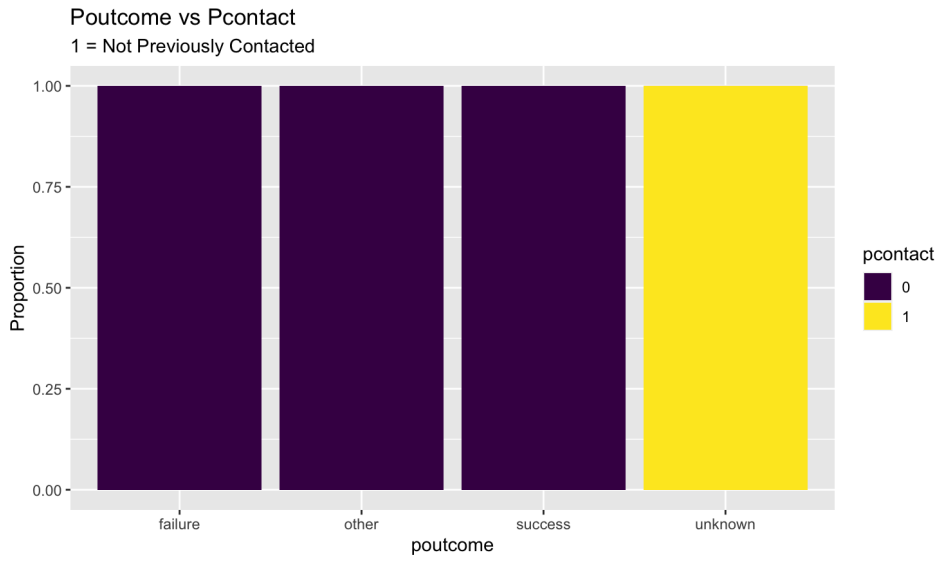

Figure 5

Correlations and Scatterplots of Continuous Variables by Term Deposit

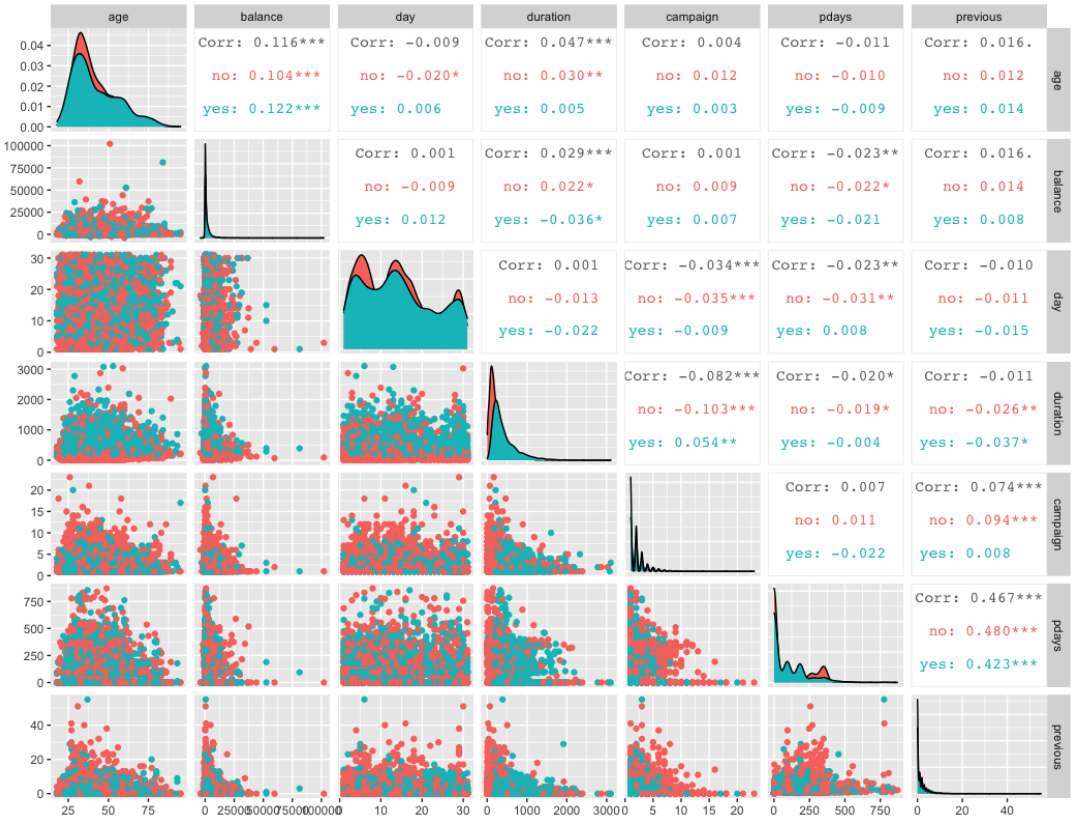

Figure 6

Correlations of Continuous Variables

