FACTORS DETERMINING TERM DEPOSIT PURCHASES

How a Bank Can Get Other People's Money

DECEMBER 31, 2016

KENNESAW STATE UNIVERSITY
THE THREE AMIGOS
GINA COLAIANNI, JONAS MAGDANGAL, MATTHEW MITCHELL

1. Introduction

Banks exist to provide monetary services to people and to make profit. With that in mind, banks devote significant resources and activity to gain capital. One way banks do this is to engage in direct marketing campaigns to sell and provide services. Our group found a data set that was the result of a Portuguese Bank direct marketing campaign to sell term deposits. We set out to determine what factors in the data set would contribute to a high volume of sales of term deposits.

2. Data

The Data Set is the Portuguese Bank Marketing Data Set in the University of California, Irvine (UCI) Machine Learning Repository located at the following URL: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. The data is a result of a direct marketing campaign performed by a Portuguese banking institution to sell term deposits/certificate of deposits. The banking institution made phone calls to potential buyers from May 2008 to November 2010. Often, more than 1 contact to the same client was required to assess whether a client would place an order. The full data set, bank-additional-full.csv, was used.

There are 41,188 observations and 21 Variables in the Data Set. There are 10 continuous measure variables and 10 categorical variables. The target response (y) is a binary response indicating whether the client subscribed to a term deposit or not. 'Yes' (numeric value 1) indicated the client subscribed to a term deposit. 'No' (numeric value 0) indicated the client did not subscribe to a term deposit. Table 1 in the Appendix describes the variables in the data set. Table 1 gives the variable name, its category, its description, and the type of variable (Continuous or Categorical.) Variable duration is listed in Table 1 but will not be used in the analysis due to the high impact on the target response (y) per the variable description. The variables are broken into 4 categories: Client Data, Last Contact Info, Other, and Social and Economic Variables.

3. Problem/Objective

The first objective of this study is to determine which variables have the highest influence on whether a client purchases a term deposit or not. The second objective is to determine the levels of those variables that produce the most term deposit purchases.

4. Data Cleaning/Validation

4.1. Continuous Variables

There are no missing continuous values in this data set. Thus, no imputation was necessary. Figure 1 through 10 in the Appendix are continuous variable histogram and boxplots by the target response (y) variable. A review of the plots reveals the following:

- Figure 1 - Age does not appear to have an impact on the target variable (y). The histograms are centered around the same region and have very similar variance regardless of the value of y.
- The following continuous variables were categorized due to the histograms, box plots, and frequency tables showing that the variables are more categorical in nature than continuous:

<table>
<tr><th>Continuous Variable</th><th>Figure</th><th>Figure Notes</th><th>Categorization Notes</th><th>Categorized Variable</th></tr>
<tr><td>campaign</td><td>3</td><td>> 97% of data in lowest 10 levels for both response levels.</td><td>Ordinal into {1,2,3,>3)</td><td>campaign_cat</td></tr>
<tr><td>pdays</td><td>4</td><td>> 79% of data in pdays = 999 (not previously contacted) for both response levels.</td><td>Binary into {‘contacted before’ and ‘never contacted’}</td><td>pdays_cat</td></tr>
<tr><td>previous</td><td>5</td><td>>68% of data in lowest level for both response levels.</td><td>Binary into {‘contacted before’ and ‘never contacted’}</td><td>previous_cat</td></tr>
<tr><td>emp_var_rate</td><td>6</td><td>Per bar chart, natural grouping seems to be {<=-1.8, (-1.8 to -0.1], > -0.1} for both response levels.</td><td>Ordinal into {<=-1.8, (-1.8 to -0.1], > -0.1}</td><td>emp_var_rate_cat</td></tr>
<tr><td>cons_price_idx</td><td>7</td><td>Data is highly multi-modal. Bucket binning used since data has even spread through the range of the histogram.</td><td>Ordinally Bucket binned into:
<table>
<tr><th>Range</th><th>Frequency</th><th>Proportion</th></tr>
<tr><td>cons_price_idx < 93.056333333</td><td>8992</td><td>0.21831601</td></tr>
<tr><td>93.056333333 <= cons_price_idx < 93.911666667</td><td>11966</td><td>0.29052151</td></tr>
<tr><td>93.911666667 <= cons_price_idx</td><td>20230</td><td>0.49116247</td></tr>
</table></td><td>cons_price_idx_cat</td></tr>
<tr><td>cons_conf_idx</td><td>8</td><td>Data is highly multi-modal. Quintile binning used since data has does not have even spread through histogram.</td><td>Ordinally Quintile binned into:
<table>
<tr><th>Range</th><th>Frequency</th><th>Proportion</th></tr>
<tr><td>cons_conf_idx < -46.19925</td><td>8866</td><td>0.21525687</td></tr>
<tr><td>-46.19925 <= cons_conf_idx < -41.99763</td><td>10311</td><td>0.25033990</td></tr>
<tr><td>-41.99763 <= cons_conf_idx < -39.99959</td><td>5679</td><td>0.13787997</td></tr>
<tr><td>-39.99959 <= cons_conf_idx < -36.39786</td><td>8528</td><td>0.20705060</td></tr>
<tr><td>-36.39786 <= cons_conf_idx</td><td>7804</td><td>0.18947266</td></tr>
</table></td><td>cons_conf_idx_cat</td></tr>
<tr><td>euribor3m</td><td>9</td><td>Data is highly multi-modal. Quintile binning used since data has does not have even spread through histogram.</td><td>Ordinally Quintile binned into:
<table>
<tr><th>Range</th><th>Frequency</th><th>Proportion</th></tr>
<tr><td>euribor3m < 1.2991788</td><td>8636</td><td>0.20967272</td></tr>
<tr><td>1.2991788 <= euribor3m < 4.1910304</td><td>8430</td><td>0.20467126</td></tr>
<tr><td>4.1910304 <= euribor3m < 4.864149</td><td>8446</td><td>0.20505973</td></tr>
<tr><td>4.864149 <= euribor3m < 4.9620732</td><td>8498</td><td>0.20632223</td></tr>
<tr><td>4.9620732 <= euribor3m</td><td>7178</td><td>0.17427406</td></tr>
</table></td><td>euribor3m_cat</td></tr>
<tr><td>nr_employed</td><td>10</td><td>Data is highly multi-modal. Quintile binning used since data has does not have even spread through histogram. SAS was only able to bin into 3 categories.</td><td>Ordinally Quintile binned into:
<table>
<tr><th>Range</th><th>Frequency</th><th>Proportion</th></tr>
<tr><td>nr_employed < 5099.10335</td><td>13498</td><td>0.32771681</td></tr>
<tr><td>5099.10335 <= nr_employed < 5191.0171</td><td>7773</td><td>0.18872002</td></tr>
<tr><td>5191.0171 <= nr_employed</td><td>19917</td><td>0.48356317</td></tr>
</table></td><td>nr_employed_cat</td></tr>
</table>

4.2. Categorical Variables

There are no missing categorical values in this data set. Thus, no imputation was necessary. Figure A through Figure Q in the Appendix show categorical variable frequency tables and mosaic plots by the target response (y). The continuous variables campaign, previous, emp_var_rate, cons_price_idx, cons_conf_idx, euribor3m, and nr_employed were coded as categorical variables and will be analyzed. A review of the plots reveals the following:

Categorical Variable	Figure	Figure Notes
job	A	Level 'unknown' of variable job, 1.6% of total observations, will be rolled into the largest category of 'admin'. It would be advantageous to reduce the number of levels, but there does not seem to be commonality among the levels that the various levels could be collapsed into. Per the Mosaic plot, job does appear to impact the response.
marital	B	Level 'unknown' of variable marital, .45% of total observations, will be rolled into the largest category of 'married'. Per the Mosaic plot, marital does appear to impact the response.
education	C	Level 'unknown' of variable education, 9.46% of total observations, will be rolled into the largest category of 'university.degree'. Also, since there are only 18 observations total for level 'illiterate', 'illiterate' level will be dropped from the observations. 18 observations is not enough to make a proper inference. Per the Mosaic plot, education does appear to impact the response.
default	D	Level 'yes' of variable default, .01% of total observations, will be deleted. 3 observations is not enough to make a proper inference. Per the Mosaic plot, default does appear to impact the response.
housing	E	Level 'unknown' of variable housing, 4.73% of total observations, will be rolled into the largest category of 'yes'. Per the Mosaic plot, housing does not appear to impact the response.
loan	F	Level 'unknown' of variable loan, 4.73% of total observations, will be rolled into the largest category of 'no'. Per the Mosaic plot, loan does not appear to impact the response.
contact, month, campaign_cat, previous_cat, poutcome, emp_var_rate_cat, cons_price_idx_cat, cons_conf_idx_cat, euribor3m, nr_employed_cat	G, H, J, K, L, M, N, O, P, Q	Per the Mosaic plot(s), these variables appear to impact the response.
day_of_week	I	Per the Mosaic plot(s), these variables do not appear to impact the response.

From Figure R, the target response (y) has 88.73% no responses and 11.27% yes responses. In modeling the data it would be best to use all of the yes response and an equal number of no responses. This is to ensure that we adequately discover what variables in the model make a difference between 'yes' and 'no' responses. Too many 'no' observations would push a model to predict that 'no' variables make a difference.

The final list of variables used in the modeling techniques is listed in Table 2 of the Appendix.

5. Analysis

Since the target response (y) variable is binary in nature, a logistic regression model and binary decision tree model was used to analyze the data. Since only 11.27% of the data has 'yes' responses for y, the models will be built on a data set composed of all the 'yes' responses (4,636) and a random sample of 4,636 'no' responses. The 'no' response random sample was stratified by job, marital, education, default, housing, loan, and month to ensure the sample mirrors the raw data as much as possible. This method gives the model greater power to detect what variables impact target response (y).

Next, the data was split 50/50 into training and validation data sets. The training and validation data sets were stratified by education and job since these 2 variables have the most categories. Figure 20 and 21 in the Appendix show that the categorical variables in the training and validation data sets have at least 29 observations in each level for both response levels. This affirms that adequate inferences can be made off of either data set. Figure 22 and 23 in the Appendix show that continuous variable age has significant representation across the range of age in the training and validation data sets.

A model was created using logistic regression since target response (y) is a Bernoulli response. In logistic regression, the odds (P[Success]/P[Failure]) are modeled instead of the response itself. The Logistic Regression model is a Generalized Linear Model that assumes the explanatory variables are linear predictors and the logit function (P[Success]/P[Failure]) as the link function that relates the mean of the response y to the linear predictors. For the Linear Regression Model, we will assume the observations are independent from each other, and that the response variable (y) is binomially distributed independent variable. Independence among responses is a reasonable assumption since this data is gathered from individual clients. Residual and influential diagnostics will then be analyzed to ensure the model does not violate assumptions and that there are no extremely influential individual data points.

A Binary Decision Tree model will also be built to determine which variables have the highest influence. Entropy will be used to make splitting decisions, and SAS cost-complexity pruning will be used to prune the tree to the optimum number of leaves. The ROC curves and fit statistics of the logistic regression model and decision tree will be compared to see which model fits the data best. Inferences will then be made from the better model.

6. Results/Generalization

6.1. Logistic Regression Model

In building our Logistical Regression Model, we performed the Likelihood Ratio Test for each individual explanatory variable at the 0.05 alpha level on the training data set. The results are shown in Appendix Figure 24. Every variable is significant per the likelihood ratio test. (The null hypothesis for each individual explanatory variable is that the different levels of each parameter have no impact on the log odds.)

Running the Logistic Regression Model with all significant variables, we found that variables age, housing, campaign_cat, and cons_price_idx_cat were not significant at the alpha = 0.05 level in the presence of other variables. See figure 25 in the Appendix for the Likelihood Ratio Statistics for Type 3 Analysis of Effects. Dropping these insignificant variables and running the Logistic Regression Model again, we find all remaining variables are significant per Type 3 Analysis of Effects. This is our final logistic regression model with 14 significant main effects. See Figure 25 for the significance of the variables.

The Standarized Pearson Residuals of the logistic regression model look normally distributed. However, there are 62 observations that are greater than 3 standard deviations from 0 (see Figure 26). From the graphs in Figure 26, the observations with high residuals occur when the predicted value is close to 0 and when the predicted value is close to 1. These high residuals occur when a predicted value of 0 coincides with an actual observation of 1 and when a predicted value of 1 coincides with an actual observation of 0. This is not unexpected behavior for the residuals of a logistic regression model. Since this occurred in only 62/4,169 = 1.5% observations, this was not of great concern. The residuals are in good shape.

The Standardized Dfbeta plots in Figure 27 also show that none of the Standardized Dfbetas for any of the levels of the explanatory variables exert great influence on the value of the model coefficient parameter estimates. No Dfbeta is greater than 0.6 standard deviations for any variable level for any observation. This further indicates that the large residuals are not having a large impact on the model. (Dfbetas describe the estimated effect on model parameters when removing observations from the data set.)

6.2. Binary Decision Tree Model

A binary decision tree was built in SAS using proc hpsplit. A maximum depth of 30 was given to allow the tree to grow until completion using entropy (SAS default). The data set was partitioned into 50 percent training data and 50 percent validation data via a random sample without stratification. (proc hpsplit does not allow stratification.) A 50/50 split between training and validation was chosen so that the same scheme was used as in logistic regression. On the first run of the decision tree, SAS chose 63 leaves as the optimal number of leaves with a Validation Misclassification rate of 0.26 and a tree depth of 11. (See Appendix Figure 32). The second run limited the depth to 11 and produced the same number of leave and Validation Misclassification rate. On the 3rd run, the depth was reduced to 9 and produced a Validation Misclassification rate of 0.26, a SAS recommended optimal leaves of 45, and a tree depth of 9. On a fourth run, the depth was reduced to 7 and produced a Validation Misclassification rate of 0.27, a SAS recommended optimal leaves of 13, and a tree depth of 6. The final run in Figure 33 was run with a maxdepth of 30 and pruned to 13 leave. This yielded a Validation Misclassification rate of 0.27 and a tree depth of 6 with 13 leaves. This final decision tree is the best model of all 4 models since its misclassification rate is very close to the most complex model but is a very simple model. There is not much of a difference between a misclassification rate of 0.26 and 0.27, but there is a large difference between 63 and 13 leaves to explain the model.

6.3. Comparison of Decision Tree and Logistic Regression Model

The ROC Curve and Misclassification rate for the Decision Tree and the logistic Regression Model Validation Data Sets are shown in Figure 31. The Logistic Regression Model is the better model with a higher Area Under the Curve of 0.8375 compared to an AUC of 0.78. The Misclassification Rate of the Logistic Regression Model is also better at 0.242 as compared to 0.2734. The Logistic Regression Model is the better model to use to draw inferences.

6.4. Inferences from the Logistic Regression Model

Per the ROC Curves in Figure 28, the model has very good predictive power. The area under the ROC Curve for the training data is 0.8383. The area under the ROC Curve for the validation data is 0.8375. ROC curve area can be from 0.5 to 1, so our model has very good predictive power. 0.5 is worthless. 1 is perfect.

Any change in the baseline levels of any of the model parameters changes the log odds of the model. Exponentiation of the log odds gives the change in odds P[yes]/P[no]. Odds Ratios for all variable levels that do not contain 1 in their 95% confidence interval are listed in Table 29 by variable category. Odds Ratios for all variable levels that contain 1 in their 95% confidence interval are listed in Table 30 by variable category. Variable levels that contain 1 in their 95% confidence interval do not impact the response.

The 3 Client Data variables that have the largest impact on odds ratios are job, default, and loan. The baseline level for job is ‘unemployed’. All but 3 job categories that varied from unemployed were 2.5 to 5 times more likely to make a term deposit purchase than an unemployed person. Housemaids, entrepreneurs, and self-employed persons were no more likely to make a purchase than an unemployed person. For Client Data variable default, a person who reports they have no credit in default is 3 times more likely to purchase a term deposit than a person who reports that it is unknown whether they have credit in default. For Client Data variable loan, a person who reports that they don’t have a personal loan is 3 times more likely to purchase a term deposit that a person that reports they have a personal loan.

All Last Contact Info variables have an impact on term deposit purchases. Month is the most important with all months making an impact except for September(baseline), October, and December. The most important month was November with clients 5 times more likely to make a purchase in November as September. Contact was the next most important

Last Contact Info variable. Clients contacted by Cell Phone are 2.4 times more likely to make a term deposit purchase than clients contacted by telephone. Wednesday was the best day_of_week level to contact clients. Clients were about 1.5 times more likely to make a purchase on Wednesday than any other day of the week.

Other category variables poutcome and previous_cat were also important. A client who previously purchased a term deposit from another campaign was 5.7 times more likely to purchase another term deposit. A client who had been contacted before was 3.9 times more likely to purchase.

The Social and Economic variables were important based on their levels. The highest level of euribor3m_cat makes some huge differences whether clients purchase term deposits or not. At the highest level, clients are 24 to 28 times more likely to purchase term deposits than at the lowest 2 levels. Social and Economic variables nr_employed_cat and emp_var_rate_cat make the greatest impact at their lowest levels. At their lowest levels a client is from 6 to 17 times more likely to make a term deposit purchase. Variable cons_conf_idx_cat at its highest level impacts term deposit purchases as 2 times more likely than at its lower levels.

7. Suggestions for Future Studies

The Social and Economic variables euribor3m_cat and nr_employed_cat contained variable levels that have very large odds ratios (17,24,28) and variable levels that don't impact the odds. It would be an interesting study to determine why there is such a different impact in the levels of these social & economic variables.

8. Conclusion

The objective of this study was to determine which variables have the highest influence on whether a client purchases a term deposit or not. A second objective was to determine the levels of those variables that produce the most term deposit purchases. Through our study, we discovered 14 variables that impact the decision of clients to purchase term deposits. The 3 Client Data variables that have the largest impact on odds ratios are job, default, and loan. The 2 Last Contact Info variables that have the largest impact on odds ratios are month and contact. The Other variables that have the largest impact on odds ratios are poutcome and previous_cat. The Social and Economic variables that have the largest impact on odds ratios are euribor3m_cat and nr_employed_cat. Knowing the variables that provide the highest odds of success are important to a bank. A bank can use these variables to target Clients that would most likely make term deposit purchases.

