Springboard Capstone Project 1: Analysis of Online Shopping Dataset

Kaggle Dataset Repository

Final Report by: Michael Thabane

March 18, 2020

Contents

1 Introduction

1.1 The Business Problem

With the development of technology and the rapid increase of online shopping it is important for fashion and clothing companies to estimate the customer's intentions of making 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 estimate 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 can be used to derive the likelihood of a customer making a purchase, or not. This business can be modeled as is a supervised classification data science problem in which we can use multiple machine learning algorithms to estimate the probability that a customer will 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 on website session data. Various exploratory methods were used to glean intuitive understanding of the dataset. The methods involved with machine learning classification models include k-nearest neighbors, logistic regression, random forests 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 analysis of the performance models built that the XGBoost model is the best classifier for a user making a purchase. The Page Values, Bounce and Exit Rates have been found to be the most important parameters in model classification by all models. Based on our findings in the report our recommendation for the website owner is 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 is Boolean but in the report we have converted it into a 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, "Weekend" which is 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 Jupyter Notebook https://github.com/thabanenm/Springboard/blob/master/Capstone%201%20Final%20Project%20-%20Online%20Shopping.ipynb). 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.

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.

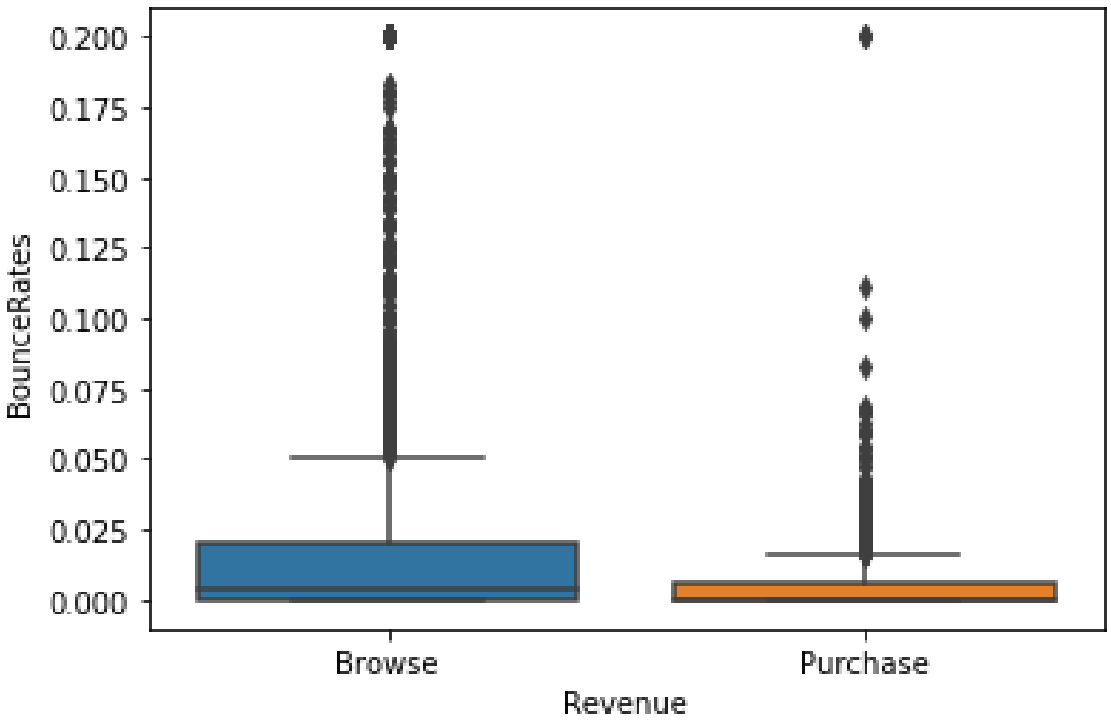


Figure 1: Boxplot of Bounce Rates

Also, Browsing Customers visit less pages on average than Purchasing Customers illustrated in the boxplot below.

