Capstone Project MLE

Lakshya Khandelwal
July 15’ 2020

Online Shopper's Conversions Prediction

Project Overview

Everytime a customer browses some product online, the search engine and visited website can store relevant session information. This data is heavily used to predict what kind of sessions will actually convert to purchases (revenue).

This can highly enhance the profit making ability of a brand/ ecommerce platform by targeting such potential sessions / individuals.

Due to improvement in machine learning algorithms as well as data storing and processing capabilities, companies can store huge amounts of data (even at individual customer’s session level). This makes usage of sophisticated machine learning algorithms very effective.

Problem Statement

Using session level user data for online shopping interaction, we try to develop intuition into how various features / attributes lead to click conversion in terms of revenue.

Evaluation Matrix

I plan to use classification accuracy of Rate of correct predictions

(No. of Correct Predictions) / (Total Cases)

Analysis

Data Exploration

The dataset consists of feature vectors belonging to 12,330 sessions.The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

Attribute Information:

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 pageviews to the page, the percentage that were the last in the session.

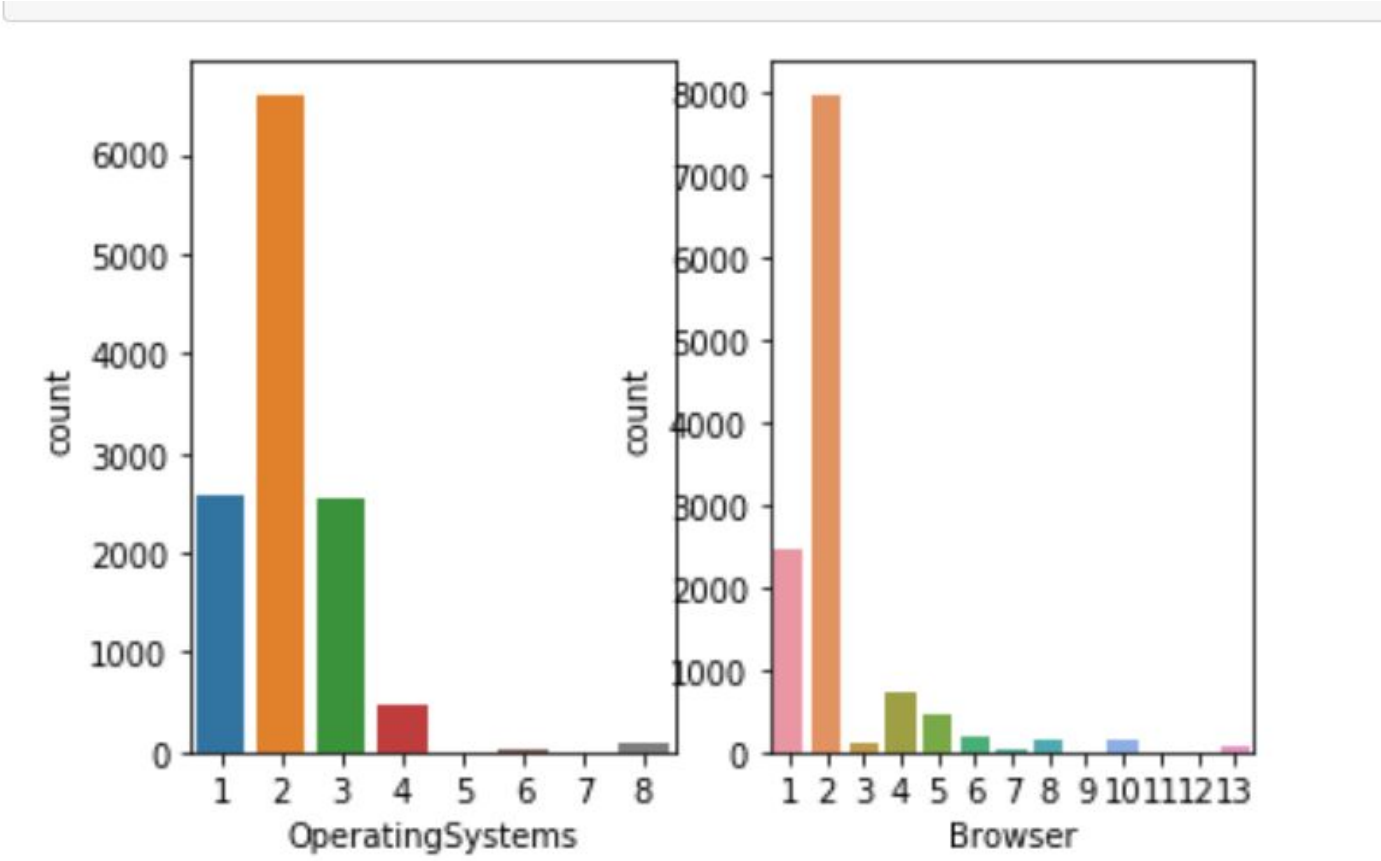
The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

