Springboard Capstone Project 1: Analysis of Online Shopping Dataset

Kaggle Dataset Repository

Report by: Michael Thabane

February 12, 2020

Contents

1 Introduction

1.1 Explain Business Problem

With the development of technology and the rapid increase of online shopping it is important to know if the customer's intention is to truly make a purchase. Since the customer is no longer going into a store to create open dialogue with the seller we must find different ways to know the intentions of the consumer. Using data from the sessions of the users on the website we need to answer the question about which variables have an effect on users making a purchase and how can we use these variables to predict the intention of the customer. This is a supervised classification data science problem in which we can use multiple machine learning algorithms to help classify if the users intention is browse or make a purchase.

1.2 Executive Summary

This report provides an analysis and classification of the online shopper's decision to make a purchase based website session data. The methods used for the analysis include a heat-map, box-plots, chi-square hypothesis test and a two-sample t-test. The methods involved with classification include k-nearest neighbors, logistic regression, random forest and xgboost algorithm. The results of the analysis show that there is a significant difference in the user's intention to make purchases on either a weekend or special day. The results of the classification show that the xboost model is the best classifier for a user making a purchase with a prediction accuracy of 90% (+/- 0.03). The Page Values, Bounce and Exit Rates have also been found to be the most important parameters in model classification. Based on our findings in the report I would recommend the website owner to examine ways to optimize sales on both special days and weekend. It also may be worthwhile to investigate how to engage users more on the website in order for the Page Values per session to increase and the Exit and Bounce Rates to decrease.

2 Approach

2.1 Data Acquisition and Wrangling

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to 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. The value of "Exit Rate" feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. 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. The data set also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the

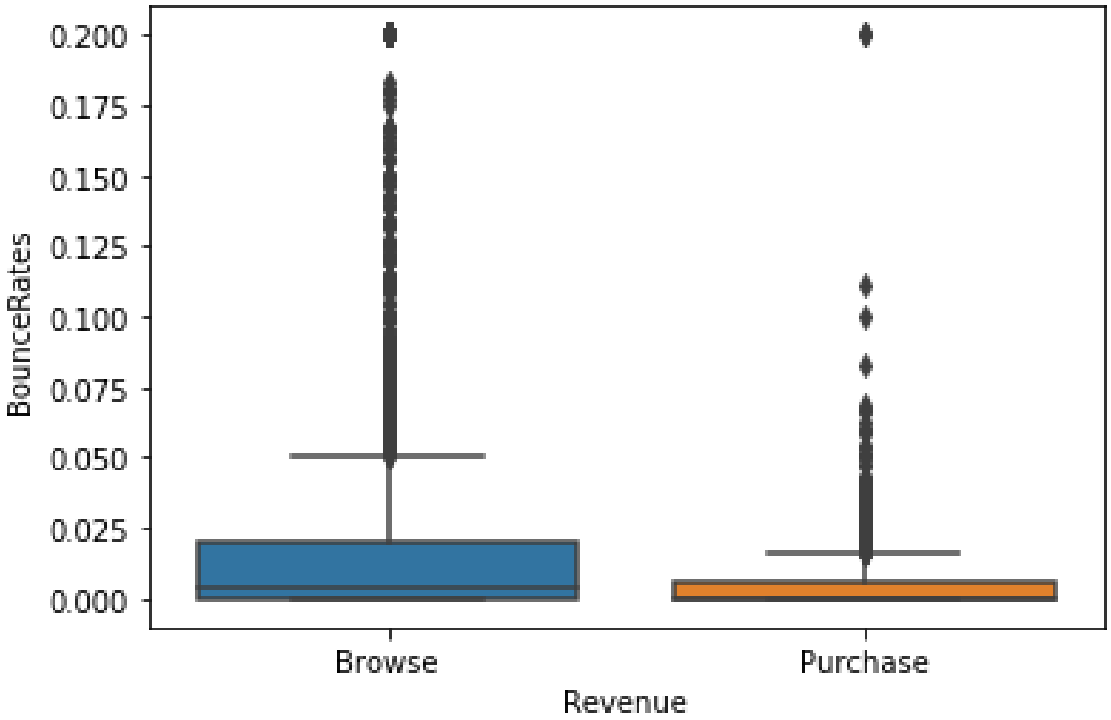
year.

The only wrangling needed was to separate the Purchase and Browsing data so we can create data stories with meaning.

