International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Prediction of Customer Churn on e-Retailing

M Jaeyalakshmi, S Gnanavel, K S Guhapriya, S Harshini Phriyaa, K Kavya Sree

Abstract*: The technology has always been an instigating factor in progress for human civilization which resulted in driving the customer services to a greater need. The enrichment of technology has amplified and embellished the customer interaction among various business to consumer sectors. These technological upgrading have a huge impact on the retail industry which is an ever-growing market with key competitors around the world. In a consortium of multiple competitors in the same business, the re-engagement of disinterested customers is essential rather than winning a new customer. The sustenance of a customer can be figure out by Churn Prediction. Churn prediction is a new promising method in customer relationship management to analyze customer retention in subscription-based business. It is the activity of identifying customer with a high probability to discontinue the company based on analyzing their past data and behavior. It looks at what kind of customer data are typically used, do some analysis of the features chosen, and initiate a churn prediction model. Thus, churn prediction is a valuable approach in identifying and profiling the customers at risk.

Keywords: Data Mining, Machine Learning, Churn Prediction, Customer Retention

I. INTRODUCTION

In the digital world, people want to purchase easily, quickly and within budget. In order to save time on shopping, they prefer online market where they can find a huge variety of products at an affordable price that can be indemnified through web and obtained at door step. Marketing ccampaign invite people to online shopping. In recent times the number of retailers in online is being increasing.

Satisfying the customer is a tedious thing, as they want best and reasonable product.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.

* Correspondence Author

Jaeyalakshmi M*, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: jaeyalakshmi.m@rajalakshmi.edu.in

Dr. S Gnanavel, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: gnanavel.s@rajalakshmi.edu.in

Guhapriya K S, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: guhaseran@gmail.com

Harshini Phriyaa S, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: harshinisns@gmail.com

Kavya Sree K, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India Email: kavyapooja1418@gmail.com

In order to get favorable product, there is a chance of switching to other online sites from classic, which leads to loss of customer. If this continues for a certain period of time it leads to customer churn. Retailing gives an approach to products to find a good pace to customer. It endures snags like investment on labour, gratifying the bargaining customer. e-retail provide muddle

free shopping to buyers, as having various preferences. To stay in super competitive online market, retailers should undergo proceedings like verifying the identity of customer, loyal and transparent to customers, follow return and refund policies, securing customer data. A customer likely to break the relationship or dwindle the purchase rate is known as churn.Customer churn occurs when a customer stops employing a retailer's product, stops visiting a specific place of business, shift to lower-tier experience or shift to the contender's products. Retailers need an abiding strategy to manage customer churn. Measuring the churn rate is kind of crucial for retail businesses because the metric reflects customer response towards the merchandise, service, price and competition.

Churn prediction envision the likelihood of customer to churn. It pares the investment on gaining new customer and helps to retain the existing customer. The marketing efforts and amount spends on attracting a new customer is high and difficult than clinging to existing customer. Customers who are unlikely to make a purchase or willing to shift the shopping site because of cautiousness with money, expecting standard and assortment in products can be convinced and clutched. The customers who are ending the relationship due to valuable and unavoidable reasons are free to leave. Result is firm, though we invest on involuntary churners.

Target marketing aids to reach the customers and connect with them. The voluntary churners can be stopped by extending discounts, amending the products to customers choice and by sending out trigger mails. Concentrating only on voluntary churners will scale down the cost of offering benefits to yet and all churned customers. Predicting the customer churn with unstructured data leads to poor results and causes class imbalance problem. Huge difference in the percentage of churn and non-churn of historical data defines class imbalance problem, to avoid these problems customer data should be preprocessed, and prediction is done with necessary attributes which is feature selection. This reduce the time and cost on prediction.

Customer attrition can be done using different classification and predictive model. Efficient algorithms based on its accuracy are subjected to soft voting, that elect ensemble model as best to follow in impending works. The paper is organized as follows with related works in Section II. The working methodology has been explained in Section III.

In section IV, the experimental results and analysis are discussed, and the conclusion and future work are discussed in section V.

Prediction of Customer Churn in e-retailing

II. RELATED WORKS

Abinash Mishra and U. Srinivasulu Reddy [1] performed a comparative study on ensemble methods. The performance metrics calculated here are accuracy, error rate, specificity (portion of negative cases that were classified correctly) and sensitivity (portion of positive cases that correctly identified). The result show that Random Forest algorithm performed better with 91.66% low error rate of less specificity 53.54 and high sensitivity 98.89 than others like Bagging, Boosting and Decision tree.

