BANK MARKETING

Babson College

Sang Won Baek

December, 2017

**Executive Summary**

How much impact would telemarketing play a role in leading the clients to subscribe a bank term deposit? To determine the success rate of telemarketing done by the banks, we analyzed the direct marketing campaign of a Portuguese banking institution by using the data obtained on the UCI Machine Learning Repository. Using a classification tree and a logistic regression on RStudio, we classified and analyzed the target variable: whether the client is going to subscribe a bank term deposit or not. A bank term deposit is a type of deposit that cannot be given any access to for a certain amount of time. In return, the clients receive a higher interest rate compared to a normal savings account. Our overall objective was to find the model with the highest specificity, which is the percentage that shows the model’s predictive accuracy in determining the non-subscribers out of all non-subscribers. In terms of cost and benefit, we wanted to find the model with highest specificity and minimum misclassification rate to maximize the efficiency of the banks’ costs of phone calls and their time; having this model will allow the banks to focus mostly on the clients that are willing to subscribe. Therefore, identifying the specific group of clients through the analysis will ensure more profit to be generated.

We utilized the classification trees to identify the variables with high leverage on predicting the subscriptions of the clients, which is done through generating various classification trees using two different methodologies (“category based” and “all variables”). Then, we tested the three different models (the four significant variables from each category, the four significant variables of all variables, and all variables) by running the logistic regression. For the comparison, we generated a confusion matrix for each testing models and analyzed one by one based on its specificity rate and its misclassification rate. Surprisingly, we obtained almost the same specificity and misclassification rate for all three models. Although the phenomenon seemed negative because we thought our attempt of narrowing down the variables did not have much impact on changing the model’s predictive accuracy, this phenomenon was a testimony for these model’s credibility because nearly the same results on confusion matrix indicate that we are able to attain the similar predictive power with fewer number of predictors. In conclusion, we recommend to the marketing department of the bank institution that focusing on the customers with the low number of employees at work (0 at it’s optimal value) at months of March, June, July, August and October, at lowest possible days of days passed after the last contact was made, at Euro Interbank Offered Rate with lowest possible value would result them to have a better success rate in getting the clients to subscribe through the telemarketing.

**Analysis: Data Preprocessing**

The data set, bank-additional, we utilized contained 4,119 records; this dataset was generated through randomly selecting 10% of the original dataset that contained 41,188 records. Within the data set, there is total of 20 input variables (10 of them are categorical and the other 10 are numerical) and 1 target variable. These input variables are categorized into four major categories as shown on Appendix F**.** The first category is “Bank Client Data”, which includes personal information pertaining to the clients that the bank is contacting. The second category is information “Related with the Last Contact of the Current Campaign”, which contains information about the previous contacts with the clients during this specific campaign. The third category is “Social and Economic Context Attributes”; this contains information about the economic conditions in Portugal at the time of the campaign. The fourth category is “Other Attributes”, which contains information about the number of contacts made to the client and the outcome of the previous marketing campaign. We analyzed each category regarding that every categories had its individual importance.

The data had to be preprocessed before in use because many of the records had missing values, which were labeled as ‘unknown”. To repair the missing data, we decided to delete all the records that were ‘unknown’. Even though we have to be cautious about removing the data because losing data will not incur a positive impact on the model, we decided to remove all of them for the pragmatic reasons. As a result, the number of records of the data set decreased from 4,119 to 3,090, the total number of subscriptions decreased from 451 subscriptions to 370 subscriptions, and the percentage of total subscriptions decreased from 12.0% to 10.9%. Although the removal of a missing data led the decrease in number of records, subscriptions, and percentage as shown above, our decision was necessary to avoid introducing any form of false results when interpreting and analyzing the data. Because the percentage of total subscriptions dropped only by 1,1%, it seemed that the missing data did not have much influence on the data as a whole, which ensures the validity of the data. More interesting aspect about the total subscription percentage was that it was a low percentage for modeling purposes. However, the low percentage is reasonable considering that the data set described a telemarketing campaign of the bank, in which most clients are likely to *hang up.*

In addition to eliminating the missing data, we decided to remove ‘default’ and ‘duration’ from the input variables. It was necessary to remove the default variable because the records of default were all ‘no’, which indicates that the bank only marketed to the clients without defaulted credit. Since a term deposit require the clients to spend a considerable amount of cash, the bank would only target the clients with good credit, making the default variable inadequate for our particular model. Moreover, ‘duration’ must be excluded from the input variables because this data can be collected only after the decision for subscriptions is made. Therefore, duration at a practical standpoint does not hold any power in making implications about the client’s subscription decision.

**Analysis: Classification Tree**

The main purpose of using the classification tree was to identify the significant variables in predicting the client’s likelihood of subscription. Later, we ran the logistic regression using these variables to find out the specific influence of the variables on predicting the target variable. To run the classification tree accurately, it was necessary to balance the number of outcome for the target variable. As described above, our total percentage of subscriptions was around 10%; this percentage indicates that the amount of non-subscriptions was 9 times the amount of subscriptions. Having such imbalanced ratio for the target variable would cause the problems on the classification tree because most results will get classified as ‘non-subscription’. Therefore, we manually augmented the number of subscriptions to balance out the number of non-subscriptions and subscriptions. Moreover, we set the complexity parameter as .01 to facilitate the visualization of the pruned tree. By doing so, most trees can be pruned to under 10 nodes that it was easy to visualize the significant variable, which is the first splitting variable shown on the top of the tree.

