Yun Kyaw
20177325
CISC 251 – FINAL PROJECT

1.0 PROPERTIES OF THE DATASET

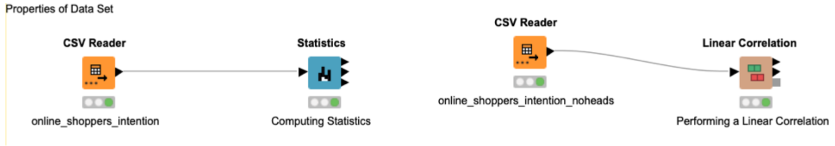
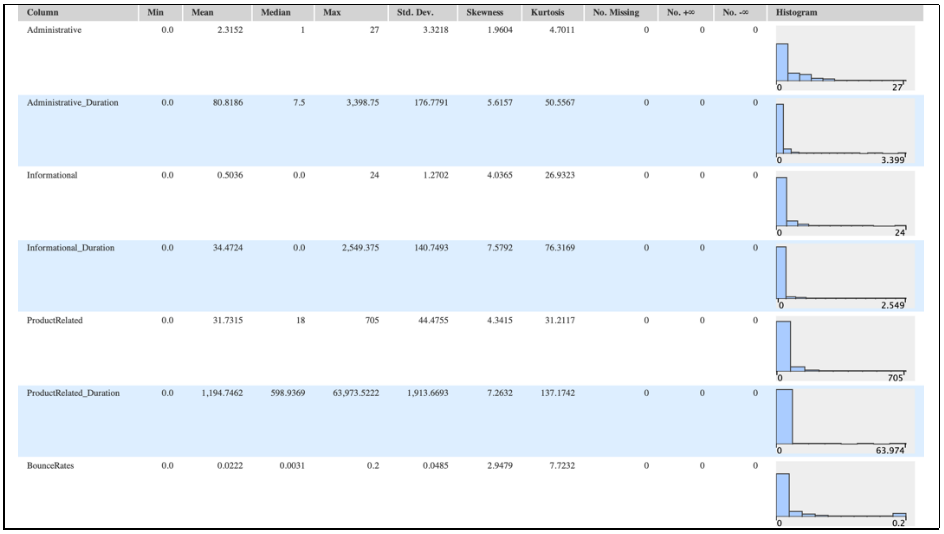


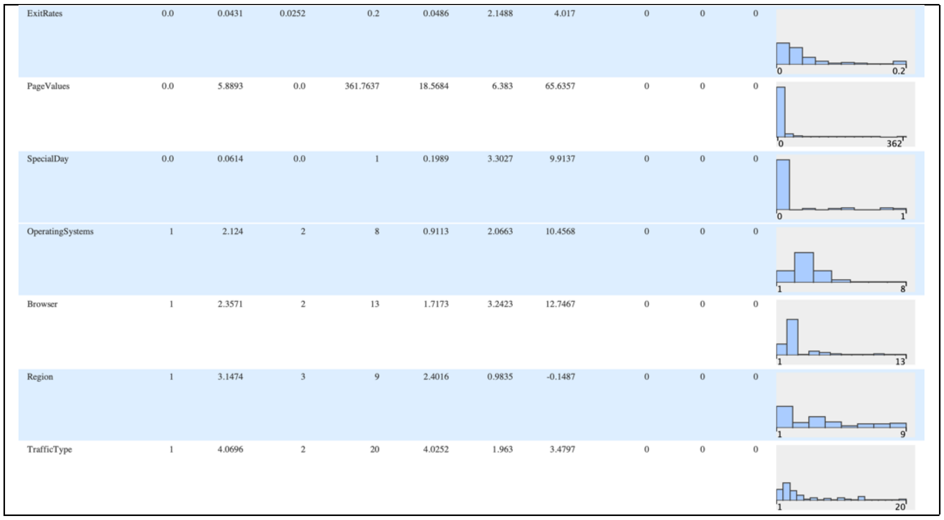
Figure 1.0: The workflow used to gain information on the dataset. The Statistics node used the online_shoppers_intention dataset while the Linear Correlation node used the online_shoppers_intention_noheads dataset.

1.1 STATISTICS

To gain information on the online_shoppers_intention.csv dataset, the statistics were computed using the statistics node in KNIME (*figure 1.0*). In computing the statistics, the goal was to gain any knowledge on the meaning of the dataset, and any tendencies of users. In performing these statistical tests, insight was gained in the different variables, such as knowing that the majority of users are from region three and use browser two (*figure 1.1*).



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Administrative	0.0	2.3152	1	27	3.3218	1.9604	4.7011	0	0	0	0 27
Administrative_Duration	0.0	80.8186	7.5	3,398.75	176.7791	5.6157	50.5567	0	0	0	0 3.399
Informational	0.0	0.5036	0.0	24	1.2702	4.0365	26.9323	0	0	0	0 24
Informational_Duration	0.0	34.4724	0.0	2,549.375	140.7493	7.5792	76.3169	0	0	0	0 2.549
ProductRelated	0.0	31.7315	18	705	44.4755	4.3415	31.2117	0	0	0	0 705
ProductRelated_Duration	0.0	1,194.7462	598.9369	63,973.5222	1,913.6693	7.2632	137.1742	0	0	0	0 63.974
BounceRates	0.0	0.0222	0.0031	0.2	0.0485	2.9479	7.7232	0	0	0	0 0.2



ExitRates	0.0	0.0431	0.0252	0.2	0.0486	2.1488	4.017	0	0	0
PageValues	0.0	5.8893	0.0	361.7637	18.5684	6.383	65.6357	0	0	0
SpecialDay	0.0	0.0614	0.0	1	0.1989	3.3027	9.9137	0	0	0
OperatingSystems	1	2.124	2	8	0.9113	2.0663	10.4568	0	0	0
Browser	1	2.3571	2	13	1.7173	3.2423	12.7467	0	0	0
Region	1	3.1474	3	9	2.4016	0.9835	-0.1487	0	0	0
TrafficType	1	4.0696	2	20	4.0252	1.963	3.4797	0	0	0

Figure 1.1: The results from computing the statistics in KNIME. The statistics node was set to having no maximum number of objects in a column and included the median.

