



REVA
UNIVERSITY

Bengaluru, India

A Project Report on
Repeat Visit Modeling in E-Commerce

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Ashish Chandra Jha
R19MBA01

Under the Guidance of
Dr.J.B.Simha
Chief Mentor, RACE,
REVA University

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

September, 2020



Candidate's Declaration

I, Ashish Chandra Jha hereby declare that I have completed the project work towards the Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Repeat Visit Modeling in E-Commerce** under the supervision of Dr.J.B.Simha, Chief Mentor. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Place: Bengaluru

Date: 13th Sep 2020

Name of the Student:

Ashish Chandra Jha

Signature of Student



Certificate

This is to Certify that the PROJECT work entitled Repeat Visit Modeling in E-Commerce carried out by Ashish Chandra Jha with SRN R19MBA01, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master of Business Administration in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

Signature of the Guide

Name of the Guide **Dr.J.B.Simha**

Guide

Signature of the Director

Name of the Director **Dr. Shinu Abhi**

Director

External Viva

Names of the Examiners

1. Indrajit Kar, Head of AI-Chief Architect & Data Scientists, Siemens
2. Pradeepta Mishra, Associate Principal & Head of AI, LTI-Larsen and Toubro

Place: Bengaluru

Date: 12th Dec 2020



Acknowledgement

Please acknowledge the role of your mentors, trainers, classmates, program office members, family and friends who have directly and indirectly supported you in this work.

Please acknowledge the support provided by Hon'ble Chancellor, Dr. P Shayma Raju, Vice Chancellor, Dr. K. Mallikharjuna Babu, and Registrar, Dr. M. Dhanamjaya, as a standard protocol.

Place: Bengaluru

Date: 12th Dec 2020



Similarity Index Report

This is to certify that this project report titled **Repeat Visit Modeling in E-Commerce** was scanned for similarity detection. Process and outcome is given below.

Software Used: **Turnitin**

Date of Report Generation: **12th Dec 2020**

Similarity Index in %: **10 %**

Total word count: **7510**

Name of the Guide: **Dr. J.B.Simha**

Place: Bengaluru

Date: 12th Dec 2020

Verified by: Andrea Brain C

Name of the Student: Ashish Chandra Jha

Signature of Student

Signature

Dr. Shinu Abhi

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	RFM	Recency, Frequency, Monetary
2	RFMT	Recency, Frequency, Monetary, Tenure
3	K-Means	K-Means Clustering

List of Figures

No.	Name	Page No.
5.1	Project Pipeline	15
8.1	Cleaning Total address or only the pin code Column	19
8.2	Cleaning State and City	20
8.3	Removing negative invoice	21
8.4	EDA Correlation Heatmap	21
8.5	EDA top 5 state revenue	22
8.6	EDA top 5 cities revenue	22
8.7	Exploratory Review of Data:-Top 5 Gross Margin Cities	23
8.8	Exploratory Data Analysis: - Top 5 Cities Gross Margin	24
8.9	Exploratory Data Analysis: - Monetary & GM Quartile wise	25
8.10	Exploratory Data Analysis: - Retain Per Quartile	25
8.11	Exploratory Data Analysis: - Customers Retain/Lost	26
9.1	K-Nearest Neighbors (KNN)	27
9.2	Decision Tree (DT)	27
9.3	Random Forest	28
12.1	KNN Result	29
12.2	Decision Tress Result	29
12.3	Random Forest Result	30

Abstract

Two new sectors that need strong support from the CRM scheme are e-retailing and website marketing. It is important for a website to keep clients interested and to come back to visit regularly. Data mining is an important and common instrument for all sectors to build successful CRM programmes to reach loyal consumers, as site data and direct marketing data are accessible in enormous quantities. Because most of these data are real buying data, models to explain and forecast consumer behaviour may also be built a step further.

Two mathematical models from the repeat purchasing action theory were used in this analysis to evaluate consumer satisfaction. The models were able to estimate the repeat-customer ratio and were able to classify marketing factors that influenced the repeat-rate.

Keywords: Repeat Purchase Modeling, Churn Analysis, KNN, Decision Tree, Random Forest,

Contents

Candidate's Declaration.....	2
Certificate.....	3
List of Abbreviations	5
List of Figures	5
List of Tables	5
Abstract.....	6
Chapter 1: Introduction.....	8
Chapter 2: Literature Review.....	10
Chapter 3: Problem Statement	12
Chapter 4: Objectives of the Study	13
Chapter 5: Project Methodology	14
Chapter 6: Business Understanding.....	16
Chapter 7: Data Understanding.....	18
Chapter 8: Data Preparation.....	20
Chapter 9: Data Modeling.....	27
Chapter 9: Data Evaluation.....	27
Chapter 10: Deployment.....	29
Chapter 11: Analysis and Results	30
Chapter 12: Conclusions and Recommendations for future work.....	31
Bibliography	34
Appendix.....	35
Plagiarism Report.....	35
Publications in a Journal/Conference Presented/White Paper	35
Any Additional Details	36

Chapter 1: Introduction

1.1: Introduction

Loyal customers are those who make repeated purchases of a product/service. Studies have Global discount campaigns are very much conducted by big business-to-consumer (B2C) websites like Zappos and Amazon, on festive days such as Black Friday and Mellandagar(Christmas holidays). In these situations, retailers are attracting new buyers. However, many new customers only binge once on these discount sales, and even the deals for the people often don't accomplish return on investment (ROI) as planned by sellers. To give custom-made ads (and promotions) against them, retailers must consider potential buyers to deduct the cost of the advertisement. For any solo merchant, foreseeing its loyal customers is tough as they won't be prone to the information of their current customers. Instead, B2C e-commerce platforms have the click path information and all shoppers' shopping details at all the stores on their networks. Thus, they will get an overview of their customers' likes and preferences using their previous history, and then decide how a new customer will most likely purchase again from the same retailer.

Based on the sales data of the 'Double 11' day of 2014, Alibaba hosted a foreign competition for repeat buyer prediction at IJCAI 2015 at Tmall.com, China's largest B2C website. In 2015, Double 11 was China's highest online shopping event with the highest sales of over 14 billion US dollars. Competition data includes the number of suppliers and their new customers hired during the incident and the customer contact log data for six months before the event. The objective of the task is to predict whether after six months the new consumers of a particular retailer will buy merchandise again from them.

One of the most critical issues for corporations is the development and retaining of potential customers. While recipient companies focus on acquiring new buyers, to build cross-selling opportunities for themselves, experienced ones try to concentrate on retaining the existing ones. According to Freeman (1999), one of the most important methods of increasing the importance of customers is to support them for prolonged periods.[1]

These new consumers are called the respective merchants' repeat buyers. We got first place in the first stage of the race. Our victory strategy comprised broad functionality and model designing engineering. Specifically, to classify consumers, vendors, goods, categories, items and their experiences, we created certain types of characteristics from different aspects. Data has portrayed that the price of recruiting a new customer is said to range from \$300 to \$600, and taking in a new client costs about 5-6 times as much as keeping a regular client. While a 5 per cent rise in client loyalty results in a 25-95 per cent profit gain, we have also used a group of approaches to combine several classifiers along to advance the performance. Like most of the competition participants did, the repeat buyer estimation dilemma can be conceived as a standard classification problem. The need to test new approaches to prediction is, thus, powerful and urgent.

