CUSTOMER BEHAVIOUR PREDICTION USING WEB USAGE MINING

Submitted in partial fulfillment of therequirements
of the degree of

B. E. Computer Engineering

By

Riya Dodthi **07**
Avril Lopes **23**
Caroline Lopes **24**

Supervisor:

Ms. Snehal Kulkarni
Assistant Professor

Department of Computer Engineering
St. Francis Institute of Technology
(Engineering College)
University of Mumbai
2018-2019

CERTIFICATE

This is to certify that the project entitled**"Customer Behaviour Prediction Using Web Usage Mining"**is a bonafide work of**"Riya Dodthi (Roll No. 07),Avril Lopes (Roll No. 23), and Caroline Lopes (Roll No. 24)"**submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

Ms. Snehal Kulkarni
Supervisor/Guide

Dr. KavitaSonawane
Head of Department

Dr. Sincy George
Principal

Project Report Approval for B.E.

This project report entitled ***Customer Behaviour Prediction Using Web Usage Mining*** by ***Riya Dodthi, Avril Lopes, and Caroline Lopes*** is approved for the degree of ***B.E. in Computer Engineering.***

Examiners

1.----------30/04/2019----------

2.----------30.04.19----------

Date: 30/04/2019

Place: Borivli

Declaration

Wedeclare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Riya Dodthi - 07

Avril Lopes - 23

Caroline Lopes - 24

Date: 30/04/2019

Abstract

Web usage mining involves first recording behavior and flow of customers on a website and then mining through this data for behavioral patterns. Ecommerce sites analyze this data in order to provide better performance and also suggest better products and services to customers. The system is tuned to record web shopping/buying patterns and track various analytics data. The system scans for user budget tracking, tallying to previous years, user bounce rates- number of users returning from payment page and other site usage factors. Ecommerce sites need to survey and mine for previously recorded data to check their website performance and constantly optimize it as per customer needs.

Contents

List of Figures

List of Tables

List of Abbreviations

Sr. No.	Abbreviation	Expanded form
i	BPNN	Back Propagation Neural Network
ii	KNIME	Konstanz Information Miner
iii	WEKA	Waikato Environment for Knowledge Analysis
iv	SPRINT	Scalable Parallelizable Induction of Decision Trees
v	DFD	Data Flow Diagram
vi	ANOVA	Analysis of Variance
vii	MSE	Mean Squared Error
viii	RMSE	Root Mean Square Error

Chapter 1

Introduction

With the expeditious growth of e-commerce or the web-based marketing system, the Internet has become the most important media for understanding consumers. Marketing managers seek to gain significant insights on consumers' web navigation behaviour allowing them to identify the most important visitors and hence derive customized marketing strategies. Business consultants, on their part, benefit from the use of Web information to evaluate key performance indicators; by doing so, they could offer more tailored recommendations and solutions to marketing managers. Hence such a methodology that allows the analysis of web data and deduction of insights would be invaluable.

1.1 Description

In a regular retail shop the behavior of customers may yield a lot to the shop assistant. However, when it comes to online shopping it is not possible to see and analyze customer behavior such as facial mimics, products they check or touch etc. In this case, clickstreams or the mouse movements of e-customers may provide some hints about their buying behavior. In this study, we have presented a model to analyze clickstreams of e-customers and extract information and make predictions about their shopping behavior on a digital marketplace. The model we present predicts whether customers will or will not buy their items added to shopping baskets on a digital marketplace.

1.2 Problem Formulation

Clickstream is a record of a user's activity on the internet; these mouse clicks a user makes when he or she is surfing may tell us a lot about the behavior of the user if it is analyzed in an appropriate way. These movements are sort of the behavior of online customers. We can name this analysis as web mining or web farming approach to discover patterns in the navigation of websites and web contents. By analyzing users' navigation patterns and their relation with web content one can redesign a website, portal or e-business along with the behavior of the online users.

1.3 Motivation

A website should be designed to entice the customers. Web Mining analyses visitor's behaviour and makes predictions on their future interaction. This can be exploited to improve website performance and to recommend products or links based on user's behaviour. Visitors entering the site exhibit different behaviour. They might just surf through or the process might end up in a purchase. For understanding customer behaviour and thus improve the performance of the web site, certain standards should be used like perform mining on web log data.

1.4 Proposed Solution

We will be using artificial neural networks model. Neural networks learn through input, a set of hidden layers and an output layer. Setting up of a dummy website is required through which the data sets will be collected. These data sets will then be used for prediction with the help of BPNN algorithm.

1.5 Scope of the project

Predicting the ever-evolving consumer behavior is one of the biggest challenges faced by marketers around the world. Well, it has always been a challenging task, but today, it is even harder as consumers are constantly being exposed to new technologies, products and even new wants! With a plethora buying options to their disposal, today's consumers' buying behavior flickers way too often. The number of online shoppers in just a single country is projected to reach 224 million in 2019. Top reasons to analyze customer behavior are gain insight – Segmenting customer database with cluster analysis to identify consumer segments, attract and engage – Targeting the segment of customers with right offers by analyzing historical purchases and profiles and improve retention – It enables companies to calculate customer value and put proactive retention approach to retain customers.

Chapter 2

Review of Literature

Literature reviews are a basis for research in nearly every academic field. A narrow-scope literature review may be included as part of a peer-reviewed journal article presenting new research, serving to situate the current study within the body of the relevant literature and to provide context for the reader.

For the analysis they used KNIME program. KNIME is an open source, Eclipse based program for data mining. It resides decision tree and artificial neural network algorithms and other open source data mining platforms such as WEKA and R. So, KNIME is quite useful to run different programs and algorithms on its own platform. For learning, decision tree and artificial neural network algorithms have been used. Decision tree algorithms generate rules from data sets. When they run they create a tree structure and if-else-then rules. In this study we used two decision tree algorithms together. This technique is called bagging or bootstrapping. The algorithms they have used are C4.5 and SPRINT (Scalable Parallelizable Induction of Decision Trees). C4.5 algorithm uses entropy function to determine the best attribute or data field to arch from. SPRINT algorithm uses Gini function to choose the best data field to create branches. Neural networks learn through input, a set of hidden layers and an output layer. [1]

Markov model is assumed to be a probability model by which users' browsing behaviors can be predicted at category level. Bayesian statistics can also be applied to present and infer users' browsing behaviors at webpage level. Bayesian data analysis is a powerful technique for fitting almost any model to data, and R is the tool that makes this easy. K-means clustering algorithm is one of the simplest unsupervised learning algorithms that solves clustering problem. The proposed EPFK-Means algorithm, on average, showed an efficiency gain of 12.39% over conventional ensemble system, indicating that the enhancement operations incorporated are

successful in improvingclustering efficiency. This paper presented an algorithm that improves the prediction accuracy. [2]

In this paper, they firstly analyze the correlation between site search query data and daily e-commerce orders theoretically. And then, by empirical analysis of orders and search data of an E-commerce site, the paper processes a search data index, verifies the co-integration relationship and builds a model combining search index with historical data. The results show that the consumers usually search the site for information one to two days before they make an order to buy the product, and there is statistically significant correlation between daily e-commerce orders and site search data, and the search data has good prediction ability for the daily e-commerce orders. They calculate the Pearson correlation coefficient of every search query volume data and the daily e-commerce orders. They use extended Dickey-Fuller test method to do stationary test for non-dummy variables. Granger causality can test whether a variable has predictive ability for another. This new model can predict daily e-commerce orders, while most traditional models can only forecast total or average e-commerce orders. [3]

The main contribution of this article is the application of sequence analysis (with complimenting concurrent use of cluster analysis) to understand user web navigation behavior. Methodology used is cluster analysis and sequence analysis. Sequence analysis methodology discriminates between different users' groups based on their web navigation behaviors across websites. The main aim of applying sequence analysis is to compare sequences in order to discover similarities and dissimilarities in compared sequences. The algorithm used in this model is Optimal Matching distance (OM) algorithm for comparing sequence patterns. The data used in this study is obtained through collaboration with an independent Internet-consulting firm in China, which has installed a client-based tracking application in each individual consumer's computer with which his/her Internet usage behavior is observed. In this study, they have shifted the views from each individual site's web activities to cross-websites' web activities. In other words, they selected users to be observed for a certain period of time, recorded their web activities, and analyzed and clustered their behaviors based on their browsing. [4]

So it's been observed that ANN is the best algorithm for prediction. There are different types of NN algorithms, out of which BPNN is the most suitable for prediction. Back Propagation Neural Network supports high speed classification and multi- classification. BPNN can be used for linear as well as nonlinear classification.

Chapter 3

System Analysis

The System Analysis phase signifies the commencement of system implementation. The objectives of this phase are to investigate and understand the user and technical requirements, to specify the new system, to detail the implementation requirements in terms of cost, effort and time. The SA Report will be produced to document the findings and recommendations of this phase.

3.1 Functional Requirements

- User should open the E-commerce website.
- Enter age and gender.
- Browse through different categories of the website.
- Click on different items.
- Put items in shopping basket.

3.2 Non- Functional Requirements

- Reliability: The data should be constantly recorded by the software for generation of rules and predictions.
- Scalability: The system should accommodate new updating if necessary.
- Maintainability: Software should be coded in a way that is easy to understand and maintain.
- Capacity: System should have enough capacity to store a huge database of different users.
- Performance: Short response time, high throughput, low utilization of computing resource.
- Durability: The system should remain functional, without requiring excessive maintenance or repair, when faced with the challenges of normal operation over its design lifetime.

3.3 Specific Requirements (Hardware and Software Requirements)

3.3.1 Technical Feasibility

- Hardware Requirements
  - Processor: 3.5 gigahertz (GHz) Quad Core
  - RAM: 4 GB or more
  - Hard Disk: Secondary storage with 100 GB available space.

- Software Requirements
  - Windows 7 or later
  - E-commerce website

3.3.2 Economic Feasibility

The financial feasibility focuses on the system's potential return on investment. It determines the benefits and the savings that are expected from the project and compare them with the costs. Being an entirely software project, the investment required to be done was computers, electricity, application software, etc. but our college itself already fulfilled all these requirements. Therefore we did not have to spend as such from our pockets as regard to the financial part of our software. This feasibility analysis led to various cost benefit analysis. The purpose of above study is to estimate the resources required for the establishment of the enterprise or service business and the cost that should be incurred to achieve this.

- Hardware Requirements

Table 3.1 – Hardware Requirements

Requirements	Details	Cost
PC	Desktop/ Laptop	30000/-
Internet Connection	Modem/ Ethernet	5000/-
	Total Cost	35000/-

➢ Software Requirements

Table 3.2 – Software Requirements

Requirements	Details	Cost
Windows	OS	6000/-

3.4 Use-Case Diagram

The figure illustrates the use case diagram for the Customer Behaviour Prediction Using Web Usage Mining system:

- Customer surfs through the website.
- The clickstream of the customer will be recorded by the system.
- The system will then classify this data according to the datasets provided.
- The system will generate rules based on the available data.
- The system will make prediction on customer's behaviour based on these rules.

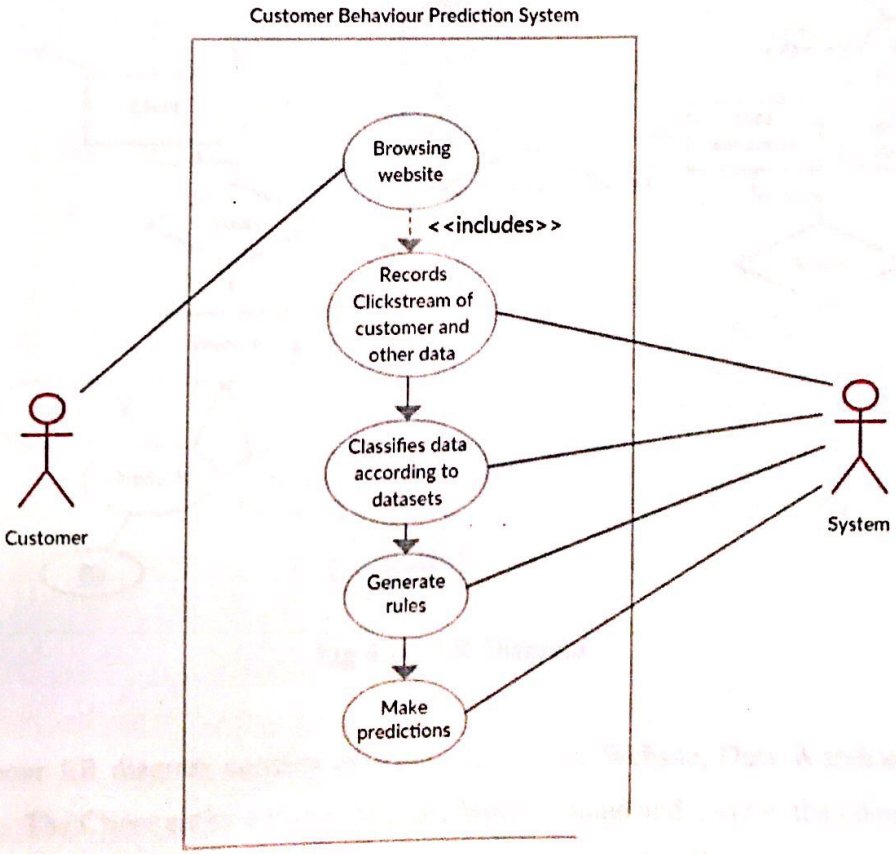

Fig 3.1 – Use Case Diagram

Chapter 4

Analysis Modeling

Structured Analysis considers data and processes transforming the data as separate entities. Data objects are modeled defining their attributes and relationships. Processes depict transformation of data objects as they flow through the system. Object Oriented focuses on the definition of classes and the way they collaborate with one another to satisfy customer requirements. (UML and Unified Process are predominantly object-oriented).

4.1 Data Modeling

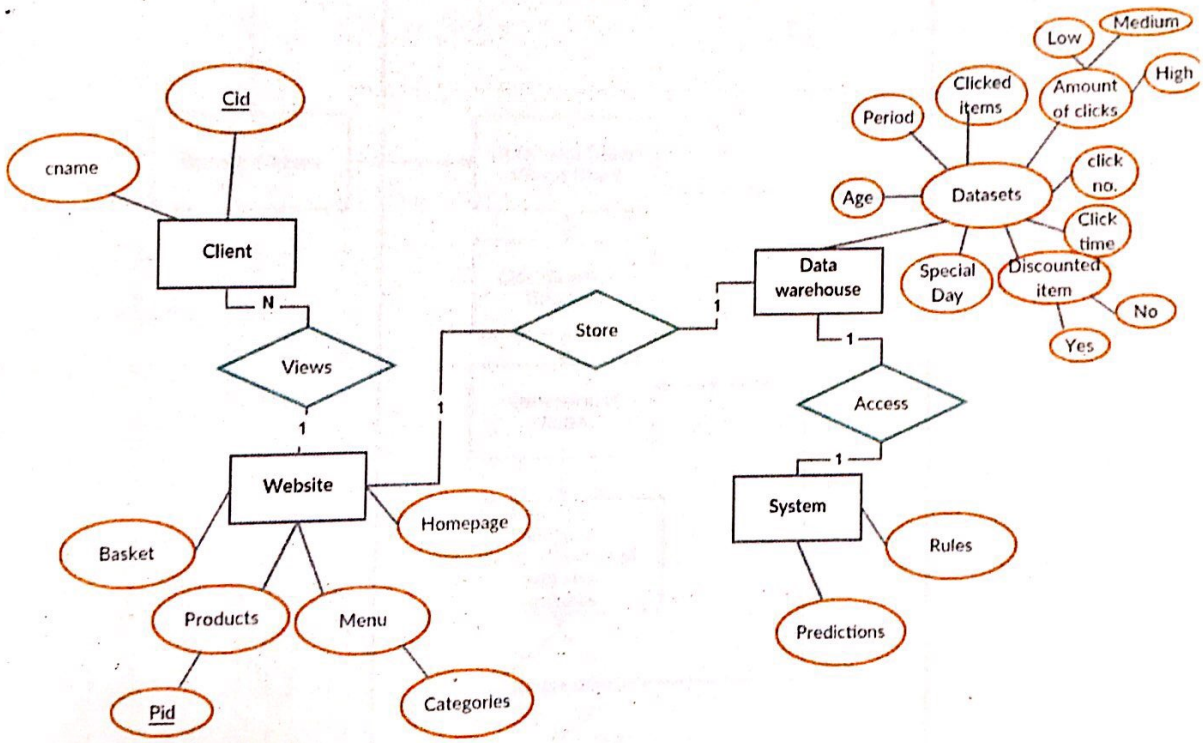

Fig 4.1 – ER Diagram

The above ER diagram consists of 4 entities- Client, Website, Data Warehouse and System. The Client entity consists of 2 attributes- Cname and Cid i.e. the client name and a unique client id for each client respectively. The Website consists of 4 attributes-Homepage, Menu which will further consist of different Categories and

Products where each product will have a unique product id (Pid) and Basket. The System would consist of Rules and Prediction. The Data Warehouse consists of Datasets which would further consist of attributes like Clicked items, Click number, click time, discounted items, Special Day, Period of day, Age, etc. The relationship between the website and the client would be- many clients can view a single website. The relationship between the website and the data warehouse would be- a website stores all its data in a single data warehouse. The relationship between the data warehouse and the system would be- system has access to the data warehouse for data retrieval.

4.2 Activity Diagram

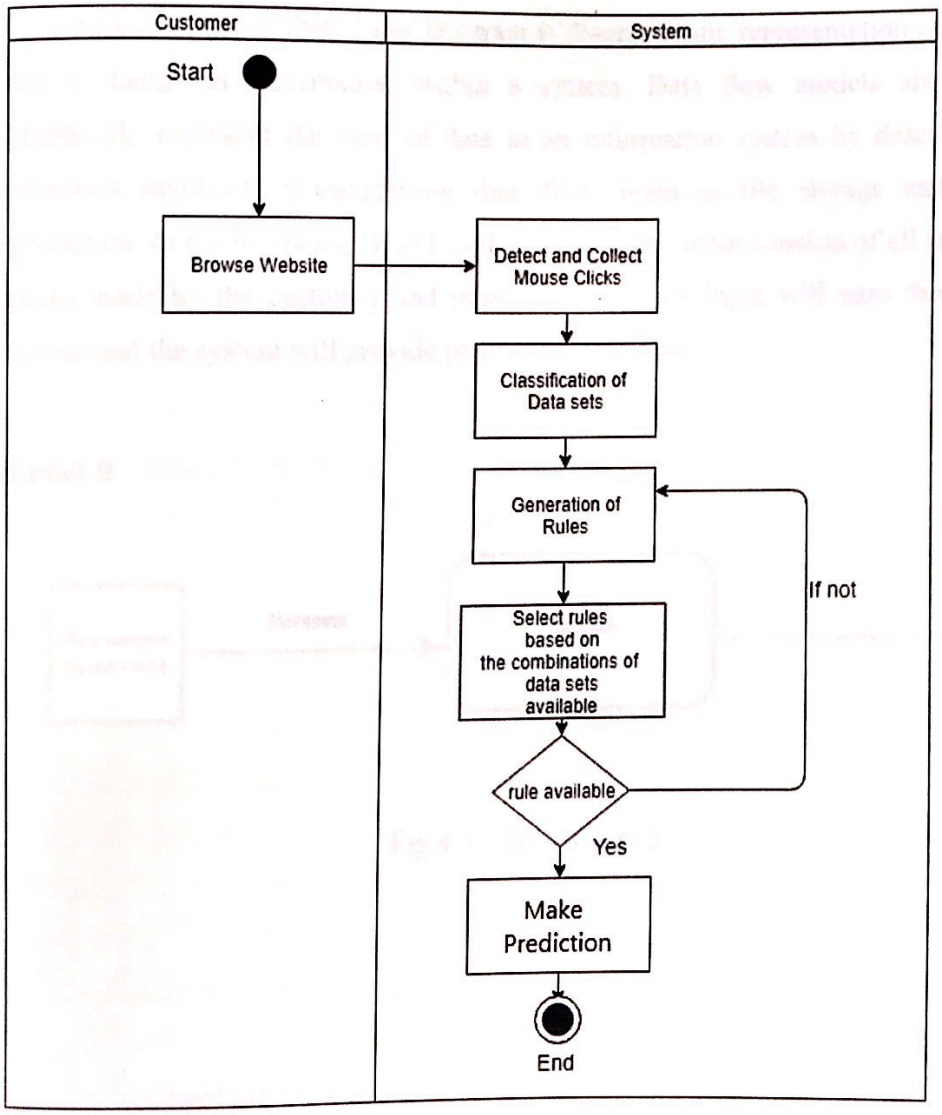

Fig 4.2 – Activity Diagram

The above Activity Diagram shows all the activity of the customer and the system throughout the prediction process. The flow of activities is as follows:

1. The customer first browses through the website.
2. The system then detects and collects all the mouse clicks of the customer.
3. This mouse clicks are then classified into different data sets (click number, clicked item, etc.).
4. Further rules are generated based on these datasets.
5. Now, select rules based on the combination of dataset available for prediction.
6. If rules are available, make predictions else generate a new rule.

4.3 Functional Modeling (DFDs)

A data flow model or Data Flow Diagram is diagrammatic representation of the flow and exchange of information within a system. Data flow models are used to graphically represent the flow of data in an information system by describing the processes involved in transferring data from input to file storage and reports generation. In the following DFD Level 0 diagram, the input consists of all the mouse clicks made by the customer and other datasets. This input will pass through the system and the system will provide predictions as output.

Level 0

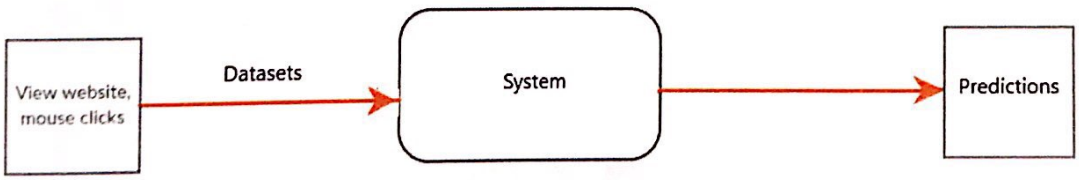

Fig 4.3 – DFD Level 0

Level 1

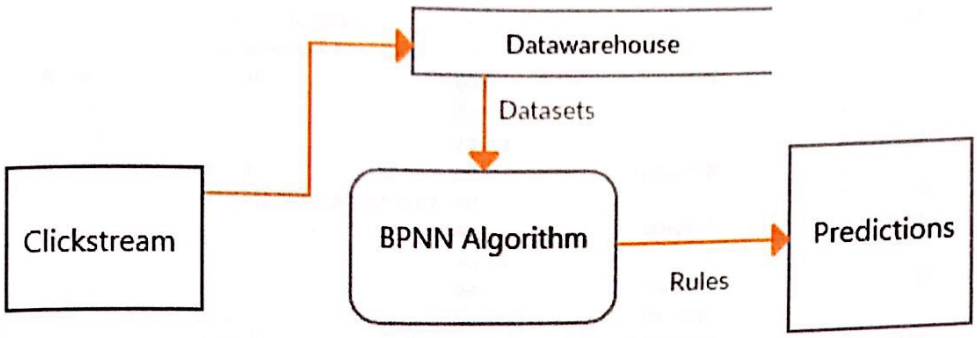

Fig 4.4 – DFD Level 1

4.4 Time Line Chart

A timeline chart is an effective way to visualize a process using chronological order. Since details are displayed graphically, important points in time can be easy seen and understood. Often used for managing project schedule, timeline charts function as a sort of calendar of events within a specific period of time. It includes all the major software development phases like Problem Definition, Requirement Gathering, Analysis Phase, Design, Implementation and Testing.

	Name	Duration	Start	Finish
1	**Problem Definition**	**12d**	**7/18/2018**	**7/31/2018**
2	Research Project	6d	7/18/2018	7/23/2018
3	Identifying Goals	5d	7/24/2018	7/30/2018
4	Project Presentation and Approval	1d	7/31/2018	7/31/2018
5	Milestone Problem Statement Description	1d	7/31/2018	7/31/2018
6	**Requirement Gathering**	**20d**	**8/6/2018**	**8/31/2018**
7	Literature Survey	9d	8/6/2018	8/16/2018
8	Feasibilty Study	2d	8/17/2018	8/20/2018
9	Information Gathering	6d	8/20/2018	8/27/2018
10	Study of Preprocessing and Algorithm used	4d	8/28/2018	8/31/2018
11	**Analysis Phase**	**8d**	**9/3/2018**	**9/12/2018**
12	UML Diagram	4d	9/3/2018	9/6/2018
13	Timeline Chart	2d	9/4/2018	9/5/2018
14	Study of Algorithms	3d	9/6/2018	9/10/2018
15	Learning Softwares	3d	9/10/2018	9/12/2018
16	**Design**	**18d**	**9/13/2018**	**10/8/2018**
17	Procedural Design	3d	9/13/2018	9/17/2018
18	New Synopsis	15d	9/18/2018	10/8/2018
19	**Implementation**	**29d**	**2/4/2019**	**3/14/2019**
20	Build System	15d	2/4/2019	2/22/2019
21	Test System	14d	2/25/2019	3/14/2019

Fig 4.5 – Timeline Chart

Fig 4.6 – Timeline Bar Chart

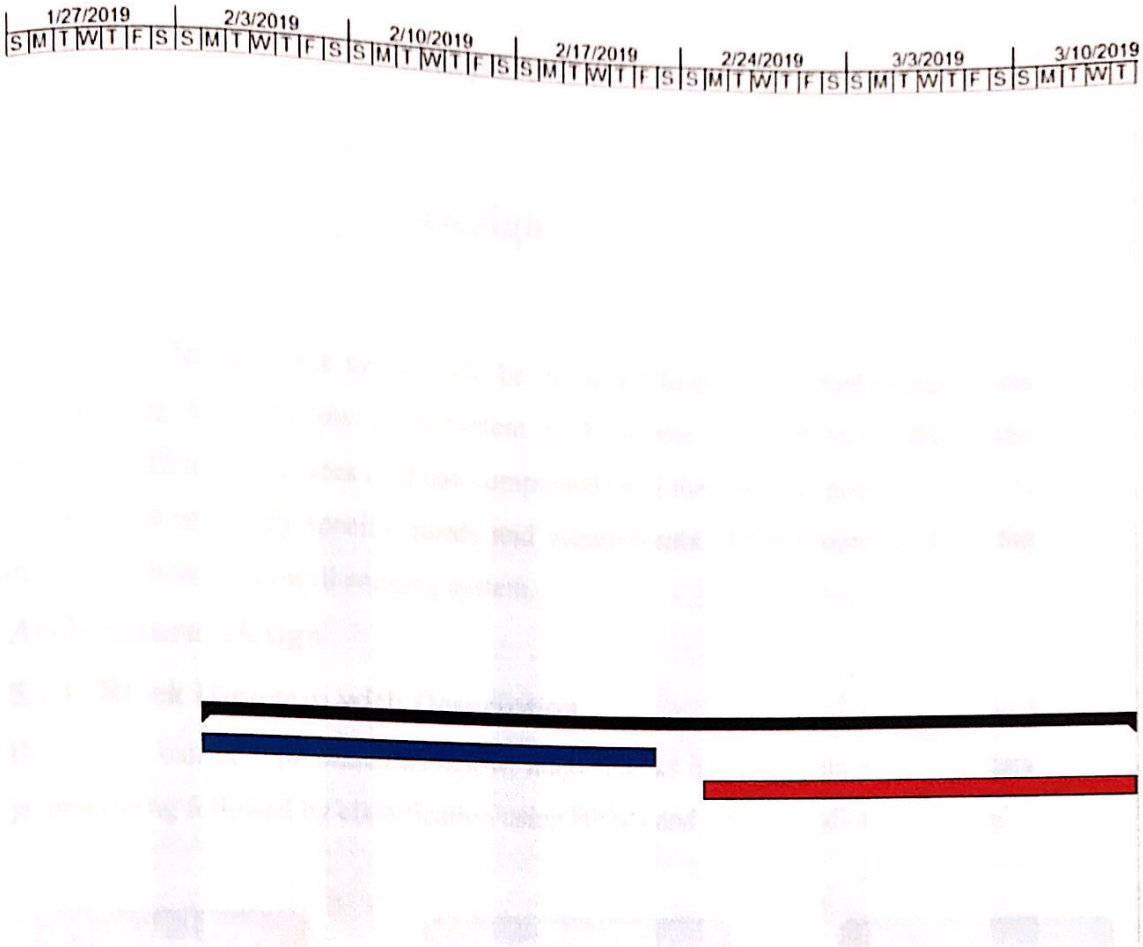

Fig 4.7 – Timeline Bar Chart continued

Chapter 5

Design

In this chapter, the design of the system will be discussed in detail. System design is the process of defining the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through that system. It is meant to satisfy specific needs and requirements of the project through the engineering of a coherent and well-running system.

5.1 Architectural Design

5.1.1 Block Diagram with Description

Our system can be represented as below, major blocks of the system start with data preprocessing followed by classification using BPNN and finally prediction is done.

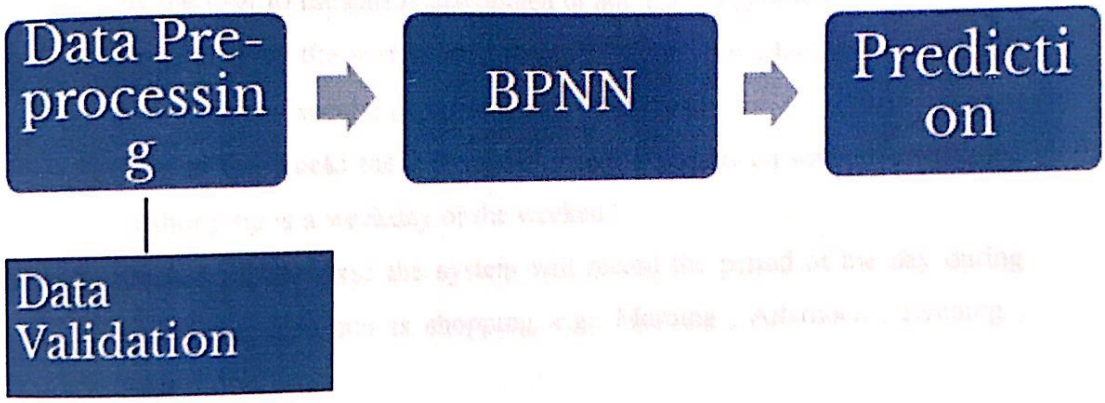

Fig 5.1 – Block Diagram

a. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.Data preprocessing prepares raw data for further processing. Different parameters are selected for the said purpose of prediction.

These parameters are decided after analyzing the current e-business scenarios and which are best suitable and impacting customers the most.

Input Parameters (With Linguistic variables used) for prediction-

1. **Amount of clicks:** the system records the clicks when the customer is surfing through the web site .These amount of clicks are categorized based on frequency/repetition. e.g. Low, Moderate, High
2. **Clicked item:** the system will record the items which the customer clicks on and will categorize the item based on the category to which it belongs. e.g.: apparel, home appliances or electronics.
3. **Click number:** the system will record that among a list of items which item was clicked on first, second, etc. e.g. First preference, Second preference or Third preference.
4. **Click time:** the system records the stickiness of the user to a particular item i.e. amount of time spent looking at an item over a given time period. e.g.: Stickiness High , Medium or Low
5. **Discounted item:** the system will keep a record of weather the item added by the user to the cart is discounted or not. e.g.: Yes or No
6. **Special day:** the system will record if the day on which the customer is shopping is a special day or not. e.g.: Yes , No
7. **Day of the week:** the system will record if the day on which the customer is shopping is a weekday or the weekend.
8. **Period of the day:** the system will record the period of the day during which the customer is shopping. e.g.: Morning , Afternoon , Evening , Night
9. **Cart:** the system will record if the item that the customer is clicking on and viewing is finally added by the customer to the cart or not. e.g.: Yes , No
10. **Gender:** the system will record if the item is viewed by a male customer or a female customer.
11. **Age:** the system will record the age of the customer viewing the item and categorize them based on their age groups. e.g.: 14-20 , 21-30 or 31 and above
12. **Demography:** the system will record the location of the customer. e.g.: Mumbai, Bangalore, etc.

b. Data Validation

Data validation is a process that ensures the delivery of clean and clear data to the programs, applications and services using it. Data validation is also known as input validation. It checks for the integrity and validity of data that is being inputted to different software and its components.

Data validation primarily helps in ensuring that the data sent to connected applications is complete, accurate, secure and consistent. This is achieved through data validation's checks and rules that routinely check for the validity of data. These rules are generally defined in a data dictionary or are implemented through data validation software.

Validation of the data will done using any of the following methods:

a. **ANOVA:**it stands for Analysis Of Variance. It is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found (hence its name). Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups. When we will be testing the different groups of each input data to see if there's a difference between them, ANOVA will help us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis.

b. **K Fold Cross Validation:**The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. It uses a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups.
3. For each unique group:
    1. Take the group as a hold out or test data set.
    2. Take the remaining groups as a training data set.

3. Fit a model on the training set and evaluate it on the test set.
4. Retain the evaluation score and discard the model.

4. Summarize the skill of the model using the sample of model evaluation scores.

c. **Evaluation on training, testing and checking errors:** Training, checking and testing errors will be checked for different data samples

d. **MSE:** Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss.

c. BPNN

Back Propagation Neural Network is a type of ANN algorithm that will be used on the data sets for the prediction.

BPNN is a supervised algorithm in which error difference between the desired output and calculated output is back propagated. The procedure is repeated during learning to minimize the error by adjusting the weights through the back propagation of error.

Purpose for using BPNN:

1. BPNN supports high speed classification.
2. BPNN can be used for linear as well as non-linear classification.
3. BPNN supports multi class classification.

d. Prediction

It is the final result of the algorithm i.e. the output whether customer will buy the product or not.

Advantages of prediction:

1. Analyze customer's behavior patterns.
2. Recognize important customers.

3. Recognize the areas of the website that need improvement.
4. Recognize the demand for items and accordingly make changes in the availability of these items.

5.2 User Interface Design

The UI of the system consists of setting up a dummy website to collect data for the above mentioned parameters,

1. Simple homepage which consists of different products, a cart button, sign in and sign up buttons and a search tab as shown in fig no. 5.2.

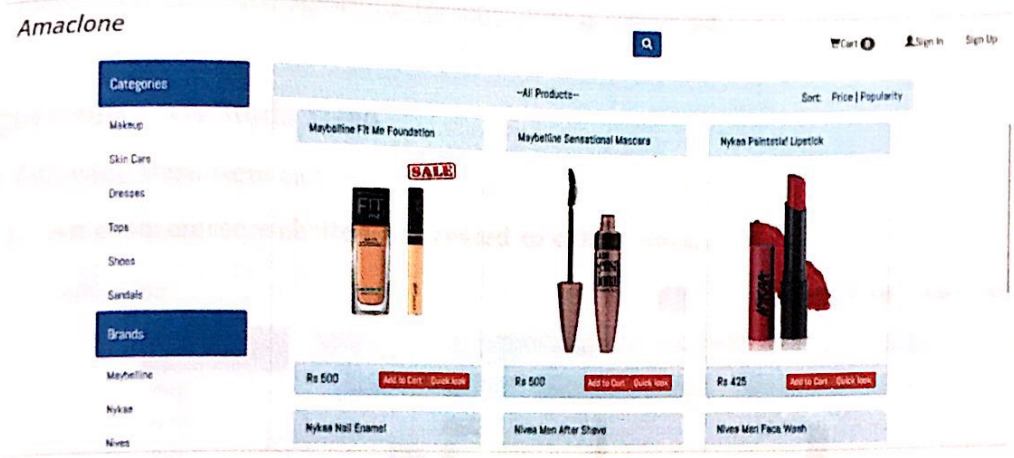

Fig 5.2 – Home page of dummy website

2. User registration for first time log in.

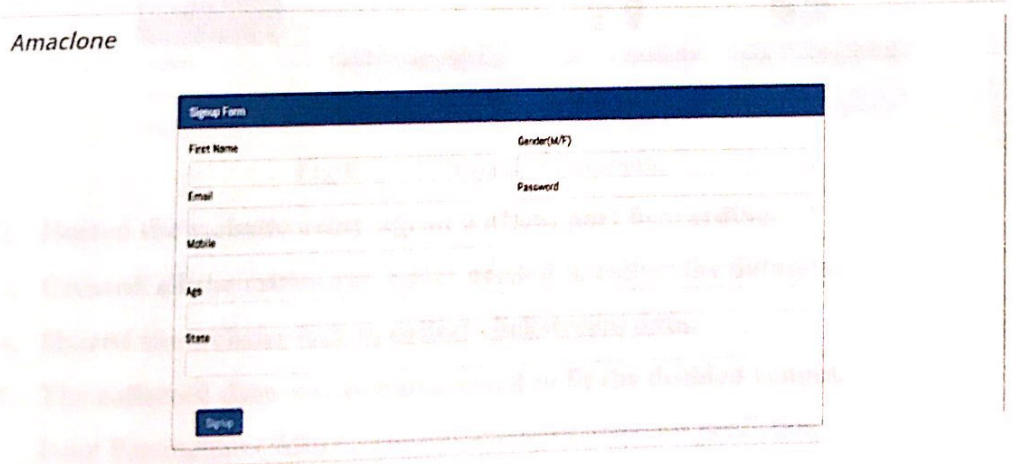

Fig 5.3 – Registration page

Chapter 6

Implementation

Implementation is a realization of a technical specification or algorithm as a program, software component, or other computer system through computer programming and deployment. Many implementations may exist for a given specification or standard. In this chapter, the proposed algorithm for execution is discussed.

6.1 Algorithms / Methods Used

The following steps were carried out during implementation:

1. **An ecommerce website was created to collect data.**

Fig 6.1 – E-commerce website

2. **Hosted the website using ngrok without port forwarding.**
3. **Created all the tables and codes needed to collect the datasets.**
4. **Shared the website link to collect clickstream data.**
5. **The collected data was pre-processed to fit the decided ranges.**

   Input Parameters (With Linguistic variables used) for prediction:

   a) Amount of clicks :- <2 (0-Low), >=2&<6 (1-Average), >=6 (2-High)
   b) Category :- Makeup(0), Clothing(1), Footwear(2)
   c) Product preference :- High(0), Medium(1), Low(2)
   d) Stickiness:- High(2) , Medium(1) , Low(0)
   e) Discounted item :- Yes(1) , No(0)

f) Special day :- Yes(1) , No(0)

g) Day of the week :- Week day(0), Week end(1)

h) Period of the day :- 12am-12pm(1-Morning), 12pm-4pm(2-Afternoon), 4pm-8pm(3-Evening), 8pm-12am(4-Night)

i) Cart :- Yes(1) , No(0)

j) Gender :- Male(0) , Female(1)

k) Age :- 14-18 (0), 19-35(1) , 35 and above(2)

l) Demography :- Urban(1), Rural(0)

6. **The pre-processed data was normalized.**

Normalization of the pre-processed data was done using the Min-Max formula.

PROID	CATEGORY	GENDER	AGE	DEMOGRAPHY	DAY	TIME	SPECIAL_DAY	STICKINESS	DISCOUNT	CLICKS	PREFERENCE	CART	BUY
0	0	0	0.5	1	1	0	0	0	1	0	0	0	1
0.25	0.5	0	0.5	1	1	0	0	0.5	1	0	0	1	0
0.795455	1	0	0.5	1	1	0	0	1	0	0	1	0	0
0.909091	1	0	0.5	1	1	0	0	0	0	0	1	0	1
0.159091	0	1	0.5	0	1	0	0	0	0	0	0	0	0
0.090909	0	1	0	1	1	0	0	0.5	0	0	0	0	1
0.113636	0	1	0	1	1	0	0	0.5	0	0	0.5	0	1
0.681818	1	1	0	1	1	0	0	0	0	0	0	1	0
0.045455	0	0	1	1	1	0	0	0.5	0	0	0	0	0
0.068182	0	0	1	1	1	0	0	0.5	0	0	0	0	0
0.204545	0.5	0	1	1	1	0	0	0	0	0.5	0	0	1
0.227273	0.5	0	1	1	1	0	0	0.5	0	0	0	0	0
0.25	0.5	0	1	1	1	0	0	0.5	1	0	0	0	0
0.840909	1	0	1	1	1	0	0	0.5	0	0.5	1	0	0
0.863636	1	0	1	1	1	0	0	1	1	0	0.5	0	0
0.068182	0	0	0.5	1	1	0.333333	0	0.5	0	0.5	0	0	0
0.136364	0	0	0.5	1	1	0.333333	0	0	0	0	0	0	0
0.181818	0	0	0.5	1	1	0.333333	0	0.5	1	0	0.5	0	0
0.613636	0.5	0	0.5	1	1	0.333333	0	0	0	0	1	1	0
0	0	0	0.5	1	1	0.333333	0	1	1	0	0	0	0
0.068182	0	0	0.5	1	1	0.333333	0	0.5	0	0	0	1	0
0.636364	0.5	0	0.5	1	1	0.333333	0	0	0	0	0.5	1	0
0.090909	0	1	0	1	1	0.333333	0	0.5	0	0	0	0	1
0.113636	0	1	0	1	1	0.333333	0	0.5	0	0	0.5	0	0

Fig 6.2 – Normalized dataset

7. **Data validation was performed on the dataset.**

Validation was done using:

a) ANOVA(Analysis Of Variance)

After performing ANOVA we found out that:

- The result is most affected by the Day parameter because users mostly tend to buy on weekends rather than weekdays.

- The result is least affected by the Cart parameter as it's not necessary that the items added to the cart will be bought by the user.

Parameter	Result (Buy)
Day	0.2498
Discount	0.2488
Special Day	0.2275
Time	0.2127
Gender	0.1875
Category	0.1678
Demography	0.1444
Stickiness	0.1148
Prodid	0.1051
Clicks	0.1047
Preference	0.0894
Age	0.0623
Cart	0.0148

Fig 6.3 – ANOVA on dataset

b) K-Fold Cross Validation

K-fold Cross Validation was done using nntool in MATLAB.

Results

	Samples	MSE	%E
Training:	1322	1.05899e-1	10.59001e-0
Validation:	283	8.48051e-2	8.48056e-0
Testing:	283	1.27207e-1	12.72084e-0

Fig 6.4 – K-Fold cross validation on data set (Fold 1)

Results

	Samples	MSE	%E
Training:	1322	1.03629e-1	10.36308e-0
Validation:	283	9.89389e-2	9.89399e-0
Testing:	283	1.23673e-1	12.36749e-0

Fig 6.5 – K-Fold cross validation on data set (Fold 2)

The error in the first fold was lesser so we used the first dataset combination.

8. **BPNN algorithm was used to do prediction using MultilayerPerceptron in WEKA.**

- BPNN stands for Back Propagation Neural Network.
- BPNN is a supervised algorithm in which error difference between the desired output and calculated output is back propagated.
- The procedure is repeated during learning to minimize the error by adjusting the weights through the back propagation of error.
- Purpose for using BPNN:
    a) BPNN supports high speed classification.
    b) BPNN can be used for linear as well as non-linear classification.
    c) BPNN supports multi class classification.
- MultilayerPerceptron in WEKA
    a) WEKA tool was used to do the classification.
    b) The function MultilayerPerceptron in WEKA uses Back Propagation to classify instances.
    c) 70% of the data was used as training set.
    d) 30% of the data was used for testing on the trained model.
    e) The prediction was obtained in the form of 'Yes'/'No' (whether the user will buy the product or not).

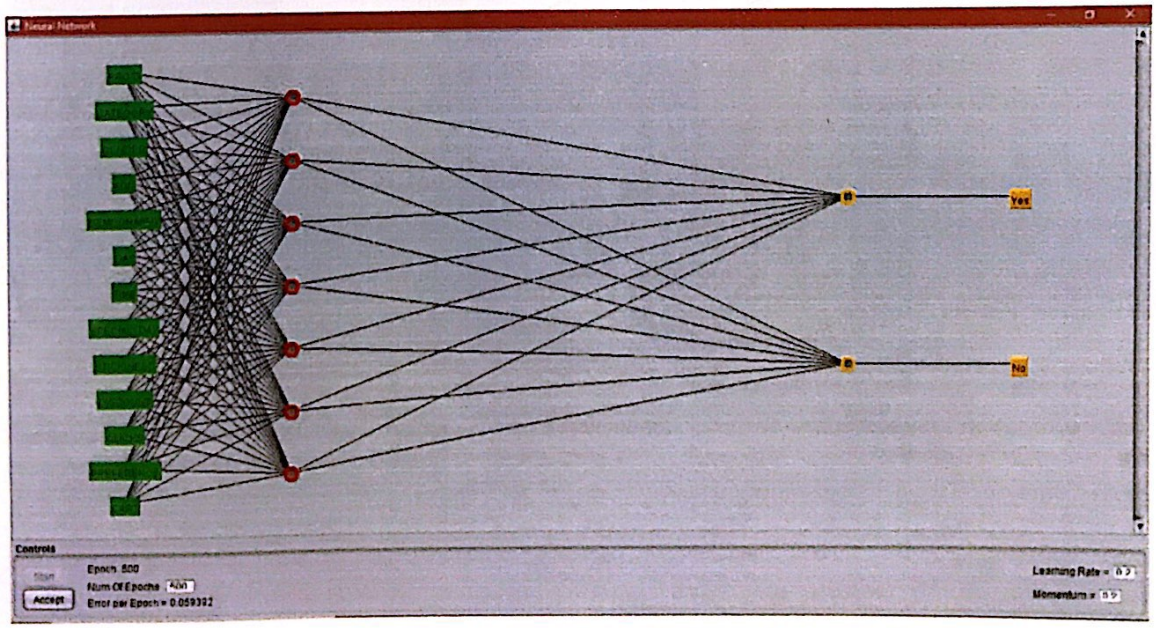

Fig 6.6 – BPNN when no. of epochs=500 & learning rate=0.2

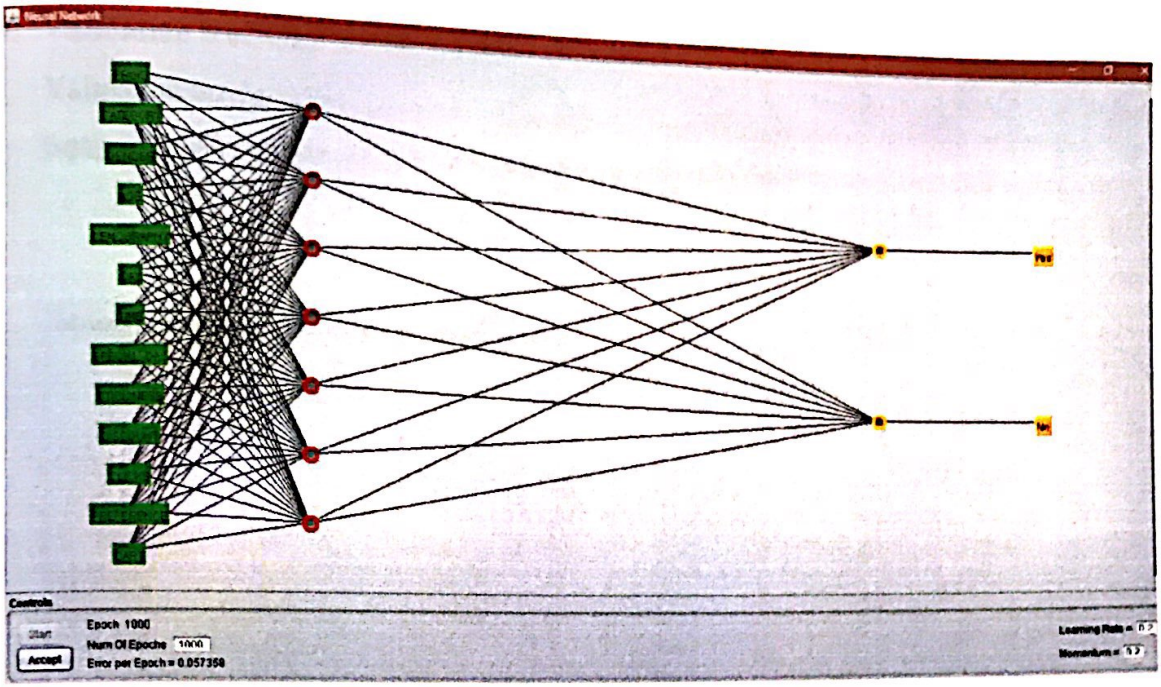

Fig 6.7 – BPNN when no. of epochs=1000 & learning rate=0.2

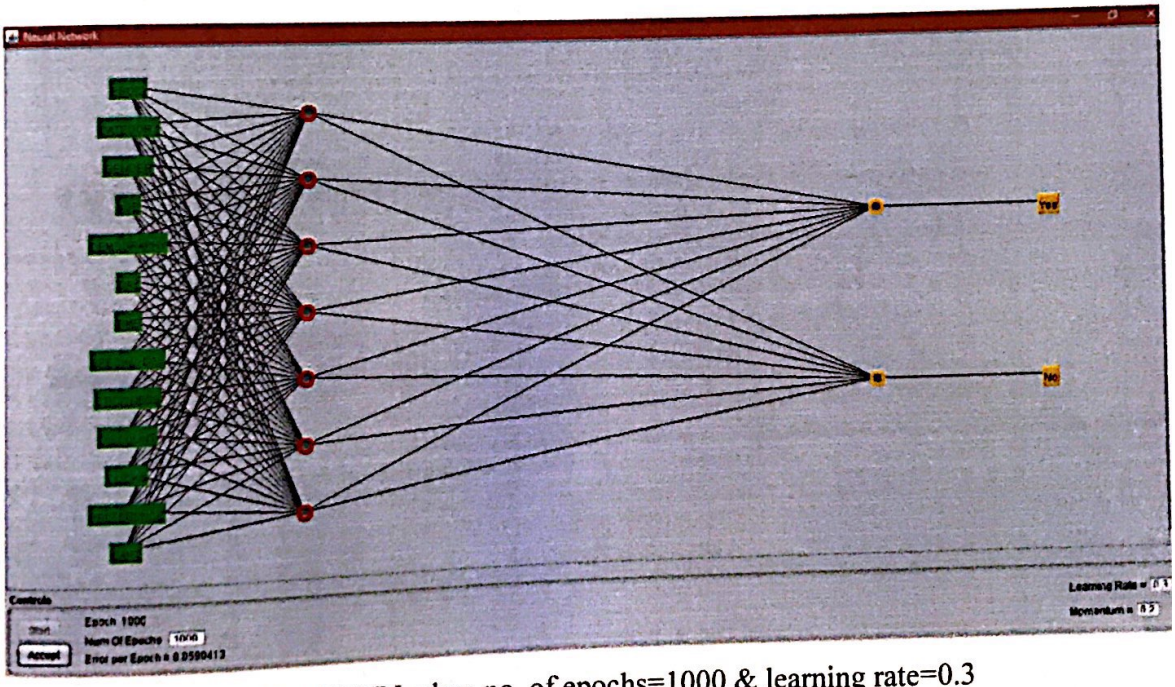

Fig 6.8 – BPNN when no. of epochs=1000 & learning rate=0.3

9. **Validation was done for the predicted result.**

Validation on the predicted result was performed using RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error)

observed value	predicted value	difference	RSME	MSE
2	2	0	0.493	0.243386
2	2	0		
2	2	0		
2	2	0	Yes =	1
2	2	0	No =	2
2	2	0		
2	2	0		
2	2	0		
2	1	1		
2	2	0		
2	2	0		
2	2	0		
2	2	0		
2	2	0		
2	2	0		
2	2	0		

Fig 6.9 – RMSE on predicted result

Actual	Forecast	Error	Absolute Value of Error	Square of Error	Absolute Values of Errors Divided by Actual Values	n	MAPE
A_t	F_t	$A_t - F_t$	$\lvert A_t - F_t \rvert$	$(A_t - F_t)^2$	$\lvert (A_t - F_t)/A_t \rvert$	567	12.87478
2	2	0	0	0	0		
2	2	0	0	0	0	Yes =	1
2	2	0	0	0	0	No =	2
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	1	1	1	1	0.5		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		

Fig 6.10 – MAPE on predicted result

6.2 Working of the project

1. **Function used in Weka:**

   public class MultilayerPerceptron

   extendsAbstractClassifier

   implementsOptionHandler, WeightedInstancesHandler, Randomizable, IterativeClassifier

2. **BPNN Algorithm**

   initialize network weights (often small random values)

   do

   forEach training example named ex

   prediction = neural-net-output(network, ex) *// forward pass*

   actual = teacher-output(ex)

   compute error (prediction - actual) at the output units

   compute {\displaystyle \Delta w_{h}} for all weights from hidden layer to output layer *// backward pass*

   compute {\displaystyle \Delta w_{i}} for all weights from input layer to hidden layer *// backward pass continued*

   update network weights *// input layer not modified by error estimate*

   until all examples classified correctly or another stopping criterion satisfied

   return the network

Chapter 7

Testing

Software Testing is evaluation of the software against requirements gathered from users and system specifications. Testing is conducted at the phase level in software development life cycle or at module level in program code. Software testing comprises of Validation and Verification.

7.1 Test cases

Table 7.1 – Test cases for Login

Test Case No.	Test Case Description	Test Input	Expected Output	Actual Output	Result
1.	To check valid username and password	Username: abc Password: abc	User should get logged in and homepage is displayed.	User is logged in and homepage is displayed.	Pass
2.	To check when invalid username or password is entered	Username: abc Password: 123	"Invalid username or password!" message should be displayed.	"Invalid username or password!" message is displayed.	Pass
3.	To check empty username	Username: Password: abc	"Please enter a valid username" message should be displayed.	"Please enter a valid username" message is displayed.	Pass

4.	To check empty password	Username: abc Password:	"Please enter a valid password" message should be displayed.	"Please enter a valid password" message is displayed.	Pass

Table 7.2 – Test cases for Cart

Test Case No.	Test Case Description	Test Input	Expected Output	Actual Output	Result
1.	To check that user is not able to add to cart without login.	Click on the add to cart button without logging in.	"Please login to add to cart" message should be displayed.	"Please login to add to cart" message is displayed.	Pass
2.	To check that the user is able to add to the cart when logged in.	Click on the add to cart button while being logged in.	The item should be added to the cart and "successfully added to cart" message should be displayed.	The item is added to the cart and "successfully added to cart" message is displayed.	Pass
3.	To check that the user is taken to the payment page after clicking on checkout.	Click on the checkout button.	The user should be navigated to the payment page.	The user is navigated to the payment page.	Pass

Table 7.3 – Test cases for Prediction

Test Case No.	Test Case Description	Test Input	Expected Output	Actual Output	Result

1.	To check if the prediction that "the user will buy the product" is true.	The user buys the product.	System should predict that the user will buy the product.	System predicted that the user will buy the product.	Pass
2.	To check if the prediction that "the user will buy the product" is true.	The user buys the product.	System should predict that the user will buy the product.	System didn't predict that the user will buy the product.	Fail
3.	To check if the prediction that "the user will not buy the product" is true.	The user does not buy the product	System should predict that the user will not buy the product.	System predicted that the user will not buy the product.	Pass
4.	To check if the prediction that "the user will not buy the product" is true.	The user does not buy the product	System should predict that the user will not buy the product.	System didn't predict that the user will not buy the product.	Fail

7.2 Type of Testing used

Black Box Testing

Black Box Testing also known as Behavioral Testing, is a software testing method in which the internal structure/design/implementation of the item being tested is not known to the tester. These tests can be functional or non-functional, though usually functional.

This method is named so because the software program, in the eyes of the tester, is like a black box; inside which one cannot see. This method attempts to find errors in the following categories:

a. Incorrect or missing functions
b. Interface errors
c. Errors in data structures or external database access
d. Behavior or performance errors
e. Initialization and termination errors
f. White Box Testing

White Box Testing

White Box Testing is a software testing method in which the internal structure/design/implementation of the item being tested is known to the tester. The tester chooses inputs to exercise paths through the code and determines the appropriate outputs. Programming know-how and the implementation knowledge is essential. White box testing is testing beyond the user interface and into the nitty-gritty of a system.

This method is named so because the software program, in the eyes of the tester, is like a white/transparent box; inside which one clearly sees.

Chapter 8

Results and Discussions

1. Validation

a. ANOVA(Analysis Of Variance)

- Analysis of Variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not.
- The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.
- ANOVA determines the influence that independent variables have on the dependent variable.
- ANOVA test was used as a way to find out if the parameters considered in the dataset are significant. In other words, they helped to figure out if we need to reject the null hypothesis or accept the alternate hypothesis.

- After performing ANOVA we found out that:
  a. The result is most affected by the Day parameter.
  b. The result is least affected by the Cart parameter.

Parameter	Result (Buy)
Day	0.2498
Discount	0.2488
Special Day	0.2275
Time	0.2127
Gender	0.1875
Category	0.1678
Demography	0.1444
Stickiness	0.1148
Prodid	0.1051
Clicks	0.1047
Preference	0.0894
Age	0.0623
Cart	0.0148

Fig 8.1 – ANOVA on dataset

b. K-Fold Cross Validation

K-fold cross validation is performed as per the following steps:

1. Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as $f_1, f_2, \ldots, f_k$.
2. For i = 1 to i = k
   a. Keep the fold f_i as Validation set and keep all the remaining k-1 folds in the Cross validation training set.
   b. Train your machine learning model using the cross validation training set and calculate the accuracy of your model by validating the predicted results against the validation set.
   c. Estimate the accuracy of your machine learning model by averaging the accuracies derived in all the k cases of cross validation.

It is used to compare the performance of different machine learning models on the same data set.

K-fold Cross Validation was done in MATLAB.

Results

	Samples	MSE	%E
Training:	1322	1.05899e-1	10.59001e-0
Validation:	283	8.48051e-2	8.48056e-0
Testing:	283	1.27207e-1	12.72084e-0

Fig 8.2 – K-Fold cross validation on data set (Fold 1)

Results

	Samples	MSE	%E
Training:	1322	1.03629e-1	10.36308e-0
Validation:	283	9.89389e-2	9.89399e-0
Testing:	283	1.23673e-1	12.36749e-0

Fig 8.3 – K-Fold cross validation on data set (Fold 2)

The error in the first fold was lesser so we used the first dataset combination.

c. RMSE

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

Validation on the predicted result was performed using RMSE (Root Mean Square Error)

observed value	predicted value	difference	RSME	MSE
0	0	0	0.493	0.243386
0	0	0		
0	0	0		
0	0	0		
0	0	0		
0	0	0		
0	0	0		
0	0	0		
0	1	-1		
0	0	0		
0	0	0		

Fig 8.4 – RMSE on predicted result

d. MAPE

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Where A_t is the actual value and F_t is the forecast value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying it by 100% makes it a percentage error.

Actual	Forecast	Error	Absolute Value of Error	Square of Error	Absolute Values of Errors Divided by Actual Values	n	MAPE
A_t	F_t	A_t-F_t	$\lvert A_t-F_t \rvert$	$(A_t-F_t)^2$	$\lvert (A_t-F_t)/A_t \rvert$	567	12.87478
2	2	0	0	0	0		
2	2	0	0	0	0	Yes =	1
2	2	0	0	0	0	No =	2
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	1	1	1	1	0.5		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		
2	2	0	0	0	0		

Fig 8.5 – MAPE on predicted result

2. Output

The final prediction obtained using WEKA is as follows:

uid	proid	category	gender	age	demography	day	time	special_day	stickiness	discount	clicks	preference	cart	'predicted BUY'
216	25	1	1	1	1	1	2	0	2	0	0	1	0	No
216	26	1	1	1	1	1	2	0	2	0	0	2	0	No
216	27	1	1	1	1	1	2	0	2	0	0	2	0	No
216	32	2	1	1	1	1	2	0	1	1	0	1	0	No
216	36	2	1	1	1	1	2	0	2	0	0	2	0	No
216	37	2	1	1	1	1	2	0	0	0	0	2	0	No
216	40	2	1	1	1	1	2	0	0	0	0	2	0	No
216	41	2	1	1	1	1	2	0	0	0	0	2	0	No
217	1	0	1	1	1	1	2	0	2	1	0	0	0	Yes
217	2	0	1	1	1	1	2	0	1	0	0	0	0	No
217	7	0	1	1	1	1	2	0	0	0	0	0	0	No
217	8	0	1	1	1	1	2	0	0	0	0	0	0	No
217	12	1	1	1	1	1	2	0	1	1	1	0	0	No
217	13	1	1	1	1	1	2	0	0	0	1	1	0	No
217	14	1	1	1	1	1	2	0	1	0	0	0	0	No
217	15	1	1	1	1	1	2	0	2	0	0	1	0	No
217	19	1	1	1	1	1	2	0	1	0	0	2	0	No
217	20	1	1	1	1	1	2	0	1	0	0	1	0	No
217	22	1	1	1	1	1	2	0	1	0	0	1	0	No
217	28	1	1	1	1	1	2	0	0	0	0	2	1	No
217	29	1	1	1	1	1	2	0	0	0	0	1	1	No
217	30	1	1	1	1	1	2	0	0	0	0	2	0	No
217	42	2	1	1	1	1	2	0	0	0	0	2	0	No
217	43	2	1	1	1	1	2	0	1	1	1	0	0	No

Fig 8.6 – Prediction obtained using WEKA

- The RMSE (Root Mean Square Error) for this predicted result was found out to be 0.493.
- The MAPE (Mean Absolute Percentage Error) for this predicted result was found out to be 12.87478.
- The result can be used for marketing strategies for the website.
- Either the user can be given some discount for a particular product that he/she didn't buy or the user can be provided with some coupons.
- That way the website owners will profit more.

Chapter 9

Conclusions & Future Scope

In this system, we will introduce an approach so as to predict customers' shopping patterns from mouse movements and website logs in order to predict if a customer will buy the items which he/she has added to his /her basket.Artificial Neural Networks model (BPNN) will be used to predict online customers' behavior patterns. The main purpose behind marketing a product is to satisfy demands and wants of the consumers. Study of consumer behaviour helps to achieve this purpose. With the help of this prediction the seller will be able to convert a web surfer to consumer.

As a future scope, we can further develop this system into a recommendation system where in based on a customer's behavior, certain products will be recommended to him/her.

Literature Cited

IEEE standard

[1] G. Sılahtaroğlu and H. Dönertaşli, "Analysis and prediction e-customers' behaviour by mining clickstream data," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 1466-1472.

[2] S. Cheriyan and K. Chitra, "Web page prediction using Markov model and Bayesian statistics," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2017, pp. 1-6.

[3] Q. Jiang, C. Tan, C. W. Phang and K. K. Wei, "Using Sequence Analysis to Classify Web Usage Patterns across Websites," *2012 45th Hawaii International Conference on System Sciences*, Maui, HI, 2012, pp. 3600-3609.

[4] L. Na, P. Geng, C. Hang and B. Jiaxing, "A prediction study on E-commerce orders based on site search data," *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Xi'an, 2013, pp. 314-318.