Title: Online shoppers purchasing prediction with classification models.

Student ID: s3961136 Student Name: Benjaporn Wongmayura email: s3961136@student.rmit.edu.au

Affiliations: RMIT University. Date of Report: 9 May 2023

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honour code by typing "Yes": Yes.

Table of Contents

| Executive Summary

This report aims to predict online visitor purchase behaviour on an e-commerce website. Data exploration identifies patterns and trends. Two classification models are used: decision tree and K-nearest neighbours (KNN). Results show that the decision tree model outperforms KNN, with higher precision, recall, f1-scores, and accuracy. The decision tree effectively predicts purchase behaviour and informs targeted marketing strategies. After data pre-processing, attribute exploration, and model evaluation, it is evident that machine learning can accurately predict online user purchase behaviour using limited attributes, surpassing human capabilities in this task. This study helps sellers identify influential factors for customer decisions, enabling targeted discounts and tailored marketing strategies. Optimising website features and providing personalised recommendations based on classification model insights improves customer satisfaction, conversion rates, and business performance.

| Introduction

The advent of e-commerce has revolutionised how people shop, bringing about significant global advantages for consumers. However, not every search for a product results in a purchase by the customer. With this in mind, the objective of this report is to forecast whether an online visitor will make a purchase transaction on an e-commerce platform.

| Methodology

Data

Data: Online Shoppers Purchasing Intention Dataset Data Set from UCI Machine Learning Repository
Dataset: 12,330 items with 18 features (12330 rows × 18 columns)
Types of data: Categorical = 8, Numerical = 10

Model

The report used supervised learning algorithms, which are the k-Nearest Neighbours algorithm and Decision Tree Model algorithm.

| Result

Data cleaning and preparation

Redundant white space

I have run **dataframe ['column name'].value_counts().** All the column with string values does not have redundant white space. Therefore, no need to use the strip function to eliminate white space. For the numerical column, since we're dealing with values that aren't strings, it's unnecessary to verify for mistakes such as whitespace or capitalization discrepancies.

Null Value
I have run **dataframe.isnull().values.any()** and **dataframe.isnull().sum()** and found that null value does not exist in this dataset. We do not need to drop the row or impute with 0 or means.

Outliers
I have plotted boxplots for all the features in the data frame, and I found that there were two outliers in Productrelated_Duratio n, which are row 8071th - **63973.52223**, and 5152th - **43171.23338**. This makes sense because sometimes people leave the page open without closing it. Hence, we will delete the outlier to make the data more reliable and truly represent the normal behaviour of online shoppers.

Impossible Value Sanity Check - Duration Value¶
We use **.min()** function to check the values in the three duration columns('Administrative_Duration', 'Informational_Duration', 'ProductRelated_Duration') cannot be negative since time spent cannot be negative.

Encoding Categorical Data
As most machine learning models exclusively handle numerical variables, it becomes crucial to pre-process categorical variables. It is necessary to transform these categorical variables into numerical representations so that the model can comprehend and extract meaningful insights from them.

Data Exploration

1. Explore each column.

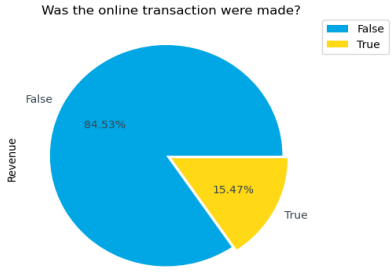


Figure 1.1: A pie chart of the total percentage of revenue

We can see that only 15.47 % of the purchase was made, which I assumed that most of the time people will search the product that they interested before making a purchase. This also implies that the data are **imbalanced.**

Hypothesis: most users do not make a purchase on their initial visit, but rather take time to research the product before making a buying decision.

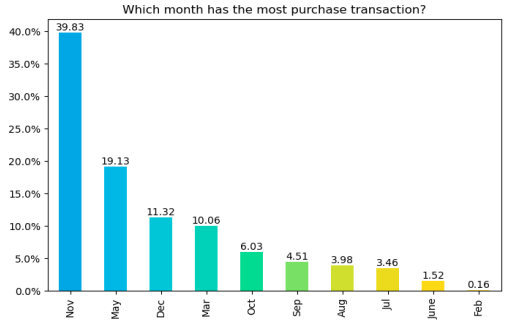


Figure 1.2: A bar chart providing insights into the month in which the majority of purchases took place.

Among completed purchase transactions, November emerges as the top month for online shopping, followed by May and December. These months witnessed significant spending by online shoppers. Now, our interest lies in determining whether these purchases were made on special days.

