Analysis and prediction of churn customers for telecommunication industry

Mrs. Ushasukhanya , mr surya vamsi saripudi , mr shivam sharma

SRM IST,KTR Chennai

Abstract

Churn Examination is one of the widespread used study on Subscription Oriented Businesses for analyzing the behavior and activities of customers in order to predict beforehand which customer is likely to exit the service agreement. Built on Machine Learning procedures and algorithms it has become very significant for companies in today's market as securing of another client is more costlier than their maintenance. The paper analyses the relevant studies on Customer Churn Analysis in Telecom Business to present overall information to readers about the commonly used data mining means, and performance of the methods. Initially, we present the details about the availability of public datasets and various customer details in each dataset for predicting customer churn. Then, we compare and contrast various analytical modeling systems and compare their performances and results. Conclusively, we review what kinds of performance metrics have been used to gauge the current churn prediction approaches. Examining all these three viewpoints is very critical for developing a more well-organized churn prediction structure for telecom businesses.

Keywords:

EDA – exploratory data analysis

CRM- customer relationship management

LRM- logistic regression model

SVM- support vector machines

Introduction

Mass customer data have been collected throughout the functional procedure of the mobile telephone company, it is quite import ant for a competitive company to collect effe ctive information in excessive data resources and then to create a combined information platform, but it seems impossible to deal wit h the useful data using the conventional method of database management,in telecom market competition coming into rough competition at times, several domestic telecommunication corporation commence to use several systems to solve the problem. Customer churn is a crucial tool during foundation and system set up predictive model based on customer conduct. With sometimes harsh competition from the

telecommunications market, the telecommun ications company starts using several system s to solve the problem. Customer churn is a crucial tool for establishing a predictive mod el based on customer behavior during the fo undation and system.In this paper we propose extrapolative models using machine learning to predict whether the customers in Communications/telecom firm will churn or not. We propose the machine learning models with different algorithms such as Naïve Bayes , Random forest. Prediction performance of each algorithms is estimated using accuracy matrix, More challenging is to set a model for Telecommunication sector as there are no contracts between a customer and Telecommunication concerning the duration of facilities/services .The telecom industry endures rising valuing burden globally. Studies to be done on customer churns is more critical for the Telecom companies nowadays.Classification problem – classification task comes under supervised learning in machine learning where the main goal is to establish models which are supervised by an external agent where the classifiers to training samples are known well in advance. The created models identify the class labels of an unseen sample using this technique.Feature Selection-Feature selection is the process of identifying and selecting the relevant features from the environment. The selected features are extracted using various feature extraction techniques . a feature vector is made which represents the set of features such that each feature vector is mapped to a class label. This helps in identifying class label . Moreover only relevant features are extracted as any unnecessary feature adds more to computational cost and random errors.

