

ANALYSIS FOR CUSTOMER REVENUE GENERATION

MOHMED HASIM KAGDI

A project report submitted in conformity with the requirements
for the degree of Master's of Science in Information Technology

Department of Mathematical and Physical Sciences
Faculty of Graduate Studies
Concordia University of Edmonton



© Copyright 2022 by Mohmed Hasim Kagdi

ANALYSIS FOR CUSTOMER REVENUE GENERATION

MOHMED HASIM KAGDI

Approved:

Rossitza Marinova, Ph.D.

Supervisor

Date

Committee Member Name, Ph.D.

Committee Member

Date

Alison Yacyshyn,, Ph.D.

Dean of Graduate Studies

Date

Abstract

Most businesses around the world have started using e-commerce. When using the e-commerce platform, it becomes easier to gather data about the users of the platform. Businesses must utilize this data collected to the best of their ability. This would result in the growth and survival of the businesses. In this project, we are going to discuss some reasons why businesses would want to invest in predicting the future using machine learning, learn the process of data mining, and identify and discuss a few different type of classifiers which can be used in prediction. To demonstrate how machine learning and different classifiers work, a dataset is acquired from the internet and analysis was performed on it. Using the analysis, a conclusion on the state of the website was prepared. The improvements the business could make to generate more revenue were identified and stated. Classifiers were used on the dataset after it was split 70-30 (train and test respectively) to check the performance of the classifiers.

Keywords: Classifiers, Analysis, K-Nearest Neighbour, Support Vector Machine, Gaussian Naive Bayes, Logistic Regression, Random Forest, Gradient Boost, Ada Boosting, Revenue

Contents

1	Introduction	1
2	Literature review	2
3	Machine Learning Methods	4
3.1	Naive Bayes	4
3.2	K-Nearest Neighbour	5
3.3	Support Vector Machine	6
3.4	Logistic Regression	6
3.5	Random Forest	7
3.6	Gradient Boost	8
3.7	AdaBoost / Adaptive Boost	9
4	Project Design and Implementation	10
4.1	Dataset	10
4.2	Dataset Feature Explanation	10
4.2.1	Numerical Features	11
4.2.2	Categorical features	13
4.3	Data Description	15
4.4	Implementation Method	15
4.5	Data Cleaning and Preprocessing	16
5	Results and Discussion	18
5.1	Data Analysis	18
5.2	Building and Testing the Classifiers	22
5.3	Different Classifiers Results	22
6	Conclusions	29
7	Future Work	29

List of Tables

1 Classifier Prediction Performance 27

List of Figures

1	The Process of Data Mining [2]	3
2	Hyperplane Representation of SVM Algorithm [9]	7
3	Logistic Regression Curve [10]	8
4	Visualization of Random Forest [11]	9
5	Visualization of Gradient Boost [12]	9
6	Dataset Information	14
7	Importing Dataset from the Google Drive	15
8	Summary of Data Used	15
9	Summary of Data Used Cont.	15
10	Checking for null Values in the Dataset	16
11	Correlation Heatmap between Features	17
12	Distribution of customer on Revenue	18
13	Revenue Per Month	19
14	Users Per Month	19
15	Informational Duration vs Revenue	20
16	Administrative Duration vs Revenue	20
17	Product-Related Duration vs Revenue	20
18	Different Visitor Types	21
19	Count Per User Type	21
20	Confusion Matrix Example [14]	22
21	Confusion Matrix for Gaussian Naive Bayes	23
22	Confusion Matrix for K-Nearest Neighbor	23
23	Confusion Matrix for Support Vector Machine	24
24	Confusion Matrix for Logistic Regression	25
25	Confusion Matrix for Random Forest	25
26	Confusion Matrix for Gradient Boosting	26
27	Confusion Matrix for Ada Boosting	27
28	ROC Curve	28

1 Introduction

E-commerce is expanding more quickly than ever before. With the recent worldwide pandemic specifically [16], businesses are shifting to a more online-focused mode of operation because e-commerce is increasingly where buyers go to buy their preferred goods. Most companies throughout the world are investing in this e-commerce platform to advertise, offer customers goods and services, and compete with other companies. To compete with other companies and stay in business, they are working hard to develop a strategic plan that will allow them to fully utilize the technology in our environment. And thanks to recent developments in machine learning and deep learning technology, organizations can now forecast a wide range of events that previously seemed unpredictable.

The way people consume goods and services is always evolving along with the rest of the world. One of the largest shifts in recent years has been the growth of data, which has given businesses unprecedented insights into their customers [17]. For instance, data can be used by a business to determine which of its present clients are most likely to purchase a specific good or service, allowing it to better cater to its most devoted clients. The same information has also been used to create predictions about the kind of consumers who are likely to exist in the future, allowing businesses to better serve their current clients.

Over the years, there have been many changes in the retail industry. Some of the most significant changes have been the introduction of online shopping and the shifting of mall space to other uses. Both changes have increased the competition between retail companies, which has forced them to become much more competitive when it comes to attracting and servicing their customers. This has resulted in the need to adopt a customer revenue prediction model that can be used to identify which of these customers are most likely to generate revenue for a company.

The project is taking the task to predict customer revenue. Predicting customer revenue is an essential aspect of running any organization. Companies may choose which investments to make, where to devote resources, and how to manage their company by knowing which customers are likely to buy products and services and which ones aren't. Data mining is a method that businesses may use to examine huge datasets and extract data that will help them make better business decisions [2]. This work shows how data mining can be utilized to anticipate a customer's future revenue in this paper.

An effective model for predicting revenue can be produced with the aid of machine learning. We need to collect enough historical data for this purpose to be able to choose the best model to use. There are seven different classifiers used in the project to help build a model best suited for this task. They are:

1. Gaussian Naive Bayes
2. K-Nearest Neighbour

3. Support Vector Machine
4. Logistic Regression
5. Random Forest
6. Gradient Boost
7. Ada Boost/ Adaptive Boost

The seven classifiers were used to determine their prediction performance. The confusion matrix, accuracy, precision, recall and F1 score were collected as prediction performance of the classifiers. These classifiers are compared using the receiver operating characteristic (ROC) curve to determine the best classifier. The accuracy, F1 score, precision, recall, confusion matrix and receiver operating characteristic (ROC) curve are further discussed with explanation in 'Results and Discussion' section of the project.

2 Literature review

Businesses are the backbones of the economy. According to the article [1], “without good forecast, businesses are horribly exposed”. Today, any business that is unable to predict, foresee and adapt; faces the risk of missing out on opportunities or, in the worst case, failing. For example, even Cisco, known for its highly regarded real-time reporting tools, paid a hefty \$2 billion price for failing to integrate operational and financial forecasting into a reliable risk management framework [1].

The ability to identify patterns in databases has been made available for businesses and organizations because of advancements in machine learning technology and artificial intelligence [2]. The business infrastructure has undergone several adjustments as a result. Since a big percentage of commerce these days occurs online, many businesses now actively spend gathering as much data as possible on anything. Innovative techniques for collecting data through e-commerce have emerged as a result. The data collected must be analyzed to discover the meaningful pattern for the improvement and the survival of the business.

Article [2] gives us a better understanding of the complexity of the process of data mining as shown in Figure 1. For the data mining process, a data set must be accessible. The data set must undergo the cleaning and preprocessing step. Improper cleaning or preprocessing of the data might lead to inconsistencies and discrepancies [2], which would be problematic. The quality of the data is also improved by this process. To create models that represent the data patterns, the data is analyzed for patterns. The predicted accuracy of the model is examined using a set of data. The model needs to be adaptable enough to be turned into a viable business plan that will probably help the organization accomplish its goal. A model that satisfies this criterion is regarded as having business knowledge [2]. The phases in the mining process are repeated up until useful business knowledge is extracted [2].

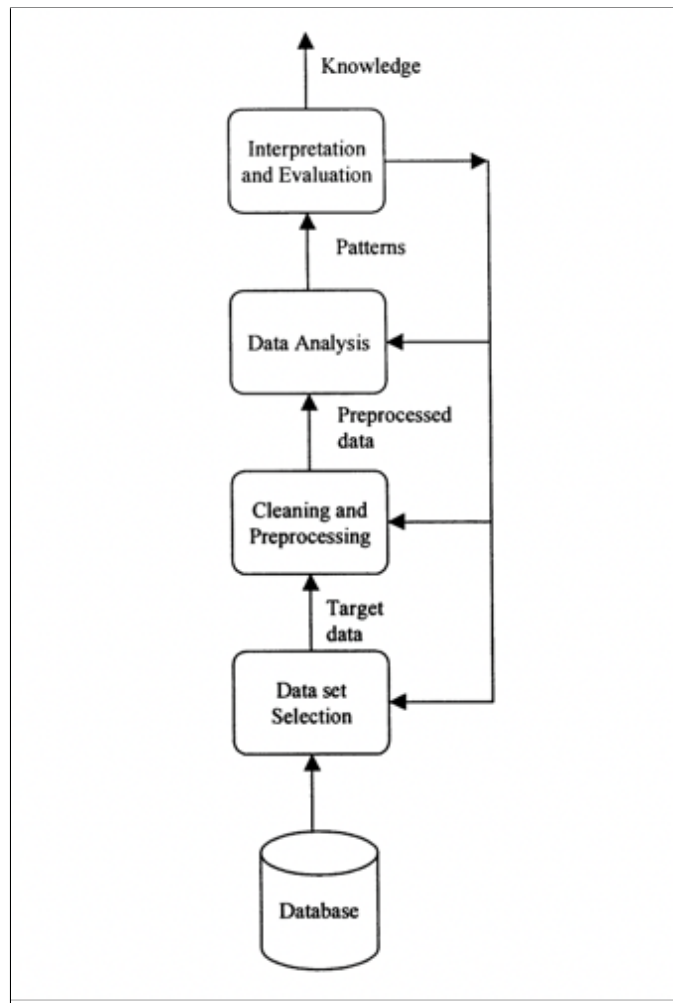


Figure 1: The Process of Data Mining [2]

There are many reasons why businesses would want to invest in predicting the future. Some of the reasons are as follow:

1. Set plans and goals [3] – Businesses may utilize prediction to create reasonable, measurable objectives based on recent and historical data. By employing accurate data and statistics, businesses may determine how much change, development, or advancement will be viewed as successful. These objectives make it simpler to evaluate the development and, when required, alter operating procedures to stay on course. Certain programmes, such as customer relationship management, that offer information on a range of topics, such as sales, opportunities, and more, may make visual forecasting simpler [3].
2. Budgeting [3] – It would provide knowledge into possible trends and changes which may help firms decide how much money and effort to spend on a task. It would also help identify the areas in which money can be used to give the best return. It makes the investment of money knowledgeable and safer.
3. Can identify changes in the market [3] – The prediction would help the business see the changes in the market and adjust accordingly. It would make it easier for the business to make a better decision before it is too late.

According to [4], there are two types of learners in classification: lazy learners and eager learners. Lazy learners keep the training data until the testing data emerges, then classify it using the data that is most related to the testing data. An example of a lazy learner would K-Nearest Neighbour. Based on the training data eager learners create a classification model. When testing data is provided it simply classifies it using the model created. Compared to lazy learners, eager learners take time to train the model, but it takes a very short time to predict. An example of an eager learner would be Naive Bayes.

3 Machine Learning Methods

For this project, seven different classifiers are used to determine the best classifier for future work. Classification is the process of predicting the class/labels of the given data. Classification algorithm in machine learning is frequently used for data mining, predictive analytics, and recognizing patterns in data and then converting them into actionable insights. In the case of this project, classification algorithms will be used to determine if a certain type of customer is going to generate revenue or not.

3.1 Naive Bayes

A Naive Bayes classifier is a classifier inspired by Bayes Theorem. It is the easiest and the fastest method of classification. This classifier is best suited for a huge amount of data. It uses a probabilistic machine learning model for classification tasks [5].

Bayes theorem formula is shown in the equation (1).

$$P(A|B) = \frac{P(B|A)P(A)}{B(B)} \quad (1)$$

By using the Bayes theorem, the likelihood that A will occur can be calculated given that B has already happened. Naive Bayes treats all the features as independent, and one feature does not impact other features.

Characteristic of Naive Bayes classifier:

1. It is an approach that operates on the premise that the predictors choose the output class equally and independently.
2. This classifier model makes the unrealistic assumption that all predictors are independent of one another, yet in the majority of cases, this assumption yields a sufficient result.
3. This classifier is used widely for text categorization.

In this project, the Gaussian Naive Bayes classifier was utilized though the results were not desirable. Gaussian Naive Bayes is based on the assumption that each class follows the Gaussian or normal distribution. It is generally used when the predictor values are continuous [5].

3.2 K-Nearest Neighbour

K-Nearest Neighbour classifier is a supervised machine learning approach [6], which means it teaches models how to create the desired output using the training data set. The symbol 'K' represents the number of nearest neighbours to the unknown variable that has to be classified or predicted [6].

To predict the class of a new data point or an unknown variable K-Nearest Neighbour employs a method of figuring out the nearest neighbour of the said new data point or an unknown variable.

K-Nearest Neighbour classifier would need to calculate the distance of the new data point to other data points. There are many methods of calculating the nearest neighbour but the most widely used one is Euclidean distance which is a true straight-line distance of two points. The function for Euclidean distance [7] is shown in the equation 2.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

In ascending order, the classifier sorts the nearest neighbours of the given point. For the classification problem, the point is classified by a vote of its neighbours, and then

the point is allocated to the class with the highest occurrence among its K nearest neighbours. The K value would determine if the prediction were good or not. To find the best value for K , it needs to learn and cross-validate.

The K value also determines the bias and variance [8]. Bias refers to the number of prejudices or assumptions your model is making against a certain problem that you're trying to frame. The more assumptions your model has the higher the bias. Variance is the sensitivity of your model to the training data so it tells us how much the output would have changed if we change the training data.

The smaller the value of K means it would have low bias and a high variance which would cause overfitting. As the name implies, overfitting accurately predicts the data. The larger the value of K means it would have high bias but a low variance which would cause underfitting. Underfitting is when the prediction is not very good.

3.3 Support Vector Machine

Like K -Nearest Neighbour, a Support Vector Machine (SVM) classifier is also a supervised machine learning approach [9]. One of the most straightforward and perhaps beautiful approaches for classification is support vector machines. The coordinates of each item that has to be classified are represented as a point in an n -dimensional space and are referred to as features. SVMs carry out the classification process by creating a hyperplane, which is a line in 2D or a plane in 3D, with all the points belonging to one category being on one side and all the points belonging to the other category being on the other.

While there may be more than one of these hyperplanes, the supporting vector machine seeks to identify the one that best distinguishes between the two groups by maximizing the distance between points in each group; this distance is known as the margin, and the points that exactly fall on the margin are known as the supporting vectors. The Figure 2 shows a hyperplane representation of the SVM algorithm [9].

SVM is referred to as a supervised learning algorithm [9] since it needs a training set, or a collection of points that have previously been labelled with the right category, to locate the hyperplane in the first place. The major advantages of his SVM are that they are simple to comprehend, use, interpret, and implement. In addition, they work well with small amounts of training data.

3.4 Logistic Regression

Logistic regression is a machine learning model which is used for binary classification. Binary classification is where the output predictions can only take one of the two possible values. for example, it can be either zero or one.

The logistic function predicts the probability that a binary result will occur. By converting data between 0 and 1 using the logistic function, nonlinearity was added

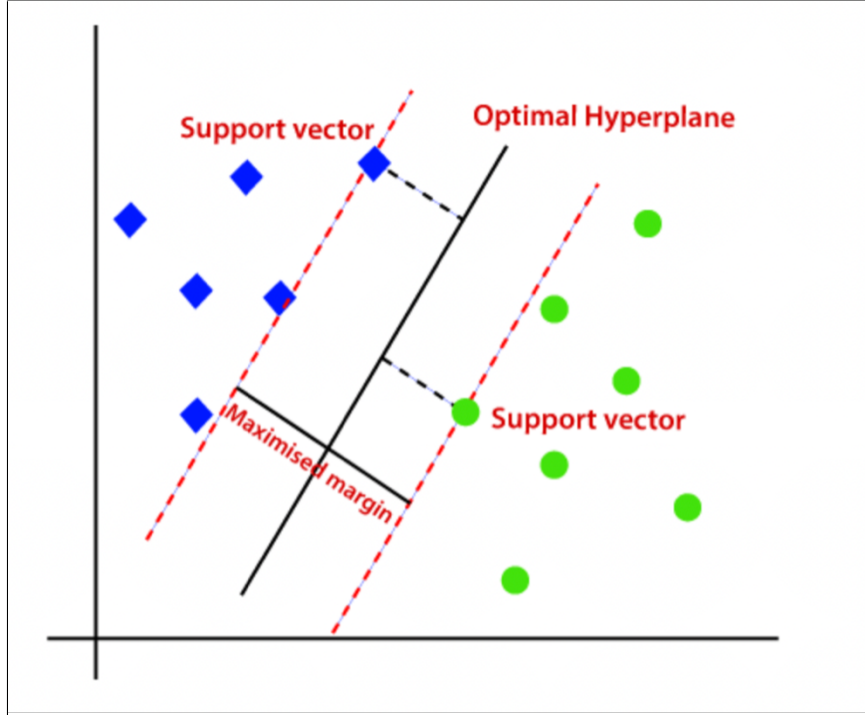


Figure 2: Hyperplane Representation of SVM Algorithm [9]

to the machine learning model [10]. The logistic function is shown in the equation 3.

$$f(x) = \frac{1}{1 + e^{-(x)}} \quad (3)$$

The data logistic regression fits an S-shaped logistic function [10] shown in the Figure 3. The curve goes from zero to 1 and it means that the curve tells you the probability that is true or false. If the probability is greater than 50% then it would categorize it as true or 1 whereas if the probability is less than 50% then it would categorize it as 0 or false.

Based on the values of an independent variable, a logistic regression classification is used to predict the category of a dependent variable. As previously established, the result of logistic regression falls under binary classification. Executing a logistic regression is simple [10]. Training is incredibly effective [10]. Overfitting is less likely to occur when using logistic regression [10].

3.5 Random Forest

The random forest, similar to the previous algorithm, is also a supervised learning algorithm used for classification and regression [11]. It is a versatile and simple algorithm. Trees comprise a forest. It is believed that the more trees a forest has, the stronger it is. Random forests build decision trees from randomly selected data

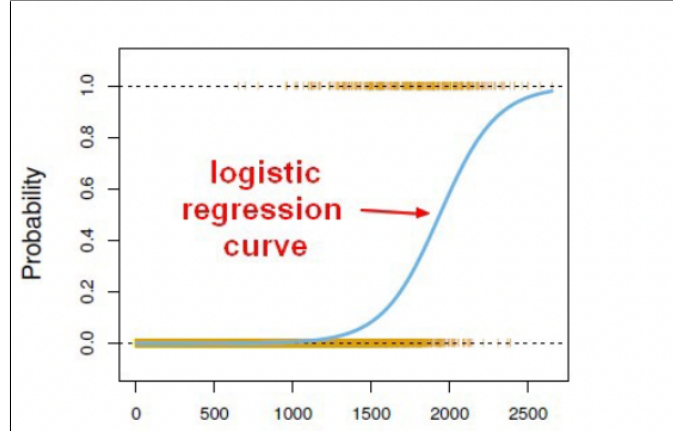


Figure 3: Logistic Regression Curve [10]

samples, then vote on the best choice based on the predictions from each tree [11]. It is also an excellent indicator of the feature’s importance.

The random forest classifiers create several decision trees from a subset of the training set. This subset of training data is selected at random for each decision tree. Each of the decision trees predicts an outcome [11]. It then combines the votes from several decision trees to get the final prediction of the random forest. The Figure 4 shows a random forest classifier making a prediction.

Methods such as Bagging, and feature randomness [11] ensure that decision trees in the random forest algorithm are not too correlated. Bagging is where random forest exploits the technique, where changing the data in the training set results in a creation of a unique tree [11]. Feature randomness is where each tree may only choose from a random subset of features [11]. This causes even more variety across the trees in the model, resulting in a weaker correlation among trees and more diversity.

3.6 Gradient Boost

Since we’re generating an immeasurable amount of data it is becoming a need to develop more advanced and complex machine learning techniques. Boosting machine learning is one such technique that can be used to solve complex data-driven real-world problems.

Boosting is an ensemble learning technique that uses a set of machine learning algorithms to convert or combine many weak learners to create a strong learner [12]. This increases the accuracy of the model. Ensemble learning is a technique that is used to enhance your model performance and its accuracy by combining several learners [12].

Gradient boosting is based on the sequential ensemble learning model where what happens is the base learners are generated sequentially in such a way that the present base learner is always more effective than the previous one. So, with each iteration,

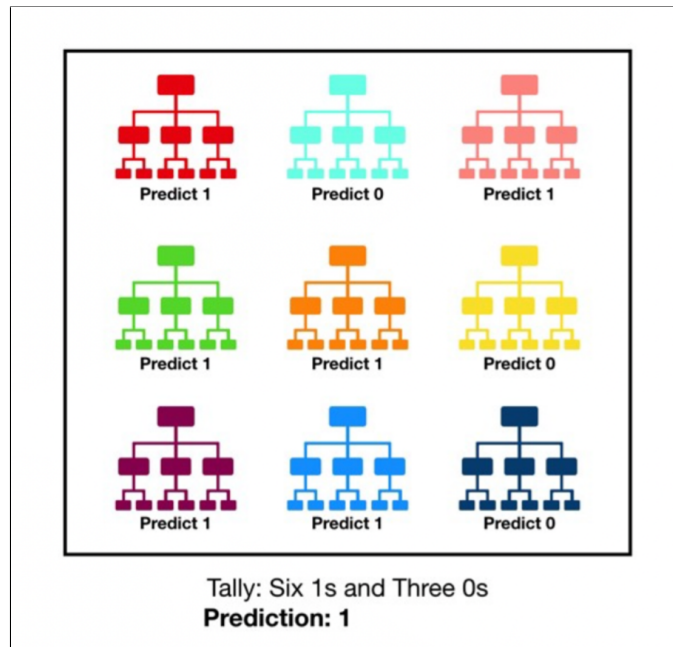


Figure 4: Visualization of Random Forest [11]

the overall model improves and tries to reduce the number of errors. Figure 5 shows the visualization of the gradient boost classifier.

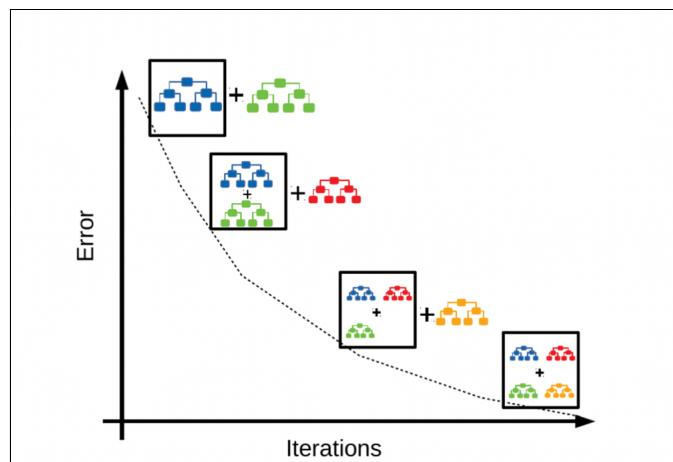


Figure 5: Visualization of Gradient Boost [12]

3.7 AdaBoost / Adaptive Boost

Like gradient boost, Adaptive boosting or AdaBoost is combining many weak learners or classifiers into a single strong learner [13].

4 Project Design and Implementation

The dataset used in the project is described and discussed in this section. The dataset is then cleansed in preparation for data analysis. This section will also go through how the dataset was used to get the results.

4.1 Dataset

Datasets can help businesses pinpoint existing issues as well as potentially predict the future. When combined with various modelling techniques, a dataset's predictive power can be rapidly enhanced. This paper will briefly discuss datasets for revenue prediction and how they can be used for predictive modelling as well as the modelling of a range of other important business metrics.

The dataset used in this project originated from UCI machine learning repository [15]. The dataset used is completely for learning purposes and does not have any relation to any real business. There are 10 numerical and 8 categorical attributes in the dataset. The target attribute from this dataset was the Revenue attribute.

The terms Administrative, Administrative Duration, Informational, Informational Duration, Product Related, and Product Related Duration describe how many different page types a visitor has accessed throughout a session and how much time they have spent on each of these pages types overall.

The metrics tracked by Google Analytics for each page of the e-commerce site are represented by the Bounce Rate, Exit Rate, and Page Value features. The percentage of users that arrive at a website through a specific page and then leave it without making any additional requests to the analytics server during that session is referred to as the bounce rate for that page [15].

The percentage of page views that were the final one of the session is used to determine the "Exit Rate" figure for a certain web page [15]. The average value of a web page that a user browsed before completing an e-commerce transaction is shown by the "Page Value" feature [15].

The "Special Day" feature shows how close a site visit is to a particular holiday when transactions are expected to be completed [15]. The importance of this attribute is established by considering e-commerce phenomena, such as the time between the order date and the delivery date. The features are explained in the 'Dataset Feature Explanation' section of the project.

4.2 Dataset Feature Explanation

There are ten numerical features and eight categorical features in the dataset.

4.2.1 Numerical Features

1. Administrative:

- It is the number of account management-related pages, such as user profiles, that were accessed by the user.
- Mean value for the feature – 2.315166
- Standard Deviation of the feature – 3.321784
- Minimum value in the feature – 0
- Maximum value of the feature – 27

2. Administrative Duration:

- It is the total time spent by the visitor on pages relating to account management, measured in seconds.
- Mean value for the feature – 80.818611
- Standard Deviation of the feature – 176.779107
- Minimum value in the feature – 0
- Maximum value of the feature – 3398.750000

3. Information:

- It is the number of pages a visitor has viewed regarding a website, communication, the address of a shopping site, seller information, etc.
- Mean value for the feature – 0.503569
- Standard Deviation of the feature – 1.270156
- Minimum value in the feature – 0
- Maximum value of the feature – 24

4. Information Duration:

- It is the total time spent by the visitor on informational pages, measured in seconds.
- Mean value for the feature – 34.472398
- Standard Deviation of the feature – 140.749294
- Minimum value in the feature – 0
- Maximum value of the feature – 2549.375000

5. Product Related:

- It is the number of product-related pages, such as product descriptions, reviews, and photos, that visitors have accessed.
- Mean value for the feature – 31.731468
- Standard Deviation of the feature – 44.475503
- Minimum value in the feature – 0
- Maximum value of the feature – 705.000000

6. Product Related Duration:

- It is the length of time, in seconds, that a visitor spends on pages related to products.
- Mean value for the feature – 1194.746220
- Standard Deviation of the feature – 1913.669288
- Minimum value in the feature – 0
- Maximum value of the feature – 63973.522230

7. Bounce Rates:

- It is the value of the visitor's average bounce rate for the pages they viewed. When a person enters a page and leaves it without viewing another one on the website or engaging with any of its components, this is known as bouncing.
- Mean value for the feature – 0.022191
- Standard Deviation of the feature – 0.048488
- Minimum value in the feature – 0
- Maximum value of the feature – 0.200000

8. Exit Rates:

- It is the visitor's average exit rate value of the pages visited. The exit rate compares the number of users who leave a website after landing on a page to the total number of views obtained by the page.
- Mean value for the feature – 0.043073
- Standard Deviation of the feature – 0.048597
- Minimum value in the feature – 0
- Maximum value of the feature – 0.200000

9. Page Values:

- It is the visitor's average page value of the pages he or she visited. The average value of a page visited by a user before completing a conversion or an eCommerce purchase is referred to as page value.
- Mean value for the feature – 5.889258
- Standard Deviation of the feature – 18.568437
- Minimum value in the feature – 0
- Maximum value of the feature – 361.763742

10. Special Day:

- It is how near a special day is to the day you're browsing the site.
- Mean value for the feature – 0.061427
- Standard Deviation of the feature – 0.198917
- Minimum value in the feature – 0
- Maximum value of the feature – 1

4.2.2 Categorical features

1. Month:

- It is the month in which the visit occurred.
- Number of values in the feature: 12

2. Operating Systems:

- It is the operating system that the user is using.
- Number of values in the feature: 8

3. Browser:

- It is the browser that the user is using.
- Number of values in the feature: 13

4. Region:

- It is the geographic area from which the visitor has begun the session.
- Number of values in the feature: 9

5. Traffic Type:

- It is the website's origin of traffic that brought the visitor. It provides a numerical number that describes how the user arrived at the page. For example, banners, SMS, direct, and advertisement link.

- Number of values in the feature: 20
6. Visitor Type:
- It is the type of visitor that is using the site.
 - There are three values recorded in this feature. They are:
 - New visitor
 - Returning Visitor
 - Other
7. Weekend:
- It is a Boolean value that represents if the visit date is on the weekend.
 - Values in this feature are true and false
8. Revenue:
- It is a class label. This feature indicates if a transaction was performed during the visit, producing revenue.
 - Values in this feature are also Boolean.

In this dataset, there are 12330 entries. And 18 data columns are listed in Figure 6. There are no null values in the dataset as shown in Figure 6.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
#   Column                               Non-Null Count  Dtype
---  -
0   Administrative                       12330 non-null  int64
1   Administrative_Duration              12330 non-null  float64
2   Informational                        12330 non-null  int64
3   Informational_Duration               12330 non-null  float64
4   ProductRelated                      12330 non-null  int64
5   ProductRelated_Duration             12330 non-null  float64
6   BounceRates                         12330 non-null  float64
7   ExitRates                           12330 non-null  float64
8   PageValues                          12330 non-null  float64
9   SpecialDay                          12330 non-null  float64
10  Month                               12330 non-null  object
11  OperatingSystems                    12330 non-null  int64
12  Browser                             12330 non-null  int64
13  Region                             12330 non-null  int64
14  TrafficType                         12330 non-null  int64
15  VisitorType                         12330 non-null  object
16  Weekend                             12330 non-null  bool
17  Revenue                             12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)

```

Figure 6: Dataset Information

4.3 Data Description

The main objective of this project is to create a classification system using machine learning that can anticipate an online customer’s intention based on the values of the provided features.

To choose the optimal classification method for the project, we test out a variety of classifiers and evaluate their performance.

For the project, I have connected google drive to the notebook so that the dataset can be accessed through it. All the necessary library needed for the project was then installed on the notebook. The dataset was imported into the notebook as shown in Figure 7. The head function from the pandas library was used for the first few rows in the dataset as shown in the Figure 8 and Figure 9. The statistical analysis of the dataset was calculated and in that statistical analysis the mean, standard deviation, minimum, maximum and the four percentiles were calculated using the described function. The statistical analysis is listed under the 'Dataset Feature Explanation' of the project.

```
[ ] df = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/FinalProjectMHK/online_shoppers_intention.csv")
```

Figure 7: Importing Dataset from the Google Drive

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues
0	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0
1	0	0.0	0	0.0	2	64.0	0.0	0.1	0.0
2	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0

Figure 8: Summary of Data Used

SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.0	Feb	4	1	9	3	Returning_Visitor	False	False

Figure 9: Summary of Data Used Cont.

4.4 Implementation Method

To deal with any dataset, the most important step is to see if the dataset is clean or not. Missing and incorrect values in the dataset can produce undesired inaccuracies in predictions. Unclean datasets can also have an impact on graphs and plots. Following cleaning and preprocessing, the dataset may be utilized for a variety of purposes. The dataset can be used for gaining some insight into the data using visualization tools. The project uses matplotlib and seaborn libraries for data visualization.

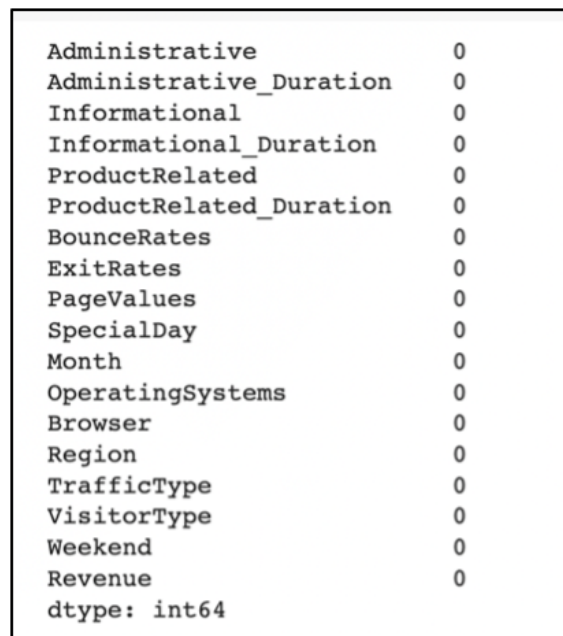
A correlation heatmap is a typical approach for visualising the relationship between different features in a dataset. Correlation analysis helps in identifying the relationship between different features and it indicates the strength of the relationship. The correlation heatmap for the dataset is analysed in the 'Data Cleaning' section.

We looked at the data's distribution using a variety of visualisation tools to gain a clear understanding of the data. Using matplotlib and seaborn libraries, we were able to visualize the distribution of features among their different values. We were also able to show the division of the data on the target (revenue) feature with the aid of the same visualisation libraries.

Once the dataset has undergone analysis, several classifiers were utilized for making predictions. To use the dataset for classifiers, it was split into 2 parts, one for training and the other for testing. The results of the classifiers are in the 'Results and Discussion' section part of the project.

4.5 Data Cleaning and Preprocessing

The first step in the data cleaning process was to check if there were any missing values in the dataset. In this dataset, there were no missing values as shown in Figure 10. The next step was to fix the data type by converting the Boolean into integers. Values of weekend and revenue were converted from Boolean to an integer.



Administrative	0
Administrative_Duration	0
Informational	0
Informational_Duration	0
ProductRelated	0
ProductRelated_Duration	0
BounceRates	0
ExitRates	0
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0
dtype: int64	

Figure 10: Checking for null Values in the Dataset

Correlation analysis between features was made using a heatmap as shown in Figure 11. The heatmap of correlation indicates that there is very little association

between the various features in the dataset. Some of the cases where the correlation is higher than 0.7 are the bounce rate and exit rate with a correlation of 0.91. Product related and product-related duration also have a high correlation of 0.86. There are also cases where the correlation between features is moderate by which it is from 0.3 to 0.7. These cases include features such as administrative and administrative duration, informational and informational duration, and page value and revenue.

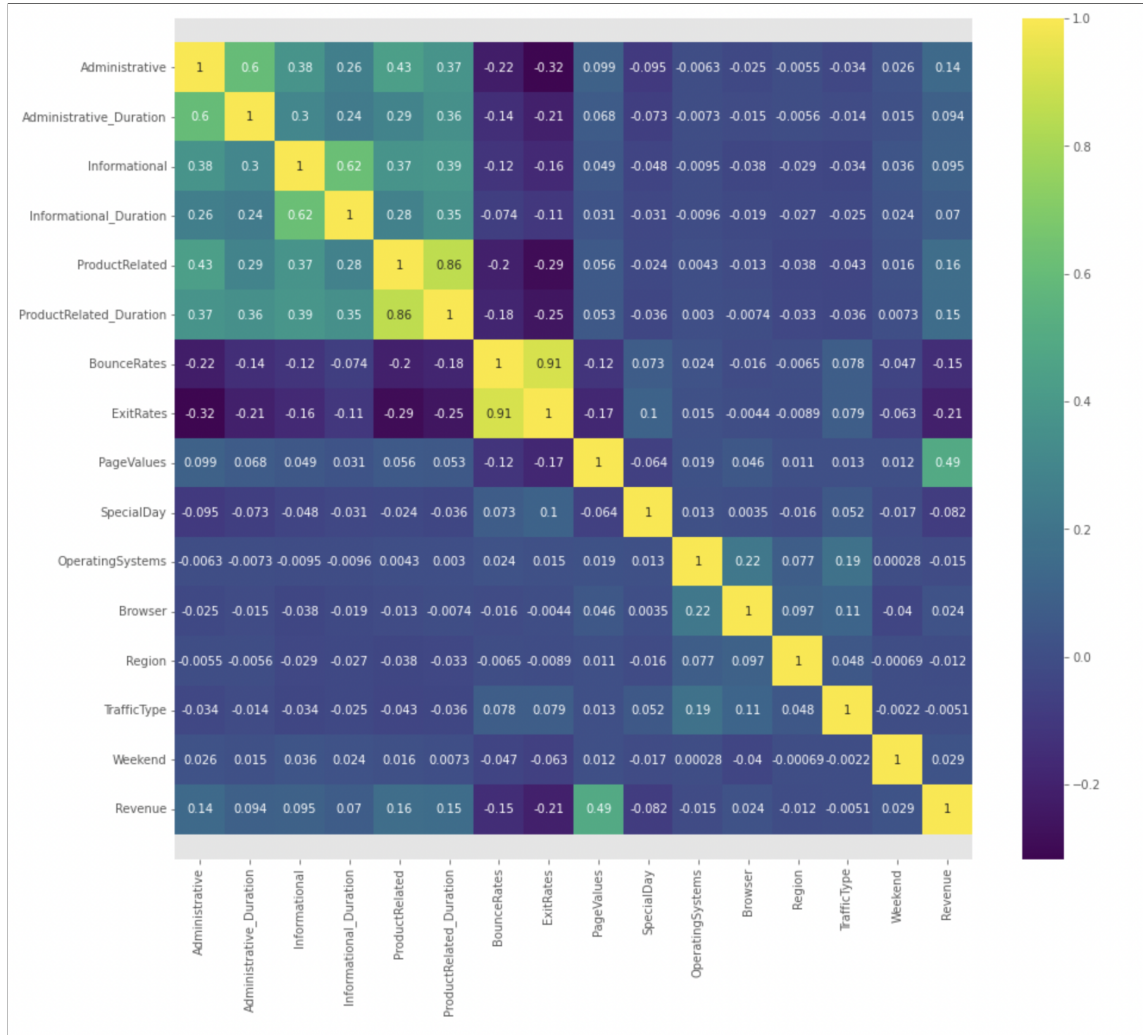


Figure 11: Correlation Heatmap between Features

5 Results and Discussion

This section provides data analysis now that the dataset has been cleansed. The data analysis contains graphs and plots created by the visualizations library, as well as their discussion. Afterwards, classifiers are used to check their performance. The results for classifiers include accuracy, precision, recall, and F1 score. In the end, the ROC curve is generated to determine the best classifier.

5.1 Data Analysis

The distribution of the dataset has more than 10000 visitors who did not contribute to the generation of revenue. Approximately 2000 visitors contributed to the generation of revenue as shown in Figure 12. This would indicate that most people who visit the site do not generate revenue and that the dataset is unbalanced. To understand why there is such disparity the dataset needs to be further analyzed.

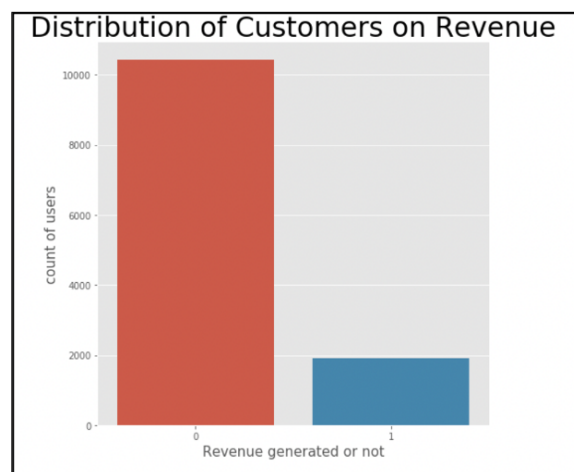


Figure 12: Distribution of customer on Revenue

Based on the dataset it can be said that the month in which most revenues were generated was in November as shown in Figure 13. The percentage of users who visited the site each month can also be seen in Figure 14. The figure suggests that there are a lot of visitors in months such as November, March, December, etc. Though the month with the highest percentage of visitors, May, had very little revenue generation. From this, we can confirm that there is a lot of visitor across the year but not a lot of buyers.

From the dataset, there is a lot of visitor on the site but not all the visitors are customers. The website's revenue must now be increased by turning all users into customers. To do this, we must comprehend how each user navigates the website, which products they view during a session, and which pages generate revenue.

We've noticed that there's a good chance we'll lose potential consumers if they spend too much time on various website pages. The bi-variate analysis of the time spent

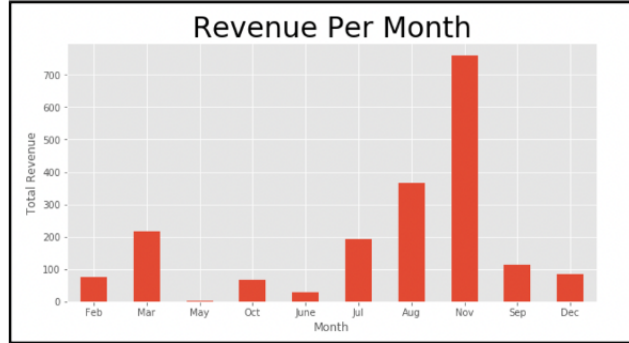


Figure 13: Revenue Per Month

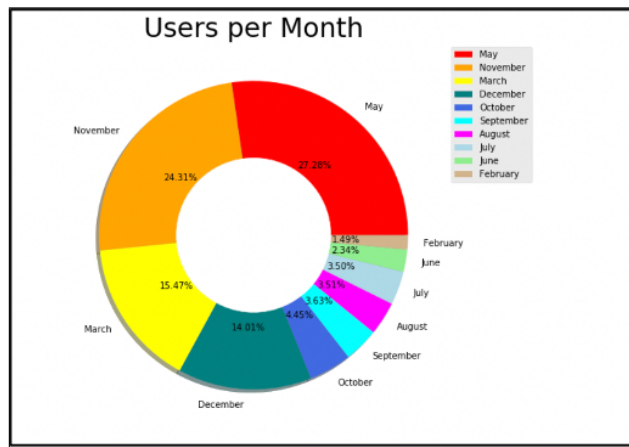


Figure 14: Users Per Month

on each type of page and the amount of revenue earned or not may be used to study this pattern which can be seen in the figures 15, 16, and 17.

Figure 15 gives the amount of time spent on information pages which generate revenue or not. Based on the plot it can be said that the more time the visitor spends on the informational pages the less likely they are to purchase anything from the website. Most of the transactions occur during the first two hundred seconds of visiting the informational pages. So, I would suggest that the informational pages need to be reworked so that it encourages the customer to buy the product on the site.

The time spent on administrative pages, whether they bring in revenue or not, is shown in the Figure 16. Like the previous plot, this one suggests that visitors are less likely to make purchases from the website the longer they stay on the administrative pages. Most transactions take place during the first 500 seconds of accessing the administration pages. To encourage customers to purchase the goods on the website, I would thus propose that the administration pages also need to be updated.

The time spent on product-related pages that either generate revenue or not is shown in the Figure 17. According to the plot, visitors spend a lot more time on product-

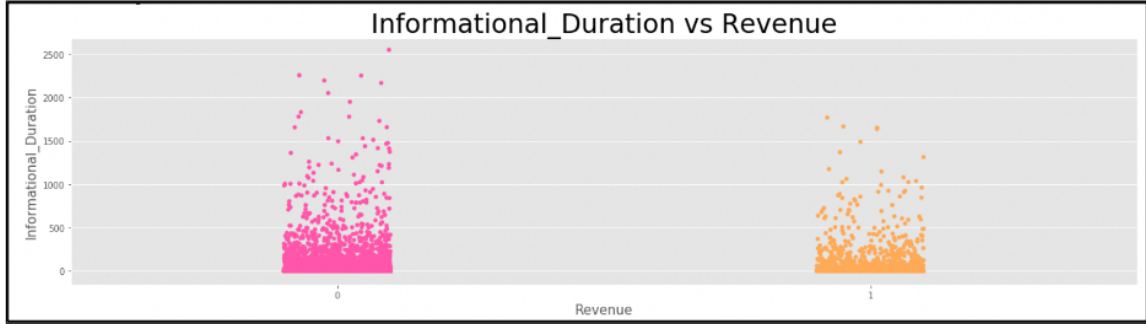


Figure 15: Informational Duration vs Revenue

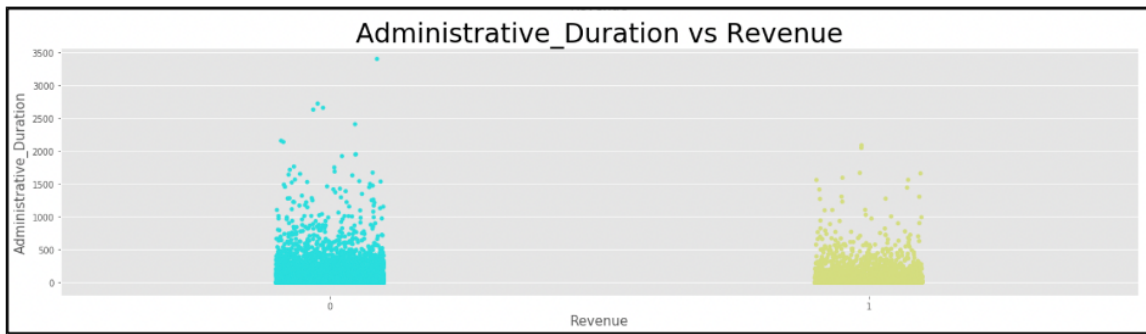


Figure 16: Administrative Duration vs Revenue

related pages than on other pages. However, much like other pages, the longer a visitor stays on the product-related pages, the less probable it is that the visitor would make a purchase from the website.

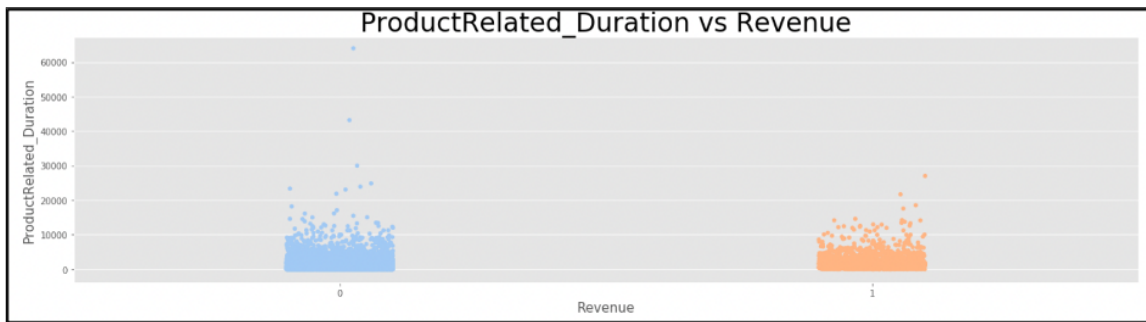


Figure 17: Product-Related Duration vs Revenue

The visitors coming to the site are mostly returning customers there are only a few newcomers and other visitors as shown in the Figure 18. The site also generates the most revenue from returning customers as shown in the Figure 19. Based on the results it can be said that businesses to improve their marketing strategy. The business should have some incentives in place for the customer to come and spend on the product and service they are offering.

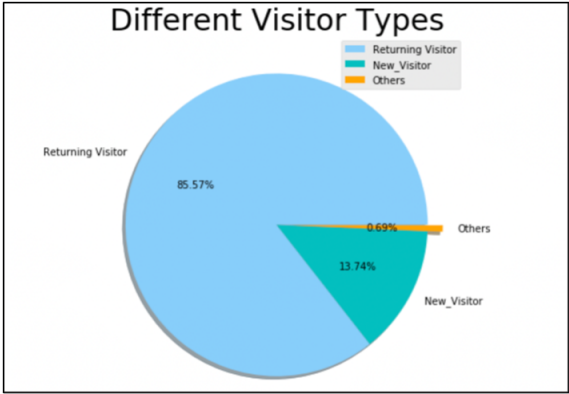


Figure 18: Different Visitor Types

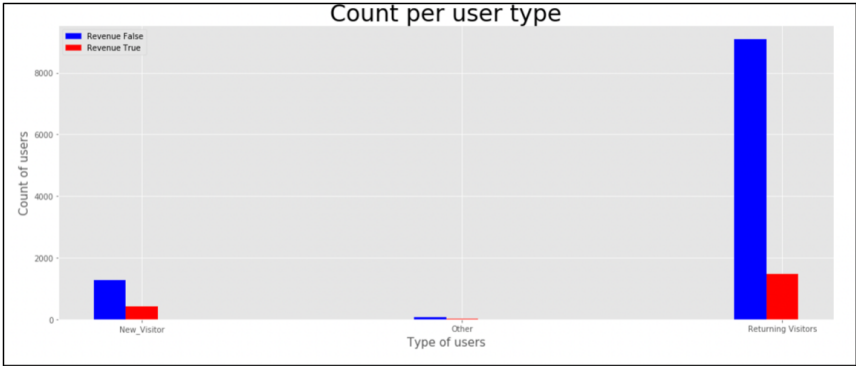


Figure 19: Count Per User Type

5.2 Building and Testing the Classifiers

For this part of the project, we have split the dataset into 2 different parts. The first part of the dataset is the training set which has 70% of all the dataset. The second part of the dataset is the testing set which has 30% of all the dataset.

For each of the classifiers their accuracy, F1 score, precision, recall and confusion matrix were requested as an output. Accuracy is how often the classifier is correct [14]. Precision is how the many true positive was predicted over all the positive predictions. The recall is how many actual positive cases the classifier was able to predict correctly. F1 score is the harmonic mean of precision and recall. Confusion matrix is a $N \times N$ matrix. Since the project is about predicting if the customer would generate revenue or not, the confusion matrix would be 2×2 as shown in the Figure 20.

n=165	Predicted: NO	Predicted: YES
	Actual: NO	50
Actual: YES	5	100

Figure 20: Confusion Matrix Example [14]

5.3 Different Classifiers Results

1. Gaussian Naive Bayes –
 - Accuracy: 0.7791294944579616
 - F1 Score: 0.48126984126984135
 - Precision: 0.37711442786069654
 - Recall: 0.6649122807017543
 - Confusion Matrix: Figure 21
2. K-Nearest Neighbor
 - Accuracy: 0.8626655852933225
 - F1 Score: 0.4253393665158371
 - Precision: 0.5987261146496815
 - Recall: 0.3298245614035088
 - Confusion Matrix: Figure 22

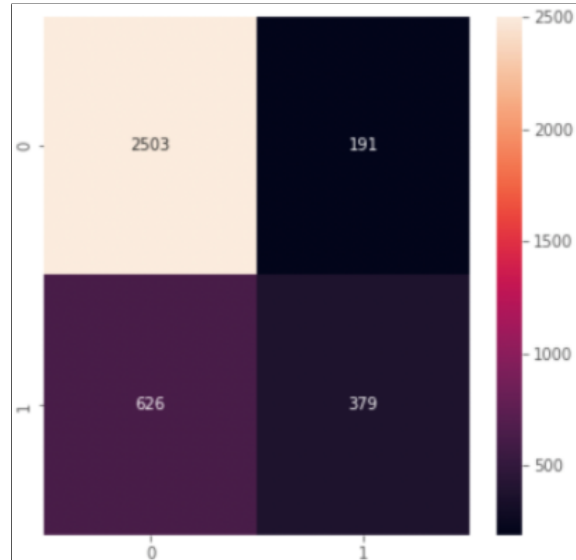


Figure 21: Confusion Matrix for Gaussian Naive Bayes

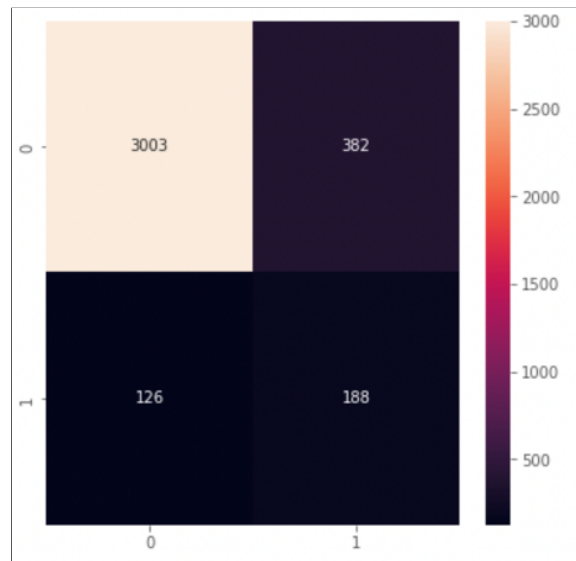


Figure 22: Confusion Matrix for K-Nearest Neighbor

3. Support Vector Machine

- Accuracy: 0.8475263584752636
- F1 Score: 0.02422145328719723
- Precision: 0.875
- Recall: 0.012280701754385965
- Confusion Matrix: Figure 23

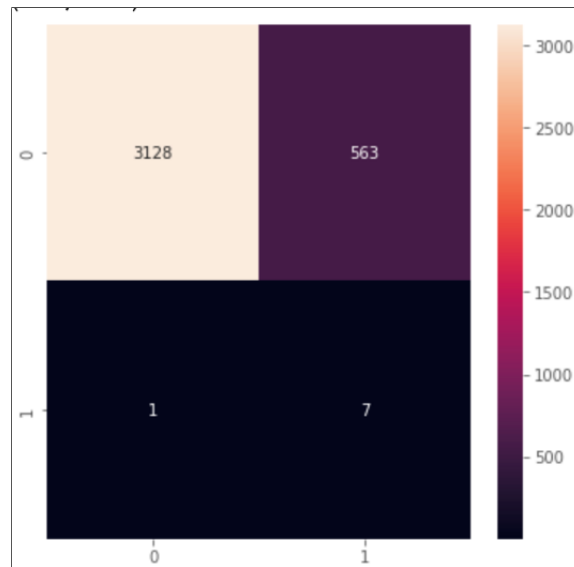


Figure 23: Confusion Matrix for Support Vector Machine

4. Logistic Regression

- Accuracy: 0.8851040821843742
- F1 Score: 0.5186862967157417
- Precision: 0.731629392971246
- Recall: 0.4017543859649123
- Confusion Matrix: Figure 24

5. Random Forest

- Accuracy: 0.9034874290348743
- F1 Score: 0.6433566433566434
- Precision: 0.7470997679814385
- Recall: 0.5649122807017544
- Confusion Matrix: Figure 25

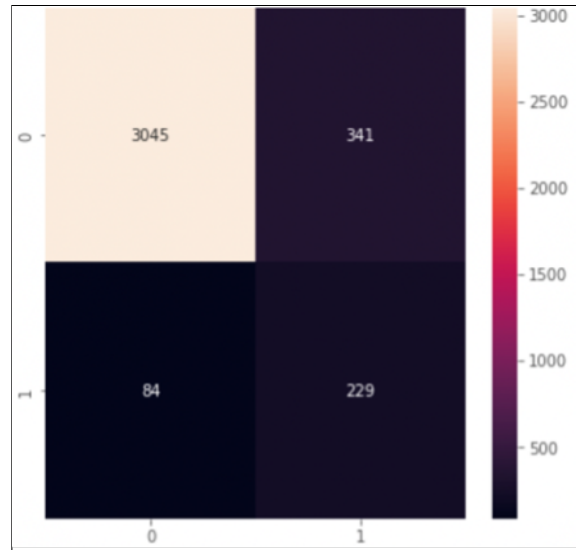


Figure 24: Confusion Matrix for Logistic Regression

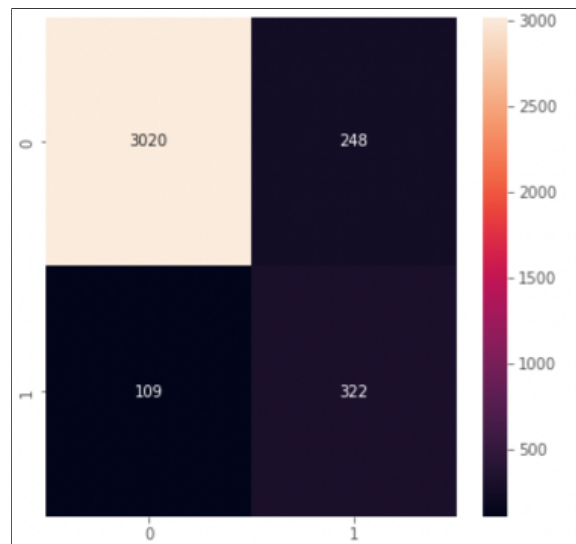


Figure 25: Confusion Matrix for Random Forest

6. Gradient Boosting

- Accuracy: 0.9056501757231684
- F1 Score: 0.6685660018993353
- Precision: 0.7287784679089027
- Recall: 0.6175438596491228
- Confusion Matrix: Figure 26

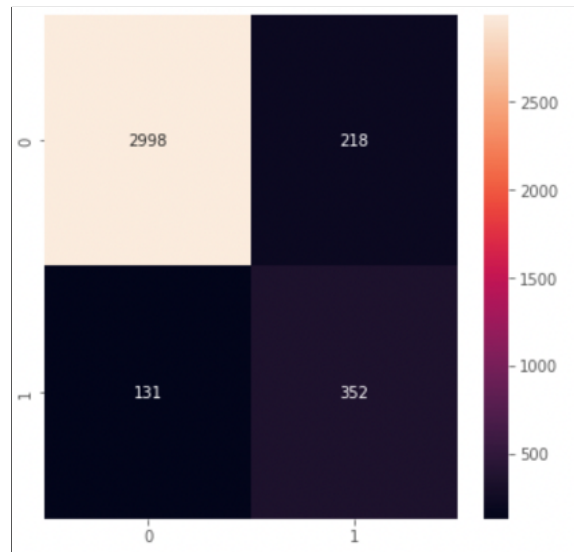


Figure 26: Confusion Matrix for Gradient Boosting

7. Ada Boosting/ Adaptive Boosting

- Accuracy: 0.8875371722087051
- F1 Score: 0.6068052930056711
- Precision: 0.6577868852459017
- Recall: 0.5631578947368421
- Confusion Matrix: Figure 27

Table 1 shows the performance of all the classifiers tested for the project. Though the accuracy of the classifiers is similar, the gradient boosting classifier has the best accuracy, with random forest coming in second. Based on the precision scores of the classifiers, the SVM classifier performed the best, while the Gaussian Naive Bayes classifier performed the poorest. The recall score was the inverse of precision, with the Gaussian Naive Bayes classifier having the best recall and the SVM having the poorest. To provide an output, the f1 score considers both precision and recall. Better F1 scores are produced by higher precision and higher recall score. This indicates that even though the SVM classifier had the greatest accuracy score, it had

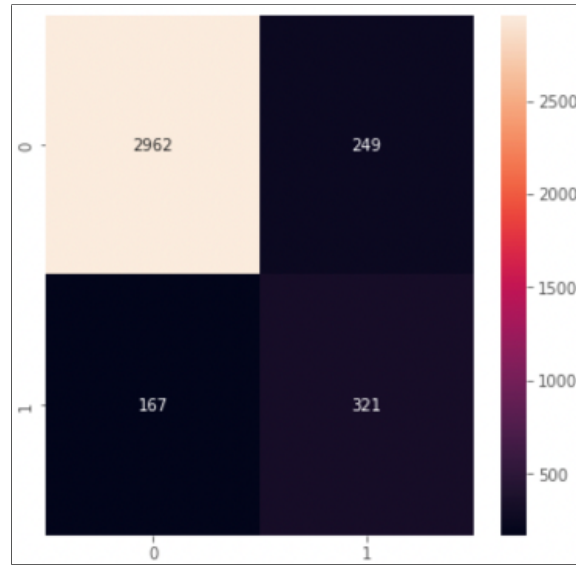


Figure 27: Confusion Matrix for Ada Boosting

Table 1: Classifier Prediction Performance

Classifier	Accuracy	Precision	Recall	F1 Score
Gaussian Naive Bayes	77.91%	37.71%	66.49%	48.13%
K-Nearest Neighbor	86.27%	59.87%	32.98%	42.53%
Support Vector Machine	84.75%	87.50%	1.22%	2.42%
Logistic Regression	88.51%	73.16%	40.18%	51.86%
Random Forest	90.35%	74.71%	56.49%	64.34%
Gradient Boosting	90.57%	72.88%	61.75%	66.86%
Ada Boosting	88.75%	65.78%	56.32%	60.68%

the poorest recall score, which is what caused the SVM F1 score to be low. The gradient boosting classifiers produced the best F1 score.

From the performance gathered of each of the classifiers it can be said that gradient boosting is the highest performing classifier. Based on the performance of the classification model at all classification levels, the receiver operating characteristic (ROC) curve was graphed as shown in the Figure 28. False positive rate is plotted on the X-axis [18] of a typical ROC curve, and true positive rate is plotted on the Y-axis [18]. The greater the region that the curve covers, the more effectively the machine learning models.

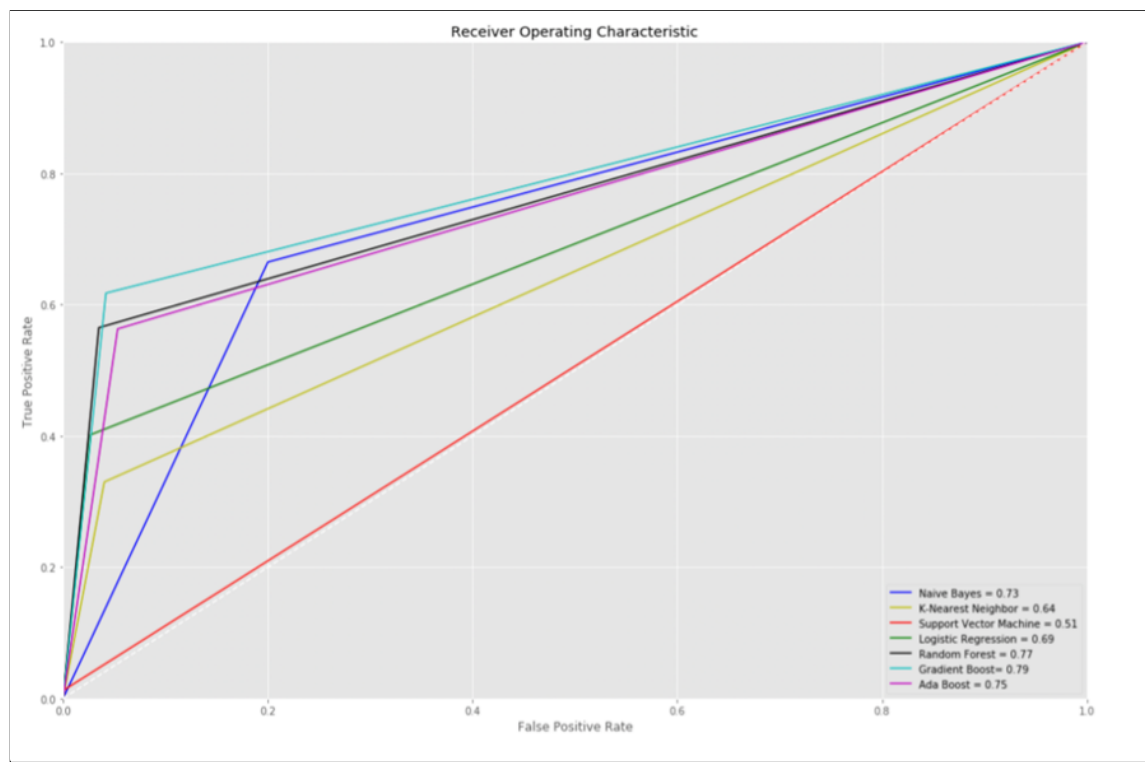


Figure 28: ROC Curve

Although most of the classifiers' performance was very similar, it can be said that based on the ROC curve, gradient boosting is more reliable compared to others. The least reliable one was the support vector machine because of its poor recall performance. It is not surprising that gradient boosting performance was the best since it is a boosting algorithm which combines many weak learners to create a strong learner.

6 Conclusions

The data analysis suggests that the business needs many improvements. Based on the dataset, it can be said that the site had a very low rate of new visitors and most of the visitors of the website did not buy any product or service. There were instances where the more the user surfed the website, the less likely he/she is to buy the product or service from the website. It can be said that the business needs to redesign the website, create an incentive to buy products or services from the business and develop a marketing strategy so that people would learn about this website.

To conclude, it is very important for businesses in this new era to collect and analyze data. Many businesses have already recognized the importance of data mining and have started to move in the direction of implementing it. Making the right move at the right time gives a business an edge over its competition and with the advancement in machine learning technologies and deep learning it has become very easy to predict future events.

7 Future Work

Following the analysis, we have determined the best classifiers from the ones mentioned above. We can now optimize this classifier to give better performance. Once optimization is complete, we can utilize the model to create a web application that the analyst may use to estimate whether or not the customer will generate revenue.

References

- [1] Morlidge, Steve, and Steve Player. Future Ready: How to Master Business Forecasting. John Wiley & Sons, 2010. <https://books.google.ca/books?hl=en&lr=&id=PZYMEAAAQBAJ>
- [2] Bose, Indranil, and Radha K. Mahapatra. 'Business Data Mining — a Machine Learning Perspective'. Information & Management, vol. 39, no. 3, Dec. 2001, pp. 211–25. ScienceDirect, [https://doi.org/10.1016/S0378-7206\(01\)00091-X](https://doi.org/10.1016/S0378-7206(01)00091-X)
- [3] Hall, Remington. Why Forecasting Is Important for Business Success. <https://www.baass.com/blog/why-forecasting-is-important-for-business-success> Accessed 19 July 2022.
- [4] Asiri, Sidath. 'Machine Learning Classifiers'. Medium, 11 June 2018, <https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623>
- [5] 'Implementing Gaussian Naive Bayes in Python'. Analytics Vidhya, 29 Nov. 2021, <https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/>
- [6] 'KNN - The Distance Based Machine Learning Algorithm'. Analytics Vidhya, 15 May 2021, <https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/>
- [7] 'Most Popular Distance Metrics Used in KNN and When to Use Them'. KDnuggets, <https://www.kdnuggets.com/most-popular-distance-metrics-used-in-knn-and-when-to-use-them.html/> Accessed 21 July 2022.
- [8] Lin, Tzu-Chi. 'Day 3 — K-Nearest Neighbors and Bias-Variance Tradeoff'. 30 Days of Machine Learning, 4 Dec. 2018, <https://medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff-75f84d515bdb>
- [9] 'Support Vector Machine(SVM): A Complete Guide for Beginners'. Analytics Vidhya, 12 Oct. 2021, <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/>
- [10] 'Logistic Regression: An Introductory Note'. Analytics Vidhya, 17 Jan. 2022, <https://www.analyticsvidhya.com/blog/2022/01/logistic-regression-an-introductory-note/>
- [11] Yiu, Tony. 'Understanding Random Forest'. Medium, 29 Sept. 2021, <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

- [12] Pal, Aratrika. 'Gradient Boosting Trees for Classification: A Beginner's Guide'. The Startup, 2 Oct. 2020, <https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea>.
- [13] Desarda, Akash. 'Understanding AdaBoost'. Medium, 17 Jan. 2019, <https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe>.
- [14] 'Simple Guide to Confusion Matrix Terminology'. Data School, 26 Mar. 2014, <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>.
- [15] Sakar, C. Okan, et al. 'Real-Time Prediction of Online Shoppers' Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks'. *Neural Computing and Applications*, vol. 31, no. 10, Oct. 2019, pp. 6893–908. DOI.org (Crossref), <https://doi.org/10.1007/s00521-018-3523-0>.
- [16] Evans, Michelle. 'Global E-Commerce Market To Expand By \$1 Trillion By 2025'. Forbes, <https://www.forbes.com/sites/michelleevans1/2021/03/25/global-e-commerce-market-to-expand-by-us1-trillion-by-2025/>. Accessed 22 July 2022.
- [17] Subramanian, Rohini. 'How Is Customer Data Used in Ecommerce Industry'. TheCommerceShop, 13 Apr. 2022, <https://www.thecommerceshop.com/blog/customer-data-and-ecommerce-how-important-is-it/>.
- [18] 'Receiver Operating Characteristic (ROC)'. Scikit-Learn, https://scikit-learn/stable/auto_examples/model_selection/plot_roc.html. Accessed 23 July 2022.