INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT046-3-M-AML

APPLIED MACHINE LEARNING

ASSIGNMENT 2 (PART - B)

HAND OUT DATE: 15 FEBRUARY 2021

HAND IN DATE: 07 MAY 2021

WEIGHTAGE: 60%

INSTRUCTIONS TO CANDIDATES:

1 **Submit your assignment at the administrative counter.**

2 **Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing).**

3 **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**

4 **Cases of plagiarism will be penalized.**

5 **The assignment should be bound in an appropriate style (comb bound or stapled).**

6 **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.**

7 **You must obtain 50% overall to pass this module.**

MACHINE LEARNING CLASSIFICATION FOR ECOMMERCE CUSTOMER BEHAVIOUR

By

Marakathavalle Muthu Chidambaram

TP062882

APDMF2101AI

CT046-3-M APPLIED MACHINE LANGUAGE

A project submitted is fulfillment of the requirements of the degree of

Master of science in Artificial Intelligence

Asia Pacific University of Technology and Innovation (APU)

MARCH 2021

Acknowledgement

While doing this paper, I spent most of the time searching for a novel dataset that will also satisfy the constraint of being one of the fields I have the interest. In the end it helped me to have the knowledge about various available datasets and different topics where machine learning was used for research. In that journey, I came across huge set of real time data as well, it was indeed a great learning process. I would like to express my gratitude and thank Prof. Dr. Mandava Rajeswari for letting us select our desired topic and being enthusiastic to guide us throughout the report and APPLIED MACHINE LEARNING module. Google dataset search, Kaggle, Rstudio and many other websites as well as software have been so useful all way, thanks to the creators and great minds. I am so grateful to all my mentors, family and god who have guided through everything.

Abstract

In this paper, real-time e-commerce customer behaviour is analysed using machine language techniques which predicts the visitor's shopping intent and Web site negligence chances. E-commerce follows customer-centric culture thus the need for the study of customer intention is highly required. E-commerce is one of the highest growing economic sectors after the COVID pandemic all around the globe. Customer behaviour pattern reorganization is required to further grow in this challenging business. This research aims to build various machine learning models from the selected dataset. Various related works that are available for the given dataset are analysed and further improvement is done. The summary of the work is shortened from the research that included gap analysis and scope. Furthermore, the exploration data analysis is carried out to plot the behaviour of the data elements specific to columns, and outliers are analysed. Next, the various models are studied determining the accuracy rate for each model which is followed by implementation of the model that results in prediction and output. Finally, the recommendation is provided to improve the Web site abandonment and purchase conversion rates by studying the customer intention.

Table of Contents

List of Tables

1. Introduction

1.1. Introduction

In today's world, everything is turning online, e-commerce and m-commerce are one such blooming industry. It has become an easy success in this era for online business due to the availability of the internet and smartphones in most people's hands in all countries. COVID-19 has caused major destruction to many people and industries, but it has been a boon to e-commerce which has contributed to the upward trajectory in the world economy. Digitalization is occurring in almost all businesses from small scale to giant companies and there are different types of customers involved for each business. Many vendors are still a novice to the adaptation of e-commerce which gives both hand lift and an obstacle to their business. Statistics states that globally e-commerce contributed to 1,336 billion US dollars in 2014 to 4,206 billion US dollars in 2020 and it is predicted to grow to 6,542 billion US dollars in 2023. (https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/). It is a huge and fast growth compared to another competing economic sector. The use of AI (Artificial Intelligence) in e-commerce is an added advantage, many AI techniques such as chatbots, voice assistants with NLP (Natural Language Processing), machine learning, and image processing impacts content writing, searching, payments, and demand prediction of the product as well as customers, all these added features can cause enormous growth to e-commerce. Since AI can analyse and foresight bi data and machine learning it can predict the customer behaviour pattern which can be triggered by advertising campaigns, promotions, and discounts. Therefore, it can recommend the customer on the products and services and more that will be relevant and useful for the customers.

"Thus, the AI-enabled data-driven business approach helped e-commerce to make accurate lead scoring and sales prospect. Besides, analysing historical and real-time data, applied machine language allows the strategists to formulate better decisions for business growth." [Stephanie, 2020]

Based on the available information, customer behaviour in online business is predicted to build the business further so that customers are satisfied and inspired to continue online shopping at the ease of their home and online retailers are benefited by posting the right advertisement which customer prefers and increasing their sales based on prediction.

1.2. Problem statement

Though the online business has spread its wings in greater aspects, it still faces major challenges to overcome. One of such challenges is to study customer behavior and customer satisfaction. As a business, an e-commerce site uses the source to sell the product. The long-time practice of customers in buying products at a brick-and-motor store is transformed into buying on online websites. So, this drastic change of the customers must be encouraged by studying and predicting their usage with relevant available data and providing them information and the right product. Customer behavior analysis is important and needed in e-commerce to increase the revenue of the vendors. The customer expects the seller to treat them better or equal to offline selling which leads to a reduction in customer churn. Customers spend many hours browsing the e-commerce website but the number of customers who purchase and make revenue to the seller is low. Determining the likelihood of purchase is required to increase sales and revenue.

1.3. Research Goal

The motive of the research is to study various machine learning models and compare them to determine the most accurate model that suits the dataset. Further, enhance the prediction using a various available algorithm to determine the likelihood of purchase between various factors that leads to revenue.

1.4. Research Objectives

- To create and compare various machine learning models to determine if the shopper is paying customer.
- To optimize and power transform the data for higher accuracy.
- To identify user behaviour patterns to effectively understand features that influence the sales.

2. Literature Review

2.1. Introduction

The Literature review establishes the various related works of a different researcher from the year 2019-2020 in the Online Shopper's Purchasing Intention Dataset. Each distinct kernel contains various machine learning models along with the analysis and pre-processing techniques and its accuracy is determined. Further, all the previous methods used are compared and a summary is derived. Finally, the gap analysis and scope are determined. Dataset is obtained from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset. The

Author of the dataset is Kelleher, John, et al. Fundamentals of Machine Learning for Predictive Data Analytics. The MIT Press, 2015.

2.2 Related works

Citation	Dataset	Size	EDA	Model	Pre -Processing	Accuracy	Comments
Kageyama (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	Plot feature distribution	Bayesian Optimization	-Extract dummies of certain columns -Manually drop revenue column -Train Test split 0.7:0.2.	ACC:0.902	-One classification model is tested -Prediction with more accuracy is determined
Vignesh Prakash (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Box plot of all columns -Outlier's analysis -PCA analysis -Univariate and multivariate analysis	Logistic regression	-Statistical test for categorical column and numerical column vs Target column. -Detecting and dropping multicollinearity.	ACC:0.8	-Full EDA analysis using PCA component is studied, and accuracy is measured
Henry Sue (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Checking for NULL value and unique value in dataset. -Visualization of data.	Naïve Bayes Random forest Extra trees	-Label and One Hot Encoding. -Feature cleaning.	ACC:0.84 ACC:0.902 ACC:0.895	-Each model is evaluated with stratifying training data.

Table 1 Related work part 1

Swapnil Bhange (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	- Univariate Analysis -Bivariate analysis -Multivariate analysis	Logistic regression Decision tree Support Vector Machine(SVM)	-Converting Outliers to NaN -MICE (Missing Imputations through Chained Equations) technique for imputing the missing values -Chi-square test	ACC:0.85 ACC:0.85 ACC:0.84	-Each classification model is studied with and without SMOTE

Beth Morrison (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-count plot to plot histogram -boxplot customers	Random Forest	-Target and feature selection -Train Test split 0.7:0.3 -Pipeline is bult to standardize -Tune the hyper parameter	ACC:0.91	-Model pipeline is built and data sampling is not performed.
				XGBoost		ACC:0.87	
Daewoongjun (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Count plots to measure customer in each feature. -Pie chart for quantity ratio - Stacked bar charts for true and false revenue	Logistic Regression	Label encoder to convert categorical columns to numerical.	ACC:0.88	-Comparison with other models is studied. -Data sampling is not employed.
				Random Forest		ACC:0.90	
				Support Vector Machine(SVM)		ACC:0.84	
				K Nearest Neighbour (KNN)		ACC:0.86	
Tushar Vij (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Statistical Analysis -Outlier Analysis	XGBoost	-Scaling and encoding. -Base model with and without transformation	ACC:0.9	-Base model with and without transformation is studied including all classification models and power transformation is deployed.

Table 2 Related work part 2.

Aurelia Sui (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Visualizing correlation between numeric values -Visualizing distribution.	Multilayer Perceptron (MLP) in Artificial Neural Network	- MLP model with single hidden layer -Train Test split.	ACC:0.89	-Single hidden layer is deployed. -Test data is not used.
Annette Catherine Paul	Online Shopper's Purchasing	12330 x18	-Correlation plot.	Decision tree	-Missing value analysis.	ACC:0.218 (Test data)	-Each category are evaluated

(2020)	Intention Dataset		-Trend line for revenue.		-Descriptive analysis. - Train Test split on datasets.		separately with details. -Only test data is evaluated.
Oscar Matias Torros (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Visual analysis. -Outlier Analysis. -Correlation plot.	Random Forest Neural Network Classification Logistic Regression	-Finding Null value -Analysing each column for strange values.	ACC:0.91 ACC:0.88 ACC:0.88	-Models are explained using Permutation. -Each column is thoroughly studied.
Saurabh Gupta (2020)	Online Shopper's Purchasing Intention Dataset	12330 x18	-Visualization of columns using plot graph. -Axes subplot.	K-Nearest Neighbour	-Data cleaning and standardisation	ACC:0.88	-ROC curve and confusion matrix are defined to study the result. -EDA and data sampling is not performed.

