Customer Behavior Analysis and Revenue Prediction System

[1]Rahul Gupta, [2]Pranil Kamble, [3]Vanshi Negandhi, [4]Ankush Hutke

[1,2,3]Student, [4]Assistant Professor, Rajiv Gandhi Institute of Technology, Mumbai, India.

[1]rg173602@gmail.com, [2]pranikamble123@gmail.com, [3]p.flora09@gmail.com, [4]ankush.hutke@mctrgit.ac.in

Abstract: In the era of e-commerce there are many organizations that have implemented customer behaviour analytics for their growth in business. It is a crucial challenge for the organizations in the e-commerce world to study and analyse the behaviour of the online buyers. The success of every organization is within the satisfaction of the customers they have and to gain new customers as well, and this is done by targeting the potential customers that can generate revenue to the organizations. RFM analysis is used to indicate recently buying customers, frequently buying customers, and huge spending customers. It is one of the best methods to segment organization's revenue generating customers around other customers. Also 80/20 rule is implemented which focuses on the 20 percent of the customers that generate 80 percent of the revenue for the organization. The model is developed using Light GBM (Gradient Boosting Method) which is a machine learning algorithm.

Keywords* — *Customer behavior analytics, frequency, light GBM, machine learning, monetary, recency.

I. Introduction

The world is more of e-commerce than the physical commerce. This is because the costumers have been provided better services from various organizations. To make growth in the e-commerce various organizations have to retain their previous customers and gain new one. For this it is very important for the organizations to understand the behavior of every customer. Customer behavior analysis is done to analyze how customers buy products on an e-commerce website. Every customer has a different approach to buy a product online. Also, there are few customers that generate a high revenue for the organization, so it is an important task for the organization to target these potential customers.

The biggest challenge for the organization in the analysis is to find all the hidden information through huge logs of data. To analyze huge amount of data machine learning's automated analytical model building is used. It is a part of AI which proposes that models can make decisions and identify patterns from the data with minimal involvement of humans.

For customer segmentation RFM analysis along with k-means clustering is used. RFM analysis defines recency, frequency and monetary value. These three attributes will help to understand the best customers that generate more revenue for the organization. This analysis will define facts like the recent the purchase is responsive the customer will be. Also, the frequency attribute indicates about the customer buying frequently. The term monetary value segments heavy spending customers from low spending customers.

The main idea is to target the potential customers that can generate a higher revenue to the organization. Every organization like Amazon, Flipkart have a system that analyzes the behavior of the customer. But they recommend products to every customer. So, there are chances that the customer may not buy it every time. So, it is essential to provide recommendations and offers only to those customers that can buy it. This is done by the 80/20 rule, the rule indicates that 20% of the customer generates 80% of the revenue. To identify those 20 percent of the customer Light GBM algorithm is used.

II. Literature Survey

A lot of research works have been done on prediction of customer's buying behavior in the past. Different data mining techniques, big data methods were used for prediction & attained different results for different data models.

Y. Xiaobing et al. [1], suggests that the cost of attracting new customer is more than retaining the old customers. This is due to marketing costs required to attract new customers. Due to this, together with the increase of competition it has become important for the organization for the development that the current customer's base is retained. An organization can analyze buying pattern of a customer and can adopt a proactive approach in predicting churn. Understanding customers' needs and patters is

DOI : 10.35291/2454-9150.2020.0245

possible as all the transactions are inserted through POS and recorded in databases.

According to [2], not all the customers generate the same amount of revenue. There are only few customers that generate a high revenue to organization. For many businesses the 80/20 rule has been proven true. 80 percent of the revenue is generated by 20 percent of the customers. To make appropriate promotional strategies is a challenge for marketing team. This 80/20 rule gives indicators and metrics that offer a truer picture of the company and allows to reach define action plans more precisely and better conclusions.

