International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Customers Churn Prediction with RFM Model and Building a Recommendation System using Semi-Supervised Learning in Retail Sector

Punya P Shetty, Varsha C M, Varsha D Vadone, Shalini Sarode, Pradeep Kumar D

Abstract: ***Customer churn or customer attrition occurs when certain customers are no longer loyal to a firm. In retail businesses, the event of churn is said to occur, if a customer's transactions terminates after a certain duration. High churn rates incur humungous losses for the businesses as it is observed that acquiring new buyers is costlier than retaining the current customer base. Hence, for calculating customer churn of companies, they should be able to monitor churn rates. These churn rates give an organization various factors to be considered to determine their customer retention success rates and identify strategies for improvement. Customer churn is predicted using Pareto/NBD model. Once the customers who are likely to churn are predicted, they need to be differentiated based on their previous purchasing history. Natural Language Processing is used to model product categorization. Semi- supervised learning does customer segmentation. This consists of assigning a score by RFM model and segmenting using k-means clustering. The prediction of clusters is then done using algorithms like logistic regression, SVM and SGD classifier. These methods are collectively used to build a suitable recommendation system, which is targeted to make the churn customers who were valuable to the company loyal again, thereby improving the business for retailers.***

Index Terms: ***NLP (Natural language processing), Pareto NBD(Pareto negative binomial distribution), RFM(Recency-Frequency-Monetary), SVM (Support vector machine), SGD (Stochastic Gradient Descent)***

I. INTRODUCTION

A customer is said to be churned in the retail sector when he stops making purchases from the company since a particular amount of time. Deciding if the customer has churned is an important task to the company. Retaining a customer is cheaper than acquiring a new one. Hence prediction of customer churn is essential for retail business.

Revised Manuscript Received on 30 May 2019.

***** Correspondence Author

Mr. Pradeep Kumar D, Assistant Professor in the Department of Computer Science and Engineering in Ramaiah Institute of Technology.

Miss Shalini Sarode Department of the Computer Science Engineering program at the Ramaiah Institute of Technology.

Miss Varsha D Vadone Department of the Computer Science Engineering program at the Ramaiah Institute of Technology.

Miss Varsha CM Department of the Computer Science Engineering program at the Ramaiah Institute of Technology.

Miss Punya Prakash Shetty Department of the Computer Science Engineering program at the Ramaiah Institute of Technology.

When the customer is predicted to churn in future, A recommendation system helps giving solutions to retain the customer from churning. The recommendation system puts all the customers in different clusters based on their past transaction history. Recommendation to retain customers is given based on the cluster the customer belongs to.

A. Objectives

The principal objectives of the paper are as follows:

- Preprocessing and Exploratory Data Analysis
- Predict if a particular customer will churn.
- Product categorization using NLP
- Using product categorization and RFM to cluster every customers
- Cluster analysis
- To analyze the results of each of the algorithms and compare their accuracies
- Predict the cluster that the churn customer belongs to.
- Give recommendation to prevent churn based on the cluster the customer belongs to.

B. Deliverables

This project will predict if the customer will churn and will offer a business strategy to the owner to keep the customers from churning. This will help in increase in the sales of the retail business.

Deliverables of the project are:

- To predict if the customer will churn using Pareto/NBD model
- Cluster the customers based on previous purchases
- Product categorization
- Recommendation based on the cluster the churn customer belongs to.

Customer churn is predicted using RMF model. Recency, Frequency and Monetary features are calculated for each customer. Using these features the RMF model will predict if the customer will churn in the future. Predicting the customer churn will help the retailers take necessary measures to prevent the customers from churning.

Recommendation System is given to the retails to help retain customers when they are expected to churn. The customers are initially divided into clusters based on their past transactions using semi-supervised learning methods. Recommendation for each customer is given based on the cluster they belong. The solution to reduce customer churn is different for each cluster.

II. LITERATURE SURVEY

Retailers are aware that a few customers' regular purchasers while others discontinue shopping at their store. Nonetheless, to predict customers that have decided to shop in a different store from the ones that is simply idle. Customer churn is the propensity of customers to cease doing business with a company in a given time period.

Renjith [1] has observed that it is significantly more profitable to retain the existing customers rather than obtaining new ones. Hence the paper has proposed an integrated platform to first identify customers who might churn, in advance. This is done through predictive analysis-namely logistic regression which takes various factors such as customer demography, usage metrics, purchase patterns etc. Once these churn rates are determined, customer profiling is done based on their traits such as behavioral and social graphics using k-means clustering. Finally, a customer retention strategy is developed to provide timely personalized recommendations to those users who are most likely to churn.

