PROJECT REPORT

ONLINE SHOPPING PURCHASING INTENT

IE 7275 Data Mining in Engineering

Prof. Srinivasan Radhakrishnan

(Group - 4)

Malavika Krishnan

Dhruv Bhalala

krishnan.ma@northeastern.edu

bhalala.d@northeastern.edu

Effort contributed by Malavika Krishnan	**: 50%**
Effort contributed by Dhruv Bhalala	**: 50%**
Submission Date	**: 20/08/2021**

Problem Setting:

Recently, a new trend has emerged among virtual shopping environments so that potential visitors are identified at the time they are browsing the website. By contrast to the near-real time model, the advantage behind that is to avoid the high risk of losing users once disconnected from the online store. Indeed, in such a model, we imitate an experienced salesperson who struggles to retain potential visitors by providing a range of customized marketing actions which are likely to encourage efficient purchases. So, the real statement is to build machine learning models to identify user behaviour patterns and determine the likelihood of purchase based on the given features for online purchase customer data.

Problem Definition:

The recent boom of online shoppers has opened up a new dimension to the business sector. The luxury of exploring items, getting the desired items and purchasing them from the comfort of one's own home has lured a lot of customers leading to a huge number of online transactions. To be a great seller, one needs to know the customer's preferences and intentions in order to recommend items to make an efficient purchase. The customer's purchase intention can be predicted by analyzing the history of the customers. Online shopping behavior's data has been analyzed to build a classification model to predict this. Different classification models have been analyzed to predict whether a customer, visiting web pages of an online shop, will eventually end up with a purchase or not.

Data Sources:

The data is collected from **UCI Machine Learning repository** which is a database of Machine Learning problems maintained by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine.

Data Source: Online Shoppers Purchasing Intention Dataset Data Set

Data Description:

Online Shoppers Purchasing Intention" – Dataset has 12330 number of records and 18 attributes. The dataset consists of 10 numerical and 8 categorical attributes. These records were the sessions in which shoppers spent some time on pages of administrative, informational, product related, and so on. Bounce Rate, Exit Rate, and Page Value attributes were collected from Google Analytics. Moreover, there are other variables such as special day, month, browser, region, traffic and visitor type, and so on. Weekend and Revenue variables have boolean values. Additionally, around 85% of user samples did not come up with shopping.

The following table depicts all the attributes in the data set along with their definitions.

1) PAGE TYPE AND TRAFFIC CHARACTERISTICS:

Administrative	Page used by account users and administrative users to add, edit, manage payment gateways of user accounts.	Integer
Administrative Duration	Total time spent on administrative page in user's session.	Continuous/ Float
Informational	Page showing general user informations.	Integer
Informational Duration	Total time spent on informational page in user's session.	Continuous/ Float
Product Related	Page type visited showing product related information	Integer
Product Related Duration	Total time spent on Product related page in user's session.	Continuous/ Float

2) GOOGLE ANALYTICS METRICS

Bounce Rates	Percentage of users who enter from a page and then leave.	Continuous/ Float
Exit Rates	Feature for a specific web page is calculated as for all page views to that page, the percentage of users leaving the site.	Continuous/ Float
Page Values	The average value for a page that the user visited before completing an e-commerce transaction.	Continuous/ Float

3) TIME CHARACTERISTICS

Special Day	Indicates the closeness of a site visiting time to a specific special day.	Continuous/ Float
Month	Month of the year.	Integer
Weekend	Boolean value indicating whether the date of visit is weekend/weekday.	Boolean/Binary

4) USER INTERFACE SYSTEMS

Operating System	Operating system(s) used by the user in the session.	Integer
Browser	Bowser(s) used by users in the session.	Integer

5) GEOGRAPHIC CHARACTERISTICS

Region	Value indicating the region of the user.	Integer
Traffic type	Value indicating the type of user activity.	Integer
Visitor Type	Value indicating type of visitor whether returning, new, or other types.	Integer

6) TARGET VARIABLE

Revenue	Value returns true if purchase occurred, else false.	Boolean/ Binary

Data Exploration:

In our dataset, we have 12330 shopping sessions and 18 attributes. Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates and PageValues are numeric in nature and others are categorical. Our response variable will be the Revenue. The dataset has no missing values.

Out of all sessions, around 15% sessions were positive cases ending with shopping. Moreover, around three fourth out of all sessions' revenue were generated on weekdays. After removing outliers from the dataset, now we have a class of interest consisting around 12% of all sessions.

Bar Plots:

Bar Plot-1: This is a count plot that shows the distribution of the percentage of customers who have shopped both on the weekdays and weekends. The trend shows that customers who purchased during the weekdays contributed to a higher revenue than the customers who purchased during the weekend.

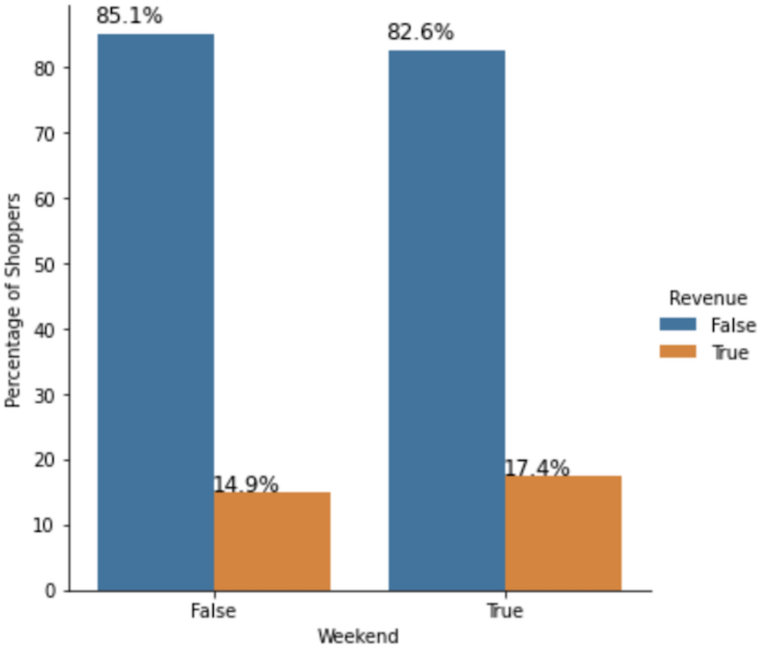


Fig 1: Percentage of shoppers on weekend/weekdays

Bar Plot-2: This graph depicts that the highest shopping sessions occurred in the month of November. In May month, it has almost half of the graph's peak value. In January, February and April almost no shopping sessions occurred.