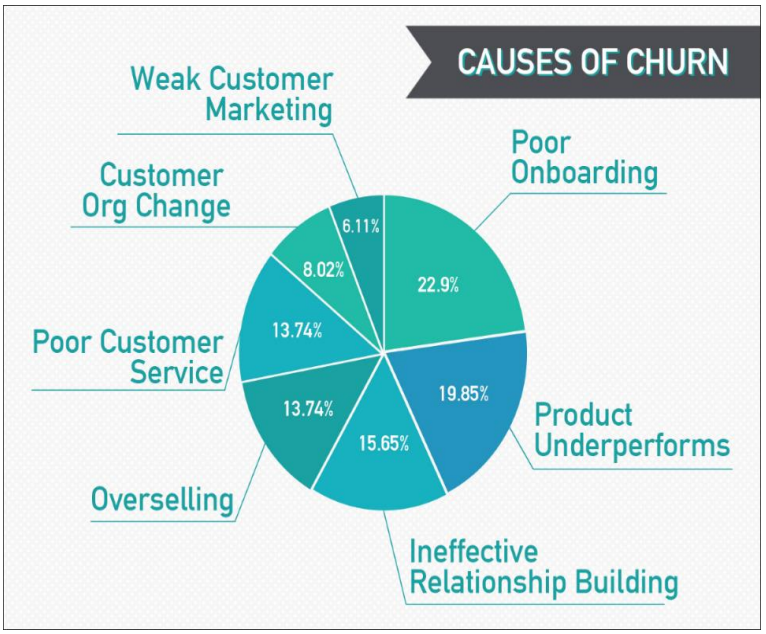


Fig 1.1

Literature survey

[1] This paper proposes a model for predicting employee churn in an organization. Employees are an important part of organization further, hiring a new employee turns out expensive for any organization and thus retaining current employees is the optimal solution. Linear support vector machine,c.5 decision tree random forest ,k-nearest neighbor and naïve bayes classifiers are used for classification. This study requires further exploration to minimize the prediction rate.[2] This paper

builds model to predict customer behavior in fitness industries it is found that a yearly gym membership allows the client to terminate their membership with hardly any advanced notice.[3]model based on logistic regression, decision trees and neural networks is built. This study finds it hard to find false positive rate for leavers at certain situations. Moreover it assumes that customers that have sporadic transactions i.e irregular gym attendance are also classified as churned customers. [3] A comprehensible customer churn prediction model is built for analyzing client behavior for determining which customer is likely to churn in the future. In order to prevent A model based on multiple kernel vector support machine approach is built .the drawback of MK-SVM is that while feature selection they can reduce some of the relevant features .This study has left the application of this framework to financial etc institutions as a future work.[4]This paper focuses on analyzing customer churn in telecom industry using logistic regression, neural network and decision trees, although neural networks perform very well for classification and prediction tasks it only does so for very large datasets. Moreover it takes the same amount of time for processing much smaller data sets.they left this work for future studies for making this model to handle large datasets.[5] This paper uses the technique of decision trees to develop a model for telecommunication subscribers churn prediction. This study can't work on diverse data, which is considered as a drawback. .Its future work is test the approach on bigger data sets containing data over a longer period of time.[6] This work made use of four different rule generation algorithms (i.e. Exhaustive, genetic ,covering and LEM2 to predict customer churn in telecom industry with and without counter. The most fou problem of which classification technique could use to tactic the churn prediction in a more appropriate manner ,remains an open exploration problem. At the same time, the black-box model generated by SVM is also considered as one of its main drawback. It is necessary to increase the number of home appliances to be controlled.

Existing System: Studies shows that acquiring new customers is about 5 to 10 times expensive than retaining their existing customers and moreover keeping the customers loyal in today's competitive conditions has become priority for any organizations , according to reports an average business loses around 25-35% of their customers every year. Many companies, realizing this situation,are strongly focused in satisfying and retaining their customers in order to prevent churn. Particularly in the subscription oriented businesses, such as telecommunications, banking sectors, insurance companies, and in general in any particular field where customer relationship management(CRM) is crucial for the organization .The revenue generated and overall profits of the companies are provided by the payments/investments made by the customers periodically. Therefore the need of hour is to be able to keep customers gratified in order to be able to sustain this profits and revenue with the least expenses and minimize loss.

Disadvantage: In today's technological conditions, large volume of data is being produced from different sources in various sector .It is very important the data extracted

for large chunks of data repositories is pre-processed properly because the useful information hidden in these data sets can't be put into use, unless they are processed properly. In order to find out this hidden information and features, data science comes in handy for information extraction using several data mining methods and machine learning algorithms.We review the existing works on churn prediction in three different perspectives: datasets, methods, and metrics for banking sectors. Initially, we present the particulars about the availability of public datasets and customer details available in each dataset for predicting customer churn. Secondly, we compare and contrast the various predictive modeling methods that have been used in the literature for predicting the churners using different categories of customer records, and then quantitatively compare their performances. Finally, we summarize what kinds of performance metrics have been used to evaluate the existing churn prediction methods. Analyzing all these three perspectives is very crucial for developing a more efficient churn prediction system for telecom industries.

