International Journal of Research in Engineering, Science and Management
Volume 5, Issue 11, November 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Performance Optimization of an Enterprise using Data-Driven Strategy

Divya Jatain*
Assistant Professor, Department of Computer Science & Engineering, Maharaja Surajmal Institute of Technology, New Delhi, India

Abstract: **In this digital age, where data is ubiquitous, it is crucial to transform it into information to obtain useful insights that power growth. Considering the need for the appropriate deployment of socio-technical systems to benefit modern businesses, this paper focuses on using customer segmentation backed by Recency-Frequency-Monetary analysis, as its basis in order to attain marketing and sales excellence, understand consumer behaviour, and transform customer success. The project encompasses analysing data of the purchases made through an online retail company over the duration of a year. Thereafter, the customers are segmented using K-Means clustering technique. It is then followed by classification of customers, for which five classification techniques are implemented. The best result is shown by Random Forest classifier which is then used to understand customer behaviour.**

Keywords: **Classification, Customer segmentation, K-Means, Machine Learning, RFM analysis.**

1. Introduction

With an increasing number of businesses being evolved every day, it has become important for the corporations to act strategically to cater to the customer needs so as to stay in this competitive market. With the advent of data science and large availability of data archives, the profitability and growth of an enterprise can not only be escalated, but also strengthened by strategically coalescing programming, data analysis and machine learning. Customer segmentation is one of the applications of data mining, which is used to create different market segments that are uniquely identified by similar patterns of the customer base. The segments are used further for customizing the market plans as per each segment, to discover new schemes like for which segment the product will be most efficient and deeply understand the customer and object interrelation. It basically provides an insight to the companies what actually a customer is investing in and an opportunity to serve them better.

This paper presents RFM (i.e., recency, frequency and monetary) based customer segmentation using K-Means clustering algorithm [1] followed by comparative performance analysis of classification of customers using various machine learning classifiers. The RFM score works by analyzing Recency i.e., how recently the purchasing has been done by a customer, Frequency i.e. what is the frequency of the purchase and Monetary i.e. how much money is spent by the customer to do the purchase to determine the most valuable customers for the business. From the dataset, three segments are identified using the K-Means algorithm [2] and label led as low value, mid value and high value customers. Furthermore, classification of customers is done by training and testing various Sci-kit classifiers and brief performance analysis is presented.

2. Data Understanding & Metric Identification

Coordinated work by all the independent teams is one of the biggest challenges of any organization. Everyone in the group must know what they are dealing with to come with effective streamlined solutions that are in unison with what the industry demands.

For effective functioning; user acquisition [3], retention, marketing, efficiency, and growth become fundamental for organizations. This roots to the data at hand, analysing different ways to approach this data, interpret and make effective observations from them to impact growth. User driven approaches require user acquisition to be long lasting and not just temporary, this makes growth a consistent variable for net profitability. User sustainable growth is influenced by improving optimization metrics.

Dataset from an Online Retail company is gathered and carefully filtered, analysed in order to implement meaningful hypotheses [4]. Numerous calculations are performed on the initial dataset to observe net revenue, monthly retention and customer ratio. The dataset, however, does not have the complete data for the month of December (201112). The dataset comprises Customer ID, Unit Price, Quantity and Invoice Date. Monthly Revenue is selected as the basis of this paper's metric to build the following equation:

Revenue = Active Customer Count * Order Count * Average Revenue per Order

Monthly Revenue (Total sales for the Online Retail every month) is visualized using the monthly revenue data frame. A steep growth in Revenue is observed after 201108 (which is Aug, 2011). This trend is represented in Figure 1.

Then, Month over Month growth is visualized where steep growth in Revenue per month is observed in March (201103), May (201105), September (201109) and November (201110). Figure 2 represents the monthly growth rate curve.