2.2 Storytelling and Inferential Statistics

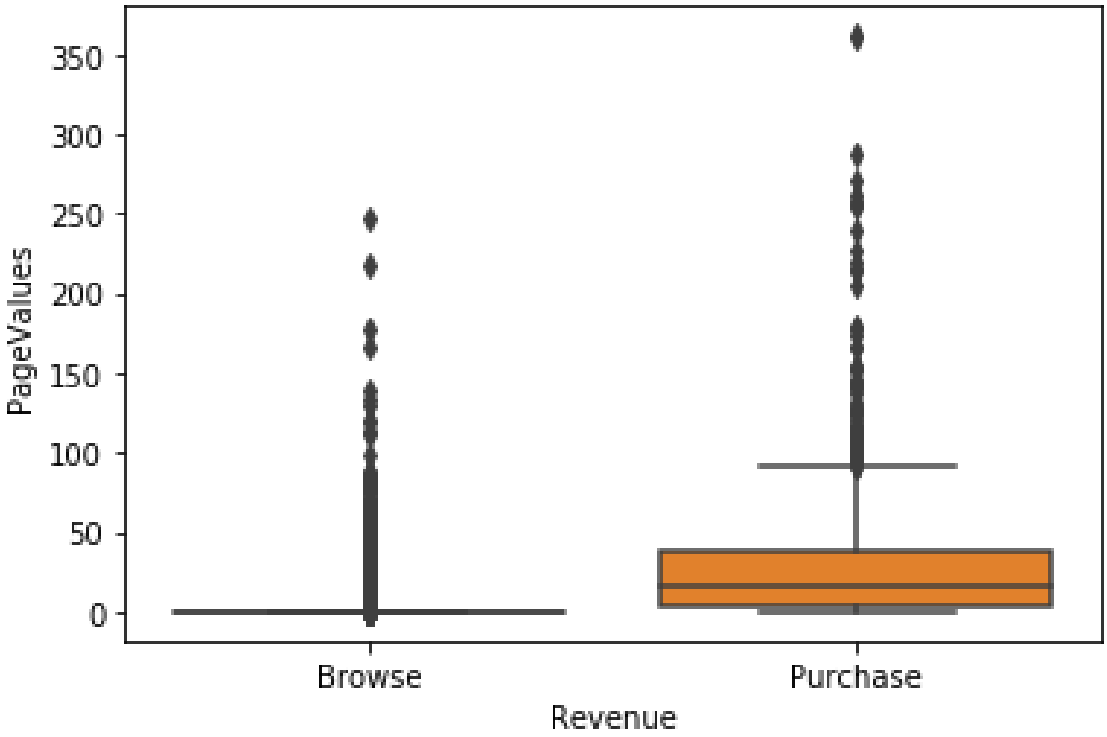
2.2.1 Descriptive Analysis

First we run descriptive analysis on both the groups of data (Purchasing, Browsing). Based on this descriptive analysis we can discover which variables are continuous and which variables are categorical and create the necessary plots to describe the data (can be found in the corresponding Appendix Jupyter Notebook). This analysis shows us that the Bounce Rates for Customer's who are just Browsing is higher than Purchasing customers on average as we can see graphically with the below boxplot.

