12-17-2014

A Comparison of Two Modeling Techniques in Customer Targeting For Bank Telemarketing

Hong Tang

A COMPARISON OF TWO MODELING TECHNIQUES IN CUSTOMER TARGETING FOR BANK TELEMARKETING

by

HONG TANG

Under the Direction of Gengsheng Qin, PhD

ABSTRACT

Customer targeting is the key to the success of bank telemarketing. To compare the flexible discriminant analysis and the logistic regression in customer targeting, a survey dataset from a Portuguese bank was used. For the flexible discriminant analysis model, the backward elimination of explanatory variables was used with several rounds of manual re-defining of dummy variables. For the logistic regression model, the automatic stepwise selection was performed to decide which explanatory variables should be left in the final model. Ten-fold stratified cross validation was performed to estimate the model parameters and accuracies. Although employing different sets of explanatory variables, the flexible discriminant analysis model and the logistic regression model show equally satisfactory performances in customer classification based on the areas under the receiver operating characteristic curves. Focusing on the predicted “right” customers, the logistic regression model shows slightly better classification and higher overall correct prediction rate.

INDEX WORDS: AUC, Discriminant analysis, KS test, Logistic regression, ROC

A COMPARISON OF TWO MODELING TECHNIQUES IN CUSTOMER TARGETING FOR BANK TELEMARKETING

by

HONG TANG

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2014

A COMPARISON OF TWO MODELING TECHNIQUES IN CUSTOMER TARGETING FOR BANK TELEMARKETING

by

HONG TANG

Committee Chair: Gengsheng Qin

Committee: Satish Nargundkar

Xin Qi

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2014

ACKNOWLEDGEMENTS

There are many people who have helped me during my thesis research. First, I would like to thank my advisor Dr. Gengsheng Qin, Professor and Graduate Director of Statistics in Department of Mathematics and Statistics at GSU, for his comprehensive guidance in sample survey, constructive suggestions in cutoff point optimizing and helpful advice in my thesis writing. I am especially grateful for his moral support and continuous encouragement during my study at GSU. Next, I want to express my sincere appreciation to Dr. Satish Nargundkar, Clinical Associate Professor in Department of Managerial Sciences at GSU, for his valuable guidance on data mining and the application of quantitative methods for strategic decision support. His instructions in discriminant analysis and logistic regression are extremely helpful for my thesis. Last but not the least, I want to thank Dr. Xin Qi, for his detailed instructions and continuous help in R coding techniques, linear regression and cross validation. The heuristic education method he used encouraged me to explore various computational methods in modeling.

Thank Dr. Guantao Chen for offering me the opportunity to enroll in Master of Science in Mathematics Program at GSU. I would also like to thank Dr. Marko Samara and Dr. James Michael Bowling for their careful instruction in SAS programming. Thank Dr. Sérgio Moro (ISCTE-IUL), Dr. Paulo Cortez (Univ. Minho) and Dr. Paulo Rita (ISCTE-IUL) for sharing the data set and thank UCI Machine Learning Repository for providing the data sharing platform.

Thank my friends at GSU. Their wisdom and enthusiasm have been a great encouragement to me. Finally, I am indebted to my family members who always believe in me and support me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Background Information

1.1.1 Bank telemarketing

Banks introduced the call center channel in the early 1980s with the aim of reducing overall servicing costs [Gupta et al, 2008]. It seemed reasonable that if the low-value, basic transactions were eliminated from the high-touch, high-cost, traditional branch banking channel, branch employees then could put more focus on revenue generation. However, in reality, revenues and profits were falling down [Gupta et al, 2008]. Therefore, banks needed to find call centers a new and profitable mission, which was telemarketing.

Marketing operationalized through a contact center, which allows communicating with customers through telephone channels, is called telemarketing due to the remoteness characteristic [Kotler and Keller, 2012]. One of the advantages of telemarketing is that it can centralize customer remote interactions in a contact center and thus ease operational management of campaigns. According to the "Bank Marketing Survey Report-2000" released by the American Banking Association/Bank Marketing Association, banks had sharply increased telemarketing [Albro and Linsley, 2001].

However, it is of vital importance to find the “right” customers to make sure the success of bank telemarketing, because if the contacted customer did not want the product, the outbound calls would be considered intrusive and inbound calls loaded with too much campaign content would also be annoying. Thus, more focus should be put on the task of selecting the best set of clients or targeting the right segments of customers, i.e., those who are more likely to subscribe a product [Moro et al, 2014].

Lau et al described the potential usefulness of data mining techniques in marketing [Lau et al, 2004]. Martens and Provost identified clients for targeting at a major bank using pseudo social networks based on relations (money transfers between stakeholders) [Martens and Provost, 2011]. However, none of them used real-data to test their results. In 2014, Sérgio Moro et al proposed a data mining approach to predict the success of telemarketing calls for selling bank long-term deposits [Moro et al, 2014]. The data collected from 2008 to 2013 by a Portuguese retail bank was addressed. For evaluation purposes, a time ordered split was initially performed, where the records were divided into training (four years) and test data (one year). The training data including all contacts executed up to June 2012, in a total of 51,651 examples, was used for model generation and selection. The test data, including the more recent 1293 contacts, from July 2012 to June 2013, was used to measure the prediction capabilities of the selected models. A large set of 150 features related with bank client, product and social-economic attributes was semi-automatic selected in the modeling phase and a final set of 22 features was achieved. Four data mining models were compared, including logistic regression (LR), decision trees (DTs), neural network (NN), and support vector machine (SVM) models. The area under the curve (AUC) and area of the LIFT cumulative curve (ALIFT) of the four models were compared on the test data using a rolling window scheme. The NN presented the best results (AUC = 0.8 and ALIFT = 0.7), allowing to reach 79% of the subscribers by selecting the half better classified clients [Moro et al, 2014].

1.1.2 Discriminant analysis

Discriminant analysis is a statistical analysis to predict a categorical dependent variable by one or more numerical or categorical independent variables.

Fisher (1936) was the first to suggest that classification should be based on a linear combination of the discriminating variables. He proposed using a linear combination which maximizes group difference while minimizing variation within the groups [Klecka, 1980]. In 1968, Edward I. Altman, who was the first to apply linear discriminant analysis for the case of the corporate credit granting problem, constructed the so-called Z value, which is a linear combination of several explanatory variables, including Sales/Total assets (TA), Working capital/TA, Retained Earnings/TA, Earnings before Interest and Taxation/TA, and Market Value of Equity/Book Value of Total Debt [Altman, 1968]. The Z value formula was designed to predict the probability that a firm will go into bankruptcy within two years as an easy-to-calculate control measure for the financial distress status of companies in academic studies. The model was found to be extremely accurate in correctly predicting bankruptcy [Altman, 1968].

Classical linear discriminant analysis requires the assumptions of normality, linearity, homoscedasticity and independence of errors [Meyers et al, 2013]. In practice, it is very rare that all these assumptions can be met. Furthermore, nonlinear boundaries can be more effective than linear decision boundaries in the real world. Hastie et al thus proposed the flexible discriminant analysis (FDA), which used nonparametric regression procedures to estimate nonlinear boundaries for classification [Hastie et al, 1994]. They demonstrated that linear discriminant analysis is equivalent to multi-response linear regression using optimal scoring to represent the groups. The linear predictors define one set of variables, and a set of dummy variables representing class membership defines the other set. Making use of the well-known fact that linear discriminant analysis is equivalent to canonical correlation analysis, the solution to the scoring problem can be found by canonical correlation analysis [Hastie et al, 1994].

1.1.3 Logistic regression

In general, logistic regression (LR) provides a method for modeling a binary dependent variable, which takes values 1 and 0, from a set of independent variables. The logit function, defined as the natural logarithm (ln) of the odds, is used to transform an 'S'-shaped curve into an approximately straight line and to change the range of the proportion from 0–1 to -∞ to +∞ [Bewick et al, 2005]. Let Y be the dependent variable and Xs be the independent variables. The LR model can be written as

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where π $= \text{Probability}\left(Y \ = outcome\ of\ interest \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right)$.

The following assumptions need to be satisfied for LR [Anderson, 2001]:

1) The dependent variable should be discrete (mostly dichotomous).
2) Observations are independent.
3) The sample size is large.
4) No severe multicollinearity among the independent variables.
5) The independent variables are linearly related to the ln odds of the event.

The main assumption required for LR is the last one, which involves two aspects. One is that logit function is the correct link function, and the other is that the logit function is a linear combination of the predictors.

The first three assumptions are often easily satisfied in the real business world. The fourth assumption can be satisfied by principle component analysis or by removing redundant independent variables if strong multicollinearity exists. LR is much more flexible than the discriminant analysis, because unlike the linear discriminant analysis, LR does not have the requirements of the independent variables to be normally distributed, linearly related to

dependent variable, or equal variance within each group [Liong and Foo, 2013]. Being free from the assumption of the linear discriminant analysis, LR is a useful tool in many situations. Therefore, LR is extensively used in various marketing related fields, such as consumer behavior, management, planning, strategy, channel of distribution, pricing, sales promotion, advertising, and educational issues [Leonard, 1998].

1.2 Purpose of the study

There are many statistical modeling techniques that can be used for prediction and customer classification. Moro et al have compared the AUC and ALIFT of LR, DT, NN, and SVM models [Moro et al, 2014] using a real bank telemarketing dataset. Both flexible discriminant analysis (FDA) and LR can also be used for classification. The first goal of this thesis is to complement Moro's study by using the similar dataset to compare the FDA and the LR in helping telemarketing campaign managers classify the "right" customers.

Moro et al used 2/3 of the whole dataset to build the models and used a rolling window scheme for model generation and selection. However, this method decreased the model accuracy estimation [Kohavi, 1995]. The second goal of this thesis is to use the 10-fold stratified cross validation method to acquire unbiased estimation of the model accuracy for model comparison.

The comparison results of the FDA and LR models in customer targeting shown in this thesis can be used for model selection in bank telemarketing.

2 DATA AND METHOD

2.1 Data preparation

The dataset used in this thesis was related with direct marketing campaigns of a Portuguese banking institution and is available at UCI Machine Learning Repository (please see details at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#). The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The data from the bank was enriched by the addition of five social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb.

