Predictive Analytics in Ecommerce

Moffat Kagiri Ngugi

R2106D12371591

Advanced Decision-Making Predictive Analytics & Decision Making

DS-7003

29th October 2023

Abstract

E-commerce enterprises seek to increase their revenues by enhancing customer conversion rates on their websites. This study employs predictive analytics and machine learning methodologies to create a model that predicts how a given customer makes a purchase decision during each user session based on a dataset from a case study site. Valuable insights are gleaned through exploratory data analysis of the customer behaviours. Random Forest and Logistic Regression models are used to predict customer conversion. They are then assessed using various metrics such as accuracy, AUC, precision, and recall. The Random Forest model stands out with its impressive performance, achieving an AUC of 0.92.

Table of Contents

Introduction

Ecommerce

Ecommerce is one of the hallmarks of the advancement of modern technology. It covers the application of highly advanced information technology through electronic customer relationship management systems, state of the art logistics operations, and automation in the fulfilment of orders. Leveraging these inventions of the 21st century, this industry has achieved exponential growth throughout the last two decades. Southern, 2017 observes that ecommerce continues to grow in all countries as the internet gains popularity in the less developed countries around the world. Ecommerce affords people and businesses the advantage of acquiring all the information they require to make their purchase decisions, as well as the convenience of making purchases remotely. Ecommerce firms also offer convenient alternatives for the delivery of commodities to the buyers. As a result, ecommerce is gradually taking over the market share held by many types of businesses ranging from motor sales to grocery stores.

Increased Data Volumes

Through the last 30 years, the worldwide web has grown from a government resource into a global resource which contributes to nearly every aspect of human life. This is one of the driving factors for the accelerated growth of ecommerce today (Huberman & Adamic, 1999). From a handful of nodes, the web has grown to include billions of devices around the world. Today, it is used by two thirds of the global population (Ring, 2023). With the ever-increasing numbers of devices and users now, the volume of data on the web is on an exponential growth trend. In addition, there are new and non-traditional sources of data on the web which generate data in a variety of formats which can be classified as structured or unstructured. In addition to this complexity, the data is not always stored in convenient or accessible locations on the web.

This creates a problem as many firms are not sufficiently equipped with the tools, skills, and finances required to take advantage of the data in the volumes and formats it comes in.

The large volumes of data available facilitate quick and effective market research to understand the trends in the consumption of various commodities. Ecommerce enterprises are, themselves, connected to the internet. This means that all interactions customers make with the ecommerce sites constitute invaluable data regarding the nature of customers and businesses seeking to use ecommerce, as well as the value they seek from the sites. This enriches the value of data analysis to ecommerce firms as they would gain insights into trends in the ecommerce industry as well as firm-specific information. With such insights, they would be able to identify areas for improvement such as cutting down on costs, improving customer satisfaction through more relevant suggestions and advertisements, and enhancing their efficiency through increased traffic and increased conversion of site visits into sales (Adnan, et al., 2011). Additionally, the firms can identify new opportunities for growth such as potential categories of products they can incorporate in their catalogues.

Objectives

This study evaluates a case study dataset from an ecommerce site. It examines the dataset in order to gain insights into the performance of the site with respect to different types of customers, different patterns in their interactions, and their purchase decisions. From these insights, the study reveals the varying likelihoods of customers in different categories to make the ‘buy now’ decision. It then applies predictive analytics to evaluate the customer conversion rate based on the dataset. The predictive models used determine whether or not customers make a purchase based on their interactions with the site and other variables in the data. The study uses cluster analysis to gain insight into the customer segments based on the data. It applies logistic regression and decision trees through a Random Forest model in order to predict the

likelihood of a customer making a purchase based on the features of the data provided which include the movement across the site's pages, the page types, page values, the customer's operating system, among others. It then evaluates these models' relative efficacies. Ultimately, the study seeks to provide recommendations on which model more accurately represents the likelihood of a customer making a purchase and how the site's performance can be improved based on these facts.

Literature Review

Predictive Analytics

Predictive analytics is a computational approach which seeks to identify trends in data and project their motion in the future to support decision making processes in business settings by providing foresight. It leverages statistical modelling, machine learning, and artificial intelligence. However, it is not a new approach to managing uncertainty. The statistical roots of this discipline were developed in the early 20th century and were developed through the years, culminating in "exploratory data analysis" by John Tukey in the 70s (Tukey, 1977). Although the discipline is often reduced to predictive modelling, it covers decision modelling and descriptive analysis. In the context of ecommerce, predictive models use data on the past performance of a website to determine how likely a random customer is to behave in a certain way. In doing this, it presents the unique advantage of weighing in all the data patterns, no matter how subtle they appear in the entire set. As such, they provide utility in the form of fraud detection and other aspects of customer performance. In addition, leveraging machine learning, they can provide the solutions to complex computations in real time. For instance, they flag potentially fraudulent transactions before they are completed (Hair, 2007).

Classification Models

There is variety in the choice of model that can be used to approach a business problem through predictive analysis. Picking a suitable model for the business problem raises the effectiveness of predictive analysis as an approach. Therefore, it is crucial to understand the choices available. Classification models can be considered as the most basic model type. They predict a target variable or class based on a set of input features in the dataset. Classification models include clustering, forecast, outliers, and time series models. Clustering models seek to categorize the data presented based on different features they may possess. This study uses the

K-Means cluster analysis method. This enables it to incorporate the concept of market segments, enabling a much better understanding of the data. Forecast models predict numeric values based on prior data. Outlier models aim at determining the unusual statistics in each dataset. Lastly, time series models are used to determine the growth patterns in data and predict the trends in a dataset, following its periodic nature.

They can be implemented through several models which all fall under the supervised machine learning category. These include logistic regression, which is based on logistic probability distribution in statistics. This model uses a logistic function to model the likelihood of a binary variable. Another example is the naïve Bayes classifier, which is based on Bayes' theorem. It is effective when modelling text data as it assumes that features are all independent of each other. Decision trees are a non-parametric approach which works by splitting the data depending on feature values. Decision trees are the building blocks of Random Forest models which are implemented in this study (James, et al., 2013).

Other examples include Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). SVMs find optimal decision boundaries between points that separate datapoints which belong in different classes. These boundaries are called hyperplanes (Chollet, 2018). Using these hyperplanes, they easily distinguish data classes and are very effective, particularly with binary classification problems. Nonlinear SVMs are effective even with complex datasets as they can mathematically transform the data into higher-dimensional space where it becomes easy to find a boundary (Tabsharani, 2023). On the other hand, KNNs use proximity of datapoints to make predictions about their grouping. They are preferable due to the simplicity they offer but are sensitive to scale and the presence of outliers.

Model Evaluation

Upon the selection and implementation of a predictive model, it is essential to determine its accuracy and usefulness on the data at hand and in the future. For this reason, there is a wide range of model evaluation techniques. Some of these techniques are domain oriented and thus subjective. These entail the scrutiny of professionals in the field of application for the model. For instance, the predictive model applied to predict customer behavior in ecommerce can be evaluated by professionals in the retail industry or ecommerce to determine the effectiveness of its predictions. Other techniques are more technical and require metrics and procedures to determine the statistical and logical integrity of a predictive model.

