International Journal of Latest Trends in Engineering and Technology
Vol.(19)Issue(2), pp.020-027
DOI: http://dx.doi.org/10.21172/1.192.03
e-ISSN:2278-621X

CLASSIFICATION OF SHOPPER'S INTENTION TO PURCHASE AND MAKING REVENUE PREDICTION

Sakshi Katara[1], Chandresh Kumar Karn[1], Manvi Agrawal[1], Devendra Jamaliya[1], Ishika Mittal[1], Dr. Ankush Verma[1]

Abstract- In recent time, the online shopping has gained huge popularity among the customers, it is very important to understand the intent of the customer. Understanding the factors affecting the intent of the customer plays the major role in the success of the online business. By using machine learning models, the behavior pattern of the customers can be tracked and based of their activity the result can be predicted that a customer will purchase a product or not. Where prior research has attempted at most a limited adaptation of the information system success model, we propose a comprehensive, empirical model that separates the 'use' construct into 'intention to use' and 'actual use'.

In this paper our results give you the whole information about the consumer's intention to use is important, and accurately predicts the usage and behavior of consumers. We describe the real-time online shopper behavior prediction system which predicts the users shopping intention as soon as the customer visit the website. To accomplish the task, we depend on session and visitor information and we use naive Bayes classifier, Logistic Regression, Neural Network, GBC, decision tree and random forest investigate the dataset. In addition, we use oversampling to improve the performance and the scalability of the classifier. The results show that random forest produces the higher accuracy and F1 Score than some other Algorithms used in this below project. Some triggering factors have been working behind this phenomenal surge of online shopper such as convenience, varieties of products, friendly return policy, customers review etc. Understanding the behavior and intention of online customers has become immensely important for marketing, improving customer's experience which, in return, increases sales..

Keywords:- Logistic Regression, GBC, Random Forest, naive Bayes, Neural Network

I. INTRODUCTION

The tremendous growth of the internet has increased the trend of online shopping and digital transactions, commonly known as Business to Customer (B2C). More than 310 million active customers have bought nearly 136 billionUSD goods in 2016 from Amazon which is the leading e-commerce company in the world. 60% of internet users in Asia region has purchased online by the year 2018. The number of active users of Alibaba.com, one of the largest online shopping platforms, was 82.67 thousand in 2017 that was increased by 43.3% from the previous year [8]. The analysis of the online shopper's purchase intention from the dataset of shopper's information has become demanding field of research in the area of computation and data mining. It has become difficult and complicated to predict the online shopper's purchase intention as no interaction happens between the buyer and the seller. It is not that complicated to predicted based on the analysis of the shopper's past data and purchase items.

Data mining techniques have been exploited to help organizations for knowledge discovery and decision making by interpreting the past data. Online shopper's behaviors analysis is also called click-stream analysis as they

[1]*Institute of Advance Computing, SAGE University, Indore, MP, India*

request a series of web pages within an ongoing session. So, analyzing these click-stream data is very important towards a better and successful online business as it can mine the behavior of the online shoppers via their web page requests. This analysis has many challenges such as user identification, session identification, path identification, transaction identification. The main task to overcome is predicting the intention of the customer in real-time is also a great challenge.

In this paper, we have experimented different data mining techniques, including different supervised classification algorithms. Moreover, we have used different ensemble methods and compared among them to find the model that gives an accurate prediction. Nowadays, a large majority of businesses are supported or carried out online. In order to generate the virtual environments, marketing offers one of the most valuable strategies which can be employed. Traditionally, these offers were indiscriminately suggested to the visitors of a given e-commerce website. Being aware about the necessity of their marketing actions to the right target, online stores opted for a real time analysis for the visitors' information. The purpose is to contact the most relevant users in order to suggest offers which are likely to induce them to go back to the website and achieve an effective purchase.

Recently, a new trend has emerged among virtual shopping environments so that potential visitors are identified at the time they are browsing the website. Comparing to the real time model, the advantage behind that is to avoid the high risk of losing users once disconnected from the online store.

