Direct Marketing Campaigns of Portuguese Banking Clients

DSC 423 Final Report

Crystal Contreras
Morgan (Masahiro) Cho
Roshen Samuel
Todd Lehky

Table of Contents

Abstract

We set out to create a logistic regression model to effectively predict the outcomes of telemarketing campaigns selling long-term bank deposits. We first identified a dataset including research between 2008 to 2013 from a Portuguese retail bank. We began our analysis with a subset of 3,090 observations containing 20 features, 10 numerical and 10 categorical. We began by normalizing the dataset by converting categorical variables to dummy numeric variables. We then built a full logistic regression model to analyze significant variables, apply selection methods to find a subset of optimal predictors, removed outliers, and monitored for multicollinearity among independent variables. After analyzing the full model we employed the backwards and stepwise method of model selection to identify our best fit model. We identified four key performance indicators to compare each model, R-square, Akaike Information Criterion, Schwarz Criterion, and Likelihood Ratio. We concluded that the backwards selection methods out performed the full and stepwise models in three of four observed metrics. The backwards selection method had the highest r-squared and LR values of .2647, and 1,314.45 respectively. In addition, the backwards selection method produced the lowest AIC score of 1,334.45. Only the stepwise selection method outperformed the backwards method with a lower SC. We then tested the possibility of removing significant outliers as our model was only predicting 26% of the variation in

our dependent variable. Following our previous methodology we performed the same steps on a smaller dataset of 1,090 observations, removing approximately 2,000 records that had a dependent variable of ‘No’ and PDAYS = 999. Again, the backwards selection method out-performed the others with the exception of the Schwarz Criterion. Having identified our best fit model we continued to validate our model by splitting the data into a training and testing set. Our final model is able to correctly classify 85.1% of total relevant results. Predicted Y’s are correct 84.3% among all cases and the amount of all true positives against predicted positives is 72.3%.

Introduction

The dataset being utilized for analysis is the Bank Marketing Dataset which focuses on variables related to a marketing campaign for a financial institution in Portugal. The goal of our analysis is to use regression techniques to identify significant parameters that impact the likelihood and success rate of clients subscribing to long-term bank deposits. Through data exploration and logistic regression analyses, we aimed to produce a model that indicates the significant nominal and categorical variables on the client’s response of ‘yes’ to making a bank term deposit from direct phone marketing campaigns.

The original dataset was obtained using a data mining approach to “...predict the success of telemarketing calls for selling bank long-term deposits. A Portugeuse retail bank was addressed, with data collected from 2008 to 2013, thus including the effects of the recent financial crisis.” (Moro et al., 1). The importance of this analysis is to find key areas of focus to help improve future marketing campaigns for financial institutions in order to allocate time and effort appropriately with comparisons among models derived from logistic regression, decision trees, neural networks, and support vector machines. With inspiration from the research conducted by Moro and his team, the success rate of the telemarketing campaign will be determined with a more simplified approach that utilizes a predictive model derived solely from logistic regression and data-driven insight.

Methodology

Obtaining the Dataset

To begin our analysis, we first examined and analyzed the “*Bank Marketing Dataset " found* via Kaggle. From Kaggle, two main datasets were offered. The first was a set of 45,211 containing 17 features (bank-additional-full.csv). The second was a

randomly selected subset of the first containing 4,119 observations & 20 features (bank-additional.csv). We chose the latter in order to fit the constraints of the assignment, as well as to avoid latency issues with SAS. From this second set, we created our own subset of the data containing 3,090 observations by removing rows that contained any null values. This final subset of 3,090 will be considered our "original dataset" and will serve as an initial platform from where further analysis is performed to create our model. The telemarketing data included the consumer information (AGE, JOB, MARITAL STATUS, EDUCATION), current campaign information (CONTACT - communication type used to contact the consumer in the past, MONTH, DAY_OF_THE_WEEK, DURATION, CAMPAIGN), campaign history (PDAYS, PREVIOUS, POUTCOME), banking information (DEFAULT status, HOUSING, LOAN), and social economic information (EMP_VAR_RATE, CONS_PRICE_IDX, ERIBOR3M, NR_EMPLOYED). (see Appendix P - Variables Definitions)

Pre-Processing the Dataset

With the use of Python's Pandas library, we performed the exploratory analysis and data cleaning processes on the initial dataset. The full source included data of 4,119 records, with 20 independent variables spanning 2 years from 2008 to 2010. We began by eliminating any records with null values, which led to a reduction of 1,029 data points. As seen from the output below, a high concentration of the null values are in the default variable, which contains information of the individual having credit in default with

the bank. No interaction variables were found in the dataset, and the total quantity of independent variables remained static prior to the creation of dummy variables.

```
age                0
job               39
marital           11
education        167
default          803
housing          105
loan             105
contact            0
month              0
day_of_week        0
duration           0
campaign           0
pdays              0
previous           0
poutcome           0
emp.var.rate       0
cons.price.idx     0
cons.conf.idx      0
euribor3m          0
nr.employed        0
y                  0
dtype: int64
```

Addressing Qualitative Variables

Of the 20 variables contained in the dataset, 10 were qualitative in nature requiring conversion to numeric fields. We employed the dummy variable approach to create numeric variables associated with each qualitative attribute.

The variable “Job” was split into 11 independent variables:

```
job1=(job='blue-collar');
job2=(job='services');
job3=(job='admin.');
job4=(job='self-employed');
job5=(job='technician');
job6=(job='management');
job7=(job='retired');
job8=(job='entrepreneur');
job9=(job='housemaid');
job10=(job='unemployed');
job11=(job='student');
```

“Martial” was split into 3 independent variables:

```
marital1=(marital='married');
marital2=(marital='single');
marital3=(marital='divorced');
```

“Education” was split into 6 independent variables:

```
ed0=(education='illiterate');
ed1=(education='basic.4y');
ed2=(education='basic.6y');
ed3=(education='basic.9y');
ed4=(education='high.school');
ed5=(education='professional.course');
ed6=(education='university.degree');
```

“Month” was split into 9 independent variables because the study was conducted between March and December:

```
month3=(month='mar');
month4=(month='apr');
month5=(month='may');
month6=(month='jun');
month7=(month='jul');
month8=(month='aug');
month9=(month='sep');
month10=(month='oct');
month11=(month='nov');
month12=(month='dec');
```