Advantage:reviewing the relevant studies about customer churn analysis observed in the telecom industry presented in last five years, particularly in the last two years, and introducing these up-to-date studies in the literature, Determining the data mining methods frequently used in churn implementations, Shedding a light on strategies that can be utilized in future works.

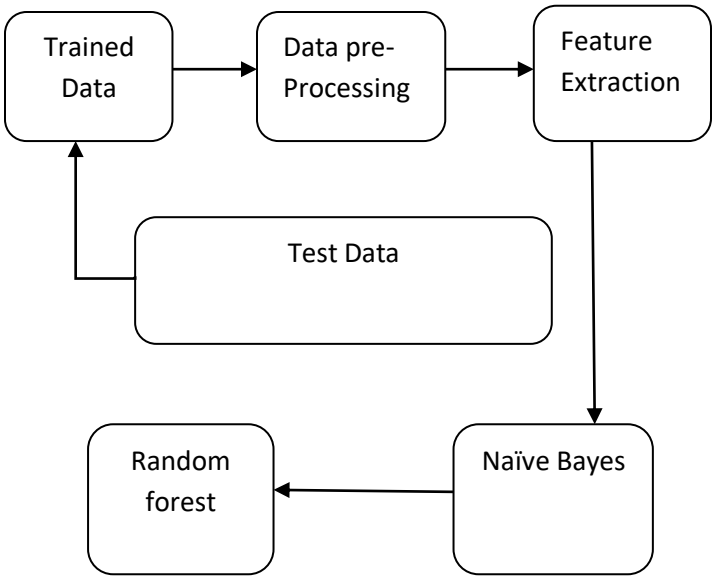


Fig 1.2 flowchart

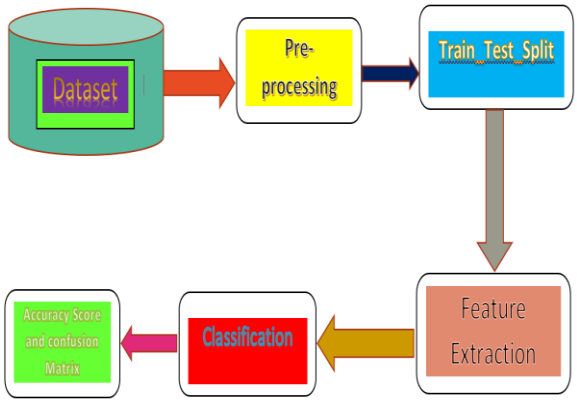


Fig 1.3 system architecture

Predicting Customer churn is a business scenario in which a company is trying to retain a customer which is more likely to

leave the services. For reducing churn rate, we have to classify which customers are most probably going to churn and which will not. Also we have some data to train our model which makes our problem as Supervised Classification problem. EDA It includes looking into the data analyzing various variables, visualization, missing value analysis, correlation analysis, chi-square test, scaling of features, Sampling. Basic Modeling Will try different machine models over preprocessed data (Random forest, Naïve Bayes). Model Evaluation & Optimization Evaluating model performances, select the best model fit for our data, optimizing hyper parameters tuning, Cost effectiveness of model. Implementation model on Final test data and to visualize the result.

Machine learning Algorithms: Customer churn prediction is a binary classification problem. Here we have to build a model which can classify if a customer will move (churn out) or not. So for deal with particular problem we will use an Classification Model here. There are lots of classification model present in the market. Here we will test four particular algorithms on our train data.

Random Forest;Random Forest build multiple decision trees and merge them together to get a more accurate and stable prediction.Random forest is a kind of supervised machine learning algorithm based on collective learning .Collective learning is a type of learning where you join different types of algorithms or same algorithm numerous times to form more powerfulpredictionmodel. The random fores t algorithm combines many similar algorith ms multiple decision trees, resulting in a tree s forest, hence the name Random Forest.For both regression & classification tasks, the ra ndom forest algorithm can be used.

