Customer Behaviour Prediction Using Machine Learning

Guided by

Prof. S. K. Gaikwad

Project by

Sheetal Gautam Jade

121742009

Contents

1. Introduction
2. Motivation
3. Research Gap
4. CBP model
5. Problem Statement
6. Objective
7. Literature Survey
8. System Requirement
9. Proposed System Design
10. Experimentation and results
11. Conclusion
12. Future Scope
13. References

Introduction

- What is customer behaviour prediction?
- It is a process of identifying common behaviour among the group of customers.

- Why it is needed?
- It is used to retain valued customers and retaining current customer of organization is cheaper as compared to attracting new customers.
- Customer Relationship Management.
- 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.

Motivation

- Nowadays people are very busy. They don't 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.

Research Gap

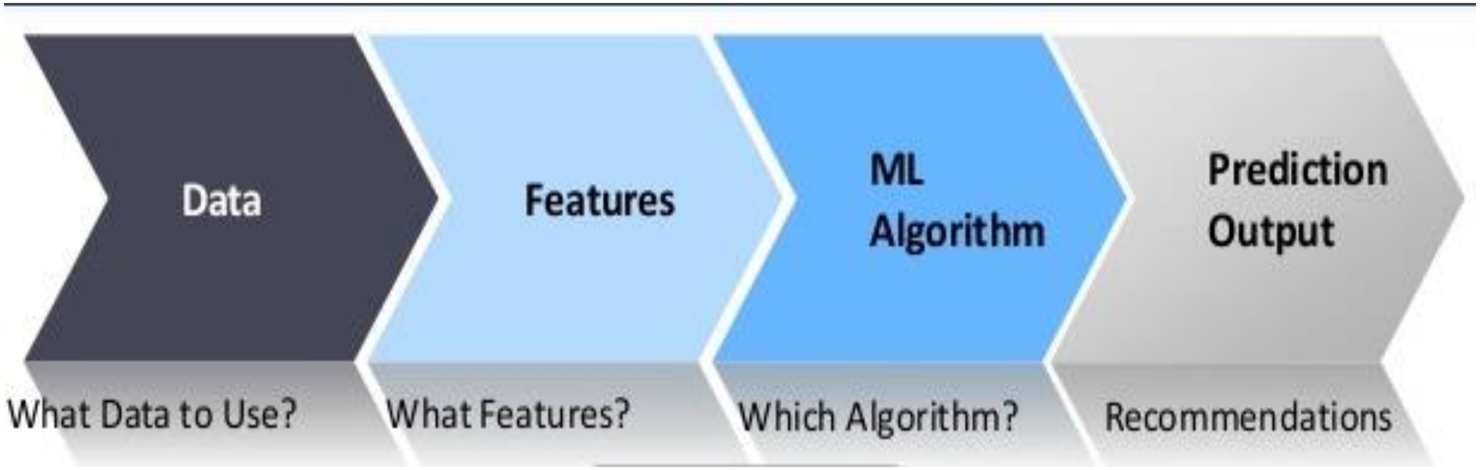
- From literature survey, it has been found that different approaches have been used in the field of Customer Behaviour Prediction. But still need to improve prediction accuracy.

- 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.

Customer Behavior Prediction Model



Problem Statement

- To propose efficient model for customer behavior mining in the domain of online shopping.

- Study and understand the current research work being done in this area.

- Improving feature engineering to improve accuracy of model.

- Selecting best suitable algorithm for classification.

Objective

- To make the literature survey in the field of customer behavior prediction.

- Selecting best suitable classification algorithm by comparing various algorithms.

- Feature selection to improve accuracy and reduce unnecessary processing overhead.

Literature Survey

Author [1] classified customers using SVM into 6 classes. These classes are customer on regular, occasion, festival, offer, window shopping customer, recent customer. Future work for this was to compare results of SVM with other classification methods.

Author [2] proposed CBMF which is divided into 2 phases first phase is customer segmentation based on socio-demographic features. Second phase is prediction behavior of customer. Author used K means clustering for first part and Decision Tree and Neural Network for Behavior prediction.

Research [3] used MLPNN(88.63) and NB(87.97) algorithms for customer behavior prediction in banking. With the help of WEKA tool, he proved that accuracy of MLPNN is better than NB.

Literature Survey

According to [9], customer behavior prediction can be used to increase sales profit. Author used R and implemented **RFA** algorithm. He did data analysis according to gender and concluded that online shopping needs to be promoted among females.

In [15], author mentioned 10 techniques for customer retention of telephonic industry. He got best Maximum accuracy(approximately 96%) for **RFA** and adaboost.

Similarly[16,7,9], did survey in field of online shoppers behavior prediction and stated that good **feature engineering** is necessary to improve accuracy of models.

Literature Survey

From literature survey, we observed that many classification algorithms can be used for customer behaviour prediction in various fields like telephonic industry, banking industry and online shopping but need to improve accuracy of classification for online shopping data.

Feature selection can be done to improve accuracy and reduce unnecessary overhead of processing features.