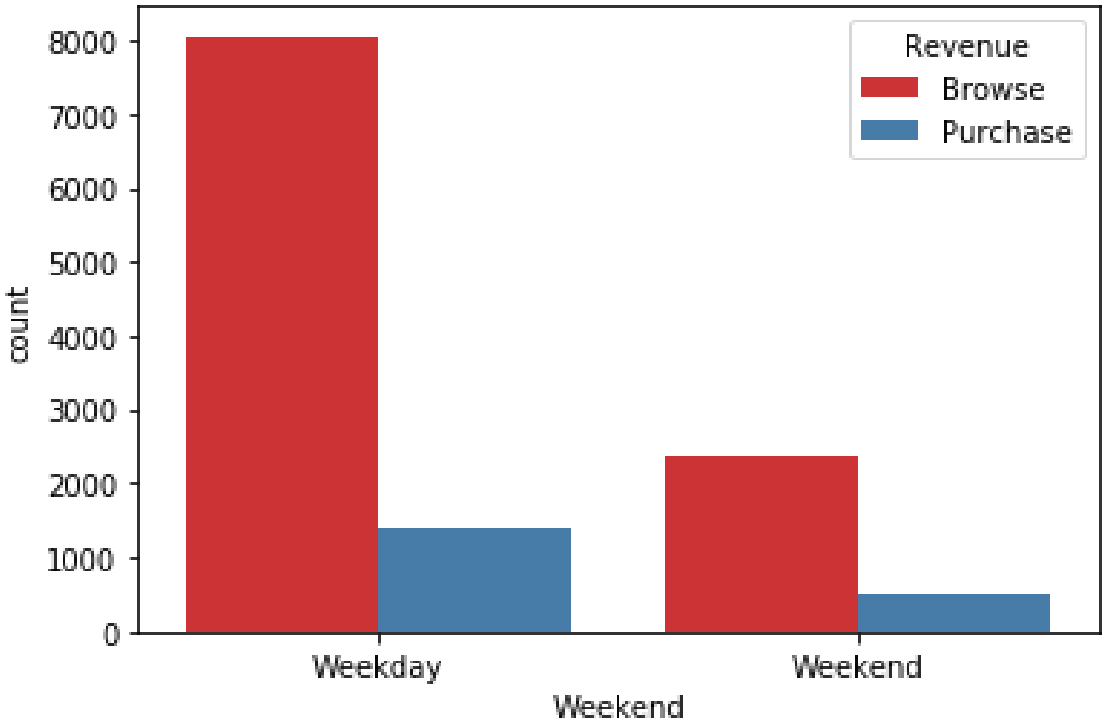


Also, Browsing Customers visit less pages on average than Purchasing Customers

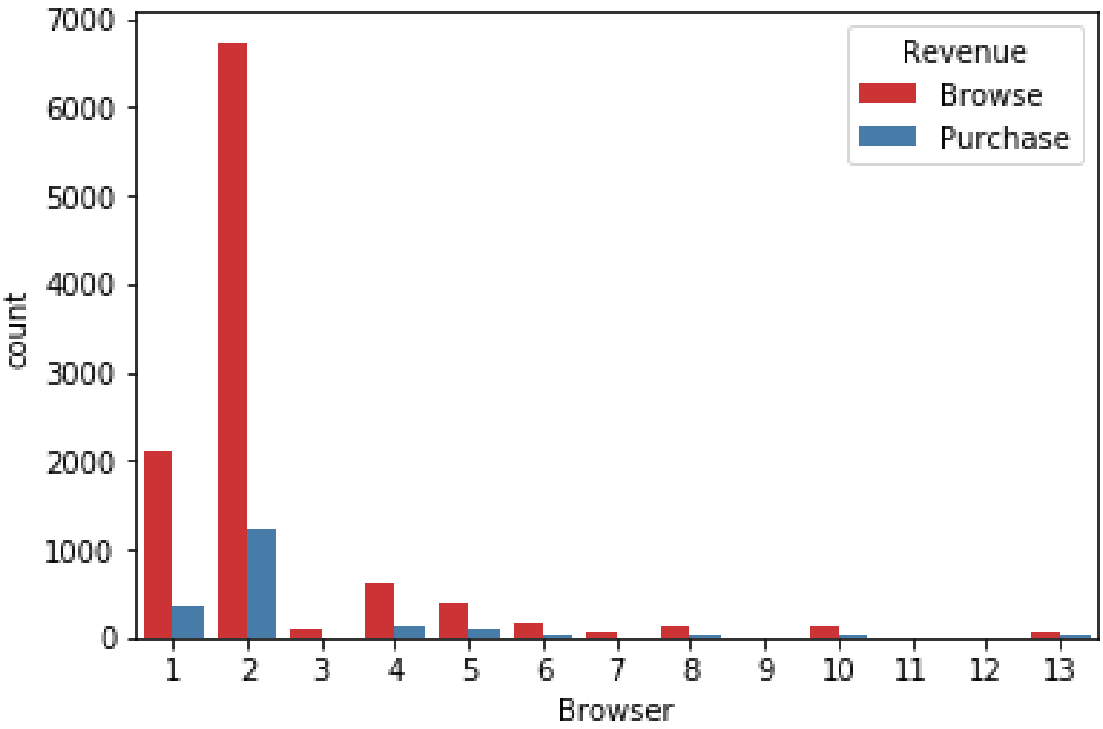
illustrated in the boxplot below.



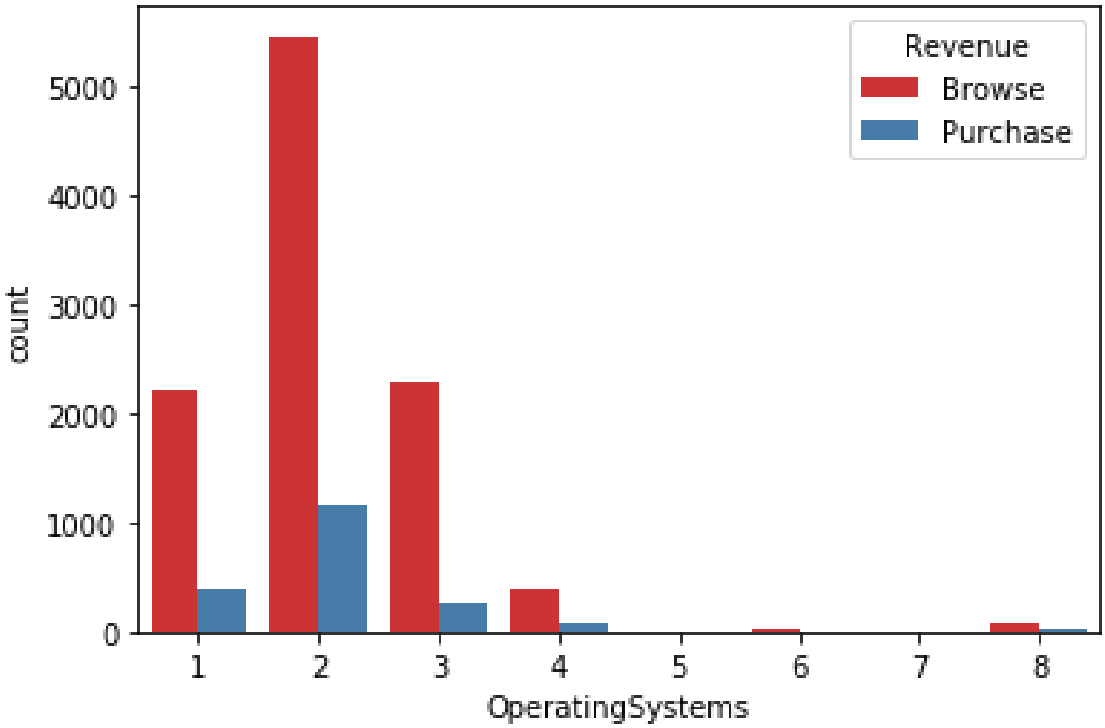
These are some insights we can gain from the descriptive statistics of the two groups of customers. Below we will examine the difference between customer groups for categorical dependent variables.



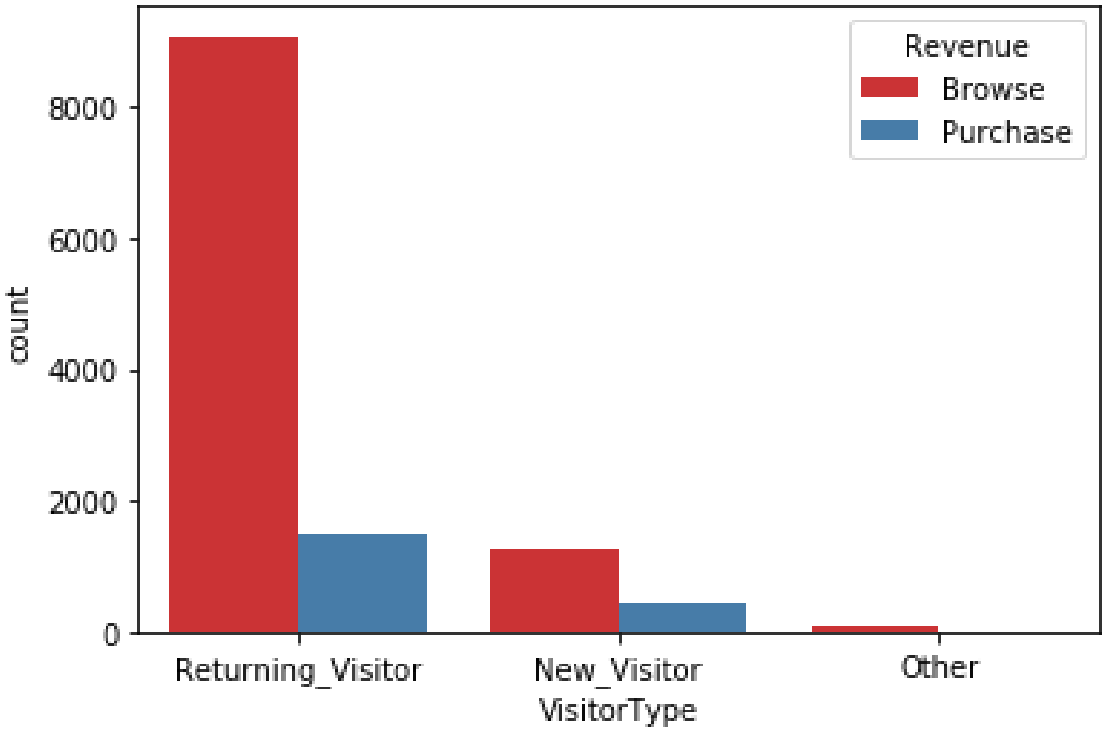
From above it seems that the ratio between browsing and and purchasing customers is lower on the weekend vs. week days. Which means that there more browsing customers for every purchasing customer on the weekday when compared to the weekend.