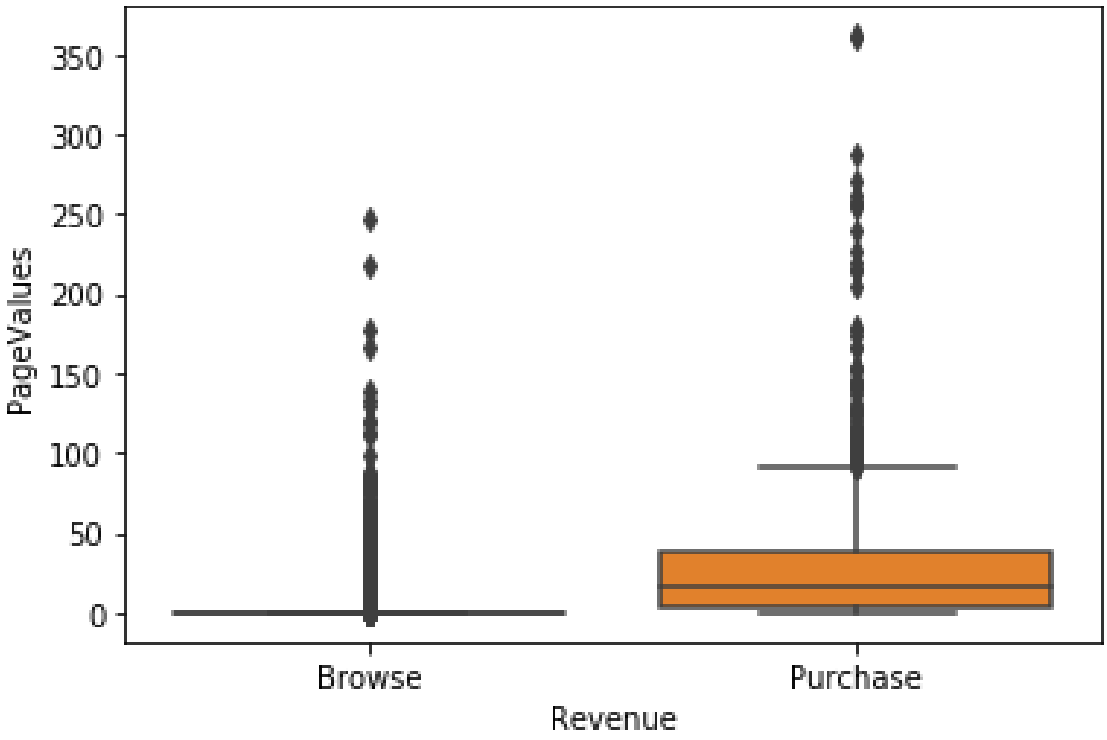


Figure 2: Boxplot of Page Values

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.

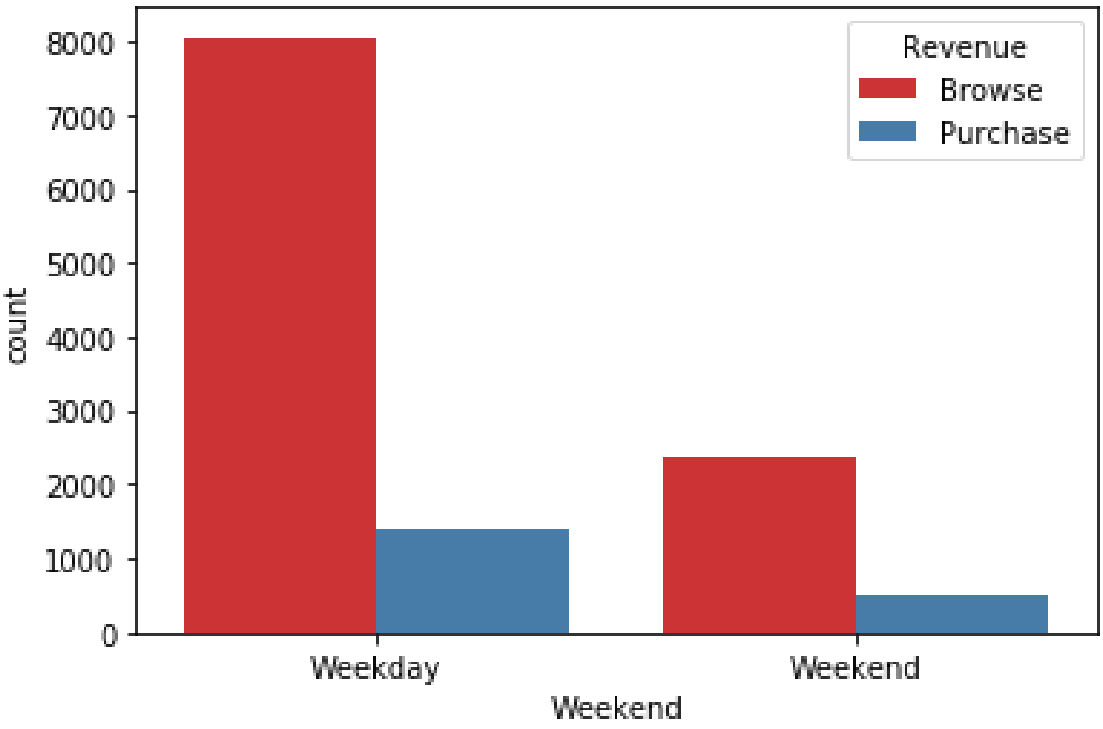


Figure 3: Histogram of Weekend

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.