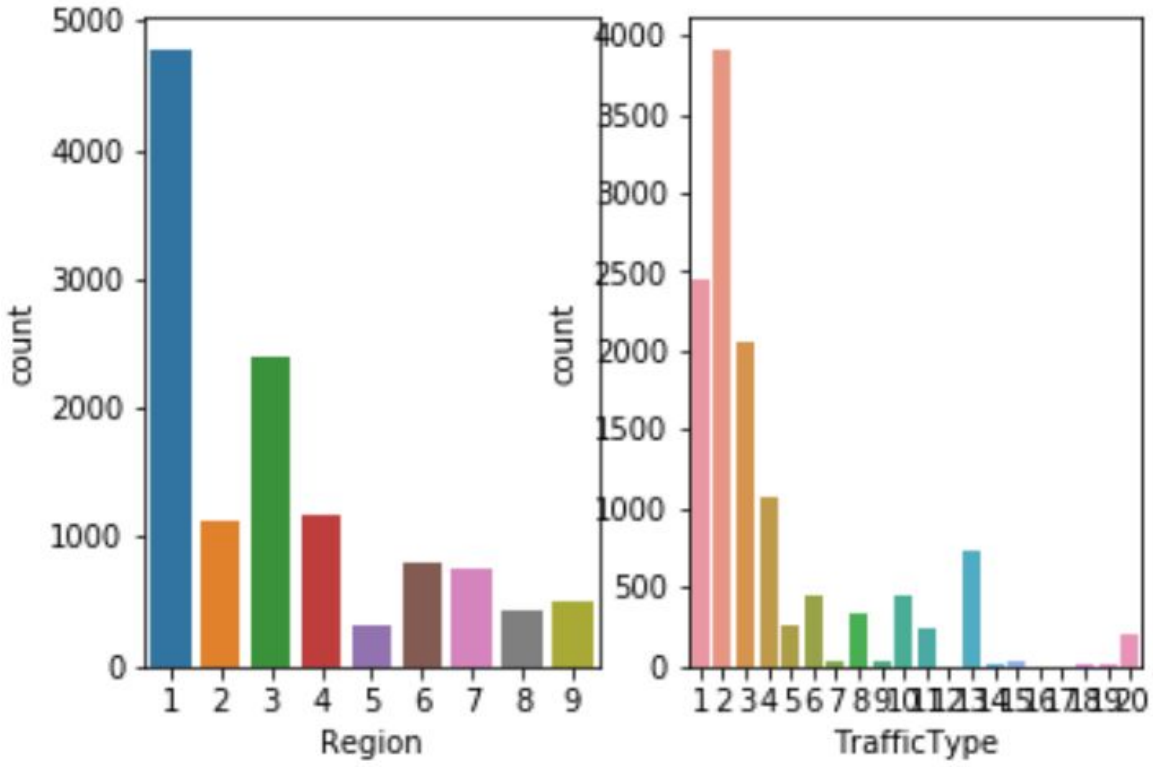
The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’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 dataset 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.