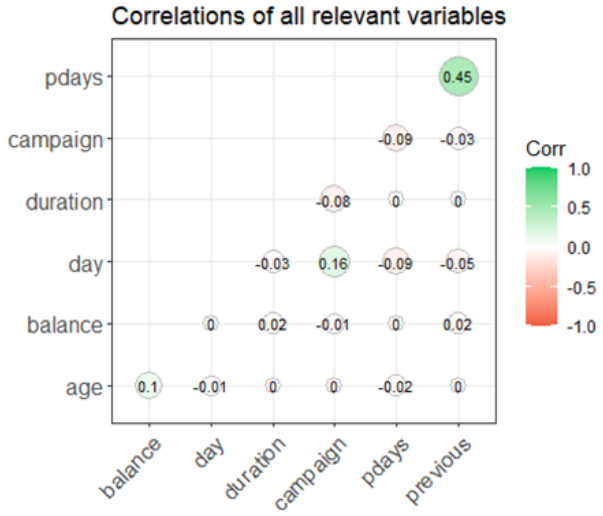

Figure 7

Boxplots of Duration by Term Deposit

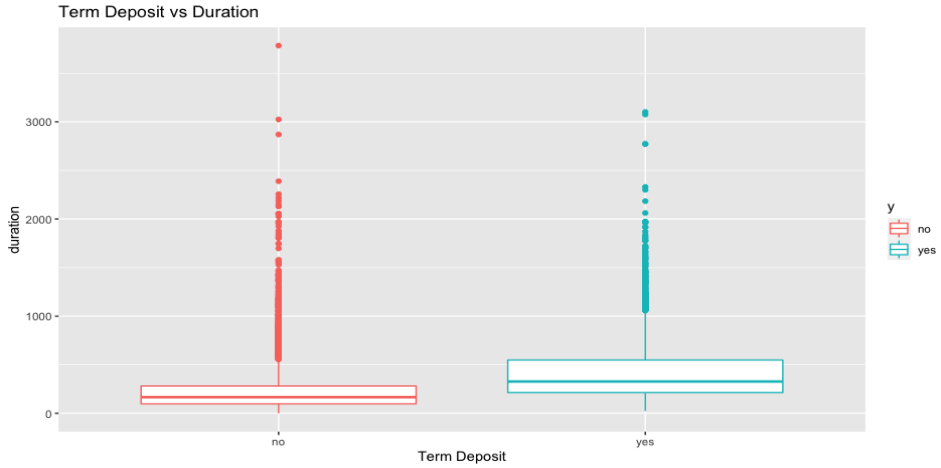

Figure 8

Proportion of Term Deposits by Poutcome