The dataset, which is close to the data analyzed in by Moro et al [Moro et al, 2014], is composed of 41,188 observations and 21 attributes in this thesis. A total of 52,944 phone contacts of a Portuguese retail bank were addressed, with data collected from 2008 to 2013. Although the observations were ordered by date (from May 2008 to November 2010), the “date” variable was not included in the online dataset and was therefore not included in this thesis.

The details about the dependent and independent variables of the dataset are listed in the following table. The independent variables were sub-grouped by their sources or practical meanings.

Table 2.1 Attribute Information

Independent			
Bank client			
1	Age	Age of the client	Numeric
2	Job	Type of job	Categorical
3	Marital	Marital status ('divorced' means divorced or widowed)	Categorical
4	Education	Education level	Categorical
5	Default	Has credit in default?	Categorical
6	Housing	Has housing loan?	Categorical
7	Loan	Has personal loan?	Categorical
Current campaign			
8	Contact	Contact communication type	Categorical
9	Month	Last contact month of year	Categorical
10	Day_of_week	Last contact day of the week	Categorical
11	Duration	Last contact duration, in seconds. Important note: (if Duration=0 then Y='no').	Numeric
12	Campaign	Number of contacts performed during this campaign and for this client (includes the last contact)	Numeric
Previous campaign			
13	Pdays	Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)	Numeric
14	Previous	Number of contacts performed before this campaign and for this client	Numeric
15	Poutcome	Outcome of the previous marketing campaign	Categorical
Social and economic context			
16	Emp.var.rate	Employment variation rate - quarterly indicator	Numeric
17	Cons.price.idx	Consumer price index - monthly indicator	Numeric
18	Cons.conf.idx	Consumer confidence index - monthly indicator	Numeric
19	Euribor3m	Euribor 3 month rate - daily indicator	Numeric
20	Nr.employed	Number of employees - quarterly indicator (thousands)	Numeric
Dependent			
1	Y	Has the client subscribed a term deposit? (binary: 'yes','no')	Categorical

The raw data was read in R from the CSV format and each column was labeled with the appropriate variable names. In the original dataset, missing values were shown as “unknown”, and this notation was kept in later analysis.

Our intention here is to compare two realistic models in predicting if a customer will subscribe the long term deposit without the last contact. The duration is not known before a call is performed. After the last call, the result is obviously known. Thus, “duration” could not be used as an independent variable and was excluded after the data was read in.

Next, to get an overview of all the attributes and to check abnormal observations and outliers, the “summary” and “boxplot” functions in R were used. To detect if the numerical independent variables follow normal distributions, the histograms for each independent variable in subscribers and non-subscribers were generated. A correlation matrix for all the numerical variables was generated. Independent variables (“Euribor3m” and “Emp.var.rate rea”) with high correlations (>=0.7) with other variables were removed. The correlation matrix and the variance inflation factors for the rest numerical variables were calculated.

The “campaign” variable indicates the number of contacts performed during this campaign for this client, including the last contact. In reality, we want to predict the result without the last contact. Thus, the value of “campaign” variable minus 1 was used in later steps.

Then, the whole dataset was randomly split into training and validation datasets (50/50) to build the models. Lastly, 10-fold stratified cross validation was performed to estimate the coefficients and model performance indices [Kohavi, 1995].

2.2 The flexible discriminant analysis model

2.2.1 Definition of the dummies

For FDA, the dependent variable Y was replaced with “good” (if Y=yes, good=1, bad=0) and “bad” (if Y=no, bad=1, good=0) variables for grouping and scoring. The observations in training set were divided into 5-10% groups if possible by the distribution of each numerical independent variable. Two-way frequency tables were generated in SAS to calculate the “Good/Bad ratio” for each numerical independent variable group. After that, the dummy breakpoints (the average ratio difference between two dummies is greater than or equal to 0.1) for each numerical independent variable were created. The dummy variable creation for a numerical variable is shown in Table 2.2. For categorical independent variables, no grouping based on Good/Bad ratio was performed. An example of dummy variable creation for a categorical variable is shown in Table 2.3.

Table 2.2 An example of dummy variable creation for a numerical variable in training set*

Nr_employed	Bad%	Good%	Total%	Good/Bad	Dummy
<=5008.7	2.68	22.31	4.92	8.32	Employ3
5008.8-5076.2	4.85	24.14	7.05	4.98	Employ2
5076.3-5099.1	20.08	23.58	20.48	1.17	Employ1
5099.2-5191	20.90	5.62	19.16	0.27	Neutral
5191.1-5195.8	9.56	5.15	9.06	0.54	
5195.9-5228.1	41.93	19.2	39.34	0.46	

* Bad%=% of (number of non-subscribers in each class/number of non-subscribers in training set)
Good%=% of (number of subscribers in each class/number of subscribers in training set)
Total%=% of (number of customers in each class/number of customers in training set)
Good/Bad=the ratio of Good%/Bad%

Table 2.3 An example of dummy variable creation for a categorical variable in training set*

Poutcome	Bad%	Good%	Total%	Dummy
Failure	10.05	13.48	10.44	Pout1
Nonexistent	88.68	66.7	86.18	Neutral
Success	1.27	19.83	3.38	Pout2

* Bad%=% of (number of non-subscribers in each class/number of non-subscribers in training set)
Good%=% of (number of subscribers in each class/number of subscribers in training set)
Total%=% of (number of customers in each class/number of customers in training set)

2.2.2 Variable selection and coefficient estimation

All of the 58 created dummy variables in the training set (50/50 split) were used to build a full model with OLS method in SAS. A backward elimination was performed manually by eliminating the variables with the highest p-values to make sure all the variables left in the final model were significant at $p<0.05$. A re-defining of the dummy variables was performed each iteration. The regression outputs of the final models are displayed in Appendix A. Once the explanatory variables were decided in the final model, the 10 training sets created by 10-fold stratified cross validation method were used to "re-train" the model, and the estimated coefficients from each training set were saved and averaged.

2.2.3 Prediction

The scoring, which was actually a prediction process, was completed with the score procedure in SAS. The score range was set from 0 to 1. The scoring procedure was performed for the 10 training data sets and the 10 validation data sets.

2.2.4 Calculate average cumulative count, average cumulative percent and average correct prediction rate

The average cumulative count, average cumulative percent and average correct prediction rates at various cutoff points were calculated using score results from the training and validation datasets in Excel.

TP, or true positive, stands for the number of the customers who predicted by the model would subscribe the product (predicted “Yes”) and actually subscribed the product in reality (Yes), too. FP, or false positive, stands for the number of the customers who predicted by the model would subscribe the product but actually did subscribe the product in reality (No). FN, or false negative, stands for the number of the customers who predicted by the model would not subscribe the product (predicted “No”) but actually subscribed the product. TN, or true negative, stands for the number of customers who predicted by the model would not subscribe the product and actually did not subscribe the product. TPR and FPR are the percentages of TP and FP divided by the total number of customers that subscribed the product. FNR and TNR are the percentages of TN and FN divided by the total number of customers that did not subscribe the product.

2.2.5 Determine the optimal score cutoff point

The optimal score cutoff point was determined by the Kolmogorov–Smirnov (KS) distance between average TPR and average FPR in the training set. At the optimal score cutoff point, the model can best separate real Yes and No in the predicted "Yes" group.

In determining the optimal cutoff point, the KS distance is basically the same as the Youden Index (J). The former is widely used in credit scoring models, while the latter is commonly used in medical diagnosis.

KS = max (Cumulative Percent of Yes in Predicted "Yes"-Cumulative Percent of No in Predicted "Yes") = max (TPR-FPR) = max (sensitivity+specificity-1) = J

2.3 The logistic regression model

2.3.1 Definition of the dummies

No dummy variables were generated for numerical independent variable for the LR. For LR scoring, the dummy variables generated for categorical independent variables in the FDA were used, and the dependent variable Y was replaced with a dummy variable, "sub", which was a short name for subscribers.

2.3.2 Variable selection and coefficient estimation

The 50/50 split training set was used to build the model using SAS proc logistic command. The automatic stepwise selection was performed in SAS. A significance level of 0.3 (SLENTRY=0.3) was required to allow a variable into the model, and a significance level of 0.05 (SLSTAY=0.05) was required for a variable to stay in the final model. The SAS regression outputs were displayed in Appendix B. Similarly to the method described in 2.2.2, the averaged coefficients were estimated from the 10 training sets.

2.3.3 Prediction

The predictions performed for the 10 training data sets and the 10 validation data sets using the final models built from the training data sets were completed by the score option of SAS proc logistic command.

2.3.4 Calculate average cumulative count, average cumulative percent and average correct prediction rate

The average cumulative count, average cumulative percent and average correct prediction rates at various cutoff points were calculated using score results from the training and validation datasets in Excel.

2.3.5 Determine the optimal score cutoff point

The optimal score cutoff point was determined by the KS distance between average TPR and average FPR in the training sets.

2.4 The comparison of the two models

2.4.1 The prediction accuracies

TPRs, predicted "Yes"/total, and overall correct prediction rates at the optimal cutoff point of the FDA and LR models from the 10 validation sets were subjected to Shapiro-Wilk test for normality test and were then compared by paired t-test. Averaged results are shown in the tables.

2.4.2 The receiver operating characteristic curves and the areas under the curves

The receiver operating characteristic (ROC) curves for the 10 training sets and 10 validation sets were plotted using the average TPRs and average TNRs at various cutoff points. The ROC curves for the FDA and LR models were plotted together for comparison. For each validation set, the ROC curve was plotted from its TPR and TNR from FDA or LR model. The

AUCs calculated from the 10 training sets and 10 validation sets were subjected to Shapiro-Wilk test for normality test. For the training sets, Wilcoxon signed-rank test was performed to see if there was statistical significant difference between the average AUC of the FDA model and that of LR model. For the validation sets, paired t-test was performed to see if there was statistical significant difference between the average AUC of the FDA model and that of LR model.

3 RESULTS

3.1 Data characteristics

3.1.1 Descriptive statistics

Table 3.1 gives an overview of all the attributes. There are no obvious abnormal observations or considerable large numbers of missing values. But skewness and outliers are found in the numerical independent variables (Figure 3.1 and Table 3.1). Also, the dependent variable is unbalanced, with only about 10% customers subscribed and about 90% did not subscribe.