Data Visualization

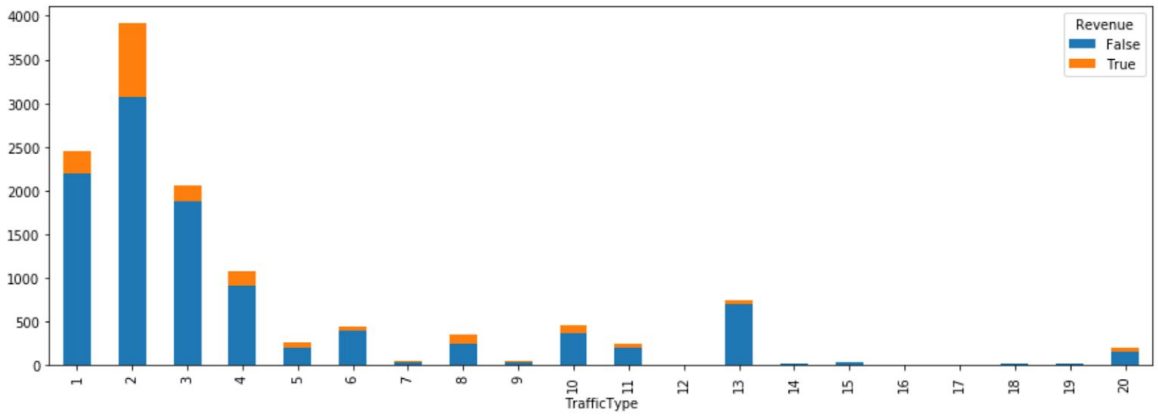
Following are plots on distribution of various features in our dataset

