Applied Statistics Project 2

An Analysis of Banking Sales Performance

Authors: Jon Paugh, Laura Lazarescou, Andrew Leppla

Table of Contents

[Introduction 2](#_Toc38102793)

[Data Description 2](#_Toc38102794)

[Data Structure: bank-additional-full.csv 3](#_Toc38102795)

[Exploratory DATA Analysis (EDA) 4](#_Toc38102796)

[Collinearity 5](#_Toc38102797)

[Missing Values 8](#_Toc38102798)

[Mosaic Plots for Logistic Regression 9](#_Toc38102799)

[Objective 1 – Interpretable logistic regression model 10](#_Toc38102800)

[Data Preparation 10](#_Toc38102801)

[Dataset Creation Method 11](#_Toc38102802)

[Logistic Regression Model Creation 11](#_Toc38102803)

[Model Selection 12](#_Toc38102804)

[Objective 1 Models 13](#_Toc38102805)

[Assumptions 13](#_Toc38102806)

[Hypothesis Tests 15](#_Toc38102807)

[Maximum Likelihood Estimates 16](#_Toc38102808)

[Odds Ratio Estimates 16](#_Toc38102809)

[Conclusion 18](#_Toc38102810)

[Objective 2 – Predictive model 19](#_Toc38102811)

[Random Forest ROC Curves (mtry=9, ntree=1000) 20](#_Toc38102812)

[Model Description for Random Forest (mtry=9, ntree=1000) 21](#_Toc38102813)

[Objective 2 Conclusion 21](#_Toc38102814)

[Compare All Models 22](#_Toc38102815)

[Final Conclusions 22](#_Toc38102816)

[References 23](#_Toc38102817)

[Appendix: Code for EDA, model development and comparison 23](#_Toc38102818)

# Introduction

Hiring a sales team and paying them to contact potential customers is an expensive investment for any organization. One way to reduce costs and increase the lead conversion rate is to use statistical and machine learning techniques to build a model of the factors that contribute to lead conversion. This model can then be used both to better understand the factors leading to lead conversion, and to predict for a given potential customer whether success is likely.

Our data science team analyzed a sales interaction dataset from a Portuguese retail bank with the goal of developing models to interpret and predict the success of these interactions. This report will review the data we have received and identify some factors that correlate with higher or lower sales outcomes. Later we will review the various models that we produced and compare the relative merits of the models based on their ability to predict, using a held-out validation dataset. We will present one interpretive model which allows us to estimate the impact for each variable on the probability of lead conversion. We will also review another model that has higher predictive success but may not be as easy to interpret.

# Data Description

We obtained the sales interaction dataset from the University of California at Irvine Machine Learning Data Repository. (See Reference 1) This dataset is a subset of the data used by the team of Moro, Cortez and Rita for the paper “A data-driven approach to predict the success of bank telemarketing”. (See Reference 2) This dataset has fewer rows than the original dataset described in the study– 41,189 rows vs 52,944 rows. Also, there are many columns that are not included from the original dataset: interest rates, gender of the agent, agent experience level. We are working with a smaller, narrower dataset and we cannot expect to get the same results as the original study. Another difference is that the original study included a full date, and the data up to June 2012 was used as training data. The data between July 2012 and June 2013 was used as test data. The dataset we are using identifies the month but not the year - so we cannot break down the train and test data in the same way.

The dataset comes from a Portuguese retail bank and contains a list of calls between sales agents and potential customers or “leads”. For each lead there is a “y” yes/no value which indicates whether the contact resulted in a sale. The dataset includes three basic types of information about the call - lead demographic data, history of the sales contact, and background economic data for the time when the contact took place. First there is demographic information about the lead; their age, job, marital status, education, whether they have a prior loan default, housing status, and loan. Then there is information about the contact method and history of the sales effort: the method of contact (phone vs. Cellular), the month, day, and day of the week of the contact, the duration of the last contact, the number of contacts performed during this campaign, the number of days since the lead was last contacted, number of days since contact from previous campaigns, and the outcome from contacts during the prior campaign. Finally, there are economic indicators from the time period of the contact: employment variation rate, consumer price index, consumer confidence index, Euribor 3-month interest rate (See reference 3) and number of employees (quarterly indicator).

The dataset includes the duration of the final call between the salesperson and the sales lead. This value is highly correlated with the success or failure of the call, but it is not predictive, because it is not known in advance of the call. For this reason, the documentation from the dataset suggests not to include this in any predictive model. We will not include duration in our interpretive or our predictive models.

## Data Structure: bank-additional-full.csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Description | Min | Max | Median | Mean |
| age | Age of the sales lead | 17.00 | 98.00 | 38.00 | 40.02 |
| job | Type of job. Categorical variable. Values : ‘admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown' | | | | |
| marital | Marital status of the sales lead. Categorical variable. Values : 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed | | | | |
| education | Education of the sales lead. Categorical variable. (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown') | | | | |
| default | is the sales lead’s credit in default? (categorical: 'no','yes','unknown') | | | | |
| housing | does the sales lead have a housing loan? (categorical: 'no','yes','unknown') | | | | |
| loan | has personal loan? (categorical: 'no','yes','unknown') | | | | |
| month | last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec') | | | | |
| day\_of\_week | last contact day of the week (categorical: 'mon','tue','wed','thu','fri') | | | | |
| duration | last contact duration, in seconds (numeric). | 0.00 | 4918.0 | 180.0 | 258.3 |
| campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) | 1.00 | 56.00 | 2.00 | 2.568 |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) | 0 | 999.0 | 999.0 | 962.5 |
| previous | number of contacts performed before this campaign and for this client (numeric) | 0 | 7.000 | 0.00 | 0.173 |
| poutcome | outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success') | | | | |
| Emp.var.rate | employment variation rate - quarterly indicator (numeric) | 3.400 | 1.400 | 1.100 | 0.08189 |
| Cons.price.idx | consumer price index - monthly indicator (numeric) | 92.20 | 94.77 | 93.75 | 93.58 |
| Cons.conf.idx | consumer confidence index - monthly indicator (numeric) | -50.8 | -26.9 | -40.5 | -41.8 |
| euribor3m | euribor 3 month rate - daily indicator (numeric) | 0.634 | 5.045 | 4.857 | 3.621 |
| Nr.employed | number of employees - quarterly indicator (numeric) | 4964 | 5228 | 5191 | 5167 |
| y | This is the response variable. Has the client subscribed to a term deposit? (binary: 'yes','no') | | | | |