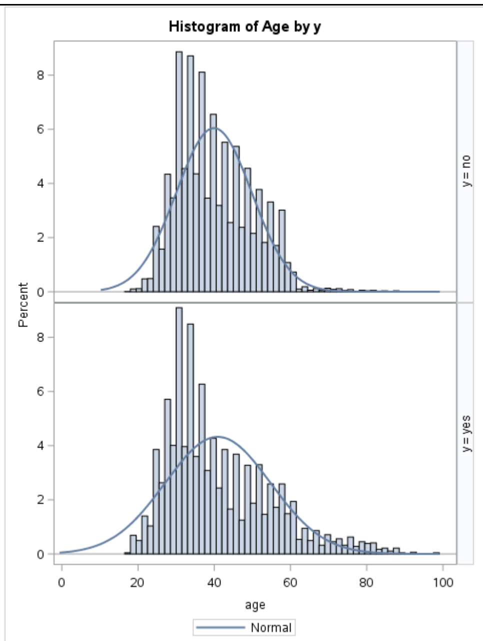
9. Appendix: Tables, Graphs, and SAS Code

9.1. Table 1 – Raw Data Set Variables

Variable	Variable Category	Description	Variable Type
age	Client Data	Clients age at time of call	Continuous
job	Client Data	Clients type of job - 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')	Categorical
marital	Client Data	Clients Marital Status at time of call - 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed	Categorical
education	Client Data	Clients educational background at time of call - 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'	Categorical
default	Client Data	Does client have credit in default? - 'no','yes','unknown'	Categorical
housing	Client Data	Does client have a house loan? - 'no','yes','unknown'	Categorical
loan	Client Data	Does client have a personal loan? - 'no','yes','unknown'	Categorical
contact	Last Contact Info	Communication type with client – 'cellular','telephone'	Categorical
month	Last Contact Info	Last contact month of year with client - 'jan', 'feb', 'mar', ..., 'nov', 'dec'	Categorical
day_of_week	Last Contact Info	Last contact day of week with client - 'mon','tue','wed','thu','fri'	Categorical
duration	Last Contact Info	Last contact duration, in seconds to Client. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.	Continuous
campaign	Other	Number of contacts performed during this campaign for this client (includes last contact)	Continuous
pdays	Other	Number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)	Continuous
previous	Other	Number of contacts performed before this campaign and for this client	Continuous
poutcome	Other	Outcome of the previous marketing campaign - 'failure','nonexistent','success'	Categorical
emp_var_rate	Social & Economic	Employment Variation Rate - Quarterly indicator	Continuous
cons_price_idx	Social & Economic	Consumer Price Index – Monthly indicator; Monthly Consumer Price Index or CPI measures changes in the prices paid by consumers for a basket of goods and services each Month.	Continuous
cons_conf_idx	Social & Economic	Consumer Confidence Index – Monthly indicator; In Portugal, the consumer confidence index is based on interviews with consumers about their perceptions of the country's current and future economic situation and their tendencies to purchase. It is estimated using the difference between the share of positive evaluation responses and negative evaluation responses, but do not include the share of neutral responses.	Continuous
nr_employed	Social & Economic	Euribor 3 Month Rate – Daily indicator; Euribor is short for Euro Interbank Offered Rate. The Euribor rates are based on the average interest rates at which a large panel of European banks borrow funds from one another that mature after 3 months.	Continuous
nr.employed	Social & Economic	Number of Employees – Quarterly indicator; Number of employed persons for a quarter.	Continuous
y	Target/Response	Has the client subscribed a term deposit? - 'yes','no'	Categorical/ Binary

9.2. Continuous Variable Plots by y

Figure 1 – Histogram and Boxplot of age by y



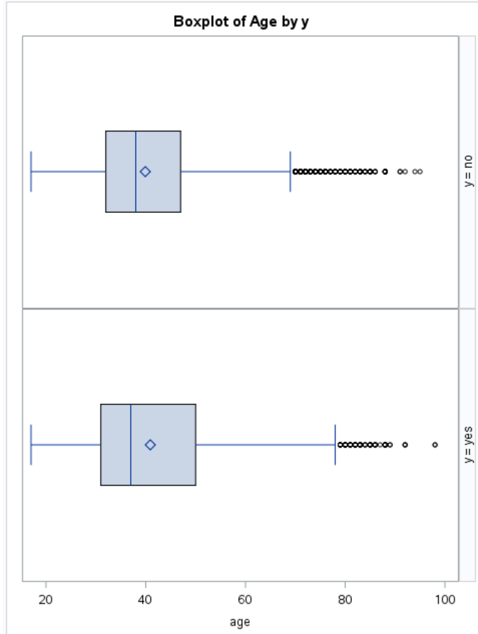
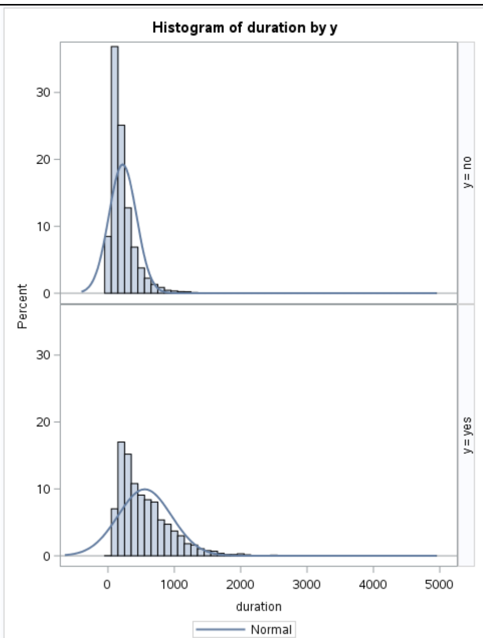


Figure 2 – Histogram and Boxplot of duration by y



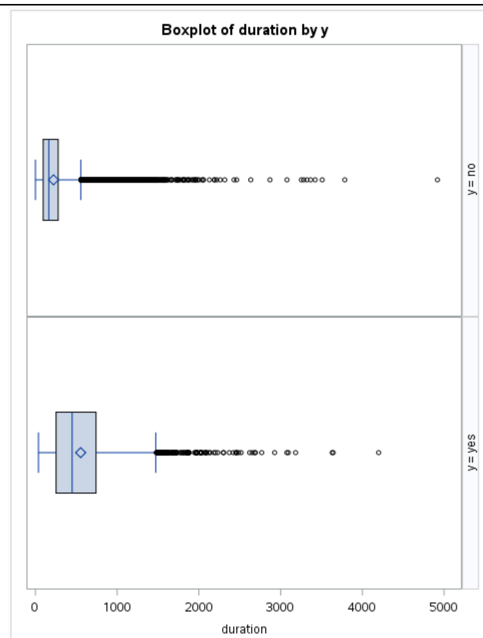
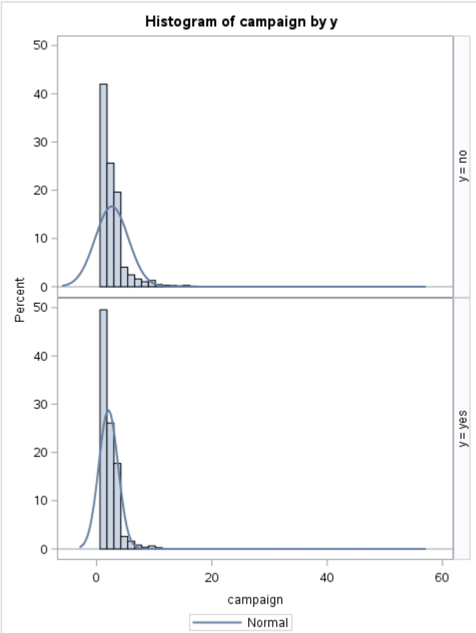


Figure 3 – Histogram, Boxplot, and Frequency Table of campaign by y



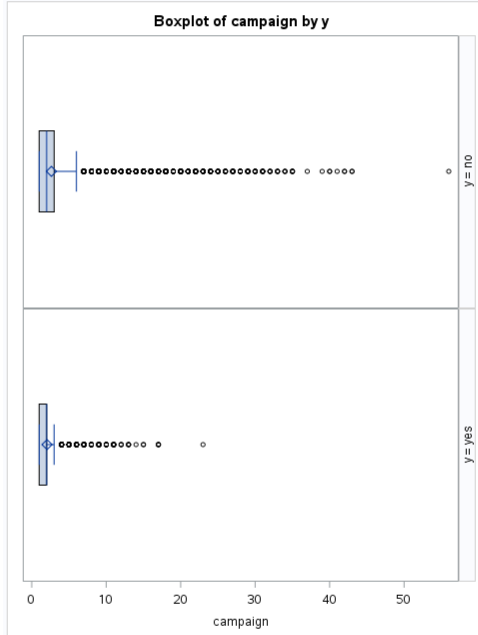


Figure 3 Frequency Tables - Cumulative Percentages of campaign by y

y=no

campaign	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15342	41.98	15342	41.98
2	9359	25.61	24701	67.59
3	4767	13.04	29468	80.63
4	2402	6.57	31870	87.20
5	1479	4.05	33349	91.25
6	904	2.47	34253	93.72
7	591	1.62	34844	95.34
8	383	1.05	35227	96.39
9	266	0.73	35493	97.11
10	213	0.58	35706	97.70

y=yes

campaign	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2300	49.57	2300	49.57
2	1211	26.10	3511	75.67
3	574	12.37	4085	88.04
4	249	5.37	4334	93.41
5	120	2.59	4454	95.99
6	75	1.62	4529	97.61
7	38	0.82	4567	98.43
8	17	0.37	4584	98.79
9	17	0.37	4601	99.16
10	12	0.26	4613	99.42

Figure 4 – Histogram and Boxplot of pdays by y

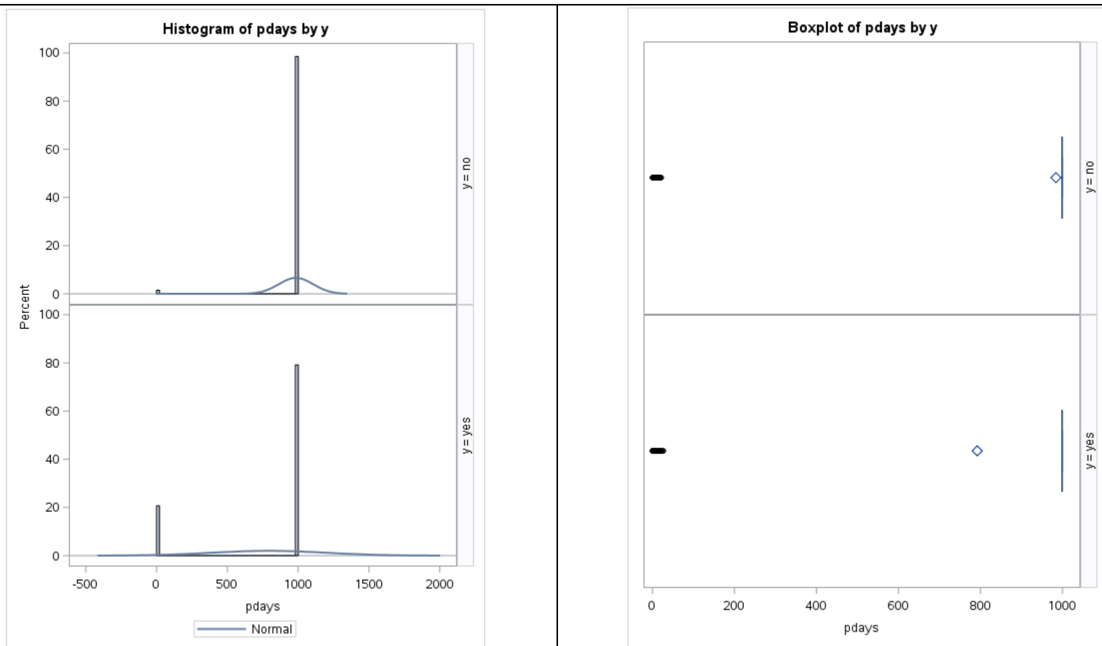


Figure 4 Frequency Tables - Cumulative Percentages of pdays by y

y=no

pdays	Frequency	Percent	Cumulative Frequency	Cumulative Percent
999	36000	98.50	36000	98.50
3	141	0.39	36141	98.89
6	123	0.34	36264	99.22
4	55	0.15	36319	99.37
12	32	0.09	36351	99.46

y=yes

pdays	Frequency	Percent	Cumulative Frequency	Cumulative Percent
999	3673	79.16	3673	79.16
3	298	6.42	3971	85.58
6	289	6.23	4260	91.81
4	63	1.36	4323	93.17
7	40	0.86	4363	94.03

Figure 5 – Histogram and Boxplot of previous by y

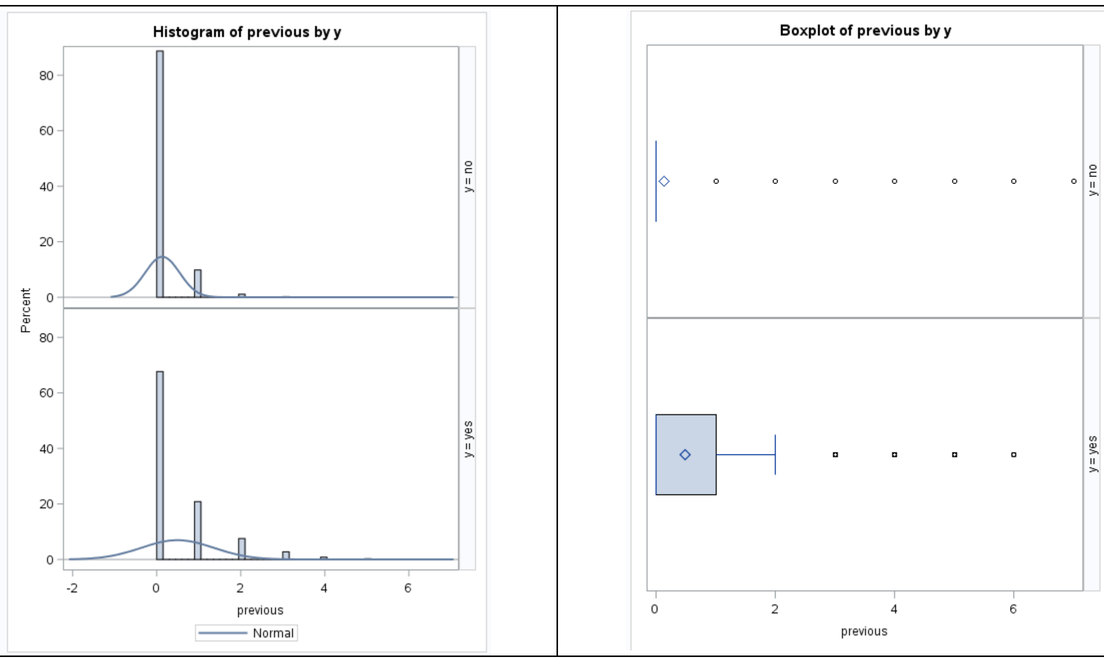


Figure 5 Frequency Tables - Cumulative Percentages of previous by y

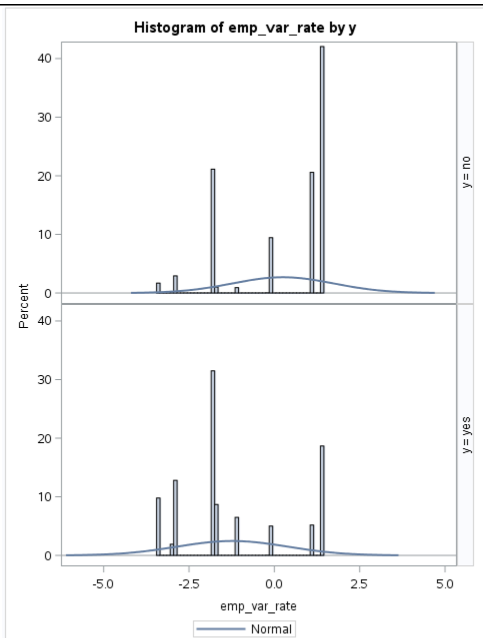
y=no

previous	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	32422	88.71	32422	88.71
1	3594	9.83	36016	98.54
2	404	1.11	36420	99.65
3	88	0.24	36508	99.89
4	32	0.09	36540	99.98
5	5	0.01	36545	99.99
6	2	0.01	36547	100.00
7	1	0.00	36548	100.00

y=yes

previous	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3141	67.69	3141	67.69
1	967	20.84	4108	88.53
2	350	7.54	4458	96.08
3	128	2.76	4586	98.84
4	38	0.82	4624	99.66
5	13	0.28	4637	99.94
6	3	0.06	4640	100.00

Figure 6 – Histogram and Boxplot of emp_var_rate by y



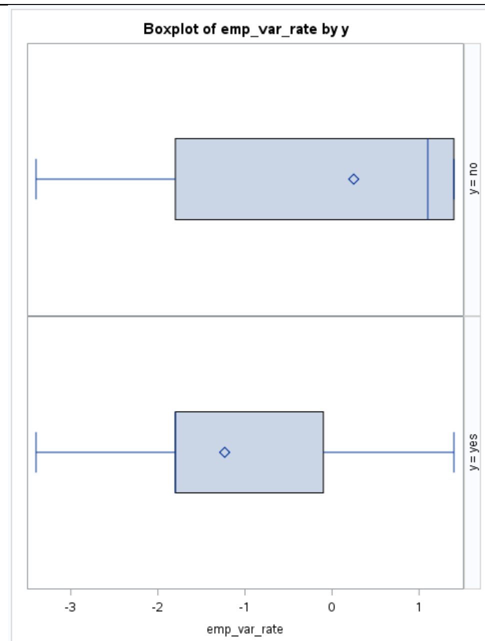


Figure 6 - Grouped Bar Chart of emp_var_rate by y

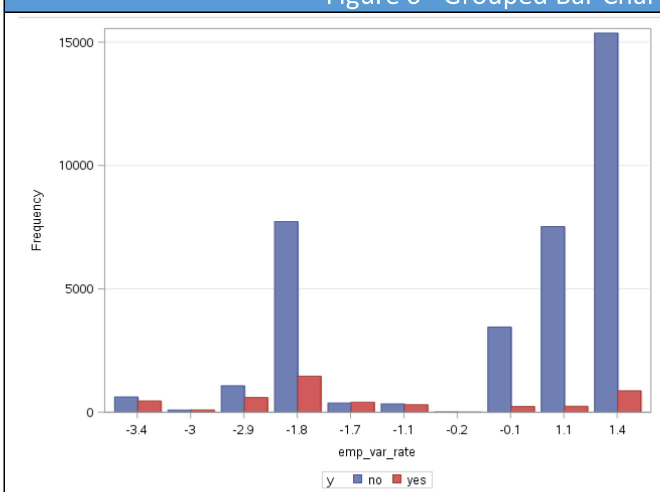
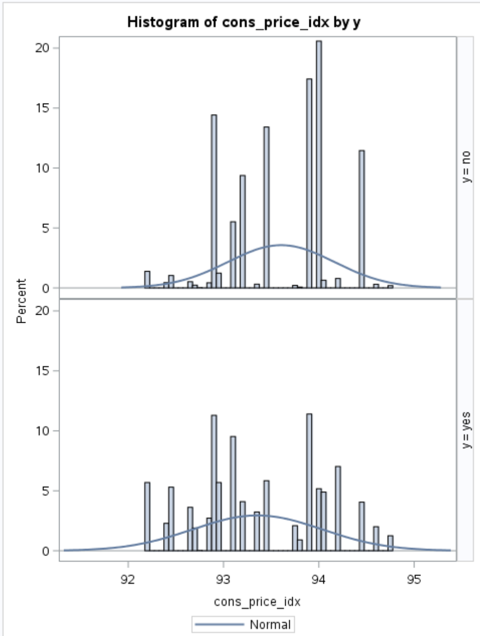


Figure 6 Frequency Tables - Cumulative Percentages of emp_var_rate by y

y=no				
emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<= -1.8	9493	25.97	9493	25.97
[-1.8,-0.1)	4164	11.39	13657	37.37
>-0.1	22891	62.63	36548	100.00

y=yes				
emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<= -1.8	2597	55.97	2597	55.97
[-1.8,-0.1)	937	20.19	3534	76.16
>-0.1	1106	23.84	4640	100.00