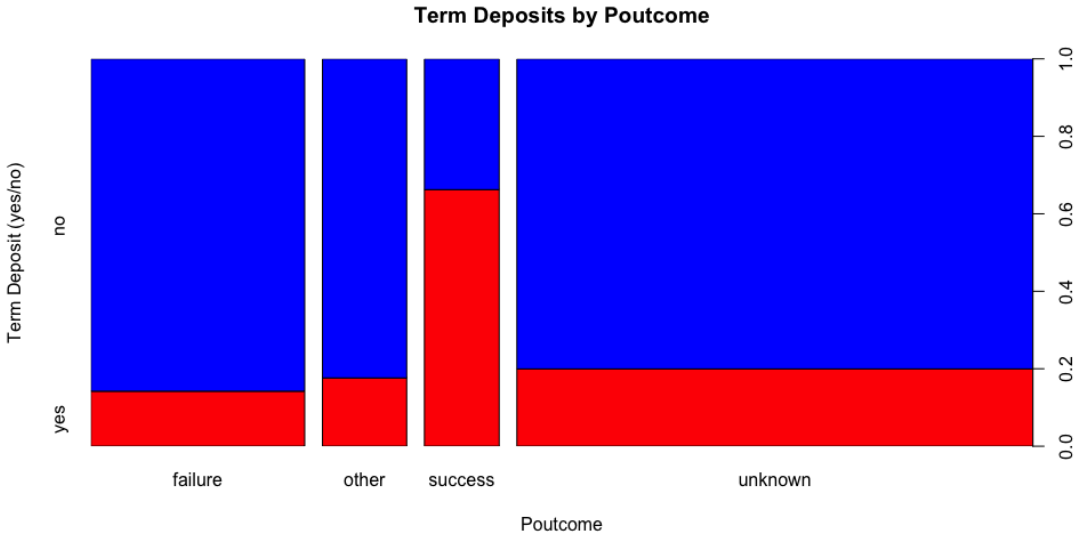

Figure 9

Evidence of an interaction between year and month variables

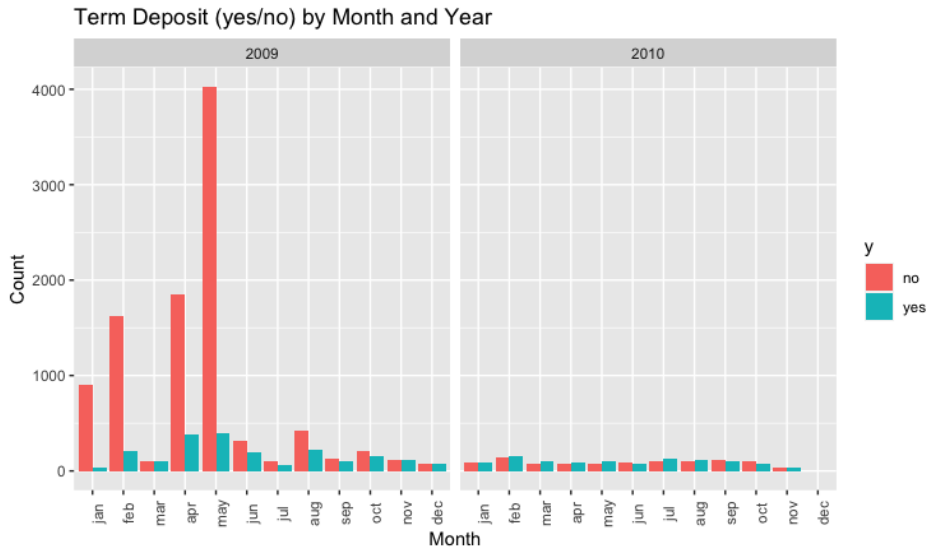

Figure 10

Simple Logistic Regression (Model 1) ROC Curve (Stepwise)