# Exploratory DATA Analysis (EDA)

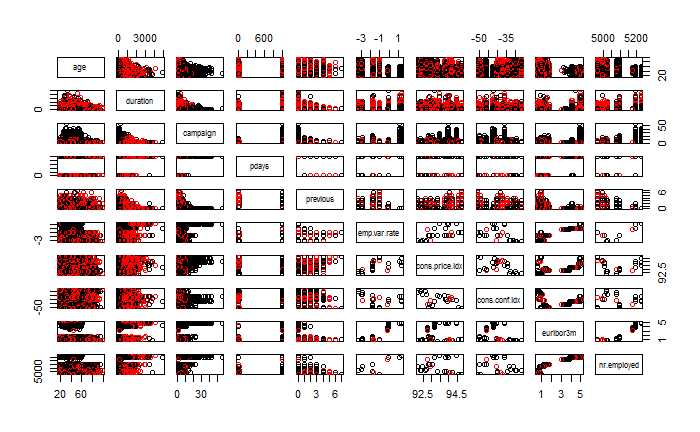
The response variable “y” is categorical with two values, “yes” and “no”. The bar plot of our response variable “y” below shows that only 10.13% of the y values are “yes”. If we use the model to identify and eliminate leads who are likely to say “no”, we expect the percent of “yes” responses to increase. That is our goal in doing this type of analysis.

If we build a model based on the full dataset, the accuracy will primarily reflect the ability to predict the “no” values. A model which simply returns “no” for all cases should get a 90% accuracy rate. That does not help us identify the yeses. To identify the yeses we will have to focus not on the overall accuracy but on the sensitivity (correct identification of yes responses) and specificity (correct identification of no responses). We have chosen to prioritize sensitivity over specificity in our models since an increase in true positives is more valuable than the potential cost of additional true negatives. The bank’s primary objective is to maximize number of depositors, and for each incremental percent of yeses our margins will cover the cost of false yes predictions.

Additional bar plots below show the number of each group within different variables. From a high level, we see most of our leads have administrative, blue collar or technician jobs. They tend to be married, have at least a high school education and have not defaulted on a loan. Month is not significant because we don’t have a complete date, so it is likely that the period includes two months of May. The variable “poutcome” has a significant proportion of “nonexistent” which is for no prior contact in the last campaign.

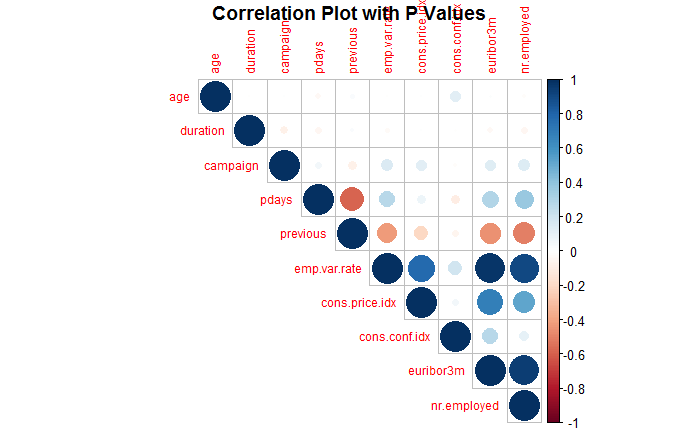
|  |  |
| --- | --- |
|  |  |

The matrix below shows scatter plots for each numerical variable with red representing yes and black representing no. We don’t see any significant linear relationships between the continuous variables, although age is right skewed and may benefit from log transformation. As we discussed, duration is not practically significant since it is not known until the decisive call is complete.



## Collinearity

The correlation plot below shows the correlation between the various continuous variables. This does not include the y response, so it is helpful for understanding correlation but not identifying differences in yes vs. no responses. Most of the demographic variables are categorical so they will not appear here. Only age appears and it has very little correlation with other variables. Pdays is negatively correlated with previous. We do see strong correlation between the various economic indicators – emp.var.var, cons.price.ifx, euribor3m, and nr.employed are highly correlated. This is not surprising since they are all reflecting the economic conditions of that time period, so they will likely trend up and down together. As we look at simple models or focus on hypothesis testing, we will need to choose which economic variables to keep, since their collinearity will prevent us from explaining the contribution of each variable.



Among the demographic variables were several that have higher or lower values for different factor levels. Job had higher success rates for retired, students in particular. Single and Unknown Marital status gave a slight boost to yes’s as well. Education of Illiterate, University, and Unknown also contributed to higher success. If the contact with the client was via cellular, it was more likely to result in success. Finally, if poutcome indicates the prior contact with the client resulted in success, then the current contact is much more likely to be successful as well. This is a typical phenomenon in brand loyalty – if a customer has purchased once, they are more likely to purchase again than a new customer.

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

Another interesting point from the EDA is that the month has a huge impact on sales success. Certain months –have a much higher success rate for calls - March, September, October and December.

Moving on to the call-related data, we have already mentioned that longer call duration is correlated with higher success – but we will not include this variable in our models because it is only known after the call is complete. The previous variable, which indicates the number of prior contacts on the current campaign, has a higher mean for success responses. Otherwise, none of the call-related data were particularly interesting.

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

Many of the economic indicator variables are correlated with the response variable. The employee variation rate, consumer price index, Euribor 3-month rate, and number of employees (quarterly indicator) were all ***lower*** for success responses. This seems to indicate that when the economy was doing worse, the success rate was higher.

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

Looking at the economic indicators, we can see that in fact there are discrete values. These indicators are assessed quarterly and we are seeing values at different points in time. There are complex relationships between these variables. The curves for yes vs. no responses vary slightly, in particular for euribor3m vs. Cons.conf.id and the employee variation rate vs. Number of employed.

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

## Missing Values

The original dataset did not contain any missing values. However, several variables had values of “unknown” or “nonexistent”. The variable pdays (number of days since the previous campaign) was set to 999 if the lead was not previously contacted. When we filter observations where pdays is not 999, the boxplot below does show a significant difference between success and failure responses (but higher right skew).

