Online Shopping Purchase Intent Analysis

By Manish Varkala (20151005), Sumaj Reddy Pannala (20147141), Goutham Kasula (20150992)

Executive Summary:

The project aimed to analyze online shopping data and predict purchase intent based on various features. A dataset comprising 12,330 entries with 18 features, including administrative, informational, and product-related metrics, was used for analysis. The goal was to build a predictive model to understand the factors influencing online purchase decisions and provide insights for marketing strategies.

Objectives:

- Understand Feature Correlations: Examine relationships among features to identify potential drivers of revenue.
- Visitor Behavior Analysis: Explore visitor types, their distribution, and the impact on revenue.
- Predictive Modeling: Utilize machine learning models to predict revenue and evaluate their performance.

Problem Statement:

The client sought to optimize website performance and revenue generation by understanding the relationships between various features and predicting revenue based on visitor behavior.

Methodologies:

The analysis employed descriptive statistics, correlation matrices, pie charts, bar charts, box plots, and pair plots to visually represent relationships among features. Additionally, machine learning models, including Naive Bayes, K-Nearest Neighbors, Logistic Regression, Random Forest, and MLP Classifier, were employed to predict revenue and evaluate model performance.

1. Data Collection:

- The dataset was obtained from a reliable source containing information about online shoppers' interactions and purchase behavior.
- Initial exploration of the dataset revealed 18 features, including administrative, informational, and product-related variables, as well as visitor details and traffic-related information.

2. Data Preprocessing:

- Data cleaning involved handling missing values, which were found to be absent in the dataset.
- Categorical variables, such as 'Month' and 'VisitorType,' were one-hot encoded to facilitate model training.
- Binary categorical variables, such as 'Weekend' and 'Revenue,' were converted to integers (0 or 1).
- The dataset was split into training and testing sets using the train_test_split function.

3. Exploratory Data Analysis (EDA):

Correlation analysis:

The correlation matrix (***Fig 1***) shows that 'PageValues' (0.492569) has the highest positive correlation with the likelihood of generating revenue for the target variable 'Revenue'. Though less prominent than 'PageValues,' 'ProductRelated,' 'ProductRelated Duration,' and 'Administrative' also have positive correlations. Negatively, 'BounceRates' and 'ExitRates' have negative correlations with 'Revenue,' meaning that higher rates decrease revenue. Weekend, Browser, TrafficType, Region, OperatingSystems, and SpecialDay have weak correlations with 'Revenue.'

Darker colours indicate stronger correlations in the heatmap. The strong positive correlations between 'Administrative' and 'Administrative Duration,' 'Informational' and 'Informational Duration,' 'ProductRelated' and 'ProductRelated Duration,' and 'BounceRates' and 'ExitRates' are notable Additionally, the heatmap shows a moderately positive correlation (0.49) between 'PageValues' and 'Revenue,' suggesting that pages with higher values generate more revenue. Additionally, 'SpecialDay' and 'Revenue' have a weak positive correlation (0.14).

To increase revenue, client can optimize pages with higher 'PageValues' and lower bounce and exit rates. Understanding feature relationships can help make strategic decisions like tailoring content on special days to boost revenue.

Features correlation with revenue - The correlation bar chart (***Fig 2***) provides insights into the relationships between various features and revenue. Noteworthy associations include 'PageValues' (correlation = 0.49) and 'ProductRelated' (correlation = 0.16), both demonstrating positive correlations with revenue. This suggests that pages featuring extensive product-related content and higher intrinsic values tend to contribute significantly to revenue generation. Additionally, 'Administrative' (correlation = 0.14) shows a positive correlation with revenue, indicating the potential for revenue generation from administrative content. While not as strong as other correlations, these findings highlight key features that play a role in influencing revenue outcomes in the online shopping environment.

Visitor Analysis - The revenue distribution chart (*Fig 3*) shows how the shopping revenue is split between two groups: purchased and Just Visited. The chart shows that out of total visitors, only 15.5 percent of purchases were made and other 84.5% just browsed the website and did not convert into revenue.

The Visitor count and Revenue distribution type chart (***Fig 4***), shows the total number of visits and revenue conversion divided by type of visitor. The graph shows that over 10,000 visits made by the returning visitors but around 1700 purchases were made by the returning visitors. There are around 1600 visits are from new visitors and less than 700 purchases were made by the new visitors, and there are minimal visits from other visitors.

Revenue by month & special occasion days (*Fig 5*) - The first chart (Count of Revenue by Month) shows the number of purchases made in each month. The chart shows that the most purchases are made in November, followed by May, and December. The fewest purchases are made in January and February.