Table 3.1 Summary of attributes

Dependent variable					
Variable	Value	Count			
Y	No	36548			
	Yes	4640			
Categorical independent variables					
Variable	Value	Count	Variable	Value	Count
Job	Admin.	10422	Loan	No	33950
	Blue-collar	9254		Unknown	990
	Technician	6743		Yes	6248
	Services	3969	Contact	Cellular	26144
	Management	2924		Telephone	15044
	Retired	1720	Month	May	13769
	(Other)	6156		Jul	7174
Day_of_week	Fri	7827		Aug	6178
	Mon	8514		Jun	5318
	Thu	8623		Nov	4101
	Tue	8090		Apr	2632
	Wed	8134		(Other)	2016
Education	University.degree	12168	Marital	Divorced	4612
	High.school	9515		Married	24928
	Basic.9y	6045		Single	11568
	Professional.course	5243		Unknown	80
	Basic.4y	4176	Poutcome	Failure	4252
	Basic.6y	2292		Nonexistent	35563
	(Other)	1749		Success	1373
Default	No	32588	Housing	No	18622
	Unknown	8597		Unknown	990
	Yes	3		Yes	21576

Numerical independent variables						
Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	17	32	38	40.02	47	98
Campaign	1	1	2	2.568	3	56
Pdays	0	999	999	962.5	999	999
Previous	0	0	0	0.173	0	7
Emp.var.rate	-3.4	-1.8	1.1	0.08189	1.4	1.4
Cons.price.idx	92.2	93.08	93.75	93.58	93.99	94.77
Cons.conf.idx	-50.8	-42.7	-41.8	-40.5	-36.4	-26.9
Euribor3m	0.634	1.344	4.857	3.621	4.961	5.045
Nr.employed	4964	5099	5191	5167	5228	5228

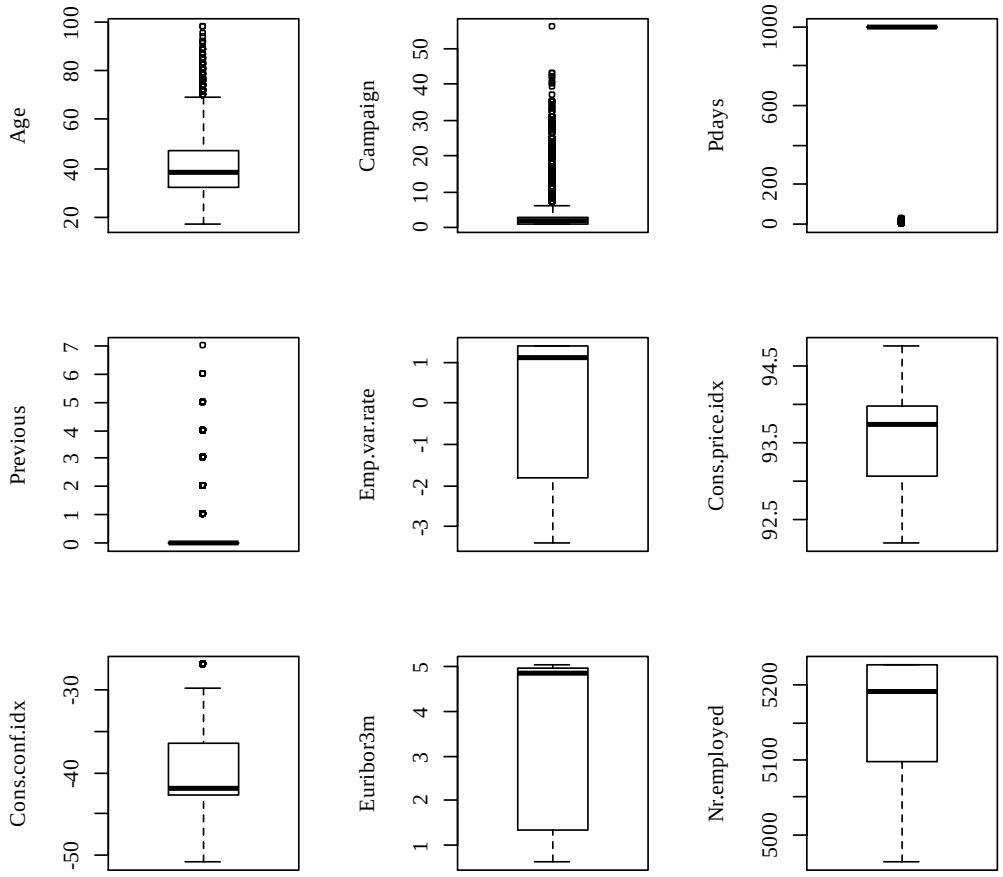

Figure 3.1 Boxplots of numerical independent variables

* Pdays: Number of days that passed by after the client was last contacted from a previous campaign. 999 means client was not previously contacted.

Campaign: Number of contacts performed during this campaign and for this client, including the last contact.

Previous: Number of contacts performed before this campaign and for this client

Poutcome: Outcome of the previous marketing campaign

Emp.var.rate: Employment variation rate - quarterly indicator

Cons.price.idx: Consumer price index - monthly indicator

Cons.conf.idx: Consumer confidence index - monthly indicator

Euribor3m: Euribor 3 month rate - daily indicator

Nr.employed: Number of employees - quarterly indicator (thousands)

3.1.2 Correlation

There is strong correlation between some numerical independent variables. For example, Emp.var.rate shows high correlation with Cons.price.idx, Euribor3m, and Nr.employed.

Euribor3m has high correlation with Emp.var.rate and Cons.price.idx. The correlation greater than or equal to 0.7 were highlighted in yellow in the following correlation matrix.

Table 3.2 Correlation matrix of all numerical independent variables

Correlation	Age	Campaign	Pdays	Previous	Emp.var.rate	Cons.price.idx	Cons.conf.idx	Euribor3m	Nr.employed
Age	1.0								
Campaign	0.0	1.0							
Pdays	0.0	0.1	1.0						
Previous	0.0	-0.1	-0.6	1.0					
Emp.var.rate	0.0	0.2	0.3	-0.4	1.0				
Cons.price.idx	0.0	0.1	0.1	-0.2	0.8	1.0			
Cons.conf.idx	0.1	0.0	-0.1	-0.1	0.2	0.1	1.0		
Euribor3m	0.0	0.1	0.3	-0.5	1.0	0.7	0.3	1.0	
Nr.employed	0.0	0.1	0.4	-0.5	0.9	0.5	0.1	0.9	1.0

Table 3.3 Correlation matrix and variance inflation factors of the numerical independent variables considered in modeling

Correlation	Age	Campaign	Pdays	Previous	Cons.price.idx	Cons.conf.idx	Nr.employed	VIF
Age	1.0							1.0
Campaign	0.0	1.0						1.0
Pdays	0.0	0.1	1.0					1.6
Previous	0.0	-0.1	-0.6	1.0				1.8
Cons.price.idx	0.0	0.1	0.1	-0.2	1.0			1.4
Cons.conf.idx	0.1	0.0	-0.1	-0.1	0.1	1.0		1.1
Nr.employed	0.0	0.1	0.4	-0.5	0.5	0.1	1.0	1.8

Euribor3m and Emp.var.rate were removed to make sure the correlation between each pair of numerical independent variables is less than 0.7 (Table 3.3) and the variance inflation factor (VIF) are less than 5.

3.1.3 The distribution of numerical independent variables

Histograms of the numerical independent variables of subscribers (customers who subscribed the long term deposit) are shown in Figure 3.2, and those of the non-subscribers (customers who did not subscribe the long term deposit) are shown in Figure 3.3. It is obvious that none of the numerical independent variables follow normal distribution.