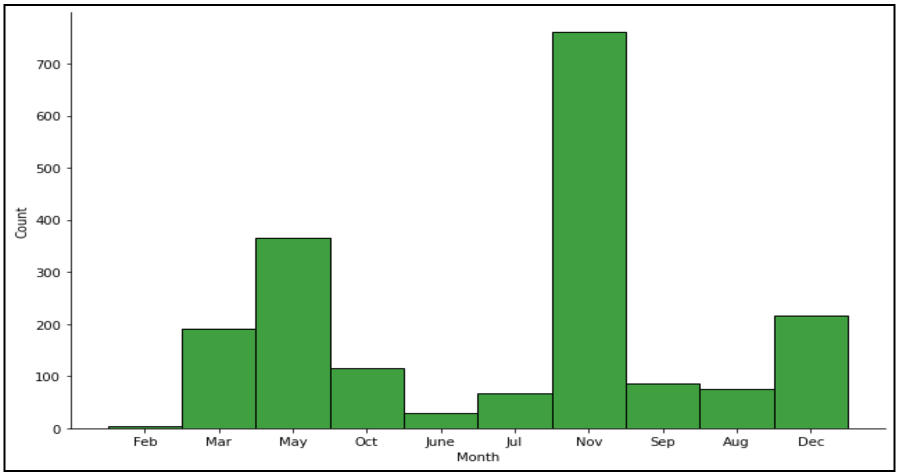


Fig 2: Bar plot of months when shopping occured

Bar Plot-3: We have six categories in SpecialDay feature and they are in numerical values of 0.0, 0.2, 0.4, 0.6, 0.8, 1.0. It can be seen from the bar plot that almost 50% of the records have value of zero. This feature might not add any value in our model building.

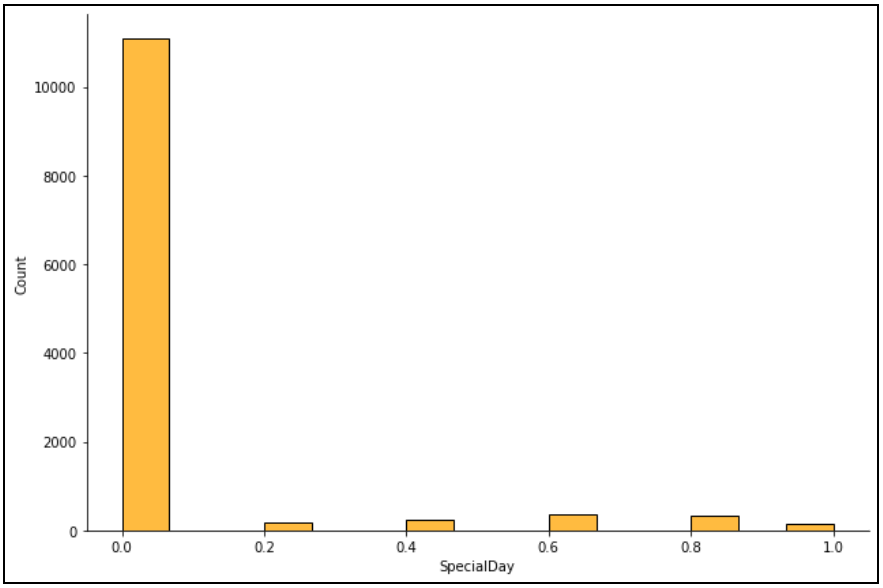
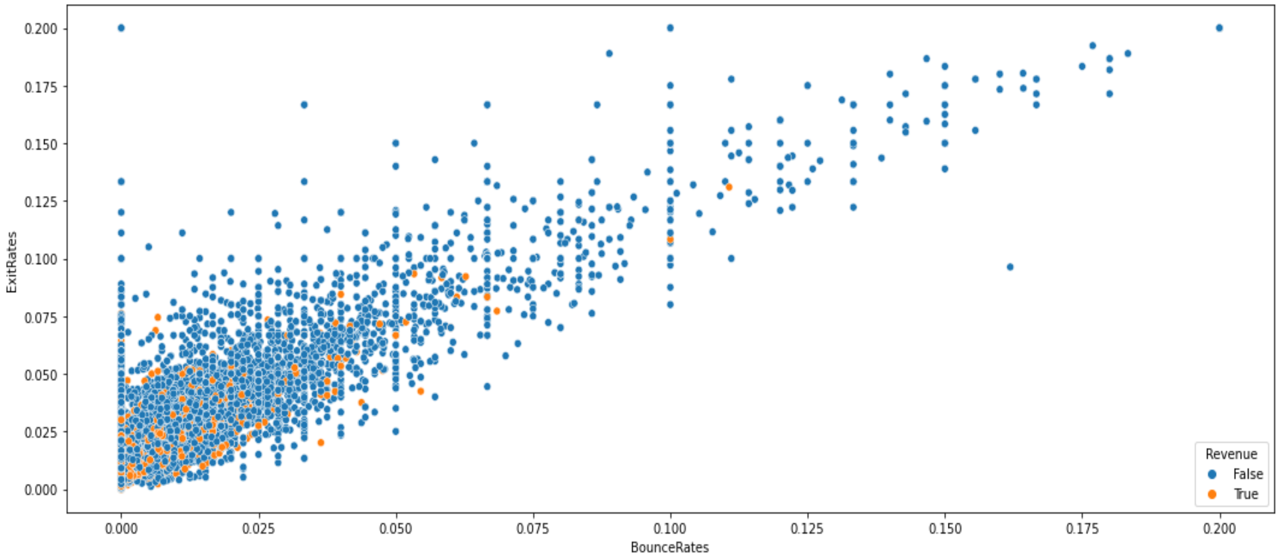