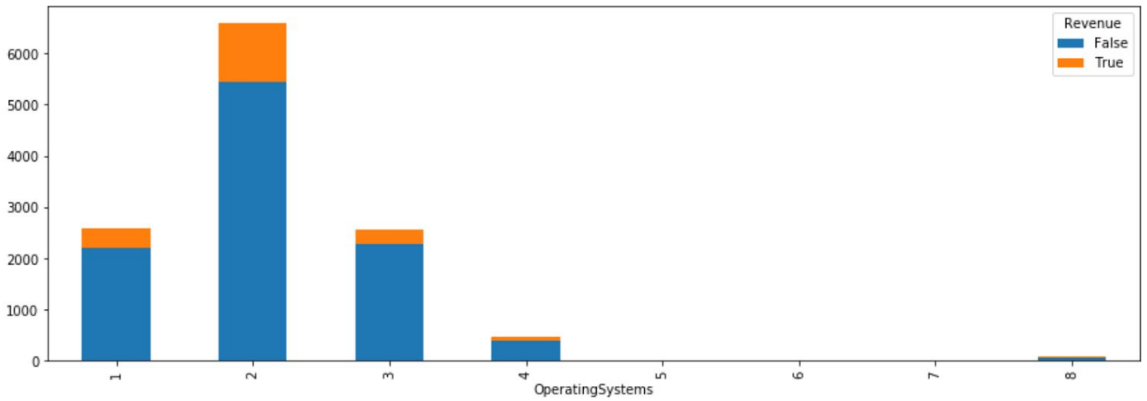


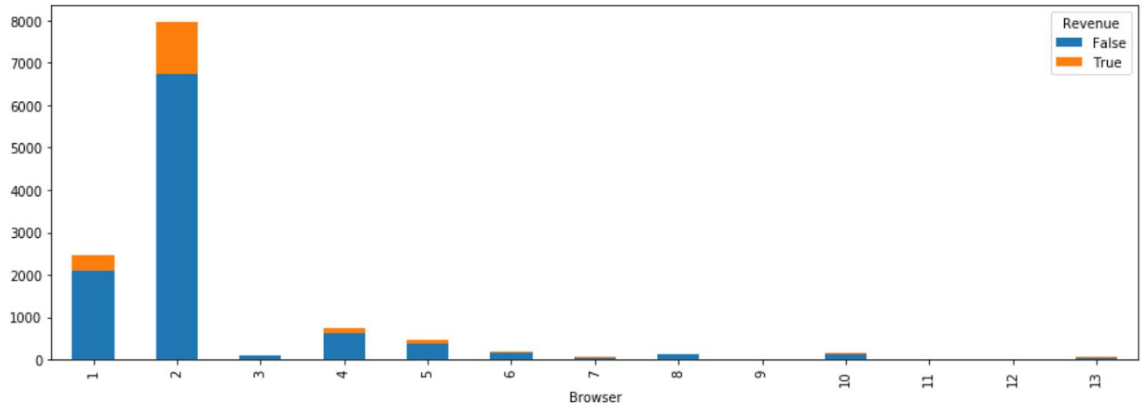


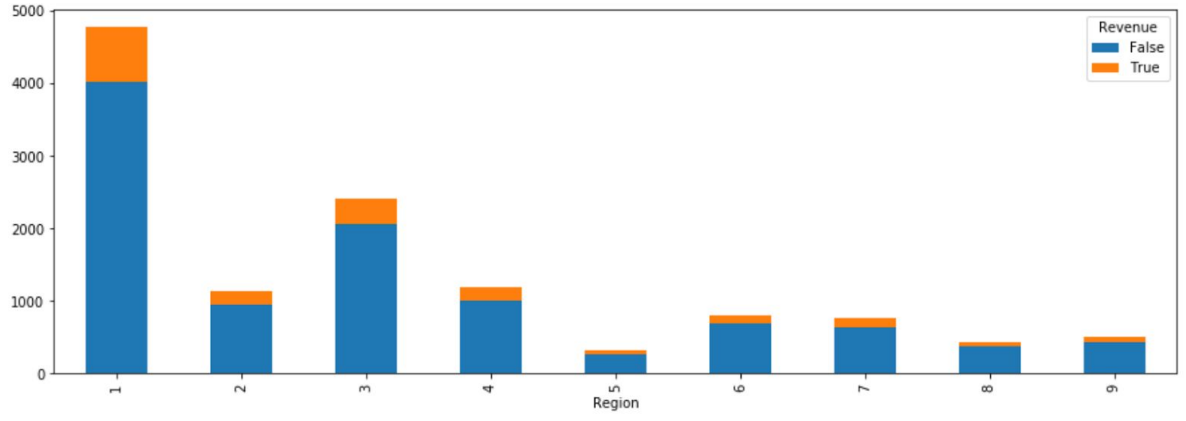
We have very skewed data for Operating systems and Browser.
Also we have a higher number of data coming from region 1.

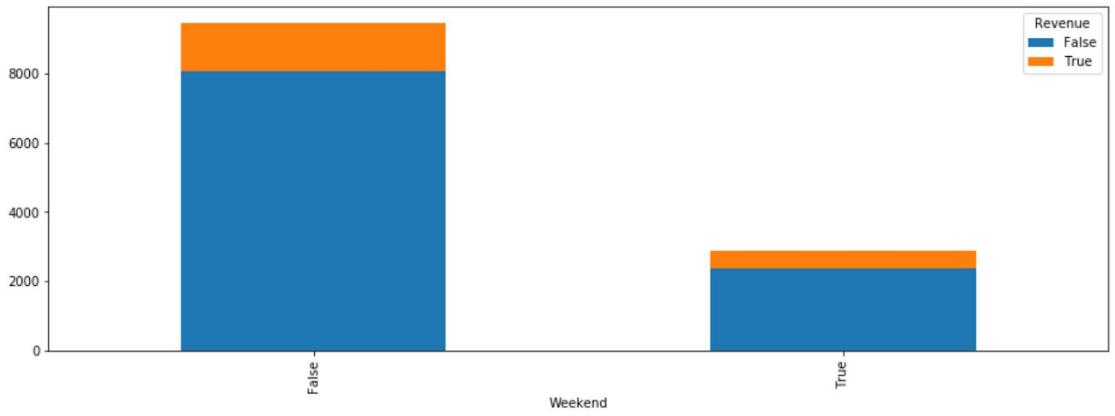
Now let us look at distribution of Revenue with few features




Revenue
False
True
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
OperatingSystems


Revenue
False
True
8000
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Browser


Revenue
False
True
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
Region



We are unable to establish any clear difference in distribution of Revenue with Weekend features. Though browser 2 appears to have slight positive biased distribution with Revenue. Also operating system 2 has higher positive rates but it also has higher overall data. So we will not make strong assumptions from this analysis.