This task's model preparation does not vary much from that of other classification tasks. Instead, the main aspect which brings a distinction of this role from others is feature engineering. Feature engineering is also the secret to the victory of a machine learning objective and is an important data science aspect. Because it is domain-specific, it can be more complex than learning, whereas machine learning algorithms are mostly general-purpose. It is challenging to build features, and most of the attempt in a machine learning project normally ends there. While in the academic community, thousands of categorization algorithms have been initiated and tested, not much progress has been outlined on feature engineering for prognostic tasks for e-commerce. We are also discussing function engineering in this article. We would explain how to produce different types of features from user interface log details and evaluate the relevance of these features through extensive researches. In all types of e-commerce applications, such as market classification, product recommendation, and user-based creation for brands, the features we have developed can be utilised. We hope that our work will be useful for data science practitioners who need to create applications for e-commerce prediction activities.

Chapter 2: Literature Review

2.1 MODELS OF REPEAT PURCHASE BEHAVIOUR

As a subfield of customer behaviour, the pioneering work of Ehrenberg[11] has opened up the field of repeat purchasing theory. For fast-moving consumer products (FMCG), the principle holds especially well.

In particular, in the present sense, two models of repeat buying behaviour are relevant: the model of the negative binomial distribution (NBD) and the model of logarithmic sequence distribution (LSD). The negative binomial distribution is used to model a heterogeneous population's purchasing activity, where the total number of sales varies across the population, whereas the distribution of the logarithmic sequence is used as a simpler form of the negative binomial distribution when consumers who have not purchased during the sample time are eliminated[12]

2.2 Online Shopping Customer Churn

Online shoppers attribute in the factors of non-contractual partnership, using the period of the purchase as a thought, the loss of online shoppers is divided into two types: interruptions and recurrent loss.

Intermittently failing: Irregular failure shows that the clients did not consume the company's goods or services during one particular period. The main key feature is the drop in trading frequency. The consumers not buying the products of a particular retailer on a particular period does not indicate abandonment by the customers. It still shows the chances of the buyer to once again continuing to buy.

Permanently Missing: Permanent Loss indicates the customers' complete abandonment of the merchant and the merchandise. E-commerce companies should not write off the user's credit, because though the client does not use the ID for a very long period, he/she might still use it to login which the company will not be able to identify. The permanent loss means that the client is a complete loss and for some purposes, contributes to the rise of other customers, such as the shift in the client's purchase preferences, the switch in the production with the client no longer finds musing, or the client's demise.

2.3. Overview of Online Customer Churn

After a thorough read on certain consumer churn literary works, it is recognised that there is no evidence on the customer churn outlook for non-contractual bonds. The thesis primarily focussed on the purchasing habit of a non-contractual commercial consumer. Schmittlien Morrison and Colombo presented the SMC model for forecasting market revenue in 1987. The

model acquired the experience of customers through mathematical inference and assessed the future actions of the clients based on customer engagement. To advance the accuracy of the client behaviour prognosis, Liu Xuewei (2006) uses the grouping of naive Bayesian and SMC models to forecast performance. The Naive Bayesian algorithm, the latest model, is equivalent to the SMC model. Dai Yisheng (2010) counted the extent of consumer service of the e-commerce site with the non-contractual relationship along with the SMC model.

The factual results show how when the involvement of the consumer is high the turnover becomes less.

Wu Hong (2015) utilises the SMC model to amplify the rewards of the consumers into the features of customer churn. The research points out that the artificial neural network model's potential advantage is higher. To count the present clients and to identify the predicted model for the assist vector machine attribute reduction analysis, Zhu Bangzhu (2010) chose active degree 0.5 for the point. For the SMC model-rough set-support vector machine, the results indicate that it is correct.

The use of the SMC model to render the future habit of the customer paves the way to a new method for advancing the estimation of non-contract customer attrition. However, when the customer behaviour is expected at an individual's level, the SMC model has the downside of precision. Around the same period, SMC Engineering Research's 149 757 models, focuses on rigid requirements, which does not permit users to be stale and then return being active. Online shoppers are casual, and losing the customer duration threshold needn't mean losing customers eternally. Therefore, this states that there some disadvantages to the SMC model hypothesis in the field of online retail customer churn. In predicting market turnover, the ability to generalise is low, producing a single classification paradigm. Additionally, it is difficult to fix similar issues.

The consolidated model would make use of the multiple forecast approaches to encourage the productivity and steadiness of the forecast performance. The initiation of the combined predicting model and the increase of forecasting correctness have been a mandatory research path in the prediction field.

Chapter 3: Problem Statement

The e-commerce company depends largely on clients who access the website and purchase the items. The approach of identifying the right buyers, getting them to the website and making them buy any of the items is a huge process by itself and requires tremendous efforts and costs to turn some random individuals to a consumer. Hence, the first goal of any company is to keep those clients.

Retention of consumers is a blindfold game where we are not sure initially who is the person who is going to churn, whether they have not bought anything for the past few days or months, whether there is any personal cause or whether they have cracked a good bargain and switched to the rivals, and eventually how to keep them.

Customers of e-commerce who have ordered few items in the past do not buy anything for quite some time. This paper introduces e-commerce consumer repeat transaction modelling based on a few machine learning approaches to boost the estimation accuracy of repeat customers as well as enhance the recognition of high-value regular customers.

This segment discusses both the history and the setting of the issue. In terms of the internal and external context where the problem is located, the problem description shall be described; i.e. entity, culture, etc. It explains the groups or organisations (stakeholders) impacted by the problem, how they are interested, where they are based, and why they have an interest in the topic. The description of the problem also defines the organisational structure, role, services, and processes related to the question.

Chapter 4: Objectives of the Study

Examination of current working models followed by most firms in online commerce.

- To define important e-commerce criteria from the point of view of consumer engagement.
- To consider the customer's perception of the level of service offered by e-commerce businesses and also to describe the areas on which businesses need to change their emphasis
- And have analysis-based feedback to boost consumer service and decrease the rising rate of customer churn.

Here we aim to figure out the clients who do not come for some time with the modelling of consumer repeat visit buy and will eventually concentrate on recovering them by providing some advice (future work).

Chapter 5: Project Methodology

The following point consists of the project methodology:-

- **Defines:-**Engages the on-site client and knows the current situation. Defines the scale, priorities, methodology and timelines.
- **Plan:-**Schedule the case and connect with consistency.
- **Design & Construction:-**The same set of data and procedures.
- **Analysis:-**The data is processed, models are developed and ideas are derived.
- **Deliver:-**Offers consumer feedback through reports and dashboards on negotiated timelines. Show and close the engagement with the final analysis.

Pipeline for Ventures

The project pipeline outlines the blueprint for the project over time. The proposed roadmap vs. the real roadmap (as seen in the following image)



The picture above illustrates how the project was designed using the technique of CRISP-DM, including

Business Comprehension:-Relies on the interpretation from a business viewpoint of the project priorities and criteria. The analyst formulates this understanding as a challenge of data mining and establishes a tentative strategy

Data Understanding:-The analyst continues with practices to get acquainted with the data, recognise data quality concerns and uncover first insights into the data, beginning with initial data collection. The analyst could also detect interesting subsets in this stage to form hypotheses for secret knowledge.

Data Planning:-The stage of data preparation includes all operations to create the final dataset from the original raw data collection.

Repeat Visit Modeling:-The researcher tests, chooses and applies the necessary methods of modelling. Since such approaches such as neural networks have particular criteria concerning the data form. There will be a loop back to data preparation here.