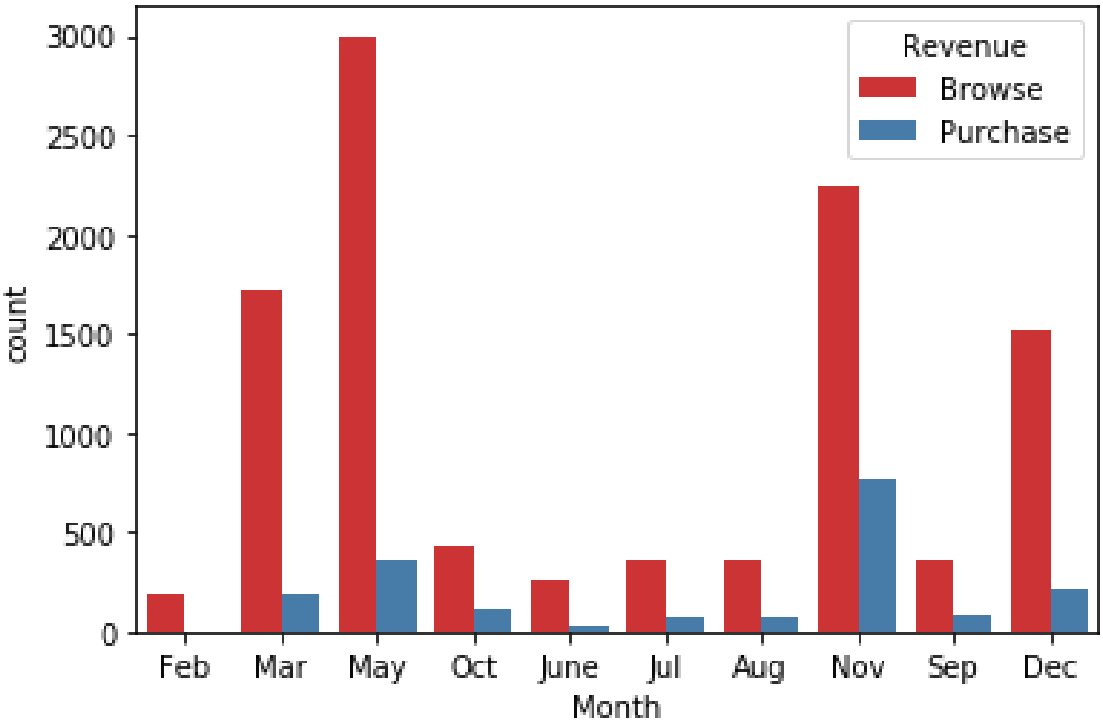
Based on the above plot we can say that Browser 1 2 is probably are the most popular browsers being used the users on this online website. Aside from that there isn't much else we can conclude from the above plot.



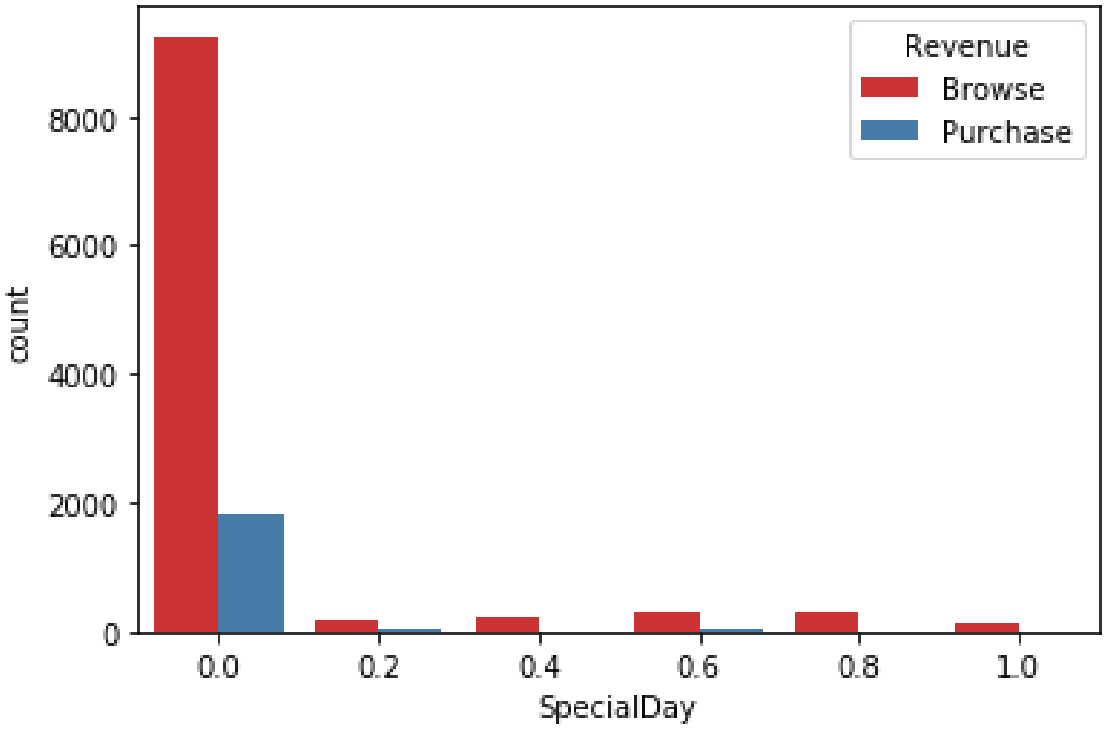
Based on the above plot we can say that Operating Systems 1, 2, 3 and 4 are the most widely used. Based on knowledge we can guess that those 4 operating systems include Windows, Linux, Apple IOS.



