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

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.

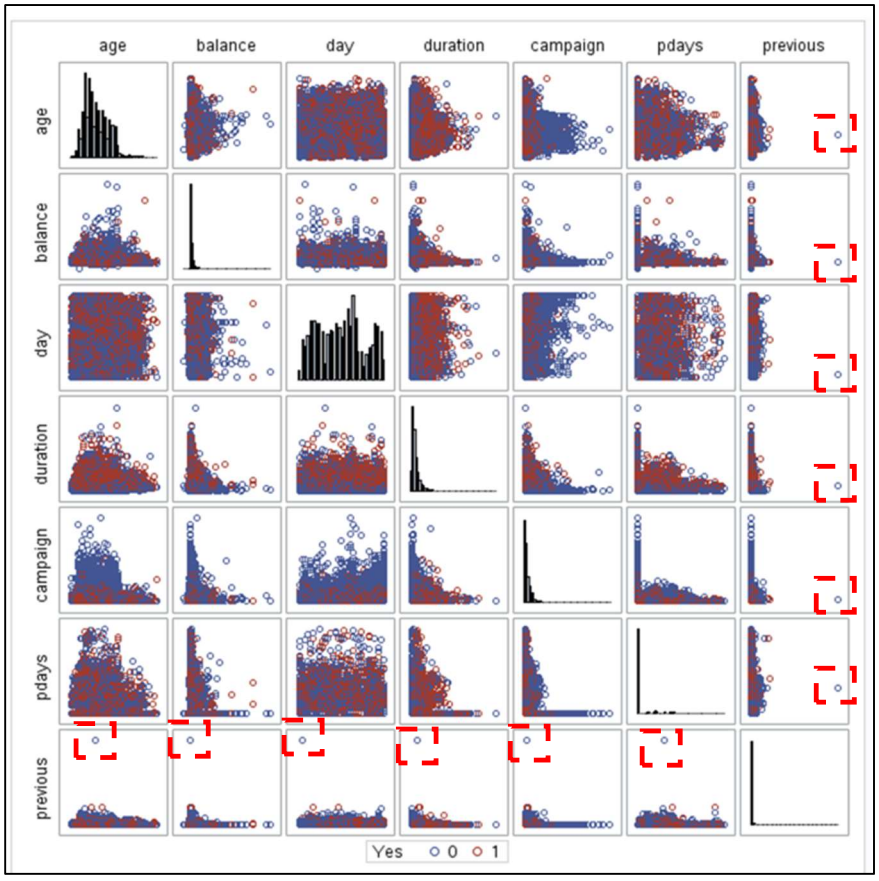


Figure 1. Matrix scatterplot of continuous variables

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.

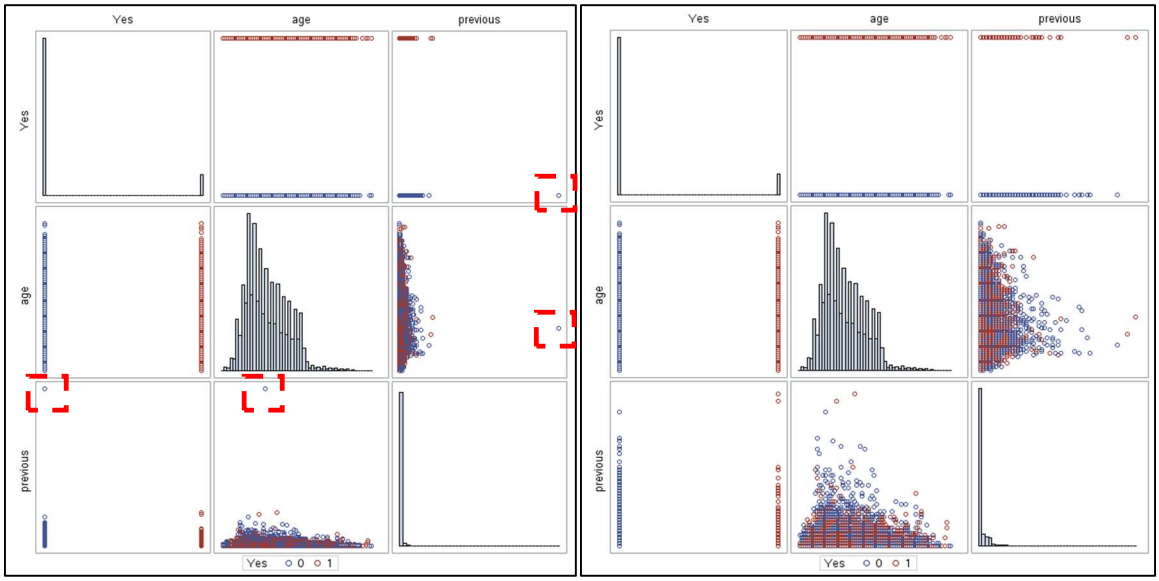


Figure 2. With outlier(left) vs without outlier(right)

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)

The MEANS Procedure

y=no

Variable	N	Mean	Maximum	Minimum	Range	Std Dev
age	39922	40.8390	95.0000	18.0000	77.0000	10.1727
balance	39922	1303.7	102127	-8019.0	110146	2974.2
day	39922	15.8923	31.0000	1.0000	30.0000	8.2947
duration	39922	221.2	4918.0	0	4918.0	207.4
campaign	39922	2.8464	63.0000	1.0000	62.0000	3.2128
pdays	39922	36.4214	871.0	-1.0000	872.0	96.7571
previous	39922	0.5022	275.0	0	275.0	2.2568

y=ye

Variable	N	Mean	Maximum	Minimum	Range	Std Dev
age	5289	41.6701	95.0000	18.0000	77.0000	13.4978
balance	5289	1804.3	81204.0	-3058.0	84262.0	3501.1
day	5289	15.1583	31.0000	1.0000	30.0000	8.5019
duration	5289	537.3	3881.0	8.0000	3873.0	392.5
campaign	5289	2.1410	32.0000	1.0000	31.0000	1.9218
pdays	5289	68.7030	854.0	-1.0000	855.0	118.8
previous	5289	1.1704	58.0000	0	58.0000	2.5533

