**Project #3 for MSDS 6372-4023**

**Logistic Regression:**

**Direct Marketing Campaigns of a Portuguese**

**Banking Institution**

**Submitted by:**

Lee Mooyoung

Thomas Wang

Timothy McWilliams

**{ mooyoungl, keyuew, tmcwilliams}@smu.edu**

**August 15, 2017**

**Table of Contents**

[**Problem Description** 2](#_Toc490594025)

[**Dataset** 2](#_Toc490594026)

[**Exploratory Data Analysis** 3](#_Toc490594027)

[**Logistic Regression: Assumptions and Analysis** 6](#_Toc490594029)

[**Analysis Procedure and Interpretation of Results** 7](#_Toc490594030)

[**Conclusion** 9](#_Toc490594031)

[**References** 10](#_Toc490594032)

[**Appendix** 10](#_Toc490594033)

[**SAS Code** 21](#_Toc490594034)

## 

## **Problem Description**

There are two main approaches for enterprises to promote products: through mass campaigns and through directed marketing. Direct marketing is where a campaign targets a specific set of contacts. “In a global competitive world, positive responses to mass campaigns are typically very low, less than 1%” (Ling and Li 1998). Data were collected from a Portuguese bank direct marketing campaign. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

The goal of this analysis is to find a model using logistic regression techniques to help better predict the success of a contact, if the client subscribes the deposit or not. Logistic regression will be used to help understand how certain explanatory variables may influence the likelihood of an event occurring and can those explanatory variables be used to predict if an event will occur or not. Such model can increase campaign efficiency by identifying the main characteristics that affect if a contact subscribed or not. This will help management make better use of the available resources and selection of a high quality and affordable set of potential buying customers.

## **Dataset**

This dataset contains data from a Portuguese Banking Institution on marketing campaigns having to do with bank deposit descriptions. The dataset is broken up into three subsets; bank client data, data related with the last contact of the current campaign, and other related data. The attributes are as follows: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, and y. The attribute y holds whether the contact subscribed or not. This project will focus of these attributes to build a model that best predicts whether a caller subscribes or not (Yes/No). For a more detailed description of all the attributes used in this analysis see section I in the appendix.

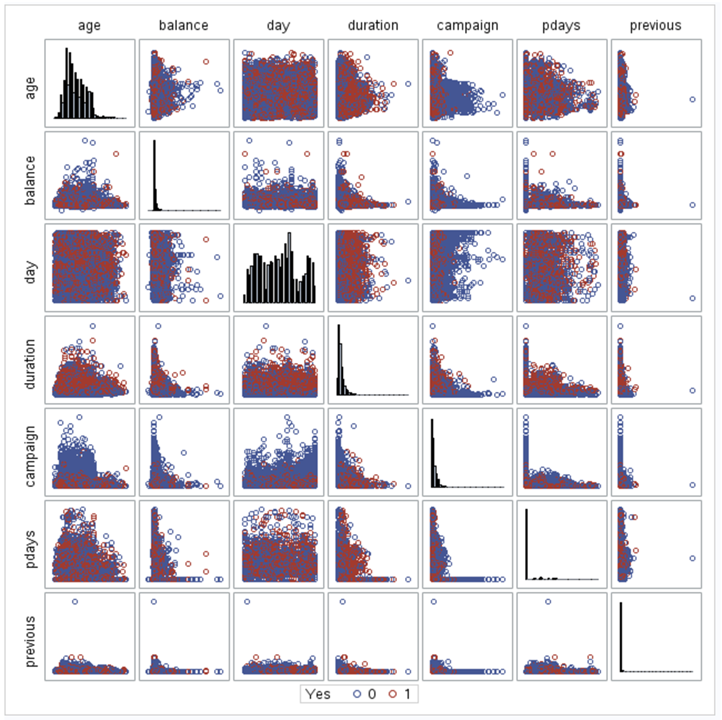
**Concerns and Limitations**

This study is an observational study, so the reader must be cautioned against generalizing the results to the larger population. Since these data were collected from a Portuguese Banking Institution any inferences that are made can only apply to the area in which the banking institution conducts business. Also, it is not known how many different persons were used to call the contacts and attempt to persuade that contact to subscribe. This could add some bias into the data since one person and be better than another person in pitching a subscription to the contact. Lastly, we must recognize that there could be additional important predictors of subscription that are not included in our explanatory variables.

## **Exploratory Data Analysis**

The data file, csv, is opened using Excel and converted from string to a table in order to make the data read process from SAS easy. There was no missing or not-a-number value in the data set. The sample size was initially 54,211 with one outlier being removed later. The size of sample was more than enough to produce a reliable outcome.

Figure 1 displays a matrix scatter plot of the continuous variables for this analysis. This figure shows that there is an outlier present in the data. The variable previous has an outlier at observation 275. This variable holds information on the number of contacts performed before this campaign and for this client. The outlier was removed from the dataset for this analysis since the variable is significant to the final model.

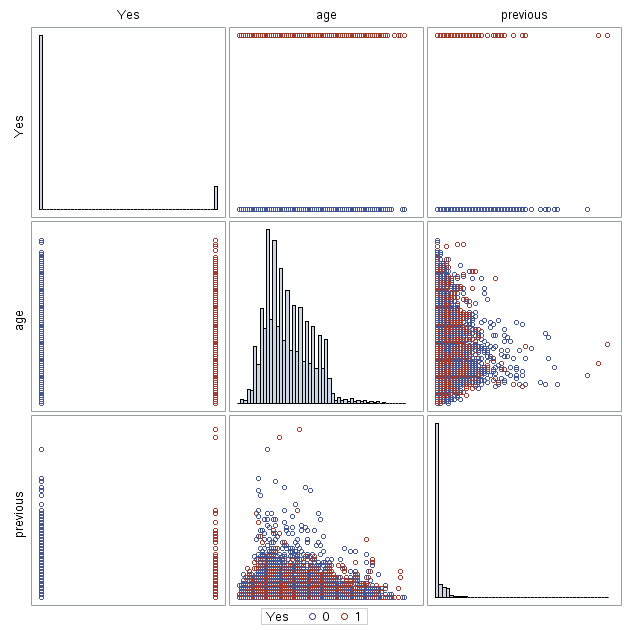
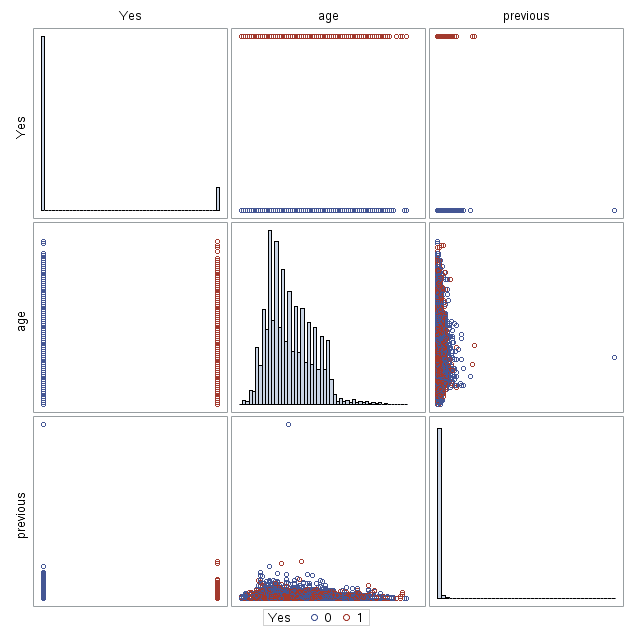