The second chart (Count of Revenue by Special Day) shows the number of purchases made on special days, such as holidays. The chart shows that the most purchases are made on Black Friday, followed by Cyber Monday, Christmas Eve, and Christmas Day. The fewest purchases are made on New Year's Eve and New Year's Day.

Pie chart 1: Weekend vs. Weekday (Total Visits) - The first pie chart (***Fig 6***) shows that 77% of visits occur on weekdays and 23% of visits occur on weekends. This suggests that people are more likely to visit the website during the week than on the weekends.

Pie chart 2: Weekend vs. Weekday (total sales) - The second pie chart (***Fig 6***) shows that 74% of visits with transactions occur on weekdays and 26% of visits with transactions occur on weekends. This suggests that people are more likely to purchase something from the website during the week than on the weekends.

Additionally, there is Very Regionally Diverse traffic from dataset and Traffic Source is also very diverse (***Fig 7.1 – 7.4***) with a few sources which didn't contribute many people are more often visiting from one region than any other region and people tend to use 1 dominant browser to access website and one operating system is responsible for more than 5000 examples followed by other Operating Systems.

Site Page Analysis:

- The Boxplot shows (***Fig 8***), Visitors who make a purchase exhibit higher page views and longer time spent per page, indicating a deeper level of interest and a stronger likelihood of conversion.
- Non-purchasing visitors generally have lower page views and shorter time spent on each page, suggesting a more cursory exploration of the website.
- Visitors dedicate significantly more time to product pages compared to account-related or informational pages, emphasizing the importance of crafting compelling product content to drive engagement and conversions.

Page metrics:

- The pair plot shows (***Fig 9***), Bounce rates and exit rates are positively correlated. This means that visitors who bounce more from a website are more likely to also exit the website.
- Page values and exit rates are negatively correlated. This means that visitors who spend more time on a website are less likely to exit the website.
- Visitors who generate higher revenue are less likely to bounce from the website and more likely to spend more time on the website.

4. Feature Engineering:

- Feature engineering involved creating dummy variables for categorical features like 'Month' and 'VisitorType.'
- The 'Revenue' column was converted to binary integers for modeling purposes.

5. Model Building:

- Five machine learning models were selected: Naïve Bayes(Bernoulli), Logistic Regression, K-Nearest Neighbors, Random Forest, and Multi-Layer Perceptron (MLP).
- Each model was trained on the preprocessed data to predict the likelihood of online shopping purchases.

Naives Bayes model:

- The model demonstrates good accuracy, but the F1 score suggests a trade-off between precision and recall, indicating potential room for improvement.
- Precision and recall values provide insights into the model's ability to make correct positive predictions and capture actual positive instances, respectively.
- The confusion matrix provides a granular view of the model's performance, showing where it excels and where it makes errors.
- In summary, the Bernoulli Naive Bayes model has reasonable performance, but further optimization may be possible to enhance its precision and recall. The confusion matrix provides a nuanced understanding of the model's strengths and weaknesses.

Bernoulli Naive Bayes Performance:

```
Bernoulli Naive Bayes Performance:
---------------------------------
Accuracy        :  0.862124898621249
F1 Score        :  0.567062818336163
Precision       :  0.5538971807628524
Recall          :  0.5808695652173913
Confusion Matrix:
  [[2855  269]
 [ 241  334]]
```

```
Bernoulli Naive Bayes Classificaton Report:
-------------------------------------
Accuracy        :  0.8537442552041092
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      3124
           1       0.56      0.27      0.36       575

    accuracy                           0.85      3699
   macro avg       0.72      0.61      0.64      3699
weighted avg       0.83      0.85      0.83      3699

Confusion Matrix:
 [[3004  120]
 [ 421  154]]
```

K- Nearest Neighbor:

- While the accuracy is relatively high, the low F1 score, and recall indicate that the model struggles to effectively identify positive instances.
- Precision is relatively higher, suggesting that when the model predicts positive instances, it tends to be correct more often than not.
- The confusion matrix provides insights into specific areas of model performance, such as the number of false positives and false negatives.
- In summary, the K-Nearest Neighbors model has decent accuracy but needs improvement in capturing true positive instances. Further tuning or exploration of different model parameters may be necessary to enhance its performance.