System Requirement

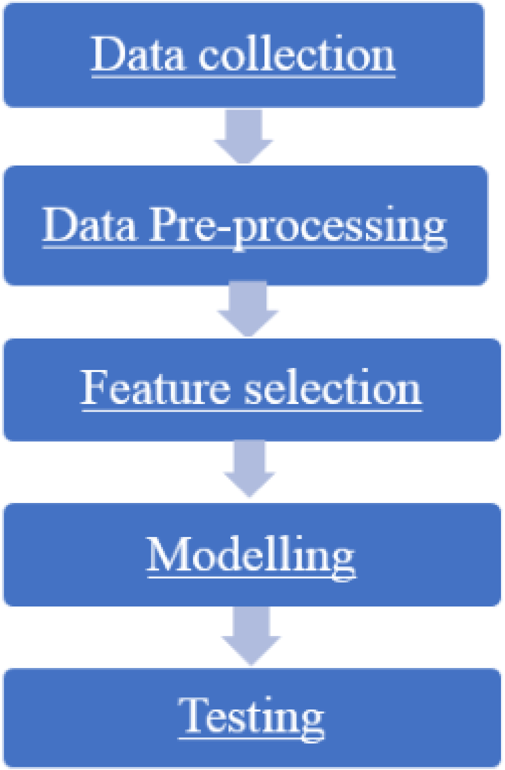
➢ Hardware Requirements

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

➢ Software Requirements

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

Proposed System Design



Data Collection and Description

- Online Shoppers Purchasing Intention dataset from UCI Machine learning repository.
- Dataset contains 12330 sessions and18 attributes. Revenue is our target variable.
- Attribute values of this dataset are integer, real.
- This dataset is donated on 2018-08-31.

Data Pre-processing

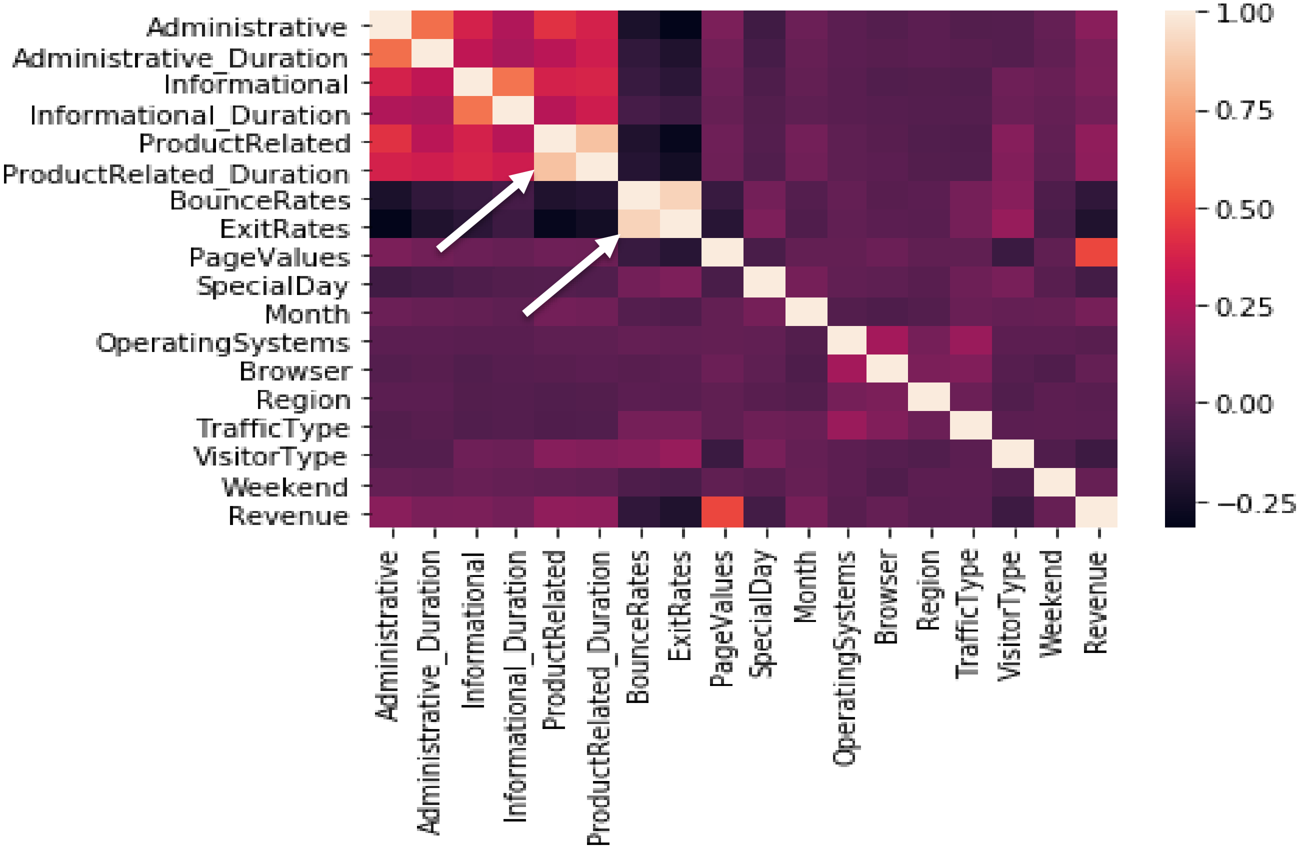
- Data Cleaning - missing values removed

