Paper 3600-2018

Enhancing Subscription Based Business by Predicting Churn Likelihood

Analytics Androids: Varsha Reddy Akkaloori, Sujal Reddy Alugubelli, Smitha Etlapur, Mounika Kondamudi
Oklahoma State University

ABSTRACT

Customer retention is a challenge faced by most businesses in today's competitive market. Predicting customer churn would help a subscription business such as KKBox in creating substantial difference in their revenue stream. This paper describes work relating to predicting churn likelihood using SAS® 9.4, SAS® Enterprise Miner for data cleaning, preparation and analysis.

INTRODUCTION

Miner KKBox provides streaming services to millions of users with over 30 million sound tracks. It provides both free and premium (paid) streaming services on various devices including wearables. The paid subscribers have the added benefit of being able to play the music offline too.

The work described by this paper includes segmentation of the paid subscribers into meaningful categories based on both transactional and listening behaviors. Insights from segmentation would help in formulating customized strategies to enhance customer retention, loyalty and profitability. The work also includes a summary of predictive model built to identify customer churn for KKBox music subscription service.

The predictive model built offers insights into potential patterns between churners and non-churners based on recent usage, usage rate, number of unique songs heard and whether they opted for auto renewal features etc. As acquisition of new customers in any business is usually expensive, subscription service like KKBox would benefit financially from investing in retention of its existing customers.

DATA

The data used for this study is obtained from "WSDM - KKBox's Churn Prediction Challenge" launched by Kaggle. The data was released by the KKBox team in two versions and both versions are considered for the study. The data includes a transaction table and a user logs table.

The transaction table has the transactional history of customers until 2/28/2017.The table includes the subscription periods, subscription plan, cancellation flag and method of payment. The User logs table contains the service usage information of the customers throughout each of their transactions, which includes, total number of unique songs listened to, total number of songs listened to for various proportions say 25%, 50%, etc. and the total duration spent listening. The types of tables along with their variables have been included as Appendix 1.

PROBLEM

The aim of this analysis is to predict premium customers' churn likelihood based on their behavior until 01/31/2017 and to understand their behavior by classifying them into various groups based upon their behavior using their transaction and listening history. The business rule of KKBox states that, a customer is said to be churned, if they do not renew or make a new service subscription transaction within 30 days from their existing plan expiration date.

DATA CLEANING AND VALIDATION

The purpose of this study was to segment the premium customers into meaningful segments and to understand the churn propensity based on their behavior and listening trends. The training and validation data has been restricted to transactions of members until January 31st, 2017. Data Cleaning/validation was performed in the two phases as described below.

TRANSACTION DATA PREPARATION

First, the duplicate transactions in the data were deleted. A Churn flag indicator was created by calculating the number of days the customer took to renew or make a new subscription. If the calculated days were greater than 30, then the user was considered to have churned. To capture reasonable amount of user's behavioral information, only the transactions that had subscription plans more than 3 weeks were considered in the analysis. Churn flag for those customers whose activity extends beyond 01/31/2017, were computed based on their activity in February. Based on KKBox's churn definition, a user can come back after they churned (after 30 days of previous expiration). When the user comes back after they churned they can still retain the same member id. The process of aggregating the transactions per user resulted in more number of churners than non-churners. In order to overcome the problem, each instance of a single user's transaction was considered as an individual record for modelling churn. However, to retain user history, flag "previous percentage of churn" was created to capture the number of times the customer previously churned or not churned as a variable. The user logs for these individual transactions were also aggregated at the transaction level, treating each transaction as a separate subscriber.

USER LOGS DATA PREPARATION

Both versions of user log tables were imported and appended. Any duplicate log entries which were present in the data were removed. User logs for the users whose transaction details were not present were removed. Since every transaction was considered as a separate user, the user logs were aggregated at user and transaction level.

ANALYSIS

Data exploration and analysis was performed to study the distribution of various variables present in the data. Appropriate data preparations were performed to aid in understanding the customer behavior based on transactions and usage.

The analysis was performed in two stages. The first phase deals with bucketing the customers into segments. The second phase deals with predicting the churn likelihood of premium customers using predictive modelling techniques. Capturing behavioral trends by using the usage metrics enables a company to better understand the churn likelihood of their customers. For this purpose, various flags relating to users listening trends were created.

Understanding customers' dissatisfaction based on usage of the service is significant in improving customer experience. In this study, a flag, "consecutive decrease" was created to indicate decrease in usage during the last two consecutive weeks. An ideal recommender system engages its users, by suggesting songs based upon users' interest. On contrary, when the users are not impressed by the recommendations or collection of songs, they might choose to listen to the same songs repeatedly. To capture the impact of this behavior on churn, another flag, "repeat mode" has been created to capture this trend. In general, consistent usage of service is a good indicator of customer satisfaction. A metric, "usage rate" was computed to capture the usage rate. Also, how recent a customer is using the service can be indicative of customers' latest impression of the service, hence a flag called "recency" has been created based on the customers' latest user log and his transaction expiration date. When a customer is disinterested with a song, there is a high probability he would not listen to it completely. So, the customers with a listening history of less than 50% of a given song predominantly, could be an indication that they do not like the songs or the recommendation. A flag, "low proportion" was created to capture if most of the songs listened to were less than 50% of the song length.

PHASE 1: SEGMENTATION ANALYSIS

Customer's listening behavior usage patterns could be highly valuable to a music streaming company. Having an understanding of such patterns and trends would help in formulating customized strategies to enhance customer retention, loyalty and profitability. Segmentation analysis is performed to cluster customers into groups where the behavior of customers within the group have similar usage patterns and

different across the groups. In this project K-means segmentation is performed in order to create the customer groups. By default, 8 segments are formed by the SAS node. As few segments had very similar characteristics, segments are regrouped and five segments are formed based on CCC criterion. "Cluster" in SAS Enterprise Miner is used for this purpose. Range transformation is performed internally on the variables and the rest of the properties are defaults. Variables, payment method, last week decrease, usage rate, plan list price, recency, total sec. used, total amount paid are the base variables used in order to form the segments. Segmentation results are provided in Appendix 2. The following illustrates the segments formed and their respective marketing strategies.

Dependable Casual Listeners Segment

- Users in this segment opt for subscription plans with low amount and majority of them auto-renew their subscription.
- This segment predominately consists of users who have consecutive decrease in their usage in the last two weeks of their subscription period and very low overall usage rate.
- Customers in this segment could be dependable because their subscription auto renewal rate is high. But since they have a lesser usage and also pay less, the company needs to encourage these users to increase their usage rate by promoting new subscription plans tailored to their interests.

Unprofitable Disinterested Listeners Segment

- Most users of this group are not listening to entire songs. Among the songs they are listening to, they repeatedly listen to same songs.
- Opting to the plans with low amounts, the overall usage for this segment is also less. They tend to have a reduced usage in the last two weeks of their subscription period.
- As the revenue brought by these users is less, in case of higher acquisition costs, they might not be considered for the new promotional offers, which incur more expense to the company.

Profitable Diversified Listeners Segment

- High revenue generating users who consistently listen to different songs.
- Along with being diverse in their respective listening activity, they also listen to most of the songs completely.
- As the customers generate high revenue, new promotional offers which bring higher profits to the company can be made to this segment.

Potential High Usage Listeners Segment

- They have very high usage rate and only few customers reduced their usage in the last two weeks. They also prefer listening to unique songs.
- These customers did not re-subscribe within 30 days on many instances and they subscribe to lower amount plans.
- Despite having high usage rate, they have less interest to subscribe back indicates bad experiences. The company needs to improve the services provided to these customers and send out promotions to them to re-subscribe.

Persuadable Listeners Segment

- These customers re-subscribed within 30 days on many instances in the past but they have a high churn rate. Majority of these customers do not auto renew their subscriptions.
- The average of actual amount paid is the highest. Also, they have the highest usage rate.
- Since these customers have a good usage rate and highest amount paid, KKBox should consider cross selling strategies to retain them.

PHASE 2: PREDICTIVE MODEL

SAS SEMMA methodology was applied to build the predictive model. As the data set was rather large, a sample data set was obtained through stratified sampling; 25% of the overall population is collected. Proper

input and target roles have been assigned to the data, and necessary input variable transformations to get symmetrical distributions have been performed (Details on variable transformations are provided in Appendix 3). The sample data is then portioned into training and validation datasets so as to train the model and validate.

Three types of classification models were built on the data. A model comparison node was then used to select the best model based on the misclassification rate for validation dataset. All the models yielded very similar results and performed equally well. Logistic regression was selected as the final model for its simplicity and interpretability.

Predictive Model Insights

The misclassification rate of the predictive model is 30%. The raw predictor variables that were used in the model are actual amount paid, total seconds listened, and number of songs listened for 25%, 50%, 75% and 100%. Some behavioral variables like recency, re-subscribe within 30 days, usage rate, decrease in usage during last two consecutive weeks, lower proportion of the songs listened got selected into the model.

From the model, Table 1 shows the variables affecting customer's churn:

Positive Impact	Negative Impact
Low to Moderate Recency.	Listening to Entire Songs.
Number of times they have churned in the past.	Total Amount Paid.
Decrease in usage in last weeks and repeat flag	Total Usage of their subscription plan.

Table 1: Variables affecting customer's churn

The degree to which predicted probabilities agree with actual outcomes is expressed in a confusion matrix as seen in Appendix 3. Observations with Churn Indicator = 1 are considered as events and Churn Indicator = 0 is considered as Non-events.

Sensitivity measures the proportion of correctly classified events, and it is equal to 86.5%. As the model has a very good sensitivity, KKBox can identify the customers who are going to churn very well. Taking appropriate actions based on the aforementioned results, improves customer retention and profitability. Model results are provided in Appendix 3.

GENERALIZATION

Since there exists an acquisition cost associated with every new customer, a company's revenue largely depends on retaining its customer base. In a business, if a customer churns before their expected customer life time value, cost involved in acquiring them is a loss to the business. In this world of Big Data, a business can collect and store massive amounts of transactions and behavioral data of its customers. Using this data, a company can understand users' interaction with the subscription along with identifying user preferences, likes and dislikes. Inclusion of demographic variables like gender, location, age, can also help better understand the behavioral patterns.

SUGGESTIONS FOR FUTURE STUDIES

In general, success of a music industry depends on the recommendation system. Suggesting accurate songs based on user's playlist and also from other user's playlists that have similar songs; giving differential weights to types of songs heard completely and songs skipped which were heard for 25% of the duration. The company could use user's location and the device they are listening on, in the model to analyze their listening behaviors based on devices and locations. Demographic details of a customer like age, gender etc.; can be used in order to segment them and provide personalized recommendations. Also, results from previous marketing campaigns can be used in future modelling to improve prediction rates.

CONCLUSION

"It takes months to find a customer and only seconds to lose one". A company would not take action to retain profitable customers, if it is ignorant of the fact that they would churn in the near future. In addition, if

it sends out attractive promotional offers to unprofitable customers, the company would incur huge losses. Therefore, understanding customer behavior from the segmentation analysis and predicting churn of a profitable customer in advance, KKBox would succeed in retaining them by being able to implement the most impactful marketing strategies. Thereby, a huge ratio of churn could be avoided.

REFERENCES

Guilherme Dinis Chaliane Junior. 2017, “Churn Analysis in a Music Streaming Service Predicting and understanding retention,” https://kth.diva-portal.org/smash/get/diva2:1149077/FULLTEXT01.pdf.

Kaggle. 2017, “Terms of Use,” https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data.

Titlow, John Paul. 2016, ”7 ways streaming music will change in 2017, after another crazy year,” https://www.fastcompany.com/3066532/7-ways-streaming-music-will-change-in-2017-after-another-crazy-year.

Aurélie Lemmens, Sunil Gupta. 2013, “Managing Churn to Maximize Profits,” http://www.hbs.edu/faculty/Publication%20Files/14-020_3553a2f4-8c7b-44e6-9711-f75dd56f624e.pdf.

https://www.pinterest.com/pin/471118810993978695.

CONTACT INFORMATION

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

Varsha Reddy Akkaloori
varsha.akkaloori@okstate.edu

Sujal Reddy Alugubelli
sujal_reddy.alugubelli@okstate.edu

Smitha Etlapur
smitha.etlapur@okstate.edu

Mounika Kondamudi
mounika.kondamudi@okstate.edu

APPENDIX 1
(Available Tables & Corresponding Variables)

1. a. Transactions.csv

Contains transactions of users up until 2/28/2017

- msno: user id
- payment_method_id: payment method
- payment_plan_days: length of membership plan in days
- plan_list_price: in New Taiwan Dollar (NTD)
- actual_amount_paid: in New Taiwan Dollar (NTD)
- is_auto_renew
- transaction_date
- membership_expire_date
- is_cancel: whether or not the user canceled the membership in this transaction.

1.b. Transactions_v2.csv

Same format as transactions.csv, contains the transactions data until 3/31/2017.

2.a. User_logs.csv

Contains daily user logs describing listening behaviors of a user until 2/28/2017.

- msno: user id
- date: format %Y%m%d
- num_25: # of songs played less than 25% of the song length
- num_50: # of songs played between 25% to 50% of the song length
- num_75: # of songs played between 50% to 75% of of the song length
- num_985: # of songs played between 75% to 98.5% of the song length
- num_100: # of songs played over 98.5% of the song length
- num_unq: # of unique songs played
- total_secs: total seconds played

2.b. User_logs_v2.csv

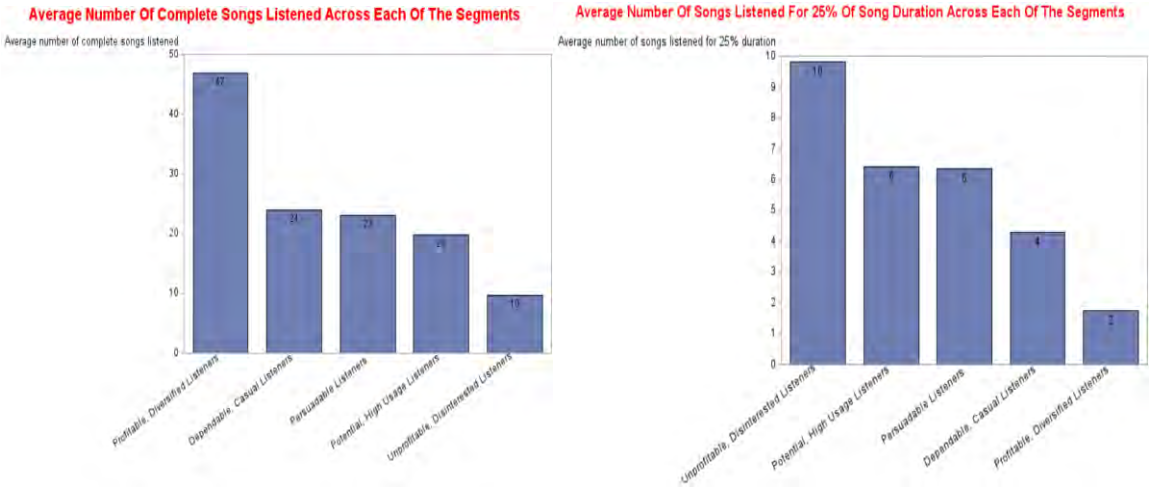
Same format as user_logs.csv, contains the user logs data until 3/31/2017.

APPENDIX 2

SEGMENTATION ANALYSIS:

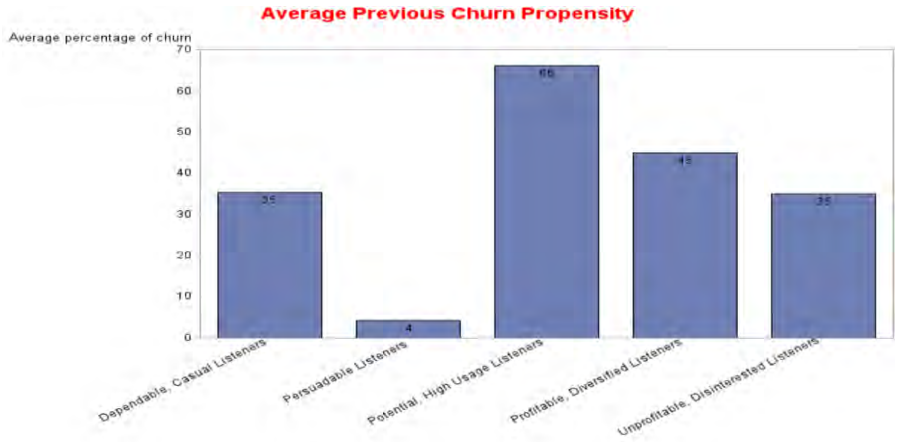
- The average amount paid by the customers is highest in Persuadable listeners group and least in Dependable, casual listeners group.

- Profitable diversified listeners group listened to more number of complete songs than any other group. Unprofitable disinterested listeners mostly listened to only 25% of the song's duration.

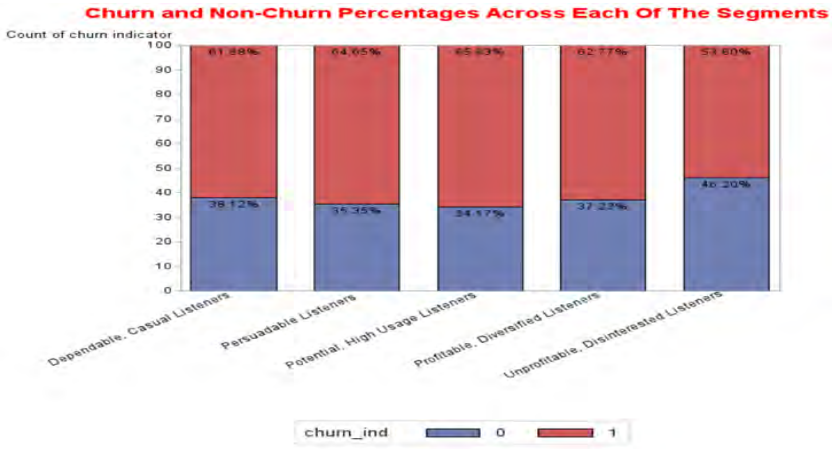


- On an average Profitable diversified listeners group listened to highest number of unique songs in a day. Unprofitable disinterested listeners listened to least number of songs.

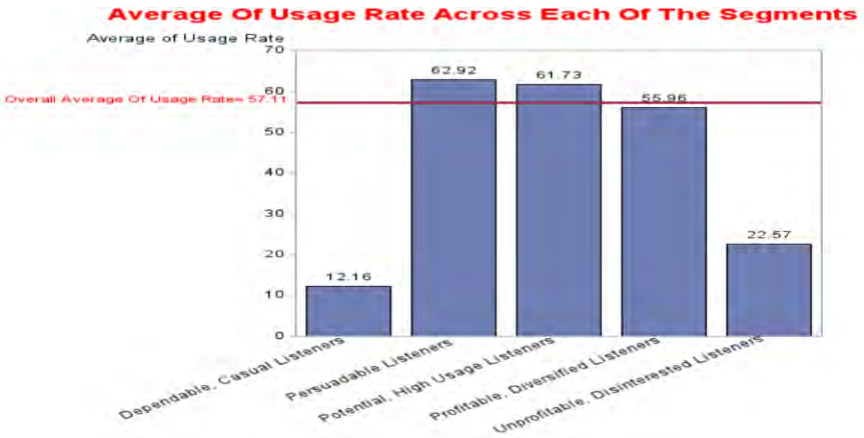
- Majority of the potential high usage customers did not re-subscribe back within 30 days of current expiration (They have the highest churn percentage). Most of the persuadable listeners' groups re-subscribed within 30 days (They have the least churn percentage).



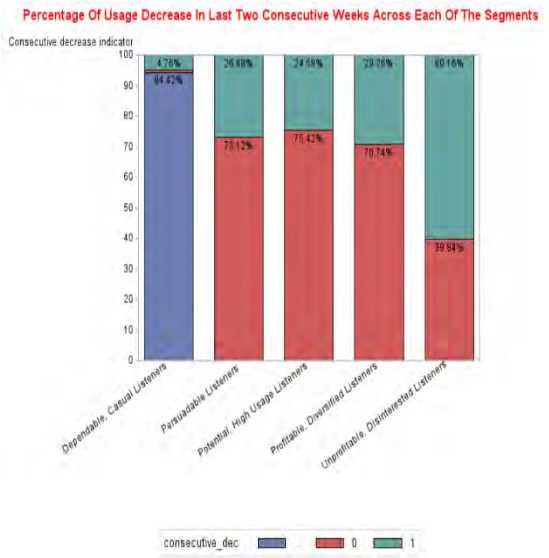
- Interestingly, potential, high usage listeners have very high churn proportion. Unprofitable listeners have least churn proportion.

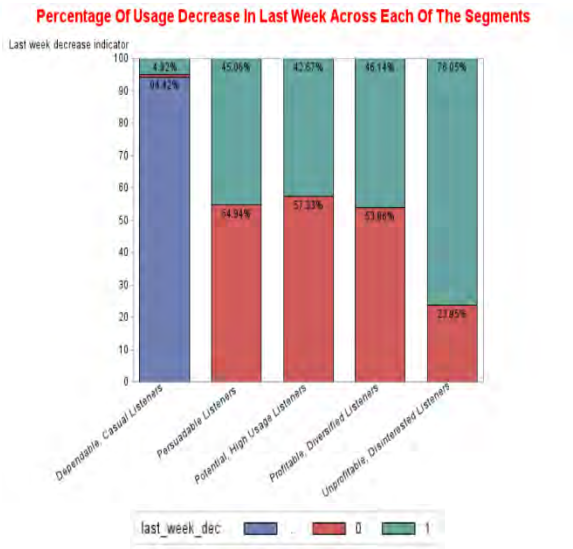


- Persuadable listeners have very high usage rate whereas dependable casual listeners have very low usage rate.

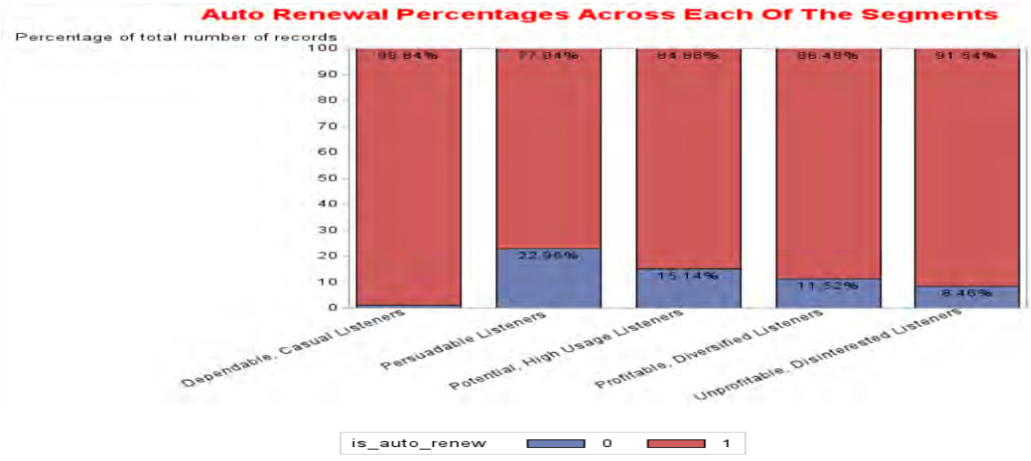


- Majority of Unprofitable disinterested customers have decreased usage in the last two weeks. Dependable, casual listener customers have very low proportion of people whose usage decreased in last week.





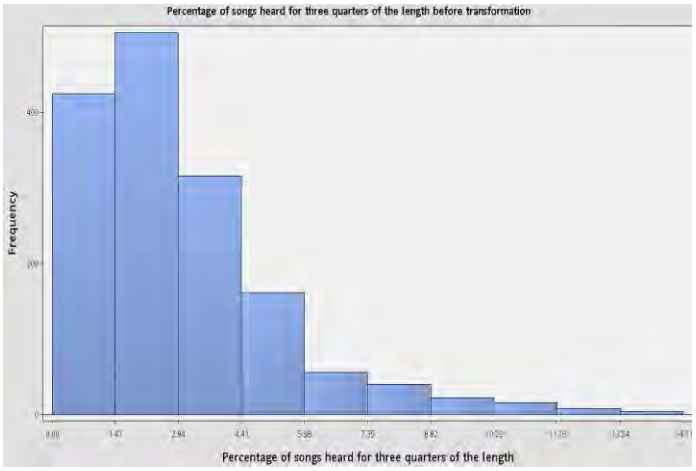
- Almost all of the dependable causal listeners' auto renew their subscription. Persuadable listeners comprise the least auto renewing customers.

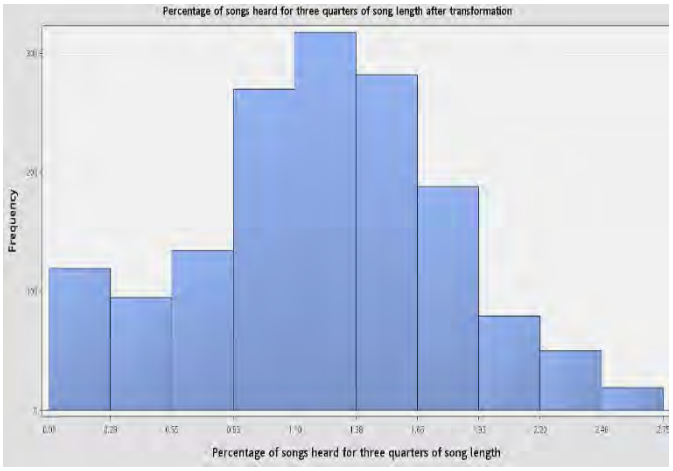


APPENDIX 3

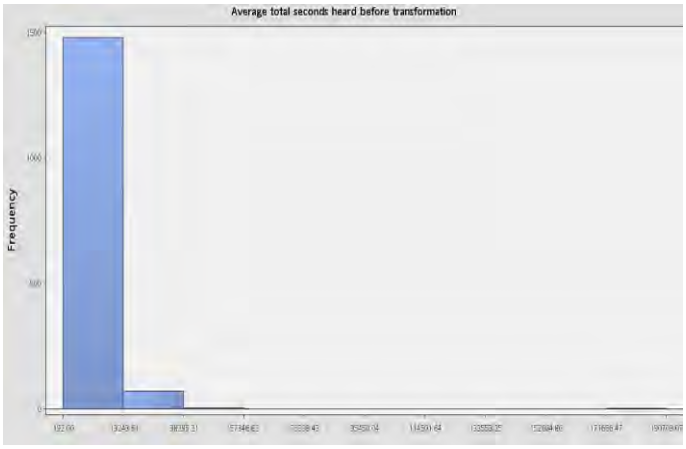
VARIABLE TRANSFORMATIONS:

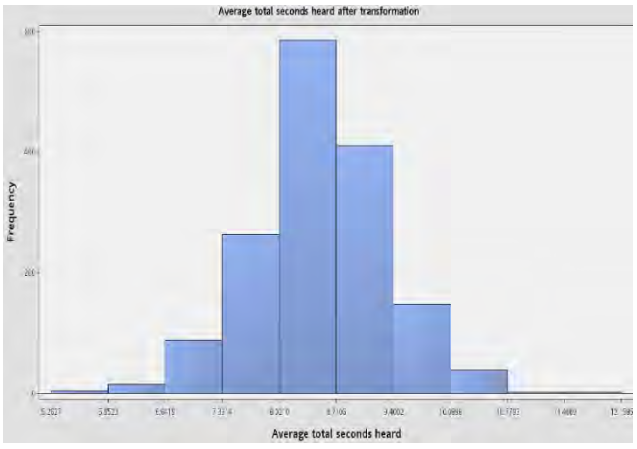
- Log transformation is performed on all the variables that are skewed right. Since the variables indicating percentage of songs heard for 75% of its total length, average total seconds heard, percentage of songs heard for 50% of its total length, percentage of songs heard for 25% of its total length, percentage of songs heard for 98.5% of its total length are skewed to right, they are Log transformed.



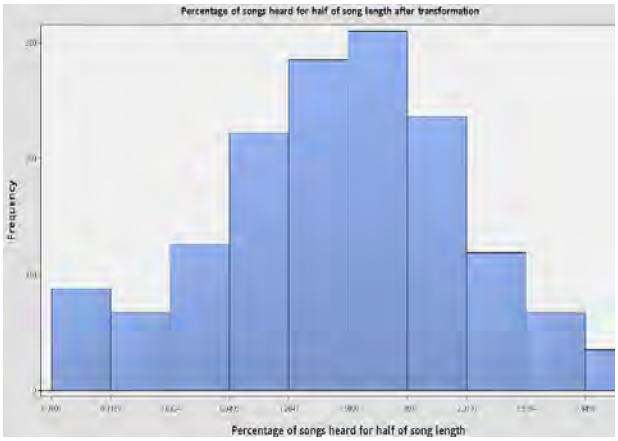


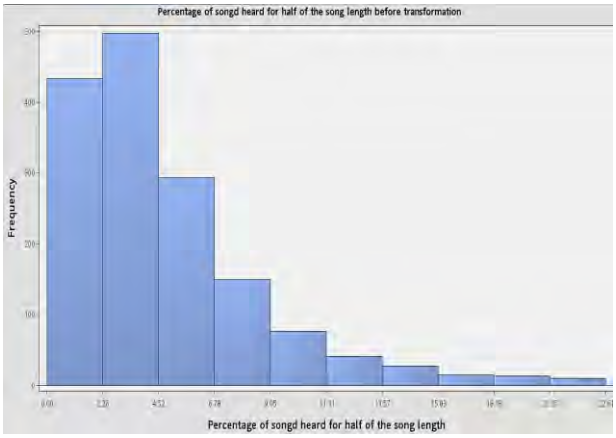
Percentage of songs heard for 75% of its total length: Before and after log transformation