Pretam Jayaswal, Bakshi Rohit Prasad, Divya Tomar, and Sonali Agarwal [11] selected Bagging and Boosting based ensembles classifiers like Random Forest, Gradient Boost Trees (GBT) and Decision tree classifier where dataset is divided in training (75%) and testing (25%) subsets using Apache Spark. This work shows that GBT outperformed other methods in terms of accuracy of 96.78% and specificity. They also employed optimization phase that made the results more accurate and refined.

Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr [13] proposed a model that predicts churners into 3 categories in using a retention strategy. They have used decision tree, Support vector machine and Neural Network through an open source software called WEKA. The data set is divided into training set of 80% instances and the testing set of 20% instances. The training data contains 80% instances are categorized as non-churners and others are 20% are categorized as churners. In this paper, both SVM and Neural networks showed the same results in terms of accuracy and error rates.

Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Sail Ul Islam, Sung Won Kim [8] proposed a hybrid model using Random Forest, Decision Stump, J48 and Random Tree with 10-fold cross-validation for classification of churners and non-churners and used k-means for clustering. The proposed model targets churn customers and distinguish the explanations for their relocation. The experimental result 88.63% correct classification through RF. In this study, for factor identification, a comparable classifier such as Attribute Selected Classifier is used for rule generation that can be easily visualized since RF is not appropriate for rule generation for factor identification as it generates complex forest which is difficult to visualize and rule inference.

Jaya Kawale, Aditya Pal, Jaideep Srivastava [9] proposed an updated diffusion model in online role-playing games (MMORPG) where they considered two aspects-decrease in player engagement of churners over time until they churn and increase in churn prosperity with increase in number of churner neighbors. Using diffusion model having two valued influence vectors in modelling accurately- negative influence and positive influence and a parameter spread factor where the portion of influence transferred to his network. Their technique outperformed both diffusion model and network and engagement feature based classification in their dataset. This scheme is able to capture social influence and player engagement effectively and resulted in a consistent prediction accuracy. However, they did not compare their approach with classification based on features that might not capture both important aspects mentioned.

The main objective of business domain is to maximize margin of profit. This can be maximized by increasing sales and preventing churn.

Churn prediction is one among the major requirements of the present competitive environment. Everything relies on data being generated. Using those data industries are now capable of predicting customer churn and take proactive steps to forestall it.

IV. METHODOLOGY

A. Tools

The tools used are Anaconda, Jupyter notebook. Anaconda is used for scientific computing like machine learning application, data analytics etc. It uses anaconda navigator as a graphical alternative to the command line interface. The Jupyter Notebook is online application that can be used to create and run live code, equations, visualizations, and text.

B. Dataset Description

A dataset of e-commerce retail store is used for predicting customers churn. It contains 10000 row and 9 columns, where each row constitutes a customer data and each column constitutes a single attribute.

C. Module description

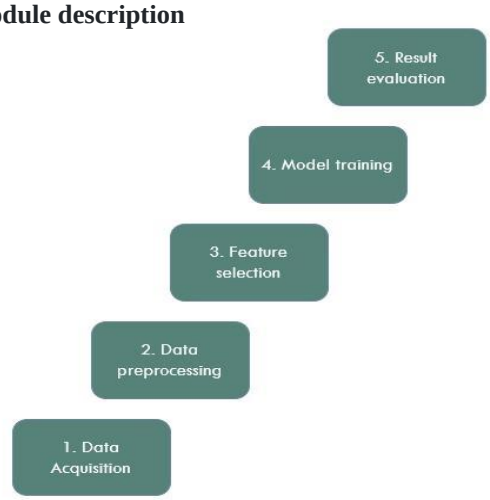


Fig 1: churn prediction modules

The paper deals with identifying and predicting churn in the retail data. The different phases of churn predictive system proposed as: Scrutinizing the customer dataset that consist of present and past records, Pre-processing the input customer records which includes cleaning and handling the missing data, Extracting the desired features for developing churn model, Construction of model using different classifiers and cross validate the model, Calculating the prediction accuracy and variableimportance, Furnish with customer retention strategies.

D. Proposed work

The walkthrough of the process followed to predict the customer churn:

We collate the required market data on online retail shop and construct a comma-separated value (CSV) file. Import all the necessary libraries we would need to proceed further in our code.

Once the data is read, some pre-processing needed to be done to test for null and outliers. Exploring the raw dataset by defining the function to visualize the feature with missing values and also the percentage of total value and data type.

