# Practical Data Science
# COSC 2670 / 2738
# Assignment 2 Report

## **Model Building**: Online Shoppers Purchasing Intention

**Mrwan Fahad A Alhandi** **s3969393@student.rmit.edu.au**

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 honor code by typing "Yes": Yes

Date : 28/05/2023

Affiliations: RMIT University.

## Table of Contents

***1. An abstract summary*** **3**

***2. Introduction*** **3**

***3. Methodology*** **3**

**3.1 Data Preparation 4**

3.1.1 Numerical Attributes 4

3.1.2 Categorical Attributes 5

**3.2 Data Exploration 5**

3.2.1 Descriptive Statistics 5

3.2.2 Discrete Attributes 6

3.2.3 Continuous Attributes 6

3.2.3 Categorical Attributes 7

3.2.4 Relationship Between Attributes 8

**3.3 Data Modelling 10**

3.3.1 Feature Selection 10

3.3.2 KNN and Decision Tree parameters 11

***4. Results 11***

***5. Discussion 11***

***6. Conclusion 12***

***References 12***

## 1. An abstract summary

- Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.
- KNN and Decision Tree model accuracies of predicting Revenue is 0.88.
- Feature engineering is required for better model performance.
- Many attributes, small data set. Therefore, it is harder to train a good model. This is due to the curse of dimensionality, where the number of possible combinations grows exponentially with the number of features, making the dataset sparse.

## 2. Introduction

Grasping the essence of human behaviour from data and decoding the critical psychological factors that influence choices is not a straightforward task. In fact, the core of many artificial intelligence, machine learning and reinforcement learning algorithms are optimization problems (minimizing a cost function) by trying to find the least and most important features that determine the decision from human behaviour in this context (Géron, 2023). Identifying these characteristics can be extremely difficult, if not unfeasible, due to constraints related to data gathering or because the intricate nature of the system in question (in this case, human behaviour). The system comprises numerous elements that impact the decision-making process. The perspective of chaos theory suggests that seemingly small life events can have a larger impact on psychology, mental health, and human behaviour (Carroll, 2022).

The study of consumer behaviour includes two parts. The first is to study consumers' emotional, mental, and behavioural responses. The second is to study the processes they use to choose, use (consume), and dispose of products and services (Radu, 2023). It is crucial for businesses to understand consumer behaviour to create effective marketing strategies that can influence consumers' decision-making processes.

In this report, a given dataset from UCI Machine Learning Repository about Online Shoppers Purchasing Intention are prepared, explored, feature engineering and selection and finally a predictor model is built to predict whether the consumer will purchase or not based on the feature selected. On other words, predicting the target variable "Revenue" which is a binary nominal attribute. Therefore, this is a supervised classification task. The raw data features include the type of pages the user/session explored and the duration on these pages. Google analytics website features which are bounce, exit rates, and page values. A feature of whether the visit is close to a special day Eg. Valentine's day. The dataset also includes operating system, browser, region, traffic type, visitor type, weekend, and month of the year. Please refer to the notebook for further information about the attributes.

## 3. Methodology

It is important for the data scientist to first make sure that the attributes given are clear and well understood from the data source. If not, then further investigations must be done before proceeding to next steps. In the dataset given, an assumption that the data scientist possesses google analytics background seems to be made. For example, in the dataset source, Bounce

Rate was not explained in detail and how it is calculated. After the data scientist understood the attributes, the task and it is goal must be clear. Then, the data scientist can proceed with the three important steps of a data science project life cycle which are data preparation, exploration, and modelling. The methodology that was taken when doing these steps is to ask questions and to have a plausible hypothesis before implementing any change on the data or to make conclusions about them. Using data summarisation: descriptive statistics and visualisation, the plausible hypothesis can be confirmed.