Figure 7 – Histogram and Boxplot of cons_price_idx by y



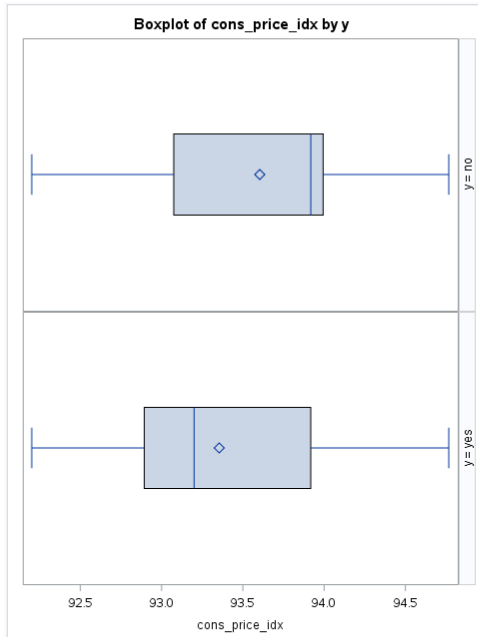


Figure 7 – Bucket Binning of cons_price_idx by y

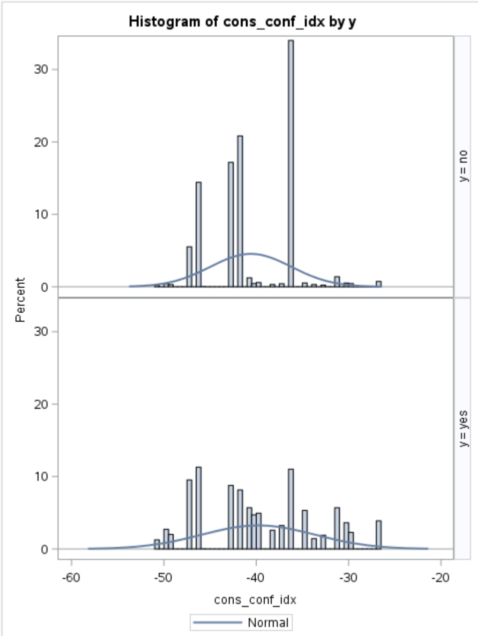
Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
cons_price_idx	BIN_cons_price_idx	cons_price_idx < 93.056333333	8992	0.21831601
		93.056333333 <= cons_price_idx < 93.911666667	11966	0.29052151
		93.911666667 <= cons_price_idx	20230	0.49116247

Figure 7 Frequency Tables - Cumulative Percentages of cons_price_idx_cat by y

y=no				
cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<93.06	7205	19.71	7205	19.71
[93.06,93.91]	10652	29.15	17857	48.86
>93.91	18691	51.14	36548	100.00

y=yes				
cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<93.06	1787	38.51	1787	38.51
[93.06,93.91]	1314	28.32	3101	66.83
>93.91	1539	33.17	4640	100.00

Figure 8 – Histogram and Boxplot of cons_conf_idx by y



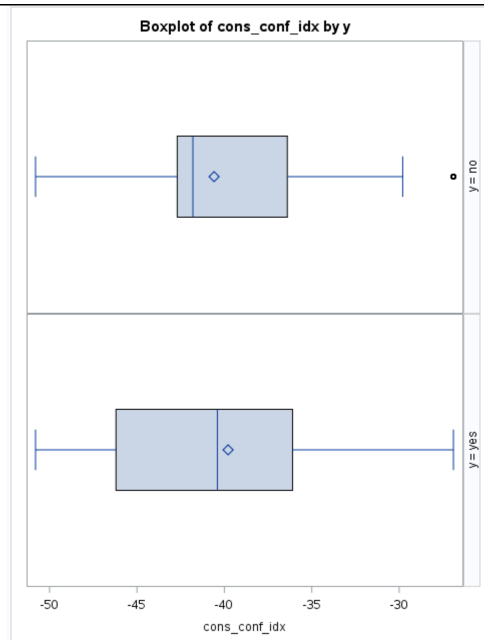


Figure 8 – Quantile Binning of cons_conf_idx by y

Mapping

Variable	Binned Variable	Range	Frequency	Proportion
cons_conf_idx	BIN_cons_conf_idx	cons_conf_idx < -46.19925	8866	0.21525687
		-46.19925 <= cons_conf_idx < -41.99763	10311	0.25033990
		-41.99763 <= cons_conf_idx < -39.99959	5679	0.13787997
		-39.99959 <= cons_conf_idx < -36.39786	8528	0.20705060
		-36.39786 <= cons_conf_idx	7804	0.18947266

Figure 8 Frequency Tables - Cumulative Percentages of cons_conf_idx_cat by y

The FREQ Procedure

y=no

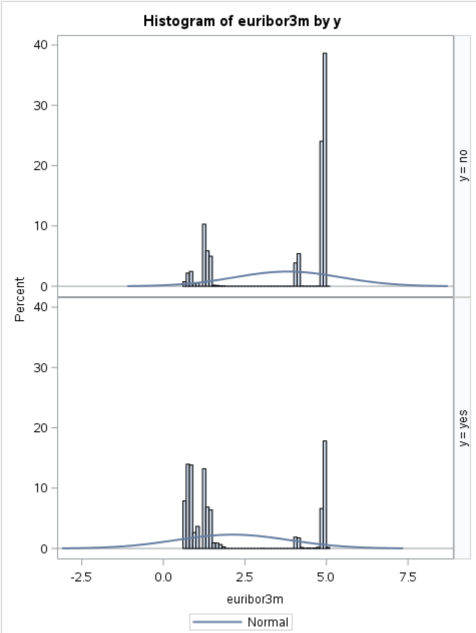
cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<-46.2	7623	20.86	7623	20.86
[-46.2,-42)	9713	26.58	17336	47.43
[-42,-40)	4887	13.37	22223	60.80
[-40,-36.4)	7911	21.65	30134	82.45
>=-36.4	6414	17.55	36548	100.00

The FREQ Procedure

y=yes

cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<-46.2	1243	26.79	1243	26.79
[-46.2,-42)	598	12.89	1841	39.68
[-42,-40)	792	17.07	2633	56.75
[-40,-36.4)	617	13.30	3250	70.04
>=-36.4	1390	29.96	4640	100.00

Figure 9 – Histogram and Boxplot of euribor3m by y



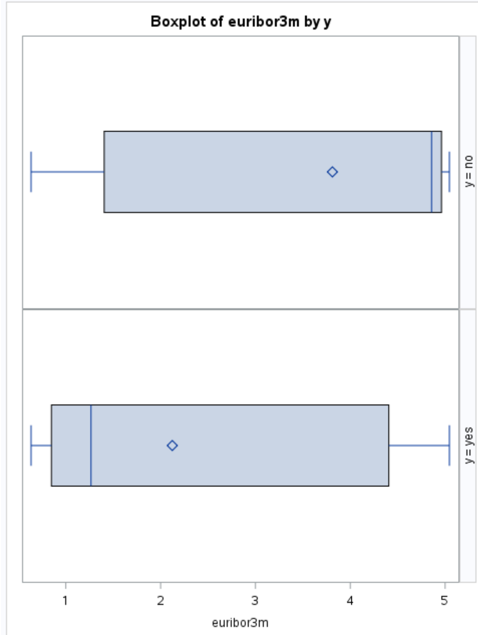


Figure 9 – Quantile Binning of euribor3m by y

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
euribor3m	BIN_euribor3m	euribor3m < 1.2991788	8636	0.20967272
		1.2991788 <= euribor3m < 4.1910304	8430	0.20467126
		4.1910304 <= euribor3m < 4.864149	8446	0.20505973
		4.864149 <= euribor3m < 4.9620732	8498	0.20632223
		4.9620732 <= euribor3m	7178	0.17427406

Figure 9 Frequency Tables - Cumulative Percentages of euribor3m_cat by y

y=no

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<1.3	6075	16.62	6075	16.62
[1.3,4.19)	7521	20.58	13596	37.20
[4.19,4.86)	8140	22.27	21736	59.47
[4.86,4.96)	8029	21.97	29765	81.44
>=4.96	6783	18.56	36548	100.00

y=yes

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<1.3	2561	55.19	2561	55.19
[1.3,4.19)	909	19.59	3470	74.78
[4.19,4.86)	306	6.59	3776	81.38
[4.86,4.96)	469	10.11	4245	91.49
>=4.96	395	8.51	4640	100.00

Figure 10 – Histogram and Boxplot of nr_employed by y

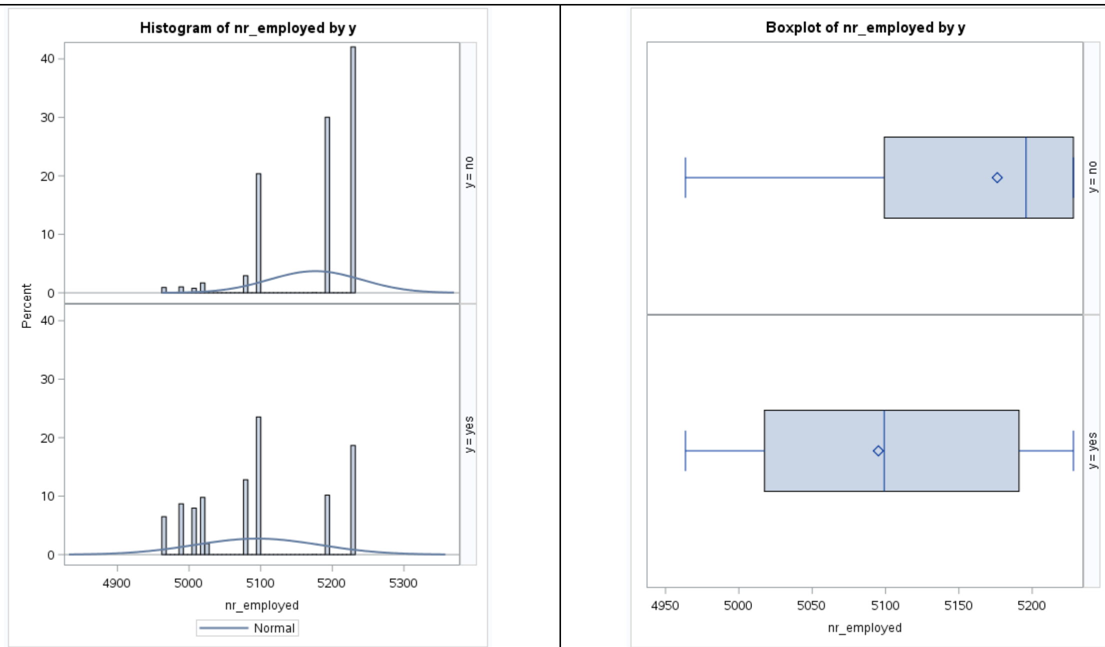


Figure 10 – Quantile Binning of nr_employed by y

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
nr_employed	BIN_nr_employed	nr_employed < 5099.10335	13498	0.32771681
		5099.10335 <= nr_employed < 5191.0171	7773	0.18872002
		5191.0171 <= nr_employed	19917	0.48356317

Figure 10 Frequency Tables - Cumulative Percentages of nr_employed_cat by y

y=no

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<5099.1	10197	27.90	10197	27.90
[5099.1,5191.02)	7532	20.61	17729	48.51
>5191.02	18819	51.49	36548	100.00

y=yes

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<5099.1	3301	71.14	3301	71.14
[5099.1,5191.02)	241	5.19	3542	76.34
>5191.02	1098	23.66	4640	100.00

9.3. Discrete Variable Plots and Tables

Figure A – Contingency Table and Mosaic plot of job by y

Frequency Row Pct Col Pct

Table of job by y

job(job)	y(y) no	y(y) yes	Total
admin.	9070 87.03 24.82	1352 12.97 29.14	10422
blue-collar	8616 93.11 23.57	638 6.89 13.75	9254
entrepreneur	1332 91.48 3.64	124 8.52 2.67	1456
housemaid	954 90.00 2.61	106 10.00 2.28	1060
management	2596 88.78 7.10	328 11.22 7.07	2924
retired	1286 74.77 3.52	434 25.23 9.35	1720
self-employed	1272 89.51 3.48	149 10.49 3.21	1421
services	3646 91.86 9.98	323 8.14 6.96	3969
student	600 68.57 1.64	275 31.43 5.93	875
technician	6013 89.17 16.45	730 10.83 15.73	6743
unemployed	870 85.80 2.38	144 14.20 3.10	1014
unknown	293 88.79 0.80	37 11.21 0.80	330
Total	36548	4640	41188

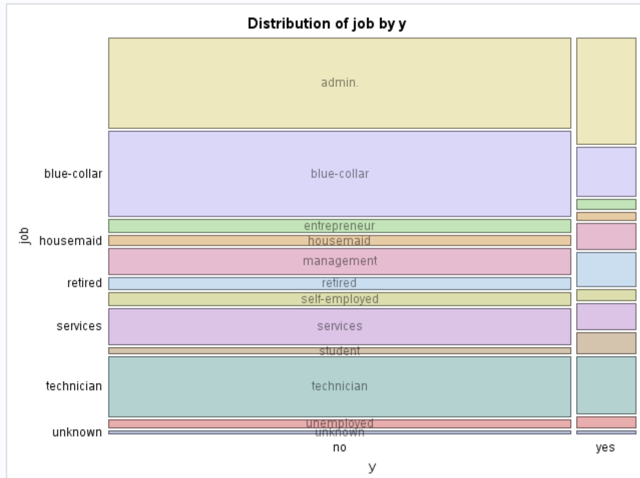


Figure B – Contingency Table and Mosaic plot of marital by y

Frequency Row Pct Col Pct

Table of marital by y

marital(marital)	y(y) no	y(y) yes	Total
divorced	4136 89.68 11.32	476 10.32 10.26	4612
married	22396 89.84 61.28	2532 10.16 54.57	24928
single	9948 86.00 27.22	1620 14.00 34.91	11568
unknown	68 85.00 0.19	12 15.00 0.26	80
Total	36548	4640	41188

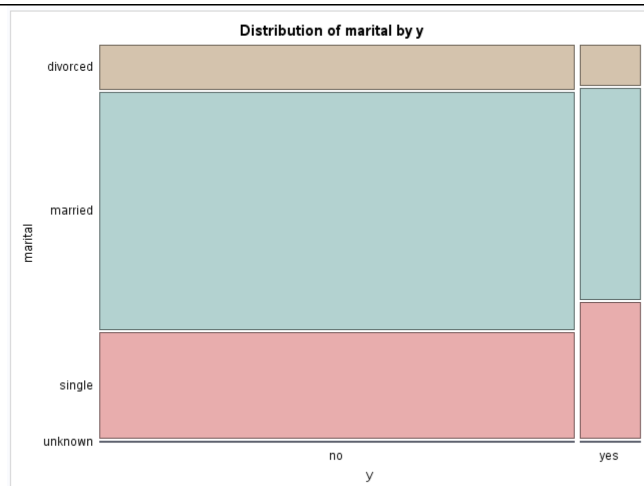


Figure C – Contingency Table and Mosaic plot of education by y

Frequency
Row Pct
Col Pct

Table of education by y

education(education)	y(y) no	y(y) yes	Total
basic.4y	3748 89.75 10.26	428 10.25 9.22	4176
basic.6y	2104 91.80 5.76	188 8.20 4.05	2292
basic.9y	5572 92.18 15.25	473 7.82 10.19	6045
high.school	8484 89.16 23.21	1031 10.84 22.22	9515
illiterate	14 77.78 0.04	4 22.22 0.09	18
professional.course	4648 88.65 12.72	595 11.35 12.82	5243
university.degree	10498 86.28 28.72	1670 13.72 35.99	12168
unknown	1480 85.50 4.05	251 14.50 5.41	1731
Total	36548	4640	41188

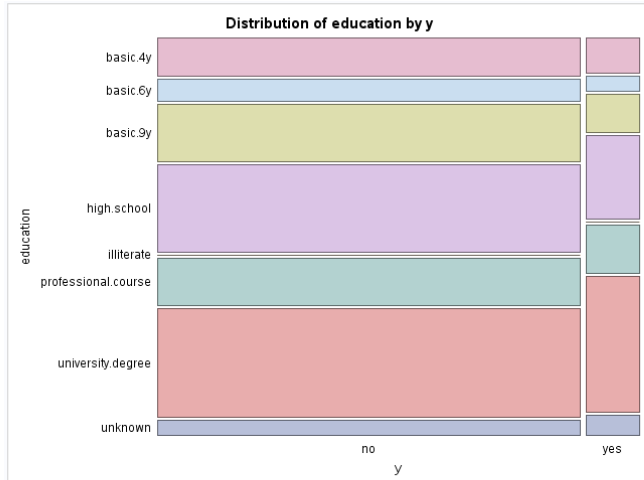


Figure D – Contingency Table and Mosaic plot of default by y

Frequency
Row Pct
Col Pct

Table of default by y

default(default)	y(y) no	y(y) yes	Total
no	28391 87.12 77.68	4197 12.88 90.45	32588
unknown	8154 94.85 22.31	443 5.15 9.55	8597
yes	3 100.00 0.01	0 0.00 0.00	3
Total	36548	4640	41188

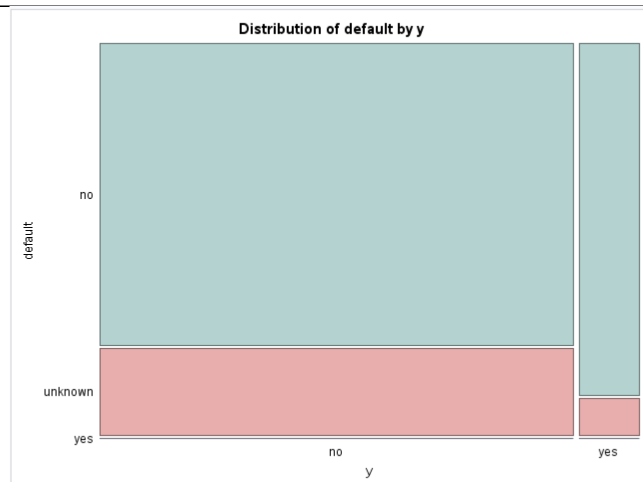


Figure E – Contingency Table and Mosaic plot of housing by y

Frequency
Row Pct
Col Pct

Table of housing by y

housing(housing)	y(y) no	y(y) yes	Total
no	16596 89.12 45.41	2026 10.88 43.66	18622
unknown	883 89.19 2.42	107 10.81 2.31	990
yes	19069 88.38 52.18	2507 11.62 54.03	21576
Total	36548	4640	41188

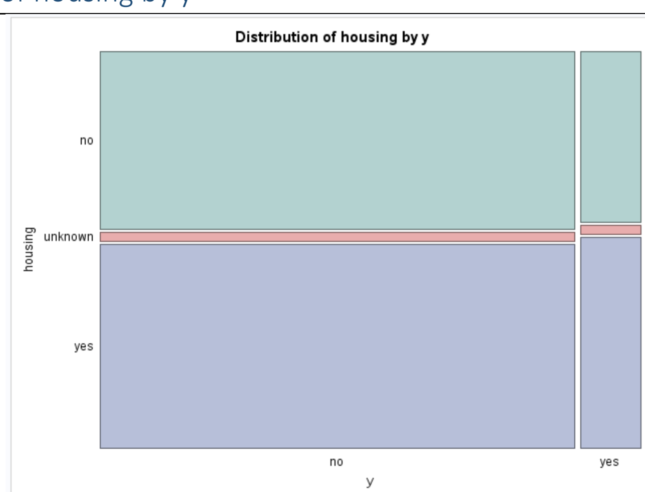


Figure F – Contingency Table and Mosaic plot of loan by y

Frequency Row Pct Col Pct

Table of loan by y

loan(loan)	y(y): no	yes	Total
no	30100 88.66 82.36	3850 11.34 82.97	33950
unknown	883 89.19 2.42	107 10.81 2.31	990
yes	5565 89.07 15.23	683 10.93 14.72	6248
Total	36548	4640	41188