Table 3 Related work part 3

2.3 Comparison of Previous Methods

Table 1, Table 2, and table 3 are the literature review that contains related work to the given Online shopper's purchasing intention dataset of different kernels and notebooks as stated in the introduction.

Kageyama (2020) has only performed model in Bayesian optimization and acquired a better accuracy rate comparatively the model was carried out in train and test split data only. Logistic regression was carried out by quite many of the researchers namely Vignesh Prakash (2020), Swapnil Bhange(2020), Daewoongjun(2020), and Oscar Matias Torros(2020) they have all obtained an accuracy rate in the range of 0.8 to 0.88 which is close. The highest accuracy in the table is obtained by random forest with an accuracy rate of 0.91 which is considered as the best model by some researchers namely Henry Sue(2020), Beth Morrison(2020), Daewoongjun(2020), Oscar Matias Torros for this particular dataset. A different classification model extra tree which is a group technique that is obtained from decorrelated decision tree termed as a forest is derived to do classification. Naïve Bayes, Multilayer perceptron, and Neuraral network classifier are other

uncommon models in the given table that is explored by Henry Sue (2020), Aurelia Sui (2020), and Oscar Matias Torros (2020) respectively. Swapnil Bhange(2020) performed a Support Vector Machine(SVM) and decision tree using SMOTE and without SMOTE to determine the best model with greater accuracy. XGBoost model obtained from Annette Catherine Paul has the lowest accuracy rate has the result was obtained from test data which had lesser and not sufficient data another same model was obtained by Tushar Vij (2020) that contained lower base and variance error then all other models put together. K-Nearest Neighbour is another model performed by Daewoongjun (2020) and Saurabh Gupta (2020) which is a decent accuracy rate.

In conclusion, all researchers have performed various exploratory data analyses using visualization techniques such as histogram, count plot, box plot, correlation plot, and trendline plot are plotted by all researchers. Another analysis such as statistical analysis, univariate analysis, bivariate analysis, multivariate analysis, outlier analysis, and PCA analysis is derived. Data pre-processing such as searching for null or missing values, one-hot encoding, label encoding, splitting the data into train and split, strange value detecting, scaling data, and MICE detection are performed.

2.4 Summary

From the above literature survey, the dataset of Online Shopper's Purchasing Intention with 122330 unique values and 18 attributes 11 classification models were developed by the researchers using the dataset along with the highest accuracy rate where the target variable is revenue is has value of true or false that belongs to categorical variable are as follows:

CLASSIFICATION MODELS	ACCURACY	AUTHORS
• **Bayesian optimization-**	ACC:0.902	Kageyama (2020)
• **Logistic regression**	ACC:0.8-.0.88	Vignesh Prakash (2020) Swapnil Bhange(2020) Daewoongjun(2020) Oscar Matias Torros(2020)
• **Random forest**	ACC:0.91-0.85	Henry Sue(2020) Beth Morrison (2020) Daewoongjun(2020) Oscar Matias Torros(2020)

• **Support Vector Machine**	ACC:0.91-0.84	Swapnil Bhange(2020) Daewoongjun(2020)
• **Naïve bayes**	ACC:0.84	Henry Sue (2020)
• **XGBoost**	ACC:0.9-0.87	Tushar Vij (2020) Beth Morrison (2020)
• **Decision tree**	ACC:0.84-0.218	Annette Catherine Paul(2020) Swapnil Bhange(2020)
• **Extra trees**	ACC:0.895	Henry Sue (2020)
• **Neural network**	ACC:0.88	Oscar Matias Torros (2020)
• **Multilayer Perceptron**	ACC:0.89	Aurelia Sui (2020)
• **K-Nearest Neighbour**	ACC:0.88-0.86	Saurabh Gupta(2020) Daewoongjun(2020)

Table 4 Summary of related works.

Table 4 addresses various classification models along with the authors. From the table, it is seen that random forest and logistic regression are the common classification model attempted by many. Thus, in related works, exploratory data analysis are focused and necessary analysis are made. Data pre-processing is also carried out to do the classification. Data splitting of train and test data are performed only by the following researchers Kageyama (2020), Beth Morrison (2020), and Annette Catherine Paul (2020). Thus, the literature review helps to determine the gap in the research and build the scope that could be carried out in the project that is stated in the next section.

2.4.1 Gap Analysis

In all the previous works research has used many models to some of the attributes of the data, but still, there is not any single work with all available models. The following objects are yet to be identified.

1) Can prediction of revenue increase be made clearer and more accurate?

2) Can comparison of 6 different classification approach be obtained?
3) Can be data be made more accurate using under sampling?

2.4.2 Scope

While obtaining the objective the scope of the research is limited to Performing various classification models such as random forest, decision tree, support vector machine, Naïve Bayes, K-nearest neighbor, and artificial neural network.

Perform data segmentation such as stratification and PCA model and analyzing the models. Perform EDA on model using different graphs and analysis.

3. Methods and Dataset

3.1. Dataset Description

3.1.1 Introduction

The data set provided had features which are more or less related to the purchases of the users. All features are mentioned below with explanations.

The data set provided for model making has a total entry of "12330" with "18" features. Among these features 9 features are numerical, continuous, and distinct, and 9 are categorical including the target feature Revenue.

The categorical variables in the data are:

Special Day, Month, Operating Systems, Browser, Region, Traffic Type, Visitor Type, and Weekend.

The numerical variables in the data are:

Administrative, Administrative Duration, Informational, Information Duration, Product Related, Product Relation duration, Bounce rates, Exit rates, and Page Values.

3.1.2 Description of Features

Numerical Features in the Dataset:

Feature Name	Data Type	Feature Description
Administrative	Integer	Number of different pages visited related to the administrative concerns of the website
Administrative Duration	Integer	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Integer	Number of different pages visited related to the information of the website and other useful contents of the website
Informational Duration	Integer	Total amount of time (in seconds) spent by the visitor on informational pages
Product Related	Integer	Number of different pages visited related to different products of the website.
Product Related Duration	Integer	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce Rate	Float	Average bounce rate value of the pages visited by the visitor
Exit Rate	Float	Average exit rate value of the pages visited by the visitor
Page Value	Float	Page Value is the average value for a page that a user visited before making a transaction.
Special Day	Float	The "Special Day" feature indicates the closeness of the site visiting time to a specific special day

Categorical Features in the Dataset:

Feature Name	Data Type	Feature Description
Browser	Integer	ID of bowsers from which the session took place.
Region	Integer	ID of Regions from which the session took place.
Traffic Type	Integer	ID of different types of sources from which the users landed on the website.
Visitor Type	String	Visitor type as "New Visitor", "Returning Visitor" and "Other"
Weekend	Boolean	Whether the session was on a weekend or not.
Operating Systems	Integer	Operating system of the visitor
Month	Boolean	Month value of the visit date
Revenue (Target Variable)	Boolean	Whether the user contributed to the revenue by purchasing or not.

Machine learning methods

There are many classification algorithms with the help of the available dataset six machine learning algorithm are performed with under sampling and cross verification. The machine learning classification models used are:

- Decision Tree.
- Random forest
- K-Nearest Neighbor (KNN)
- Artificial Neural Network (ANN)
- Naïve bayes classification
- Support Vector Machine (SVM)

Random forest is one of the supervised machine learning algorithms that uses ensemble technique. It is a technique that combines all algorithm of previous version to produce the desired output. Random forest can be used both in classification and regression models, while using in classification the new entry of data is classified based on each tree in the forest. Random forest is a combination of large number of decision tree. The best advantage is that is can be used for category and numerical data but the training time is comparatively more in this algorithm.

Artificial neural networks have a structure that is close to that of biological neural networks and are designed to replicate neural networks in the human brain. The human brain is a highly dynamic, nonlinear network of billions of closely connected neurons with trillions of synapses. Dendrites, axons, cell bodies, synapses, soma, and nucleus are the key components of a neural network. Similarly, in ANN there are collection of nodes which are called as artificial neurons that is helpful in processing information.

Naïve bayes is another classification algorithm in machine learning that uses Bayes theorem and has the presumption of independence for predictor. Predictor independence reflects is assessing each of the variable dependently. Naïve bayes algorithm is a simple algorithm that outperforms other algorithms. Naïve bayes algorithm are very helpful in multi-class and it is equally efficient with categorical data.

Performance validation Method

The Evaluation of the models will be done by performing the F Score ("Micro") metric. First the F score is defined as follows:

F_{score} = 2 * (precision * recall) / (precision + recall)

Where precision and recall are:

Recall=**TP/(TP+FN)**

Precision=**TP/(TP+FP)**

The values are obtained through confusion matrix, that is shown as below.

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

confusion matrix

Which states,

- TP is true positive that is the number of instances that are positive.
- FN is false negative that is the number of instances that are positive but predicted as negative.
- TN is true negative that is the number of instances that are negative.
- FP is false positive that is the number of instances that are negative but predicted as positive.