Based on the above plot we notice that ratio of Returning Visitors Browse more per Purchase compared New Visitors.



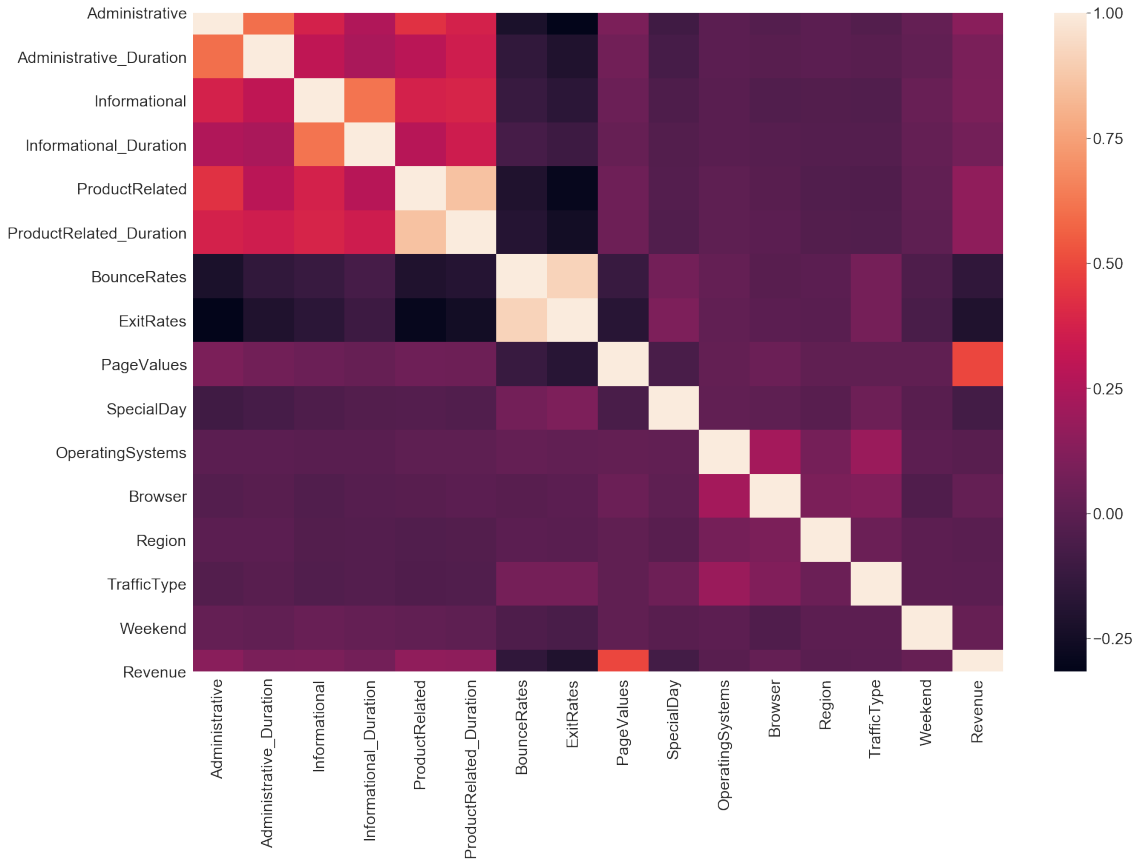
The above plot shows that most purchases and overall site activity happens in the months of November and December which is around Christmas time. Also there is an increase in activity in the month of May which is during the time of Mother's Day. These "Special Days" can be attributed to this increased activity and purchase which is shown in the below histogram.

