National College of Ireland

Customer Behaviour Prediction

MSc Research Project
MSc Data Analytics

Nikhil Kulkarni
Student ID: X18201130

School of Computing
National College of Ireland

Supervisor: Manaz Kaleel

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Nikhil Kulkarni ……. ……

Student ID: X18201130 ……

Programme: …Msc Data Analytics…..…… **Year:** 2019-2020 ……..

Module: Research Project ……

Supervisor: Manaz Kaleel ……

Submission Due Date: 28/09/2020 ……

Project Title: Customer Behaviour Prediction ……

Word Count: ……8012…… **Page Count**……..…21…..

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Nikhil Kulkarni …… ……

Date: 28/09/2020 …… ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Behaviour Prediction

Nikhil Kulkarni
X18201130

Abstract

In this paper customer behaviour prediction models are applied on data which is real-world e-commerce data over a period of 4.5 months, The main aim of the research is to predict customer buying habits and recommend items according to behavioural data of customers. For predictive research, insight into consumer actions may be obtained to improve business decision-making. Comparing the statistical approach to data mining in predicting customer behaviour, In this study Comparison analysis is performed in between Logistic regression, LightFM and Multilayer perceptron with label encoding and hyperparameter optimization. As a result, MLP achieved 92.4 % accuracy, An accuracy of LightFm is 81.6% and For Logistic 79.40. Altogether MLP displayed better results but execution time could have been reduced.

1 Introduction

Customer Behaviour Prediction is a method of defining rising customer behaviour. This is used to maintain valuable customers because maintaining the current organization's clients is cheaper than gaining new customers. This helps to understand how consumers spend their time on websites of online shopping, how much time they spent looking for the products, most commonly purchased products and what quantity of items they bought. People are quite busy nowadays. (Bradlow et al., 2017) Due to a lack of time to go for shopping, consumers approach online shopping. Online shopping has now become the third most common internet activity, trailing by email messaging and web surfing. The primary reason to focus on customer behaviour prediction is the relationship structure between buyer and retailer depends on learning consumer behaviour in online environments for improvement of the online sellers of products(Be *et al.*, 2019). To increase the client base and recommend accurate products customer behaviour prediction and understanding is the key factor. The data regarding viewed products, added into cart products and purchased products, login time, logout time, categories of the products and price range are crucial factors in understanding customer buying habits and recommend desired similar products. Macro-level predictions such as average consumer behaviour are focused by most of the e-commerce companies, but depending on such features is not helpful as the behaviour varies from one consumer to another. It is logical to rely on predictions at the micro-level such as individual customer interactions, which gives them a suitable idea of customer choices. One challenge emerging in the research is behavioural prediction, namely the ability of predictive analytics to predict individual decisions and behaviours accurately. Also, predicting actions can help to inform the theory of actions. Prediction of behaviour is increasingly important, but there are still drawbacks to the current traditional modeling approach, namely high reliance on observable objective data and inability to consider micro-level decisions and behaviours that collectively facilitate macro-level behaviour. There are no descriptive variables or insufficient to

distinguish differences and similarities between customers. This research data is used from real-world, e-commerce websites over a period of 4.5 months and that would help web shoppers to make better decisions by recommending better products over time through multiple channels all over the world. Research contains a comparison analysis between several models like Logistic Regression, Light FM model which is top for any recommender system and MLP. In the comparison of these models proposed recommender system is designed and would recommend user better products as per collected data. When a customer visits an e-commerce website, certain criteria are followed for finding a perfect product like product type, price range, vender etc. In this research better algorithm is proposed for predicting properties of products by "viewed" items and "inCart" items for every visited user. And another important aim of this research is to find abnormal users of e-commerce websites so better data can be collected for building a better model for customer behaviour prediction.

2 Research Question

1. How well Predictive analysis help E-commerce retailers to understand customer behaviour?
2. How to recommend products to customers by products visit history?

3 Literature Review

In this section, relevant literature is reviewed on customer behaviour and prediction techniques, various approaches and methodologies are used to solve problems and measuring performance to test models.

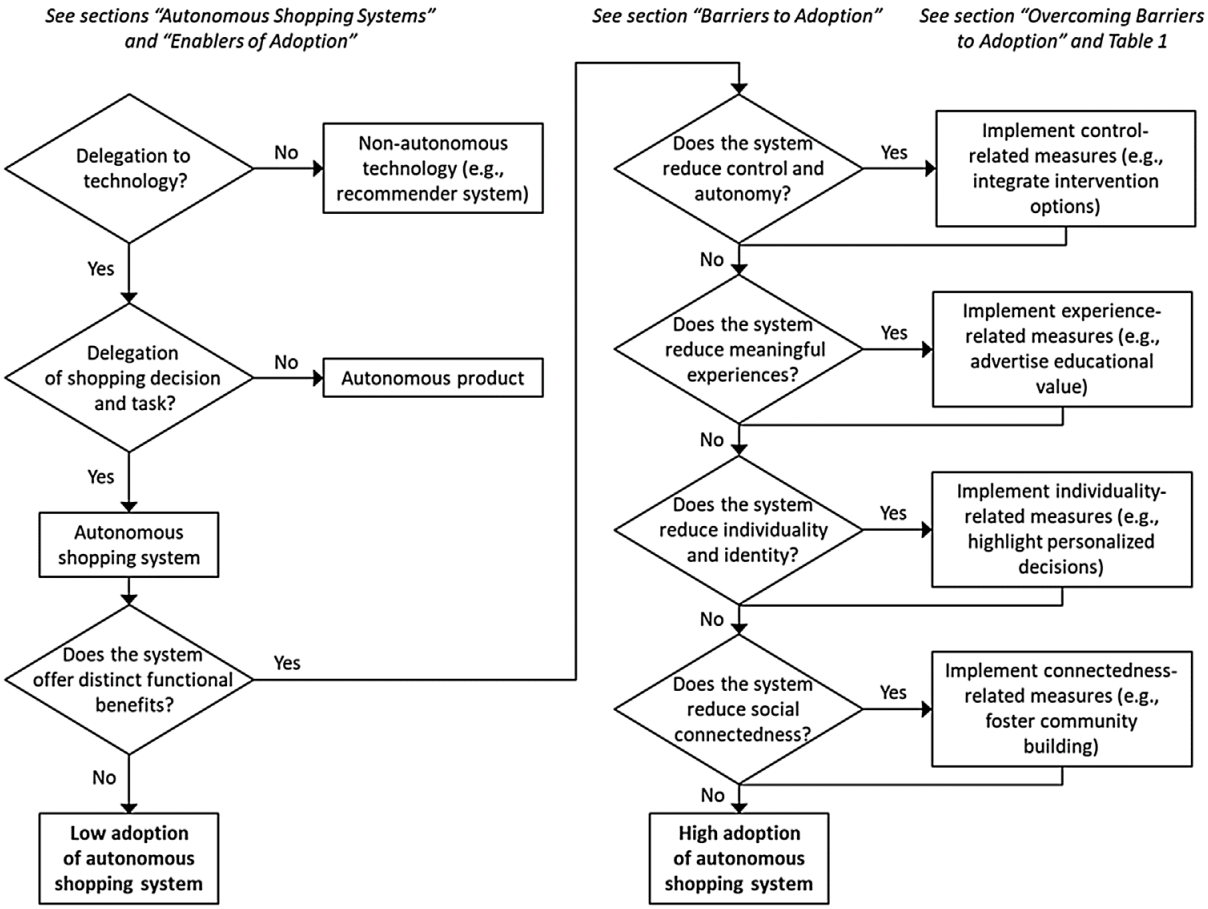
According to the research of (Bradlow *et al.*, 2017)Supermarkets are one of the most comprehensive sectors of today. They give their customers the right to select their preferred product based on their needs and satisfaction and thus lead to customer loyalty during shopping. Customer satisfaction plays a critical role in the shop and this is obtained by studying most customer's shopping habits. In this study researcher identified some demographic variables to evaluate the customer's shopping pattern, variables are the economic status of the market, the daily item of the customer sets the purchasing pattern, the number of times the customer visits the store, the location of the store, the home of the customer, etc. The apriori algorithm used regression to get results. According to the research author, a conclusion was drawn that supermarket stores would have a clear image of the types of customers who are most likely to come to their supermarkets and favourite items. Prediction of regular geographically distant things may be utilized to develop. Classification and categorization of consumers is a vital aspect of Bradlow 's research. Customers are classified and categorized according to their shopping patterns and their age groups, sex, economic status etc. The pattern of consumer buying behaviour is grouped into five relationships, such as Purchased Products, Buying Region, Purchasing Time and Purchasing Frequency and Sales Response Promotions.