- Data Transformation – Label Encoding

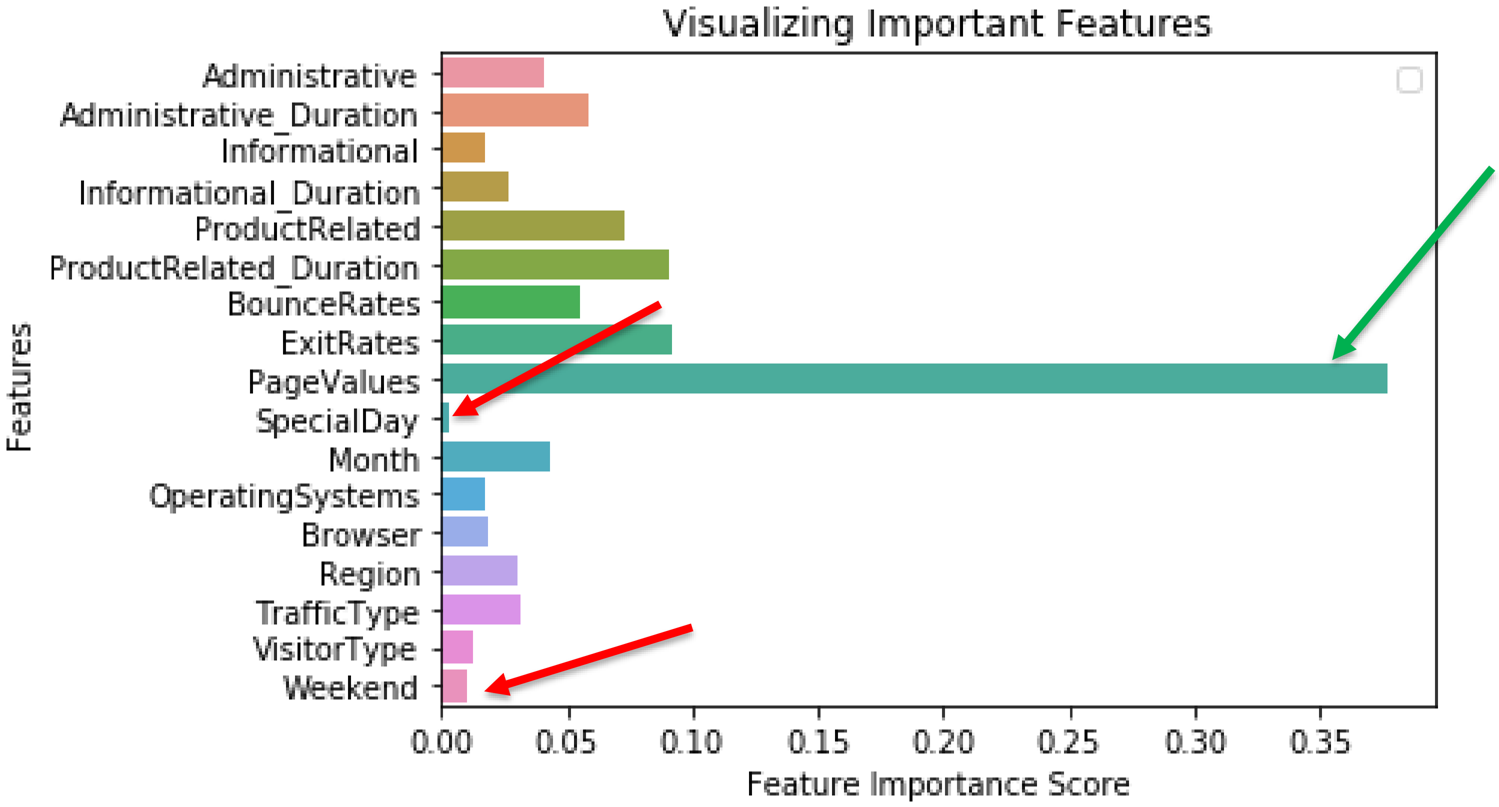
- It is a process of converting data from one format to another format.
- Alphabetical order