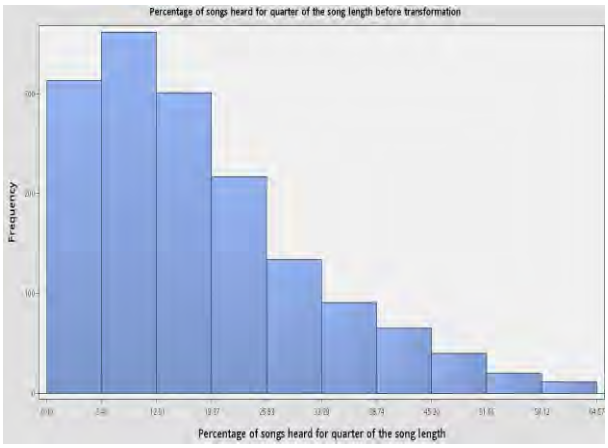


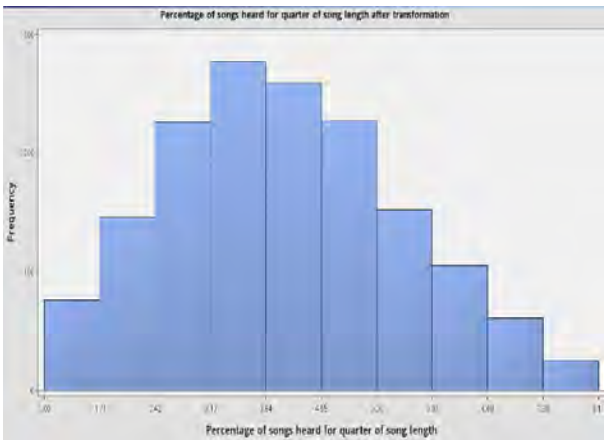
Average Total seconds heard: Before and after log transformation



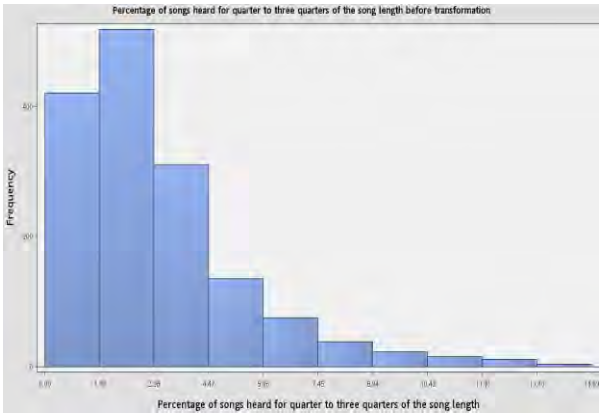


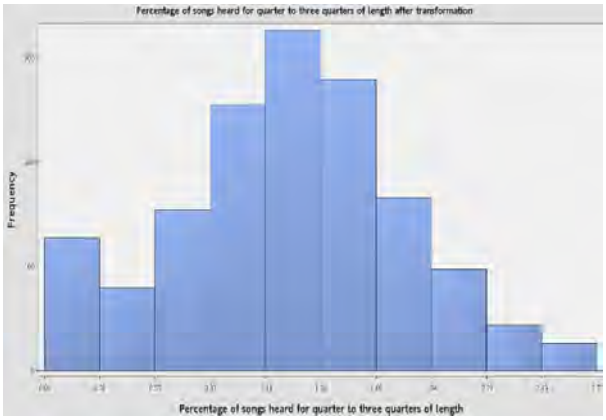
Percentage of songs heard for 50% of its total length: Before and after log transformation





Percentage of songs heard for 25% of its total length: Before and after log transformation

