MIS S381N
Project report

Anurag Arakala
Sanyam Jain
Jui-Jia Lin
Geer Zhang

9th Aug’23

Predicting Online Shoppers Purchasing Intention

1. Dataset Description

The dataset we are employing for our study is the Online Shoppers Purchasing Intention Dataset. This rich dataset was created to explore the purchasing intentions of online shoppers, and it includes different types of attributes drawn from online session records of consumers. The dataset contains ten numerical and eight categorical attributes, with a total of approximately 12,330 sessions. The variables relate to different aspects such as the type of visitor, whether the visit was on a weekday or weekend, the month of the visit, and so forth. The “Revenue” attribute is a binary variable that indicates whether or not the session concluded with a transaction.
We would like to explore which customer features decide the online purchase behavior in different models.

2. Importance of Analyzing the Dataset

Online shopping has revolutionized the way consumers purchase products, and with this shift, an enormous amount of data is generated which can provide vital insights. Analyzing this dataset holds great significance as it can help understand customers' online behavior and purchasing patterns. Understanding these patterns can lead to more effective marketing strategies, better customer service, and improved overall business performance. By predicting customer purchasing intentions, businesses can tailor their services to increase conversion rates and ultimately, revenues.

3. Basic Statistics and Intuitions About the Dataset (EDA)

The dataset includes sessions that resulted in both transactions (purchases made) and non-transactions. Approximately 85% of the sessions did not result in a purchase, whereas around 15% of sessions led to a transaction.

This indicates that the majority of online browsing sessions do not lead to a purchase.

In terms of visitor type, approximately 86% of sessions were by returning visitors, while the rest were new visitors. This indicates that most online shopping is done by users who have previously interacted with the site.

From the correlation heatmap (*Figure 1*), we can tell that there is only a high correlation between “ProductRelated_Duration” and “ProductRelated”, “ExistRates” and “BounceRates”. As the more pages the customer visits, the more time they will spend on the product; the higher percentage of visitors who enter the website through that page and exit without triggering any additional tasks, the higher the percentage of pageviews on the website that end at the specific page. Those correlations are normal and common.

Furthermore, a preliminary exploration of the data indicates that the month, the type of visitor, and the time spent on the website may be correlated with the likelihood of a purchase being made.

Exploratory Data Analysis reveals the distribution of online sessions by Month. March, May, November, and December months have more online sessions than the other months and this might be due to more festive seasons occurring during these months (*Figure 2*). November has the highest purchase conversion rate, possibly due to Black Friday, Cyber Monday, and Thanksgiving sales.

Contrary to what one might believe, weekdays have more sessions than weekends. However, the conversion rate on weekends is higher than that on weekdays (*Figure 3*). Returning users have more online sessions than new users (*Figure 4*), although there is a much higher conversion rate for the latter (*Figure 5*), which provides an opportunity to generate more revenue from new users.

4. Solution to the Problem

a) Features Used and Why

In our analysis, we chose to focus on the following features: “ProductRelated”, “ProductRelated_Duration”, “BounceRates”, “ExitRates”, “PageValues”, “VisitorType”, “Weekend”, and “Month”. These features provide a comprehensive overview of user behavior during their browsing session. For instance, “ProductRelated” and “ProductRelated_Duration” give insights into the interaction of the customer with the products.

“BounceRates” and “ExitRates” help to understand the quality of the user session. “PageValues” can show the potential of the page in terms of generating revenue. “VisitorType”, “Weekend”, and “Month” are demographic and temporal attributes that can impact shopping behavior.

b) Classifier(s) Used

We decided to use Random Forest, Gradient Boosting Classifier, KNN model, and Logistic Regression and Neural Network. These are powerful machine-learning models capable of handling complex datasets with high dimensionality.

c) Summary of the Results Obtained

Our models performed reasonably well in predicting whether a session would end in a purchase or not. The Random Forest Classifier had an accuracy of 90%, the KNN model had an accuracy of 88%, and the Gradient Boosting Classifier reached an accuracy of 91%. The Logistic Regression model had an accuracy of 89%. The Neural Network best model had an accuracy of 90%. The Gradient Boosting Classifier was our best-performing model (*Table 1*).

d) Analysis of the Model Learned by the Classifier

In our analysis, we employed three models: Random Forest, Gradient Boosting Classifier, K-Nearest Neighbors (KNN), Logistic Regression, and Neural Network.