*Please note that the dataset only includes data for 10 months. Since the data collection spanned a period of one year, it can be assumed that the information for the remaining two months, namely April and January, is missing from the dataset

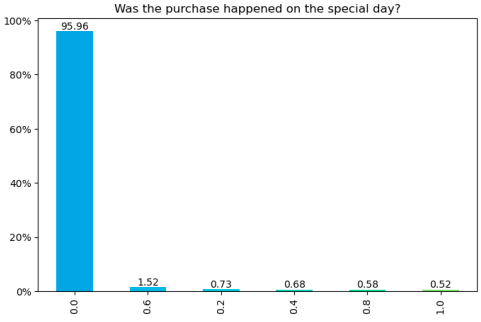
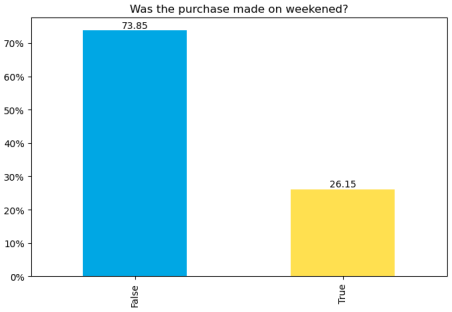
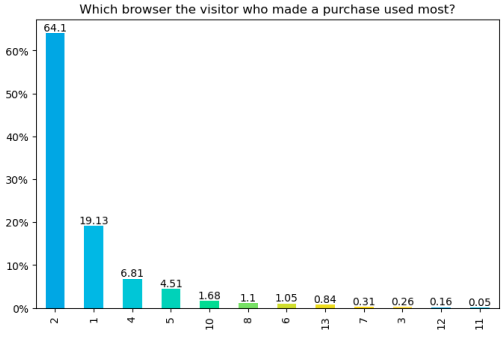


Figure 1.3: A bar chart of the total percentage of the purchase occurred during each special day.

Based on the bar chart, **the majority of purchases (96.96%) were not made on special days**. Our next question is whether these purchases were made on weekends.





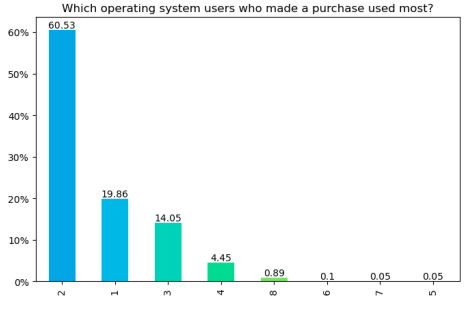


Figure 1.4: A bar chart illustrating the overall proportion of purchases made on weekends versus weekdays.

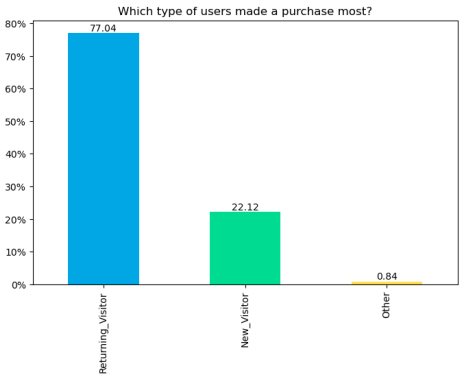
According to the findings, the majority of purchase transactions (73.85%) did not occur on weekends.

Figure 1.5: A bar chart illustrating the browser preference of purchasers.

The analysis reveals that the highest percentage of purchase transactions (64.1%) was attributed to browser no.2. In comparison, browser no.1 accounted for 19.13% of the purchases, while browser no.4 constituted 8.61% of the total transactions.

Figure 1.6: A bar chart illustrating the operating system preference of purchasers.

The bar chart shows that a majority of purchasers (60.53%) favoured operating system no.2, with 19.86% opting for operating system no.1 and 14.05% choosing operating system no.3.



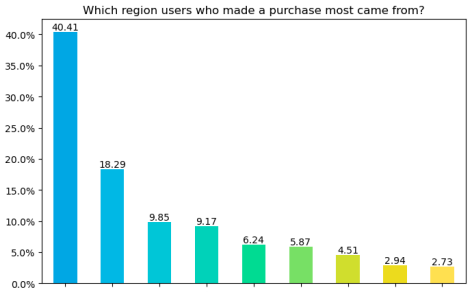


Figure 1.7: A bar chart depicting the user type that made the highest number of purchases.

The analysis reveals that returning visitors accounted for the majority of purchases (77.04%), followed by new visitors (22.12%), and others (0.84%). This finding supports the hypothesis that **most users do not make a purchase on their initial visit**, but rather take time to research the product before making a buying decision.