## 

## 

***Figure 1. Matrix scatterplot of continuous variables***

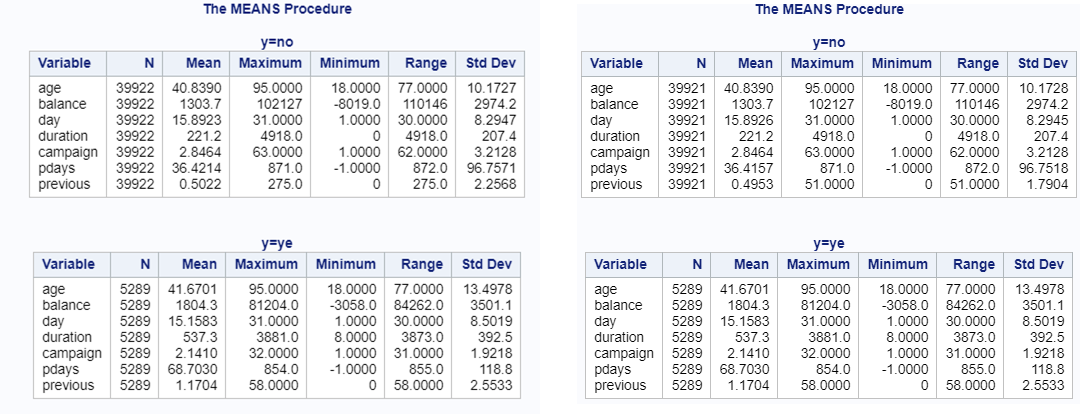
Figure 2 provides a visualization of the difference in scatter before and after removing the outlier. As expected, the scatter distribution is much more appropriate once the outlier is removed. We will continue our analysis without the outlier.



***Figure 2. With outlier(left) vs without outlier(right)***

Figure 3 shows our summary statistics split into four categories (yes with outlier, no with outlier, yes without outlier, no without outlier). The outlier data point resulted in a “no” response, therefore we can expect the “yes” summary statistics to be identical despite the omission of the outlier. The largest difference we see once the outlier is removed is the “maximum” statistic for the “previous” variable. There is a significant change from 275 to 51, which explains the differences in the other statistics as well. 275 previous phone calls to a potential client does seem a bit out of the ordinary so it may have just been bad recorded data point. Nevertheless, this is not a concern as we are not including the outlier in our study.

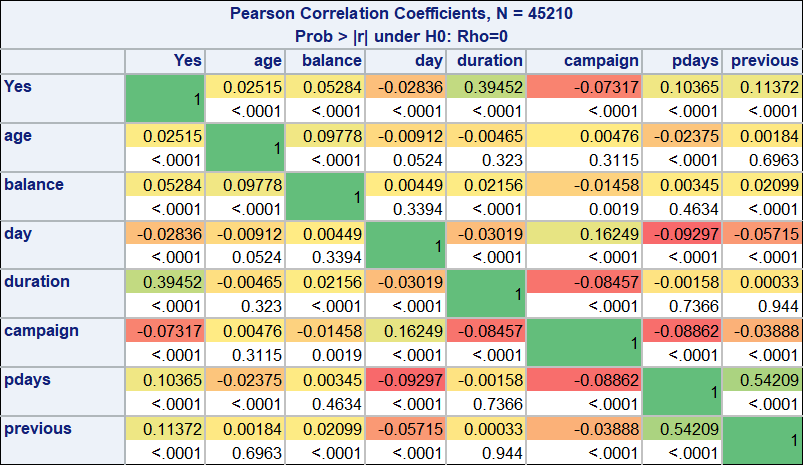
***Figure 3. Summary statistics with outlier(left) vs without outlier(right)***

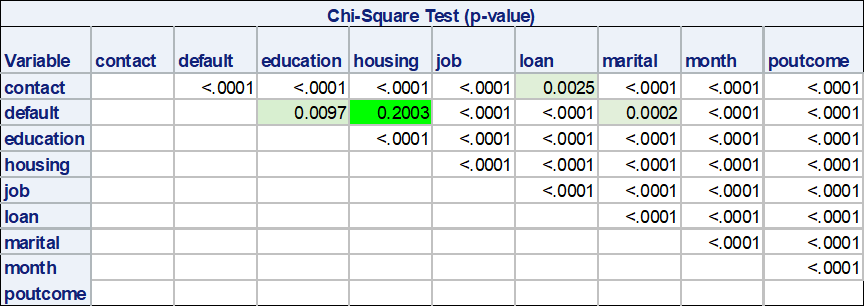


**Association Check**

Table 1 portrays a table of Pearson Correlation Coefficients for all continuous variables. Here we can see that the strongest point of correlation is between “previous” and “pdays”. This correlation is expected as both variables depend on previous contacts of a client in terms of number of times and days. Variables “yes” and “duration” also exhibit slight correlation as we can presume that a longer conversation is more likely to result in a subscription.

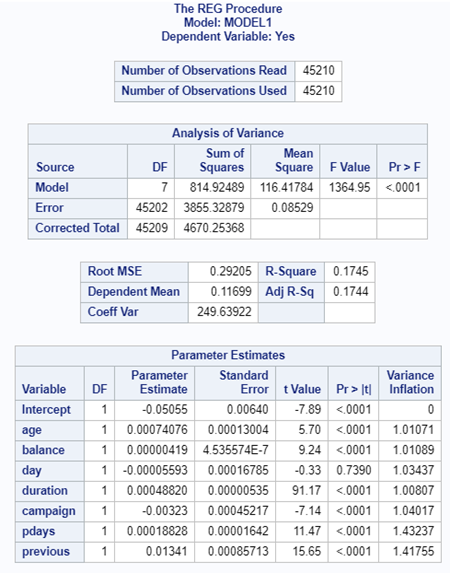
***Table 1. Correlation Coefficients between continuous variables***



Table 2 presents a Chi-Square Test table of the associations between our categorical variables. In theory, we would prefer to see insignificant p-values across the board. However, it appears that only “default” and “housing” show no strong association. This is a result of a large dataset (sample size-54,210) that makes a test significant with any small differences. We presume that there is no practical significance, and we will pay attention to the association in the categorical variables throughout the analysis.