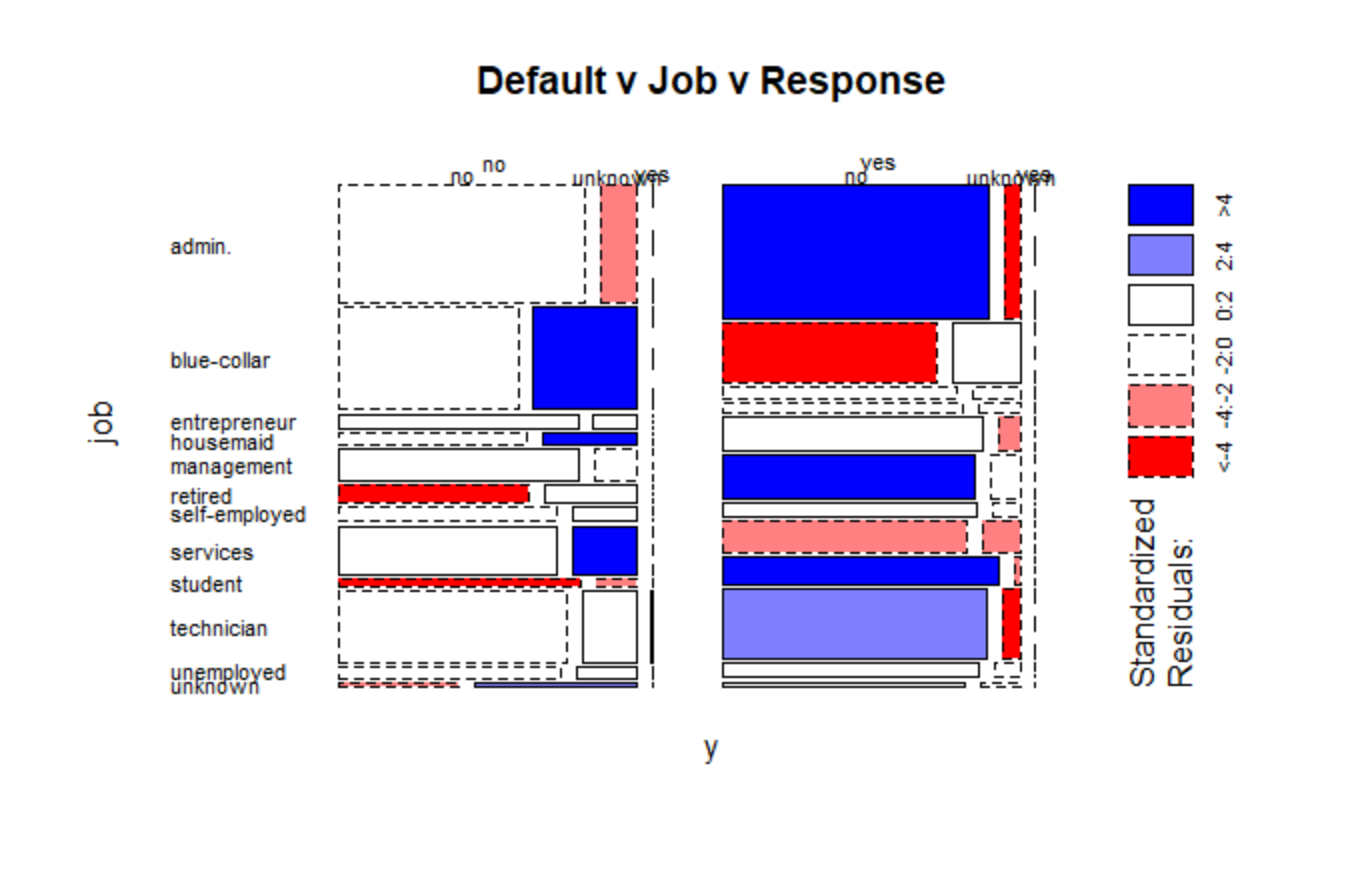
We have included all observations in our analysis since it is likely that future campaigns will encounter leads who have similar missing values. For example, it is difficult to know if a lead has previously defaulted on a loan, so if we removed all observations where default=“unknown” we would restrict our model to interpretation and prediction of results where the loan default history of a person is known. This is not a practical or a desirable constraint. The variable previous is another case where the majority of the observations are 999, which means there was no previous contact. If we transform the 999 to 7 so it will plot consistently with the non-999 values, we see that the number of yes values is more than 10 percent in all categories (1-6 prior calls). This tells us that contacting the person more than once will produce a greater than average yes response. It is also consistent with the practice of lead nurturing where leads are gradually marketed a new product or service, then contacted by the sales team once they have been contacted by marketing staff.

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

|  |
| --- |
|  |

## Mosaic Plots for Logistic Regression

We evaluated a number of combinations of categorical variables to see if any significant interactions existed. None of them showed anything conclusive. An example is given below:



# Objective 1 – Interpretable logistic regression model

For the object 1 model we decided to use the “Manual with EDA Model”

## Data Preparation

Our basic strategy for model creation is to create a training dataset for training the model, and a test dataset for testing it. We did this initially by splitting the data into separate groups of “yes” response rows and “no” response rows, and then splitting those groups 70/30 between train and test data sets. Then we combined the yes and no training data to get a training file “bank\_train\_90\_10”. And we combined the yes and no testing data sets to get a test file ”bank\_train\_90\_10”. Both of these files reflect the original distribution of the data (approximately 90% no and 10% yes).