Correlation Matrix

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



Feature Selection



Modelling

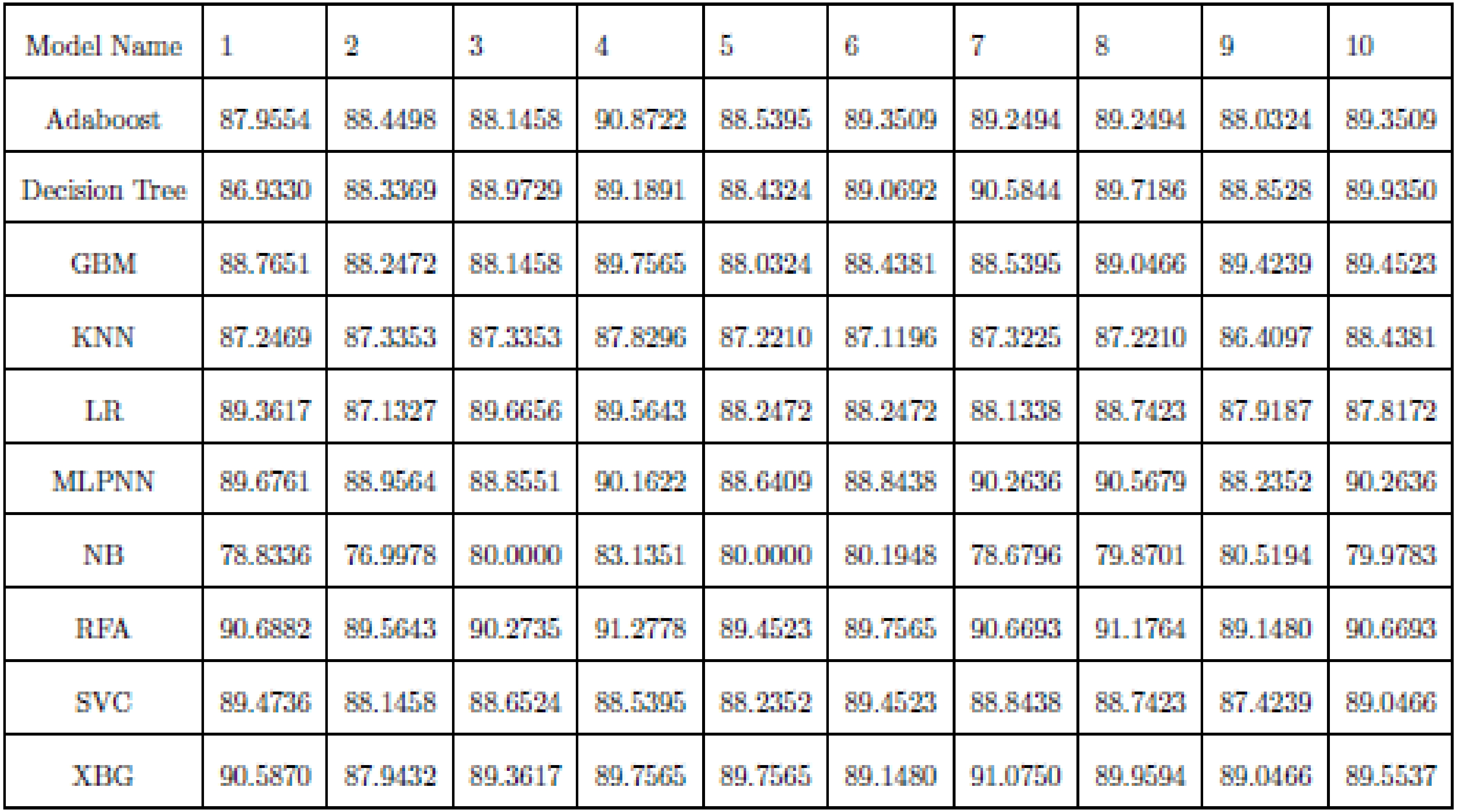
- Adaboost classifier
- Decision Tree classifier
- GBM(Gradient Boosting Machine)
- KNN (K-Nearest Neighbor)
- Logistic Regression
- Multi-Layered Perceptron Neural Network (MLPNN)
- Nave Bayes (NB)
- Random Forest Algorithm (RFA)
- Support Vector Classification algorithm (SVC)
- XGB (XGBoost)