***Table 2. Chi-Square Test between categorical variables***

Table 3 shows the variance inflation factors for continuous variables. While the VIF is not especially large, we can assume that “pdays” and “previous” are inflating each other from our previous Pearson Correlation Coefficients table.

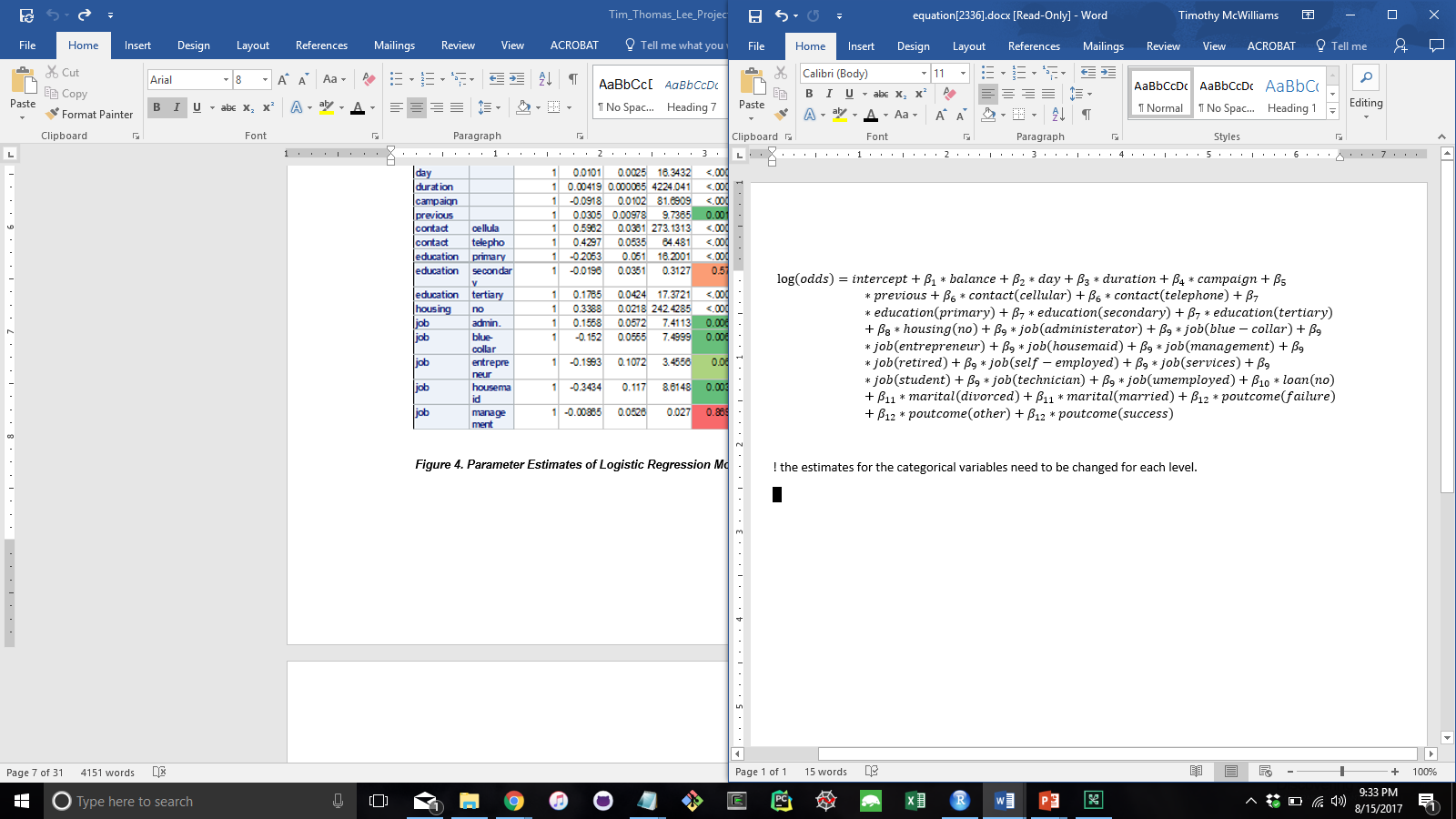


***Table 3. Variance Inflation factors for continuous variables***

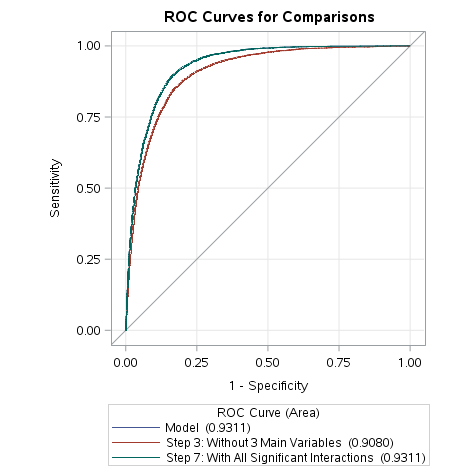
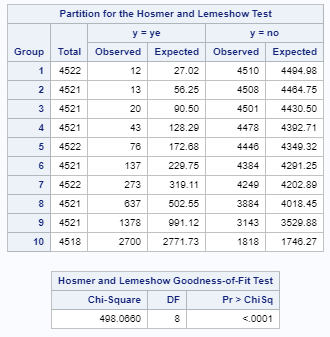
For the final logistic regression model, “age”, “pdays”, and “defaults” variables are removed due to insignificant p-values (0.971, 0.8198, 0.9169 accordingly) in the Chi-Squared table (Appendix II). The odds ratios for the removed three variables were near zero and the confidence intervals were crossing 1, which reflects the insignificant p-values in the chi-square test mentioned above (Appendix II). The “balance” variable is not removed, even though the odds ratio is very close to 1. This is due to the significant p-value (0.01), and the area under the Receiver Operating Characteristic (ROC) curve dropping slightly without the “balance” variable.