K-Nearest Neighbour Performance:

```
K-Nearest Neighbor Performance:
---------------------------------------
Accuracy        :  0.8537442552041092
F1 Score        :  0.3627797408716136
Precision       :  0.5620437956204379
Recall          :  0.2678260869565217
Confusion Matrix:
 [[3004  120]
 [ 421  154]]
```

```
KNN Classificaton Report on Test Set:
---------------------------------------
Accuracy        :  0.8537442552041092
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      3124
           1       0.56      0.27      0.36       575

    accuracy                           0.85      3699
   macro avg       0.72      0.61      0.64      3699
weighted avg       0.83      0.85      0.83      3699

Confusion Matrix:
 [[3004  120]
 [ 421  154]]
```

Logistic Regression:

- The model has a relatively high accuracy, but the low recall suggests that it misses a significant number of positive instances.
- Precision is decent, indicating that when the model predicts positive instances, it is correct more often than not.
- The ROC AUC value of 0.67 suggests moderated power, but further tuning may required to improve recall.
  In summary, The Logistic Regression model demonstrates good accuracy but has a trade-off between precision and recall. It achieves higher precision at the cost of lower recall, making it suitable for scenarios where minimizing false positives is a priority.

```
Logistic Regression Performance:
---------------------------------
Accuracy        :  0.8832116788321168
F1 Score        :  0.4881516587677725
Precision       :  0.7657992565055762
Recall          :  0.3582608695652174
Confusion Matrix:
  [[3061   63]
 [ 369  206]]
```

```
Logistic Regression Classificaton Report:
--------------------------------------
Accuracy        :  0.8832116788321168
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.98      0.93      3124
           1       0.77      0.36      0.49       575

    accuracy                           0.88      3699
   macro avg       0.83      0.67      0.71      3699
weighted avg       0.87      0.88      0.86      3699

Confusion Matrix:
 [[3061   63]
 [ 369  206]]
```

Random Forest:

- The Random Forest model for classification demonstrates an impressive accuracy of 90.08%, showcasing its effectiveness in correctly classifying instances into their respective classes. The F1 Score, a harmonic mean of precision and recall, is 0.6391, indicating a balanced performance between precision and recall. The precision of 73.53% suggests the model's ability to avoid false positives, while a recall of 56.52% highlights its capability to capture actual positive cases. The confusion matrix breaks down the results, revealing 325 true positives, 3007 true negatives, 117 false positives, and 250 false negatives.

- The model's AUC (Area Under the Curve) score is 0.76, denoting the overall performance of the classifier across various classification thresholds. This score reaffirms the model's strong discriminatory power and its ability to distinguish between the positive and negative classes.

- Comparing these metrics collectively, the Random Forest model emerges as a robust performer, excelling in both accuracy and the ability to balance precision and recall.

```
Random Forest Performance:
---------------------------------
Accuracy        :  0.9007839956745066
F1 Score        :  0.6391347099311702
Precision       :  0.7352941176470589
Recall          :  0.5652173913043478
Confusion Matrix:
  [[3007  117]
 [ 250  325]]
```

```
Random Forest Classificaton Repor:
-------------------------------------
Accuracy        :  0.9002433090024331
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      3124
           1       0.74      0.55      0.63       575

    accuracy                           0.90      3699
   macro avg       0.83      0.76      0.79      3699
weighted avg       0.89      0.90      0.89      3699

Confusion Matrix:
 [[3012  112]
 [ 257  318]]
```

MLP Classifier:

- The model demonstrates good accuracy, but there is a notable difference in precision and recall between class 0 and class 1.
- Class 0 has higher precision, recall, and F1-score, indicating better performance in predicting instances of class 0.
- Class 1 has lower precision, recall, and F1-score, suggesting that the model struggles to correctly identify instances of class 1.
- The weighted average provides an overall summary, considering the class imbalance.
- In summary, the MLPClassifier performs reasonably well, but there is room for improvement, especially in correctly identifying instances of class 1. Further model tuning or adjustment of class weights may be considered to address these issues.

```
MLP Classifier Performance:
--------------------------------------
Accuracy        :  0.8788861854555285
F1 Score        :  0.5889908256880734
Precision       :  0.6233009708737864
Recall          :  0.5582608695652174
Confusion Matrix:
 [[2930  194]
 [ 254  321]]
```

```
MLP Classification Report:
--------------------------------------
Accuracy        :  0.8788861854555285
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      3124
           1       0.62      0.56      0.59       575

    accuracy                           0.88      3699
   macro avg       0.77      0.75      0.76      3699
weighted avg       0.87      0.88      0.88      3699

Confusion Matrix:
 [[2930  194]
 [ 254  321]]
```

Model Evaluation:

- Model performance was evaluated using various metrics, including accuracy, precision, recall, F1 score, FP rate, and ROC AUC.
- Confusion matrices provided a detailed breakdown of model predictions.
- Bernoulli Naive Bayes Achieves 86% accuracy with balanced precision and recall (F1: 0.57), effectively predicting positive instances.
- K-Nearest Neighbor (KNN) Shows 85% accuracy, but struggles with low recall (0.27), indicating limitations in identifying actual positive instances.
- Logistic Regression Demonstrates 88% accuracy with a balanced trade-off between precision and recall (F1: 0.49), showcasing overall strong predictive ability.
- Random Forest Impressive 90% accuracy, balanced precision and recall (F1: 0.64), indicating robust performance in classification tasks.
- MLP Classifier Achieves 88% accuracy, with high precision (0.76) but lower recall (0.35), suggesting good identification of positive instances but room for improvement in recall.

Area Under Curve(fig 10):

AUC for KNN	0.65
AUC for NBM	0.75
AUC for LR	0.67
AUC for RF	0.76
AUC for MLP	0.75

1. **KNN:** AUC of 0.65 suggests moderate discriminatory power but room for improvement compared to other models.
2. **Naive Bayes:** AUC of 0.75 indicates good overall performance, with a balanced trade-off between sensitivity and specificity.
3. **Logistic Regression:** AUC of 0.67 suggests moderate discriminatory ability, falling behind RF and MLP.
4. **Random Forest:** AUC of 0.76 signifies strong discriminatory power and good overall performance.
5. **MLP:** AUC of 0.75 showcases excellent discriminatory ability, competing closely with RF.

Based on the AUC scores and overall performance metrics, Random Forest emerges as the best-performing model for predicting online shopper intention in this project. It demonstrates strong discriminatory power, high accuracy, and a well-balanced trade-off between precision and recall. The AUC score of 0.76 further supports its effectiveness in distinguishing between different classes.

Key Findings:

1. **Feature Correlations**

- 'PageValues' exhibited the highest positive correlation (0.49) with 'Revenue,' suggesting that higher page values are associated with increased revenue.
- Positive correlations were observed for 'ProductRelated,' 'ProductRelated_Duration,' and 'Administrative,' emphasizing their potential impact on revenue.
- Negative correlations were noted for 'BounceRates' and 'ExitRates,' indicating a potential revenue decrease with higher bounce and exit rates.
- Random Forest Model demonstrated an accuracy of 90.02% and balanced performance across precision, recall, and F1 score.
- Multi-Layer Perceptron demonstrated an accuracy of 88.16%, competed very well with random forest model but short on F1 Score, if precision and avoiding false positives are crucial MLP is a good choice, But for a balanced performance across multiple metrics random forest model will me more effective based on analysis.

2. **Visitor Behavior**

- Visits by Returning visitor are higher than new visitors, it means more visits were made by returning visitors rather than new visitors.
- The majority of revenue was generated by customers who visited the website before rather than new new visitors, underlining the significance of converting visits into transactions.

Recommendations

- **Feature Optimization** - Focus on optimizing pages with higher 'PageValues' and enhancing content related to 'ProductRelated' and 'Administrative' to positively impact revenue.
- **Visitor Engagement** - Implement strategies to retain repeat visitors, such as targeted promotions or loyalty programs.
- **Minimize Bounce and Exit Rates** - Prioritize reducing bounce and exit rates through improved website design, content, and user experience.
- **Strategic Content Tailoring** - Leverage knowledge of special days to tailor content and promotions for maximum revenue generation.
- **Model Improvement** - Further tune machine learning models, especially focusing on recall improvement for positive revenue instances.

Conclusion

The analysis provides valuable insights into feature correlations, visitor behavior, and predictive modeling for revenue generation. By implementing the recommended strategies, the client can optimize website performance, enhance user engagement, and ultimately increase revenue. Continuous monitoring and refinement of these strategies will be crucial for sustained success.