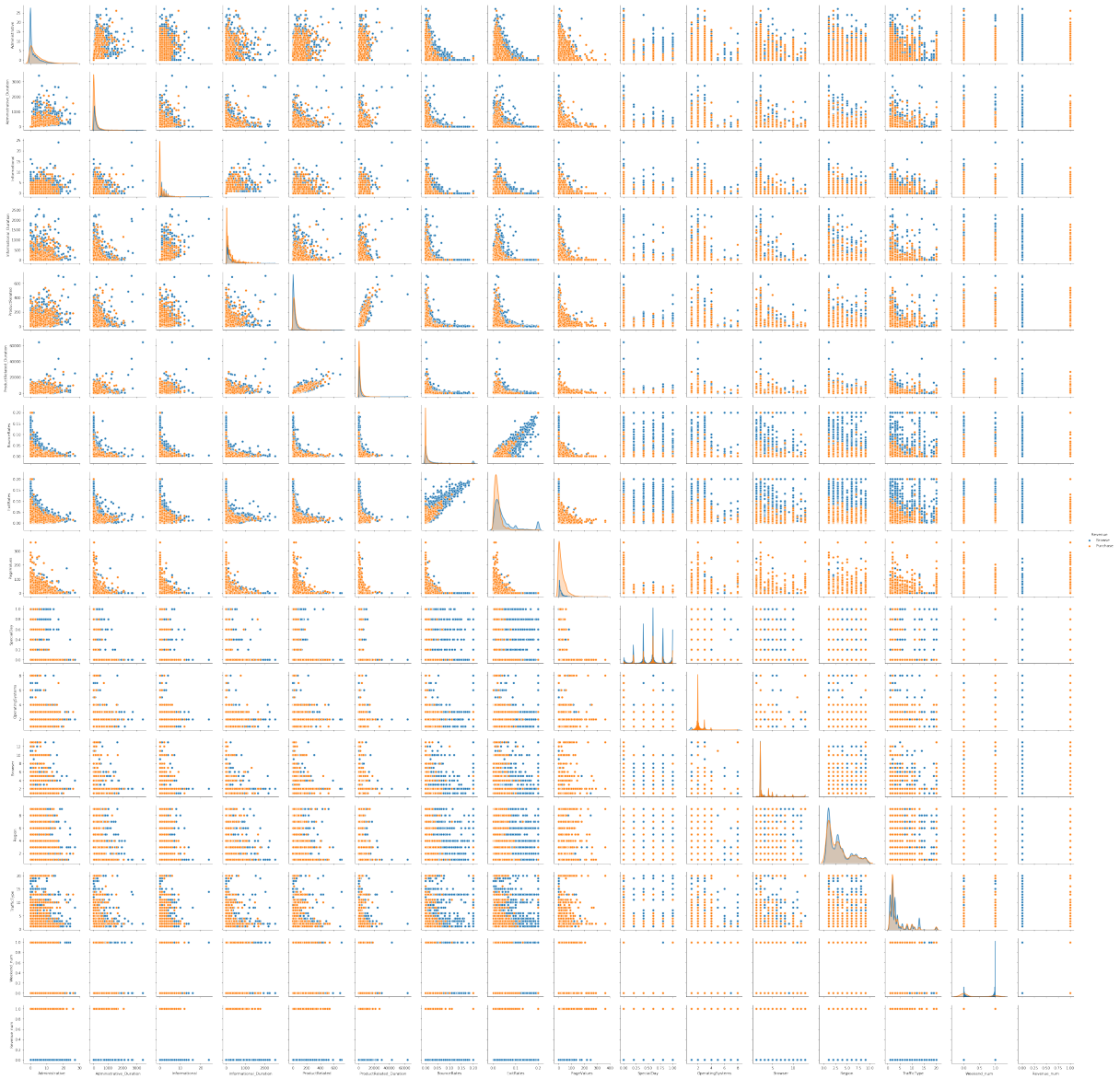


More analysis and work have been done to discover the relationship between (Administrative, Administrative Duration), (Informational, Informational Duration), (ProductRelated, ProductRelated Duration) and Revenue using scatter plots. However these graphs did not give us any information to make any insight and therefore was not included in the report but can be viewed in the Jupyter Notebook Located in the Appendix.

2.2.2 Correlation

Below we will look at both a pair plot which plots all the variable against each other and a heatmap to analyze the correlation between variables.


Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
OperatingSystems
Browser
Region
TrafficType
Weekend
Revenue
1.00
0.75
0.50
0.25
0.00
−0.25



Based on the above plots we can see that the page value rate is the most strongly correlated variable to the a customer purchasing on the website.

2.2.3 Hypothesis Testing

Performing a hypothesis check is a way to verify information about any parameter. In this method, we use two different hypotheses i.e. H0, referred to as the null

hypothesis and alternate hypothesis which is the opposite as our null hypothesis. Our main objective is to verify whether our null hypothesis is true or false based on the data. We try to change the null hypothesis in such a way that the resulting null hypothesis is true. First we will run a hypothesis test on two of the assumed categorical variables which are believed to have significant difference between the Browsing and Purchasing data. The three hypotheses we will test with a chi-square test is: 1) a Special Day is independent of a user making a Purchase, 2) a Returning User is independent of the user making a purchase and 3) Weekends are independent of a user making a Purchase. Below are the results:

1) Chi-square hypothesis results: Degrees of Freedom = 1 , p-value = 0,
test statistic = 10, critical value = 3

Weekend	Browse	Purchase
Weekend	8053	1409
Weekday	2369	499

2) Chi-square hypothesis results: Degrees of Freedom = 1 , p-value = 0,
test statistic = 132, critical value = 3

Visitor Type	Browse	Purchase
Other	1341	438
Returning Visitor	9081	1470

3) Chi-square hypothesis results: Degrees of Freedom = 1 , p-value = 0,
test statistic = 91, critical value = 3

Special Day	Browse	Purchase
Regular Day	1174	77
Special Day	9248	1831

Since all of the hypothesis tests show a significant difference we reject all three null hypotheses. Now we will test the hypotheses on the Informational Duration means

for the two groups of data (Purchase, Browse) are equal. Since the size of the Browse data is much larger and we need the samples to have the same size we will have to pull a random sample of 1908 sessions (size of Purchase data) from the Purchase data. In order to minimize sample bias we will re-run the test on 5 different Purchase data samples. Below are the results:

1) t-test on informational duration for browsers sample 1 results:

Degrees of Freedom = 12329 , p-value = 0, test statistic = -4, critical value = 1

2) t-test on informational duration for browsers sample 2 results:

Degrees of Freedom = 1 , p-value = 0, test statistic = -5, critical value = 1

3) t-test on informational duration for browsers sample 3 results:

Degrees of Freedom = 1 , p-value = 0, test statistic = -5, critical value = 1

4) t-test on informational duration for browsers sample 4 results:

Degrees of Freedom = 1 , p-value = 0, test statistic = -5, critical value = 1

5) t-test on informational duration for browsers sample 5 results:

Degrees of Freedom = 1 , p-value = 0, test statistic = -5, critical value = 1

Since all of the t-test show there is a significant difference between the Informational Duration of the Purchasing Customer group and Informational Duration of the Browsing Customer group we reject the null hypothesis.

