Report for PA04

Group DA-02

Assumptions and instructions :

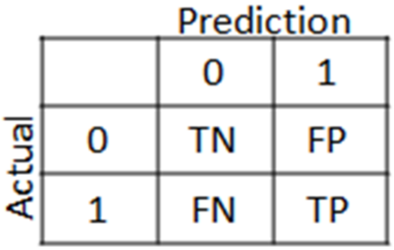
- The ratio for training set to test set is kept at 9:1 currently . It can be changed on the line $trainIndex <- createDataPartition(tree\$churn,\ p = 0.9,\ list = FALSE)$
- Packages need to be pre-installed by running sudo R in terminal and then installing each package in churn.r with install.packages('package_name'). The packages required are mentioned at the start of the code.

1. By observing all the 22 variables we see that the variable variable area_code num does not affect churn in any way, so it is dropped.

2. Most of the code is self-explanatory. For binary decision tree, we are opting for rpart method. The output for one of the runs is stored in terminal.out.txt, with plots saved in Rplots.pdf in the order they are executed. Note that the output may vary slightly each time since train and test sets are chosen arbitrarily.

3. In the problem statement, churn variable is false in majority cases and hence true negative value will be very high. However, we will not be changing the basic formulae as we are trying to develop a model to predict the customers to switch the company with sufficient accuracy.

4. After developing the three classification models,we then find confusion matrices for each model,matrices are calculated on test data set. We then calculate accuracy,precision and recall for each model,using the following formulae.
   $Accuracy = (TP + TN)/(TP + TN + FP + FN)$
   $Precision = TP/(TP + FP)$
   $Recall = TP/(TP + FN)$



		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

-Confusion Matrix

5. After obtaining right set of predictor variables,we choose the best model by plotting ROC curves and comparing them by finding the model which has larger area in the graph.

6. After comparing the three models, by comparing the ROC's of each of them, we observe that(for majority of the runs), the ROC for SVM model has the maximum AUC (Area Under Curve), and hence is the best model to be selected out of the three.

7. (Part 5) We see that total_eve_minutes is a subset of total_eve_calls and total_eve_charge. So we can remove it and introduce a new variable,say, total_eve_rate, which related to the above two, as :

   total_eve_rate = total_eve_charge / total_eve_calls.

   Similarly we can introduce 2 new variables, total_day_rate and total_night_rate, for the variables, {total_day_calls, total_day_charge}, and, {total_night_calls, total_night_charge}, respectively, while removing variables total_day_minutes and total_night minutes as well, to optimise accuracy of all the models.

8. The maximum accuracy obtained was 0.953 (95.33%)

Gudepu Prithviraj Reddy	14CS10016
Kinsuk Das	14CS10025
Ragireddy S. Charan Reddy	14CS10037
Kaustubh Hiware	14CS30011
Surya Midatala	14CS30017