1.2 LINEAR CORRELATION

To learn more about the dataset, a linear correlation was performed to understand any relationship between attributes, particularly considering Revenue as the independent variable, to gain any insight into any attributes that may suggest a user is a consumer. While the linear correlation did not show many strong correlations for Revenue (column 17), it does show some relationships that should be considered in attribute selection, such as column 8 which has the strongest correlation to column 17 (*figure 1.2*).

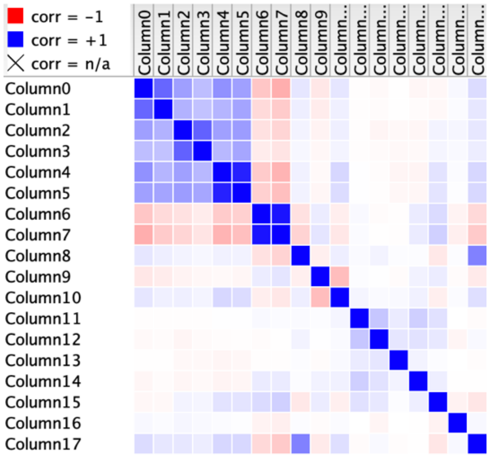


Figure 1.2: A correlation matrix from finding the linear correlation of the dataset. Computed using all column pairs and taking a right-side p-value.

2.0 CLUSTERING

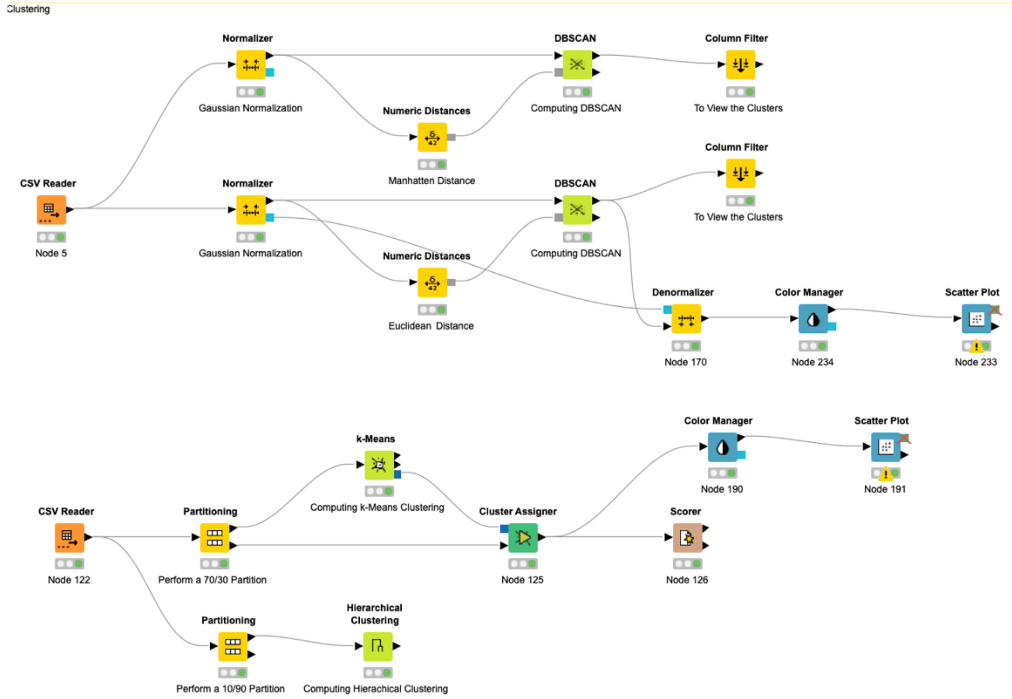
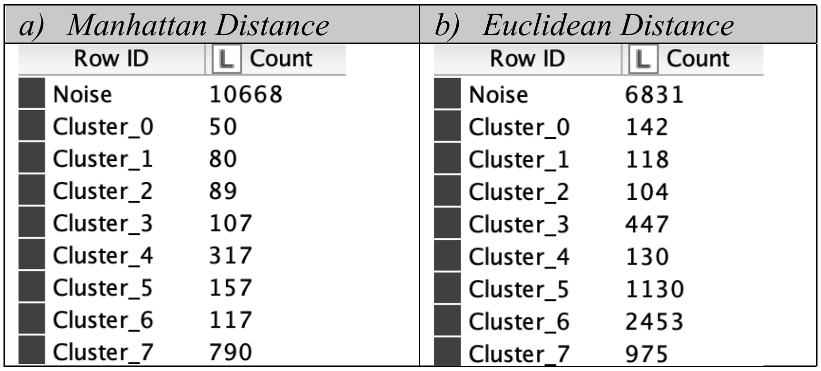


Figure 2.0: The workflow used to perform clustering on the online_shoppers_intention dataset. DBSCANs were performed using Gaussian normalization and calculated with both Manhattan and Euclidean distance. The k-Means clustering was calculated with a 70/30 partition, and the Hierarchical Clustering was calculated using 10 per cent of the data.

The goal of performing clustering tests on the dataset was to understand any similarities in the dataset. Three types of clustering tests were performed to compare any similarities found among the clusters. As clustering is an unsupervised task and the results are unlabeled, scatter plots were created of the k-Means clustering and DBSCAN to compare the clusters formed.

2.1 DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

The first clustering test used on the dataset was a density-based spatial clustering of applications with noise (DBSCAN). DBSCAN was selected as it performs well for large datasets, can handle outliers, and can separate clusters of different densities well. The dataset was normalized using a Z-score and the numeric distance was calculated using both Manhattan and Euclidean distances. The purpose of using Manhattan and Euclidean distances was to compare the clusters formed by both distances. While both did not perform very well, as they considered a large portion of the data to be noise, they both created eight clusters, suggesting eight types of users (*figure 2.1*). Using Manhattan distance, less than half of the dataset was clustered; when using Euclidean distance, approximately half of the dataset was clustered (figure 2.2b). Issues in the DBSCANs ability to cluster more than half of the dataset may lie in the data potentially being in a higher dimension, or the data truly being mostly noise.



a) Manhattan Distance

Row ID	Count
■ Noise	10668
■ Cluster_0	50
■ Cluster_1	80
■ Cluster_2	89
■ Cluster_3	107
■ Cluster_4	317
■ Cluster_5	157
■ Cluster_6	117
■ Cluster_7	790

