



Predicting Online Shoppers Purchasing Intentions

Group 2

Anurag Arakala, Sanyam Jain, Jui-Jia Lin, Geer Zhang



Dataset

- The dataset was created to explore the purchasing intentions of online shoppers.
- It contains different types of attributes drawn from online session records of customers
- The dependent variable indicates whether or not the session concluded with a transaction
- In this project, we explored which customer feature decides the online purchase behavior through different models



Variables

- BounceRates and ExitRates: A high bounce rate indicates that visitors quickly leave the website without interacting, while a high exit rate suggests users are leaving from a specific page.
- PageValues: This feature assigns a value to each page based on its contribution to conversions. It's a useful metric to identify pages that are most influential in driving sales.



Data Preprocessing

- For categorical variables , we created dummy variables for each value
- Removed rows with Null values from the dataset
- Converted Boolean Values to 0 and 1



Importance of Analyzing Online Shopping Behavior

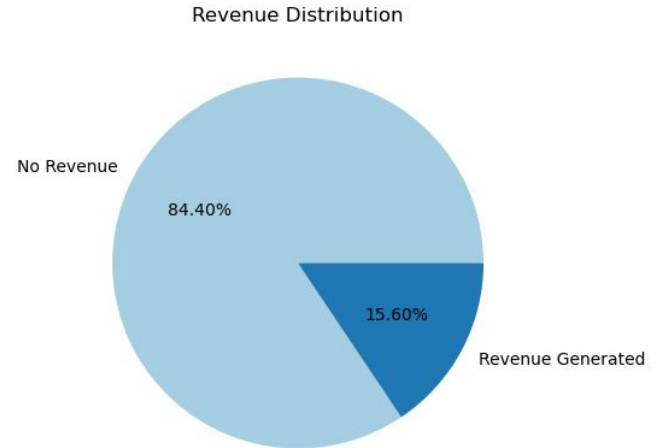
Analyzing online shopping behavior and purchasing patterns can lead to targeted marketing strategies, better customer service and better overall performance.

By understanding the online shoppers behaviors, businesses can increase their conversion rates, thereby increasing revenues and profits.

Basic Statistics about Dataset

Approximately, 85% of the online sessions did not result in a purchase, whereas only 15% resulted in a purchase. Most sessions do not convert into a purchase

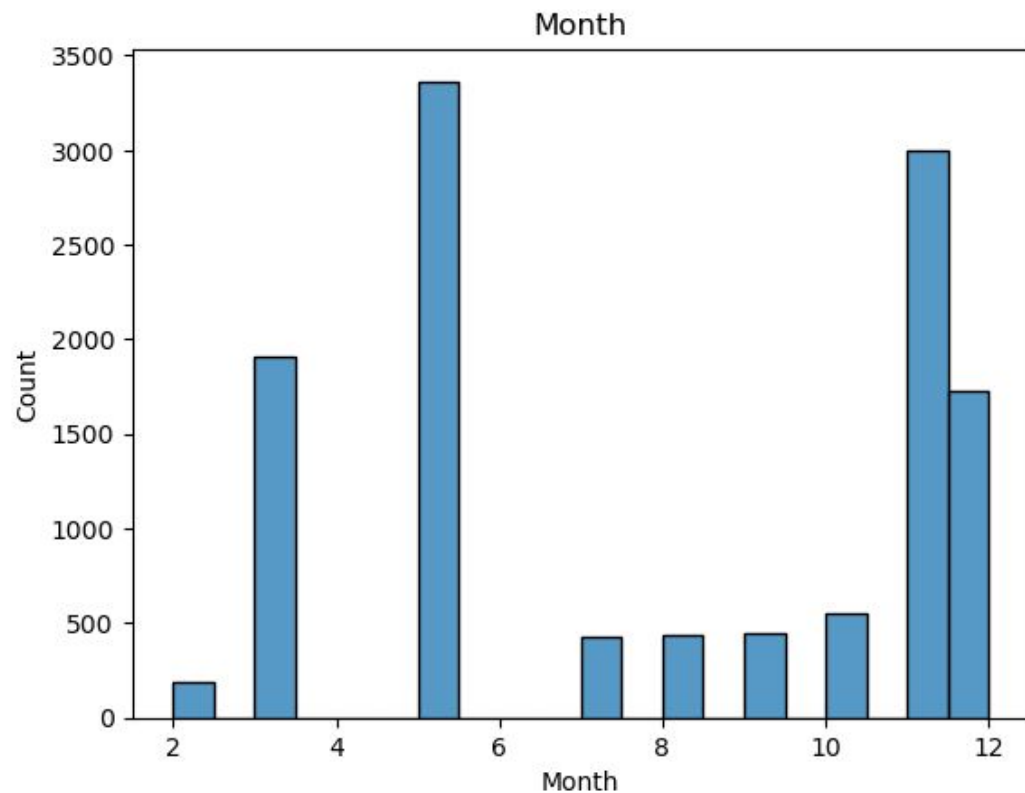
About 86% of the sessions were by returning visitors and about 14% were by new visitors. Majority of the online shopping is done by returning visitors.