Appendix

Fig 1

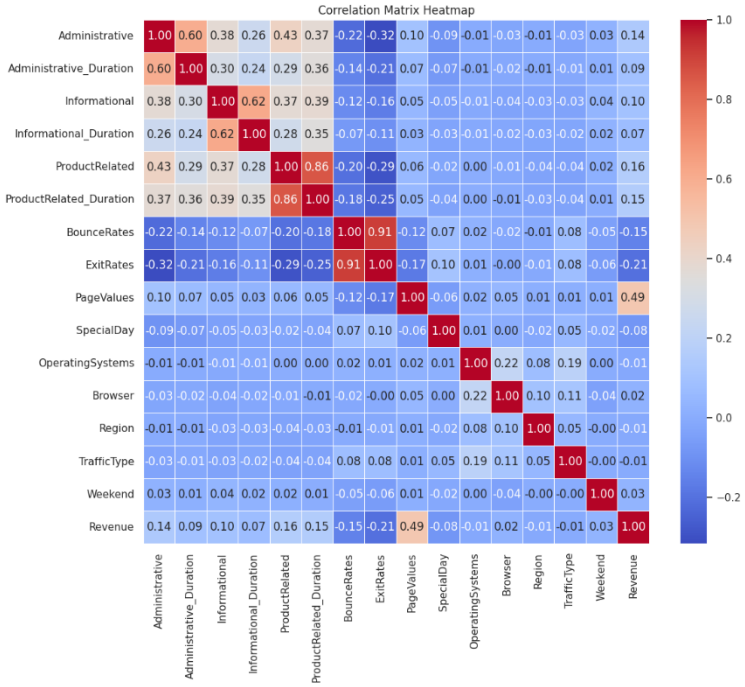


Fig 2

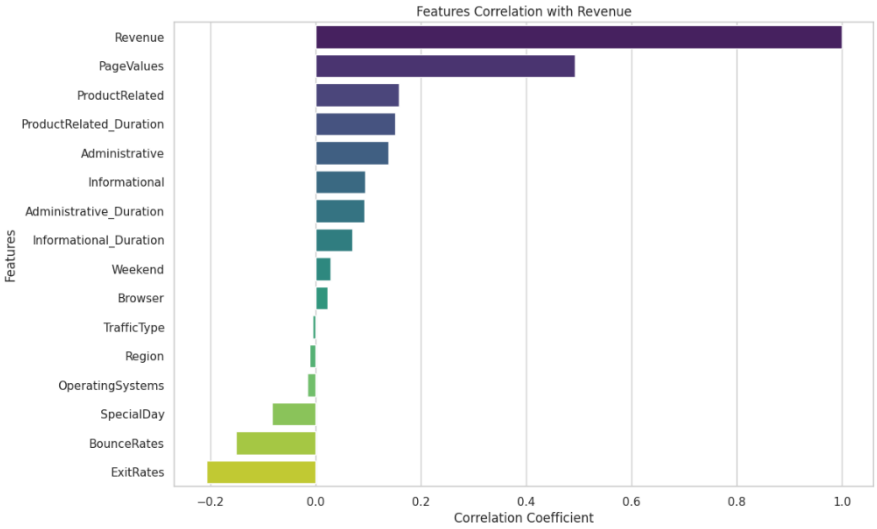


Fig 3

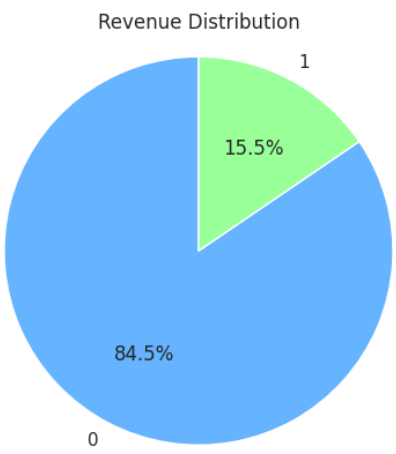


Fig 4

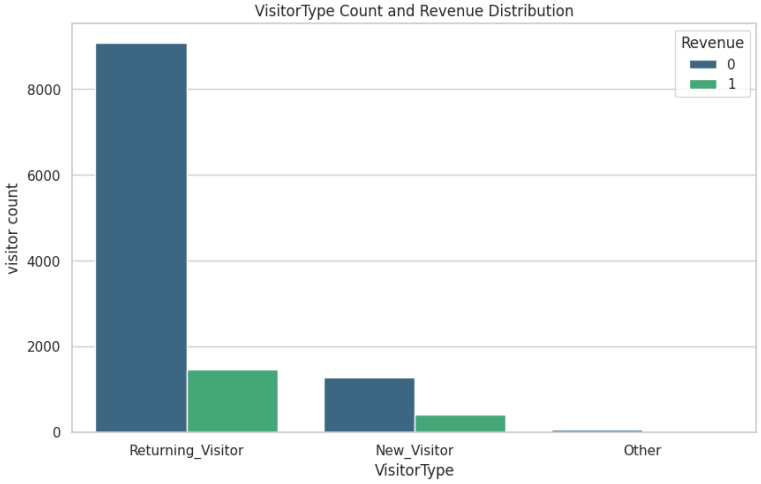

VisitorType Count and Revenue Distribution
Revenue
0
1
visitor count
8000
6000
4000
2000
0
Returning_Visitor
New_Visitor
Other
VisitorType