3.2 Dataset Preparation

3.2.1 Data Collection

Data is collected from the Kaggle dataset that is derived from the following source:

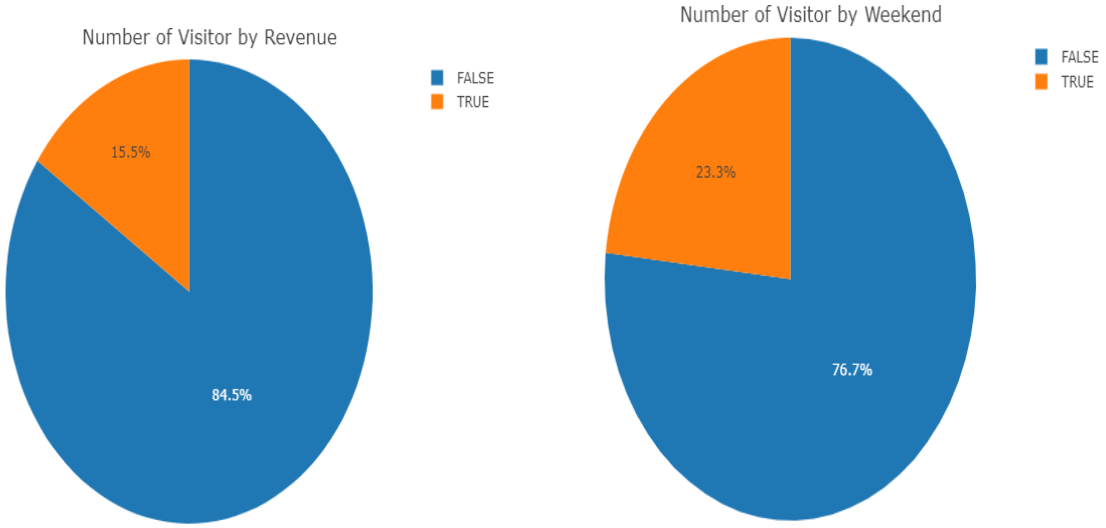
C. Okan Sakar Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey

Yomi Kastro Inveon Information Technologies Consultancy and Trade, 34335 Istanbul, Turkey*

3.2.2 Data Exploration

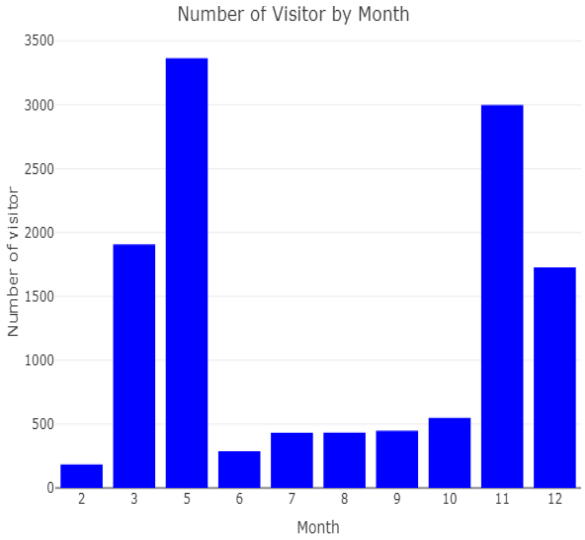
Data exploration is done through different types of graphs and charts to explore each category feature with all type of data analysis. Pie chart are derived for revenue and weekend category to the number of visitors and the result is as follows:

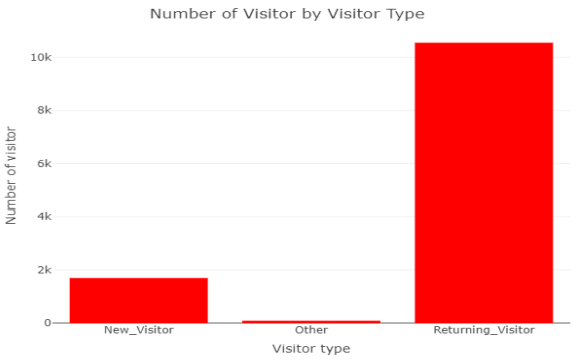
Pie chart for Number of visitors by revenue



The analysis from the graph is the number of customers turning positive revenue is 1908 that is 15.5% and number of not turning revenue are 10422 that is 84.5% . The rate of number of customers not turning revenue is very high comparatively. The number of visitors by weekend is 2868 that is 23.3% and number of not turning revenue are 9462 that is 76.7%.

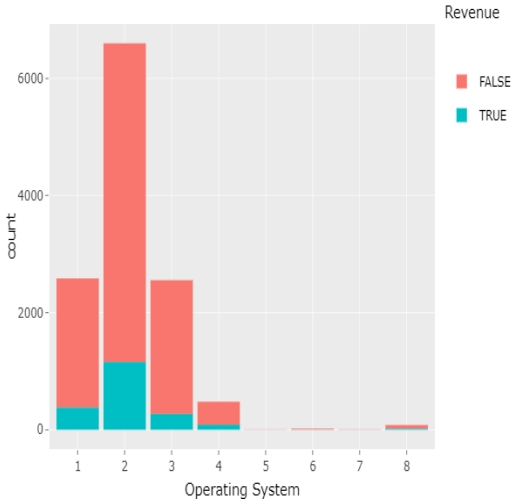
Bar chart is obtained for the number of visitors by month and visitor type.

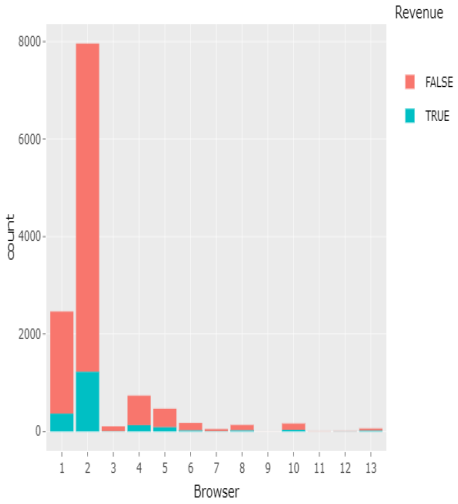


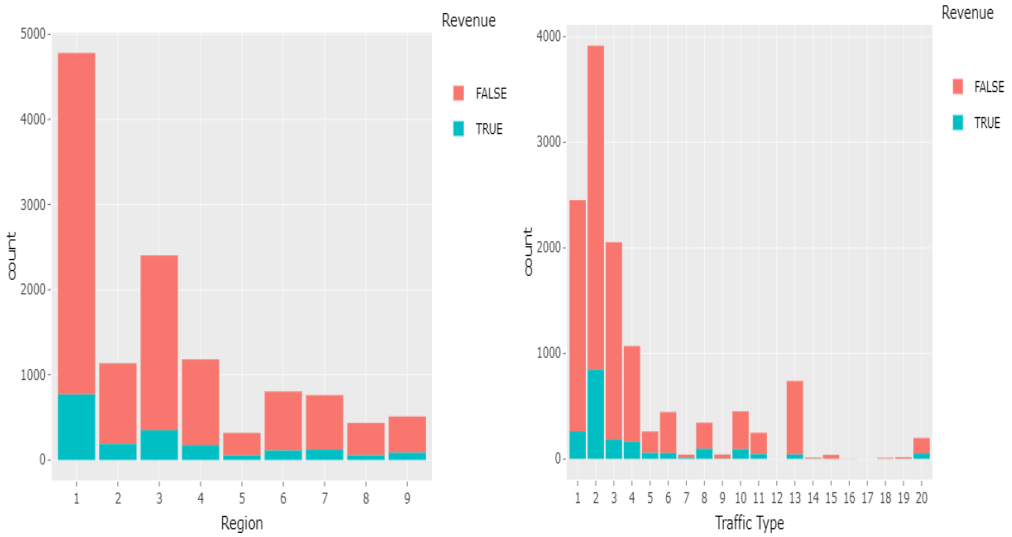


From the graph in month 1 and 4 that is January and April there are no data available. The maximum sales happen in month of March and minimum sales occurs in moth of February. From the visitor type chart returning visitor are maximum compared to new visitor and other visitors.

Bar chart for specific columns such as operating system, browsers, region and traffic type turning revenue are obtained and studied.