Some of the technical methods available for model evaluation are statistics which can be computed upon model implementation. These statistics are, at times presented as singular indicators of the goodness of the predictive models. However, these metrics are unable to demonstrate the purpose, application, and the appropriateness of the purpose of the predictive model to an audience beyond the scope of the business directly utilizing the model. In fact, these models are often in conflict with each other, and with the observations of the industry experts as some context is needed to effectively make use of a model and communicate its purpose, and its goodness. Winters, (2017) argues that the ability to convey a model's purpose and application to a larger audience is an integral part of the effectiveness of a model. Therefore, regardless of how an evaluation metric rates a given model, a model remains bad. Similarly, he posits that if a model is too slow to be practically applied in the real world, it is a poor model regardless of the model evaluation statistics obtained, or its effectiveness when applied on sample data. Therefore, it is crucial that predictive models are evaluated from all angles to determine any potential points of failure.

Model evaluation metrics include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These two metrics are primarily used when selecting a model and dealing with the issue of overfitting. AIC was developed in the 1970s by Hirotsugu Akaike. It penalizes a model's complexity, which is determined by the number of parameters and is calculated as:

$$AIC = 2k - 2ln(L)$$

Where k is the number of parameters and L is the highest value of the likelihood function. In evaluating the efficacy of a model, a lower AIC value implies that the model is closer to the unknown true model (which perfectly represents the relationships in the data) and therefore more robust. On the other hand, BIC penalizes free parameters more, given that it incorporates the sample size. It is calculated as:

$$BIC = ln(n) * k - 2ln(L)$$

Where k is the number of parameters and n is the sample size. By favoring simpler models, BIC overcomes the challenge of overfitting. The basic tradeoff between the two models is that while AIC selects a model with minimum information loss, and is therefore preferrable for predictive accuracy, BIC better explains the data and is, therefore, a better metric for selecting the true generative model (McQuarrie & Tsai, 1998).

Besides AIC and BIC, there are other metrics which offer better alternatives for evaluating classification models. They include accuracy, recall, precision, F-1 score, and AUC. Accuracy is the proportion of correct predictions given by a model. This is often presented as a percentage of the sample size. Precision is a similar metric to the accuracy as it is a proportion of the correct predictions. However, precision only considers the positive cases. It is the proportion of the predicted positive cases which are actually positive. That is, the proportion of true positives to all positives. Recall is a measure of a model's ability to find all relevant cases in

the sample it is presented with. It is calculated as the proportion of a model's true positives to the total positives in its predictions. That is, the true positives divided by the sum of true positives and false negatives. While recall and precision can be easily confused, it is important to note that their main difference lies in the aspects of accuracy they pursue. Recall focuses on a model's completeness, that is, its ability to identify all true positives, while precision is aimed at determining the exactness of a model, that is, the identification of positives as positives (Drummond & Holte, 2006).

Combining the two metrics enables one to determine whether a model is conservative, in which case it has high precision but low recall leading to many false negatives, or liberal, in which case it has high recall and low precision, reducing the number of false negatives. The F-1 score is defined as a harmonic mean of these two metrics. It is, therefore, used to find an agreeable balance between the completeness and exactness of a model as defined by the recall and precision respectively. As such, from a technical point of view, the F-1 score is often preferred. However, the business context is a better approach to determine which aspect of accuracy is more valuable in the model evaluation (Müller & Guido, 2016).

The area under the receiver operating curve, abbreviated as AUC-ROC or simply AUC, is another balancing metric for model evaluation. It was developed during World War II and was initially termed as the receiver operating curve as its original use was to tell whether a signal on the radar screen indicated an enemy ship or was simply random noise (Zou, et al., 2007). AUC describes the goodness of a model in several aspects. It is a ratio whose value ranges between 0 and 1. A value of 0.5 represents random classification and higher values represent more accurate modelling. AUC can also be seen as a ratio of the true positives to the false positives. It, therefore, provides a good assessment of the tradeoffs which arise from classification errors. However, it cannot be considered as an absolute measure of the accuracy

of a model. Therefore, this calls for an assessment of the model efficacy through other metrics and approaches.

Case Study Dataset

This study utilizes a dataset with 12,330 rows which all represent customer session visits to the ecommerce site. The data has been cleaned such that it does not present any trend to a particular day, user profile, campaign, or period. The dataset has 18 attributes, 10 of which are numerical and 8 categorical. These include the "Revenue" as a binary variable which has TRUE or FALSE values which represent the purchase decision made by a customer. "Administrative" and "Administrative Duration" are the number of administrative pages and the total time spent in each page of this category. The number of pages and durations define the feature pairs of "Informational" and "Informational Duration" for informational pages in the site, and "Product Related" and "Product Related Duration" for product related pages. The "Bounce Rate" variable denotes the proportion of visitors who enter the site from a given page and then leave the site without interacting with any other feature in the site or making any other request to the analytics server in that session. "Exit Rate" is the percentage of exits on a given page. "Page Value" is the average value for every page that a customer visited before making a transaction. The "Special Day" variable is defined by the proximity of the time when a customer visited the site to a precise special day (Sakar & Kastro, 2018).

These variables all have a potential to infer where the company can invest efforts to improve. For instance, it follows from the customer visits that if a customer visits an administrative page, they are more likely to be following up on a delivery or giving one form of feedback or another, but if a customer visits a product related page, they are more likely to be seeking a product to

buy. These correlations exist between the likelihood of a customer making a purchase and the rest of the variables based on the definitions of the variables.

Business Problem

Customer conversion refers to the proportion of the purchases or signups made on a site to the total number of customer visits. It defines one of the most important business problems in ecommerce. A low conversion rate implies poor performance of the ecommerce site while a higher one implies a better performance. Given that more visitors pose a maintenance challenge for the firm, creating more costs for website design, development, and maintenance, while the purchases made generate revenue for the ecommerce site, a low conversion rate indicates a lower profit margin.

This study develops and evaluates models which facilitate the identification of high intent customers. These are customers who are more likely to make purchases on the site. Targeting this kind of customers would lead to an improvement in the conversions. Typical ecommerce conversion rates range between 2 and 5% as most people use the sites as a point of reference for prices (Zumstein & Kotowski, 2020). Given this low conversion rate, even slight improvements can lead to considerable revenue gains.

By considering variable such as the operating system, browser, region, traffic type, and visitor type, this study factors in the effect of the website speed, promotions, ease of navigation, and even product availability in different contexts. These factors weigh in considerably on the conversion rate on an ecommerce platform (Gabir & Karrar, 2018). The predictions made are, therefore, more valuable in guiding the firm on what investments to make regarding the website speed, design, maintenance, and even the promotions it can run to improve the conversion.

Methodology

This study entails the creation of two predictive models to determine whether a customer will buy on the ecommerce platform or not. To achieve this, the study follows a structured approach to facilitate efficient analysis along with a clear documentation of the analysis in the code and the paper. This structure is best defined by the CRISP-DM framework and is clearly visible in the python script. This chapter explains the implementation of this structure in the code. The code is adapted from Medium article and updated for compatibility with the newest version of Python as well as a more unique effect on the visualizations (Sahu, 2021).

CRISP-DM Framework

This framework requires business understanding to build the context for the descriptive and predictive analysis that follows. In addition to the theoretical review of the ecommerce industry discussed in the literature review, the script used for the python analysis starts by importing the dataset provided. This introduces the business context of the analysis.

