# Customer Churn Prediction in Telecom Industry

**Bhupesh Sudhakar janwalkar [1] Ms. Priyanka Chaudhari[2]**
[2]Professor
[2]Department of MCA
[1,2]IMCOST, Thane, Maharashtra, India

*Abstract*— Customer churn refers to a decision made by the customer about stop subscribing to service, also known as customer attrition. It is also referred as loss of clients or customers. Churn prediction aims to detect customers intended to leave a service provider. Customer loyalty play major Role. As per 80/20 customer profitability rule, 20% of customers are generating 80% of revenue. So, it is very important to predict the users likely to churn from business relationship and the factors affecting the customer decisions. Predictive models can provide correct identification of possible churners in the near future in order to provide a retention solution. Retaining one customer costs an organization from 5 to 10 times than gaining a new one.

***Key words:*** Churn Prediction, Classification, Predictive Models, Retention Solution, Confusion Matrix

## I. INTRODUCTION

Churn prediction process is a highly debated research area for more than ten years. Researchers from different disciplines have tried to analyse this problem from their own perspectives to figure out a clear understanding and to recommend an effective solution for churners in many business areas. Abbasimehr et al. [1] state that churn prediction is a useful tool to predict customer at churn risk.

Conventional churn prediction techniques have the advantage of being simple and robust with respect to defects in the input data, they possess serious limitations to the interpretation of reasons for churn. Therefore, measuring the effectiveness of a prediction model depends also on how well the results can be interpreted for inferring the possible reasons of churn [2]. The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events [3]. Most Machine Learning techniques derive their predictions from the value of a set of variables associated with the entities in a database. There are many Machine Learning techniques that can be used in classification and clustering customer data to predict churners in the near future. These techniques may use Decision Tree, Logistic Regression, and Random Forest to predict churners. This paper is organized as follows. Section 2 describes the types of churners. Section 3 shows the existing prediction models. Section 4 escribes the proposed churn prediction model. Finally; conclusion and future work are presented.

## II. TYPES OF CHURNERS

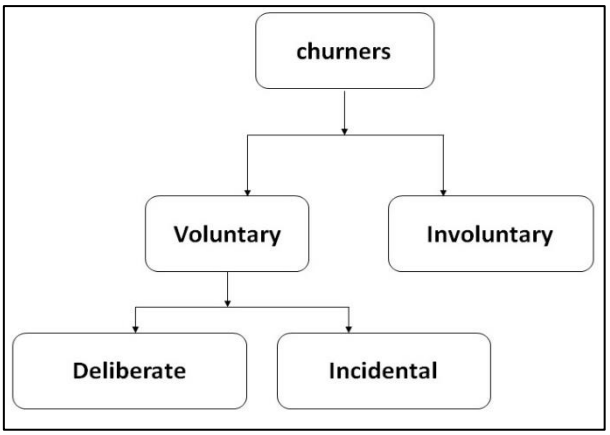


Fig. 1: Churn Taxonomy

The churn rate, also known as the rate of attrition, is the percentage of subscribers to a service who discontinue their subscriptions to that service within a given time period. For a company to expand its clientele, its growth rate, as measured by the number of new customers, must exceed its churn rate. As figure 1 depicts; There are two main categories of churners which are voluntary and involuntary [4]. Involuntary churners are the easiest to identify. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider. Involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. Figure 1: churn taxonomy Voluntary churn can be sub-divided into two main categories, incidental churn and deliberate churn. Incidental churn occurs, not because the customers planned on it but because something happened in their lives. For example, change in financial condition churn, change in location churn, etc. Deliberate churn happens for reasons of technology (customers wanting newer or better technology), economics (price sensitivity), service quality factors, social or psychological factors, and convenience reasons.

Deliberate churn is the problem that most churn management solutions try to solve [4] [5].

## III. PREDICTIVE MODELS

Predictive modelling is mainly concerned with predicting how the customer will behave in the future by analysing their past behaviour. Predicting customers who are likely to churn is one example of the predictive modelling [6]. Predictive modelling is used in analysing Customer Relationship Management (CRM) data to produce customer-level models that describe the likelihood that a customer will take a particular action. The actions are usually sales, marketing and customer retention related. There are many models that can used to define distinguish between churners and non-churners in an organization.