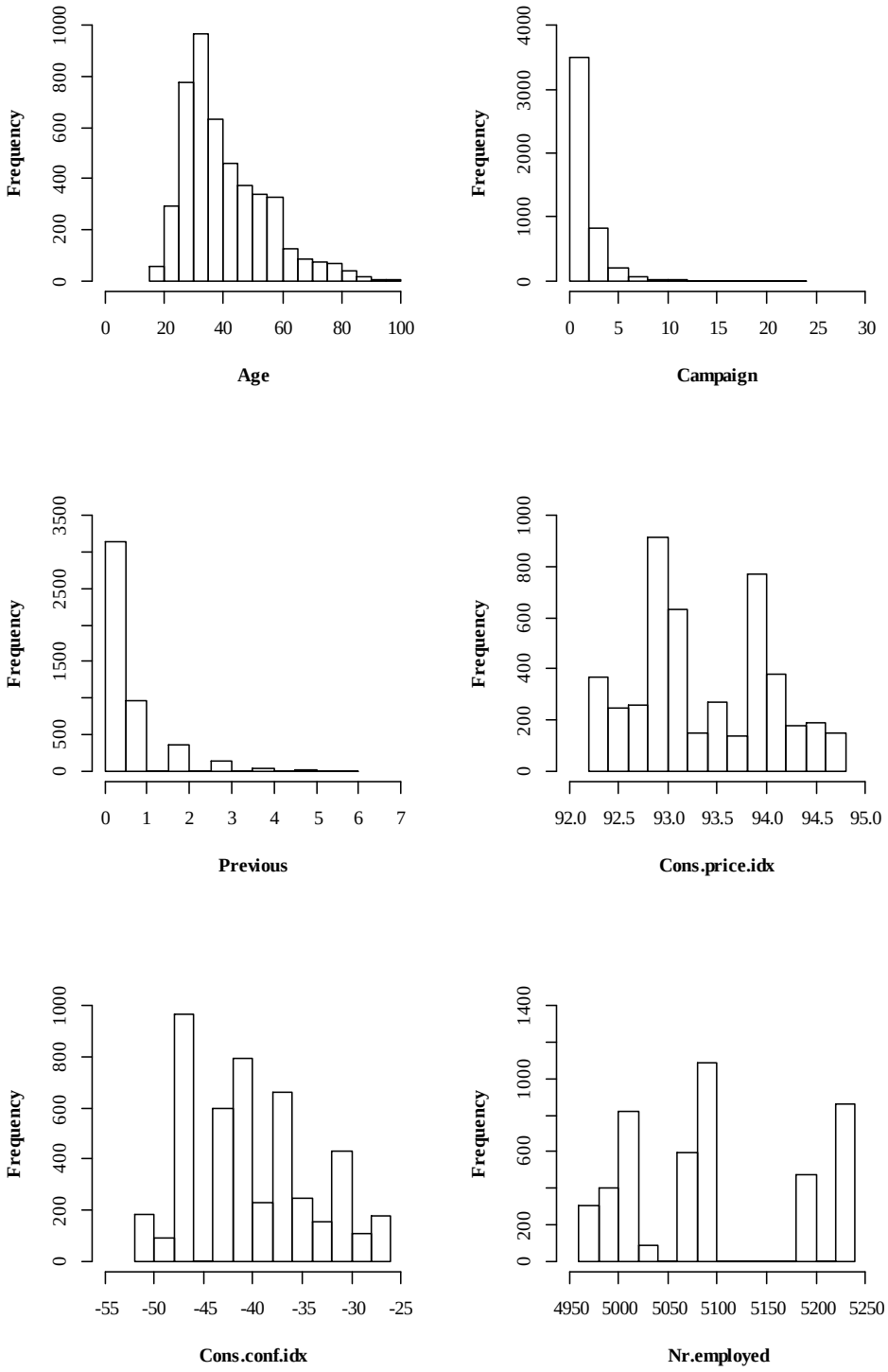

Figure 3.2 The distribution of numerical independent variables in subscribers

* Histograms of numerical independent variables in subscribers

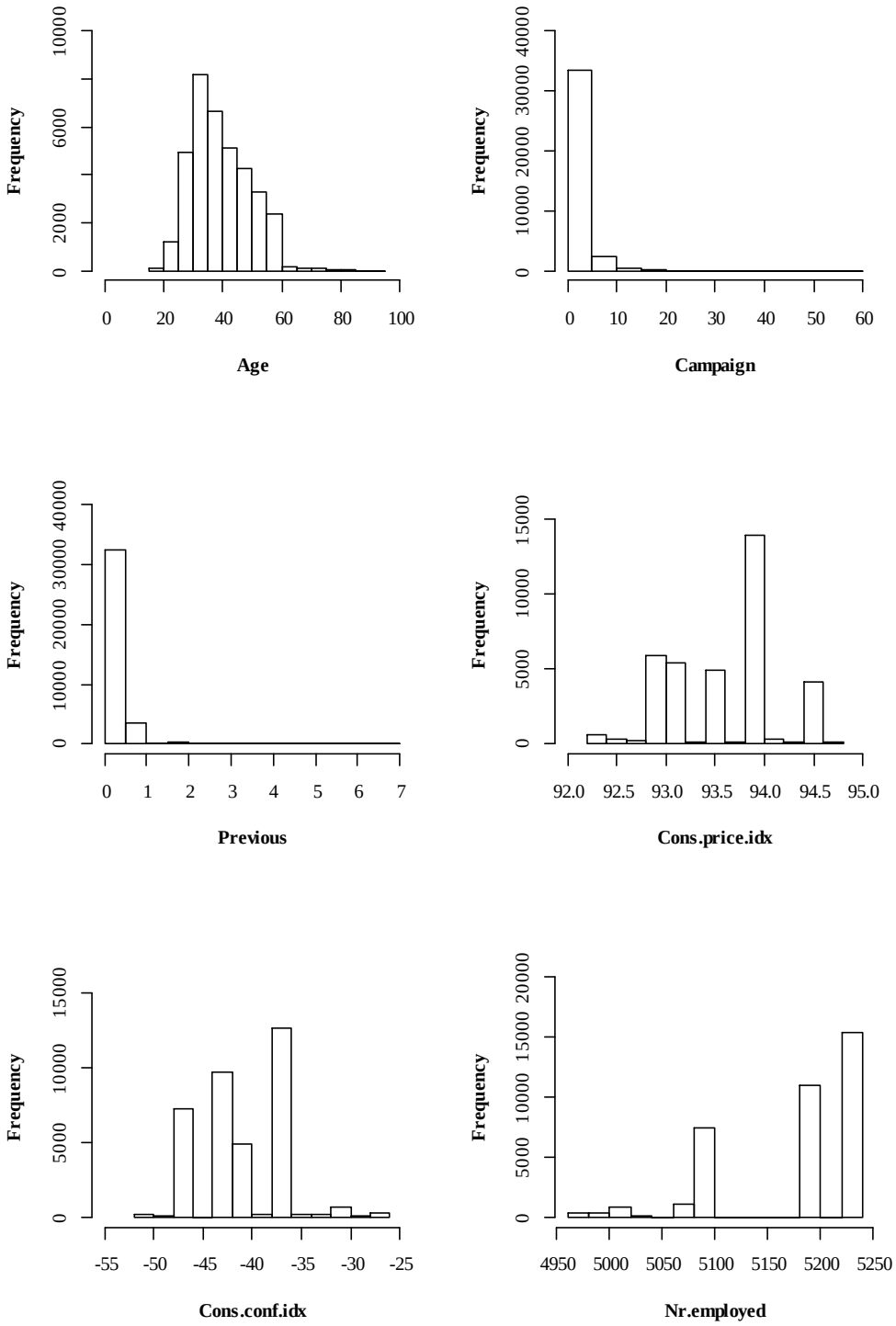

Figure 3.3 The distribution of numerical independent variables in non-subscribers

* Histograms of numerical independent variables in non-subscribers

3.2 Model comparison

3.2.1 The explanatory variables

The explanatory variables in the final FDA model and LR model are shown in Table 3.4 and Table 3.5, respectively. The two models use two slightly different sets of explanatory variables for prediction. The FDA model includes “Age” and “Education”, while the LR model uses “Cons_conf_idx” and “Job”. The average of 10 estimated coefficients from the 10 training sets are listed.

Table 3.4 The explanatory variables in the final flexible discriminant analysis model

Variable	Range	Coefficient
Intercept		0.082
Age	27-58	0.000
	<=26	0.027
	>58	0.048
Campaign	<5	0.000
	>=5	-0.018
Pdays	Not contacted before	0.000
	Contacted before	0.288
Nr_employed	>5099	0.000
	5076.3-5099.1	0.036
	5008.8-5076.2	0.261
	4963.6-5008.7	0.320
Eduction	Not basic9y	0.000
	Basic9y	-0.009
Default	Known	0.000
	Unknown	-0.014
Contact	Cellular	0.000
	Telephone	-0.038
Month	May, Jun, Jul, Sep, Nov	0.000
	Oct	0.055
	Apr	0.079
	Aug	-0.020
	Dec	0.102
	Mar	0.217
Day_of_week	Tue-Fri	0.000
	Mon	-0.023
Poutcome	Success or nonexistent	0.000
	Failure	-0.064

Table 3.5 The explanatory variables in the final logistic regression model

Variable	Range	Coefficient
Intercept		54.776
Campaign	0 to 55	-0.047
Pdays	0 to 999	-0.001
Cons_conf_idx	-50.8 to -26.9	0.022
Nr_employed	4964 to 5228	-0.011
Job	Admin.	0.000
	Blue-collar	-0.215
	Entrepreneur	-0.077
	Housemaid	-0.148
	Management	-0.050
	Retired	0.220
	Self-employed	-0.059
	Services	-0.174
	Student	0.222
	Technician	-0.055
	Unemployed	-0.067
	Unknown	-0.129
Default	No	0.000
	Unknown or Yes	-0.263
Contact	Cellular	0.000
	Telephone	-0.513
Month	May	0.000
	Nov	0.281
	Oct	0.445
	Sep	0.130
	Apr	0.701
	Aug	0.469
	Dec	0.808
	Jul	0.815
	Jun	0.910
	Mar	1.495
Day_of_week	Wed	0.000
	Tue	-0.094
	Thu	-0.075
	Mon	-0.354
	Fri	-0.132
Poutcome	Nonexistent	0.000
	Failure	-0.520
	Success	0.274

3.2.2 The prediction accuracies

For scoring, the FDA model assigned each customer a score ranging from 0 to 1, which was an indicator of the Yes/No ratio calculated according to the groups where his or her explanatory attributes are located. The LR model calculated the probability that a customer would subscribe the product based on his or her explanatory attributes in the model.

To compare the performances of the two models, the average cumulative count (TP, FP, TN, FN, and TP+FP), average cumulative percent (TPR, FPR, TNR, FNR and “Yes”/Total), average correct rate (Yes, No and overall), and the results of average TPR minus average FPR at various cutoff points calculated from the FDA or the LR model using the 10 different training and validation sets are listed in the following tables.