Figure 1.8: A bar chart displaying the distribution of purchasers across different regions.

The majority of purchasers (40.41%) were from region no.1, while region no.3 accounted for 18.29% of the purchasers, and region no.2 comprised 9.85% of the total.

PageValues Density

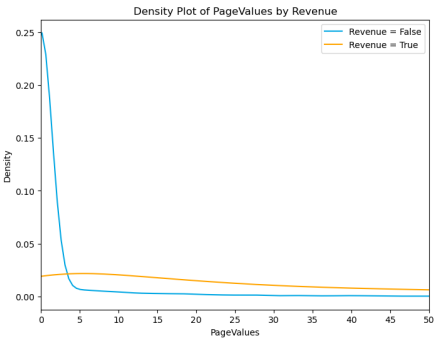


Figure 1.9: A density chart of the total percentage of regions where primary school students can access the Internet.

From the density graph, it is evident that as the PageValues increase beyond 5, the likelihood of a visitor making a purchase also increases.

*Page value is calculated based on the notion that pages with higher values are typically associated with critical pages like checkout pages or pages that come before the checkout process.

Right-skewed distribution

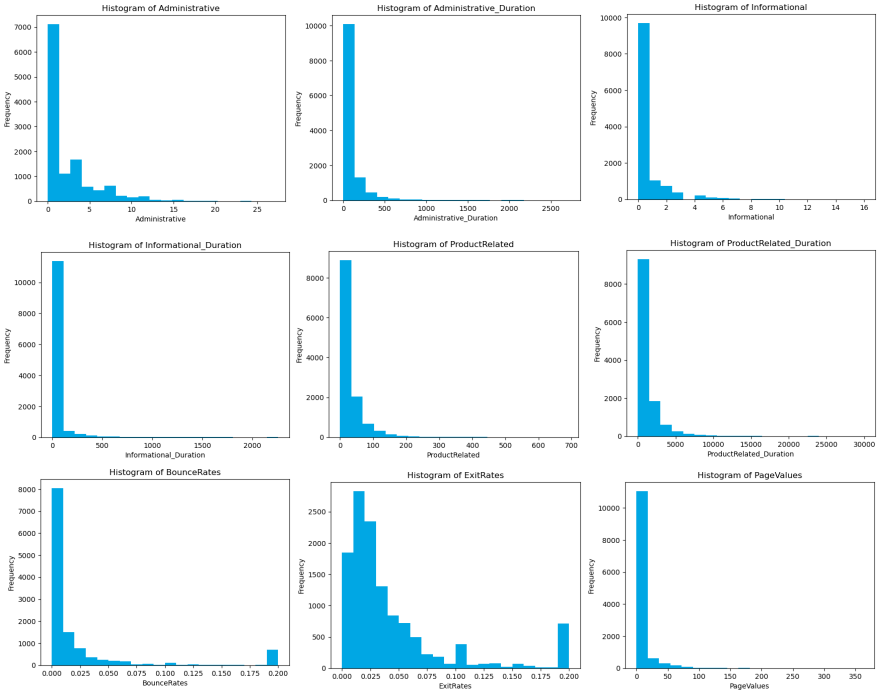


Figure 1.10: histograms of numerical features

This right-skewed distribution will affect the performance of classification models in different ways by causing imbalanced classes, misclassifying outliers, violating assumptions of certain algorithms, and affecting the optimal decision threshold.

Nevertheless, we can mitigate this by using various techniques.

2. Explore relationships between all pairs of attributes.

```
Revenue                    1.000000
PageValues                 0.492560
ProductRelated_Duration    0.165053
ProductRelated             0.161482
Administrative             0.139215
Informational              0.097431
Administrative_Duration    0.096968
Informational_Duration     0.073033
Weekend                    0.029344
Browser                    0.023970
TrafficType               -0.005047
Region                    -0.011658
OperatingSystems          -0.014677
SpecialDay                -0.082328
BounceRates               -0.150705
ExitRates                 -0.207108
Name: Revenue, dtype: float64
```