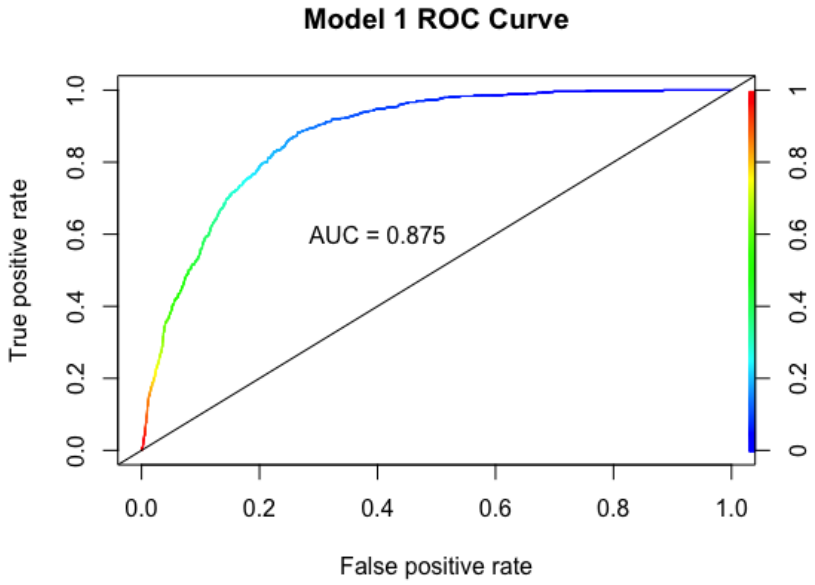

Figure 11

Simple Logistic Regression (Model 1) Residual Plots

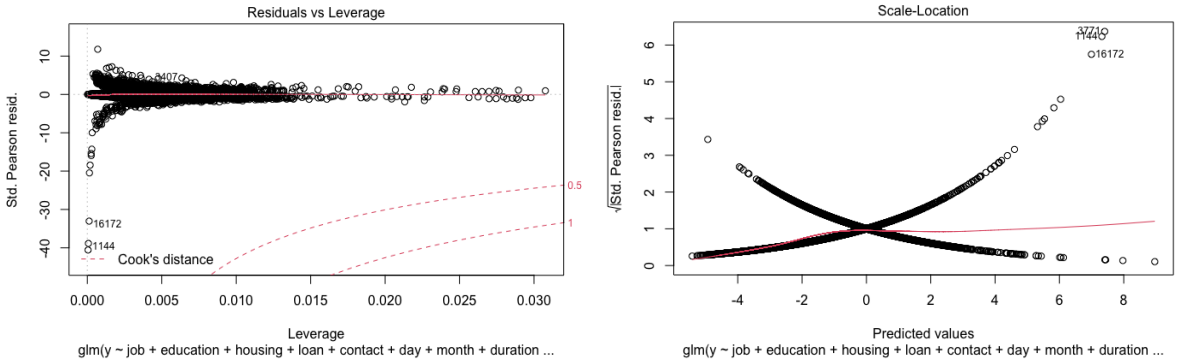

Figure 12

Complex Logistic Regression (Model 2) ROC Curve

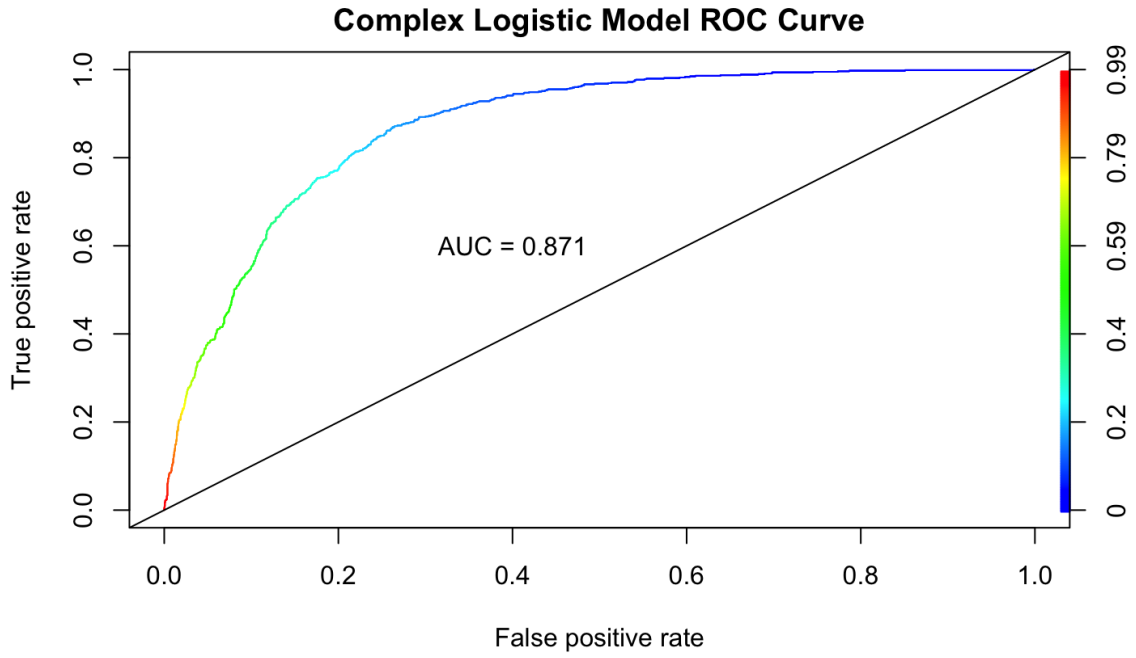

Figure 13
Distribution of continuous variables:

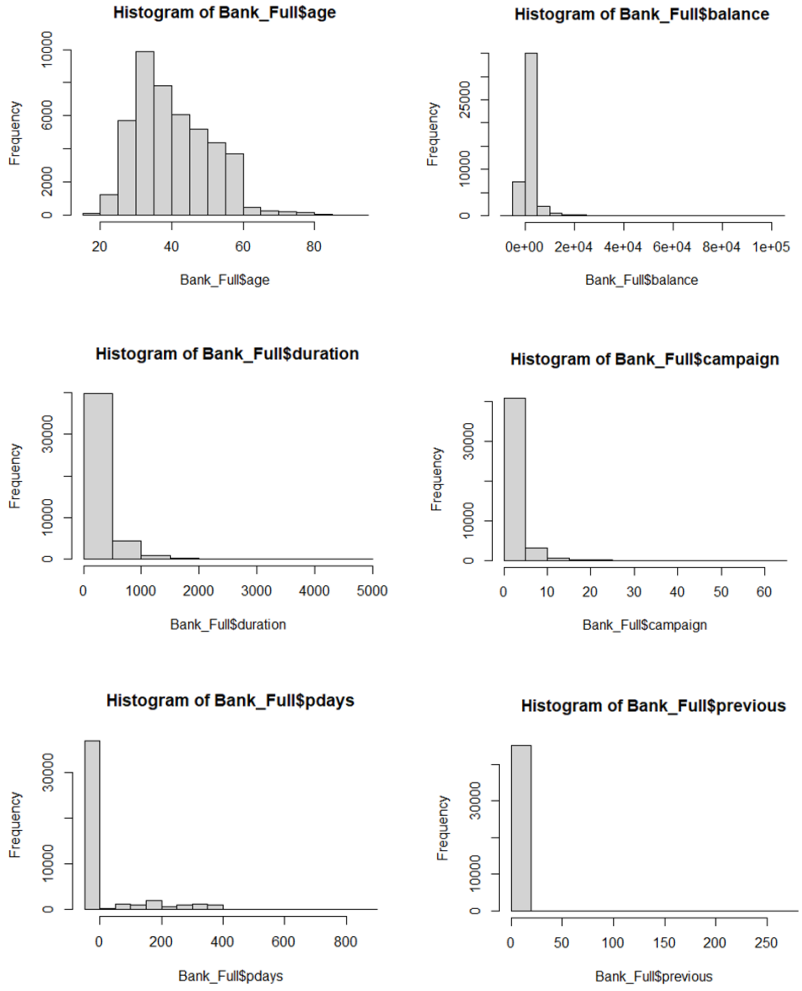

Figure 14

Scatterplots of pairs of the continuous variables color coded by response

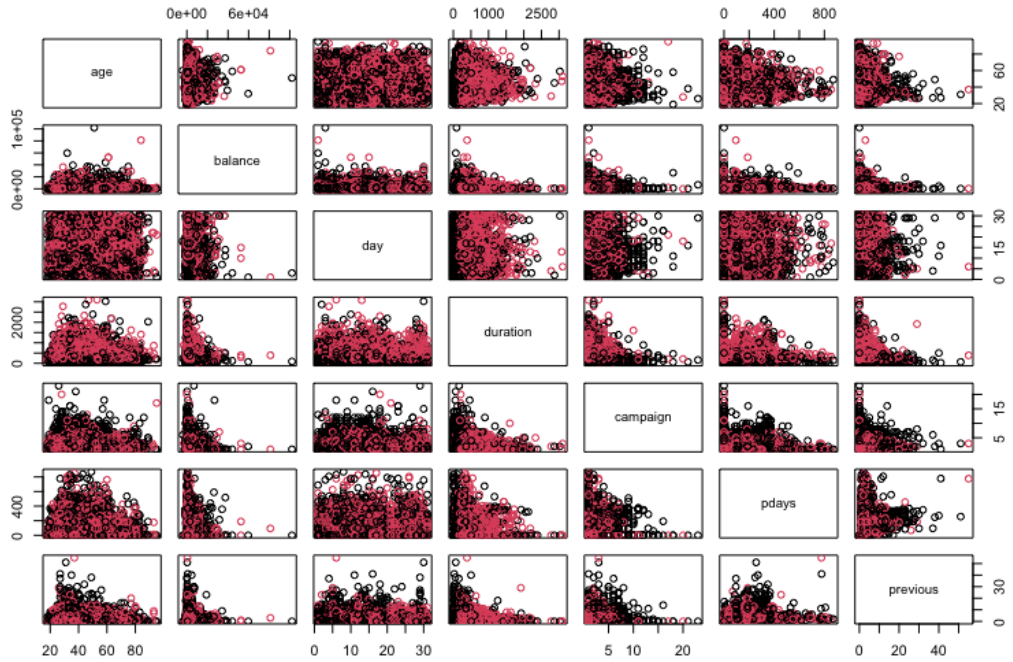

