Data Mining Practical Exercise (Homework)

Deadline: April 21, 2024, 11:59 p.m. (GMT-5)

Developing Models for Customers Likely to Churn

Churn is a term used to indicate a customer leaving the service of one company in favor of another company. The aim of this exercise is to develop models that predict whether a customer is likely to churn, i.e., the company would like to know how to characterize the customers who may soon churn.

You should analyze the obtained models and compare them. Moreover, based on the knowledge gained from analyzing the data and the built models, you are required to propose several practical measures that the company can put into action to try to retain those customers likely to churn.

The data set contains 20 attributes, plus indication of whether the customer churned, and about 3333 customers. A description of the meaning of each attribute is given in the appendix.

The points described below should be explicitly addressed in your report. Notice that you can also build other models and try other strategies you may find suitable. Use your imagination.

The suggested data mining software for the described exercise is RapidMiner and excel (you can also use Python with sklearn, matplotlib and other libraries). Consider also using graphs and images in your report since they can improve the clarity and the strength of your conclusions. Refer and motivate in your report the use of any data pre-processing prior to the application of any data mining technique.

State and Area Code attributes have a lot of erroneous data; you must remove them from the dataset.

1 Exploratory Data Analysis

The first step in approaching a data mining problem is to delve into the data, identify any interesting relationships between the attributes, and formulate some initial hypothesis, i.e., possible associations between the attributes and the class. Graphical tools can aid you in this phase.

(1) Find possible correlated variables. For instance, find out whether the data shows that the number of minutes and amount charged tend to increase as the number of calls increases. Use the correlation matrix operator and explain why you think some variables are correlated.

(2) Are there any variables that can be eliminated? Justify your answer and motivate the possible benefits of doing so (if any).

(3) Investigate the proportion of churners and non-churners among customers who have (not) selected an international plan (Inter Plan). What can you conclude? (use graphs and/or tables).

(4) Investigate possible relationships between the No of Calls Customer Service and Churn. What can you conclude? (use graphs and/or tables)

(5) Investigate possible relationships between the Total Day Min and Churn. What can you conclude? (use graphs and/or tables)
(6) Investigate possible relationships between the Total Evening Min and Churn. What can you conclude? (use graphs and/or tables)
(7) Investigate possible relationships between the remaining variables and Churn. (use graphs and/or tables)
(8) Summarize in a table your findings so far, concerning the predictive value of each attribute with respect to churn.
(9) Compare your conclusions with the results obtained by using a feature weights operator in RapidMiner. Do not forget to indicate which operator you have used.

Exploring Multivariate Relationships

Next, you are asked to investigate possible multivariate associations of numerical attributes with churn.

(10) Study the scatter plot of No of Calls Customer Service versus Total Day Minutes. Identify possible high-churn areas (if any).
(11) Study the scatter plot of Total Day Min versus Total Evening Min. Identify possible high-churn areas (if any) and try to quantify the churn rate in these areas with respect to the entire data set.

2 Building a Classifier: Decision Trees

Build a decision tree in RapidMiner. Include in your report a figure with the decision tree you have obtained.

(12) What can you conclude from the model you have obtained? Compare your conclusions with the ones you have obtained previously (section 1).
(13) Select some performance measures and evaluate the model with cross validation. Justify the choice of performance measures.
(14) Compare the performance of this classifier with a classifier that always predicts the majority class (default model in RapidMiner, ZeroR in Weka or DummyClassifier in sklearn).

3 Building a Classifier: Rules

Build a set of rules with Rule Induction operator and cross validation. You may consider eliminating a few more attributes. If so, indicate which attributes you have eliminated and why. Include in your report the rules you have obtained.

(15) What can you conclude from the model you have obtained? Compare your conclusions with the ones you have obtained previously (section 1 and section 2).
(16) Using the performance measures, you have selected for the evaluation of the decision tree model, evaluate the current model, and compare its performance with the previous one (obtained in section 2).
(17) Compare the performance of this classifier with a classifier that always predicts the majority class.

4 Cost-sensitive Learning

Consider that the cost for a company of losing a customer is higher than the cost of offering some incentives to a customer, even when he is not likely to churn anyway.

(18) Does your data pose any problem in this perspective? If so, describe the problem.
(19) Describe how the problem can be tackled. Use then the tools available in RapidMiner to build another model according to the ideas you have described.
(20) What can you conclude from the model you have obtained? Compare your conclusions with the ones you have obtained previously (section 1, 2, and 3).
(21) Compare the performance of the model you have obtained with the previous ones. Investigate whether any differences are statistically significant.

5 Clustering

Investigate the use of clustering techniques (for instance with K-means), to segment the customers to get groups of customers with similar service usage characteristics.

(22) Profile the clusters, i.e., what you can learn about the types of records falling into each cluster. Justify the number of clusters you have chosen.
(23) Investigate whether you can use the information obtained by clustering to assist you in the churn classification problem.

Appendix

You can find below a brief description of the meaning of each attribute.

- State: discrete variable that indicates the state where the customer lives.
- Account Length: integer variable that indicates how long the account has been active.
- Area Code.
- Phone Number.
- Inter Plan: binary variable indicating whether the customer has an international plan.
- VoiceMail Plan: binary variable indicating whether the customer has a voice mail plan.
- No of Vmail Mesgs: integer variable that indicates the number of voice mail messages.
- Total Day Min: continuous variable that indicates number of minutes the customer used the service during daytime.
- Total Day Calls: integer variable that indicates the number of calls during daytime.
- Total Day Charge: continuous variable that indicates how much was charged for using the service during daytime.
- Total Evening Min: continuous variable that indicates number of minutes the customer used the service during evening time.
- Total Evening Calls: integer variable that indicates the number of calls during evening time.
- Total Evening Charge: continuous variable that indicates how much was charged for using the service during evening time.
- Total Night Min: continuous variable that indicates number of minutes the customer used the service during nighttime.

- Total Night Calls: integer variable that indicates the number of calls during nighttime.
- Total Night Charge: continuous variable that indicates how much was charged for using the service during nighttime.
- Total Int Min: continuous variable that indicates the number of minutes the customer used the service to make international calls.
- Total Int Calls: integer variable that indicates the number of international calls.
- Total Int Charge: continuous variable that indicates how much was charged for international calls.
- No of Calls Customer Service: integer variable that indicates the number of calls to customer support service.
- Churn: binary class variable.