## **Logistic Regression: Assumptions and Analysis**

## 



***Figure 4. Parameter Estimates of Logistic Regression Model with “pdays”, “defaults”, and “age” variables removed***

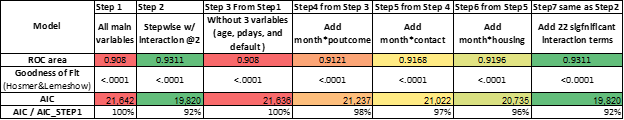


***Figure 5. ROC Curve Table 4. Goodness-of-Fit Test***

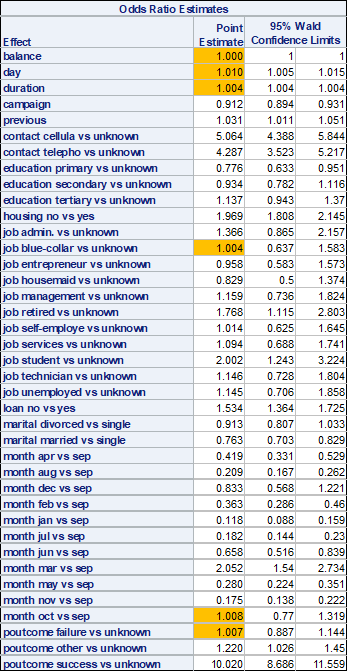
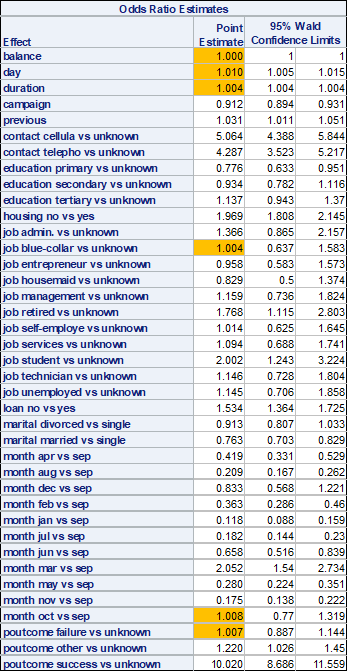
Figure 4 shows the parameter estimates of Logistic Regression Model that was chosen for this analysis. The Hosmer and Lemeshow Test in table 4 shows the lack of fit that this model has (p-value = <0.0001). An important note here is that this could be a direct result of the dataset being too large. The practically insignificant differences can be detected with a large data set. The influence diagnostic plots are not provided due to a lack of memory in SAS to display the plots. Thus, we will continue with our analysis. Based on the area under the ROC curve being above 0.9, the model classifies the binary response pretty well (Figure 8). There are several categorical variable levels that are not significant in this model, which are highlighted red in Figure 7.

## **Analysis Procedure and Interpretation of Results**

The model from Step 3 has the minimum number of variables without any interaction terms. The model from Step 2 or 7 is a more complicated model that includes many interaction terms. This may cause a more complex and difficult interpretation of the model. In fact, Step 7 includes all 22 interaction terms found from the stepwise selection process in Step 2. With 22 additional interaction terms, the area under the ROC curve changed from 0.908 to 0.931, while lowering the AIC by 8%. However, for the purpose of this analysis, the model from Step 3 was chosen in terms of simplicity.



***Table 5. Table displaying models for this analysis.***

 ***Figure 6. Odds Ratio Estimates (including removed variables)***

Based on the parameter estimates (Figure 4) and the odds ratio (Figure 6), the most influential inputs to accept the bank term deposit are previously accepted term deposit(poutcome), contact communication type(contact), student/retired person(job), a particular month of a year(month), and ownership of a house(housing).

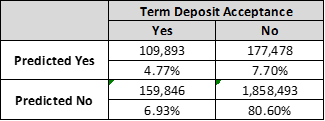
The odds of accepting a term deposit by those who previously accepted a term deposit is 10.0 times higher than the unknown group; this includes a 95% confidence interval of (8.68, 11.5), assuming all other factors are fixed. Out of the four groups of people who were contacted previously (unknown, other, failure, success), the bank should strategically focus their attention on those who have previously experienced their term deposit product.

The bank can get 5.0 times higher odds of opening a new term deposit account, with a 95% confidence interval of (4.38, 5.84), if they call the customer’s cellular phone as opposed to the unknown method, assuming all other factors are fixed. Similarly, the odds will increase by 4.2 times when calling to a customer’s telephone over unknown method. Thus, we can conclude that contacting customers by a phone call is more important than the unknown contacting method. The bank should review what the unknown method is and avoid that form of communication in favor of the more effective method.

Job wise, the student group and retired group show higher odds to accept the term deposit by 2.0 and 1.7 times, respectively, compared to those of unknown job status (assuming all other factors are fixed). To maximize the odds, we recommend the bank prioritize the job groups in the following order: student, retired, admin, management, technician, unemployed, services, self-employed, and lastly, blue-collar workers.

The odds of opening a new term account during the month of March is higher than any other months. It is more than twice the odds of opening an account in September, assuming all other factors are fixed. Those who do not own a house also increase the odds by 1.9 times than those who own a house, with a 95% confidence interval of (1.8, 2.1), assuming all other effects are fixed.

“Campaign” is the most significant continuous variable that can be plotted over an effect plot (Appendix VII). “Campaign” represent the number of contacts during a campaign period. The odds of opening an account change by 0.91 times as one additional call made to a same customer.

The classification accuracy of the model is 85%. See below table 6, summary from the original classification table (Appendix IX).

***Table 6. Classification Summary***

## **Conclusion**

When promoting their term deposit product, the bank may want to focus their attention on specific groups and methods in order to maximize their return. Based on our logistic regression analysis, we recommend contacting a student or retired person who already had a term deposit account with the bank, and also does not own a house. To further increase their odds, we advise the contact be made by a phone call during the month of March. All these factors, integrated together, should provide the bank with the best plan of action when targeting potential clients. The logistic regression model produced in this analysis has a classification accuracy of 85%.

## **References**

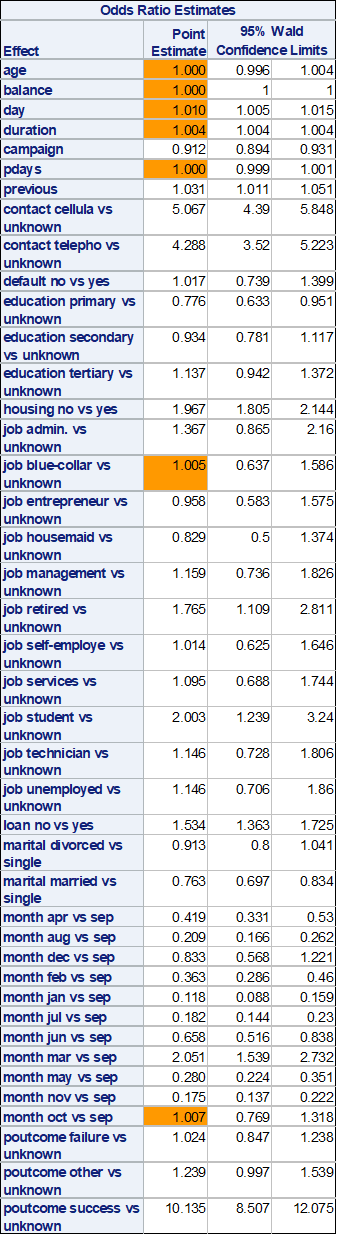
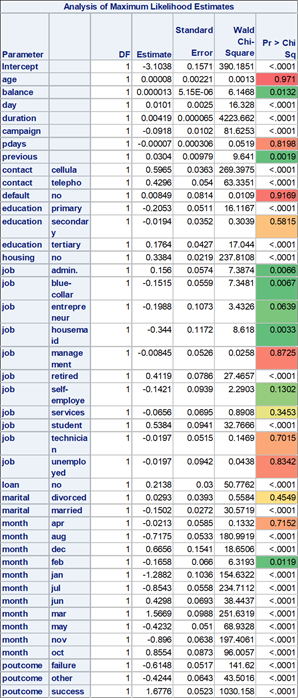
Ling and Li 1998