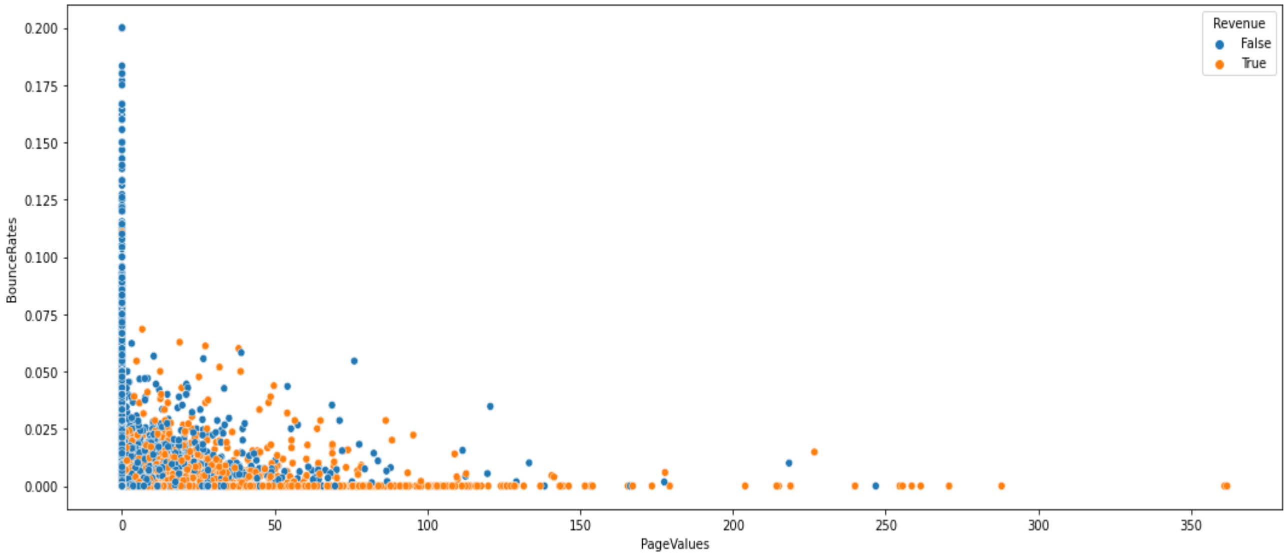
Fig 3: Bar plot of SpecialDay

Scatter Plots:

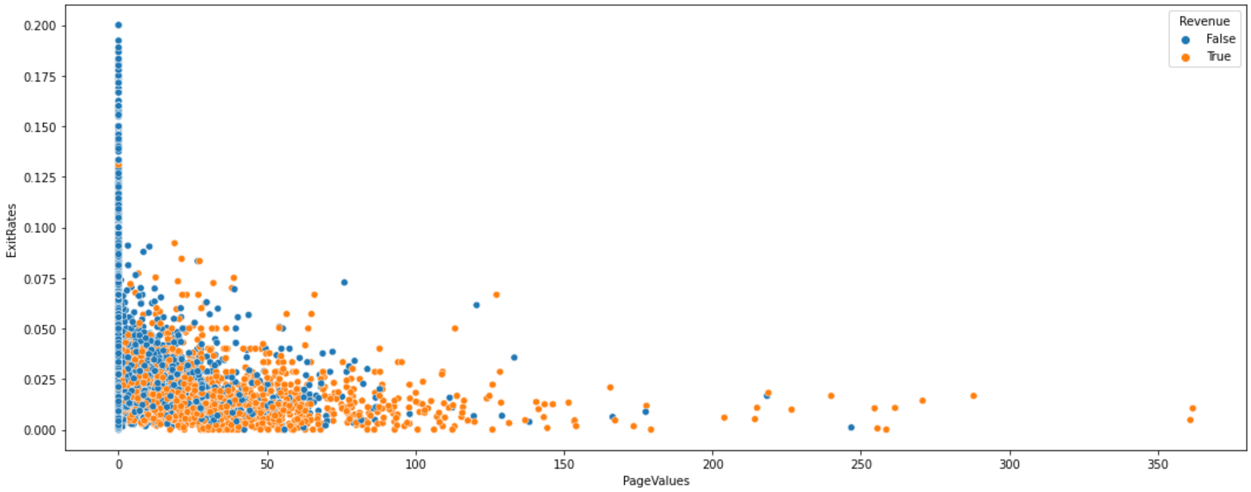
Scatter plot-1: Here, we see a good linear relationship in the plot. It also shows that revenue is mostly false and true value for revenue is concentrated between 0-1. Moreover, it can be seen that higher the exit rates and bounce rates, higher the chances of revenue being false.



Scatter plot-2: The plot shows a good non-linear relationship. Though there are concentrations of Bounce rates and Page values in certain areas, it seems that all page values towards the extremes translate into revenue equals true and all the Bounce rates translate into revenue equals false.



Scatter plot-3: This plot also shows a similar trend of strong nonlinear relationship. If exit rates are between 0-1 and page values between 0-max, there is a high chance that revenue could be equal to true.



Heat Map:

It can be seen from the heat map that ProductRelated and ProductRelated_Duration attributes are highly correlated. So are the other two variables- BounceRates and ExitRates. So we have dropped ProductRelated_Duration and ExitRates.

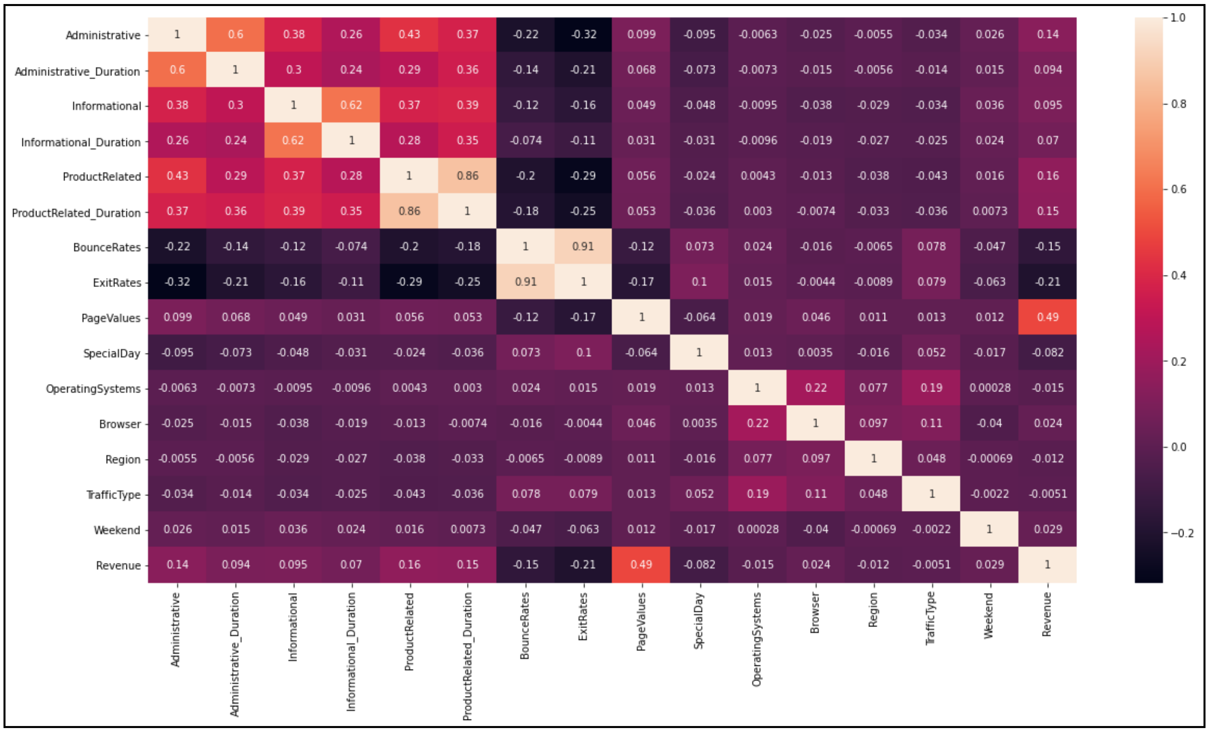


Fig 1: Heat Map

Box Plots:

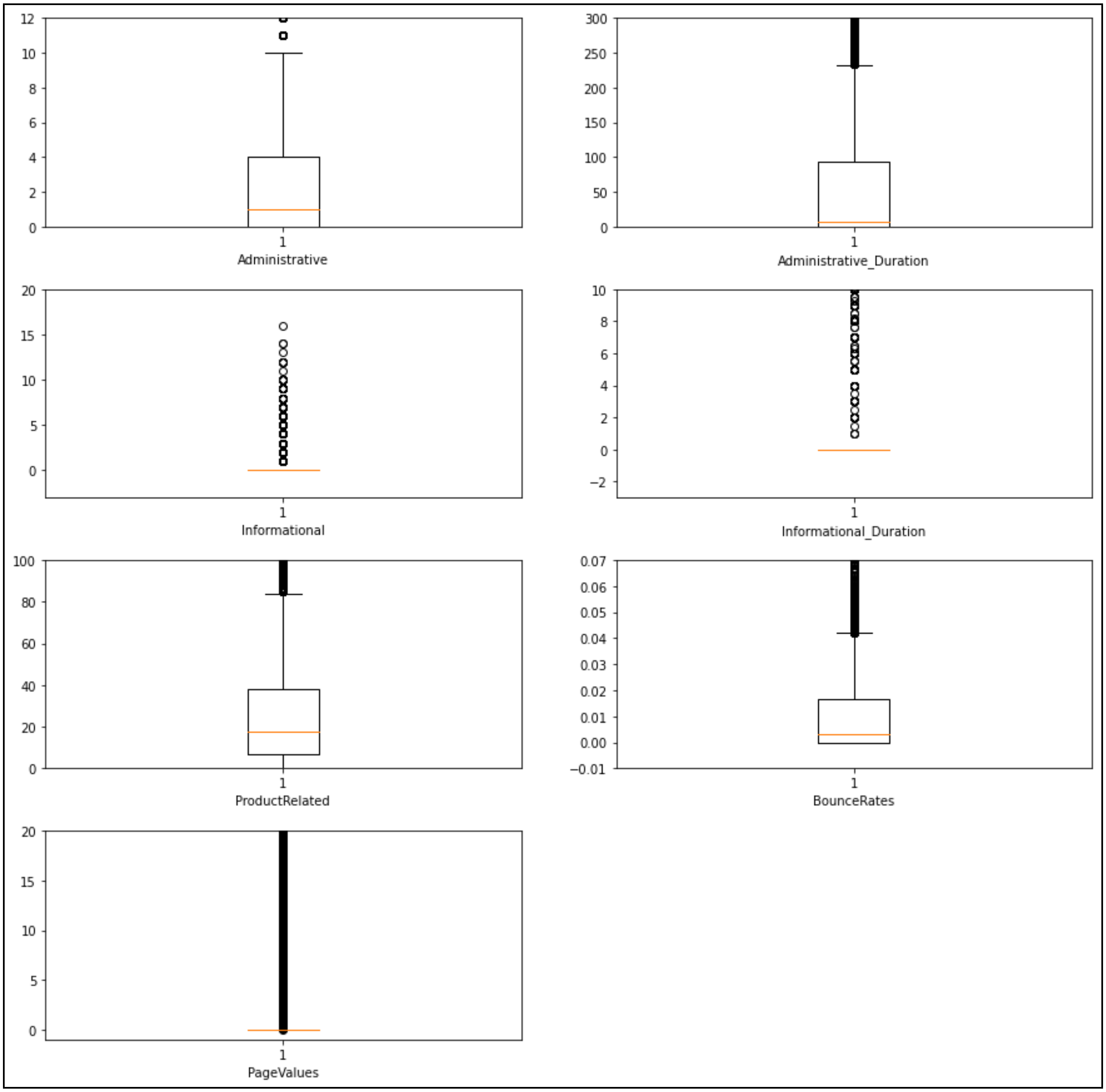


Fig 3: Box plots

From these plots, we observed that the Administrative, Administrative_Duration, ProductRelated and BounceRates attributes have left skewed distributions. It can been seen that outliers are large in numbers and we need to control them without changing the meaning of the data set altogether.

Density Plots:

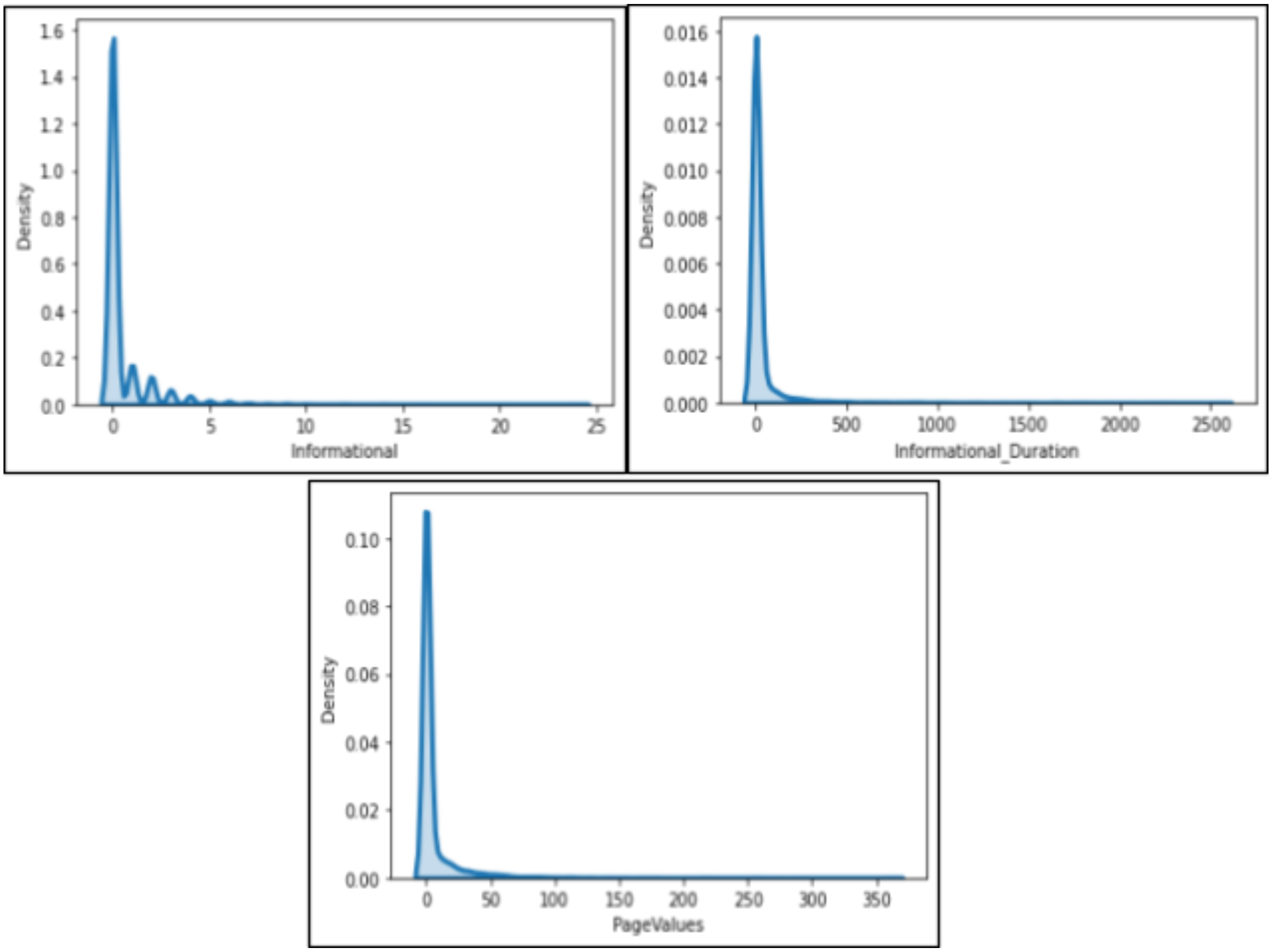


Fig 4: Density plots

For the attributes, Informational, Informational_Duration and PageValues, we have created density plots since distribution was not distinguishable for these with the box plot. It can be observed from the plots that the majority of the PageValues are concentrated in the zero region.

Data Mining Tasks:

1) Missing Value Imputation

The ‘Online Shoppers purchasing intention’ dataset was free of any missing values. So we never had to pre-process the data set to remove any of the same. However in case we encounter missing values, we would have opted to go with the following methods:

1) Imputation using mean/median/mode values
2) Randomised Imputation
3) Imputation using linear regression
4) Imputation using k-nn

2) Outlier Detection

As we know that outliers are anomalistic behaviour of a data point. We confirm from the box plots that there are many outliers and by removing or altering them we would introduce significant errors to our data. We have eliminated the outliers that are beyond three standard deviations (we checked with mean(+/-) standard deviation for each of the values).

3) One-Hot Encoding

One-hot encoding is the method where the integer encoded variable is removed and a new binary variable is added for each unique integer value. Here, we converted the nominal variables into (n-1) dummy variables and chose the most frequent categories. Moreover, the ordinal variables were label encoded.

4) Feature Scaling

Feature scaling is a method that is used to normalize or standardise the range of independent variables or features of data when there is a huge range difference within the data points. In our data processing, we have standardized our data and determined the distribution mean and standard deviation for each feature and calculated the new data point with mean at zero and standard deviation equals to one.

5) Dimension Reduction (PCA)

With the result set we obtained after standardisation, we performed dimensionality reduction. The most common method used for dimensionality reduction is Principal Component Analysis or PCA. The PCA performed data set was used to test the accuracy of the model, and was found to not show a better accuracy result and hence we continued to implement modelling techniques without the PCA performed data set.

Data Mining Models/Method:

In our project, we have our response variable as Revenue. Since it has a boolean datatype, the models are classification based. We have performed various model implementations such as Logistic Regression, Neural Networks, k-Nearest Neighbors, Decision Trees, Random Forest Trees and Boosted Trees. For evaluating the classification performance, we created a confusion matrix for analysing the accuracy, recall, and f1 scores associated with it. The interest class under consideration has a 15% proportion with the whole dataset and therefore, oversampling of the training set was done in order to improve accuracy.

1) K-Nearest Neighbors Algorithm:

First we deployed the k-NN method to our dataset. We split the dataset first into test and train. We plotted a graph with k vs Accuracy in order to get maximum accuracy when k is between 0 to 40. Moreover, we did trial and error method for test size and eventually got a maximum accuracy (0.8982) when k=11, test size is 20% and measurement attribute is Euclidean.

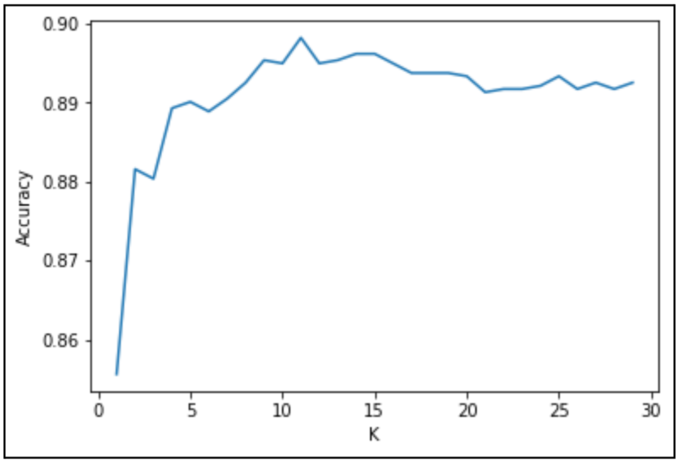
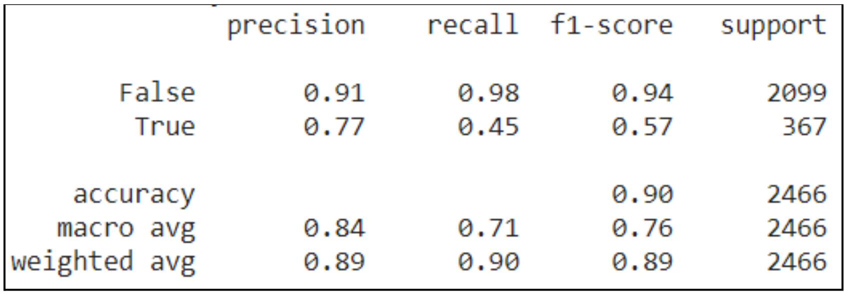


