**TELCO CUSTOMER CHURN PREDICTIVE MODELING**

**MAGGIE TANG**

**The College of Saint Rose**

**Spring 2020**

Group Team roles

* Data Cleansing; Data Visualization; Telco Customer Churn data
* Technical Lead; Data Capture; Analytics

INTRODUCTION

The primary project of this analysis was to predict customer churn for a telco company.

The methodology used in this project is Logistic Regression and Classification Tree. The process includes: Data Understanding, Data Preparation, Model Building, Testing and Evaluation, Business Understanding.

A close up of a device

Description automatically generated

Findings

We predicted that tenure, contract, InternetService, MonthlyCharge would influence the customer churn most, we also found classification tree analysis is stronger than logistic regression to predict this model.

BUSINESS UNDERSTANDING

In the telecommunication industry, customers are able to choose among multiple service providers and switch from one service provider to another. Telco companies often use customer churn analysis and customer churn rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. If the teleco companies know which customers are at high risk of churn, they are able to design customized treatment programs and reduce the churn rate. Therefore, finding factors that increase customer churn is important to take necessary actions for business decision making.

There’re different forms for customer churn, such as customers switching to a competitor, reducing the number of services they used, or switching to a lower cost service. This customer churn model enables you to predict which customers will churn and the reason what makes customers churn.

Our objective of this project is to predict the churn based on conditions such as contract, tenure, InternetService and MonthlyCharges. Other factors that may influence the churn are PaperlessBilling, PaymentMethod, MultipleLines, SeniorCitizen and OnlineSecurity.

DATA UNDERSTANDING

The dataset we used combined by 7043 records. Each record includes CustomerID, Churn, SeniorCitizen, Dependents, tenure, PhoneService, MultipleLines, InternetService, Contract, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBillings, PaymentMethod, MonthlyCharges, TotalCharges. The target variable is churn, the company’s current churn rate is 27%.

**Distribution for Churn**

A screenshot of a cell phone

Description automatically generated

DATA PREPARATION

Tools - SAS JMP

JMP was very helpful in processing our data so that we can put significant predictors into our model. JMP was used to format, clean, and filter non-significant variables. Many researches confirmed that predictive modeling technology is highly efficient to predict this situation. This technique is applied through learning from previous data to make a prediction for the future.

Since our goal is to predict customer churn, We first reset the customer value order, set 0 stand for not-churn, and 1 stand for churn.

We then checked there’re no typographical errors in category variables; no outliers in continuous variables; 11 missing value in TotalCharge , we excluded it from the dataset; CustomerID should not effect to the target variable, we dropped it; SeniorCitizen just has 2 values (0 and 1), we change its data type to nominal, 0 stand for not SeniorCitizen, 1 stand for SeniorCitizen; Gender and PhoneService both have a p-value higher than 0.05 with churn, we excluded it; TotalCharges has high correlation with tenure, we excluded it.

Data Summary

Before we progressed with modeling, it became clear that our data need to be cleaned more and more. To do this, we used step-wise personality to filter the data. There’re 9 significant variables appeared in our model: Chur, SeniorCitizen, tenure, InternetService, Contract, MultipleLines,

PaperlessBillings, PaymentMethod, OnlineSecurity, and MonthlyCharges.

**Stepwise fit for churn**

A close up of text on a white background

Description automatically generated

MODEL BUILDING (Predictive Model)

Since our response variable churn is categorical, both logistic regression and classification tree can work for it. We built two models use the reduced variables from step-wise personality at last step.

**Logistic Regression**

**Classification Tree**

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generated

Both models are highly significant with a Prob>ChiSq <.0001. Logistic regression model has a misclassification of 0.197, while we reached a misclassification of 0.1967 at 26 splits with classification tree model. So, it indicates in this case classification tree model works better than logistic regression model.

**Logistic Regression**

**Classification Tree**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

From the logistic regression Column Contributions and classification tree variable importance above, we can see the ranking for the predictors are very similar, the most important variables are tenure, contract, InternetService and MonthlyCharges, some other variables in the model are not that important.

EVALUATION

Findings and results

We initially predicted that tenure, contract, InternetService and MonthlyCharges would influence the customer churn, this model has a 80.33% accuracy.

We took a deep look into these 4 variables.

**Tenure** plays a very important role in the model, look at the tenure distribution we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for more than 70 months. This could be potentially because different customers have different contracts. Thus, based on the contract they are into it could be more or less easy for the customers to stay or leave the telecom company. In the Parameter Estimate, the value for tenure is negative, that indicates recent clients are more likely to churn.

**Logistic Fit of Churn By tenure**

**Tenure Distribution**

A screenshot of a cell phone

Description automatically generatedExhibit 5


**Contract** is the second main effort. There’re 3 values in contract: Month-to-Month (55%); One year (21%); Two year (24%). From the Mosaic Plot We found short term contracts have much higher churn rates.

From Odds Ratios for Contract, we also can see that odds of churn for customers with Month-to-Month contract was 3.86 times higher than those with Two-year contract, and 1.89 times higher than those with One-year contract.

The majority of this company’s customers have Month-to-Month contract, this might be the main reason why customer churn rate in this company is so high.

**Mosaic Plot for Contract**

**Odds Ratios for Contract**

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**MonthlyCharges** has a positive relationship with churn, that indicates there’s a higher churn when the MonthlyCharges is high.

**Logistic Fit of Churn By MonthlyCharges**

A close up of text on a white background

Description automatically generated

**InternetService** is another important predictor appears in the Top4 efforts in both models, it has 3 levels: No, indicates no internet service (22%); DSL, indicates slow internet service (34%); Fiber Optic, indicates fast internet service (44%).

From Odds Ratios for InternetService, we found odds of churn for customers with Fiber Optic was 3.16 times higher than those without internet, and 2.15 times higher with those who have DSL internet service.

We also found customers without internet have a very low churn rate. There’s something interesting that although Fiber optic services are faster, customers are more likely to churn because of it. It might because its price is high, or the company doesn’t have long term contract for this service.

**Mosaic Plot for InternetService**

**Mosaic Plot for InternetService**

A screenshot of a cell phone

Description automatically generated

A screenshot of text

Description automatically generated

There’re other factors like SeniorCitizen, OnlineSecurity, PaperlessBilling, PaymentMethod and MultipLines be observed that some variables have a positive relation to our target variable, and some have a negative relation.

SUGGESTIONS

The goal of this study was to build a predictive model for customer churn. We worked on both logistic regression and classification tree, these two modules gave us a very slight difference on results. We predicted that tenure, contract, InternetService, and MonthlyCharges would influence the customer churn most.

We made a classification tree model with a 80.33% accuracy. We predicted that new customers, customers with short term contract, customers with Fiber Optic internet service, and high MonthlyCharges have a high rate to churn.

In case to keep customers, the telco company should focus on developing long term contract, reducing monthly charges, set promotion for newly joined customers, and so on.