We can remove the additional whitespace using strip function. Perform EDA (exploratory data analysis) to grasp how the data and its features are interrelated & correlated, evaluate presence of outliers and its effects. By using various box plots, kde plot and bar plots helps to understand what are the features which majorly impact on our target variable.

Perform K-S test on each feature to check whether the distributions of each feature of churn or not are drawn from the identical distribution.

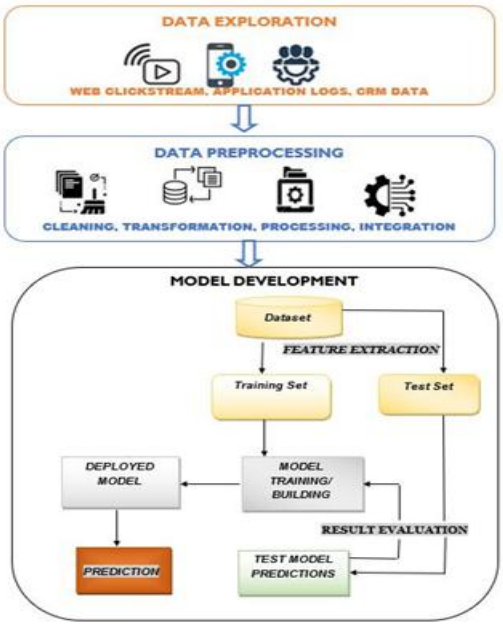


Fig 2: Architecture Diagram

Explore and visualize each feature to have an idea about how each feature behaves for churned/stayed customers. Check the correlation between features and with the target variable by the linear correlations in a heatmap.

The heatmap below represents the collinearity of the multiple variables in the dataset. Dark red area in heatmap constitutes a positive correlation, while light white is a negative correlation. It is also normal that the darkest areas are a 1:1 ratio since Avg Session Length= Avg Session Length, Time on App=Time on App, etc. By adjusting the color and adding annotation, makes it easier to derive a conclusion and what possible actions can be taken.

Once the pre-processing is done, the subsequent step is to induce the relevant features to use in our model for the prediction. A model cannot take categorical data as input; hence the features are encoded into numbers to be used in prediction. Take the features which has more influence on churn prediction. The data is then scaled and split it into training and test data set. This gets easier to make judgement as you train and build more and more models. The dataset is split to an 80–20 ratio.

The classifier algorithms are fitted to the new scaled data. Compare the popular best classifiers and evaluate their performance using a stratified kfold cross validation procedure. The number of folds is 5.

Perform some hyperparameter tuning on each model to choose the most promising model. We decide to choose: SVC, XGBoost, Gradient Boosting, Random Forest, KNN and logistic Regression for further fine-tuning. Generate a simple plot for test and training learning curve. Learning curves are a best way to diagnose the overfitting effect on the training set and the effect of the training size on the accuracy.

Gradient Boosting, KNN, XGBoost tend to overfit the training set. According to the growing cross-validation curves KNN and Gradient Boosting could perform better with more training examples. SVC, Logistic Regression and Random Forest classifiers seem to better generalize the prediction since the training and cross-validation curves are close together.

Calculate the confusion matrix to find the Precision, Recall and Accuracy.

Heatmap makes it easy to identify which features are most related to the target variable, we will concatenate all classifier and plot heatmap of correlated features using the seaborn library.

The prediction seems to be quite similar for the 6 classifiers except when Logistic Regression is compared to the others classifiers. The 6 classifiers give more or less the same prediction but there are some differences.

