Data Mining Modeling Workshop

A telecommunication provider is concerned about the number of customers it is losing to competitors. They want to retain as many customers as possible. It will be good to predict who are more likely to churn, so that they can target on these customers and provide them attractive offers. Over the years, they have accumulated numerous customer data, including customer personal data, transactional and usage data.

Description of the data

The TelcoChurn data set has 36 variables, 1000 customer records. The variable definitions are provided below.

Variable	**Description**
customerID	Customer ID
Region	Geographic indicator
Tenure	Months with service
Age	Age in years
marital	Marital status
address	Years at current address
income	Household income in thousands
Ed	Level of education
employ	Years with current employer
retire	Retired
gender	Gender
reside	Number of people in household
tollfree	Toll free service
equip	Equipment rental
callcard	Calling card service
wireless	Wireless service
longmon	Long distance last month
tollmon	Toll free last month
equipmon	Equipment last month
cardmon	Calling card last month
wiremon	Wireless last month
longten	Long distance over tenure

tollten	Toll free over tenure
equipten	Equipment over tenure
cardten	Calling card over tenure
wireten	Wireless over tenure
multline	Multiple lines
voice	Voice mail
internet	Internet
callid	Caller ID
callwait	Call waiting
forward	Call forwarding
confer	3-way calling
ebill	Electronic billing
custcat	Customer category
churn	Churn within last month

Instructions:

1. Start RStudio and Rattle.

   > library(rattle)

   > rattle()

2. Locate the TelcoChurn dataset: TelcoChurn.csv
3. Partition the dataset as default: 70/15/15, meaning 70% for training, 15% for validation, and 15% for testing. (or try 50/0/50)
4. Click the Execute button to load the data.
5. Check out the different roles assigned to Variables: input, target, or identification.

6. Click the Explore Tab to explore the data. (Optional)

7. Click the Model tab; Select Tree model; Select Traditional algorithm: rpart();

   Check out the default settings:

   Min Split, Min Bucket, Max Depth, Complexity, etc.

8. Click the Execute button to build a Decision Tree model.

9. Browse the Modeling results.

   Click the Rules and Draw buttons to explore the decision tree and rule set.

10. Experiment with different parameter settings on MinSplit, MinBucket, MaxDepth, etc.
11. Check out the R code generated in the Log tab.

12. Click the Evaluate tab.
13. Tick Tree model; select different evaluation types such as Error Matrix (Confusion Matrix), Lift, ROC, etc.
14. Identify the dataset for evaluation, usually the testing dataset.
15. Click the Execute button to view the evaluation results.

16. Select the evaluation Type as Score, Model as Tree, Data as Testing, click the Execute button to generate and save a CSV file.
17. Open the CSV file in EXCEL to view the result for each record in the dataset.

18. Click the Model tab; Select Neural Net type model; Specify Hidden Layer Nodes as 10.
19. Click the Execute button to build a Neural Net model.
20. Browse the NN Modeling results.
21. Experiment with different number of Hidden Layer Nodes.

22. Click the Model tab; Select SVM model.
23. Click the Execute button to build a SVM model.
24. Browse the SVM Modeling results.

25. Click the Evaluate tab; Compare the performance of all models built by using Error Matrix (confusion matrix) and ROC chart. Identify the best model.

26. Score all models on the testing data; view the prediction of each model on each record in the dataset.
27. Save your project.