II. PROPOSED ALGORITHM

Solution Development Methodology: We have designed our study to analyze the performance of different classification algorithms as well as various ensemble methods. Firstly, we need to pre-process our dataset so that it is ready to be used. Preprocessing includes the replacing of any categorical text with numbers. Then, we will apply different algorithms and ensemble methods. We will calculate the values of evaluation metrics for further comparison.

Classification Algorithms:

A. Decision Tree:

The decision tree performs classification it represents the classified data by creating a tree-like structure. It breaks the data set into smaller in each step and it creates the rules for doing breakdown to form trees. Follows a top-down approach as the top side presents all the observations at a single place which splits into two or more branches that further. This approach is also called a greedy approach as it only considers the current node between the worked on without focusing on the other nodes [2].

B. Random Forest:

The main concept of random forest is to create many correlated decision trees where all the decision trees act as an ensemble model [1]. The output of the Random forest algorithm is more accurate than the decision tree algorithm because each of the trees evaluates the error of other trees [6]. All the trees act as a committee to make the decision. It creates a set of decision trees from randomly selected subset of training set. It then collects the votes from different decision trees to decide the final class of the test object.

C. SVM:

The approximation of the support vector machine Algorithm is to find a line separates the data in two different classes. Data are projected in a high dimensional space also the data are mapped by the kernel. The kernel can be linear or nonlinear [3]. In our case, we have used nonlinear kernel. Then the algorithm creates hyper plane to separate data from one class to another. The hyperplane will be generated with an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane [4].

D. Naive Bayes:

The naive Bayes algorithm is a simple and straightforward algorithm based on Bayes theorem for conditional probability. It classifies data depending on the frequency of occurrence of the data descriptors of the training set. Assumption of Naive Bayes algorithm is all the data are equally independent. This classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which it predicts on the basis of the probability of an object.

E. Boosting:
Boosting is an amazing method of converting a weak learner to a strong learner. In boosting first a tree is generated where each of the data points is given an equal weight. Then after evaluating the first tree the weight of those data points those are difficult to classify their weight are increased and those are easy to classify is decreased. The second tree is made using the weight of the first tree. Boosting Machine Learning is one such technique that can be used to solve complex, real-world problems [5].

F. Neural Network:
A neural network is the algorithms that used to recognize the relationships in a set of data through a process try to mimic the way the human brain operates. The neural networks refer to the systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input, so the neural network be able to generates the best possible result without the need of redesigning the output. A "neuron" in an artificial neural network is a mathematical function that collects the information according to a specific architecture.

G. AdaBoost Classifier:
An AdaBoost classifier is a meta-estimator that use for fitting a classifier on the original dataset and then fits additional copies of the classifier in the same dataset. Among all it is best algorithm used to boost the performance of decision trees on binary classification problems. It is an iterative ensemble method. AdaBoost classifier builds the strong classifier by combining many poorly performing classifiers so that the accuracy of the output will be high as strong classifier. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration so that accurate predictions of unusual observations done.

H. Histogram Based Gradient Boost:
It is one of the most popular machine learning algorithms is histogram based gradient boosting. This estimator has the feature that use for missing values. If there will be no missing values in the given features, then samples with missing values are mapped to the child that has most samples. It is a graphical representation that organizes a group of data into user-specified ranges. The histogram contains a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

I. Gradient Boosting Classifier:
Gradient boosting classifiers (GBC) is group of machine learning algorithms that combine weak learning models together to create a strong machine learning model. Decision trees are used while doing gradient boosting. The idea behind introducing the "gradient boosting" is to take a weak hypothesis or weak learning algorithm and make a series of tweaks so that it will improve the strength of the hypothesis/learner. This type of Hypothesis is Boosting technique based on the idea of Probability Approximately Correct Learning (PAC).

J. *Logistic Regression*: Linear regression used to predict the data by finding a linear straight-line equation in model or predict data points. Logistic regression (LR) does take care of the relationship between the two variables as a straight line. Instead of that logistic regression uses the natural logarithm function to find the relationship between the variables and by using testing data to find the coefficients. The function can use to predict the future results using these coefficients in the logistic regression equation. It is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis.