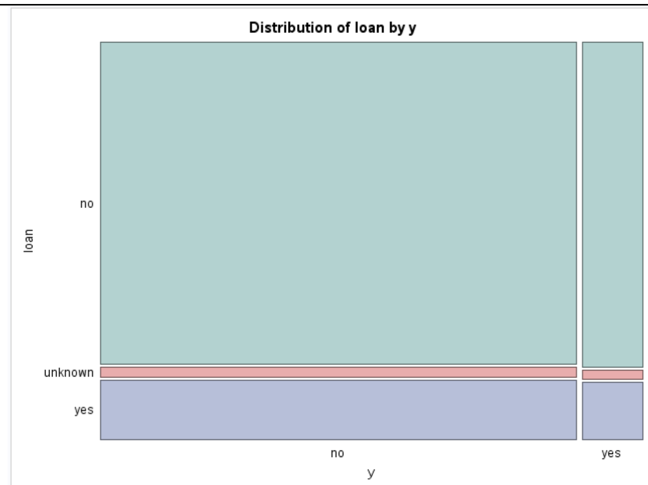


Figure G – Contingency Table and Mosaic plot of contact by y

Frequency Row Pct Col Pct

Table of contact by y

contact(contact)	y(y): no	yes	Total
cellular	22291 85.26 60.99	3853 14.74 83.04	26144
telephone	14257 94.77 39.01	787 5.23 16.96	15044
Total	36548	4640	41188

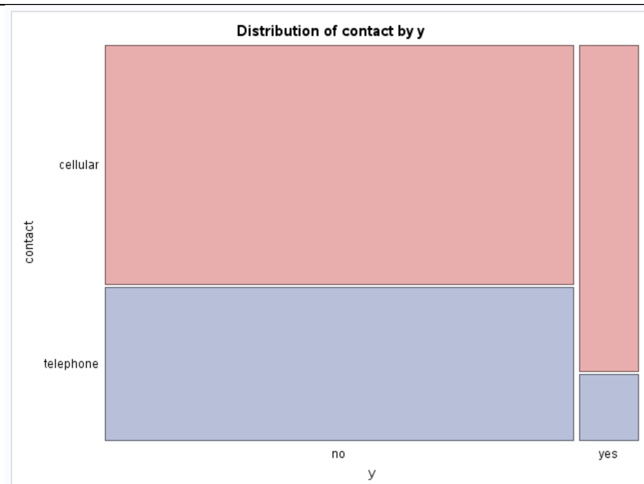


Figure H – Contingency Table and Mosaic plot of month by y

Frequency
Row Pct
Col Pct

Table of month by y			
	y(y)		
month(month)	no	yes	Total
apr	2093 79.52 5.73	539 20.48 11.62	2632
aug	5523 89.40 15.11	655 10.60 14.12	6178
dec	93 51.10 0.25	89 48.90 1.92	182
jul	6525 90.95 17.85	649 9.05 13.99	7174
jun	4759 89.49 13.02	559 10.51 12.05	5318
mar	270 49.45 0.74	276 50.55 5.95	546
may	12883 93.57 35.25	886 6.43 19.09	13769
nov	3685 89.86 10.08	416 10.14 8.97	4101
oct	403 56.13 1.10	315 43.87 6.79	718
sep	314 55.09 0.86	256 44.91 5.52	570
Total	36548	4640	41188

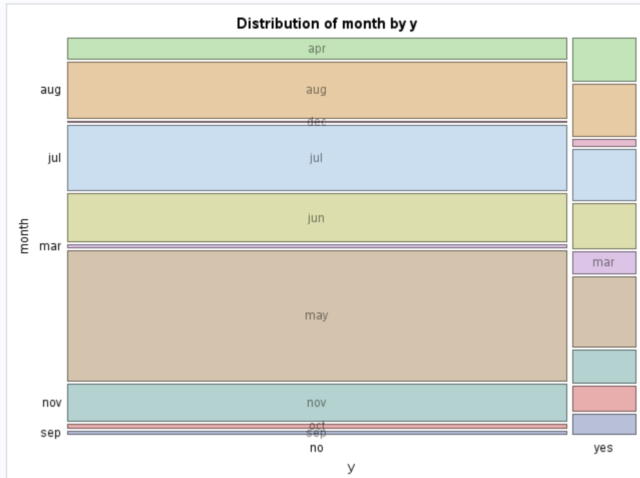


Figure I – Contingency Table and Mosaic plot of day_of_week by y

Frequency
Row Pct
Col Pct

Table of day_of_week by y			
	y(y)		
day_of_week(day_of_week)	no	yes	Total
fri	6981 89.19 19.10	846 10.81 18.23	7827
mon	7667 90.05 20.98	847 9.95 18.25	8514
thu	7578 87.88 20.73	1045 12.12 22.52	8623
tue	7137 88.22 19.53	953 11.78 20.54	8090
wed	7185 88.33 19.66	949 11.67 20.45	8134
Total	36548	4640	41188

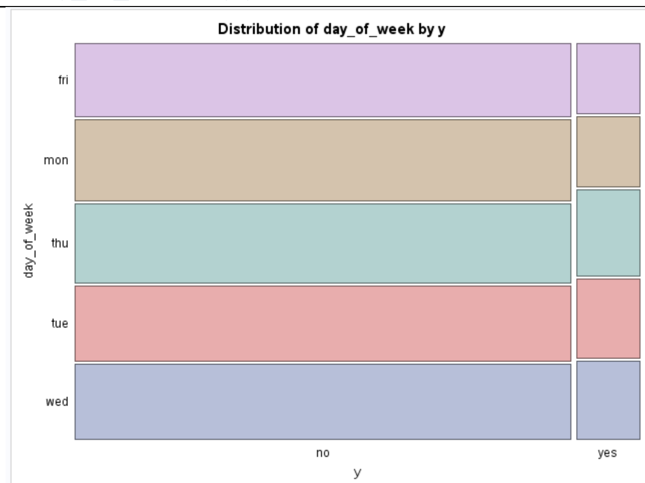


Figure J – Contingency Table and Mosaic plot of campaign_cat by y

Frequency
Row Pct
Col Pct

Table of campaign_cat by y

campaign_cat	y(y) no	y(y) yes	Total
1	15342 86.96 41.98	2300 13.04 49.57	17642
2	9359 88.54 25.61	1211 11.46 26.10	10570
3	4767 89.25 13.04	574 10.75 12.37	5341
>3	7080 92.73 19.37	555 7.27 11.96	7635
Total	36548	4640	41188

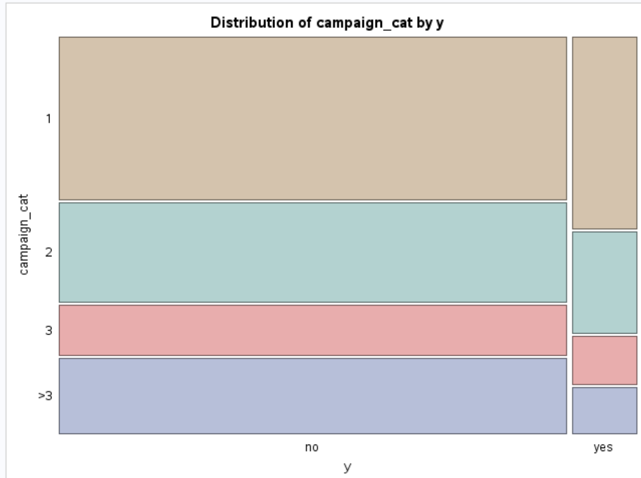


Figure K – Contingency Table and Mosaic plot of previous_cat by y

Frequency
Row Pct
Col Pct

Table of previous_cat by y

previous_cat	y(y) no	y(y) yes	Total
never contacted	32422 91.17 88.71	3141 8.83 67.69	35563
contacted before	4126 73.35 11.29	1499 26.65 32.31	5625
Total	36548	4640	41188

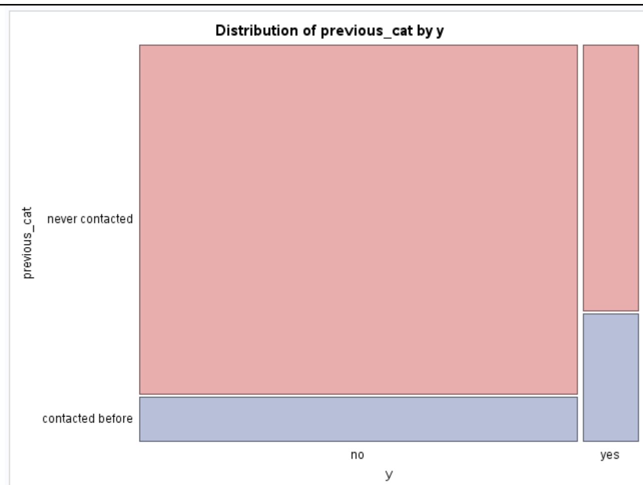


Figure L – Contingency Table and Mosaic plot of poutcome by y

Frequency
Row Pct
Col Pct

Table of poutcome by y

poutcome(poutcome)	y(y) no	y(y) yes	Total
failure	3647 85.77 9.98	605 14.23 13.04	4252
nonexistent	32422 91.17 88.71	3141 8.83 67.69	35563
success	479 34.89 1.31	894 65.11 19.27	1373
Total	36548	4640	41188

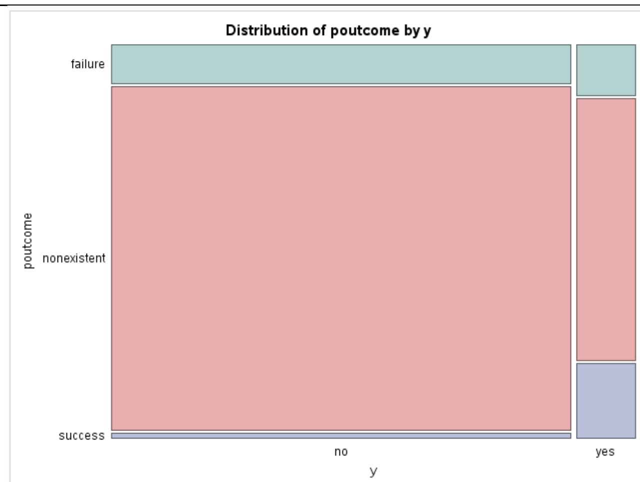


Figure M – Contingency Table and Mosaic plot of emp_var_rate_cat by y

Frequency
Row Pct
Col Pct

Table of emp_var_rate_cat by y

emp_var_rate_cat	y(y) no	y(y) yes	Total
<= -1.8	9493 78.52 25.97	2597 21.48 55.97	12090
[-1.8,-0.1)	4164 81.63 11.39	937 18.37 20.19	5101
>-0.1	22891 95.39 62.63	1106 4.61 23.84	23997
Total	36548	4640	41188

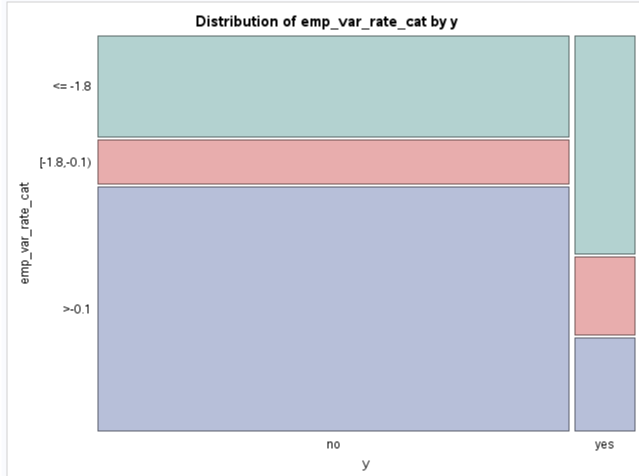


Figure N – Contingency Table and Mosaic plot of cons_price_idx_cat by y

Frequency
Row Pct
Col Pct

Table of cons_price_idx_cat by y

cons_price_idx_cat	y(y) no	y(y) yes	Total
<93.06	7205 80.13 19.71	1787 19.87 38.51	8992
[93.06,93.91]	10652 89.02 29.15	1314 10.98 28.32	11966
>93.91	18691 92.39 51.14	1539 7.61 33.17	20230
Total	36548	4640	41188

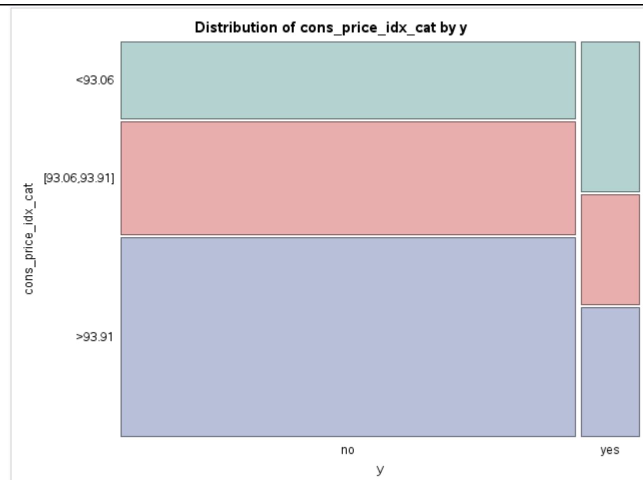


Figure O – Contingency Table and Mosaic plot of cons_conf_idx_cat by y

Frequency
Row Pct
Col Pct

Table of cons_conf_idx_cat by y

cons_conf_idx_cat	y(y) no	y(y) yes	Total
<-46.2	7623 85.98 20.86	1243 14.02 26.79	8866
[-46.2,-42)	9713 94.20 26.58	598 5.80 12.89	10311
[-42,-40)	4887 86.05 13.37	792 13.95 17.07	5679
[-40,-36.4)	7911 92.77 21.65	617 7.23 13.30	8528
>=-36.4	6414 82.19 17.55	1390 17.81 29.96	7804
Total	36548	4640	41188

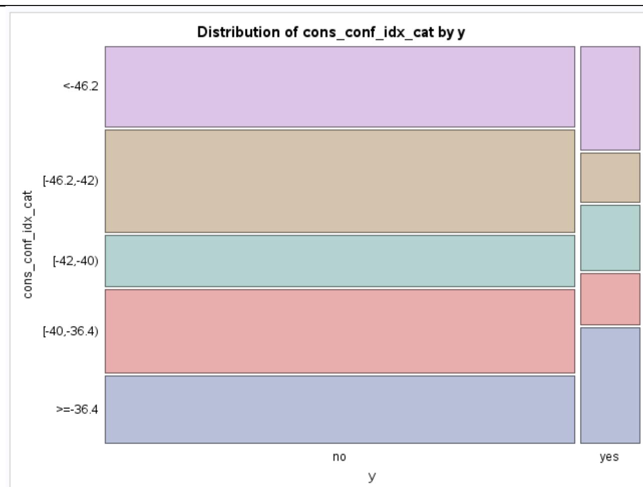


Figure P – Contingency Table and Mosaic plot of euribor3m_cat by y

Frequency
Row Pct
Col Pct

Table of euribor3m_cat by y			
	y(y)		
euribor3m_cat	no	yes	Total
<1.3	6075 70.35 16.62	2561 29.65 55.19	8636
[1.3,4.19)	7521 89.22 20.58	909 10.78 19.59	8430
[4.19,4.86)	8140 96.38 22.27	306 3.62 6.59	8446
[4.86,4.96)	8029 94.48 21.97	469 5.52 10.11	8498
>=4.96	6783 94.50 18.56	395 5.50 8.51	7178
Total	36548	4640	41188

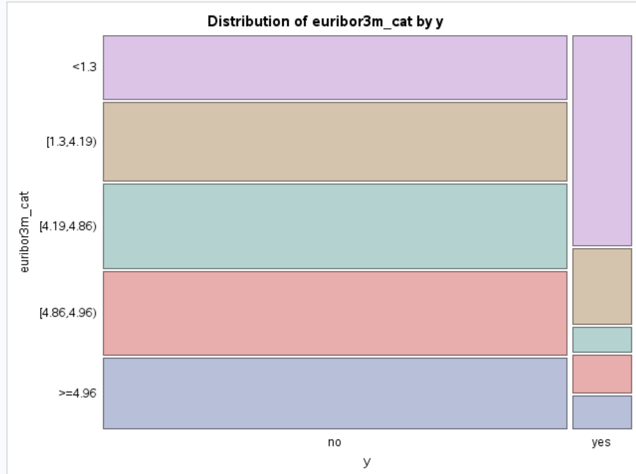


Figure Q – Contingency Table and Mosaic plot of nr_employed_cat by y

Frequency
Row Pct
Col Pct

Table of nr_employed_cat by y			
	y(y)		
nr_employed_cat	no	yes	Total
<5099.1	10197 75.54 27.90	3301 24.46 71.14	13498
[5099.1,5191.02)	7532 96.90 20.61	241 3.10 5.19	7773
>5191.02	18819 94.49 51.49	1098 5.51 23.66	19917
Total	36548	4640	41188

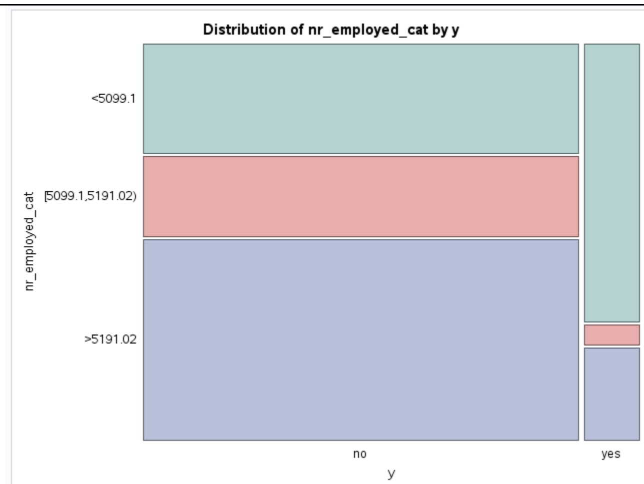
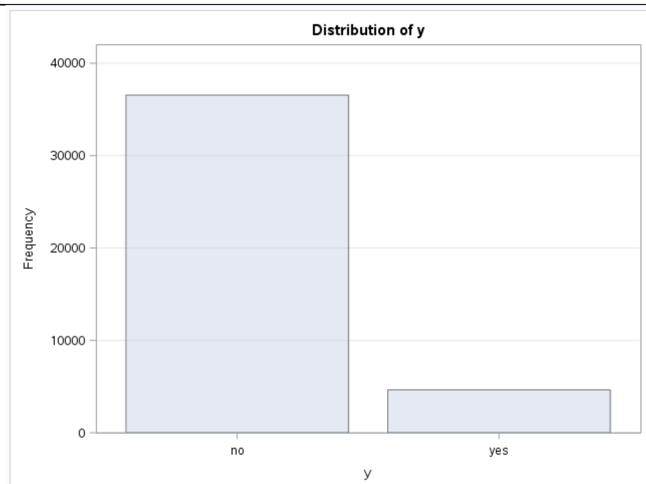