Correlation Heatmap of Various Features

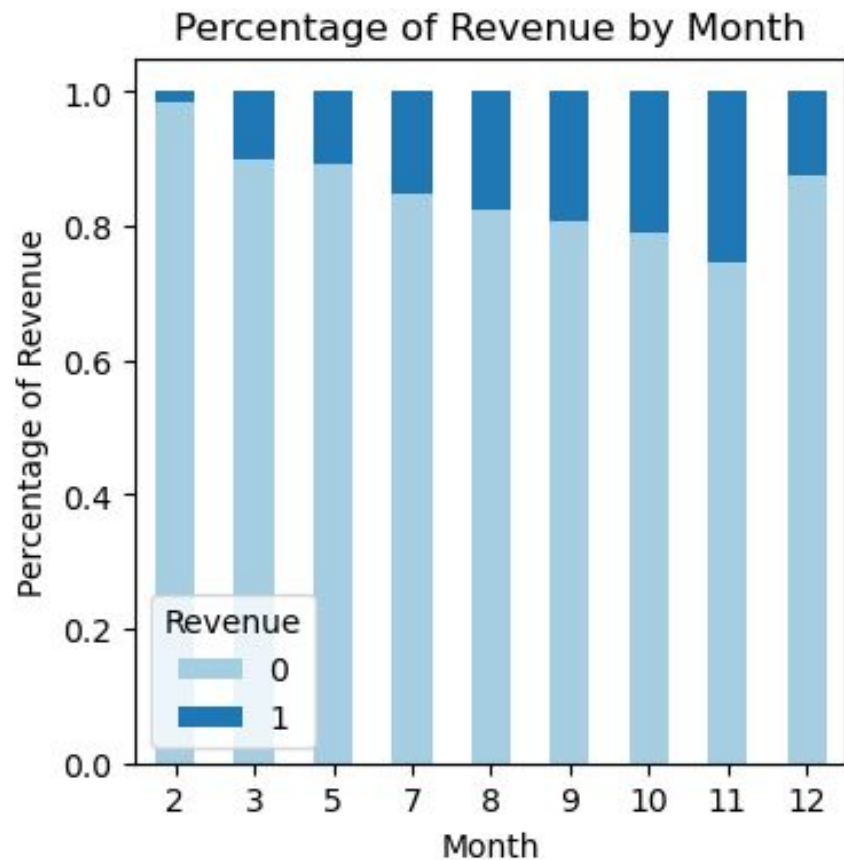


Exploratory Analysis



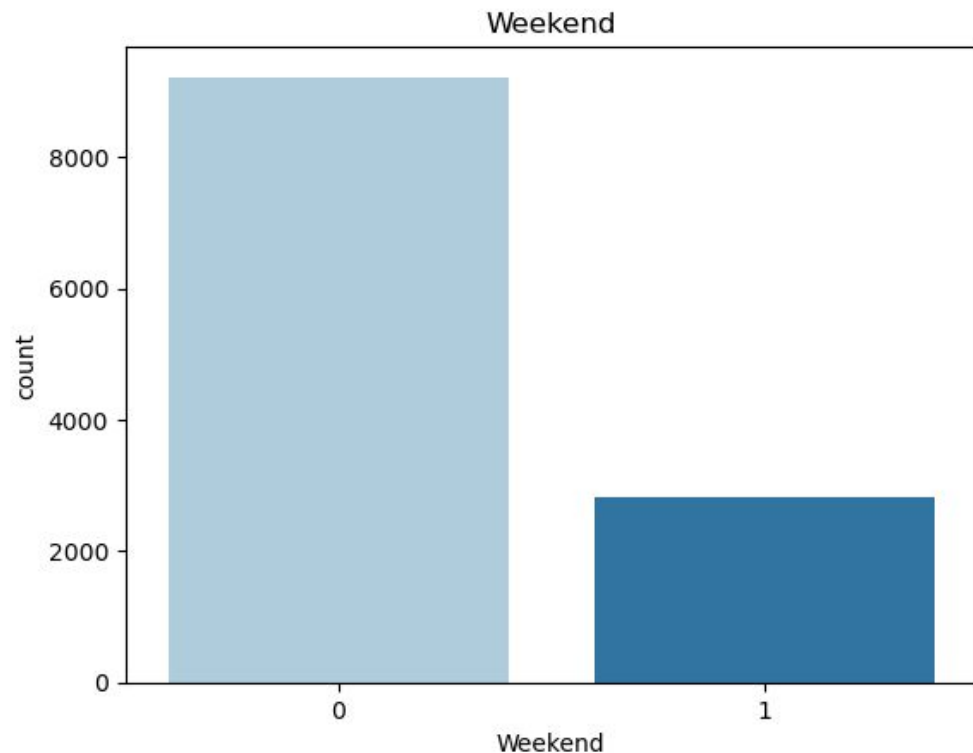
We can see that Months 3(March), 5(May), 11(November), 12(December) have more online sessions than other months

