ONLINE SHOPPERS PURCHASING INTENTION

05-DEC-2021

BY KRISCHELLE JOYNER

Background

As the majority of businesses today operate online, marketing promotion is one of the most effective ways to drive traffic to an online virtual ecosystem. As visitors browse a website, the most relevant users are identified, and can be connected with offers that will encourage them to return to a store after they have left, avoiding the high probability of losing customers after they leave.

Marketing data is becoming a powerful tool for companies to not only understand what products customers are interested in, but to also explore some of the reasons why a product is not purchased. The goal of using marketing data is to identify the underlying patterns of behavior.

Motivation

The motivation for using this dataset is to understand visitors' shopping intent. A growing number of people search for items to buy online and make purchases through online transactions; this has made their lives more convenient and easier. Yet sellers must also be aware of the patterns and intentions of different types of online visitors. Customer behavior insights can help sellers to target advertising, marketing, and deals to potential customers to further increase their sales and revenue.

Specifically, I am interested in purchases as an indicator of good marketing on products. The analysis will primarily focus on:

Which factors heavily influenced visitors to purchase an item?

Using *Revenue* as the dependent variable (TRUE – made a purchase, FALSE – did not make a purchase), I will explore 18 independent variables such as bounce rate, the type of visitor and several other characteristics to see which variables worked best for indicating a purchase.

Hypothetical research questions:

How long does a visitor stay on a page?

Understanding the attention span for most individuals is very important when thinking about the layout of the page and number of steps it takes from visiting the site to making a purchase. I can get a better idea of this by viewing the average page time per visitor.

What type of pages does the customer visit?

This behavior can help to understand if the customer knew what they were shopping for initially, or were browsing and open to other items as well.

How many visitors are repeat customers?

Future marketing campaigns can benefit greatly from knowing the type of visitor on the site.

Are customers more likely to buy on or around holidays?

If this is the case, stores would have more incentive to change prices around high demand periods.

Description of the Data

The data set consists of feature vectors from 12,330 sessions. Each session was allocated to a unique user within a 1-year window in order to avoid a tendency toward a specific campaign, special day, or user profile. Within the dataset, 84.5% of sessions (10,422) were examples of negative classes that did not end with shopping, while the remaining 15.5% (1,908) were examples of positive classes that did end with shopping.

The data was gathered a few different ways. The page type and duration were gathered from the URL, while the remaining variables, other than *Special Day*, were gathered from Google Analytics.

The three different page types, Administrative, Informational, and Product Related information were derived from the URL information of the pages the user visited and updated in real time when the user took action. The duration for each of these page types was calculated in hours by month.

1. **Administrative:** A person who determines the site policies, appoints moderators, and manages the technical operations of the website.

2. **Administrative_Duration:** Hours spent on administrative pages.

3. **Informational:** An informational website basically provides information about a business and its products and services.

4. **Informational_Duration:** Hours spent on informational pages.

5. **Product_Related:** A product related page is a web page on an eCommerce website that provides information on a specific product. This information includes size, color, price, shipping information, reviews, and other relevant information customers want to know before purchasing.

6. **ProductRelated_Duration:** Hours spent on product-related pages.

7. **Bounce Rate:** The percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session. A website’s bounce rate is calculated by dividing the number of single-page sessions by the number of total sessions on the site.

   Website Bounce Rate = single-page sessions / total sessions

8. **Exit Rate:** The value for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. You can calculate exit rate by dividing the total amount of exits from a page by the total amount of visits to that page.

   Exit Rate: # of page exits / # of page visits

9. **Page Value:** Page Value is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both). This value is intended to give you an idea of which page in your site contributed more to your site's revenue. If the page wasn't

involved in an ecommerce transaction for your website in any way, then the Page Value for that page will be $0 since the page was never visited in a session where a transaction occurred.

Page value = (Total page value + Transaction revenue) / Total unique pageviews

10. **Special Day:** The *Special Day* feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentine's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

11. **Month:** The month of the year visited.

12. **OperatingSystems:** The computer operating system of the visitor.

13. **Browser:** The computer browser of the visitor.

14. **Region:** Geographic region from which the session has been started by the visitor.

15. **TrafficType:** Traffic source by which the visitor has arrived at the website.

16. **VisitorType:** The visitor type is "New Visitor", "Returning Visitor" or "Other".