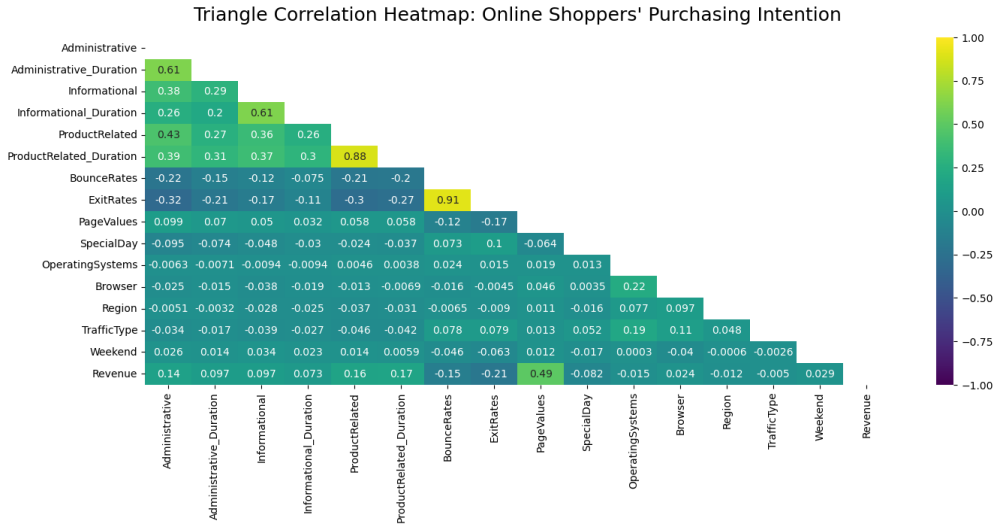


Figure 2.1: Triangle Correlation Heatmap

From PageValues density graph, I was curious if Pagevalues has a positive correlation with other features. Then, I constructed the correlation heatmap and found that PageValues are positively correlated with Administrative_Duration, Informational_Duration, and ProductRelated_Duration. The correlation analysis reveals the following noteworthy relationships. Especially higher page values, longer durations on product-related pages, and more visits to administrative and informational pages tend to be associated with higher revenue. On the other hand, higher bounce rates and exit rates are linked to lower revenue. It's important to remember that correlation does not imply causation and further analysis is needed to determine the underlying factors influencing revenue.

Hypothesis: The higher the PageValues, the longer the pageview duration.

PageValues and Revenue

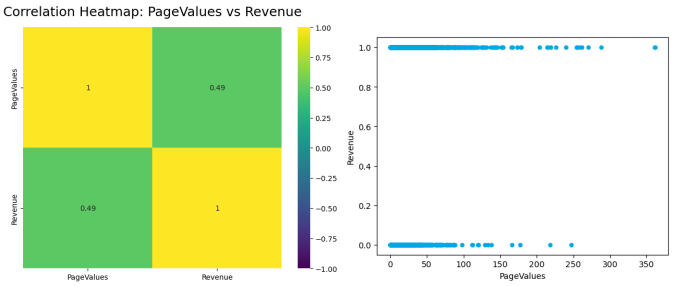


Figure 2.2: A correlation heat map and a scatter plot of PageValues and Revenue

Hypothesis: higher page values lead to higher revenue.

Explanation: The positive correlation (**0.492560**) between PageValues and Revenue suggests that the revenue is also likely to increase as the page values increase. This implies that engaging and valuable content on web pages may attract more customers and result in

PageValues and ProductRelated_Duration

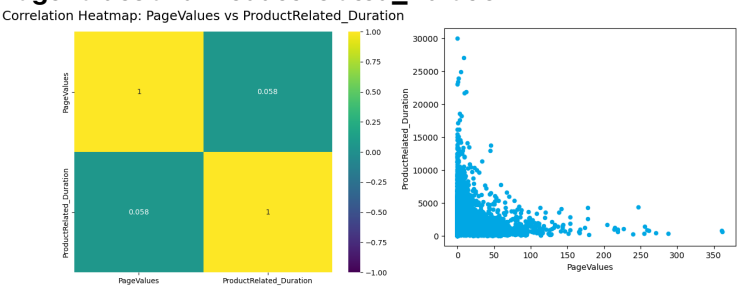
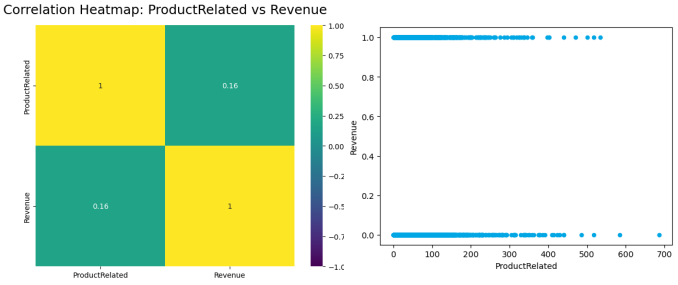


Figure 2.3: A correlation heat map and a scatter plot of PageValues and ProductRelated_Duration

Hypothesis: as 'PageValues' increase, there is a tendency for the 'ProductRelated_Duration' to increase as well.