Exploratory Analysis



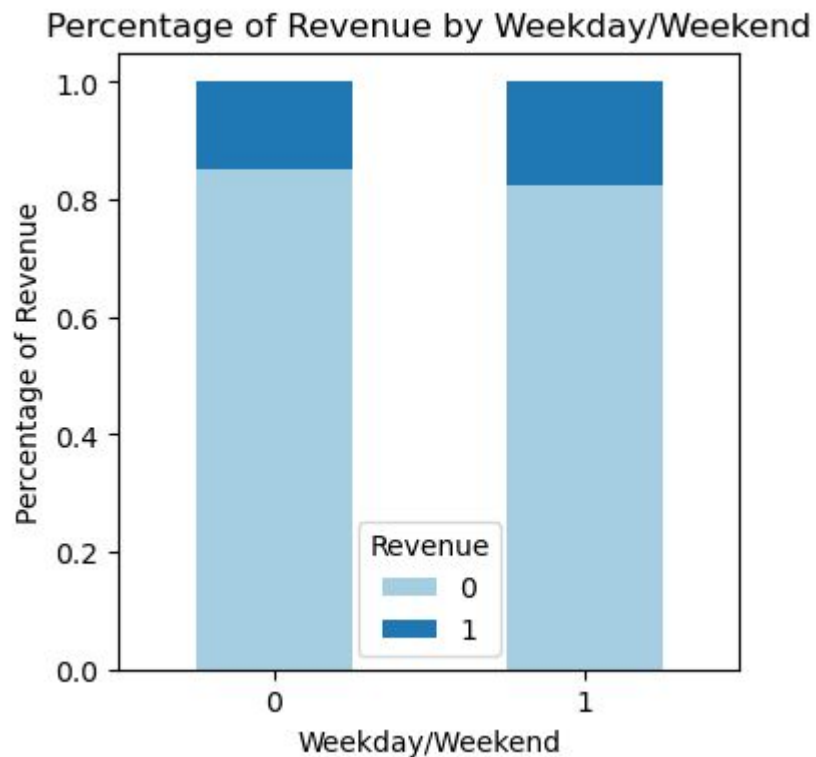
We can see that Month 11(November) has the highest purchase conversion rate