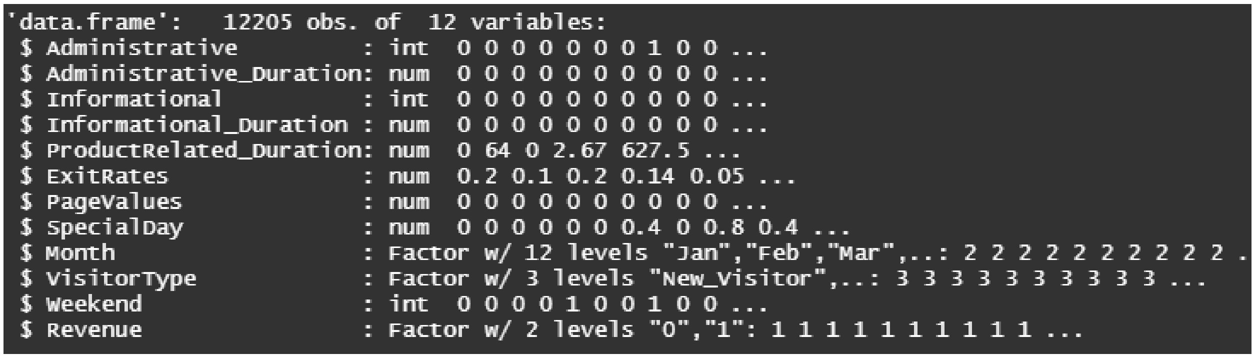
17. **Weekend:** The weekend value is a Boolean value of TRUE or FALSE of whether it is the weekend or not.

18. **Revenue:** Class label indicating whether the visit has been finalized with a transaction.

I performed the following tasks to pre-process and clean the data, and adjust for its limitations.

1. I removed 125 duplicate records from the data set leaving a total of 12205 records.
2. As far as the data quality is concerned, there were no missing data points to remove.
3. I changed the Weekend variable to an integer, and the following variables to factors as they have multiple classes:
    - OperatingSystems
    - Browser
    - Region
    - TrafficType
    - VisitorType
    - Revenue
    - Month
4. I noticed a correlation between the "BounceRates" and "ExitRates" variables, so I removed the "BounceRates" variable from the dataset. I also noticed a correlation between the "ProductRelated" and "ProductRelated_Duration" variables, so I removed the "ProductRelated" variable from the dataset.