Figure 15

PCA of training data set: PC1 vs. PC2

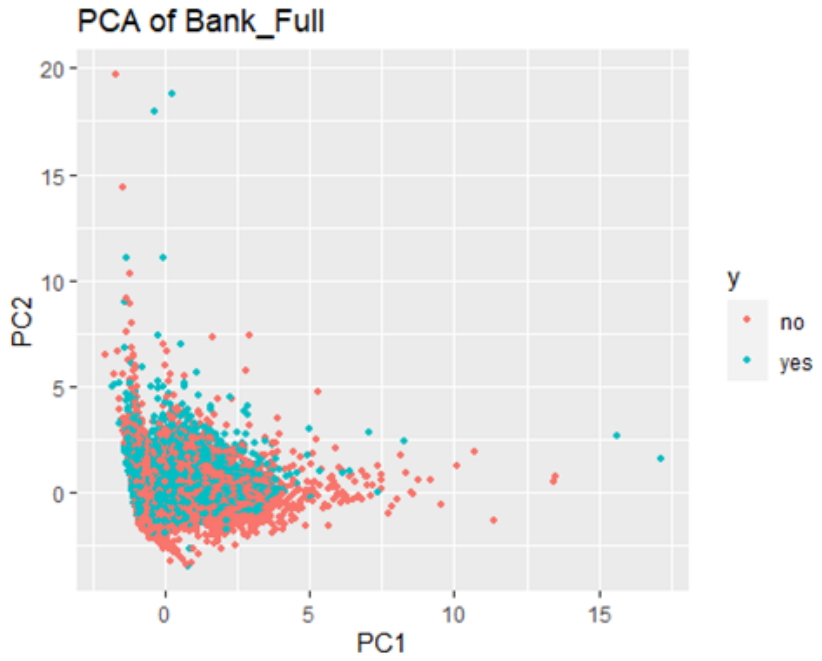

Figure 16
PCA Scree Plot

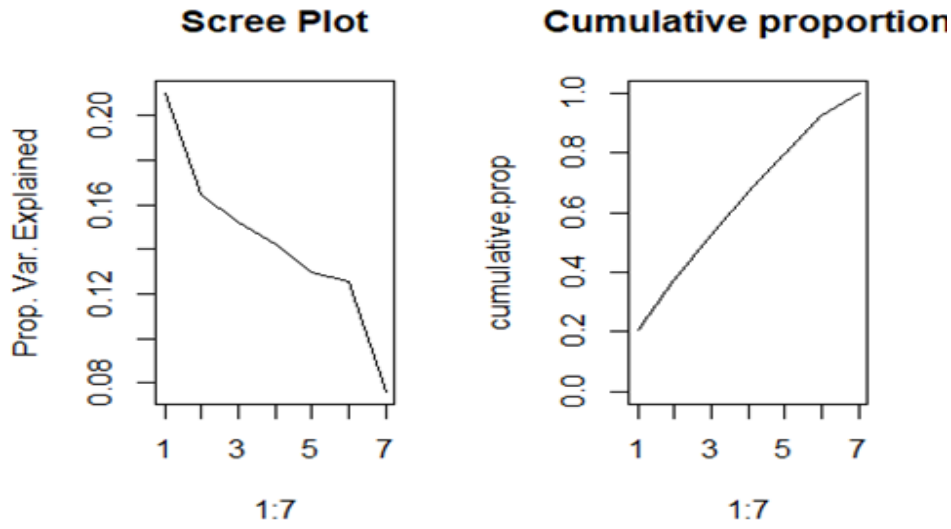