Exploratory Analysis



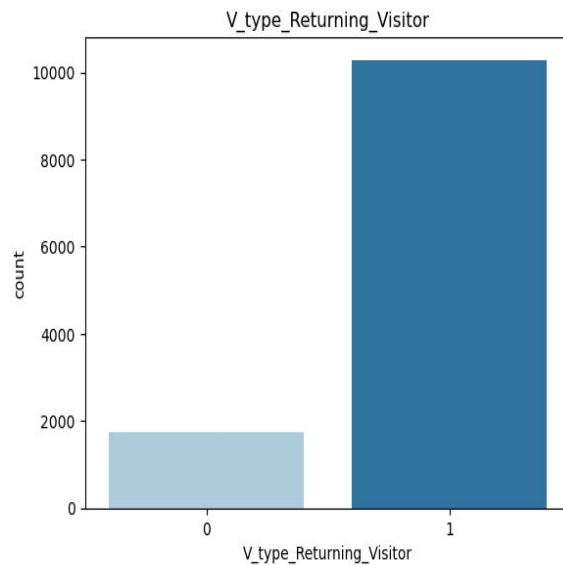
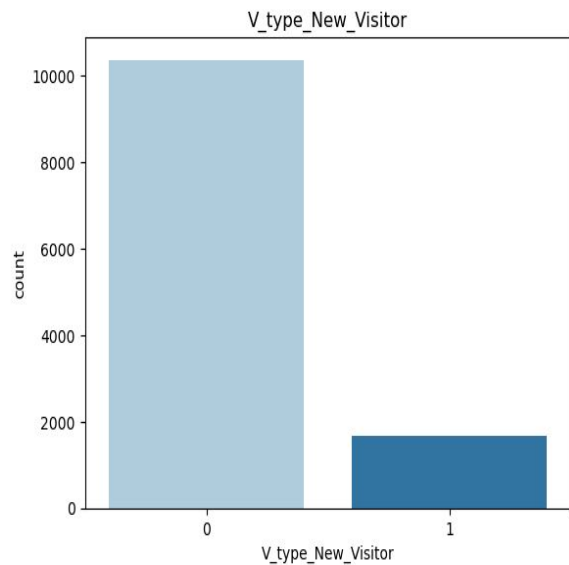
Contrary to expectations,
weekdays receive more online
sessions than weekends

Exploratory Analysis



But weekends have a slightly higher purchase conversion rate

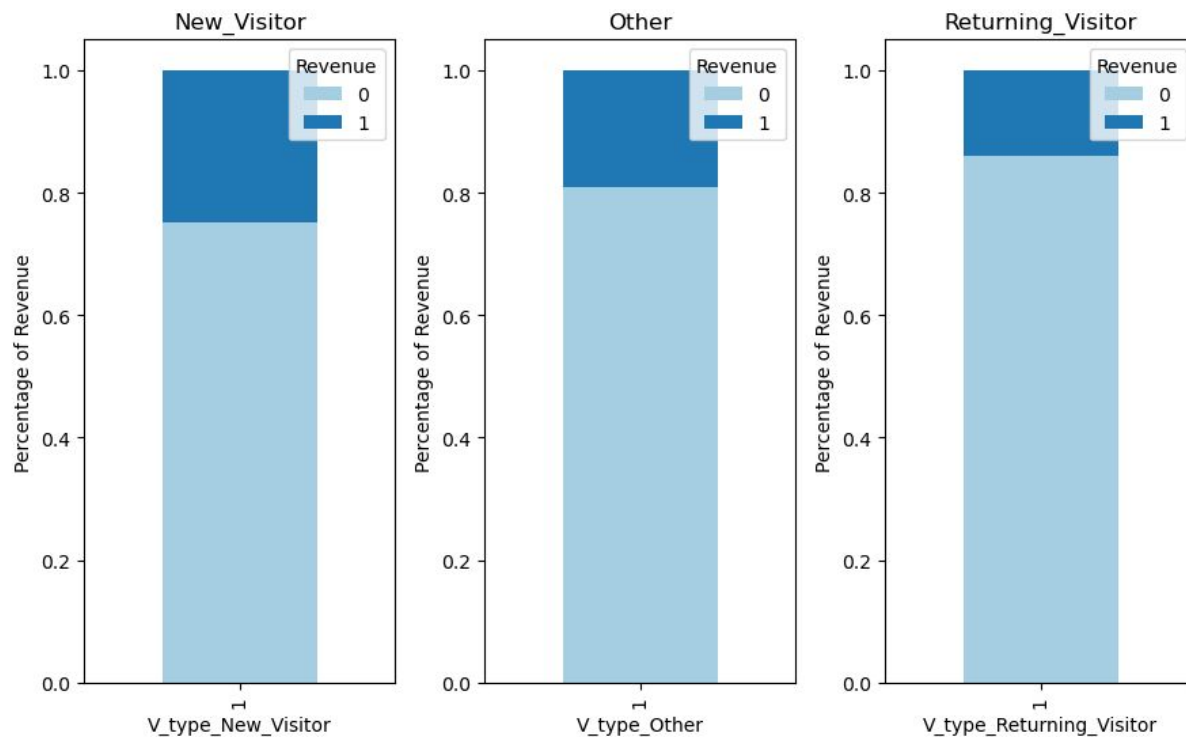
Exploratory Analysis



As stated before,
returning users have
more sessions than new
users

Exploratory Analysis

Revenue Percentage by Visitor Type



Though new visitors have less online sessions than returning visitors, new visitors have a higher purchase conversion rate

