DWH & Data Mining

Semester Project: Fall 2018 Marks: 50

Submission: Tue, 11 Dec 2018 at 10:00pm

Predicting Customer Churn

Customer churn is the loss of customers. Many businesses use predictions of customer churn as a key business metric because the cost of acquiring new customers is much higher than the cost of retaining existing customers.

Dataset (Telco-Customer-Churn.csv) from a telecommunications company is provided to you for this project. This data includes demographic and account information on 7,043 customers.

Here is a brief description of the 17 variables used in the dataset:

- **customerID**: uniquely identifies a customer. Not needed in any model.
- **gender**: the gender of the customer (Female or Male).
- **SeniorCitizen**: 1 if the customer is a senior citizen, 0 otherwise
- **Partner**: whether the customer has a partner
- **Dependents**: whether the customer has dependents
- **tenure**: how long the customer has been with the company (in months)
- **PhoneService**: whether the customer has phone service
- **MultipleLines**: whether the customer has multiple lines
- **InternetService**: what type of internet service the customer has (DSL, Fiber optic, or No)
- **StreamingTV**: whether the customer has streaming TV service
- **StreamingMovies**: whether the customer has streaming movies service
- **Contract**: what type of contract the customer has (Month-to-month, One year, Two year)
- **PaperlessBilling**: whether the customer has set up paperless billing
- **PaymentMethod**: how the customer makes payments (Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check)
- **MonthlyCharges**: how much money the customer is charged a month
- **TotalCharges**: how much the customer has been charged over their tenure
- **Churn**: 1 if the customer has left the business, 0 otherwise. This is the class label.

Project Requirements

In this project, you are required to use three classification methods to predict customer churn. You are required to use “R” only for building the models and for getting answers to the questions mentioned below:

1. Exploratory Data Analysis

Answer the following questions:

1.1. What is the proportion of males/females in the dataset?
1.2. What is the proportion of senior citizens in the dataset?
1.3. What is the most common type of contract in the dataset?
1.4. What is the most common type of internet service in the dataset?
1.5. What is the least common payment method in the dataset?
1.6. What is the average tenure of male and female customers?
1.7. Is Streaming TV favored over the Streaming Movies service?
1.8. How many customers churned out of the total customers?
1.9. What is the mean “monthly charges” amongst customers with month-to-month contracts?
1.10. What is the gender wise average of “total charges” amongst customers?

2. Baseline Model

Ensure that there are no missing values in the dataset. If missing values exist, replace them with either the average or with the most frequently occurring values in the column.

Create a training and test split using the sample.split() function in the caTools library, with 70% of the observations in the training set and 30% in the testing set. Answer the following questions.

2.1. Our baseline model in classification is to always predict the most frequent outcome in the training set. What is the most frequent outcome?
2.2. What is the accuracy of this baseline model on the training set?
2.3. What is the accuracy of this baseline model on the test set?
2.4. What is the true positive (TP) rate of the baseline model on the test set?
2.5. What is the false positive (FP) rate of the baseline model on the test set?

3. CART Models

Create a CART classification model on the training dataset. Use the model to make predictions on the test set. Make the confusion matrix. Answer the following questions.

3.1. Plot the tree and identify which variables appear in the tree.
3.2. Write all the decision rules that you get from the tree.
3.3. What is the accuracy of your CART model on the test set?
3.4. What is the true positive rate of the CART model on the test set?
3.5. What is the false positive rate of the CART model on the test set?
3.6. What does the CART model predict for customer with a one-year contract and a tenure of 12?

4. Naïve Bayesian Classification

Create a Naïve Bayesian classification model on the training dataset. Use the model to make predictions on the test set. Make the confusion matrix. Answer the following questions.

4.1. What is the accuracy of your model on the test set?
4.2. What is the true positive rate of the model on the test set?
4.3. What is the false positive rate of the model on the test set?

5. Random Forest Models

Create a random forest classification model on the training dataset. Use the model to make predictions on the test set. Make the confusion matrix. Answer the following questions.

5.1. What is the accuracy of your model on the test set?
5.2. What is the true positive rate of the model on the test set?
5.3. What is the false positive rate of the model on the test set?
5.4. Make a summary table showing accuracy, TP and FP rates of the three algorithms. Compare the three algorithms and draw conclusions.

Extra Credit (optional): Build a logistic regression model on the training set and make predictions on the test set. What is the accuracy? Which variables are significant in the model?

Submission: Submit a word document using Google Classroom before the deadline. Write answers to all questions (with the question statement too). Write all R code (with comments) used to get the answers in the Appendix to your report. All answers are to be written in exactly 4 decimal places.

Poorly formatted documents not meeting project requirements and lacking necessary details will not get due credit. Start early.

Note: There will be a demo to grade the project. Late or copied projects will get no credit.