2.3 Baseline Modeling

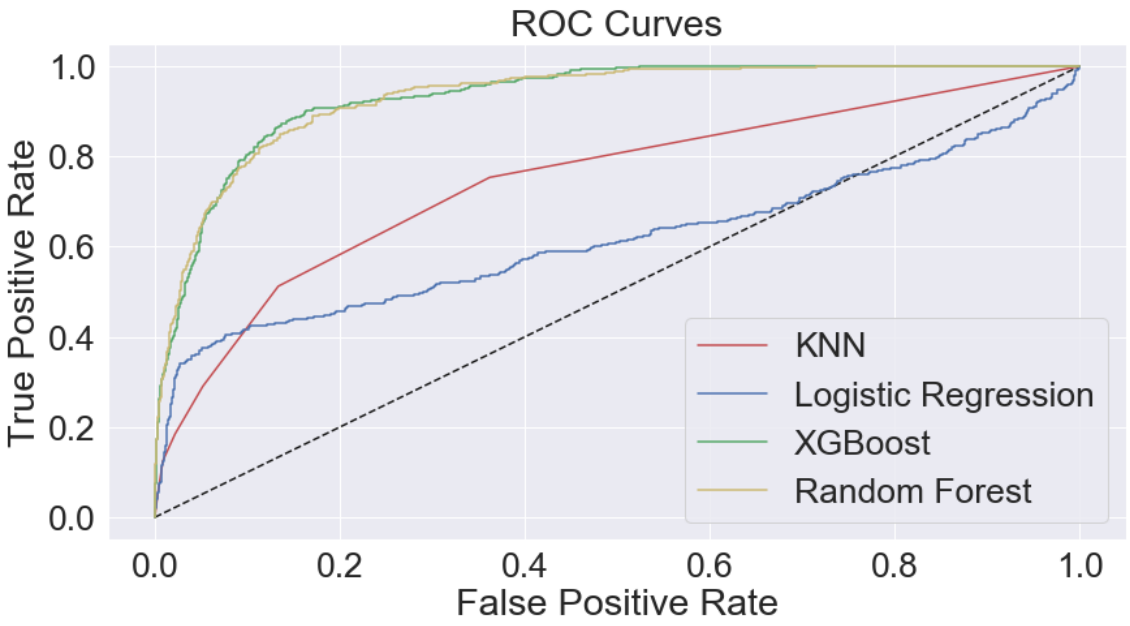
I have chosen to choose 4 different types of prediction models and test which model will perform the best. The k nearest neighbour and logistic regression problems are the simplest methods as the baseline. I would like to show the difference between the model prediction accuracy for those two models compared to the more intensive ML algorithms with the random forest and xgboost models in the extensive modeling section. The training and test data split was 80-20. Using 5-fold cross-validation the KNeighbors Accuracy = 0.86 (+/- 0.03) and the Logistic Regression Accuracy = 0.88 (+/- 0.01).

2.4 Parameter Tuning

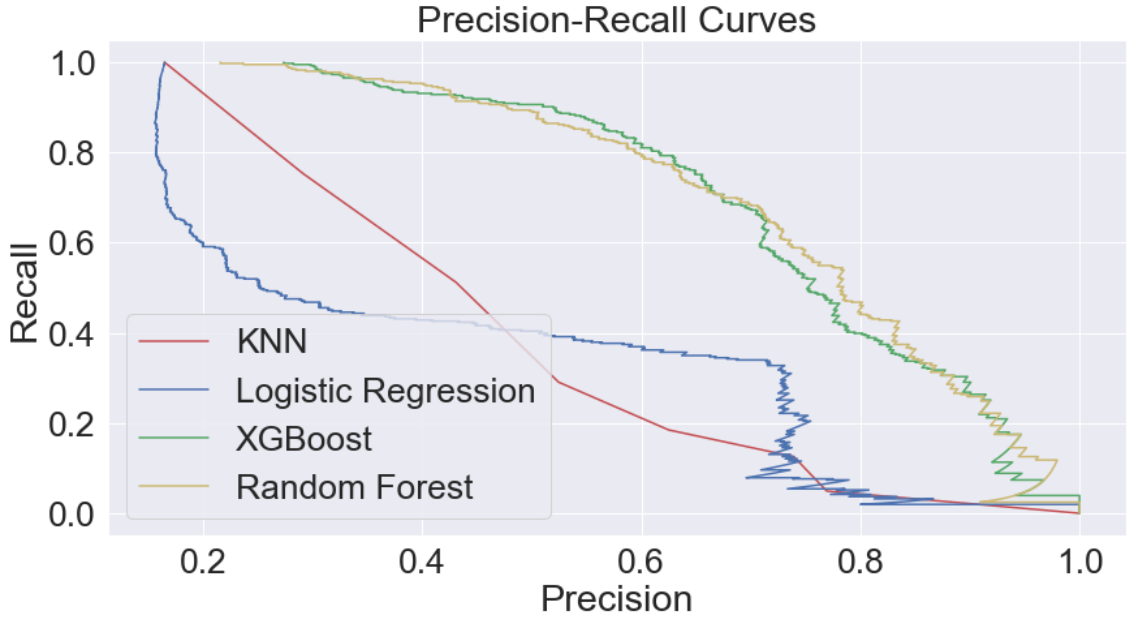
Hyper-parameters optimization is a way to obtain model values that we set before training any model using a machine learning method. Hyper-parameters are described as model characteristics that cannot be estimated from the data, unlike model parameters. In order to better have confidence in our model output, we need to be sure that we use the best possible model at our disposal with high model characteristics. Hyper-parameters can, however, be used to estimate the value of model parameters. Since there is no methodology to estimate the hyper-parameters, we are left with no choice but to use trial error and come up with a value that has better model characteristics as opposed to choosing some other value. In the below code, I've used the grid search algorithm in python to estimate my various model parameters such as nodes of a tree, learning rate, etc. which uses all the possible combinations available in the range of my input to come up with the optimal hyper-parameter values with best model accuracy.