Evaluation:-The analyst develops and selects models that tend to have good output depending on the chosen loss functions. To ensure that they can generalise the models against unseen evidence, the analyst checks them. The analyst therefore subsequently validates that all main

market problems are properly resolved by the templates. The result is the champion model(s) range.

Once the project roadmap was measured based on expected vs. real, we found that it took around one month extra from the planned timetable to prepare results. This is because, before going to the modelling level, multiple areas need to be pre-processed. In this text, the pre-processing steps are explained under the heading of data preparation.

Chapter 6: Business Understanding

ABC is a global apparel company with a turnover of more than one billion dollars. It has 40 factories worldwide.

ABC is a profitable enterprise that uses the e-commerce platform and a few other markets, such as Amazon and Flipkart, to make profits.

They also link up with the few brands with which the goods are made.

Victoria's Secrets, Marks & Spencer, Victory, Speedo, Calvin Klein, among others, are their clients.

The first and foremost challenge for any new business entrant is to gain the confidence of consumers, particularly when you are in the field of e-commerce. Gaining exposure is gradually becoming an uphill challenge, with so many players already present and more and more entrants coming into the market almost every day. Identifying a market and consciously moving for it largely helps. Starting from a niche and spreading is faster and more realistic than the other way around.

The second factor that is a big obstacle is to win the confidence of a prospective client and build a reputation. When the product can not be touched or sensed by the consumer, it is very difficult to persuade them of its consistency, which is the USP in our situation.

High commissions that eventually impact the price of the goods are paid by logistics and payment gateways. Another challenge that is frequently faced by e-commerce setups is the logistics of Cash on Demand as a business. Last but not least, the client should refuse or decline any order, as per our rule, just in case he changes his mind. This raises return rates and total costs and are more prominent for COD orders.

We strive to have a brainstorming session every day to prepare and implement innovative initiatives and new capabilities, not being so acquainted with the dynamic arena of the ever-growing digital space. It is something that we constantly aim to inculcate in our framework to create a digital approach to change the marketing emphasis from brand and services/products to the end-user.

Another significant challenge we faced was the introduction of the well-established payment network and user verification. Digitisation entails hacking and cyber threats that can be effectively exploited against ransomware users by enterprises. We ensure increased focus to try to be protected from these harmful practises, whether it be data manipulation or malicious client targeting brands. Not only in terms of time usage and financial losses but even in terms of consumer reputation, they are a threat.

One of the most difficult elements for any start-up is having client support-oriented services. It is something that start-ups need to reach to inform clients in a situation to discourage them from pressing the panic button and coming up with negative comments on social media.

Customers come first and everything we do must be in line with their wishes and convenience. There may be daunting scenarios that one may encounter with a few clients, but the safest way forward is to take them as examples. [10]

We will concentrate on consumer turnover and retention research in this project. They have noticed over the last few months that consumers who have ordered goods in the last few months are not going again.

Here we aim to find out about consumers who do not come with consumer repeat buy visit modelling for some time and will eventually concentrate on recovering them by providing some advice (future work).

Chapter 7: Data Understanding

Sales data from Mar 2018 to Feb 2020 are included in the dataset.

We took 1-year data for Repeat Purchasing Modeling, i.e. Jan-The-Dec-2019

Data includes 35 features: Order No, External Order No, Order Date, Order Form, Status, Customer Name, Country, State, Area, Email, SKU Code, Design, SKU Definition, Category 1, Sub Category, Scale, Colour, Type, Quantity, Return Qty, Currency Order, Price, Ship Cost, Packaging Cost, Discount, Discount Code, Tax, Invoiced, COGS, Base Currency Invoiced, Gross Margin, GM Percent, Primary Currency

Below Data Dictionary:-

- Order No: -Auto generated Order number in order management system
- External Order No: - Shopify order number sent to the customer
- Order Date: - Exact time of the order
- Order Type: - Pre paid & Cash of Delivery
- Status: - Shipped complete - order shipped to the customer | Cancelled - Cancelled ordered before shipping
- Customer Name: - name of the customer
- Country: -Orders are shipped only in India
- State: - Order placing state
- City: - Order placing city
- Email: - Email id which has placed the order
- SKU Code: - The SKU code (product code) ordered by the customer
- Style: - Style/Colour & size details of the SKU
- SKU Description: - Style/Colour & size details of the SKU
- Category1: - Main category
- Sub Category: - Silhouette
- Size: - Product Size
- Colour: - colour name given by the designer
- Type: -Basic colour
- Quantity: - Quantity ordered per SKU
- Return Qty: -Quantity returned
- Order Currency: -in INR
- Price: -price after product discounts & before cart level discounts
- Ship Cost: -Cost of shipping the product
- Packing Cost: - cost of packaging the product
- Discount: -Discount amount cart level
- Discount Code: -Discount code applied at cart level

- Tax: -tax amount
- Invoiced: -invoiced to the customer
- COGS: -cost of goods sold - approx.
- Invoiced In Base Currency: -invoiced to the customer
- Gross Margin: -Profit made on invoiced amount
- GM Percent: -percentage of profit made on invoiced amount
- Primary Vendor: -Primary vendor who is selling the product
- On Hold Status: -Product's status
- Replacement Order: -Orders which are getting replaced

Order number, Order Date, City, SKU Code, Scale, Colour, Quantity, Invoiced, Gross Margin are some significant features in the results. When modelling the repeat visit for the organisation, these features will be included.

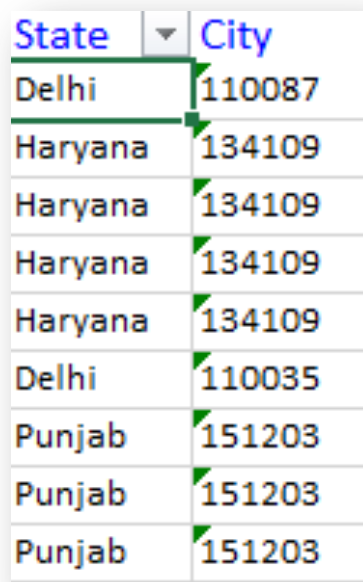
Chapter 8: Data Preparation

Data planning is the most critical step for any enterprise, as it enhances the accuracy of data and increases overall efficiency.

The only methods that have been used for data cleaning are below:

- Creating Specific Order No.: -Dataset includes several order numbers for the purchase of the same product (as it takes one row for each buying unit) The solution is to generate a new order number by combining Order No and SKU Code
- Removing Outlier: -Excluded the outlier (these are retailers) depending on the invoiced number (will handle them separately)
- Federal column washing
- A considerable amount of time has been spent on cleaning the area. The key difficulties with the town columns are:-
 - Specific names for the same cities. Examples include Bangalore, Bengaluru

- Total address or only the pin code in the town column



State	City
Delhi	110087
Haryana	134109
Haryana	134109
Haryana	134109
Haryana	134109
Delhi	110035
Punjab	151203
Punjab	151203
Punjab	151203

Fig 8.1

- Some towns like Bangalore are branded with Tamil Nadu, Delhi, etc.

State	City
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Tamil Nadu	BANGLORE
Karnataka	Banglore
Karnataka	Benglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Banglore
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru
Karnataka	Benguluru

Fig 8.2

- Creating Unique Order No.: - Dataset contain multiple Order number for same product purchase (as it's taking one row for every unit purchase) Solution to this is creating a new order number by matching Order No and SKU Code
- Removing Outlier: - Based on Invoiced amount removed the outlier (these are retailer. Will treat them separately)
- Cleaning State column
- The invoicing function has some negative values: -We found that some disappointed consumers have been compensated with gift products upon checking with the company. It has been recommended to exclude the negative value from the invoiced value.