With a sufficient background of the case study in the business context, the study then seeks to complete data understanding through exploratory data analysis. This is understanding is enhanced by generating visualizations of the dataset and the statistical analysis of the data which reveals the correlations between various features. The exploratory data analysis also reveals the data distributions in various variables in the dataset. The Code below entails the import of various libraries which are required for predictive modelling in python, the import of the dataset, and the exploratory analysis.

```
#Import Libraries for Predictive Modelling
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy
import sklearn
import seaborn as sns

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import roc_curve, auc, roc_auc_score, RocCurveDisplay
from sklearn.linear_model import LogisticRegression
from sklearn.cluster import KMeans

#Import Dataset
raw =pd.read_csv("C:/Users/kagir/OneDrive/Desktop/online_shoppers_intention.csv")

#Get Descriptive Statistics for the Dataset
raw.describe()
#Get a count of the missing values
mv=raw.isnull().sum()
mv
```

Figure 1 Exploratory analysis for data understanding.

The following code continues with the exploratory analysis and generates graphs of customers who have brought revenues as a percentage, and the distribution of visitor type. These are only two of the plots created for this stage of the analysis.

```
#Plot the percentage of customers who have brought revenue
sns.set_palette('Paired_r')
sns.set_style("darkgrid")
plt.figure(figsize=(8,5))
total = float(len(raw))
ax = sns.countplot(x="Revenue", data=raw)
for p in ax.patches:
  percentage = '{:.1f}%'.format(100 * p.get_height()/total)
  x = p.get_x() + p.get_width()
  y = p.get_height()
  ax.annotate(percentage, (x,y), ha='center')
plt.show()

#Distribution of Visitor Type
raw['VisitorType'].value_counts()
sns.set_palette("Paired")
plt.figure(figsize=(8,5))
total = float(len(raw))
ax = sns.countplot(x="VisitorType", data=raw)
for p in ax.patches:
  percentage = '{:.1f}%'.format(100 * p.get_height()/total)
  x = p.get_x() + p.get_width()
  y = p.get_height()
  ax.annotate(percentage, (x, y), ha= 'center')
plt.show()

#Distribution of Visitor Type Over the Weekend
```

Figure 2 Exploratory analysis plots

The last step in the data understanding is an assessment of the relationship between the variables. This is achieved through conducting linear regression analysis as shown below.

```
#Linear Regression plot between Administrative and Informational
sns.lmplot(x = 'Administrative', y = 'Informational', data = raw, x_jitter = 0.05)
plt.show()
#Multi-variate analysis: Month vs Pagevalues wrt Revenue
sns.boxplot(x = raw['Month'], y = raw['PageValues'], hue = raw['Revenue'], palette = 'inferno')
plt.title('Mon. vs PageValues w.r.t. Rev.', fontsize = 30)
plt.show()

#month vs bouncerates wrt revenue
sns.boxplot(x = raw['Month'], y = raw['BounceRates'], hue = raw['Revenue'], palette = 'Oranges')
plt.title('Mon. vs BounceRates w.r.t. Rev.', fontsize = 30)
plt.show()
#visitor type vs exit rates w.r.t revenue
sns.boxplot(x = raw['VisitorType'], y = raw['BounceRates'], hue = raw['Revenue'], palette = 'Purples')
plt.title('Visitors vs ExitRates w.r.t. Rev.', fontsize = 30)
plt.show()
```

Figure 3 Linear regression and multivariate analysis.

Having created data understanding through the exploratory analysis, the next step in the framework is data preparation. This starts with the checking of the number of null values in the data and replacing them. The preparation also entails the preparation of the data for cluster analysis in order to determine the customer segments according to the dataset's features. It also includes the preprocessing in order to prepare the training and test datasets for the Random Forest and Logistic Regression. The code for this step is shown below.

```
# checking the no. of null values in data after imputing the missing value
raw.fillna(0, inplace = True)
raw.isnull().sum().sum()

##Cluster Analysis
#Elbow method is a graph between WCSS and No.of Clusters.
# preparing the dataset and checking the shape
x = raw.iloc[:, [1, 6]].values
x.shape
```

```
#Data Preprocessing to build Random Forest classifier and Logistic Regression
# one hot encoding
data1 = pd.get_dummies(raw)
data1.columns
le = LabelEncoder()
raw['Revenue'] = le.fit_transform(raw['Revenue'])
raw['Revenue'].value_counts()
# getting dependent and independent variables
x=data1
# removing the target column revenue from
x = x.drop(['Revenue'], axis = 1)
y = data1['Revenue']
# checking the shapes
print("Shape of x:", x.shape, "Shape of y:", y.shape)
#Splitting the data between train and test sets
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 0)
# checking the shape
print("Shape of x_train :", x_train.shape, "Shape of y_train :", y_train.shape)
print("Shape of x_test :", x_test.shape, "Shape of y_test :", y_test.shape)
```

Figure 4 Data preparation

After the data preparation, the next step in the CRISP-DM structure is modelling. The models developed include the elbow method, which is a clustering algorithm aimed at determining the natural groups of customers within the dataset. It is a graph of the number of clusters against the within cluster sum of squares (WCSS). It seeks to reduce the Euclidean distance between the points within one cluster by using the sum of squares (Kel'manov, et al., 2019). In a way, the cluster analysis falls under both the data preparation step and modelling. This is because

the number of clusters and their characteristics are just as much a point of consideration when building the other models as they are answers by themselves. The graph is generated as follows:

```
154  x.shape
155  wcss = []
156  for i in range(1, 11):
157      km = KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_state=0,
158                  algorithm='lloyd',tol = 0.001)
159      km.fit(x)
160      labels = km.labels_
161      wcss.append(km.inertia_)
162  plt.rcParams['figure.figsize'] = (15, 7)
163  plt.plot(range(1, 11), wcss)
164  plt.grid()
165  plt.tight_layout()
166  plt.title('The Elbow Method', fontsize = 20)
167  plt.xlabel('No. of Clusters')
168  plt.ylabel('wcss')
169  plt.show()
170
171  #The maximum bend is at third index, that is the number of Optimal no. of Clusters for
172  # Adminstrative Duration and Revenue is Three.
173  # plotting the clusters
174  km = KMeans(n_clusters = 3, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
175  y_means = km.fit_predict(x)
176  plt.scatter(x[y_means == 0, 0], x[y_means == 0, 1], s = 100, c = 'red', label = 'Un-interested Customers')
177  plt.scatter(x[y_means == 1, 0], x[y_means == 1, 1], s = 100, c = 'yellow', label = 'General Customers')
178  plt.scatter(x[y_means == 2, 0], x[y_means == 2, 1], s = 100, c = 'green', label = 'Target Customers')
179  plt.scatter(km.cluster_centers_[:,0], km.cluster_centers_[:, 1], s = 50, c = 'blue' , label = 'centeroid')
180  plt.title('Administrative Duration vs Duration', fontsize = 20)
181  plt.grid()
182  plt.xlabel('Administrative Duration')
183  plt.ylabel('Bounce Rates')
184  plt.legend()
185  plt.show()
186
```

Figure 5 Cluster analysis

Upon the determination of the optimal number of clusters, the Random Forest and logistic regression models are built on the training data previously prepared as follows.

```
267  #RandomForest classifier model Building
268  # MODELLING
269  rfmodel = RandomForestClassifier()
270  rfmodel.fit(x_train, y_train)
271  y_pred = rfmodel.predict(x_test)
```

Figure 6 Random Forest Model

```
#Building a logistic regression model
##MODELLING
lrmodel = LogisticRegression(solver="liblinear", random_state=0)
lrmodel.fit(x_train, y_train)
y_pred1 = lrmodel.predict(x_test)
```