## 3.1 Data Preparation

Data Preparation is an important step and the task of it is to make the data clean for the model. Doing so is tremendously important because the models will perform better, and the data scientist will lose less time trying to fix strange errors. The first step here is to make sure the data imported in the notebook matches the data source. Then, the focus is on the content of the variables and whether errors exist. The errors are typos, entry errors (unlogical errors), missing values. When using functions like value_counts(), the dataset studied does not seem to have these type of errors. Therefore, the focus in this step was on two parts: to investigate outliers for numerical attributes and to encode categorical attributes.

### 3.1.1 Numerical Attributes

Outliers are data points that significantly differ from the other observations in a dataset. They must be addressed because they can heavily skew statistical measure and data modelling. They can influence the mean, impact model accuracy, statistical analysis, or misinterpretation of data.

The following steps were taken to deal with outliers:

1- Using boxplot visualisation to determine whether outliers exist.

2- Finding those outliers using interquartile range method (IQR).

3- Determining from these outliers how many True and False of the target variable “Revenue”.

4- Calculating the percentage of False instances of outliers by: (False instances)/(False instances + True instances).

5- If the percentage of False instances from the outliers are significantly different than the percentage of False instances of the whole dataset which is 84.5%, then this mean that the outliers might influence Revenue. Therefore, the outliers must be left as it is. Otherwise, proceed to step 6.

6- If the percentage of False instances from the outliers are the same/close percentage of False instances of the whole dataset, then this mean that the outliers have no influence on Revenue and must be dealt with.

7- Whether to replace the outlier with the mean and the median depends on the distribution of the attribute. If the distribution is right or left skewed, the outlier is replaced with the median. If it is normally distributed, the outlier is replaced with the mean.

| Attribute | # of Outliers instances | False Outliers % | Replacement method |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Administrative | 404 | 72% | Median |
| Administrative_Duration | 1172 | 75% | Median |
| Informational | 2631 | 77% | Median |
| Informational_Duration | 2405 | 77% | Median |
| ProductRelated | 987 | 71% | Median |
| ProductRelated_Duration | 961 | 70% | Median |
| BounceRates | 1551 | 98% | Nothing |
| ExitRates | 1099 | 99% | Nothing |
| PageValues | 2730 | 56% | Nothing |

**Table 1**: Outliers Information and Replacement

### 3.1.2 Categorical Attributes

Encoding provides multiple advantages such as machine learning algorithms compatibility, improve in efficiency and performance and better feature representation. There are numerical and text categorical attributes. Numerical attributes such as SpecialDay, OperatingSystems, Browser, Region and TrafficType were not encoded. The attributes that are encoded are VisitorType, Weekend and Month.

There are different methods to encode such as label, one-hot, binary, dummy, target (Mean) and hashing encoding. The choice of encoding depends on the model algorithm and the preferred representation of the attribute. Since the two models that will be used are K Nearest Neighbours and Decision Tree, one-hot encoding is the best choice here (Malla, 2018).

## 3.2 Data Exploration

The approach taken for every attribute is to first have a descriptive statistics information to help making conclusions after producing a visualisation. Then, depending on the type of the attribute, an appropriate visualisation is used. For attributes that are continuous a violin plot is used. Violin plots are great because they combine a box plot and a density plot (Ngo, 2018). For attributes that are discrete, a line chart is used. For categorical attributes, either frequency distribution or pie charts are used. If the attributes have small number of categories, pie charts are used.

### 3.2.1 Descriptive Statistics

| Attribute | mean | std | min | 25% | 50% | 75% | max | True revenue median | False revenue median |
|---|---|---|---|---|---|---|---|---|---|
| Administrative | 1.9 | 2.6 | 0 | 0 | 1 | 3 | 10 | 2 | 0 |
| Administrative_Duration | 35.6 | 55.7 | 0 | 0 | 7.5 | 54.8 | 233.1 | 17.2 | 0 |
| Information | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Informational_Duration | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ProductRelated | 21.4 | 18.5 | 0 | 7 | 18 | 29 | 84 | 20 | 16 |
| ProductRelated_Duration | 772.6 | 765 | 0 | 184.1 | 598.9 | 1109.2 | 3382.3 | 742.1 | 510.2 |
| BounceRates | 0.02 | 0.05 | 0 | 0 | 0.003 | 0.02 | 0.2 | 0 | 0.004 |
| ExitRates | 0.04 | 0.05 | 0 | 0.01 | 0.03 | 0.05 | 0.2 | 0.02 | 0.03 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| PageValues | 5.9 | 18.6 | 0 | 0 | 0 | 0 | 361.8 | 16.8 | 0 |
| SpecialDay | 0.06 | 0.2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

**Table 2**: Descriptive Statistics

True and false revenue median refers to the median of those instances of the attribute which resulted in true or false revenue.