“Day” was split into 5 independent variables (the study was conducted Monday through Friday):

```
day1=(day_of_week='mon');
day2=(day_of_week='tue');
day3=(day_of_week='wed');
day4=(day_of_week='thu');
day5=(day_of_week='fri');
```

“Previous” was split into 3 independent variables:

```
prev_outcome1=(poutcome='failure');
prev_outcome2=(poutcome='nonexistent');
prev_outcome3=(poutcome='success');
```

Dummy variables were also created for the binary qualitative attributes as outlined below:

```
credit_default=(default='yes');
housing_loan=(housing='yes');
has_loan=(loan='yes');
cellphone=(contact='cellular');
```

Model Approach and Validation

With our dependent variable being “Target” (output being either ‘Yes’ or ‘No’), we proceeded to: use logistic regression analysis to first fit and analyze our model with all significant variables, apply selection methods to find a subset of optimal predictors, remove outliers and strong influential points as needed, and analyze for multicollinearity. Based on this analysis, we also explored the effect of our model on a smaller subset of data by removing approximately 2,000 variables that had a “Target" output of “No” and PDAYS value of 999 that signifies that the client had not previously been contacted.

Analysis, Results and Findings

Exploratory Analysis

For the original dataset, we first evaluated the frequencies of the output of our dependent variable of interest to confirm that we have enough observations to perform our analysis. With our understanding of needing at least 10 - 30 observations per variable, we determined that at a minimum we would require 530 variables (since we have 53 variables now with the dummy variables we incorporated). We also noticed that the breakdown between the frequencies of ‘No’ vs ‘Yes’ is significant (88% vs 12% respectively). (see Appendix A - Dependent Variables Estimates)

Next, we performed logistic regression analysis on the full model, as follows:

MODEL target (event='1') = age duration campaign pdays previous emp_var_rate cons_price_idx cons_conf_idx euribor3m nr_employed job1 job2 job3 job4 job5 job6 job7 job8 job9 job10 job11 marital1 marital2 marital3 ed0 ed1 ed2 ed3 ed4 ed5 ed6 credit_default housing_loan has_loan cellphone month3 month4 month5 month6 month7 month8 month9 month10 month11 month12 day1 day2 day3 day4 day5 prev_outcome1 prev_outcome2 prev_outcome3

*log(target=1/target=0) = -140.5 + 0.00374*AGE + 0.00495*DURATION – 0.0983*CAMPAIGN – 0.00051*PDAYS + 0.0936*PREVIOUS – 0.8707*EMP_VAR_RATE + 1.3895*CONS_PRICE_IDX + 0.0574*CONS_CONF_IDX – 0.1219*EURIBOR3M + 0.00166*NR_EMPLOYED + 0*JOB1 – 0.140*JOB2 + 0.0103*JOB3 + 0*JOB4 + 0*JOB5 + 0*JOB6 -0.1491*JOB7 + 0*JOB8 + 0*JOB9 + 0*JOB10 + 0.1511*JOB11 + 0.2928*MARITAL1 + 0.3349*MARITAL2 + 0*MARITAL3 + 0*ED0 – 0.3319*ED1 + 0.02*ED2 – 0.0675*ED3 + 0*ED4 + 0*ED5 + 0*ED6 – 7.2688*CREDIT_DEFAULT + 0.0448*HOUSING_LOAN – 0.1193*HAS_LOAN + 1.1239*CELLPHONE + 2.0781*MONTH3 – 0.3086*MONTH4 - 0.7321*MONTH5 + 0.1957*MONTH6 – 0.4074*MONTH7 + 0.1525*MONTH8 – 0.2815*MONTH9 + 0.003*MONTH10 – 0.73*MONTH11 + 0*MONTH12 + 0.2922*DAY1 + 0.1155*DAY2 + 0.3240*DAY3 + 0.2963*DAY4 + 0*DAY5 – 0.5637*PREV_OUTCOME1 + 0*PREV_OUTCOME2 + 0.8015*PREV_OUTCOME3*

Where job1=1 when job='blue-collar',

job2=1 when job='services',

job3=1 when job='admin.',

job4=1 when job='self-employed',

job5=1 when job='technician',

job6=1 when job='management',

job7=1 when job='retired',

job8=1 when job='entrepreneur',

job9=1 when job='housemaid',

job10=1 when job='unemployed',

job11=1 when job='student',

marital1=1 when marital='married',

marital2=1 when marital='single',

marital3=1 when marital='divorced' (which could also mean widowed),

ed0=1 when education='illiterate',

ed1=1 when education='basic.4y',

ed2=1 when education='basic.6y',

ed3=1 when education='basic.9y',

ed4=1 when education='high.school',

ed5=1 when education='professional.course',

ed6=1 when education='university.degree',

credit_default=1 when default='yes',

housing_loan=1 when housing='yes',

has_loan=1 when loan='yes',

cellphone=1 when contact='cellular',

cellphone=0 when contact='telephone',

month3=1 when month='mar',

month4=1 when month='apr',

month5=1 when month='may',

month6=1 when month='jun',

month7=1 when month='jul',

month8=1 when month='aug',

month9=1 when month='sep',

month10=1 when month='oct',

month11=1 when month='nov',

month12=1 when month='dec',

day1=1 when day_of_week='mon',

day2=1 when day_of_week='tue',

day3=1 when day_of_week='wed',

day4=1 when day_of_week='thu',

day5=1 when day_of_week='fri',

prev_outcome1=1 when poutcome='failure',

prev_outcome2=1 when poutcome='nonexistent',

prev_outcome3=1 when poutcome='success',

target=1 when y='yes'.

From this analysis, we observed that there are a few variables that have a high p-value and would be considered insignificant (e.g. age, jobs, education, etc). The

parameters with the highest influence on the outcome are DURATION with a standard estimate of 0.7197 and EMP_VAR_RATE (employee variable rate) with a standard estimate of -0.7663. The top 2 predictors that have a significant effect on the odds of TARGET=1 (client subscribing to a long-term deposit) was CREDIT_DEFAULT with a parameter estimate of -7.2688 & MONTH3 (March) with estimate 2.078 (see Appendix B - Analysis of Maximum Likelihood Estimates). According to our results, the top 2 predictors with the highest influence are not considered insignificant.

Checking for Collinearity

We also did a preliminary analysis to check for any multicollinearity amongst the non-categorical variables. “The Pearson correlation coefficient is a linear correlation coefficient, which is used to reflect the linear correlation of two normal continuous variables” (Honghui and Yong, 11635) (see Appendix C - Pearson Correlation Coefficients).