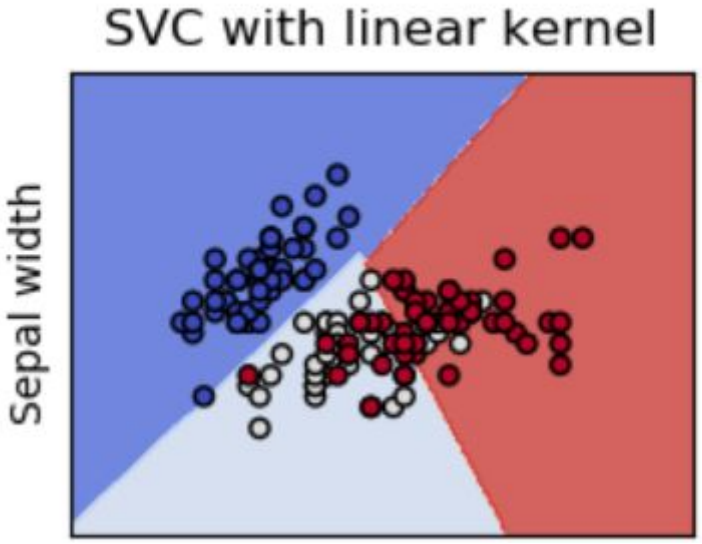
Algorithms and Techniques

Supervised learning approach is taken in order to predict revenue conversion using given features.

I use linear SVC (Support Vector Machine for Regression) to create a base model.

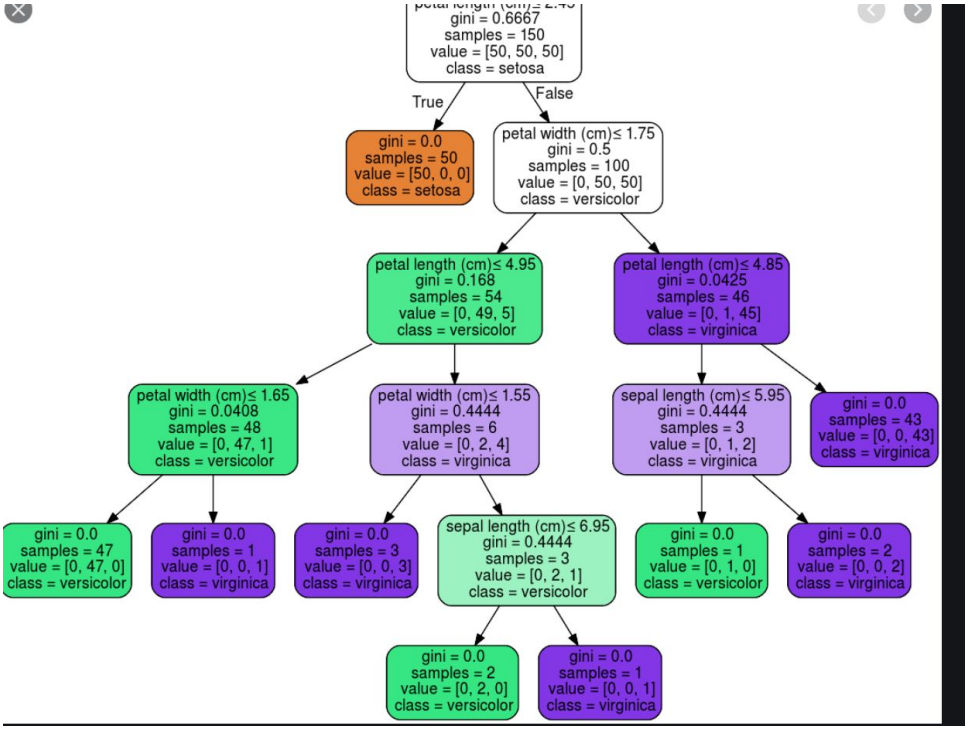
(https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)

This method is helpful in creating a hyperplane in feature dimension in order to classify



* (Image taken from Sklearn webpage)

Next I use non linear method (Decision Tree)