### 3.2.2 Discrete Attributes

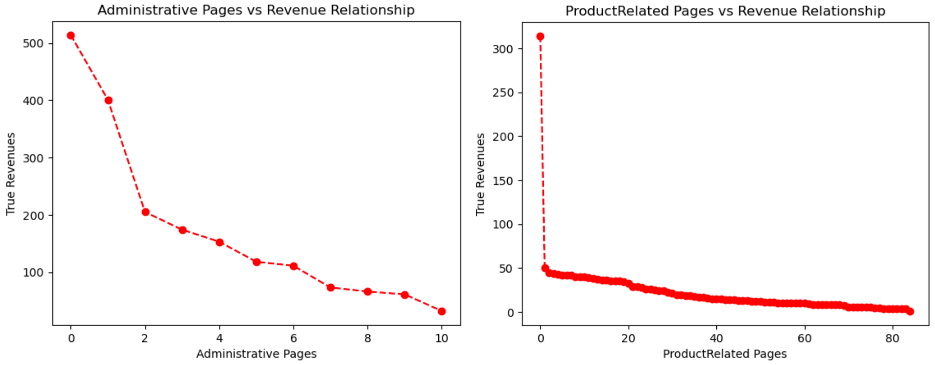


**Figure 1**: Discrete Attributes Relationship with Revenue

From the figure, it can be observed that both product and administrative pages have an inverse relationship. As the number of pages increases, as the number of true revenues decreases.

### 3.2.3 Continuous Attributes

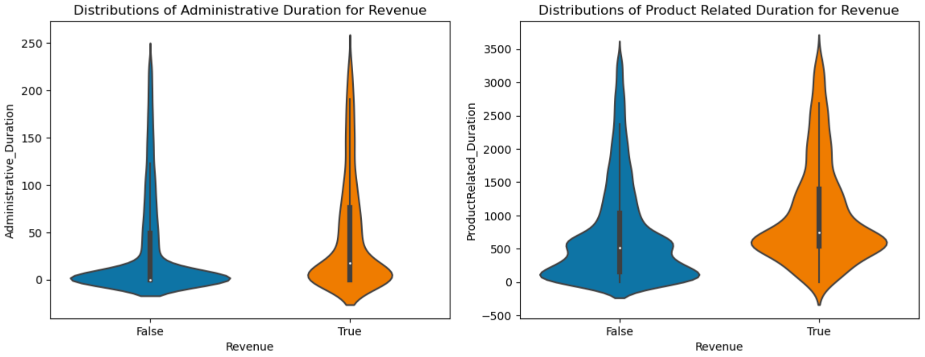


**Figure 2**: Pages Duration Relationship with Revenue

For administrative pages, most of the false values are distributed around the median 0 while most of the true values are around the median 17.2. Comparing this with product pages, most of the false values are also around 0 but there are large number of data points around the false median which is 16 and most of the true values are around the median 20. In general, it can be observed that an increase to product pages result in more true revenue than an increase in administrative pages.

