CHURN PREDICTION

Submitted by
KARAN RAJ

TABLE OF CONTENTS

OBJECTIVE

The project aims to solve Churn Prediction Problem

INTRODUCTION

Customer churn refers to when a customer (player, subscriber, user, etc.) ceases his or her relationship with a company. Online businesses typically treat a customer as churned once a particular amount of time has elapsed since the customer's last interaction with the site or service. The full cost of customer churn includes both lost revenue and the marketing costs involved with replacing those customers with new ones. Reducing customer churn is a key business goal of every online business.
Customer acquisition is much more expensive than retention. The primary focus of any small business will move from retention to conversion to traffic and back. Many larger enterprises will need to pursue a retention strategy at all times.

Churn prediction is the use of statistics and data about your customers to try to model who might leave for another service. Most businesses are concerned with this type of churn (voluntary), but some -- like banks -- are concerned with involuntary churn where a customer runs out of money.
Churn prediction can refer to a couple of different concepts in marketing analytics:

1. Techniques drawn from machine learning and predictive modeling to estimate likelihood those customers will churn;

2. Techniques drawn from time-series forecasting and regression analysis to project the future churn rate for a segment of customers.

The first concept of predicting the likelihood of customer to churn is often used to power marketing campaigns. In order to maximize return on investment, marketers are often interested in extending discounts or incentives ONLY to those customers that are at-risk of churning or unlikely to make a purchase. By estimating the likelihood of churn for each user, marketers can segment their customer base and target specific marketing communications to those segments that they deem eligible for a discount. In this way, marketers can promote a discount or other offer without indiscriminately spamming loyal users or incurring the costs associated with offering a discount to users who did not really need it in order to continue their relationship with the product, brand, and business.

The second concept of forecasting the churn rate for a segment or all of your customers is used more often in the context of business intelligence and analytics. Analysts are typically interested in knowing how tweaks to their business or product strategy — from pricing, to packaging, to adding or removing specific features — will impact their key performance indicators (KPIs) including revenue, profitability, and customer churn. By projecting a future churn rate using time-series forecasts, analysts can get a quick read on how product changes are impacting customer behavior and then act to either change what they're doing or keep things the same.

APPROACH

A customer data of 3333 customers with 21 variables was given. A part of this dataset was used to create a prediction model & then the model was fed by the other part of the dataset for prediction. This prediction calculation was completely done using R. First, the variables with most information values out of the 21, were found out. Now a sampled dataset called training data was created with 80% of the data. The rest 20% was chosen as the testing data. The prediction model was created on the training dataset and this model was fed with the testing data for churn prediction. The prediction result and the real results were compared for accuracy checkup. Suitable changes were incorporated for increasing the prediction accuracy of the model. Tableau was connected to R to create the prediction visualization.

❖ Firstly, after loading the churn.csv file, the **information value** of all the variables was found out using iv.mult function. **Day. Charge, Day.Mins & State** has the highest IV (screenshots attached).

❖ The data was split and into 80-20 ratio for training and testing respectively.

❖ ‘Glm_model’ command was used to make the model & the model was used to predict on the test data using ‘predict’.

❖ The predicted table had more no. of **False positives**. This value should be reduced, because false positives are more important than false negatives in churn analysis.

	0	1
0	537	28
1	**76**	26

❖ More false positives, more the people who were predicted to continue, is actually going to churn. This ultimately makes the analysis a failure. This problem is because of the imbalanced dataset which comes from initial dataset with more 0s than 1s.

0	1
613	54

❖ A **balanced dataset** was created using equal samples of 1s and 0s. It reduces the false positives in the prediction.

	0	1
0	423	142
1	**23**	79

❖ Accuracy was calculated using formula

Accuracy = (TP+TN) / (TP+FP+FN+TN)

❖ Sensitivity was calculated using the formula

Sensitivity = TP / TP+FN

❖ Specificity was calculated using the formula

Specificity = TN / TN+FP

Where, TP is true positive value
FP is false positive value
TN is true negative value
FN is false negative value

RESULTS AND OBSERVATIONS

- The final predicted outcome vs original outcome is as below

	0	1
0	423	142
1	**23**	79

- Accuracy measures were calculated using confusion matrix & using some formulas like
  - Accuracy =(TP+TN) / (TP+FP+FN+TN)= 0.752
  - Sensitivity = TP / TP+FN = 0.748
  - Specificity = TN / TN+FP = 0.774
- ROC Curve for different probability cutoff's were found using prediction & performance.
- Area under Curve (AUC) was found using measure ='auc'
  - AUC= 0.84
- Optimum threshold and maximum accuracy was found from accuracy vs cutoff plot as
  - Max. Accuracy = 0.776
  - Optimum cutoff = 0.468

- Tableau was connected to R using **Rserve** package to do the visualization.
- All supporting files of the project are as below (these files are also attached)
  - .R file 'Churn_prediction'
  - .RData file 'churn'
  - Word file with Rcode.
  - Images files of graphs of ROC & Accuracy-cutoff curve
  - Screenshots of R & Tableau outputs
  - Tableau Packaged Workbook (.twbx) 'Tableau-R'