From evaluating the Pearson Correlation Coefficients, we found a few noteworthy variables with strong multicollinearity: EURIBOR3M & EMP_VAR_RATE (Pearson correlation coefficient of 0.96753), EURIBOR3M & NR_EMPLOYED (Pearson correlation coefficient of 0.94228) and NR_EMPLOYED & EMP_VAR_RATE (Pearson correlation coefficient of 0.8906).

Selection Methods

After analyzing the full model, we ran a few selection methods to find the optimal model for our data set. We utilized backward and stepwise selection methods. The final outputs for these selection methods are as follows:

Backwards: DURATION, EMP_VAR_RATE, CONS_PRICE_IDX, CONS_CONF_IDX, CELLPHONE, MONTH3, MONTH6, MONTH8, PREV_OUTCOME3

Stepwise: DURATION, CONS_CONF_IDX, NR_EMPLOYED, CELLPHONE, MONTH3, MONTH5, MONTH6, PREV_OUTCOME3

Model Selection & Selection Criteria

In order to determine which model would be better to use, we compared specific selection criteria between the top 2 contending models: M1 (variables chosen by the backwards selection method) & M2 (variables chosen by stepwise selection method). The model returned by forward selection performed slightly worse than M1 & M2 in terms of selection criteria, & was hence dismissed from the comparison for best model.

The selection criteria we used to determine the best model consisted of R-Square (R2), Akaike Information Criterion (AIC), Schwarz Criterion (SC), and Likelihood Ratio (LR). We also compared the most influential predictors each model chose to analyze them at a more granular level.

For R2 and LR, we want the model with the highest value. The higher the R2 value, the more the predictors selected are said to explain the variation in the model. LR is similar to an F-test, which tests the global null hypothesis (H_0). This tests our Goodness-of-Fit. Parameters with P-values < 0.05 are considered significant in the prediction of the outcome of 1 for the Dependent Variable, and therefore reject the H_0.

For AIC and SC we look for the lowest value (see Appendix D - Comparison of Noteworthy Parameters). AIC (Akaike Information Criterion) and SC (Schwarz Criterion) are additional measures of goodness of fit. AIC and SC are affected (penalized) by the number of insignificant predictors in the model. The best fit model using these criteria explains the greatest amount of variation using the fewest possible variables.

Most of the predictors overlapped, with the exception of the ones highlighted in red above. MONTH5 (May) was seen as significant in the stepwise selection method. This is biased because it is the month with the most calls (981).

MONTH3 (March) followed as significant. This was the first month of the campaign and had the 2nd lowest frequency (42). Finally, MONTH6 (June) followed as significant. This month was close to the median, with frequency of calls at 365.

In order to check for multicollinearity, we used the CORRB option in our Logistic Regression's MODEL statement, which checks "the correlation of the coefficients of these variables in the model" (Slide 40, lecture 7). The final model chosen (M1) did not present multicollinearity among predictors. All the predictors selected had a correlation value of < 0.7.

Estimated Correlation Matrix										
Parameter	Intercept	duration	emp_var_rate	cons_price_idx	cons_conf_idx	cellphone	month3	month6	month8	prev_outcome3
Intercept	1.0000	-0.1395	0.6430	-0.9992	-0.1566	-0.1272	-0.0855	0.1918	-0.2040	0.2362
duration	-0.1395	1.0000	-0.3132	0.1316	0.0634	0.0603	0.1334	0.0799	0.1307	0.0467
emp_var_rate	0.6430	-0.3132	1.0000	-0.6330	0.1400	0.2620	0.0054	0.1965	-0.3663	0.2606
cons_price_idx	-0.9992	0.1316	-0.6330	1.0000	0.1917	0.1288	0.0886	-0.1954	0.1860	-0.2409
cons_conf_idx	-0.1566	0.0634	0.1400	0.1917	1.0000	0.3684	0.1638	0.0562	-0.4148	-0.1132
cellphone	-0.1272	0.0603	0.2620	0.1288	0.3684	1.0000	0.0738	0.2909	-0.2734	-0.0322
month3	-0.0855	0.1334	0.0054	0.0886	0.1638	0.0738	1.0000	0.0960	0.0351	0.0234
month6	0.1918	0.0799	0.1965	-0.1954	0.0562	0.2909	0.0960	1.0000	0.0240	0.1154
month8	-0.2040	0.1307	-0.3663	0.1860	-0.4148	-0.2734	0.0351	0.0240	1.0000	-0.0507
prev_outcome3	0.2362	0.0467	0.2606	-0.2409	-0.1132	-0.0322	0.0234	0.1154	-0.0507	1.0000

Backwards:

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-134.2	15.3303	76.5990	<.0001	
duration	1	0.00483	0.000276	305.2364	<.0001	0.7027
emp_var_rate	1	-0.9219	0.0731	159.2522	<.0001	-0.8113
cons_price_idx	1	1.4043	0.1644	72.9584	<.0001	0.4636
cons_conf_idx	1	0.0649	0.0149	19.0724	<.0001	0.1700
cellphone	1	1.2008	0.2495	23.1612	<.0001	0.3083
month3	1	2.5969	0.3971	42.7622	<.0001	0.1658
month6	1	0.7439	0.2348	10.0429	0.0015	0.1324
month8	1	0.6644	0.2438	7.4246	0.0064	0.1344
prev_outcome3	1	1.4984	0.2440	37.7175	<.0001	0.1646

Stepwise:

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	60.7606	4.8630	156.1134	<.0001	
duration	1	0.00480	0.000273	309.1623	<.0001	0.6985
cons_conf_idx	1	0.0485	0.0136	12.6810	0.0004	0.1269
nr_employed	1	-0.0124	0.000957	167.6207	<.0001	-0.5223
cellphone	1	1.0821	0.2322	21.7272	<.0001	0.2778
month3	1	2.0956	0.4874	26.4646	<.0001	0.1338
month5	1	-0.4122	0.1970	4.3793	0.0364	-0.1058
month6	1	1.2212	0.2456	24.7327	<.0001	0.2173
prev_outcome3	1	1.5582	0.2435	40.9466	<.0001	0.1712

Based on our selection criteria, the variables selected by Backwards selection returned t

		Full Model	Backwards Selection Method (M1)	Stepwise Selection Method (M2)
	n	3090	3090	3090
	k	54	9	
highest	RSQUARE	0.2697	0.2647	0.2637
lowest	AIC	1371.033	1334.452	1336.414
lowest	SC	1606.434	1394.811	1390.738
highest	LR	971.3739	1314.452	945.9924
(Most Influential Predictor) Standardized Estimates for:				
EMP_VAR_RATE		-0.7663	-0.8113	
CONS_PRICE_IDX		0.4488	0.4536	0.1269
MONTH5		-0.1879		-0.1058
MONTH8		0.0308	0.1344	
NR_EMPLOYED		0.0701		-0.5223

Outliers & Influential Points

That being said, our R2 value for the backward selection output is still considerably low (0.2647). In an attempt to improve the R2 value, we analyzed the possibility of removing significant outliers and influential points using SAS's IPLOTS and INFLUENCE options. With the amount of observations that we have in our dataset, we removed outliers that had a Pearson and Deviance Residual that was greater than 3.

