ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

Cost Analysis of Misclassification of Customer churning in Telecom Industry

Nikita Arora, Atul Mishra
BML Munjal University
nikita.arora.17csc@bml.edu.in, atul.mishra@bmu.edu.in

Abstract
Organizations facing a high rate of customer churn suffer a huge loss both in terms of finances and number of customers, which then hinders the telecom industry's development. Loss and cost become a major concern for the telecom industry. Several studies have shown that during the expansion phase of its life cycle, the company focuses on increasing its sales to increase the number of customers than churned clients. On the other hand, during a mature phase of their life cycle, the company focuses more on reducing the customer churn rate by adopting some strategy to retain the old customers. The proposed methodology provides an insight of the loss on the part of the company if a customer churns, whether it may or may not be expected to be a churned customer and the cost of the customer's services based on their projected churn status. Results show that loss and cost analysis is of greater value for the marketing and management departments of any telecom company to plan their strategy to reduce the value of the loss.

Keyword: *Churn analysis, business services, telecommunication, customer, logistic, random forest.*

1. Introduction

Customer churn is the major and typical problem that arises when a customer stops using the services of the enterprise. It obstructs the growth of the companies in the market. Business leaders believe that it is often easier for the company to keep the customers it already has than to get the new one. The companies that do not prefer maintaining strong and reliable business relationships with their clients risk high churn rates which then jeopardize the company's future.

To overcome and minimize the loss of the company due to customer churn, enterprises follow strategies or methods to predict whether the customer will stop using the company's services or not soon. As mentioned in [5], the performance of predictive churn model is generally improved by using decision tree which has been used widely and some optimal parameters (the time of sub-period being 10 days, misclassification cost being 1:5, and random sample method for train set) of models that are found under the help of three research experimentations (changing sub-periods for training data sets, changing misclassification cost in churn model, changing sample methods for training data sets).

Customer churns when either he is being provided with poor customer service (like poor product value, lack of interest in a product) or when he is receiving a better deal from some other organization, maybe at a better price and with a bigger profit. As stated in [4], the number of non - churned clients in the dataset of any telecom company is always likely to be more in number as compared to the number of churned users. So, misclassification brings huge losses to the company. The losses are in the form of marketing costs and various offers that the company provides to its subscribers/clients to retain them and when despite these customer retention efforts by the company, if the customer still churns then the cost of these offers and marketing aid becomes the loss for the company. When the subscriber or clients of the company begins to churn, the company should pay special attention towards the number of churns in the specific time frame, the recurring business value that is lost and the most importantly, modification in the retention strategies so as not to lose them in large numbers.

Negative reviews by the churned customer for a company are a big gift for its competitors in the market. These churned customers may come back to prevent the company from closing the deal and hitting the goals. To minimize the losses in the form of misclassification costs, the company comes up with various performance metrics to minimize the loss.

One of the performance metrics that have been highlighted in the paper is the confusion matrix. It is a very powerful tool that breaks model performance

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

into true positives, false positives, true negatives, and false negatives. A higher value of false positives and false negatives indicate the higher value of the loss incurred by the company on the part of churn customer. But to minimize the loss, we need to calculate approximate loss that the company may bear due to misclassification of its customers as churned users or non-churned customers.

2. Related work

Comparison between the cost-sensitive and cost insensitive predictive analytics models [1] and the effectiveness of a 2voluntary churn campaign is evaluated taking into account the type of offers made by the company to retain its customers, their financial cost and probability of acceptance of offer based on the customer profile.

Liu et.al [2] have reported that the churn prediction model with customer segmentation gives higher accuracy as compared to the one without considering the misclassification cost. A customer churn model has been established based on the misclassification cost and customer segmentation to analyze the customer behavior dataset of the telecom company. The hard clustering technique k-means clustering is used to cluster the customers into customer groups based on their consumption preferences, behaviour and demand.

Fridrich et.al [3] discussed the performance of the classification model is assessed by its confusion matrix and most commonly through the area under the curve metric. AUC is the simplest and threshold independent cost metric method but does not perform well in the context of the impact of misclassification errors. To overcome the problem, several other cost-benefit metrics have been proposed in recent researches like Customer Lifetime Value, Hand, etc.

In a churn prediction context [4], the number of non-churned customers are often higher than the churned customers present in the dataset. Hence, wrong predictions for the churned users are very costly; however, they do not exceedingly influence the error rate. A churn prediction model minimizing the error rate usually results in a futile model, because it predicts customers as non-churned customers.

The performance of predictive churn model is generally improved by using decision tree [5] which has been used widely and some optimal parameters like the time of sub-period being 10 days, the misclassification cost being 1:5, and random sample method for train set of models that are discovered under the help of experimentations like changing sub-periods for training datasets, changing the cost of misclassification in churn model and changing the sample methods for training datasets.