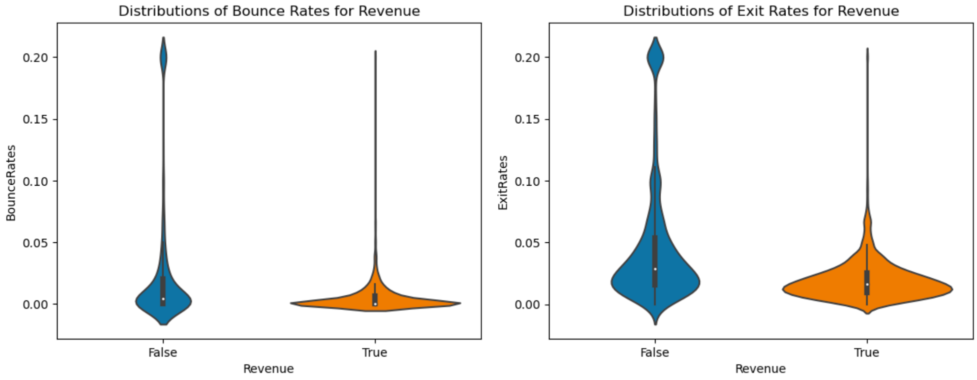


**Figure 3**: Bounce and Exit Rates Relationship with Revenue

For bounce rate, most of the true values are around the median 0. This is expected because as bounce rate increases, as the percentage of visitors who enter the site from that page and then leave without triggering any other requests increases. A lower value might indicate that the user is interested and is further exploring the website. The same thing for exit rate can be seen. The true values for exit rate are distributed in lower values than false distribution. A lower exit rate indicates that users are more likely to continue browsing or navigating to other pages before leaving (Vasile, 2019).

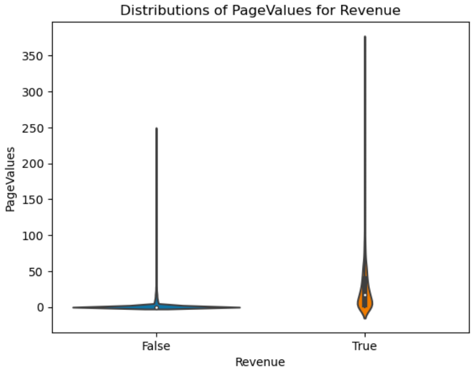


**Figure 4**: PageValues Relationship with Revenue

This is expected. If page value is 0, then this means that the user is most likely not interested in a particular webpage and so the user will not visit it again before a transaction. But if the user is interested in the webpage, more than one visit is expected. As the number of re visits increases, as the number of true revenue increases.