## **Appendix**

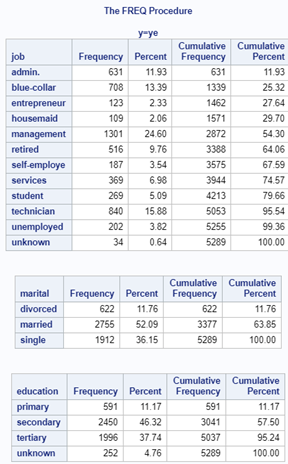
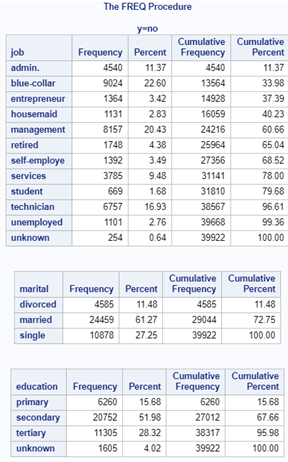
**I. Dataset**

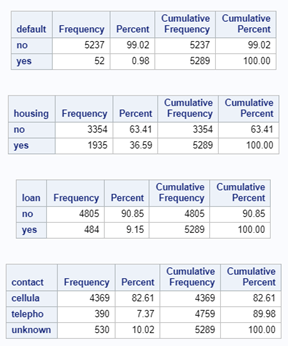
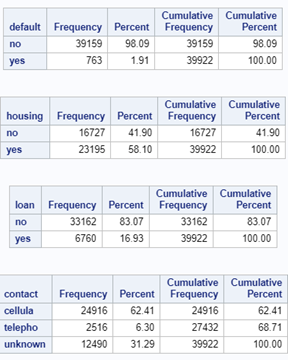
|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Data Type** |
| Age | The age of the caller | Continuous |
| Job | Type of job | Categorical |
| Marital | Marital status | Categorical |
| Education | Level of education (Unknown/Primary/Secondary/Tertiary) | Categorical |
| Default | Has credit in default (Yes/No) | Categorical |
| Balance | Average yearly balance, in euros | Continuous |
| Housing | Has housing loan? (Yes/No) | Categorical |
| Loan | Has personal loan? (Yes/No) | Categorical |
| Contact | Contact communication type (Unknown/Telephone/Cellular) | Categorical |
| Day | Last contact day of the month | Continuous |
| Month | Last contact month of the year | Continuous |
| Duration | Last contact duration | Continuous |
| Campaign | Number of contacts performed during this campaign and for this client (Includes last contact) | Continuous |
| pdays | Number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted) | Continuous |
| previous | Number of contacts performed before this campaign and for this client | Continuous |
| poutcome | Outcome of the previous marketing campaign (Unknown/Other/Failure/Success) | Continuous |
| y | Has the client subscribed a term deposit? (Yes/No) | Categorical |

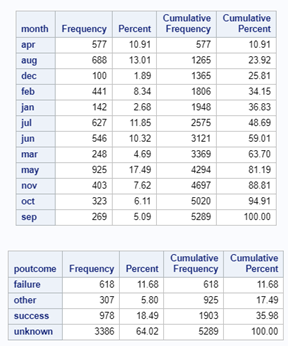
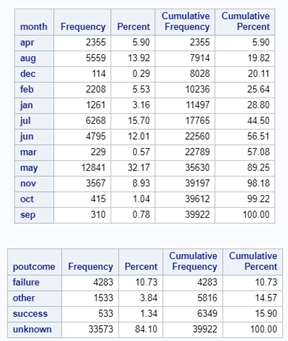
**II. Parameter Estimates and Odd Ratio Estimates**

****

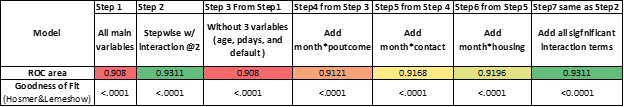
**III. Frequency Table**



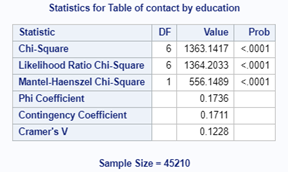
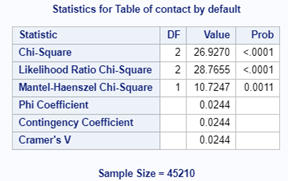


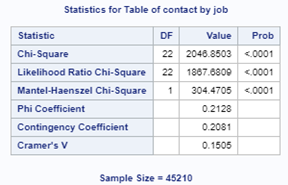
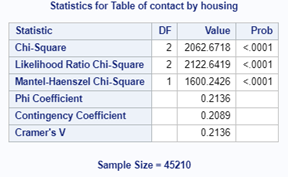


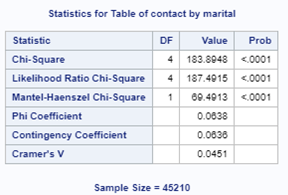
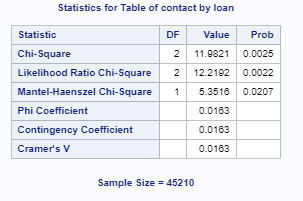
**IV. Model Selection Methods**

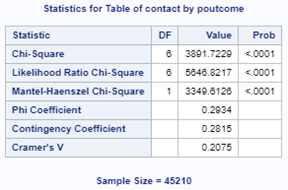
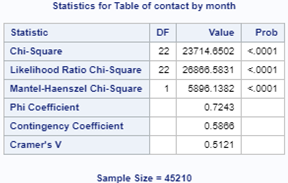