Figure 7 Logistic Regression Model

After the modelling stage, the evaluation of the models comes next. The models are evaluated using the metrics of accuracy, the confusion matrices, and classification report. The code for the evaluations is as follows.

```
272
273   # evaluating the model
274   print("Training Accuracy :", rfmodel.score(x_train, y_train))
275   print("Testing Accuracy :", rfmodel.score(x_test, y_test))
276
277   #Confusion Matrix.
278   fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 6), sharey=False)
279   cm = confusion_matrix(y_test, y_pred)
280   sns.heatmap(cm, ax=ax1, annot=True)
281   ax1.set_title('Confusion Matrix 1')
282   ax1.set(xlabel='Predicted Label', ylabel='True Label')
283
284   cm2 = confusion_matrix(y, rfmodel.predict(x))
285   ax2.imshow(cm2)
286   ax2.grid(False)
287   ax2.set_title('Confusion Matrix 2')
288   ax2.xaxis.set(ticks=(0, 1), ticklabels=('Predicted 0s', 'Predicted 1s'))
289   ax2.yaxis.set(ticks=(0, 1), ticklabels=('Actual 0s', 'Actual 1s'))
290   ax2.set_ylim(1.5, -0.5)
291   for i in range(2):
292       for j in range(2):
293           ax2.text(j, i, cm2[i, j], ha='center', va='center', color='red')
294   fig.tight_layout()
295   plt.show()
296   # classification report
297   cr = classification_report(y_test, y_pred)
298   print(cr)
299
```

Figure 8 Accuracy, confusion matrix, and classification report for the Random Forest model

The ROC curve is also used to evaluate the Random Forest and logistic regression models. Its code is structured as follows.

```
##Plot the ROC for RandomForest
# Calculate the ROC curve
y_scores = rfmodel.predict_proba(x_test)[:, 1]
rf_fpr, rf_tpr, _ = roc_curve(y_test, y_scores)
rf_auc = roc_auc_score(y_test, y_scores)
# Plot the curve
plt.plot(rf_fpr, rf_tpr, label='Random Forest (auc = %0.3f)' % rf_auc)
plt.legend()
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Random Forest')
plt.show()
#Saving the predictions as a dataframe
df=pd.DataFrame(y_pred,columns=["Revenue"])
df
```

Figure 9 ROC curve for the random forest model

With this structure, the study is very easy to understand and even replicate its results. This enhances its credibility and, therefore, its value to the ecommerce site. In addition, the CRISP-DM structure generally increases the effectiveness of analytics in solving business problems (Palacios, et al., 2017). For better comparison between the two predictive models, the study also generates the ROC curves of both models on the same axes as shown below. Following this thorough analysis, the models are ready for deployment.

```

#Plotting ROC curve for both Random Forest and Logistic Regression
plt.plot(logit_fpr, logit_tpr, label='Logistic (auc = %0.3f)' % lr_auc)
plt.plot(rf_fpr, rf_tpr, label='Random Forest (auc = %0.3f)' % rf_auc)
plt.legend()
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve Comparison')
plt.show()
```

Figure 10 ROC curves on the same axes

Discussion

Exploratory Analysis

Upon building and running the models using the provided dataset, the python output creates a deeper understanding of the ecommerce business and the data in the case study. This facilitates much better interpretation and understanding of the data, the models, and the efficacies of the models as per the different evaluation criteria. Upon the exploratory analysis of the shopping records for the 12,330 sessions, the study uncovers the distributions of the customers according to different variables. The graph below is a simple snapshot of the site's conversion rate, suggesting that the rate is about 15.5%.

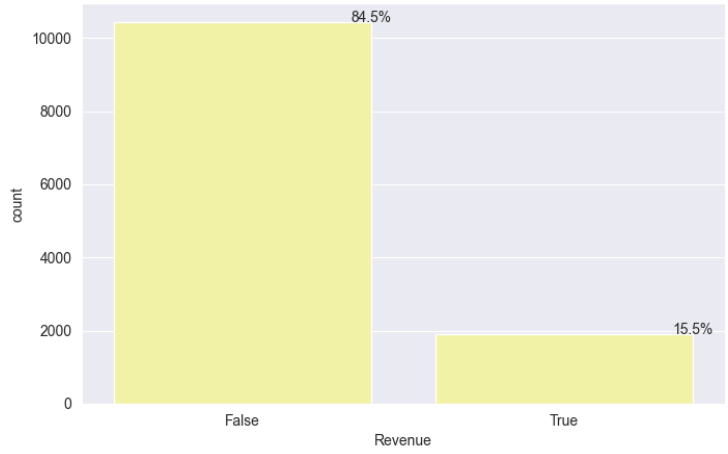

Figure 11 Percentage of customers who made a purchase

The business problem facing the case study ecommerce site is to identify signals in the data which predict the likelihood of a visitor to the site makes a purchase. In essence, the predictive analytics of the data should essentially identify the trends and patterns in the 17 variables in the data and how they seem to affect the outcome in the revenue variable. This way, the site can identify the high intent users as a cluster and encourage a higher conversion rate among them. The study achieves this through the exploratory analysis. For instance, the graph below

shows the distribution of revenue for different traffic types. The disparity between the traffic types suggests the need for an assessment of the website's performance on the non-performing traffic types, particularly those that scored 100% false on the revenue variable, to determine whether the challenge lies within the firm or whether the traffic types simply represent low intent users.