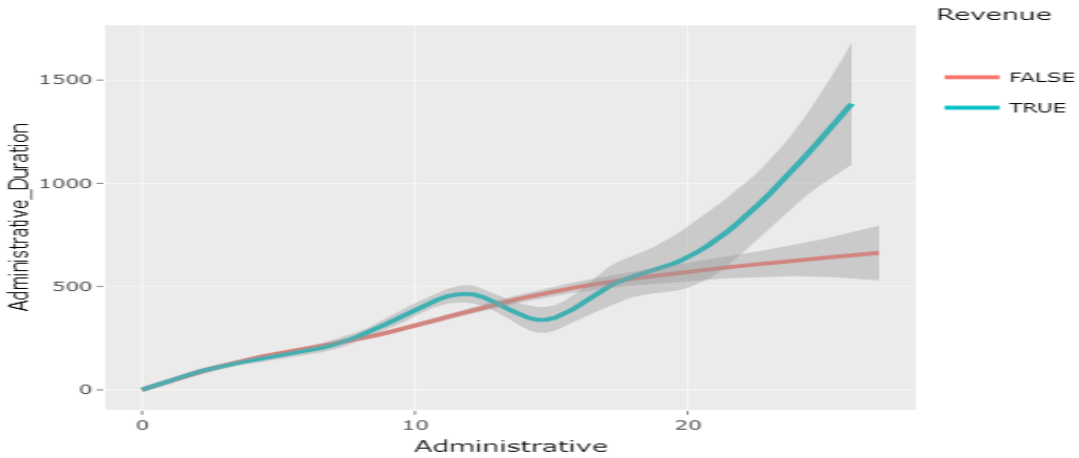




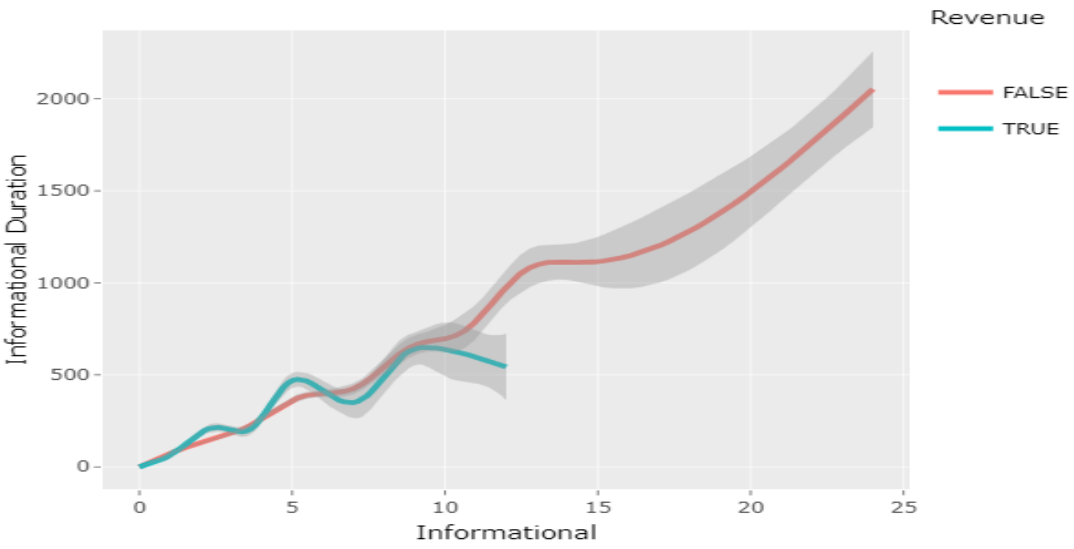
From the above four graphs it is seen that all four factors region, traffic type, operating system, and browser contribution towards revenue turning true factor is very less and many zero value is also observed.

QQ plot is plotted for three sectors with each of its duration which are compared with the revenue factor. The three categories are :

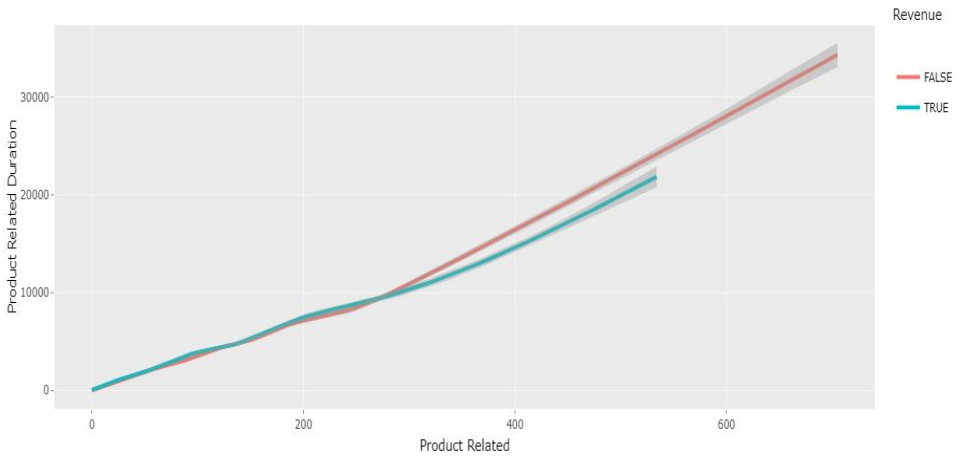
- Administrative and Administrative duration.



- Information vs Informational duration.

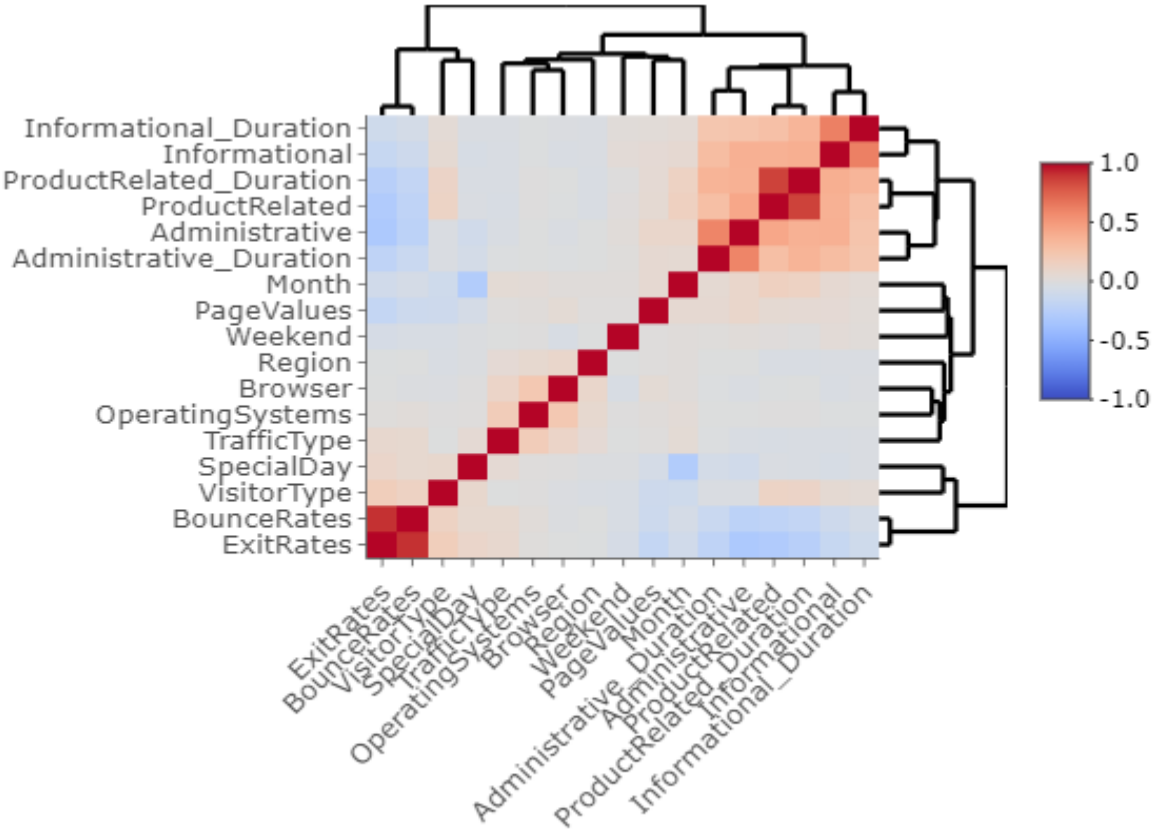


- Product related vs Product related duration.

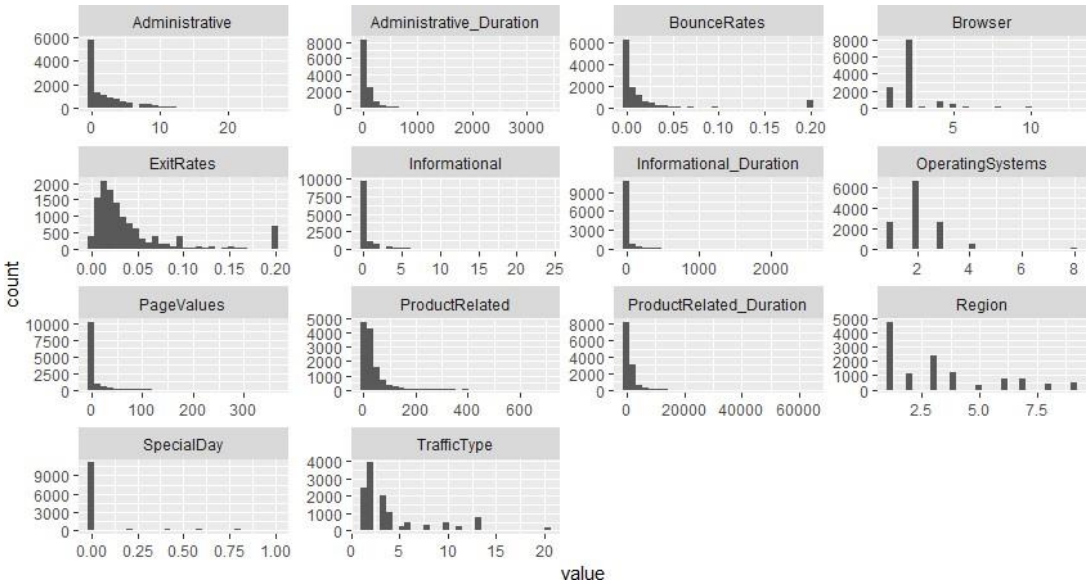


From the graph it is seen that product and informational true revenue are skewed right that is the value is not available on right. In administrative duration as the count increases the revenue increases indicating a good sign.

Heat map of correlation matrix is obtained as follows:



Histogram for each feature:



From the plot it is seen that there are many zero values for in each variable and some values maybe the outliers. Normalization is the solution to prevent any disturbance for the machine learning algorithms.

3.2.3 Data Splitting

After data splitting data are divided into two sets training and test data where training data consist of 70% and test data consist of 30% of the preprocessed data. K fold cross validation is used for data splitting where data are converted into sample data using random splitting of five samples. In each sample data are split into training and test data.