*Corresponding author: divyajatain@msit.in

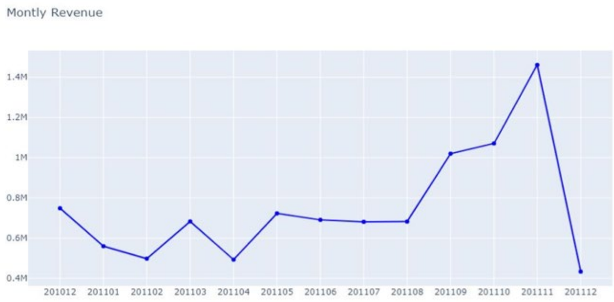

Fig. 1. Monthly revenue curve

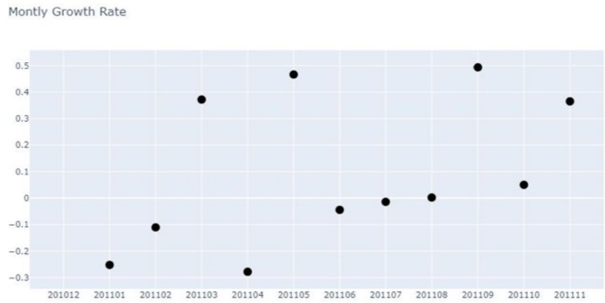

Fig. 2. Monthly growth rate curve

However, April (201104) observed its lowest Revenue in 2011. An in-depth exploration is conducted to find the reason that led to the huge depression in the curve by visualizing Monthly Active customers and Monthly Order Count. Unique Customer IDs from the dataset are used to visualize the Active Customers count per month. While there is a consistent increase in customers from September (201109) to November (201111).

April, however, observed a 11.5% decline in customer count and an 8% decline in the order count per month as observed in Figure 3(a) and Figure 3(b) respectively.

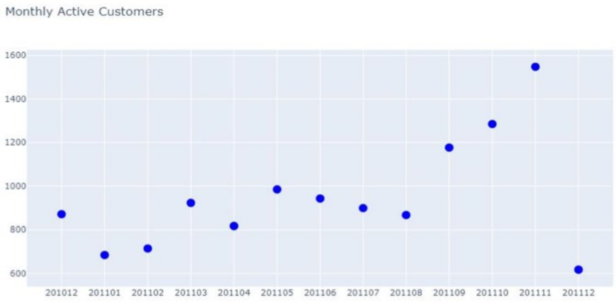

Fig. 3(a). Monthly active customers curve

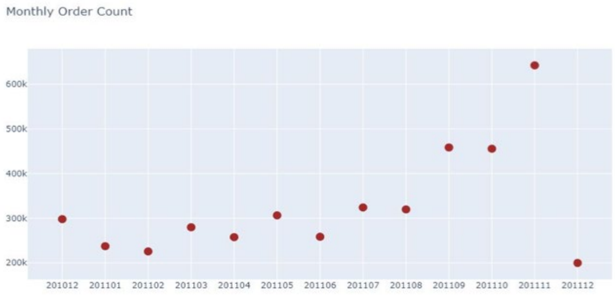

Fig. 3(b). Monthly order count curve

Active Customer Count per month directly influencing Order Count per month which led to the decline in April is observed. Finally, Average of Revenue for each month is calculated to visualize Monthly Order Average which further confirmed the above relation by expressing a decline in April. This is represented in Figure 4 below.

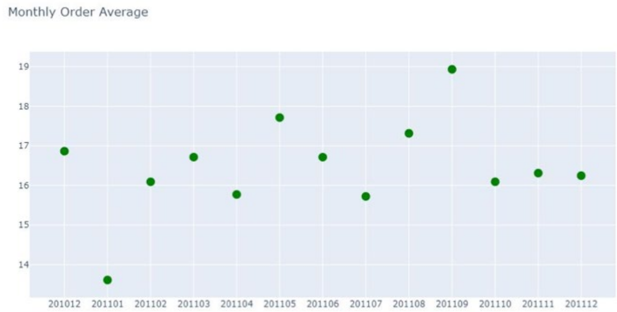

Fig. 4. Monthly order average curve

3. Customer Segmentation

The RFM analysis is the most dominant and classical method for customer segmentation as it covers the behavioural aspect of the customer database [5]. It is a preliminary data mining technique performed before the actual segmentation process by extracting customers' recency, frequency and monetary values. The consumption interval is shown by recency, frequency shows the usage regularity whereas monetary expenditure is depicted by the monetary value. Segmentation based on RFM analysis provides an effective decision-making capability to target the customer tank and establish appropriate strategies like promotional campaigns, discount coupons and vouchers or seasonal discounts to improve customer retention, distribution strategies and sales.