 

### *A. Decision Trees*

Decision trees is most popular type of predictive model. It has become an important knowledge structure, used for the classification of future events [7]. Decision trees usually consists of two main steps, tree building and tree pruning. The tree-building step consists of recursively partitioning the training sets according to the values of the attributes. The partitioning process continues until all, or most of the records in each of the partitions contain identical values. Some branches may be removed because it could consist of noisy data. The pruning step involves selecting and removing the branches containing the largest estimated error rate. Tree pruning is known to enhance the predictive accuracy of the decision tree, while reducing the complexity [8].

### *B. Random Forest*

In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. An error estimate is made for the cases which were not used while building the tree. That is called an OOB (Out-of-bag) error estimate which is mentioned as a percentage. [9]

### *C. Logistic Regression*

The Logistic Regression is a regression model in which the response variable (dependent variable) has categorical values such as True/False or 0/1. It actually measures the probability of a binary response as the value of response variable. [10]

## IV. THE PROPOSED MODEL

The proposed model is composed of five steps. As shown in figure 2, these steps are: Define the problem, Data cleansing/preprocessing and feature engineering, Model selection and Training, Evaluate and test the model result and performance, Deploy the model.

As figure 2 depicts; step 1 will do classification and within Step 1, step 2 and step 3 Exploratory Data Analysis is done. Step 3 Will give Model selection from Logistic Regression, Decision Tree and Random Forest. Step 4 produce Confusion Matrix precision and recall.[11]

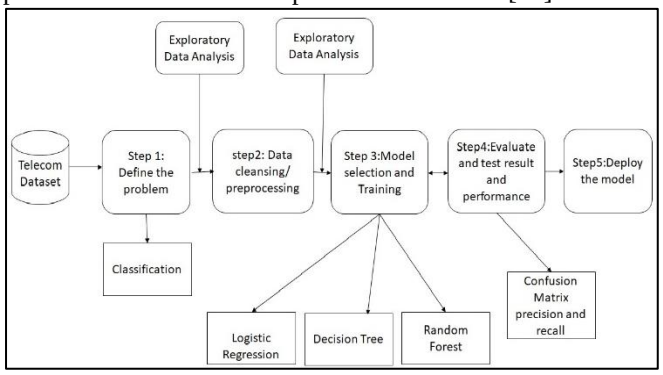


Fig. 2: The Proposed Churn Prediction Model

### *A. Churn Prediction with the Proposed Model*

During this case study R Studio tool is used with the data was downloaded from IBM Sample Data Sets. Each row represents a customer, each column contains that customer's attributes:

| Attribute | Attribute Description |
|---|---|
| customerID | Customer unique identification |
| gender | (female, male) |
| SeniorCitizen | (Whether the customer is a senior citizen or not (1, 0)) |
| Partner | (Whether the customer has a partner or not (Yes, No)) |
| Dependents | (Whether the customer has dependents or not (Yes, No)) |
| Tenure | (Number of months the customer has stayed with the company) |
| PhoneService | (Whether the customer has a phone service or not (Yes, No)) |
| MultipleLines | (Whether the customer has multiple lines r not (Yes, No, No phone service) |
| InternetService | (Customer's internet service provider (DSL, Fiber optic, No) |
| OnlineSecurity | (Whether the customer has online security or not (Yes, No, No internet service) |
| OnlineBackup | (Whether the customer has online backup or not (Yes, No, No internet service) |
| DeviceProtection | (Whether the customer has device protection or not (Yes, No, No internet service) |
| TechSupport | (Whether the customer has tech support or not (Yes, No, No internet service) |
| StreamingTV | (Whether the customer has streaming TV or not (Yes, No, No internet service) |
| streamingMovies | (Whether the customer has streaming movies or not (Yes, No, No internet service) |
| Contract | (The contract term of the customer (Month-to-month, One year, Two year) |
| PaperlessBilling | (Whether the customer has paperless billing or not (Yes, No)) |
| PaymentMethod | (The customer's payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic))) |
| MonthlyCharges | (The amount charged to the customer monthly—numeric) |
| TotalCharges | (The total amount charged to the customer—numeric) |
| Churn | (Whether the customer churned or not (Yes or No)) |