As per (IbukunT *et al.*, 2016) research, most of the studies examined the prediction of customer retention and institutional datasets were used more for prediction compared to other types of datasets. Besides, this study showed that data mining is primarily used for analytical purposes by contrasting the statistical approach to data mining in predicting customer behaviour. comparatively, the Artificial Neural Network is the most widely used method for data mining to predict consumer habits. The research study concludes that work is underway to predict consumer behaviour and is of utmost importance to the company. Many of the research discussed the modeling of customer retention and this was done primarily by using organizational data as a predictive dataset. Data mining techniques have been more

commonly used to predict customer behaviour compared to statistical approaches and Artificial Neural Network is the most widely used data mining method.
Research of (Xu, Yang and Ma, 2018) states that the evolving exposure of department stores and e-commerce organizations to customer actions and satisfaction and the growth of social media, as well as online systems, has stimulated the development and research of user experience pattern recognition, user network analysis, consumer feedback topic identification, text-based sentiment analysis etc. Ethical problems have also played a significant role in the implementation of big data, user behaviour and input mining, after the rise of the Internet of Things and the social media network. Deng's work focuses on tracking related emotional trajectories over time across all the big customer feedback problems at Skype. However, through using unstructured textual customer feedback data and consolidated user behaviour telemetry data, it also provides a medium for both the study of consumer emotions and tracking how emotions change over time concerning various important issues. The author's conclusion from the research is that text-based customer feedback is user-generated content that can include dirty words and sounds and identify both of these conditions by filtering out biases that may lead to misleading results.
(Agrawal *et al.*, 2018) notes that customer retention is also a significant consideration for future consumer base development in the retail sector. The rate of retention used to predict growth is as important as financial revenue. Despite rising market pressure, companies are keen to keep the turnover rate as low as possible. Built to construct a non-linear classification model, the Multilayer Neural Network. The model of churn prediction works on user characteristics, service, usability and contextual features. The probability of retention is predicted, as are the deciding factors. The trained model then applies final weights to these functions and calculates the customer's retention probability. Since the model also points out the retention causes, businesses may analyse the reasons behind these causes and take measures to eliminate them. Have seen the holes in the process of implementing a solution. This has also been used to derive several parameters that have been seen to affect retention. The multi-layer Artificial Neural Network model used to solve this problem has resulted in 80.03 percent accuracy. The model also shows a list of attributes that can be used to distinguish retention parameters directly and inversely related to the retention rate. This research seems to be an efficient technique for companies to evaluate which metrics can be used to maintain clients and avoid their loss to competitors.
(Doan, Veira and Keng, 2019)Their study finds retailers make decisions based on comprehensive consumer and product databases to fully understand customer behaviour and buying trends to better engage consumers. It has been a challenging job, as customer modeling is a multifaceted, time-dependent problem. The most effective way of tackling this problem is obliquely through task-specific supervised learning prediction issues, with some customer modeling literature specifically simulating their potential transactions. The author recommends simulating their potential transactions as a solution to this. In research, the author has proposed a method for generating realistic basket sequences that a given client is likely to buy over some time. Using Recurrent Neural Network (RNN), the entire sequence of transaction data, customer embedding representations are learned. Given the customer status at a specific point in time, the Generative Adversarial Network (GAN) is trained to produce a consolidation basket. The freshly formed basket is then directly fed into the RNN to update the status of the customer. Upon analysing the customer's embedding with the LSTM, the author also concludes that they generate a product basket conditioned on the customer's embedding using the GAN generator. The generated product basket is redirected to the LSTM to generate embedding of new customers and the process repeats. Researchers have tested as further tests that there are specific sequential correlations between the items in the

generated and the actual data and orders produced are hard to distinguish from the actual orders.
(de Bellis and Venkataramani Johar, 2020)Developed a cross-disciplinary strategy, building on research in marketing, psychology and human-computer interaction, to identify obstacles to the implementation of automated shopping systems. They identify different psychological and cultural barriers and recommend solutions to address these barriers across the customer's entire online retail environment. The author explains and distinguishes independent shopping networks from related technologies. Second, they identify the main obstacles to the implementation of automated shopping systems. Second, they propose practical measures for companies to tackle these problems to adoption.

Figure 1



(de Bellis and Venkataramani Johar, 2020)