Customer segmentation is one of the significant applications of K-Means algorithm which is a prototype based exploratory data analysis technique [6]. It divides the given dataset into a user specified number of clusters (group of objects having similar characteristics). It works well on large datasets, decreases misclassification data rate and is computationally faster. The present work in this section makes use of the K-Means algorithm for clustering.

In this section, at first, the distribution of recency, frequency and monetary over the processed dataset is analysed and depicted through various visualizations to get a better understanding. Furthermore, individual scores (clusters) for recency, frequency and revenue are created using K-Means clustering algorithm. Thereafter, overall RFM score-based segmentation is performed and analysed using visualizations.

1) Recency

For recency calculation, the most recent purchase date and number of days of inactivity is found for each customer and a data-frame is created. The max invoice date in the dataset is considered as the observation point for every customer. It is found that the average recency is 91 days with maximum recency being 374 days while the median is 52 days. The visualization for the recency distribution across the customer dataset is shown in the figure 5.

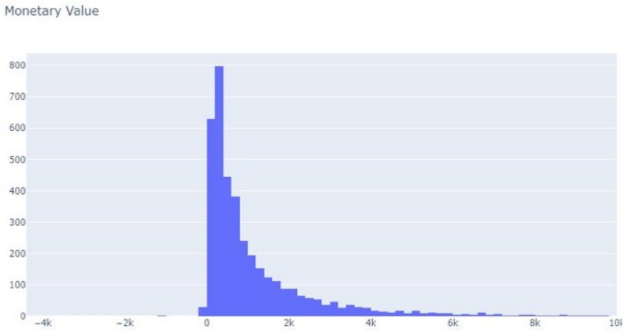

Fig. 5. Recency distribution graph

2) *Frequency*

To calculate frequency, total order count for each customer starting from the first day is found out. A data frame is created and merged with each new customer in the dataset. A frequency of maximum 7984 order count is observed. High frequency number clearly indicates valuable customers whereas low frequency indicates inactive customers. The distribution of frequency across the dataset is depicted through a histogram represented by Figure 6.

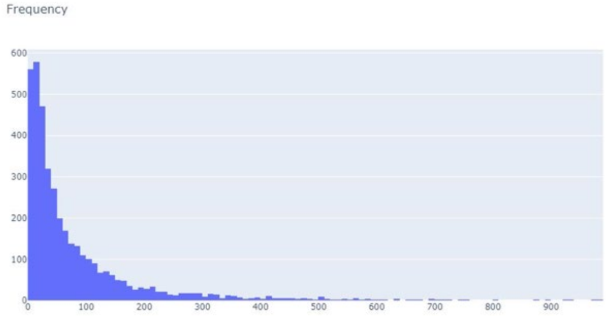

Fig. 6. Frequency distribution graph

3) *Monetary*

Revenue calculation is done using the equation given below:

Revenue (for each customer) = Unit Price * Quantity

Data-frame for revenue is created for each customer and a maximum revenue of 256439 is observed. Visualization depicting revenue division is shown in Figure 7.

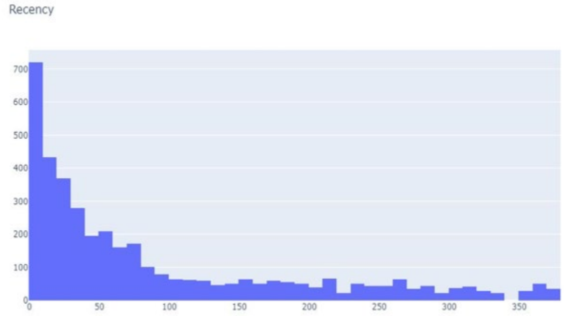

Fig. 8. k-Means elbow curve

After analysing the distribution of RFM on the dataset, K-Means algorithm is applied to generate recency, frequency and monetary score (clusters) individually for each customer. The factor 'K' used to specify the number of clusters is decided using the Elbow method which determines optimal cluster number for ideal inertia.