Parikh et al. [2] has analyzed the techniques to implement a recommendation system for e-commerce site. The paper suggests the use association rule mining to provide recommendation based on customer profile, social networking site profile and based on their cart. However, it is observed that this technique is computationally infeasible for real-time systems with large databases. Therefore, it has proposed the use of a system with a combination of association rule mining as well as clustering methods.

Xiaojun Wu et al. [3] noted that in the usual scenario, the e-commerce customer churn datasets are imbalanced. It is found that the count of churn customers greatly exceeds nonchurn consumers. Due to this traditional algorithm favor the majority set and does not give accurate results. Hence, the paper proposes improved SMOTE technique to balance the data. It then uses AdaBoost on this data to classify the customers.

Shahriar Akter et.al [4] present an interpretive framework that explores the definitional aspects, distinctive characteristics, types, business value and challenges of Big Data Analytics in the e-commerce landscape. They illustrate the business value of big data in e-commerce and provide guidelines for tackling the challenges of big data application within e-commerce. The paper aims to provide a general taxonomy to broaden the understanding of Big Data Analytics and its role in creating business value.

Qiu Yanfang et.al [5] use a logistic regression model to establish an e-commerce user churn prediction model through preliminary research on e-commerce customer churn behavior. Different metrics like the user's online duration, number of logins, attentions were analyzed to determine the factors that result in loss of users and propose an efficient EBURM model to predict user churn behavior over a period of time in a high confidence level.

Adnan Amin et.al [6] propose an intelligent rule-based decision-making technique, based on rough set theory to extract important decision rules related to customer churn and non-churn. The paper provides an effective approach for classification of churn from non-churn customers, along with prediction of those customers who will churn possibly in future.

Shini Renjith[7] proposed an E-Commerce industry to dig deep into the customer base of their competitors and concurrently not let their current customers to churn. Retaining the customers is crucial for these enterprises as the cost incurred to acquire new customer base is huge due to high competition. Identification of customers who might churn and preventing the churn with immediate retention plans is needed to utilize resources efficiently. It is critical to also understand why the customer is leaving to develop better strategies. Data mining and machine learning can aid in evaluating this massive amount of data involved in E-commerce sites, to interpret the customer behavior and identify probable candidates who will part ways. This paper proposes a framework built on SVM to forecast E-Commerce customer churn and a novel recommendation strategy to recommend customized retention plans.

Joy et.al [8] have proposed a new method of segmenting the customers. The customers are segmented into groups based on Recency, Frequency and Monetary values of the customers. Segmentation helps in understanding the need of the customer and also increases the revenue. The customer churn can be calculated by finding the clusters of customers with similar needs fuzzy and k-means algorithms are used to find the clusters. The k-means algorithm is proved to better segmenting the customers.

LG Pee et.al [9] have proposed a new method for predicting customer loyalty. The perceived website usefulness is calculated to find how the longitudinal changes and satisfaction affect the customer satisfaction. Perceived Usefulness (PU) is very high during the first transaction and the loyalty on the future transactions depend on the customer satisfaction. When the satisfaction of the customer increases then the loyalty of the customer increases. PU helps find the longitudinal approach to retain customers and increase satisfaction. Retaining of customers is as important as attracting new customers. Hence these methods proposed helps greatly to sustain the business.

Niccolò Gordini et.al [10] has proposed a new method for churn prediction. A model based on SVMMauc is used to predict customer churn. The parameters are optimized and the marketing data that is noisy, nonlinear and imbalanced. This method outperforms the other methods like classic SVM, neural networks and logistic regression. Churn prediction is very important for the e-commerce business. Hence this model outperforms the other traditional methods and helps find the churn rate better.

Yogesh Patil et.al [11] have done a comparative study using three algorithms: Ensemble methods, Support vector machines and boosting to determine the factors affecting the costs incurred by the company in to attract new customer and also costs to retain the existing customers. They have inferred that reducing customer churn is a crucial business goal of every online business. The churn value is determined based on the customer's transactions.

IMPLEMENTATION

A. Tools

The tools used are Python Anaconda, Jupyter notebook, and AngularJS with HTML and CSS. Python Anaconda is a tool that is widely used by developers for data analytics. It has more that 1,400 packages this eliminates the needs to install every library separately. Jupyter notebook is an open source application where live code can be run. The front-end application is built using AngularJS with a REST API connecting the front end with the backend.