Explanation: There is a weak positive correlation (**0.058**) between PageValues and ProductRelated_Duration', suggesting that the revenue is also likely to increase as the page values increase.

ProductRelated and Revenue



ProductRelated_Duration and Revenue

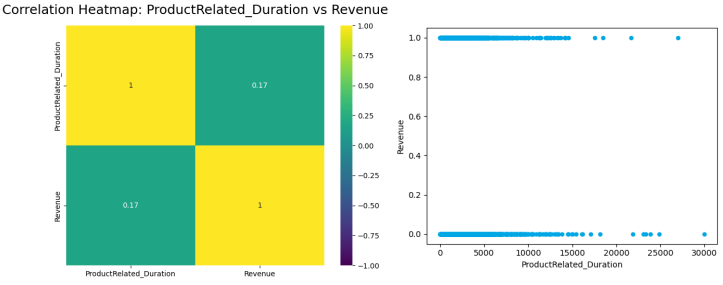


Figure 2.4: A correlation heat map and a scatter plot of ProductRelated and Revenue

Hypothesis: longer durations on product-related pages positively affect revenue.
Explanation: The positive correlations of ProductRelated_Duration (**0.165053**) and ProductRelated (**0.161482**) with Revenue indicate that spending more time on product-related pages and visiting more product-related pages could contribute to higher revenue. This suggests that customers who spend more time exploring products are more likely to purchase, thus positively influencing revenue.

Figure 2.5: A correlation heat map and a scatter plot of ProductRelated_Duration and Revenue.

Hypothesis: higher the PageValues, the longer the ProductRelated_Duration
Explanation: There is a positive correlation (**0.17**) between PageValues and ProductRelated_Duration', suggesting that the revenue is also likely to increase as the ProductRelated_Duration increase.

BounceRates and ExitRates

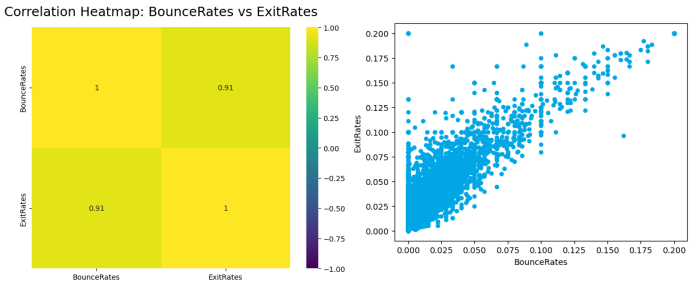


Figure 2.6: A correlation heat map and a scatter plot of BounceRates and ExitRates

Hypothesis: higher bounce rates are associated with higher exit rates
Explanation: The correlation between BounceRates and ExitRates is strong (**0.913004**), suggesting that visitors who quickly leave the website after landing on a page are more likely to continue exiting the website after navigating through multiple pages. This correlation implies that a high initial bounce rate has a negative impact on user engagement, potentially leading to lower conversion rates and revenue.

BounceRates and Revenue

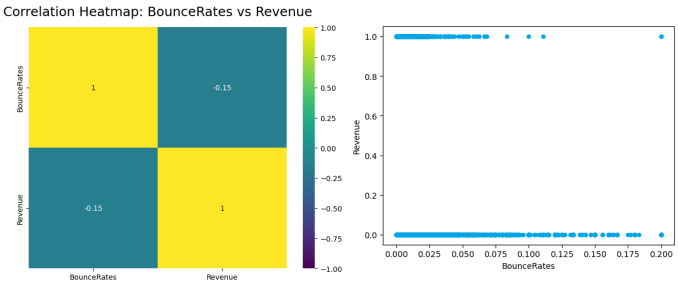


Figure 2.7: A correlation heat map and a scatter plot of BounceRates and Revenue

Hypothesis: higher bounce rates have a slight negative impact on revenue.

Explanation: The correlation coefficient between BounceRates and Revenue is **-0.150705**.Websites with higher bounce rates tend to have slightly lower revenue. Improving user engagement and minimizing bounce rates can potentially increase revenue.