This algorithm is not prejudiced or biased, as, it consists of multiple decision trees and each tree is basically undergoes training under some sort of data. In other words, this supervised learning algorithm relies on the vote of all the trees and thus features with maximum votes is selected and thereby reducing the overall biasness of the system.This algorithm is quite sturdy and robust , i.e Even in case some new or unknown data point is introduced in the dataset the overall efficiency and performance of the algorithm is less affected since newly introduced sample might impact not more than 1 or 2 trees.

This algorithm works fairly well in the case of both categorical as well as numerical data.Less affected by noisy data or missing values i.e null entities. Making it quite suitable for prediction in this scenario.

Naive Bayes:The Naive Bayesian classifier is based on Bayes' theorem based on conditional probability. This model is quite easy to deploy and build, as it does not require any complicated parameter estimation iteratively and thus making it fairly convenient for larger data samples. Despite its simplicity, the Naive Bayesian classifier often does predicts unknown class labels surprisingly well compared to even some of the most sophisticated classification algorithms .A classifier built on this model performs comparatively better than

other models like logistic and linear regression and requires less time in training data and testing the model.It performance is better in the case of categorical data as compared to numerical data.

This model makes it easy to predict class labels in a very short time with good accuracy. When multiple class prediction is required it performs well in that case as well.Another restriction of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are totally independent.

Supprt Vector Machine: This is a kind of classification algorithm, where the independent data and dependent data points are basically are separated by a line or a hyperplane depending upon whether the SVM is linear or non linear. The separation line or plane is selected such that; the two sides formed from the division by the line or plane makes 2 classes.When an unknown tuple/data comes its task is to correctly predict and identify which side/class of the line it belongs. The margin between the hyperplane and the support is kept as large as it can be possible In order to reduce the errors and improve the overall efficiency.

Dataset analysis The data set is obtained from Kaggle Website, is used in this paper for the churn analysis and prediction. This data set comprises 18 attributes and 7044 records or tuples. The bar graph (fig1.4) and pie chart (fig1.5) respectively shown below gives the number of customers that will churn.

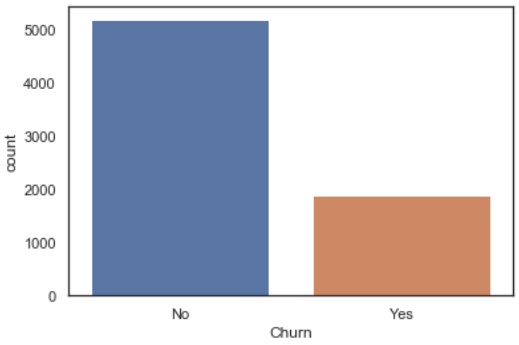


Fig-1.4 count vs churn

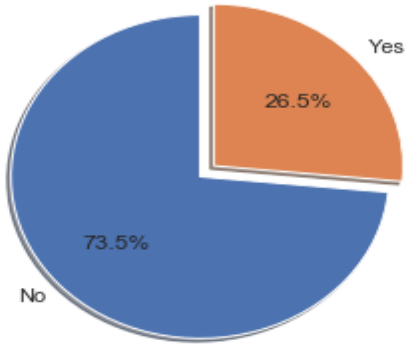
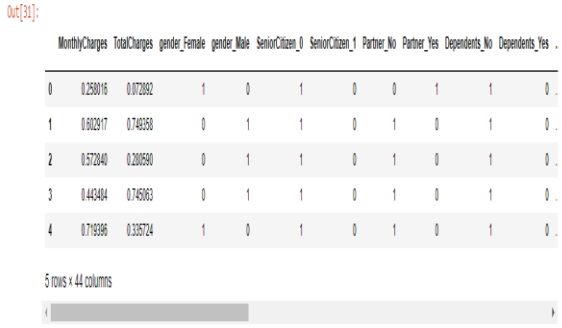


Fig-1.5 pie chart

The categorical data values are converted into numerical values in order to make the classification task more convenient and efficient.