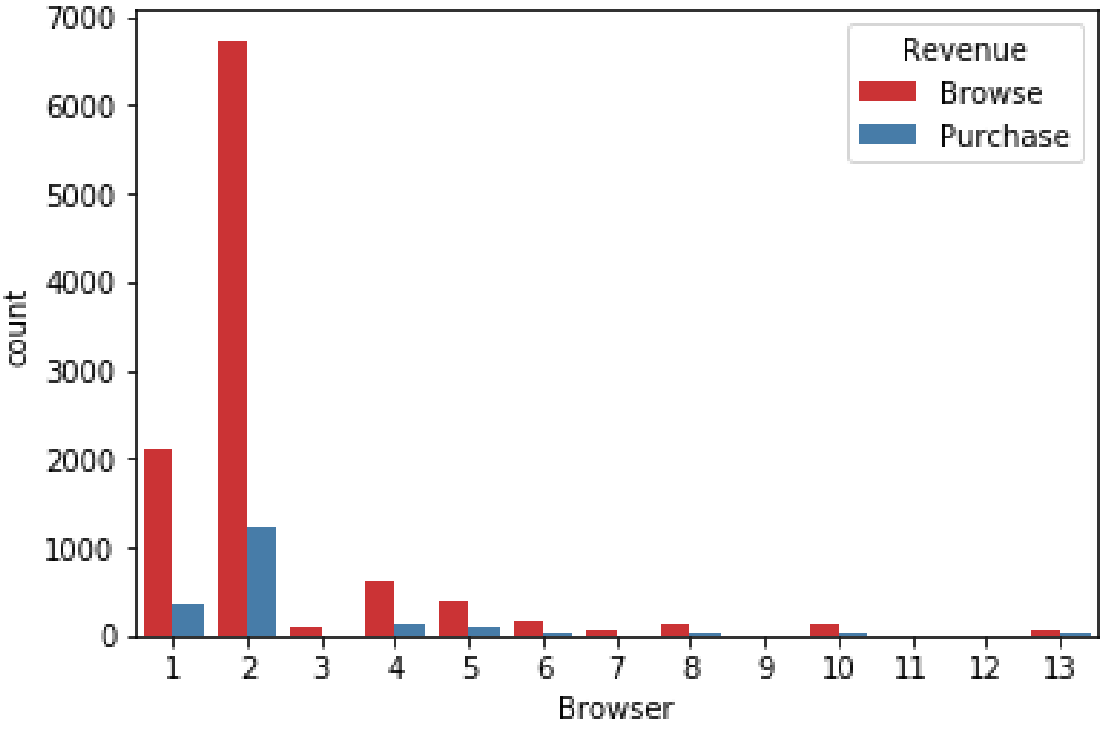


Figure 4: Histogram of Browser

Based on the above plot we can say that Browser 1 and 2 is probably are the most popular browsers being used the users on this online website.

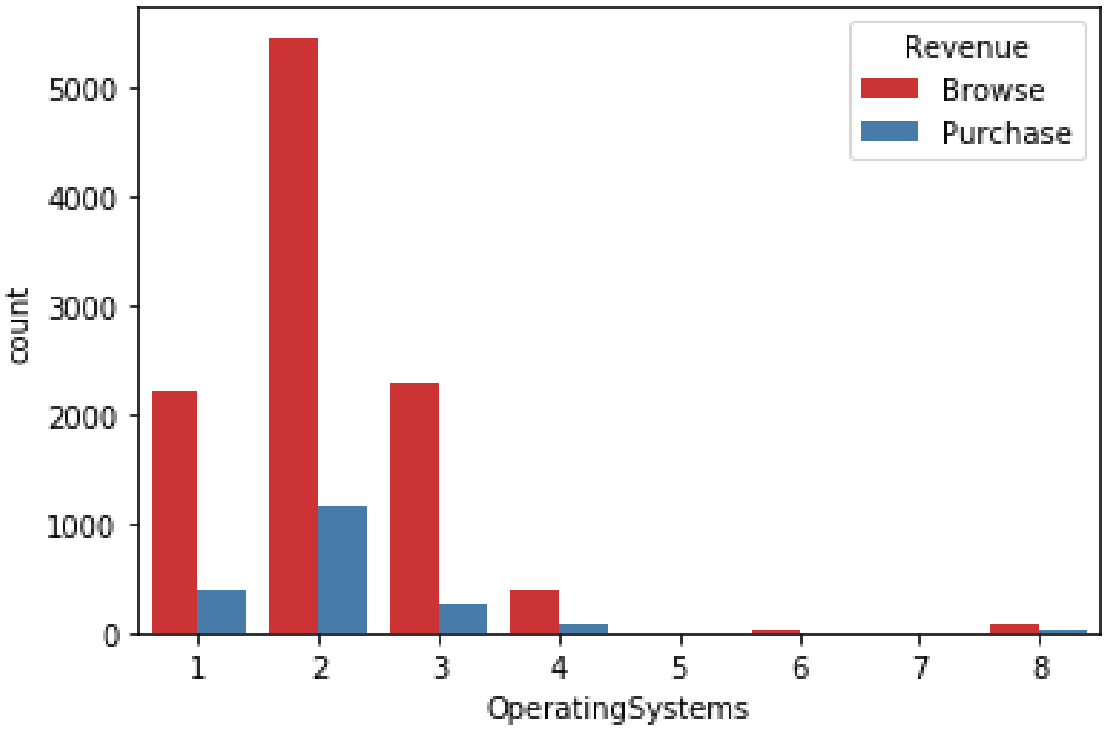


Figure 5: Histogram of Operating Systems

Based on the above plot we can say that Operating Systems 1, 2, 3 and 4 are the most widely used.