For the current work, 4 clusters are the optimal one as shown in Figure 8.

Thereafter, overall RFM scores are generated for each customer using recency, frequency and revenue scores (clusters). The customer dataset is segmented in total three segments based on the overall RFM scores with score of 8 depicting best customers and 0 is the worst ones. The meaning assigned to each segment are given below:

- High Value (5+ RFM score)
  This segment consists of the most valuable customers who are highly active and have high monetary values.
- Mid Value (3-4 RFM score)
  This segment includes customers who are not frequent buyers but have moderate monetary values.
- Low Value (0-2 RFM score)
  Customers who buy a few times and spend low fall under this segment of inactive customers.
  The distribution of the customer segments differentiated in terms of RFM is depicted below through different scatter plots.

Figure 9(a), 9(b) and 9(c) represent the plots for Revenue vs. Frequency, Frequency vs. Recency & Revenue vs. Recency for the given problem scenario.

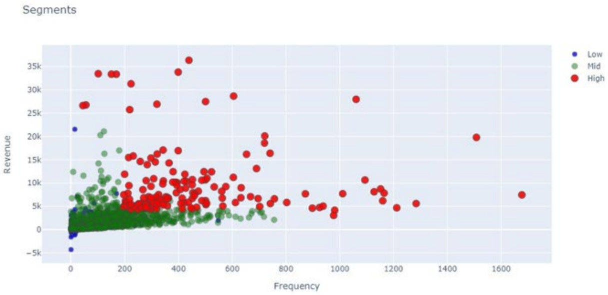

Fig. 9(a). Revenue vs. Frequency graph

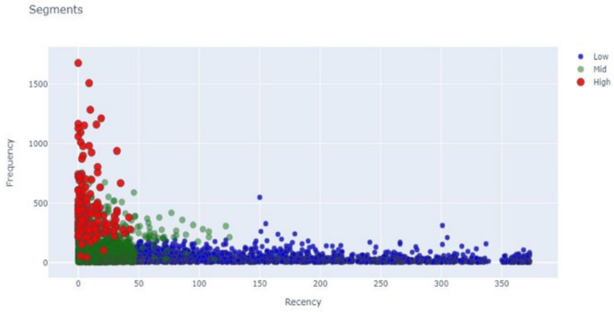

Fig. 9(b). Frequency vs. Recency graph

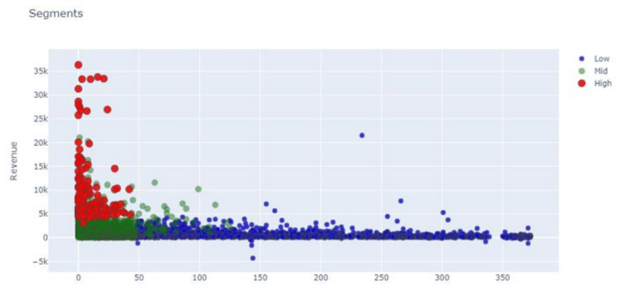

Fig. 9(c). Revenue vs. Recency graph

4. Classification of Customers

Five different classifiers are implemented to classify the customers into the categories discussed in the previous section. These are Support Vector Machine, Logistic Regression, k-Nearest Neighbour, Decision Tree and Random Forest. All these methods have their corresponding pros and cons which are discussed in Table 1.

Out of the above mentioned methods, only Linear Regression is limited to classification problems, rest other methods, viz., Support Vector machines, k-Nearest Neighbours, Random Forest and Decision Tree work both for classification and regression. In the next section, results for the implementation of these different classifiers will be discussed.