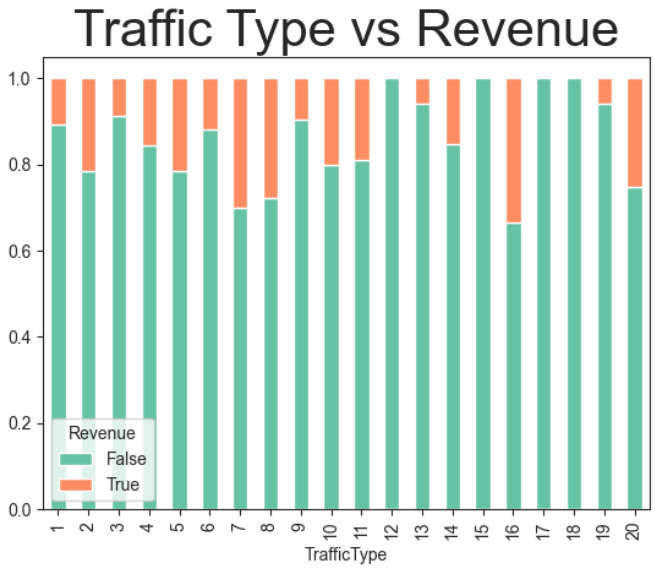

Figure 12 Distribution of revenue for different traffic types

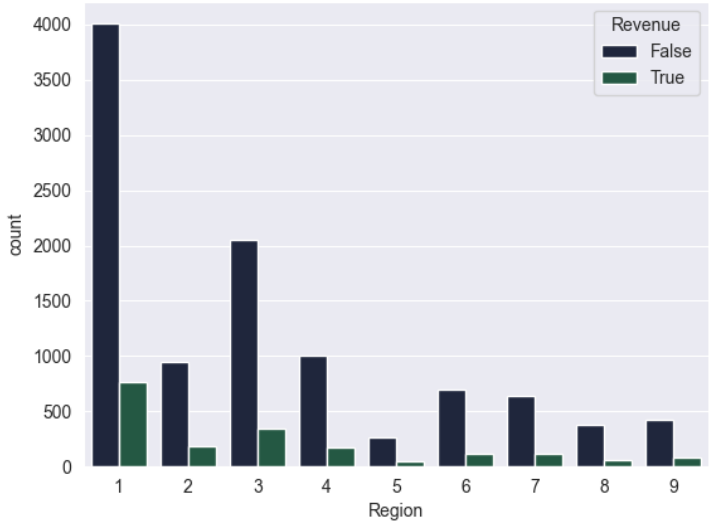

Figure 13 Distribution of revenue for different regions

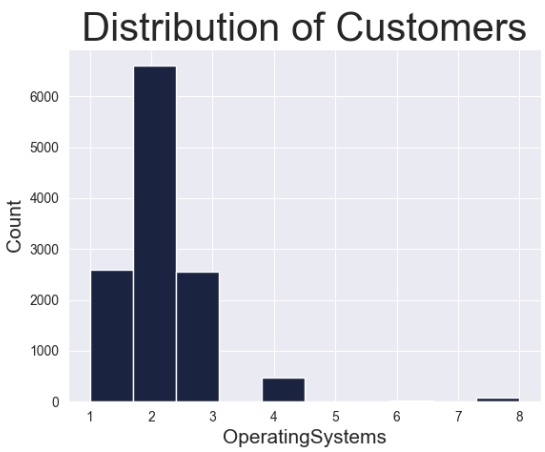

Figure 15 Distribution of customers using different operating systems.

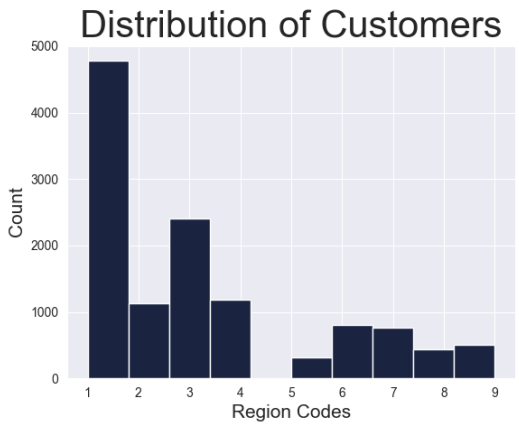

Figure 15 Distribution by region codes

Another outcome of the exploratory analysis is the understanding of the distribution of the site visitors with respect to different variables, regardless of their choice to purchase or not to. For instance, the graphs shown below indicate the usage of different operating systems and the distribution of customers in different regions. This is a robust potential guide for the company's web development team as they would need to enhance the website to perform seamlessly on operating systems 1,2, and 3.

The exploratory analysis also includes the determination of the correlation between the informational and administrative visits. As the graph below shows, the analysis reveals a positive correlation in that, the more informational pages a customer visited, the more administrative ones they did as well. This indicates a need to enhance the informational pages to make self service more convenient for customers.

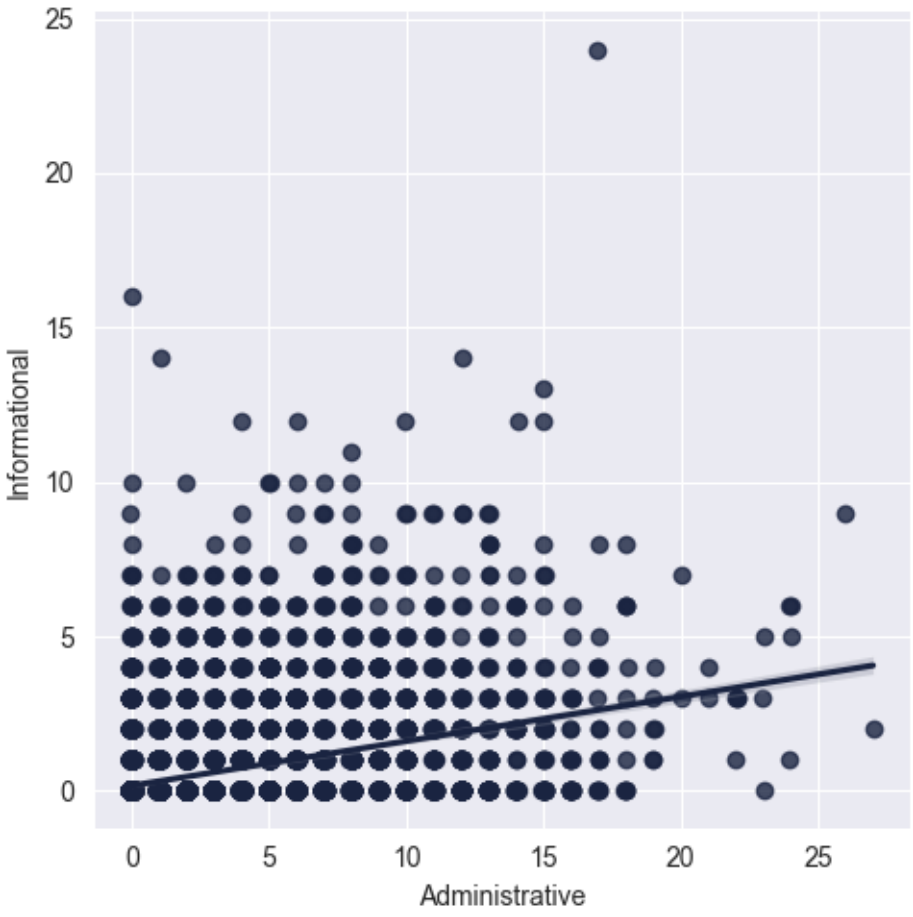

Figure 16 Linear regression plot between informational and administrative sites

The multivariate analysis of the monthly customer visits reveals a pattern which may arise even in clustering. As shown in the diagram below, different periods have different conversion rates. For instance, the month of December has higher page values and higher sales than the rest of the months, albeit with a seemingly low number of visits.

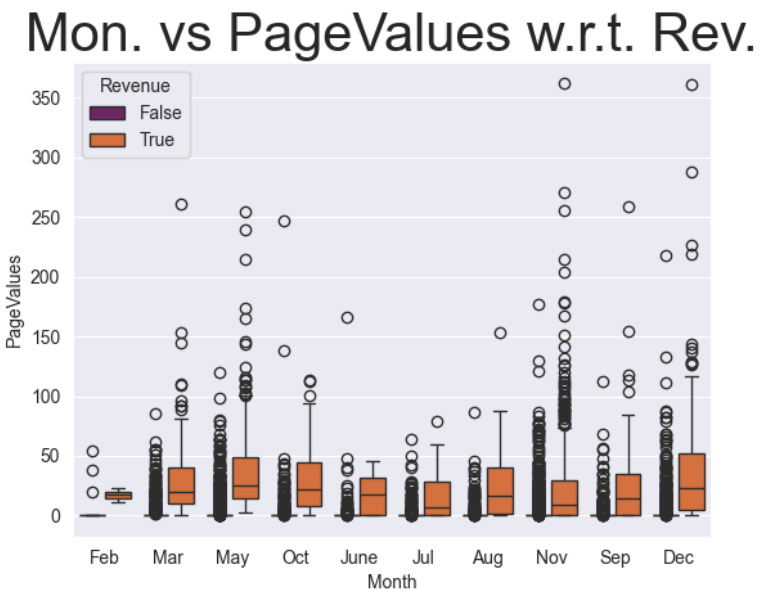

Figure 17 Multivariate analysis of the month vs page values with respect to revenue

Clustering Analysis

The exploratory analysis exhibits the need for clustering analysis in order to expound on the grouping of customers, and hopefully identify the most promising segments with regard to the level of purchase intent. As the graph below shows, the clustering analysis helps to separate the target customers from the general and uninterested ones. With high bounce rates and low durations on the administrative pages, it follows that the customers indicated in red may have visited the site but failed to set up accounts. The clustering analysis, therefore, suggests that the ecommerce site would increase the customer conversion by streamlining its administrative pages to make activities such as signing up much easier.

