STAT 515 Applied Statistics and Visualization for Analytics (Fall 2016)

Assignment 3

Due Date: 10/27/2016 Thursday 11:59 PM

1. Customer Churn Study (50pts)

Customer Retention

Telecommunications companies providing cell phone service are interested in customer retention. In particular, identifying customers who are about to churn (cancel their service) is potentially worth millions of dollars if the company can proactively address the reason that customer is considering cancellation and retain the customer. The file **Churn.csv** contains customer data to be used to classify a cell phone customer as a churner or not. A brief description for the variables are provided below:

Churn	= 1, if the customer churns (cancels the service); = 0, if not.
AccountWeeks	Number of weeks that since the customer's account was open
ContractRenewal	= 1, if the customer has renewed his/her contract; = 0, if not;
DataPlan	= 1, if the customer has participated the data plan; = 0, if not.
DataUsage	Data usage (GB) of the customer per month
CustServCalls	Number of times that the customer calls the service center
DayMins	Number of minutes of service usage per day
DayCalls	Number of calls per day
MonthlyCharge	Monthly charge for the current service for the customer
OverageFee	Fee for the extra minutes
RoamMins	Usage (in minutes) of roaming data

The average net loss resulting from classifying customers into churning and not churning categories is given in the following table:

		Predicted	
		Churn	Not Churn
Actual	Churn	$0	$200
	Not Churn	$50	$0

NOTE: You do not need to use R codes to solve every question. The R-Codes uploaded on blackboard, under the in-class exercise panel in Lecture 7 (Logistic Regression Models) will be useful for this assignment.

a. What is the overall churning rate in the original dataset?

b. Partition the data into training (60%) and validation (40%) (set the random seed to 12345). Using the **training** dataset, generate the scatterplot of **Churn** (Y axis) vs. **RoamMins** (X axis) colored by **ContractRenewal**.

c. Build a logistic regression model by including only the numerical variables using the training data, then apply the constructed model to the validation set and set the cutoff level at 0.5. Report the classification/confusion table (frequency counts), the overall accuracy rate, misclassification error rate, sensitivity value for the **validation** dataset. (Don't include classification measures of the training dataset).

d. Using all the predictors, build a full logistic regression model using the training data. Which categorical predictor has the most impact on the response variable? How does the odds of churning change with a change (e.g. changing from 0 to 1) in this categorical variable? Please report the change in odds. Also, by setting the cutoff level at 0.5, report the classification/confusion table (frequency counts), the overall accuracy rate, misclassification error rate and sensitivity value for the **validation** dataset. (Don't include classification measures of the training dataset).

e. Using all the predictors, build a logistic regression model with backward selection method. Write down the logistic regression equation (with only the significant predictors). By setting the cutoff level at 0.5, report the classification/confusion table (frequency counts), the overall accuracy rate, misclassification error rate and sensitivity value for the **validation** dataset. (Don't include classification measures of the training dataset).

f. Utilizing the average net loss table (provided at the beginning of question context) along with the classification/confusion matrix, compare the performance of models developed in parts (c), (d) and (e) with respect to the average net loss in the **validation** dataset using cutoff level at 0.5. Which model is the best (i.e. which model has the lowest average net loss)? Show your calculation work and explain briefly.

g. Provide the **validation** ROC curves, AUC values and Lift charts for all the models developed in parts (c), (d) and (e). Which model has the best performance with respect to the AUC value.

h. For the best model you identified in part (g), find the best selection of cutoff level with respect to the overall accuracy rate in the **validation** dataset. In doing so, change the cutoff level from 0.1 to 0.9 with increment of 0.1, report the accuracy rate associated with each cutoff level and find the cutoff level with the highest accuracy rate. (Note: changing the cutoff level will change the classification/confusion matrix, and therefore will also change the accuracy rate.)

Submission:

1. Prepare a pdf (or MS word) file with answers to above questions, brief explanations for the analysis you perform. Please provide your R codes at the end of the submitted file as appendix.
2. Name the file as "LastName, FirstName-HW3.pdf" (for example, "Ji,Ran-HW3.pdf") and submit it on blackboard.