Training and Testing

- 80% training 20% testing
- 70% training 30% testing
- 60% training 40% testing

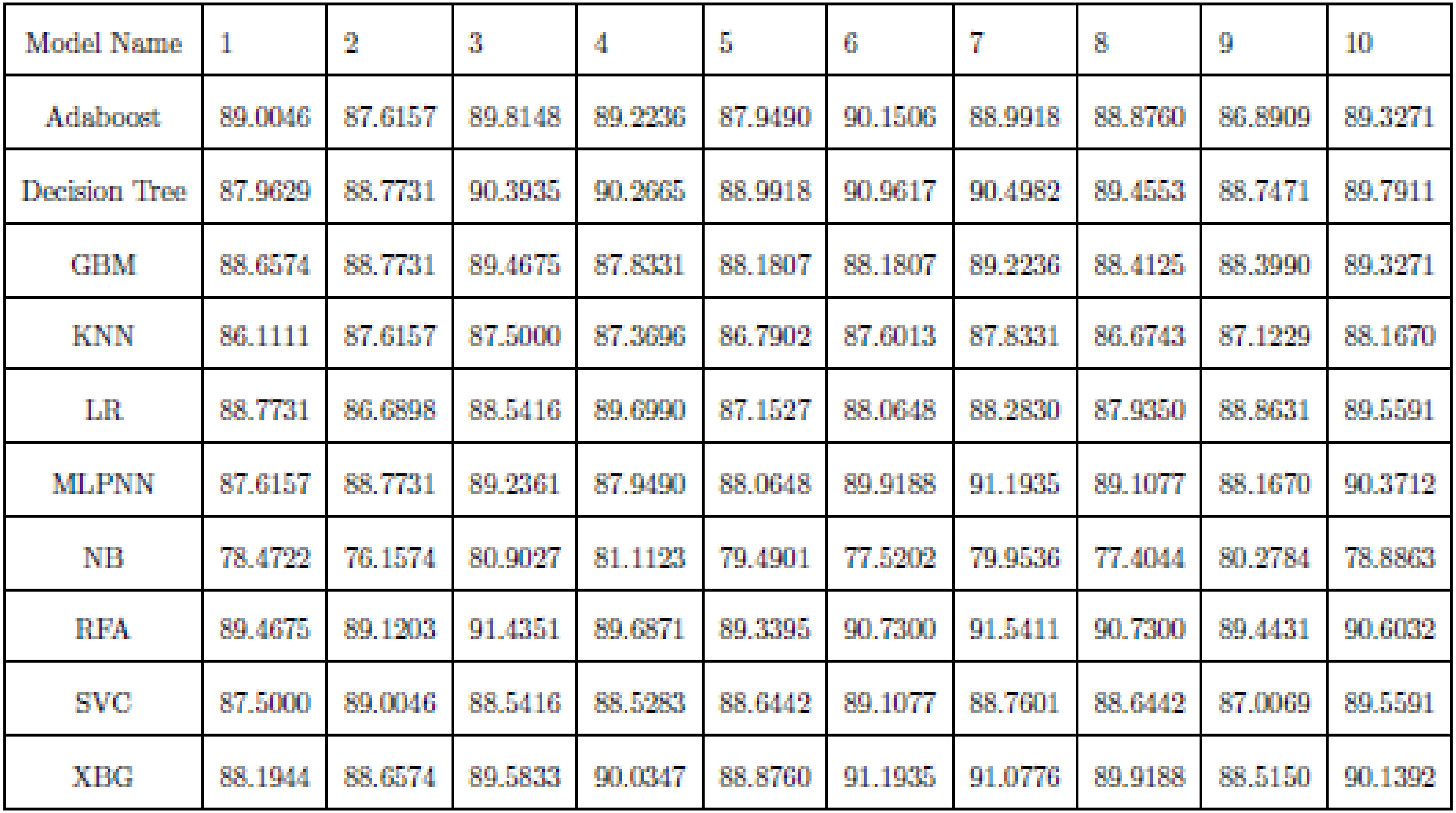
Experimentation and Results

- Accuracy of models during 10 folds with 80-20 split of data
- Accuracy of models during 10 folds with 70-30 split of data
- Accuracy of models during 10 folds with 60-40 split of data
- Minimum, Maximum and Mean Accuracy of models
- Confusion Matrix for RFA
- Classification Report for RFA



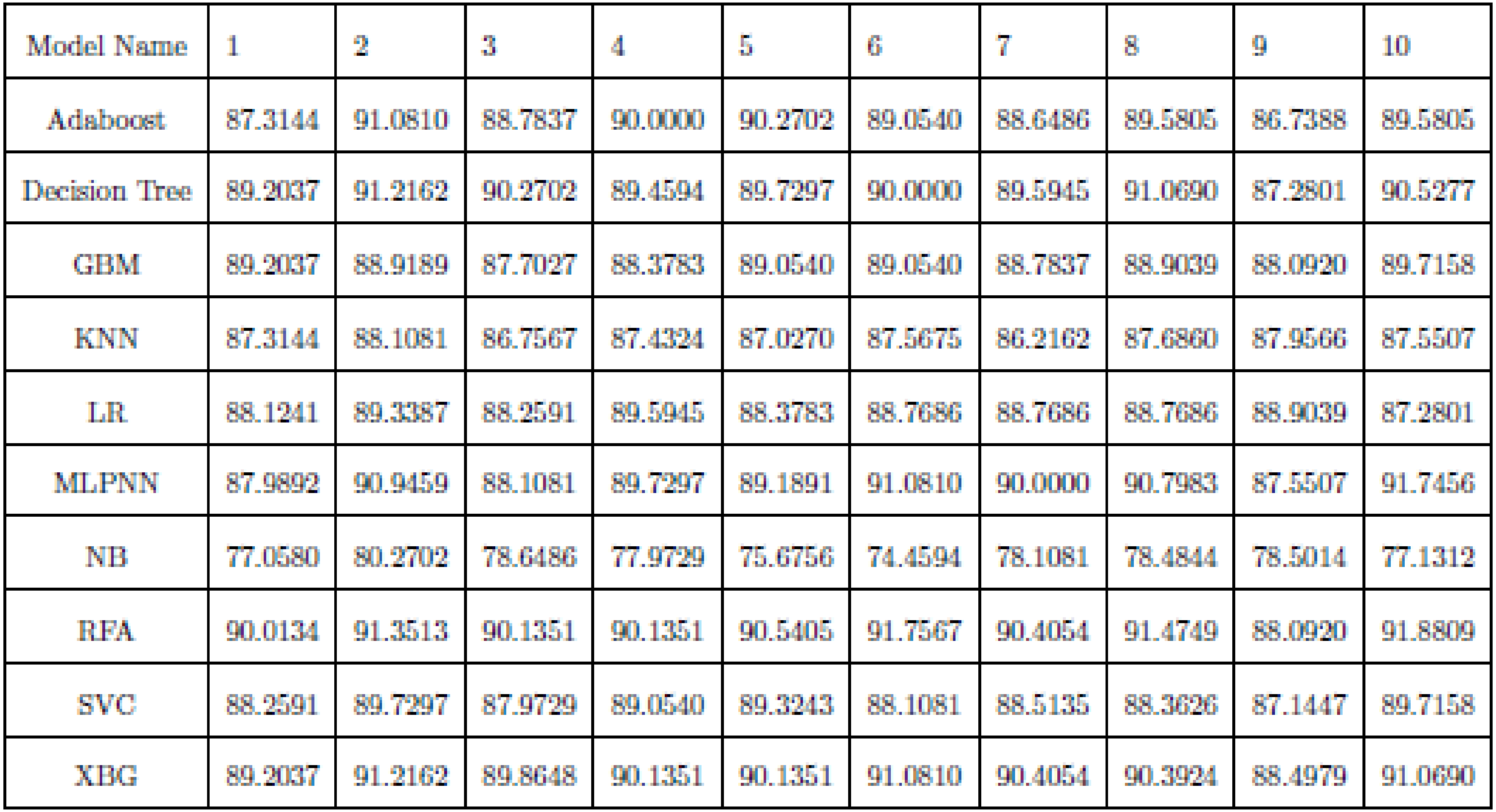
Model Name	1	2	3	4	5	6	7	8	9	10
Adaboost	87.9554	88.4498	88.1458	90.8722	88.5395	89.3509	89.2494	89.2494	88.0324	89.3509
Decision Tree	86.9330	88.3369	88.9729	89.1891	88.4324	89.0692	90.5844	89.7186	88.8528	89.9350
GBM	88.7651	88.2472	88.1458	89.7565	88.0324	88.4381	88.5395	89.0466	89.4239	89.4523
KNN	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



