Customer Behaviour Prediction Using Machine Learning

A Project Report

Submitted by

Sheetal Jade 121742009

in partial fulfilment for the award of the degree

of

M.Tech

(Information Security)

Under the guidance of

Prof. Siddharth K. Gaikwad

College of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY,
COLLEGE OF ENGINEERING, PUNE-5

August, 2019

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY, COLLEGE OF ENGINEERING, PUNE

CERTIFICATE

Certified that this project, titled "Customer Behaviour Prediction Using Machine Learning" has been successfully completed by

Sheetal Jade 121742009

and is approved for the partial fulfilment of the requirements for the degree of "M.Tech. Information Security".

SIGNATURE

Prof. Siddharth K. Gaikwad
Project Guide
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

SIGNATURE

Dr. V. Z. Attar
Head
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

Acknowledgement

Apart from the efforts of myself, the success of any project depends largely on direct-indirect help, encouragement, motivation and support of many others. First and foremost, I would like to express my gratitude towards my guide Prof. S. K. Gaikwad for their valuable guidance and advice throughout the project work. I am lifetime grateful to my family for their all-time encouragement, support, assistance and believing in me, which always keep me motivated and revitalized. Finally, special thanks go to my friends for their understandings and cooperation in the completion of this project.

Abstract

In the era of e-commerce, Customer relationship management became necessity to improve sell of products of company. Because retaining old customer is easy and cheaper than acquiring new customers. So, we successfully predicted behaviour of online shopper with the help of machine learning algorithms. In this work, we compared traditional machine learning techniques like logistic regression, K-Nearest Neighbour, SVM, Nave Bayes, Decision tress with some ensemble methods like Random Forest, Adaboost, GBM, XBG. High values of accuracy, precision score, recall score, f1 score highlights that the random forest algorithm performs very well as compared to other classification algorithms. Finally, we ranked features according to their importance and removed unwanted features so that processing overhead gets reduced for random forest.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Day by day use of internet is increasing. As people started trusting e-commerce, need of predicting customer behaviour is increased. Marketing opportunities are improved due to continuously developing information technology, advancement of internet and explosion in customer data. It changed the way of customer relationship management. Customer relationship management can be done with the help of customer behaviour prediction. It can be used to improve sells of product by giving some discounts, offers to people who prefer less to buy products online [13].

1.1 Consumer Behaviour Prediction

Customer behaviour prediction is process of identifying common behaviour among the group of customers. It is used in customer relationship management. Need of customer behaviour prediction is mentioned below:

- Literature survey have shown that retaining current customer of organization is cheaper as compared to attracting new customers. So, Customer behaviour prediction is used to improve decision making process for retaining valued customers.

- Customer relationship management.

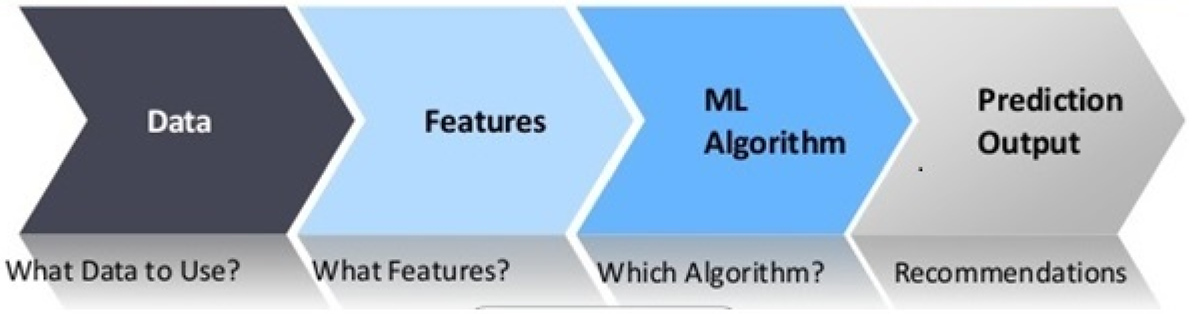


Figure 1.1: Customer Behaviour Prediction model

- Finding how customer spends their time on online shopping websites, how much time it spends on searching for items, most frequent items bought, quantity of items bought.

Figure 1.1 shows that customer behaviour prediction have 4 important steps like data collection, feature selection, ML algorithm and output prediction. First we have to select proper dataset and then do feature engineering to remove unimportant features. Finally, apply proper machine learning algorithm for prediction of behaviour. In-detailed description of these steps is discussed in next chapter. From literature survey, we observed that lots of research has been done in customer behaviour prediction in domain of banking, telecommunication industry. In this paper, we have done customer behaviour prediction in online shopping. The objective of our paper is to improve overall accuracy of the customer behaviour prediction.