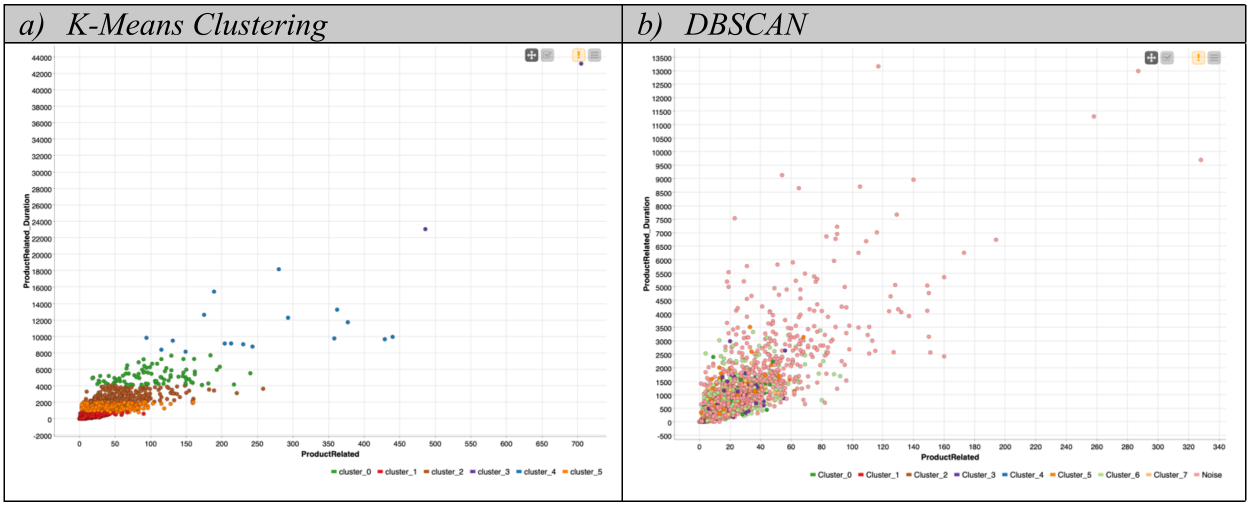
b) Euclidean Distance

Row ID	Count
■ Noise	6831
■ Cluster_0	142
■ Cluster_1	118
■ Cluster_2	104
■ Cluster_3	447
■ Cluster_4	130
■ Cluster_5	1130
■ Cluster_6	2453
■ Cluster_7	975

Figure 2.1a, b: The clusters and their count formed by using a *a)* Manhattan distance and *b)* DBSCAN.

2.2 K-MEANS CLUSTERING

The second clustering method selected was a k-means clustering, as it performs well for large datasets, and can forms clusters of different shapes and sizes. Using trial-and-error, a k-value of six was taken. This method performed better than the DBSCAN in forming clusters (*figure 2.2*) as it could cluster all the data, though the use of this clustering method is limited by the need to manually pick the number of clusters.



a) K-Means Clustering

b) DBSCAN

Figure 2.2 a, b: The clusters formed by *a)* K-Means Clustering and *b)* DBSCAN. Taking the Product Related attribute as the independent variable and Product Related Duration as the dependent variable.

2.3 HIERARCHICAL CLUSTERING

The third clustering method performed was a hierarchical clustering. Hierarchical clustering was selected as it can form clusters based on an independent variable and can create clusters that relate dependent variables to the independent variable. As hierarchical clustering requires much computational power and is ideal for smaller datasets, a stratified partition of 10 per cent of the data was used to perform the clustering. Despite taking a 10 per cent subset of the dataset, the clustering was still too large to be well-interpreted (*figure 2.3*), thus leading to a dead end.

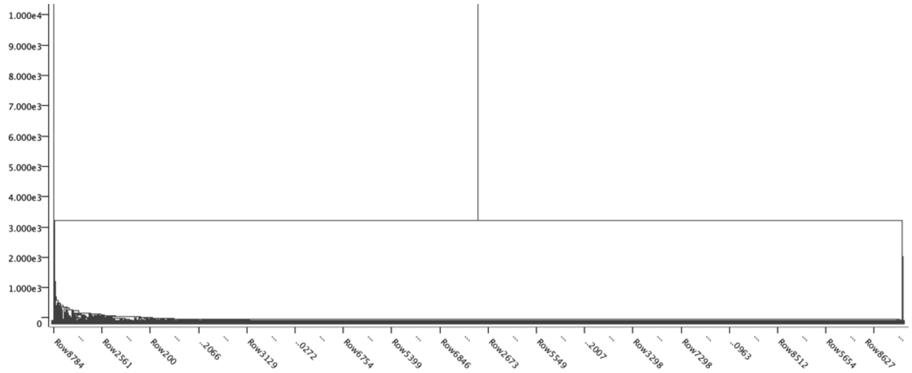


Figure 2.3: The dendrogram formed from the hierarchical clustering.

