Understanding Online Shopper Behaviour: A Data-Driven Analysis of E-Commerce Sessions

Dillen Perumal | u5590917

1 Introduction

The dataset contains information about online shopping sessions carried out by anonymous shoppers over the span of one year. A total of 12330 sessions were recorded throughout the year, each recording data throughout the time spent by the visitor on the e-commerce website. 18 variables were observed for each session:

"Administrative", **"Informational"** and **"Product Related"** variables represent the number of different types of pages visited by the user, and their respective **"Administrative Duration"**, **"Informational Duration"** and **"Product Related Duration"** variables show the time spent on each of these pages.

The **"Bounce Rate"**, **"Exit Rate"** and **"Page Value"** variables represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of **"Bounce Rate"** shows how often the shopper would enter and exit specific web pages without interacting with them. The value of **"Exit Rate"** shows that of the pages visited by the user, how many were frequently the last visited pages for users in a session. The **"Page Value"** variable measures for the pages visited by the user, the average values of the purchases typically made on those pages.

The **"Special Day"** variable ranges from 0 to 1 and ranges the closeness of the visit to a specific special day (e.g. Mother's Day/Valentine's Day) when sessions are more likely to be finalised with a purchase.

The dataset also includes categorical variables: **"Operating system"** representing the device the user is accessing the site from; **"'Browser"** representing the internet browser being used; **"Region"** showing where the user if from; **"Traffic type"** representing different ways in which the user entered the site (e.g. through ads, directly entering the URL to their browser or from a search engine); **"Visitor Type"** showing if the user if a returning or a new visitor; **"Weekend"** showing whether or not the user is visiting the site on a weekend; and a **"Month"** variable.

Lastly, our target variable **"Revenue"** was monitored which classifies whether the user made a purchase during their session or not.

2 Exploratory Data Analysis

Firstly, the dataset was inspected for any rows with missing or incorrect values, where 85 sessions were removed after I saw they had visitor types of "other" rather than returning or true in a count plot. As these were a small subset of our total dataset removing them was not detrimental to the quality of our data. Numerical features were standardised by subtracting the means from each respective feature and dividing by their standard deviations. This allows for comparison between features of different units and magnitudes. Categorical data underwent one-hot-encoding to allow them to be used with numerical features, and boolean data was also converted to binary, with TRUE set to 1 and FALSE set to 0.

An interesting correlation heatmap is shown below displaying the correlations between all pairs of numeric data, from it we can see some interesting relationships between revenue and different features. "Page Values" have the highest correlation with revenue of 0.49 showing that users which

visited pages with higher page values were more likely to make purchases, this is intuitive because non-product pages will have page values of 0 and no purchase will be made on them, whereas purchases can only be made on product pages (in a further analysis one could filter out rows where not enough time was spent on product pages, such that the page value feature could be used to see whether customers were more likely to purchase cheaper or more expensive products). We can also see how variables such as “Bounce Rates” have a negative correlation of -0.15 with revenue since if users are clicking through multiple pages without interacting with them, they are likely struggling to find a product they want.

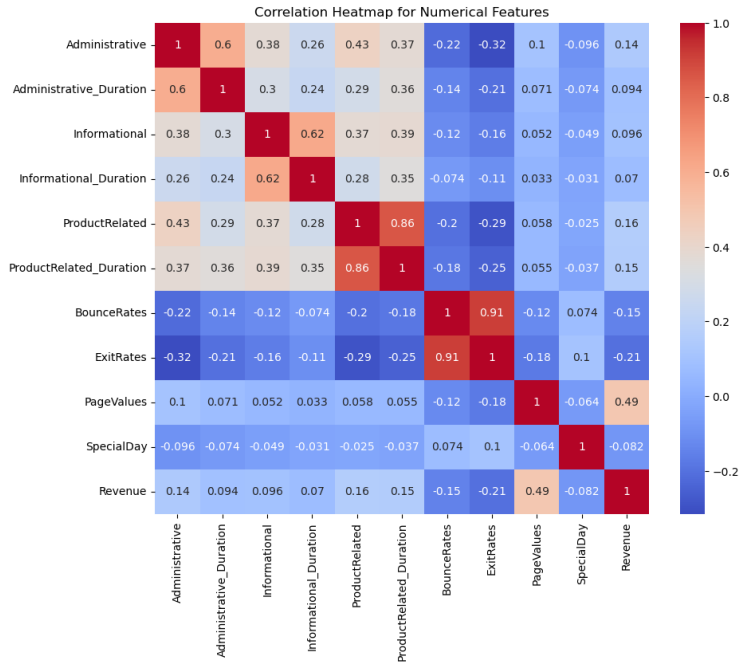

Figure 1: Heatmap showing correlations between all pairs of numerical features found in code under reference location 1