This led to us removing the following data observations: 94, 190, 219, 406, 416, 860, 940, 969, 1012, 1093, 1012, 1119, 1323, 1448, 1499, 1500, 1580, 1693, 1797, 2143, 3082. Indeed removing these observations improved our R2, AIC, SC and LR values (see Appendix E - Logistic Regression on Backwards selection method's Predictors).

Improving the Model

We also explored the possibility of improving our model by testing it on a smaller version of the dataset. To do this, we removed approximately 2,000 records that had a dependent variable of 'No' and pdays = 999. We chose these parameters to remove observations from because we didn't have enough DV = 'yes' (our target class), and had an abundance of pdays for which the client had not been previously contacted (symbolized by the value 999). This decreased our breakdown of 'No' vs 'Yes' from a 88% to 12% (approximately) split to a 66% to 34% split. The final count for the smaller subset is 1,090 records.

Following the same steps as we did before, we performed full logistic regression on the variables within this smaller dataset (see Appendix F - Full Logistic Regression on Smaller Dataset and Appendix G - Analysis of Maximum Likelihood Estimates Smaller Dataset). Full model for the new subset:

*log(target=1/target=0) = -401.3 + 0.0136*AGE + 0.006*DURATION – 0.117*CAMPAIGN + 0.0015*PDAYS + 0.276*PREVIOUS – 1.759 *EMP_VAR_RATE + 3.16 *CONS_PRICE_IDX + 0.08*CONS_CONF_IDX – 0.55*EURIBOR3M + 0.02*NR_EMPLOYED + 0*JOB1 – 0.21*JOB2 0.09*JOB3 + 0*JOB4 + 0*JOB5 + 0*JOB6 -0.47*JOB7 + 0*JOB8 + 0*JOB9 + 0*JOB10 - 0.0009*JOB11 + 0.48*MARITAL1 + 0.63*MARITAL2 + 0*MARITAL3 + 0*ED0 – 1.0567*ED1 - 0.11*ED2 – 0.2*ED3 + 0*ED4 + 0*ED5 + 0*ED6 – 9.3*

**CREDIT_DEFAULT – 0.04 *HOUSING_LOAN – 0.04*HAS_LOAN + 1.45*CELLPHONE + 2.97*MONTH3 – 0.34*MONTH4 - 0.56*MONTH5 - 0.41*MONTH6 – 0.61*MONTH7 + 0.89*MONTH8 + 1.29 *MONTH9 + 0.9*MONTH10 – 0.5*MONTH11 + 0*MONTH12 + 0.39*DAY1 + 0.24*DAY2 + 0.43*DAY3 + 0.31*DAY4 + 0*DAY5 – 0.89*PREV_OUTCOME1 + 0*PREV_OUTCOME2 + 0.94*PREV_OUTCOME3*

Where job1=1 when job='blue-collar',

job2=1 when job='services',

job3=1 when job='admin.',

job4=1 when job='self-employed',

job5=1 when job='technician',

job6=1 when job='management',

job7=1 when job='retired',

job8=1 when job='entrepreneur',

job9=1 when job='housemaid',

job10=1 when job='unemployed',

job11=1 when job='student',

marital1=1 when marital='married',

marital2=1 when marital='single',

marital3=1 when marital='divorced' (which could also mean widowed),

ed0=1 when education='illiterate',

ed1=1 when education='basic.4y',

ed2=1 when education='basic.6y',

ed3=1 when education='basic.9y',

ed4=1 when education='high.school',

ed5=1 when education='professional.course',

ed6=1 when education='university.degree',

credit_default=1 when default='yes',

housing_loan=1 when housing='yes',

has_loan=1 when loan='yes',

cellphone=1 when contact='cellular',

cellphone=0 when contact='telephone',

month3=1 when month='mar',

month4=1 when month='apr',

month5=1 when month='may',

month6=1 when month='jun',

month7=1 when month='jul',

month8=1 when month='aug',

month9=1 when month='sep',

month10=1 when month='oct',

month11=1 when month='nov',

month12=1 when month='dec',

day1=1 when day_of_week='mon',

day2=1 when day_of_week='tue',

day3=1 when day_of_week='wed',

day4=1 when day_of_week='thu',

day5=1 when day_of_week='fri',

prev_outcome1=1 when poutcome='failure',

prev_outcome2=1 when poutcome='nonexistent',

prev_outcome3=1 when poutcome='success',

target=1 when y='yes'.

We then performed the same stepwise (M12) and backwards (M13) selection methods to find the optimal predictors for the model (see Appendix H - Comparison of Noteworthy Parameters Small Dataset). We found in our analysis the parameters improved significantly. The backwards selection method (M13) returned the best statistics of the 2 models, with the exception of SC value. We continued our analysis on each model in order to get a well-rounded idea of which model to choose.

Next, we determined if any multi-collinearity issues existed among the independent variables chosen. CONS_PRICE_IDX & EMP_VAR_RATE had a correlation value of -0.95. NR_EMPOLYED & CONS_PRICE_IDX had a correlation

value of 0.91. EMP_VAR_RATE & CONS_PRICE_IDX had a correlation value of 0.96. Since the number-of-employees predictor had one of the highest non-significant p-value (0.019), & is a quarterly indicator, we removed that one and reran our model. After that we no longer had a correlation value >.9 amongst our selected predictors, and the SC value increased slightly.