Figure 17
QDA (Model 3) ROC Curve

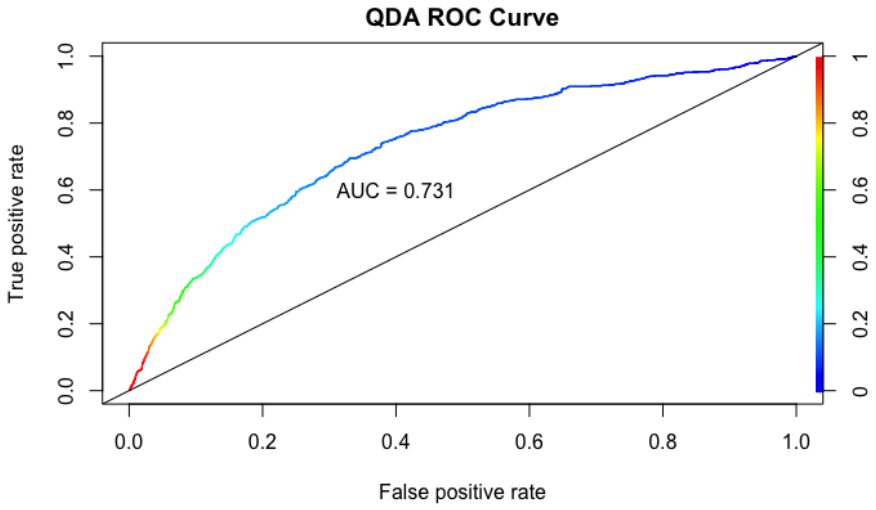

Figure 18

Variable Importance Plot for Random Forest (Model 4)