Zuo Y et al. [3], deep learning approach for the Prediction of Retail Store Sales system. The model predicts the increase and decrease in the sales of a retail store and tests its usability. The system used three years of POS data from supermarkets for the analysis, from which the data of 29 months was used for learning, while the remaining of data was used for verification. Predictive accuracy of the system varied between 75% to 86% according to the changes in the number of product attributes.

Authors [4], discuss that revenues and margins will increase if the old customers are retained. The cost of attracting new customers is very high and the margin also decreases in this process. Organizations can identify hidden trends and patterns by applying statistical techniques and machine learning algorithms. To improve the relationship with customers companies, need to implement data mining techniques to predict churn. Using these models' companies have improved their relationship with the customers.

Liu Bing et al. [5], naive Bayesian algorithm has the advantages of high-class efficiency and simple implementation. However, this method has disadvantages and potential of instability. It depends on the Distribution of samples in the sample space. Due to this, decision tree method was introduced to deal with the problem of interest classification. It used local storage technology in HTML5 to obtain the required experimental data. Information entropy of the training data is used to build the classification model.

Yi Zuo et al. [6], This model is doing a prediction on the basis of behavior of consumer purchasing in a grocery store using machine learning techniques. The system employs two representative machine learning methods. Bayes classifier and support vector machine (SVM) is used and the performance of them with the data in the real world is measured. It also executed a module which predicted consumer purchasing behavior.

III. DATA SET

This is a transnational data-set which have all the transactions happened during 01/12/2010 and 09/12/2011 for a UK-based online retail shop. Many customers of the company are wholesalers. The dataset is consisting of total 541909 transactions out of which only 4338 transactions are unique. Quantity, Unit Price and Date are taken into consideration while developing a revenue model.

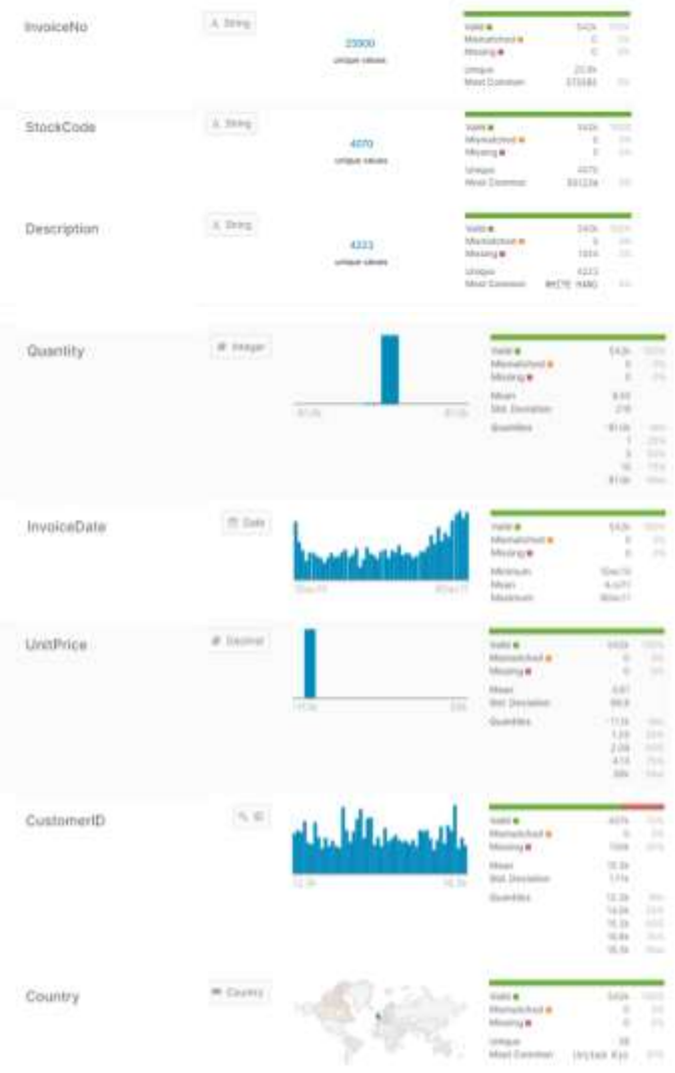


Fig 1: Dataset

IV. DATA ANALYSIS TECHNIQUES USED FOR PREDICTION

The analysis in done in three parts, First is 80/20 rule of marketing followed by the Gradient Boosting algorithm for decision making and lastly RFM analysis to figure out the potential customers.