Fig 5

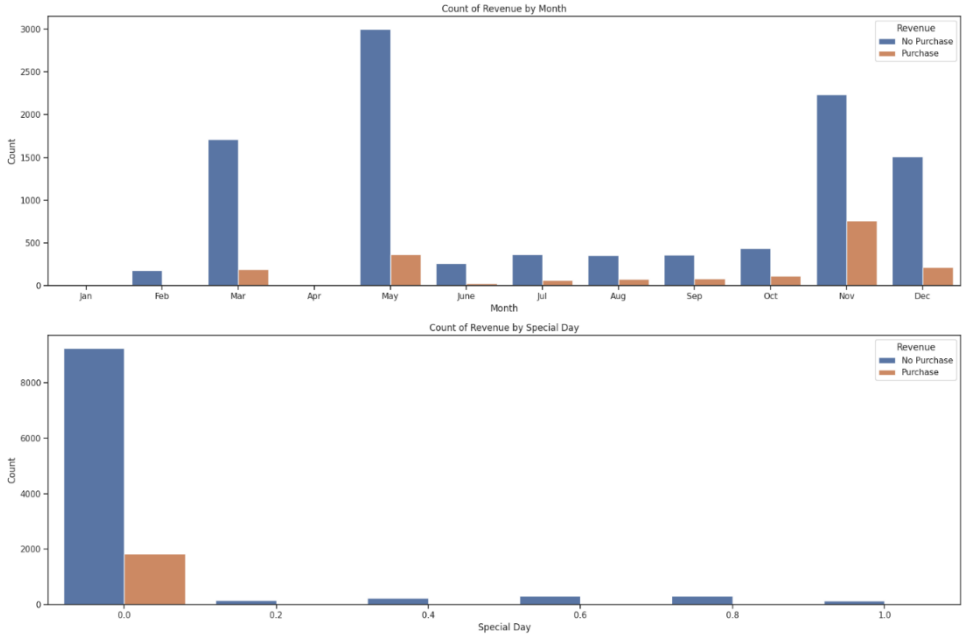

Count of Revenue by Month
Revenue
No Purchase
Purchase
Count
3000
2500
2000
1500
1000
500
0
Jan
Feb
Mar
Apr
May
June
Jul
Aug
Sep
Oct
Nov
Dec
Month
Count of Revenue by Special Day
Revenue
No Purchase
Purchase
Count
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
Special Day