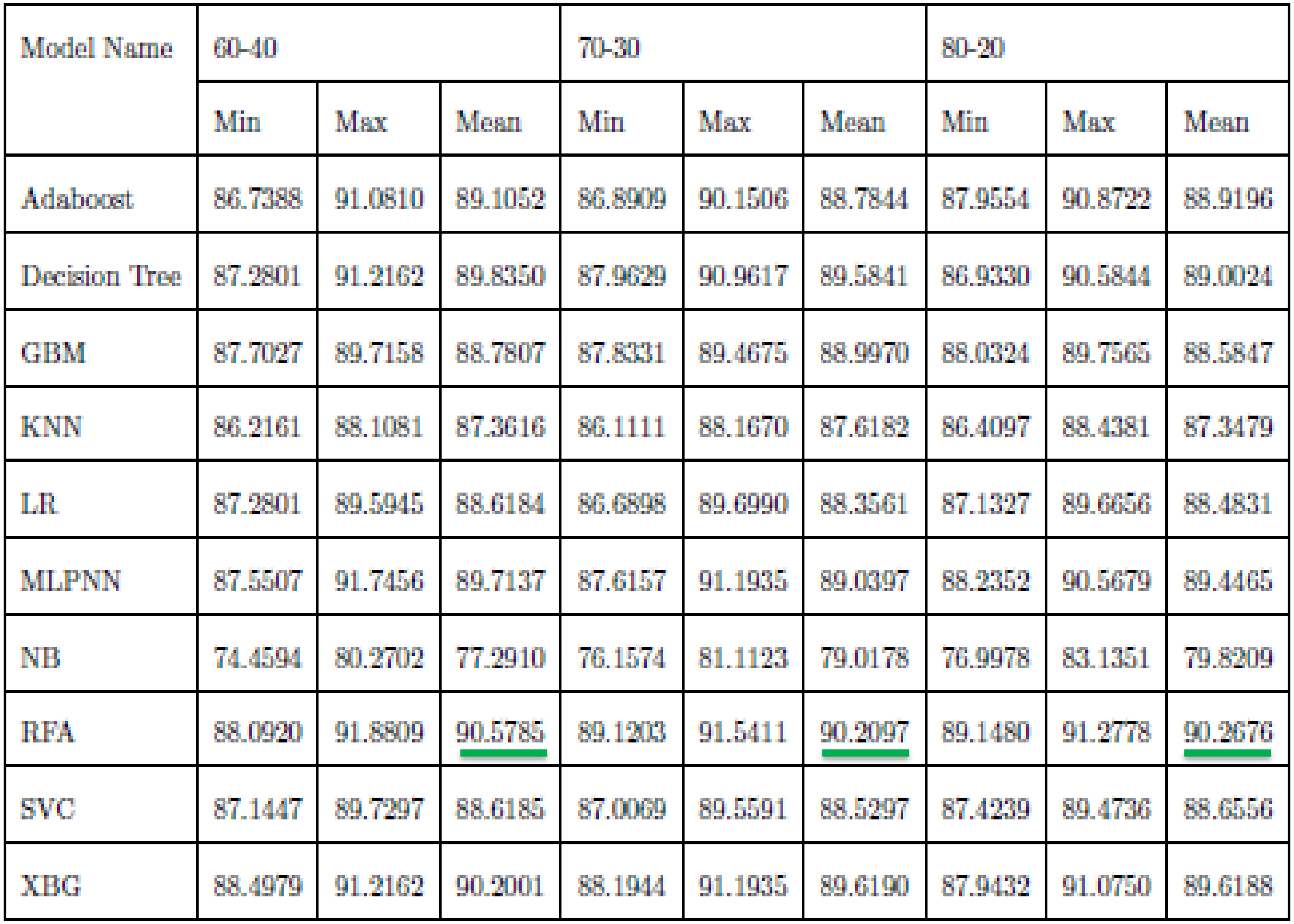
Model Name	1	2	3	4	5	6	7	8	9	10
Adaboost	89.0046	87.6157	89.8148	89.2236	87.9490	90.1506	88.9918	88.8760	86.8909	89.3271
Decision Tree	87.9629	88.7731	90.3935	90.2665	88.9918	90.9617	90.4982	89.4553	88.7471	89.7911
GBM	88.6574	88.7731	89.4675	87.8331	88.1807	88.1807	89.2236	88.4125	88.3990	89.3271
KNN	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