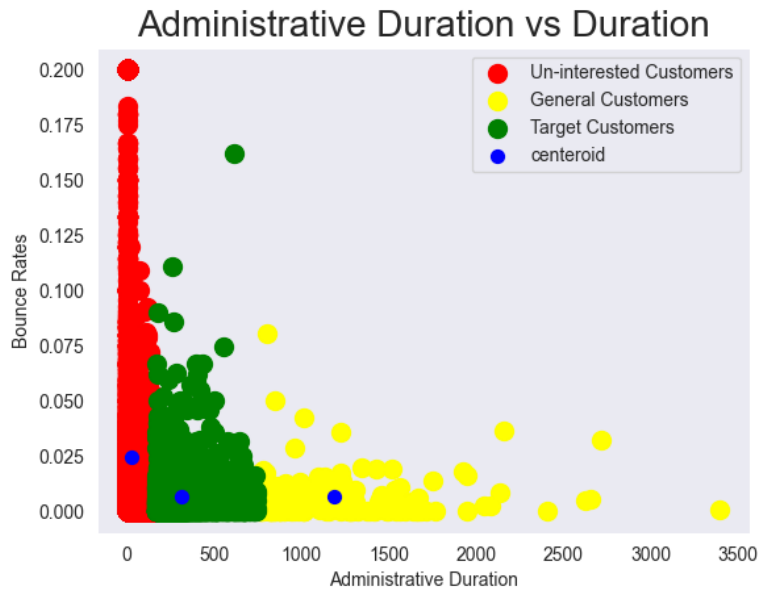

Figure 18 Clustering analysis for administrative duration and revenue

The clustering analyses reveals that the optimal number of clusters was 3 as it balances shorter distances between the respective centroids and the datapoints in their clusters with the highest distances between the centroids. This is shown in the graph of the elbow method below.

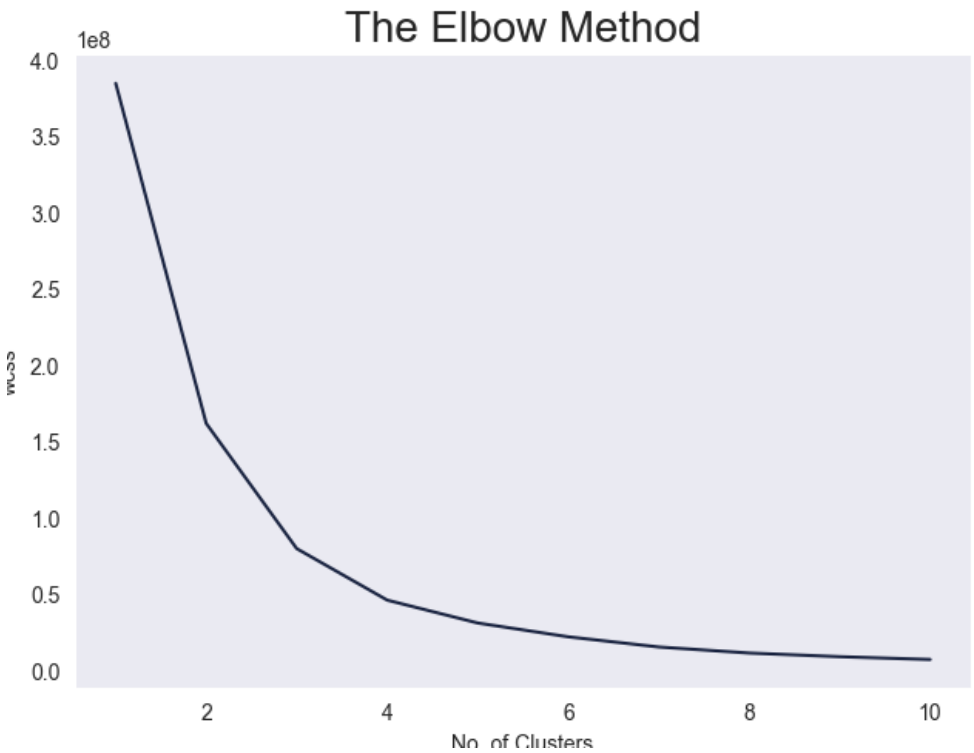

Figure 19 The elbow method

Random Forest Model

Both models achieve reasonably high accuracy, with Random Forest at 89% as shown in the classification report below. However, accuracy is not a reliable indicator of model efficacy, particularly when the class distribution is highly imbalanced as is the case with this dataset.

	precision	recall	f1-score	support
False	0.91	0.96	0.94	3077
True	0.75	0.54	0.63	622
accuracy			0.89	3699
macro avg	0.83	0.75	0.78	3699
weighted avg	0.88	0.89	0.88	3699

Figure 20 Random Forest classification report

In order to make up for the shortcomings of the measure of accuracy, the study relies on other indicators of efficacy such as the AUC. The ROC curve below shows the AUC for the Random Forest Model with the value of 0.92.

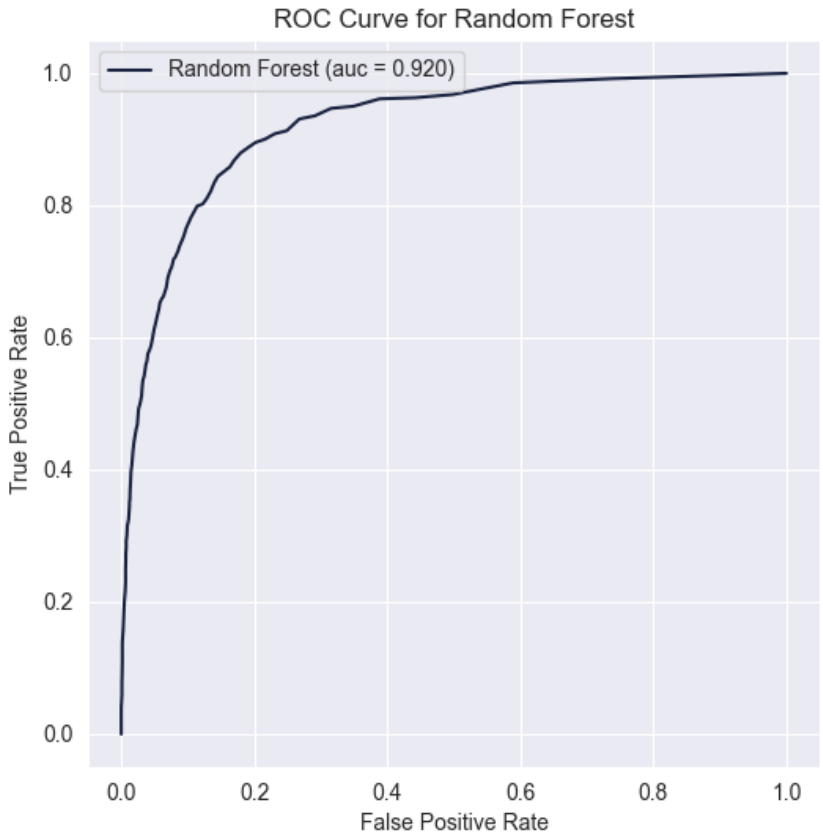

Figure 21 ROC for random forest model

The confusion matrix also highlights the performance of the model with regard to its proportion of correct and incorrect predictions.

