Final Report

- ONLINE PURCHASE INTENTION -

By: Tanvi Gandhi

- ## PROPOSAL

- ### MOTIVATION

In this project, we will develop a classification model measuring the user's intention to finalize an online purchase transaction. By analyzing the various behavioral attributes of the user while on an online shopping platform can give important insights to target revenue generating customers.

- ### WHO MIGHT CARE?

Due to the boom of the online e-commerce industry in recent years, our problem stands relevant for various segments who want to monitor user behavior. The clientele in question are the various digital shopping websites/vendors/platforms who would like to maximize their revenue by taking into consideration the user intent by the various actions he performs while on the website. Once model is built, it can be applied in real time so as to focus on those customers more who have a higher probability to purchase a product. For those customers, whose probability is less likely, extensive recommendations can be provided to increase their website time and conversion marketing can be applied to turn them into revenue generating customers. It would also be possible to identify the times of the year when sales increase and thus help in providing better response time to customers. The vendor can focus on the season special products and highlight them during the peak season. There can be various application of the model and more input feature vectors would aid in tailoring a better model.

- ### DATA

The data is extracted from the UCI machine Learning Repository (link: *https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset*). It is a

recently donated dataset and is used in the research study by Sakar et al (link: *https://doi.org/10.1007/s00521-018-3523-0*).

The data is collected for different users (12,330 sessions specifically) over a period of 1 year so as to avoid any tendency to a specific campaign, special day, user profile, or period. It contains 18 features vectors like "Revenue", "Product Related", "Product Related Duration", "Bounce Rate", "Exit Rate", "Page Value" etc.

File: online_shoppers_intention.csv

This file consists of various informational values related to a customer behavior in an online shopping session.

• DATA WRANGLING

• DATA INFORMATION

The data has 12330 rows and 18 feature vectors including the categorical target feature. Among them there are 10 numerical and 8 categorical features.

Below is a brief explanation of the variables: -

Numerical features

Administrative: Total number of administration related pages visited by the visitor in that session like account management.

Informational: Total number of Informational pages visited by the visitor in that session like information about offers, cart etc.

ProductRelated: Total number of product related pages visited by the visitor in that session.

Note: The values of the above 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.

Administrative Duration: Total amount of time (in seconds) spent by the visitor on adminitration related pages.

Informational Duration: Total amount of time (in seconds) spent by the visitor on information related pages.

ProductRelated Duration: Total amount of time (in seconds) spent by the visitor on product related pages.

Note: The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

BounceRates: Average bounce rate value of the pages visited by the visitor. 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.

ExitRates: Average exit rate value of the pages visited by the visitor. 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.

PageValues: Average page value of the pages visited by the visitor. The Page Value feature represents the average value for a web page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both).

SpecialDay: Closeness of the site visiting time to a 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.

Categorical features

Month: Month value of the visit date

OperatingSystems: Operating system of the visitor

Browser: Browser of the visitor

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

TrafficType: Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)

VisitorType: Visitor type as 'New Visitor', 'Returning Visitor' and 'Other'

Weekend: Boolean value indicating whether the date of the visit is weekend

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

- DATA CLEANING

The data had some inconsistency and unavailable data. Both of these errors could not be handled. Below are the explanations:

- The data is available for 10 months (i.e. Jan and April not included). This discrepancy cannot be handled.

- There seems to be some sort of inconsistency in the special days data since it does not have any values for December month - which has Christmas and New Year as the biggest holidays. This cannot be imputed since no data is available.
- The target vector (binary True and False) is highly imbalanced as the False values are as much as 5 times of True Values. To handle this, we will perform SMOTE analysis in the pre-processing phase.
- The data contains 125 duplicate rows. However, we can keep them since it is only around 1% of the data and isn't necessarily erroneous. Further, we also need to realize the reasons due to which the data can seem duplicate:
    1. It is possible that two different sessions of two different users can be same i.e. they visited the exact same pages and may have same browser or OS type and reside in the same region. This is rare event but not impossible.
    2. It is likely that the same user (or a relative) could perform the same set of page flow from other device with same configuration of OS and browser at around the same time.

• MISSING VALUES

Surprisingly, there were no missing values in the dataset.