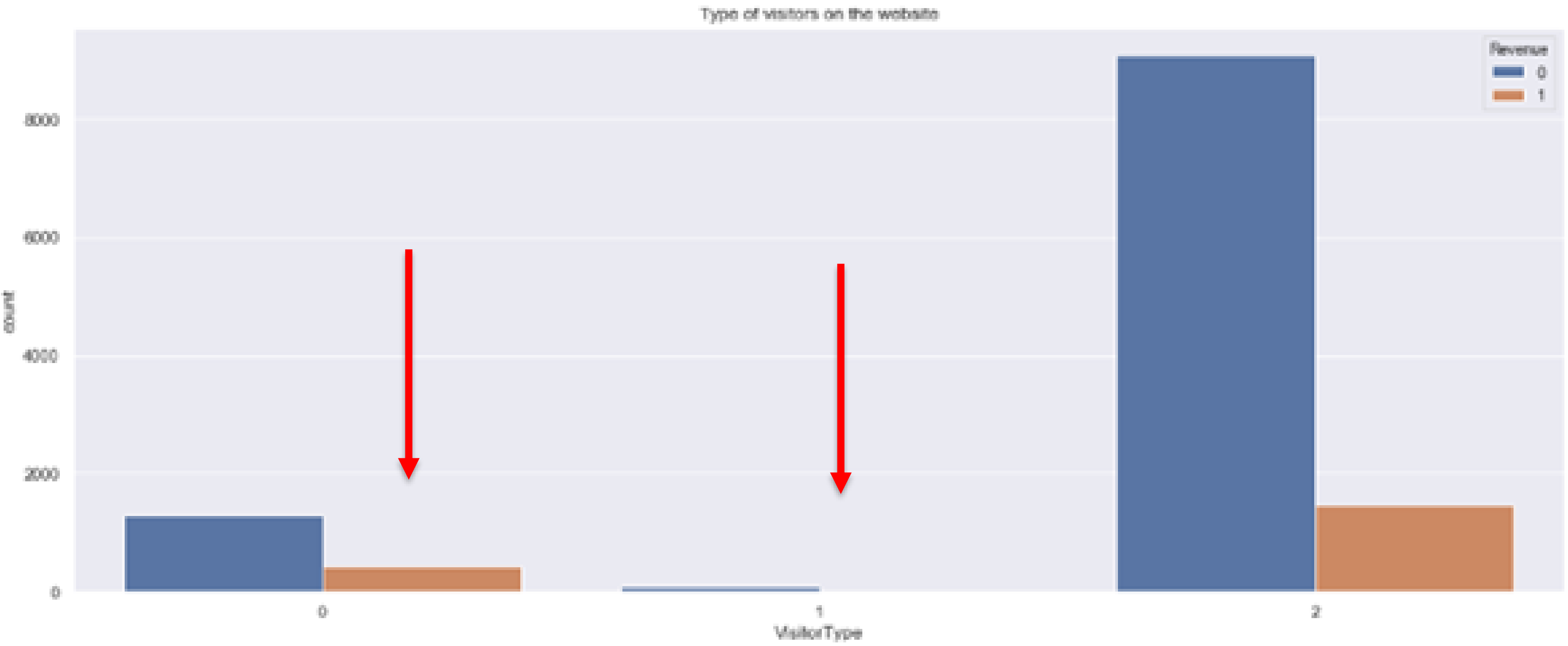
Table 5.4: Minimum, Maximum and Mean Accuracy of models