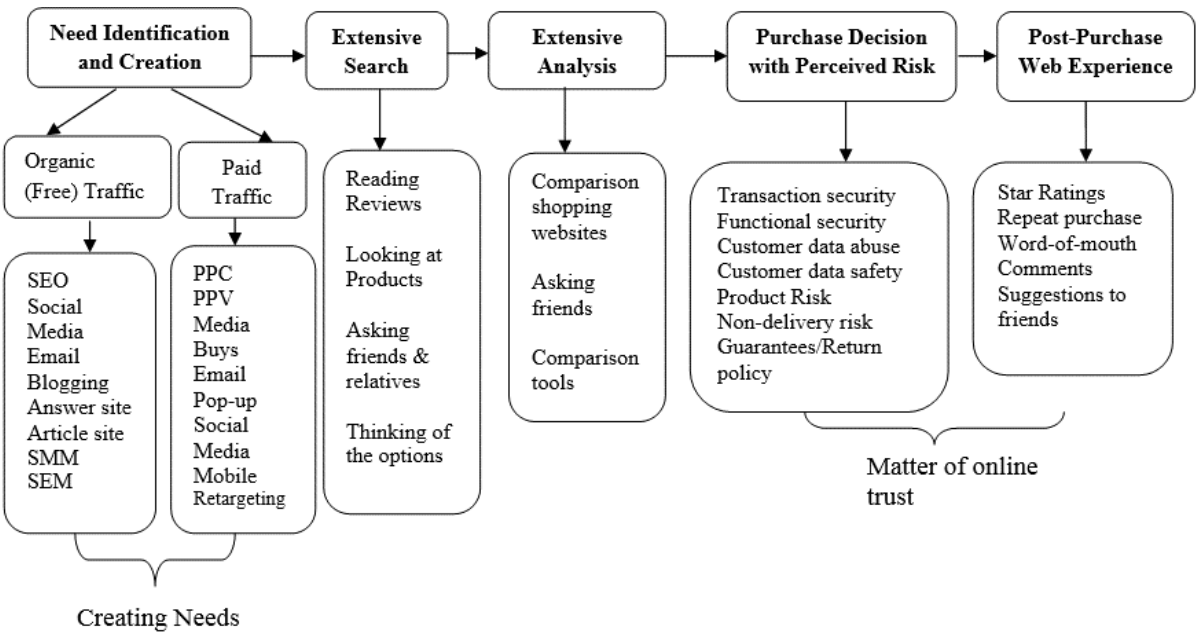
(Wang, Zhang and Ren, 2019)The author has done comprehensive work on consumer behaviour learning. V carious groups of customers with different patterns of energy consumption in the new Smart Grid. Customer's energy use behaviours are defined as customer behaviour. Load forecasting throughout the grid might have a significant benefit if customer intervention could be acknowledged. Research suggests an innovative method that combines various types of customers with their specified activities and then forecasts a load of customer regions to improve the accuracy of the ultimate grid load forecast. The suggested solution to load forecasting in the Smart Grid has three important advantages. Identifying customer behaviour not only improves predictive efficiency but also has low computational

costs. SCCRF can accurately model the load forecasting issue with one customer, at the same time, identify key features to determine the energy consumption pattern. This research will theoretically encourage studies in related fields. Furthermore, learning customer behaviour about aggregate customers provides a particular approach to support better decision-making for particular products in a complex market environment. It is more constructive to investigate further in other areas of business. The results of the assessment also suggest that the proposed SCCRF has been successful in selecting and predicting features. SCCRF can be used in other relevant fields of analysis.

Research on (Be *et al.*, 2019) explains that e-commerce is growing rapidly and people don't prefer walking around shopping malls. Great competition in e-commerce attracts consumers by giving them even more. To attain this need to categorize current customers on shopping websites and to provide them with further sales, the income from the e-commerce sector is increased by extracting the shopping dataset. This work concerns an approach that defines the key customer using the centrality measures by considering the transaction information of the shopping dataset. Machine learning techniques are now in place to overcome the shortcomings of the above methods. A significant customer could be identified. More services can be offered to the user. The centrality calculation approach is also much more effective than the web-based mining process. Accuracy of details on customer behaviour is more important to the method of evaluation than to the method of web use.

(Krishnan *et al.*, no 2017)The author has studied the marketing goal of fulfilling the needs and desires of the expected customers better than the competitor and more competitive. Consumer behaviour is a study of how individual consumers, organizations and businesses select, buy and use products and services to fulfil their desires and wishes. Consumers ' buying behaviour has also become a hot subject of marketing. The consumer decision-making cycle consists of a compilation of five phases of operation subsequently offline customers in the case of physical stores. Issue Identification, Information Search, Alternative evaluation, Purchasing decision, Post-purchase behaviour.

Figure 2



(Krishnan *et al.*, 2017)

Figure 2 explains the customer behaviour 5 stage model in early 2000.

1. Need Identification to Need creation – To understand the need for buyer author has adopted two different approaches 'Organic' and 'Paid'. While online shopping customer uses different digital platforms. Potential customers visit occurred without driving them to the website is organic traffic, whereas In paid traffic buyers are intentionally moved to the specific products. These are attempts of triggering buyer's intention of buying.
2. Extensive Search – With analytics tool like Google Analytics, provides detailed information about potential customers. This search provides accurate information about the search history of the buyer. In this research putting relevant products forward triggers customers' attention.
3. Extensive Analysis - When choosing from his awareness environment, there are many analytical methods used by the client. Check for the need at the very first customer satisfactory object, then he goes searching for some advantages from the product and eventually each product, as a related profit to meet his desires behind it.
4. Purchased Decision and perceived Risk - Online shopping entails a high perceived risk that includes transaction security: financial risk, vulnerability to card information fraud, misuse of consumer data, risk of non-delivery. Product risk includes a product that might not be as defined. Functional risk involves a product that does not work as expected. Unlike the policy of guarantee and return, all these risks produce a question mark on confidence on the website.
5. Post Purchase Web Experience - A happy customer implies that others are worth knowing about their value and can cause traffic to the same web store. For some different goods or services, also he may be a frequent and repeat customer of the same web store.

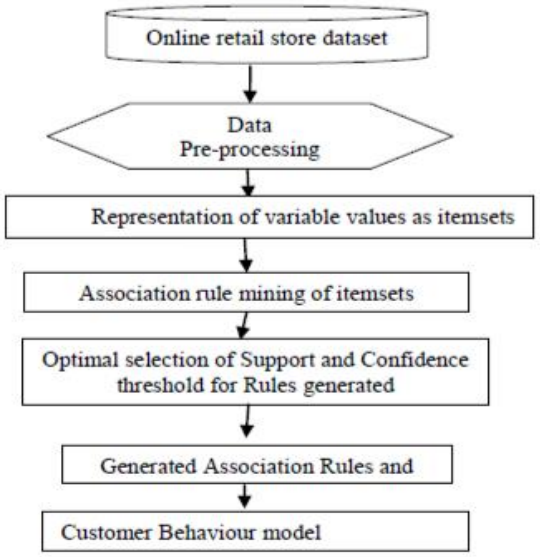
Customer purchase patterns and factors that influence the decision-making process of online consumers have shown similarities as well as differences between online and offline customers. ungovernable variables affect the actions of both types of consumers. In the case of offline customers, 4P marketing mixes are considered to be the most controllable devices influencing customer behaviour. Where in, web users, the set of various factors observed during virtual interaction appears to be controllable factors that affect online shopping behaviour.

(Badea (Stroie), 2014a) has researched that artificial neural networks incorporate a range of features that enable their use in financial and economic applications. Backed by flexibility in coping with different types of data and high accuracy in forecasting, these approaches give substantial benefits to business activities. This paper discusses how consumer behaviour can be represented using artificial neural networks, based on information derived from traditional surveys. Results point out that neural networks have a powerful discriminatory power and typically show better results than traditional discriminatory study. The research author has shown that the results of classification with ANN are superior to those obtained by classical discriminant analysis. Although the data set available was very small, neural networks produced high detection levels for the target category and delivered good results when tested on out-of-sample data, making them a good option for improving marketing strategies and

decision-making processes. The fact that ANN is more time consuming concerning the configuration steps of the model is counterbalanced by fewer prior data transformations and assumptions that are needed to be checked compared to the discriminant analysis. Nevertheless, caution must be taken while training ANN, because it may be subjected to an over-adapting phenomenon. Future directions in the study of the performance of ANN in classification matters that suggest using second derivative optimization algorithms when changing network weights, as these may provide superior results compared to the classical gradient descent back-propagation process.