### 3.2.3 Categorical Attributes

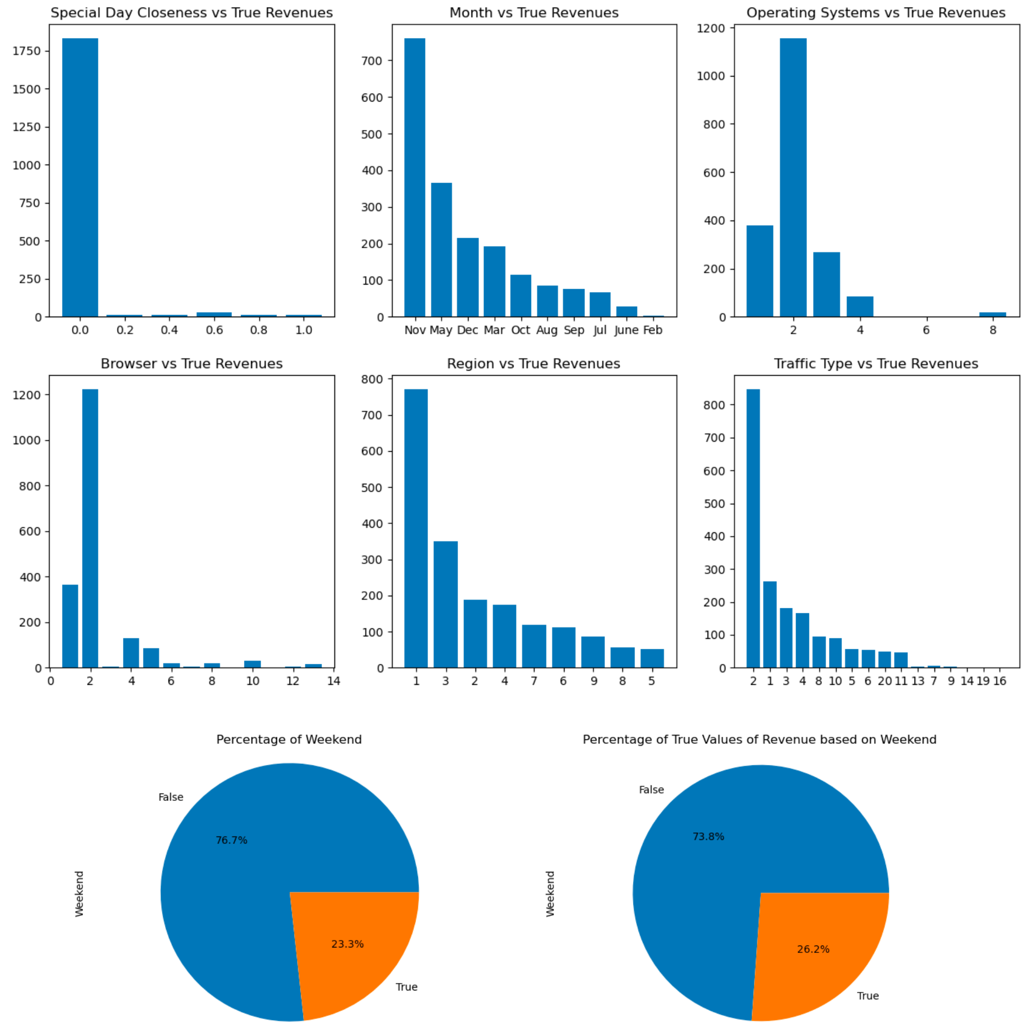


**Figure 4**: Categorical Attributes Relationship with Revenue

All categorical attributes seem to lean toward one category more than others. All of them are not normally distributed.