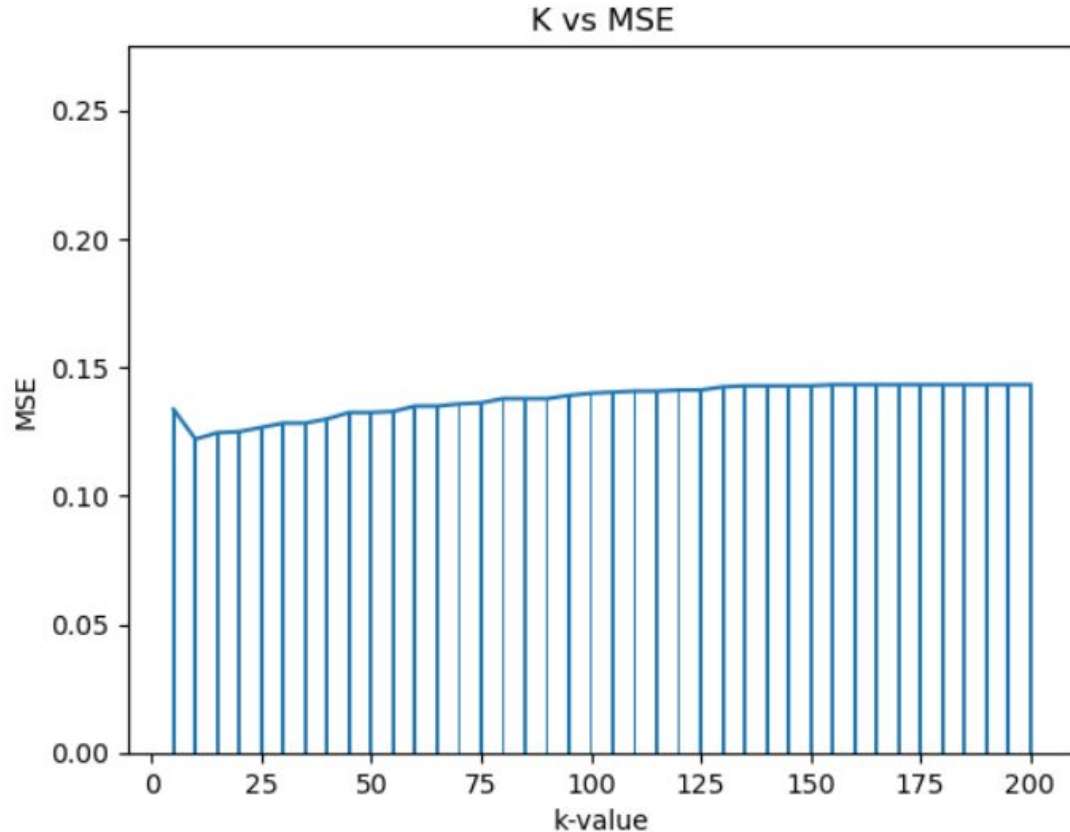


Models Used

We used the following classifiers for our dataset -

- KNN model
- Logistic Regression
- Gradient Boosting Classifier
- Random Forest
- Neural Network