The Random Forest Classifier highlighted the importance of features like “PageValues”, “ProductRelated_Duration”, and “ExitRates”. High “PageValues” and “ProductRelated_Duration” scores were indicative of a higher likelihood of a purchase. “ExitRates”, however, had an inverse relationship, indicating that sessions with higher exit rates were less likely to result in a purchase (*Figure 6*).

The Gradient Boosting Classifier, similar to the Random Forest model, indicated “PageValues” and “BounceRates” as significant features for predicting purchasing behavior. Additionally, it assigned importance to “ProductRelated_Duration” indicating that the longer a user interacts with the product-related pages, the more likely they are to make a purchase (*Figure 7*).

Our KNN model, being a distance-based algorithm, does not provide an explicit feature importance ranking like tree-based models. However, by observing the model's performance with the inclusion and exclusion of certain features, we could deduce that “V_type_Returning_Visitor”, “V_type_Other”, and “V_type_New_Visitor” significantly impacted its prediction accuracy. We can also find the optimal value of K by checking which value has the least MSE (*Figure 8*).

Overall, all three models pointed towards “PageValues” and “ProductRelated_Duration” as crucial indicators of a consumer's purchasing intention. High engagement with product-related pages and a high potential for generating revenue from certain pages were consistent signs of a session likely ending in a transaction.

The “ExitRates” were also a common important feature across the models but had a negative correlation with the purchasing intention, demonstrating that sessions with higher exit rates often don't result in a purchase. In terms of time variables, “Month” was observed to have more importance in predictions than the “Weekend” variable.

5. Conclusion

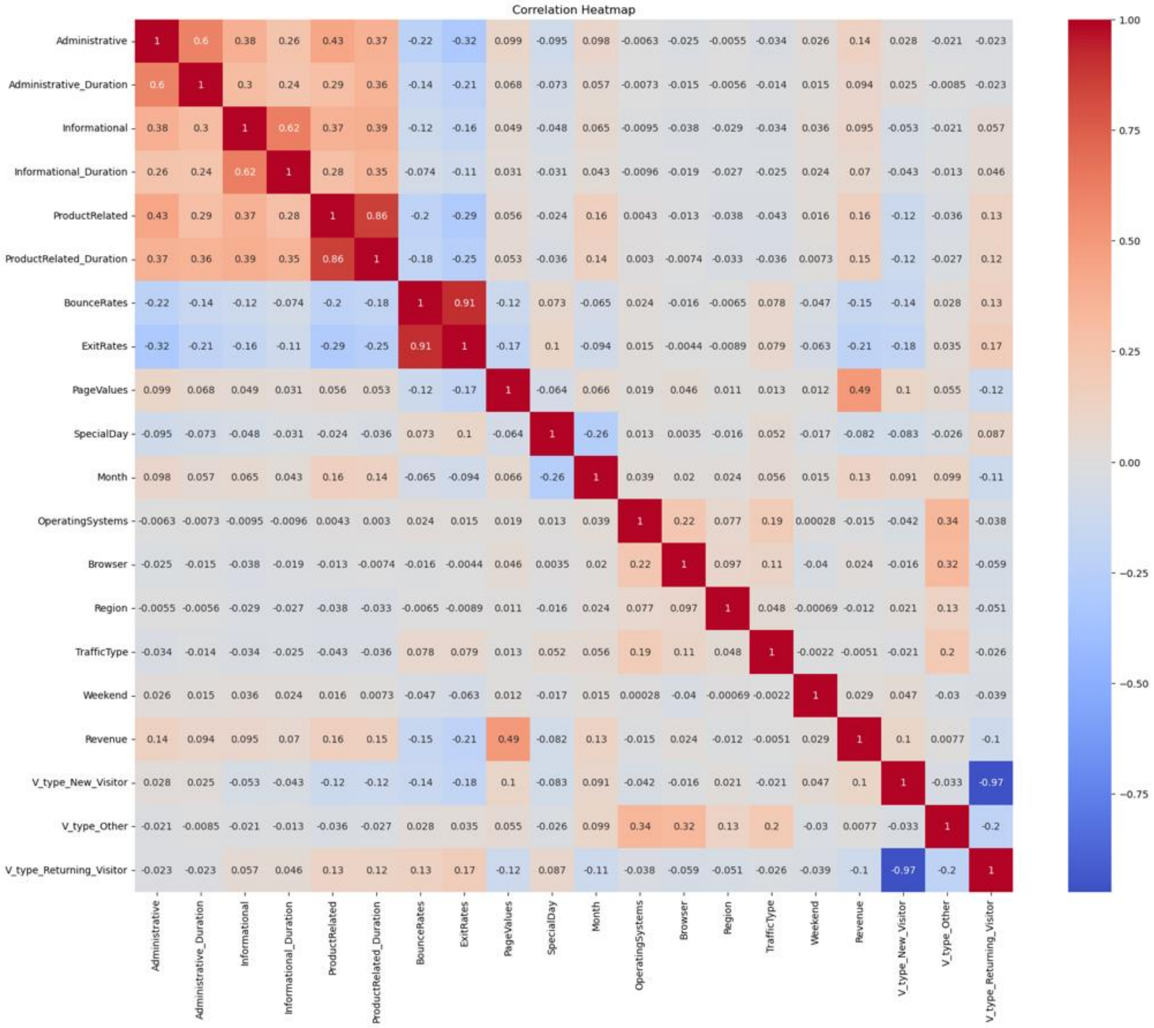
These findings reinforce the idea that understanding customer interactions with product pages and their session quality are key to predicting online purchase intentions.