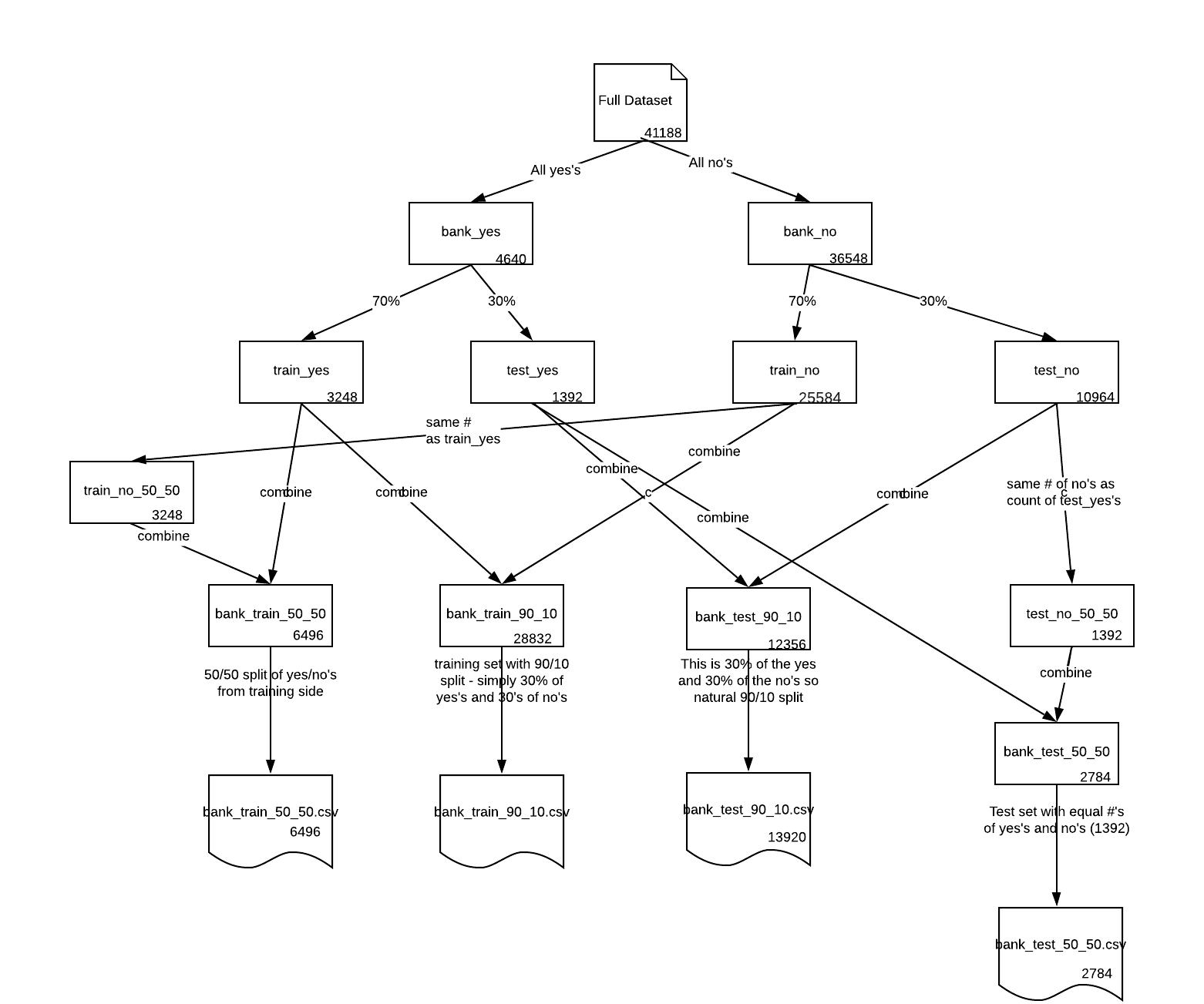
Giving that the “no’s” are overrepresented in the dataset, and we want the model to be equally tuned to yes and no’s we also tried creating a training dataset composed of equal numbers of yes’s and no’s. To do this we took all the yes’s from the training set and combined them with and equal number of “no’s” from the training set. This resulting dataset, bank\_train\_90\_10, has less overall “no’s” then the unbalanced bank\_train\_9010, and therefore less data overall. But it is balanced in terms of yes’s and no’s.

Our initial models were run on both bank\_test\_90\_10 and bank\_test\_50\_50. When we did not see significant differences in the test model results, we decided to perform all test set modeling using bank\_test\_90\_10.

The data sets used for training are testing are:

* Bank\_train\_50\_50.csv - taken from training sets, equal combinations of yes and no’s, 6496 rows.
* Bank\_train\_90\_10.csv - Follows the same distribution as the original full data set at approximately 10% yes’s.
* Bank\_test\_90\_10.csv - same distribution as the Bank\_train\_90\_10 – except for test data.
* Bank\_test\_50\_50.csv - this contains an even number of yes’s and no’s from the testing data set. We did not use this dataset for testing as the bank\_test\_90\_10 better shows the accuracy, and the sensitivity and specificity can be used to focus on yes’s and no’s.

## Dataset Creation Method



## Logistic Regression Model Creation

We created two manual models “Manual” and “EDA Manual”. We also used feature selection to create three additional models using forward, backward, and stepwise selection processes. Each model was created using the bank\_train\_50\_50 dataset, and the bank\_train\_90\_10 dataset.

We created the “Manual” model by first starting with all parameters and progressively removing parameters with high p-values. As we removed parameters, we assessed the Area Under the Curve (AUC) for the test dataset (bank\_test\_90\_10) to see the impact – eliminating variables that caused negative results. Once all the remaining parameters had significant p-values the model was complete.

We created the “EDA Manual” model by first starting with all the parameters that looked useful from the EDA: pdays, previous, emp\_var\_rate, cons\_price\_idx, cons\_conf\_idx, euribor3m, nr\_employed, Blue\_Collar, Retired, Student, MarriedUnknown (status), MarriedSingle, EducationIlliterate, EducationUniversity, EducationUnknown, DefaultNo, contact, HighMonth. In this case Blue\_Collar, Retired, Student, MarriedUnknown, MarriedSingle, EducationIlliterate, University, EducationUnknown, DefaultNo represent one-hot encoding of these particular factors. HighMonth has a value of “yes” for the month with higher yes prevelence.   
We ran logistic regression with these initial “EDA Manual” parameters and then remove those that had low p values ; previous, MaritalUnknown, EducationIlliterate, Blue\_Collar, Euribor3m, poutcome, Student, and then after checking the VIF and seeing colinearity, emp\_var\_rate. This simplified model minus variables with non-significant p values, and VIF eliminated, was kept.

# Model Selection

For the Objective 1 model the goal is interpretability which led us to choose from among the logistic regression models. The choices then are among the manual chosen model, the manual model based on the EDA, and then the models creating feature selection; forward, backward and stepwise. We also need to consider the models trained using the 90\_10 dataset vs. the 50/50 training set (even number of yes’s and no’s in the set).

There was not a huge difference in the performance of these models. They all scored close to 73 sensitivity and specificity when the values for those two were close to equal. We used the SAS output for finding these values and they were not always completely equal but we can see the equality point is generally around 73. The models training with the 90\_10 data had a higher overall accuracy than the 50\_50 models ( 75 vs 72.5 so so). This likely is due to having a much higher set of data for the no’s, to increase the accuracy of their prediction.

