Churn Prediction

**CHURN PREDICTION**

Project completed by

SUJEET KUMAR

Data Analytics

Churn Prediction

**TABLE OF CONTENTS**

1. Objective .........................................................................................................................page 3

2. Introduction .....................................................................................................................page 3

3. Approach..........................................................................................................................page 4 and 5

4. Challenges faced..............................................................................................................page 5

5. Results and Observations.................................................................................................page 6

Churn Prediction

**OBJECTIVE**

The project aims to solve Churn Prediction Problem

**INTRODUCTION**

Churn prediction can refer to a couple of different concepts in marketing analytics:

1. Techniques drawn from machine learning and predictive modeling to estimate

Likely hood that customers will churn;

2. Techniques drawn from time-series forecasting and regression analysis to project

the future churn rate for a segment of customers.

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 the ultimate aim of churn prediction.

Predicting the likely hood 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.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*pto\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Churn Prediction

**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

Churn Prediction

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

**CHALLENGES FACED**

The initial model had large number of false positives. This was because, the model was

biased as the training data had more number of 0’s than 1’s. More false positives means,

more the people who were predicted to continue, is actually going to churn. To avoid this a

balanced training dataset was created for modeling.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*pto\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**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.