Data Analysis Problem 18.11.2017

Customer Churn Prediction

Customer Relationship Management (CRM) is a comprehensive strategy for building, managing and strengthening loyal and long-lasting customer relationships. It is broadly acknowledged and extensively applied to different fields, for example, telecommunications, banking and insurance, retail market, etc.

In any business strategy, one of its main objectives is customer retention. For example, Vodafone should asses the customers' activities (from their call logs) and hence predict the customers' churn rate, so that a measure (a special incentive to customers) can be taken to retain them accordingly. The importance of this objective is obvious, given the fact that the cost for customer acquisition is much greater than the cost of customer retention (in some cases it is 20 times more expensive).

Thus, tools to develop and apply customer retention models (churn models) are required and are essential Business Analytics applications. In the dynamic market environment, churning could be the result of low-level customer satisfaction, aggressive competitive strategies, new products, regulations, etc. Churn models aim to identify early churn signals and recognize customers with an increased likelihood to leave voluntarily.

A customer call data with 5000 records is known to us related to the customers' call logs of a mobile phone service provider. We call this data as Customer Churn Data (and abbreviated as CCD for brevity). The data set is synthetic and based on claims similar to a real world service provider. The meta data of the CCD data is as follows.

Variable	Name	Type
Account_length	(number of months active user)	Num
Total_eve_charge	(total charge of evening calls)	Num
area code Num	(area code of customer)	Num
total_night_minutes	(total minutes of night calls)	Num
international_plan	(local/international call)	Binary (Yes/No)
total_night_calls	(total number of night calls)	Num
voice_mail_plan	(voice mail or normal)	Binary (Yes/No)
total_night_charge	(total charge of night calls)	Num
number_vmail_messages	(number of voice-mail messages)	Num
total_intl_minutes	(total minutes of international calls)	Num
total_day_minutes	(total minutes of day calls)	Num
total_intl_calls	(total number of international calls)	Num
total_day_calls	(total number of day calls)	Num
total_intl_charge	(total charge of international calls)	Num
total_day_charge	(total charge of day calls)	Num

number_customer_service-calls	(number of calls to customer service)	Num
total_eve_minutes	(total minutes of evening calls)	Num
total_eve_calls	(total number of evening calls)	Num
churn	(customer churn - target variable)	Binary (Yes/No)

A data set is attached for your input to this project. A paper related to this project is also attached

Problems

1. Review the predictor variables and guess what their roles in customer churn. You are free to create new derived variables from these predictors. [Hint: You have solved similar problems in Project #3]

2. Divide the data into training and test set. Make sure relative proportions of true and false in the target variable are maintained in training and test set. [Hint: Use stratified rando sampling : You can use data partition function of R Caret package.]

3. Using training data set, develop classification models using at least 3 classification techniques (1) Naïve Bayes' Classifier, 2) Any one decision tree classifier and 3) SVM classifier. You can get the classifier in R Caret package, for example, R-SVM for SVM classifier.]

4. Construct confusion matrices using test data set for each model. Compute a) Accuracy, b) Precision and c) Recall for each model.

5. Try to improve classification accuracy by choosing right set of predictor variables and model parameters and choose the best model. [Hint: Follow ROC curve, or any other if you think suitable.]