Looking closer at different variables produced unexpected insights, for example highlighting in which months users visited the website. Visitors frequented the sight in the winter period, with high counts in November and December, as expected due to special holidays such as Christmas leading to increased demand. However, there was also a surge of activity in March and particularly in May where 3364 users visited the sited, the highest out of all the months. For all the other months besides these four. visits were relatively low and below 500.

Figure 2 shows the proportion of sessions which resulted in a purchase across each month, we can see towards the later months purchases were more likely to be made, and the error bars do not overlap here with the earlier months. Again, this is likely due to the demand for presents increasing during the festive seasons.

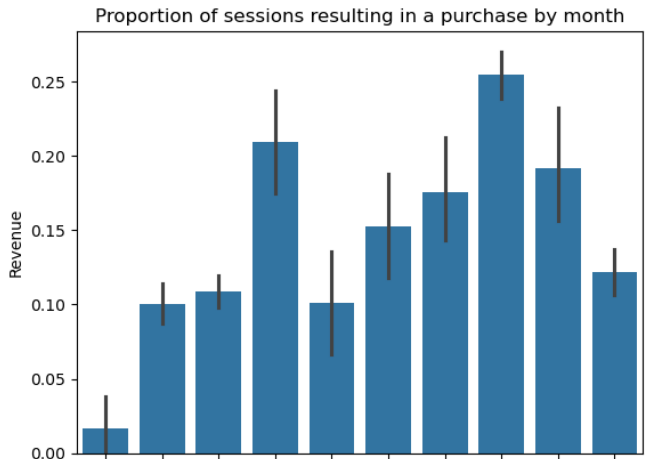

Figure 2: Bar plot of revenue against month from cell 55

A few more interesting statistics were obtained from the data, such as that 15% of all sessions resulted in a purchase and 23% of all visits were on the weekend. We can also see the total number of visits to each page type for all the session, with 6194

visits to informational pages, 28421 visits to administrative pages and 390189 visits to shopping pages.

3 Technical Analysis

In order to predict whether or not the user made a purchase, the following classification models were selected for the analysis: Logistic Regression, Random Forest and a Neural Network Binary Classifier. Initially all models were trained using all features of the processed dataset, then recursive feature selection was carried out individually for each dataset to reduce the features upon which each model was trained and tested and to avoid overfitting.

The training/test data split was 80/20 to provide sufficient data for both training and testing. This was possible as our dataset was very large with 12245 datapoints.

3.1 Logistic Regression

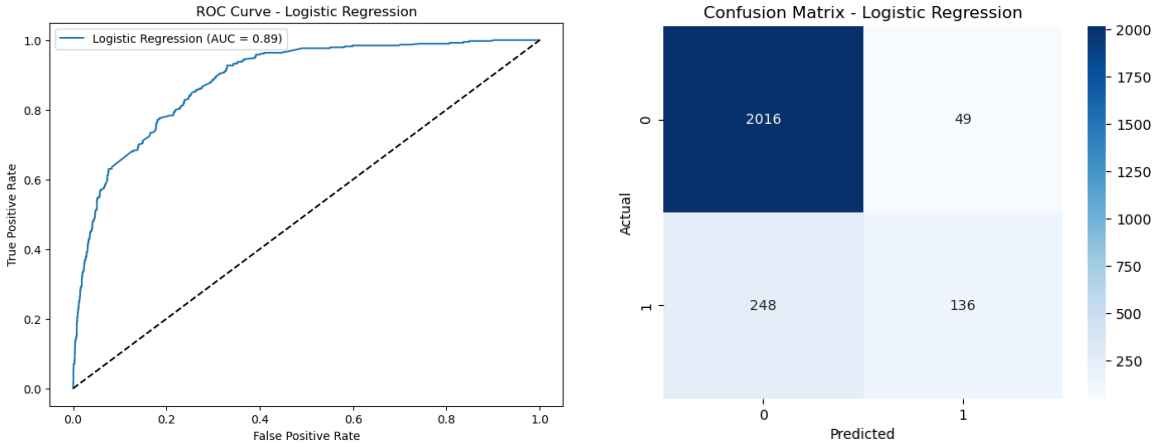

Figure 3: Model evaluation plots for LR models found in code under reference location 3

The first LR model achieved an AUC value for its ROC curve of 0.887 and after recursive feature selection to the 4 most significant features this value dropped only to 0.881. After experimenting with different parameters the model performed best with C, the inverse of regulation strength, set to 0.1. This produced an AUC value of 0.886 closer to the original model with all the features. The confusion matrix for the final model on the right shows a strong F1 score of 0.93 in predicting where there was no revenue, but a poorer score of 0.48 in predicting revenue points, struggling to identify where customers made a purchase.