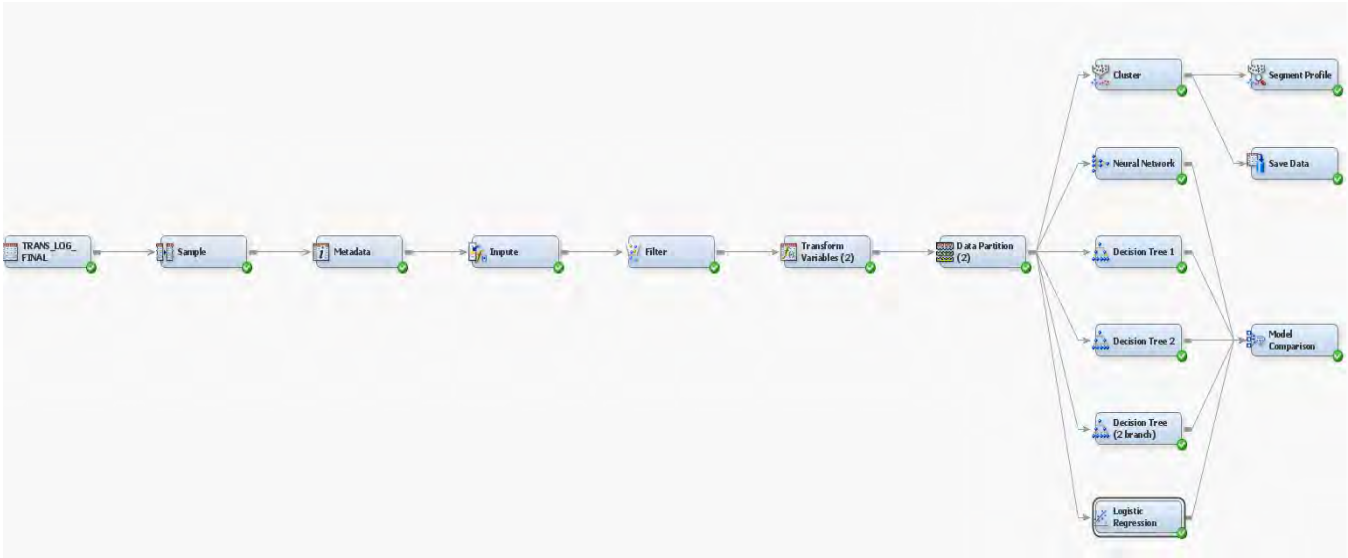




Percentage of songs heard for 98.5% of its total length: Before and after log transformation

PREDICTIVE MODEL FOR CHURN PREDICTION:

The following is the process flow diagram that is used for building the predictive model



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Valid: Roc Index ▼	Valid: Kolmogorov-Smirnov Statistic	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Root Average Squared Error	Valid: Average Squared Error
Y	Neural	Neural	Neural Network (3)	churn_ind		0.308194	0.705	0.284	0.9362...	15938...	0.4488...	0.2014...
	Reg3	Reg3	Regression (3)	churn_ind		0.309417	0.704	0.287	0.9734...	16060...	0.4505...	0.2030...
	Tree3	Tree3	Decision Tree (4)	churn_ind		0.311119	0.698	0.288	0.9236...	15963...	0.4492...	0.2018...
	Tree4	Tree4	Decision Tree (6)	churn_ind		0.311119	0.698	0.288	0.9236...	15963...	0.4492...	0.2018...
	Tree5	Tree5	Decision Tree (2 branch)	churn_ind		0.311119	0.698	0.288	0.9236...	15963...	0.4492...	0.2018...

Table 2: Result of model comparison node in Enterprise Miner

The confusion matrix is as shown below. It has the details of true positives true negatives, false positives and false negatives. These are used to calculate the accuracy, sensitivity.

Event Classification Table

Data Role=TRAIN Target=churn_ind Target Label=' '

False Negative	True Negative	False Positive	True Positive
78912	133483	207445	502976

Data Role=VALIDATE Target=churn_ind Target Label=' '

False Negative	True Negative	False Positive	True Positive
33654	57395	88719	215727

The fit statistics of the model are as shown in Table 3. The misclassification rate is 30% in the validation dataset.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
churn_ind		AIC	Akaike's Information Criterion	1096583	.
churn_ind		ASE	Average Squared Error	0.203237	0.203038
churn_ind		AVERR	Average Error Function	0.593919	0.59351
churn_ind		DFE	Degrees of Freedom for Error	922602	.
churn_ind		DFM	Model Degrees of Freedom	214	.
churn_ind		DFT	Total Degrees of Freedom	922816	.
churn_ind		DIV	Divisor for ASE	1845632	790990
churn_ind		ERR	Error Function	1096155	469460.8
churn_ind		FPE	Final Prediction Error	0.203331	.
churn_ind		MAX	Maximum Absolute Error	0.969845	0.973416
churn_ind		MSE	Mean Square Error	0.203284	0.203038
churn_ind		NOBS	Sum of Frequencies	922816	395495
churn_ind		NW	Number of Estimate Weights	214	.
churn_ind		RASE	Root Average Sum of Squares	0.450818	0.450598
churn_ind		RFPE	Root Final Prediction Error	0.450922	.
churn_ind		RMSE	Root Mean Squared Error	0.45087	0.450598
churn_ind		SBC	Schwarz's Bayesian Criterion	1099095	.
churn_ind		SSE	Sum of Squared Errors	375100.3	160601.3
churn_ind		SUMW	Sum of Case Weights Times Freq	1845632	790990
churn_ind		MISC	Misclassification Rate	0.310308	0.309417

Table 3: Fit statistics of logistic regression model

The ROC plot can be seen in the below figure. The area under the curve, which is proportional to the performance of the models built, suggests that they perform better from that the base line model.