III. QUALITY ASSURANCE

System Quality: The customer websites that are easy to use as well as easy to navigate. This quality of usability should still be an additional variable. The usability of a website refers to the easy accessibility of the website when consumer uses to achieve goal. Reliability refers to the dependency on the website's operations. Adaptability refers to the systems that use to adjust their content with respect to the changing demands of the customer. The response time refers how quickly the system responds to requests for information or action.

Potential Risk: Although there are not only one of the risk sub-dimensions identified in traditional channels, privacy risk has received growing attention as market research shows that 75 percent online shoppers report concerns regarding the security of their user's credit card information, they have to provide the details to complete online transactions. The study examines the impact of product risk, financial risk, and privacy risk on shoppers' online

purchase intentions for two product categories: digital products and non- digital products.

Product Risk/Performance Risk: It is defined as the probability of the item failing to meet the performance requirements originally focused. Product risk has been reported as the most frequently faced challenge and this is the main reason for not shopping online.

Financial Risk: There are many reasons why online shoppers suffer some loss while shopping online. It is hard for online to predict the price of the item purchased at an online is lowest or highest available compared to others.

Customers had to rely on different information sources to confirm product quality and enhance the chances of satisfaction while purchasing different kind of products. They differ in their preferences for online and traditional outlets based on the importance with different product attributes. It is possible that some types of risks arrive, like privacy risk, it may increase as online shopping experience, whereas possibilities of other types of risk, and financial risks decrease with online shopping experience.

Privacy Risk: It is defined as the probability of having personal information disclosed as a result of online transactions. Despite the growing online sales volume, concerns regarding privacy remain high among many online shoppers.

- Online shopping experience is negatively associated with perceived product risk.
- Online shopping experience is negatively associated with perceived financial risk.
- Online shopping experience is positively associated with perceived privacy risk.

In the nutshell the online shopping experience is positively associated with online shopping intentions. The risk perceptions increase with the online intentions.

Table1- Numerical Features Used in the user behavior analysis model

Feature name	Feature description	Min. value	Max. value	SD
Administrative	Number of pages visited by the visitor about account management	0	27	3.32
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	-1	3398.75	176.85
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	0	24	1.27
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages	-1	2549.38	140.81
Product related	Number of pages visited by visitor about product related pages	0	705	44.48
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages	-1	63973.5	1914.29
Bounce rate	Average bounce rate value of the pages visited by the visitor	0	0.2	0.04
Exit rate	Average exit rate value of the pages visited by the visitor	0	0.2	0.04
Page value	Average page value of the pages visited by the visitor	0	361.764	18.56
Special day	Closeness of the site visiting time to a special day	0	1.0	0.19

IV. IMPLEMENTATION

A: Dataset Description

In our test, we have utilized session information of the clients [7]. There are two classes of clients: who bought anything and who didn't. The objective worth is subsequently clear cut. There are both Numerical data and Categorical data. The Numerical and Categorical data is depicted in table 1 and table 2 respectively. The informational index comprises 12,330 columns where every one of the lines represents one user's session data. The information was gathered in one year to maintain a strategic distance from the effect of extraordinary days. events, client profile, or period. Among the 12,330 meetings,84.5% (10,442) sessions have a negative objective worth that is the customer wound up with no buy. The rest 15.5% (1908) sessions positive objective worth. In Table I the mathematical highlights have appeared with their insights. Among the features, Administrative, Administrative Duration, Informational, Informational Duration, Product Related, and Product Related Duration talks about the number of various kinds of pages visited by the guest in that session and all-out time spent in every one of these page classes. These qualities have been found from the URL of the pages visited by the clients during their meetings and the exercises on the pages.

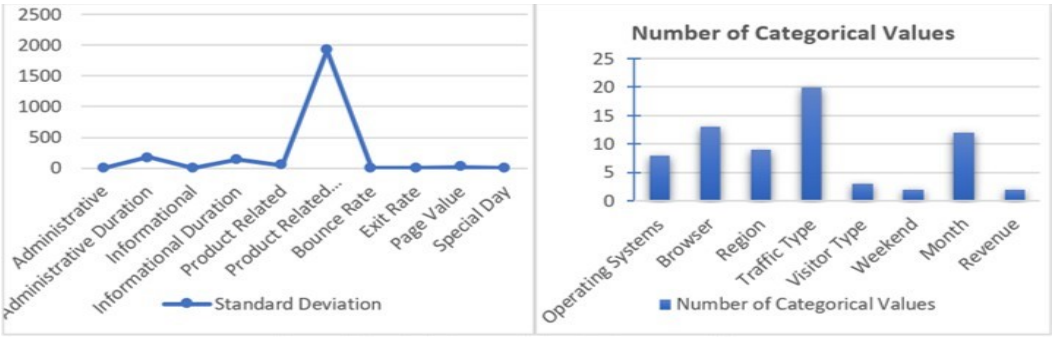

Figure 1. Graph between different Columns and Revenue

The Bounce Rate, Exit Rate, and Page Value highlights given in Table 1 speak to the measurements which have been estimated by Google Analytics. In the created framework these qualities can be put away in the application information base of the apparent multitude of pages at a standard span consequently.
The bounce rate feature shows the percentage of visitors who have placed a site on those pages and left the website out of use. The exit rate is the percentage of users who have their last session on that page. The page value indicates the average value of the page visitors visited before purchasing any product. A special day indicates the proximity of any special event where users are likely to purchase the product. Its value is determined by considering the relationship between the date of the order and the date of delivery. Its value ranges from 0 to 1. The standard deviations of the information can be pictured in from figure 1.
There are eight categories of fields called an operating system, browser, region, traffic type, visitor type, weekend, month, revenue. This explains some details of browsing history. Revenue indicates whether the user purchased the item or not. The number of categories in each category can be found in the figure.

B: Trends in dataset Analysis of categorical data:
In below graph we have plotted the graphs where x-axis denotes the features and y-axis denotes the number of visitors. By analyzing graphs we found out that number of visitors are more at the end days of the sale, in the month of May and November there are a greater number of visitors, higher number of visitors are from category 2 o.s, category 2 browser is most likely to be used, region one and traffic type 2 has maximum number of visitors, returning visitors site the more and less people visit site at weekends.

Analysis of categorical data with respect to revenue:
When features were plotted with respect to revenue, we found out that revenue is higher at end days of sale , November month and category 2 os have higher revenue, category two traffic type has more revenue, region one and browser category 2 contributes more in revenue, returning visitors are more likely to purchase and revenue is generated more on weekdays.

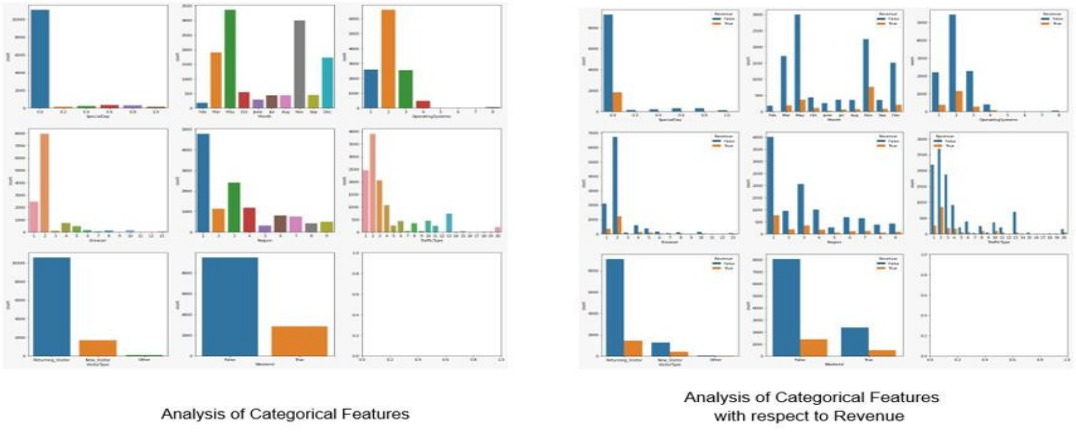

Figure 2. Analysis of categorical data with respect to revenue

Analysis of numerical data –
In each graph 0 shows that user does not visit that page. In last three graphs bounce rate, exit rate, and page values 0 stands for people neither exit the page nor move to another page.

Analysis of numerical page with respect to revenue*:*
Administrative and Administrative duration - These graphs show that if visitors do not visit admin page revenue is high. *Information and information duration*- these graphs show that the revenue is less is visitors visit the informational page frequently.
Product related and Product related duration-These graph shows that the revenue is high if user visits the product page more.
Bounce rate and exit rate- This graph shows that revenue is high if bounce rate and exit rate is high.
Page value- This graph shows that revenue is less affected by page value.

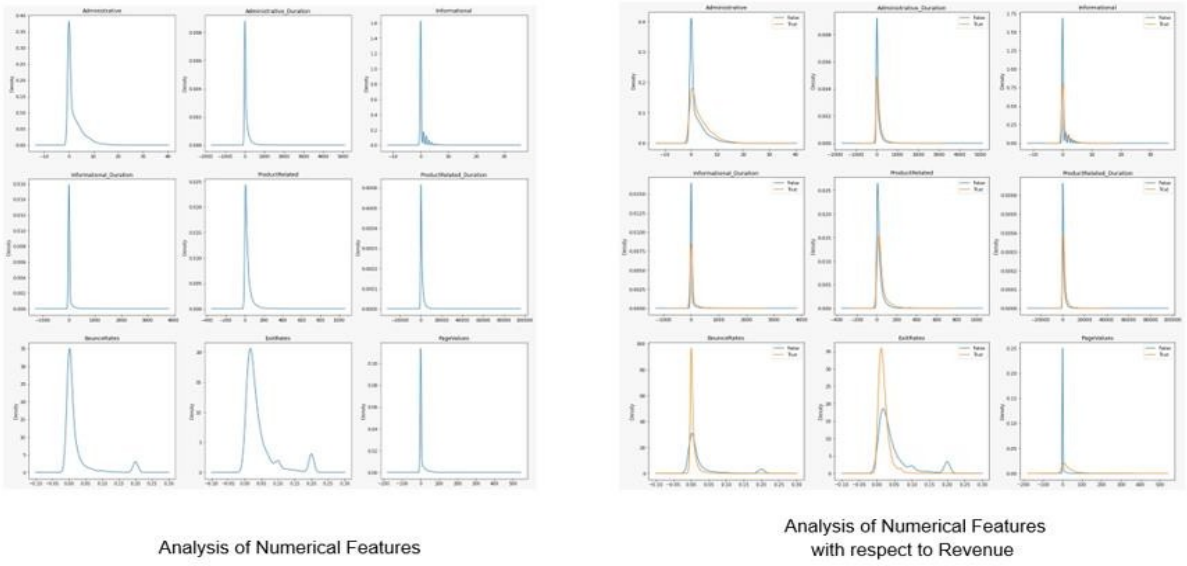

Figure 3. Analysis of numerical data with respect to revenue

V. ANALYSIS

We have performed Logistic Regression, GBC, Random Forest Classifier, Gaussian NB, Neural Network and Support Vector Classifier on our dataset to compare their performance which is why the dataset needed to be pre-processed before experimentation. We have counted the total numbers of True and False values of Revenue column as Revenue is our Target Column. We have plotted the graph of different variables to see the variations in values.

Also, plotted the graph between different columns versus target column i.e., Revenue (figure 1 shows the graph between different columns and Revenue). Figure 2 showing the Heat Map that we plot to check the correlation between different columns. We have run our experiments on a machine that runs on AMD Ryzen 7 processor with 16GB ram. For each of the experiments, we have calculated the values of different evaluation metrics for comparison.

Testing/Result Analysis
The most important thing in the development of any model is the feature selection. There are many techniques which can be used for feature selection. In this model we have used Correlation technique for feature selection. We plotted heatmap for the correlation values and then analyzed the correlation between the values. The range of correlation is (-1,1).

The positive values of the correlation indicate that if the value of one feature increases the value of the other feature also increases. The negative correlation We ran the algorithms with all the features, value indicates that if value of one feature increases then the value of other feature decreases. If the value of correlation is near to 0 this indicates that the two features have no effect on each other. We have removed the features which were highly correlated as they may act as duplicate values and affect the accuracy of the model [8].

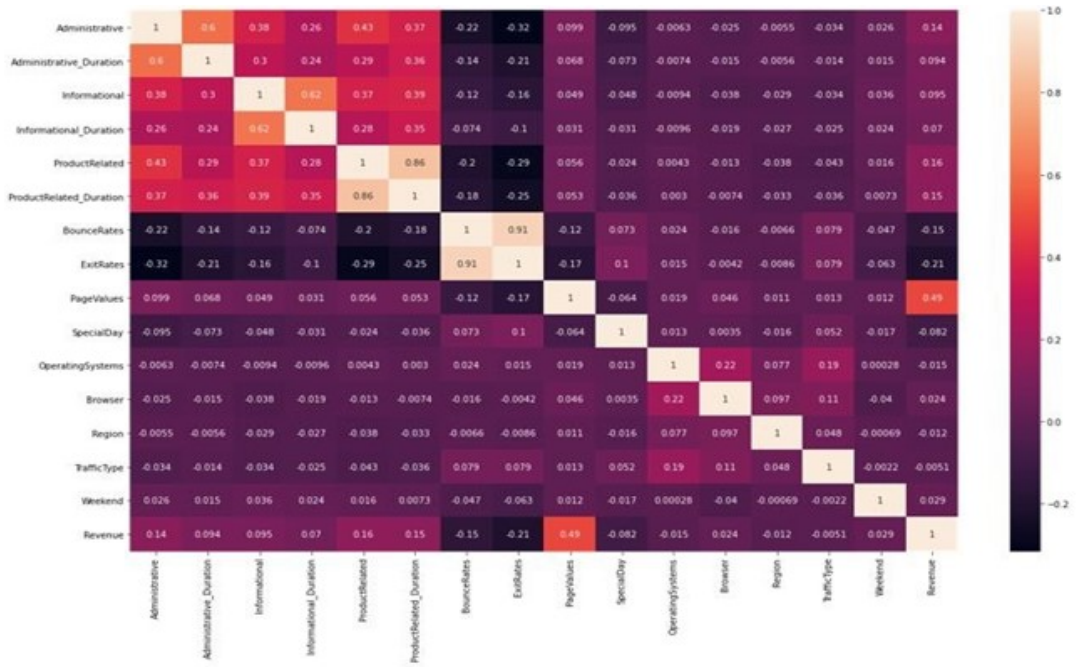

Figure 4. Correlation among features

We have performed correlation on the training set to avoid the overfitting of the model. We ran the algorithms with all the features, we observed that the amount of time taken in training the model was very large. After performing feature selection when the models were trained the amount of time taken dropped significantly. We observed that the amount of time taken in training the model was very large. After performing feature selection when the models were trained the amount of time taken dropped significantly.

We took two cases in the feature selection for analyzing the how feature selection affects the accuracy of the model. In Case-1 the features which were dropped from the dataset are Bounce Rates, Administrative Duration, Informational Duration, Product Related Duration and in Case 2 we dropped Operating Systems feature along with that of the Case 1. The results are displayed in the form of bar graph Fig.3. We observe that difference between the accuracies is not significant. Accuracy of GradientNB is affected the most by the difference in feature selection.

After feature selection we have run different algorithms on our dataset and evaluated the performance of the algorithms using matrices which are commonly used to analyze the performance of various algorithms. We have used confusion matrix to do analysis of the different algorithms. Our target or the dependent variable was Revenue whose value was either True or False. Our result included various terms to evaluate the outcome which are explained

True Positive - This means that the actual output of Revenue is True, and the model has also predicted it correctly i.e., the predicted output is also True.

False Positive - This means that the actual output Of Revenue is False, and the model has predicted it True.

True Negative - This means that the actual output of Revenue is False, and the model has also predicted it correctly i.e., the predicted output is also False.

False Negative - This means that the actual output Of Revenue is True, and the model has predicted it False.

The above concepts were used to evaluate the models. By using these concepts, we have calculated following aspects which helped in comparison of different models-

Accuracy tells us the percentage of correctly predicted outcomes i.e.; the Revenue was correctly predicted either true or false. Precision gives the ratio of the correctly predicted positive outcomes upon the total number of positive outcomes. Recall gives the ratio of the correctly predicted positive outcomes upon the actual positive outcomes.