We chose the EDA based manual model for objective 1 because it is based on the results from the EDA, and it lends itself to interpretability. We created this EDA model by identifying columns in the dataset that were correlated with higher “yes” responses in the dataset. We then used SAS proc logistic to create a logistic regression model of these data points. We then removed any data columns that were not significant and did not overly decrease the model performance (test ROC AUC in particular).

This model includes varables for includes a combination of call related data (pdays, contact, HighMonth), demographic data (Retired, MaritalSingle, EducationUniversity, EducationUnknown, DefaultNo), and economic data (cons\_price\_idx, cons\_conf\_idx, nr\_employed). These variables all had noticeable effects in the EDA, and also have significant p values individually.

We chose to use the 50/50 train split model because the test AUC for the 50/50 split data should be more attuned to finding the “yes” responses. On the 90/10 train set, the threshold for even sensitivity/specificity is around 0.10 whereas for 50/50 it’s about 0.38. This difference in threshold location is the main difference – there was no a big difference in the accuracy, sensitivity, specificity of the models.

# Objective 1 Models

|  |  |
| --- | --- |
| Selection Algorithm | Model Found |
| Manual | Log(odds) = age Retired Entrepre HouseMaid contact month Monday campaign cons\_price\_idx euribor3m poutcome |
| EDA Manual | Log(odds)= pdays cons\_price\_idx cons\_conf\_idx nr\_employed Retired MaritalSingle EducationUniversity EducationUnknown DefaultNo contact HighMonth |
| Forward | Log(odds)= job default contact month day\_of\_week campaigh pdays emp\_var\_rate cons\_price\_idx euribor3m nr\_employed poutcome |
| Backward | Log(odds)= job default contact month day\_of\_week campaign pdays emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m poutcome |
| Stepwise | Log(odds)= job default contact month day\_of\_week campaign pdays emp\_var\_rate cons\_price\_idx cons\_conf\_idx euribor3m poutcome |
| Manual with DownSample | Log(odds) = age Retired Entrepre HouseMaid contact month Monday campaign cons\_price\_idx euribor3m poutcome |
| EDA Manual with Downsample | Log(odds) = pdays cons\_price\_idx cons\_conf\_idx nr\_employed Retired MaritalSingle EducationUniversity EducationUnknown DefaultNo contact HighMonth |
| Forward with DownSample | Log(odds)= default contact month day\_of\_week campaigh emp\_var\_rate cons\_price\_idx euribor3m nr\_employed poutcome |
| Backward with DownSample | Log(odds)= default contact month day\_of\_week campaign emp\_var\_rate cons\_price\_idx euribor3m poutcome |
| Stepwise with DownSample | Log(odds) = default contact month day\_of\_week campaign emp\_var\_rate cons\_price\_idx euribor3m poutcome |
| LASSO with DownSample | Log(odds) = job marital education housing contact month day\_of\_week campaign pdays poutcome emp.var.rate cons.price.idx cons.conf.idx nr.employed |

## Assumptions

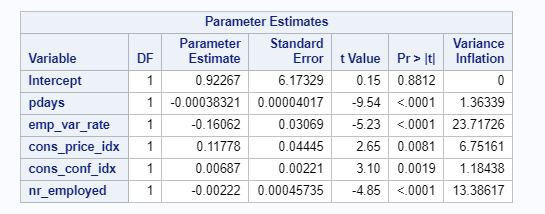
**Linear relationship**

Logistic regression requires a linear relationship between the log(response) variable and the predictive variables. We can see that this is the case based on the Global Null Hypothesis test results, whose p value gives an overwhelming probability that at least one of the variables in significant.   
We can also see this by looking at the box plots and pie charts showing that for some variables, there is a significant different in the distribution of “yes” vs. “no” in the response variable for the variables we have chosen for this model.

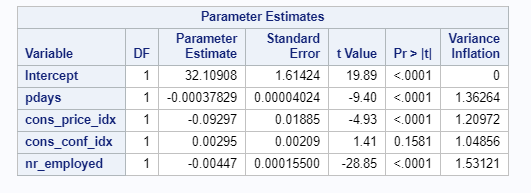
**Categorical response**  
Our response variable is binary – “yes” or “no” responses

**Independence of observations**  
We assume that our observations are independent. This is difficult to assess, e.g. some observations may have occurred sequentially by the same sales person. But overall with such a high number of observations, we believe this assumption is fulfilled.

**Multivariate normality**   
Many of our continuous variables exhibited right skew, however, our training datasets is a very large dataset. Based on the central limit theorem we expect this assumption to be satisfied by the large data set.

**Multicollinearity**  
One of the assumptions for a logistic regression model is lack of collinearity. We addressed this assumption by checking the VIF for the model. For the model including emp\_var\_rate the model shows a high VIF for emp\_var\_rate and also for nr\_employed:  


Therefore, we removed emp\_var\_rate from the model. After this change, the VIF of the model parameters was adequate



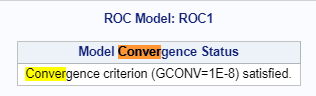
**Outliers**  
The CI Displacement vs. Leverage Plot (bottom left below) allows us to identify data points that have high CI displacement. We don’t see any significant values, all are well below 1.

**Residual plots**

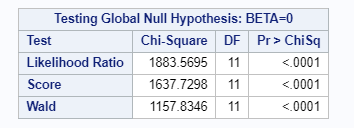
|  |  |
| --- | --- |
|  |  |
|  |  |

## Hypothesis Tests

**Model Convergence Status**

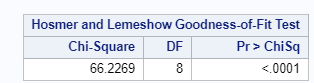
First, checking the output for this logistic regression model, we see that the maximum likelihood estimator has converged (so our results are likely to represent a successful fit):  


**Model Global Null Hypothesis**

We can also see that the Global Null Hypothesis is getting p values < 0.05 so we reject the null hypothesis that all of the predictors in the model are equal to zero. This suggests that at least one of the predictors in the model is significant. All 3 of the tests are giving the same result here with high significance levels.   


**Hosmer and Lemeshow Goodness-of-Fit Test**

The Hosmer and Lemeshow Goodness-of-Fit Test’s null hypothesis checks that the observed event rates (yes/no) match the expected event rates in subsets of the population. The p value is very significant indicating that the model is likely well calibrated.