However, it is also important to note that machine learning models are probabilistic and work on patterns in the dataset. As such, these interpretations should be used in conjunction with other business insights to form a comprehensive strategy.

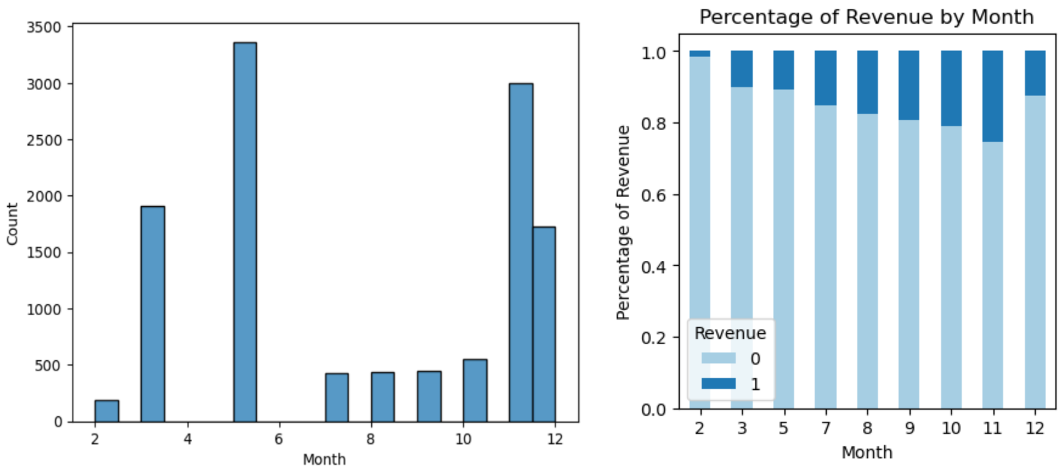
These models have proven valuable in understanding the online shopping behaviors and purchase intentions of consumers, allowing for more targeted and effective business strategies. However, they also highlight the need for more data and features that provide a more holistic view of the customer's online shopping journey, such as personal preferences, previous purchase history, and more specific demographic information.

Appendix

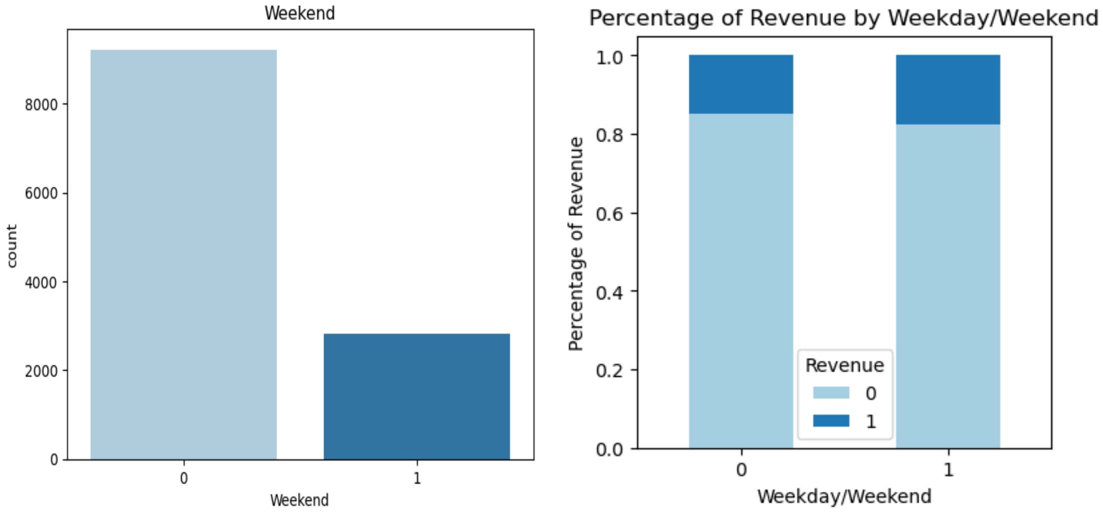
1. *Figure 1 (Correlation heatmap of the variables in the dataset)*