Out[31]:

	MonthlyCharges	TotalCharges	gender_Female	gender_Male	SeniorCitizen_0	SeniorCitizen_1	Partner_No	Partner_Yes	Dependents_No	Dependents_Yes	.
0	0.258016	0.072892	1	0	1	0	0	1	1	0	.
1	0.602917	0.749358	0	1	1	0	1	0	1	0	.
2	0.572840	0.280590	0	1	1	0	1	0	1	0	.
3	0.443484	0.745963	0	1	1	0	1	0	1	0	.
4	0.719396	0.335724	1	0	1	0	1	0	1	0	.

5 rows × 44 columns

Fig-1.6 catagorical to numerical .

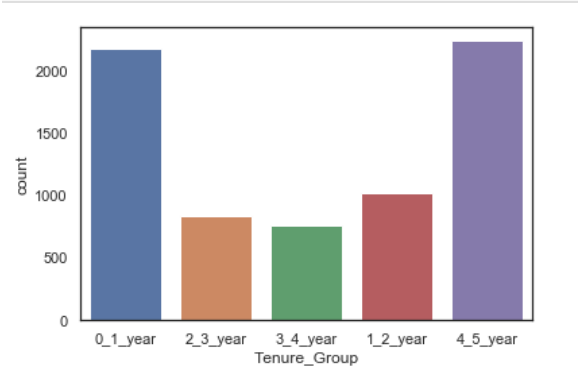


Fig-1.7 tenure_group vs count

This bar chart fig(1.7) demonstrates the comparision of the services that are being provided by a particular company in the particular years (tenure_group).The no. of services (Count) being provided by the company vary from year to year tenure_group).By this comparison we can understand what is lagging and predict the churn by the means of changes in the amount of services that are being provided by the company to the customers. Performance measures-the figure shown below gives performance measure of classification algorithm used

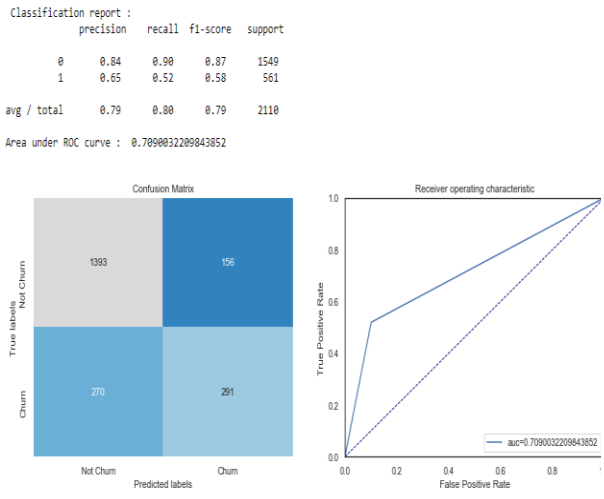


Fig 1.8 classification performance

Factors considered in measuring performance are accuracy , sensitivity, specificity and precision .accuracy in the sense how good the classifier is performing sensitivity means how perfect the classifier is with respect to positive entries. There are 2 kinds tuples ,positive and negative tuples. positive tuples obey some specific rules whereas negative tuples do not . 4 factors are taken into consideration .**true positive**(TP) tuples under consideration is perfectly positively classified i.e expected and observed tuple were positive.**True negative(TN)**:if tuple was observed negative and expectation was same .**False positive(FP):** data tuples were mistakenly classified as positive however expected outcome was the opposite.**False negative(FN)** were actually supposed to be classified as positive but were shown to be negative.We can calculate the performance measures of the created classifiers by using the following equations representing the relationship b/w the tuples wrt to the performance measuring factors.

Accuracy=TP+TN/P+N ----------(1)

Sensitivity=TP/P----------(2)

Specificity=TN/N----------(3)

Precision=TP/TP+FP----------(4)

Confusion matrix:simple tool to check whether tuples belonging to a class are perfectly classified.

conclusionThis paper find out which machine learning model is efficient and performs well in order to predict and analyse on the customer data with the help of the

predictive algorithm and is used to find whether the customer will churn or not. it got better prediction result with each algorithms .Finally, we summarize what kinds of performance metrics have been used to evaluate the existing churn prediction methods. This study might help telecom companies what are the factors that causes churn in customers and can take the necessary steps to minimize that.