Invoiced
-0.07
-0.60
-0.12
-0.03
-0.77
-1.97
-0.03
-0.07
-0.05
-1.00
-50.00
-5.00
-125.00

Fig 8.3

- Exploratory Data Analysis:-Using Heatmap to figure out the connection between the features

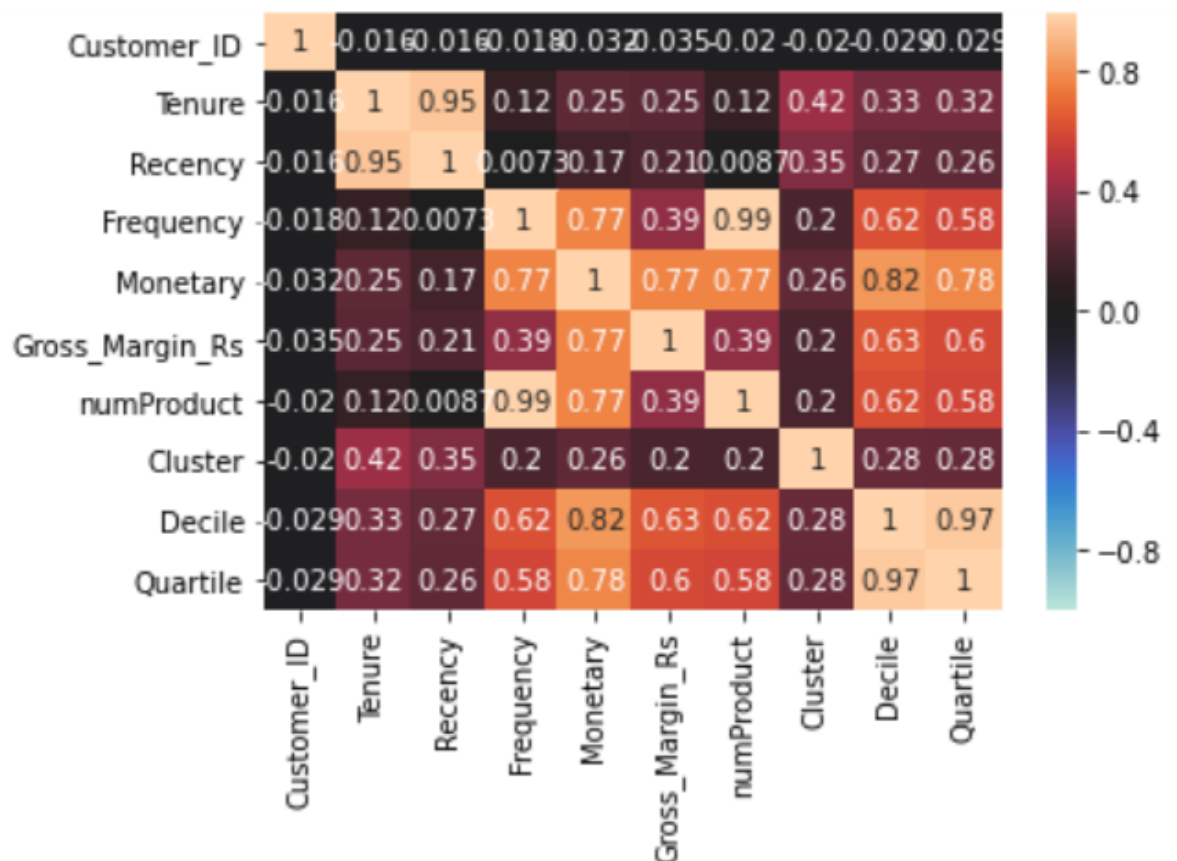


Fig 8.4

- Exploratory Data Analysis:-Top 5 States with the largest revenue

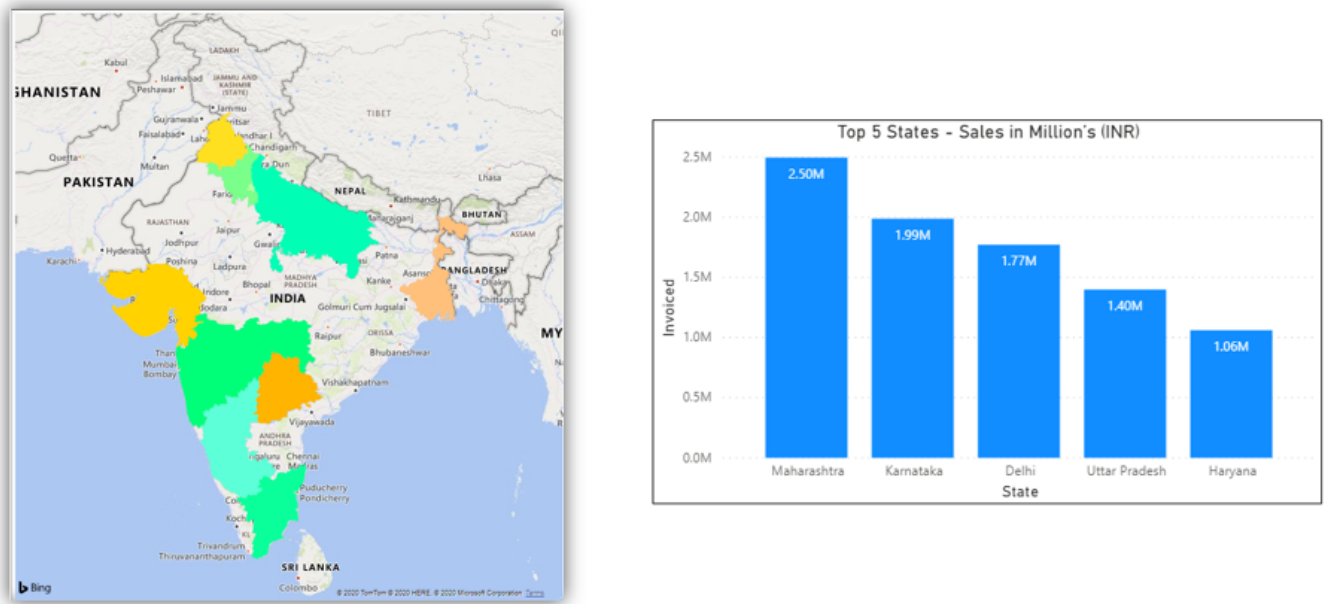


Fig 8.5

- Exploratory Review of Data:-Top 5 Sales & GM Cities

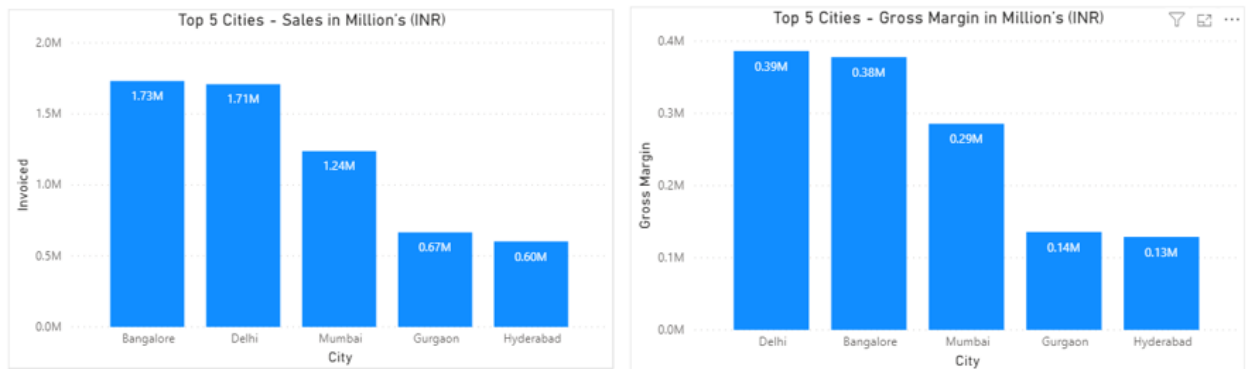


Fig 8.6

- Exploratory Review of Data:-Top 5 Gross Margin Cities

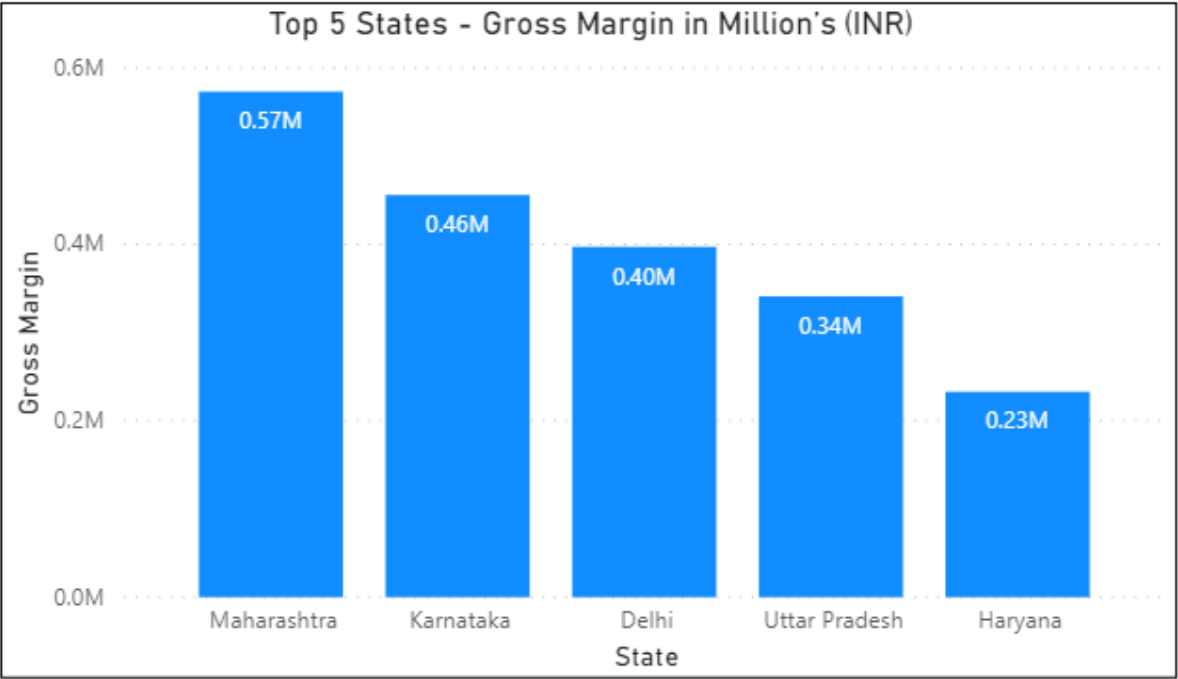


Fig 8.7

- Exploratory Data Analysis: - Top 5 Cities Gross Margin

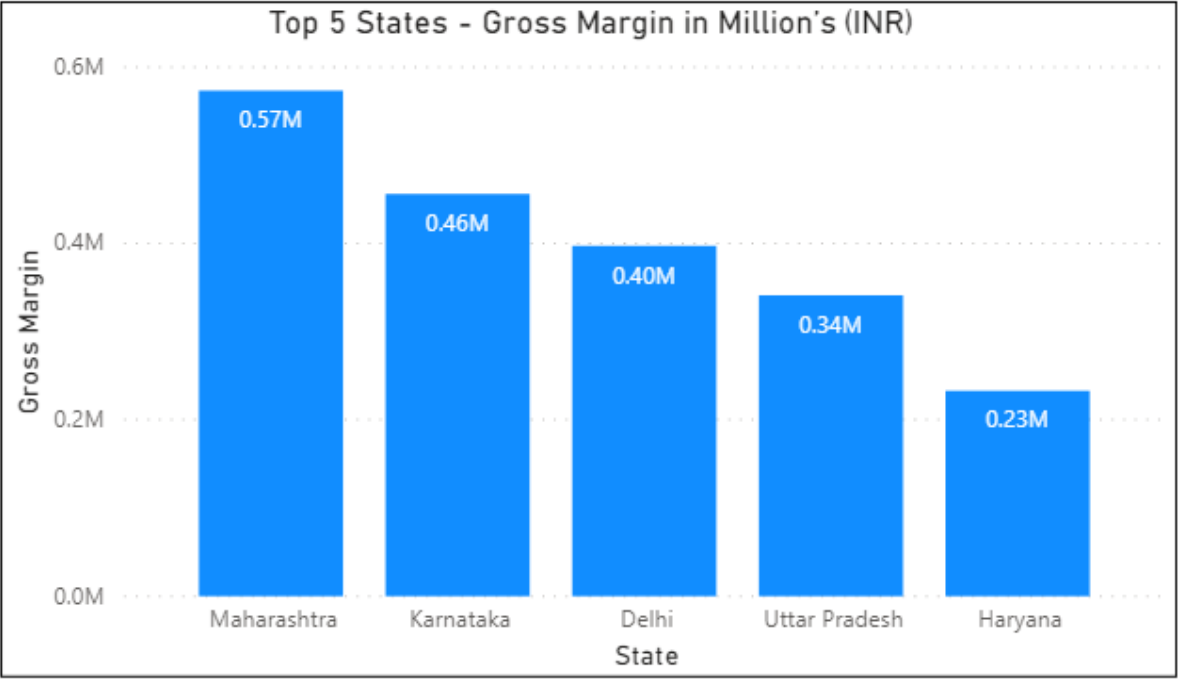


Fig 8.8

- Exploratory Data Analysis: - Monetary & GM Quartile wise

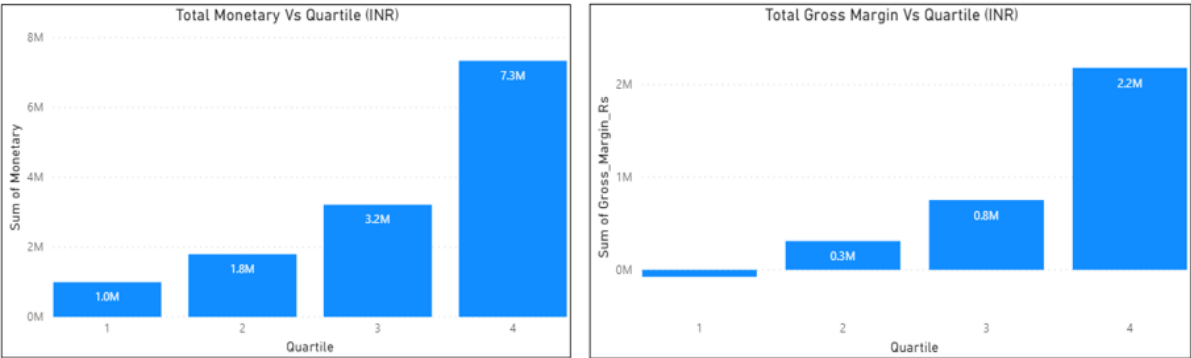


Fig 8.9

- Exploratory Data Analysis: - Retain Per Quartile

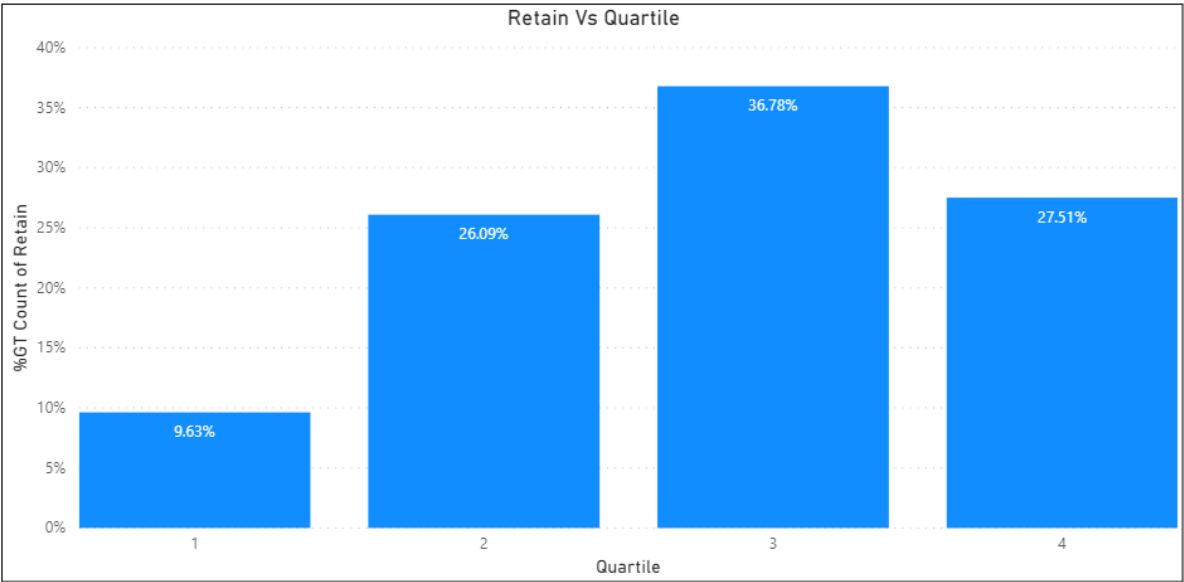


Fig 8.10

- Exploratory Data Analysis: - Customers Retain/Lost

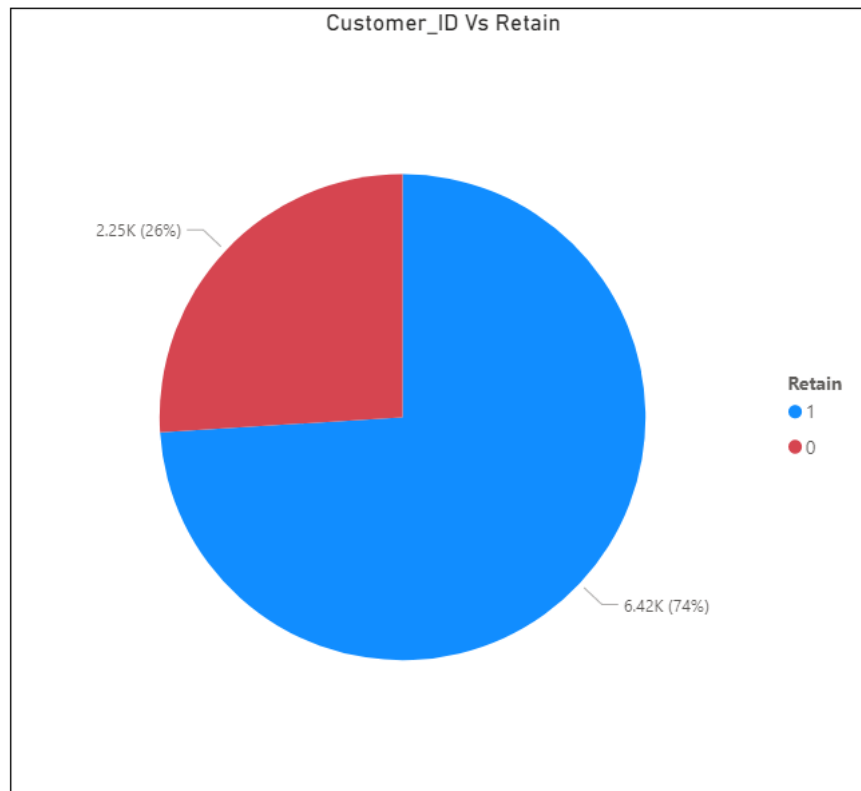


Fig 8.11

Chapter 9: Data Modeling

Three data modelling approaches have been considered to create a more efficient and reliable repeat purchasing model. Data modelling aims to determine which clients are going to purchase the item or not. It is a process by which from a given data set, various models, summaries and derivative values are discovered. The following explains the main pointers that can advance the purposes of this report, which are KNN, Random Forest and Decision Tree.

The three models were used to estimate the retention of consumers and the model assessment was conducted using the consistency score and the uncertainty matrix. The information below is

9.1 K-Nearest Neighbors (KNN): - An easy algorithm that contains every available report and categorises new reports based on similarities. In the early 1970s, KNN worked as a non-parametric instrument to provide statistical estimation and pattern recognition.

Training Accuracy Score – 97.64%

Test Accuracy Score – 97.00%

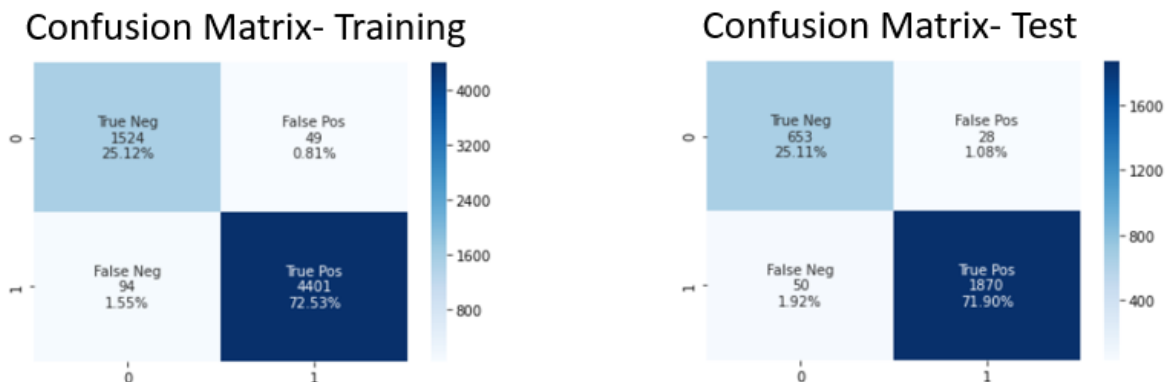


Fig 9.1

9.2. Decision Tree (DT): - A decision tree is a decision support system that uses a model which looks like a tree to decide its possible effects, implications of chance events, energy cost, and utility. One way to show an algorithm that requires control statements is to imagine it.

Training Accuracy Score – 100%

Test Accuracy Score – 97.42%

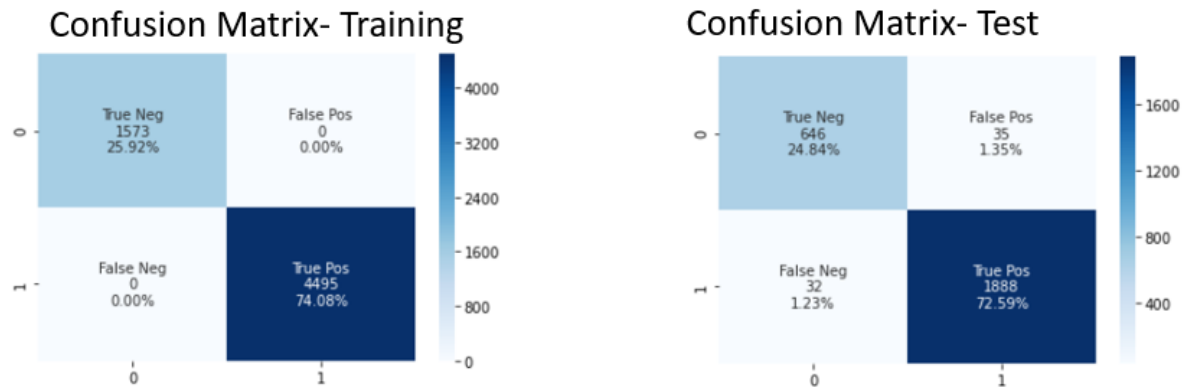


Fig 9.2

9.3. Random Forest: - These forests are a group learning method for categorization, retrogression and various operations that work by generating training time and forming a class that is classification and retrogression of trees. The habit of overfitting their preparation set for decision trees is corrected by random forests of decision making.

Training Accuracy Score – 100%

Test Accuracy Score – 97.92%

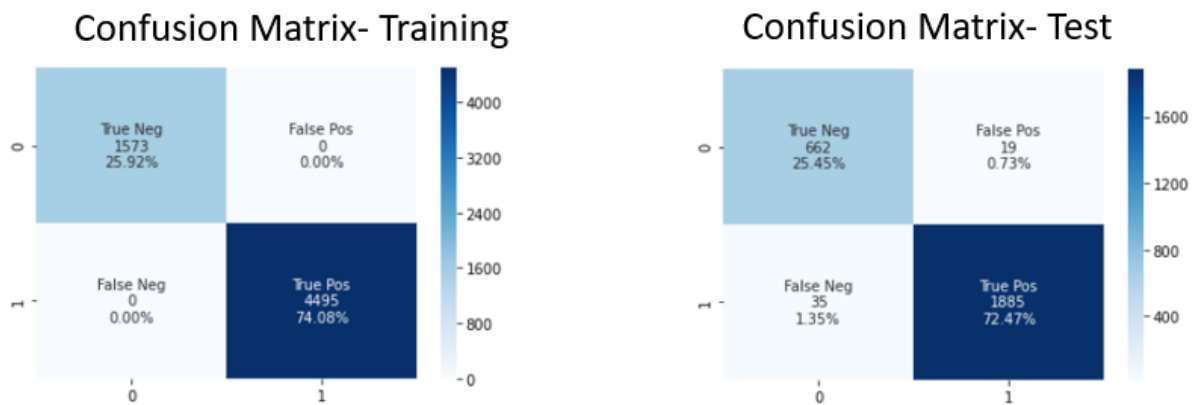


Fig 9.3

Chapter 10: Data Evaluation

We also discussed the three modelling approaches for repeat consumer visits under the modelling section. For these models, the measurement criteria are the accuracy score and the uncertainty matrix. The findings for the three models are found below:-

- The KNN algorithm offers accurate train data which is 97.64%, while test data accuracy is 97%.
- The Decision Tree algorithm provides 100 % accuracy on train data, while 97.42% is accurate on test data.
- The Random Forest algorithm train data is accurately 100 %, while 97.92% test data is precise.

We find that Random Forest and decision tree function very well for training data (100 per cent accuracy score), while KNN, Decision Tree and Random Forest do the same for test data. Yet Random Forest performs well upon digging down to decimal values. Hence, as our modelling tool, we will take up the Random Forest algorithm.

Chapter 11: Deployment

In this project, we developed a model for repeated visits. To find the most reliable solution, the model has been developed and tested through several algorithms such as KNN, Decision Tree, Random Forest. We haven't deployed the model for the time being.

Chapter 12: Analysis and Results

In this section, the outcome of the project and review are discussed. The product of the technique is included in this section, i.e. the numerical or descriptive work that was done. For eg, the findings are included in this section if a SWOT analysis is done. Similarly, if a scientific study is done, this section includes the computational outcomes.

- **KNN Result**

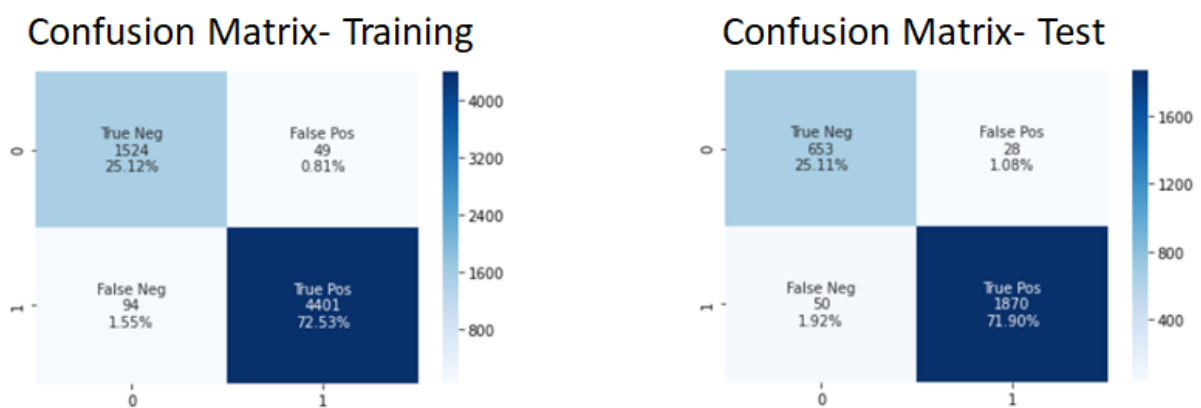


Fig 12.1

- **Decision Tree Result**

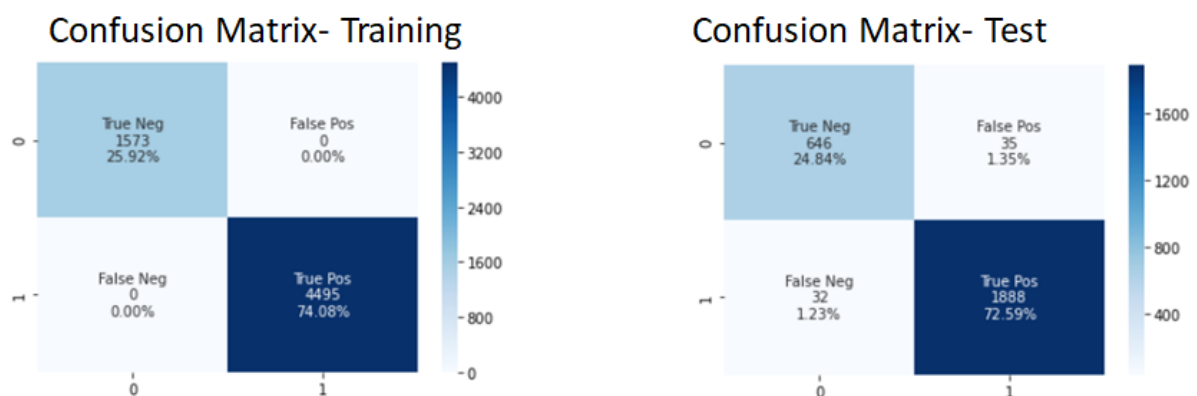


Fig 12.2

- **Random Forest Result**

Training Accuracy Score – 100%

Test Accuracy Score – 97.92%

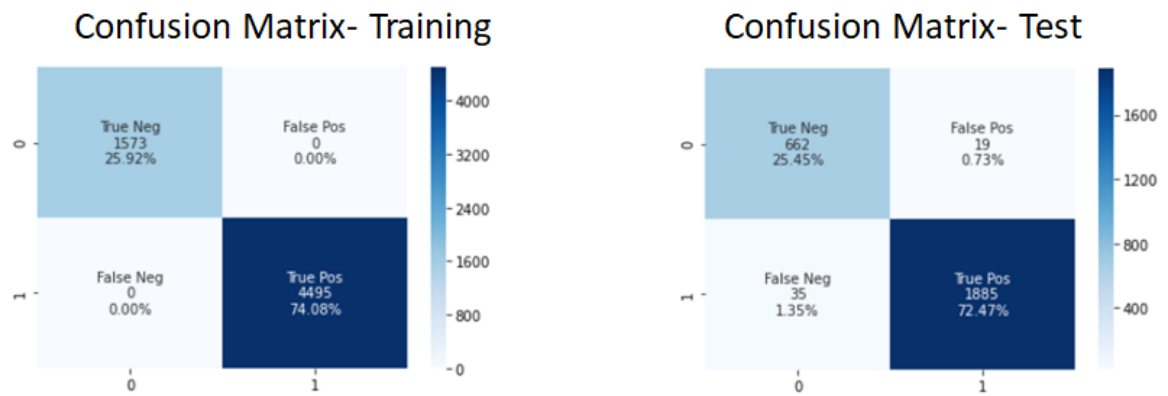


Fig 12.3

We find that Random Forest can help forecast the replicate modelling analysis using the test matrix such as accuracy score and uncertainty matrix model contrast.

Chapter 13: Conclusions and Recommendations for future work

Now a days, E-commerce is at boom. Companies are investing considerable amount on getting customers on their site. Using various coupons and value added products, they are attracting customers to buy from them. Soon they become addicted to particular brand or product and will start buying at normal price. But, this habit might get stopped with experience. There are few brands which started making loss because of these bad experiences.

To overcome customer's bad experiences and retain the loyal customers, companies started investing in repeat visit modelling. This has worked so well that popular E-commerce companies like Amazon, Myntra and flipkart started no question asked return policies.

no question asked return policies is an example of repeat visit modeling where companies making more loyal customers.

Hence repeat visit modelling is the backbone of E-commerce business and companies has to work seriously towards it.

E-commerce is here to stay. And since it is a fairly new market place, everyone is busy to have a go. This has boosted the careers of such people who did not dream of careers. There is place for everyone here. Customer base analytics is the only measure with which one can dominate the industry. It is also necessary to add new varieties, but in a manner that is appealing. For instance, HM has been at the receiving end of social media censure for its unethical approach of fast fashion. Fast fashion has been responsible for a considerable amount of environmental waste. The brand's environmentally conscious line has gathered censure because they are only 5% of the total of the brand's clothing lines. Hence proved that variety is not everything since the very concept of fast fashion is variety.

Bibliography

- [1] Ali Tamaddoni Jahromi, “Predicting customer churn in Telecommunication service providers”
- [2] Athanassopoulos A D, “Customer satisfaction cues to support market segmentation and explain switching behaviour”, Journal of Business Research, vol. 47, pp.191-207, 2000.
- [3] Bhattacharya C B, “When customers are members: Customer retention in paid membership contexts”, Journal of the Academy of Marketing Science, vol. 26, pp.31-44, 1998.
- [4] Wu Hengliang, Zhang Weiwei, “A Customer Churn Analysis Model in E-business Environment ”, Online – Semantic Scholar.org
- [5] K.A. Venkatesh, Mihir Dash, “REPEAT-PURCHASE MODELLING FOR E-COMMERCE WEBSITES”
- [6] K Nearest Neighbors – Classification,
https://www.saedsayad.com/k_nearest_neighbors.htm#:~:text=K%20nearest%20neighbors%20is%20a,as%20a%20non%20parametric%20technique.
- [7] Decision tree, “https://en.wikipedia.org/wiki/Decision_tree”
- [8] Random forest,
[https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,prediction%20\(regression\)%20of%20the%20individual](https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,prediction%20(regression)%20of%20the%20individual)
- [9] Rahul Bhagat, Alex Lobzhanidze, “Buy It Again: Modeling Repeat Purchase Recommendations”
- [10] Fashion E-commerce Start-ups: Challenges in the Indian scenario,
<https://www.indiaretailing.com/2020/02/25/fashion/fashion-e-commerce-start-ups-challenges-in-the-Indian-scenario/#:~:text=The%20first%20and%20foremost%20challenge,in%20the%20e%20commerce%20sphere.&text=When%20the%20customer%20can't,our%20case%20C%20is%20the%20USP>
- [11] Ehrenberg, A.S.C. (1988), Repeat Buying 2nd Ed., New York: Oxford University Press
- [12] Lilien, G.L., Kotler, P. and Moorthy, K.S. (2003), Marketing Models, New Delhi: Prentice-Hall India
- [13]. Liu Xuewei. The prediction of e-commerce customer churn based on Pareto/NBD+ naïve bayesian combined model [D]. Sichuan University, 2006.
- [14]. Dai Yisheng, Shen Peilan, Sun Hongxia. The research of customer churn prediction based on the Pareto/NBD model [J]. Science and technology and engineering, 2010.
- [15]. Wu Hong. The Comparison and empirical study on the model of e-commerce customer churn [D]. Capital Economics and trade university, 2015.
- [16]. Zhu Bangzhu. The prediction of e-business customer churn based on SMC-rs-live model [J]. Systems engineering theory and practice, 2010(11):1960-1967.
- [17]. Ren Jianfeng, zhang Xinxiang. The modelling and prediction of the e-commerce customer churn[J]. Computer simulation, 2012(05):363-366.

Appendix

Plagiarism Report¹

Repeat visit modeling in E-commerce			
ORIGINALITY REPORT			
10%	10%	0%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	download.atlantis-press.com Internet Source	3%	
2	www.indiaretailng.com Internet Source	2%	
3	Submitted to novi Student Paper	2%	
4	Submitted to Sogang University Student Paper	1%	
5	www.termpaperwarehouse.com Internet Source	1%	
6	digitalcommons.calpoly.edu Internet Source	1%	
7	Submitted to University of Edinburgh Student Paper	<1%	
8	www.inmybangalore.com Internet Source	<1%	
9	www.scribd.com Internet Source	<1%	

¹Turnitn report to be attached from the University.