**Customer Churn Analysis using Telecommunications Data**

J.K.D.B.G.Jayaneththi (209335F)

P.K.D.N.M.Alwis (209305N)

*Department of Computer Science and Engineering*

**Abstract**

***With the rapid development of communication technology, the field of telecommunication faces complex challenges due to the number of vibrant competitive service providers. Customer Churn is the major issue that faces by the Telecommunication industries in the world. Churn is the activity of customers leaving the company and discarding the services offered by it, due to the dissatisfaction with the services. The main areas of this project contend with the ability to identify potential churn customers and mine the relevant patterns embedded in the collected data. For analysis and model building, Pearson chi-square test association rule mining and logistic regression algorithm was used.***

***Keywords: Pearson chi-square test, association rule mining, logistic regression***

1. **Introduction**

The telecommunication sector is one of the main industries in developed countries. The level of competition has risen due to technical progress and the increasing number of operators. So, companies are working hard to survive in this competitive market depending on mainly three main strategies. They are 1) acquiring new customers, 2) increase the retention period of the customers and 3) upsell the existing customers. Among them it has been identified that second strategy is the most powerful one because retaining an existing customer costs much lower than acquiring a new one. Hence, to apply the second strategy companies have to decrease the potential of customer’s churn. . Customer churn is the action of the customer who is like to leave the company and it is one of the mounting issues of today’s rapidly growing and competitive telecommunication industry. Customers’ churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase. The data used in this project contains customer information which is collected through a survey. Descriptive, diagnostic and predictive analysis has been done on the data set to identify the factors which are affecting for customer churn in telecommunication industry. Also this predicts the churn rate of customers.

1. **Problem identification**

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up with the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning In telecommunication industry each company provides the customers with huge incentives to attract them to switch to their services, it is one of the reasons that customer churn is a big problem in the industry nowadays. To prevent this, the company should know the reasons for which the customer decides to move on to another telecom company.

1. ***Objectives***

The main objective of this project is to assess customer churn rate of telecommunication companies.

The supporting objectives examined are to:

1. Classify customers into various categories to enhance marketing and promotional activities.
2. Identify the factors which have a high effect on customer churn of the telecommunication companies.
3. Mine the relevant patterns embedded in the collected data have a huge influence on the revenues and growth of the Telecommunication companies.
4. **Methodology and Experimental Design**
5. **Data Collection**

The questionnaire was used as the tool to collect the data primarily from customers. The Google drive plug-in was used to design the questionnaire. Training data was collected from the 200 respondents and 50 responses were received from respondents on the questionnaire for testing data. The data was collected during the period (October – November) of 2017.

Table 1: The variables used in dataset for this research

|  |  |  |
| --- | --- | --- |
| **No** | **Variable Name** | **Description** |
| 1 | Age, Gender, Occupation | Demographic variables considered |
| 2 | The number of networks | Identifies the number of mobile networks a customer is connected to and actively using |
| 3 | Frequently used network | Identifies the most frequently used mobile network by the consumer |
| 4 | Tariffs | The type of customer, whether a prepaid or post-paid customer |
| 5 | Tenure | Length of time a customer has been with a particular subscriber |
| 6 | Credit purchase amount (CpM) | Approximates the amount used to purchase call credits a month in rupees |
| 7 | Data purchase amount (DpM) | Approximates the amount used to purchase data bundles a month in rupees |
| 8 | Internet usage | Identifies whether customer have used internet facility or not |
| 9 | Product innovation | Determines whether product innovation is necessary for sustaining customers |
| 10 | Churn | Identifies whether customer have changed networks or not |

1. **Data Pre-processing**

To handling missing data and removing duplicated data values data pre-processing is done.

In doing the Pearson chi-square and predictive analysis the data types are needed to be converted into numerical values.

Table 2: Codes for alternatives

|  |  |  |
| --- | --- | --- |
| **Variable** | **Alternative** | **Code** |
| Gender | Female | 0 |
| Male | 1 |
| Occupation | Student | 1 |
| Government Employee | 2 |
| Private Employee | 3 |
| Own Business | 4 |
| Others | 5 |
| Network often used | Dialog | 1 |
| Mobitel | 2 |
| Airtel | 3 |
| Hutch | 4 |
| Etisalat | 5 |
| Tenure | Less than 1 year | 1 |
| 1-3 | 2 |
| 3-5 | 3 |
| Above 5 | 4 |
| Churn | No | 0 |
| Yes | 1 |
| Tariffs | Pre-paid | 1 |
| Post-paid | 2 |
| Both | 3 |
| Usage of Internet | No | 0 |
| Yes | 1 |
| Product innovation | No | 0 |
| Yes | 1 |
| Not sure | 2 |

After pre-processing the data set was analysed by using descriptive, diagnostic and predictive analysis methods to identify the factors which are affecting for customer churn and to predict the churn rate of the customers.