Fig 5: kNN Accuracy



```
              precision    recall  f1-score   support

       False       0.91      0.98      0.94      2099
        True       0.77      0.45      0.57       367

    accuracy                           0.90      2466
   macro avg       0.84      0.71      0.76      2466
weighted avg       0.89      0.90      0.89      2466
```

Fig 6: k-NN Classification Report

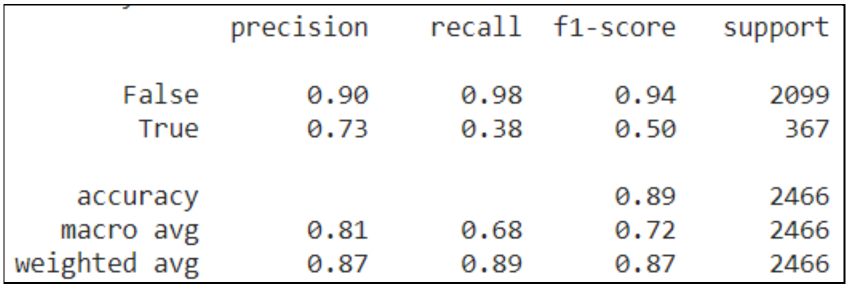
We also tried to implement a pipeline model in which we used standardization and PCA, but it did not increase the accuracy level significantly.

2) Introducing Oversampling method to test whether the accuracy will increase or not

We oversampled the training set using the SMOTE (Synthetic Minority Over-sampling Technique) method as our interest class with less proportion in number. Here we did not observe any improvement in the accuracy.

3) Logistic Regression

Our second approach is the Logistic Regression model. We went for the same test size (20%) And performed Logistic regression with 450 maximum iterations. Logistic Regression provided with a maximum accuracy of 88.64% (0.8864) with 'lbfgs' solver which gave a higher accuracy than with other solvers like Saga, Sag and Newton etc.



```
              precision    recall  f1-score   support

       False       0.90      0.98      0.94      2099
        True       0.73      0.38      0.50       367

    accuracy                           0.89      2466
   macro avg       0.81      0.68      0.72      2466
weighted avg       0.87      0.89      0.87      2466
```

Fig 7: Logistic Regression Classification Report

Furthermore, we implemented a similar pipeline model that was used in the kNN classifier and SMOTE oversampling method. In both the cases, there was no significant increase in the accuracy level.

4) Decision Tree

In this model, we used the same test size and plotted a one line chart which shows depth vs accuracy with a specific purity method.

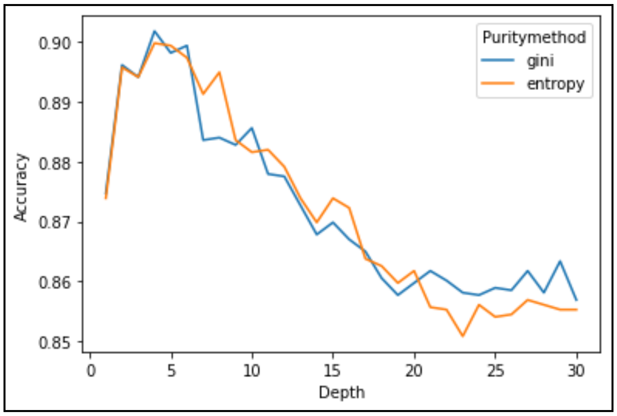
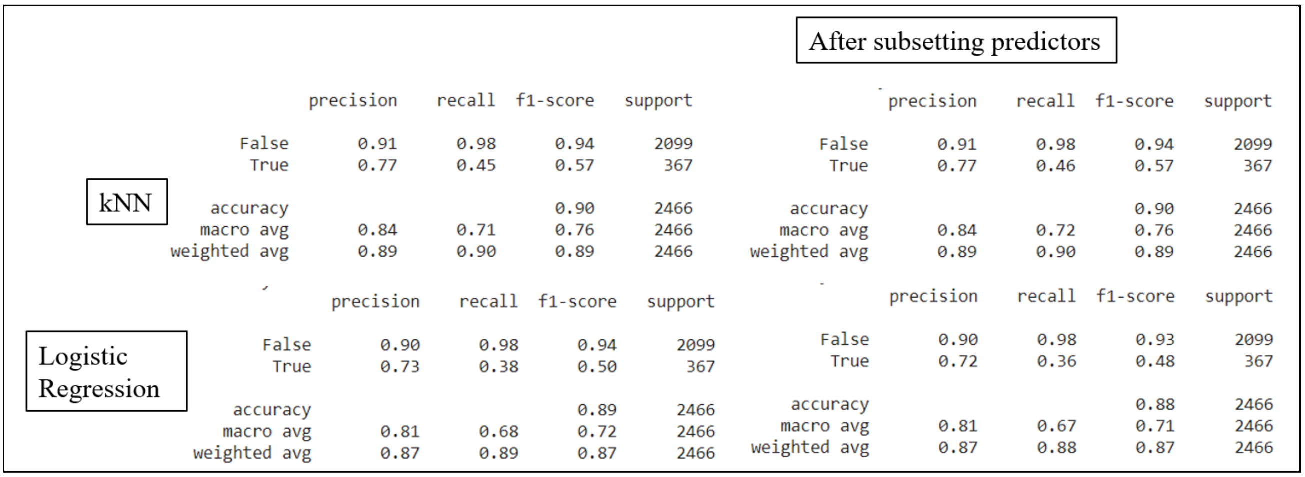


Fig 8: Depth vs Accuracy chart of Decision Tree

The maximum accuracy obtained was 90.18% (0.9018) when the maximum depth is 4 and the purity method is Gini. We also performed using the Random Forest classifier where we got a maximum accuracy of 90.02% (0.9002)..

Subsetting Predictors: The optimized predictors are Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, BounceRates, PageValues and SpecialDay. After subsetting the predictors, we carried out k-NN method and Logistic Regression in which we followed the same procedure as we contrived before and the results are as follows:



After subsetting predictors

kNN

```
              precision    recall  f1-score   support                    precision    recall  f1-score   support

       False       0.91      0.98      0.94      2099          False       0.91      0.98      0.94      2099
        True       0.77      0.45      0.57       367           True       0.77      0.46      0.57       367

    accuracy                           0.90      2466       accuracy                           0.90      2466
   macro avg       0.84      0.71      0.76      2466      macro avg       0.84      0.72      0.76      2466
weighted avg       0.89      0.90      0.89      2466   weighted avg       0.89      0.90      0.89      2466
```

Logistic Regression

```
              precision    recall  f1-score   support                    precision    recall  f1-score   support

       False       0.90      0.98      0.94      2099          False       0.90      0.98      0.93      2099
        True       0.73      0.38      0.50       367           True       0.72      0.36      0.48       367

    accuracy                           0.89      2466       accuracy                           0.88      2466
   macro avg       0.81      0.68      0.72      2466      macro avg       0.81      0.67      0.71      2466
weighted avg       0.87      0.89      0.87      2466   weighted avg       0.87      0.88      0.87      2466
```

Fig 9: Comparison result after subsetting predictors

5) Random Forest

For this model, we used the same test size of 20% and we optimized the number of estimators and obtained a maximum accuracy of 90% with 45 number of estimators. Below represents the plot of Accuracy vs Estimators.

