**Chapter 1**

**INTRODUCTION**

* 1. **Introduction**

In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.

* 1. **Problem Statement**
  2. **Scope of the Project**

In general, the methods used to gather the data for this project can easily be extended into other relevant contexts/analyses. While there is clear value in using the same data to investigate purchasing patterns or to build an item-based collaborative filtering recommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four intertwined Goals:

* To cluster customers based on common purchasing behaviours for future operations/marketing projects
* To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines
* To investigate how similar data and algorithms could be used in future data mining projects.
* To create an understanding and inspiration of how data science can be used to solve real-world problems. Before delving into the details of the project and its implications, the next chapter discusses what customer segmentation analysis actually is and the reasons for its importance.
  1. **Methodology**

When we are provided with raw data extracted from database, it might be messy and non-informative to look at individual records.

* RFM analysis is applied to present data at aggregate level and is used to segment customers into homogenous groups.
* These three values are important as **F** and **M** indicate value of customers, and **R** indicate customers’ engagement and satisfaction.
* The values are easy to obtain from the basic set of information for each purchasing history.
* FM technique is a **cost-efficient marketing strategy** based on customer behavior segmentation.
* With this system the accuracy can be increased by about **20.4%** than the existing system.

|  |  |  |
| --- | --- | --- |
| **Recency** | **Frequency** | **Monetary** |
| When is your last purchase? | How many times have you place or purchased? | How much have you spent? |
| **Example:**  Length of duration since last purchase. | **Example:**  Number of orders over selected analysis period. | **Example:**  Sum of total amount spent over the period. |

**Chapter 2**

**LITERATURE REVIEW**

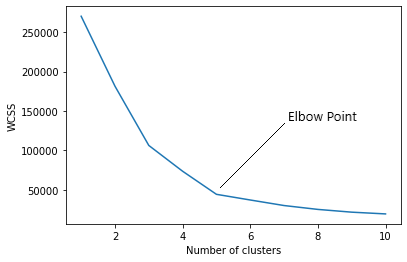
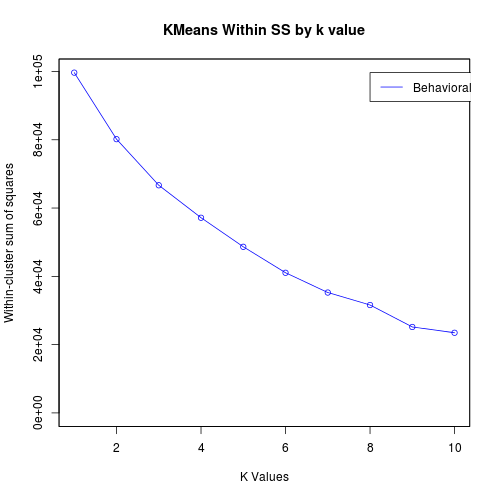
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **TITLE** | **AUTHOR** | **YEAR** | **APPROACH** | **ADVANTAGES** | **DISADVANTAGES** |
| 1. | User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis | W. Jun, S. Li, Y. Liping, N. Xiaxia, L. Yuanyuan, C. Xiaodong | 2021 | K Means Initialization method with RFM analysis | RFM analysis used for better customer history analysis | Sihoultee method with RFM model can perform better |
| 2. | Customer Segmentation using K-Means Clustering | Yash Kushwaha and Deepak Prajapati | 2021 | K Means Clustering  Elbow method | Finding the optimal no of clusters (k) | This method doesn't always work well, especially if the data is not very clustered. |
| 3. | Customer Segmentation using K-Means Clustering | Jaswanth Reddy Vulchi | 2021 | Data pre-processing and K Means Clustering  Elbow method | Same as Literature paper 2 | Same as Literature paper 2 |
| 4. | Customer Purchase Prediction Based on Improved Gradient Boosting Decision Tree Algorithm | H. Wu and B. Li | 2022 | Using Gradient Boosting Decision Tree Algorithm (GBDT) | - | The Clustering Accuracy is less compared to RFM model |
| 5. | K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data | Kayalvily Tabianan, Shubashini  Velu and Vinayakum-  ar Ravi | 2022 | Data Mining and K Means Clustering  Elbow method used | Same as Literature paper 2 | Same as Literature paper 2 |
| 6. | Customer Segmentation using K-Means Algorithm | Hemashree Kilari, Sailesh Edara, Guna Ratna Sai Yarra, Dileep Varma Gadhiraju | 2022 | K Means Clustering  Silhouette method | Better than the K Means Elbow method | No techniques used for improving the clustering. |