ExitRates and Revenue

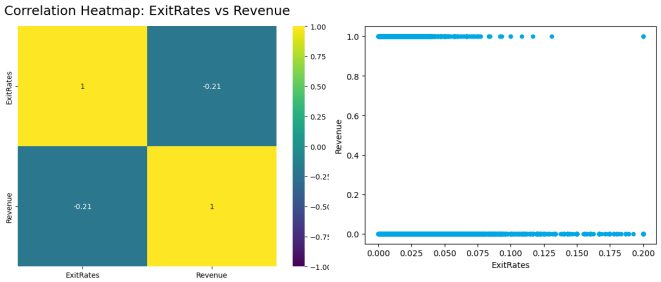


Figure 2.8: A correlation heat map and scatter plot of ExitRates and Revenue

Hypothesis: higher bounce rates and exit rates negatively impact revenue.

Explanation: The negative correlations of BounceRates (**-0.150705**) and ExitRates (**-0.207108**) with Revenue indicate that higher bounce rates and exit rates are associated with lower revenue. This suggests that visitors who quickly leave the website without further interaction or exit after visiting a page are less likely to convert into customers, resulting in reduced revenue.

Informational_Duration and Revenue

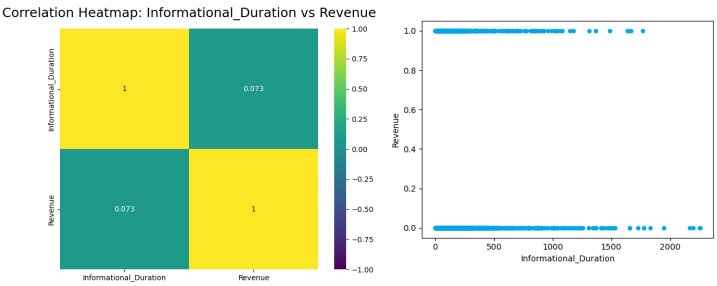


Figure 2.9: A correlation heat map and a scatter plot of Information_Duration and Revenue

Hypothesis: longer durations on informational pages have a positive impact on revenue

Explanation: The correlation coefficient between Informational_Duration and Revenue is 0.073033. Visitors who spend more time engaged with informational content are more likely to generate higher revenue. Providing valuable and engaging informational content can lead to increased revenue.

ProductRelated and ProductRelated_Duration

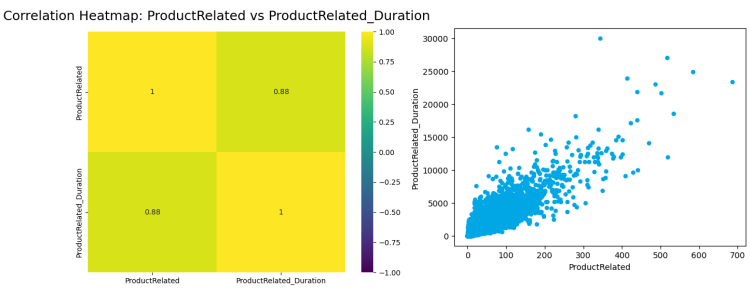


Figure 2.10: A correlation heat map and a scatter plot of PageValues and Revenue

Hypothesis: higher engagement with product-related pages is positively associated with longer durations spent on those pages.
Explanation: The correlation coefficient between ProductRelated and ProductRelated_Duration is 0.88. Visitors who navigate through more product-related pages tend to spend more time exploring and evaluating the products, indicating a strong positive correlation between ProductRelated and ProductRelated_Duration. This suggests that offering a wide range of appealing product-related content can enhance visitor engagement and revenue

Informational and Informational_Duration

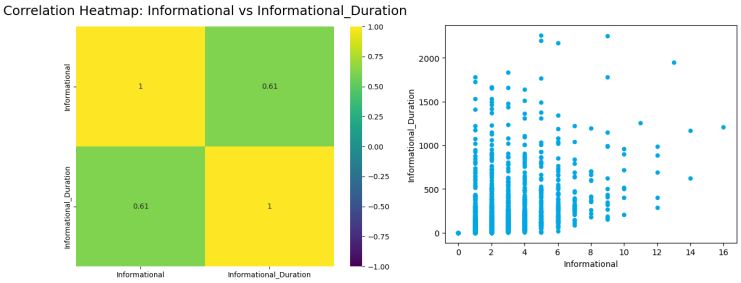


Figure 2.11: A correlation heat map and a scatter plot of Informational and Informational_Duration

Hypothesis: Higher engagement with informational pages is positively correlated with longer durations spent on those pages

Explanation: The correlation coefficient between Informational and Informational_Duration is 0.61. Visitors who navigate through more informational content tend to spend more time consuming the information. Providing valuable and informative content can enhance visitor engagement, leading to increased time spent on informational pages and potentially higher understanding and interest in the subject matter.

Data Modelling

By visualising data, I found various issues in the dataset. It's time to resolve it before running the model.