Table 3.6 The average cumulative count, average cumulative percent and average correct rate calculated from the flexible discriminant analysis model using the training sets

Training Set	Total:	37069.20												
	Cumulative Count					Cumulative Percent					Correct Rate			
	Predicted "Yes"			Predicted "No"		Predicted "Yes"			Predicted "No"					
Cutoff	Yes (TP)	No (FP)	Sum	Yes (FN)	No (TN)	Yes (TPR)	No (FPR)	Sum ("Yes"/ Total)	Yes (FNR)	No (TNR)	Yes (TP/Total)	No (TN/Total)	Overall (TP+TN) /Total	TPR-FPR
>=0.95	5.60	0.70	6.30	4175.30	32887.60	0.13%	0.00%	0.02%	99.87%	100.00%	0.02%	88.72%	88.73%	0.13%
>=0.9	33.30	7.90	41.20	4142.70	32885.30	0.80%	0.02%	0.11%	99.20%	99.98%	0.09%	88.71%	88.80%	0.77%
>=0.85	56.20	15.20	71.40	4119.80	32878.00	1.35%	0.05%	0.19%	98.65%	99.95%	0.15%	88.69%	88.85%	1.30%
>=0.8	66.50	18.90	85.40	4109.50	32874.30	1.59%	0.06%	0.23%	98.41%	99.94%	0.18%	88.68%	88.86%	1.54%
>=0.75	116.50	34.30	150.80	4059.50	32858.90	2.79%	0.10%	0.41%	97.21%	99.90%	0.31%	88.64%	88.96%	2.69%
>=0.7	261.70	80.20	341.90	3914.30	32813.00	6.27%	0.24%	0.92%	93.73%	99.76%	0.71%	88.52%	89.22%	6.02%
>=0.65	571.50	189.30	760.80	3604.50	32703.90	13.69%	0.58%	2.05%	86.31%	99.42%	1.54%	88.22%	89.77%	13.11%
>=0.6	786.50	328.90	1115.40	3389.50	32564.30	18.83%	1.00%	3.01%	81.17%	99.00%	2.12%	87.85%	89.97%	17.83%
>=0.55	850.40	391.70	1242.10	3325.60	32501.50	20.36%	1.19%	3.35%	79.64%	98.81%	2.29%	87.68%	89.97%	19.17%
>=0.5	880.90	439.30	1320.20	3295.10	32453.90	21.09%	1.34%	3.56%	78.91%	98.66%	2.38%	87.55%	89.93%	19.76%
>=0.45	984.70	533.60	1518.30	3191.30	32359.60	23.58%	1.62%	4.10%	76.42%	98.38%	2.66%	87.30%	89.95%	21.96%
>=0.4	1188.10	804.70	1992.80	2987.90	32088.50	28.45%	2.45%	5.38%	71.55%	97.55%	3.21%	86.56%	89.77%	26.00%
>=0.35	1547.90	1388.40	2936.30	2628.10	31504.80	37.07%	4.22%	7.92%	62.93%	95.78%	4.18%	84.99%	89.16%	32.85%
>=0.3	2053.30	2406.30	4459.60	2122.70	30486.90	49.17%	7.32%	12.03%	50.83%	92.68%	5.54%	82.24%	87.78%	41.85%
>=0.25	2176.20	2747.10	4923.30	1999.80	30146.10	52.11%	8.35%	13.28%	47.89%	91.65%	5.87%	81.32%	87.19%	43.76%
>=0.2	2276.10	2946.10	5222.20	1899.90	29947.10	54.50%	8.96%	14.09%	45.50%	91.04%	6.14%	80.79%	86.93%	45.55%
>=0.15*	2510.80	4082.40	6593.20	1665.20	28810.80	60.12%	12.41%	17.79%	39.88%	87.59%	6.77%	77.72%	84.49%	47.71%
>=0.1	2818.90	6972.20	9791.10	1357.10	25921.00	67.50%	21.20%	26.41%	32.50%	78.80%	7.60%	69.93%	77.53%	46.31%
>=0.05	3545.30	17731.10	21276.40	630.70	15162.10	84.90%	53.91%	57.40%	15.10%	46.09%	9.56%	40.90%	50.47%	30.99%
>=0	4176.00	32893.20	37069.20	0.00	0.00	100.00%	100.00%	100.00%	0.00%	0.00%	11.27%	0.00%	11.27%	0.00%

* Highlighted row: the optimal cutoff point determined by the KS distance between TPR and FPR

Table 3.7 The average cumulative count, average cumulative percent and average correct rate calculated from the logistic regression model using the training sets

Training Set	Total:	37069.20												
	Cumulative Count					Cumulative Percent					Correct Rate			
	Predicted "Yes"			Predicted "No"		Predicted "Yes"			Predicted "No"					
Cutoff	Yes (TP)	No (FP)	Sum	Yes (FN)	No (TN)	Yes (TPR)	No (FPR)	Sum ("Yes"/ Total)	Yes (FNR)	No (TNR)	Yes (TP/Total)	No (TN/Total)	Overall (TP+TN) /Total	TPR-FPR
>=0.95	0.00	0.00	0.00	4176.00	32893.20	0.00%	0.00%	0.00%	100.00%	100.00%	0.00%	88.73%	88.73%	0.00%
>=0.9	12.60	1.30	13.90	4163.40	32891.90	0.30%	0.00%	0.04%	99.70%	100.00%	0.03%	88.73%	88.77%	0.30%
>=0.85	55.60	11.70	67.30	4120.40	32881.50	1.33%	0.04%	0.18%	98.67%	99.96%	0.15%	88.70%	88.85%	1.30%
>=0.8	174.00	41.60	215.60	4002.00	32851.60	4.17%	0.13%	0.58%	95.83%	99.87%	0.47%	88.62%	89.09%	4.04%
>=0.75	354.80	93.90	448.70	3821.20	32799.30	8.50%	0.29%	1.21%	91.50%	99.71%	0.96%	88.48%	89.44%	8.21%
>=0.7	490.80	150.70	641.50	3685.20	32742.50	11.75%	0.46%	1.73%	88.25%	99.54%	1.32%	88.33%	89.65%	11.30%
>=0.65	615.40	218.50	833.90	3560.60	32674.70	14.74%	0.66%	2.25%	85.26%	99.34%	1.66%	88.15%	89.81%	14.07%
>=0.6	734.60	293.30	1027.90	3441.40	32599.90	17.59%	0.89%	2.77%	82.41%	99.11%	1.98%	87.94%	89.93%	16.70%
>=0.55	848.70	381.10	1229.80	3327.30	32512.10	20.32%	1.16%	3.32%	79.68%	98.84%	2.29%	87.71%	90.00%	19.16%
>=0.5	880.80	486.00	1366.80	3295.20	32407.20	21.10%	1.48%	3.69%	78.90%	98.52%	2.38%	87.42%	89.80%	19.62%
>=0.45	1117.70	656.90	1774.60	3058.30	32236.30	26.76%	2.00%	4.79%	73.24%	98.00%	3.02%	86.96%	89.98%	24.77%
>=0.4	1270.40	897.90	2168.30	2905.60	31995.30	30.42%	2.73%	5.85%	69.58%	97.27%	3.43%	86.31%	89.74%	27.69%
>=0.35	1528.40	1282.70	2811.10	2647.60	31610.50	36.60%	3.90%	7.58%	63.40%	96.10%	4.12%	85.27%	89.40%	32.70%
>=0.3	1787.30	1811.60	3598.90	2388.70	31081.60	42.80%	5.51%	9.71%	57.20%	94.49%	4.82%	83.85%	88.67%	37.29%
>=0.25	2033.90	2284.20	4318.10	2142.10	30609.00	48.70%	6.94%	11.65%	51.30%	93.06%	5.49%	82.57%	88.06%	41.76%
>=0.2	2331.50	3078.30	5409.80	1844.50	29814.90	55.83%	9.36%	14.59%	44.17%	90.64%	6.29%	80.43%	86.72%	46.47%
>=0.15*	2496.80	3888.80	6385.60	1679.20	29004.40	59.79%	11.82%	17.23%	40.21%	88.18%	6.74%	78.24%	84.98%	47.97%
>=0.1	2768.50	6317.40	9085.90	1407.50	26575.80	66.30%	19.21%	24.51%	33.70%	80.79%	7.47%	71.69%	79.16%	47.09%
>=0.05	3629.00	19354.10	22983.10	547.00	13539.10	86.90%	58.84%	62.00%	13.10%	41.16%	9.79%	36.52%	46.31%	28.06%
>=0	4176.00	32893.20	37069.20	0.00	0.00	100.00%	100.00%	100.00%	0.00%	0.00%	11.27%	0.00%	11.27%	0.00%

* Highlighted row: the optimal cutoff point determined by the KS distance between TPR and FPR

Table 3.8 The average cumulative count, average cumulative percent and average correct rate calculated from the flexible discriminant analysis model using the validation sets

Validataion Set	Total:	4118.80												
	Cumulative Count					Cumulative Percent					Correct Rate			
	Predicted "Yes"			Predicted "No"		Predicted "Yes"			Predicted "No"					
Cutoff	Yes (TP)	No (FP)	Sum	Yes (FN)	No (TN)	Yes (TPR)	No (FPR)	Sum ("Yes"/ Total)	Yes (FNR)	No (TNR)	Yes (TP/Total)	No (TN/Total)	Overall (TP+TN) /Total	TPR-FPR
>=0.95	0.50	0.10	0.60	463.90	3654.30	0.11%	0.00%	0.01%	99.89%	100.00%	0.01%	88.72%	88.74%	0.10%
>=0.9	3.40	1.10	4.50	460.60	3653.70	0.73%	0.03%	0.11%	99.27%	99.97%	0.08%	88.71%	88.79%	0.70%
>=0.85	6.20	1.80	8.00	457.80	3653.00	1.34%	0.05%	0.19%	98.66%	99.95%	0.15%	88.69%	88.84%	1.30%
>=0.8	7.30	2.10	9.40	456.70	3652.70	1.58%	0.06%	0.23%	98.42%	99.94%	0.18%	88.68%	88.86%	1.53%
>=0.75	12.50	4.00	16.50	451.50	3650.80	2.71%	0.11%	0.40%	97.29%	99.89%	0.30%	88.64%	88.94%	2.60%
>=0.7	28.70	9.00	37.70	435.30	3645.80	6.20%	0.25%	0.92%	93.80%	99.75%	0.70%	88.52%	89.21%	5.96%
>=0.65	63.50	21.10	84.60	400.50	3633.70	13.71%	0.58%	2.05%	86.29%	99.42%	1.54%	88.22%	89.77%	13.14%
>=0.6	87.10	36.60	123.70	376.90	3618.20	18.81%	1.00%	3.00%	81.19%	99.00%	2.12%	87.85%	89.96%	17.81%
>=0.55	93.90	43.80	137.70	370.10	3611.00	20.28%	1.20%	3.34%	79.72%	98.80%	2.28%	87.67%	89.95%	19.08%
>=0.5	98.10	49.00	147.10	365.90	3605.80	21.18%	1.34%	3.57%	78.82%	98.66%	2.38%	87.55%	89.93%	19.84%
>=0.45	108.60	59.40	168.00	355.40	3595.40	23.44%	1.63%	4.08%	76.56%	98.37%	2.64%	87.29%	89.93%	21.82%
>=0.4	131.30	89.40	220.70	332.70	3565.40	28.35%	2.45%	5.36%	71.65%	97.55%	3.19%	86.56%	89.75%	25.90%
>=0.35	170.80	155.30	326.10	293.20	3499.50	36.84%	4.25%	7.92%	63.16%	95.75%	4.15%	84.96%	89.11%	32.60%
>=0.3	227.70	266.90	494.60	236.30	3387.90	49.08%	7.30%	12.01%	50.92%	92.70%	5.53%	82.25%	87.78%	41.78%
>=0.25	241.90	305.00	546.90	222.10	3349.80	52.14%	8.35%	13.28%	47.86%	91.65%	5.87%	81.33%	87.20%	43.80%
>=0.2	252.70	327.90	580.60	211.30	3326.90	54.48%	8.97%	14.10%	45.52%	91.03%	6.14%	80.77%	86.91%	45.51%
>=0.15*	278.80	454.40	733.20	185.20	3200.40	60.11%#	12.43%	17.80%&	39.89%	87.57%	6.77%	77.70%	84.47%&	47.67%
>=0.1	313.40	776.50	1089.90	150.60	2878.30	67.57%	21.25%	26.46%	32.43%	78.75%	7.61%	69.88%	77.49%	46.33%
>=0.05	393.70	1968.70	2362.40	70.30	1686.10	84.86%	53.87%	57.36%	15.14%	46.13%	9.56%	40.94%	50.49%	31.00%
>=0	464.00	3654.80	4118.80	0.00	0.00	100.00%	100.00%	100.00%	0.00%	0.00%	11.26%	0.00%	11.26%	0.00%