V. RESULTS

A. Snapshot of results

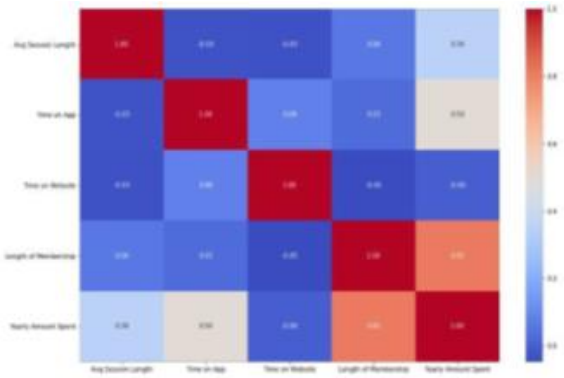


Fig 3 : Heatmap of features

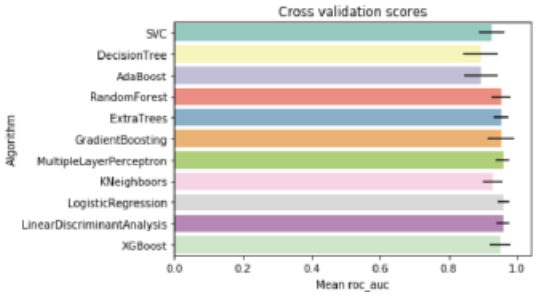


Fig 4 : Cross Validation Scores

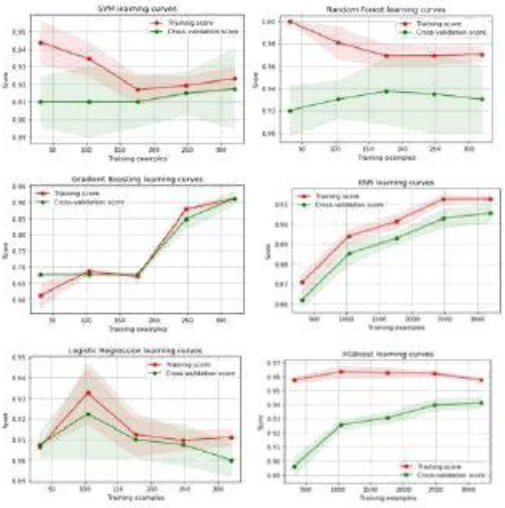


Fig 5 : Learning Curve

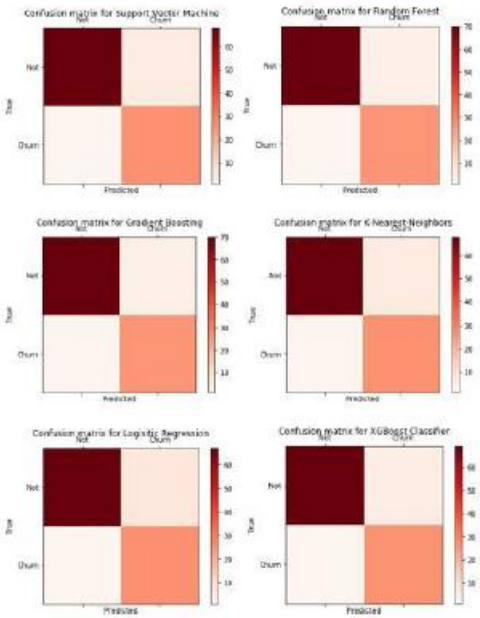
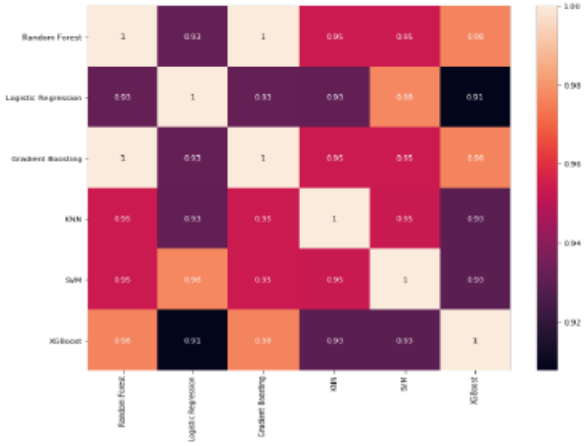


Fig 6 : Confusion Matrix

Fig 7 : Heatmap of classifiers



B. **Comparison of algorithms and prediction Results**

Classifier name	Accuracy	Precision	Recall	F1 Score	ROC AUC
Support Vector Machine	0.94	0.8387	0.9629	0.8965	0.9619
Random Forest	0.96	0.8965	0.9629	0.9285	0.9545
K Nearest Neighbors	0.94	0.8387	0.9629	0.8965	0.9596
Logistic Regression	0.93	0.8125	0.9625	0.8813	0.9624
Gradient Boosting	0.96	0.8965	0.9629	0.9285	0.9685
XGBoost	0.95	0.8666	0.9629	0.9122	0.96321

VI. CONCLUSION

In existing model predictions done with unstructured or semi-structured data. Related works stated results with one particular machine learning algorithm. A comparative study on various algorithms come out with best model based on accuracy. Accuracy of the algorithm varies depend on customer data. This project proposed churn prediction in e-retail with structured data, imperative features, machine learning techniques. On cross validation better algorithms are picked and elected based on voting. Ensemble algorithm is the closing result and suggest appropriate preventive measures on churners.