We adopted two different methodologies to run the classification tree. The first method was to select the most significant variables from each of the four categories. As shown in Appendix A**,** we generated four different classification trees using the variables of each respective categories. As a result, the four significant variables that we identified from each categories are: ‘job’ for a ‘Bank Client Data’ category; ‘month’ for a ‘Related with the Last Contact of the Current Campaign’ category; ‘pdays’ for a ‘Other Attributes” category; ‘euribor3m’ for a ‘Social and Economic context attributes’. The second method was to select four most significant variables by generating the tree using all input variables at once, which is shown in Appendix B. We started with including all 18 variables to generate the classification tree. Then, we identified the significant variable of the tree and kept generating the trees with the remainder of the variables. For instance, we generated the tree with 18 variables, then 17 variables, then 16 variables, and so on. Through this process, the four significant variables we identified are: ‘nr.employed’, ‘euribor3m’, ‘emp.var.rate’, and ‘pdays’ (listed in an order of identification). Overall, we obtained the different sets of significant variables as the result of generating the trees through two different methodologies. Utilizing these sets of significant variables to create the models for logistic regression, we aim to find the model with the highest predictive accuracy because that model would ensure the Bank Institution to identify the most leveraging factors affecting the success rate of its telemarketing. Moreover, interestingly, two of the variables, ‘euribor3m’ and ‘pdays’, of the significant variables generated above were identical. This concurrence may imply that ‘euribor3m’ and ‘pdays’ are the variables that influence heavily on the likelihood of client’s subscription. With taking account of this observation of the variables, we later used the logistic regression to verify the validity of the observation’s potential implication along with the predictive accuracy of the models.

**Analysis: Logistic Regression**

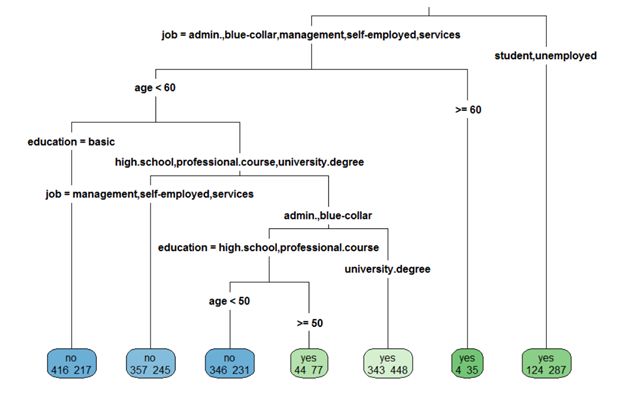
After identifying the two sets of four significant variables using the classification tree, we ran the logistic regression to analyze the model’s predictive accuracy through comparing the specificity and the misclassification rate of each models. Thus, to verify whether using these significant variables improve the model or not, we also ran the logistic regression using all input variables (shown in Appendix E), making it serve as the control model for the models created through a use of significant variables. We used .5 as our cutoff value and used the original ‘bank.csv’ with 3090 records as our dataset for running the logistic regression.

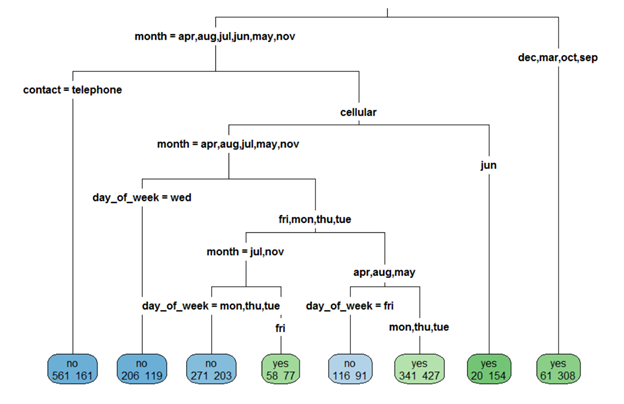
Looking at the confusion matrix for all models, the specificity was 98.6% and the misclassification rate was 10.6%for the model obtained through Appendix A method, which is shown on Table C2; the specificity was 98.8% and the misclassification rate was 10.6%for the model obtained through Appendix B method, which is shown on Table D2; and the specificity was 97.8% and the misclassification rate was 11%for the model with all variables shown on Table E2. Analyzing these results, we could conclude that all of the models had similar predictive accuracies. These specificity values of over 90% for all models verify that our model is efficient in identifying the clients that are not going to subscribe to the bank term deposit. Also, the misclassification rate around 11% for all models indicate that the model only misclassifies the actual results 11% of the time. Most importantly, the fact that the models generated through using significant variables identified through classification tree have almost identical predictive accuracy results compared to the model with all variables signify that these significant variables are, indeed, influential for predicting the likelihood of a client’s subscription because these models were able to perform similar predictive power with fewer predictors. Considering these specificities and misclassification rates shown above, it becomes evident that the models generated through Appendix A and B method are beneficial for the bank’s telemarketing because they will be able to use the model for not reaching the non-subscribing clients; instead, they can focus on marketing towards the potential clients that are going to subscribe.

Which attributes should the marketing department of the Bank especially take account when making decisions for telemarketing? After validating the models, we looked into the specific influence of the significant variables on predicting the likelihood of subscription or non-subscription. We considered all variables with the p-value less than .05 (indicated by the star sign next to the variable shown in Table C1 and D1) as an important influencer. In table C1, we found that ‘euribor3m’, ‘pdays’, and ‘months’ especially March, June, July, and August were significant variables. When the ‘euribor3m’, which is the 3 month Euro Interbank Offered Rate, increases by 1% 3 month rate, then the odds of the client subscribing to the bank term will decrease by a factor of 2e^-.5061483, all else constant. This phenomenon is because the clients are likely to keep their money instead of investing their money since the value of holding their money increases as the result of increase in the Euro Interbank Offered Rate. When ‘pdays’ increase by 1 day, then the odds of the client subscribing to the bank term will decrease by a factor of 7.78e^-9. This phenomenon indicates that the likelihood of client subscribing to the bank term decreases every day after the contact is made; therefore, the bank must demand the clients for the subscriptions as soon as possible to attain a higher success rate of subscription. Furthermore, table C1 shows that telemarketing on March, June, July, August, and September have a high correlation with the high subscription likelihood of the clients. Therefore, reaching the customers during those months listed above will allow the Banks to have better chance of being successful. Interesting aspect about the variable ‘month’ is that all of the significant months happen between spring and summer; this may imply that certain confounding variables affect the clients become more comfortable in subscribing the bank term deposit during spring and summer season. In table D1, we identified ‘nr.employment’ and ‘pdays’ as the significant variables. As number of employees increase by 1 member, then the odds of the client subscribing to the bank term will decrease by a factor of 3.79e^08. This means that the clients with more number of employees are not likely to subscribe possibly due to the high wage for employees. Therefore, from a strategical standpoint, targeting the clients with low number of employees would incur a higher success rate of telemarketing for the bank. In fact, the variable ‘pdays’ is the only variable that is recognized as the significant variable in both models. This result testifies that the bank’s effort on making the clients to subscribe right away is really important for them to be successful at telemarketing. When the clients are given more time to consider about a bank term, they are not likely to subscribe. Looking at these specific type of impact that each variable had on the client’s likelihood of subscription, it seems plausible that targeting the right clients at a right time through a right method is essential for the telemarketing to be successful.

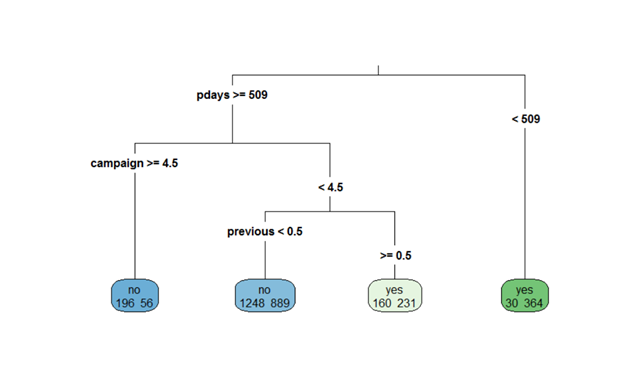
Lastly, we created the ROC curve and the lift chart to visualize the results of our models, which are shown in Image C3, D3, and E3. Since all models had almost identical predictive accuracy results, we only analyzed C3 for an interpretation of ROC curve and lift chart. For an ROC curve, our initial point (.227,.106) was located near bottom left, which was created with cutoff value of .5. For an ROC curve, the better performance is reflected by the curve closer to the top-left corner. To identify the optimal point that is closest to the top-left corner, we changed the cutoff values around. As a result, the optimal point was obtained at a cutoff value of .14 with (.587,.108) coordinate. Although the optimal point of the curve could not reach the top left corner of sensitivity near .9 and ‘1-specificity’ near 0, at a pragmatic standpoint, our model was still a statistically significant model for predicting the subscription of the clients because there exist many confounding variables in the reality. The lift chart is utilized for testing the effectiveness of the model. Looking at our lift chart, when we go through 300 clients with the model, we can identify 100 clients, but without the model we can only identify 30 clients. This result indicates that our model (ranked in a decreasing order) is useful compared to a random assignment because we could identify 70 more clients by going through the same number of clients.

**Appendix A**

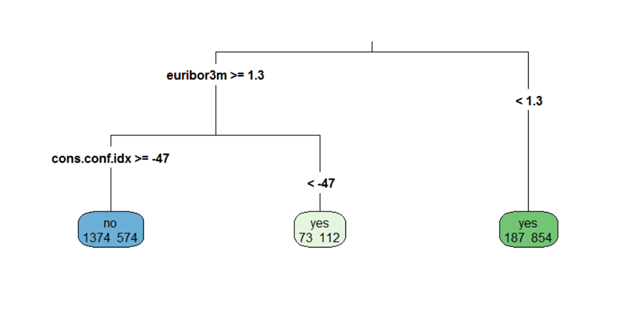
**Image A1: Classification Tree for Bank Client Data**

**Image A2: Classification Tree for Related with the Last Contact of the Campaign**

**Image A3: Classification Tree for Other Attributes**

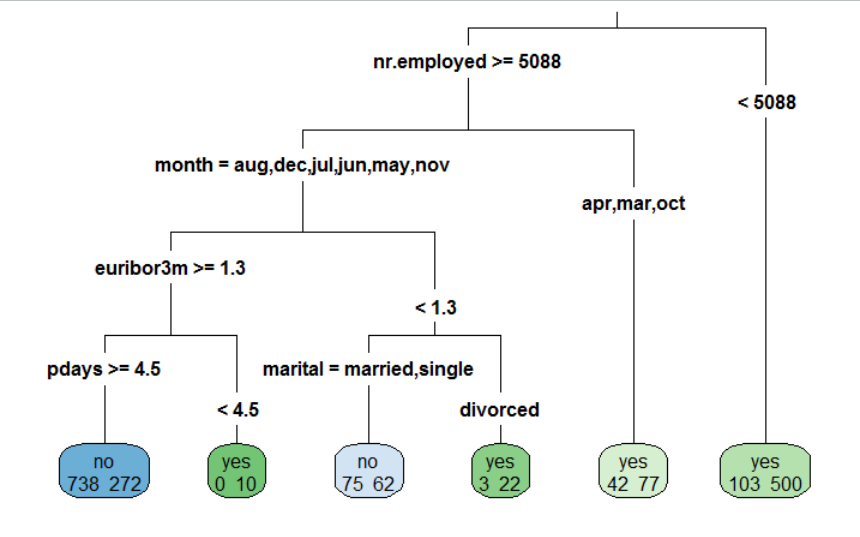


**Image A4: Classification Tree for Social and Economic Context Attributes**

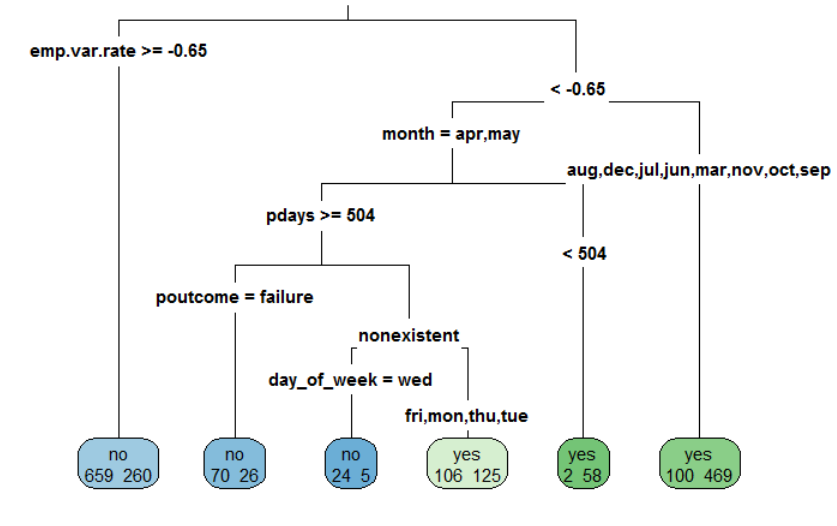


**Appendix B**

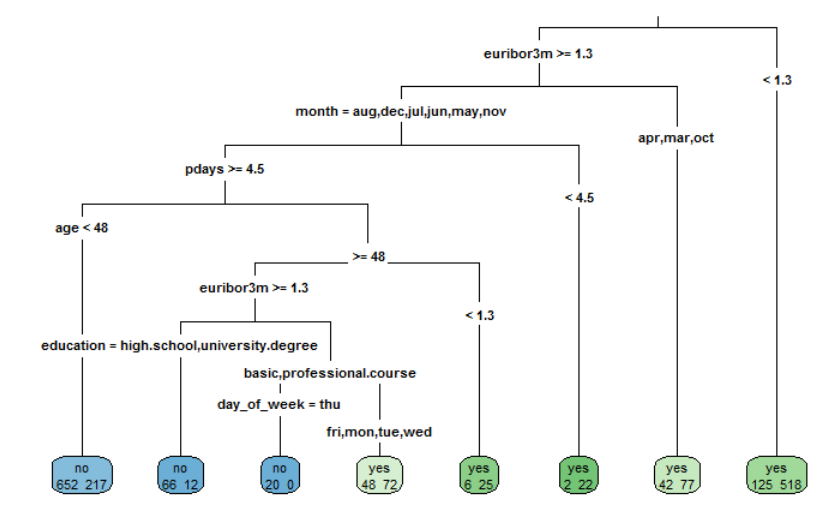
**Image B1: Classification Tree for All Variables**



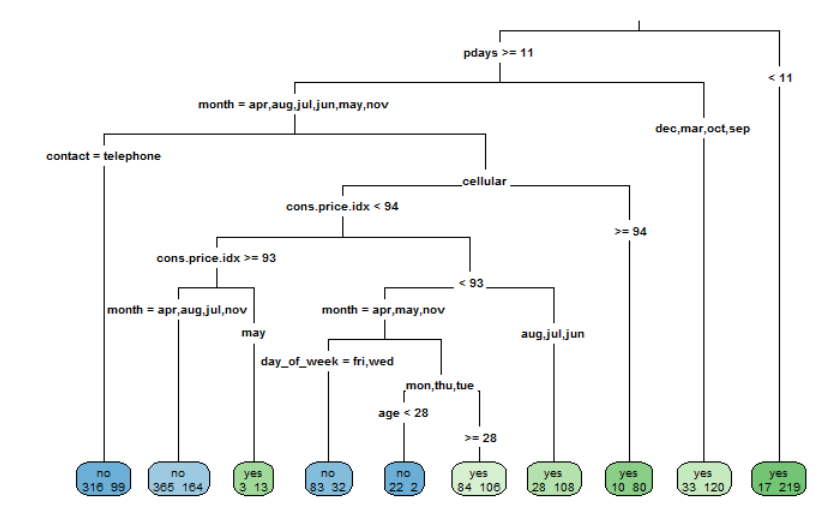
**Image B2: Classification Tree for All Variables - “nr.employed”**



**Image B3: Classification Tree for All Variables - “nr.employed” - “emp.var.rate”**

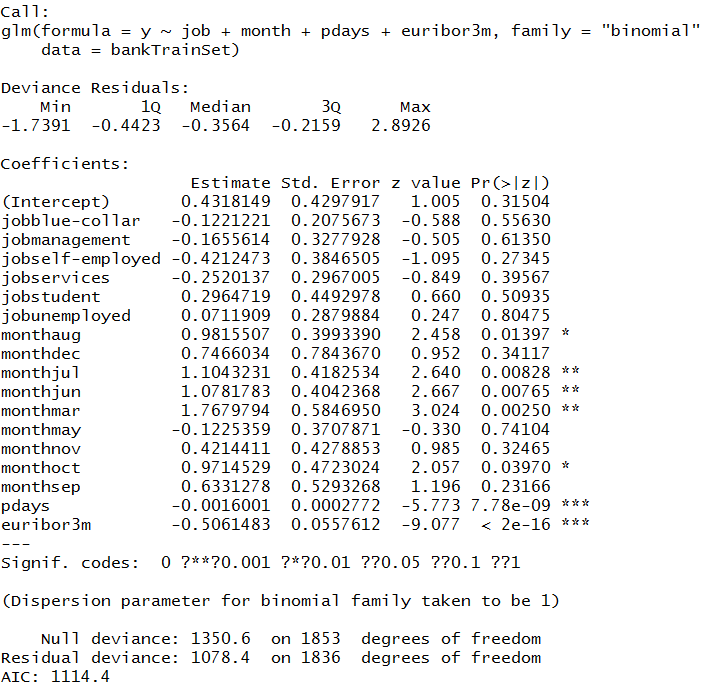


**Image B4: Classification Tree for All Variables - “nr.employed” - “emp.var.rate” - “euribor3m”**

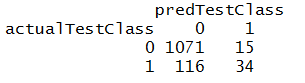


**Appendix C**

**Table C1: Log Regression table for significant variables obtained from Appendix A result.**



**Table C2: Log Regression confusion matrix for significant variables obtained from Appendix A result.**

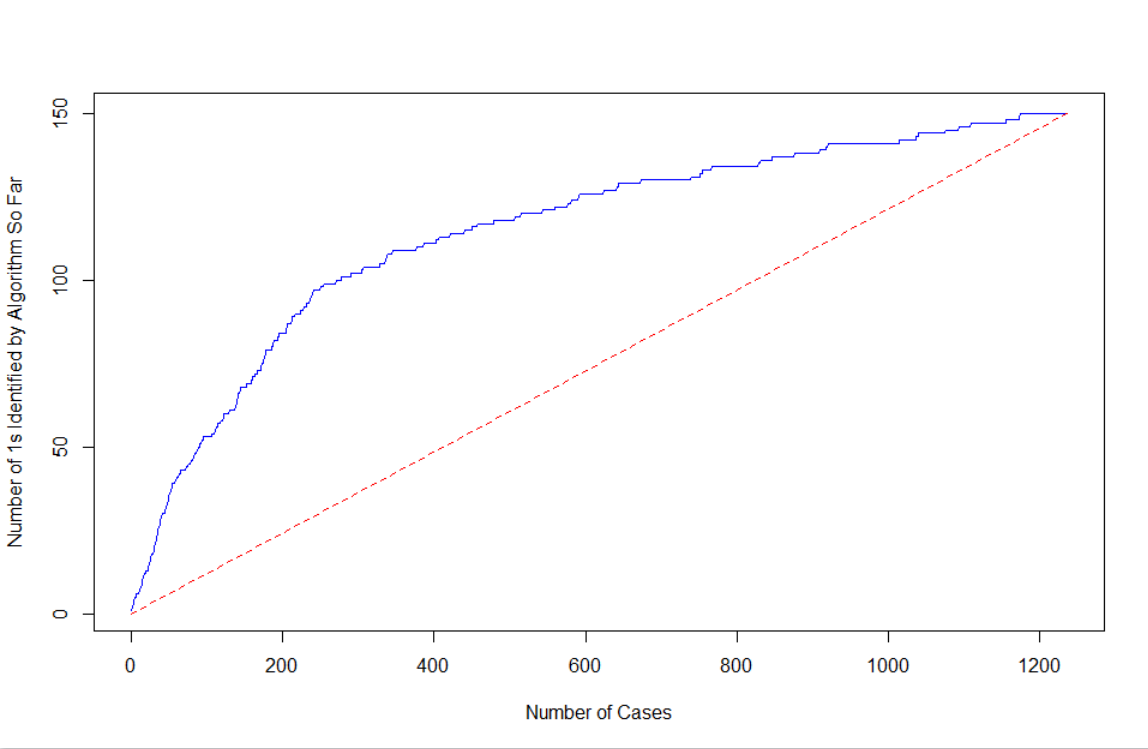
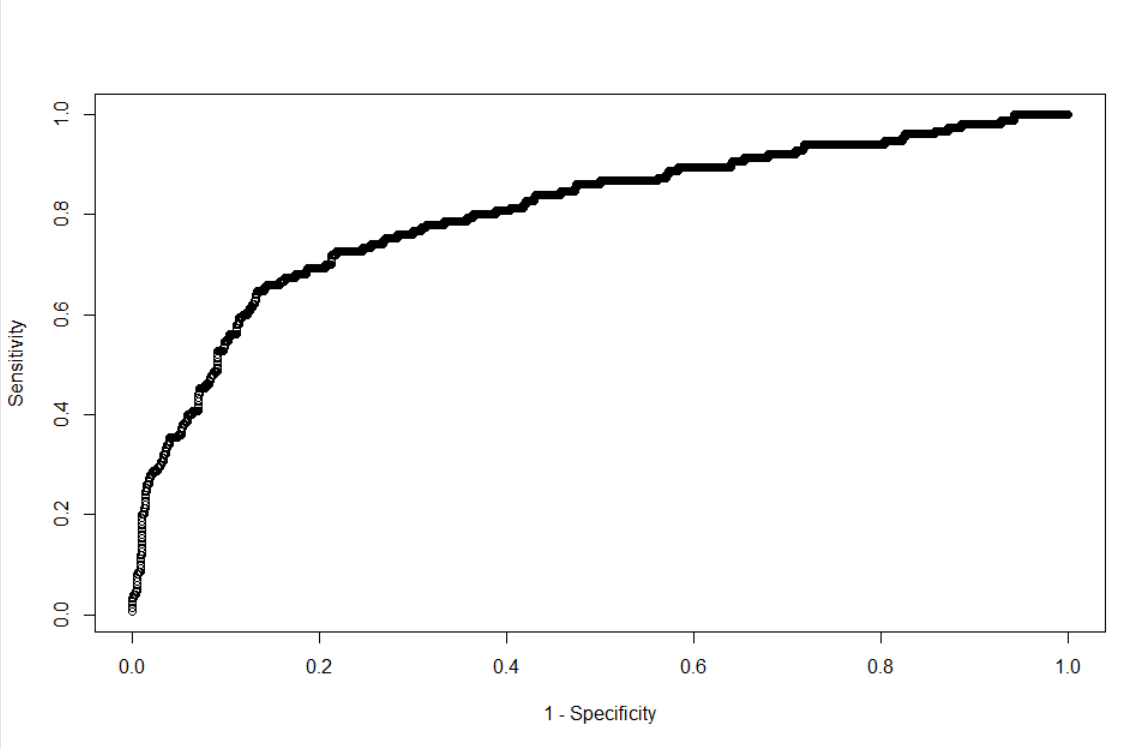


**Sensitivity = 34/(116+34)\*100 = 22.7%**

**Specificity = 1071/(1071+15) = 98.6%**

**Misclassification = (116+15)/1236= 10.6%**

**Image C3: Lift Chart and ROC Curve for log regression by using the significant variables obtained from Appendix A result.**



**(30,300)**

**At a cut-off value of .14**

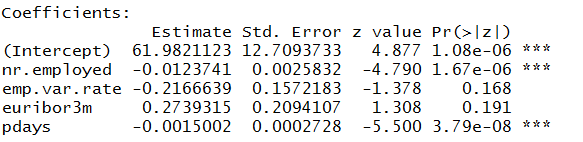
**(.587,.108)**

**At a cut-off value of .14**

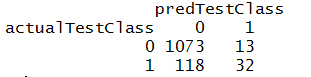
**(100,300)**

**Appendix D**

**Table D1: Log Regression table for significant variables obtained from Appendix B result.**



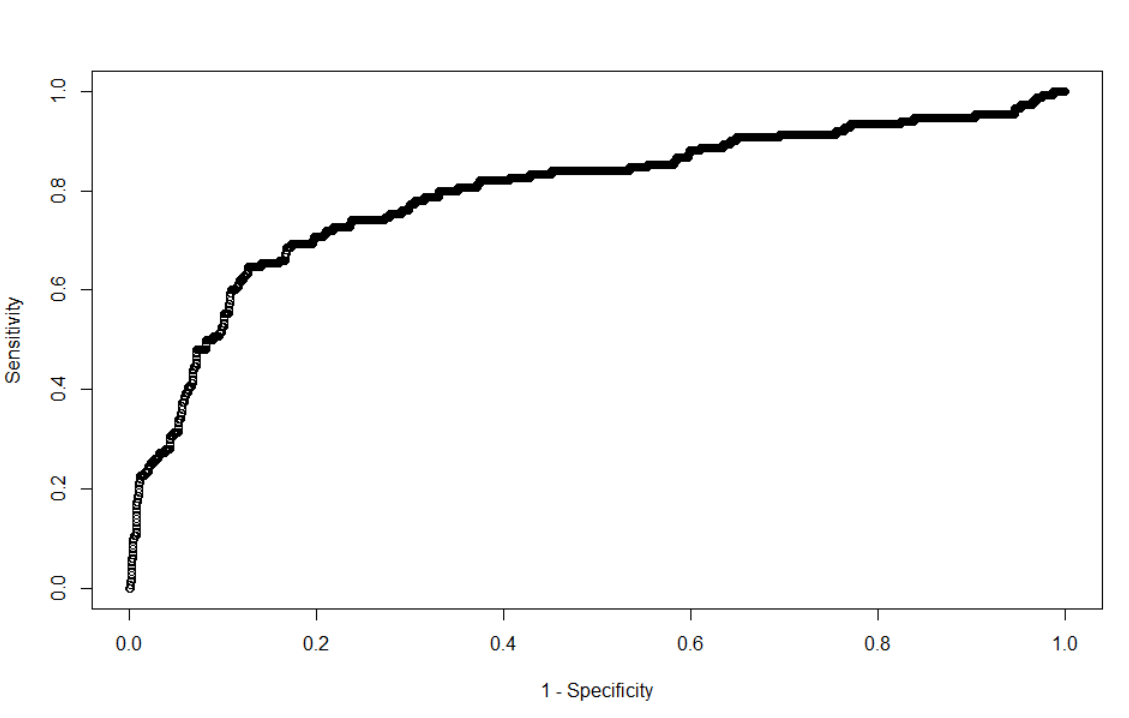
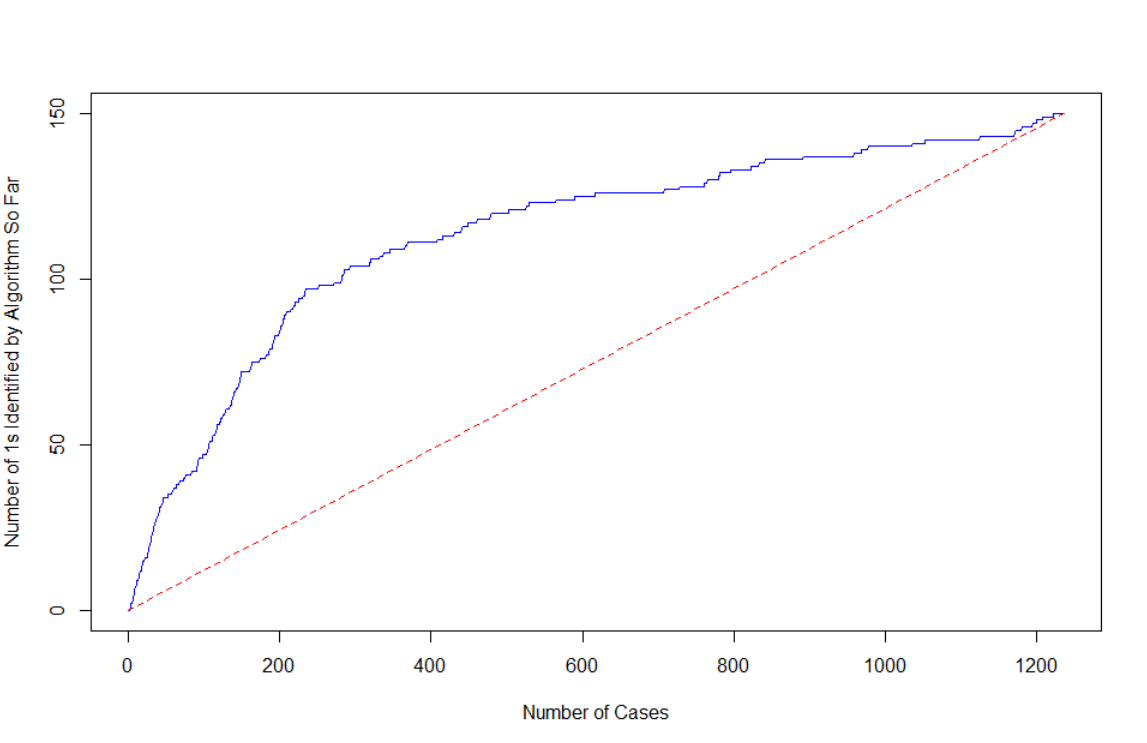
**Table D2: Log Regression confusion matrix for significant variables obtained from Appendix B result.**



**Sensitivity = 32/(118+32)\*100 = 21.3%**

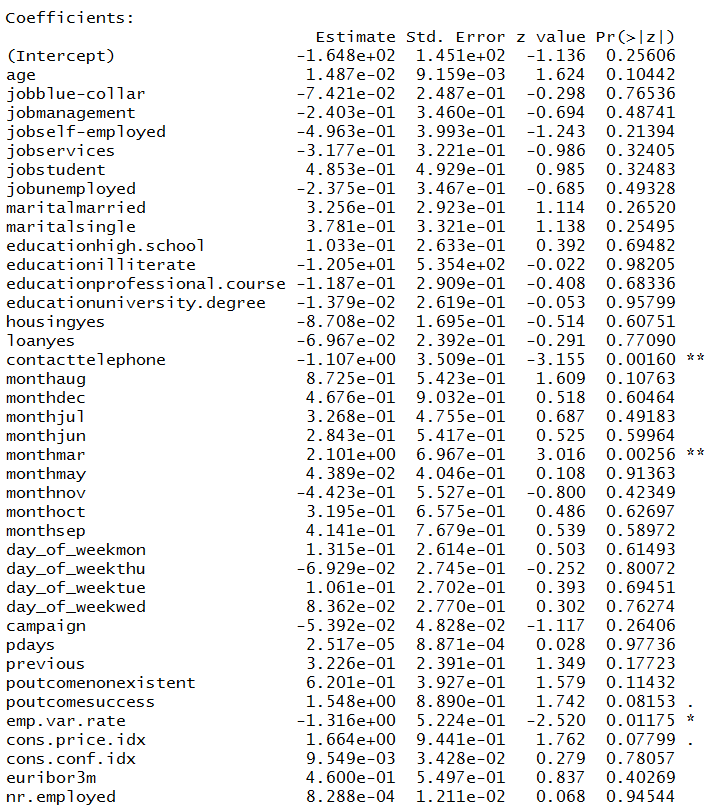
**Specificity = 1073/(1073+13) = 98.8%**

**Misclassification Rate = (13+118)/1236=10.6%**

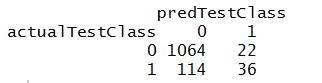
**Image D3: Lift Chart and ROC Curve for log regression by using the significant variables obtained from Appendix B result.** 

**Appendix E**

**Table E1: Log Regression table using all variables.**



**Table E2: Log Regression confusion matrix using all variables.**

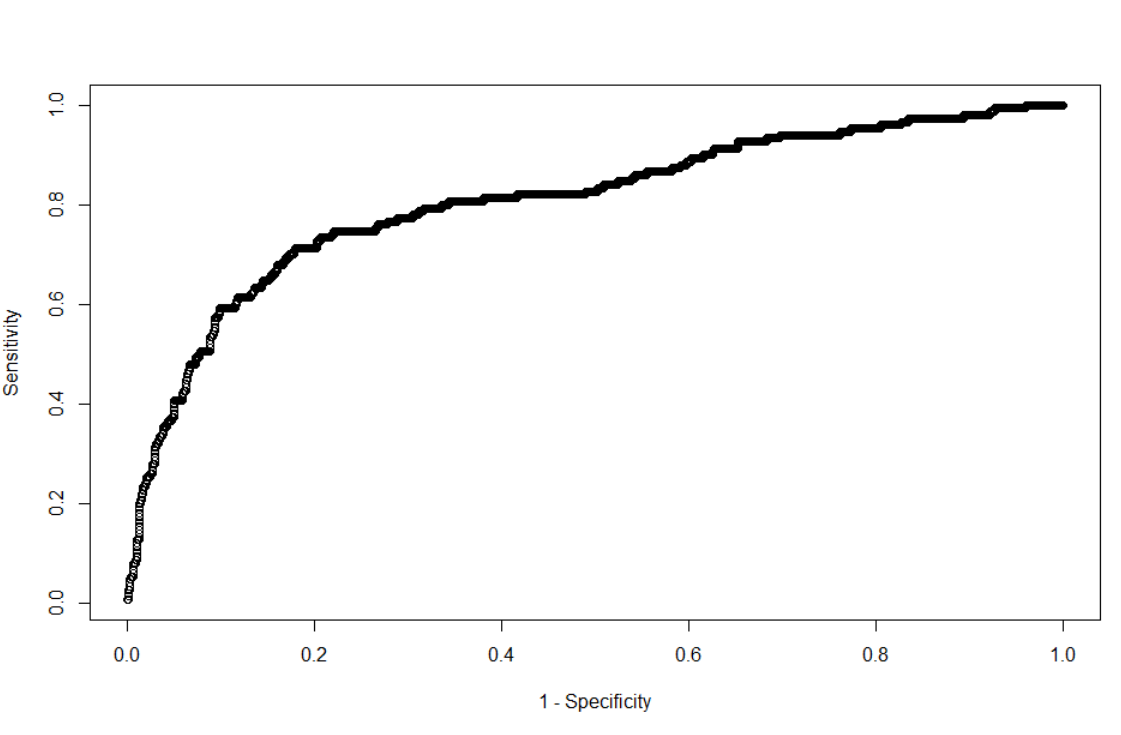
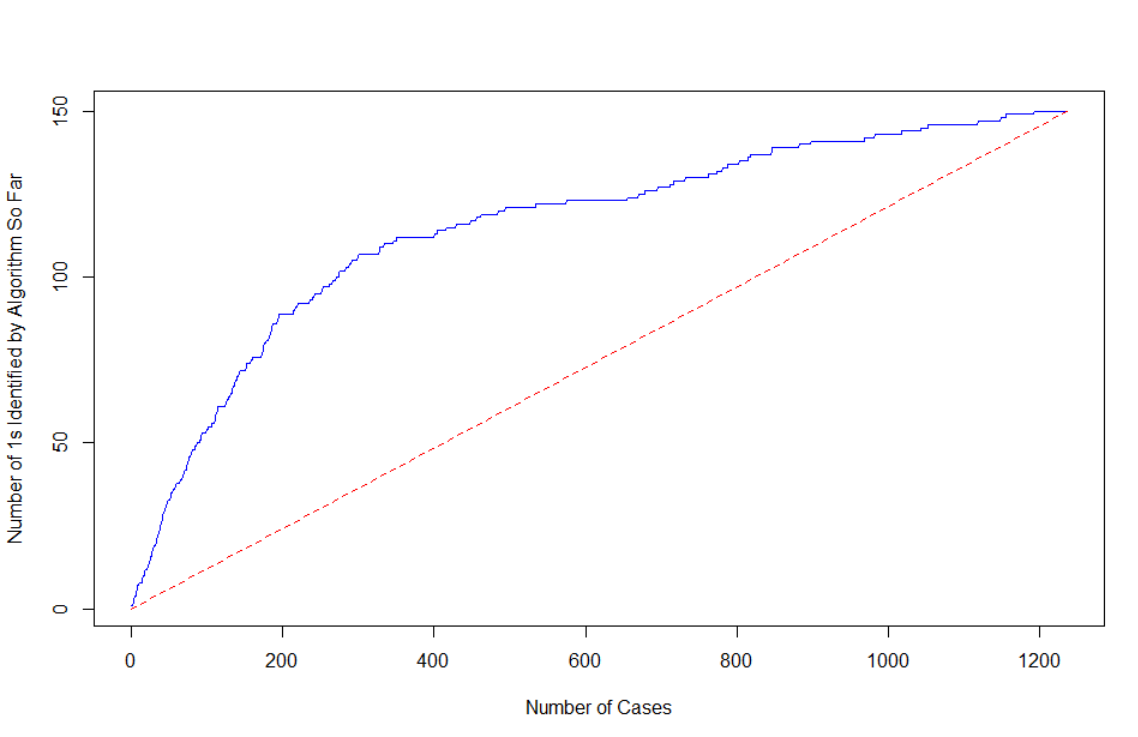


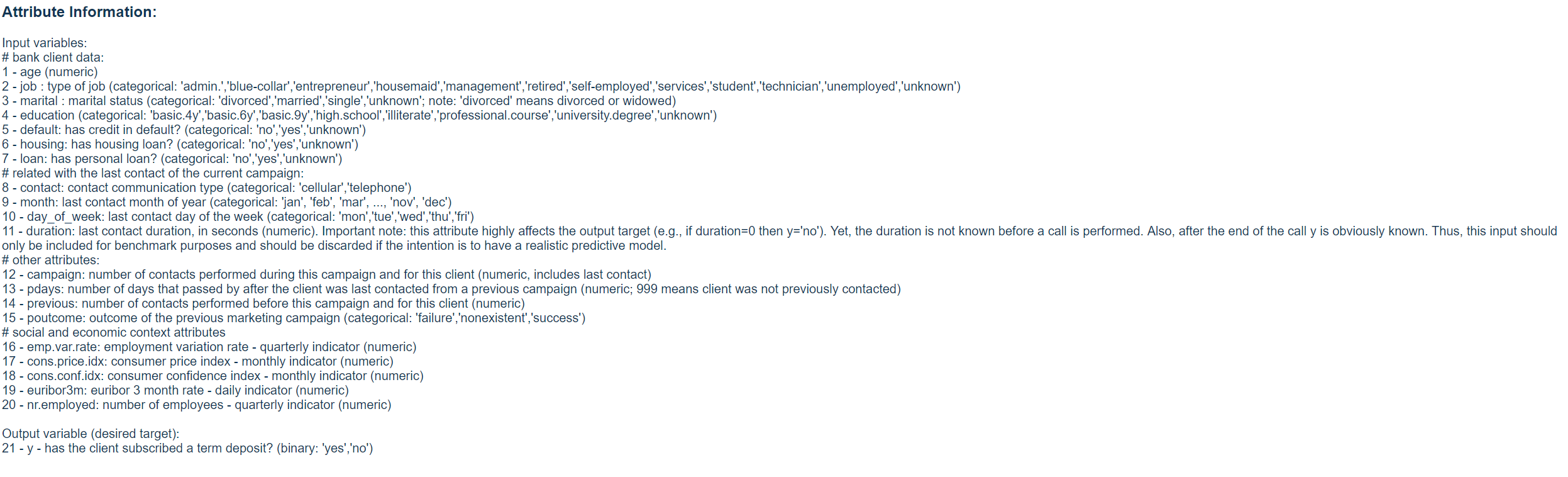
**Sensitivity =36/(114+36)\*100=24%**

**Specificity=1064/(1064+22)=97.8%**

**Misclassification=(22+114)/1236=11%**

**Table E3: Lift Chart and ROC Curve for log regression by using all variables.**



**Appendix F**