(Kanade, 2018) Customer behaviour models are typically based on data extraction of customer data and each model is designed to respond to one question at a time. Predicting customer behaviour is an unpredictable and daunting activity. Developing models of customer behaviour requires the right approach and methodology. Once a prediction model is developed, it is hard to modify it for the benefit of the marketer to determine precisely what marketing measures should be taken for customers. Most of the customer models are basic. To predict customer behaviour using a traditional online retail store for data collection and to extract important patterns from consumer behaviour data association models are created.

Figure 3



1: Customer behaviour prediction model using rule mining approach architecture.

(kanade , 2018)

The optimum rule generation occurred at a minimum support and trust level of 0.1 and 0.2 are results from the regular pattern of mining. This research has been able to develop and apply an association rule mining model to predict customer behaviour. This has established interesting periodic market buying patterns that have appeared in the dataset of the online retail store and have weakened strong association rules. The performance is affected by the accuracy and dimensionality of the dataset and the complexity of the selected features of the model. A major part of customer service is played by Customer retention management. Everywhere now, telecommunications companies are focused on seeking high value and potential customers churn to raise sales and market share. It is understood that attracting new customers is more competitive than holding existing customers. There is growing concern

that customers are leaving the company for unknown reasons. The client's churn operation by using a variety of data mining techniques is predicted by the author. At the end of the day, that will help to assess the behaviour of the customer and to decide if he is a churning customer. In a nutshell, the researcher agrees with the fact that the Bagging and SMO algorithm is more than 99.8 percent accurate, but if we develop a model using SMO from data collection, it takes more than 10 times more time to construct a classifier, so using the elegance concept, Bagging is the better choice.

(Valecha et al., 2018) Initially analyse the relationship between consumer behaviour and the purchase of product based emerging factors like environmental, organizational, individual factor and the interpersonal factor. The study proposed a time-evolving random forest classifier that optimizes complex feature engineering to predict customer behaviour that has a significant impact on the choice of buying a product. Random forest classifier results are more robust than any other simple machine learning algorithms. It also shows that the customer's buying behaviour depends on personality characteristics and the environment in various regions. Machine learning approaches are used to predict customer behaviour that provides acceptable accuracy. This research was based on a random forest algorithm that achieved 94 percent accuracy.

(Orogun and Onyekwelu, 2019) The study indicates with behavioural computing and analytics approach in such a way that deeper insight into customer behaviour can be obtained to enable a predictive analysis to improve business decision-making that predictive analytics can be used to predict customer behaviour through. Predictive analysis is used to predict future trends, events and behaviour based on data intended to facilitate better decision-making. Predictive analytics is used mainly for marketing purposes, to predict customer behaviour as companies begin to monitor consumer reactions and transactions and then to use it to enhance their marketing strategies. Customer behaviour can move rapidly to new needs and changes in life. quantitative data can be less predictive than behavioural data. Through a behavioural approach to informatics and analytics, where there is a behavioural data framework that maps transaction data to behavioural data and behavioural analysis that identifies behavioural patterns, it is expected that a deeper understanding of consumer behaviour can be obtained in order to make customer behaviour more accurate.

(Xu, Yang and Ma, 2018) the study explains that to gain more market, traders also carry out large promotions. Unfortunately, many of the attracted buyers are one-time traders and these offers can have no long-lasting impact on sales. Traders must recognize who can be converted to multiple buyers to fix this problem. By targeting these potential loyal customers, traders will significantly reduce their marketing costs and increase their return on investment. The proposed methodology is a two-layer model fusion algorithm (TMFBG) based on GBDT for predicting repeat purchasers. The algorithm is verified by publishing data from the behaviour data of some of the annual "Double 11" customers on the website. Research studies have shown that this fusion algorithm can enhance predictive accuracy and model robustness. The report presents the idea of a two-layer fusion based on Ensemble Learning in Machine Learning and designs a two-layer fusion algorithm based on GBDT (TMFBG) and applies it to acquisition predictions in e-business. The results indicate that the TMFBG has more robustness and more accurate predictive efficiency, which is 2.1% higher than the single base classifier. In Comparison with the voting process, the average F1 score of the

TMFBG is 1.1% higher and the mean square error of the F1 score is 7.2% lower, indicating that the TMFBG is more accurate. The accuracy and F1 performance of the GBDT are the highest in the second layer of the fusion algorithm.

The study of (Wang, Zhang and Ren, 2019) states that precisely, more than 50 million original data are obtained and pre-processed for mining correlation. 60 percent was selected randomly as the training set and 20 percent as the evaluation set and 20 percent as the assessment package. Logistic regression and XGBoost algorithms have been used to set up two models based on the advantages of each. Proof shows that the logistic regression of the XGBoost system is feasible and that the model evaluation index is greater than any methodology used on its own. The user's consumer behaviour model is intended to predict the user's purchasing behaviour of a given product for a certain period in the future. The data file generated by the user modeling training includes the purchase data of the specified product in the verification data set table, which provides applied technology for the accurate marketing of large data analysis on the e-commerce platform.

Research (Tong et al.) explains that large amounts of data being made available due to the modern retail sector, increasing expectations, automation and technology, but business decision-making have become complex and difficult. The conventional database system has not been able to satisfy the user's request for intelligent analysis and forecasting of mass data. How to change the current situation of 'mass data, poor knowledge,' encourage better business decision-making and help businesses increase profitability and become topics of common interest in the field of industry and IT. Tong's article offers an overview of Business Intelligence, the key business intelligence tools and the growth and implementation of the Business Intelligence System in the retail sector. Company theme and dimensional design, ETL device design, Data Display middleware design and key developments are key elements of the program.

Research (Bradlow *et al.*, 2017) mentions according to some estimates, Walmart collects about 2.5 petabytes of information every hour about revenue, customer purchases, location and equipment. Where online and offline retail data provide a full view of customer buying behaviour and if data is connected at the individual customer level to allow 'true' customer value estimates Imagine a day when data is thought to exist only in online retail, e.g. consumer path data occurs inside the store due to RFID and other GPS tracking technologies.

(McCarthy et al.) Predictive analytics help understand customer behaviour as opposed to other methods. In this technique, many approaches help to define the actions of the customer. One of the approaches is a regression that could be used to measure the relationship between the variables involved in the purchase of the consumer. These methods include data analysis, data mining and deep learning algorithms to understand the various behaviours of clients. Predictive analytics technology is used to predict the potential behaviour of customers concerning a specific product, which allows retailers to narrow their search for profitability.