To further improve the models, we searched for and removed any outliers that had a Pearson and Deviance Residual greater than 3. This led to us to remove the following data observations: 1082, 1045, 418, 320, 318, 261, 248, 230, 26, 66, 11 (see Appendix I - Final Model - Influence Diagnostics). We repeated this process until we reached satisfactory selection criteria values (R2, AIC, SC and LR respectively). By this time, the statistics measuring which model was best became pretty close. We selected M12 as our final model since it had less predictors with about equal accuracy.

Model Validation

Satisfied with the improvement of our accuracy metrics on the final model, we proceeded to split the data into a training and testing set. 60% of the data was used for the training set, and the remainder for the test set. After comparing the metrics produced by the backwards & stepwise selection methods done on the training set, we proceeded to move forward with the Stepwise model for the training set since the scores between the 2 models were similar but the Stepwise model it used has less

observations, which is preferred. (see Appendix J - Comparison of Noteworthy Parameters Training Set).

The training dataset, after cleaning, contains 1,070 total observations, 364 of which are Y=1. The probability of getting 'Y' in the entire test set is P(Y=1/N) = 364/1,070 = 0.34 = 34%. That will be our cutoff/threshold value (see Appendix L - Target Variables Frequency). We have 706 samples of 'no' = 65.98% probability, and 364 samples of 'yes' = 34.02% probability. Odds(y=1) = 0.34/0.66, which means that the odds of event Y = 1 occuring is 0.515-to-1. Odds(y=0) = 0.66/0.34, results in the odds that event Y=0 occurs is 1.94-to-1. This means we have a higher chance of failure, in other words, it increases the odds of the client not subscribing to a long-term bank deposit.

Computing the predicted probability on SAS confirmed that the range 0.34-0.35 was optimal for getting the highest combined score of sensitivity & specificity (see Appendix M - ClassificationTable). This model is able to make true predictions for customers that will put in a deposit about 85.1% of the time based on the independent variables associated.

Sensitivity or Recall is the percentage of total relevant results correctly classified by the model. The formula is *TP/(TP+FN), which translates into 120/(120+21), which is approximately equal to 0.851, or 85.1%.*

Accuracy tells us how many predicted y's were correct among all cases.

Formula: *(TP+TN)/(TP+TN+FN+FP)*

(120 + 241)/(241+46+21+120) ≈ 0.843 = 84.3%

Precision is our “exactness”, or the amount of all the true positives against all the predicted positives.

Formula: *TP/(TP+FP)*

120/(120+46) ≈ 0.7229 = 72.3%

See Appendix N - Table of target by Predicted Y.

General Equation for Final Model:

*log(target=1/target=0) = 102.2 + 0.009*DURATION + 0.08*CONS_CONF_IDX – 0.02*NR_EMPLOYED + 1.08*CELLPHONE + 5*MONTH3 + 1.8*MONTH6 + 0.9*MONTH8*

Where cellphone=1 when contact='cellular',

cellphone=0 when contact='telephone',

month3=1 when month='mar',

month6=1 when month='jun',

month8=1 when month='aug',

target=1 when y='yes'.

Analysis of Coefficients

In order to interpret the effect of each independent variable on the target in the final logistic regression model we must first retransform the beta coefficients. In addition, we can only consider a single parameter’s effect on the dependent variable

holding all other variables constant. Consider marketing campaigns where the contact equals CELLPHONE, the beta coefficient in the final model is 1.0795. We would then calculate the constant e raised to the power of the coefficient, $e^{1.0795} = 2.9432$. We can then take the result less 1 multiplied by 100 to get, (2.9432 - 1)*100 = 194.32. We can now say that the percentage increase in likelihood that marketing campaigns target will result in a yes when contact equals CELLPHONE is 194.32% , holding all other variables constant, with a 95% confidence that the average increase is between 28% (1.28-1) * 100 and 576% (6.768 - 1) * 100.

Parameter	Estimate	$(e^B -1)*100$	95% Wald Confidence Limits		Re-transfromed Confidence Limits	
Duration	0.00935	0.94	1.008	1.011	0.80	1.10
cons_conf_idx	0.08370	8.73	1.033	1.145	3.30	14.50
nr_employed	(0.02030)	(2.01)	0.976	0.984	(2.40)	(1.60)
cellphone	1.07950	194.32	1.280	6.768	28.00	576.80
month3	4.99630	14,686.50	14.418	999.999	1,341.80	99,899.90
month6	1.80970	510.86	2.493	14.967	149.30	1,396.70
month8	0.93170	153.88	1.091	5.906	9.10	490.60

Future Work

Extracting the most beneficial characteristics of a dataset proves to be difficult due to the necessity of an in-depth analysis and investigation into the various independent variables and correlation values. However, in future works, valuable insight can be further derived from a more comprehensive dataset and sampling rate. Cohesively, avenues worth exploring may be those that expand on the applicability to

the broader financial markets. “...predicting customer behavior is one of the challenges of indirect marketing analysis. They also discussed big data visualization methods for the marketing industry…[that may] solve problems such as purchase behavior, review ratings, customer loyalty, customer, lifetime value, sales, profit, and brand visibility” (Alaa Abu-Srhan 1)”. Insights derived from the analysis performed on this dataset can have broad effects that benefit different facets of customer-oriented interactions. By expanding on this established framework, we can pursue ventures that bring enrichment towards decision-making processes to provide tangible business solutions. The authors that collected the original dataset noted their interest in future work of collecting “more client based data, in order to check if high quality predictive models can be achieved without contact-based information.” (Cortez et al. 5). Additionally, with the implementation of sophisticated classification algorithms such as random forests and artificial neural networks, a more accurate model may be produced to aid in the prediction of subscription rates for term deposits. In conclusion, the advancement of the modeling efforts can be taken down a plethora of paths, as long as the end-goal remains the focal point: finding the most optimal way of selling long-term bank deposits to customers.

Research and References

Alaa Abu-Srhan, Sanaa A zghoul, Bara'a Alhammad, Rizik Al-Sayyed. "Visualization and Analysis in Bank Marketing Prediction." Thesai.org, International Journal of Advanced Computer Science and Applications, 7 Nov. 2019, https://thesai.org/Downloads/Volume10No7/Paper_85-Visualization_and_Analysis_in_Bank_Direct_Marketing.pdf

Cortez, Paulo, et al. "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology". Conference Paper. October 2011. *Departamento de Sistemas de Informação DSI - Engenharia da Programação e dos Sistemas Informáticos*, EUROSIS-ETI, http://hdl.handle.net/1822/14838.

Honghui, Xu, and Deng Yong. *Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient*. 18 December 2017. *IEEE Xplore*, https://ieeexplore.ieee.org/abstract/document/8218753/citations.