* Highlighted row: the optimal cutoff point determined from training set #: p>0.05 vs that of LR model &: p<0.05 vs that of LR model

Table 3.9 The average cumulative count, average cumulative percent and average correct rate calculated from the logistic regression model using the validation sets

Validation Set	Total:	4118.80												
	Cumulative Count					Cumulative Percent					Correct Rate			
	Predicted "Yes"			Predicted "No"		Predicted "Yes"			Predicted "No"					
Cutoff	Yes (TP)	No (FP)	Sum	Yes (FN)	No (TN)	Yes (TPR)	No (FPR)	Sum ("Yes"/ Total)	Yes (FNR)	No (TNR)	Yes (TP/Total)	No (TN/Total)	Overall (TP+TN) /Total	TPR-FPR
>=0.95	0.00	0.00	0.00	464.00	3654.80	0.00%	0.00%	0.00%	100.00%	100.00%	0.00%	88.74%	88.74%	0.00%
>=0.9	1.30	0.20	1.50	462.70	3654.60	0.28%	0.01%	0.04%	99.72%	99.99%	0.03%	88.73%	88.76%	0.27%
>=0.85	6.10	1.20	7.30	457.90	3653.60	1.32%	0.03%	0.18%	98.68%	99.97%	0.15%	88.71%	88.86%	1.29%
>=0.8	18.60	5.00	23.60	445.40	3649.80	4.03%	0.14%	0.57%	95.97%	99.86%	0.45%	88.61%	89.07%	3.89%
>=0.75	40.20	10.90	51.10	423.80	3643.90	8.68%	0.30%	1.24%	91.32%	99.70%	0.98%	88.47%	89.45%	8.39%
>=0.7	54.60	17.10	71.70	409.40	3637.70	11.80%	0.47%	1.74%	88.20%	99.53%	1.33%	88.32%	89.65%	11.34%
>=0.65	68.30	25.00	93.30	395.70	3629.80	14.75%	0.68%	2.27%	85.25%	99.32%	1.66%	88.13%	89.79%	14.07%
>=0.6	81.70	33.20	114.90	382.30	3621.60	17.64%	0.91%	2.79%	82.36%	99.09%	1.98%	87.93%	89.91%	16.74%
>=0.55	93.90	42.50	136.40	370.10	3612.30	20.28%	1.16%	3.31%	79.72%	98.84%	2.28%	87.70%	89.98%	19.12%
>=0.5	107.30	53.90	161.20	356.70	3600.90	23.17%	1.47%	3.91%	76.83%	98.53%	2.61%	87.43%	90.03%	21.69%
>=0.45	122.80	73.50	196.30	341.20	3581.30	26.51%	2.01%	4.77%	73.49%	97.99%	2.98%	86.95%	89.93%	24.50%
>=0.4	140.20	101.50	241.70	323.80	3553.30	30.26%	2.78%	5.87%	69.74%	97.22%	3.40%	86.27%	89.68%	27.49%
>=0.35	169.20	144.80	314.00	294.80	3510.00	36.50%	3.96%	7.62%	63.50%	96.04%	4.11%	85.22%	89.33%	32.53%
>=0.3	197.80	200.70	398.50	266.20	3454.10	42.65%	5.49%	9.68%	57.35%	94.51%	4.80%	83.86%	88.67%	37.16%
>=0.25	225.60	254.40	480.00	238.40	3400.40	48.63%	6.96%	11.65%	51.37%	93.04%	5.48%	82.56%	88.04%	41.67%
>=0.2	257.90	342.20	600.10	206.10	3312.60	55.59%	9.36%	14.57%	44.41%	90.64%	6.26%	80.43%	86.69%	46.23%
>=0.15*	277.20	432.30	709.50	186.80	3222.50	59.76%#	11.83%	17.23%&	40.24%	88.17%	6.73%	78.24%	84.97%&	47.93%
>=0.1	307.50	702.10	1009.60	156.50	2952.70	66.31%	19.21%	24.51%	33.69%	80.79%	7.47%	71.69%	79.15%	47.09%
>=0.05	401.60	2152.20	2553.80	62.40	1502.60	86.58%	58.89%	62.00%	13.42%	41.11%	9.75%	36.48%	46.23%	27.69%
>=0	464.00	3654.80	4118.80	0.00	0.00	100.00%	100.00%	100.00%	0.00%	0.00%	11.26%	0.00%	11.26%	0.00%

* Highlighted row: the optimal cutoff point determined from training set #: p>0.05 vs that of FDA model &: p<0.05 vs that of FDA model

Focusing on the predicted "right" customers, it is straightforward that higher TPR and lower FPR will be ideal. Thus, the KS distance between TPR and FPR in the training set is used to determine the optimal cutoff point, which is 0.15 for both FDA and LR models.

With the help of FDA model, at the optimal cutoff point determined from training sets, we only need to contact 17.80% of the total customers in the validation sets and can capture 60.11% of the real “right” customers. The overall correct prediction rate of FDA model is 84.47%.

With the help of LR model, at the optimal cutoff point determined from training sets, we only need to contact 17.23% (p= 0.00000547 vs that of FDA model, paired t-test) of the total customers in the validation sets and can capture 59.76% (p= 0.2391 vs that of FDA model, paired t-test) of the real “right” customers. The overall correct prediction rate is 84.97% (p= 0.00001228 vs that of FDA model, paired t-test).

Thus, focusing on the predicted "right" customers, LR has a slightly higher efficiency, i.e. with a lower contact rate while capture similar number of true customers. LR also shows higher overall prediction rate.

3.2.3 The receiver operating characteristic curves and the areas under the curves

The ROC curves and AUCs of the two models generated by the average TPR and average FPR from training and validation data sets are shown in the following two figures. Note that the ROC curves of the FDA and LR models almost overlap with each other in both situations. This indicates that the two models have similar performance, which is further confirmed by the fact that the AUCs for both models using training or validation set are not significantly different ($p>0.05$, FDA vs LR, Wilcoxon signed-rank test for training sets and paired t-test for validation sets).

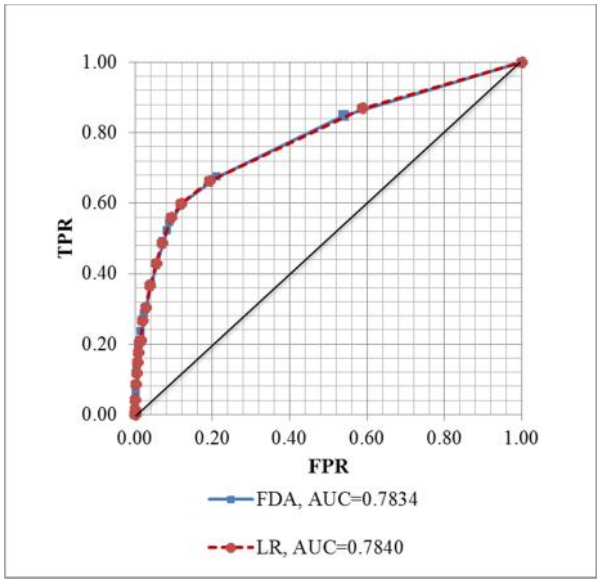

Figure 3.4 The receiver operating characteristic curves and the areas under the curves of the two models using the training data sets

*The ROC curves and AUCs of the FDA model (blue solid line) and the LR model (red dash line) with the training set were plotted together. FPR: false positive rate, which is 1-speciticity. TPR: true positive rate, which is sensitivity.

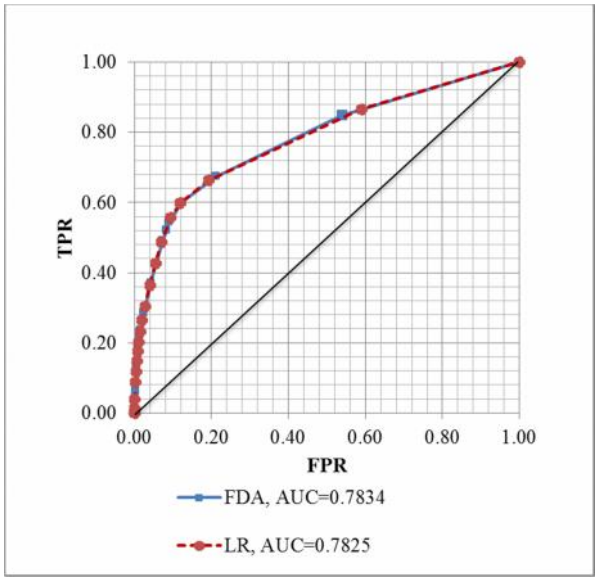

Figure 3.5 The receiver operating characteristic curves and the areas under the curves of the two models using the validation data sets

*The ROC curves and AUCs of the FDA model (blue solid line) and the LR model (red dash line) with the training set were plotted together. FPR: false positive rate, which is 1-speciticity. TPR: true positive rate, which is sensitivity.