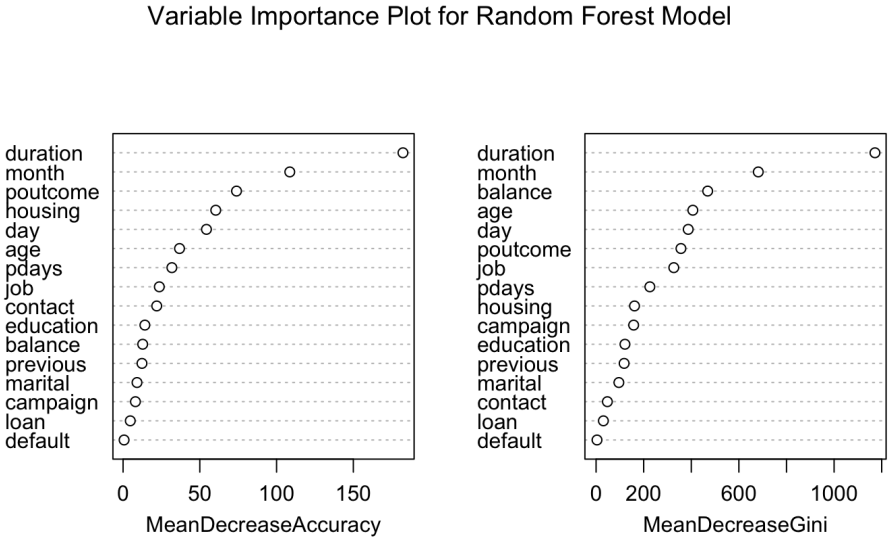

Figure 19

Random Forest (Model 4) ROC Curve

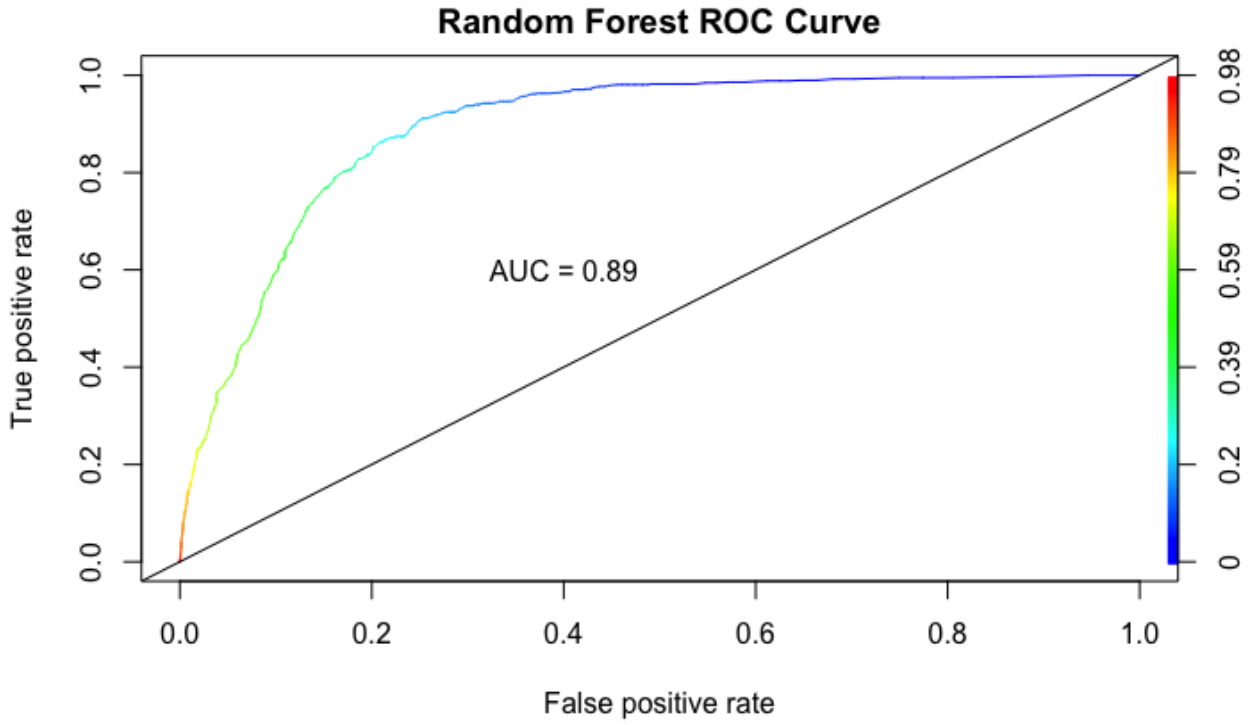

Figure 20

Model Comparisons of ROC