Janioe. "Bank Marketing Dataset Predicting Term Deposit Subscriptions." *Kaggle*, Bank Marketing Dataset Bachmann https://www.kaggle.com/janiobachmann/bank-marketing-dataset.

Moro, Sergio, et al. “A data-driven approach to predict the success of bank telemarketing.” *Decision Support Systems*, vol. 62, no. June 2014, 2014, pp. 22-31. *Science Direct*, https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X?via%3Dihub

Onwuegbuzie, Anthony J., and Larry G. Daniel. *Uses and Misuses of the Correlation Coefficient*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. 1999. *Institute of Education Sciences*, https://files.eric.ed.gov/fulltext/ED437399.pdf.

Appendices

Appendix A - Dependent Variable's Frequency

Dependent Variable's Frequency

The FREQ Procedure

target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2720	88.03	2720	88.03
1	370	11.97	3090	100.00

Appendix B - Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-140.5	135.6	1.0740	0.3001	
age	1	0.00374	0.00882	0.1798	0.6715	0.0210
duration	1	0.00495	0.000287	297.7014	<.0001	0.7197
campaign	1	-0.0983	0.0515	3.6468	0.0562	-0.1408
pdays	1	-0.00051	0.000693	0.5317	0.4659	-0.0580
previous	1	0.0936	0.1846	0.2569	0.6122	0.0289
emp_var_rate	1	-0.8707	0.4984	3.0516	0.0807	-0.7663
cons_price_idx	1	1.3895	0.8884	2.4460	0.1178	0.4488
cons_conf_idx	1	0.0574	0.0286	4.0343	0.0446	0.1503
euribor3m	1	-0.1219	0.4826	0.0638	0.8006	-0.1190
nr_employed	1	0.00166	0.0111	0.0223	0.8814	0.0701
job1	0	0	.	.	.	.
job2	1	-0.1410	0.2991	0.2221	0.6374	-0.0222
job3	1	0.0103	0.1807	0.0033	0.9544	0.00255
job4	0	0	.	.	.	.
job5	0	0	.	.	.	.
job6	0	0	.	.	.	.
job7	1	-0.1491	0.3652	0.1668	0.6830	-0.0158
job8	0	0	.	.	.	.
job9	0	0	.	.	.	.
job10	0	0	.	.	.	.
job11	1	0.1511	0.4317	0.1226	0.7263	0.0110
marital1	1	0.2928	0.2654	1.2167	0.2700	0.0797
marital2	1	0.3349	0.2983	1.2609	0.2615	0.0852
marital3	0	0	.	.	.	.
ed0	0	0	.	.	.	.
ed1	1	-0.3319	0.3197	1.0773	0.2993	-0.0493
ed2	1	0.0200	0.4126	0.0023	0.9614	0.00237
ed3	1	-0.0675	0.2467	0.0749	0.7843	-0.0126
ed4	0	0	.	.	.	.
ed5	0	0	.	.	.	.
ed6	0	0	.	.	.	.
credit_default	1	-7.2688	445.6	0.0003	0.9870	-0.0721
housing_loan	1	0.0448	0.1508	0.0883	0.7663	0.0123
has_loan	1	-0.1193	0.2071	0.3320	0.5645	-0.0244
cellphone	1	1.1239	0.3071	13.3918	0.0003	0.2886
month3	1	2.0781	0.7223	8.2764	0.0040	0.1327
month4	1	-0.3086	0.7192	0.1841	0.6678	-0.0387
month5	1	-0.7321	0.6559	1.2457	0.2644	-0.1879
month6	1	0.1957	0.8207	0.0569	0.8115	0.0348
month7	1	-0.4074	0.7040	0.3349	0.5628	-0.0837
month8	1	0.1525	0.6435	0.0562	0.8126	0.0308
month9	1	-0.2815	0.7018	0.1609	0.6883	-0.0209
month10	1	0.00331	0.6612	0.0000	0.9960	0.000254
month11	1	-0.7337	0.6502	1.2735	0.2591	-0.1339
month12	0	0	.	.	.	.
day1	1	0.2922	0.2349	1.5468	0.2136	0.0654
day2	1	0.1155	0.2442	0.2238	0.6362	0.0254
day3	1	0.3240	0.2471	1.7194	0.1898	0.0718
day4	1	0.2963	0.2375	1.5562	0.2122	0.0658
day5	0	0	.	.	.	.
prev_outcome1	1	-0.5637	0.3162	3.1781	0.0746	-0.0997
prev_outcome2	0	0	.	.	.	.
prev_outcome3	1	0.8015	0.6997	1.3125	0.2519	0.0881

Appendix C - Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 3090
Prob > |r| under H0: Rho=0

	age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed
age	1.00000	0.04736 0.0085	-0.02728 0.1295	-0.05882 0.0011	0.06825 0.0001	-0.05613 0.0018	-0.03228 0.0728	0.08502 <.0001	-0.04769 0.0080	-0.07032 <.0001
duration	0.04736 0.0085	1.00000	-0.07997 <.0001	-0.05019 0.0053	0.02673 0.1374	-0.02916 0.1051	0.01943 0.2802	-0.03565 0.0475	-0.03044 0.0907	-0.04283 0.0173
campaign	-0.02728 0.1295	-0.07997 <.0001	1.00000	0.05558 0.0020	-0.08754 <.0001	0.18544 <.0001	0.14771 <.0001	0.02235 0.2142	0.16768 <.0001	0.16434 <.0001
pdays	-0.05882 0.0011	-0.05019 0.0053	0.05558 0.0020	1.00000	-0.58903 <.0001	0.27235 <.0001	0.05701 0.0015	-0.10262 <.0001	0.30306 <.0001	0.38053 <.0001
previous	0.06825 0.0001	0.02673 0.1374	-0.08754 <.0001	-0.58903 <.0001	1.00000	-0.40232 <.0001	-0.14689 <.0001	-0.02919 0.1047	-0.44674 <.0001	-0.50199 <.0001
emp_var_rate	-0.05613 0.0018	-0.02916 0.1051	0.18544 <.0001	0.27235 <.0001	-0.40232 <.0001	1.00000	0.74106 <.0001	0.16428 <.0001	0.96753 <.0001	0.89064 <.0001
cons_price_idx	-0.03228 0.0728	0.01943 0.2802	0.14771 <.0001	0.05701 0.0015	-0.14689 <.0001	0.74106 <.0001	1.00000	0.01022 0.5703	0.63030 <.0001	0.43335 <.0001
cons_conf_idx	0.08502 <.0001	-0.03565 0.0475	0.02235 0.2142	-0.10262 <.0001	-0.02919 0.1047	0.16428 <.0001	0.01022 0.5703	1.00000	0.25184 <.0001	0.09455 <.0001
euribor3m	-0.04769 0.0080	-0.03044 0.0907	0.16768 <.0001	0.30306 <.0001	-0.44674 <.0001	0.96753 <.0001	0.63030 <.0001	0.25184 <.0001	1.00000	0.94228 <.0001
nr_employed	-0.07032 <.0001	-0.04283 0.0173	0.16434 <.0001	0.38053 <.0001	-0.50199 <.0001	0.89064 <.0001	0.43335 <.0001	0.09455 <.0001	0.94228 <.0001	1.00000