1.2 Machine Learning

It is an application of artificial intelligence (AI). It gives ability to automatically learn and improve from experience without being explicitly programmed. Machine learning tries to develop a computer program which can access data and use it to learn for themselves. Some ML algorithms are

- Supervised machine learning algorithms
- Unsupervised machine learning algorithms
- Semi-supervised machine learning algorithms
- Reinforcement machine learning algorithms

In this project, Supervised machine learning algorithms are used for prediction of customer behaviour.

1.3 Classification

Classification tries to predict the class of given data points. Classes can be called as targets or labels or categories. Classification is used when we want to find out category like product is Red or Blue. Classification algorithm classifies data based on training dataset and then tries to predict class of new or testing data. There are number of classification models like Logistic Regression, Decision Tree, Nave Bayes, neural network etc.

1.4 Motivation

Nowadays people are very busy. They dont have time to go to shop for shopping. Customers are approaching towards online shopping. Online shopping has become the third most popular Internet activity, following e-mail using/instant messaging and web browsing. Consumer-retailer relationship structure is dependent on understanding consumer behaviour in online environments. So, customer behaviour prediction has gained attention to improve sell of products. It is influenced by many external and internal factors, but the company can also influence the final process of buyer decision-making significantly by its activities.

1.5 Research Gap

From literature survey, it has been found that different approaches have been used in the field of customer behaviour prediction. But still need to improve prediction accuracy. Challenges presented in literature survey gave a scope to work in the domain of customer behaviour prediction based on what type of data user is visiting, time spent on web pages, which type of page is visited by user. Results of existing approaches can be optimized further for accurate customer behaviour prediction.

1.6 Problem Statement

To propose efficient model for customer behaviour prediction in the domain of online shopping. We aim at predicting online shoppers purchasing intention that whether shopper will purchase the product or not. More in detail, we aim at studying and understanding the current research work being done in this area. Improving feature engineering to improve accuracy of model. Selecting best suitable classification algorithm for customer behaviour prediction.

1.7 Objectives

- To make the literature survey in the field of customer behaviour prediction.
- Selecting best suitable classification algorithm by comparing various algorithms.
- Feature selection to improve accuracy and reduce unnecessary processing overhead.

1.8 Organization of Report

This process consists of five chapters. Second chapter contains literature survey in area of customer behaviour prediction. It explains the factors influencing consumers behaviour, comparison of different approaches for prediction of customer behaviour and performance of Forest Trees algorithm. Third chapter is about hardware-software requirements. The proposed system model is explained in forth chapter. Chapter five contains experimentation and results of proposed model in detail. It contains performance measures of classification algorithms as well as prediction accuracy of model.

Chapter 2

Literature Review

We have done literature survey with the aim to find out the different techniques used in customer behaviour prediction of online shopper. Literature survey shown that lot of work has been done in this area but still there is a scope for accuracy improvement. In previous research, there have been attempts to analyze the effects of customer behaviour on online shopping. Dr. K. Maheswari [1] proposed an approach of prediction of customer behaviour in online shopping using SVM classifier. Author used inventory data and sales data which is available on internet along with R language. Data cleaning and data transformation is done as part of data preprocessing. According to author, the customer who has age less than 7 was attracted more to buy online products in recent years [11]. As a part of future scope, the results of this approach are analyzed with other classification methods.

Mahendra Pratap Yadav [2] discussed about customer behaviour prediction using web usage mining in E-commerce. The Author collected usable customer data of 4263 e-commerce transactions from client log, server log and agent logs. Author used K-means clustering to divide customers into groups. Mine set tool is used for this. As a result, data segmentation is done based on age, gender, marital status, salary. Farshid

Abdi [3] developed technique and framework for Customer behaviour Mining. It has two phases clustering and classification. Clustering is done for grouping of similar customers with K-means clustering algorithm. Best number of clusters can be found out using Davis-Bouldin Index. Feature selection is done with information gain. Second phase is Classification for prediction of customer behaviour with decision tree. Accuracy of classification with Neural network was 0.67 and with decision tree was 0.69.

Femina Bahari T [4], An Efficient CRM-Data Mining Framework built for the prediction of customer behaviour. Data used in this paper is bank marketing data from the University of California at Irvine (UCI). He compared Multilayer Perception Neural Network (MLPNN) and Nave Bayes algorithm. He found that that MLPNN works better than Nave Bayes algorithm with classification accuracy of 0.8863. Weka tool is used to compare these algorithms. The main purpose of Pooja Sharma et al [5] is to study patterns of online grocery shopping in India. Author explored factors and trends influencing online grocery shopping [9,10]. They collected 96 filled forms received from online questionnaire and used as dataset. The author identified patterns of online grocery shopping and completed customer segmentation.

Ge Yunshengi et al [6] done research on the prediction of user behaviour based on neural network with social networking data [7]. Based on the Gauss Newton method, the improved Levenberg-Marquardt algorithm (LM algorithm) belongs to the optimization theory algorithm and has both the local characteristics of the Gauss Newton method and the global characteristic of the gradient method. So, author used LM neural network algorithm for prediction of user behaviour. According to author, Neural network is slow in processing. So, feature engineering [8] should be improved increase speed and accuracy.

RanaAlaa El-DeenAhmeda, M. ElemamShehaba, ShereenMorsya and NermeenMekawiea [11] compared various classification algorithms like Nave bayes, Bayes Net, K Star, classification via clustering, Decision Tree etc. with performance measures TP rate, FP rate, precision, recall, f1 measure, ROC area with WEKA tool. Results have shown that decision table gives best accuracy of 0.8713. Research [14] indicates that machine learning can be used to predict customers who are expected to churn and what are reasons behind their churn. This predictions results can be used to make future market strategies. According to author [17, 18], prediction customer behaviour prediction can be done with Decision Tree classifier, KNN (K-Nearest Neighbour), Logistic Regression,Multi-Layered Perceptron Neural Network (MLPNN), Nave Bayes (NB), Random ForestAlgorithm (RFA), Support Vector Classification algorithm (SVC) . They used this prediction for finding customer churn.

Chapter 3

System Requirement

3.1 Hardware Requirements

- Processor: Intel(R) Core(TM) i3-2350M CPU @ 2.30GHz 2.30 GHz
- RAM: 4 GB RAM
- Disk: 500 GB

3.2 Software Requirements

- Operating System: Windows 8.1 Pro
- OS type: 64-bit Operating System, x64-based processor
- Python 3.7
- Jupyter notebook

Chapter 4

Proposed System

Following figure 4.1 shows the overall system design of the proposed system. It has 5 main blocks shown in figure.

4.1 Data Collection and its description

Data collection is important step for research in area. It is a process of gathering and measuring information related to our area. If we collect faulty data, it results in difficulty in prediction of customer. So, accurate data collection is important for maintaining integrity of prediction. It enables us to predict customer behaviour accurately. We have semi structured data in the form of comma separated file. It does not fit into formal structure of data models associated with relational databases, but It has self-describing structure. Here we have collected Online Shoppers Purchasing Intention dataset from UCI Machine learning repository [16].

Online Shoppers Purchasing Intention Dataset contains 12330 sessions in the dataset. out of which, 84.5% (10,422) were negative class samples that indicates customer did not Purchase the product and the rest (1908) were positive class samples that end with

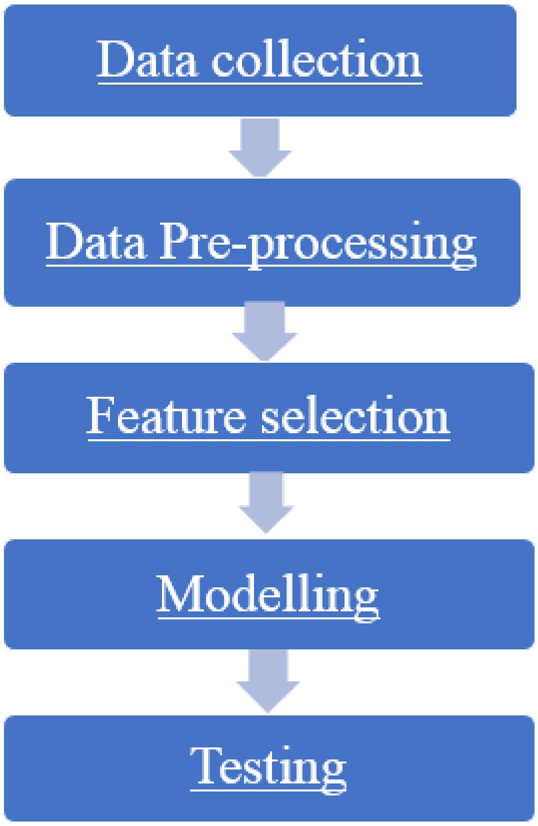


Figure 4.1: Proposed System design

Purchase of the product. Attribute values of this dataset are integer, real. This dataset is donated on 2018-08-31 and have 18 attributes. Revenue is our target variable. Details of these features is described below:

1. Administrative : It indicates the number of different types of administrative pages visited by the visitor in that session.

2. Administrative Duration : Includes total time spent in administrative pages.

3. Informational : It indicates the number of different types of informational pages visited by the visitor in that session.

4. Informational Duration : Includes total time spent in informational pages.

5. Product Related : It indicates the number of different types of product related pages visited by the visitor in that session.

6. Product Related Duration : Includes total time spent in product related pages.

7. Bounce Rate : The value of ”Bounce Rate” feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (”bounce”) search without purchase of product.

8. Exit Rate : The value of ”Exit Rate” feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.

9. Page Value : It is the average value for a web page that a user visited before completing an e-commerce transaction.

10. Special Day : It contains details of the closeness of the site visiting time to a specific special day (e.g. Mothers Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction

11. Operating system : Includes details of type of operating system used.

12. Browser : Includes details of type of browser used.

13. Region : Includes details of region of customer.

14. Traffic type : Includes traffic type details.

15. Visitor type : Indicates type of visitor like returning_visitor, new_visitor or other.

16. Month : Month of the year

17. Weekend : Boolean value indicating whether the date of the visit is weekend.

18. Revenue : Boolean value indicating whether customer purchase the product.

4.2 Data Preprocessing

Second most important step of proposed system is data pre-processing. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and it may contain many errors. So, data pre-processing is a method to resolve such issues.Data pre-processing prepares raw data for further processing. Data goes through a series of pre-processing steps like data cleaning, data integration, data transformation, data reduction, data discretization etc. In proposed model, data cleaning and data transformation is done to prepare data for further processing.

Data Cleaning

It is a process of finding and removing corrupt or incorrect data from dataset. Data of proposed system is cleansed by finding missing values and then filling in missing values.

Data Transformation

It is a process of converting data from one format to another format. Dataset contains two or more labels in one or more than one column. These labels are in the form of words or numbers. Labels in the form of words or numbers are human readable but not machine readable. So, to make data machine readable label encoding is done. Label encoding is process of converting human readable labels into machine readable form [15]. Online shoppers purchasing intention dataset contains some features like Weekend, Revenue which have binary values like True-False. Some features like Month have many labels like Jan, Feb, Mar, Apr etc. So, we converted it to machine readable form with label encoding. For example, True/False converted to 1/0 (Here, label encoding is done according to alphabetical order or numerical order).

4.3 Feature Selection

Feature selection is alternative name of variable selection or attribute selection. It is the process of selection of subset of attributes in data (such as columns in dataset) that are relevant to the proposed model. Feature selection and dimensionality reduction are different. Both methods reduce the number of attributes in the dataset, but a dimensionality reduction method reduce number of features by creating new combinations of attributes, whereas attribute selection method includes necessary feature and excludes unnecessary features present in the data without changing them.

Dataset contains many features that may be relevant or irrelevant, important or unimportant. Feature selection method helps to remove features that are irrelevant or unimportant for model. It helps to create an accurate predictive model by removing unnecessary overhead of processing of unimportant features. Simpler model is easy to understand and explain. So, Feature selection makes model simpler by selecting fewer attributes which

```
<matplotlib.axes._subplots.AxesSubplot at 0x2e13c1ff28>
```

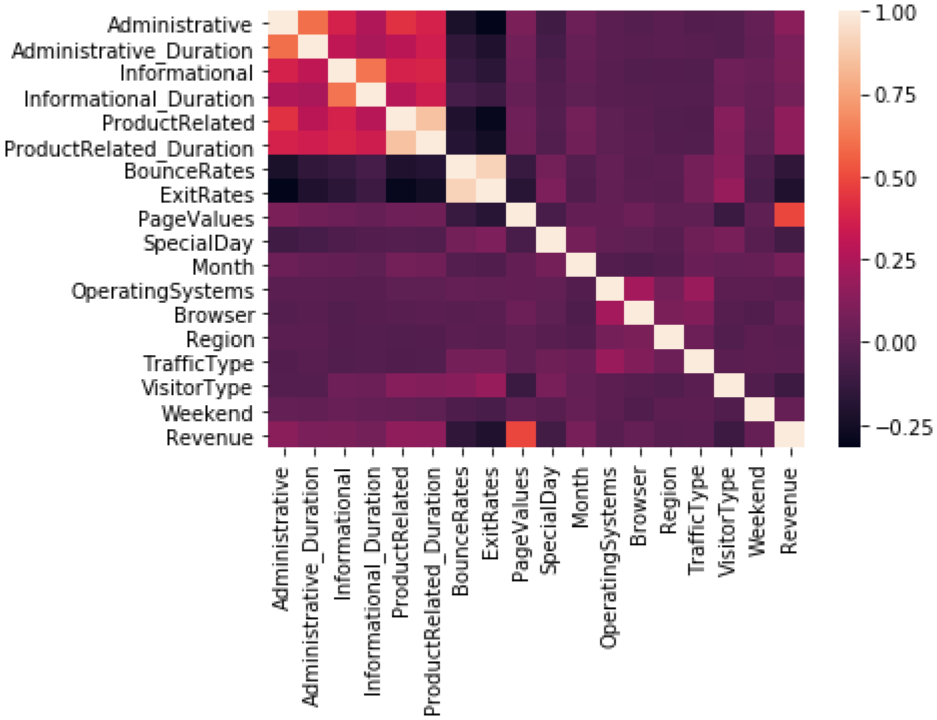


Figure 4.2: Correlation Matrix for features of dataset

reduce complexity of the model.

Correlation

Correlation is process of finding mutual relationship or association among each other. It can be positive or negative. Figure 4.2 shows correlation among features of online shoppers purchasing intention dataset. It shows that ProductRelated_Duration and ProductRelated are positively correlated to each other. Similarly, BounceRates and ExitRates are positively correlated to each other.

After Correlation Matrix, feature ranking is done according to feature importance as shown in figure 4.3. According to feature ranking, we removed less important features and noted results. Figure 4 shows that PageValues (with feature importance score 0.377216) is most important feature among all features. SpecialDay is least important feature with

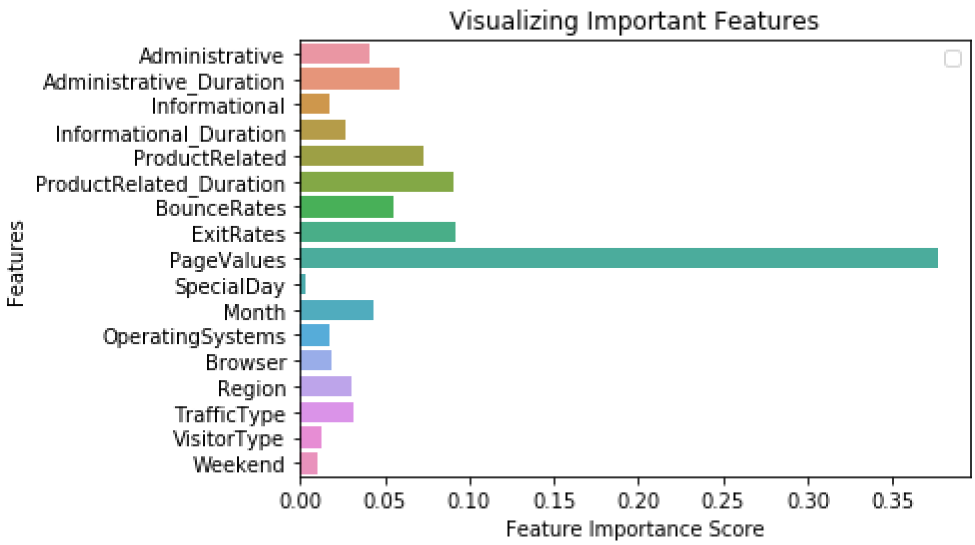


Figure 4.3: Feature Importance graph for features of dataset

feature importance score 0.003404. Some other less important features that can be removed are Weekend (0.010041), VisitorType (0.012600).

4.4 Modelling

The last module in proposed system is the prediction algorithm. In this module, we predicted behaviour of customer that whether he will purchase or not purchase. We applied 10 classification algorithms like adaboost classifier, Decision Tree classifier,GBM (Gradient Boosting Machine), KNN (K-Nearest Neighbour), Logistic Regression, Multi-Layered Perceptron Neural Network (MLPNN), Nave Bayes (NB), Random Forest Algorithm (RFA), Support Vector Classification algorithm (SVC), XGB (XGBoost) [17, 18]. Random forest algorithm performed very well in all cases of data splits i.e. 80-20, 70-30, 60-40 splitting. It gave best prediction accuracy even after removal of less important features. So, Random Forest algorithm is used in proposed system.

Figure 4.4 shows that VisitorType Other (i.e. 1 in figure) did not end with purchase and Purchase done by New_visitor (i.e. Type 0 in figure). To improve Purchase, company

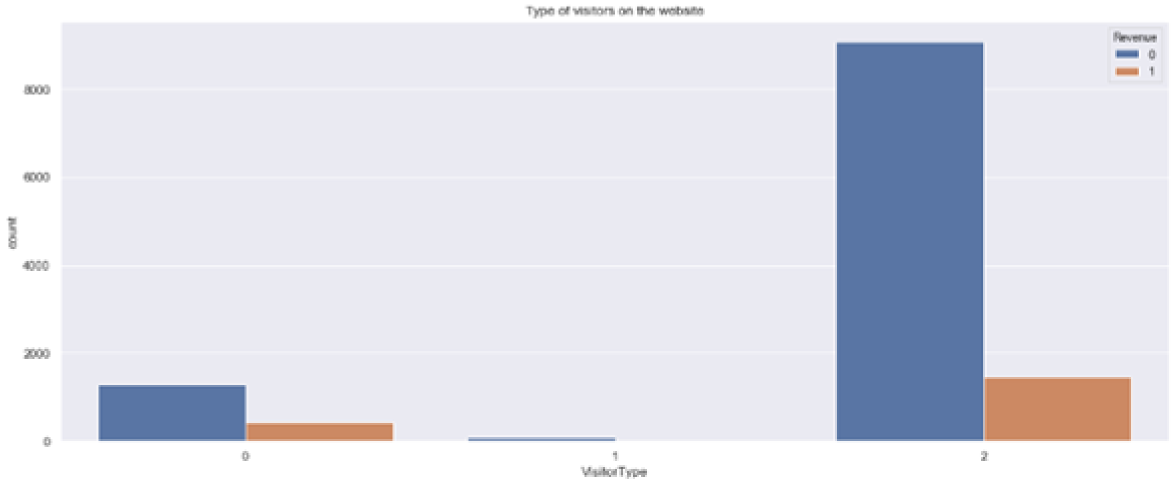


Figure 4.4: Visitor Type and Revenue

needs to focus on these categories of visitor type.

4.5 Training and Testing

Finally training of model is done by splitting data in the form of

80% training 20% testing

70% training 30% testing

60% training 40% testing

Accuracy calculated for all 10 folds and mean accuracy is considered as final accuracy.

Chapter 5

Experimentation and Results

The experimentation of proposed model has been done in python programming language with Online Shoppers Purchasing Intention dataset of UCI Machine Learning Repository. Label encoding is done on dataset to make it machine readable. It is the most important preprocessing step in machine learning. Once data becomes ready, 10 different classification algorithms are applied and their accuracy is tested to get better prediction results. These classification algorithms are adaboost classifier, Decision Tree classifier, GBM (Gradient Boosting Machine), KNN (K-Nearest Neighbour), Logistic Regression, Multi-Layered Perceptron Neural Network (MLPNN), Nave Bayes (NB), Random Forest Algorithm (RFA), Support Vector Classification algorithm (SVC), XGB (XGBoost). Accuracy of models during 10 folds with 80% training data and 20% testing data is shown in table 5.1. Similarly, Accuracy of models during 10 folds with 70% training data and 30% testing data is shown in table 5.2 and accuracy of models during 10 folds with 60% training data and 40% testing data is shown in table 5.3.

Finally, to compare all classification algorithms and to find out one with better prediction accuracy. We calculated minimum accuracy, maximum accuracy and mean accuracy for all 10 classification algorithms. It is summarized in table 5.4. From summery table, we conclude that Random Forest Algorithm performs better in all cases of data split.

Model Name	1	2	3	4	5	6	7	8	9	10
Adaboost	87.9554	88.4498	88.1458	90.8722	88.5395	89.3509	89.2494	89.2494	88.0324	89.3509
Decision Tree	86.9330	88.3369	88.9729	89.1891	88.4324	89.0692	90.5844	89.7186	88.8528	89.9350
GBM	88.7651	88.2472	88.1458	89.7565	88.0324	88.4381	88.5395	89.0466	89.4239	89.4523
KNN	87.2469	87.3353	87.3353	87.8296	87.2210	87.1196	87.3225	87.2210	86.4097	88.4381
LR	89.3617	87.1327	89.6656	89.5643	88.2472	88.2472	88.1338	88.7423	87.9187	87.8172
MLPNN	89.6761	88.9564	88.8551	90.1622	88.6409	88.8438	90.2636	90.5679	88.2352	90.2636
NB	78.8336	76.9978	80.0000	83.1351	80.0000	80.1948	78.6796	79.8701	80.5194	79.9783
RFA	90.6882	89.5643	90.2735	91.2778	89.4523	89.7565	90.6693	91.1764	89.1480	90.6693
SVC	89.4736	88.1458	88.6524	88.5395	88.2352	89.4523	88.8438	88.7423	87.4239	89.0466
XBG	90.5870	87.9432	89.3617	89.7565	89.7565	89.1480	91.0750	89.9594	89.0466	89.5537

Table 5.1: Accuracy of models during 10 folds with 80-20 split of data

Model Name	1	2	3	4	5	6	7	8	9	10
Adaboost	89.0046	87.6157	89.8148	89.2236	87.9490	90.1506	88.9918	88.8760	86.8909	89.3271
Decision Tree	87.9629	88.7731	90.3935	90.2665	88.9918	90.9617	90.4982	89.4553	88.7471	89.7911
GBM	88.6574	88.7731	89.4675	87.8331	88.1807	88.1807	89.2236	88.4125	88.3990	89.3271
KNN	86.1111	87.6157	87.5000	87.3696	86.7902	87.6013	87.8331	86.6743	87.1229	88.1670
LR	88.7731	86.6898	88.5416	89.6990	87.1527	88.0648	88.2830	87.9350	88.8631	89.5591
MLPNN	87.6157	88.7731	89.2361	87.9490	88.0648	89.9188	91.1935	89.1077	88.1670	90.3712
NB	78.4722	76.1574	80.9027	81.1123	79.4901	77.5202	79.9536	77.4044	80.2784	78.8863
RFA	89.4675	89.1203	91.4351	89.6871	89.3395	90.7300	91.5411	90.7300	89.4431	90.6032
SVC	87.5000	89.0046	88.5416	88.5283	88.6442	89.1077	88.7601	88.6442	87.0069	89.5591
XBG	88.1944	88.6574	89.5833	90.0347	88.8760	91.1935	91.0776	89.9188	88.5150	90.1392

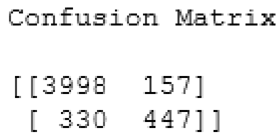
Table 5.2: Accuracy of models during 10 folds with 70-30 split of data

Model Name	1	2	3	4	5	6	7	8	9	10
Adaboost	87.3144	91.0810	88.7837	90.0000	90.2702	89.0540	88.6486	89.5805	86.7388	89.5805
Decision Tree	89.2037	91.2162	90.2702	89.4594	89.7297	90.0000	89.5945	91.0690	87.2801	90.5277
GBM	89.2037	88.9189	87.7027	88.3783	89.0540	89.0540	88.7837	88.9039	88.0920	89.7158
KNN	87.3144	88.1081	86.7567	87.4324	87.0270	87.5675	86.2162	87.6860	87.9566	87.5507
LR	88.1241	89.3387	88.2591	89.5945	88.3783	88.7686	88.7686	88.7686	88.9039	87.2801
MLPNN	87.9892	90.9459	88.1081	89.7297	89.1891	91.0810	90.0000	90.7983	87.5507	91.7456
NB	77.0580	80.2702	78.6486	77.9729	75.6756	74.4594	78.1081	78.4844	78.5014	77.1312
RFA	90.0134	91.3513	90.1351	90.1351	90.5405	91.7567	90.4054	91.4749	88.0920	91.8809
SVC	88.2591	89.7297	87.9729	89.0540	89.3243	88.1081	88.5135	88.3626	87.1447	89.7158
XBG	89.2037	91.2162	89.8648	90.1351	90.1351	91.0810	90.4054	90.3924	88.4979	91.0690

Table 5.3: Accuracy of models during 10 folds with 60-40 split of data

Model Name	60-40			70-30			80-20		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Adaboost	86.7388	91.0810	89.1052	86.8909	90.1506	88.7844	87.9554	90.8722	88.9196
Decision Tree	87.2801	91.2162	89.8350	87.9629	90.9617	89.5841	86.9330	90.5844	89.0024
GBM	87.7027	89.7158	88.7807	87.8331	89.4675	88.9970	88.0324	89.7565	88.5847
KNN	86.2161	88.1081	87.3616	86.1111	88.1670	87.6182	86.4097	88.4381	87.3479
LR	87.2801	89.5945	88.6184	86.6898	89.6990	88.3561	87.1327	89.6656	88.4831
MLPNN	87.5507	91.7456	89.7137	87.6157	91.1935	89.0397	88.2352	90.5679	89.4465
NB	74.4594	80.2702	77.2910	76.1574	81.1123	79.0178	76.9978	83.1351	79.8209
RFA	88.0920	91.8809	90.5785	89.1203	91.5411	90.2097	89.1480	91.2778	90.2676
SVC	87.1447	89.7297	88.6185	87.0069	89.5591	88.5297	87.4239	89.4736	88.6556
XBG	88.4979	91.2162	90.2001	88.1944	91.1935	89.6190	87.9432	91.0750	89.6188

Table 5.4: Minimum, Maximum and Mean Accuracy of models



```
Confusion Matrix

[[3998  157]
 [ 330  447]]
```

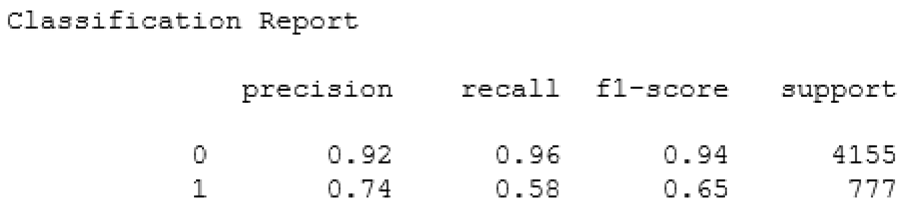
Figure 5.1: Confusion Matrix for RFA

Feature importance graph shown in Figure 4.3 states that PageValue is most important feature as well as SpecialDay and Weekend are less important feature. So, we removed SpecialDay and Weekend. As a result, unnecessary overhead of system gets reduced and accuracy of system increased to 90.42986861119473.

Figure 5.1 shows confusion matrix for random forest algorithm. Confusion matrix is used to describe performance of classification algorithm [10]. In our case, it describes performance of Random Forest Algorithm. First row indicates negative class (Not Purchase) and Second row indicates positive class (Purchase). The value 447 indicates that customer end with Purchase and model predicted it as Purchase. The value 3998 indicates that customer did Not Purchase, and model predicted it as Not Purchase. The value 157 indicates that customer did not purchase but model predicted it as Purchase (FP). The value 330 indicates that customer purchased product, but model predicted it as Not Purchase (FN).

Figure 5.2 shows classification report for Random Forest Algorithm. Classification report for Random Forest Algorithm shows the scores of precision, recall, f1 and support for Random Forest Algorithm[10, 12]. Here 0 indicates value of Not Purchase class and 1 indicates value of Purchase class.

- Precision : It is the ability of classifier to label Positive as Positive and Negative as



```
Classification Report

               precision    recall  f1-score   support

          0       0.92      0.96      0.94      4155
          1       0.74      0.58      0.65       777
```

Figure 5.2: Classification Report for RFA

Negative (i.e Negative sample should not be labelled as Positive).

$$\text{Precision} = TP/(TP+FP)$$

- Recall : It is the ability of finding all positive samples.

$$\text{Recall} = TP/(TP+FP)$$

- F1-score : It is a mean of precision and recall.
- Support : It indicates the number of samples of true responses that is present in class.

Conclusion

To conclude with the analysis, we have understood that customers purchase chances are more if Bounce Rate is below 0.050 and exit Rate below 0.075. The Chances of product purchase is high if productRelated_Duration is between 0-30000 seconds and ProductRelated pages are between 0-300. During data analysis, we observed that customers have preferred Operating System 1,2,3,4 is most frequently used in all region. Browser 2 is used by many customers. Finally, we understood that online purchasing must be emphasized and improved more among New customers (Type 0) and other (Type 1) customers whereas the use of promo codes must be emphasized with both Visitor Types. The future work of this paper can be suggesting promotional tools for improving the sales profit, predicting which products the customer buy most and providing marketing strategies for improving the sales.

Bibliography

[1] Dr.K. Maheswari and P.Packia Amutha Priya, ‘‘Predicting Customer Behavior in Online Shopping Using SVM Classifier”, 2017 IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL, OPTIMIZATION AND SIGNAL PROCESSING.

[2] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav, ‘‘Mining the customer behavior using web usage mining in e-commerce”, ICCCNT’12 26th_2Sdl July 2012, Coimbatore, India.

[3] Farshid Abdi and Shaghayegh Abolmakarem, ‘‘Customer Behavior Mining Framework (CBMF) using clustering and classification techniques”, Journal of Industrial Engineering International. Received: 4 June 2017 / Accepted: 2 August 2018.

[4] Femina Bahari T and Sudheep Elayidom M., ‘‘An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour”, International Conference on Information and Communication Technologies (ICICT 2014).

[5] Pooja Sharma, Vidyalakshmi Nair, Amalendu Jyotishi, ‘‘Patterns of Online Grocery Shopping in India: An Empirical Study”, CONIAAC ’14, October 10 - 11 2014, Amritapuri, India.

[6] Ge Yunshengi, Zhang Qianqian and Kong Jie, ‘‘Research on the prediction of user behavior based on neural network”, ICIIP ’18, May 1920, 2018, Guilin, China.

[7] Zoltan Balogh, ‘‘Analysis of Public Data on Social Networks with Ibm Watson”, Acta Informatica Malaysia, 2(1):10-11(2018).

[8] Zhang R, ‘‘Research on big data feature analysis based on kernel discriminant analysis and neural network[J]”, Multimedia Tools and Applications,2018(10).

[9] https://stats.stackexchange.com/questions/117654/what-does-the-numbers-in-the-classification-report-of-sklearn-mean

[10] https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

[11] RanaAlaa El-DeenAhmeda, M. ElemamShehaba, ShereenMorsya and Nermeen-Mekawiea, ‘‘Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining”, (CSNT), 2015 Fifth International IEEE Conference on 4-6 April 2015, Electronic ISBN: 978-1-4799-1797-6, Printon Demand (PoD) ISBN: 978-1-4799-1798-3.

[12] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.

[13] M.N.Saroja, S.Kannan,K.R. Baskaran, ‘‘Analysing the Purchase Behavior of a Customer for Improving the Sales of a Product”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018.

[14] Mr. Shrey Harsh Baderiya, Prof. Pramila M. Chawan, ‘‘Customer buying Prediction Using Machine-Learning Techniques: A Survey”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 10 — Oct 2018.

[15] https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

[16] https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[17] Harsh Valecha, Aparna Varma and Ishita Khare, ‘‘Prediction of Consumer Behaviour using Random Forest Algorithm”, 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).

[18] Sahar F. Sabbeh, ‘‘Machine-Learning Techniques for Customer Retention: A Comparative Study”, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018..