Addressing right-skewed distribution
To address the issues caused by right-skewed distribution in the k-NN model, I suggested using techniques such as oversampling or undersampling the minority class, using weighted distance metrics to adjust for the imbalanced class distribution, or using feature scaling to normalise the data and reduce the impact of outliers. It's important to note that the best approach will depend on the specific dataset and the goals of the analysis. However, due to the time limit of the report, I have not applied these techniques to the dataset.

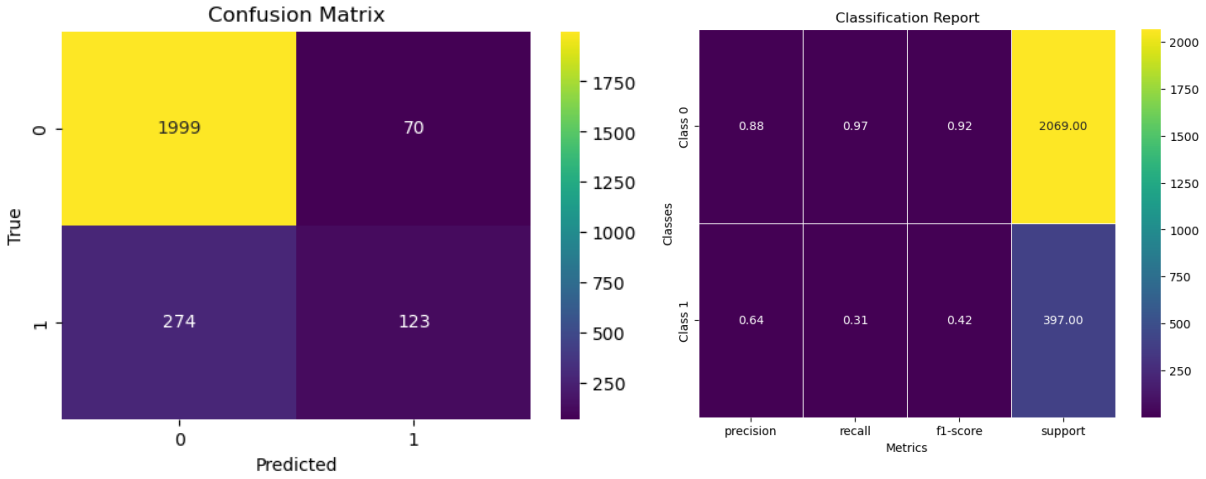
Addressing imbalanced data
To address the imbalanced data issue, I suggest using resample method. Resampling is a technique where we draw multiple samples from the training dataset to further analyse and refine a specific model. However, due to the time limit of the report, I have not applied this technique to the dataset. This process helps us gain additional insights, enhance accuracy, and estimate uncertainty by iteratively re-fitting the model with the collected samples.

Generating Train and Test Set
Training Data: Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, PageValues, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend
Target Variable: Revenue

Metrics
When selecting a metric for our model, we should consider the implications of Type I and Type II errors.
Type I error occurs when we incorrectly predict a customer will make a purchase, while **Type II** error occurs when we incorrectly predict a customer will not make a purchase.

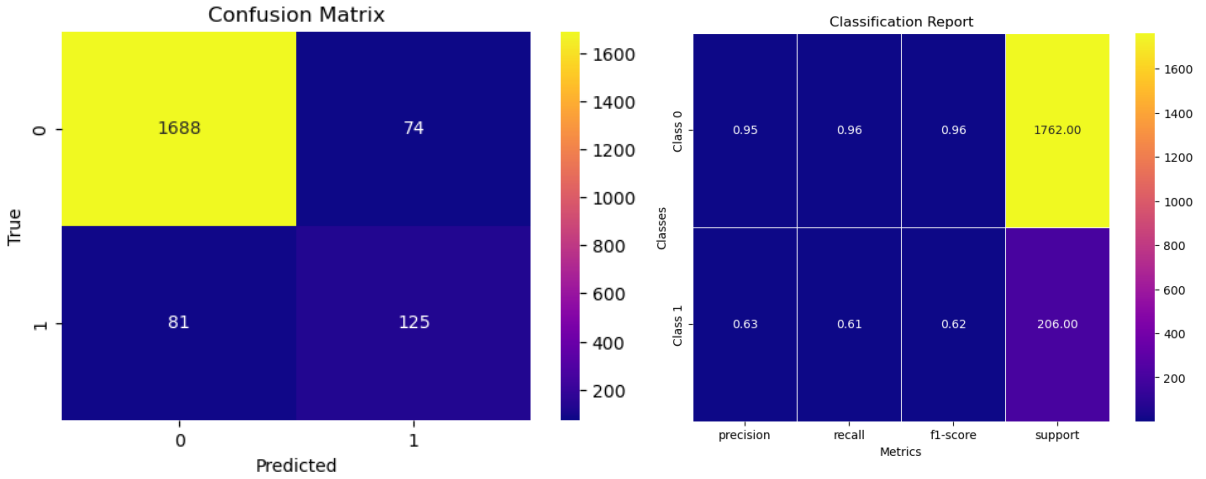
Classification Models
1. **K-nearest**

The K-nearest neighbours (KNN) classifier is a simple, yet powerful algorithm used for classification tasks. It assigns a class label to a new data point based on the majority vote of its K nearest neighbours in the training data. KNN is non-parametric, lazy-learning, and can handle both binary and multi-class classification problems.



2. Decision Tree

The decision tree classifier is a popular machine-learning algorithm for classification tasks. It creates a tree-like model where each internal node represents a feature, and each leaf node represents a class label. The algorithm splits the data based on feature values to make decisions and predict the class labels. Decision trees are interpretable, can handle both categorical and numerical data, and can capture complex relationships between features and classes.



| Discussion

Why I chose classification over clustering techniques?

Classification models are preferred for predicting online shoppers' purchasing behaviour because they can provide targeted predictions and personalised marketing strategies. They can be evaluated using metrics and trained with labelled data, enabling accurate predictions on unseen instances. Classification models also offer clear decision boundaries, aiding informed decision-making for marketing campaigns and increasing conversion rates. Clustering, although useful for exploring patterns, lacks direct insight into predicting specific outcomes like purchase behaviour.

Why do I think that the Decision Tree Model is better? What Criteria did I use?

We can compare the performance metrics of the two models based on the classification report to determine which classification method predicts if a customer will purchase a product online better.

From the provided classification report:

KNN: Class 0 (not purchasing): precision = 0.88, recall = 0.97, f1-score = 0.92 Class 1 (purchasing): precision = 0.64, recall = 0.31, f1-score = 0.42 Accuracy = 0.86

Decision Tree: Class 0 (not purchasing): precision = 0.95, recall = 0.96, f1-score = 0.96 Class 1 (purchasing): precision = 0.63, recall = 0.61, f1-score = 0.62 Accuracy = 0.92

Based on the metrics, the Decision Tree model performs better in predicting if a customer will purchase a product online. It has a higher precision, recall, and f1-scores for both classes compared to the KNN model. Additionally, the overall accuracy of the Decision Tree model is higher (0.92) compared to the KNN model (0.86). Therefore, the Decision Tree model is expected to provide better predictions regarding customer online purchases in this scenario.

After comparing the two models, we find out that the Decision Tree model performs better in predicting customer online purchases compared to the KNN model. It has higher precision, recall, and f1-scores for both classes and higher overall accuracy.

| Conclusion

After pre-processing the data, exploring the attributes and relationships, and testing and evaluating the dataset using two classification models, we have confirmed that machine learning can provide relevant predictions on whether an online user will make a purchase. This predictive power surpasses human capabilities and can be achieved using only the available electronic records and a limited set of attributes. Therefore, machine learning models can effectively predict online user purchase behaviour based on the given dataset.

Recommendation

This study helps the seller to identify what factors impact customers' decisions the most. Later, they can offer a discount or tailor their marketing strategies towards these influential factors to increase the likelihood of online purchases. Additionally, the seller can optimise website features, improve user experience, and provide personalised recommendations based on the insights gained from the classification model. This targeted approach can effectively enhance customer satisfaction and conversion rates, ultimately leading to improved business performance.

| References

1. Bibor, S. (2021). "How to Create a Seaborn Correlation Heatmap in Python." Medium. Available at: https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e.
2. Python Graph Gallery. (n.d.). "Density Plot of Several Variables." Available at: https://www.python-graph-gallery.com/74-density-plot-of-several-variables.
3. Sharma, A. (2022). "Pandas Resample Tricks You Should Know for Manipulating Time Series Data." Towards Data Science. Available at: https://towardsdatascience.com/pandas-resample-tricks-you-should-know-for-manipulating-time-series-data-7e9643a7e7f3.
4. Nikolov, B. (2023). "The Role of Resampling Techniques in Data Science." KDnuggets. Available at: https://www.kdnuggets.com/2023/02/role-resampling-techniques-data-science.html.
5. Stack Overflow. (2020). "Can someone explain to me how MinMaxScaler works?" Available at: https://stackoverflow.com/questions/62178888/can-someone-explain-to-me-how-minmaxscaler-works.