1. **Results**
2. ***Descriptive Analysis***

Descriptive analysis is the first step for conducting statistical analysis. It gives an idea about the distribution of the data set and helps to identify the associations among variables and makes the data set ready for further statistical analyses.

1. ***Identify the types of variables***

The total number of 200 observations of the dataset was used to identify churn customers.

Table 3: Summary of the Types of Variables

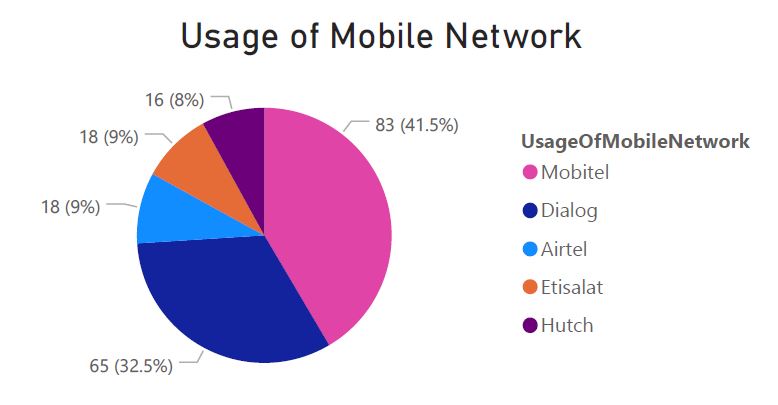
|  |  |  |
| --- | --- | --- |
| **No** | **Variable Name** | **Type of the variable** |
| 1 | Age | Discrete |
| 2 | Gender | Nominal |
| 3 | Occupation | Ordinal |
| 2 | Monthly income | Discrete |
| 3 | Frequently used network | Ordinal |
| 4 | Tariffs | Ordinal |
| 5 | Tenure | Ordinal |
| 6 | Credit purchase amount (CpM) | Continuous |
| 7 | Data purchase amount (DpM) | Continuous |
| 8 | Internet usage | Ordinal |
| 9 | Product innovation | Ordinal |
| 10 | Churn | Ordinal |

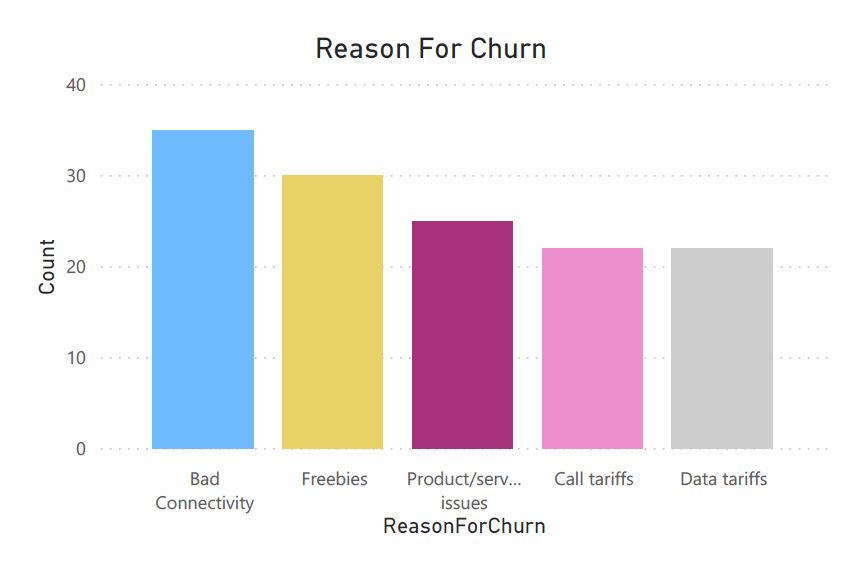
According to the data set 51.5 percent of the respondents were male while 48.5 percent were female.

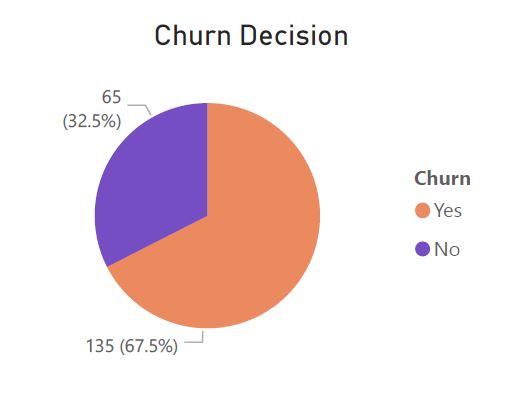
Majority of the respondents are government employees. Nearly 79 percent of the respondents use their mobile phone for the purpose of phone calls and internet usage. However, more respondents (41.5 percent) use Mobitel as their mobile network. Dialog follows with the second highest respondents of 32.5 percent.

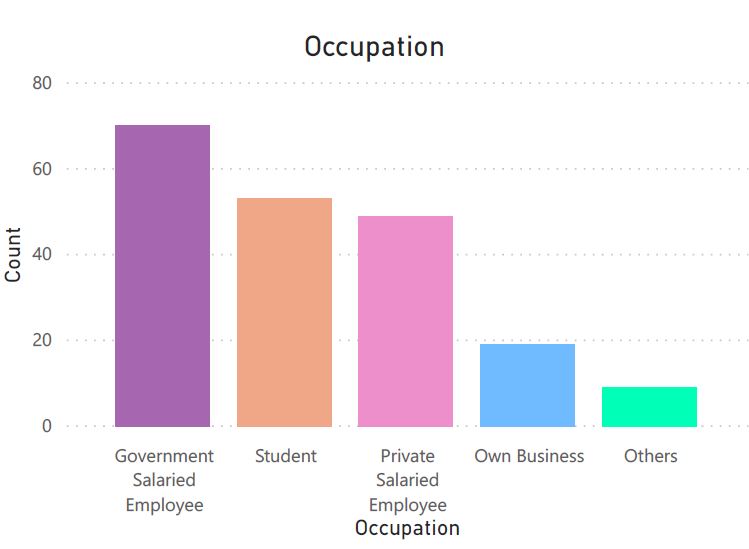
Out of the total number of (200) respondents, 80 percent indicated that they use the internet facility and the 76 percent of respondents used prepaid service package to both voice calls and internet. Nearly 52 percent of respondents need product innovation for their loyalty to a telecommunication network.

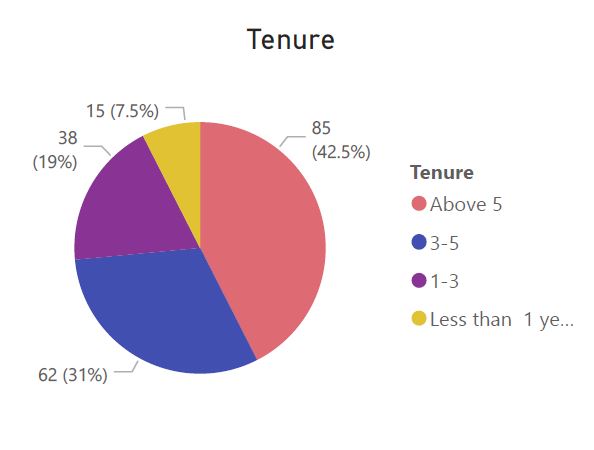
1. ***Representation of some Descriptive Analysis***

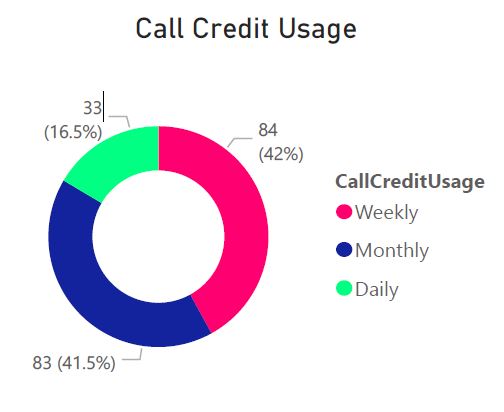
****

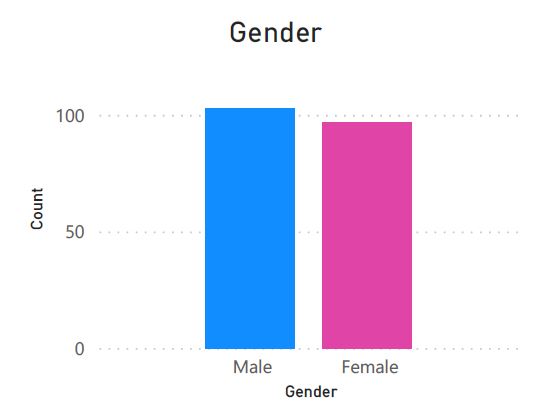
****

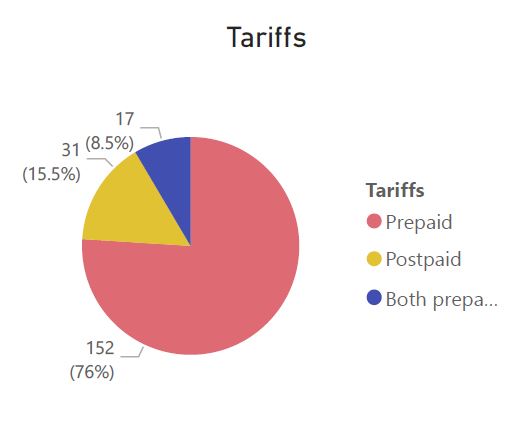
****

****

****

****

****

****

1. ***Diagnostic Analysis***

In descriptive analysis we have gained an idea on what has happened. And the next step is to identify “why something has happened?” For that diagnostic analysis has to be done.

In descriptive analysis it has been identified that 67.5% of the customers churn the network. So, in diagnostic analysis the main causes for that churn need to be identified.

1. **Pearson Chi-square Test**

Pearson Chi-square test is used to evaluate the variables which are associated with the decision of churn that can be used in the predictive model building. The test produced significant results (p–value is less than α level of 0.05) to indicate that some of the variables have an association with the decision to churn.

Table 4: Summary of association of each attributes

With the churn decision

|  |  |  |
| --- | --- | --- |
| **Variable** | **P-Value** | **Association** |
| Marital Status | 0.645 | No |
| Gender | 0.038 | Yes |
| Age | 0.005 | Yes |
| Occupation | 0.011 | Yes |
| Monthly Income | 0.001 | Yes |
| Purpose of mobile phone usage | 0.107 | No |
| No of mobile network connected | 0.003 | Yes |
| Mobile network often used | 0.006 | Yes |
| Tenure | 0.000 | Yes |
| CpM | 0.004 | Yes |
| Tariffs | 0.029 | Yes |
| Internet usage | 0.020 | Yes |
| DpM | 0.105 | No |
| Product Innovation | 0.021 | Yes |

According to the results of the Pearson Chi-square test it has been identified that:

* Gender
* Age
* Occupation
* Monthly Income
* Number of mobile networks connected
* Mobile network often used
* Tenure
* Credit usage per month
* Tariff
* Internet usage
* Product innovation

are the main attributes which are associated with customer churn decision.

1. **Association Rule Mining**

Association rule mining used to determine interestingness patterns and trends between variables in the dataset. It is contracted to identify strong rules explored in the dataset using some measures of interestingness. In creating the model, the Frequent Pattern Growth (FP-Growth) algorithm was used to mine associations between variables that result in a churn decision with particular interest and focus on confidence.

Table 5 shows ten (10) generated association rules were selected based on filtering the conclusion as the decision of churn is yes and sorted in descending order in line with confidence. The sorted rules have a maximum confidence of 95.5 % and a minimum of 84.6%.

Table 5: Top (10) generated rules

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Premises** | **Conclusion** | **Support** | **Confidence** | **Laplace** |
| InternetUsage=Yes, Gender =Male and Tenure = 3-5 years | Churn\_Yes | 0.105 | 0.955 | 0.995 |
| Tariffs=Prepaid, Gender=Male and Tenure= 3-5 years | Churn\_Yes | 0.100 | 0.952 | 0.995 |
| Gender=Male and Tenure=3-5 years | Churn\_Yes | 0.130 | 0.929 | 0.991 |
| MobileNetworkOftenUsed=Mobitel and Tenure=3-5 years | Churn\_Yes | 0.115 | 0.920 | 0.991 |
| InternetUsage=Yes and MobileNetworkOftenUsed=Mobitel | Churn\_Yes | 0.105 | 0.913 | 0.991 |
| Tariffs=Prepaid and Tenure= 3-5 years | Churn\_Yes | 0.230 | 0.902 | 0.980 |
| InternetUsage=Yes, Tariffs=Prepaid and Tenure= 3-5 years | Churn\_Yes | 0.190 | 0.884 | 0.979 |
| Tenure= 3-5 years | Churn\_Yes | 0.270 | 0.871 | 0.969 |
| Tariffs=Prepaid, Gender=Female and Tenure= 3-5 years | Churn\_Yes | 0.130 | 0.867 | 0.983 |
| InternetUsage=Yes, Tariffs=Prepaid and Gender=Female | Churn\_Yes | 0.110 | 0.846 | 0.982 |

1. ***Predictive Analysis***

After performing diagnostic analysis we have to identify “what will happen in the future?” for that predictive analysis techniques have to be applied. For that Logistic regression algorithm was used in this project.

As the result of generating the logistic regression model, it built up a statistical model which consists of two mathematical equations to calculate the ability of a person being churner or non-churner.

Equation 1: Calculating Y’

Y’ = 0.067\*Gender+(-0.0014)\*Age

+0.047\*Occupation

+0.000014\*MonthlyIncome

+(-0.727)\*Tenure+(-0.197)\*Tariffs

+0.000103\*CpM+2.666

Equation 2: Calculating P(1)

P(1) = exp(Y')/(1 + exp(Y'))

Equation 1 consists of most relevant variables which are most affected by the churn decision. The variables values should be replaced by this equation and then the value of Y' can be calculated. Then the calculated Y' value should be applied to equation 2 and calculate the value of P(1). Prediction of being a churn or non-churn customer is depending on this P(1) value.

If the P(1) value is equal or greater than 0.5, then the prediction result is positive and the person will be a churner. If the P(1) value is less than 0.5, the result is close to 0 (zero). It means the prediction result is negative and the person will be a non-churner.

To test the validity of the built model a confusion matrix and accuracy was calculated. The model produced 72% accuracy.

Table 6: Confusion Matrix with training data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LR** |  | no | yes | % correct |
| no | 27 | 38 | 41.5 |
| yes | 18 | 117 | 86.6 |
| Overall Percentage | | | | **72%** |

1. **Discussion and Conclusion**

Customer churn is a major problem and one of the most important concerns for telecommunication industry. The companies should identify why the customers are churning and they have to take respective measures to reduce the churn because retaining an existing customer costs much lower than acquiring a new one. This study focuses on identifying the highly affecting factors for customer churn and developing a predictive model based on those factors. Descriptive, diagnostic and predictive analysis was applied on the collected data set for getting the results. In descriptive analysis all the attributes were categorized into nominal, ordinal, discrete and continuous. And data set was studied how the data was distributed among different attributes. During diagnostic analysis Pearson’s chi square test was applied for the data set and it was identified that Gender, Age, Occupation, Monthly Income, Mobile network often used, Tenure, Credit usage per month, Tariff, Internet usage and Product innovation are the factors which are highly associated with customer churn. Further association rule mining also used and 10 rules were identified which has a high effect on customer churn. In predictive analysis logistic regression algorithm was used and based on the results an equation was generated which can be used by the telecommunication companies to identify customer churn. So, the telecommunication companies can use the results presented by this study to identify the factors affecting for customer churn and they can reduce the amount of customer churn and retain successfully in the telecommunication industry.

**References**

Amal M. Almana, Mehmet Sabih Aksoy, Rasheed Alzahran, 2014. *A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry,* Riyadh, Saudi Arabia: s.n.

Amjad Khan, Zahid Ansar, n.d. *Comparative Study Of Data Mining Techniques In Telecommunications-A Survey.* Mangalore, India, Dept of Electronics and Communication, P.A. College of Engineering.

Essam Shaaban, Yehia Helmy, Ayman Khedr and Mona Nasr, 2012. A Proposed Churn Prediction Model. *International Journal of Engineering Research and Applications (IJERA),* June-July.2(4).

K.Dahiya,S.Bhatai, 2015. Customer churn analysis in telecom industry. *4th International Conference on Realibility, Infocom Tehnilogies and Optimization(ICRITO).*

M.Balasubramanian, M.Selvarani, 2014. *Churn Prediction in Mobile Telecom Systems Using Data Mining Techniques,* Chidambaram: s.n.

Manpreet Kaur and Prerna Mahajan, 2015. Churn Prediction in Telecom Industry Using R. *International Journal of Engineering and Technical Research (IJETR),* May.3(5).

Rahul J. Jadhav and Usharani T. Pawar, 2011. Churn Prediction in Telecommunication Using Data Mining Technology. *International Journal of Advanced Computer Science and Applications,,* February.2(2).

Shin-Yuan Hung, Hsiu-Yu Wang, 2015. *Applying Data Mining to Telecom Churn Management,* Taiwan, ROC: s.n.

V. Umayaparvathi, K. Iyakutti, April 2016. A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. *International Research Journal of Engineering and Technology (IRJET),* Volume 03, p. 04.