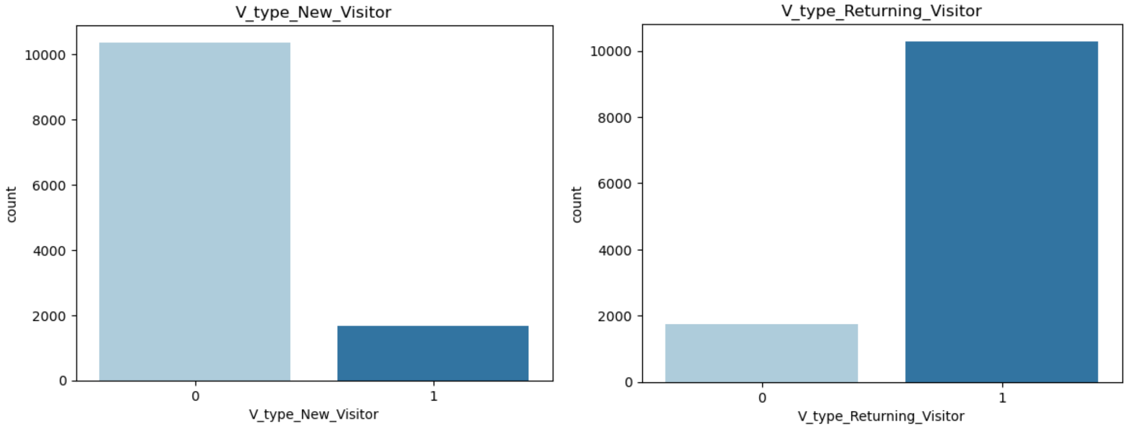
2. *Figure 2 (EDA: Sessions distribution and conversion by month)*



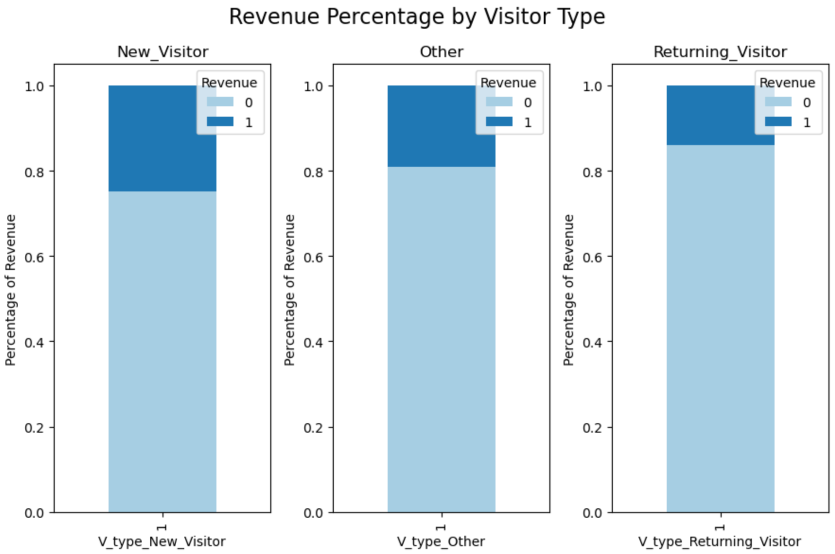
3. *Figure 3 (EDA: Sessions distribution and conversion by Weekend)*