• OUTLIERS

Below is a screenshot of detected outliers using IQR method:

Administrative	404
Administrative_Duration	1172
BounceRates	1551
Browser	4369
ExitRates	1099
Informational	2631
Informational_Duration	2405
Month	0
OperatingSystems	111
PageValues	2730
ProductRelated	987
ProductRelated_Duration	961
Region	511
Revenue	1908
SpecialDay	1251
TrafficType	2101
VisitorType	0
Weekend	2868

We can exclude the categorical features from the outlier's consideration since it doesn't make sense to have outliers for them. For e.g., Weekend column might detect outliers

since there are too many False values as there are True values. But this scenario is natural in real life as there are a greater number of non-weekends days than weekend days.

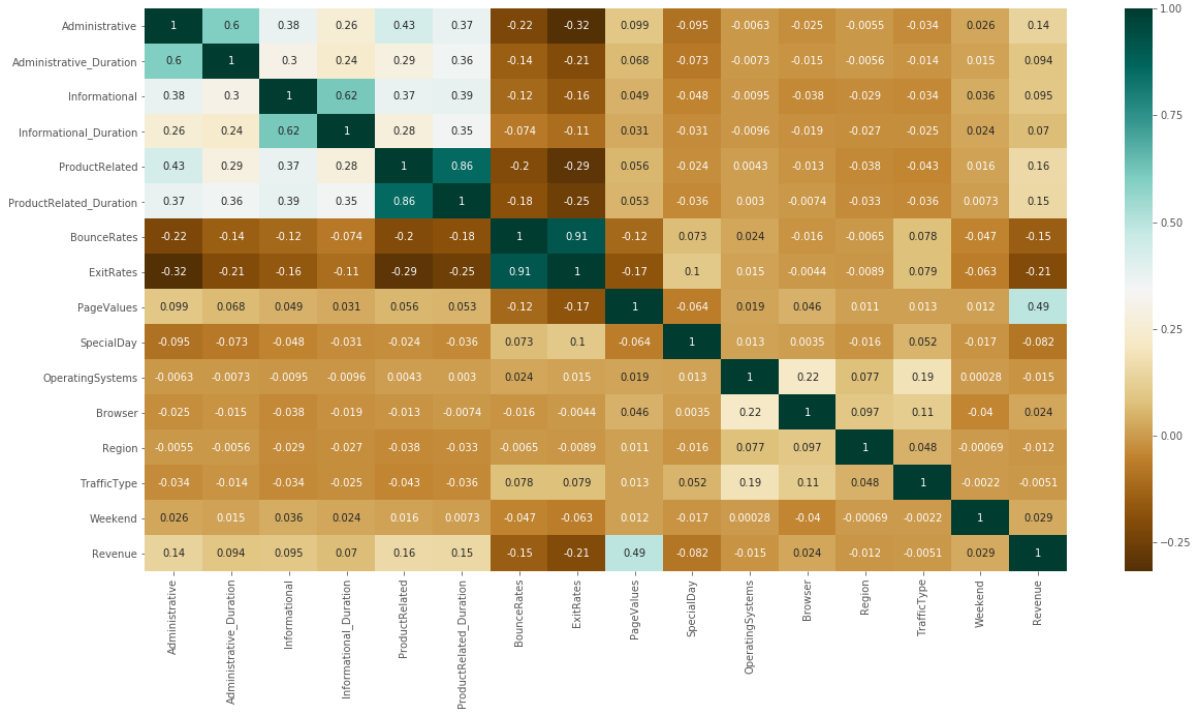
Coming to the categorical data, the high numbers of outliers suggest that they might be natural outliers (as opposed to measurement error) so it would be okay to keep them. For e.g. outliers for BounceRates are 1551. It's not surprising to think that a frequently bounced page might sometime yield revenue. Also, a page with low bounce rate might not always result in revenue. So, these measurements might be detected as outliers but are normal if considered in real-life.

In our case, we will do nothing with outliers since all the numerical features are metrics and are suggestive of natural outliers. Thus, removing them can result in loosing important information. We can later on apply Standardization to bring the values to the same scale.

• INFERENTIAL ANALYSIS

• DATA INFORMATION

We analyzed the data for identifying any relationships between the different variables by constructing boxplots, scatter plots, pair plots etc. The heatmap, as shown below, was able to realize the correlation between the various variables.