Figure R – Frequency Table and Distribution of y

y		
y	Frequency	Percent
no	36548	88.73
yes	4640	11.27



9.4. Table 2 – Modified Data Set Variables – to be used for modeling

Variable	Variable Category	Description	Variable Type
age	Client Data	Clients age at time of call	Continuous
job	Client Data	Clients type of job - 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed')	Categorical
marital	Client Data	Clients Marital Status at time of call - 'divorced','married','single'; note: 'divorced' means divorced or widowed	Categorical
education	Client Data	Clients educational background at time of call - 'basic.4y','basic.6y','basic.9y','high.school','professional.course','university.degree'	Categorical
default	Client Data	Does client have credit in default? - 'no','unknown'	Categorical
housing	Client Data	Does client have a house loan? - 'no','yes'	Categorical
loan	Client Data	Does client have a personal loan? - 'no','yes'	Categorical
contact	Last Contact Info	Communication type with client – 'cellular','telephone'	Categorical
month	Last Contact Info	Last contact month of year with client - 'jan', 'feb', 'mar', ..., 'nov', 'dec'	Categorical
day_of_week	Last Contact Info	Last contact day of week with client - 'mon','tue','wed','thu','fri'	Categorical
campaign_cat	Other	Number of contacts performed during this campaign for this client (includes last contact). Levels of campaign = {1,2,3,>3}	Categorical
previous_cat	Other	Was the client previously contacted – Levels of previous_cat – 'never contacted', 'contacted before'.	Categorical
poutcome	Other	Outcome of the previous marketing campaign - 'failure','nonexistent','success'	Categorical
emp_var_rate_cat	Social & Economic	Employment Variation Rate - Quarterly indicator. Levels - {-1.8 or less, [-1.8 to -0.1], greater than -0.1}	Categorical
cons_price_idx_cat	Social & Economic	Consumer Price Index – Monthly indicator; Monthly Consumer Price Index or CPI measures changes in the prices paid by consumers for a basket of goods and services each Month. Levels - {<93.06, [93.06,93.91], >93.91}	Categorical
cons_conf_idx_cat	Social & Economic	Consumer Confidence Index – Monthly indicator; In Portugal, the consumer confidence index is based on interviews with consumers about their perceptions of the country's current and future economic situation and their tendencies to purchase. It is estimated using the difference between the share of positive evaluation responses and negative evaluation responses, but do not include the share of neutral responses. Levels - {<-46.2, [-46.2,-41.2), [,-41.2,-40), [-40,-36.4),>-36.4}	Categorical
euribor3m_cat	Social & Economic	Euribor 3 Month Rate – Daily indicator; Euribor is short for Euro Interbank Offered Rate. The Euribor rates are based on the average interest rates at which a large panel of European banks borrow funds from one another that mature after 3 months. Levels - {<1.3, [1.3,4.19), [4.19,4.86), [4.86,4.96),>4.96}	Categorical
nr_employed_cat	Social & Economic	Number of Employees – Quarterly indicator; Number of employed persons for a quarter. Levels - {<5099.1, [5099.1,5191.02), >5191.02}	Categorical
y	Target/Response	Has the client subscribed a term deposit? - 'yes'(1),'no'(0)	Categorical/ Binary

Figure 20 – Frequency Tables for variables in the Training data set by y

y = 0

y=0

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	347	14.76	347	14.76
blue-collar	342	14.55	689	29.31
technician	269	11.44	958	40.75
services	249	10.59	1207	51.34
retired	210	8.93	1417	60.27
management	179	7.61	1596	67.89
unemployed	176	7.49	1772	75.37
entrepreneur	173	7.36	1945	82.73
self-employed	166	7.06	2111	89.79
housemaid	163	6.93	2274	96.72
student	77	3.28	2351	100.00

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
married	1139	48.45	1139	48.45
single	677	28.80	1816	77.24
divorced	535	22.76	2351	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
university.degree	606	25.78	606	25.78
high.school	464	19.74	1070	45.51
basic.9y	369	15.70	1439	61.21
professional.course	369	15.70	1808	76.90
basic.4y	326	13.87	2134	90.77
basic.6y	217	9.23	2351	100.00

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1555	66.14	1555	66.14
unknown	796	33.86	2351	100.00

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	1238	52.66	1238	52.66
no	1113	47.34	2351	100.00

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1588	67.55	1588	67.55
yes	763	32.45	2351	100.00

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	1488	63.29	1488	63.29
telephone	863	36.71	2351	100.00

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
may	530	22.54	530	22.54
jul	414	17.61	944	40.15
jun	372	15.82	1316	55.98
nov	293	12.46	1609	68.44
aug	280	11.91	1889	80.35
apr	213	9.06	2102	89.41
oct	87	3.70	2189	93.11
sep	72	3.06	2261	96.17
mar	57	2.42	2318	98.60
dec	33	1.40	2351	100.00

day_of_week	Frequency	Percent	Cumulative Frequency	Cumulative Percent
mon	512	21.78	512	21.78
thu	484	20.59	996	42.36
fri	456	19.40	1452	61.76
tue	453	19.27	1905	81.03
wed	446	18.97	2351	100.00

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
nonexistent	2011	85.54	2011	85.54
failure	285	12.12	2296	97.66
success	55	2.34	2351	100.00

campaign_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1004	42.71	1004	42.71
2	580	24.67	1584	67.38
4	468	19.91	2052	87.28
3	299	12.72	2351	100.00

previous_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2011	85.54	2011	85.54
1	340	14.46	2351	100.00

emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1292	54.96	1292	54.96
1	674	28.67	1966	83.62
2	385	16.38	2351	100.00

cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1174	49.94	1174	49.94
2	740	31.48	1914	81.41
1	437	18.59	2351	100.00

cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	671	28.54	671	28.54
1	459	19.52	1130	48.06
5	437	18.59	1567	66.65
4	395	16.80	1962	83.45
3	389	16.55	2351	100.00

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	578	24.59	578	24.59
4	570	24.25	1148	48.83
1	474	20.16	1622	68.99
3	377	16.04	1999	85.03
5	352	14.97	2351	100.00

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1230	52.32	1230	52.32
1	777	33.05	2007	85.37
2	344	14.63	2351	100.00

y=1

y=1

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	680	29.55	680	29.55
technician	364	15.82	1044	45.37
blue-collar	320	13.91	1364	59.28
retired	214	9.30	1578	68.58
management	166	7.21	1744	75.79
services	163	7.08	1907	82.88
student	149	6.48	2056	89.35
self-employed	75	3.26	2131	92.61
unemployed	70	3.04	2201	95.65
entrepreneur	54	2.35	2255	98.00
housemaid	46	2.00	2301	100.00

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
married	1264	54.93	1264	54.93
single	797	34.64	2061	89.57
divorced	240	10.43	2301	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
university.degree	939	40.81	939	40.81
high.school	520	22.60	1459	63.41
professional.course	291	12.65	1750	76.05
basic.9y	250	10.86	2000	86.92
basic.4y	202	8.78	2202	95.70
basic.6y	99	4.30	2301	100.00

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	2093	90.96	2093	90.96
unknown	208	9.04	2301	100.00

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	1287	55.93	1287	55.93
no	1014	44.07	2301	100.00

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1965	85.40	1965	85.40
yes	336	14.60	2301	100.00

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	1917	83.31	1917	83.31
telephone	384	16.69	2301	100.00

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
may	444	19.30	444	19.30
aug	335	14.56	779	33.85
jul	329	14.30	1108	48.15
jun	278	12.08	1386	60.23
apr	238	10.34	1624	70.58
nov	219	9.52	1843	80.10
oct	152	6.61	1995	86.70
mar	135	5.87	2130	92.57
sep	127	5.52	2257	98.09
dec	44	1.91	2301	100.00

day_of_week	Frequency	Percent	Cumulative Frequency	Cumulative Percent
thu	528	22.95	528	22.95
wed	467	20.30	995	43.24
tue	464	20.17	1459	63.41
fri	430	18.69	1889	82.09
mon	412	17.91	2301	100.00

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
nonexistent	1528	66.41	1528	66.41
success	465	20.21	1993	86.61
failure	308	13.39	2301	100.00

campaign_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1133	49.24	1133	49.24
2	628	27.29	1761	76.53
3	280	12.17	2041	88.70
4	260	11.30	2301	100.00

previous_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1528	66.41	1528	66.41
1	773	33.59	2301	100.00

emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1255	54.54	1255	54.54
3	559	24.29	1814	78.84
2	487	21.16	2301	100.00

cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	893	38.81	893	38.81
3	805	34.98	1698	73.79
2	603	26.21	2301	100.00

cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	672	29.20	672	29.20
1	600	26.08	1272	55.28
3	398	17.30	1670	72.58
4	327	14.21	1997	86.79
2	304	13.21	2301	100.00

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1279	55.58	1279	55.58
2	433	18.82	1712	74.40
4	230	10.00	1942	84.40
5	192	8.34	2134	92.74
3	167	7.26	2301	100.00

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1625	70.62	1625	70.62
3	546	23.73	2171	94.35
2	130	5.65	2301	100.00

Figure 21 – Frequency Tables for variables in the Validation data set by y

y = 0

y=0

job

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
blue-collar	342	14.97	342	14.97
admin.	318	13.92	660	28.88
technician	263	11.51	923	40.39
services	249	10.90	1172	51.29
retired	203	8.88	1375	60.18
management	182	7.96	1557	68.14
unemployed	167	7.31	1724	75.45
self-employed	165	7.22	1889	82.67
entrepreneur	156	6.83	2045	89.50
housemaid	145	6.35	2190	95.84
student	95	4.16	2285	100.00

marital

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
married	1165	50.98	1165	50.98
single	628	27.48	1793	78.47
divorced	492	21.53	2285	100.00

education

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
university.degree	557	24.38	557	24.38
high.school	470	20.57	1027	44.95
basic.9y	389	17.02	1416	61.97
professional.course	351	15.36	1767	77.33
basic.4y	296	12.95	2063	90.28
basic.6y	222	9.72	2285	100.00

default

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1518	66.43	1518	66.43
unknown	767	33.57	2285	100.00

housing

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	1213	53.09	1213	53.09
no	1072	46.91	2285	100.00

loan

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1524	66.70	1524	66.70
yes	761	33.30	2285	100.00

contact

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	1423	62.28	1423	62.28
telephone	862	37.72	2285	100.00

month

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
may	504	22.06	504	22.06
jul	384	16.81	888	38.86
jun	366	16.02	1254	54.88
nov	302	13.22	1556	68.10
aug	254	11.12	1810	79.21
apr	244	10.68	2054	89.89
oct	76	3.33	2130	93.22
mar	63	2.76	2193	95.97
sep	63	2.76	2256	98.73
dec	29	1.27	2285	100.00

day_of_week

day_of_week	Frequency	Percent	Cumulative Frequency	Cumulative Percent
mon	519	22.71	519	22.71
thu	516	22.58	1035	45.30
fri	434	18.99	1469	64.29
tue	419	18.34	1888	82.63
wed	397	17.37	2285	100.00

poutcome

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
nonexistent	1981	86.70	1981	86.70
failure	267	11.68	2248	98.38
success	37	1.62	2285	100.00

campaign_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	956	41.84	956	41.84
2	589	25.78	1545	67.61
4	429	18.77	1974	86.39
3	311	13.61	2285	100.00

previous_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1981	86.70	1981	86.70
1	304	13.30	2285	100.00

emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1255	54.92	1255	54.92
1	665	29.10	1920	84.03
2	365	15.97	2285	100.00

cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1116	48.84	1116	48.84
2	775	33.92	1891	82.76
1	394	17.24	2285	100.00

cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	662	28.97	662	28.97
1	459	20.09	1121	49.06
5	410	17.94	1531	67.00
4	379	16.59	1910	83.59
3	375	16.41	2285	100.00

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	623	27.26	623	27.26
4	592	25.91	1215	53.17
1	400	17.51	1615	70.68
3	364	15.93	1979	86.61
5	306	13.39	2285	100.00

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1211	53.00	1211	53.00
1	739	32.34	1950	85.34
2	335	14.66	2285	100.00

y=1

y=1

job

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	709	30.36	709	30.36
technician	366	15.67	1075	46.04
blue-collar	318	13.62	1393	59.66
retired	218	9.34	1611	68.99
management	162	6.94	1773	75.93
services	160	6.85	1933	82.78
student	126	5.40	2059	88.18
unemployed	74	3.17	2133	91.35
self-employed	73	3.13	2206	94.48
entrepreneur	69	2.96	2275	97.43
housemaid	60	2.57	2335	100.00

marital

marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
married	1277	54.69	1277	54.69
single	823	35.25	2100	89.94
divorced	235	10.06	2335	100.00

education

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
university.degree	982	42.06	982	42.06
high.school	511	21.88	1493	63.94
professional.course	304	13.02	1797	76.96
basic.4y	226	9.68	2023	86.64
basic.9y	223	9.55	2246	96.19
basic.6y	89	3.81	2335	100.00

default

default	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	2101	89.98	2101	89.98
unknown	234	10.02	2335	100.00

housing

housing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	1324	56.70	1324	56.70
no	1011	43.30	2335	100.00

loan

loan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1989	85.18	1989	85.18
yes	346	14.82	2335	100.00

contact

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	1933	82.78	1933	82.78
telephone	402	17.22	2335	100.00

month

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
may	442	18.93	442	18.93
jul	320	13.70	762	32.63
aug	318	13.62	1080	46.25
apr	300	12.85	1380	59.10
jun	281	12.03	1661	71.13
nov	196	8.39	1857	79.53
oct	163	6.98	2020	86.51
mar	141	6.04	2161	92.55
sep	129	5.52	2290	98.07
dec	45	1.93	2335	100.00

day_of_week

day_of_week	Frequency	Percent	Cumulative Frequency	Cumulative Percent
thu	515	22.06	515	22.06
tue	488	20.90	1003	42.96
wed	481	20.60	1484	63.55
mon	435	18.63	1919	82.18
fri	416	17.82	2335	100.00

poutcome

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
nonexistent	1610	68.95	1610	68.95
success	428	18.33	2038	87.28
failure	297	12.72	2335	100.00

campaign_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1165	49.89	1165	49.89
2	583	24.97	1748	74.86
4	295	12.63	2043	87.49
3	292	12.51	2335	100.00

previous_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1610	68.95	1610	68.95
1	725	31.05	2335	100.00

emp_var_rate_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1339	57.34	1339	57.34
3	547	23.43	1886	80.77
2	449	19.23	2335	100.00

cons_price_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	892	38.20	892	38.20
3	734	31.43	1626	69.64
2	709	30.36	2335	100.00

cons_conf_idx_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	716	30.66	716	30.66
1	642	27.49	1358	58.16
3	394	16.87	1752	75.03
2	293	12.55	2045	87.58
4	290	12.42	2335	100.00

euribor3m_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1280	54.82	1280	54.82
2	474	20.30	1754	75.12
4	239	10.24	1993	85.35
5	203	8.69	2196	94.05
3	139	5.95	2335	100.00

nr_employed_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1673	71.65	1673	71.65
3	551	23.60	2224	95.25
2	111	4.75	2335	100.00

Figure 22 – Histogram and Boxplot of age by y in the Training Data Set

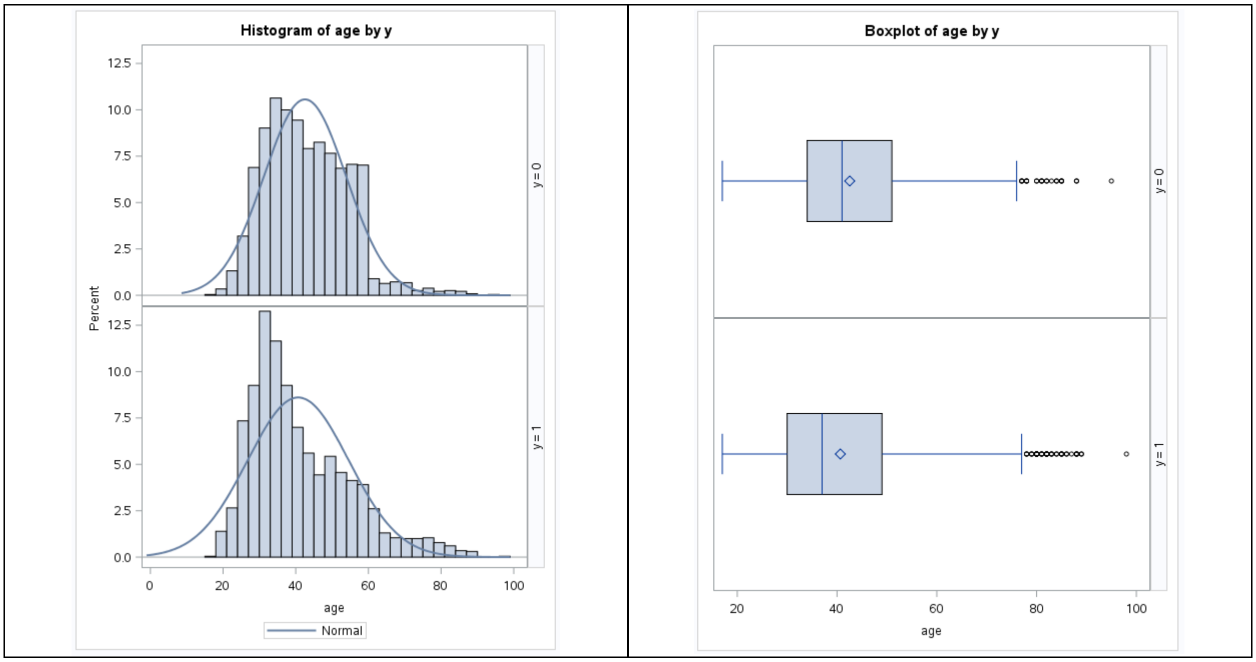


Figure 23 – Histogram and Boxplot of age by y in the Validation Data Set

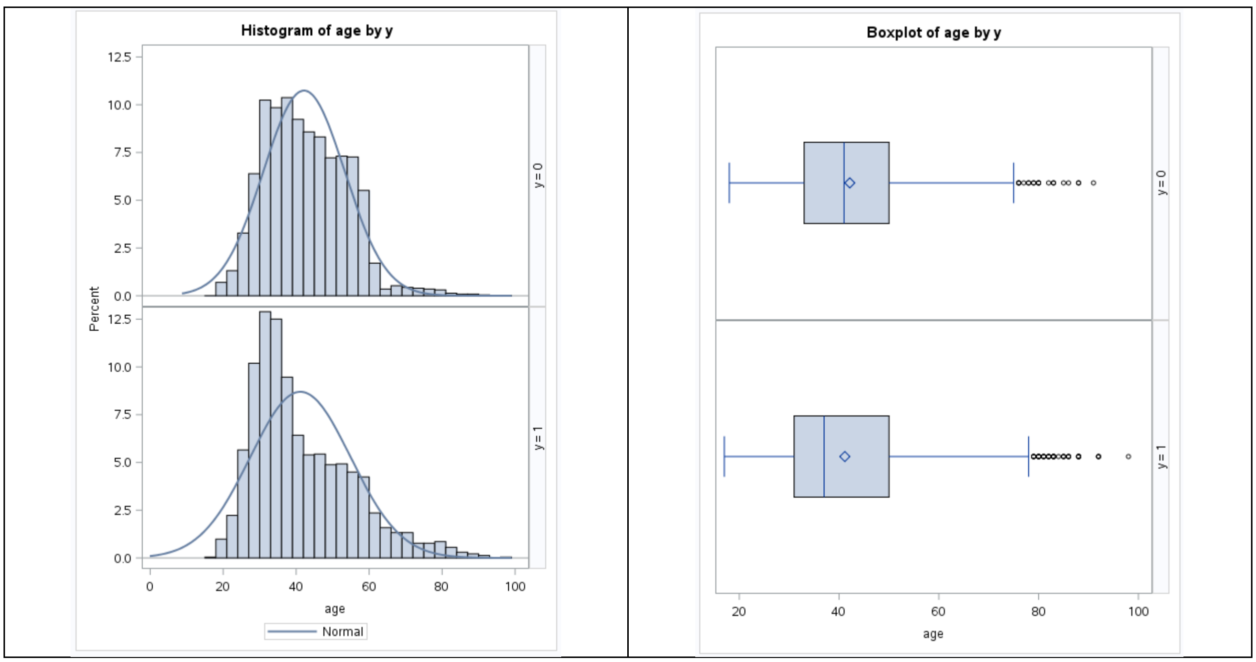


Figure 24 – Logistic Regression Likelihood Ratio Tests for each Training individual variable

Testing Global Null Hypothesis: BETA=0

Variable	Test	Chi-Square	DF	Pr > ChiSq
age	Likelihood Ratio	25.7415	1	<.0001
job	Likelihood Ratio	383.8910	10	<.0001
marital	Likelihood Ratio	130.9205	2	<.0001
education	Likelihood Ratio	181.7945	5	<.0001
default	Likelihood Ratio	446.4846	1	<.0001
housing	Likelihood Ratio	5.0231	1	0.0250
loan	Likelihood Ratio	209.8937	1	<.0001
contact	Likelihood Ratio	242.4661	1	<.0001
month	Likelihood Ratio	114.9999	9	<.0001
day_of_week	Likelihood Ratio	13.5981	4	0.0087
campaign_cat	Likelihood Ratio	70.0505	3	<.0001
previous_cat	Likelihood Ratio	238.5715	1	<.0001
poutcome	Likelihood Ratio	436.2731	2	<.0001
emp_var_rate_cat	Likelihood Ratio	487.5370	2	<.0001
cons_price_idx_cat	Likelihood Ratio	242.2293	2	<.0001
cons_conf_idx_cat	Likelihood Ratio	216.5903	4	<.0001
euribor3m_cat	Likelihood Ratio	684.3988	4	<.0001
nr_employed_cat	Likelihood Ratio	675.9557	2	<.0001

Figure 25 – Logistic Regression Likelihood Ratio Statistics for type 3 analysis of Effects for all significant variables

With all main effects

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	1	0.68	0.4096
job	10	212.60	<.0001
marital	2	111.57	<.0001
education	5	74.30	<.0001
default	1	128.87	<.0001
housing	1	1.85	0.1733
loan	1	167.22	<.0001
contact	1	31.11	<.0001
month	9	67.20	<.0001
day_of_week	4	12.63	0.0132
campaign_cat	3	6.33	0.0966
previous_cat	1	71.51	<.0001
poutcome	2	96.86	<.0001
emp_var_rate_cat	2	20.92	<.0001
cons_price_idx_cat	2	5.05	0.0799
cons_conf_idx_cat	4	15.15	0.0044
euribor3m_cat	4	49.61	<.0001
nr_employed_cat	2	17.23	0.0002

After dropping age, housing, campaign_cat, and cons_price_idx_cat due to insignificant at alpha=0.05.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
job	10	220.56	<.0001
marital	2	111.90	<.0001
education	5	74.41	<.0001
default	1	136.14	<.0001
loan	1	166.83	<.0001
contact	1	38.17	<.0001
month	9	65.65	<.0001
day_of_week	4	13.02	0.0112
previous_cat	1	70.76	<.0001
poutcome	2	103.32	<.0001
emp_var_rate_cat	2	19.45	<.0001
cons_conf_idx_cat	4	10.90	0.0278
euribor3m_cat	4	47.92	<.0001
nr_employed_cat	2	15.90	0.0004

Figure 26 – Logistic Regression Residual plots

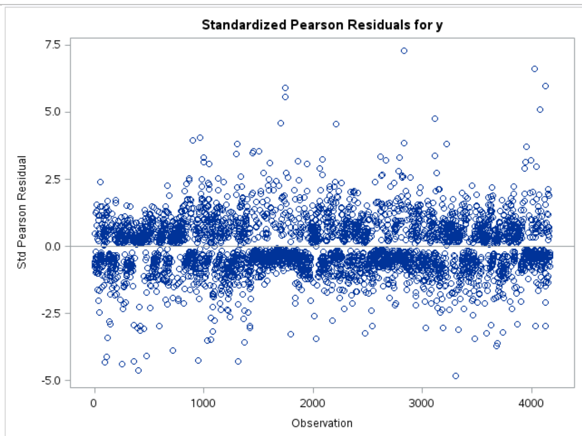
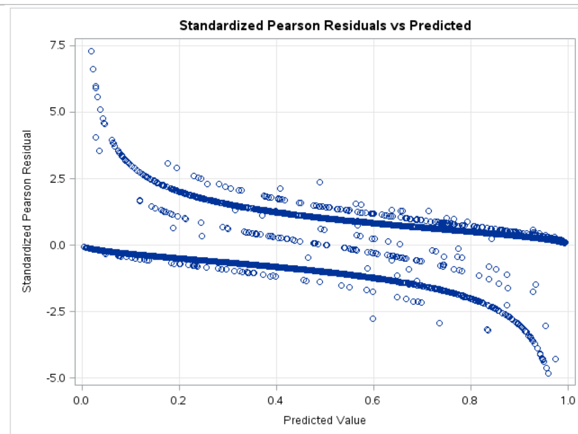
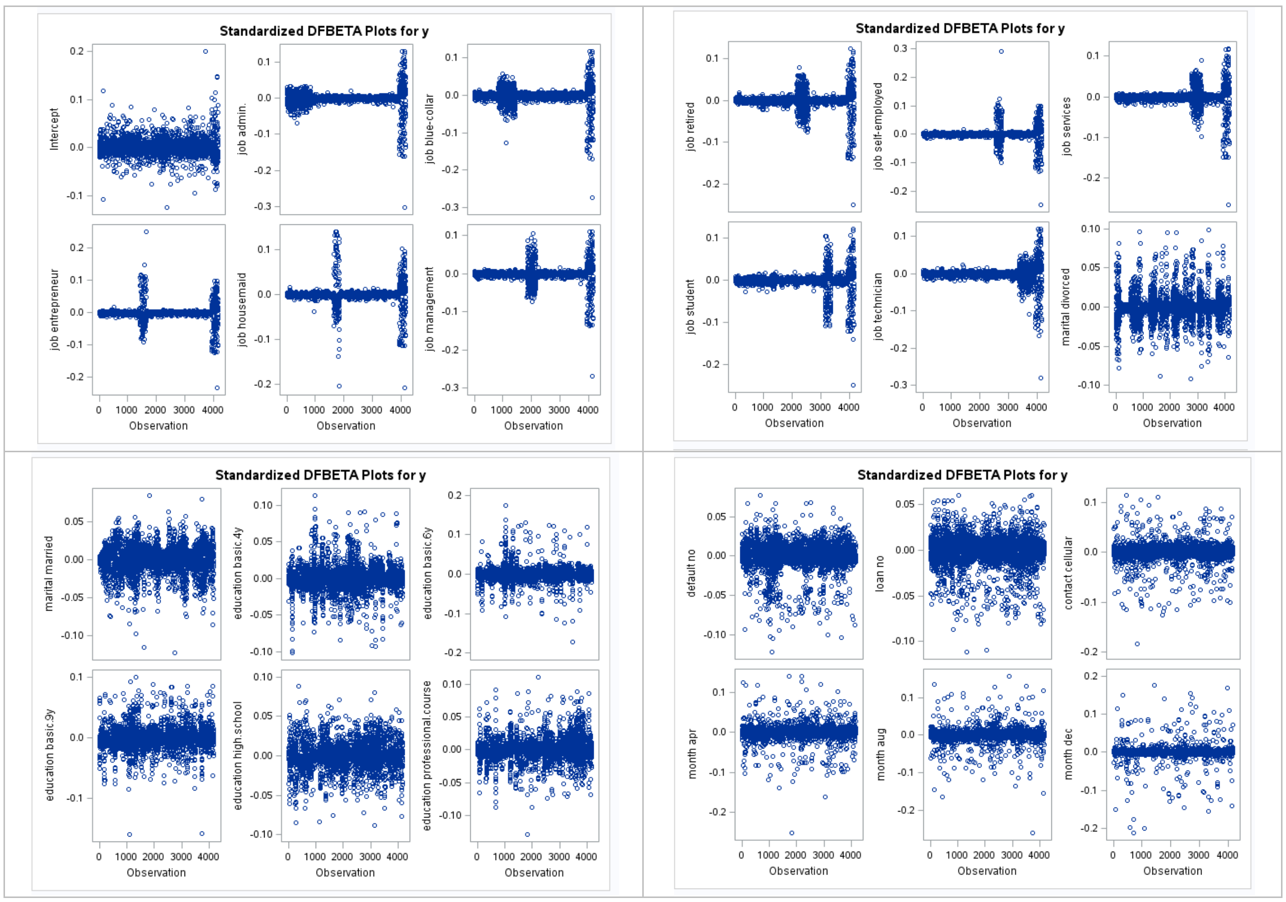
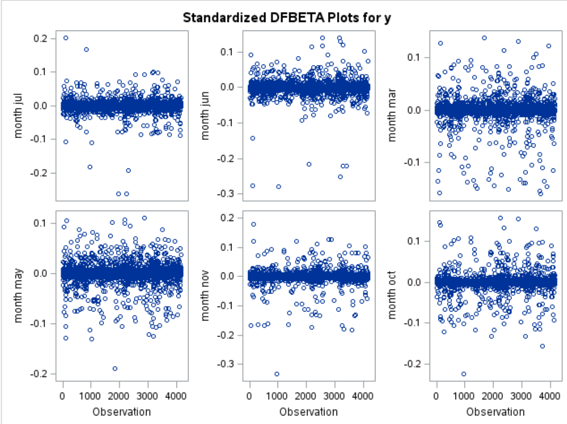
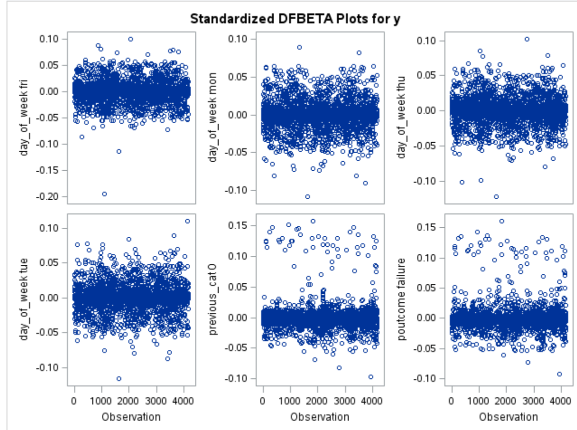
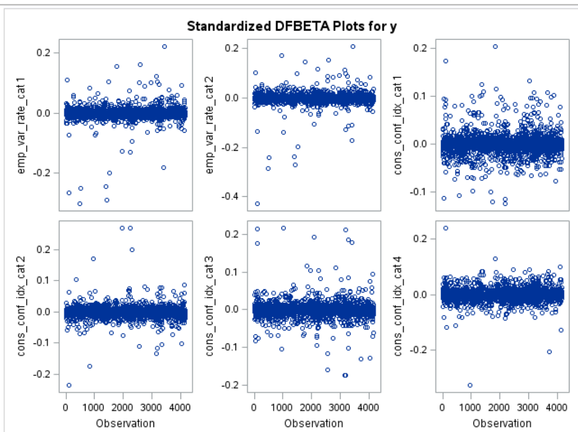
Standardized Pearson Residuals vs Observations	Standardized Pearson Residuals vs Predicted
	

Figure 27 – Logistic Regression Dfbeta plots









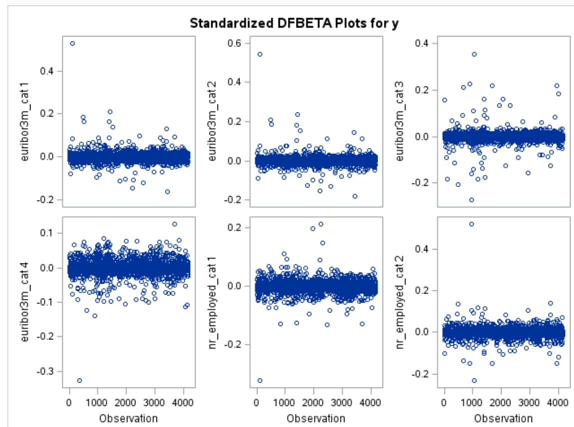
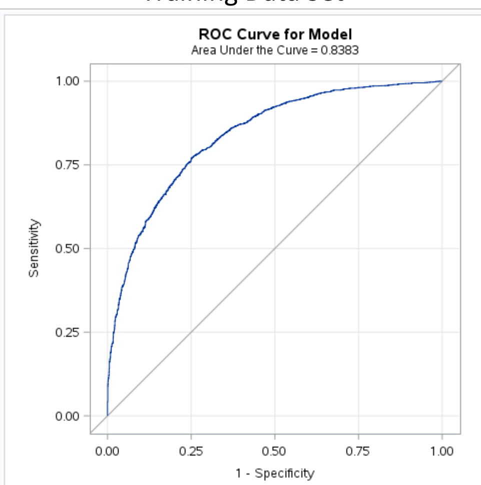
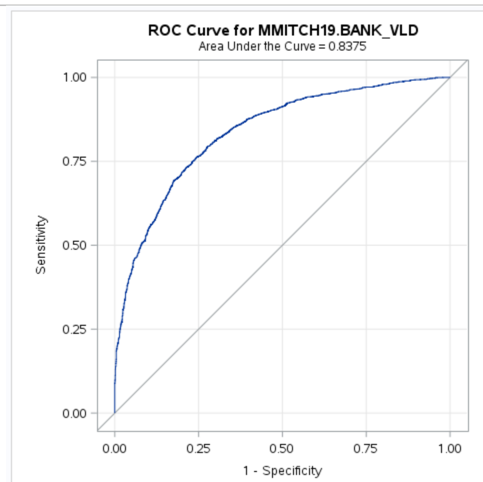


Figure 28 – Logistic Regression ROC Plots

Training Data Set	Validation Data Set
	

ROC Curve Area Under the Curve:
<=0.5 Worthless,
(0.5,0.6) Poor,
[0.6,0.7) Fair,
[0.7,0.8) Good,
[0.8,0.9) Very Good,
[0.9, 1] Excellent.

Table 29 – Logistic Regression Odds Ratio Tables without 1 in the 95% CI

Client Data

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
job	admin.	unemployed	4.95774335	3.407477481	7.280807402
job	blue-collar	unemployed	4.917463269	3.332706866	7.32699882
job	technician	unemployed	4.707912723	3.178154288	7.038298898
job	student	unemployed	3.309642093	2.06563352	5.343643906
default	no	unknown	3.055480129	2.518959559	3.71787627
loan	no	yes	3.035126258	2.553556135	3.61552711
job	services	unemployed	2.902399492	1.917090049	4.426235449
education	university.degree	basic.6y	2.755613855	2.009777564	3.800110307
job	retired	unemployed	2.48681259	1.643377995	3.789528935
job	management	unemployed	2.416814569	1.57557284	3.731341328
education	university.degree	professional.course	2.066805691	1.62991736	2.625250415
education	university.degree	basic.4y	1.895916037	1.45394934	2.477292339
education	university.degree	basic.9y	1.873789545	1.470044252	2.392216325
marital	single	divorced	1.864615243	1.483667707	2.347758899
marital	married	single	1.576092971	1.329244547	1.870639277

Last Contact Info

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
month	nov	sep	5.34096703	2.796776711	10.43870424
month	may	sep	4.559012842	2.335173543	8.992535601
month	jul	sep	4.01449121	1.800143963	9.259929772
month	jun	sep	3.827701733	1.866745025	8.130323966
month	mar	sep	3.424727667	1.879906639	6.308834163
month	aug	sep	3.296406083	2.04411237	5.341211168
month	apr	sep	2.941757182	1.57755399	5.536861978
contact	cellular	telephone	2.363433946	1.795477088	3.120963037
day_of_week	wed	mon	1.424191389	1.131107423	1.794320417

Other

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
poutcome	success	failure or nonexistent	5.68318066	3.956168016	8.278842749
previous_cat	1	0	3.920610615	2.800251296	5.577040547

Social & Economic

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
euribor3m_cat	5	2	28.15204356	7.914191174	112.0247351
euribor3m_cat	5	1	24.91007463	6.145853821	112.1805658
nr_employed_cat	1	3	17.52526337	3.877957182	79.45394151
emp_var_rate_cat	2	3	12.16495584	3.864626574	41.73169737
emp_var_rate_cat	1	3	6.822154396	2.079293864	23.86981474
cons_conf_idx_cat	5	1	1.957229249	1.059630519	3.651487447

Table 30 – Logistic Regression Odds Ratio Tables with 1 in the 95% CI

Client Data

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
education	university.degree	high.school	1.212965689	0.989126454	1.487562848
job	unemployed	housemaid	0.993081125	0.594027156	1.669763466
job	unemployed	entrepreneur	0.809223987	0.496540372	1.321230637
job	unemployed	self-employed	0.718267079	0.447316203	1.151316424

Last Contact Info

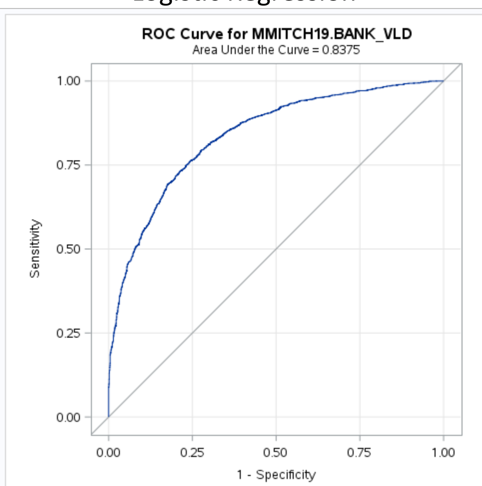
Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
day_of_week	wed	thu	1.046799174	0.836860806	1.309597502
day_of_week	wed	tue	1.051613838	0.836000487	1.322912237
day_of_week	wed	fri	1.001391359	0.795103503	1.261193532
month	sep	oct	0.682033005	0.391609736	1.186049515
month	sep	dec	0.706369849	0.353236702	1.400662231

Social & Economic

Variable	Numerator	Denominator	OddsRatioEst	LowerCL	UpperCL
cons_conf_idx_cat	5	4	1.115239629	0.432923739	2.87194656
cons_conf_idx_cat	5	2	0.929697569	0.412898784	2.155126093
cons_conf_idx_cat	5	3	0.755179292	0.373296965	1.543479781
euribor3m_cat	5	4	0.997038796	0.71776858	1.381565849
euribor3m_cat	5	3	0.540512422	0.277534096	1.061286569
nr_employed_cat	3	2	1.005792798	0.355626835	2.876751751