B. Dataset Description

Online retail dataset is taken from UCI machine learning repository. The dataset is a transactional data from 01/12/2010 to 09/12/2011 for a UK based online store. This store sells all occasion gifts. There are 541,909 transactional entries with 8 attributes.

C. Overall View of Implementation

Customer churn prediction and recommendation is done using an online retail data set. Since this data captures actual transactions of an e-commerce retail store it must be pre-processed before further analysis. The customers who may churn are predicted using a probabilistic model. The customers are further segmented into various clusters and the churn customers' clusters are determined. A recommendation is provided to the customers that belong to he clusters of most value.

D. Algorithms

K Means Clustering: K-means clustering is used to cluster data in data mining. This method divides the data into k clusters. Each observation is put in the cluster nearest to the cluster mean. This acts as a sample to the cluster. This algorithm is similar to k-nearest classifier.

Pareto/NBD: Pareto/NBD model that is used to predict future activity of customers. It simulates two events. It determines whether the customer will churn or not and determines the many times a customer will order. Past transaction data, recency and frequency of orders are used as input parameters to train the model. The output given by the model becomes more accurate with the increase in input data.

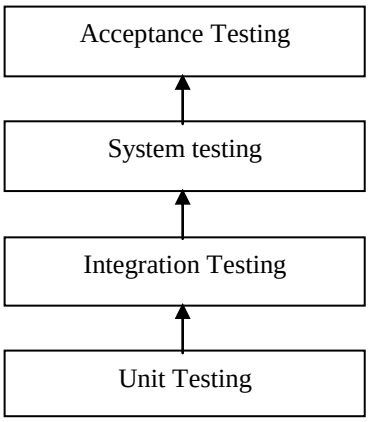
E. Modules

There are four main modules in the paper namely data preprocessing, predict churn customers, cluster the churn customers and recommendation.

Data preprocessing: The data is cleaned to be used for further analysis. The missing entries in the customerID field, duplicate entries, cancelled transactions that completely got cancelled and the invoices which do not relate to customers are removed. Predict churn customers: The customers are split into test and train data and the model is trained using Pareto/NBD. The summary of customer transaction log which is the Recency, Frequency and T is calculated for each customer. Recency is the time elapsed between the most recent transaction and the end of observation period. Frequency is the number of times the customer has bought items from the store. T is calculated to be the time elapsed between first transaction and end of calibration period. The model is fine-tuned by these parameters to get churn prediction. Cluster the churn customers: Product categorization is obtained using the models developed with the training data set. This will give us the specifications of the products by using NLP. Clustering of customers is done based on product categorization used above and RFM. The recency, frequency and monetary are calculated and the products with the lowest RFM values are considered. This reduces the number of clusters which are taken into consideration. The prediction set is plotted graphical to get a better visualization. The selected clusters are further analyzed to predict churn patterns. The clusters which have a high number of customers and are closely related to each other are chosen. Every customer is categorized into a particular cluster by using the above algorithms. Predicting which cluster the customer belongs to is the most important task to predict if churn will occur in the future and suggest churn prevention strategies. Recommendation: This module takes the predicted churn customers and their respective cluster as input. This information is collectively analyzed to give an appropriate recommendation. The churn customers who are of low value to the business are not given any recommendation. These recommendations are in the form of images of products that they are likely to buy. This can be combined with various business strategies like discounts to maximize the customer retention. The predicted churn customers with their respective clusters are sent to the front-end via an API call. The recommendations are shown in the dashboard which is used by the retailers.

III. TESTING

Testing involves the execution of a program or application to find any errors or bugs in it. It is a way of validating the application to ensure that it meets the technical requirements of the design. Each module is individually tested. After this, all the modules are integrated. The whole module is then again evaluated on various parameters such as accuracy. The different phases are split using various system designs like iterative and incremental model. After the implementation, the model is tested and unified. Once the functional and nonfunctional testing is done the product is released to the market.



A. Types of testing performed

Unit Testing: In this level of testing, individual units/ components of software are tested. Unit testing gives the programmers a thorough understanding of program's internal design. The main purpose is to validate that each unit of the software performs as designed. A unit is the smallest testable part of any software.

Integration Testing: In integration testing, all the units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units.

System Testing: After each component is tested and the entire program is integrated into a single working application, system testing is performed. The main purpose of this test is to evaluate the compatibility of the hardware with the specified requirements.