In future we'll enhance the customers who are predicted as churn are again analyzed based on their reason of churn. Offers and benefits are given only to voluntary churners, which leads to target specific marketing. The churn prediction will be implemented in mobile application. The application will automatically send notification to churned customers about the offers and benefits. It will remind the customer to do shopping and inform them about updates. It will serve the marketers as user friendly and cost efficient to predict their churn with the data and to proactively take retention actions on their customers.

REFERENCE

1. Abinash Mishra and U. Srinivasulu Reddy. “A Comparative Study of
2. Customer Churn Prediction in Telecom Industry Using Ensmeble Based
3. Classifiers”, Proceedings of the International Conference on Incentive
4. Computing and Informatics (ICICI), 10.1109/ICICI.2017.8365230, 2017.
5. Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, andHossamFaris. “Hybrid Data Mining Models for Predicting CustomerChurn”. International Journal of Communications, Network and SystemSciences, vol. 8, no. 05, pp. 91, 2015.
6. Annapurna P Patil, Deepshika M P, Shantam Mittal, Savita Shetty, Samarth S Hiremath, Yogesh E Patil, “Customer churn prediction for retailbusiness”, International Conference on Energy, Communication, DataAnalytics and Soft Computing (ICECDS),10.1109/ICECDS.2017.8389557,2017.
7. Chih-Fong Tsai and Yu-Hsin Lu., “Customer churn prediction by hybridneural networks”. Expert Systems with Applications, vol. 36, No. 10, pp.12547-12553, 2009.
8. Guoen Xia, Hui Wang, Yilin Jiang, “Applications of customer churnprediction based on weighted selective ensemble”, 3rdInternational conference on systems and informatics (ICSAI), 2016.
9. Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholie, “Predicting PotentialBanking Customer Churn using Apache Spark ML and MLlib Packages: AComparative Study”, (IJACSA) nternational Journal of Advanced Computer Science and Applications, Vol. 9, No. 11, 2018.
10. Ibrahim Onuralp Yigit and Hamed Shourabizadeh. “An approach forPredicting Employee Churn Using Data Mining”, 978-1-5386-1880-6/172017 IEEE.
11. Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, SailUl Islam, Sung Won Kim, “A Churn Prediction Model Using RandomForest: Analysis of Machine Learning Techniques for Churn Prediction andFactor Identification I Telecom Sector”, Digital Object Identifier 10.1109/ACCESS.2019.2914999.
12. Jaya Kawale, Aditya Pal, Jaideep Srivastava, “Churn prediction in MMORPGs: A social influence-based approach”, Conference paper 2009 DOI 10.1109/CSE.2009.80.
13. Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, and V. A. Kanade. “Analysis of customer churn prediction in telecom industry using decision trees and logistic regression.” In Symposium onColossal Data Analysis and Networking (CDAN), pp. 1-4. IEEE, 2016.
14. Pretam Jayaswal, Bakshi Rohit Prasad, Divya Tomar and SonaliAgarwal, “An Ensemble Approach for Efficient Churn Prediction inTelecom Industry”, International Journal of Database Theory and Application, Vol. 9, No.8, pp. 211-232, 2016.
15. Xiaojun Wu, Sufang Meng, “E-commerce customer churn prediction based on improved SMOTE and AdaBoost”, 13th International Conference on Service Systems and Service Management (ICSSSM),10.1109/ICSSSM.2016.7538581,2016.
16. Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr, “A Proposed Churn Prediction Model”, International Journal of Engineering Research and Applications (IJERA): Volume 2, Issue 4, pp.693-697, 2012.
17. Qiu Yihui and Zhang Chiyu, “Research of indicator in customer churn prediction for telecommunication industry”, 2016 11th International Conference on Computer Science and Education (ICCSE), pp. 123-130, IEEE, 2016.
18. Qiuhua Shen, Hong Li, Qin Liao, Wei Zhang, and KoneKalailou, “Improving churn prediction in multilayer features based on factorization and construction”, 26th Chinese Control and Decision Conference (2014 CCDC), pp. 2250-2255, IEEE, 2014.
19. Jas Semrl, Alexandru Matei, “Churn prediction model for effective gym customer retention”, 2017 International ConferenceonBehavioral,Economic, Socio-cultural Computing (BESC), 2017.
20. Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, “Evaluation of machine learning models for employee churn prediction”, 2017 International Conference on Inventive Computing and Informatics