5. Result and Discussion

This paper focuses on driving the growth of corporations by following the proposed gradational data-driven methodology. The online retail company dataset consists of above 5 lakh instances of customer purchase information with 8 attributes

Table 1
Description & Details of different classifiers used in this work

#Classifier Name	#Classifier Description	#pros	#cons
Support Vector Machine	It is a supervised machine learning technique that can be used for classification and regression both. In this project, the SVM was deployed for classifying the customers. The algorithm aims at finding a hyperplane that differentiates a dataset into two classes in the best manner [7].	i) It handles high dimensional data effectively. ii) It works well when the classes are well-separated.	i) The model does not work well for large data sets as it takes more training time. ii) SVM is a complex model, i.e, it is difficult to interpret unlike other models.
Logistic Regression	Logistic regression used to predict an outcome (classification) in case of problems having dichotomous nature. It is a predictive analysis technique based on a set of independent variables, where either the event occurs (outcome: 1) or does not occur(outcome:0) [8].	i) It is easy to interpret and implement. ii) If the dataset has features that are linearly separable, Logistic Regression is very efficient to train which also makes the model fast.	i) For high dimensional dataset, it might run into the problem of overfitting on the training set which, in turn, would produce inaccurate results on the test set. ii) It assumes linear relationship between the dependent and the independent variables. Moreover, owing to its linear decision surface, it cannot be used to solve nonlinear problems like k-Nearest Neighbours.
k-Nearest Neighbour	It is a supervised machine learning technique that used for both classification and regression predictive problems. It works on the principle of feature similarity, i.e., it classifies data points based on points in the training set that are most similar to it [9].	i) It is a non-parametric algorithm as it does not make assumptions about the underlying data. Thus, it is useful for non-linear data. ii) Unlike SVM, it works well for datasets that have more training data compared to the number of features. iii) Since it uses all the data for training while classification, it does not have a specialized training period and has low calculation time.	i) It is computationally expensive and for large datasets, as the cost of calculating the distance between the new and existing data points is high, it does not work well. It also has limitations while working with high dimensional data. ii) It is sensitive to outliers and missing values. iii) It requires more memory storage as compared to other supervised learning algorithms. As the size of the dataset increases, the speed of the algorithm decreases, making it slower than other classifiers like Decision Tree.
Decision Tree	It is a supervised machine learning technique which is non-parametric in nature. It can solve both classification problems and regression problems by transforming the data into a tree-like representation of decisions and the likely outcomes. By making the use of the features of the data, the model learns decision rules and then is able to predict the target variable's value [10].	i) Unlike most of the other algorithms discussed above, decision trees are easy to understand and require less code and analysis during the data pre-processing stage. ii) It neither requires normalization of data nor the scaling of data.	i) In most cases, the mathematical calculations for a Decision tree require more time and memory. ii) Decision trees are not stable because a small change in the data can cause the generation of an entirely different tree.
Random Forest	This is a supervised machine learning algorithm deployed for both classification and regression. Here, multiple decision trees are created on various data samples of the dataset and selects the most optimal solution by taking the average prediction of individual trees [11].	i) It is more robust than an individual decision tree. Since it is an aggregate of decision trees, it reduces overfitting and helps to improve the accuracy. ii) Since it follows a rule-based approach, data normalization is not required. Also, it handles the missing values by default.	i) Owing to the aggregation they build upon, random forests require more computational power as well as resources. ii) It has a longer training period as it combines a lot of decision trees to determine the class.

collected for the duration of a year. After carefully analysing the behavioural trends of customers and metric identification, RFM based segmentation using K-Means clustering algorithm is performed. Overall RFM scores in the range 0-8 (which implies 9 total clusters) are generated for each customer. Thereafter, 3 segments are created based on the scores.

Following are the customer segments created and strategies for the future.

- High Value (most valuable customers) - improve retention.
- Mid Value (moderately active customers) - improve retention, increase frequency.
- Low Value (inactive customers) - improve frequency After customer segmentation, classification of customers into the above-mentioned segments is performed using various machine learning classifiers. Same dataset as mentioned earlier is used. The classification result run on the training dataset is depicted below through precision values using a scatter plot.

Training set precision:

The above result plot clearly shows that the Random Forest classifier gives the highest precision value of 91.69 while 75.81 is the lowest given by Support Vector Machine classifier on the training set. Visualization of precision obtained for classification results run on the test set is shown below.