2.5 Extended Modeling

Using 5-fold cross-validation the XGBoost Accuracy = 0.90 (+/- 0.03) and the Random Forest Accuracy = 0.89 (+/- 0.03). Even though there are much more Browsing sessions compared to Purchasing sessions our models did not have an issue over performing for the purchase data. The ROC and Precision-Recall curves on the tuned models will further illustrate the performance differences between these models.



Model Performance can be examined by the ROC curve which plots the True Positive Rate against the False Positive Rate. The ROC curve of the best model will be high fitted closest to the top left corner of the ROC plot. As we can see the XGBoost and Random Forrest Models perform the best when predicting the customers activity on the online website.



The Precision-Recall Curve is also another method of showing model accuracy. It plots the recall (number of true positives over the sum of true positives and false negatives) precision (number of true positives over the sum of true positives and false positives). The most accurate model will have the line located as far to the top right corner of the plot. From the above plot we can see the XGBoost model and the Random Forrest model are very similar and strong in accuracy of prediction the purchasing customer.

2.6 Feature Influence Analysis

Coefficients for the Logistic Regression:

Variable	Coefficients
Administrative Duration	-0.001481
Product Related	-0.005028
Page Values	0.037146
Month	-0.0690007
Region	-0.031102
Traffic Type	-0.071842

** All other variables have coefficient values of zero*

Looking at variable coefficients it seems that none of the variables have a strong influence on the classification of a session being a purchase. However all coefficients shown above expect for page values contribute to the session being classified as browse. The page value attribute contributes to the session concluding with a purchase, the higher the page values in the session the more likely a purchase will be made.

Feature Importance Table for XGBoost:

Feature	Importance
Administrative	0.04255423
Administrative Duration	0.026304374
Browser	0.009115764
Bounce Rates	0.065999046
Exit Rates	0.045520242
Informational	0.01348208
Informational Duration	0.01609661
Product Related	0.041459087
Product Related Duration	0.04068779
Page Values	0.5072776
Month	0.078832574
Operating Systems	0.010195555
Region	0.011102518
Special Day	0.024347754
Traffic Type	0.016063038
Visitor Type	0.041039262
Weekend	0.009922537

Based on this table the most influential variable would be the Page Values of the session. Below we will examine the random forest model to ensure the strength of the page values variable.

Feature Importance Table for Random Forest:

Feature	Importance
Administrative	0.03991472916277499
Administrative Duration	0.05326477626464433
Browser	0.015186731436891426
Bounce Rates	0.05503031342519937
Exit Rates	0.08472902192739108
Informational	0.014253705557205114
Informational Duration	0.023699276751093012
Product Related	0.06511307261725152
Product Related Duration	0.08185991646079405
Page Values	0.44407117877760555
Month	0.04194614441640582
Operating Systems	0.01315369412357316
Region	0.02250333618743712
Special Day	0.0031824215381647033
Traffic Type	0.023561369487226275
Visitor Type	0.011165971098807854

Based on the this we can confirm that the page value variable is the most influenctial variable on the outcome of a purchase session.

3 Summary of Findings

3.1 Summary Table

Model	User Decision	Precision	Recall	F1 Score	Comments
K-Nearest Neighbors	Browse	0.86	0.98	0.91	Optimized parameters: n = 6
	Purchase	0.62	0.18	0.29	
Logistic Regression	Browse	0.88	0.98	0.92	Optimized parameters: Penalty = l1
	Purchase	0.73	0.3	0.43	C = 0.00021544346900318823
XGBoost	Browse	0.92	0.96	0.94	Optimized parameters: Learning rate = 0.05, Max Depth = 3, Min Child Weight = 1,
	Purchase	0.73	0.56	0.63	gamma = 0.1 , Colsample by tree = 0.6, n-estimators = 300
Random Forest	Browse	0.92	0.96	0.94	Optimized parameters: n-estimators = 100, Max Depth = 20
	Purchase	0.75	0.56	0.64	Min Sample Split = 12, Min Sample Leaf = 3

4 Conclusions and Future Work

4.1 Conclusion

Based on our hypothesis tests we can conclude that the type of day whether it is a weekend or special day as an impact on the users decision to make a purchase. Also, a returning user is not mutually exclusive on the decision to make a purchase. Tying our hypothesis tests with our descriptive analysis means that there is more site activity (sessions) for returning users, and also on special days and weekends. Based on our model prediction analysis we can conclude that the XGBoost and random forest models are the best in predicting whether a session will result in a purchase.

4.2 Future Work

Future work could include how to encourage users to make a purchase on Special Days and Weekends as those are the times when site activity is at its highest. This would allow for maximization of sales on these days. Also, a new or improved product recommendation system can be implemented in order to encourage sessions to have more page values and lower exit rates.

5 Recommendation for the Client

My recommendation to the client would be to use an XGBoost model when looking to predict purchases by session as this model performs the best. If computing power or time restraints are present and there is a need for quicker parameter tuning the random forest model can be implemented without losing much accuracy. If the client is looking to maximize sales I would recommend they explore ways to optimize sales on Special Days and weekend. It would also be a good idea to encourage users to explore more products on the site by implementing a product recommendation system. This would increase the page values per session which has a positive impact on sessions resulting in purchases.