Fig 6

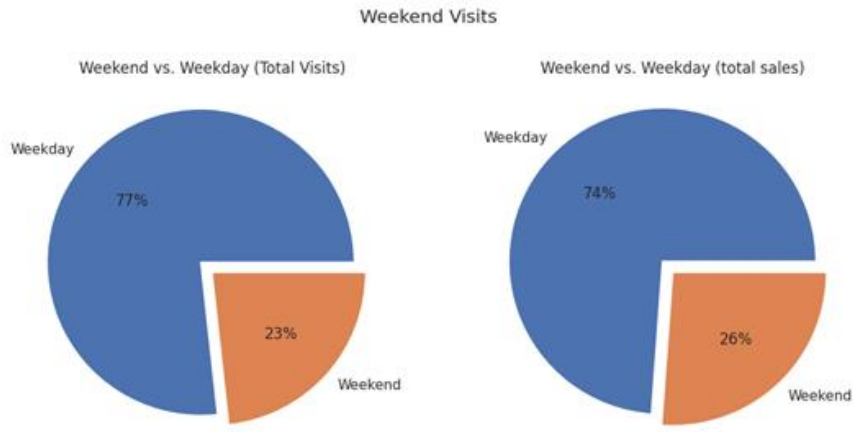


Fig 7.1

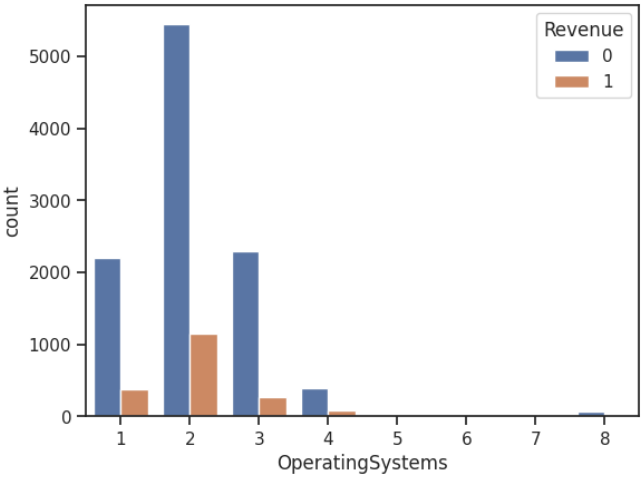
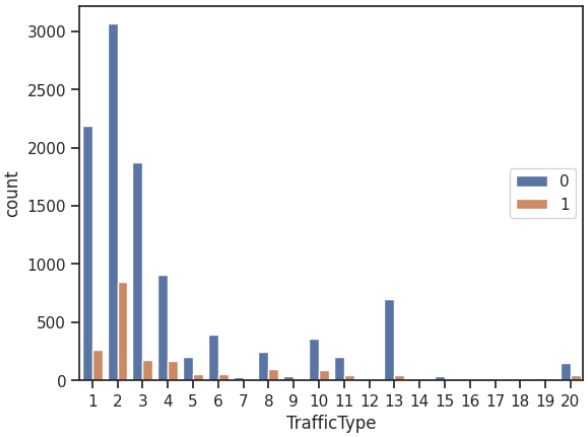


Fig 7.2



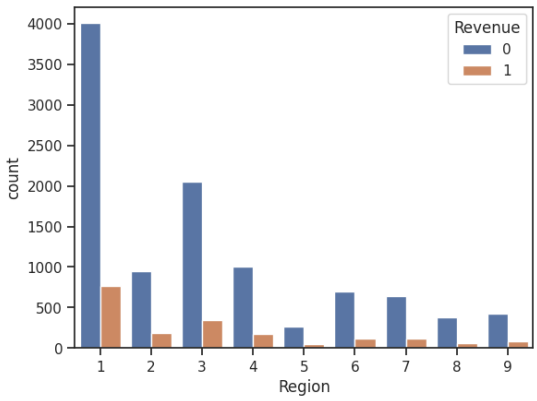


Fig 7.3

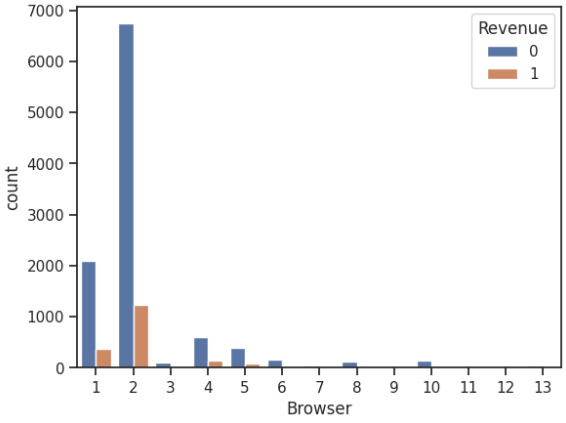


Fig 7.4

Fig 8

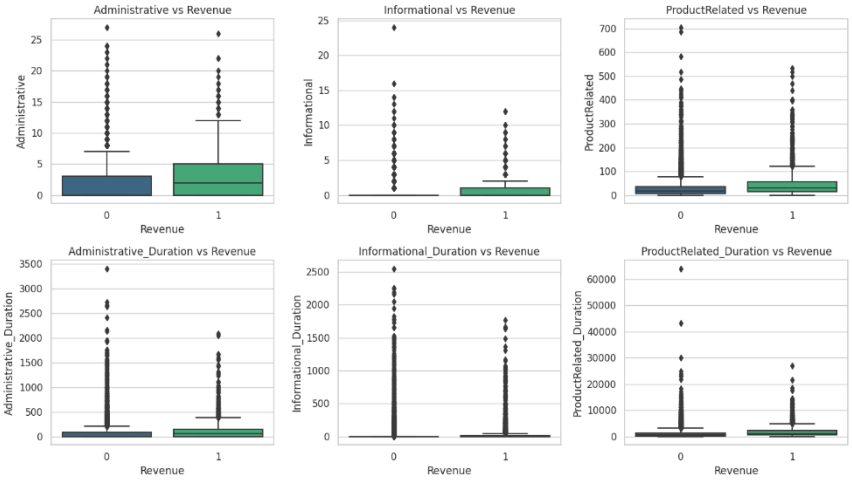

Administrative vs Revenue
Informational vs Revenue
ProductRelated vs Revenue
Administrative_Duration vs Revenue
Informational_Duration vs Revenue
ProductRelated_Duration vs Revenue
Revenue