4.1 80/20 Rule

According to the Pareto Principle, almost 80% of the result come from the 20% of the causes. Also, by looking at the Distribution of world GDP it is proved that 20% richest produce 80% income.

To develop a revenue model for an organization, instead of targeting all the customers, it is better to target only those 20% customers who are most responsible. To implement this 80/20 rule on a dataset first do sum up of transaction revenue at user level and take a log then do a scatter plot.

DOI : 10.35291/2454-9150.2020.0245

This scatter plot will help in analysis and finding those customers who generates revenue and then discard the non-revenue or customers with no transactions. Now further analysis and testing will be done only on those who generates revenue for an organization.

4.2 LightGBM

LightGBM is a high-performance gradient boosting framework based on decision tree algorithm and used for machine learning tasks such as ranking, prediction, classification etc. This boosting algorithm can become better with each added new case. In this boosting algorithm, when modifications are done on original data set new tree is added to the existing tree. Gradient Boosting trains many models in a stepwise, additive, gradual and sequential manner. Important feature of GBM prediction is that user can define number of trees which leads to more optimized solution.

LightGBM uses histogram-based algorithms, continuous feature values are bucket into discrete bins. This reduces memory usage and speed up training. Due to this feature time complexity decreases also reduce the memory usage. Other advantages of LightGBM over other training algorithms are: its faster training speed & better efficiency, Capable of handling large scale data, has better accuracy and support GPU and parallel learning.

Parameters of LightGBM used for the model:

num_leaves: Selecting accurate number of leaves in important as having a large number of leaves will improve accuracy, but will also lead to excessive number of parameters.

min_child_samples: The parameter can greatly deal with excessive number of parameters: larger sample sizes per leaf will reduce excessive number of parameters.

learning_rate: Increasing the number of repetitions may improve accuracy for that learning rate should be smaller.

bagging_fraction & bagging_freq: enables sub-sampling of the training dataset. The frequency controls how often bagging is used.

feature_fraction: sub-sampling of the attributes used in the training.

4.3 RFM Analysis

RFM stands for Recency, Frequency, and Monetary_value, each corresponding to some distinguishing qualities of customer. These three RFM metrics are important indicators of a customer's behavior because frequency & monetary value is responsible for customer's lifetime value, and recency is important retention, a measure of engagement.

Importance of RMS analysis:

- The customer with most recent purchase is more responsive to promotions
- Customers who are frequent buyers are more engaged and satisfied.
- Heavy spenders and low value purchasers are differentiated by monetary value.

V. Results and Discussion

IPython i.e. Interactive Python programming is used for the implementation of algorithms. The UK based online retail shop dataset consist of total 541909 transactions. 80/20 rule is first applied on this dataset to discard no-revenue customers from training and testing. Now the new dataset which consist of only customers who generate revenue. For the analysis and development of model the dataset is been divided into training dataset and testing dataset in the ratio of 70:30 respectively.

In this study, the log of the customers is predicted which consist of the customerID and their revenue. The goal is to generate this prediction log as accurate as possible.

In machine learning LightGBM is one of the most efficient algorithms while handling a huge data. Comparing LightGBM with other machine learning algorithms like Random forest, in terms of working with large datasets or real-time data, LightGBM is faster as well as produce more accurate results. Also works better with anomalous data as it is an algorithm which learns from the errors. This allows it to make more precise and accurate predictions. Comparing LightGBM with Deep Learning algorithms, studies claim that Deep learning algorithms are better. Deep learning algorithms are usually use to analyze image, speech or text data but when it comes to tabular data no deep learning algorithms can work as efficiently as LightGBM.