## Maximum Likelihood Estimates

This table gives the estimates for the various coefficients in our logistic regression model, the standard error, the wal chi-square estimate, and the corresponding p value for the significance of the coefficient, including the intercept. All of the coefficients have a p value less than 0.05, and based on this we conclude they are all significant in predicting the response variable. The last column represents a chi-square test of the significance of each parameter in the model, and all the parameters are significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | Wald Chi-Square | Pr > ChiSq |
| (Intercept) | 30.2036 | 5.6193 | 28.8899 | < 0001 |
| pdays | -0.00145 | 0.000161 | 81.4902 | <.0001 |
| cons\_price\_idx | 0.1727 | 0.0684 | 6.3693 | 0.0116 |
| cons\_conf\_idx | 0.0180 | 0.00616 | 8.5167 | 0.0035 |
| nr\_employed | -0.00848 | 0.000503 | 284.3327 | <.0001 |
| Retired (yes) | 0.3613 | 0.0724 | 24.9217 | <.0001 |
| MaritalSingle (yes) | 0.0883 | 0.0319 | 7.6685 | 0.0056 |
| EducationUniversity (yes) | 0.0882 | 0.0315 | 7.8604 | 0.0051 |
| EducationUnknown (yes) | 0.1525 | 0.0740 | 4.2464 | 0.0393 |
| DefaultNo (yes) | 0.1626 | 0.0410 | 15.7142 | <.0001 |
| Contact (cellular) | 0.3457 | 0.0407 | 72.1677 | <.0001 |
| HighMonth (yes) | 0.4986 | 0.0673 | 54.9086 | <.0001 |

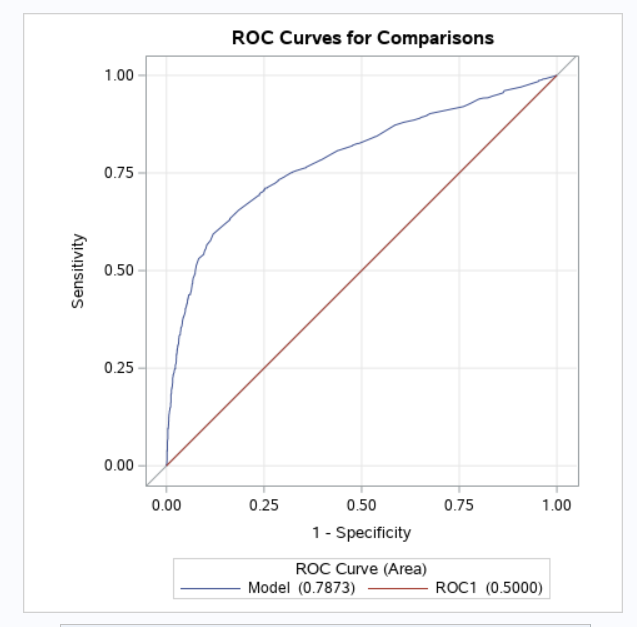
## Odds Ratio Estimates

|  |  |  |  |
| --- | --- | --- | --- |
| Effect | Point Estimate | 2.5 % | 97.5% |
| pdays | 0.999 | 0.998 | 0.999 |
| cons\_price\_idx | 1.189 | 1.039 | 1.359 |
| cons\_conf\_idx | 1.018 | 1.006 | 1.031 |
| nr\_employed | 0.992 | 0.991 | 0.993 |
| Retired yes vs. no | 2.058 | 1.550 | 2.732 |
| MaritalSingle yes vs. no | 1.193 | 1.052 | 1.351 |
| EducationUniversity yes vs. no | 1.193 | 1.055 | 1.400 |
| EducationUnknown yes vs. no | 1.357 | 1.015 | 1.815 |
| DefaultNo yes vs no | 1.386 | 1.179 | 1.625 |
| Contact cellular vs. telephone | 1.996 | 1.702 | 2.342 |
| HighMonth yes vs. no | 2.710 | 2.083 | 3.536 |

For each 1 unit increase in pdays, the odds of sales success decrease by a factor of 0.1%. For a 999 unit increase in pdays (unsuccessful or no prior contact vs. successful prior contact), the odds of sales success decrease by a factor of (1 - exp(-0.00145\*999))\*100% = (1-0.235)\*100% = 76.5%.

For each 1 unit increase in cons\_price\_idx the odds of sales success increase by a factor of 18.9%.  
For each 1 unit increase in cons\_conf\_idx the odds of sales success increase by a factor of 1.8%.  
For each 1 unit increase in nr\_employed, the odds of sales success decrease by a factor of 0.8%  
If the sales client is retired vs. not retired, the odds of sales success increase by a factor if 206%.  
If the sales client is single vs. married or divorced, the odds of sales success increase by a factor of 19.3%.  
If the sales client’s education is “University”, the odds of sales success increases by a factor of 19.3%.  
If the sales client’s education is “Unknown” (I.e. it is now known), the odds of sales success increases by a factor of 35.7%.  
If the sales client is in default, the odds of sales success increases by a factor of 38.6%.  
If the sales client contact is via cellular (vs. Telephone), the odds of success increase by a factor of 99.6%.  
If the HighMonth is true (it is December, March, October, or September) then the odds of success are increased by a factor of 271.0%.

The ROC curve for Objective 1 Model below describes the (Manual Model based on EDA). This ROC curve is for the testing dataset which contains 90/10 mix of data.



## Conclusion

The chosen model, based on the EDA and manually removing non-significant variables, shows some interesting features of the dataset. The demographic data tells a story – the sales results are more successful if the caller is Retired, Single, University education (or unknown), and they have not defaulted on a previous loan. Each of these variables was suggested by the EDA. If one is Retired they are much more likely to purchase or invest in a term deposit – 205% more likely to succeed!

The “contact” variable was also very important to this model as well as in the EDA– again almost doubling the odds of success. Is there a way to encourage calls to be over cellular vs. land lines? This data was captured between 2011 and 2013 when cell phones were not as commonplace. If we repeated this study today, the use of cellular would likely be even more prevalent given many people have stopped subscribing to land lines or “home phones”.

The economic variables were significant and show that as the economy changes, calls may be more or less successful. As the consumer-related indices increased, the odds of success increased. So success is apparently more successful in a good economy. It was strange that the nr\_employed variable increasing causes a decrease in success. Why could this be? It seems to contradict the index variable results? Below is a graph of consumer price index vs. number employed. It shows the relationship is not straight linear. We can conclude that during a better economy the sales success will be higher, but this may not be as practical to the sales team and their targeting, unless we want to reduce the size of the sales team during economic downturns.  