Acceptance Testing: This is a level of software testing where a system is tested for acceptability. The purpose of this test is to evaluate the system's compliance with the business requirements and assess whether it is acceptable for delivery.

IV. RESULT

A. Results Snapshots

```
trace = pm.sample(n_draws,chains=2, cores=1, init=None)

Sequential sampling (2 chains in 1 job)
NUTS: [mu, lambda, beta, s, alpha, r]
100%|██████████| 3000/3000 [08:03<00:00,  1.33it/s]
100%|██████████| 3000/3000 [08:05<00:00,  6.28it/s]
```

Fig. 2. Pareto/NBD model training

```
index = 255
# show purchasing behavior
des = rfm.iloc[index]
```

```
des
```

```
frequency_cal         1.0
recency_cal           3.0
T_cal                16.0
frequency_holdout     0.0
duration_holdout     25.0
Name: 12845.0, dtype: float64
```

Fig 3. Snapshot of calculation of duration holdout for a particular customer

```
lambda_post = trace['lambda']
mu_post = trace['mu']

# Select distributions of lambda and mu for a customer
lambda_individual = lambda_post[:,customer_index]
mu_individual = mu_post[:,customer_index]

# predict purchases for the user at t = 25
t = 25
predict(t, lambda_individual, mu_individual, x[customer_index],
        t_x[customer_index], T[customer_index]).mean()
```

```
0.799729087643
```

Fig 4. Prediction for the particular customer whether they will churn or not

Since this value is closer to 0, they are considered as a churn customer. This process is iterated over all test data and is used as an input for customer segmentation.

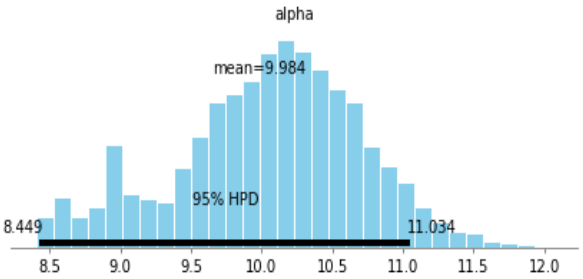


Fig 5. Performance of the model with alpha parameter.

CustomerID	min_recency	max_recency	frequency	monetary_value
12346.0	325.0	325.0	1	0.00
12347.0	2.0	367.0	7	4310.00
12348.0	75.0	358.0	4	1437.24
12349.0	18.0	18.0	1	1457.55
12350.0	310.0	310.0	1	294.40

Fig 6. RFM scores

Product clustering using NLP and k-means

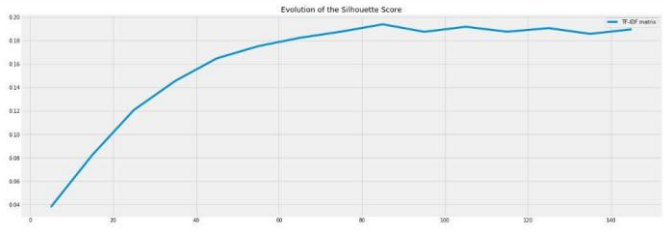


Fig 7. TFIDF matrix

The curve peaks at 135 clusters hence 135 is chosen as the number of clusters.

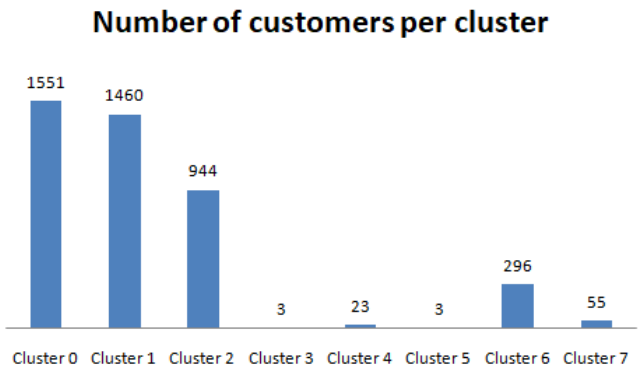


Fig 8. Clusters with different number of customers

B. Comparison results table

Algorithm	Accuracy
Baseline Accuracy	30.680
Logistic Regression classifier	95.155
SGD Classifier	93.656
Linear SVC	94.694

V.CONCLUSION

Various algorithms are compared and contrasted in predicting customer churn for a retail business is done and recommendation is given based on the cluster the customer belongs to. Different prediction algorithms were analyzed and studied and the best among them was chosen. Comparative study of each churn predictive algorithm was done and Pareto/NBD was chosen among them. The input parameters for the model was fine tuned to suit the needs of the model. The products are categorized, and the customers are put into a cluster based on their RMF scores. The recommendation is given based on the cluster they belong to.