5. I removed the following variables as the numbers do not have good references for the outcomes:
    - OperatingSystems
    - Browser
    - Region
    - TrafficType



```
'data.frame':   12205 obs. of  12 variables:
 $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
 $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
 $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
 $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
 $ Month                  : Factor w/ 12 levels "Jan","Feb","Mar",..: 2 2 2 2 2 2 2 2 2 2 .
 $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Weekend                : int  0 0 0 0 1 0 0 1 0 0 ...
 $ Revenue                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Proposed Analysis

The online shopper's intent dataset will be analyzed using logistic regression, random forest, and naive bayes models, and include visual representations of the data analysis to aid in the interpretation of the results. These are great models since the features are categorical.

I will run several logistic regression models to see which one yields the best results, including: no cross-validation, 5-fold cross-validation, and 5-fold repeated cross-validation. The logistic model should be easy to train and interpret. Using the cross-validation approach should reduce statistical noise and produce a better accuracy than the model without cross validation. Logistic regression is a good model for categorical independent variables and one dichotomous dependent variable, making it a good choice for this data.

Random forest models are also expected to be used on the data since they analyze the implications of the decisions made by customers. Random forest creates several iterations of decision trees based on different decisions, it produces the model with the best attributes and highest accuracy.

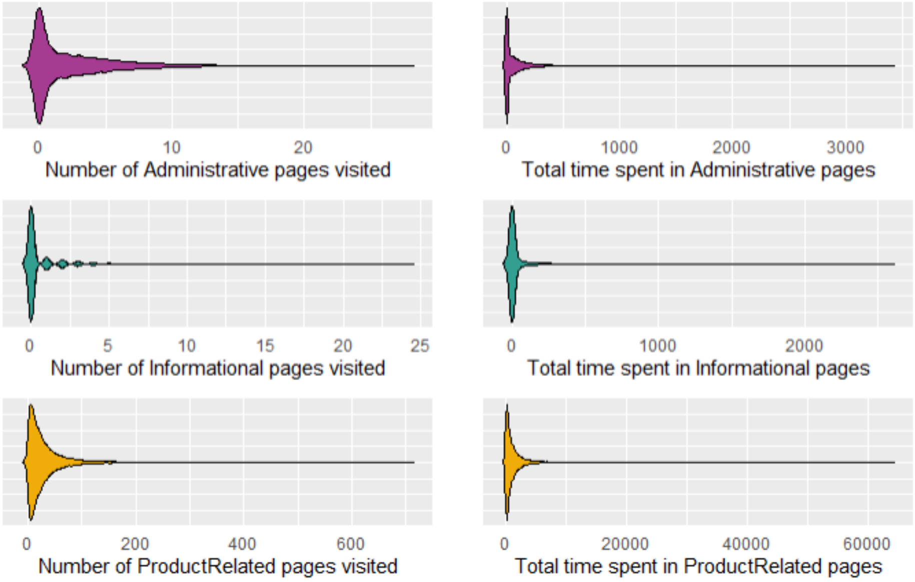