The final interesting point is that sale success is 270% higher odds during certain months - March, September October and December. This does not make a lot of sense. For example, if this is a seasonal effect, why would November be left out? And why March? The effect is so significant that we cannot discount it. Perhaps there are certain events in these months that are outside of the economic variables (e.g. bonuses paid then?). This is worth more investigation. As mentioned earlier, the data we have here is a subset of the data from a larger study, covering multiple years. Could the data have been incorrectly sampled from the original larger dataset so that these months are from different years, for example? This is an area for further investigation. In any case, the economic data is not useful for selecting individual sales leads to pursue.

We did not include any interaction terms in the object 1 interpretive model, nor was there any basis in the EDA for doing so.

At a threshold of 0.40, the chosen EDA based manual model produces an accuracy of 72.6, sensitivity of 73.2 and specificity of 71.2. We could have chosen a different threshold for the model in order to get a different sensitivity and specificity but we preferred to take a balanced approach to manage cost of sales and opportunity to improve sales.

# Objective 2 – Predictive model

A number of more complex models were developed and tested with the objective of gaining better predictions at the cost of model interpretation. The criteria used to assess prediction performance were AUC for the training and test sets, and the Balanced Accuracy/Sensitivity/Specificity for the test set. The same test set (test\_90\_10) was used for all models. Training set AIC could be used as a tie-breaker if needed for the logistic regression models.

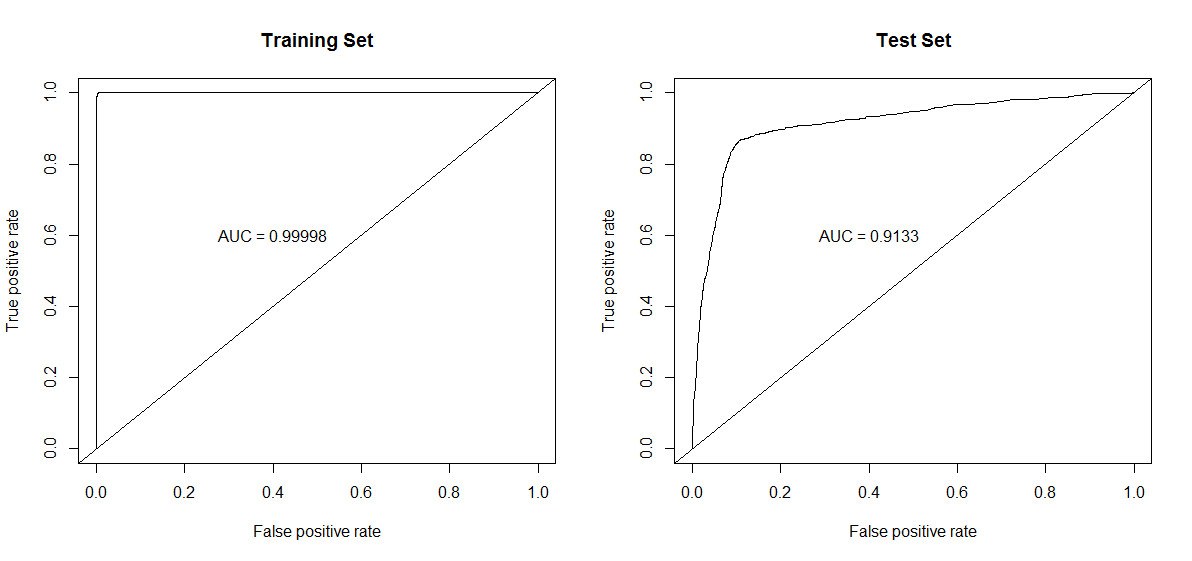
Several variable selection methods were used with logistic regression, none of which produced a meaningful improvement in prediction even when interaction terms were included (LASSO). It’s possible that there are no meaningful 2-way interactions for this data, the interactions are highly confounded, or the selection methods may have too high of bias.

Linear Discriminant Analysis (LDA) produced a comparable model to Objective 1 with the numeric variables (except for duration, previous, and emp.var.rate) and all of their 2-way interactions. One of the numeric variables used, pdays, was not really continuous and likely violated the model assumptions, but it improved the prediction which was the goal. Quadratic Discriminant Analysis (QDA) did not improve the model.

There was no practical difference in the models or their prediction performance when training on the even data split (train\_90\_10) vs. the down-sampled data (train\_50\_50). This is likely because of the relatively large data set (n=41,188) that didn’t have too extreme of a yes/no split (90/10). Down-sampling is likely more critical with smaller data sets and/or more extreme category proportions.

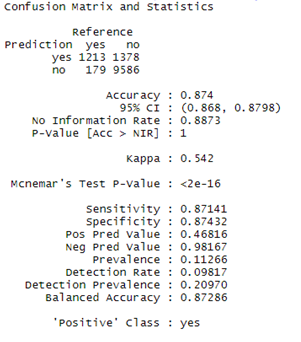
The only higher complexity model that did significantly better was random forest. The maximum performance was achieved with mtry=9 and ntree=1000. This is definitely overfitting the down-sampled training set (AUC=0.99998) and is likely “over-fitting” the test set as well (AUC=0.9136) which could lead to poor prediction of any new data. That said, mtry and ntree can be dialed way down and random forest still easily outperforms the other models for test set AUC: mtry=1 and ntree>100, mtry=2 and ntree>3, mtry=3 and ntree>2, etc. Random forest is clearly the best choice here to improve model prediction with a “black box” that loses model interpretation.