Predicting the Best Value of K for KNN model



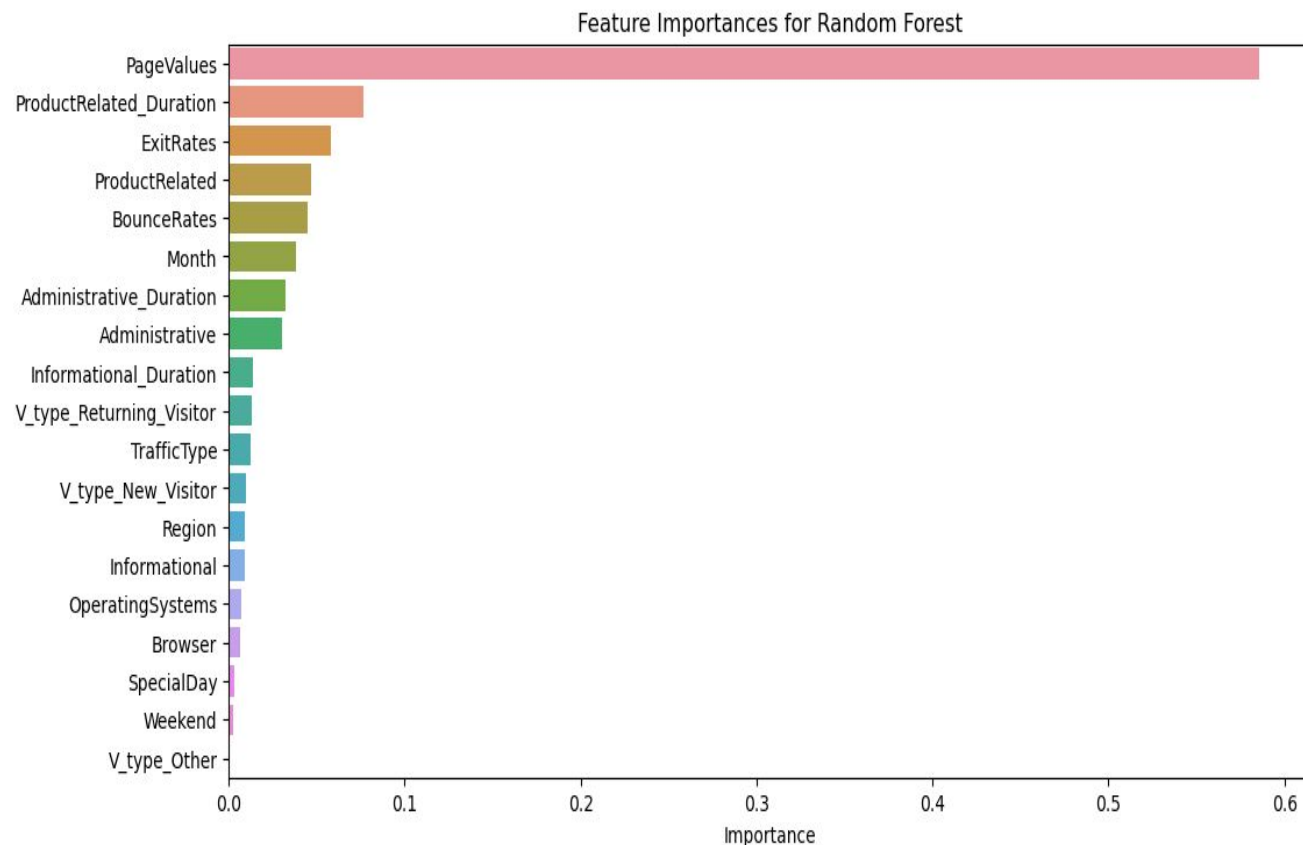
We tried to find the best value of K by optimizing the test MSE. We got best accuracy for $K = 10$