4 DISCUSSIONS

Although Moro et al used a similar dataset to build predictive models for customer targeting, their focus was on logistic regression, decision trees, neural network, and support vector machine models. This thesis compares the LR and FDA models, which is a complement of their work.

Using AUC and ALIFT as classification metrics, Moro et al concluded that the neural network model was the best among the four models. However, in their study, the model parameters and classification metrics were estimated either by the 2/3-1/3-hold-out or a rolling window method, both of which decrease accuracy estimation. Because the observations in the training set and validation set are independent, the information in the validation set cannot be reflected in the training model [Kohavi, 1995]. Therefore, the hold-out or rolling window method makes inefficient use of the data and decreases accuracy estimation. This thesis uses the 10-fold stratified cross validation method, which efficiently includes all the information from the whole data set and does not induce bias in accuracy estimation [Kohavi, 1995]. Thus the model performance comparison result from this thesis is more convincing.

In this thesis, the optimal cutoff point was determined by the KS distance between TPR and FPR which is equivalent to Youden Index. In practice, if the cost and revenue generated by each TP, TN, FP and FN are known, the final profits at various cutoff scores can be easily calculated. Managers can decide the optimal cutoff point to maximize the final profits based on their sales goals, the call center resource capacities and the total number of customers. For a more general way to decide the optimal cutoff point, managers can set different weights for TPR, TNR, FPR and FNR based on prior experience. Once the weights are determined, the sum of the

weighted TPR, TNR, FPR and FNR can be used as an index. The optimal cutoff point can be found at the maximum of this index.

5 CONCLUSIONS

The LR model and the FDA model show equally satisfactory performances in customer classification based on their AUCs and ROC curves. Focusing on the predicted “right” customers, the LR model shows slightly higher classification efficiency.

REFERENCES

Albro W and Linsley C (2001) ABA/BMA Survey Shows Big Increase in Use of Bank Telemarketing. Bank Marketing 33 (2): 9

Altman EI (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23: 589-609

Anderson S (2001) Logistic Regression. [Web log post] Retrieved from http://schatz.sju.edu/multivar/guide/logistic.pdf

Bewick V, Cheek L, Ball J (2005) Logistic regression. Statistics review 9 (1):112–118

Cohen J, Cohen P, West SG, Aiken LS (2003) Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences. L. Erlbaum Associates

Eisenbeis RA (1977) Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics. Journal of Finance 32:875-900

Gupta A, McMahon S, Jain A, Kanagasabai K (2008) Redefining the Mission for Banks' Call Centers Cut Costs, Grow Sales, or Both. [Web log post] Retrieved from http://www.strategyand.pwc.com/media/file/Redefining_Mission_for_Banks_Call_Centers.pdf

Hand DJ and Henley WE (1997) Statistical Classification Methods in Consumer Credit Scoring. Journal of the Royal Statistical Society, Series A (Statistics in Society) 160: 523-541

Hastie T, Tibshirani R, Buja A (1994) Flexible discriminant analysis by optimal scoring. J. Amer. Statist. Assoc. 89: 1255-1270

Klecka WR (1980) Discriminant analysis. Sage Publications

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. 9: 1137–1143

Kotler P and Keller KV (2012) Framework for Marketing Management. Pearson

Lau K, Chow H, Liu C (2004) A database approach to cross selling in the banking industry: practices, strategies and challenges. Journal of Database Marketing and Customer Strategy Management 11 (3): 216–234

Leonard M (1998) Marketing literature review. Journal of Marketing 62 (3):128-140

Liong CY and Foo S (2013) Comparison of linear discriminant analysis and logistic regression for data classification. AIP Conference Proceedings 1522 (1): 1159-1165

Martens D and Provost F (2011) Pseudo-social network targeting from consumer transaction data. NYU Working Papers Series CeDER-11-05

Meyers LS, Gamst G, Guarino AJ (2013) Applied Multivariate Research: Design and Interpretation. SAGE Publications

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decision Support Systems 62: 22–31

Pohar M, Blas M, Turk S (2004) Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodološki zvezki 1 (1): 143-161

Rust RT, Moorman C, Bhalla G (2010) Rethinking marketing. Harvard Business Review 1: 1–8

APPENDICES

Appendix A SAS outputs for the flexible discriminant analysis model using the 50/50 split training set

The SAS System

The REG Procedure
Model: bgscore
Dependent Variable: good

Number of Observations Read	20594
Number of Observations Used	20594

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	417.64440	24.56732	303.89	<.0001
Error	20576	1663.42314	0.08084		
Corrected Total	20593	2081.06754			

Root MSE	0.28433	**R-Square**	0.2007
Dependent Mean	0.11406	**Adj R-Sq**	0.2000
Coeff Var	249.27492		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	0.08492	0.00435	19.50	<.0001
age3	**1**	0.02220	0.00868	2.56	0.0106
age4	**1**	0.05495	0.01067	5.15	<.0001
camp4	**1**	-0.01632	0.00617	-2.64	0.0082
pday1	**1**	0.26956	0.01191	22.63	<.0001
employ1	**1**	0.03870	0.00653	5.92	<.0001
employ2	**1**	0.25570	0.00927	27.58	<.0001
employ3	**1**	0.31549	0.01117	28.24	<.0001
edu5	**1**	-0.01313	0.00566	-2.32	0.0203
def1	**1**	-0.01180	0.00502	-2.35	0.0187

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
con1	**1**	-0.03918	0.00488	-8.03	<.0001
mon2	**1**	0.08993	0.01662	5.41	<.0001
mon4	**1**	0.07703	0.00939	8.20	<.0001
mon5	**1**	-0.02088	0.00624	-3.34	0.0008
mon6	**1**	0.11952	0.03007	3.97	<.0001
mon9	**1**	0.20547	0.01825	11.26	<.0001
day3	**1**	-0.02370	0.00496	-4.78	<.0001
pout1	**1**	-0.06568	0.00716	-9.17	<.0001

Appendix B SAS outputs for the logistic regression model using the 50/50 split training set

The SAS System

The LOGISTIC Procedure

Model Information

Data Set	SAVE.TRAINSET
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	20594
Number of Observations Used	20594

Response Profile

Ordered Value	Y	Total Frequency
1	yes	2349
2	no	18245

Probability modeled is Y='yes'.

Stepwise Selection Procedure

Class Level Information

Class	Value	Design Variables										
Job	**admin**	1	0	0	0	0	0	0	0	0	0	0
	bluecollar	0	1	0	0	0	0	0	0	0	0	0
	entrepreneur	0	0	1	0	0	0	0	0	0	0	0
	housemaid	0	0	0	1	0	0	0	0	0	0	0
	management	0	0	0	0	1	0	0	0	0	0	0
	retired	0	0	0	0	0	1	0	0	0	0	0
	selfemployed	0	0	0	0	0	0	1	0	0	0	0
	services	0	0	0	0	0	0	0	1	0	0	0
	student	0	0	0	0	0	0	0	0	1	0	0
	technician	0	0	0	0	0	0	0	0	0	1	0
	unemployed	0	0	0	0	0	0	0	0	0	0	1
	unknown	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Marital	**divorced**	1	0	0								

Class Level Information

Class	Value	Design Variables								
	married	0	1	0						
	single	0	0	1						
	unknown	-1	-1	-1						
Education	**basic4y**	1	0	0	0	0	0	0		
	basic6y	0	1	0	0	0	0	0		
	basic9y	0	0	1	0	0	0	0		
	highschool	0	0	0	1	0	0	0		
	illiterate	0	0	0	0	1	0	0		
	professionalcourse	0	0	0	0	0	1	0		
	universitydegree	0	0	0	0	0	0	1		
	unknown	-1	-1	-1	-1	-1	-1	-1		
Default	**no**	1								
	unknown or yes	-1								
Housing	**no**	1	0							
	unknown	0	1							
	yes	-1	-1							
Loan	**no**	1	0							
	unknown	0	1							
	yes	-1	-1							
Contact	**cellular**	1								
	telephone	-1								
Month	**apr**	1	0	0	0	0	0	0	0	0
	aug	0	1	0	0	0	0	0	0	0
	dec	0	0	1	0	0	0	0	0	0
	jul	0	0	0	1	0	0	0	0	0
	jun	0	0	0	0	1	0	0	0	0
	mar	0	0	0	0	0	1	0	0	0
	may	0	0	0	0	0	0	1	0	0
	nov	0	0	0	0	0	0	0	1	0
	oct	0	0	0	0	0	0	0	0	1
	sep	-1	-1	-1	-1	-1	-1	-1	-1	-1
Day_of_week	**fri**	1	0	0	0					
	mon	0	1	0	0					
	thu	0	0	1	0					
	tue	0	0	0	1					

Class Level Information

Class	Value	Design Variables
	wed	-1 -1 -1 -1
Poutcome	**failure**	1 0
	nonexistent	0 1
	success	-1 -1

Step 0. Intercept entered:

Model Convergence Status

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

-2 Log L = 14618.662

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4118.9725	48	<.0001

Step 1. Effect Nr_employed entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	12447.808
SC	14628.595	12463.673
-2 Log L	14618.662	12443.808

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2174.8541	1	<.0001
Score	2519.4468	1	<.0001
Wald	1989.7268	1	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
758.6537	47	<.0001

Note: No effects for the model in Step 1 are removed.

Step 2. Effect Month entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	12097.088
SC	14628.595	12184.349
-2 Log L	14618.662	12075.088

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2543.5735	10	<.0001
Score	3293.6202	10	<.0001
Wald	2393.9329	10	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
408.8466	38	<.0001

Note: No effects for the model in Step 2 are removed.

Step 3. Effect Poutcome entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11850.652
SC	14628.595	11953.778
-2 Log L	14618.662	11824.652

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2794.0101	12	<.0001
Score	3907.9527	12	<.0001
Wald	2479.2581	12	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
153.9470	36	<.0001

Note: No effects for the model in Step 3 are removed.