### 3.2.4 Relationship Between Attributes

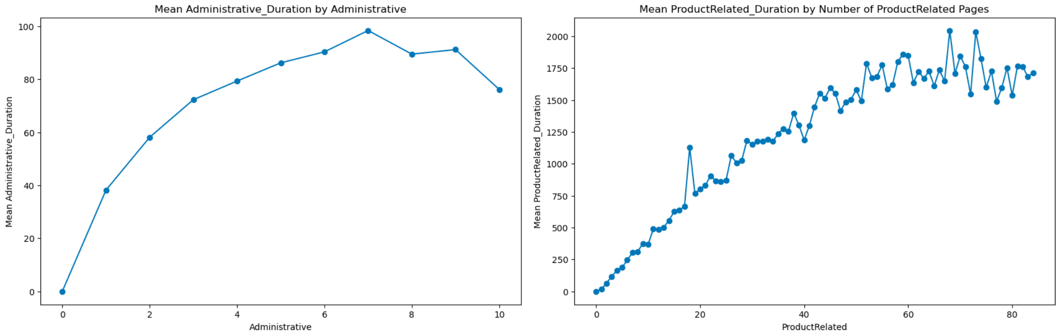


**Figure 5**: Number of Pages vs Time Spent

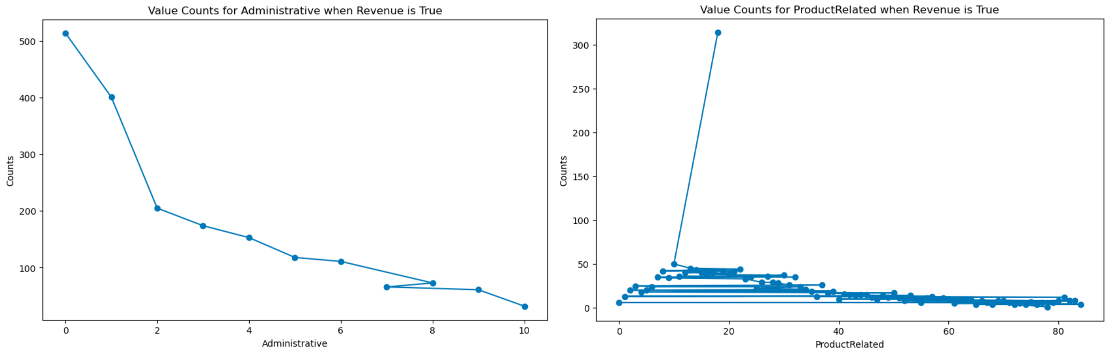


**Figure 6**: Number of Pages vs True Revenue

In figure 5, both administrative and product related pages seems to have an upward trend with the mean duration of a particular number of pages. But does this increase in duration result in more revenue? Based on figure 6, the answer is no. For administrative, there's a downward decrease for the number of true revenues as pages increases. For product related pages, most of high number true revenue are between 0 and 40 pages.

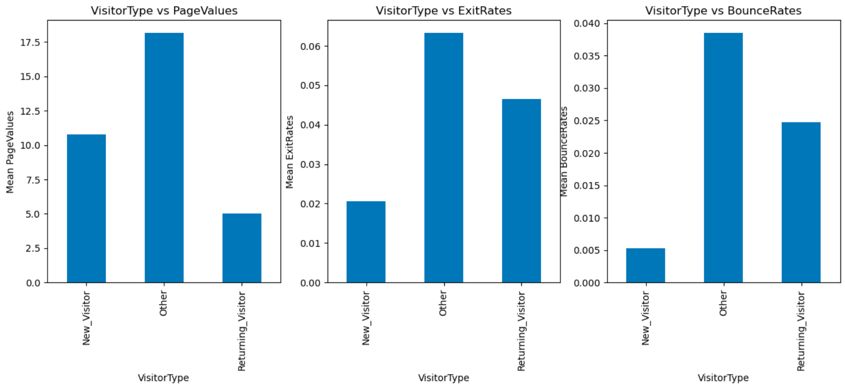


**Figure 7**: VisitorType vs PageValues, ExitRates and BounceRates

In figure 7, the goal was to investigate the behaviour of different visitor types. New and other visitors seem to review same pages more than returning visitors. Also, new visitors seem to get more interested in viewing more pages as exit and bounce rate are the smallest.

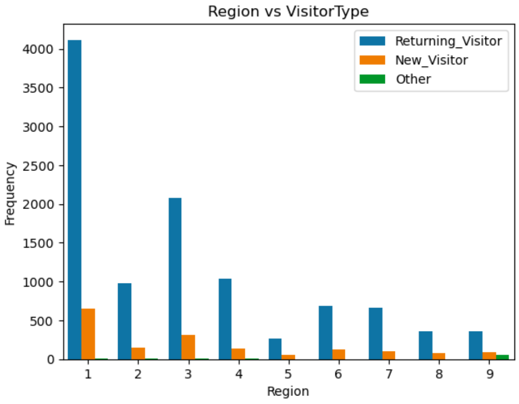


**Figure 8**: Region vs VisitorType

Returning visitors have the highest true values overall. Regions 1 and 3 have the most of them. For new visitors, regions 1 and 3 also have the highest of them. For others, region 9 have the highest.

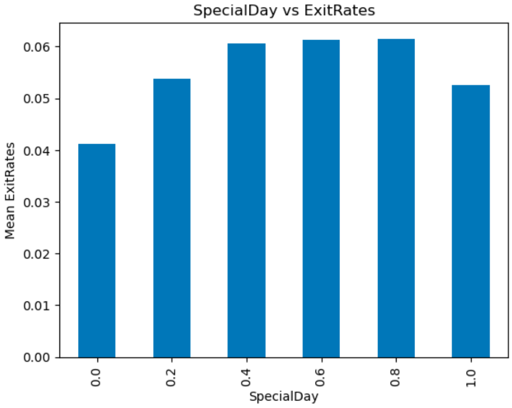


**Figure 9**: Special Day vs ExitRates

The hypothesis was that as it gets close to a special day, as the number of pages that the user will visit before exiting will increase. In figure 9 however, this is not the case. There's an increase in mean exit rates as it is closer to a special day.