3.2.4 Data Cleaning
Duplicate and Missing Value

Duplicate data needed to be deleted in the dataset as it will result in wastage of time and cost of the analysis. It might also be the reason of false analysis in case of many duplicate data. For finding the duplicate data instance the function duplicated () is used to solve the problem. This function checks the duplicate in each instance instead of multiple instances so that it replaces many duplicate data to one data or instance.

```
# Defines same instance. If there is any same instance, It returns True.

isDuplicated <- duplicated(online_shopper)

duplicatedValues <- online_shopper[isDuplicated, ] #Same instance' values

num_duplicatedRow <- nrow(duplicatedValues) #Number of same instances

duplicated_online_shopper <- online_shopper[!duplicated(online_shopper), ] #Diffrent instance' values

num_datasetRow <- nrow(duplicated_online_shopper) #Number of distinct instances
```

This program executes the output result of 125 observation that are duplicated. Now, the dataset does not have any duplicated value.

Missing of the dataset is to be obtained to clean the dataset. Null value or no value in any of the instances are obtained using is.na function. The missing data is carried out in the dataset after removing duplicate data. The R program for the above-mentioned statement is as follows:

```
is.na(duplicated_online_shopper)

apply(is.na(duplicated_online_shopper),2,sum)

which (is.na(duplicated_online_shopper)) # Which one is NA?

duplicated_online_shopper[! complete.cases(duplicated_online_shopper),] #The function complete.cases() returns a logical vector indicating which cases are complete.  *
```

The result shows there are no missing value or NA in the whole dataset.

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

3.2.5 Data Transformation

Dataset has 18 descriptive features. These descriptive features include 3 features categorical except the target feature. Normalization technique was applied for converting descriptive feature' values in a dataset to common a scale without breaking the difference in range of values. But, before normalization was done, categorical features in the dataset were converted dummy attributes. Library 'dummy' was used for this process. No changes were made to the target feature in this use. target feature has been converted to dummy attribute in required models. Target feature has been converted to dummy attribute in required models.

```
dummy_onlineShoppers<- dummy.data.frame(duplicated_online_shopper,
names =
c("Month","VisitorType","Weekend"),omit.constants=FALSE,dummy.cla
sses  =getOption("dummy.classes"))
```

These three categorical features were converted into binary form by using function dummy (). In this way, all of the features in the dataset became continuous features. Normalization techniques were applied in the created new dataset. There are two normalization techniques: range normalization and Z -score normalization. The range normalization technique guarantees all features will have the same scale but does not handle outliers well. But, the Z-score normalization technique, handles outliers but does not produce normalized data with the same scale. Due to these results, the range normalization technique was used in this data set for having more consistent results.

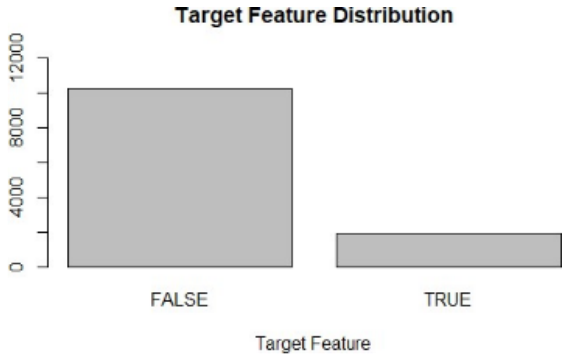
```
prepared_onlineShopper<as.data.frame(apply(dummy_onlineShoppers[,
1:num_descriptiveFeat ure],  2,  function(x)  (x -  min(x))/(max(x)-
min(x)))) prepared_onlineShopper[, 'Revenue']  <-
as.factor(dummy_onlineShoppers[, 'Revenue'])
```

3.2.6 Data sampling

There are 2 types in target feature of dataset that are chosen. These types are true and false values. Distributions of target feature values in the data set have been analysed. Results are shown below.

```
FALSE  TRUE
10297  1908
```

Distribution matrix



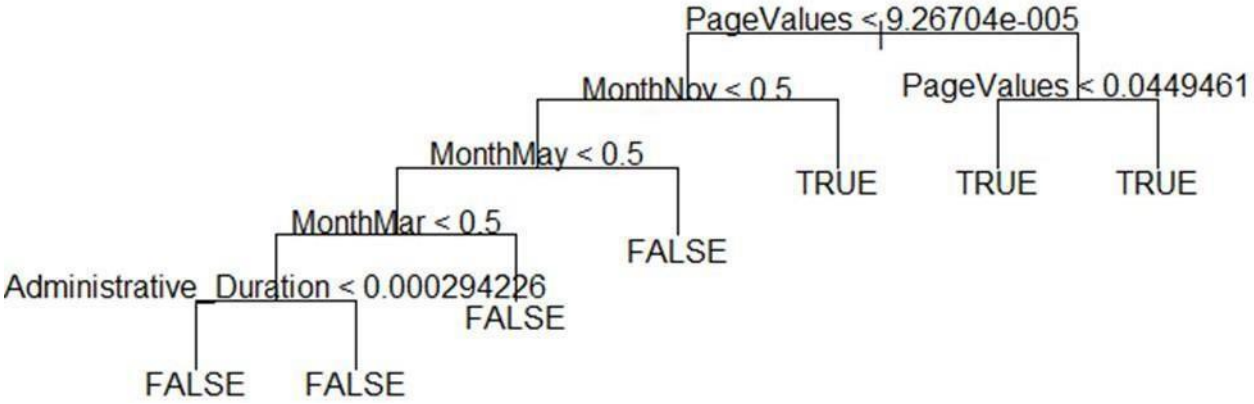
Distribution bar plot

In obtained results, irregular distribution is to be seen. Under-sampling is applied to be equal distribution for2 types of value in a dataset. The number of true is 1908. The number of falseis10297. So, 1908 instances that have target feature is false are chosen randomly. The remaining data has not been used. So, under-sampling has been applied two twice to use different instances. But the result has not changed much.

4 Chapter 4: Model Implementation and validation

4.1 Decision Tree

Datasets were trained with a decision tree and the classification tree shown below is created.



Four features were just used when creating a decision tree although there are thirty features. Other features do not influence training. Besides total path of the tree is 3.28.

All metrics of datasets and means are as follows.

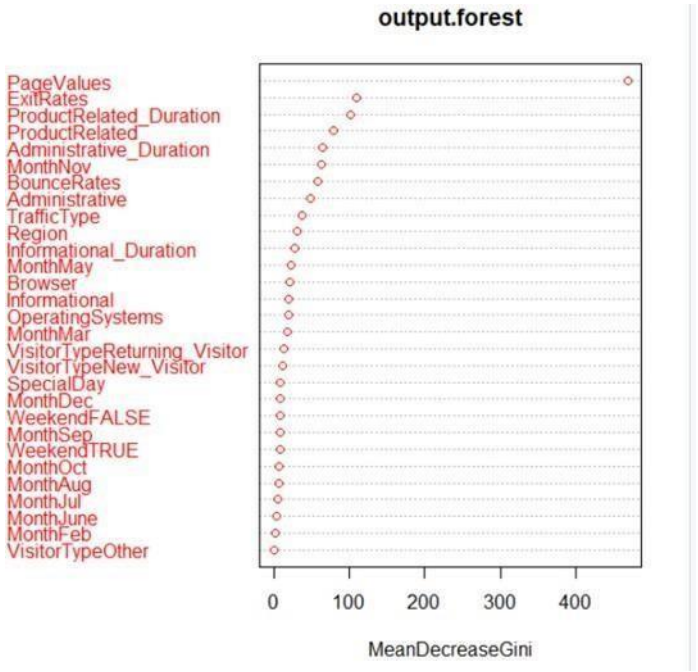
Decision tree	Accuracy	Precision	Recall	Fscore	AUC	Runtime(s)	F/F	F/T	T/F	T/T
Sample1	0.8829	0.7902	0.9260	0.8527	0.8953	0.05686	388	31	103	623
Sample2	0.8777	0.7849	0.9192	0.8468	0.8664	0.051	387	34	106	618
Sample3	0.8672	0.7651	0.9130	0.8326	0.8549	0.0548	378	36	116	615
Sample4	0.8637	0.8993	0.8172	0.8329	0.8820	0.0568	389	87	69	600
Sample5	0.8689	0.7958	0.8792	0.8354	0.8740	0.0568	386	53	99	607
Mean	0.8720	0.8070	0.8909	0.84	0.8745	0.0552				

Table 5 Decision tree output

4.2 Random Forest

Hyperparameters influence the performance of random forest algorithms and especially mtry and ntrees parameters are very popular in random forest. Mtry is the count of variables that are obtained randomly sampled in each split as a candidate. Mtry has a default value that is different for classification and regression. So firstly, optimum mtry was selected as 5. Because the decision tree algorithm uses only 5 features and selects greater than 5 is senseless. The number of trees was selected as 500 because the runtime is too high when we choose the number of trees too large. Datasets were trained with random forest. After training a random forest, the importance of each predictor is calculated by using the import function. The outcome value is affected by the variables that have high importance. The other variables that have low importance can be discarded in the model.

Importance of each predictor is showed at the following graphic.



According to graphic, the first value is highly significant on the model when examined the graphic. Besides ExitRates and ProductRelated_Duration is significant on the model but Values close to 0 like MonthFeb, Visitor_TypeOther, MonthJune are not significant on the model and this feature should drop from the model.