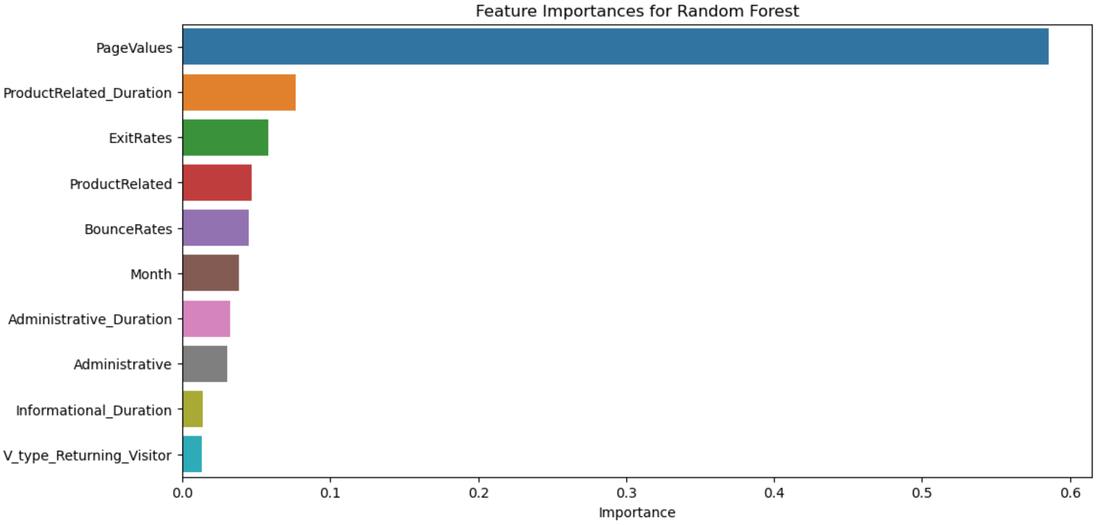
4. *Figure 4 (EDA: Sessions distribution by Visitor type (New Visitor and Returning Visitor))*



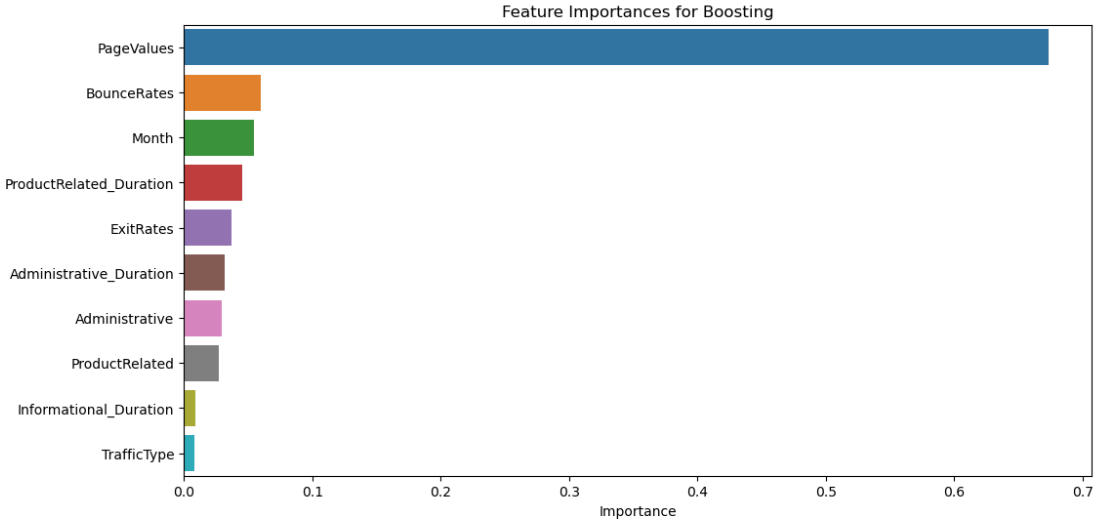
5. *Figure 5 (EDA: Sessions conversion by Visitor type (New Visitor, Returning Visitor, and Other))*



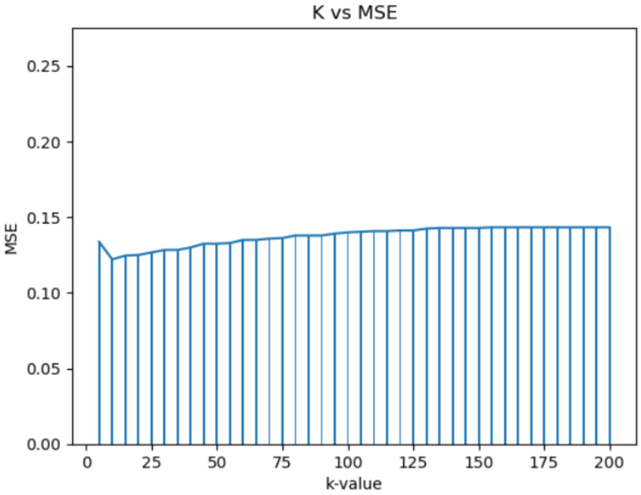
6. *Figure 6 (Feature Importance Graph for Random Forest Model)*