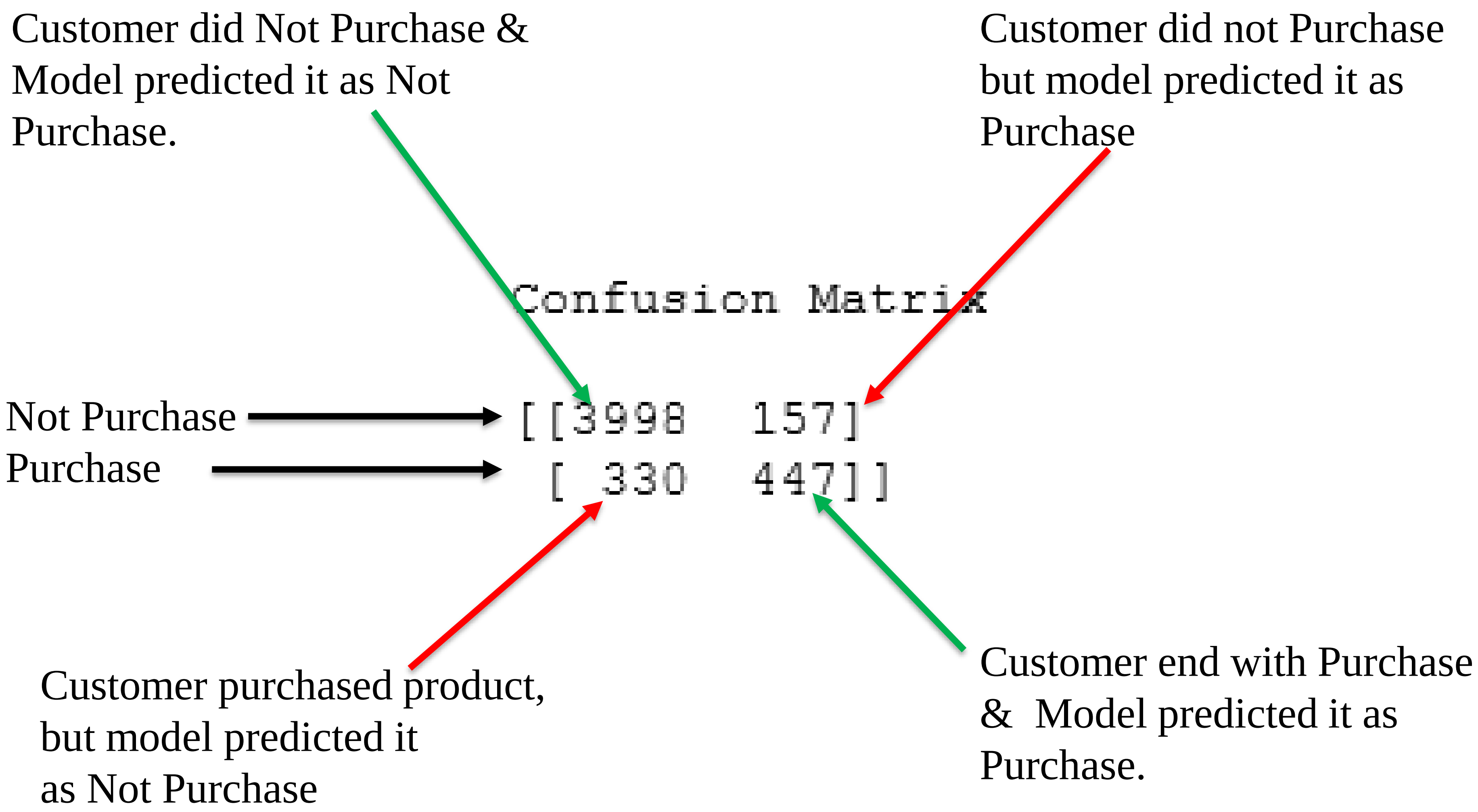
Visitors Type vs Revenue

New

Other

Returning





Customer did Not Purchase & Model predicted it as Not Purchase.

Customer did not Purchase but model predicted it as Purchase

```
Confusion Matrix
[[3998  157]
 [ 330  447]]
```

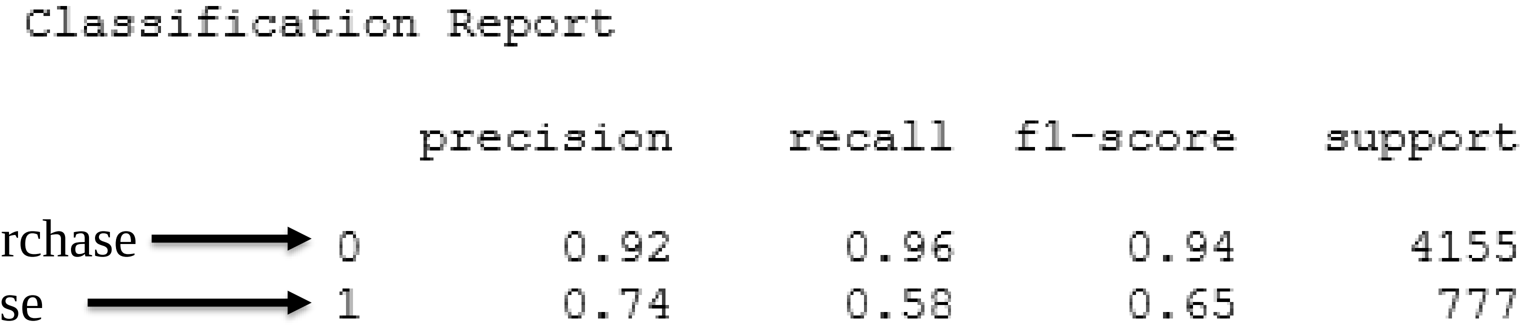
Not Purchase

Purchase

Customer purchased product, but model predicted it as Not Purchase

Customer end with Purchase & Model predicted it as Purchase.

Accuracy of Proposed Model : 90.42986861119473



```
Classification Report

              precision    recall  f1-score   support

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

Not Purchase

Purchase

❑ Precision

- It is a ability of classifier to label Positive sample as positive and negative as negative.

Precision= TP/(TP+FP)

❑ Recall

- It is a ability of finding all positive samples of the class.

Recall = TP/(TP+FP)

❑ F1-Score

- Mean of Precision and Recall

❑ Support

- It is the number of samples of true responses present in that class. (i.e. sum of both Positive and Negative samples of 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.

Conclusion

- 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.

Future Scope

- Suggesting promotional tools for improving the sales profit.

- Predicting which products the customer buy most and providing marketing strategies for improving the sales.

References

1. Dr.K. Maheswari and P.Packia Amutha Priya, “Predicting Customer Behavior in Online Shopping Using SVM Classier", 2017 IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL, OPTIMIZATION AND SIGNAL PROCESSING.
2. 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.
3. Femina Bahari T and Sudheep Elayidom M., “An Eficient CRM-Data Mining Framework for the Prediction of Customer Behaviour", International Conference on Information and Communication Technologies (ICICT 2014).
4. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
5. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

References

6. Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav, “Mining the customer behaviour using web usage mining in e-commerce", ICCCNT'12 26th 2Sdl July 2012, Coimbatore, India.
7. Pooja Sharma, Vidyalakshmi Nair, Amalendu Jyotishi, “Patterns of Online Grocery Shopping in India: An Empirical Study", CONIAAC '14, October 10 - 11 2014, Amritapuri, India.
8. 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.24
9. 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.
10. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

References

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.html
13. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
14. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
15. 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.

References

16. 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.
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).

Thank you