(Borisyak, Zykov and Noskov, 2015) research makes a point that there is a difference in the most common situation in e-commerce that requires reactive algorithms focused on a short-term study of user behaviour. This paper provides a specific mathematical framework for the short-term identification of user interest developed in terms of the properties of the object and its implementation to improve suggested systems. The paradigm is based on the basic principle of knowledge theory — Kullback — Leibler divergence.

As previous research, many of the researchers have worked efficiently on predicting customer behaviour using various machine learning, deep learning techniques but there is a gap between short term prediction and long term prediction for e-commerce users. Few models have given efficient results but the execution time is far more than expectations and few are less time-consuming but less efficient than the other one.
To balance this phenomenon In my research Proposed models are Logistic Regression, Light FM and MLP and by comparison with these modules, Better results can be achieved.
The main objective of this study is to predict the sell of items from the users transaction history.

4 Research Methodology

4.1 Dataset Resource and Explanation-

Dataset Link - https://www.kaggle.com/retailrocket/ecommerce-dataset
Kaggle is the source of the dataset and its public data so any permission and authorization are not needed to use it. The data is a collection of 4.5 months of real-world operations one commences a website and its raw data without any content transformation. Due to confidentiality issues, all values are hashed and the main purpose of usage of this dataset is the build an implicit recommender system with implicit feedback and predict customer purchases. The dataset is divided into 3 files.

1. Events.csv – In events file the behavioural data is stored, events like “click”, “view”, “add to cart” and “transaction”. Overall there are 2756101 events in which 2664312 are “view” events, 69332 are “add to cart” events and 22457 are “transaction” events. All these transactions are produced by 1407580 unique customers.
   Example - “1439694000000,1,view,100,” means visitorId = 1, clicked the item with id = 100 at 1439694000000 (Unix timestamp)

2. Category.csv – There are 1669 rows in the category tree file. Each row in the file specifies the child category Id and the corresponding parent.
   - Example - Line “100,200” means that categoryid=1 has a parent with categoryid=200.
   - The line “300,” means that category id hasn’t parent in the tree.

3. Item_properties.csv – The item properties file has 20275902 rows, explaining distinct properties of 417053 unique items. The file is split into two partitions due to file size limitations. Since the property of a product can differ in time (e.g. price changes over time), each row in the file has the corresponding timestamp. the file consists of concatenated snapshots for each week in the behaviour data file. However, if the property of the item is constant over the period observed, only a single snapshot value will be present in the file. The Item Properties File includes timestamp columns since all of them are time-dependent because properties will change over time, e.g. price, category, etc. Initially, this file consisted of snapshots every week in an event file containing over 200 million rows. We've combined consecutive constant property

values and we've changed the snapshot method to change the log shape. As a result, constant values would only appear once in the file. This action has significantly reduced the number of rows by 10 times. All values in the file "item properties.csv" except for "categoryid" and "available" properties have been hashed. The value of the "categoryid" property contains an identifier for the category object. The value of the "available" property includes the availability of the object, i.e. 1 indicates that the product was available, or 0. Both numerical values have been labelled with "n" at the beginning and have a precision of 3 digits after the decimal point, e.g. "5" will become "n5.000," "-3.67584" will become "n-3.675".

4.2 Cross-Industry Standard for Data Mining –

(Galleta and Carpio, 2019) CRISP-DM methodology is an open-source framework consisting of six phases of the mining cycle. Phases include business understanding, data interpretation, data preparation, modeling, evaluation, implementation. CRISP-DM is extensively used in data mining.

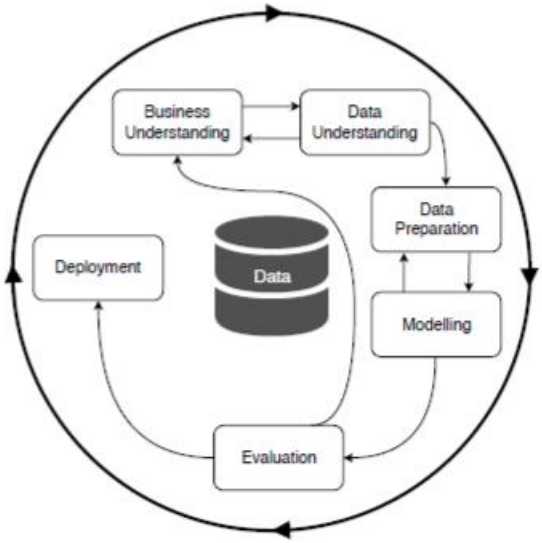


Figure 4- CRISP-DM Process(Martinez-Plumed *et al.*, 2019)

1. Business Understanding – In this phase significance, motivation and requirement of research are identified.
2. Data Understanding - In the data understanding phase, the recognition of different data or values is used to assess competitiveness.
3. Data preparation phase - The most important step in the processing of data is the preparation of data. In general, data cleaning and pre-processing take about 80 % of the time. The processing of data is time-consuming and challenging. The data preparation process deals with incomplete data where certain attribute values were missing, where some significant features were missing. In the data preparation process, data outliers and anomalies were also addressed and data irregularities were discussed in the preparation of the data. The data processing process generates a smaller dataset than the original data collection.
4. Modeling Phase – In this phase based on processed data relevant model is designed to build and assess.

5. Evaluation Phase – In this phase Results achieved from the model are evaluated, monitored and reviewed to determine the next step.
6. Deployment -In this phased strategy is designed and deployed to use developed model for business use.

4.3 Data Collection/Description/Pre-processing –

The raw data needs to be processed for any model and for an accurate prediction that's why data pre-processing is the most important aspect of research. Data pre-processing is a cycle in which data gets transformed, structured and encoded in a manner where algorithms can predict easily. The dataset used for this study has numerical and categorical variables. For pre-processing data pandas, the library is used.
Pandas- Pandas library is used for data wrangling, processing and analysis. Particularly pandas are used for manipulating numeric data.

Steps of Data Pre-processing-

- Data Quality Assessment – The dataset used for this research is a raw dataset there are numerous impurities like missing values, inconsistent values and null values. For making data consistent and handling null values from the dataset I have dropped the records which have null values by using dropna (). There were no duplicate and inconsistent values in the dataset.
- Feature Aggregation – The better perspective of data is achieved by feature aggregation, for better consistency of data several data frames as customer_purchased, all_customers, Customer borrowed so that model interpretation becomes comparatively easy. Better feature aggregation expects a reduction in processing time and memory consumption.
- Feature Encoding – In pre-processing, 2 types of encoders are being used.
    1. Label encoder class is used o convert categorical variables into understandable numeric data. And it helps to execute the model faster. To convert categorical data into numerical data I have used sklearn library. For timestamp and category id column I have used label encoder after preparing datasets x and y before training.
    2. One Hot Encoder- One hot encoder is also called a binary one. It selects a column that has categorical data that has been encoded by the label and then divides the column into several columns. Numbers are replaced by 1s and 0s, depending on which column has which value. One hot encoding is applied to the transaction column for better execution in this study.

- Train/Test split – Dataset is divided into train/test after feature encoding, which makes data ready for training with the model. Data is divided into (70:15:15) ratio. Train data is used for training models(ANN, Logistic, LightFM). Test data is 15% of data and it is used to measure the accuracy of the model in real-world data. Validation data is also 15% and it is used for improving hyperparameters.
- Feature Scaling – Feature scaling is a method normalize the fixed range of independent features of data. On dataset I have applied standard scaling, it re-scales the feature that has a distribution of 0 mean value and 1 variance. For feature scaling standardscaler library is imported from sklearn.preprocessing.

4.4 Model-

In this section, all applied models are described and explained

4.4.1 Model Description –

1. Logistic Regression –
   Logistic regression is popular for binary classification problems. The idea of logistic regression is extended from multiple linear regression. In logistic regression discontinue variable is the dependant variable. To predict the probability of a particular situation logistic regression is mainly used. (Bahrami, Bozkaya and Balcisoy, 2020) The researcher explains that Logistic is mostly used to process the variable with dependant two values.'1' and '0' are used to predict the results. The logistic function is also called a sigmoid function. Inputs(x) are linearly combined by weights or coefficient values to the prediction of result(y). The equation of logistic regression is as follows.
   y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))
   In the equation y is the output, b0 is bias and b1 is the coefficient of a single value(x).

2. LightFM-
   LightFM is a hybrid matrix factorization model that describes users and objects as linear combinations of the latent factors of their content features. (Kula, 2015) The model overtops both collaborative and content-based models in cold-start or sparse interaction data scenarios (using both user and item metadata) and works at least as well as a strictly collaborative matrix factorization model where interaction data is ample. In LightFM, as in a collaborative filtering model, users and objects are defined as latent vectors (embeddings). Even then, as in the CB model, these are primarily determined by the linear combinations of the content embedding features identified by and product or consumer. The equation of the LightFM model is as follows.
   $$r_{ui}\hat{} = f(q_u \cdot p_i + b_u + b_i)$$
   The reason behind using LightFM is the LightFM performs as well as pure content-based models in both cold-start and low-density scenarios, achieved a major when either collaborative information is available in the training set or user features are included in the model. It is very helpful for our customer behaviour forecasting program, as we're going to have a lot of new questions and students that make a really good environment for a cold start challenge.
3. ANN-
   Artificial neural networks are non-parametric techniques used for pattern recognition and optimization. In this research, a multilayer perceptron is used which is consists of 3 layers. I/P layer, hidden layer, output layer. (Badea (Stroie), 2014b) explains that in MLP the information flow is distributed in a feed-forward way and all the elements in the layer are completely connected to the nodes of the next layer. The training cycle within the MLPs is performed by re-propagating the errors and changing the network weights accordingly, to reduce the output deviations from the target values. To the weighted number of inputs, the activation function is applied to the node to obtain a certain result.

4.4.2 Model Building –

In this section model construction is explained.

1. Logistic Regression Model –
   Before building a model to get insights from what we can gather from viewed items, added to cart and sold Here perhaps cluster the visitors for that new data frame is created and engineer a few features for it. First added all visitor id in a single array

and sorted them ascendingly. So out of 1407580 visitors, 11719 bought something and 1395861 visitors just viewed items.
Now create a new data frame with new features: visitorid, a number of items viewed, total view count and bought something or not. Plotted the data using seaborn.

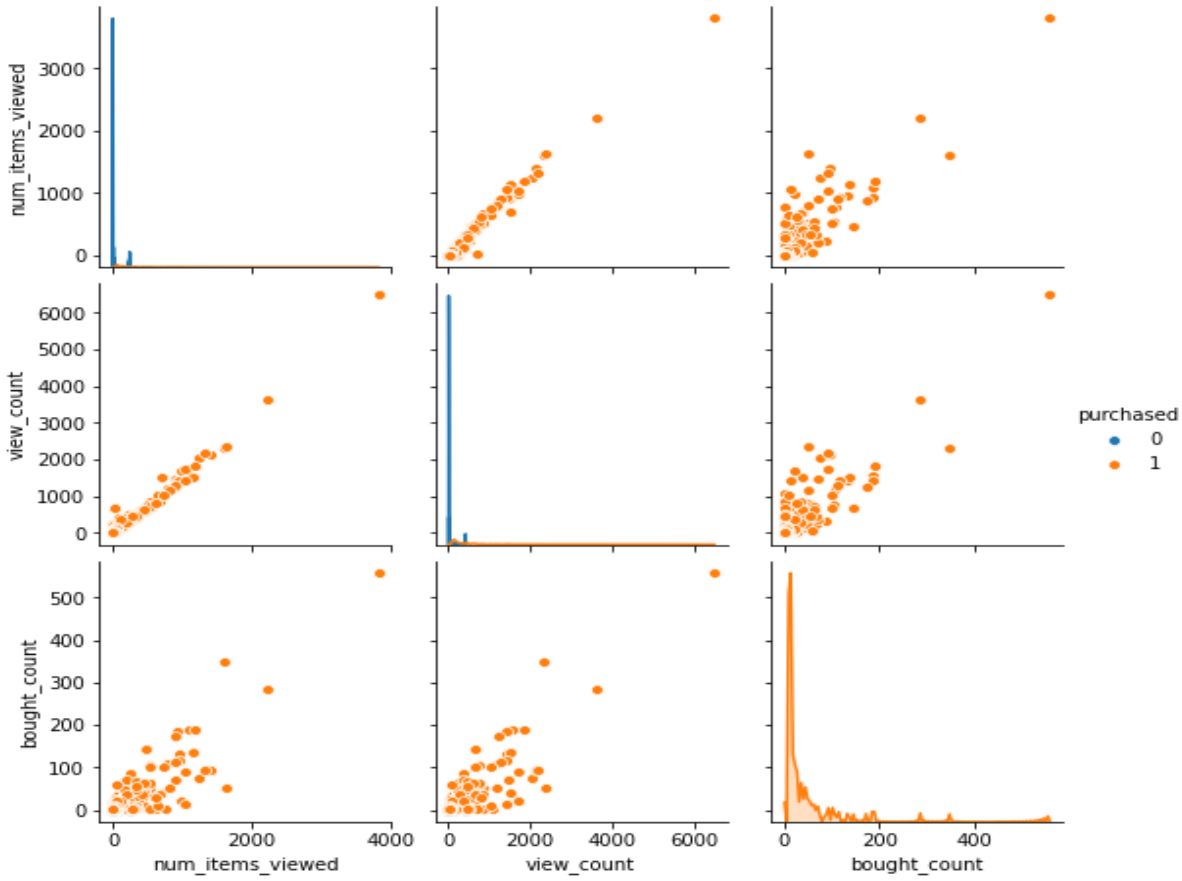


Figure 5

The plot above clearly indicates that the higher the view count and the higher the chances of a visitor buying something.
The next step is Label encoding for encoding categorical data and encoding independent variables. For label encoding, LabelEncoder library is imported from sklearn and for categorical encoding, the OneHotEncoder library is used.
After encoding data is divided into train and test data using x_train,x_test, y_train,y_test and division ratio has kept 70:30.
The logistic model applied to train data for that from sklearn.linear.model LogisticRegression was imported. Once a model is a trained model. fit() method is called. Prediction and accuracy are discussed in the implementation section.

2. LightFM-
   Importing the LightFM library is the first step towards building a model.
   Parameters used of LightFM model and are description is as follows-
   - no_componant(5)- Is the dimensionality of feature latent embeddings.
   - Loss='warp'- This is loss function which is weighted Approximate Rank pairwise. The reason behind using this loss function is, it significantly

increases the rank of positive examples by consistently sampling negative examples before the rank of violating one is found. Useful when only positive interactions are present and the top of the recommendation list is to be optimized. (Kula, 2015).

- Epoch = Number of epochs are 100, To train data appropriately 100 epochs are applied because of a large dataset.

Model is applied by model.fit() and AUC score is printed to predict accuracy as data is binary, multiclass and multilabel classification data.
The results are discussed in the Implementation section.

3. ANN (Multilayer Perceptron Model)-
   For MLP Created a classifier and in that Keras is imported and from Keras layers Input layer and Dense layer, dropout layer libraries also imported.
   - The input layer is the input layer of the network, this layer works as input it accepts input values and passes to the next layer.
   - The dense layer is fully connected, each neuron receives input from all the neurons in the previous layer. We have sequential classifier Layers are added by classifier.add().
   - The first layer is added with activation function "relu" and 1024 units will be passed to the first layer. ReLU is a rectified linear unit reason behind using ReLU is using ReLU function over the activation function it does not activate all neurons at a time, which helps in processing speed. Neurons will be deactivated if the output of linear transformation is less than 0. Used kernel initializer is uniform, it distributes weights uniformly. Units are 1024 inputs to the first layer because the dataset is large.
   - After the first layer dropout layer is added of 0.20 to prevent overfitting.
   - For hyperparameter, training is used it implements fit and scoring methods. Parameters are as follows –
     Estimator – Classifier is used as an estimator; the classifier provides score function.
     Param_grid – parameters.
     Scoring- accuracy because we need to see accuracy from it.
     CV- none it means five cross-validations would be used.
   - The third layer is like the first layer but just units are decreased to 512.
   - In the last output layer, Adam optimizer is used and for loss measure, binary crossentropy is used reason behind using adam is there is no need to mention learning rate, the learning process is configured in this last layer. And for accuracy, I have passed metrics.
   - In classifier.fit() x_train, y_train are passed and epoch, the batch size is given and the model started running, results will be discussed in the results section.

5 Implementation

In this section implementation of the proposed solution is discussed.
For the implementation of models python 3.6 is used and for data processing pandas and NumPy are used and for plotting and visualization seaborn and matplotlib.

5.1 Logistic Regression-

To check the output of this model accuracy and AUC ROC curve are obtained. According to the logistic regression model accuracy achieved is 79.4%, It means that 79.4% of buying visitors are around in the data. Below snippet of code is added of achieved accuracy.

```
#use the model to predict the test features
y_pred_class = logreg.predict(X_test)
```

```
[41] print('accuracy = {:7.3f}'.format(metrics.accuracy_score(y_test, y_pred_class)))

     accuracy =   0.794
```

Figure 6- Accuracy of Logistic Regression

ROC curve -

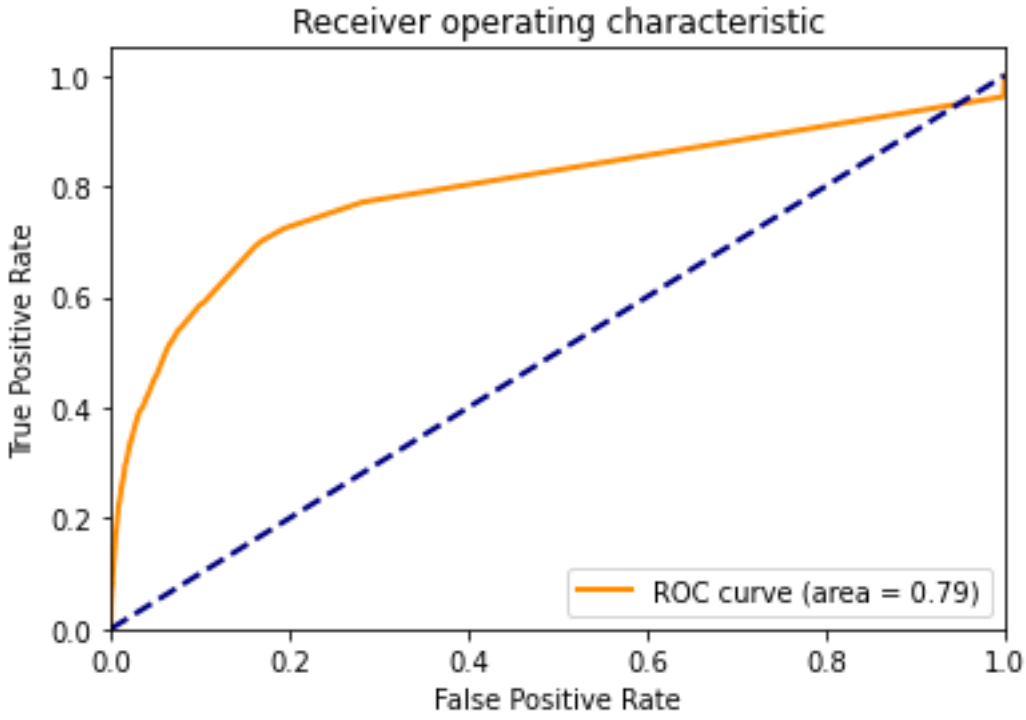


Figure 7- ROC Curve.

As the above graph shows the accuracy of binary classification. The more closer the orange line to the top left corner of the graph mentions more accuracy.

5.2 LightFM

To check the accuracy of the LightFM model Train AUC and Test AUC is calculated. Achieved training accuracy is 98.40% and Testing accuracy is 81.60%,

```
    '''
    test_set = ratings.copy() # Make a copy of the original set to be the test set.
    test_set[test_set != 0] = 1 # Store the test set as a binary preference matrix
    training_set = ratings.copy() # Make a copy of the original data we can alter as our training set.
    nonzero_inds = training_set.nonzero() # Find the indices in the ratings data where an interaction exists
    nonzero_pairs = list(zip(nonzero_inds[0], nonzero_inds[1])) # Zip these pairs together of user,item index into list
    random.seed(0) # Set the random seed to zero for reproducibility
    num_samples = int(np.ceil(pct_test*len(nonzero_pairs))) # Round the number of samples needed to the nearest integer
    samples = random.sample(nonzero_pairs, num_samples) # Sample a random number of user-item pairs without replacement
    user_inds = [index[0] for index in samples] # Get the user row indices
    item_inds = [index[1] for index in samples] # Get the item column indices
    training_set[user_inds, item_inds] = 0 # Assign all of the randomly chosen user-item pairs to zero
    training_set.eliminate_zeros() # Get rid of zeros in sparse array storage after update to save space
    return training_set, test_set, list(set(user_inds)) # Output the unique list of user rows that were altered
```