The Naïve Bayes Classifier uses the Bayes fusion rule to make a probabilistic classifier. It assumes that input features are independent and each contributes independently to the probability of a class and ignores any correlations that might exist. The final decision is then assigned to the class with the highest probability. Naive Bayes can outperform more sophisticated classification techniques despite its simplicity, so it is a good consideration for the data.

Data Exploration

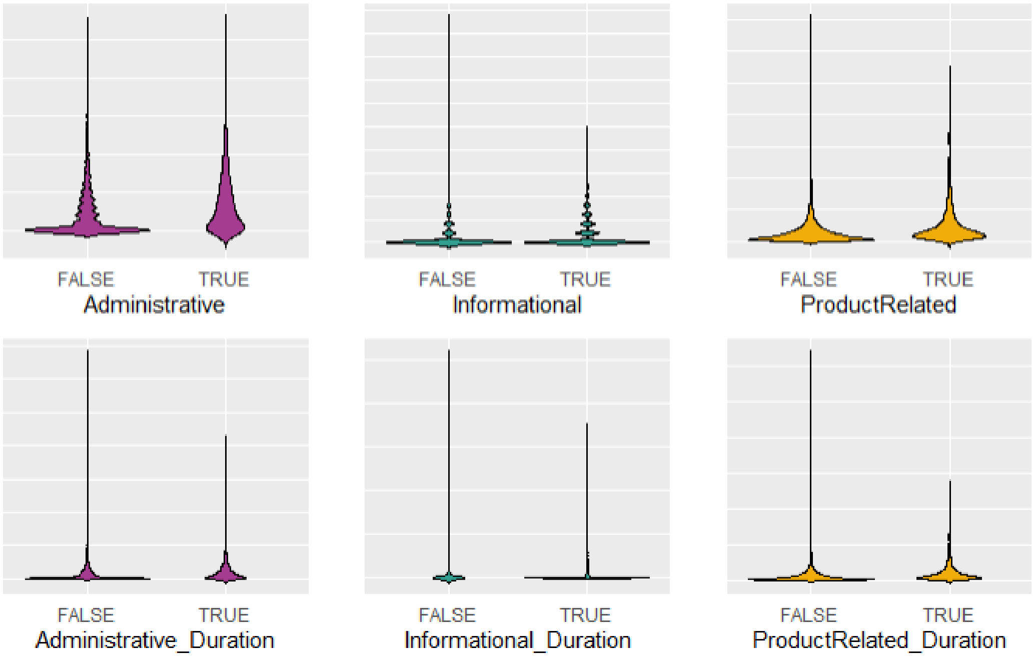
According to our page attributes, customers visited different types of pages and spent most of their time looking at related products. Only a small fraction of visitors chose to dig into information about one product.

Median # of pages visited
Administrative pages: 1
Informational pages: 0
ProductRelated pages: 18

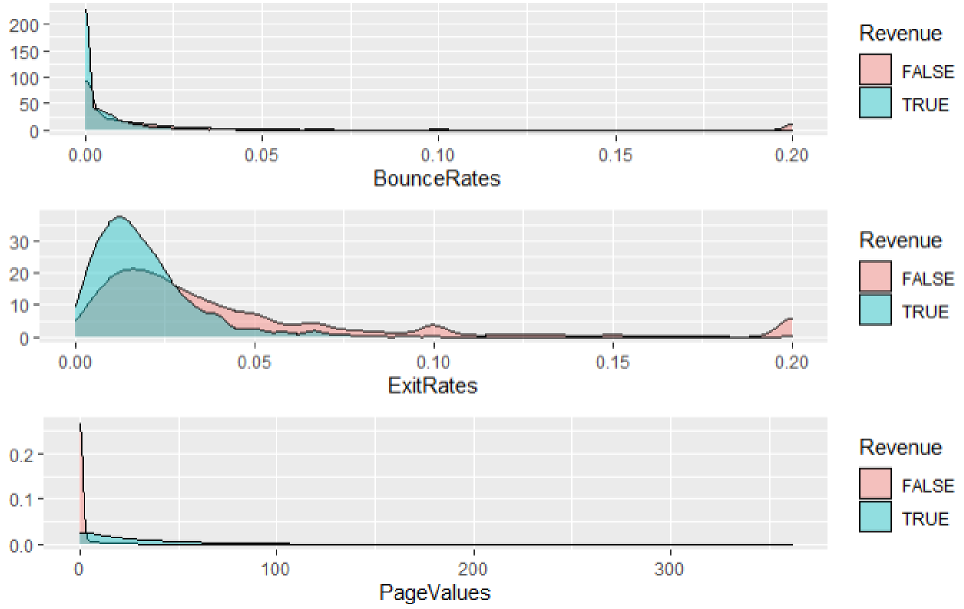
Median Time spent on pages
Administrative pages: 9
Informational pages: 0
ProductRelated pages: 608.9



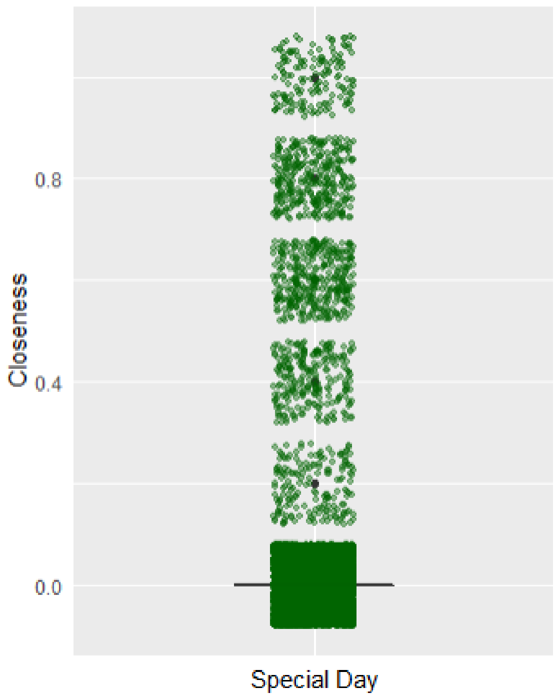
It is surprising that most of the customers who completed transactions spent more time browsing Administrative and ProductRelated pages than Informational pages, since this indicates that they are loyal customers who checked out after adding items to their carts.

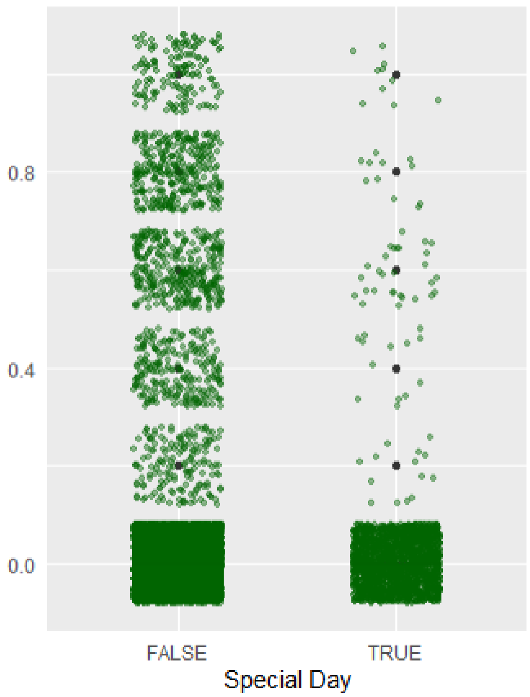


BounceRates, ExitRates, and PageValues provide insights into the behavior of customers viewing pages. There doesn't appear to be a significant difference in revenue when examining customers' BounceRates. On the other hand, when looking at ExitRates, the plot shows that customers who bought an item are in general less likely to leave than customers who did not buy an item, since they spent more time on the pages. The Pagevalues of non-purchasing customers are much lower than those of purchasing customers because they spend less time on related pages.

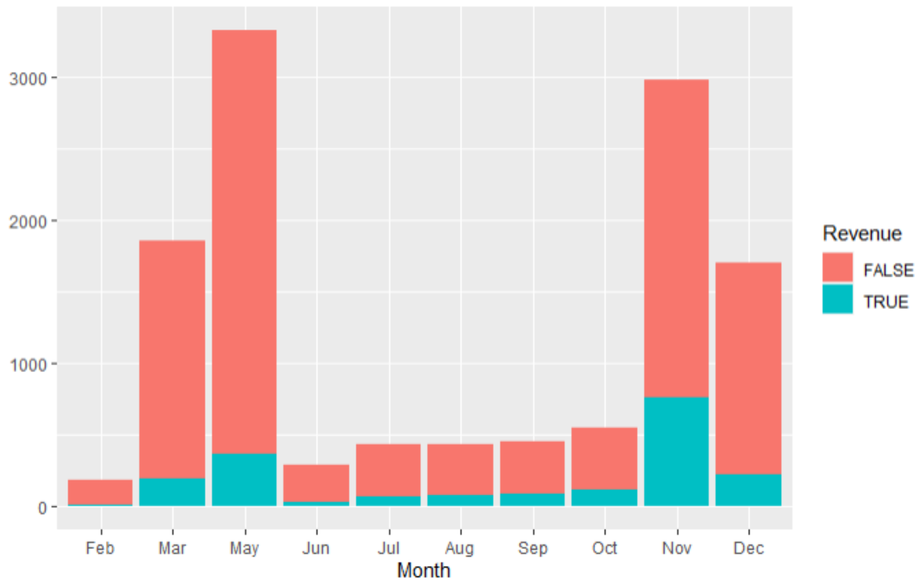