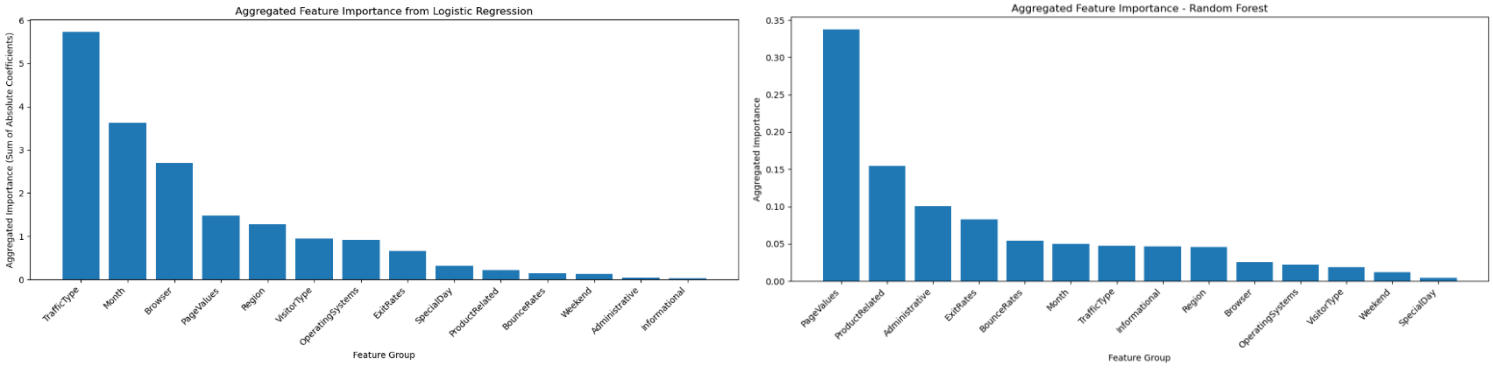

Figure 4: Plots showing feature importance for LR and RF models prior to feature selection found in code under reference location 2

3.2 Random Forest

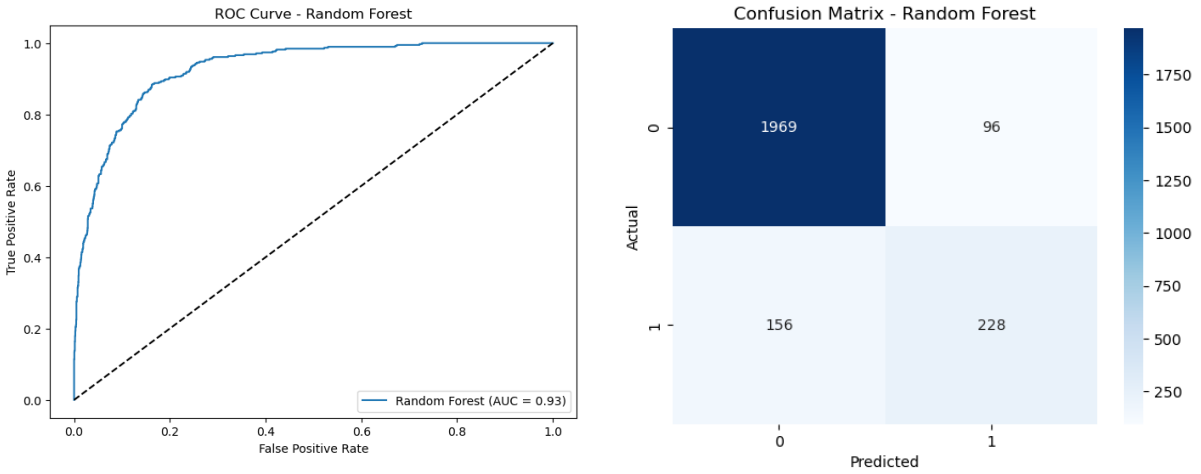

Figure 5: Model evaluation plots for RF models found in code under reference location 3

The first RF model achieved an AUC value for its ROC curve of 0.928, which was quite good and after recursive feature selection to the 7 most significant features this value dropped only to 0.9186. Slightly tweaking the minimum samples required to be at a leaf node for 3 increased the AUC to 0.928, which was quite good. As we can see the random forest model does a better job at recalling points where customers made a purchase, with an F1 score of 0.64 for the points.