+ Code + Markdown

```
X_train, X_test, item_users_altered = make_train(user_to_item_matrix, pct_test = 0.1)
```

```
no_comp, lr, ep = 100, 0.01, 10
model = LightFM(no_components=no_comp, learning_rate=lr, loss='warp')
model.fit_partial(
        X_train,
        item_features=item_to_property_matrix_sparse,
        epochs=ep,
        num_threads=4,
        verbose=True)
model.summary(accuracy)
```

Figure 8- LightFM Model

5.3 Multilayer Perceptron Model

To achieve better prediction and accuracy Hyperparameter optimization technique is used.
In below diagram structure of the model is shown mainly there are 3 layers

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 1024)              12288
_________________________________________________________________
dropout (Dropout)            (None, 1024)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               524800
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 1026
=================================================================
Total params: 538,114
Trainable params: 538,114
Non-trainable params: 0
```

You will of course want to discuss the implementation of the proposed solution. Only the final stage of the implementation should be described.

It should describe the outputs produced, e.g. transformed data, code written, models developed, questionnaires administered. The description should also include what tools and languages you used to produce the outputs. This section must not contain code listing or user manual description.

6 Evaluation

This section obtains evaluation of all models, The main aim behind comparison analysis was to study and recommend which model is better for short term and long term customer behaviour prediction, as per the findings Logistic Regression model achieved an accuracy of 79.40% and to view that ROC curve is plotted. The second model is the LightFM. This model is especially recommended for recommender systems. This system model is used to recommend and predict products to customers by their behavioural data and LightFM has obtained an accuracy 81.6% testing accuracy and this is achieved by label encoding and hyperparameters optimization techniques and customization of data frames. The third model is the multilayer perceptron model. In that hyperparameters optimization, GridsearchCv is used which resulted in better accuracy and prediction and validation accuracy achieved is 92.4%.

Comparison of Models

Model name	Accuracy	Execution Time
Logistic Regression	79.40%	5.14
LightFM	81.6%	18.32
Multilayer perceptron	92.4%	36.25

Table 1

.

6.1 Discussion

In the discussion, it can be concluded that MLP demonstrated better accuracy than logistics and LightFM, but It was time-consuming and Logistic gave better accuracy and executed in less time also and LightFM has given good accuracy with just label encoding. Overall MLP performed better and This comparison could have been better if there were more models.

7 Conclusion and Future Work

The research question was How well Predictive analysis helps E-commerce retailers to understand customer behaviour? And How to recommend products to the customer by-products visit history? The accuracy acquire by all models is beyond expectation few things like label encoding and hyperparameter optimization worked and proved if these techniques associated with any model better results can be expected. This research has fulfilled its research questions successfully predicted customer behaviour and recommended items to customers by behavioural data.
MLP takes time to execute If future work use of distributed processing techniques to fasten the execution is possible. For better comparison analysis multiple models can be concluded in future work.

8 Acknowledgment –

I would like to thank every who guided and supported me throughout this process, Special thank you for my supervisor Mr. Manaz Kaleel for all support and inspiration.

References

Agrawal, S. *et al.* (2018) 'Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning', *2018 International Conference on Smart Computing and Electronic Enterprise, ICSCEE 2018*. IEEE, pp. 1–6. doi: 10.1109/ICSCEE.2018.8538420.

Badea (Stroie), L. M. (2014a) 'Predicting Consumer Behaviour with Artificial Neural Networks', *Procedia Economics and Finance*. Elsevier BV, 15, pp. 238–246. doi: 10.1016/s2212-5671(14)00492-4.

Badea (Stroie), L. M. (2014b) 'Predicting Consumer Behaviour with Artificial Neural Networks', *Procedia Economics and Finance*. doi: 10.1016/s2212-5671(14)00492-4.

Bahrami, M., Bozkaya, B. and Balcisoy, S. (2020) 'Using Behavioural Analytics to Predict Customer Invoice Payment', *Big Data*, 8(1), pp. 25–37. doi: 10.1089/big.2018.0116.

Be, K. H. *et al.* (2019) *Online Shopping Customer Behaviour Analysis using centrality measures*.

de Bellis, E. and Venkataramani Johar, G. (2020) 'Autonomous Shopping Systems: Identifying and Overcoming Barriers to Consumer Adoption', *Journal of Retailing*. Elsevier Ltd, 96(1), pp. 74–87. doi: 10.1016/j.jretai.2019.12.004.

Borisyak, M., Zykov, R. and Noskov, A. (2015) 'Application of Kullback-Leibler divergence for short-term user interest detection', (July). Available at: http://arxiv.org/abs/1507.07382.

Bradlow, E. T. *et al.* (2017) 'The Role of Big Data and Predictive Analytics in Retailing', *Journal of Retailing*. Elsevier Ltd, 93(1), pp. 79–95. doi: 10.1016/j.jretai.2016.12.004.

Doan, T., Veira, N. and Keng, B. (2019) 'Generating realistic sequences of customer-level transactions for retail datasets', in *IEEE International Conference on Data Mining Workshops, ICDMW*. IEEE Computer Society, pp. 820–827. doi: 10.1109/ICDMW.2018.00122.

Galleta, D. T. and Carpio, J. T. (2019) 'Predicting regional development competitiveness index using naive bayes: Basis for recommender system', *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*. IEEE, (Cmci), pp. 115–119. doi: 10.1109/CCOMS.2019.8821666.

IbukunT, A. *et al.* (2016) *A Systematic Review of Consumer Behaviour Prediction Studies*, *Covenant Journal of Business & Social Sciences (CJBSS)*.

Krishnan, N. *et al.* (no date) *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) : 2017 December 14-16 : venue: Tamilnadu College of Engineering, Coimbatore-641 659, Tamilnadu, India*.

Kula, M. (2015) 'Metadata embeddings for user and item cold-start recommendations', *CEUR Workshop Proceedings*, 1448, pp. 14–21.

Martinez-Plumed, F. *et al.* (2019) 'CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories', *IEEE Transactions on Knowledge and Data Engineering*, 4347(c), pp. 1–1. doi: 10.1109/tkde.2019.2962680.

Orogun, A. and Onyekwelu, B. (2019) 'Predicting Consumer Behaviour in Digital Market : A Machine Learning Approach', (Hauser 2007), pp. 8391–8402. doi: 10.15680.

Predicting Consumer Behaviour in online purchase (no date). Available at: www.worldwidejournals.com.

Wang, X., Zhang, M. and Ren, F. (2019) 'Learning customer behaviours for effective load forecasting', *IEEE Transactions on Knowledge and Data Engineering*. IEEE Computer Society, 31(5), pp. 938–951. doi: 10.1109/TKDE.2018.2850798.

Xu, D., Yang, W. and Ma, L. (2018) 'Repurchase Prediction Based on Ensemble Learning', in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, pp. 1317–1322. doi: 10.1109/SmartWorld.2018.00229.