## Random Forest ROC Curves (mtry=9, ntree=1000)



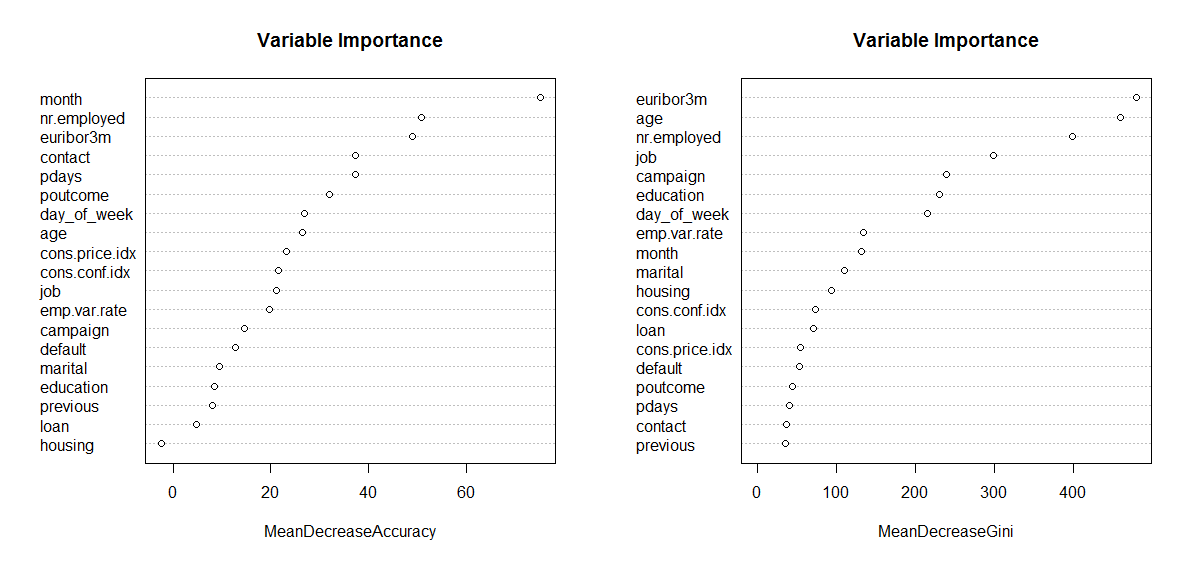
Random Forest Balanced Confusion Matrix

(mtry=9, ntree=1000, prediction threshold=0.42)



## Model Description for Random Forest (mtry=9, ntree=1000)

Several variables in the random forest are correlated. Random forest can deal with those correlations, but it makes interpretation of Variable Importance difficult.



# Objective 2 Conclusion

In conclusion, our random forest model provided the best predictive models with reduced model interpretation. Logistic regression with variable selection tools was not successful in finding interaction terms. LDA and QDA with numeric interaction terms were comparable to logistic regression models. Down-sampling did not improve model predictions.

# Compare All Models

Objective 1 final model = Manual EDA, Objective 2 final model = Random Forest

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Train Set | Selection Method | AIC | Train AUC | Test AUC | Balanced  Threshold (Test) | Accuracy (Test) | Spec. (TN, Test) | Sens. (TP, Test) |
| 90/10  Y/N | Manual\* | 15,952 | 0.7902 | 0.7827 | 0.08 | 73.0 | 70.5 | 73.3 |
| **Manual EDA** | **16,423** | **0.7799** | **0.7755** | **0.08** | **71.1** | **72.1** | **71.0** |
| Forward | 17,157 | 0.7954 | 0.7902 | 0.08 | 75.3 | 70.1 | 76.0 |
| Backward | 15,922 | 0.7956 | 0.7903 | 0.08 | 75.6 | 70.0 | 76.3 |
| Stepwise | 17,157 | 0.7956 | 0.7903 | 0.08 | 75.6 | 70.0 | 76.3 |
| LASSO | 15,927 | 0.7959 | 0.7890 | 0.075 | 71.3 | 71.3 | 71.3 |
| LDA | N/A | 0.7883 | 0.7837 | 0.025 | 71.5 | 71.5 | 71.5 |
| 50/50  Y/N | Manual | 6,930 | 0.7987 | 0.7830 | 0.40 | 72.7 | 72.6 | 72.0 |
| **Manual EDA** | **7,065** | **0.7873** | **0.7781** | **0.40** | **72.6** | **73.2** | **71.2** |
| Forward | 7,478 | 0.7995 | 0.7884 | 0.38 | 72.5 | 73.2 | 71.9 |
| Backward | 6,951 | 0.7955 | 0.7883 | 0.38 | 72.5 | 73.1 | 71.9 |
| Stepwise | 7,478 | 0.7955 | 0.7883 | 0.38 | 72.5 | 73.1 | 71.9 |
| LASSO | 6,955 | 0.8016 | 0.7894 | 0.39 | 71.8 | 71.8 | 71.8 |
| LDA | N/A | 0.7978 | 0.7849 | 0.34 | 72.3 | 72.3 | 72.3 |
| **Random Forest\*\*** | **N/A** | **.99998** | **0.9133** | **0.42** | **87.4** | **87.4** | **87.1** |

\*Manual = manual backward selection starting with all variables. Manual EDA is manual backward selection starting with variables identified from EDA.

\*\*mtry=9, ntree=1000

# Final Conclusions

Objective 1 demonstrated our ability to use EDA and derive a model that could be understood, although some of the relationships were surprising. We did not expect retirees and administrative roles to be such a significant portion of the population who would choose to invest in a term deposit.

Objective 2 showed us that a predictive model can be much more accurate in predicting future values, but we cannot explain exactly which interactions it was able to detect, or what trees and internal mechanisms allow it to be so high-performing. We are also unable to verify the statistical probability of each component within the model. However, using minimal code and tuning a few parameters (mtry, ntree) we were able to quickly develop an alternative approach that is more effective for predictions.

In both models we tried to eliminate factors that were not practically significant, and our EDA helped us understand where collinearity may exist – it just was not a problem for the “black box” predictive approach.

With this two-fold approach (two objectives) we have demonstrated that our audience or clients are best served by having a variety of models – one that reveals the relationships between different market factors, and another to support forecasting of future business results.

# References

1. <https://archive.ics.uci.edu/ml/datasets/Bank%20Marketing>Bank Marketing Data Set
2. Initial Article Studying this Data Set  
   <http://media.salford-systems.com/video/tutorial/2015/targeted_marketing.pdf>[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
3. <https://www.global-rates.com/interest-rates/euribor/euribor-interest-3-months.aspx> The **3 month Euribor interest rate** is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of **3 months**. Alongside the 3-month Euribor interest rate we have another 14 Euribor interest rates with different maturities (see the links at the bottom of this page). The Euribor interest rates are the most important European interbank interest rates. When the Euribor interest rates rise or fall (substantially) there is a high likelihood that the interest rates on banking products such as mortgages, savings accounts and loans will also be adjusted.

# Appendix: Code for EDA, model development and comparison

All code and data files can be found at:

<https://github.com/aleppla/SMU-Stats2-Project2>