All metrics of datasets and means are as follows.

random	Accuracy	Precision	Recall	Fscore	AUC	Runtime(s)	F/F	F/T	T/F	T/T	mtry
Sample1	0.9030	0.8370	0.9298	0.8796	0.8948	2.599	411	31	80	623	5
Sample2	0.8917	0.8438	0.8984	0.8702	0.8858	3.28	416	47	77	605	5
Sample3	0.8925	0.8299	0.9131	0.8695	0.8850	3.3535	400	37	94	614	5
Sample4	0.8873	0.8449	0.8694	0.8571	0.8802	3.4908	387	58	71	629	5
Sample5	0.8847	0.8412	0.8812	0.8607	0.8789	3.0444	408	55	77	605	5
Real	0.8918	0.8393	0.8983	0.8674	0.8849	3.1535					

Table 6 Random forest output

4.3 K-Nearest Neighbour

KNN algorithm is applied 100 different times in the dataset. This number of repetitions is defined according to the number of instances in the dataset. The same test and train sets are applying 100 different times with KNN algorithms. The best k value is analyzed according to different evaluation metrics and is to visualize with the plot.

The evaluation metric values for the alliteration of KNN algorithm are shown in a table. The average of metric values for 100 iterations has been calculated. Run time has been calculated for training data.

#Fold	Accuracy	Precision	Recall	F-score	AUC	F/F	F/T	T/F	T/T	Run Time (secs)
1	0.6112664	0.4986151	0.55217	0.5190288	0.597228	253	190	238	464	0.13212
2	0.6208	0.4817	0.5717	0.5147	0.6038	244	175	249	477	0.12691
3	0.6248	0.4918	0.5766	0.5277	0.6088	246	178	248	473	0.12731
4	0.6165	0.5285	0.5202	0.5157	0.6019	248	214	210	473	0.1274016
5	0.6376	0.491	0.587	0.5252	0.6182	246	155	239	505	0.1332205

Table 7 KNN output

When calculating the value of the metrics for the knn algorithm, 100 different k values have been calculated. Different accuracy values have been found according to these k values. Maximum accuracy has been found to determine the most efficient k value.

As a result of the analysis, the metrics values of the 5 iterations are below. These values have been determined by taking averages of all accuracies because of 5 iterations.

	Accuracy	Precision	Recall	F-score	AUC
Mean	0.6221432	0.498443	0.5614429	0.520507	0.520507

Averages of results and run times of evaluation metrics (with undersampling)

K value is important for this algorithm and this value has different values for different metrics. Metrics have been had a maximum k value in 100 iterations in knn algorithm. This k value and maximum metric values in each iteration are in the table that is below.

K-fold	K values	Max Accuracy	K values	Max Precision	K values	Max Recall	K values	Max Fscore	K values	Max AUC
1	98	0.628821	52	0.5173116	98	0.5753425	1	0.5203762	98	0.6144173
2	54	0.6401747	2	0.515213	54	0.599022	1	0.5292887	54	0.6227118
3	51	0.641048	1	0.5121457	55	0.6004843	1	0.5458468	51	0.6250163
4	88	0.631441	79	0.5611354	88	0.5378151	1	0.5329018	88	0.6193595
5	76	0.6567686	75	0.5154639	80	0.6167513	91	0.5150502	76	0.6374883

Table 8 KNN output without under sampling

In the results obtained from the table, the model applied to the under-sampling data has not shown an effective performance for all metrics. The algorithm has been tried for 100 neighbours, but it was analysed that the maximum results that are found in different k values have not shown good performance.

KNN algorithm has been applied dataset without under sampling. The results obtained in the modeling without Under sampling are shown in the table below.

	Accuracy	Precision	Recall	F-score	AUC	F/F	F/T	T/F	T/T	Run Time (secs)
1	0.8467	0.9922	0.8511	0.9158	0.5238	3088	559	5	10	1.200419

2	0.8444	0.9891	0.8509	0.9141	0.5259	3075	559	15	13	1.200419
3	0.8471	0.9895	0.8533	0.9158	0.5288	3086	554	10	12	1.160909
4	0.8466	0.9914	0.8513	0.9156	0.5302	3079	560	8	15	1.158105
5	0.8471	0.991	0.8525	0.9157	0.5229	3094	553	5	10	1.144377

Results and run times of evaluation metrics (without under sampling)

Data were analysed in 2 species such as with and without under sampling. As a result of this analysis, the rate of knowing the false target is increasing due to the high number of false in the data set. This causes the value of accuracy and other parameters to increase.

The results of the evaluation metric without under sampling because of 5 iterations are shown in the table.

	Accuracy	**Precision**	**Recall**	**F-score**	**AUC**
Mean	0.6221432	0.498443	0.5614429	0.520507	0.520507

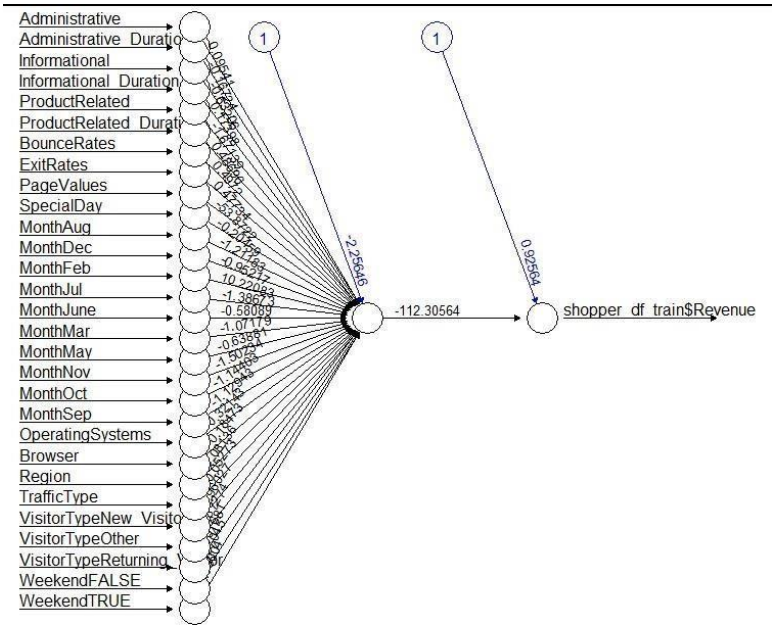
Results' means of evaluation metrics (without undersampling)

4.4 Artificial Neural Network

In the first part, data cleaning, normalization, categorical to dummy, processes are done, which means that our dataset is ready to apply ANN algorithm. In addition to it, to make probabilistic predictions (0 and 1) target feature should be represented with 0s and 1s.

Dataset is under-sampled, but to compare how efficient the process of under- sampling, we must perform ANN algorithm with a complete dataset. Train and Test splits are selected as 70% and 30% respectively.

Firstly, it is a good start to make 1 hidden layered neural network with one neuron to make a first impression.



There are 2920+384 = 3304 ***correctly predicted*** instances, however, 382 ***misclassified*** instances for this model and train/test samples.

Then, the ***accuracy*** obtained was nearly 90%, but there are other metrics to measure our models like precision, fscore, recall, and auc values.

After the model creation is done, metrics are obtained as Precision: 0.9332, Recall: 0.9401, F-Score: 0.9366 and AUC: 0.7938354. AUC is quite low for this accuracy value because of in balance ratio between the accuracy of True and False separately. The accuracy of False predictions as well, but the accuracy of True is not good, it is nearly 0.64%. To prevent this imbalanced situation, dataset must under sampled.

	Prediction: ***False***	Prediction: ***True***
Actual: ***False***	2920	186
Actual: ***True***	209	384

Accuracy	AUC	F-Score	Precision	Recall	F/F	F/T	T/F	T/T	Runtime(s)	*Hidden 1*	*Hidden 2*	*Hidden 3*
88.995	0.8831	0.8668	0.9011	0.835	410	81	45	609	4.7355 secs	*1*	-	-
89.869	0.8922	0.8779	0.9125	0.8458	417	76	40	612	3.1709 secs	*1*	-	-
88.558	0.8777	0.8608	0.9060	0.8198	405	89	42	609	3.5609 secs	*1*	-	-
89.08297	0.8843	0.8619	0.8725	0.8515	390	68	57	630	3.3630 secs	*1*	-	-
88.38428	0.8782	0.8599	0.8793	0.8412	408	77	56	604	4.0202 secs	*1*	-	-
89.08297	0.8841	0.868	0.9013	0.8371	411	80	45	609	42.4827 secs	*5*	-	-
78.51528	0.7750	0.7377	0.7775	0.7018	346	147	99	553	39.8284 secs	*5*	-	-
87.51092	0.8667	0.8477	0.8944	0.8057	398	96	47	604	11.5870 secs	*5*	-	-
77.99127	0.7718	0.7267	0.722	0.7314	335	123	129	558	3.17957 mins	*5*	-	-
88.38428	0.8782	0.8599	0.8793	0.8412	408	77	56	604	30.8176 secs	*5*	-	-
89.34498	0.8887	0.8732	0.8917	0.8554	420	71	51	603	3.00483 mins	*2*	*3*	-
89.95633	0.8935	0.8793	0.9109	0.8499	419	74	41	611	59.1360 secs	*2*	*3*	-
88.29694	0.8825	0.8571	0.9054	0.8137	387	107	55	596	1.2354 mins	*2*	*3*	-

Table 9 ANN Output

After the under-sampling process, our models are needed to re-computed. There are several hyperparameters to tune; learning rate, threshold, and stepmax, and all hyperparameters of models are set to the same values except hidden. The parameters hidden is used for determining the number of hidden layers and neurons in each.