On non-special days, purchasing customers were more likely to make purchases. This supports my observation that most customer decisions are not influenced by whether it is a special day or not.

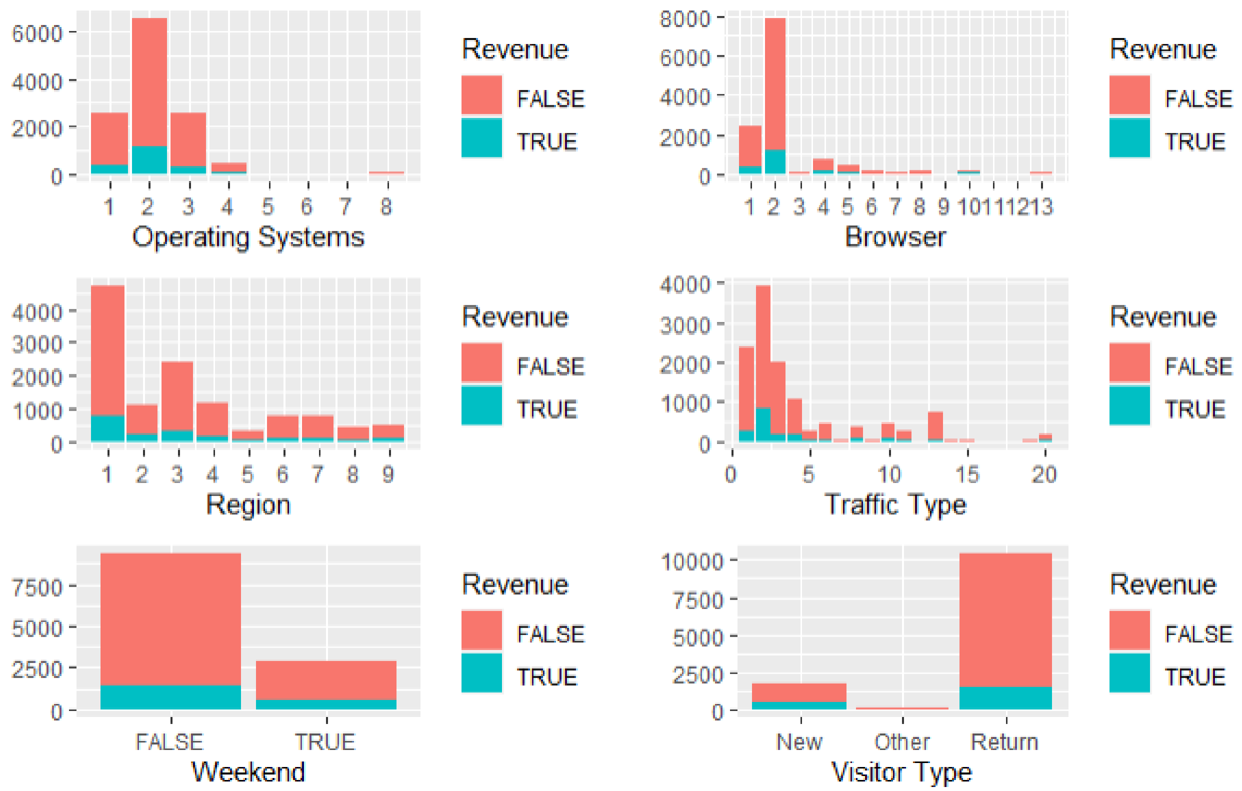




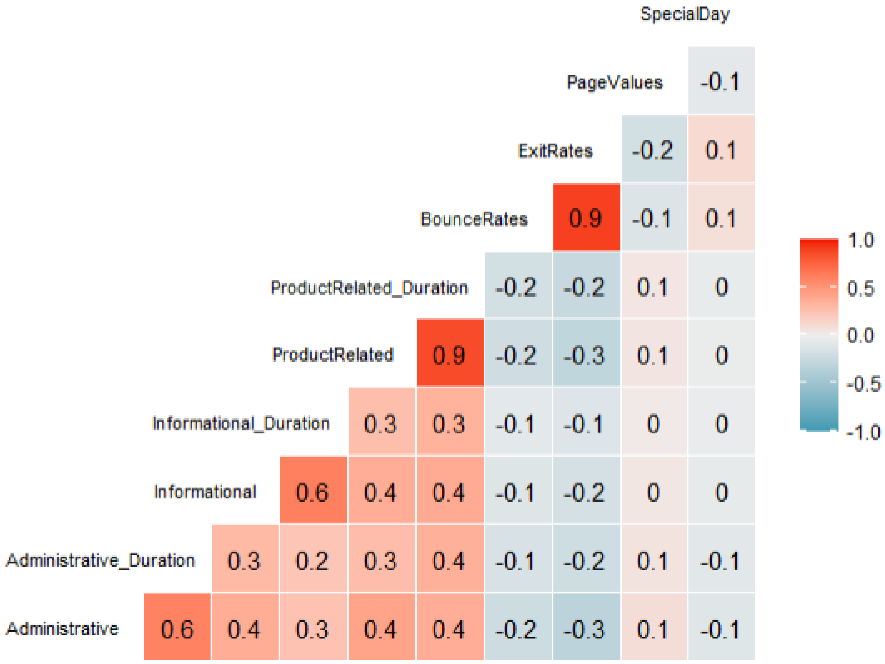
It should be noted that the month attribute displays only 10 of the 12 months, excluding January and April. The majority of shopping took place during March, May, November, and December. This might be due to these months being right before a new season.



Without knowing the numbers, the Operating Systems, Browser, Region, and Traffic Type attributes are not much help. The Weekend attribute shows more purchases on weekdays, while the Visitor Type is mostly return customers.



Checking correlations among variables in the data set. I see several high-correlated pairs like BounceRates/ExitRates and ProductRelated/ProductRelated_Duration that are evident; one of each pair might be dropped because of its importance to our model.

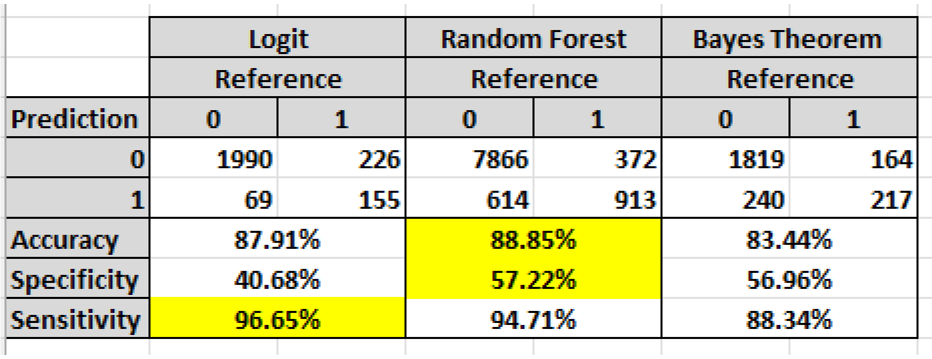


Results

Variable Importance Logistic Regression models
PageValues
ExitRates
MonthNov
ProductRelated_Duration
VisitorTypeReturning_Visitor
MonthJul
MonthAug

Variable Importance Random Forest model
Page Values
ExitRates
ProductRelated_Duration
Administrative
MonthNov
Administrative_Duration
VisitorTypeReturning_Visitor

Naïve Bayes
Naive Bayes classifiers do not offer an intrinsic method for evaluating feature importance. Instead, they compute the conditional and unconditional probabilities associated with the features and predict the class with the highest likelihood.



	Logit		Random Forest		Bayes Theorem	
	Reference		Reference		Reference	
Prediction	0	1	0	1	0	1
0	1990	226	7866	372	1819	164
1	69	155	614	913	240	217
Accuracy	87.91%		88.85%		83.44%	
Specificity	40.68%		57.22%		56.96%	
Sensitivity	96.65%		94.71%		88.34%	