In RFM analysis the score is assigned to each unique transaction. The RFM scoring method depends on the most important feature in the dataset. Feature importance is plotted using LightGBM algorithm. In this retail store transaction dataset monetary_value is turned out to be the most important feature, therefore considering it as the most important RFM scores are assigned to each unique transaction. Using LigntGBM feature importance and RFM score final revenue log is predicted.

- 80/20 Rule

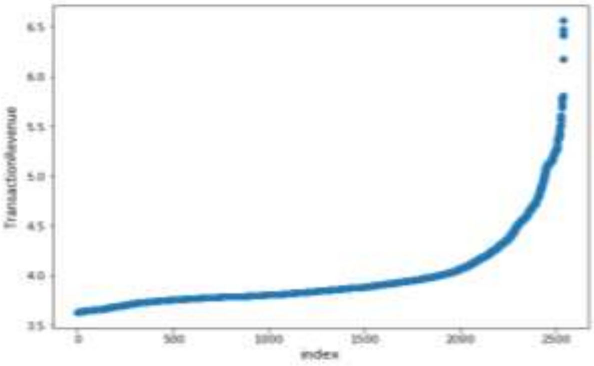
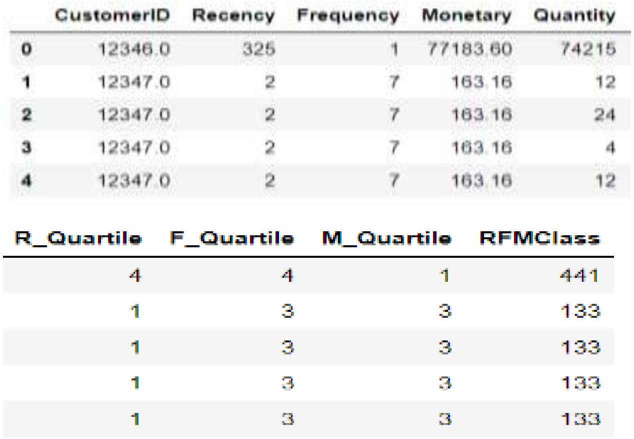


fig 2: Scatter Plot

DOI : 10.35291/2454-9150.2020.0245

It can be inferred from this plot Number of instances in train set with non-zero revenue: 115151 which is 21.24% Number of unique customers with non-zero revenue 4338

- RFM Analysis



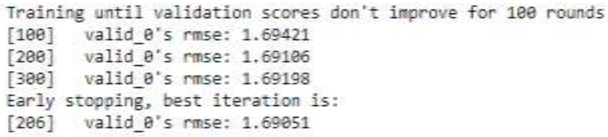
	CustomerID	Recency	Frequency	Monetary	Quantity
0	12346.0	325	1	77183.60	74215
1	12347.0	2	7	163.16	12
2	12347.0	2	7	163.16	24
3	12347.0	2	7	163.16	4
4	12347.0	2	7	163.16	12

R_Quartile	F_Quartile	M_Quartile	RFMClass
4	4	1	441
1	3	3	133
1	3	3	133
1	3	3	133
1	3	3	133

Fig 4: RFM Analysis

RFM score is assigned to each customer. There are 125 different RFM scores as each Recnecy, Frequency & Monetary are scored in 1-5 range. The RFM scores ranging from 111(lowest) to 555(highest).

- LightGBM



```
Training until validation scores don't improve for 100 rounds
[100]   valid_0's rmse: 1.69421
[200]   valid_0's rmse: 1.69106
[300]   valid_0's rmse: 1.69198
Early stopping, best iteration is:
[206]   valid_0's rmse: 1.69051
```

Fig 3: Validation Score

The validation score of the model is turned out to be 1.70.

LightGBM feature extraction helped analysing the most important features of a data set. Monetary (i.e. the intention of customer to spend or purchasing power of a customer) is turned out to be the most important feature.