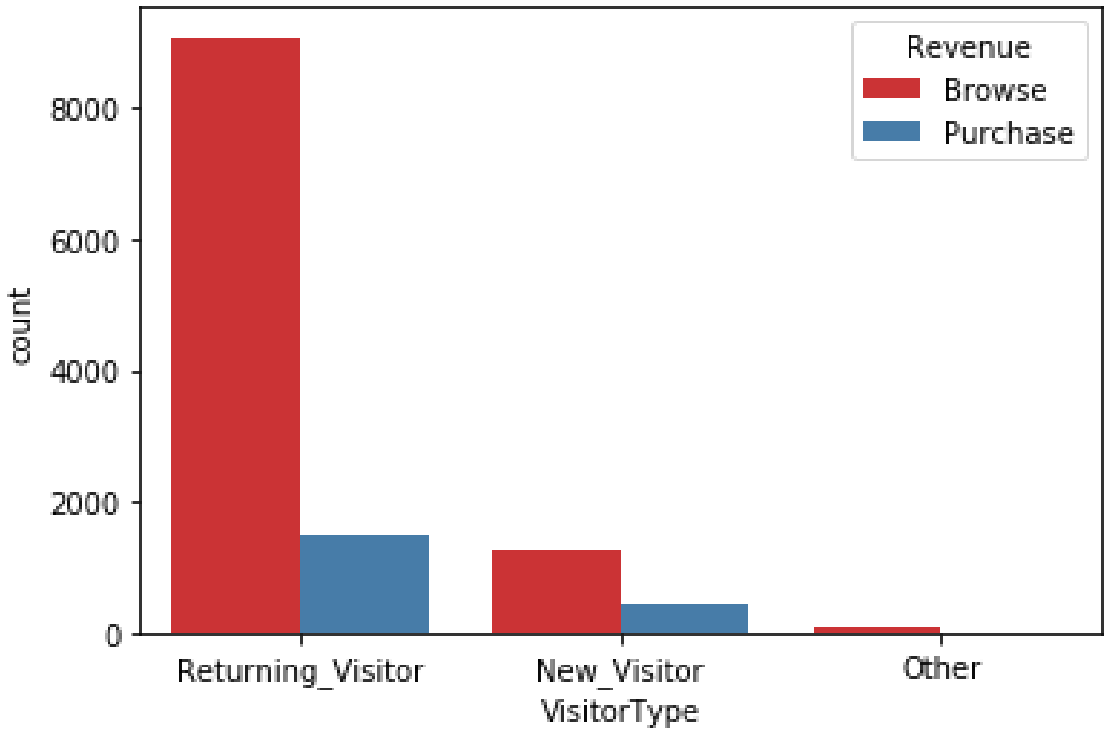


Figure 6: Histogram of Visitor Type

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

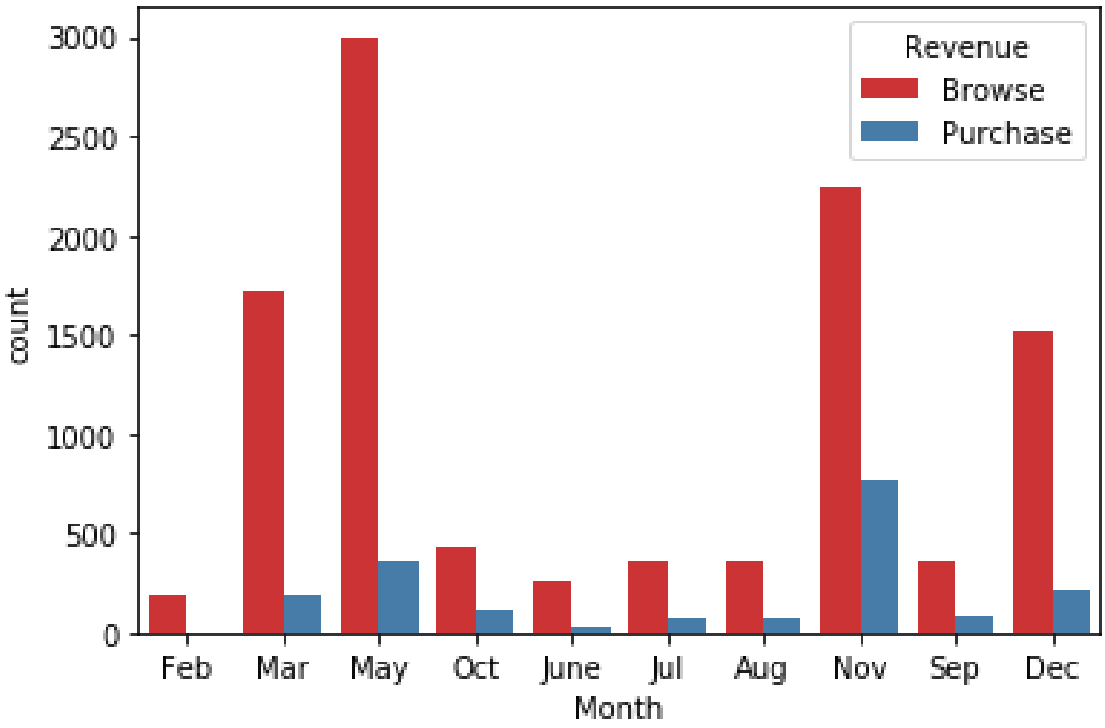


Figure 7: Histogram of Month

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.

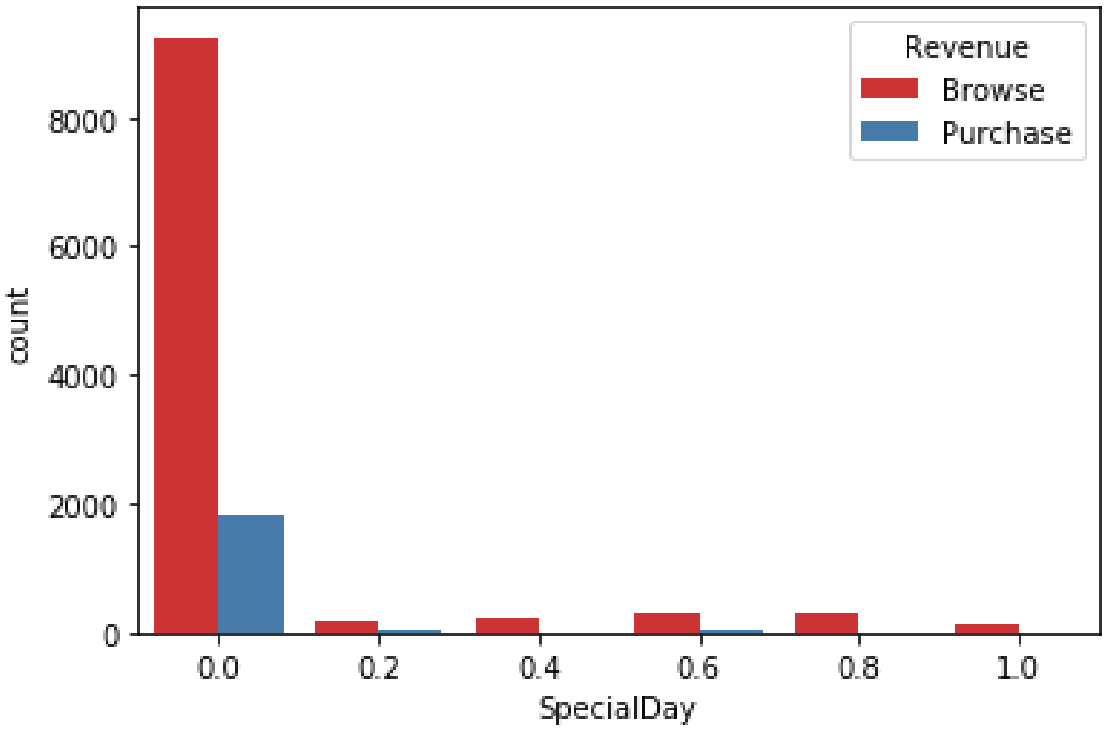
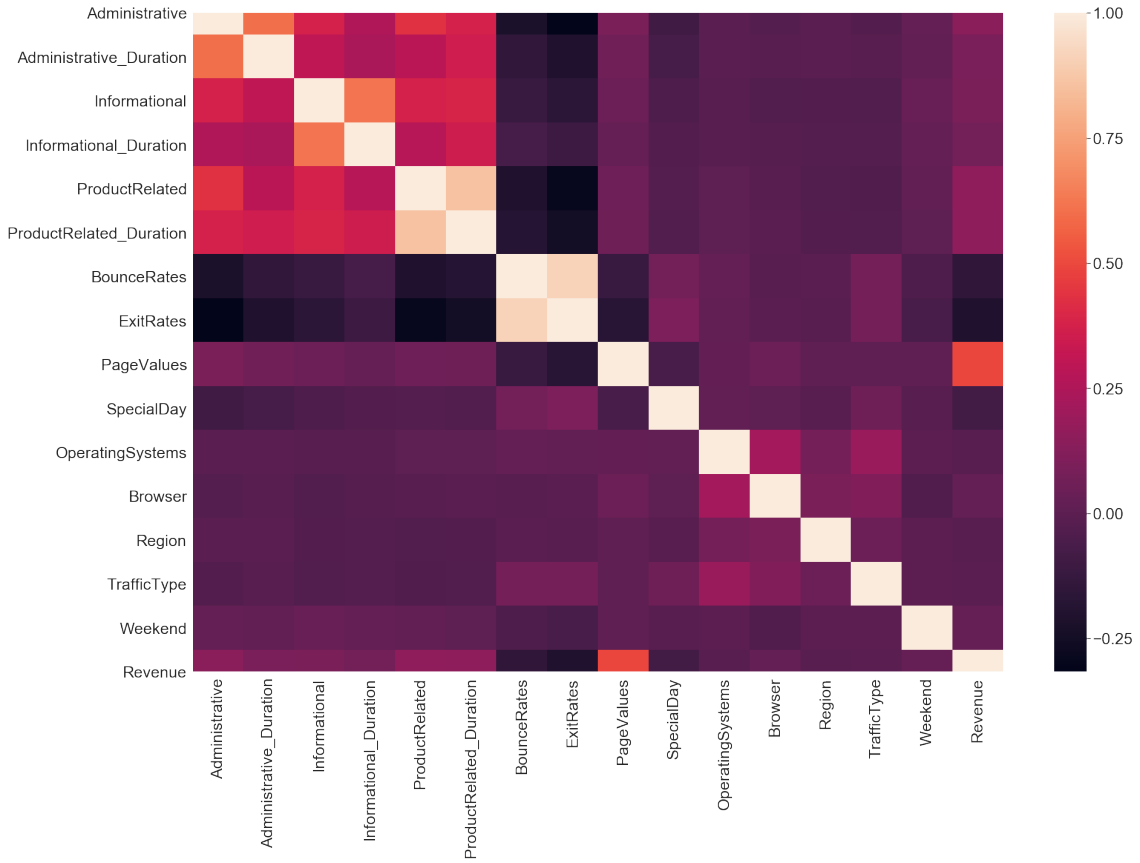
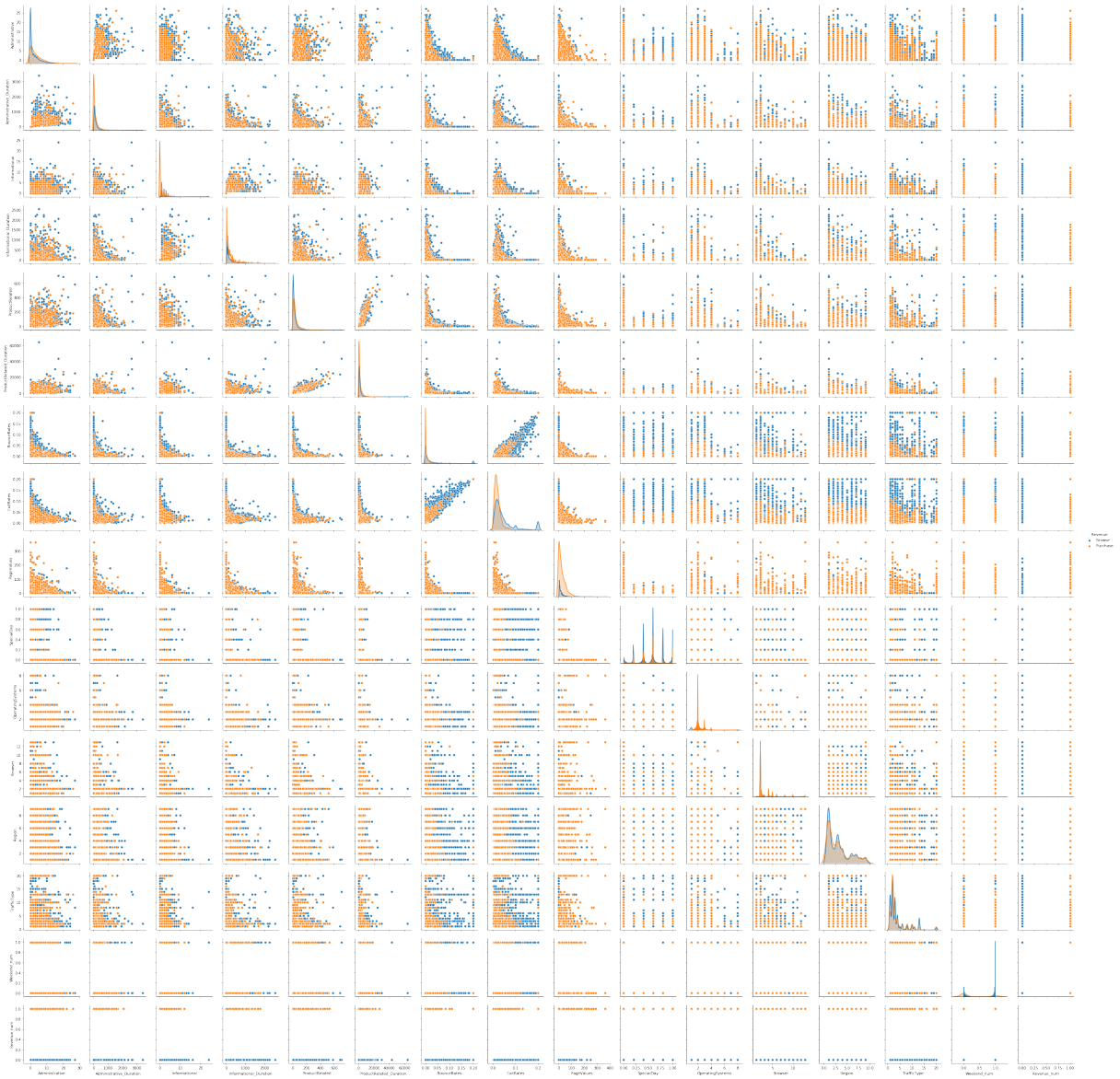


Figure 8: Histogram of Special Day

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.


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

In hypothesis testing, we use two different hypotheses i.e. H0, referred to as the null hypothesis, and Ha, alternate hypothesis, which is the opposite as our null

hypothesis. The main objective is to estimate the probability of finding values of the statistic under scrutiny that are at least as large as the observed one in the sample, provided that H0 is assumed to hold. 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 are: 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 p-values obtained are below the threshold we used (alpha= 0.05), we can reject the Null hypothesis, and conclude that the differences observed are statistically significant. 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 = 12329 , p-value = 0, test statistic = -5, critical value = 1

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

Degrees of Freedom = 12329 , p-value = 0, test statistic = -5, critical value = 1

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

Degrees of Freedom = 12329 , p-value = 0, test statistic = -5, critical value = 1

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

Degrees of Freedom = 12329 , p-value = 0, test statistic = -5, critical value = 1

Since all p-values obtained are below the threshold we used (alpha= 0.05), we can reject the Null hypothesis and conclude that the differences 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

Tried 4 different machine learning classification algorithms, analyzed their collective performances. The k-nearest neighbours and logistic regression algorithms are the simplest algorithms, and were used as the baseline. Will show the difference between the model prediction accuracy for those two models compared to the more advanced machine learning algorithms 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

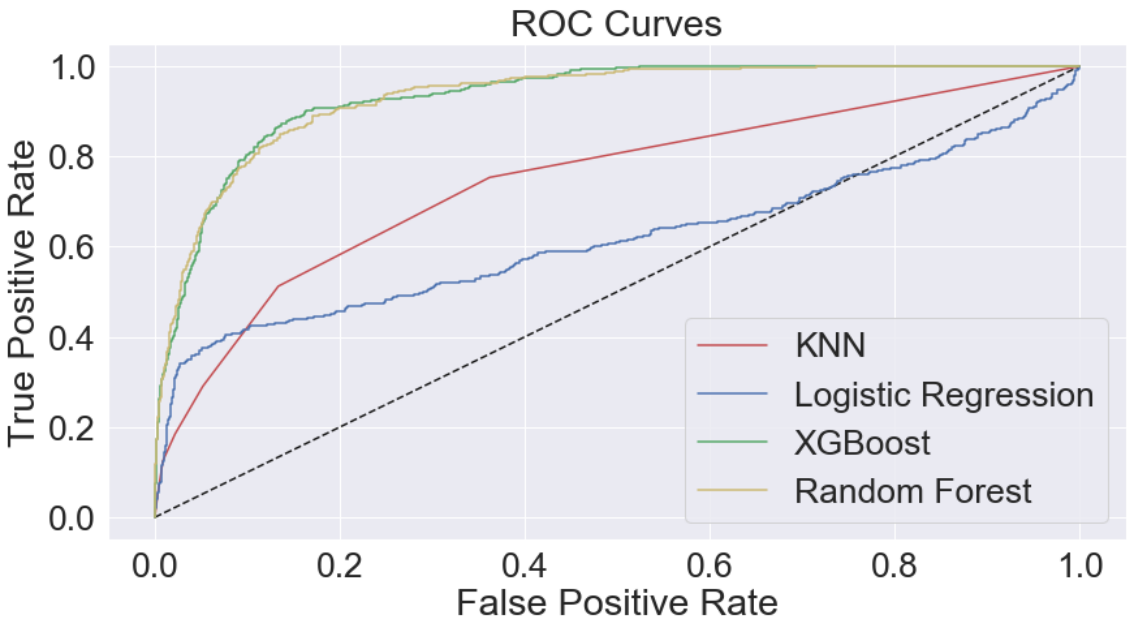
Accuracy = 0.88 (+/- 0.01).