REFERENCES

[1] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction" , 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017

[2] Jas Semrl, Alexandru Matei. "Churn prediction model for effective gym customer retention" , 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), 2017

[3] zhenyu chen ; zhiping fan(2011) building comprehensible customer churn prediction model.

[4] Preeti K. Dalvi; Siddhi K.Khandge ;Ashish Deomore; Aditya Bankar ;V. A. Kanade (2016) Analysis of customer churn prediction in telecom industry

[5] Saad Ahmed Qureshi ; Ammar Saleem Rehman ; Ali Mustafa Qamar; Aatif Kamal(2013) . Telecommunication subscribers' churn prediction model using machine learning

[6] Adnan Amin ;Changez Khan ;Imtiaz Ali ;Sajid Anwar(2014) Customer Churn Prediction in Telecommunication Industry: With and without Counter

[7] Alpaydin, E. (2010). *Introduction to Machine Learning.* London, England: The MIT Press.

[VIII] Archaux, C., Laanaya, H., Martin, A., & Khenchaf, A. (2004). An SVM based Churn Detector in Prepaid Mobile Telephony. *IEEE* .

[8] Berson, A., Smith, S., & Thearling, K. (2000). *Building Data Mining Applications for CRM.*

New York, NY: McGraw-Hill.

[9] Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th international conference on World Wide Web*, (pp. 107-117). Brisbane, Australia.

[10] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction.

Elsevier , 4626-4636.

[11] Cetnik, B. (1990). Estimating Probabilities: A crucial task in machine learning. *Ninth European Conference on Artificial Intelligence*, (pp. 147-149).

London.*CRISP-DM - Process Model*. (n.d.). Retrieved April 29, 2011, from CRISP-DM - Home:www.crisp-dm.org

[12] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukheejea, S., & Nanavati, A. A. (2008). Social ties and their relevance to churn in mobile telecom networks. *EDBT '08 Proceedings of the 11th international conference on Extending database technology: Advances in database technology* (pp. 668-677). Nantes, France: ACM.

[13] Dash, M., & Liu, H. (1997). Feature Selection for Classification. *Inrelligent Data Analysis 1* , 131-156.

[14] A Rawat and A Choubey,"A survey on classification techniques in internet environment," IJSRSET,vol.2,no.3,2016

[15] Customer prediction in mobile telecom system using data mining techniques,DR M.balasubramanium,M.selvarani, international journal of scientific and research publications , volume 4,issue 4, April 2014.

[16] D. Nguyen, N. A. Smith, and C. P. Rose. Author age prediction from text using linear regression. *Optimum Learning Rate for Classification Problem with MLP in Data MiningLaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011.

[17] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane. Com-paring data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers,* 11(4):449–460, September 2009.

[18] L. Rokach and O. Maimon. Data mining with decision trees: theory and applications. *World Scientific Pub Co Inc*, 2008.

[19] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *32nd Annual Conference of the Gesellschaft Fur*, 1980.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[21] W. Y. Loh and Y. S. Shih. Split Selection Methods for Classification Trees. *Statistica Sinica*, 1997.

[22] L. J. S. M. Alberts. Churn prediction in the mobile telecommunications industry, chapter 2. Master's thesis, Maastricht University, 2006.

[23] V. Lazarov and M. Capota. Churn prediction, Technische Universitat¨ Munchen¨. *Eighth ACM SIGKDD International Conference*, 2007.

[24] T. Mutanen, S. Nousiainen, and J. Ahola. Customer churn prediction – A case study in Retail Banking. *2010 conference on Data Mining for Business Applications, Amsterdam, Netherlands*, 2010.

[25] V. Umayaparvathi and K. Iyakutti. Applications of data mining tech-niques in telecom churn prediction. *International Journal of Computer Applications*, 2012.