Fig 9

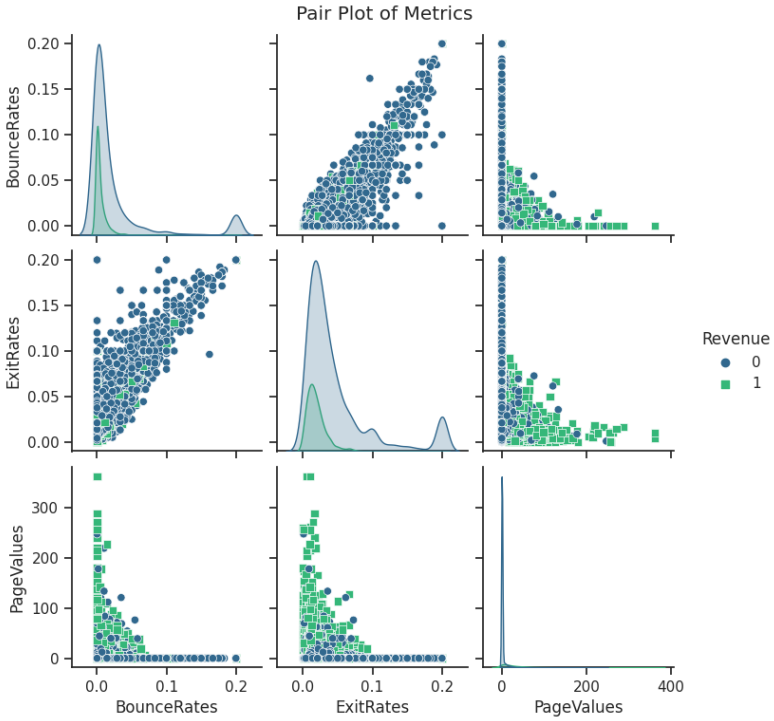
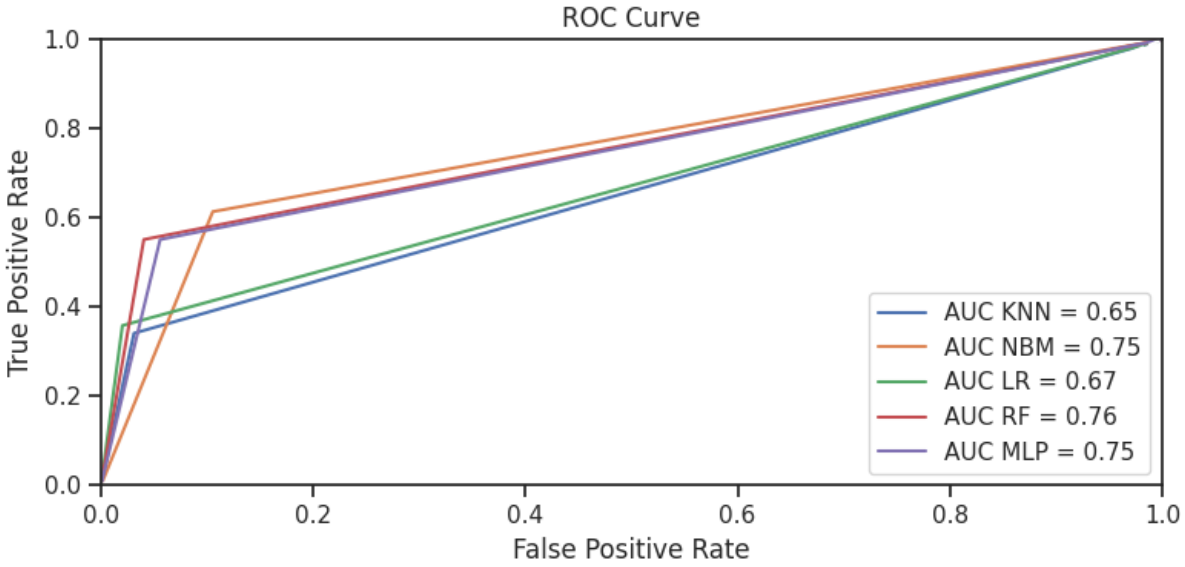

Pair Plot of Metrics
BounceRates
ExitRates
PageValues
Revenue
0
1

Fig 10


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC KNN = 0.65
AUC NBM = 0.75
AUC LR = 0.67
AUC RF = 0.76
AUC MLP = 0.75