$(a_x,b_y,c_z\ldots.m_n)$ *: xth layer has a neurons…, nth layer has m neurons*

To find out how many layers and neurons with each layer respectively, a couple of trials should calculate. Selected combinations of the layer and neurons are (1), (5), and (2,3). Obtained results of the combinations can analyse below.

First, you can see that there is critical change in the accuracy of true portions of predictions. It affects AUC in a good way, which means that under sampling version of the dataset must use to build a model. On the other hand, in the calculations there are no major changes between combinations by measurement metrics, hence one neuron with one hidden layer is sufficient for efficiency of computing performance.

In conclusion, with one neuron and one hidden layer, the requirement is meet properly, and there is no need to add a hidden layer or neurons to the model. Average of each metrics are Accuracy: 88.97785, Precision: 0.89428, Recall: 0.83866, F-Score: 0.86546, Runtime: 3.7701 and the most important observation is AUC: 0.8831. These values are compared with other algorithms in analysis part.

4.5 Naïve Bayes Classifier

Under sampling model has been modelled with naïve Bayes classifier algorithm. Function naiveBayes() has been used. This algorithm is to calculate a probability for each feature (Bayes Theorem) on this data set. After calculation, dataset that is divided for the test has been given to the model and the test dataset was analysed according to this probability. The model creates the conditional probability for each feature separately. We also have the prior probabilities which indicate the distribution of our data.

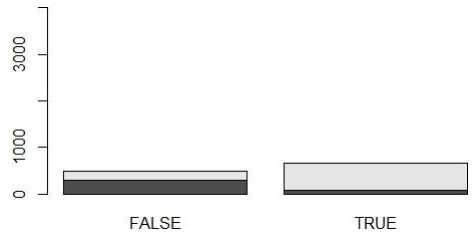
```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
    FALSE      TRUE
0.4223137 0.5776863

Conditional probabilities:
       X
Y           [,1]     [,2]
  FALSE 4948.303 2981.565
  TRUE  7054.016 3239.819

       Administrative
Y             [,1]      [,2]
  FALSE 0.07945889 0.1199765
  TRUE  0.13064977 0.1399996

       Administrative_Duration
Y             [,1]       [,2]
  FALSE 0.02144386 0.04603095
  TRUE  0.03702901 0.06315896

       Informational
Y             [,1]       [,2]
  FALSE 0.01769356 0.04659872
  TRUE  0.03415965 0.06554594

       Informational_Duration
Y             [,1]       [,2]
  FALSE 0.01271515 0.05709263
  TRUE  0.02353357 0.07084040

       ProductRelated
Y             [,1]       [,2]
  FALSE 0.03970751 0.05428380
  TRUE  0.06855302 0.08553329

       ProductRelated_Duration
Y             [,1]       [,2]
  FALSE 0.01604951 0.02481135
  TRUE  0.02926409 0.03739325

       BounceRates
Y             [,1]       [,2]
  FALSE 0.11138832 0.23833627
  TRUE  0.02650823 0.06152321

       ExitRates
Y             [,1]       [,2]
  FALSE 0.22323242 0.23988054
  TRUE  0.09789671 0.08328147

       PageValues
Y             [,1]       [,2]
  FALSE 0.005347226 0.02708661
  TRUE  0.074981538 0.09527485

       SpecialDay
Y             [,1]      [,2]
  FALSE 0.08191489 0.2248751
  TRUE  0.02488658 0.1278485

       MonthAug
Y             [,1]      [,2]
  FALSE 0.03546099 0.1850239
  TRUE  0.04471808 0.2067512
```

To analyse the performance of data in dataset that was calculated with conditional probability has been applied to predict () function. So, the result of the test data in the model has been shown as a matrix. In the first iteration, the result of data that is predicted is below.

```
          Actual
Prediction FALSE TRUE
     FALSE   302   84
     TRUE    189  570
```

Naïve Bayes Classifier Confusion matrix

In this iteration, 84 instances of 386 instances that target feature are false to have been misclassified. At the same time, 189 instances of 759 instances that target feature are to true have been misclassified. Predictions of target features are visualized as follows.



Prediction of Naïve Bayes Classifiers Algorithm

The other metrics have been calculated in a dataset that was modelled, too. The results of analyse are in the below table. For each iteration, the results of all evaluation metrics and run time are defined in the table.

	Accuracy	Precision	Recall	F-score	AUC	F/F	F/T	T/F	T/T	Run Time (secs)
1	0.7616	0.6151	0.7824	0.69	0.743	302	84	189	570	0.090756
2	0.7703	0.5862	0.8305	0.69	0.75	289	59	204	593	0.1156609
3	0.7467	0.581	0.7757	0.66	0.73	287	83	207	568	0.08576894
4	0.7415	0.5328	0.7485	0.62	0.71	244	82	214	605	0.1233358
5	0.7581	0.5711	0.8006	0.6667	0.7333	277	69	208	591	0.114979

Table 10 Naive bayes output

Values and run time of all of metrics (with under sampling method)

After 5-fold cross validation, optimum values have been analysed by calculating averages of all metrics.

	Accuracy	Precision	Recall	F-score	AUC
Mean	0.75564	0.57724	0.78754	0.67134	0.73326

Mean values of all of metrics (with under sampling method)

Results of models that are worked in dataset which has not been under sampling are specified in below table. When analyse is done between two table, for naïve Bayes classifier algorithm, dataset that is applied under sampling has been seen that has high performance values more.

	Accuracy	Precision	Recall	F-score	AUC
Mean	0.6895685	0.6700398	0.9455473	0.7841801	0.7841801

Mean values of all of metrics (without under sampling method)

4.6 Discriminative Classifier – Support Vector Machine

Support vector machine linear, polynomial, and the radial kernel were applied to five under sampling datasets, respectively. Different hyper parameters were used to optimize our algorithms in svm.

The most common parameters are as follows:

- Degree: parameter needed for a kernel of type polynomial (default: 3)
- Gamma: parameter needed for all kernels except linear (default: 1/(data dimension))
- Cost: cost of constraints violation (default: 1) - it is the ‘C’-constant of the regularization term in the Lagrange formulation.

The value of gamma and C should not be extremely high because it leads to overfitting or it should not be very small (underfitting). Thus, the optimal value of C is to be chosen and Gamma to get a good fit.

4.6.1 Support Linear Kernel

Firstly, the datasets were trained with svm linear. A Tuned function is used to optimum cost value and it is found as 4. All metrics of datasets and means are as follows.

Svm/linear	Accuracy	Precision	Recall	Fscore	AUC	Runtime(s)	F/F	F/T	T/F	T/T	cost
Sample1	0.8532	0.8289	0.8289	0.8532	0.8502401	1.2712	407	84	84	570	4
Sample2	0.8471	0.8215	0.8231	0.8480	0.8440	1.9164	405	87	88	565	4
Sample3	0.8244	0.8187	0.7606	0.7941	0.8235	1.0262	375	118	83	569	4
Sample4	0.8393	0.8016	0.8215	0.8496	0.8347	1.2673	396	86	98	565	4
Sample5	0.8410	0.8288	0.8023	0.8274	0.8394	1.70	402	99	83	561	4
Mean	0.8410	0.8199	0.8072	0.8344	0.8383	1.4362					4

Table 11 SVM linear output

4.6.2 Support Polynomial Kernel

Firstly, optimum gamma and cost were found by using the tune function for svm polynomial, and datasets were trained with optimum gamma, cost and 3 different degrees.

```
tuned_parameters <- tune.svm(Revenue~., data
online_shopper_training,kernel="polynomial" ,gamma = 2^(-
5:-5), cost = 2^(-5:5))
```

All metrics of datasets and means are as follows.

Svm/polynomial	Accuracy	Precision	Recall	Fscore	AUC	Runtime(s)	F/F	F/T	T/F	T/T	gamma	cost	degree
Sample1	0.8524	0.8268	0.8285	0.8492	0.8492	0.6971	406	84	85	570	0.03125	32	1
Sample1	0.7816	0.7209	0.7580	0.7390	0.7740	1.2676	354	113	137	541	0.03125	32	2
Sample1	0.8183	0.7637	0.8029	0.7828	0.8115	0.7280	375	92	116	562	0.03125	32	3
Sample2	0.8445	0.8215	0.8181	0.8198	0.8417	0.6686	405	90	88	562	0.03125	16	1
Sample2	0.7720	0.7079	0.7489	0.7278	0.7642	1.0259	349	117	144	535	0.03125	16	2
Sample2	0.8183	0.7586	0.8077	0.7824	0.7824	0.8110	374	89	119	563	0.03125	16	3
Sample 3	0.8366	0.8016	0.8164	0.8089	0.8324	0.6680	396	89	98	562	0.03125	16	
Sample 3	0.7720	0.6943	0.7571	0.7243	0.7626	1.08212	343	110	151	541	0.03125	16	2
Sample 3	0.8157	0.7510	0.8082	0.7785	0.8079	0.7240	371	88	123	563	0.03125	16	1
Sample 4	0.8235	0.8187	0.7591	0.7878	0.8227	0.9414	375	119	83	568	0.03125	32	1
Sample 4	0.7563	0.7096	0.6996	0.6906	0.7485	1.2776	325	146	133	541	0.03125	32	2
Sample 4	0.7903	0.7445	0.7349	0.7396	0.7827	0.6721	341	123	117	564	0.03125	32	3
Sample 5	0.8410	0.8305	0.7988	0.8165	0.8425	0.7188	405	102	80	558	0.03125	32	
Sample 5	0.7729	0.6969	0.7494	0.7222	0.7628	1.0351	338	113	147	547	0.03125	32	
Sample 5	0.8139	0.7381	0.8063	0.7707	0.8039	0.7393	358	86	127	574	0.03125	32	
Mean	0.8072	0.8579	0.7795	0.7693	0.7992	0.8704							

Table 12 SVM Polynomial output

The highest values of all metrics are equal to degree 1. this data set is best represented by a linear model.