Table 1: Attributes of the Data Set. [11]

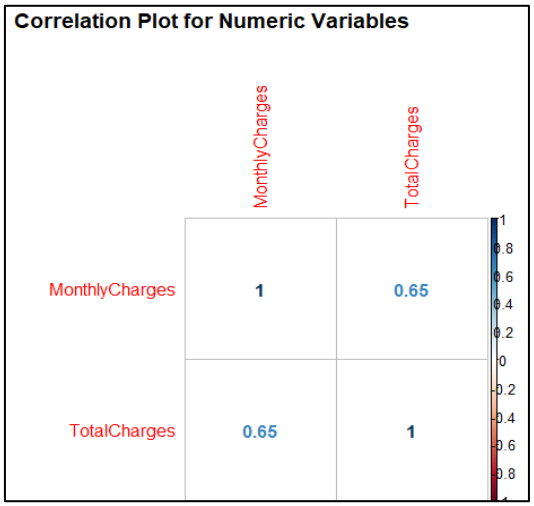


Fig. 3: Exploratory Data Analysis & Feature Selection Correlation between Numeric Variables

The Monthly Charges and Total Charges are correlated. So, one of them will be removed from the model. We remove Total Charges.

### *B. Interpretation of Results*

The result of the Model Selection and Training step is represented in table 3.

| Decision Tree | Prediction | No | Yes |
|---|---|---|---|
| | No | 1428 | 342 |
| | Yes | 120 | 218 |
| Logistic Regression | No | 1402 | 146 |
| | Yes | 298 | 262 |
| Random Forest | No | 1371 | 286 |
| | Yes | 117 | 274 |

Table 2: Confusion Matrix Precision Result

The accuracy of the predicted results is shown is table 3

| Technique | Accuracy |
|---|---|
| Decision Tree | 0.7808 |
| Logistic Regression | 0.7893 |
| Random Forest | 0.7813 |

Table 3: Accuracy Comparison for Decision Tree, Logistic Regression, Random Forest Techniques

## V. CONCLUSIONS

The IBM dataset we use and apply logistic regression decision tree and random forest techniques for customer churn analysis, throughout the analysis I have learned several important things:

- customer of month-to-month contract having paperless billing and within 12-month tenure are more likely to churn
- Customers of one or two-year contract that are not using paperless billing with longer than 12 months tenure are less likely to churn.
- There is no relationship between gender and churn.
- Attributes such as paperless billing, tenure group, monthly charges, internet service and contract appear to play a role in customer churn.

## REFERENCES

[1] H. Abbasimehr, M. Setak, M. Tarokh. A Neuro-Fuzzy Classifier for Customer Churn Prediction. International Journal of Computer Applications, vol. 19, no. 8, pp. 35-41, April, 2011
[2] V. Lazarov and M. Capota. Churn Prediction. Business Analytics Course. TUM Computer Science, December 2007.http://home.in.tum.de/~lazarov/files/research/papers/churn-prediction.pdf
[3] Carlo Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, Ltd. 2009 ISBN: 978-0-470-51138-1
[4] S. Gotovac. "Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services," vol. 51, no. 3, pp. 275-283, 2010
[5] H. Kim, and C. Yoon, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market."Telecommunications Policy. Vol. 28 No.: PP. 751-765, 2004.
[6] M. Hassouna, Agent Based Modelling and Simulation: An Examination of Customer Retention in the UK Mobile Market. PhD thesis, Brunel University, UK, 2012.
[7] K. Muata, and O. Bryson, Evaluation of Decision Trees: AMulti Criteria Approach, Computers and Operational Research, 31, 1933-1945, 2004
[8] W. Au, C. Chan, and X. Yao, A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction, IEEE transactions on evolutionary computation, 7, 6, 532-545, 2003.
[9] http://www.tutorialspoint.com/r/r_random_forest.htm
[10] https://www.tutorialspoint.com/r/r_logistic_regression.htm
[11] https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest

 