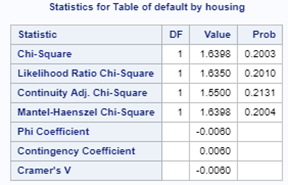
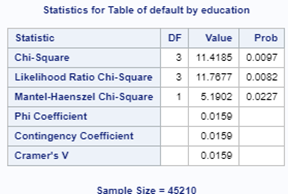
**V. Paired Chi-Square Test Results**

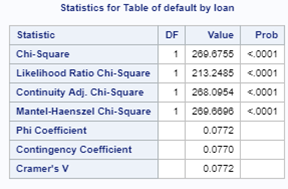
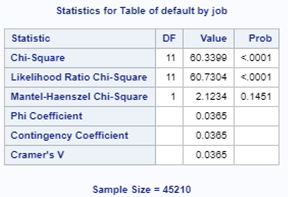


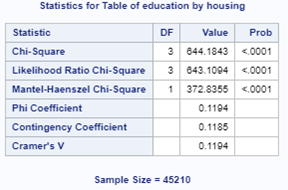
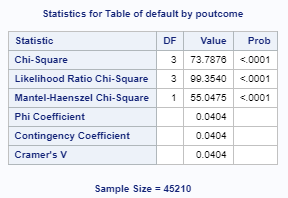


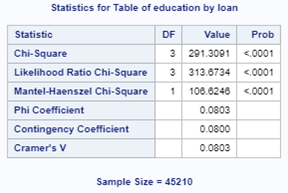
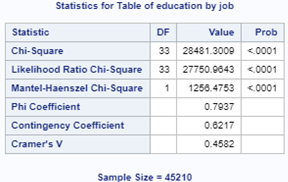


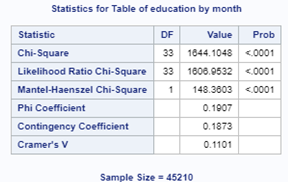
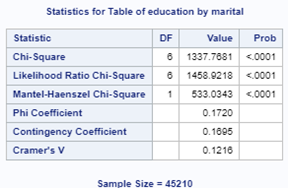


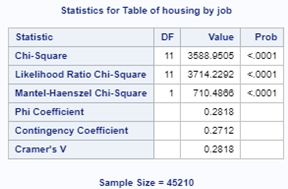
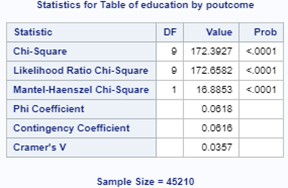


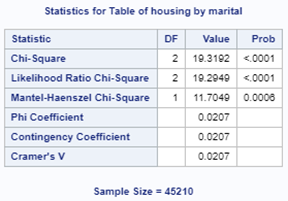
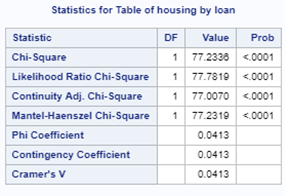


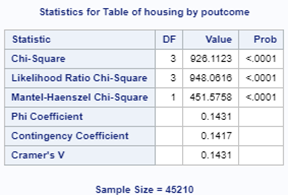
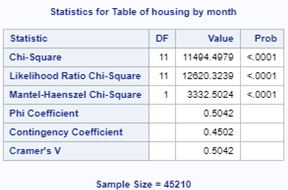


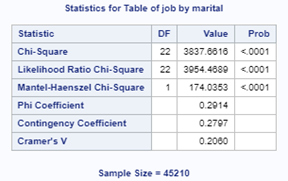
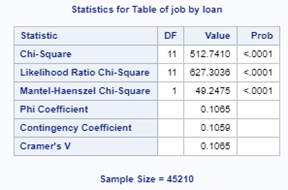


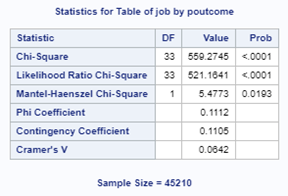
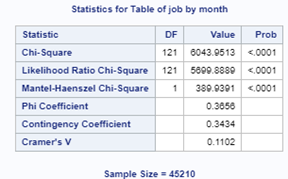


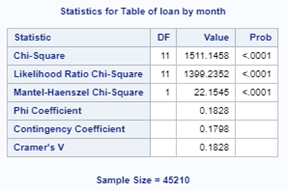
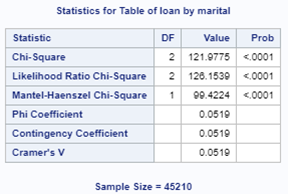


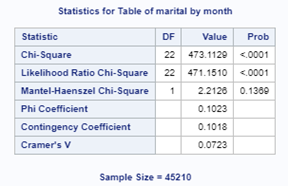
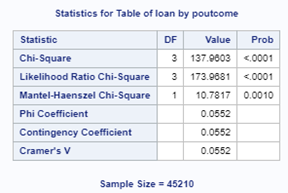


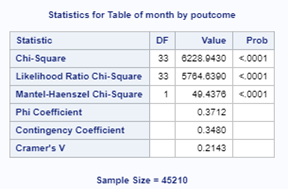
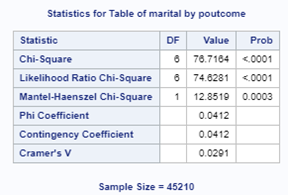




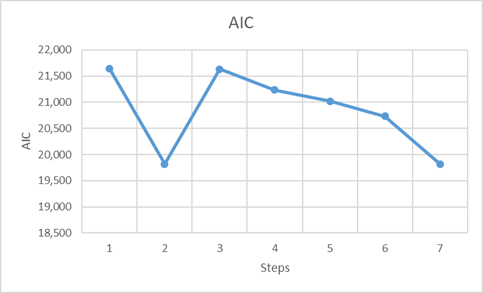




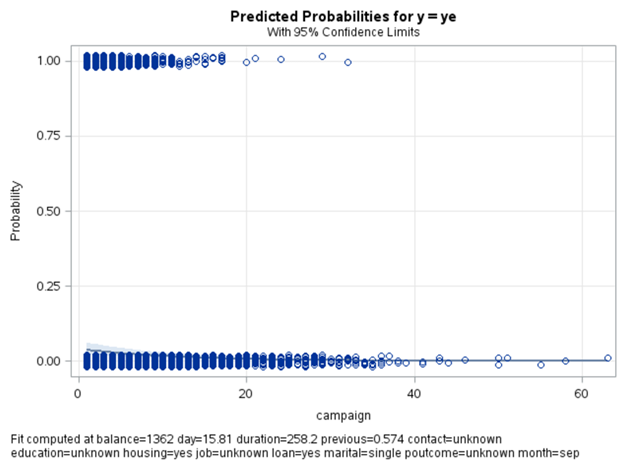




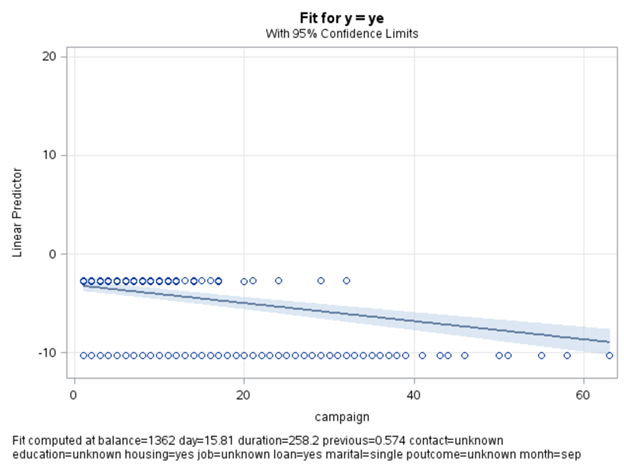
**VI. Model Steps vs AIC**



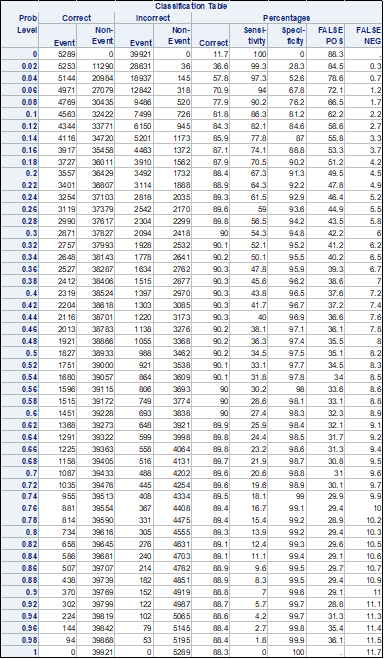
**VII. Effect Plot Probability vs. Number of Contacts During a Campaign**



**VIII. Effect Plot Linear Predictor vs. Number of Contacts During a Campaign**



**IX. Classification Table**



## **SAS Code**

|  |
| --- |
| /\* Project #3: Logistic Regression using a bank product data \*/  /\* Date: 8/15/2017 \*/  FILENAME REFFILE '/home/mooyoungl0/MSDS 6371 STAT1/bank-full2.csv';  PROC IMPORT DATAFILE=REFFILE  DBMS=CSV  OUT=bank;  GETNAMES=YES;  RUN;  /\* Data Transformation \*/    /\* !! Add a colume transforming 'y' into numerical \*/  data bank2;  set bank;  if y = 'ye' then Yes = 1;  else Yes = 0;  run;  /\* !! Remove one outlier from 'previous' \*/  data bank3;  set bank2;  if previous = 275 then delete;  run;    /\* !! Add ID variable \*/  data bank4;  set bank3;  ident = \_n\_;  run;  /\* Convert Month \*/  data bank10;  set bank3;  if month = "jan" then monthNum = 1;  if month = "feb" then monthNum = 2;  if month = "mar" then monthNum = 3;  if month = "apr" then monthNum = 4;  if month = "may" then monthNum = 5;  if month = "jun" then monthNum = 6;  if month = "jul" then monthNum = 7;  if month = "aug" then monthNum = 8;  if month = "sep" then monthNum = 9;  if month = "oct" then monthNum = 10;  if month = "nov" then monthNum = 11;  if month = "dec" then monthNum = 12;  run;    /\* Data exploration \*/  PROC CONTENTS DATA=bank; RUN;  /\* summary stat for numerical variables\*/    /\* with outlier \*/  proc sort data = bank;  by y; run;  proc means data = bank n mean max min range std fw=8;  var \_numeric\_ ;  output out = meansout mean = mean std = std;  by y;  title 'Summary Stat';  run;    /\* without outlier \*/  data bank5;  set bank3;  drop Yes;  run;  proc sort data = bank5;  by y; run;  proc means data = bank5 n mean max min range std fw=8;  var \_numeric\_ ;  output out = meansout mean = mean std = std;  by y;  title 'Summary Stat';  run;  /\* summary stat for categorical variables for BEFORE Outlier\*/  data bank6;  set bank;  drop age balance day duration campaign pdays previous;  run;    proc freq data = bank6 ;  by y;  run;  /\* Scatter Plots \*/    /\* Scatter Plot (!Takes long time to run)\*/  proc sgscatter data=bank2 ;  matrix Yes age balance day duration campaign pdays previous  / diagonal=(histogram) ;  run;    /\* Scatter by Yes (!Takes long time to run)\*/  proc sgscatter data = bank2;  matrix age balance day duration campaign pdays previous/ diagonal=(histogram) group = Yes;  run;quit;        /\* Scatter Plot BEFORE Removing ONE Outlier\*/  proc sgscatter data=bank2 ;  matrix Yes age previous  / diagonal=(histogram) group = Yes;  run;    /\* Scatter Plot AFTER Removing ONE Outlier\*/  proc sgscatter data=bank3 ;  matrix Yes age previous  / diagonal=(histogram) group = Yes;  run;  /\* Association among variable check \*/  proc reg data = bank3 plots= all;  model Yes = age balance day duration campaign pdays previous / vif partial;  run;quit;    proc corr data = bank3 plots = all;  var Yes age balance day duration campaign pdays previous;  run;    proc freq data = bank3;  table contact default education housing job loan marital month poutcome / chisq ;  run;quit;    proc freq data = bank3;  table contact\*default/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*education/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*housing/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*job/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*loan/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*month/ chisq ;  run;quit;  proc freq data = bank3;  table contact\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table default\*education/ chisq ;  run;quit;  proc freq data = bank3;  table default\*housing/ chisq ;  run;quit;  proc freq data = bank3;  table default\*job/ chisq ;  run;quit;  proc freq data = bank3;  table default\*loan/ chisq ;  run;quit;  proc freq data = bank3;  table default\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table default\*month/ chisq ;  run;quit;  proc freq data = bank3;  table default\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table education\*housing/ chisq ;  run;quit;  proc freq data = bank3;  table education\*job/ chisq ;  run;quit;  proc freq data = bank3;  table education\*loan/ chisq ;  run;quit;  proc freq data = bank3;  table education\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table education\*month/ chisq ;  run;quit;  proc freq data = bank3;  table education\*poutcome/ chisq ;  run;quit;      proc freq data = bank3;  table housing\*job/ chisq ;  run;quit;  proc freq data = bank3;  table housing\*loan/ chisq ;  run;quit;  proc freq data = bank3;  table housing\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table housing\*month/ chisq ;  run;quit;  proc freq data = bank3;  table housing\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table job\*loan/ chisq ;  run;quit;  proc freq data = bank3;  table job\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table job\*month/ chisq ;  run;quit;  proc freq data = bank3;  table job\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table loan\*marital/ chisq ;  run;quit;  proc freq data = bank3;  table loan\*month/ chisq ;  run;quit;  proc freq data = bank3;  table loan\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table marital\*month/ chisq ;  run;quit;  proc freq data = bank3;  table marital\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table month\*poutcome/ chisq ;  run;quit;    proc freq data = bank3;  table month\*month/ chisq ;  run;quit;    /\* glm \*/  proc glm data = bank3 plots=all;  class contact default education housing job loan marital month poutcome;  model  Yes =  contact default education housing job loan marital month poutcome  age balance day duration campaign pdays previous  / solution;  run;quit;  /\* Manova/Anova \*/  Proc GLM Data = bank3;  class y;  model age balance day duration campaign pdays previous = y;  Manova H=\_All\_ / PrintE PrintH Canonical;  Run  /\* LDA w/ priors from yes/no frequency \*/  proc discrim data=bank3 pool=test crossvalidate;  class y;  var age balance day duration campaign pdays previous;  priors "ye"=.1170 "no"=.8830;  run;  /\* Logistic Regression \*/  /\* Step1: All Main Variates \*/  proc logistic data = bank3 plots = all;  class contact default education housing job loan marital month poutcome;  model y(event = 'ye') = age balance day duration campaign pdays previous contact default education housing job loan marital month poutcome;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;  /\* Same as above w/ different plot option \*/  proc logistic data = bank3 plots = all;  class contact default education housing job loan marital month poutcome;  model y(event = 'ye') = age balance day duration campaign pdays previous contact default education housing job loan marital month poutcome  /lackfit;  effectplot;  effectplot slicefit(sliceby=Y) / noobs;  run;quit;    /\* Step2: Auto Select Logistic Regression (Stepwise) \*/  proc logistic data = bank3 plots = all;  class contact default education housing job loan marital month poutcome;  model y(event = 'ye') = age balance day duration campaign pdays previous contact default education housing job loan marital month poutcome  age |balance |day |duration |campaign |pdays |previous |contact |default |education |housing |job |loan |marital |month |poutcome @2  /selection = stepwise;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;    /\* Auto Select Logistic Regression (Forward) \*/  proc logistic data = bank3 plots = all;  class contact default education housing job loan marital month poutcome;  model y(event = 'ye') = age balance day duration campaign pdays previous contact default education housing job loan marital month poutcome  age |balance |day |duration |campaign |pdays |previous |contact |default |education |housing |job |loan |marital |month |poutcome @2  /lackfit selection = forward;  effectplot;  effectplot slicefit(sliceby=Y) / noobs;  run;quit;      /\* Step3: Logistic Regression WITHOUT 3 INSIGNIFICANT VARIABLE,age, pdays, and default (02) \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit ctable;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;    /\* Logistic Regression WITHOUT 4 INSIGNIFICANT VARIABLE,age, pdays, and default (02) \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = day duration campaign previous contact education housing job loan marital month poutcome  /lackfit;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;      /\* Step4: Logistic Regression : Add month\*poutcome \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = month\*poutcome balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;    /\* Step5: Logistic Regression : Add month\*contact \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = month\*poutcome month\*contact balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;    /\* Step6: Logistic Regression : Add month\*housing \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = month\*poutcome month\*contact month\*housing balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;    /\* Step7: Logistic Regression : Add all that found from stepwise \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = month\*poutcome month\*contact month\*housing month\*duration  month\*job contact\*housing day\*month day\*duration duration\*contact  duration\*loan duration\*housing day\*housing marital\*poutcome day\*poutcome  duration\*campaign job\*poutcome duration\*job duration\*education  duration\*marital marital\*month campaign\*job education\*marital  balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit;  effectplot fit / obs(jitter(y=0.02));  effectplot fit / obs(jitter(y=0.02)) link;  run;quit;      /\* ROC Curve Comparison: Step3 vs. Step7 \*/  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = month\*poutcome month\*contact month\*housing month\*duration  month\*job contact\*housing day\*month day\*duration duration\*contact  duration\*loan duration\*housing day\*housing marital\*poutcome day\*poutcome  duration\*campaign job\*poutcome duration\*job duration\*education  duration\*marital marital\*month campaign\*job education\*marital  balance day duration campaign previous contact education housing job loan marital month poutcome  /lackfit scale = none ;  roc 'Step 3: Without 3 Main Variables' balance day duration campaign previous contact education housing job loan marital month poutcome;  roc 'Step 7: With All Significant Interactions' month\*poutcome month\*contact month\*housing month\*duration  month\*job contact\*housing day\*month day\*duration duration\*contact  duration\*loan duration\*housing day\*housing marital\*poutcome day\*poutcome  duration\*campaign job\*poutcome duration\*job duration\*education  duration\*marital marital\*month campaign\*job education\*marital  balance day duration campaign previous contact education housing job loan marital month poutcome;  roccontrast reference('Step 3: Without 3 Main Variables')/  estimate e;  run;quit;  /\* Diagnostics Plot Trial 1 (!!! Keep failing to load output plots after running 20 mins) \*/  /\* Step3: Logistic Regression WITHOUT 3 INSIGNIFICANT VARIABLE,age, pdays, and default (02) \*/  ods graphics on;  proc logistic data = bank3 plots = all;  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = balance day duration campaign previous contact education housing job loan marital month poutcome  /influence iplots;  run;  ods graphics off;  /\* Diagnostics Plot Trial 2 (!!! Keep failing to load output plots after running 20 mins) \*/  /\* Step3: Logistic Regression WITHOUT 3 INSIGNIFICANT VARIABLE,age, pdays, and default (02) \*/  ods graphics on;  proc logistic data = bank3 plots(only label)=(phat leverage dpc);  class contact education housing job loan marital month poutcome;  model y(event = 'ye') = balance day duration campaign previous contact education housing job loan marital month poutcome;  run;  ods graphics off;      /\* Effect plot (!!! Not showing effect plot) \*/  /\* Step3: Logistic Regression WITHOUT 3 INSIGNIFICANT VARIABLE,age, pdays, and default (02) \*/  proc logistic data = bank10 plots = all;  class contact education housing job loan marital poutcome month;  model y(event = 'ye') = balance day duration campaign previous contact education housing job loan marital month poutcome;  /\* effectplot fit(x=month) / at(housing =all); \*/  effectplot fit(x = campaign) / obs(jitter(y=0.02));  effectplot fit(x = campaign) / obs(jitter(y=0.02)) link;  run;quit;      /\* Cluster (!Never stop running)\*/  proc cluster method = complete outtree = bank4;  var age balance day duration campaign pdays previous;  id ident;  run; |