3.3 Neural Network Binary Classifier

The neural network built for the analysis is a simple feedforward model with two hidden layers, the input layer maps our variables to 64 features and in the first hidden layer ReLU activation is used followed by a dropout layer, introducing non-linearity. The second hidden layer reduces the dimensionality from 64 to 32 and ReLU is used again followed by another dropout layer. The output layer maps our 32 features back to one output and classifies each datapoint as 1 or 0 using a logistic regression function.

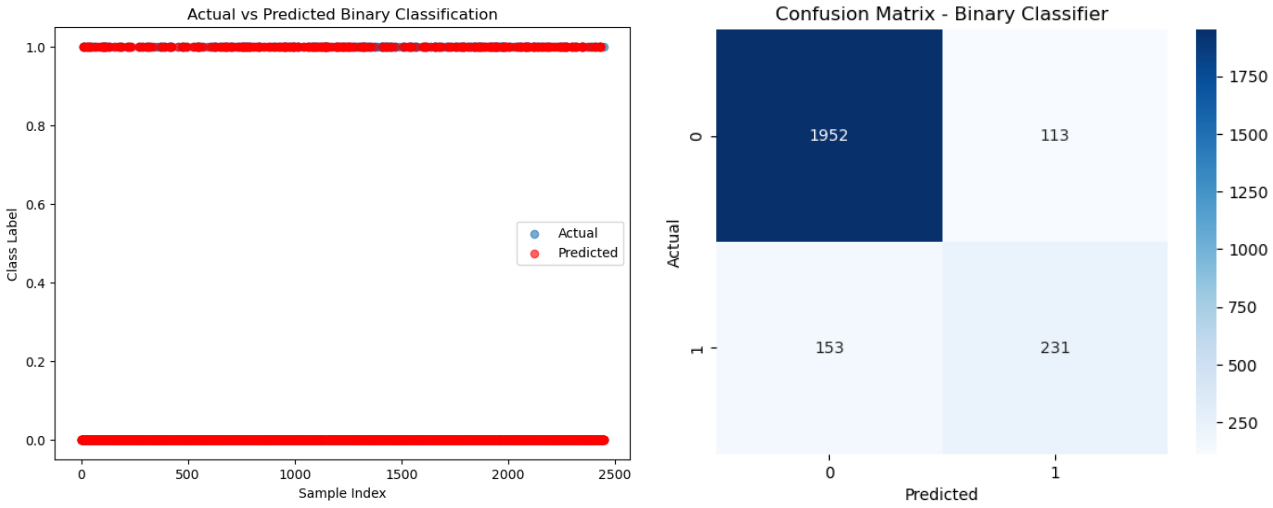

Figure 6: Model evaluation plots for the NN BC model found in code under reference location 4

The model was run for 50 epochs, a good middle ground to avoid overfitting. It yielded a final weighted average accuracy of 0.89, which is decent however looking at the confusion matrix again we

can see that the model struggled to identify datapoints where the customer made a purchase, with an F1 score of only 0.63 for these points.

4 Conclusions

Overall, we can see that all three models were effective in predicting when a customer would not make a purchase, but struggled to identify all points at which customers did make a purchase. The Random Forest was most effective in this task however still only obtained an F1 score of 64 for said points. The difficulty in accurately identifying these points is likely due to the large amount of false revenue points compared to true revenue points.

Inspecting the ROC graphs for the random forest and logistic regression models, we can also see visually how much smoother and higher the random forest curve is than the latter. The binary neural network used here significantly increased computational cost and model training times but did not perform better than the random forest model regardless, more feature engineering was likely necessary to optimise this model.

There are various potential methods which could be used to improve the dataset and model predictions, for example, trying a decision tree model more specialised for binary classifying such as the Histogram-based Gradient Boosting Classification Tree from scikit-learn, which also integrates gradient boosting allowing it to capture more complex patterns. Alternatively, one could generate new features, such as dividing the time spent on a page type by the number of that page type visited to obtain the average time spent on each page type in a session.