Accuracy and Recall Values of Various Models



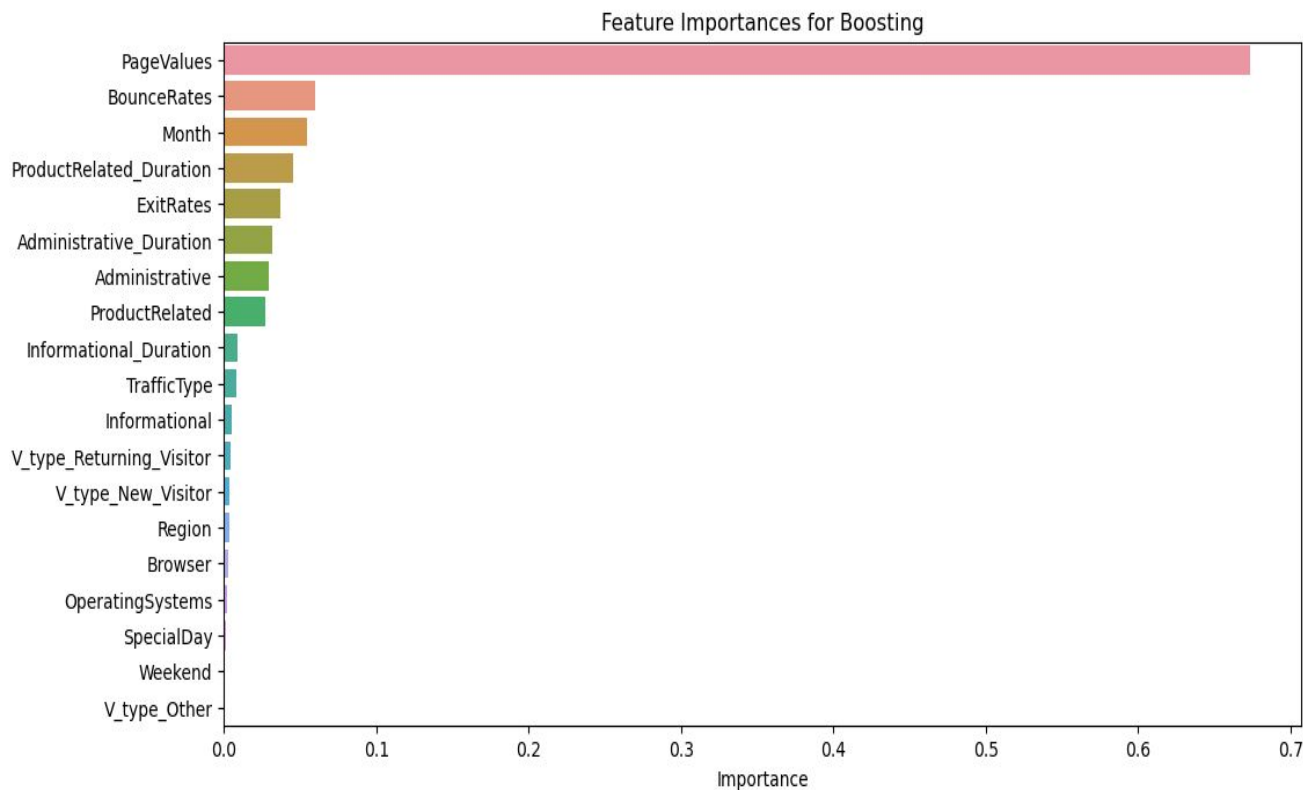
Model	Accuracy	Recall	Best Parameter
Random Forest	0.90	0.54	{'max_depth': 8, 'min_samples_split': 2, 'n_estimators': 50}
Boosting	0.91	0.62	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}
KNN	0.87	0.22	K = 10
Logistic Regression	0.89	0.37	{'C': 1, 'penalty': 'l1'}
Neural Network	0.895	0.58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

Most Important Features : Random Forest



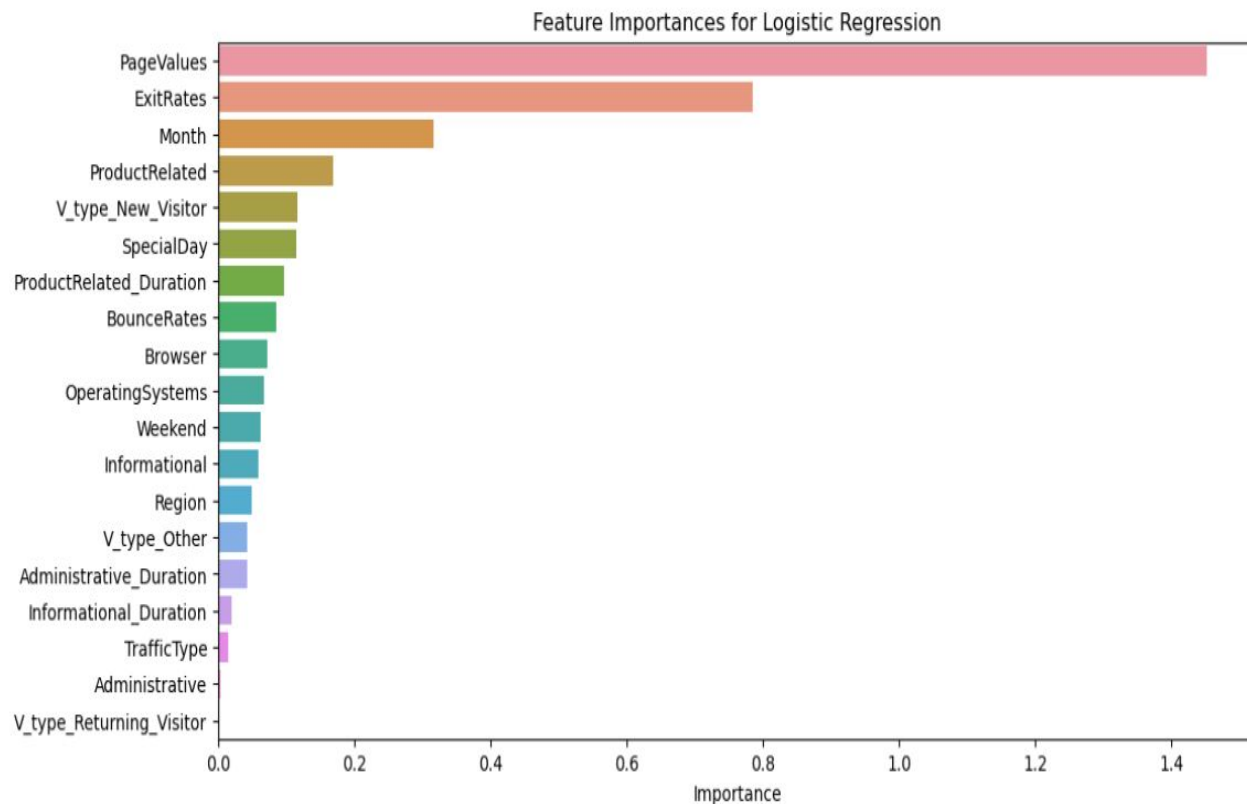
PageValues and ProductRelated_Duration are the most important features according to Random Forest Model.