3.0 PREDICTORS

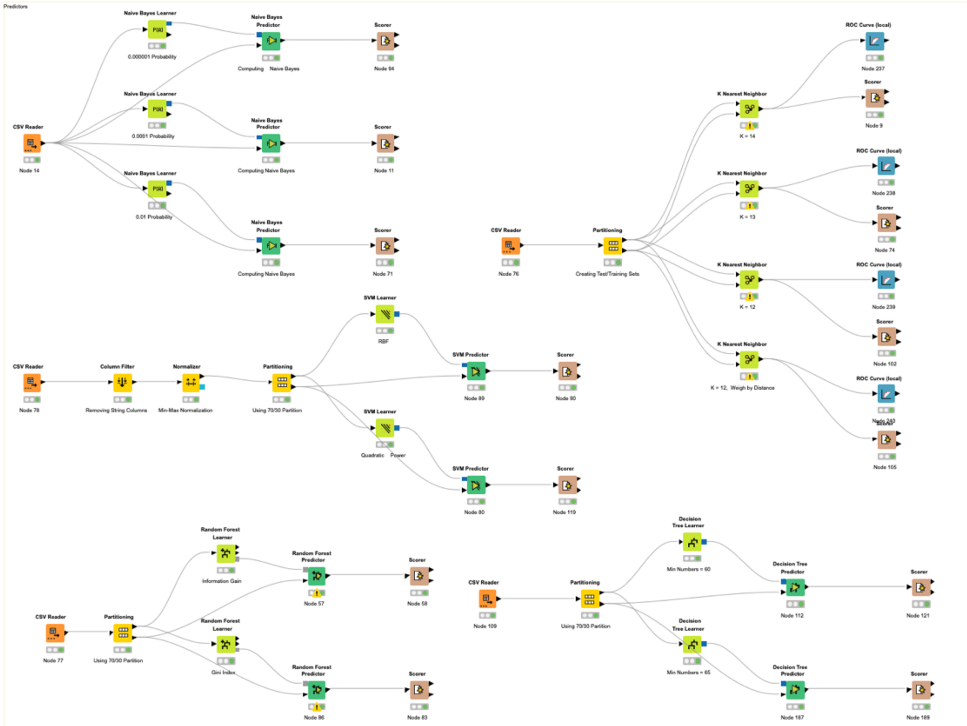


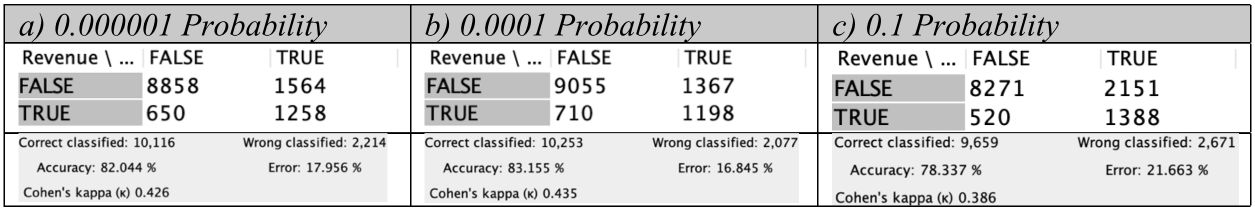
Figure 3.0: The workflow used to perform predictors on the online_shoppers_intention dataset. The predictors were partitioned with a 70/30 split. The Naïve Bayes Learner

node was adjusted taking three default probabilities. The SVM was normalized with Min-Max normalization and the SVM Learner node was adjusted with two kernels. The Random Forest Learner node was adjusted with two splitting criteria. The k-NN node was adjusted with three different k-values. The decision tree was adjusted with two minimum numbers.

The goal of performing numerous predictors was to find the three best predictors to be used for attribute selection. Multiple predictors were performed more than once with an adjustment in the test to maximize the accuracy.

3.1 NAÏVE BAYES CLASSIFIER

The first predictor implemented was a Naïve Bayes Classifier. This classifier was selected as it is a simple and fast predictor that performs well for different types of data. This classifier was performed three times, taking three different probabilities. Adjusting the default probability affected the accuracy as it changed the formula, with the given probability being used in lieu of a zero probability. As well, the probability was adjusted as the attributes are not continuous, and thus their probability is not smaller than the default probability of zero. From calculating the Naïve Bayes using different probabilities, taking a probability of 0.0001 had the best accuracy (*figure 3.1*). The accuracy could likely be improved by adjusting the probability to other values around 0.0001 and/or adjusting the other variables, such as minimum standard deviation.



a) 0.000001 Probability

Revenue \ ...	FALSE	TRUE
FALSE	8858	1564
TRUE	650	1258

Correct classified: 10,116 Wrong classified: 2,214
Accuracy: 82.044 % Error: 17.956 %
Cohen's kappa (κ) 0.426

b) 0.0001 Probability

Revenue \ ...	FALSE	TRUE
FALSE	9055	1367
TRUE	710	1198

Correct classified: 10,253 Wrong classified: 2,077
Accuracy: 83.155 % Error: 16.845 %
Cohen's kappa (κ) 0.435

c) 0.1 Probability

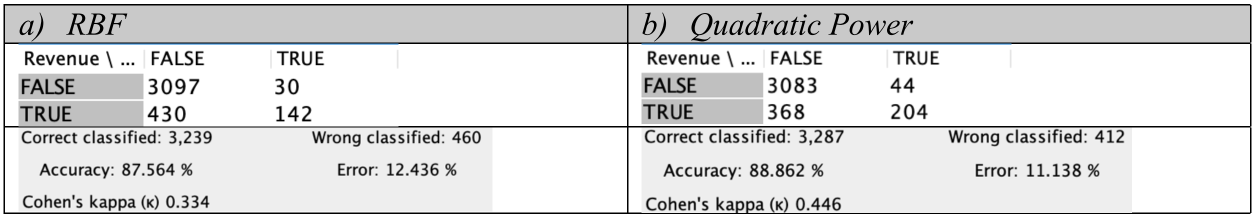
Revenue \ ...	FALSE	TRUE
FALSE	8271	2151
TRUE	520	1388

Correct classified: 9,659 Wrong classified: 2,671
Accuracy: 78.337 % Error: 21.663 %
Cohen's kappa (κ) 0.386

Figure 3.1 a, b, c: The confusion matrix (above) and the statistics (below) of Naïve Bayes Classifier given *a)* 0.000001 default probability, *b)* 0.0001 default probability, and *c)* 0.1 default probability.

3.2 SUPPORT VECTOR MACHINE

The second predictor used was a Support Vector Machine (SVM). This classification was selected as it is effective in higher dimensional spaces and works well for binary classifications. The SVM was performed taking the RBF kernel and the quadratic power kernel. While both had high accuracies, the quadratic power produced a higher accuracy (*figure 3.2*). Despite the SVM having calculated one of the best accuracies, it required a lot of computational power and time, thus it was not among the more effective predictors to use on the dataset.



a) RBF

Revenue \ ...	FALSE	TRUE
FALSE	3097	30
TRUE	430	142

Correct classified: 3,239 Wrong classified: 460
Accuracy: 87.564 % Error: 12.436 %
Cohen's kappa (κ) 0.334

b) Quadratic Power

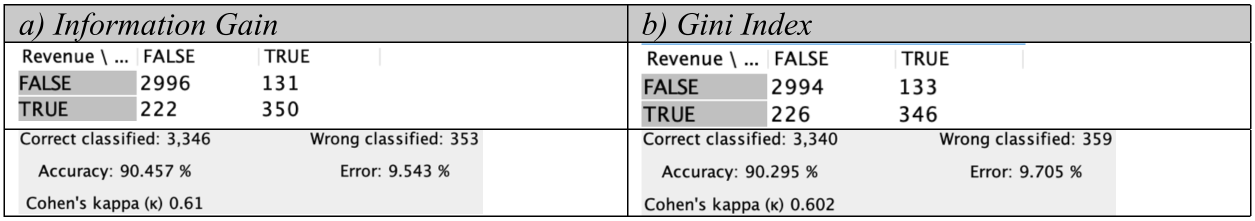
Revenue \ ...	FALSE	TRUE
FALSE	3083	44
TRUE	368	204

Correct classified: 3,287 Wrong classified: 412
Accuracy: 88.862 % Error: 11.138 %
Cohen's kappa (κ) 0.446

Figure 3.2 a, b: The confusion matrix (above) and the statistics (below) of SVM given *a)* RBF kernel, and *b)* quadratic kernel.

3.3 RANDOM FOREST

The third predictor used was a random forest classification. This method was selected as it reduces overfitting while producing an accurate prediction and works well for categorical and continuous data. The random forest was calculated using information gain and Gini index to split the trees. Both adjustments produced high prediction accuracy, although information gain produced a slightly higher accuracy (*figure 3.3*). The increased accuracy is likely due to information gain favouring smaller partitions with various distinct values.



a) Information Gain

Revenue \ ...	FALSE	TRUE
FALSE	2996	131
TRUE	222	350

Correct classified: 3,346 Wrong classified: 353
Accuracy: 90.457 % Error: 9.543 %
Cohen's kappa (κ) 0.61

b) Gini Index

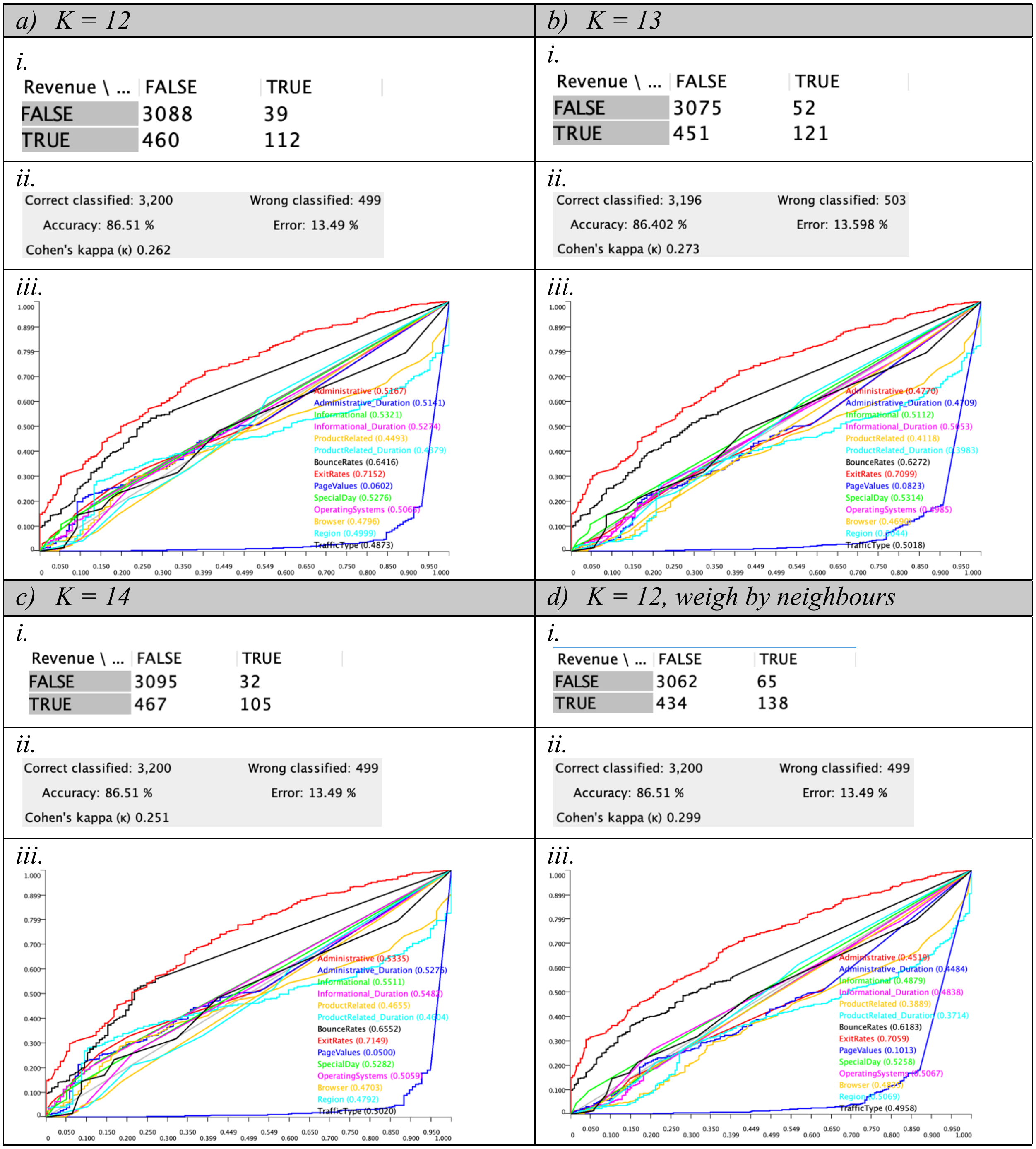
Revenue \ ...	FALSE	TRUE
FALSE	2994	133
TRUE	226	346

Correct classified: 3,340 Wrong classified: 359
Accuracy: 90.295 % Error: 9.705 %
Cohen's kappa (κ) 0.602

Figure 3.3 a, b: The confusion matrix (above) and the statistics (below) of Random Forest given *a)* Information Gain to split nodes, and *b)* Gini Index to split nodes.

3.4 k-NEAREST NEIGHBOURS

The fourth predictor used was a k-Nearest Neighbours. This classification was used as it is a simple and effective predictor which doesn’t make assumptions about any underlying data distributions. In selecting k-values, values 12, 13, and 14 were chosen using trial and error. Using k-values of 12 and 14 provided the highest accuracy, with the k-value of 12 being able to accurately identify some more true positive results than the k-value of 14 (figure 3.4). Hence, a k-value of 12 was used again for another k-NN, now weighing by neighbours. This produced the same accuracy but identified more true objects. To verify the accuracy of the k-NN, an ROC Curve was used, calculating the area of the true positive rate to the false positive rate when Revenue is False. Performing the ROC Curve confirmed the accuracy of the tests and showed that the Exit Rates attribute is consistently highly tested while Page Values is consistently very poorly tested (*figure 3.4.iii*).



a) K = 12

i.

Revenue \ ...	FALSE	TRUE
FALSE	3088	39
TRUE	460	112

ii.

Correct classified: 3,200 Wrong classified: 499
Accuracy: 86.51 % Error: 13.49 %
Cohen's kappa (κ) 0.262

iii.

b) K = 13

i.

Revenue \ ...	FALSE	TRUE
FALSE	3075	52
TRUE	451	121

ii.

Correct classified: 3,196 Wrong classified: 503
Accuracy: 86.402 % Error: 13.598 %
Cohen's kappa (κ) 0.273

iii.

c) K = 14

i.

Revenue \ ...	FALSE	TRUE
FALSE	3095	32
TRUE	467	105

ii.

Correct classified: 3,200 Wrong classified: 499
Accuracy: 86.51 % Error: 13.49 %
Cohen's kappa (κ) 0.251

iii.

d) K = 12, weigh by neighbours

i.

Revenue \ ...	FALSE	TRUE
FALSE	3062	65
TRUE	434	138

ii.

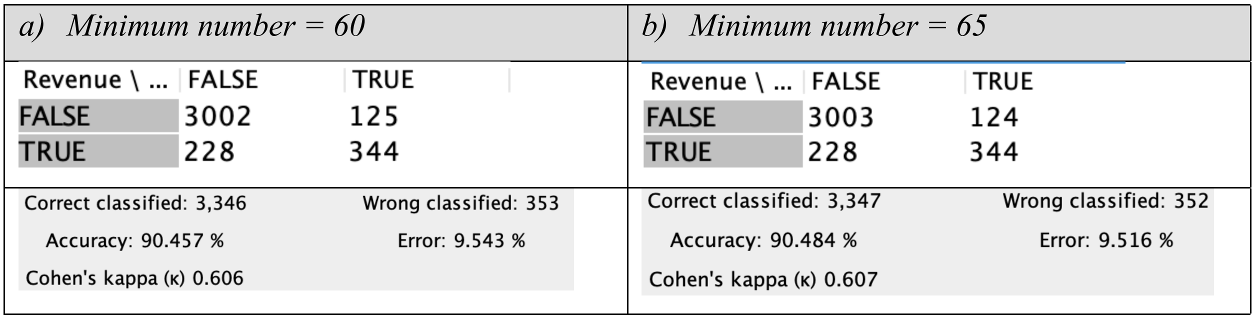
Correct classified: 3,200 Wrong classified: 499
Accuracy: 86.51 % Error: 13.49 %
Cohen's kappa (κ) 0.299

iii.

Figure 3.4 a, b, c, d: The confusion matrix (*i*), the statistics (*ii*), and the ROC Curve (*iii*) of k-Nearest Neighbours given *a)* k-value of 12, *b)* k-value of 13, *c)* k-value of 14, and *d)* k-value of 12 and weighing by neighbours.

3.5 DECISION TREE

The fifth predictor used was a decision tree. Decision trees are quick and effective, hence why it was selected. The decision tree used was a classification tree as the outcome was discrete, producing revenue or not. Two decision trees were performed, adjusting the minimum number of objects in a node with a minimum of 60 and a minimum of 65. While both decision trees provided a high accuracy, correctly identifying many true attributes, taking a minimum of 65 objects provided a slightly higher accuracy (*figure 3.5*).



a) Minimum number = 60

Revenue \ ...	FALSE	TRUE
FALSE	3002	125
TRUE	228	344

Correct classified: 3,346 Wrong classified: 353

Accuracy: 90.457 % Error: 9.543 %

Cohen's kappa (κ) 0.606

b) Minimum number = 65

Revenue \ ...	FALSE	TRUE
FALSE	3003	124
TRUE	228	344

Correct classified: 3,347 Wrong classified: 352

Accuracy: 90.484 % Error: 9.516 %

Cohen's kappa (κ) 0.607

Figure 3.5: The confusion matrix (above) and statistics (below) of decision trees given *a)* minimum number of 60 and *b)* minimum number of 65 objects per node.

4.0 ATTRIBUTE SELECTION

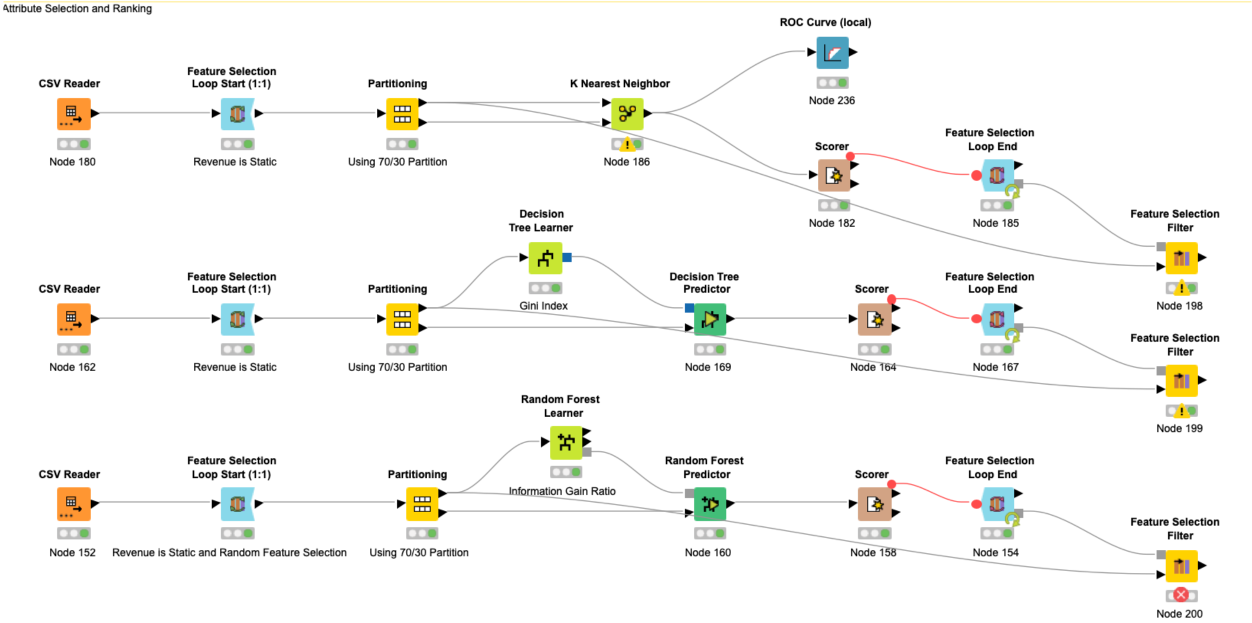


Figure 4.0: The workflow used to perform attribute selection on the online_shoppers_intention dataset. The k-NN was performed with a k-value of 12, the Decision Tree was performed with a minimum number of 65 objects peer node, and the Random Forest was performed using Information Gain Ratio.

To determine the attributes that should be tested on, feature selection was performed using a forward selection and backward elimination. This feature selection was performed on the three best performing predictors, considering the accuracy and run time of the predictors from section 3.0. The different selected attributes from the different predictors were compared to find the best attributes through the attributes that most commonly occur.

4.1 FEATURE SELECTION AND RANDOM FOREST

The first feature selection was performed using Random Forest as this predictor had the second highest accuracy in section 3.0. This feature selection was performed taking the Information Gain Ratio when splitting nodes. Performing the feature selection on the dataset, it identified Administrative, Exit Rates, Special Day, Traffic Type, and Visitor Type as the most important attributes (*figure 4.0a*).

4.2 FEATURE SELECTION AND DECISION TREE

The second feature selection was performed using decision tree. This predictor was selected as it had the highest accuracy in section 3.0. The feature selection was performed with a minimum number of 65 objects per node. This predictor found Administrative, Month, and Region to be the most important attributes (*figure 4.0b*).

4.3 FEATURE SELECTION AND k-NN

The third feature selection was performed using k-NN. This predictor was selected for its high accuracy in section 3.0 and quick run time. The feature selection was performed taking a k-value of 12. This predictor found ExitRates and VisitorType to be the more important attributes (*figure 4.0c*).

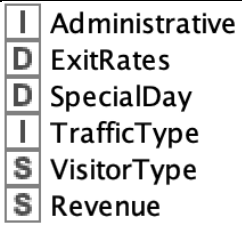
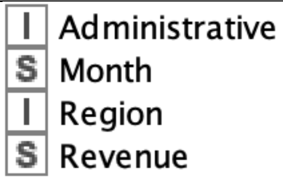
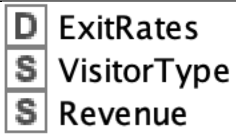
a) Random Forest	*b) Decision Tree*	*c) k-NN*
 I Administrative D ExitRates D SpecialDay I TrafficType S VisitorType S Revenue	 I Administrative S Month I Region S Revenue	 D ExitRates S VisitorType S Revenue

Figure 4.0 a, b, c: The different attributes found from a feature selection on *a)* Random Forest, *b)* Decision Tree, and *c)* k-NN

5.0 VERIFICATION OF ATTRIBUTES

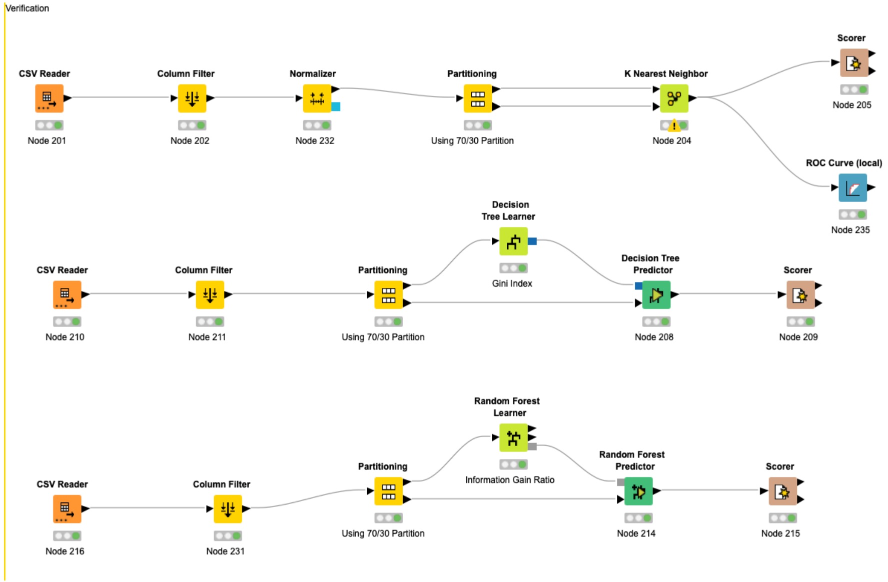
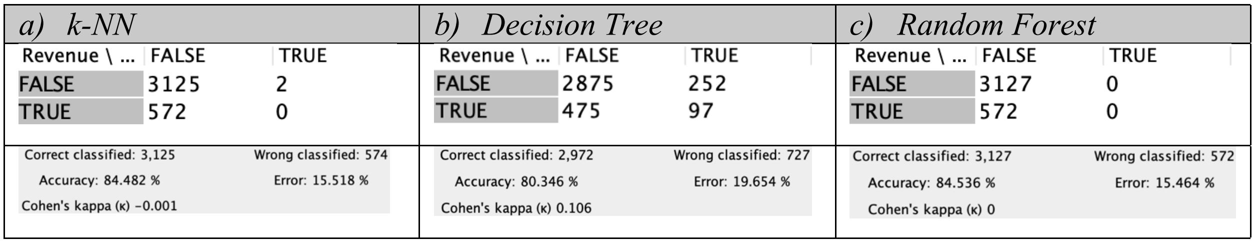


Figure 5.0: The workflow used to perform verification on the attributes in the dataset.

To verify that the selected attributes from Section 4.0 had a strong relationship to Revenue, the attributes were placed through the strongest predictors of Section 3.0 (*figure 5.0*). While all the workflows had decreased in accuracy, they maintained high overall accuracies larger than 80 per cent (*figure 5.1*). As well, the ROC Curve maintained that the Administrative and Exit Rates attributes were important as the area from their curve was large (*figure 5.2*). To assess the distribution and correlation of the different attributes given revenue, the relationship of revenue and the different attributes were visualized using R, showing the relationship of the different attributes to the Revenue (*figure 5.3*). To determine the properties of the attributes that are largely related to consumers, Cross Tables were performed on the different attributes. The Cross Table on the Administrative attribute found that most consumers use type 0 (*figure 5.4a*). The Cross Table of the Exit Rates attribute was used by forming categories from the Exit Rates based on the interquartile ranges, and found that half of the consumers have exit rates of less than 0.02516. Lastly, the Cross Table of the Visitor Type attribute found that the vast majority (0.856) of consumers are returning visitors.



a) k-NN

Revenue \ ...	FALSE	TRUE
FALSE	3125	2
TRUE	572	0

Correct classified: 3,125 Wrong classified: 574

Accuracy: 84.482 % Error: 15.518 %

Cohen's kappa (κ) −0.001

b) Decision Tree

Revenue \ ...	FALSE	TRUE
FALSE	2875	252
TRUE	475	97

Correct classified: 2,972 Wrong classified: 727

Accuracy: 80.346 % Error: 19.654 %

Cohen's kappa (κ) 0.106

c) Random Forest

Revenue \ ...	FALSE	TRUE
FALSE	3127	0
TRUE	572	0

Correct classified: 3,127 Wrong classified: 572

Accuracy: 84.536 % Error: 15.464 %

Cohen's kappa (κ) 0

Figure 5.1: The confusion matrix (above) and statistics (below) found from performing *a)* Random Forest, *b)* Decision Tree, and *c)* k-NN on the select attributes.

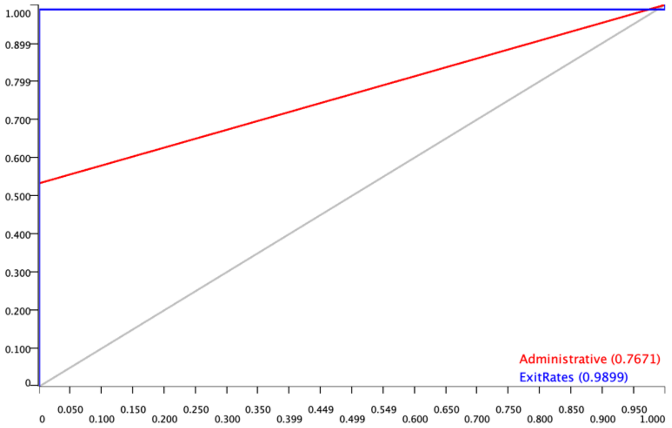


Figure 5.2: The ROC Curve performed on the k-Nearest Neighbours using the select attributes.

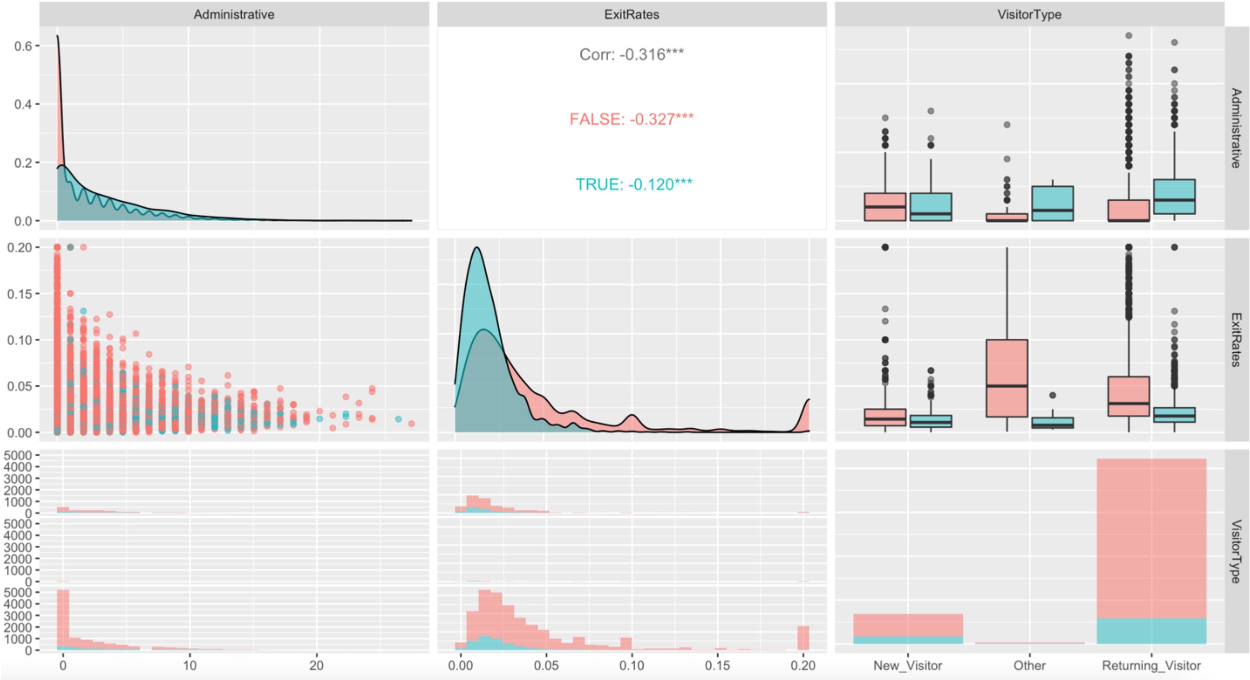


Figure 5.3: A graph matrix of correlations and counts given Revenue, where red is False and red is True. Graphed using R.

6.0 ACTIONABLE CONCLUSIONS

To make predictions on users most likely to buy from the site, predictors were used on the online_shoppers_intention.csv dataset to find the most accurate predictors to use. The three most accurate predictors were then used for attribute selection using a process of forward selection and backward elimination to find the attributes, or columns, most associated with a high prediction. The three most accurate predictors were found to be k-Nearest Neighbours, Decision Tree Classification, and Random Forest Classification. These three predictors then found the most accurate attributes to be Administrative, Exit Rates, and Visitor Type. Through performing Cross Tables on the different attributes, it was found that consumers are most likely to use Administrative type 0, have Exit Rates of less than 0.02516, and be a returning visitor.