**Chapter 3**

**PROJECT DESCRIPTION**

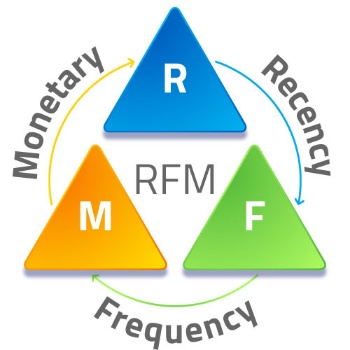
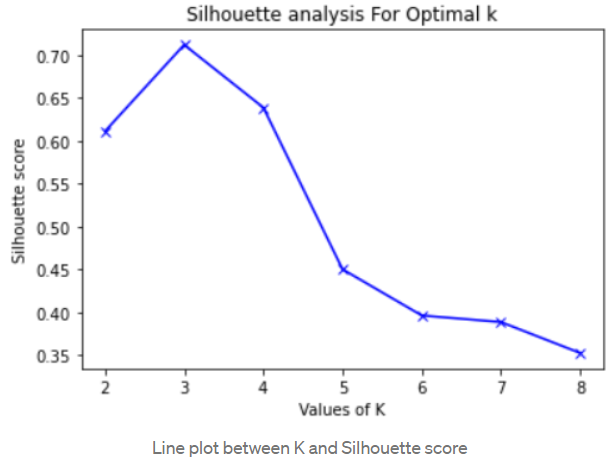
* 1. **Existing System**

In the current system, the K Means Initialization method and Elbow method are used to find the optimal no of clusters (k).

* The elbow method doesn't always work well, especially if the data is not very clustered.
* Because of that it will produce a smooth curve which makes difficult to find the optimal no of clusters (k).
* Also RFM technique (which is specifically used in customer segmentation) is not implemented in the current system.
  1. **Proposed System**

In our system we are using **K Means silhouette method** to find the optimal clusters (k) and **RFM technique** to improve the accuracy of the clustering the based on the customer history, and then it’s visualized.



* RFM technique is a cost-efficient marketing strategy based on customer behavior segmentation.
* RFM stands for **Recency, Frequency, Monetary**. Simply, it groups customers based on their purchase history.
  + Recency - How recent was the customer's last purchase?
  + Frequency - How often did this customer make a purchase in a given period?
  + Monetary - How much money did the customer spend in a given period?
* With this system the accuracy can be increased by about **20.4%** than the existing system.

**3.2.1 Advantages**

* 1. **System Specification**
     1. **Hardware Specification**

|  |  |
| --- | --- |
| PROCESSOR | Intel i5-8250 @ 3.40GHz |
| STORAGE | 512 GB SSD |
| RAM | 16 GB |
| GPU | Nvidia GTX 1650 Ti |
| OPERATING SYSTEM | Windows 11 x64 Bit |

* + 1. **Software Specification**

|  |
| --- |
| Windows 10, 11 |
| Anaconda Jupiter Notebook |
| Python 3.10 |
| Machine Learning Modules |

**REFERENCES**

[1] W. Jo-Ting, L. Shih-Yen and W Hsin-Hung, "**A review of the application of RFM model**", 2010.

[2] K. Su-Yeon, J. Tae-Soo, Suh. Eui-Ho and Hyun-Seok Hwang, "**Customer segmentation and strategy development based on customer lifetime value: A case study**", 2006.

[3] H. Fadly, B. Harry, de R. Mark and K. Maarten, "**Mobile customer segmentation based on smartphone measurement**", 2014.

[4] R. Haidi, S. Xianzhang, Kouassi R. Ahoussou, F. Juanjuan, X. Yingchun, E. Mohamed, et al., "**Feature selection based on artificial bee colony and gradient boosting decision tree**", 2019.

[5] P. Anitha and M. P Malini, "**RFM model for customer purchase behavior using K-Means algorithm**", 2019.

[6] C. Ching-Hsue and C. You-Shyang, "**Classifying the segmentation of customer value via RFM model and RS theory**", 2009.

[7] F. Safari, N. Safari and G.A. Montazer, "**Customer lifetime value determination based on RFM model**", 2016.

[8] Mahboubeh Khajvand et al., "**Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study**", 2010.

[9] W. Jun, S. Li, Y. Liping, N. Xiaxia, L. Yuanyuan, C. Xiaodong, et al., "**User Value Identification Based on Improved RFM Model and -Means++ Algorithm for Complex Data Analysis**", 2021.

[10] J H. Friedman, "**Greedy function approximation: a gradient boosting machine**", 2001.

[11] C Guo and F. Berkhahn, "**Entity embeddings of categorical variables**", 2016.

[12] T Chen and C. Guestrin, "**Xgboost: A scalable tree boosting system**", 2016.

[13] T M Oshiro, P S Perez and J A. Baranauskas, "**How many trees in a random forest?**", 2012.

[14] H. Wu and B. Li, "**Customer Purchase Prediction Based on Improved Gradient Boosting Decision Tree Algorithm**", 2022.