2.4 Parameter Tuning

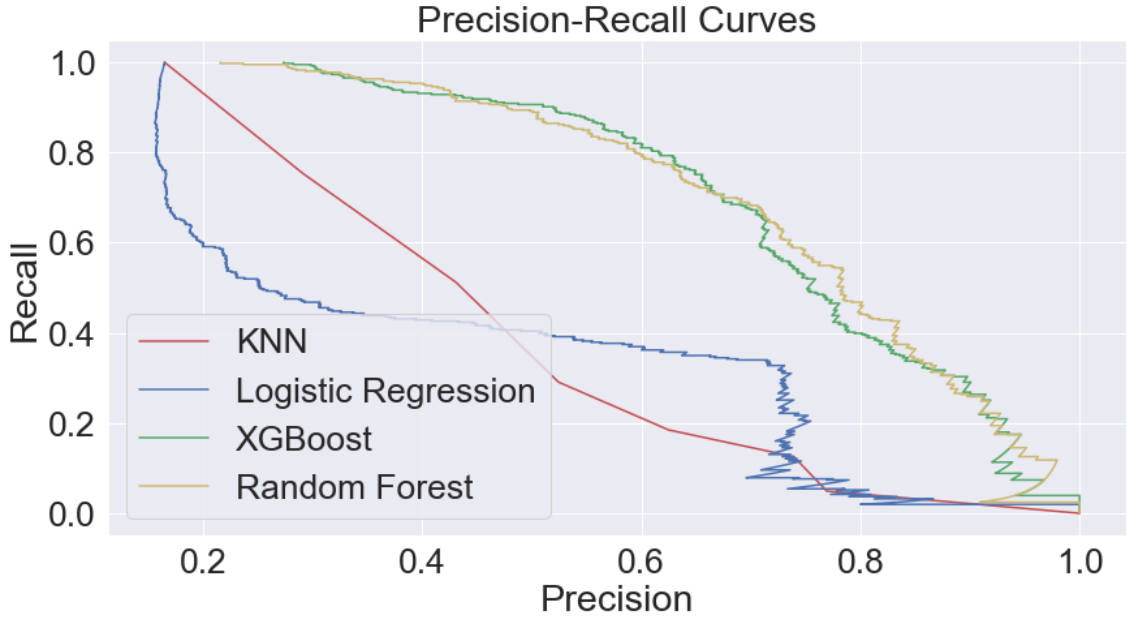
Hyper-parameter optimization is a way to determine the values of parameters associated with models that yield the best performance according to a given validation set, or according to cross-validation. Hyper-parameters are described as model characteristics that cannot be estimated from the data, unlike 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 the highest performing hyper-parameters. However, hyper-parameters are used to estimate the value of model parameters. In the below code, I've used the grid search algorithm in python to estimate various model parameters such as "nodes" of a tree, "learning rate", etc. which uses all the possible combinations available in the range of the input to come up with the optimal hyper-parameter values for the space of parameters explored.

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 the two classes under consideration differ in size by a factor of 5.5, we did not observe a bias in performance of these algorithms towards the larger class (Browse). 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), and precision (number of true positives over the sum of true positives and false positives). The most accurate model will have the curve leaning towards 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 show a strong performance.

2.6 Feature Influence Analysis

We can now use the models to identify which variables have the most significant impact on the user making a purchase. For the Linear Regression model we can examine this by looking at the values of the coefficients. The closer the absolute value of the coefficient is to one the more impactful that variable is in influencing our outcome. The positive coefficient means the variable positively influences the outcome which in our case would make a positive impact on a user making a purchase. The opposite applies for a negative coefficient (i.e negative impact).

Coefficients for the Logistic Regression:

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

All other variables have coefficient values of zero

Looking at the coefficients it seems that none of the variables have a strong influence on the classification of a session being a purchase. However all coefficients except for the "Page Values" coefficient are negative meaning they negatively contribute to the session being classified as purchase. The page value attribute positively contributes to the session concluding with a purchase, the higher the page values in the session the more likely a purchase will be made.

For the XGBoost and Random Forrest we must run a Feature Importance analysis to observe the impact of variables. The larger the values more impact those variables make on the outcome. This analysis is below: 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 is Page Values of the session. Below we examine the random forests model to ensure the strength of the page values variable.

Feature Importance Table for Random Forests:

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 influential variable.

3 Summary of Findings

The training and test data split was 80%-20%

3.1 Summary Table

Model	User Decision	Precision	Recall	F1 Score	Support	Comments
K-Nearest Neighbors	Browse	0.86	0.98	0.91	2060	Optimized parameters: n = 6
	Purchase	0.62	0.18	0.29	406	
Logistic Regression	Browse	0.88	0.98	0.92	2060	Optimized parameters: Penalty = l1
	Purchase	0.73	0.3	0.43	406	C = 0.00021544346900318823
XGBoost	Browse	0.92	0.96	0.94	2060	Optimized parameters: Learning rate = 0.05, Max Depth = 3, Min Child Weight = 1, gamma = 0.1 , Colsample by tree = 0.6, n-estimators = 300
	Purchase	0.73	0.56	0.63	406	
Random Forest	Browse	0.92	0.96	0.94	2060	Optimized parameters: n-estimators = 100, Max Depth = 20
	Purchase	0.75	0.56	0.64	406	Min Sample Split = 12, Min Sample Leaf = 3

4 Conclusions and Future Work

4.1 Conclusions

- Hypothesis tests shows that the "Weekend" and "Special Days" makes a statistically significant difference on the outcome of a session ending with a purchase
- Hypothesis tests also show that there is a statistically significant difference between outcome groups for the variable "Informational Duration"
- From the model prediction analysis we can conclude that the XGBoost and Random Forest models are the best performing when predicting if a user session will include a purchase
- From the Feature Influence Analysis we know that "Page Value" has a positive impact on the outcome of a session resulting in a purchase

4.2 Future Work

- A new or improved product recommendation system can be implemented in order to engage users so the average session will have more page values and lower exit rates.

5 Recommendations

- 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
- Based on the report we know that the "Page Values" per session has a positive impact on the session concluding with a purchase
- Recommend implementing a product recommendation system to encourage users to explore more products on the site and increase the "Page Values" per session