The differences in accuracy are close, but the Random Forest model performed the best.

Accuracy
I found the PageValues, Exit Rates, MonthNov, ProductRelated_Duration, VisitorTypeReturning_Visitor, MonthJul, and MonthAug variables to be statistically significant with an alpha of .05 in the logistic regression models. This means that several of my independent variables influenced visitors to make a purchase. Although the logistic regression accuracy, specificity, and sensitivity is the same for all three models, I would use the repeated cross-validation if I were to use it as my final model. However, my Random Forest accuracy was higher, so I'll be using it as my final model. Bayes performed the lowest among Accuracy, Specificity, and Sensitivity.

Specificity and Sensitivity
It is apparent that the Sensitivity percentages are high for all three models and the Specificity is low. It seems to be a cause of unbalanced data. If you remember, 84.5% of the sessions contained in the dataset (10,422) do not end with shopping, while the remaining 15.5% (1,908) ended with shopping which creates this imbalance. This is because the model guesses no shopping on most occurrences and is correct.

Since my final model will be the Random Forest model, I'll go into the details for Specificity and Sensitivity. The sensitivity is 94.7% and the specificity is 57.2%. This model will correctly identify 94.7% of the visitors who did not make a purchase, but it will also fail to identify 5.3%. The model will correctly identify 57.2% visitors who made a purchase, but it will also identify 42.8% of visitors as making a purchase when they will not. These numbers aren't looking very good, indicating the importance of balancing data.

Conclusion

The dataset gives us a lot of insights into how consumers behave. It is important to remember that our generalizations can sometimes be inaccurate based on our day-to-day experiences. The major takeaways from 'Online Shoppers Purchasing Intent' are the following:

- The factors that heavily influenced visitors to purchase an item were PageValues, Exit Rates, ProductRelated_Duration, Administrative, MonthNov, Administrative_Duration, VisitorTypeReturning_Visitor, from the Random Forest model.
    - The page values and exit rates variables show how important certain pages are in completing a purchase.
    - The type of pages reviewed the longest were administrative and similar products to the shopper's search.
    - The most popular times to buy were near season changes rather than holidays (or special days from my analysis). Shoppers spent most in the month of November.
    - The majority of purchases came from our loyal customers.

Recommendations

This data set suffers from unbalanced data, so if anyone wants to use it in the future, I suggest that they fix that when they clean the data.

Each of the three models works well for classification data. They were all close in prediction percentages, and they are efficient to train and easy to interpret. I would recommend using them in future classification problems.

References

Sakar, C. Okan, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. “Real-Time Prediction of Online Shoppers’ Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks.” *Neural Computing and Applications* 31, no. 10 (October 2019): 6893–6908. https://doi.org/10.1007/s00521-018-3523-0.

Liu, Yuming. Online Shoppers Purchasing Intention. https://rstudio-pubs-static.s3.amazonaws.com/588410_b71a2f1ad47c4eafa145c424f4fc0faf.html.

Baati K., Mohsil M. (2020) Real-Time Prediction of Online Shoppers’ Purchasing Intention Using Random Forest. In: Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1_4

UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data set. (n.d.). Retrieved December 5, 2021, from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.