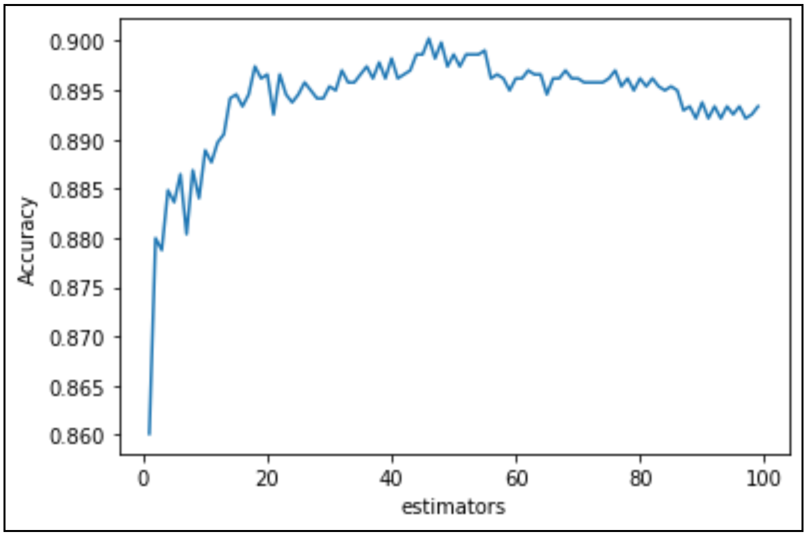


Fig 10: Accuracy vs Estimators for Random Forest

In this model, we got a lesser area under the curve as we can see from the below graph.

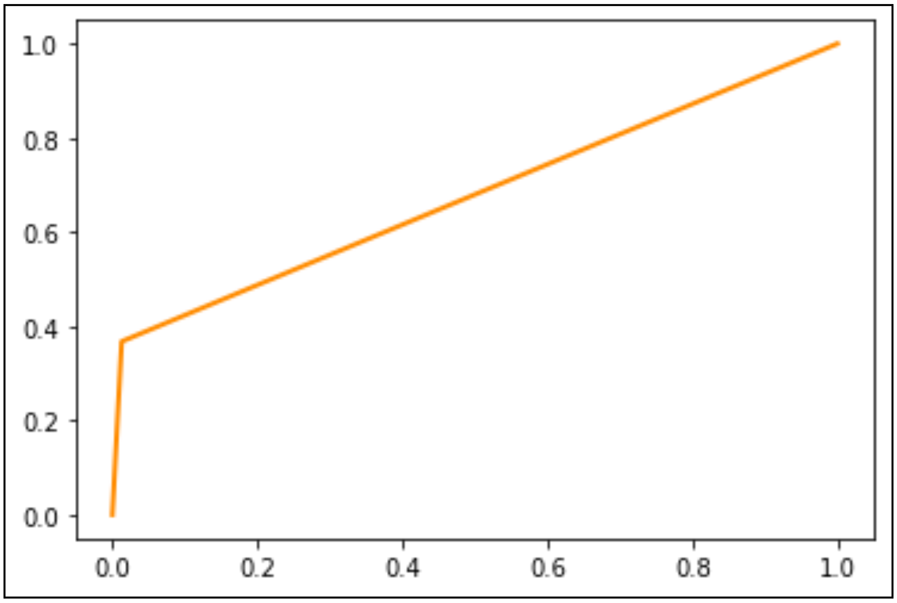
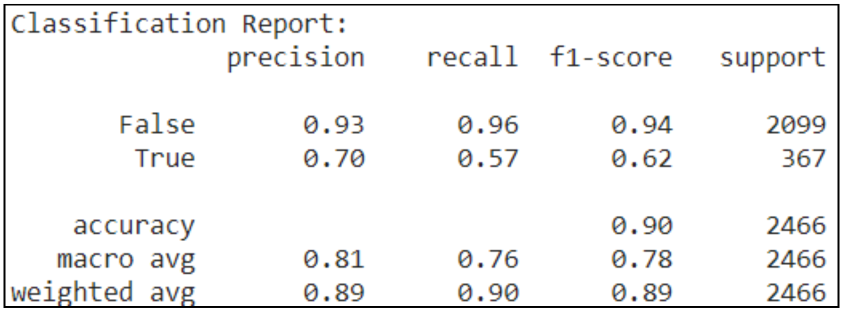


Fig 11: AUC for Random Forest

6) Neural Networks

We have implemented Neural Networks and with 450 maximum iterations, 0.001 learning rate and logistic transfer function, we obtained 0.8986 accuracy which is almost the same as what was obtained with most other models. The classification report is shown below:



```
Classification Report:
              precision    recall  f1-score   support

       False       0.93      0.96      0.94      2099
        True       0.70      0.57      0.62       367

    accuracy                           0.90      2466
   macro avg       0.81      0.76      0.78      2466
weighted avg       0.89      0.90      0.89      2466
```

Fig 12: Classification report for Neural Nets

The significant rise with the AUC curve is shown below:

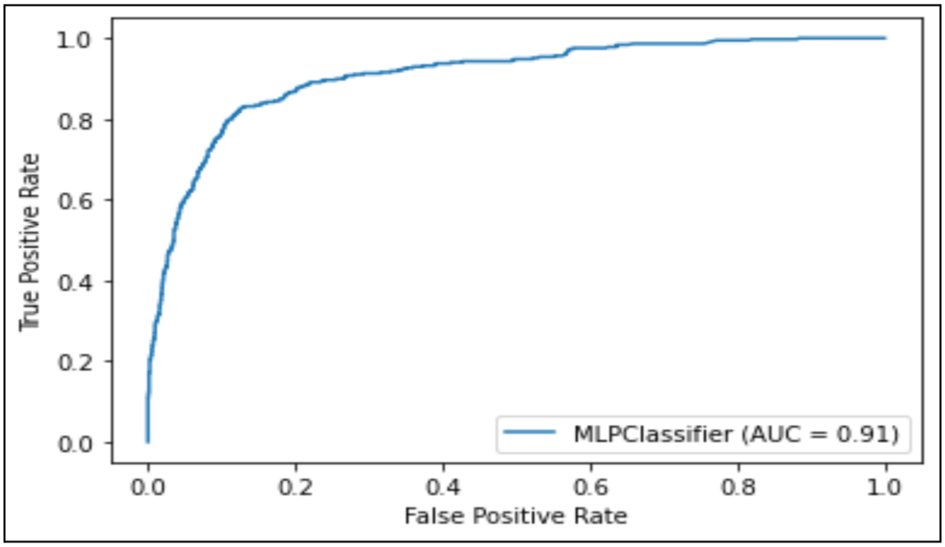
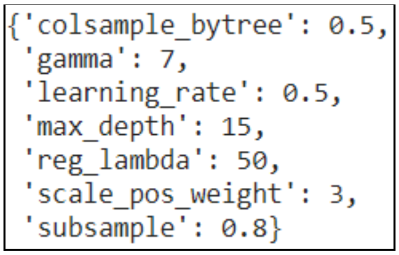


Fig 13: AUC for Neural Nets

In hyperparameter tuning feature, we obtained different learning rates (0.0001, 0.001, 0.01, 0.03, 1, 3, 6) and transfer functions ('identity', 'logistic', 'tanh', 'relu') and got a maximum accuracy of 90.06% (**0.9006)** when Learning rate = 0.0001 and transfer function = Logistic. Also, we tried to implement hyperparameter tuning gridsearch with cross validation method but it did not impact our accuracy at all. We got 0.8927 maximum accuracy.

7) XG Boosting

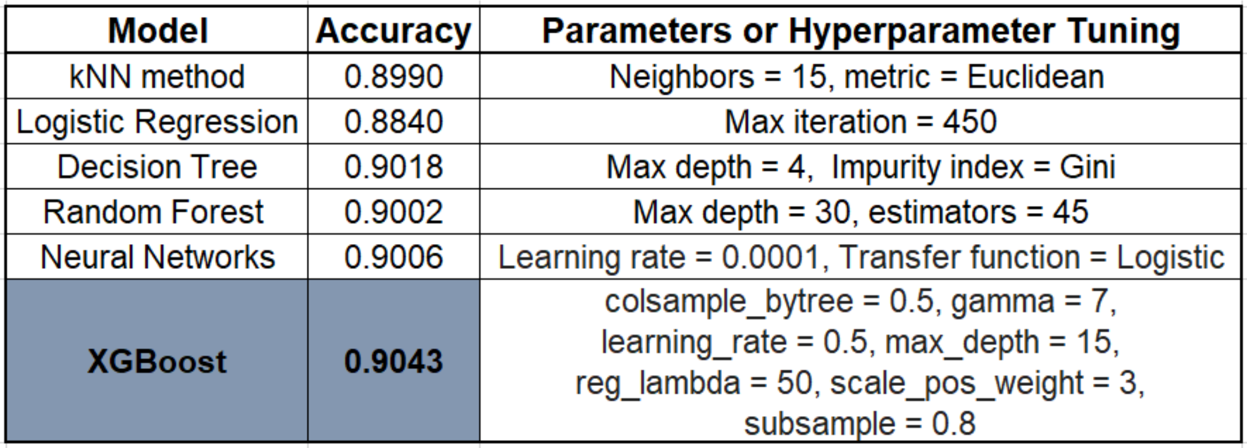
Initially, we implemented a basic XG boosting classifier and got a maximum accuracy of 0.8986. We also deployed two parametric models with different parameters like max depth, learning rate, gamma, lambda, positive weight, subsample and colsample bytree in hyperparameter tuning gridsearch with cross validation. These are the parameters in which we got a maximum accuracy of 90.043% (**0.9043)** which was fairly higher than what we obtained with all other models.



```
{'colsample_bytree': 0.5,
 'gamma': 7,
 'learning_rate': 0.5,
 'max_depth': 15,
 'reg_lambda': 50,
 'scale_pos_weight': 3,
 'subsample': 0.8}
```

Fig 14: Parameters for XG Boost model

Performance Evaluation:



Model	Accuracy	Parameters or Hyperparameter Tuning
kNN method	0.8990	Neighbors = 15, metric = Euclidean
Logistic Regression	0.8840	Max iteration = 450
Decision Tree	0.9018	Max depth = 4, Impurity index = Gini
Random Forest	0.9002	Max depth = 30, estimators = 45
Neural Networks	0.9006	Learning rate = 0.0001, Transfer function = Logistic
XGBoost	**0.9043**	colsample_bytree = 0.5, gamma = 7, learning_rate = 0.5, max_depth = 15, reg_lambda = 50, scale_pos_weight = 3, subsample = 0.8

Project Results:

From the EDA, the number of purchases in November and May is greater compared to the other months. There is a higher percentage of shoppers purchasing goods during the weekdays than on the weekends. Moreover, from the bar plot -3, we can see that though there are special days in a year, most purchases happened during the non-special days.

From the performance evaluation, it is observed that model implementation performed with XGBoost method provides the highest accuracy of 90.43% followed by Neural Network method giving 90.04% and Random Forest method giving 90.02%.

Impact of the Project Outcomes:

Since the number of purchases in November and May is greater compared to the other months. So we can therefore leverage this and obtain better turnouts by offering gift coupons , offers and special discounts to the customers. Bounce rates and exit rates depict negative influences on customers' behaviour. So we must be cautious of users with high exit rates or bounce rates.

We have a 90.43% accuracy result which means that there is a 90.43% certainty that the user will make an online purchase or not. Moreover, the model correctly classified revenue false with 93% precision, and instances of Revenue true with 70% precision. This means that we have a rather low false positive rate for Revenue false class and rather high false positive rate for revenue true class. The recall obtained is 0.96 and 0.57 which means the algorithm returned most of the relevant results and therefore, the model is good with identifying false negatives.

LIMITATIONS AND CHALLENGES: As we were performing various tasks on our data set, we encountered different challenges. There are limitations with this dataset since it focuses only on short-term user activity which shows whether a given user session will result in a purchase or not. But it is important to understand the fact that the purchase intent of a consumer may slowly build up over time, and may not instantaneously lead to a purchase. Moreover, this data set targeted user behavior on a single e-commerce platform alone, while users may use several different e-commerce platforms when deciding which product to purchase and where. Thus, what is missing from the picture is a **cross-platform analysis** of how customer purchase intent varies over time**.** To this end, it is important to contrast the population of purchasing users with the population of non- purchasing users, and then also identify how purchasers' online behavior changes over time from the norm as a result of impending purchases.