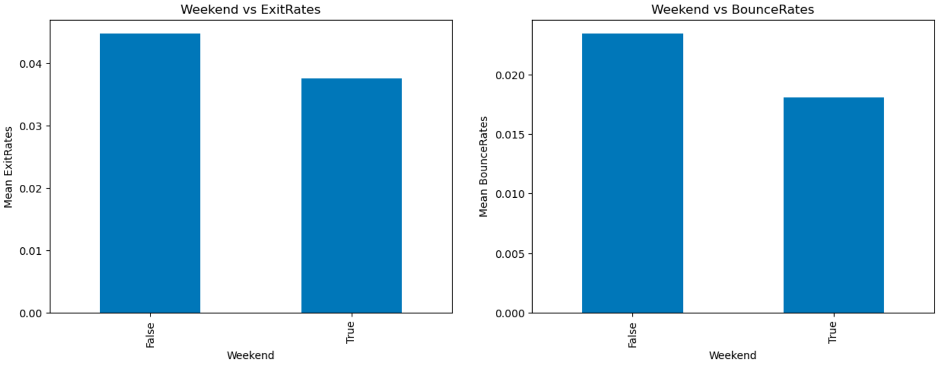


**Figure 10**: Weekend vs Bounce and Exit Rates

The hypothesis was that when it is weekend, people visit more pages. In figure 10 however, this is not the case. The mean bounce and exit rate are insignificantly different.

## 3.3 Data Modelling

The two models that were considered are Decision Tree and K nearest neighbours.

### 3.3.1 Feature Selection

For feature selection, a basic greedy search algorithm is used. It iteratively selects a subset of features from the original dataset and evaluates the performance of a K-Nearest Neighbors classifier using the selected features. The algorithm aims to find the subset of features that achieves the highest score (accuracy) on a held-out test set. The resulted selected features are: ['Month_Mar', 'Month_May', 'VisitorType_Other', 'Region', 'Month_Sep', 'PageValues', 'SpecialDay', 'VisitorType_Returning_Visitor', 'Administrative', 'Month_Jul', 'Weekend_True', 'Weekend_False','OperatingSystems'].

### 3.3.2 KNN and Decision Tree parameters

For both K nearest neighbours and Decision Tree parameters, K-fold cross validations has been used to determine the best parameters. The value of the parameter that minimises the misclassification error is chosen.

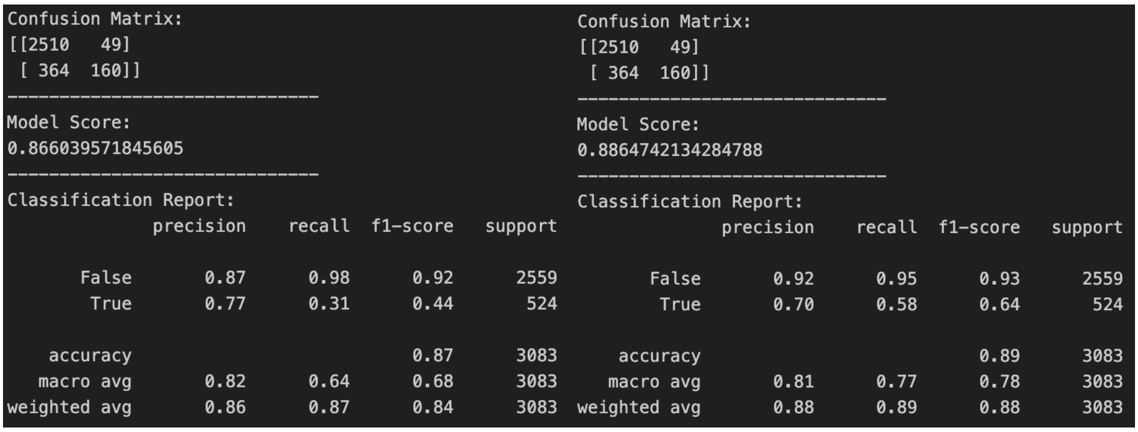
For KNN:

- The optimal number of neighbours is 13.
- The optimal number of p is 9.
- The optimal weight is uniform.

For Decision Tree:

- The optimal number of max depths is 3.

## 4. Results



```
Confusion Matrix:
[[2510   49]
 [ 364  160]]
------------------------------
Model Score:
0.866039571845605
------------------------------
Classification Report:
              precision    recall  f1-score   support

       False       0.87      0.98      0.92      2559
        True       0.77      0.31      0.44       524

    accuracy                           0.87      3083
   macro avg       0.82      0.64      0.68      3083
weighted avg       0.86      0.87      0.84      3083
```

```
Confusion Matrix:
[[2510   49]
 [ 364  160]]
------------------------------
Model Score:
0.8864742134284788
------------------------------
Classification Report:
              precision    recall  f1-score   support

       False       0.92      0.95      0.93      2559
        True       0.70      0.58      0.64       524

    accuracy                           0.89      3083
   macro avg       0.81      0.77      0.78      3083
weighted avg       0.88      0.89      0.88      3083
```

**Figure 11**: KNN and Decision Tree Results

The one on the right is KNN and on the left is decision tree. Decision tree model score is better than KNN by just 0.02. On the other hand, KNN have higher precision by 0.07 for True revenues. A model which produces no false positives have precision of 1. Decision tree recall is better than KNN by 0.27 which is large. This means that decision tree model performed much better when classifying True revenues.

## 5. Discussion

Most of the features selected from the search algorithm are categorical attributes. During the exploration stage, we have seen the distributions of continuous variable. A feature engineering by setting condition to those distribution has been tested but it worsens the model. Please refer to the feature engineering section in the notebook. It was found that almost half of the true values from the total 1908 are when bounce rate is equal to 0 and page value is larger than 0. If appropriate engineered features from continuous attributes are found, there is a possibility for a better model accuracy. In addition, compared to how many features there are in the dataset, 12,330 observations might be small to train a good model.

## 6. Conclusion

In this report, the steps of a data science project have been followed. Each of these steps contributed to the final solution. The research goal was determined. Then, the data was imported and checked that it matches the source. Then, during data preparation, the numerical attributes has been cleaned from outliers and the categorical attributes has been encoded. After that, all the attributes were explored using descriptive statistics and visualisations. Finally, two supervised machine learning algorithms has been used to predict Revenue.

## References

Géron, A. (2023) *Hands-on machine learning with scikit-learn, Keras, and tensorflow concepts, tools, and techniques to build Intelligent Systems*. Beijing: O'Reilly.

Carroll, K. (2022) *Take online courses. earn college credit. Research Schools, Degrees & Careers*, *Study.com | Take Online Courses. Earn College Credit. Research Schools, Degrees & Careers*. Available at: https://study.com/learn/lesson/chaos-theory-psychology-overview-application-examples.html#:~:text=The%20perspective%20of%20chaos%20theory%20suggests%20that%20seemingly%20small%20life,ways%20to%20adapt%20and%20grow. (Accessed: 27 May 2023).

Radu, V. (2023) *Consumer behavior in marketing - patterns, types, segmentation - Omniconvert blog*, *Omniconvert Ecommerce Growth Blog*. Available at: https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/ (Accessed: 27 May 2023).

Malla, S. (2018) *Categorical features encoding in decision trees and Knn*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/categorical-features-encoding-decision-trees-knn-sravan-malla- (Accessed: 27 May 2023).

Ngo, L. (2018) *5 reasons you should use a violin graph*, *BioTuring's Blog*. Available at: https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/ (Accessed: 28 May 2023).

Vasile, A. (2019) *Bounce rate vs exit rate - a simple visual explanation*, *Canonicalized*. Available at: https://canonicalized.com/bounce-rate-vs-exit-rate/#:~:text=A%20high%20exit%20rate%20doesn,site%20before%20they%20leave)%3B (Accessed: 28 May 2023).

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018).