F1-score shows the harmonic mean of the precision and the recall. We have analyzed different algorithms which include Logistic Regression, GBC, Random Forest, GaussianNB, SVC, Decision Tree, Neural

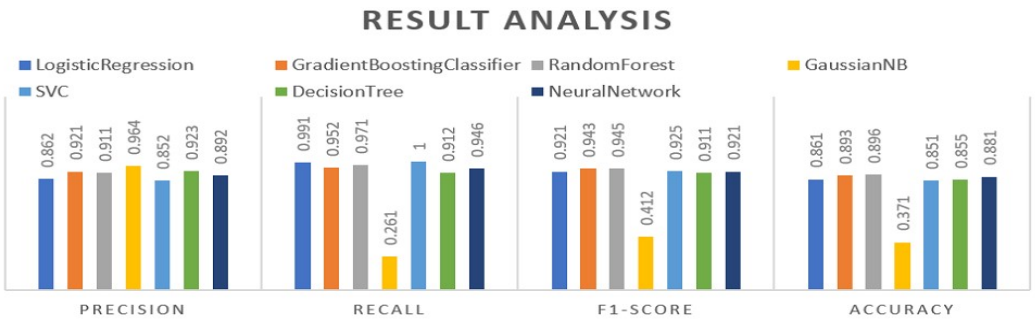

Figure 5. Result Analysis

Network. Fig.5 shows all the values which are obtained by the algorithms. In terms of accuracy Random Forest is the best algorithm as it has highest accuracy i.e., 89.6%. It shows that Random Forest is more accurate in predicting the Revenue based on customer's activity. F1-Score also indicates that Random Forest is more accurate.

Algorithm	Accuracy (%)
Random Forest	89
Logistic Regression	86
Gaussian Naive Bayes	37
SVC	85
Neural Network	88
Gradient Boosting Classifier	89

If consider precision then GaussianNB is more accurate but we consider factors like accuracy, recall and F1-score, GaussianNB is the least accurate so we cannot consider it to be the suitable algorithm for prediction of the Revenue. SVC is most accurate in terms of recall, but it fails to be most accurate in other terms. We also analyzed the dataset by using Neural Network. We trained the model using 100 epochs and as the epochs increased the accuracy also increased from 87% to 94%. But the average accuracy resulted to be 88.1% which was less as compared to Random Forest. So, we can conclude that Random Forest is the best algorithm to predict the shopper's intent from empirical data. We also used different algorithms to increase the accuracy of the model which includes Histogram based Gradient Descent and Adaboost Classifier. The accuracy of the prediction was increased to 88% and 89%respectively. By this result we concluded that Adaboost Classifier is better in predicting the intent of the customer in online purchase.

VI CONCLUSION

In this paper, we have analyzed the performance of various algorithms that came under supervised learning. Our goal was to study algorithms and find out which algorithm is more accurate in predicting our target variable Revenue. We found that Random Forest comes up with the highest accuracy of 89%. By using the result of the analysis shopper's will now be able to know where they are lacking and they can update their strategies for the promotion and sales so that they can have high revenue.

REFERENCES

[1] Breiman, Leo." Random forests." UC Berkeley TR567 (1999).
[2] Llora, Xavier, and Josep M. Garrell. " Evolution of decision trees." Forth Catalan Conference on Artificial Intelligence (CCIA2001). 2001.
[3] Plewczynski, Dariusz, Stphane AH Spieser, and Uwe Koch. "Assessing different classification methods for virtual screening." Journal of chemical information and modeling 46.3 (2006).
[4] Benediktsson, Jon Atli, and Philip H. Swain. " Consensus theoretic classification methods." IEEE transactions on Systems, Man, and Cybernetics 22.4 (1992): 688-704. Syarif, Iwan, et al. "Application of bagging, boosting and stacking to intrusion detection." International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg, 2012.
[5] Clifton, Brian. Advanced web metrics with Google Analytics. John Wiley Sons, 2012.
[6] Rayhan Kabir, Rasif Ajwad, Raisal BinAshraf, BARC university, "Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data" conference paper, December 2019.
[7] Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.
[8] Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. "Data mining techniques for customer relationship management." Technology in society 24.4 (2002).