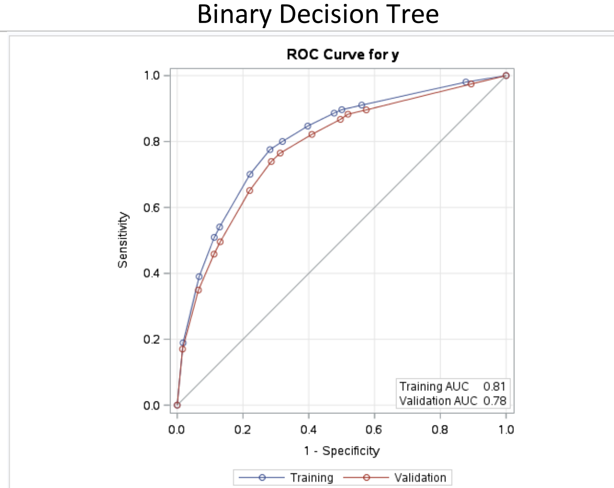
Figure 31 – Model Comparison – ROC Curve and Misclassification Rate for Validation Data

Logistic Regression



Fit Statistics for SCORE Data						
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC
MMITCH19.BANK_VLD	4620	-2297.7	0.2420	4691.428	4692.457	5000.459

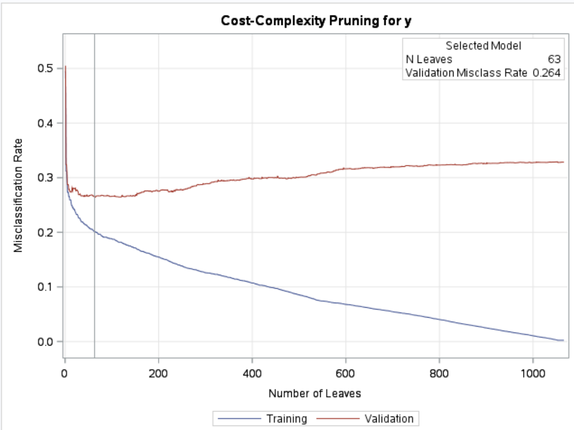
Binary Decision Tree



Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	13	0.1773	0.2529	0.7756	0.7180	0.7699	0.3545	1591.2	0.8061
Validation	13	0.1882	0.2734	0.7394	0.7140	0.8039	0.3749	1800.3	0.7820

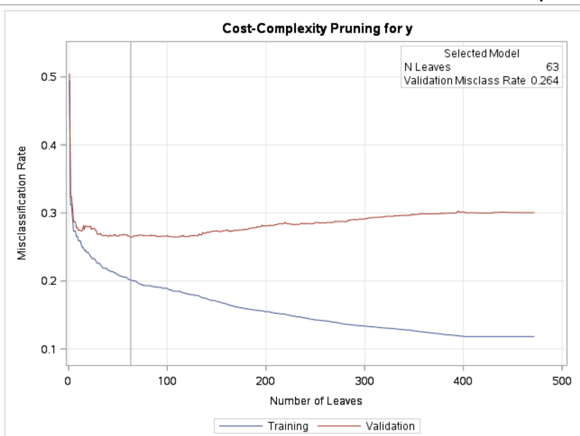
Figure 32 – Decision Tree Optimal Pruning Plots by Misclassification Rate

First Run
Maxdepth = 30
SAS Optimal Leaves = 63



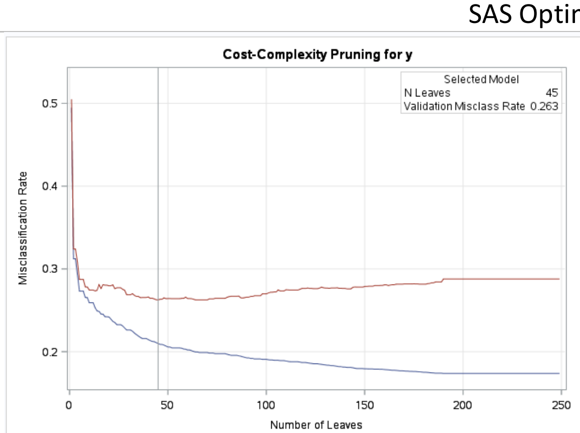
Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	30
Maximum Tree Depth Achieved	23
Tree Depth	11
Number of Leaves Before Pruning	1066
Number of Leaves After Pruning	63
Model Event Level	1

Second Run
Maxdepth = 11
SAS Optimal Leaves = 63



Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	11
Maximum Tree Depth Achieved	11
Tree Depth	11
Number of Leaves Before Pruning	472
Number of Leaves After Pruning	63
Model Event Level	1

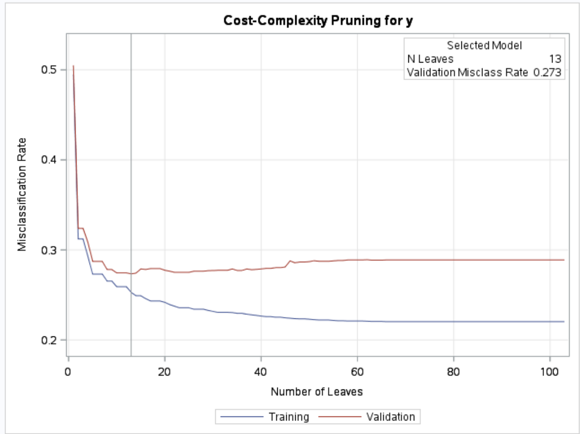
Third Run
Maxdepth = 9
SAS Optimal Leaves = 45



Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	9
Maximum Tree Depth Achieved	9
Tree Depth	9
Number of Leaves Before Pruning	249
Number of Leaves After Pruning	45
Model Event Level	1

Fourth Run

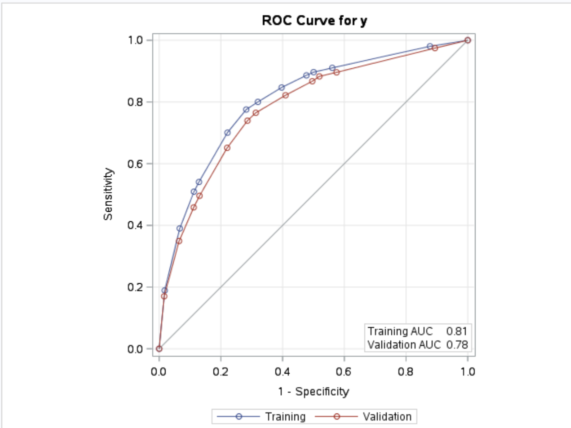
Maxdepth = 7 SAS Optimal Leaves = 13	



Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	7
Maximum Tree Depth Achieved	7
Tree Depth	6
Number of Leaves Before Pruning	103
Number of Leaves After Pruning	13
Model Event Level	1

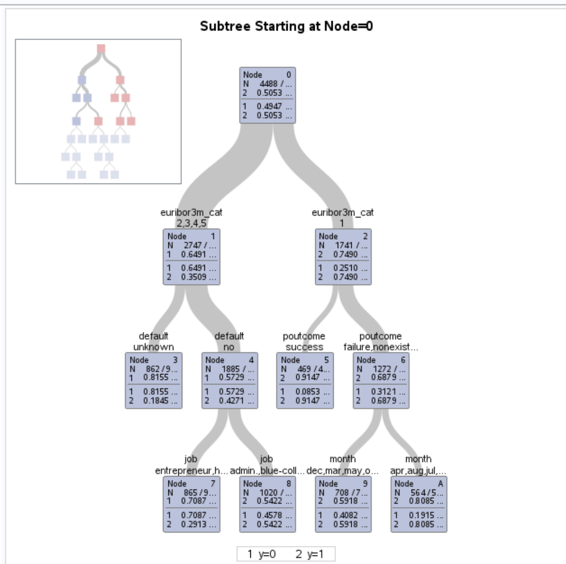
Figure 33 – Decision Tree – ROC Curve, Fit Statistics, Variable Importance and Tree Plots

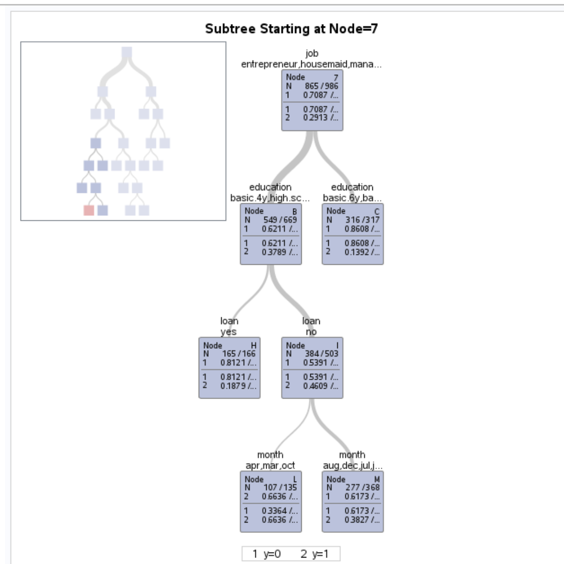
Final Run
Maxdepth = 30
Prune to 13 leaves

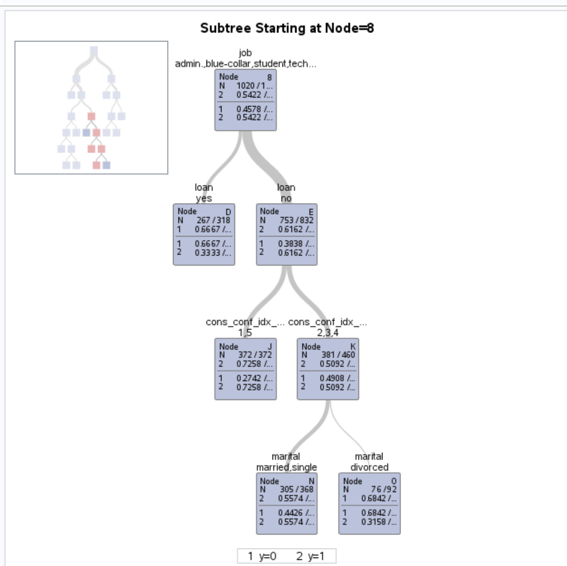


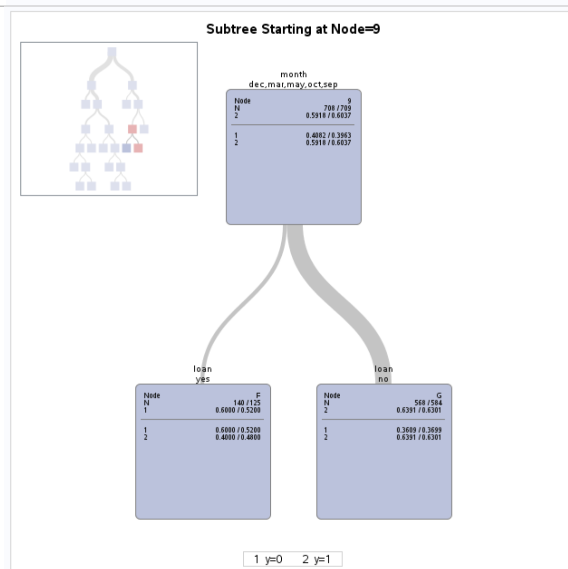
Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	13	0.1773	0.2529	0.7756	0.7180	0.7699	0.3545	1591.2	0.8061
Validation	13	0.1882	0.2734	0.7394	0.7140	0.8039	0.3749	1800.3	0.7820

Variable Importance							
		Training		Validation			
Variable	Variable Label	Relative	Importance	Relative	Importance	Relative Ratio	Count
euribor3m_cat		1.0000	18.3769	1.0000	17.8636	1.0000	1
default	default	0.4541	8.3443	0.4874	8.7076	1.0735	1
loan	loan	0.4271	7.8482	0.4045	7.2260	0.9472	3
job	job	0.4176	7.6744	0.3815	6.8153	0.9136	1
month	month	0.3512	6.4547	0.3363	6.0067	0.9573	2
poutcome	poutcome	0.3231	5.9378	0.3296	5.8881	1.0201	1
education	education	0.2612	4.7993	0.2226	3.9759	0.8522	1
cons_conf_idx_cat		0.2287	4.2029	0.1695	3.0287	0.7413	1
marital	marital	0.1450	2.6649	0.1000	1.7867	0.6897	1









SAS Code

```
*/ Import the file/*;
FILENAME REFFILE "/gpfs/user_home/mmitch19/Bank.xlsx" TERMSTR=CR;

PROC IMPORT DATAFILE=REFFILE
      DBMS=XLSX
      Replace
      OUT=mmitch19.bank;
      GETNAMES=YES;
      SHEET='Data';
RUN;


PROC CONTENTS DATA=mmitch19.bank; RUN;

*/ Look at missing data groups/*;
PROC MI Data=mmitch19.bank simple; run;

*/ Histogram and Boxplots of Continuous Variables/*;

proc sgpanel data=mmitch19.bank;
  title "Histogram of age by y";
  panelby y / layout=rowlattice;
  histogram age;
  density age;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of age by y";
  panelby y / layout=rowlattice;
  hbox age;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of duration by y";
  panelby y / layout=rowlattice;
  histogram duration;
  density duration;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of duration by y";
  panelby y / layout=rowlattice;
  hbox duration;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of campaign by y";
  panelby y / layout=rowlattice;
  histogram campaign;
  density campaign;
run;
proc sgpanel data=mmitch19.bank;
```

```
  title "Boxplot of campaign by y";
  panelby y / layout=rowlattice;
  hbox campaign;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of pdays by y";
  panelby y / layout=rowlattice;
  histogram pdays;
  density pdays;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of pdays by y";
  panelby y / layout=rowlattice;
  hbox pdays;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of previous by y";
  panelby y / layout=rowlattice;
  histogram previous;
  density previous;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of previous by y";
  panelby y / layout=rowlattice;
  hbox previous;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of emp_var_rate by y";
  panelby y / layout=rowlattice;
  histogram emp_var_rate;
  density emp_var_rate;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of emp_var_rate by y";
  panelby y / layout=rowlattice;
  hbox emp_var_rate;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of cons_price_idx by y";
  panelby y / layout=rowlattice;
  histogram cons_price_idx;
  density cons_price_idx;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of cons_price_idx by y";
  panelby y / layout=rowlattice;
  hbox cons_price_idx;
run;
```

```
proc sgpanel data=mmitch19.bank;
  title "Histogram of cons_conf_idx by y";
  panelby y / layout=rowlattice;
  histogram cons_conf_idx;
  density cons_conf_idx;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of cons_conf_idx by y";
  panelby y / layout=rowlattice;
  hbox cons_conf_idx;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of euribor3m by y";
  panelby y / layout=rowlattice;
  histogram euribor3m;
  density euribor3m;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of euribor3m by y";
  panelby y / layout=rowlattice;
  hbox euribor3m;
run;

proc sgpanel data=mmitch19.bank;
  title "Histogram of nr_employed by y";
  panelby y / layout=rowlattice;
  histogram nr_employed;
  density nr_employed;
run;
proc sgpanel data=mmitch19.bank;
  title "Boxplot of nr_employed by y";
  panelby y / layout=rowlattice;
  hbox nr_employed;
run;

proc format;
     value campaign 1 = '1' 2 = '2' 3='3' 4='>3';
     value previous 0 = 'never contacted' 1='contacted before';
     value emp_var_rate 1 = '<= -1.8' 2='[-1.8,-0.1)' 3='>-0.1';
     value cons_price_idx 1 = '<93.06' 2='[93.06,93.91]' 3='>93.91';
     value cons_conf_idx 1='<-46.2' 2='[-46.2,-42)' 3='[-42,-40)' 4='[-40,-36.4)'
5='>=-36.4';
     value euribor3m 1='<1.3' 2='[1.3,4.19)' 3='[4.19,4.86)' 4='[4.86,4.96)'
5='>=4.96';
     value nr_employed 1 = '<5099.1' 2='[5099.1,5191.02)' 3='>5191.02';

*/ Frequency Table of campaign for recoding/*;
proc sort data=mmitch19.bank out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
```

```
      tables campaign / plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;


*/ Frequency Table of pdays/*;

proc sort data=mmitch19.bank out=Work.SortTempTableSorted;
      by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
      tables pdays / missing plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

*/ Frequency Table of previous/*;

proc sort data=mmitch19.bank out=Work.SortTempTableSorted;
      by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
      tables previous / missing plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

/*--Set output size--*/
ods graphics / reset imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=mmitch19.bank;
      /*--TITLE and FOOTNOTE--*/
      title 'Grouped Bar Chart of emp_var_rate by y';

      /*--Bar chart settings--*/
      vbar emp_var_rate / group=y groupdisplay=Cluster name='Bar';

      /*--Response Axis--*/
      yaxis grid;
run;

ods graphics / reset;
title;
```

```
*/ Frequency Table of cons_price_idx /*;

proc sort data=mmitch19.bank out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
     tables cons_price_idx  / missing plots=none;
     by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

/*--Set output size--*/
ods graphics / reset imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=mmitch19.bank;
     /*--TITLE and FOOTNOTE--*/
     title 'Grouped Bar Chart of cons_price_idx by y';

     /*--Bar chart settings--*/
     vbar cons_price_idx / group=y groupdisplay=Cluster name='Bar';

     /*--Response Axis--*/
     yaxis grid;
run;

ods graphics / reset;
title;

ods noproctitle;

/*Bucket binning for cons_price_idx*/;

proc hpbin data=MMITCH19.BANK numbin=3 bucket computestats computequantile;
     input cons_price_idx;
run;

*/ Frequency Table of cons_conf_idx  /*;

proc sort data=mmitch19.bank out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
     tables cons_conf_idx   / missing plots=none;
     by y;
run;

proc delete data=Work.SortTempTableSorted;
```

```
run;

/*--Set output size--*/
ods graphics / reset imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=mmitch19.bank;
      /*--TITLE and FOOTNOTE--*/
      title 'Grouped Bar Chart of cons_conf_idx by y';

      /*--Bar chart settings--*/
      vbar cons_conf_idx / group=y groupdisplay=Cluster name='Bar';

      /*--Response Axis--*/
      yaxis grid;
run;

ods graphics / reset;
title;


/*Quantile binning for cons_conf_idx*/;

ods noproctitle;

proc hpbin data=MMITCH19.BANK numbin=5 pseudo_quantile computestats
           computequantile;
      input cons_conf_idx;
run;

ods noproctitle;


/*Quantile binning for euribor3m*/;
proc hpbin data=MMITCH19.BANK numbin=5 pseudo_quantile computestats
           computequantile;
      input euribor3m;
run;



/*Quantile binning for nr_employed*/;
ods noproctitle;

proc hpbin data=MMITCH19.BANK numbin=4 pseudo_quantile;
      input nr_employed;
run;



/*Creating categorical variables for continuous variables that require them*/;

data mmitch19.bank_rcd;
```

```
	set mmitch19.bank;
	pdays_cat = 'never contacted';
	if pdays ^= 999 then pdays_cat = 'contacted before';
	campaign_cat = 999;
	if campaign = 1 then campaign_cat = 1;
	if campaign = 2 then campaign_cat = 2;
	if campaign = 3 then campaign_cat = 3;
	if campaign > 3 then campaign_cat = 4;
	previous_cat = 1;
	if previous = 0 then previous_cat = 0;
	emp_var_rate_cat = 999;
	if emp_var_rate LE -1.8 then emp_var_rate_cat = 1;
	if (emp_var_rate > -1.8) and (emp_var_rate LE -0.1) then emp_var_rate_cat =
2;
	if emp_var_rate > -0.1 then emp_var_rate_cat = 3;

	cons_price_idx_cat = 999;
	if cons_price_idx < 93.056333333 then cons_price_idx_cat = 1;
	if (cons_price_idx GE 93.056333333) and (cons_price_idx LE 93.911666667)
then cons_price_idx_cat = 2;
	if cons_price_idx > 93.911666667 then cons_price_idx_cat = 3;

	cons_conf_idx_cat = 999;
	if cons_conf_idx < -46.19925 then cons_conf_idx_cat = 1;
	if (cons_conf_idx GE -46.19925) and (cons_conf_idx LE -41.99763) then
cons_conf_idx_cat = 2;
	if (cons_conf_idx GE -41.99763) and (cons_conf_idx LE -39.99959) then
cons_conf_idx_cat = 3;
	if (cons_conf_idx GE -39.99959) and (cons_conf_idx LE -36.39786) then
cons_conf_idx_cat = 4;
	if cons_conf_idx > -36.39786 then cons_conf_idx_cat = 5;

	euribor3m_cat = 999;
	if euribor3m < 1.2991788 then euribor3m_cat = 1;
	if (euribor3m GE 1.2991788) and (euribor3m LE 4.1910304) then euribor3m_cat
= 2;
	if (euribor3m GE 4.1910304) and (euribor3m LE 4.864149) then euribor3m_cat =
3;
	if (euribor3m GE 4.864149) and (euribor3m LE 4.9620732) then euribor3m_cat =
4;
	if euribor3m > 4.9620732 then euribor3m_cat = 5;

	nr_employed_cat = 999;
	if nr_employed < 5099.10335 then nr_employed_cat = 1;
	if (nr_employed GE 5099.10335) and (nr_employed LE 5191.0171) then
nr_employed_cat = 2;
	if nr_employed > 5191.0171 then nr_employed_cat = 3;

run;

/*Frequency table for emp_var_rate_cat*/;
proc sort data=mmitch19.bank_RCD out=Work.SortTempTableSorted;
	by y;
```

```
run;

proc freq data=Work.SortTempTableSorted;
      format emp_var_rate_cat emp_var_rate.;
      tables emp_var_rate_cat / plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

/*Frequency table for cons_price_idx_cat*/;
proc sort data=mmitch19.bank_RCD out=Work.SortTempTableSorted;
      by y;
run;

proc freq data=Work.SortTempTableSorted;
      format cons_price_idx_cat cons_price_idx.;
      tables cons_price_idx_cat / plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

/*Frequency table for cons_conf_idx_cat*/;
proc sort data=mmitch19.bank_RCD out=Work.SortTempTableSorted;
      by y;
run;

proc freq data=Work.SortTempTableSorted;
      format cons_conf_idx_cat cons_conf_idx.;
      tables cons_conf_idx_cat / plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;


/*Frequency table for euribor3m_cat */;
proc sort data=mmitch19.bank_RCD out=Work.SortTempTableSorted;
      by y;
run;

proc freq data=Work.SortTempTableSorted;
      format euribor3m_cat euribor3m.;
      tables euribor3m_cat / plots=none;
      by y;
run;

proc delete data=Work.SortTempTableSorted;
run;
```

```
/*Frequency table for nr_employed_cat*/;
proc sort data=mmitch19.bank_RCD out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted;
     format nr_employed_cat nr_employed.;
     tables nr_employed_cat / plots=none;
     by y;
run;

proc delete data=Work.SortTempTableSorted;
run;

********************************************************;
*******************Categorical Variables***************;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (job) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (marital) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (education) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (default) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (housing) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;
```

```
ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (loan) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (contact) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (month) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (day_of_week) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     format campaign_cat campaign.;
     tables  (campaign_cat) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     format previous_cat previous.;
     tables  (previous_cat) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
     tables  (poutcome) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;
```

```
proc freq data=MMITCH19.BANK_RCD;
      format emp_var_rate_cat emp_var_rate.;
      tables  (emp_var_rate_cat) *(y) / missing nopercent nocum
plots(only)=(freqplot mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
      format cons_price_idx_cat cons_price_idx.;
      tables  (cons_price_idx_cat) *(y) / missing nopercent nocum
plots(only)=(freqplot mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
      format cons_conf_idx_cat cons_conf_idx.;
      tables  (cons_conf_idx_cat) *(y) / missing nopercent nocum
plots(only)=(freqplot mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
      format euribor3m_cat euribor3m.;
      tables  (euribor3m_cat) *(y) / missing nopercent nocum plots(only)=(freqplot
mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
      format nr_employed_cat nr_employed.;
      tables  (nr_employed_cat) *(y) / missing nopercent nocum
plots(only)=(freqplot mosaicplot);
run;

ods noproctitle;

proc freq data=MMITCH19.BANK_RCD;
      tables (y) / missing nocum plots(only)=(freqplot mosaicplot);
run;

/*Collapsing categorical variables unknown level and levels that can be
collapsed; Recode y into 1=yes, 0=no*/;

data mmitch19.bank_rcd2;
      set mmitch19.bank_rcd;
      if job = 'unknown' then job = 'admin.';
      if marital = 'unknown' then marital = 'married';
      if education = 'unknown' then education = 'university.degree';
      if education = 'illiterate' then DELETE;
```

```
     if default = 'yes' then DELETE;
     if housing = 'unknown' then housing = 'yes';
     if loan = 'unknown' then loan = 'no';
     if y = 'yes' then y2 =1;
     if y = 'no' then y2=0;
     drop y;
     rename y2=y;
run;


/*Make a yes and no data set.  Randomn sample 4,636 obs from the no data set.
Append the 2 data sets for modeling.*/;
proc sql noprint;
     create table MMITCH19.BANK_yes as select * from MMITCH19.BANK_RCD2 where(y
EQ
          1);
quit;

proc sql noprint;
     create table MMITCH19.BANK_no as select * from MMITCH19.BANK_RCD2 where(y EQ
          0);
quit;

proc sort data=MMITCH19.BANK_NO out=WORK.SORTTempTableSorted;
     by job marital education default housing loan month;
run;

proc surveyselect data=WORK.SORTTempTableSorted out=MMITCH19.BANK_SAMPLE_NO
          method=srs sampsize=4636;
     strata job marital education default housing loan month / alloc=prop;
run;

proc delete data=WORK.SORTTempTableSorted;
run;

data mmitch19.bank_modeling;
     set mmitch19.bank_yes mmitch19.bank_sample_no;
run;
/*******************************************************************************
*************************************/;
/***********************Partition the data set into 50/50 Training and
Validation.************************************/;
/*******************************************************************************
*************************************/;


proc sort data=MMITCH19.BANK_MODELING out=work._sorted_;
     by education job;
run;

proc means data=work._sorted_ noprint;
     by education job;
     output out=work._meansOut_(drop=_type_ _freq_) n=__nobs__;
```

```
run;

proc sql noprint;
      select max(__nobs__) into :count from work._meansOut_;
quit;

data mmitch19.bank_trn mmitch19.bank_vld;
      set work._sorted_;
      by education job;
      retain __tmp1-__tmp%trim(&count) __nobs__ __nobs1__ __nobs2__;
      retain __nobs__ __seed__ _n1_;
      drop __k__;
      drop _i_ __seed__ __tmp1-__tmp%trim(&count);
      drop _n1_ __nobs__ __nobs1__ __nobs2__;
      array __tmp(*) __tmp1-__tmp%trim(&count);

      if (_n_=1) then
            do;
                  __seed__=9889;
                  __nobs__=&count;
            end;

      if first.job then
            do;
                  set work._meansOut_;
                  by education job;

                  do _i_=1 to __nobs__;
                        __tmp(_i_)=_i_;
                  end;

                  if (__nobs__ < dim(__tmp)) then
                        do;

                              do _i_=__nobs__+1 to dim(__tmp);
                                    __tmp(_i_)=0;
                              end;
                        end;
                  call ranperm(__seed__, of __tmp(*));

                  if (__nobs__ < dim(__tmp)) then
                        do;
                              * Move non-zero values to beginning of list;

                              do _i_=1 to dim(__tmp);

                                    if (__tmp(_i_)=0) then
                                          do;

                                                if (_i_ < dim(__tmp)) then
                                                      do;
                                                            __k__=_i_ + 1;
```

```
                                                  do while(__k__ < dim(__tmp)
and __tmp(__k__)=0);
                                                       __k__=__k__+1;
                                                  end;

                                                  if (__k__ <=dim(__tmp))
then
                                                       do;

     __tmp(_i_)=__tmp(__k__);
                                                            __tmp(__k__)=0;
                                                       end;
                                             end;
                                   end;
                         end;
                    end;
               _n1_=0;
               __nobs1__=round(0.5*__nobs__);
               __nobs2__=round(0.5*__nobs__)+__nobs1__;
          end;
     _n1_=_n1_ + 1;

     if (_n1_ <=dim(__tmp)) then
          do;

               if (__tmp(_n1_) > 0) then
                    do;

                         if (__tmp(_n1_) <=__nobs1__) then
                              do;
                                   output mmitch19.bank_trn;
                              end;
                         else if (__tmp(_n1_) <=__nobs2__) then
                              do;
                                   output mmitch19.bank_vld;
                              end;
                    end;
          end;
run;

proc delete data=work._sorted_;
run;

proc delete data=work._meansOut_;
run;

/*********************************************************************************
**************************************/;
/***************** END: Partition the data set into 50/50 Training and
Validation.*************************************/;
/*********************************************************************************
**************************************/;
```

```
/*****************************************************************************************
*************************************/;
/***************** Check the number of observations in each level of the training
data set by response level **********/;
/*****************************************************************************************
*************************************/;

proc sort data=MMITCH19.BANK_TRN out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
     tables job marital education default housing loan contact month day_of_week
          poutcome campaign_cat previous_cat emp_var_rate_cat cons_price_idx_cat
          cons_conf_idx_cat euribor3m_cat nr_employed_cat / plots=none;
     by y;
run;

proc delete data=Work.SortTempTableSorted;
run;
/*****************************************************************************************
*************************************/;
/***************** END: Check the number of observations in each level of the
training data set by response level *****/;
/*****************************************************************************************
*************************************/;

/*****************************************************************************************
*************************************/;
/***************** Check the number of observations in each level of the
validation data set by response level ********/;
/*****************************************************************************************
*************************************/;

proc sort data=MMITCH19.BANK_VLD out=Work.SortTempTableSorted;
     by y;
run;

proc freq data=Work.SortTempTableSorted order=freq;
     tables job marital education default housing loan contact month day_of_week
          poutcome campaign_cat previous_cat emp_var_rate_cat cons_price_idx_cat
          cons_conf_idx_cat euribor3m_cat nr_employed_cat / plots=none;
     by y;
run;

proc delete data=Work.SortTempTableSorted;
run;
/*****************************************************************************************
*************************************/;
/***************** END: Check the number of observations in each level of the
validation data set by response level ***/;
/*****************************************************************************************
*************************************/;
```

```
/********************************************************************************
**************************************/;
/***************** Histogram and Boxplot of age in training and validation data
sets **********************************/;
/********************************************************************************
**************************************/;
proc sgpanel data=mmitch19.bank_trn;
  title "Histogram of age by y";
  panelby y / layout=rowlattice;
  histogram age;
  density age;
run;
proc sgpanel data=mmitch19.bank_trn;
  title "Boxplot of age by y";
  panelby y / layout=rowlattice;
  hbox age;
run;

proc sgpanel data=mmitch19.bank_vld;
  title "Histogram of age by y";
  panelby y / layout=rowlattice;
  histogram age;
  density age;
run;
proc sgpanel data=mmitch19.bank_vld;
  title "Boxplot of age by y";
  panelby y / layout=rowlattice;
  hbox age;
run;

/********************************************************************************
**************************************/;
/***************** END:  Histogram and Boxplot of age in training and validation
data sets ****************************/;
/********************************************************************************
**************************************/;

/********************************************************************************
**************************************/;
/***************** Perform LRT on each individual variable in training data set
to weed out variables *****************/;
/********************************************************************************
**************************************/;

proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = age ;
run;
proc logistic data=mmitch19.bank_trn;
```

```
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = job;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = marital;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = education;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = default;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = housing;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = loan;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = contact;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
```

```
model y = month;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = day_of_week;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = campaign_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = previous_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = poutcome;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = emp_var_rate_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = cons_price_idx_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = cons_conf_idx_cat ;
run;
proc logistic data=mmitch19.bank_trn;
```

```
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = euribor3m_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = nr_employed_cat;
run;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = pdays;
run;
/*********************************************************************************
*************************************/;
/***************** END: Perform LRT on each individual variable in training data
set to weed out variables ************/;
/*********************************************************************************
*************************************/;

/*********************************************************************************
*************************************/;
/***************** Perform tye 3 LRT on significant main effects in training data
set *********************************/;
/*********************************************************************************
*************************************/;
proc genmod data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
param=ref;
model y = age job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat /
     dist=bin link=logit TYPE1 type3;
run;

/**************** Drop age, housing, campaign_cat, and cons_price_idx_cat due to
insignificant ***********************/;
proc genmod data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
     cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
model y = job marital education default loan contact month day_of_week
previous_cat poutcome emp_var_rate_cat
     cons_conf_idx_cat euribor3m_cat nr_employed_cat /
```

```
      dist=bin link=logit TYPE1 type3;
run;

/*******************************************************************************
*************************************/;
/***************** END: Perform tye 3 LRT on significant main effects in training
data set ***************************/;
/*******************************************************************************
*************************************/;

/*******************************************************************************
*************************************/;
/***************** Group the data so that residual and influence diagnostics can
be computed *************************/;
/*******************************************************************************
*************************************/;

proc sql;
      create table mmitch19.bank_trn_count as
      select job, marital, education, default, loan, contact, month, day_of_week,
previous_cat, poutcome, emp_var_rate_cat,
      cons_conf_idx_cat, euribor3m_cat, nr_employed_cat, sum(y) as y, count(*) as
n
      from mmitch19.bank_trn
      group by job, marital, education, default, loan, contact, month,
day_of_week, previous_cat, poutcome, emp_var_rate_cat,
      cons_conf_idx_cat, euribor3m_cat, nr_employed_cat;
quit;
ods output ParameterEstimates = mmitch19.ParameterEstimates_trn;
proc genmod data=mmitch19.bank_trn_count plots=(STDRESCHI DFBETAS);
      class y(desc) job marital education default loan contact month day_of_week
previous_cat poutcome emp_var_rate_cat
      cons_conf_idx_cat euribor3m_cat nr_employed_cat;
      model y/n = job marital education default loan contact month day_of_week
previous_cat poutcome emp_var_rate_cat
      cons_conf_idx_cat euribor3m_cat nr_employed_cat /
            dist=bin link=logit TYPE1 type3 residuals cl;  *aggregate influence CL
diagnostics;
            output out=mmitch19.resid_dfbeta_trn stdreschi=reschi p=predicted
resraw=resraw  dfbetas=_all_;
run;
data mmitch19.PE_zero_trn;
      set mmitch19.parameterestimates_trn;
      if DF = 0;
run;

proc sql noprint;
      create table mmitch19.resid_trn_outliers as select * from
mmitch19.resid_dfbeta_trn
            where(reschi GT 3 OR reschi LT
                  -3);
quit;
```

```
/*Scatter Plot of Standardize Perason Residuals vs Predicted Values*/
ods graphics / reset imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=MMITCH19.RESID_DFBETA_TRN;
	/*--TITLE and FOOTNOTE--*/
	title 'Standardized Pearson Residuals vs Predicted';

	/*--Scatter plot settings--*/
	scatter x=predicted y=reschi / transparency=0.0 name='Scatter';

	/*--X Axis--*/
	xaxis grid;

	/*--Y Axis--*/
	yaxis grid;
run;

ods graphics / reset;
title;

/*****************************************************************************
*************************************/;
/***************** END:  Group the data so that residual and influence
diagnostics can be computed *******************/;
/*****************************************************************************
*************************************/;


*/Rerun the model with PROC Logistic to get odds ratio and ROC curve/*;

ods graphics on;
ods output CLOddsPL=mmitch19.OddsRatiosPL;
proc logistic data=mmitch19.bank_trn;
class y(desc) job marital education default housing loan contact month
day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
	cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat
/param=ref;
model y(event="1") = job marital education default loan contact month day_of_week
previous_cat poutcome emp_var_rate_cat
	cons_conf_idx_cat euribor3m_cat nr_employed_cat / outroc=mmitch19.troc
	cl clodds=both clparm=both ctable plcl ;
score data=mmitch19.bank_vld out=mmitch19.valpred outroc=mmitch19.vroc fitstat;
roc; roccontrast;
run;

/*****************************************************************************
*************************************/;
/***************** Decision Tree
*****************************************************************************
*****/;
/*****************************************************************************
*************************************/;
```

```
ods graphics on;

proc hpsplit data=mmitch19.bank_modeling maxdepth=30;
   class y job marital education default housing loan contact month day_of_week
campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat ;
   model y(event="1") = age job marital education default housing loan contact
month day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
   prune costcomplexity;*(leaves=17) ;
   partition fraction(validate = 0.5 seed=9889) ;
   *code file='hpsplexc.sas';
   *rules file='rules.txt';
run;

ods graphics on;

proc hpsplit data=mmitch19.bank_modeling maxdepth=11;
   class y job marital education default housing loan contact month day_of_week
campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat ;
   model y(event="1") = age job marital education default housing loan contact
month day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
   prune costcomplexity;*(leaves=15) ;
   partition fraction(validate = 0.5 seed=9889) ;
   *code file='hpsplexc.sas';
   *rules file='rules.txt';
run;


ods graphics on;

proc hpsplit data=mmitch19.bank_modeling maxdepth=9;
   class y job marital education default housing loan contact month day_of_week
campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat ;
   model y(event="1") = age job marital education default housing loan contact
month day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
   prune costcomplexity;*(leaves=15) ;
   partition fraction(validate = 0.5 seed=9889) ;
   *code file='hpsplexc.sas';
   *rules file='rules.txt';
run;

ods graphics on;

proc hpsplit data=mmitch19.bank_modeling maxdepth=7;
   class y job marital education default housing loan contact month day_of_week
campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat ;
```

```
   model y(event="1") = age job marital education default housing loan contact
month day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
   prune costcomplexity;*(leaves=15) ;
   partition fraction(validate = 0.5 seed=9889) ;
   *code file='hpsplexc.sas';
   *rules file='rules.txt';
run;

ods graphics on;

proc hpsplit data=mmitch19.bank_modeling maxdepth=30 plots=zoomedtree(nodes=('0')
depth=3)
      plots=zoomedtree(nodes=('7') depth=3) plots=zoomedtree(nodes=('8') depth=3)
plots=zoomedtree(nodes=('9') depth=3);
   class y job marital education default housing loan contact month day_of_week
campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat ;
   model y(event="1") = age job marital education default housing loan contact
month day_of_week campaign_cat previous_cat poutcome emp_var_rate_cat
      cons_price_idx_cat cons_conf_idx_cat euribor3m_cat nr_employed_cat;
   prune costcomplexity(leaves=13) ;
   partition fraction(validate = 0.5 seed=9889) ;
   *code file='hpsplexc.sas';
   *rules file='rules.txt';
run;
```