4.6.3 Support Radial Kernel

In the same way, data sets were trained with optimum gamma and cost by using svm radial kernel. All metrics of datasets and means are as follows.

Svm/radial	Accuracy	Precision	Recall	Fscore	AUC	Runtime(s)	F/F	F/T	T/F	T/T	gamma	cost
Sample1	0.8331	0.7963	0.8112	0.8036	0.8285	1.5378	391	91	100	563	0.03125	4
Sample2	0.8454	0.7971	0.8361	0.7824	0.8395	1.2586	393	77	100	575	0.03125	4
Sample3	0.8427	0.7854	0.8398	0.8117	0.8358	1.1300	388	74	106	577	0.03125	4
Sample4	0.8366	0.8050	0.7901	0.7978	0.8315	1.0920	369	98	89	589	0.03125	4
Sample5	0.8331	0.8	0.8024	0.8024	0.8287	1.1471	388	94	97	566	0.03125	4
Real	0.8331	0.7962	0.8159	0.7995	0.8328	1.2331						

Table 13 SVM Radial Output

5 ANALYSIS & RECOMMENDATION

5.1 Model experiment analysis:

All selected machine learning techniques are evaluated; information based, similarity based, error based, probabilistic approach, and discriminative classifier. To compare these models, there are several metrics: Accuracy, Precision, Recall, FScore, and AUC. It is a mistake to take only one metric as Accuracy, because there can be in balance accuracy between true/positive and false/positive values.

There is a comparison table of each algorithm below. These values are the average of each metric.

	Accuracy	Precision	Recall	F-Score	AUC	Runtime
Decision Tree	0.8720	0.8070	0.8909	0.84	0.8745	0.0552
KNN	0.6221432	0.498443	0.5614429	0.520507	0.520507	1.1902
RF	0.8918	0.8393	0.8983	0.8674	0.8849	3.1535
ANN	0.88.97785	0.89428	0.83866	0.86546	0.8831	3.7701
Naïve B.	0.75564	0.57724	0.78754	0.67134	0.73326	0.0974
SVM	0.8410	0.8199	0.8072	0.8344	0.8383	1.4362

Table 14 Output comparison

From the results obtained, it can be derived that KNN, and Naïve Bayes algorithms require less computational power than other algorithms. However, the average values of metrics are quite low than other models. It means that we should apply more complex computation than these techniques. It is seen that all metrics of random forest and ANN algorithms are quite high. Especially accuracy and AUC metrics are close together, but the random forest is a bit faster than ANN so random forest is better than others.

5.2 Related work comparison analysis

Related work of other authors and our experiment is compared as below:

SNO	Author	Pre-processing	Accuracy
Decision tree			
1	Annette Catherine Paul	Data splitting	0.21(under sample)
2	Our experiment	Data splitting, cleaning, under sampling, normalization	0.8720
Random forest			
1	Henry Sue	Label and one hot encoding, cleaning	0.90
2	Beth Morrison	Hyper parameter and splitting	0.91
3	Oscar Matias Torros	Data cleaning	0.91

4	Our experiment	Data splitting, cleaning, under sampling, normalization	0.89
KNN			
1	Saurabh Gupta	Data cleaning and standardization	0.88
2	Our experiment	Data splitting, cleaning, under sampling, normalization	0.622
SVM			
1	Swapnil Bhange	MICE and outlier analysis	0.84
2	Daewoongjun	Label encoder	0.84
3	Our experiment	Data splitting, cleaning, under sampling, normalization, PARAMETER TUNING	0.84
Naïve Bayes			
1	Henry Sue	Label and one hot encoding, cleaning	0.84
2	Our experiment	Data splitting, cleaning, under sampling, normalization	0.75
ANN			
1	Oscar Matias Torros	Data cleaning	0.88
2	Our experiment	Data splitting, cleaning, under sampling, normalization	0.88

Table 15 Related work comparison

From the above table, other authors have achieved more accuracy for the random forest model, KNN and naïve bayes whereas the same accuracy is obtained for other machine learning models such as ANN and SVM.

5.3 Recommendation

Based on the above results obtained, following recommendations are made:

- There are more imbalanced data in the dataset, the dataset with more proper data can be helpful in predicting more accurately.
- Another recommendation in dataset is more distinct details of the soppers that is more variables needed for better analysis.
- There must also be proper law implemented in each country to collect necessary details and other personal details must be collected with clause included. As the data collection

and processing of data may include customers personal detail which can lead to malpractice.

- As the computation is carried out automatically the information given for analyses must be filtered beforehand maybe with the use of encoding or other securities.
- Law enforcement is practiced for artificial intelligence in ecommerce yet certain country like Ukraine does not have any law for customer information.

6 Conclusion

The topic selected was based on e-commerce where real time customer data was obtained with several categorical and numerical data. A Detailed literature review was carried out using various works available in Kaggle. Pre-processing techniques were performed in the data such as data splitting including under sample, data cleaning, data sampling, data transformation, and data scaling. The data set was completely analysed through exploratory data analysis using various data visualisation methods of bar chart, pie chart, histogram, qq plot, and heat map correlation. Then various machine learning algorithms were performed in five samples of the dataset finally mean value is derived for each of the models. Parameter tuning was done in three methods of SVM model. A thorough analysis of the result was made, and related work was compared. PCA analysis was not performed which will be the future work of the project. Another improvement in the implementation includes stratifying the data has the dataset contains a greater number of imbalanced data. Performance was validated using an F score and accuracy. The confusion matrix was obtained for each of the five samples in each model. Finally, the random forest is the best approach that can be used to obtain the result for this dataset as it contains both categorical and numerical data. A clear understanding of EDA, pre-processing techniques, different machine learning algorithm, and their working are obtained from the project.

7 Reference

Annette Catherine Paul (2020) **Online Shoppers Behavior Prediction.** Available from https://www.kaggle.com/annettecatherinepaul/online-shoppers-behavior-prediction [Accesses: 22/3/2021].

Aurelia Sui (2020) **Online Shoppers Purchasing Intention** (online). Available from https://www.kaggle.com/yufengsui/online-shoppers-purchasing-intention/code [Accesses: 22/3/2021].

Beth Morrison (2020) Online_Shopper_Analysis. Available from https://www.kaggle.com/bethmorrison/online-shopper-analysis [Accesses: 22/3/2021].

Dae Woong Jun (2020) How Can We Convince More Customers To Buy?. (online). Available from https://www.kaggle.com/daewoongjun/how-can-we-convince-more-customers-to-buy [Accesses: 22/3/2021].

Herny Sue (2020) **Online Shoppers Intention UCI Machine Learning.** (online). Available from https://www.kaggle.com/henrysue/online-shoppers-intention [Accesses: 22/3/2021].

Herny Sue (2020) **Classifying Online Shopper Intention.** (online). Available from https://www.kaggle.com/henrysue/classifying-online-shopper-intention [Accesses: 22/3/2021].

Kageyama (2020) **[LGBM] Online Shopper's EDA and Classification.** (online). Available from https://www.kaggle.com/kageyama/lgbm-online-shopper-s-eda-and-classification [Accesses: 22/3/2021].

Kelleher, John (2015) **Fundamentals of Machine Learning for Predictive Data Analytics.** (online). Available from The MIT Press, 2015 {Accesses 7/5/2021].

Oscar Matias Torres (2020) **Clients Classification, Neural Networks explained.** (online). Available from **https://www.kaggle.com/oscarmatiastorres/clients-classification-neural-networks-explained** [Accesses: 22/3/2021].

Roshan Sharma (2020) **Online Shopper's Intention (online).** Available from **https://www.kaggle.com/roshansharma/online-shoppers-intention/code** [Accesses: 22/3/2021].

Saurabh Gupta (2020) **Online shopping-KNN-ROC.** (online). Available from **https://www.kaggle.com/saurabhgupta09/online-shopping-knn-roc** [Accesses: 22/3/2021].

Sheetal Sharma (2017) **Artificial Neural Network (ANN) in Machine Learning** (online). Available from **https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning** [Accesses: 8/5/2021].

Swapnil Bhange (2020) **Online Shoppers Intention EDA, ML, etc...** (online). Available from https://www.kaggle.com/swapnilbhange/online-shoppers-intention-eda-ml-etc [Accesses: 22/3/2021].

Tusharviji (2020) **Online Shopper's Intention – Classification.** (online). Available from https://www.kaggle.com/tusharvij/online-shopper-s-intention-classification [Accesses: 22/3/2021].

Vignesh Prakash (2020) **Online Shoppers Purchasing Intention (PCA, SMOTE).** (online). Available from https://www.kaggle.com/vigneshprakash/online-shoppers-purchasing-intention-pca-smote [Accesses: 22/3/2021].

Vincent Granville (2017) **Prediction Algorithms in One Picture** (online). Available from **https://www.datasciencecentral.com/profiles/blogs/prediction-algorithms-in-one-picture** [Accesses: 8/5/2021].

Yu Feng Sui (2020) Using MLP to Predict Online Purchasing Intention. Available from https://www.kaggle.com/yufengsui/using-mlp-to-predict-online-purchasing-intention [Accesses: 22/3/2021].