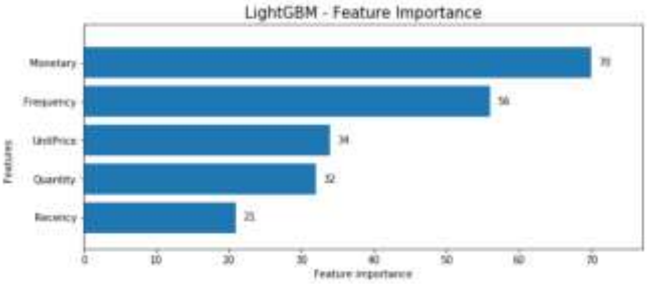
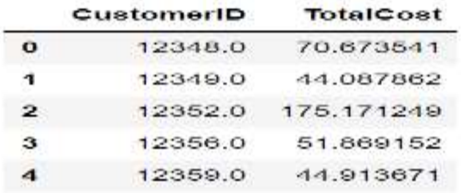


Fig 4: LightGBM Feature Importance

- Revenue Log



	CustomerID	TotalCost
0	12348.0	70.673541
1	12349.0	44.087862
2	12352.0	175.171249
3	12356.0	51.869152
4	12359.0	44.913671

Fig 2: Prediction Log

Log file is of predicted revenue is generated which is consist of the customerID and TotalCost.

VI. Conclusion

Using logs of various customers and the purchasing history a model can be developed which is used to analyse the behaviour of every customer also used to predict revenue generating customers. The model developed can identify all types of customer for better analysis of the behaviour. From the experimental results, it has been found out that Light GBM algorithm predicts the revenue generating customer with the help of feature importance developed on the basis of RFM analysis result and RFM scores. The result of LightGBM algorithm is with the validation score of 1.70 and accuracy of 80.147%. The performance of Light GBM algorithm can be improved with more accurate dataset along with modifications in the algorithm. Using this model an organization can easily make decisions about various customers.

Acknowledgment

We would like to thank MCT's Rajiv Gandhi Institute of Technology for providing research facilities and to Prof. Ankush Hutke (Information Technology) department for their support and encouragement.

References

[1] C. Jie, Y. Xiaobing, and Z. Zhifei, "Integrating OWA and data mining for analysing customers churn in ecommerce," The Editorial Office of JSSC and Springer-Verlag Berlin Heidelberg, vol. 28, pp. 381-391 2015.

[2] Google Analytics Customer Revenue Prediction. [Online], Available from:https://www.kaggle.com/c/ga-customer-revenue-prediction/overview (Accessed: 7 July 2017)

[3] Zuo Y, Yada K, Ali AS. Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques. In Computer Science and Engineering (APWC on CSE), 2016 3rd Asia-Pacific World Congress on 2016 Dec 5 (pp. 18-25). IEEE.

[4] IBM. [Online]. Available from:

https://www.ibm.com/developerworks/library/badatamining-techniques/

Developer works, Accessed: November 2016

[5] Bing L, Yuliang S. Prediction of User's Purchase Intention Based on Machine Learning. In Soft Computing & Machine Intelligence (ISCMI), 2016 3rd International Conference on 2016 Nov 23 (pp. 99-103). IEEE.

[6] Kaneko Y, Yada K. A deep learning approach for the prediction of retail store sales. In Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on 2016 Dec 12 (pp. 531-537). IEEE.

[7] Simple content-based recommendation engine. [Online], Available from: https://www.kaggle.com/cclark/simple-content-based-recommendation-engine (Accessed: 10 December 2017)

[8] Market Basket Analysis. [Online], Available from: https://github.com/sharmaroshan/Market-Basket-Analysis (Accessed: 02 September 2018)

[9] Sheela Gole and Bharat Tidke, proposed a system Frequent Item set Mining for Big Data in social media using ClustBigFIM algorithm, International Conference on Pervasive Computing in 2012.

DOI : 10.35291/2454-9150.2020.0245