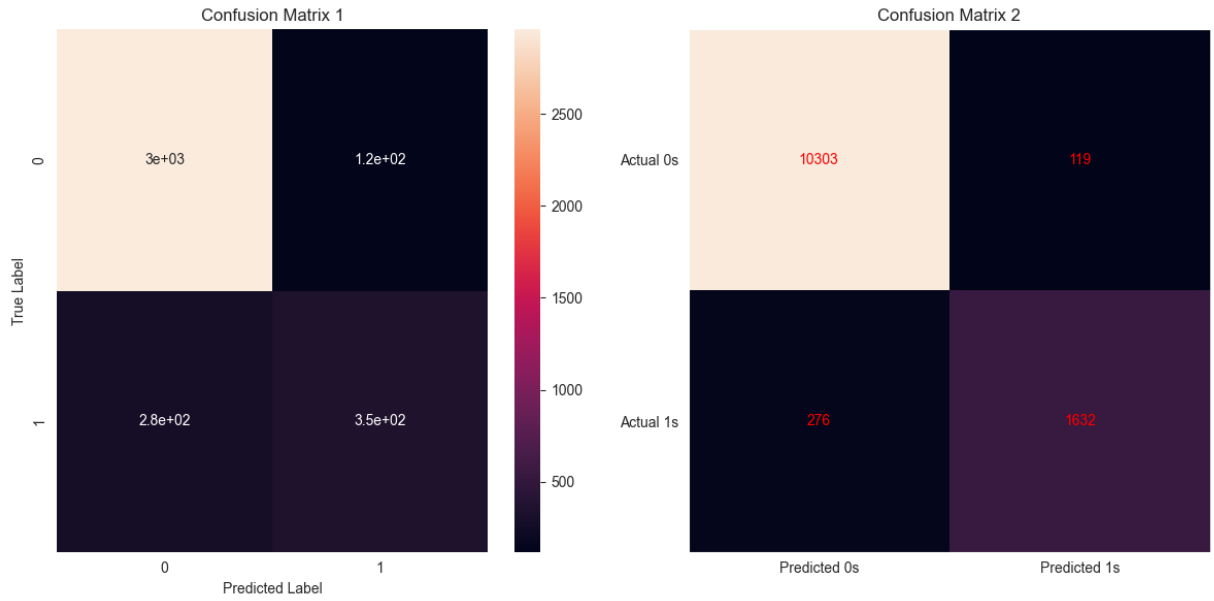

Figure 22 Confusion Matrix for the random forest model

Logistic Regression Model

The Logistic Regression model has an accuracy of 87% on the test data as shown in the classification report below.

	precision	recall	f1-score	support
False	0.88	0.98	0.93	3077
True	0.76	0.37	0.50	622
accuracy			0.87	3699
macro avg	0.82	0.67	0.71	3699
weighted avg	0.86	0.87	0.86	3699

Figure 23 Classification report for logistic regression model

The ROC curve below shows the AUC for the logistic regression Model with the value of 0.89.

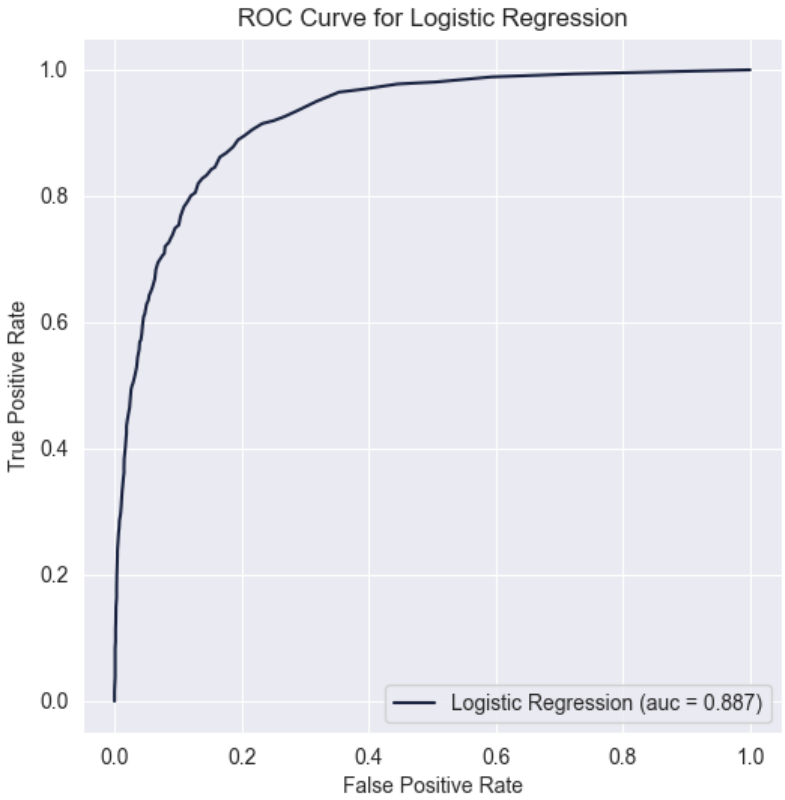

Figure 24 ROC for logistic regression model

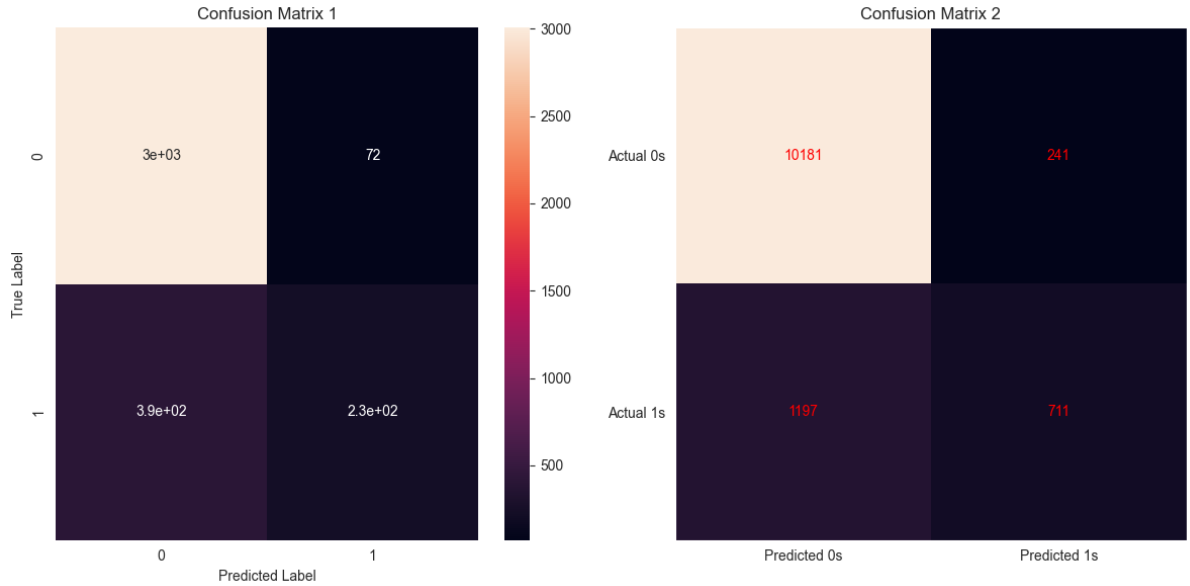

Figure 25 Confusion Matrices for logistic regression

In summary, Random Forest delivers slightly better performance than Logistic Regression for this dataset and business problem. AUC of 0.92 indicates reliable identification of converting user sessions. Recommendations to boost the conversion rate include focusing marketing on high-intent segments identified by the clustering analysis and the model. The site could also customize product recommendations using predicted user preferences. Testing different website versions and measuring impact on conversion is another strategy. This analysis demonstrates the value of predictive analytics in gaining actionable insights from customer behaviour data to meet business goals.