Most Important Features : Boosting



PageValues , BounceRates
come out as the most
important features in
Boosting Model.

Most Important Features : Logistic Regression



Apart from PageValues ,
ExitRates is also an
important feature for
prediction.


Business Analysis



Assumptions:

- When the model predicts “intention to buy” as False , we place an ad/offer to retain the transaction. The ad/offer incurs some cost.
- There is a fixed cost associated for the product transaction journey on website.
- We assume a value for probability of successful retention of transaction due to the ad/offer. (for our case $P(\text{retention}) = 0.5$).

Confusion Matrix



“Intention to Buy”	Predicted TRUE	Predicted FALSE
Actual TRUE	True Positive	False Negative
Actual FALSE	False Positive	True Negative

Costs

“Intention to Buy”	Predicted TRUE	Predicted FALSE
Actual TRUE	True Positive Cost of Placement = \$3 Cost of Ad/Offer = 0	False Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5
Actual FALSE	False Positive Cost of Placement = \$3 Cost of Ad/Offer = 0	True Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5

Revenue

“Intention to Buy”	Predicted TRUE	Predicted FALSE
Actual TRUE	True Positive Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$40	False Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = \$40
Actual FALSE	False Positive Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$0	True Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = 0\$/40\$ Probability of Retention : 0.5

Net Profit

“Intention to Buy”	Predicted TRUE	Predicted FALSE
Actual TRUE	True Positive Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$40 Net Profit = $(40-3) = \$37$	False Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = \$40 Net Profit = $(40-3-5) = \$32$
Actual FALSE	False Positive Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$0 Net Profit = $(0-3) = -\$3$	True Negative Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = 0\$/40\$ Probability of Retention : 0.5 Net Profit = $(0.5*40-3-5) = \$12$

Overall Profit

“Intention to Buy”	Predicted TRUE	Predicted FALSE
Actual TRUE	True Positive (129) Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$40 Overall Profit = $(40-3)*129 = \text{\$4,773}$	False Negative (220) Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = \$40 Overall Profit = $(40-3-5)*220 = \text{\$7,040}$
Actual FALSE	False Positive (45) Cost of Placement = \$3 Cost of Ad/Offer = 0 Avg Item Revenue = \$0 Overall Profit = $(0-3)*45 = \text{\$-135}$	True Negative (2015) Cost of Placement = \$3 Cost of Ad/Offer = \$5 Avg Item Revenue = 0\$/40\$ Probability of Retention : 0.5 Overall Profit = $(0.5*40-3-5)*2015 = \text{\$24,180}$

Model Value



- **Profit with Model:**

$$Rev(TP) + Rev(FN) + Rev(FP) + Rev(TN)$$

$$4773 + 7040 + (-135) + 24180 = \$35,858$$

- **Profit without model** (loss from true negatives / less costs for false negatives):

$$4773 + 8140 + (-135) + (-6045) = \$6,733$$

Model adds value !

Conclusion



- “PageValues” and “ProductRelated_Duration” are crucial indicators of a consumer's purchasing intention.
- High engagement with product-related pages and 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.
- Relating to time period frequencies, Month was observed to be a much more important feature than Weekend



Conclusion

- models : Random Forest, Gradient Boosting Classifier, and K-Nearest Neighbors (KNN), Logistic Regression and Neural Network.
- Exploratory Analysis
- Business Analysis : Profit increases with model



Thank you!