The MEANS Procedure

y=no

Variable	N	Mean	Maximum	Minimum	Range	Std Dev
age	39921	40.8390	95.0000	18.0000	77.0000	10.1728
balance	39921	1303.7	102127	-8019.0	110146	2974.2
day	39921	15.8926	31.0000	1.0000	30.0000	8.2945
duration	39921	221.2	4918.0	0	4918.0	207.4
campaign	39921	2.8464	63.0000	1.0000	62.0000	3.2128
pdays	39921	36.4157	871.0	-1.0000	872.0	96.7518
previous	39921	0.4953	51.0000	0	51.0000	1.7904

y=ye

Variable	N	Mean	Maximum	Minimum	Range	Std Dev
age	5289	41.6701	95.0000	18.0000	77.0000	13.4978
balance	5289	1804.3	81204.0	-3058.0	84262.0	3501.1
day	5289	15.1583	31.0000	1.0000	30.0000	8.5019
duration	5289	537.3	3881.0	8.0000	3873.0	392.5
campaign	5289	2.1410	32.0000	1.0000	31.0000	1.9218
pdays	5289	68.7030	854.0	-1.0000	855.0	118.8
previous	5289	1.1704	58.0000	0	58.0000	2.5533

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

Pearson Correlation Coefficients, N = 45210 Prob > |r| under H0: Rho=0								
	Yes	age	balance	day	duration	campaign	pdays	previous
Yes	1	0.02515 <.0001	0.05284 <.0001	-0.02836 <.0001	0.39452 <.0001	-0.07317 <.0001	0.10365 <.0001	0.11372 <.0001
age	0.02515 <.0001	1	0.09778 <.0001	-0.00912 0.0524	-0.00465 0.323	0.00476 0.3115	-0.02375 <.0001	0.00184 0.6963
balance	0.05284 <.0001	0.09778 <.0001	1	0.00449 0.3394	0.02156 <.0001	-0.01458 0.0019	0.00345 0.4634	0.02099 <.0001
day	-0.02836 <.0001	-0.00912 0.0524	0.00449 0.3394	1	-0.03019 <.0001	0.16249 <.0001	-0.09297 <.0001	-0.05715 <.0001
duration	0.39452 <.0001	-0.00465 0.323	0.02156 <.0001	-0.03019 <.0001	1	-0.08457 <.0001	-0.00158 0.7366	0.00033 0.944
campaign	-0.07317 <.0001	0.00476 0.3115	-0.01458 0.0019	0.16249 <.0001	-0.08457 <.0001	1	-0.08862 <.0001	-0.03888 <.0001
pdays	0.10365 <.0001	-0.02375 <.0001	0.00345 0.4634	-0.09297 <.0001	-0.00158 0.7366	-0.08862 <.0001	1	0.54209 <.0001
previous	0.11372 <.0001	0.00184 0.6963	0.02099 <.0001	-0.05715 <.0001	0.00033 0.944	-0.03888 <.0001	0.54209 <.0001	1

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.

Chi-Square Test (p-value)									
Variable	contact	default	education	housing	job	loan	marital	month	poutcome
contact		<.0001	<.0001	<.0001	<.0001	0.0025	<.0001	<.0001	<.0001
default			0.0097	0.2003	<.0001	<.0001	0.0002	<.0001	<.0001
education				<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
housing					<.0001	<.0001	<.0001	<.0001	<.0001
job						<.0001	<.0001	<.0001	<.0001
loan							<.0001	<.0001	<.0001
marital								<.0001	<.0001
month									<.0001
poutcome									

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.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.05055	0.00640	-7.89	<.0001	0
age	1	0.00074076	0.00013004	5.70	<.0001	1.01071
balance	1	0.00000419	4.535574E-7	9.24	<.0001	1.01089
day	1	-0.00005593	0.00016785	-0.33	0.7390	1.03437
duration	1	0.00048820	0.00000535	91.17	<.0001	1.00807
campaign	1	-0.00323	0.00045217	-7.14	<.0001	1.04017
pdays	1	0.00018828	0.00001642	11.47	<.0001	1.43237
previous	1	0.01341	0.00085713	15.65	<.0001	1.41755

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

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept		1	-3.1029	0.0818	1439.38	<.0001
balance		1	0.000013	5.13E-06	6.2447	0.0125
day		1	0.0101	0.0025	16.3432	<.0001
duration		1	0.00419	0.000065	4224.041	<.0001
campaign		1	-0.0918	0.0102	81.6909	<.0001
previous		1	0.0305	0.00978	9.7365	0.0018
contact	cellula	1	0.5962	0.0361	273.1313	<.0001
contact	telepho	1	0.4297	0.0535	64.481	<.0001
education	primary	1	-0.2053	0.051	16.2001	<.0001
education	secondar y	1	-0.0196	0.0351	0.3127	0.576
education	tertiary	1	0.1765	0.0424	17.3721	<.0001
housing	no	1	0.3388	0.0218	242.4285	<.0001
job	admin.	1	0.1558	0.0572	7.4113	0.0065
job	blue-collar	1	-0.152	0.0555	7.4999	0.0062
job	entrepre neur	1	-0.1993	0.1072	3.4556	0.063
job	housema id	1	-0.3434	0.117	8.6148	0.0033
job	manage ment	1	-0.00865	0.0526	0.027	0.8694
job	retired	1	0.4137	0.0686	36.3993	<.0001
job	self-employe	1	-0.1422	0.0938	2.2953	0.1298
job	services	1	-0.066	0.0692	0.9089	0.3404
job	student	1	0.538	0.0907	35.1659	<.0001
job	technicia n	1	-0.0198	0.0513	0.1492	0.6993
job	unemplo yed	1	-0.0202	0.0941	0.046	0.8303
loan	no	1	0.214	0.0299	51.1293	<.0001
marital	divorced	1	0.0294	0.0381	0.5975	0.4396
marital	married	1	-0.1498	0.0268	31.2314	<.0001
month	apr	1	-0.0219	0.0584	0.1401	0.7081
month	aug	1	-0.7175	0.0533	181.1598	<.0001
month	dec	1	0.6657	0.1542	18.6465	<.0001
month	feb	1	-0.1653	0.0659	6.2881	0.0122
month	jan	1	-1.288	0.1036	154.6149	<.0001
month	jul	1	-0.8548	0.0557	235.5496	<.0001
month	jun	1	0.4298	0.0693	38.439	<.0001
month	mar	1	1.5671	0.0968	251.7486	<.0001
month	may	1	-0.4244	0.0507	70.1052	<.0001
month	nov	1	-0.8947	0.0635	198.7061	<.0001
month	oct	1	0.8559	0.0872	96.2567	<.0001
poutcome	failure	1	-0.6203	0.0456	184.6667	<.0001
poutcome	other	1	-0.429	0.0613	49.0107	<.0001
poutcome	success	1	1.6769	0.0521	1036.19	<.0001

$$\log(odds) = intercept + \beta_1 * balance + \beta_2 * day + \beta_3 * duration + \beta_4 * campaign + \beta_5 * previous + \beta_6 * contact(cellular) + \beta_6 * contact(telephone) + \beta_7 * education(primary) + \beta_7 * education(secondary) + \beta_7 * education(tertiary) + \beta_8 * housing(no) + \beta_9 * job(administerator) + \beta_9 * job(blue - collar) + \beta_9 * job(entrepreneur) + \beta_9 * job(housemaid) + \beta_9 * job(management) + \beta_9 * job(retired) + \beta_9 * job(self - employed) + \beta_9 * job(services) + \beta_9 * job(student) + \beta_9 * job(technician) + \beta_9 * job(umemployed) + \beta_{10} * loan(no) + \beta_{11} * marital(divorced) + \beta_{11} * marital(married) + \beta_{12} * poutcome(failure) + \beta_{12} * poutcome(other) + \beta_{12} * poutcome(success)$$

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

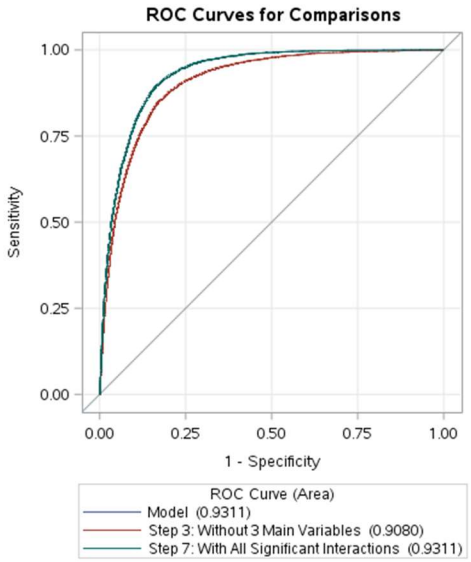


Figure 5. ROC Curve

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
498.0660	8	<.0001

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.

Model	Step 1	Step 2	Step 3 From Step1	Step4 from Step 3	Step5 from Step 4	Step6 from Step5	Step7 same as Step2
	All main variables	Stepwise w/ Interaction @2	Without 3 variables (age, pdays, and default)	Add month*poutcome	Add month*contact	Add month*housing	Add 22 signficant interaction terms
ROC area	0.908	0.9311	0.908	0.9121	0.9168	0.9196	0.9311
Goodness of Fit (Hosmer&Lemeshow)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<0.0001
AIC	21,642	19,820	21,636	21,237	21,022	20,735	19,820
AIC / AIC_STEP1	100%	92%	100%	98%	97%	96%	92%

Table 5. Table displaying models for this analysis.

Figure 6. Odds Ratio Estimates (including removed variables)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
balance	1.000	1	1
day	1.010	1.005	1.015
duration	1.004	1.004	1.004
campaign	0.912	0.894	0.931
previous	1.031	1.011	1.051
contact cellula vs unknown	5.064	4.388	5.844
contact telepho vs unknown	4.287	3.523	5.217
education primary vs unknown	0.776	0.633	0.951
education secondary vs unknown	0.934	0.782	1.116
education tertiary vs unknown	1.137	0.943	1.37
housing no vs yes	1.969	1.808	2.145
job admin. vs unknown	1.366	0.865	2.157
job blue-collar vs unknown	1.004	0.637	1.583
job entrepreneur vs unknown	0.958	0.583	1.573
job housemaid vs unknown	0.829	0.5	1.374
job management vs unknown	1.159	0.736	1.824
job retired vs unknown	1.768	1.115	2.803
job self-employe vs unknown	1.014	0.625	1.645
job services vs unknown	1.094	0.688	1.741
job student vs unknown	2.002	1.243	3.224
job technician vs unknown	1.146	0.728	1.804
job unemployed vs unknown	1.145	0.706	1.858
loan no vs yes	1.534	1.364	1.725
marital divorced vs single	0.913	0.807	1.033
marital married vs single	0.763	0.703	0.829
month apr vs sep	0.419	0.331	0.529
month aug vs sep	0.209	0.167	0.262
month dec vs sep	0.833	0.568	1.221
month feb vs sep	0.363	0.286	0.46
month jan vs sep	0.118	0.088	0.159
month jul vs sep	0.182	0.144	0.23
month jun vs sep	0.658	0.516	0.839
month mar vs sep	2.052	1.54	2.734
month may vs sep	0.280	0.224	0.351
month nov vs sep	0.175	0.138	0.222
month oct vs sep	1.008	0.77	1.319
poutcome failure vs unknown	1.007	0.887	1.144
poutcome other vs unknown	1.220	1.026	1.45
poutcome success vs unknown	10.020	8.686	11.559

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).

	Term Deposit Acceptance	
	Yes	No
Predicted Yes	109,893	177,478
	4.77%	7.70%
Predicted No	159,846	1,858,493
	6.93%	80.60%

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

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept		1	-3.1038	0.1571	390.1851	<.0001
age		1	0.00008	0.00221	0.0013	0.971
balance		1	0.000013	5.15E-06	6.1468	0.0132
day		1	0.0101	0.0025	16.328	<.0001
duration		1	0.00419	0.000065	4223.662	<.0001
campaign		1	-0.0918	0.0102	81.6253	<.0001
pdays		1	-0.00007	0.000306	0.0519	0.8198
previous		1	0.0304	0.00979	9.641	0.0019
contact	cellula	1	0.5965	0.0363	269.3975	<.0001
contact	telepho	1	0.4296	0.054	63.3351	<.0001
default	no	1	0.00849	0.0814	0.0109	0.9169
education	primary	1	-0.2053	0.0511	16.1167	<.0001
education	secondar y	1	-0.0194	0.0352	0.3039	0.5815
education	tertiary	1	0.1764	0.0427	17.044	<.0001
housing	no	1	0.3384	0.0219	237.8108	<.0001
job	admin.	1	0.156	0.0574	7.3874	0.0066
job	blue-collar	1	-0.1515	0.0559	7.3481	0.0067
job	entrepre neur	1	-0.1988	0.1073	3.4326	0.0639
job	housema id	1	-0.344	0.1172	8.618	0.0033
job	manage ment	1	-0.00845	0.0526	0.0258	0.8725
job	retired	1	0.4119	0.0786	27.4657	<.0001
job	self-employe	1	-0.1421	0.0939	2.2903	0.1302
job	services	1	-0.0656	0.0695	0.8908	0.3453
job	student	1	0.5384	0.0941	32.7666	<.0001
job	technicia n	1	-0.0197	0.0515	0.1469	0.7015
job	unemplo yed	1	-0.0197	0.0942	0.0438	0.8342
loan	no	1	0.2138	0.03	50.7762	<.0001
marital	divorced	1	0.0293	0.0393	0.5584	0.4549
marital	married	1	-0.1502	0.0272	30.5719	<.0001
month	apr	1	-0.0213	0.0585	0.1332	0.7152
month	aug	1	-0.7175	0.0533	180.9919	<.0001
month	dec	1	0.6656	0.1541	18.6506	<.0001
month	feb	1	-0.1658	0.066	6.3193	0.0119
month	jan	1	-1.2882	0.1036	154.6322	<.0001
month	jul	1	-0.8543	0.0558	234.7112	<.0001
month	jun	1	0.4298	0.0693	38.4437	<.0001
month	mar	1	1.5669	0.0988	251.6319	<.0001
month	may	1	-0.4232	0.051	68.9328	<.0001
month	nov	1	-0.896	0.0638	197.4061	<.0001
month	oct	1	0.8554	0.0873	96.0057	<.0001
poutcome	failure	1	-0.6148	0.0517	141.62	<.0001
poutcome	other	1	-0.4244	0.0643	43.5016	<.0001
poutcome	success	1	1.6776	0.0523	1030.158	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.000	0.996	1.004
balance	1.000	1	1
day	1.010	1.005	1.015
duration	1.004	1.004	1.004
campaign	0.912	0.894	0.931
pdays	1.000	0.999	1.001
previous	1.031	1.011	1.051
contact cellula vs unknown	5.067	4.39	5.848
contact telepho vs unknown	4.288	3.52	5.223
default no vs yes	1.017	0.739	1.399
education primary vs unknown	0.776	0.633	0.951
education secondary vs unknown	0.934	0.781	1.117
education tertiary vs unknown	1.137	0.942	1.372
housing no vs yes	1.967	1.805	2.144
job admin. vs unknown	1.367	0.865	2.16
job blue-collar vs unknown	1.005	0.637	1.586
job entrepreneur vs unknown	0.958	0.583	1.575
job housemaid vs unknown	0.829	0.5	1.374
job management vs unknown	1.159	0.736	1.826
job retired vs unknown	1.765	1.109	2.811
job self-employe vs unknown	1.014	0.625	1.646
job services vs unknown	1.095	0.688	1.744
job student vs unknown	2.003	1.239	3.24
job technician vs unknown	1.146	0.728	1.806
job unemployed vs unknown	1.146	0.706	1.86
loan no vs yes	1.534	1.363	1.725
marital divorced vs single	0.913	0.8	1.041
marital married vs single	0.763	0.697	0.834
month apr vs sep	0.419	0.331	0.53
month aug vs sep	0.209	0.166	0.262
month dec vs sep	0.833	0.568	1.221
month feb vs sep	0.363	0.286	0.46
month jan vs sep	0.118	0.088	0.159
month jul vs sep	0.182	0.144	0.23
month jun vs sep	0.658	0.516	0.838
month mar vs sep	2.051	1.539	2.732
month may vs sep	0.280	0.224	0.351
month nov vs sep	0.175	0.137	0.222
month oct vs sep	1.007	0.769	1.318
poutcome failure vs unknown	1.024	0.847	1.238
poutcome other vs unknown	1.239	0.997	1.539
poutcome success vs unknown	10.135	8.507	12.075

III. Frequency Table

The FREQ Procedure

y=no

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	4540	11.37	4540	11.37
blue-collar	9024	22.60	13564	33.98
entrepreneur	1364	3.42	14928	37.39
housemaid	1131	2.83	16059	40.23
management	8157	20.43	24216	60.66
retired	1748	4.38	25964	65.04
self-employe	1392	3.49	27356	68.52
services	3785	9.48	31141	78.00
student	669	1.68	31810	79.68
technician	6757	16.93	38567	96.61
unemployed	1101	2.76	39668	99.36
unknown	254	0.64	39922	100.00

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
divorced	4585	11.48	4585	11.48
married	24459	61.27	29044	72.75
single	10878	27.25	39922	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
primary	6260	15.68	6260	15.68
secondary	20752	51.98	27012	67.66
tertiary	11305	28.32	38317	95.98
unknown	1605	4.02	39922	100.00

The FREQ Procedure

y=ye

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	631	11.93	631	11.93
blue-collar	708	13.39	1339	25.32
entrepreneur	123	2.33	1462	27.64
housemaid	109	2.06	1571	29.70
management	1301	24.60	2872	54.30
retired	516	9.76	3388	64.06
self-employe	187	3.54	3575	67.59
services	369	6.98	3944	74.57
student	269	5.09	4213	79.66
technician	840	15.88	5053	95.54
unemployed	202	3.82	5255	99.36
unknown	34	0.64	5289	100.00

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
divorced	622	11.76	622	11.76
married	2755	52.09	3377	63.85
single	1912	36.15	5289	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
primary	591	11.17	591	11.17
secondary	2450	46.32	3041	57.50
tertiary	1996	37.74	5037	95.24
unknown	252	4.76	5289	100.00

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	39159	98.09	39159	98.09
yes	763	1.91	39922	100.00

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	5237	99.02	5237	99.02
yes	52	0.98	5289	100.00

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	16727	41.90	16727	41.90
yes	23195	58.10	39922	100.00

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	3354	63.41	3354	63.41
yes	1935	36.59	5289	100.00

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	33162	83.07	33162	83.07
yes	6760	16.93	39922	100.00

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	4805	90.85	4805	90.85
yes	484	9.15	5289	100.00

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellula	24916	62.41	24916	62.41
telepho	2516	6.30	27432	68.71
unknown	12490	31.29	39922	100.00

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellula	4369	82.61	4369	82.61
telepho	390	7.37	4759	89.98
unknown	530	10.02	5289	100.00

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
apr	2355	5.90	2355	5.90
aug	5559	13.92	7914	19.82
dec	114	0.29	8028	20.11
feb	2208	5.53	10236	25.64
jan	1261	3.16	11497	28.80
jul	6268	15.70	17765	44.50
jun	4795	12.01	22560	56.51
mar	229	0.57	22789	57.08
may	12841	32.17	35630	89.25
nov	3567	8.93	39197	98.18
oct	415	1.04	39612	99.22
sep	310	0.78	39922	100.00

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
apr	577	10.91	577	10.91
aug	688	13.01	1265	23.92
dec	100	1.89	1365	25.81
feb	441	8.34	1806	34.15
jan	142	2.68	1948	36.83
jul	627	11.85	2575	48.69
jun	546	10.32	3121	59.01
mar	248	4.69	3369	63.70
may	925	17.49	4294	81.19
nov	403	7.62	4697	88.81
oct	323	6.11	5020	94.91
sep	269	5.09	5289	100.00

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
failure	4283	10.73	4283	10.73
other	1533	3.84	5816	14.57
success	533	1.34	6349	15.90
unknown	33573	84.10	39922	100.00

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
failure	618	11.68	618	11.68
other	307	5.80	925	17.49
success	978	18.49	1903	35.98
unknown	3386	64.02	5289	100.00

IV. Model Selection Methods

Model	Step 1	Step 2	Step 3 From Step1	Step4 from Step 3	Step5 from Step 4	Step6 from Step5	Step7 same as Step2
	All main variables	Stepwise w/ Interaction @2	Without 3 variables (age, pdays, and default)	Add month*poutcome	Add month*contact	Add month*housing	Add all sigfnlficant Interaction terms
ROC area	0.908	0.9311	0.908	0.9121	0.9168	0.9196	0.9311
Goodness of Flt (Hosmer&Lemeshow)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<0.0001

V. Paired Chi-Square Test Results

Statistics for Table of contact by default

Statistic	DF	Value	Prob
Chi-Square	2	26.9270	<.0001
Likelihood Ratio Chi-Square	2	28.7655	<.0001
Mantel-Haenszel Chi-Square	1	10.7247	0.0011
Phi Coefficient		0.0244	
Contingency Coefficient		0.0244	
Cramer's V		0.0244	

Sample Size = 45210

Statistics for Table of contact by education

Statistic	DF	Value	Prob
Chi-Square	6	1363.1417	<.0001
Likelihood Ratio Chi-Square	6	1364.2033	<.0001
Mantel-Haenszel Chi-Square	1	556.1489	<.0001
Phi Coefficient		0.1736	
Contingency Coefficient		0.1711	
Cramer's V		0.1228	

Sample Size = 45210

Statistics for Table of contact by housing

Statistic	DF	Value	Prob
Chi-Square	2	2062.6718	<.0001
Likelihood Ratio Chi-Square	2	2122.6419	<.0001
Mantel-Haenszel Chi-Square	1	1600.2426	<.0001
Phi Coefficient		0.2136	
Contingency Coefficient		0.2089	
Cramer's V		0.2136	

Sample Size = 45210

Statistics for Table of contact by job

Statistic	DF	Value	Prob
Chi-Square	22	2046.8503	<.0001
Likelihood Ratio Chi-Square	22	1867.6809	<.0001
Mantel-Haenszel Chi-Square	1	304.4705	<.0001
Phi Coefficient		0.2128	
Contingency Coefficient		0.2081	
Cramer's V		0.1505	

Sample Size = 45210

Statistics for Table of contact by loan

Statistic	DF	Value	Prob
Chi-Square	2	11.9821	0.0025
Likelihood Ratio Chi-Square	2	12.2192	0.0022
Mantel-Haenszel Chi-Square	1	5.3516	0.0207
Phi Coefficient		0.0163	
Contingency Coefficient		0.0163	
Cramer's V		0.0163	

Sample Size = 45210

Statistics for Table of contact by marital

Statistic	DF	Value	Prob
Chi-Square	4	183.8948	<.0001
Likelihood Ratio Chi-Square	4	187.4915	<.0001
Mantel-Haenszel Chi-Square	1	69.4913	<.0001
Phi Coefficient		0.0638	
Contingency Coefficient		0.0636	
Cramer's V		0.0451	

Sample Size = 45210

Statistics for Table of contact by month

Statistic	DF	Value	Prob
Chi-Square	22	23714.6502	<.0001
Likelihood Ratio Chi-Square	22	26866.5831	<.0001
Mantel-Haenszel Chi-Square	1	5896.1382	<.0001
Phi Coefficient		0.7243	
Contingency Coefficient		0.5866	
Cramer's V		0.5121	

Sample Size = 45210

Statistics for Table of contact by poutcome

Statistic	DF	Value	Prob
Chi-Square	6	3891.7229	<.0001
Likelihood Ratio Chi-Square	6	5646.8217	<.0001
Mantel-Haenszel Chi-Square	1	3349.6126	<.0001
Phi Coefficient		0.2934	
Contingency Coefficient		0.2815	
Cramer's V		0.2075	

Sample Size = 45210

Statistics for Table of default by education

Statistic	DF	Value	Prob
Chi-Square	3	11.4185	0.0097
Likelihood Ratio Chi-Square	3	11.7677	0.0082
Mantel-Haenszel Chi-Square	1	5.1902	0.0227
Phi Coefficient		0.0159	
Contingency Coefficient		0.0159	
Cramer's V		0.0159	

Sample Size = 45210

Statistics for Table of default by housing

Statistic	DF	Value	Prob
Chi-Square	1	1.6398	0.2003
Likelihood Ratio Chi-Square	1	1.6350	0.2010
Continuity Adj. Chi-Square	1	1.5500	0.2131
Mantel-Haenszel Chi-Square	1	1.6398	0.2004
Phi Coefficient		-0.0060	
Contingency Coefficient		0.0060	
Cramer's V		-0.0060	

Statistics for Table of default by job

Statistic	DF	Value	Prob
Chi-Square	11	60.3399	<.0001
Likelihood Ratio Chi-Square	11	60.7304	<.0001
Mantel-Haenszel Chi-Square	1	2.1234	0.1451
Phi Coefficient		0.0365	
Contingency Coefficient		0.0365	
Cramer's V		0.0365	

Sample Size = 45210

Statistics for Table of default by loan

Statistic	DF	Value	Prob
Chi-Square	1	269.6755	<.0001
Likelihood Ratio Chi-Square	1	213.2485	<.0001
Continuity Adj. Chi-Square	1	268.0954	<.0001
Mantel-Haenszel Chi-Square	1	269.6696	<.0001
Phi Coefficient		0.0772	
Contingency Coefficient		0.0770	
Cramer's V		0.0772	

Statistics for Table of default by poutcome

Statistic	DF	Value	Prob
Chi-Square	3	73.7876	<.0001
Likelihood Ratio Chi-Square	3	99.3540	<.0001
Mantel-Haenszel Chi-Square	1	55.0475	<.0001
Phi Coefficient		0.0404	
Contingency Coefficient		0.0404	
Cramer's V		0.0404	

Sample Size = 45210

Statistics for Table of education by housing

Statistic	DF	Value	Prob
Chi-Square	3	644.1843	<.0001
Likelihood Ratio Chi-Square	3	643.1094	<.0001
Mantel-Haenszel Chi-Square	1	372.8355	<.0001
Phi Coefficient		0.1194	
Contingency Coefficient		0.1185	
Cramer's V		0.1194	

Sample Size = 45210

Statistics for Table of education by job

Statistic	DF	Value	Prob
Chi-Square	33	28481.3009	<.0001
Likelihood Ratio Chi-Square	33	27750.9643	<.0001
Mantel-Haenszel Chi-Square	1	1256.4753	<.0001
Phi Coefficient		0.7937	
Contingency Coefficient		0.6217	
Cramer's V		0.4582	

Sample Size = 45210

Statistics for Table of education by loan

Statistic	DF	Value	Prob
Chi-Square	3	291.3091	<.0001
Likelihood Ratio Chi-Square	3	313.6734	<.0001
Mantel-Haenszel Chi-Square	1	106.6246	<.0001
Phi Coefficient		0.0803	
Contingency Coefficient		0.0800	
Cramer's V		0.0803	

Sample Size = 45210

Statistics for Table of education by marital

Statistic	DF	Value	Prob
Chi-Square	6	1337.7681	<.0001
Likelihood Ratio Chi-Square	6	1458.9218	<.0001
Mantel-Haenszel Chi-Square	1	533.0343	<.0001
Phi Coefficient		0.1720	
Contingency Coefficient		0.1695	
Cramer's V		0.1216	

Sample Size = 45210

Statistics for Table of education by month

Statistic	DF	Value	Prob
Chi-Square	33	1644.1048	<.0001
Likelihood Ratio Chi-Square	33	1606.9532	<.0001
Mantel-Haenszel Chi-Square	1	148.3603	<.0001
Phi Coefficient		0.1907	
Contingency Coefficient		0.1873	
Cramer's V		0.1101	

Sample Size = 45210

Statistics for Table of education by poutcome

Statistic	DF	Value	Prob
Chi-Square	9	172.3927	<.0001
Likelihood Ratio Chi-Square	9	172.6582	<.0001
Mantel-Haenszel Chi-Square	1	16.8853	<.0001
Phi Coefficient		0.0618	
Contingency Coefficient		0.0616	
Cramer's V		0.0357	

Sample Size = 45210

Statistics for Table of housing by job

Statistic	DF	Value	Prob
Chi-Square	11	3588.9505	<.0001
Likelihood Ratio Chi-Square	11	3714.2292	<.0001
Mantel-Haenszel Chi-Square	1	710.4866	<.0001
Phi Coefficient		0.2818	
Contingency Coefficient		0.2712	
Cramer's V		0.2818	

Sample Size = 45210

Statistics for Table of housing by loan

Statistic	DF	Value	Prob
Chi-Square	1	77.2336	<.0001
Likelihood Ratio Chi-Square	1	77.7819	<.0001
Continuity Adj. Chi-Square	1	77.0070	<.0001
Mantel-Haenszel Chi-Square	1	77.2319	<.0001
Phi Coefficient		0.0413	
Contingency Coefficient		0.0413	
Cramer's V		0.0413	

Statistics for Table of housing by marital

Statistic	DF	Value	Prob
Chi-Square	2	19.3192	<.0001
Likelihood Ratio Chi-Square	2	19.2949	<.0001
Mantel-Haenszel Chi-Square	1	11.7049	0.0006
Phi Coefficient		0.0207	
Contingency Coefficient		0.0207	
Cramer's V		0.0207	

Sample Size = 45210

Statistics for Table of housing by month

Statistic	DF	Value	Prob
Chi-Square	11	11494.4979	<.0001
Likelihood Ratio Chi-Square	11	12620.3239	<.0001
Mantel-Haenszel Chi-Square	1	3332.5024	<.0001
Phi Coefficient		0.5042	
Contingency Coefficient		0.4502	
Cramer's V		0.5042	

Sample Size = 45210

Statistics for Table of housing by poutcome

Statistic	DF	Value	Prob
Chi-Square	3	926.1123	<.0001
Likelihood Ratio Chi-Square	3	948.0616	<.0001
Mantel-Haenszel Chi-Square	1	451.5758	<.0001
Phi Coefficient		0.1431	
Contingency Coefficient		0.1417	
Cramer's V		0.1431	

Sample Size = 45210

Statistics for Table of job by loan

Statistic	DF	Value	Prob
Chi-Square	11	512.7410	<.0001
Likelihood Ratio Chi-Square	11	627.3036	<.0001
Mantel-Haenszel Chi-Square	1	49.2475	<.0001
Phi Coefficient		0.1065	
Contingency Coefficient		0.1059	
Cramer's V		0.1065	

Sample Size = 45210

Statistics for Table of job by marital

Statistic	DF	Value	Prob
Chi-Square	22	3837.6616	<.0001
Likelihood Ratio Chi-Square	22	3954.4689	<.0001
Mantel-Haenszel Chi-Square	1	174.0353	<.0001
Phi Coefficient		0.2914	
Contingency Coefficient		0.2797	
Cramer's V		0.2060	

Sample Size = 45210

Statistics for Table of job by month

Statistic	DF	Value	Prob
Chi-Square	121	6043.9513	<.0001
Likelihood Ratio Chi-Square	121	5699.8889	<.0001
Mantel-Haenszel Chi-Square	1	389.9391	<.0001
Phi Coefficient		0.3656	
Contingency Coefficient		0.3434	
Cramer's V		0.1102	

Sample Size = 45210

Statistics for Table of job by poutcome

Statistic	DF	Value	Prob
Chi-Square	33	559.2745	<.0001
Likelihood Ratio Chi-Square	33	521.1641	<.0001
Mantel-Haenszel Chi-Square	1	5.4773	0.0193
Phi Coefficient		0.1112	
Contingency Coefficient		0.1105	
Cramer's V		0.0642	

Sample Size = 45210

Statistics for Table of loan by marital

Statistic	DF	Value	Prob
Chi-Square	2	121.9775	<.0001
Likelihood Ratio Chi-Square	2	126.1539	<.0001
Mantel-Haenszel Chi-Square	1	99.4224	<.0001
Phi Coefficient		0.0519	
Contingency Coefficient		0.0519	
Cramer's V		0.0519	

Sample Size = 45210

Statistics for Table of loan by month

Statistic	DF	Value	Prob
Chi-Square	11	1511.1458	<.0001
Likelihood Ratio Chi-Square	11	1399.2352	<.0001
Mantel-Haenszel Chi-Square	1	22.1545	<.0001
Phi Coefficient		0.1828	
Contingency Coefficient		0.1798	
Cramer's V		0.1828	

Sample Size = 45210

Statistics for Table of loan by poutcome

Statistic	DF	Value	Prob
Chi-Square	3	137.9603	<.0001
Likelihood Ratio Chi-Square	3	173.9681	<.0001
Mantel-Haenszel Chi-Square	1	10.7817	0.0010
Phi Coefficient		0.0552	
Contingency Coefficient		0.0552	
Cramer's V		0.0552	

Sample Size = 45210

Statistics for Table of marital by month

Statistic	DF	Value	Prob
Chi-Square	22	473.1129	<.0001
Likelihood Ratio Chi-Square	22	471.1510	<.0001
Mantel-Haenszel Chi-Square	1	2.2126	0.1369
Phi Coefficient		0.1023	
Contingency Coefficient		0.1018	
Cramer's V		0.0723	

Sample Size = 45210

Statistics for Table of marital by poutcome

Statistic	DF	Value	Prob
Chi-Square	6	76.7164	<.0001
Likelihood Ratio Chi-Square	6	74.6281	<.0001
Mantel-Haenszel Chi-Square	1	12.8519	0.0003
Phi Coefficient		0.0412	
Contingency Coefficient		0.0412	
Cramer's V		0.0291	

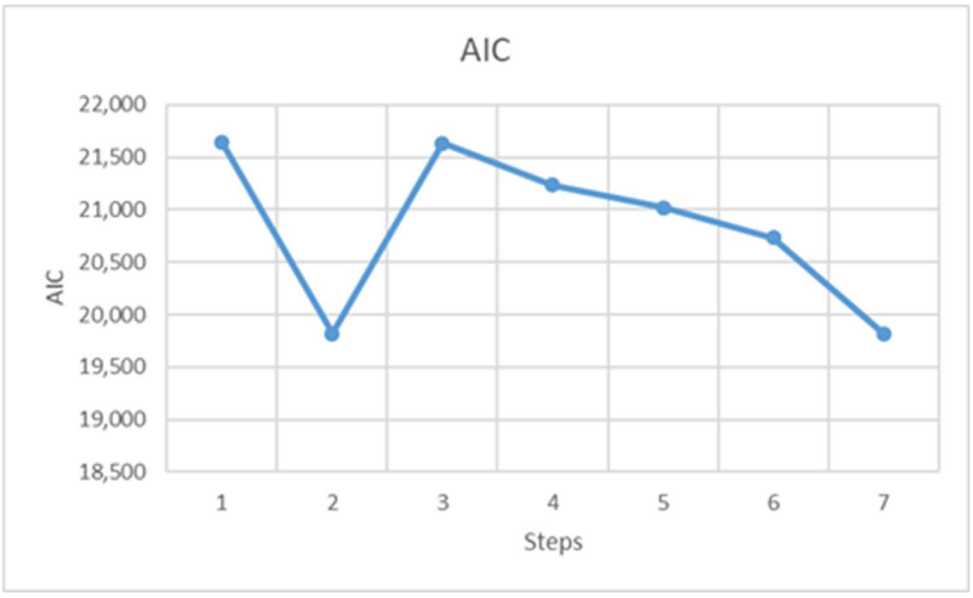
Sample Size = 45210

Statistics for Table of month by poutcome

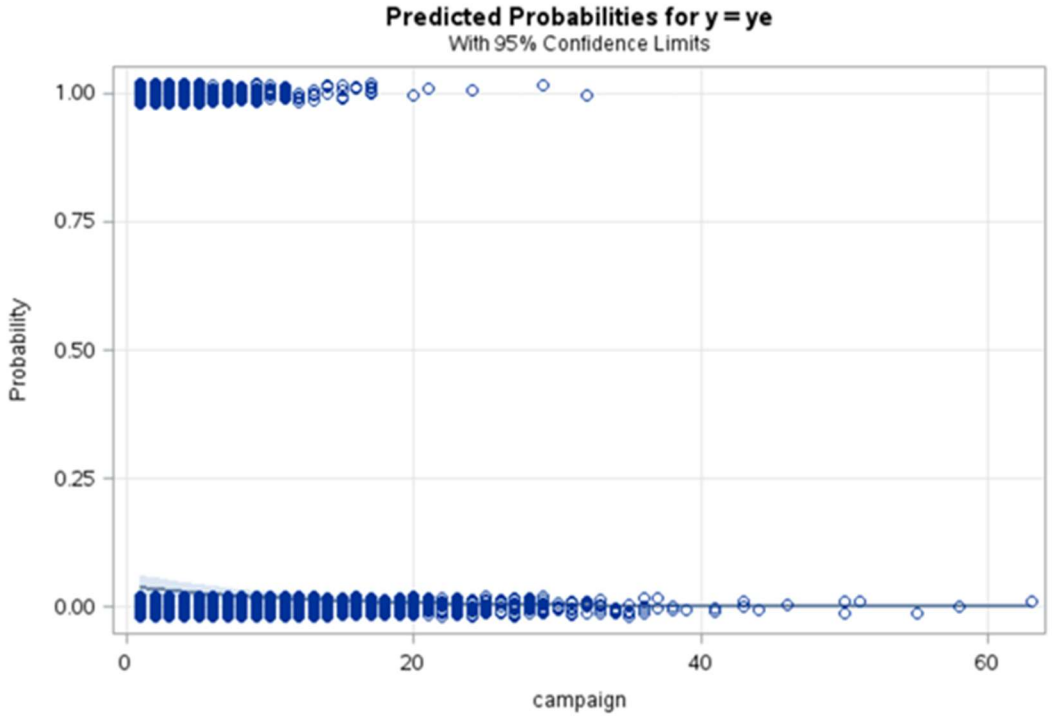
Statistic	DF	Value	Prob
Chi-Square	33	6228.9430	<.0001
Likelihood Ratio Chi-Square	33	5764.6390	<.0001
Mantel-Haenszel Chi-Square	1	49.4376	<.0001
Phi Coefficient		0.3712	
Contingency Coefficient		0.3480	
Cramer's V		0.2143	

Sample Size = 45210

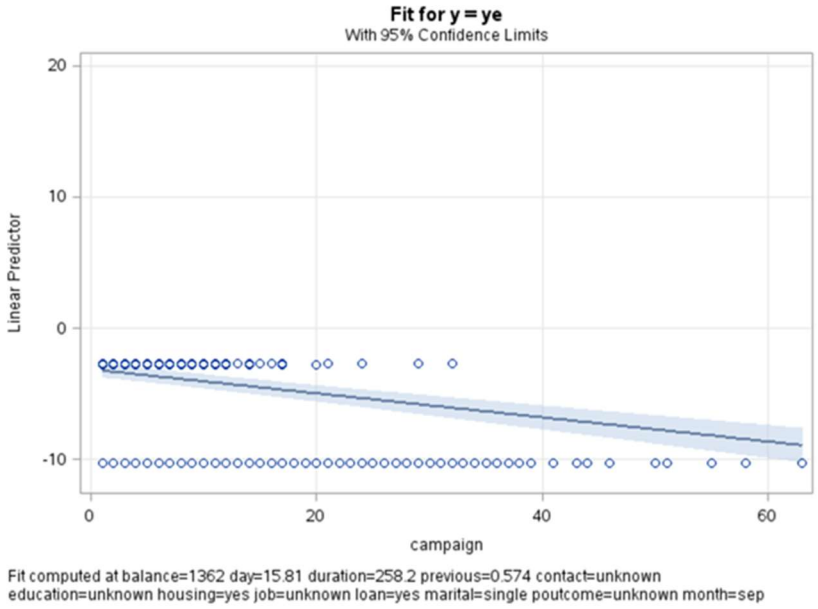
VI. Model Steps vs AIC



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

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	FALSE POS	FALSE NEG
0	5289	0	39921	0	11.7	100	0	88.3	.
0.02	5253	11290	28631	36	36.6	99.3	28.3	84.5	0.3
0.04	5144	20984	18937	145	57.8	97.3	52.6	78.6	0.7
0.06	4971	27079	12842	318	70.9	94	67.8	72.1	1.2
0.08	4769	30435	9486	520	77.9	90.2	76.2	66.5	1.7
0.1	4563	32422	7499	726	81.8	86.3	81.2	62.2	2.2
0.12	4344	33771	6150	945	84.3	82.1	84.6	58.6	2.7
0.14	4116	34720	5201	1173	85.9	77.8	87	55.8	3.3
0.16	3917	35458	4463	1372	87.1	74.1	88.8	53.3	3.7
0.18	3727	36011	3910	1562	87.9	70.5	90.2	51.2	4.2
0.2	3557	36429	3492	1732	88.4	67.3	91.3	49.5	4.5
0.22	3401	36807	3114	1888	88.9	64.3	92.2	47.8	4.9
0.24	3254	37103	2818	2035	89.3	61.5	92.9	46.4	5.2
0.26	3119	37379	2542	2170	89.6	59	93.6	44.9	5.5
0.28	2990	37617	2304	2299	89.8	56.5	94.2	43.5	5.8
0.3	2871	37827	2094	2418	90	54.3	94.8	42.2	6
0.32	2757	37993	1928	2532	90.1	52.1	95.2	41.2	6.2
0.34	2648	38143	1778	2641	90.2	50.1	95.5	40.2	6.5
0.36	2527	38287	1634	2762	90.3	47.8	95.9	39.3	6.7
0.38	2412	38406	1515	2877	90.3	45.6	96.2	38.6	7
0.4	2319	38524	1397	2970	90.3	43.8	96.5	37.6	7.2
0.42	2204	38618	1303	3085	90.3	41.7	96.7	37.2	7.4
0.44	2116	38701	1220	3173	90.3	40	96.9	36.6	7.6
0.46	2013	38783	1138	3276	90.2	38.1	97.1	36.1	7.8
0.48	1921	38866	1055	3368	90.2	36.3	97.4	35.5	8
0.5	1827	38933	988	3462	90.2	34.5	97.5	35.1	8.2
0.52	1751	39000	921	3538	90.1	33.1	97.7	34.5	8.3
0.54	1680	39057	864	3609	90.1	31.8	97.8	34	8.5
0.56	1596	39115	806	3693	90	30.2	98	33.6	8.6
0.58	1515	39172	749	3774	90	28.6	98.1	33.1	8.8
0.6	1451	39228	693	3838	90	27.4	98.3	32.3	8.9
0.62	1368	39273	648	3921	89.9	25.9	98.4	32.1	9.1
0.64	1291	39322	599	3998	89.8	24.4	98.5	31.7	9.2
0.66	1225	39363	558	4064	89.8	23.2	98.6	31.3	9.4
0.68	1158	39405	516	4131	89.7	21.9	98.7	30.8	9.5
0.7	1087	39433	488	4202	89.6	20.6	98.8	31	9.6
0.72	1035	39476	445	4254	89.6	19.6	98.9	30.1	9.7
0.74	955	39513	408	4334	89.5	18.1	99	29.9	9.9
0.76	881	39554	367	4408	89.4	16.7	99.1	29.4	10
0.78	814	39590	331	4475	89.4	15.4	99.2	28.9	10.2
0.8	734	39616	305	4555	89.3	13.9	99.2	29.4	10.3
0.82	658	39645	276	4631	89.1	12.4	99.3	29.6	10.5
0.84	586	39681	240	4703	89.1	11.1	99.4	29.1	10.6
0.86	507	39707	214	4782	88.9	9.6	99.5	29.7	10.7
0.88	438	39739	182	4851	88.9	8.3	99.5	29.4	10.9
0.9	370	39769	152	4919	88.8	7	99.6	29.1	11
0.92	302	39799	122	4987	88.7	5.7	99.7	28.8	11.1
0.94	224	39819	102	5065	88.6	4.2	99.7	31.3	11.3
0.96	144	39842	79	5145	88.4	2.7	99.8	35.4	11.4
0.98	94	39868	53	5195	88.4	1.8	99.9	36.1	11.5
1	0	39921	0	5289	88.3	0	100	.	11.7

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;
```