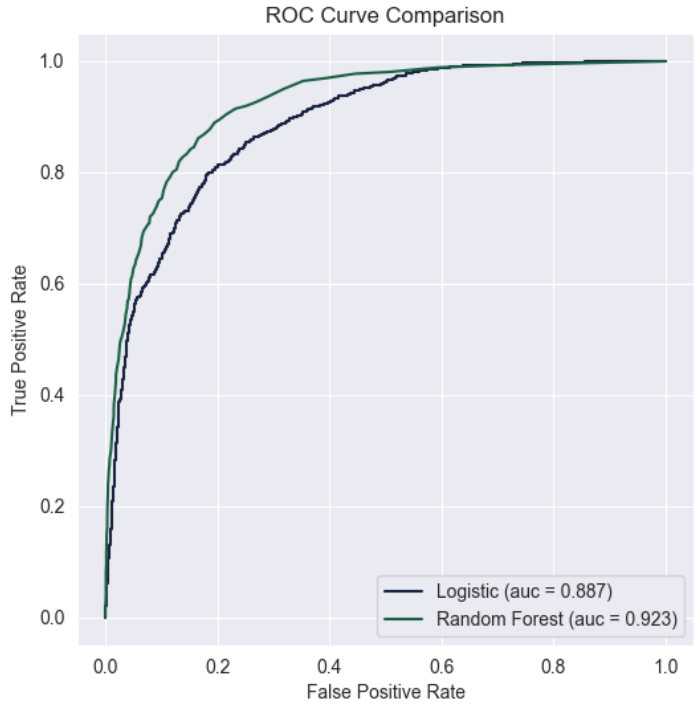

Figure 26 Comparison between ROCs of random forest and logistic regression models

Conclusion

This study demonstrates the value of predictive analytics in gaining actionable insights from customer behaviour data to improve business processes and outcomes such as conversion rate and profitability. The analysis of the ecommerce site dataset reveals thought-provoking patterns related to traffic sources, customer segments, device types and their correlation with purchase decisions.

Clustering analysis provides a data-driven approach to segment customers. Basing the clustering on their engagement levels, this study groups them as target, general or uninterested groups. The two predictive models, Random Forest and Logistic Regression, classify user sessions as converting, when the customer in the corresponding row makes the decision to purchase, or decides not to, based on attributes like pages visited, durations, bounce rates, among others.

The Random Forest model achieves a better performance as compared to Logistic Regression. The former has an accuracy of 89% and AUC of 0.92, indicating reliable identification of the outcomes in the revenue variable. However, some false positives arise in its predictions from optimizing for recall over precision. Further tuning of this model's parameters can potentially improve its precision and F1-score, making it more effective to solve the business problem.

One of the practical recommendations this study raises entails focusing marketing efforts on high-intent segments identified by the models. Another strategy to improve conversions would be customizing product recommendations, perhaps using predicted user preferences in order to create a personalized user experience. Testing different website versions and content is also recommended to optimize conversion rate.

While this research demonstrates the tremendous potential of predictive analytics in ecommerce, some limitations persist. First, the models discussed herein are built on a sample

dataset which may not represent the full population. Second, the user behaviour exhibited online evolves over time, making the models unreliable on the long term. The models would need regular re-training and tuning on up-to-date data. Extensions to the models could even explore entirely different algorithms such as neural networks and incorporate new data sources.

This study discusses how predictive analytics can utilise customer behaviour data to generate actionable insights which guide business strategy and key decisions. Features such as targeted marketing and personalized recommendations continue to improve the performance of ecommerce platforms and websites in general through data-guided experiments. With the exponential growth in data discussed before, the scope for data-driven decision making will continue expanding in this and other sectors.

References

Adnan, M. et al., 2011. Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide ecommerce business promotions. *Social Network Analysisand Mining,* 1(1), pp. 173-185.

Chollet, F., 2018. *A Tour of Machine Learning Algorithms.* [Online]
Available at: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/
[Accessed 23 October 2023].

Drummond, C. & Holte, R., 2006. Cost curves: An improved method for visualizing classifier performance. *Machine Learning,* Volume 65, pp. 95-130.

Gabir, H. & Karrar, A., 2018. *The Effect of Website's Design Factors on Conversion Rate in E-commerce.* s.l., s.n., pp. 1-6.

Hair, J. F., 2007. Knowledge creation in marketing: the role of predictive analytics. *European Business Review,* 19(4), pp. 303-315.

Huberman, B. & Adamic, L., 1999. Internet: Growth dynamics of the World-Wide Web. *Nature,* Issue 401, pp. 131-131.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning.* s.l.:Springer.

Kel'manov, A., Pyatkin, A. & Khandeev, V., 2019. NP-Hardness of Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with Constraints on the Cluster Sizes. *Doklady Mathematics,* Issue 100, pp. 545-548.

McQuarrie, A. & Tsai, C. L., 1998. *Regression and time series model selection.* Singapore: World Scientific.

Müller, S. & Guido, S., 2016. *Classifier performance visualization using ROCR and model scoring in Python.* [Online]
Available at: https://joss.theoj.org/papers/10.21105/joss.00135
[Accessed 24 Octoer 2023].

Palacios, H., Toledo, R., Pantoja, G. & Navarro, A., 2017. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal,* Volume 2, pp. 598-604.

Ring, J., 2023. *30 years ago, one decision altered the course of our connected world.* [Online]
Available at: https://www.npr.org/2023/04/30/1172276538/world-wide-web-internet-anniversary
[Accessed 23 October 2023].

Sahu, S., 2021. *PREDICTIVE WEB ANALYTICS: A CASE STUDY.* [Online]
Available at: https://medium.com/analytics-vidhya/predictive-web-analytics-a-case-study-f30feda45002
[Accessed 23 October 2023].

Sakar, C. & Kastro, Y., 2018. Online Shoppers Purchasing Intention Dataset. *UCI Machine Learning Repository.*

Southern, L., 2017. The attraction and expansion of e-commerce during the recent economic downturn. *Problems and perspectives in management,* Issue 10.

Tabsharani, F., 2023. *support vector machine (SVM).* [Online]
Available at: https://www.techtarget.com/whatis/definition/support-vector-machine-SVM
[Accessed 23 October 2023].

Tukey, J. W., 1977. *Exploratory Data Analysis.* Massachusetts : Addison-Wesley.

Winters, R., 2017. *Practical Predictive Analytics : Make Sense of Your Data and Predict the Unpredictable.* Birmingham: Packt Publishing Limited.

Zou, K. H., O'Malley, A. J. & Mauri, L., 2007. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation,* Issue 115, pp. 654-657.

Zumstein, D. & Kotowski, W., 2020. *SUCCESS FACTORS OF E-COMMERCE – DRIVERS OF THE CONVERSION RATE AND BASKET VALUE.* s.l., s.n.