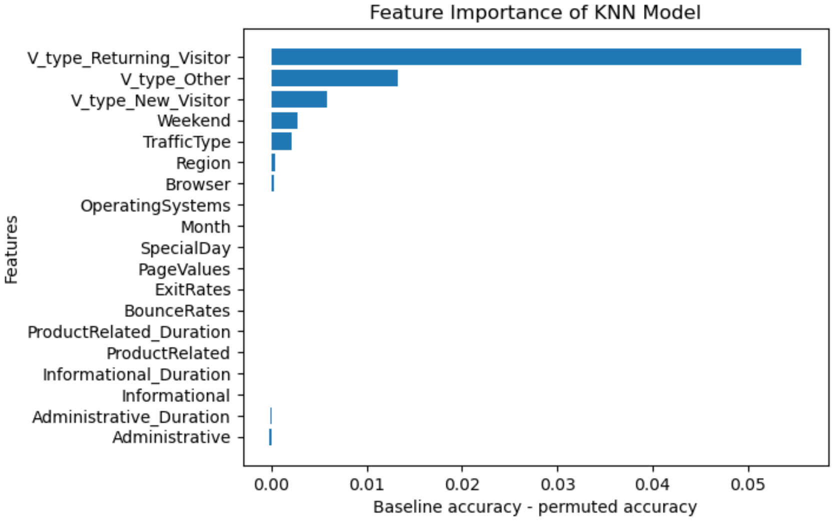


7. *Figure 7 (Feature Importance Graph for Boosting Model)*

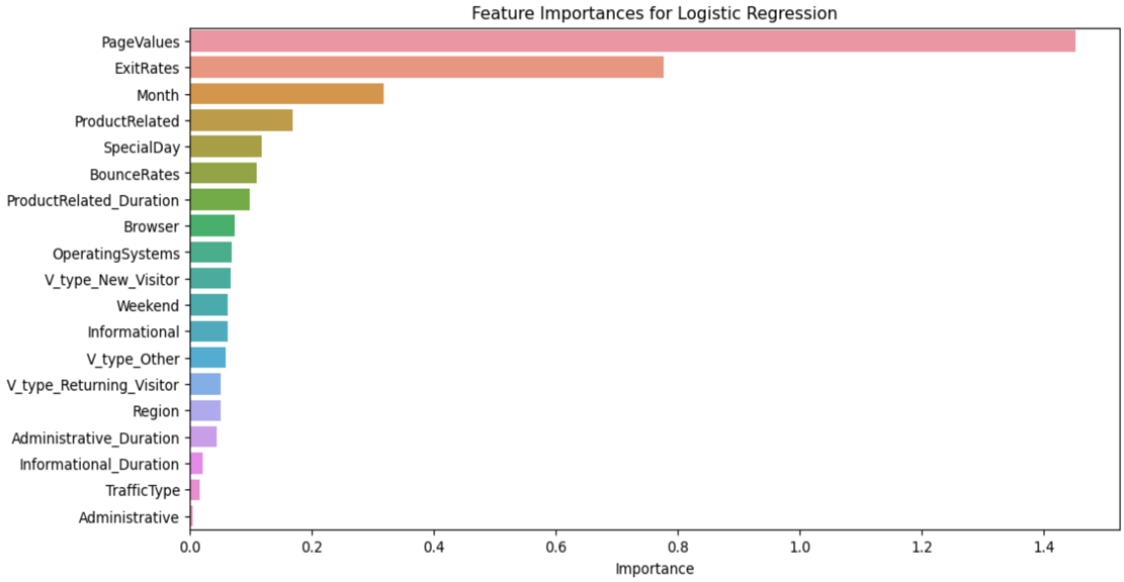


8. *Figure 8 (K vs MSE graph of the KNN Model and its Feature Importance Graph using feature permutation*





9. *Figure 9 (Feature Importance Graph for Logistic Regression Model)*



10. *Table 1* (Models Details)

Model	Accuracy	Recall	Best Parameter
Random Forest	0.90	0.54	{'max_depth':8, 'min_samples_split':2, 'n_estimators': 50}
Boosting	0.91	0.62	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}
KNN	0.87	0.22	K = 10
Logistic Regression	0.89	0.37	{'C': 1, 'penalty': 'l1'}
Neural Network	0.895	0.58	{'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (100,)}

Reference:

1. Dataset:https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset