Paper 177-2019

Telecom Industry: Customer Churn Prediction

Aakash Dwivedi, Oklahoma State University; Miriam McGaugh, PhD, Oklahoma State University

ABSTRACT

Nowadays, the telecom industry faces fierce competition in satisfying its customers. With the advent of newer technology, the services offered by telecom companies have increased from being only calls to calls, data and web services. This means a constant struggle to strike a perfect balance among services and pricing of these services. In order to survive this market, telecom companies need to innovate, offer better services and increase their customer base. With newer companies entering the market and increasing freedom of customers to switch telecom companies, it's now becoming increasingly important to focus resources in retaining existing customers. According to an article in Harvard Business Review (Gallo, 2014), it was determined that the cost of acquiring a customer is five to twenty-five times more than retaining an existing one. Furthermore, by increasing retention by five percent can lead to an increase in profits by twenty-five to ninety-five percent.

This paper aims to segment customers and find the factors contributing to churn in each customer segment. Additionally, this paper also aims to build a churn prediction model and use that model to identify customers likely to churn. The customer churn rate measures the percentage of customers who end their relationship with a company during a particular period. For the analysis, SAS Enterprise Miner was used.

INTRODUCTION

Customer churn is one of the biggest fears of any industry. From various studies in the past, we know that the cost of acquiring a new customer has been far greater than retaining one. Churn or churn rate is defined as the percentage of customers who stop subscribing to a service or percentage of employees leave a job. Churn has affected industries such as banking, insurance, internet streaming and telecommunications to just name a few. Although there are many reasons for customer churn, some of the major reasons are service dissatisfaction, costly subscription, and better alternatives. Hence, in this paper the problem of churning is addressed and data factors affecting the churn are analyzed for their effect on the rate.

PROBLEM STATEMENT

Using the data provided, this paper aims to analyze the data to determine what variables are correlated with customer churn, if any. Additionally, a prediction model, to identify the people that might churn, will also be built. To build a prediction model, we will make different models using techniques such as logistic regression, decision tree, and neural network. These models will then be compared on the number of parameters obtained and the model optimized for final use. Furthermore, this paper also aims to calculate customer churn cost by identifying the total cost of customers who churned to date and how much money could be saved if we were able to improve our identification of customer churn. After the churn rate, we will also identify a subset of customers who will be offered retention plans.

DATA DESCRIPTION

The data was taken from Kaggle. It had 51,000 rows and 58 columns. Most columns related to subscriber personal information ranging from income to number of children. Other column was indicative of service usage by the subscriber. Based on the business understanding of the data 14 columns was chosen to analyze the data.

Sno.	Variables	Description
1	MonthInService	Months for which subscriber has been with company
2	CurrentEquipmentDays	Current Headset use in days
3	MonthyMinutes	Minutes the subscriber uses service
4	OutboundCalls	Number of outbound calls by subscriber
5	RecievedCalls	Number of received calls by subscriber
6	AverageMinutes	Average minute a call last
7	CostumerCareCalls	Number Of customer care calls
8	BlockedCalls	Number Of customer care calls
9	CreditRatings	Credit Rating of customer(4 Categories)
10	RoamingCall	Number of Roaming calls
11	DroppedCall	Number Of dropped calls
12	Occupation	Occupation of subscriber(8 Categories)
13	PrizmCode	Residential Region of Subscriber(7categories)
14	Recurring Charge	Total bill for each month

Table 1. Data Dictionary

METHODOLOGY

For the analysis and modeling, the SEMMA(Sample, Explore, Modify, Model, Access) methodology was followed as shown in Figure 1.

In order to identify the clusters, present in the data, we first sampled the data with equal proportion and did some data preparation in order to impute missing data. Furthermore, range standardization was chosen to identify clusters better, with every variable having equal weight in cluster formation.

In the second portion, building a predictive churn model, the data was divided into training and validation datasets with 70/30 split. The training data was used to train various models and the validation dataset was used to assess the model performance. For assessment of models, misclassification rate was used because the target was a categorical variable.

The tool used for the analysis was SAS Enterprise Guide.

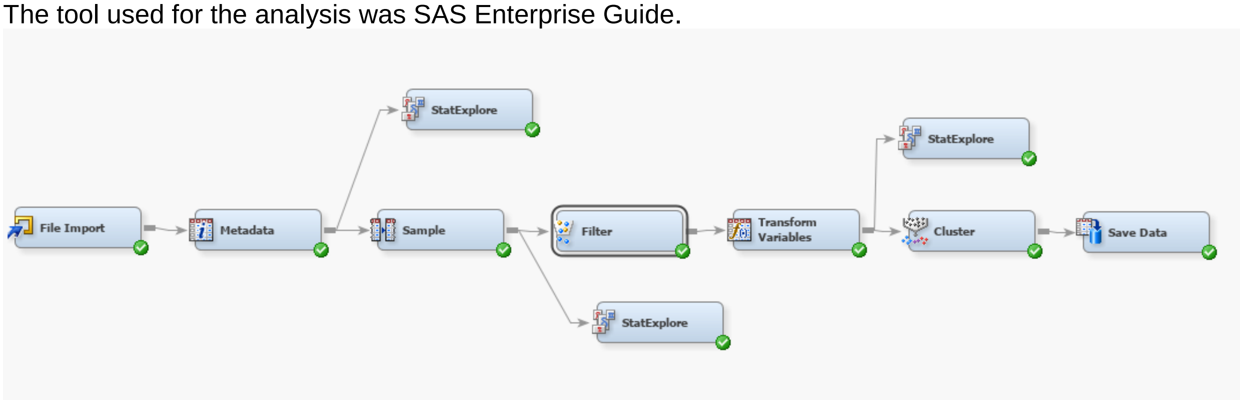

Figure 1: SEMMA Model

APPROACH

Since this was churn prediction model, the first step was to analyze the clusters in the data. The second step was to run different models and identify best predicting models for each cluster. For identifying different clusters, we took the following steps:

1. First was the variable selection, in the course of this analysis, variables of business importance and relevance were chosen.
2. Our data had more cases of churn then not churn, so stratified sampling with equal proportion was used to remedy that.

3. Then the data was filtered for outliers and the data was transformed in order to bring each variable to the same level so that they had equal impact on the clusters.
4. Then cluster analysis was done, and three different clusters were identified from the CCC plot.

RESULTS

From the cluster analysis, we found two clusters (figure 2) relevant to the business problem, which were distinguished by variables such as MonthlyMinutes, MonthlyRevenue, Credit Score, Occupation and Prizm code. Below are tables to demonstrate the difference between clusters.

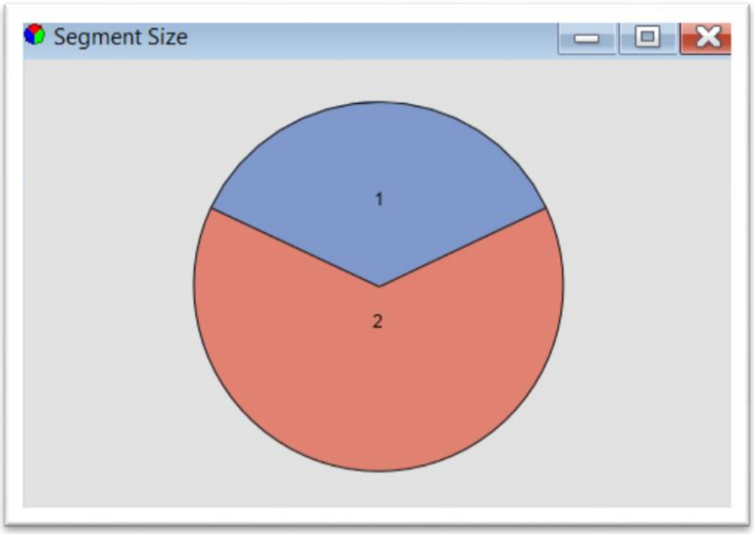

Figure 2: Cluster Size

Variables	Cluster 1	Cluster 2
Average Monthly Revenue	$47.48	$119.55
Average Monthly Minutes	350.88	1408.56
Average Current Equipment Days	419.32	242.40
Average Roaming Calls	0.91	2.66
Average Months in Service	19.03	17.34
Prizm Code		
Rural	173	284
Suburban	1358	22
Town	615	179
Other	1657	55
Credit Rating		
1 – Highest	2492	247
2 – High	86	16
3 - Good	660	130
4 – Medium	258	58
5 – Low	160	51
6 – Very Low	116	30
7 - Lowest	31	8

Table 2: Summary Information by Cluster

The second step of analysis was to apply different predictive modeling techniques to both clusters and see which model was better each cluster using the misclassification rate as metric for model assessment.

For both the clusters we used three predictive modeling algorithms decision tree, logistic regression and artificial neural network. Since the data had high value of skewness and kurtosis, various transformations and log transformations were chosen as the best options.

MODEL ASSESSMENT

Baseline Statistics

The statistical models were run on the data without any transformation, imputations or segmentation. The best model came out as decision tree with misclassification rate of 41.69%. The neural network model had the same rate, but other diagnostics were not as good as decision tree. The regression model has the highest misclassification rate of all at 41.67%.

Selected Model	Model Node	Model Description	Train: Misclassification Rate	Average Squared Error	Average Squared Error	Valid: Misclassification Rate
Y	Tree2	Decision Tree	0.39168	0.23461	0.23926	0.41678
	Neural2	Neural Network	0.39644	0.23303	0.24084	0.41678
	Reg2	Regression	0.41876	0.24044	0.24394	0.43528

Figure 4 : Baseline Statistics

Segmented Model Statistics

After that, models were run on both clusters to assess if the misclassification rate was reduced for the different techniques. All models had improved misclassification rates after imputation and transformation.

For cluster one, based on misclassification rate, the decision tree came out as the best model. Current Equipment Days was the most important variable in the decision tree model followed by Monthly Minutes used, Months in Service and Retention Calls.

Fit Statistics
Model Selection based on Train: Misclassification Rate (_MISC_)

Selected Model	Model Node	Model Description	Train: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Tree	Decision Tree	0.28129	0.19557	0.19581	0.28103
	Neural	Neural Network	0.28652	0.19640	0.19701	0.28501
	Reg	Regression	0.28812	0.19896	0.19817	0.28528

Figure 5 : Comparison statistics for Cluster One

ariable Name	Label	Number of Splitting Rules	Importance	Validation Importance
OG_CurrentEquipmentDays_1	Transformed CurrentEquipmentDays_1	1	1.0000	1.0000
OG_MonthlyMinutes_1	Transformed MonthlyMinutes_1	2	0.4253	0.5164
OG_MonthsInService	Transformed MonthsInService	1	0.4243	0.4122
OG_RetentionCalls	Transformed RetentionCalls	1	0.1766	0.0909
OG_AgeHH1_1	Transformed AgeHH1_1	1	0.1474	0.1448

Figure 6 : Variable Importance for Cluster One

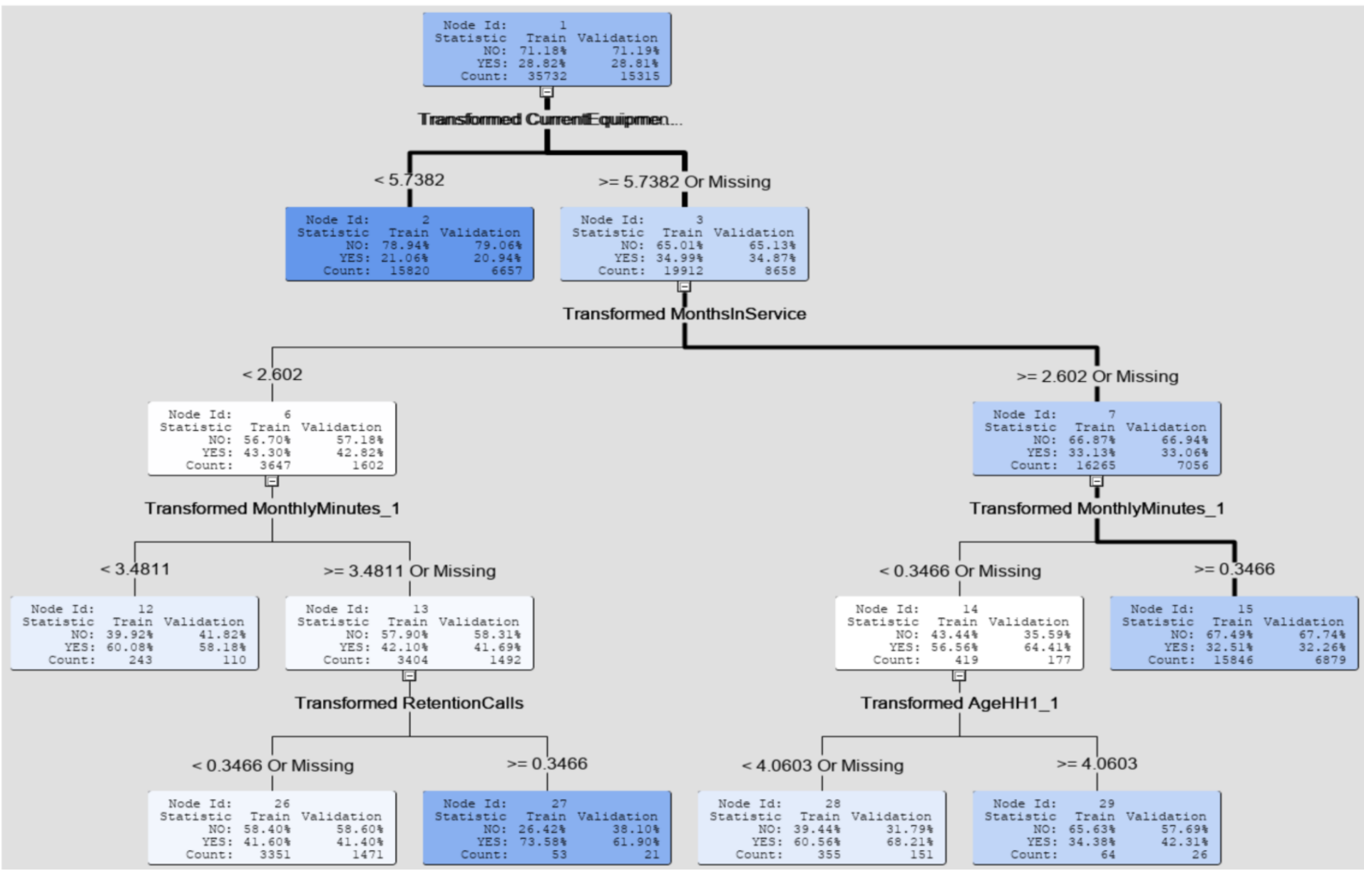

Figure 7: Decision Tree for Churn Rate of Cluster One

For cluster two, the neural network was the best model according to the misclassification rate of 0.27997.

'it Statistics
Iodel Selection based on Train: Misclassification Rate (_MISC_)

Selected Model	Model Node	Model Description	Train: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural2	Neural Network	0.27955	0.19226	0.19286	0.27997
	Reg2	Regression	0.28282	0.19701	0.19629	0.28102
	Tree2	Decision Tree	0.28319	0.20299	0.20299	0.28319

Figure 8 : Comparison statistics for Cluster Two

In order to better explain author has used decision tree to neural network. From the decision tree, variable importance was taken, which is displayed below

'ariable Importance

'ariable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
OG_CurrentEquipmentDays_1	Transformed: CurrentEquipmentDays_1	1	1.0000	1.0000	1.0000
13	Hidden: H1=3	4	0.8969	0.7897	0.8805
11	Hidden: H1=1	1	0.4776	0.3554	0.7441
12	Hidden: H1=2	3	0.4445	0.3728	0.8387
OG_MonthsInService	Transformed: MonthsInService	1	0.3138	0.3957	1.2613
OG_TotalRecurringCharge_1	Transformed: TotalRecurringCharge_1	1	0.2126	0.2590	1.2182
OG_MonthlyMinutes_1	Transformed: MonthlyMinutes_1	1	0.1987	0.0000	0.0000
OG_AgeHH1_1	Transformed: AgeHH1_1	1	0.1350	0.0484	0.3587
OG_DroppedBlockedCalls	Transformed: DroppedBlockedCalls	1	0.1205	0.1336	1.1086

Figure 6 : Variable Importance for Cluster Two

From the variable importance table, it can be seen that the most important variable in this model were LOG_CurrentEquipmentDays_1, H13, H11, H12 in the same order.

CONCLUSION

There were two main objectives of the analysis. First, to segment the customers based on an unsupervised learning method. Second, was to predict churn in the respective segment and assess if segmentation helped improve our prediction, which was demonstrated by a considerable improvement on our baseline. The metric for model performances was misclassification rate, which decreased significantly from our baseline of 41% to approximately 28% for both the segments.

Companies can use this process to segment and benchmark processes to determine who is at risk for leaving or discontinuing services. Benchmarking improvement in your models is important to ensure that companies are meeting the needs of all their clients uniquely.

REFERENCES

Lu, Junxiang, Detecting Churn Triggers for Early Life Customers in the Telecommunications Industry – An Applications of Interactive Tree Training, Proceedings of the 2nd Data Mining Conference of DiaMondSUG 2001, Chicago, IL, 2001.

Krutharth Peravalli, Dr. Dmitriy Khots, Churn Prevention in Telecom Services Industry- A systematic approach to prevent B2B churn using SAS, 2017

ACKNOWLEDGMENTS

The author would like to thank Amritha for her continued support and guidance throughout this paper. The author would also like to thank Dr Goutam Chakraborty and Dr. Miriam McGaugh for their support and guidance throughout the course of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Aakash Dwivedi, Oklahoma State University

Email : aakash.dwivedi@okstate.edu

Dr. Miriam McGaugh, Oklahoma State University

Email : miriam.mcgaugh@okstate.edu