Appendix D - Comparison of Noteworthy Parameters

Parameter	Parameter Objective (Highest or Lowest)	Full Model	Backwards Selection Method	Stepwise Selection Method
RSQUARE	highest	0.2697	0.2647	0.2637
AIC	lowest	1371.033	1334.452	1336.414
SC	lowest	1606.434	1394.811	1390.738
LR	highest	971.3739	1314.452	945.9924

Appendix E - Logistic Regression on Backwards selection Method's Predictors

Logistic Regression on Backward selection method's predictors

The LOGISTIC Procedure

Model Information	
Data Set	WORK.BANK_NEW2
Response Variable	target
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	3079
Number of Observations Used	3079

Response Profile		
Ordered Value	target	Total Frequency
1	0	2716
2	1	363

Probability modeled is target=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2235.572	1248.691
SC	2241.605	1309.015
-2 Log L	2233.572	1228.691

R-Square	0.2785	Max-rescaled R-Square	0.5398

Appendix F - Full Logistic Regression on Smaller Dataset

Number of Observations Read	1090
Number of Observations Used	1090

Response Profile		
Ordered Value	target	Total Frequency
1	0	720
2	1	370

Probability modeled is target=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1398.660	781.140
SC	1403.654	846.061
-2 Log L	1396.660	755.140

R-Square	0.4449	Max-rescaled R-Square	0.6159

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	641.5199	12	<.0001
Score	509.8412	12	<.0001
Wald	283.5306	12	<.0001

Appendix G - Analysis of Maximum Likelihood Estimates Smaller Dataset

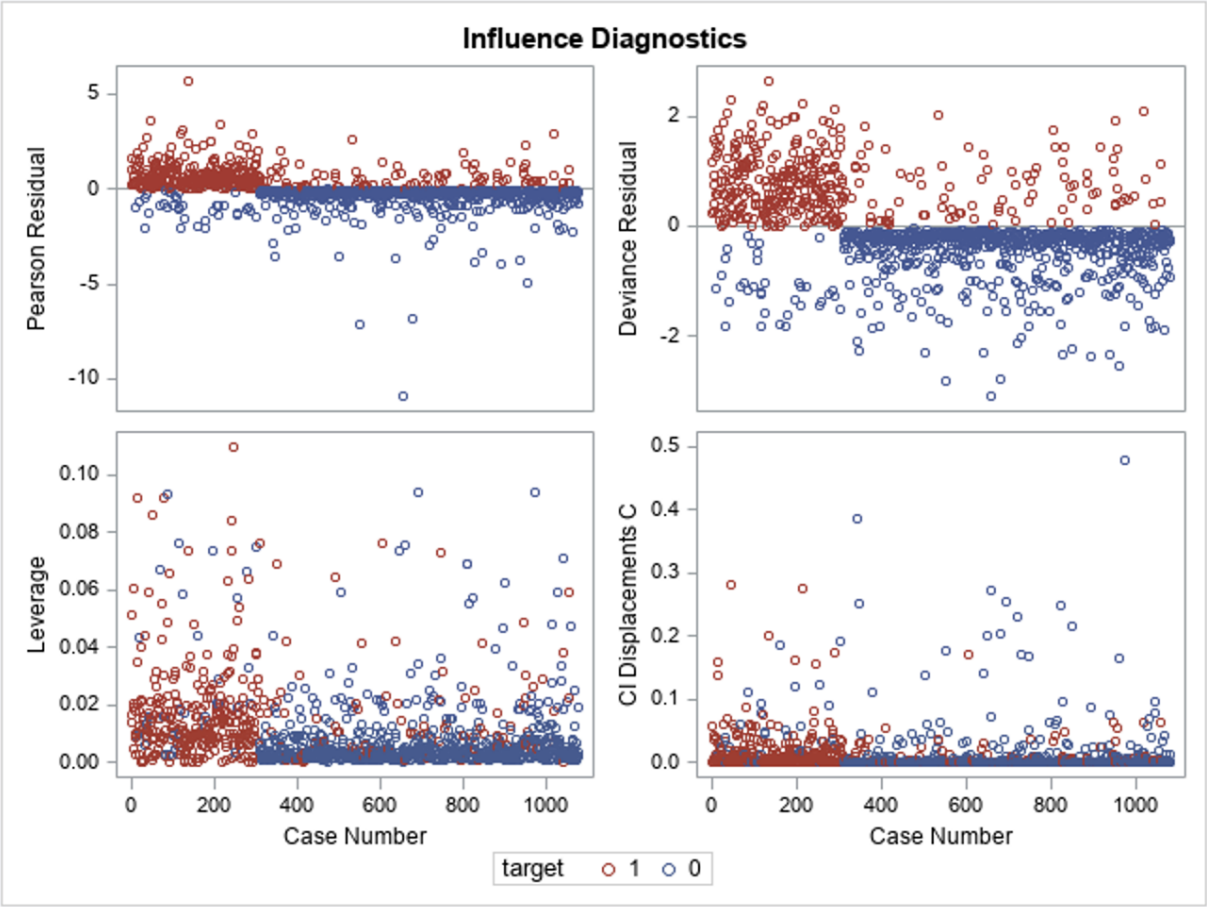
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-329.1	74.5783	19.4769	<.0001	
duration	1	0.00586	0.000432	184.0559	<.0001	1.0908
emp_var_rate	1	-1.9231	0.3476	30.6161	<.0001	-1.8057
cons_price_idx	1	2.8345	0.5346	28.1160	<.0001	0.9893
cons_conf_idx	1	0.0476	0.0209	5.2109	0.0224	0.1377
nr_employed	1	0.0119	0.00506	5.4985	0.0190	0.5694
ed1	1	-0.8356	0.3559	5.5115	0.0189	-0.1312
cellphone	1	1.3158	0.3069	18.3806	<.0001	0.3231
month3	1	3.2083	0.5762	31.0053	<.0001	0.2848
month8	1	1.5384	0.3654	17.7240	<.0001	0.3094
month9	1	1.4696	0.5583	6.9295	0.0085	0.1368
month10	1	1.1709	0.4816	5.9113	0.0150	0.1154
prev_outcome1	1	-0.6191	0.2610	5.6280	0.0177	-0.1139

