Project 2.1

ACADGILD

Project - Churn Prediction

Table of Contents

1. Introduction

Typical information that is available about customers' concerns demographics, behavioral data, and revenue information. At the time of renewing contracts, some customers do and some do not: they churn. It would be extremely useful to know in advance which customers are at risk of churning, as to prevent it – especially in the case of high revenue customers.

This is a prediction problem. Starting with a small training set, where we can see who has churned and who has not in the past, we want to predict which customer will churn (churn = 1) and which customer will not (churn = 0).

attr 1, attr 2, …, attr n => churn (0/1)

2. Objective

The project aims to solve Churn Prediction Problem

3. Prerequisites

N/A

4. Associated Data Files

https://drive.google.com/file/d/0Bxr27gVaXO5sQUlPTmZfMHdIRGM/view?usp=sharing

5. Problem Statement

The project involves you doing the below steps:

Pre-processing

Prepare each attribute in the right format and to train a machine learning model to predict churn as 0 or 1 depending on all other customer attributes.

Training

As usual, we are spoiled for choice when it comes to choosing a machine learning algorithm for training. For use cases, we will use Logistic Regression model. The model has to be optimized to suit the particular training data. Optimize means adjust the probability threshold for maximum precision and recall.

Whatever machine learning algorithm you choose, you always need to train it and evaluate it. For this reason, the Partitioning is required to partition most of the data (80%) for training and the small remaining amount (20%) for evaluation.

Evaluation

So, we trained a model. But what if the model has not learned anything useful? We need to evaluate it before running it for real on real data. For the evaluation, we use that 20% of data we have kept aside and not used in the training phase, to feed the trained model. This trained model is applied to all data rows one by one and produces the likelihood that that customer has of churning given his/her contract and operational data (P(Churn=0/1)). Depending on the value of such probability, a predicted class will be assigned to the data row (Prediction (Churn) =0/1).

The number of times that the predicted class coincides with the original churn class is the basis for any measure for the model quality as it is calculated by the Scorer node.

Notice that the customers with churn=0 are, hopefully, many more than the customers with Churn=1. If you want to take this fact into account and give more weight to the error made on the class Churn=1, then you can introduce an Equal Size Sampling node on the test set to under-sample the more numerous class Churn=0

1. Use R to do the preprocessing and training. Your final submission should have all the R code and results coming from Training Model. Show all the accuracy measures and ROC curve for different probability cut-off. Find the optimal Threshold.
2. Use R for giving all the results and statistics of the Trained Model performance on the Test Dataset. Your final report should contain all these results.
3. Use Tableau for Visualization Reporting.
    a. Perform the training and threshold optimization in tableau. Prepare required worksheet to perform the same. You should be able to change the probability threshold using a slider. When changing the threshold, you must show the accuracy measures accordingly.
    b. Once fixed on a particular threshold, you should use R_script to run the testing on the test data from Tableau.
    c. One you get the test result. Show the accuracy measure as well.
    d. How the odds ratio on the viz, i.e. show the effect of change of input variables to probability/chances of churn. You should provide option such that, we can choose any input variable and change the value to see the effect. Change on numerical variable would be offset change (some unit change), and categorical variable would be change from one category to other (Example: If binary then change from 0 to 1).

Your Final Deliverable:

1. R code
2. Final report with approach, results and observations.
3. Tableau that is as per mentioned in the requirement

6. Approximate Time to Complete Task