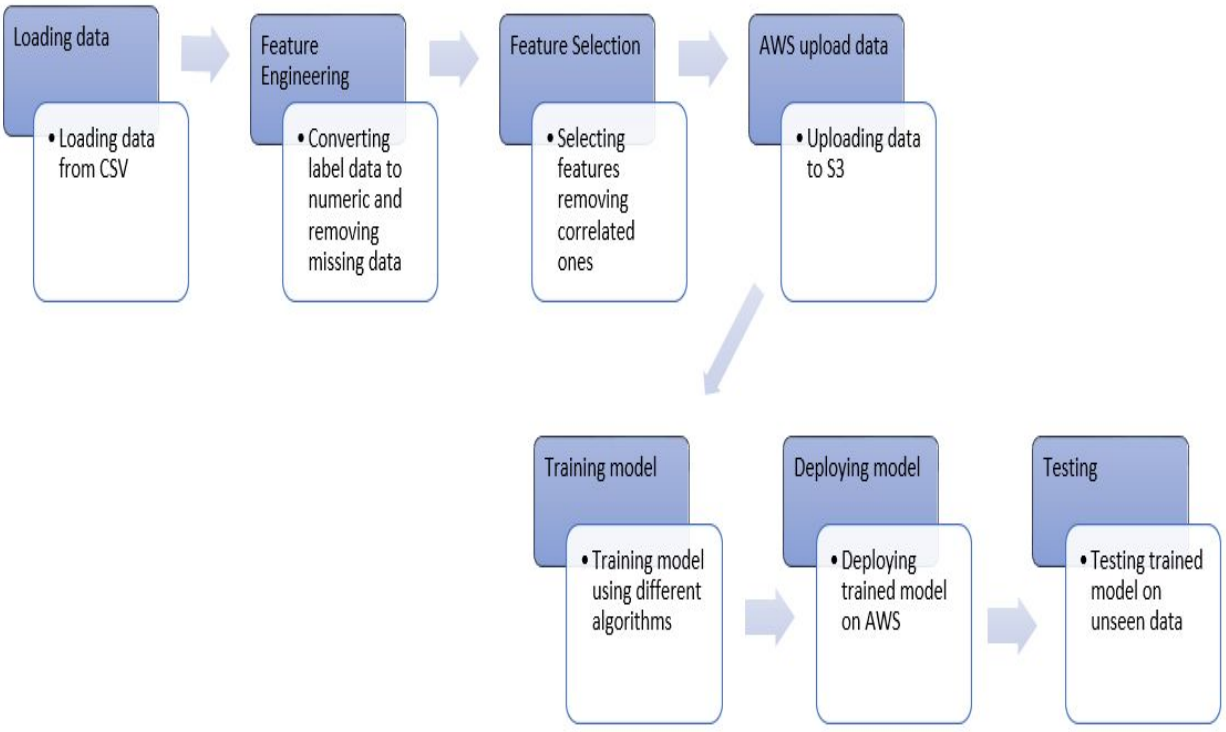
* image taken from Sklearn website

BenchMark

I compare my results with different results on kaggle Submissions for this dataset.

An accuracy of minimum 85 percent is required (on test set)

Design and Results



For feature selection, I try to remove correlated features.

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates
Administrative	1.000000	0.601466	0.376782	0.255757	0.430832	0.373647	-0.223474
Administrative_Duration	0.601466	1.000000	0.302647	0.237959	0.288869	0.355238	-0.144128
Informational	0.376782	0.302647	1.000000	0.618965	0.374098	0.387446	-0.116071
Informational_Duration	0.255757	0.237959	0.618965	1.000000	0.279966	0.347300	-0.074077
ProductRelated	0.430832	0.288869	0.374098	0.279966	1.000000	0.860868	-0.204469
ProductRelated_Duration	0.373647	0.355238	0.387446	0.347300	0.860868	1.000000	-0.184409
BounceRates	-0.223474	-0.144128	-0.116071	-0.074077	-0.204469	-0.184409	1.000000
ExitRates	-0.316192	-0.205618	-0.163539	-0.105205	-0.292219	-0.251645	0.913436
PageValues	0.098771	0.067463	0.048539	0.030787	0.056067	0.052623	-0.119357
SpecialDay	-0.095054	-0.073472	-0.048328	-0.030658	-0.024190	-0.036598	0.073088
Month	0.048543	0.029048	0.019737	0.005986	0.070289	0.061174	-0.023806
OperatingSystems	-0.006459	-0.007425	-0.009435	-0.009596	0.004193	0.002885	0.023965
Browser	-0.025243	-0.015525	-0.038257	-0.019346	-0.013326	-0.007549	-0.016009
Region	-0.005680	-0.005681	-0.029442	-0.027244	-0.038318	-0.033263	-0.006731
TrafficType	-0.033748	-0.014487	-0.034510	-0.024731	-0.043235	-0.036538	0.078894
VisitorType	-0.025506	-0.023743	0.055972	0.044781	0.126995	0.119640	0.135470
Weekend	0.026404	0.014987	0.035557	0.024054	0.016097	0.007293	-0.046870
Revenue	0.138631	0.093395	0.095085	0.070250	0.158280	0.152130	-0.150621

BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
-0.223474	-0.316192	0.098771	-0.095054	0.048543	-0.006459	-0.025243	-0.005680	-0.033748	-0.025506	0.026404	0.138631
-0.144128	-0.205618	0.067463	-0.073472	0.029048	-0.007425	-0.015525	-0.005681	-0.014487	-0.023743	0.014987	0.093395
-0.116071	-0.163539	0.048539	-0.048328	0.019737	-0.009435	-0.038257	-0.029442	-0.034510	0.055972	0.035557	0.095085
-0.074077	-0.105205	0.030787	-0.030658	0.005986	-0.009596	-0.019346	-0.027244	-0.024731	0.044781	0.024054	0.070250
-0.204469	-0.292219	0.056067	-0.024190	0.070289	0.004193	-0.013326	-0.038318	-0.043235	0.126995	0.016097	0.158280
-0.184409	-0.251645	0.052623	-0.036598	0.061174	0.002885	-0.007549	-0.033263	-0.036538	0.119640	0.007293	0.152130
1.000000	0.913436	-0.119357	0.073088	-0.023806	0.023965	-0.016009	-0.006731	0.078894	0.135470	-0.046870	-0.150621
0.913436	1.000000	-0.174397	0.102899	-0.039103	0.014745	-0.004407	-0.008836	0.078998	0.178928	-0.062942	-0.206886
-0.119357	-0.174397	1.000000	-0.063660	0.021768	0.018466	0.045510	0.011233	0.012471	-0.111098	0.011993	0.492494
0.073088	0.102899	-0.063660	1.000000	0.079332	0.012609	0.003412	-0.016188	0.052273	0.085713	-0.016792	-0.082468
-0.023806	-0.039103	0.021768	0.079332	1.000000	-0.029595	-0.045932	-0.032572	0.041778	0.026502	0.029163	0.080140
0.023965	0.014745	0.018466	0.012609	-0.029595	1.000000	0.222916	0.076785	0.189072	0.001568	0.000278	-0.014740
-0.016009	-0.004407	0.045510	0.003412	-0.045932	0.222916	1.000000	0.097297	0.111985	-0.021756	-0.040212	0.023869
-0.006731	-0.008836	0.011233	-0.016188	-0.032572	0.076785	0.097297	1.000000	0.047266	-0.036094	-0.000966	-0.011717
0.078894	0.078998	0.012471	0.052273	0.041778	0.189072	0.111985	0.047266	1.000000	-0.002751	-0.002575	-0.005212
0.135470	0.178928	-0.111098	0.085713	0.026502	0.001568	-0.021756	-0.036094	-0.002751	1.000000	-0.043687	-0.104548
-0.046870	-0.062942	0.011993	-0.016792	0.029163	0.000278	-0.040212	-0.000966	-0.002575	-0.043687	1.000000	0.029293
-0.150621	-0.206886	0.492494	-0.082468	0.080140	-0.014740	0.023869	-0.011717	-0.005212	-0.104548	0.029293	1.000000

Because of couple of high correlation pairs (Product_related_duration <-> Product_related And Exit_rates <-> Bounce_rates), we can safely remove ‘ExitRates’ and ‘Product_related_duration’

After shuffling data and splitting in 70-30 train test ratio, we obtain following results.

SVC ->

```
Training seconds: 58
Billable seconds: 58
CPU times: user 527 ms, sys: 24.9 ms, total: 552 ms
Wall time: 3min 41s
```

Accuracy = 0.85304

Refinement

Decision Tree ->

```
Training seconds: 47
Billable seconds: 47
CPU times: user 478 ms, sys: 14.9 ms, total: 493 ms
Wall time: 3min 11s
```

Accuracy = 0.87036

Decision Tree clearly has better accuracy. Though there is definitely scope for improvement.

Conclusion and Future Scope

Though linear and Decision tree based models have worked decently well, there still is a huge scope for improvement. More sophisticated feature selection algorithms along with Ensembled models like XGBoost might improve the accuracy.

Also a little more research in the literature might help with custom computed features augmentation in our dataset.

Also, one can try deep learning based methods which might be able to capture pair relations amongst features.

Acknowledgement

I would like to extend my sincere thanks to the Udacity team for helping students at every step of different projects.