Appendix H - Comparison of Noteworthy Parameters Small Dataset

Parameters	bank_small2 Full Model	bank_small2 Backwards (M7)	bank_small2 Stepwise (M8)
n	1090	1090	1090
k	54	12	10
RSQUARE	0.4538	0.4449	0.4402
AIC	815.37	781.14	786.36
SC	1010.135	846	841.294

LR	659.29	641.5	632.299

Appendix I - Final Model - Influence Diagnostics



Appendix J - Comparison of Noteworthy Parameters Training Set

	M11 (all data)	M12 (stepwise on training set)	M13 (backwards on training set)
n	1070	642?	
k	11	7	10
RSQUARE	0.51	0.53	0.5377
AIC	623	360	355.86
SC	682.95	395.78	405
LR	772.8	485	495.3

Appendix K - Final Model - Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	102.2	11.0797	85.0452	<.0001
duration	1	0.00935	0.000857	119.1233	<.0001
cons_conf_idx	1	0.0837	0.0263	10.1668	0.0014
nr_employed	1	-0.0203	0.00222	83.8033	<.0001
cellphone	1	1.0795	0.4249	6.4554	0.0111
month3	1	4.9963	1.1877	17.6962	<.0001
month6	1	1.8097	0.4572	15.6655	<.0001
month8	1	0.9317	0.4308	4.6782	0.0305

Appendix L - Target Variables Frequency

target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	706	65.98	706	65.98
1	364	34.02	1070	100.00

Appendix M - Classification Table

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.100	219	310	109	4	82.4	98.2	74.0	66.8	98.7
0.150	216	323	96	7	84.0	96.9	77.1	69.2	97.9
0.200	208	331	88	15	84.0	93.3	79.0	70.3	95.7
0.250	207	334	85	16	84.3	92.8	79.7	70.9	95.4
0.300	204	344	75	19	85.4	91.5	82.1	73.1	94.8
0.350	200	355	64	23	86.4	89.7	84.7	75.8	93.9
0.400	192	363	56	31	86.4	86.1	86.6	77.4	92.1
0.450	191	370	49	32	87.4	85.7	88.3	79.6	92.0
0.500	182	375	44	41	86.8	81.6	89.5	80.5	90.1
0.550	174	383	36	49	86.8	78.0	91.4	82.9	88.7
0.600	168	390	29	55	86.9	75.3	93.1	85.3	87.6
0.650	157	396	23	66	86.1	70.4	94.5	87.2	85.7
0.700	146	402	17	77	85.4	65.5	95.9	89.6	83.9
0.750	137	405	14	86	84.4	61.4	96.7	90.7	82.5
0.800	125	407	12	98	82.9	56.1	97.1	91.2	80.6

Classification table with calculated total for specificity & sensitivity

prob level	sensitivity (Y=1)	specificity	sensitivity + specificity
0.1	98.2	74	172.2
0.15	96.9	77.1	174
0.2	93.3	79	172.3
0.25	92.8	79.7	172.5
0.3	91.5	82.1	173.6
0.35	89.7	84.7	174.4
0.4	86.1	86.6	172.7
0.45	85.7	88.3	174
0.5	81.6	89.5	171.1

0.55	78	91.4	169.4
0.6	75.3	93.1	168.4
0.65	70.4	94.5	164.9
0.7	65.5	95.9	161.4
0.75	61.4	96.7	158.1
0.8	56.1	97.1	153.2

Appendix N - Table of Target Predicted Y

Table of target by pred_y

target	pred_y 0	pred_y 1	Total
0	241 TN	46 FP	287
1	21 FN	120 TP	141
Total	262	166	428

Appendix O - Odds Ratio Estimates

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
duration	1.009	1.008	1.011
cons_conf_idx	1.087	1.033	1.145
nr_employed	0.980	0.976	0.984
cellphone	2.943	1.280	6.768
month3	147.870	14.418	>999.999
month6	6.109	2.493	14.967
month8	2.539	1.091	5.906

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	95.0	Somers' D	0.901
Percent Discordant	5.0	Gamma	0.901
Percent Tied	0.0	Tau-a	0.409
Pairs	93437	c	0.950

Appendix P - Variables Definitions

Variable Name	Variable Type	Variable Definition
Age	Numerical	The age of the client
Duration	Numerical	The duration of the last contact with the client in seconds
Campaign	Numerical	Number of contacts performed during this campaign and for the client
pdays	Numerical	Number of days since the client was last contacted from another campaign
Previous	Numerical	Number of contacts performed before this campaign and for the client
Emp.var.rate	Numerical	The variation rate of employment (quarterly)
Cons.price.idx	Numerical	The consumer price index (monthly)
Euribor3m	Numerical	Euribor three month rate (daily)
Nr.employed	Numerical	Number of employees (quarterly)
Job	Categorical	Type of job the client has • admin • blue-collar • Entrepreneur • Housemaid • Management • retired • self-employed • Services • Student • Technician • Unemployed • Unknown

Marital Status	Categorical	Marital status of the client • divorced || widowed • married • single • unknown
Education	Categorical	Education status of the client • basic.4y • basic.6y • basic.9y • high.school • illiterate • professional.course • university.degree • unknown
Default	Categorical	Whether the client has credit in default • no • yes • unknown
Housing	Categorical	Whether the client has a housing loan? • no • yes • unknown
Loan	Categorical	Whether the client has a personal loan? • no • yes • unknown

Contact	Categorical	Communication type for the client • cellular • telephone
Month	Categorical	Month of the last contact • [jan, feb, mar, ..., nov, dec]
Day_of_week	Categorical	Last contact day of the week • [mon, tue, wed, thu, fri]
Poutcome	Categorical	Outcome of the previous campaign • failure • nonexistent • success