Test set precision:

The above plot clearly depicts that the precision value is highest i.e., 75.54 for Random Forest classifier while 67.19 is the lowest value given by Support Vector Machine classifier on the test set. A performance comparison table of the classifiers used is shown below.

Thus, the Random Forest classifier gives the best result of customer classification on both the training set (91.69) and the test set (75.54).

6. Conclusion and Future Scope

This paper discussed the strategic implementation of data analysis and machine learning techniques in order to drive the growth of an organization. Right from figuring out the appropriate metric that best predicts the company's long-term success, and growth, in general, to understanding the precise customer behaviours and attributes which signal the risk and timing of customer churn, the project covered a detailed analysis as well as comparison of suitable machine learning techniques.

Using comparative visualizations of the dataset used, monthly customer count directly influences monthly order count. Segmentation of customers was implemented using the K-Means algorithm based on RFM analysis. Customers were finally classified in a comparative analysis of different classifiers amongst which Random Forest provided the best results.

The previous sections discussed the segmentation of the customers as High Value, Mid Value, and Low Value. For the customers lying in the High Value category, the focus needs to be on improving their retention, for those in the Mid Value, improving the retention as well as incrementing the frequency, while for those in Low Value, only increasing the frequency would be a good first step. Hence, the approach discussed in this paper can be furthered by predicting the lifetime value of the customers. A machine learning model can be deployed which could predict the lifetime value of the customers using the RFM score of the customers as the feature set [12].

More so, once the lifetime values are obtained, it would facilitate the understanding of the customers, i.e, which customers are profitable and which are the ones that do not add as much to the company's revenue. Using these, the customer behaviour can be understood and the strategy can be extended by working towards retaining the customers that have a high lifetime value. This can be obtained by creating a Churn Prediction model that would, in turn, help in enhancing [13] the retention rate.

Acknowledgement

I wish to extend my gratitude to the Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, for believing in me and providing motivation all through the development of this project. Without the department's guidance this work would not have been a success.

References

[1] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using K-means Clustering," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 135–139.

[2] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno, and E. A. Pratama, "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm," *2018 Third Int. Conf. Inform. Comput. ICIC*, pp. 1–6, 2018.

[3] L. Ang and F. Buttle, "Managing for Successful Customer Acquisition: An Exploration," *J. Mark. Manag*., vol. 22, pp. 295–317, 2010.

[4] M. Abirami and V. Pattabiraman, "Data Mining Approach for Intelligent Customer Behavior Analysis for a Retail Store," 2016, pp. 283–291.

[5] Y. S. Cho, S. C. Moon, S.-P. Jeong, I.-B. Oh, and K.-H. Ryu, "Clustering Method Using Item Preference Based on RFM for Recommendation System in U-Commerce," 2013.

[6] C. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," *Int. J. Adv. Res. Artif. Intell*., vol. 4, 2015.

[7] P. Albuquerque, S. Alfinito, and C. V. Torres, "Support vector clustering for customer segmentation on mobile TV service," *Commun. Stat. Simul. Comput*., vol. 44, no. 6, pp. 1453–1464, 2015.

[8] J. Mccarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," *J. Bus. Res*., vol. 60, pp. 656–662, 2007.

[9] M. Abessi, "Marketing Data Mining Classifiers: Criteria Selection Issues in Customer Segmentation," *Int. J. Comput. Appl*., vol. 106, no. 10, pp. 5–10, 2014.

[10] G. Tirenni, C. Kaiser, and A. Herrmann, "Applying decision trees for value-based customer relations management: Predicting airline customers' future values," *J. Database Mark. Cust. Strategy Manag*., vol. 14, no. 2, pp. 130–142, 2007.

[11] A. Shaik and S. Srinivasan, "A Brief Survey on Random Forest Ensembles in Classification Model: Proceedings of ICICC 2018, Volume 2," in *Lecture Notes in Networks and Systems*, 2019, pp. 253–260.

[12] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Comput*. Sci., vol. 3, pp. 57–63, 2011.

[13] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, 2019.