The customer churn predicts if the customer will churn or not and gives recommendation to retain the churn customers. The paper can be extended to be included in the online e-commerce business models or a web application where the retailer can get a report of the churn customers along with the business plan to retain the churn customers.

APPENDIX

Variables used:
1. x: number of transactions observed during the calibration period;
2. T: observation period- the time between the first purchase of a customer and the end of the calibration period;
3. tx: time between the first and the last purchase in observation period (0, tx, T).
4. α, β, μ, r and λ are explained in [12].

REFERENCES

1. Renjith, Shini. (2015), "An Integrated Framework to Recommend Personalized Retention Actions to Control B2C E-Commerce Customer Churn", In International Journal of Engineering Trends and Technology (IJETT) – Volume 27 Issue 3 - September 2015.
2. Parikh, Vishal & Shah, Parth. (2015)."E-commerce Recommendation System using Association Rule Mining and Clustering". In Conference: International Journal of Innovations & Advancement in Computer Science, Volume: Volume 4, Special Issue.
3. Xiaojun Wu and Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost," (2016) 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, 2016, pp. 1-5.
4. Shahriar Akter, Samuel Fosso Wamba, "Big data analytics in Ecommerce: a systematic review and agenda for future research", The International Journal on Networked Business May 2016, Volume 26, Issue 2, pp 173–194.
5. Qiu Yanfang, Li Chen, "Research on E-commerce user churn prediction based on logistic regression", IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) 2017
6. Adnan Amina , Sajid Anwara , Awais Adnana , Muhammad Nawaza , Khalid Alawfib , Amir Hussainc , Kaizhu Huangd, "Customer Churn Prediction in Telecommunication Sector using Rough Set Approach", Neurocomputing journal, Volume 237, 2017, pp. 242-254
7. Shini Renjith, "B2C E-Commerce Customer Churn Management: Churn Detection using Support Vector Machine and Personalized Retention using Hybrid Recommendations" , International Journal on Future Revolution in Computer Science & Communication Engineering, Volume: 3, Issue: 11.
8. Joy, Christya, A.Umamakeswaria, L.Priyatharsinib, A.Neyaab "RFM ranking – An effective approach to customer segmentation" Journal of King Saud University - Computer and Information Sciences ,5 September 2018
9. LG Pee, JJ Jiang ,G Klein "E-store loyalty: Longitudinal comparison of website usefulness and satisfaction",International Journal of Market Research 1–17
10. Niccolò Gordini ,Valerio Veglio "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B ecommerce industry" Industrial Marketing Management 2016 IMM07404
11. Annapurna P Patil , Savita Shetty ,Deepshika M P , Samarth S Hiremath , Shantam Mittal, Yogesh E Patil , "Customer Churn Prediction for Retail Business", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).
12. Peter S. Fader, Bruce G. S. Hardie, Ka Lok Lee, ""Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model", Marketing Science Vol. 24, No. 2, Spring 2005, pp. 275–284

AUTHORS PROFILE

Mr. Pradeep Kumar D is an Assistant Professor in the Department of Computer Science and Engineering in Ramaiah Institute of Technology. His research areas of interest are Data mining and Data Analytics.

Miss Shalini Sarode is a soon to be graduate student within the Computer Science Engineering program at the Ramaiah Institute of Technology. She will graduate with a BE in Computer Science and Engineering in 2019. Her research area of interest includes Data Analytics and Data mining.

Miss Varsha D Vadone is a soon to be graduate student within the Computer Science Engineering program at the Ramaiah Institute of Technology. She will graduate with a BE in Computer Science and Engineering in 2019. Her research area of interest includes Data Analytics and Data mining.

Miss Varsha CM is a soon to be graduate student within the Computer Science Engineering program at the Ramaiah Institute of Technology. She will graduate with a BE in Computer Science and Engineering in 2019. Her research area of interest includes Data Analytics and Data mining.

Miss Punya Prakash Shetty is a soon to be graduate student within the Computer Science Engineering program at the Ramaiah Institute of Technology. She will graduate with a BE in Computer Science and Engineering in 2019. Her research area of interest includes Data Analytics and Machine Learning.