Ahmad et.al [6] developed a churn prediction model and observed a significant increase in its performance after using SNA (Social Network Analysis) features. The model experimented four algorithms: Decision Tree, Gradient Boosted Machine Tree GBM, Extreme Gradient Boosting XGBOOST and Random Forest.
However, the best results were obtained by applying XGBOOST algorithm. This algorithm was used for classification in this churn predictive model.

Amin et.al [7] mentioned the grouping of a dataset into various zones based on a specific criterion to develop the churn prediction model with enhanced accuracy. The dataset has been grouped into different zones to get the approximate single-line phone service charges, multi-line phone service charges, and internet service charges. The first group contains clients that are being provided with only single-line phone service and no other service. The second group contains clients that are being provided with only multi-line phone service and the third group contains only those customers who receive only internet services.

The main reasons for customer churn in the fast growing industry i.e. telecommunication sector in Macedonia. The methodology [8] for the customer churn prediction includes data processing, data analysis, implementing various algorithms, comparing the accuracy of the classifiers and then choosing the best one with higher accuracy for the prediction of customers to be the churned customer or the non-churner.

Another study on marketing research is represented by Min, Sungwook, et al. [9] which develops an analytical model to investigate the type of investments in customer acquisition and retention for product and services. The research represented analytical training and testing with firm-level operating data of telecom markets from 41 countries during 1999–2007. The research shows that a company's acquisition cost per customer is more responsive to market position and competition than retention cost per customer.

Kayaalp et.al[11] evaluates the relevant studies on customer churn analysis on the telecommunication company to mention the frequently used data mining

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

techniques, their accuracy, performance and results. It is evident that though there exist many feature selection methods in the literature, the features should be selected carefully as they have a remarkable effect on the performance of the analysis. Moreover, it can be suggested that the diversification of the fields on which data mining is used can surely provide great benefits to the organizations, as well as to the people who receive services and products from these organizations.

Gürsoy et.al [12] determines the customers who want to stop using the services of the organization and create specific campaigns to them by using the customer data of a huge telecom firm. Logistic regression and decision trees analysis helps the author identify the reasons for the customer to churn. It is often observed that logistic regression and random forest classifier provides results in customer churn analysis with accuracy that may or may not be higher than that of the decision trees depending upon the customer data.

Dahiya et.al [13] proposes a new framework for customer churn prediction model and implements it using WEKA data mining software. The efficiency of logistic regression and decision trees techniques have been compared. Effective methods need to be developed and the existing old methods need to be improved to provide churn prediction results with greater accuracy. The active and passive nature of the industry ensures that data mining techniques have become a remarkable aspect of the telecom industry.

Liu et.al [14] proposes the systematic method of predicting the possible customer churn from the imbalanced dataset of an organization. Binary logistic regression model, a stratified sampling method has been applied to deal with the problem of imbalanced data prediction. Binary LRM turns out to be effective to forecast the telecom customer churn.

Esteves et.al [15] compares the performance of six distinct algorithms that identify the customers who are more probable to stop using the services from their telecom provider. The algorithms are Random Forest, KNN, Ada Boost, Naive Bayes, ANN and C4.5.The models are evaluated based on three criteria: sensitivity, area under the curve, and specificity, with special importance to the first two criteria. Random forest proves to be the most acceptable in all the test cases.

3. Proposed Methodology

We have used telco customer churn dataset from Kaggle [10]. Most relevant services which are mentioned in the dataset are the phone service considering single-line phone service and multi-line phone service separately and the internet service. We treated single-line phone service, multi-line phone service and internet service columns(variables) as the categorical variables with possible values as 0,1 or 2.

We have used classic predictive modelling techniques logistic regression and random forest for modelling the binary categorical variable churn in the dataset with the descent accuracy. The model works with different accuracy for predicting the customers which are at higher risk of churning i.e. cancelling their subscriptions or services being offered by the company resulting in huge savings for the enterprise. Finally, the best one is chosen to make customer churn prediction.

Higher the cost of misclassification of churn variable, higher is the loss for the company. The misclassification cost can be represented in two ways. Firstly, in terms of the loss that the company had to bear and secondly in terms of the average cost that the company needs to pay to provide the services to the customers.

Loss and average cost on the part of the company is computed for four different situations.

- If the customer churns but was predicted not to churn.
- If the customer churns and was even predicted to churn.
- If the customer was predicted to be a churned customer but turns out to be a non-churned customer.
- If the customer was predicted to be a non-churned customer and turns out to be the same.

Confusion matrix being a performance metric gives a clear idea of the number of customers falling in any of the four categories (true positives, false positives, true negatives, and false negatives). It gives us an idea about various ways in which our classification logistic regression model is confused with predictions.

To compute the loss and cost for the company, cost for the internet service and the phone service is computed with the help of division of customers into

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

4 categories- true positive, true negative, false positive and false negative.

We computed the loss and average cost for the company using logistic regression classifier as well as the random forest classifier. Both the classifiers work with different accuracy, but the best one can be used for prediction of customer churn.

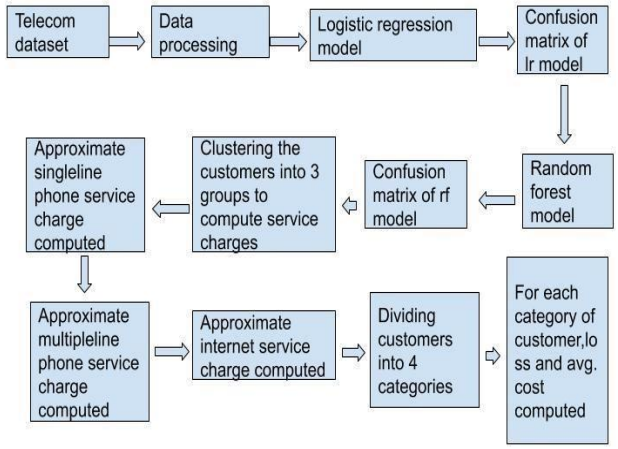

Figure 1: The Process

4. Performance

As mentioned above, we treated single-line phone service, multi-line phone service and internet service columns(variables) of the dataset as the categorical variables with possible values as 0,1 or 2.

We preprocessed the data by importing libraries, using the dataset, encoding the categorical data and splitting it into test and train set. The dataset is divided into train and test set in the ratio of 80:20.

Different classifiers like logistic regression and random forest classifier have been implemented on the churn customer dataset and the best out of two can be referred to by the telecom company based on the accuracy score and other performance metric scores like recall, precision and f-score for the customer churn prediction.

The logistic regression model classifies the customer into churn customer and the non-churn customer with an accuracy of 79% whereas the random forest classifier works with lower accuracy of 77%. The two most important assumptions which form the basis of our conclusion in the paper are as follows:

Firstly, we have used a parameter to represent the type of phone service for every customer in such a way that value is equal to 0 if the customer does not receive phone service and 1 if the customer opts for single-line phone service and 2 if a customer opts for multi-line phone service.

Secondly, for every customer, if internet service is provided then the value of categorical variable internet variable is turned 1 else 0 without considering which sub-services are being provided to him under the category internet service.

Loss and average cost in the above four situations are computed as follows.
1. If the customer churns but was predicted not to churn, the loss to the company will be equal to the average cost i.e. summation of internet and phone services charges provided by the company to the client.
2. If the customer churns and was even predicted to churn, the loss to the company would be zero and the average cost will also turn out to be zero for no internet and phone services will be provided to the client by the company.
3. If the customer was predicted to be a churned customer but turns out to be a non-churned customer, then the loss to the company will be zero in terms of money involved but might lose out the promising customer and the average cost will be equal to zero for the company would not provide him with the services.
4. If the customer was predicted to be a non-churned customer and turns out to be the same, then the loss to the company would be zero and the average cost also becomes zero.

Each row of the dataset carries information about a unique customer with different phone and internet services opted by him from the company and the total charges involved. To compute the loss and the average cost for the company in all the four situations as mentioned above, we cluster the customers into three groups based on the services that the customer is receiving from the company to find out the approximate mean value of their total charges to get the single-line phone service charge, the multiple line phone service charge and the internet service charge.

Group 1: Group of customers who receive only the single-line phone service and no other service (no multi-line phone service and internet service). The mean value of the total charges corresponding to the group of such customers is considered as the approximate single-line phone service charges per year.

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

Group 2: Group of customers who receives multiline phone service and no other service (no single-line phone service and internet service). The mean value of total charges for such customers gives us the approximate value of multiple line phone service charges per year.

Group 3: Group of customers who receives only the internet service among all the other services being provided by the company and then the approximate value of internet service charges per year is computed as the mean value of their total charges.

Considering the approximate annual charge for single-line phone service, multi-line phone service and the internet service, the approximate loss and average cost for the company are calculated for each of the above mentioned four situations.

5. Result

We have applied Logistic regression and random forest for classification of churned customers. Below table represents the comparison of both the classifiers.

Table 1 : Comparison between classifiers

	Recall	**Precision**	**f-score**
Logistic regression	0.83	0.90	0.86
Random Forest	0.91	0.80	0.85

The customer will surely fall under one of the four categories as mentioned below:

Case 1: The customer did not churn and was even predicted to be the loyal customer

Case 2: The customer churns and was even predicted to be the churn customer

Case 3: The customer churns but was predicted to be the loyal customer

Case 4: The customer did not churn but was predicted to be the churn customer

The cost and loss incurred by the company in terms of the services (phone service and internet service) provided to the customer as calculated from the telco dataset is the approximate figure and can be represented as the below matrix:

Table 2 : Loss and Cost in term of four categories of customers (All figures in INR)

	Case 1	**Case 2**	**Case 3**	**Case 4**
Cost	3287.5	0	3287.5	0
Loss	0	0	3287.5	0

Table 3 : Confusion matrix obtained from logistic regression

	Predicted to be a non-churn customer	**Predicted to be a churn customer**
Loyal customer	3232	347
Customer churns	680	672

Total cost incurred by the company as per the logistic regression classifier = (3232*3287.5) + (680*3287.5) = 10,625,200+2,235,500 = INR 12,860,700

The total monetary loss to the company will be INR 2,235,500 (3287.5*680) (Approx.). As the number of customers who were predicted to be the churned customer but remains the potential customer to the company is 672, so these customers may end receiving the services from the company and therefore, reduce the number of customers that the company may have. This turns out to be another loss to the company other than the monetary loss.

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

Table 4 : Confusion matrix obtained from Random forest classifier

	Predicted to be a non churn customer	**Predicted to be a churn customer**
Loyal Customer	3299	280
Customer churns	812	540

The total cost incurred by company as per the random forest classifier = (3299*3287.5) + (812*3287.5)=10,845,462.5+2,669,450 =13,514,912.5 INR

The total monetary loss to the company will be 2,669,450 INR (=3287.5 * 812) (Approx.). As the number of customers who were predicted to be the churned customer but remains the potential customer to the company is 812, so these customers may end receiving the services from the company and therefore, reduce the number of customers that the company may have. This turns out to be another loss to the company other than the monetary loss.

6. Conclusion

We have tried to represent the cost of misclassification that occurs during a classification process considering the case of telecom industry. Case study and the performance shows a comparison between two different classifiers in terms of loss and cost and the expense of the company for retaining the customer based upon services and projected churn status. Finally, we conclude, that the cost due to misclassification of churned customer is higher for the management and marketing departments.

7. References

[1] Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). A novel cost-sensitive framework for customer churn predictive modeling. Decision Analytics, 2(1), 5.

[2] Liu, Y., & Zhuang, Y. (2015). Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data. *Journal of Computer and Communications*, *3*(06), 87.

[3] Fridrich, M. (2018). Cost-benefit metrics in customer churn prediction: A review. International Masaryk Conference.

[4] Kristof Casement, (2014),"Improving customer retention management through cost-sensitive learning", European Journal of Marketing, Vol. 48 Iss 3/4 pp. 477 - 495

[5] Bin, L., Peiji, S., & Juan, L. (2007, June). Customer churn prediction based on the decision tree in personal handyphone system service. In *2007 International Conference on Service Systems and Service Management* (pp. 1-5). IEEE.

[6] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 28.

[7] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290-301.

[8] Petkovski, A. J., Stojkoska, B. L. R., Trivodaliev, K. V., & Kalajdziski, S. A. (2016, November). Analysis of churn prediction: A Case study on Telecommunication Services in Macedonia. In *2016 24th Telecommunications Forum (TELFOR)* (pp. 14). IEEE.

[9] Min,Sungwook, et al. "Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets." Journal of marketing research 53.5 (2016): 728-744.

[10] Telco Customer Churn Dataset : https://www.kaggle.com/blastchar/telco-customerchurn

[11] Kayaalp, F. (2017). Review of Customer Churn Analysis Studies in Telecommunications Industry. *Karaelmas Science and Engineering Journal*, *7*(2), 696-705.

[12] Gürsoy, U. Ş. (2010). Customer churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, *39*(1), 35-49.

[13] Dahiya, K., Bhatia, S. 2015. Customer Churn Analysis in Telecom Industry.4th International Conference on

ISSN: 2278-4632
Vol-10 Issue-5 No. 9 May 2020

Reliability, Infocom Technologies and Optimization (ICRITO),1-6

[14] Li, P., Li, S., Bi, T., & Liu, Y. (2014). Telecom customer churn prediction method based on cluster stratified sampling logistic regression.

[15] Esteves, G., & Mendes-Moreira, J. (2016, September). Churn perdiction in the telecom business. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)* (pp. 254-259). IEEE.