Step 4. Effect Contact entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11818.210
SC	14628.595	11929.269
-2 Log L	14618.662	11790.210

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2828.4515	13	<.0001
Score	3909.1150	13	<.0001
Wald	2452.5594	13	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
120.9028	35	<.0001

Note: No effects for the model in Step 4 are removed.

Step 5. Effect Day_of_week entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11799.953
SC	14628.595	11942.743
-2 Log L	14618.662	11763.953

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2854.7088	17	<.0001
Score	3931.3821	17	<.0001
Wald	2462.9590	17	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
95.3296	31	<.0001

Note: No effects for the model in Step 5 are removed.

Step 6. Effect Pdays entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11786.136
SC	14628.595	11936.858
-2 Log L	14618.662	11748.136

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2870.5261	18	<.0001
Score	3964.8494	18	<.0001
Wald	2478.5719	18	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
77.9445	30	<.0001

Note: No effects for the model in Step 6 are removed.

Step 7. Effect Cons_conf_idx entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11773.090
SC	14628.595	11931.745
-2 Log L	14618.662	11733.090

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2885.5721	19	<.0001
Score	4052.3163	19	<.0001
Wald	2517.0749	19	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
62.8263	29	0.0003

Note: No effects for the model in Step 7 are removed.

Step 8. Effect Default entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11764.037
SC	14628.595	11930.625
-2 Log L	14618.662	11722.037

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2896.6247	20	<.0001
Score	4057.2895	20	<.0001
Wald	2520.3937	20	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
52.2538	28	0.0036

Note: No effects for the model in Step 8 are removed.

Step 9. Effect Campaign entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11755.931
SC	14628.595	11930.452
-2 Log L	14618.662	11711.931

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2906.7306	21	<.0001
Score	4062.3778	21	<.0001
Wald	2522.9519	21	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
43.3245	27	0.0243

Note: No effects for the model in Step 9 are removed.

Step 10. Effect Job entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11755.105
SC	14628.595	12016.886
-2 Log L	14618.662	11689.105

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2929.5571	32	<.0001
Score	4087.1470	32	<.0001

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Wald	2537.6287	32	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
20.1455	16	0.2137

Note: No effects for the model in Step 10 are removed.

Step 11. Effect Previous entered:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11755.041
SC	14628.595	12024.755
-2 Log L	14618.662	11687.041

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2931.6207	33	<.0001
Score	4087.5862	33	<.0001
Wald	2540.8954	33	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
18.0606	15	0.2595

Step 12. Effect Previous is removed:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14620.662	11755.105
SC	14628.595	12016.886

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	14618.662	11689.105

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2929.5571	32	<.0001
Score	4087.1470	32	<.0001
Wald	2537.6287	32	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
20.1455	16	0.2137

Note: No effects for the model in Step 12 are removed.

Note: Model building terminates because the last effect entered is removed by the Wald statistic criterion.

Summary of Stepwise Selection

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	**Nr_employed**		1	1	2519.4468		<.0001
2	**Month**		9	2	361.7902		<.0001
3	**Poutcome**		2	3	252.8938		<.0001
4	**Contact**		1	4	33.6145		<.0001
5	**Day_of_week**		4	5	25.5125		<.0001
6	**Pdays**		1	6	16.8670		<.0001
7	**Cons_conf_idx**		1	7	15.1515		<.0001
8	**Default**		1	8	10.6505		0.0011
9	**Campaign**		1	9	9.0330		0.0027
10	**Job**		11	10	23.2403		0.0163
11	**Previous**		1	11	2.0833		0.1489
12		**Previous**	1	10		2.0754	0.1497

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Campaign	1	9.6098	0.0019
Pdays	1	16.1069	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Cons_conf_idx	1	11.3899	0.0007
Nr_employed	1	610.8373	<.0001
Job	11	23.1716	0.0167
Default	1	8.4843	0.0036
Contact	1	41.9551	<.0001
Month	9	194.7264	<.0001
Day_of_week	4	22.9727	0.0001
Poutcome	2	51.6383	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	53.8707	2.1370	635.4668	<.0001
Campaign		1	-0.0387	0.0125	9.6098	0.0019
Pdays		1	-0.00101	0.000251	16.1069	<.0001
Cons_conf_idx		1	0.0204	0.00606	11.3899	0.0007
Nr_employed		1	-0.0105	0.000426	610.8373	<.0001
Job	**admin**	1	0.0134	0.0575	0.0539	0.8163
Job	**bluecollar**	1	-0.1761	0.0685	6.6096	0.0101
Job	**entrepreneur**	1	-0.0632	0.1344	0.2213	0.6380
Job	**housemaid**	1	-0.0817	0.1560	0.2745	0.6003
Job	**management**	1	-0.0920	0.0945	0.9473	0.3304
Job	**retired**	1	0.2997	0.0947	10.0209	0.0015
Job	**selfemployed**	1	0.00914	0.1272	0.0052	0.9427
Job	**services**	1	-0.1017	0.0888	1.3134	0.2518
Job	**student**	1	0.2344	0.1152	4.1421	0.0418
Job	**technician**	1	0.00508	0.0684	0.0055	0.9408
Job	**unemployed**	1	0.0183	0.1364	0.0181	0.8931
Default	**no**	1	0.1136	0.0390	8.4843	0.0036
Contact	**cellular**	1	0.2398	0.0370	41.9551	<.0001
Month	**apr**	1	0.0717	0.0841	0.7256	0.3943
Month	**aug**	1	-0.1400	0.0805	3.0203	0.0822
Month	**dec**	1	0.2354	0.2035	1.3383	0.2473
Month	**jul**	1	0.1999	0.0765	6.8363	0.0089
Month	**jun**	1	0.1679	0.0825	4.1360	0.0420

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Month	**mar**	1	0.8022	0.1273	39.7017	<.0001
Month	**may**	1	-0.5877	0.0642	83.6948	<.0001
Month	**nov**	1	-0.3214	0.0830	14.9911	0.0001
Month	**oct**	1	-0.0140	0.1196	0.0137	0.9067
Day_of_week	**fri**	1	-0.0102	0.0508	0.0407	0.8401
Day_of_week	**mon**	1	-0.2300	0.0512	20.1684	<.0001
Day_of_week	**thu**	1	0.0899	0.0473	3.6206	0.0571
Day_of_week	**tue**	1	0.0479	0.0491	0.9496	0.3298
Poutcome	**failure**	1	-0.4210	0.0884	22.6632	<.0001
Poutcome	**nonexistent**	1	0.1078	0.0991	1.1815	0.2770

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Campaign	0.962	0.939	0.986
Pdays	0.999	0.999	0.999
Cons_conf_idx	1.021	1.009	1.033
Nr_employed	0.990	0.989	0.990
Job admin vs unknown	1.082	0.617	1.897
Job bluecollar vs unknown	0.895	0.508	1.577
Job entrepreneur vs unknown	1.002	0.539	1.862
Job housemaid vs unknown	0.984	0.518	1.868
Job management vs unknown	0.974	0.543	1.747
Job retired vs unknown	1.440	0.804	2.582
Job selfemployed vs unknown	1.077	0.584	1.987
Job services vs unknown	0.964	0.540	1.722
Job student vs unknown	1.349	0.740	2.460
Job technician vs unknown	1.073	0.608	1.892
Job unemployed vs unknown	1.087	0.584	2.023
Default no vs unknown or yes	1.255	1.077	1.462
Contact cellular vs telephone	1.615	1.397	1.868
Month apr vs sep	1.625	1.176	2.247
Month aug vs sep	1.315	0.965	1.792
Month dec vs sep	1.915	1.161	3.158
Month jul vs sep	1.848	1.340	2.548

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Month jun vs sep	1.790	1.298	2.468
Month mar vs sep	3.375	2.313	4.925
Month may vs sep	0.841	0.622	1.136
Month nov vs sep	1.097	0.797	1.510
Month oct vs sep	1.492	1.058	2.104
Day_of_week fri vs wed	0.893	0.765	1.044
Day_of_week mon vs wed	0.717	0.613	0.839
Day_of_week thu vs wed	0.988	0.852	1.145
Day_of_week tue vs wed	0.947	0.814	1.102
Poutcome failure vs success	0.480	0.293	0.787
Poutcome nonexistent vs success	0.814	0.486	1.366

Association of Predicted Probabilities and Observed Responses

Percent Concordant	78.1	**Somers' D**	0.570
Percent Discordant	21.1	**Gamma**	0.575
Percent Tied	0.9	**Tau-a**	0.115
Pairs	42857505	**c**	0.785

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	2349	0	18245	0	11.4	100.0	0.0	88.6	.
0.050	2055	7077	11168	294	44.3	87.5	38.8	84.5	4.0
0.100	1549	14567	3678	800	78.3	65.9	79.8	70.4	5.2
0.150	1357	16060	2185	992	84.6	57.8	88.0	61.7	5.8
0.200	1270	16518	1727	1079	86.4	54.1	90.5	57.6	6.1
0.250	1091	16968	1277	1258	87.7	46.4	93.0	53.9	6.9
0.300	950	17263	982	1399	88.4	40.4	94.6	50.8	7.5
0.350	806	17525	720	1543	89.0	34.3	96.1	47.2	8.1
0.400	693	17727	518	1656	89.4	29.5	97.2	42.8	8.5
0.450	623	17858	387	1726	89.7	26.5	97.9	38.3	8.8
0.500	529	17940	305	1820	89.7	22.5	98.3	36.6	9.2
0.550	444	18016	229	1905	89.6	18.9	98.7	34.0	9.6

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.600	362	18080	165	1987	89.6	15.4	99.1	31.3	9.9
0.650	313	18120	125	2036	89.5	13.3	99.3	28.5	10.1
0.700	248	18159	86	2101	89.4	10.6	99.5	25.7	10.4
0.750	166	18198	47	2183	89.2	7.1	99.7	22.1	10.7
0.800	57	18226	19	2292	88.8	2.4	99.9	25.0	11.2
0.850	21	18240	5	2328	88.7	0.9	100.0	19.2	11.3
0.900	0	18245	0	2349	88.6	0.0	100.0	.	11.4
0.950	0	18245	0	2349	88.6	0.0	100.0	.	11.4
1.000	0	18245	0	2349	88.6	0.0	100.0	.	11.4