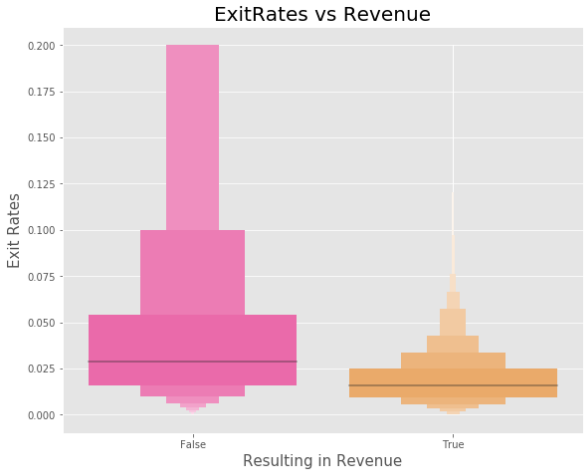


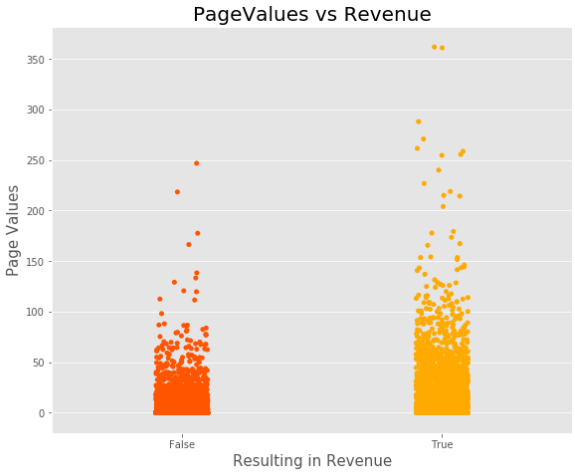
The significant observations are:

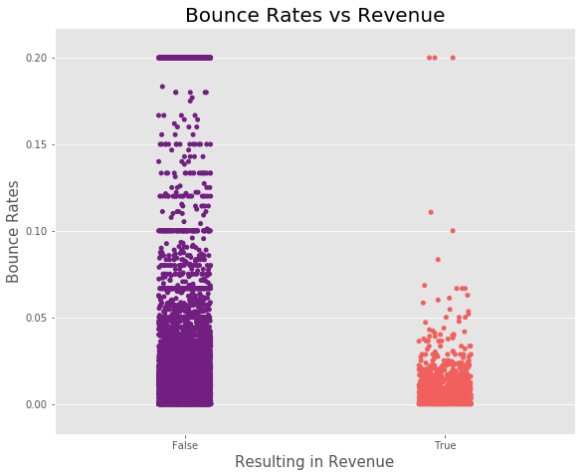
- ProductRelated and PoductRelated_Duration are highly correlated which may be an expected scenario since more the number of product pages the user will visit, the more will be the cumulative duration.
- From the above logic, we can also explain the significant relationship between Administrative & Administrative_Duration, and Informational & Informational_Duration.
- There is very high correlation between ExitRates and BounceRates since both terms intrinsically define the percentage of user visits which end up leaving a page, but in different ways.
- PageValues is a significant variable to predict the target Revenue.

• INTERESTING PLOTS

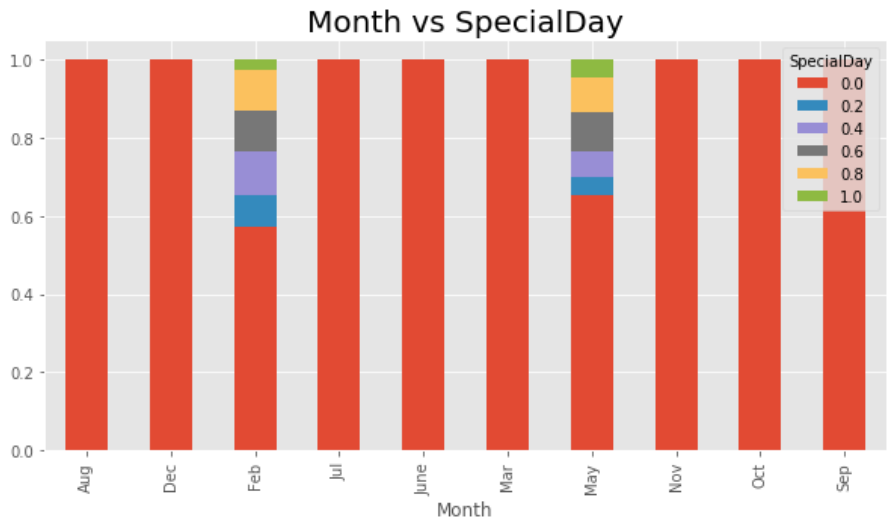
Below are some of the many interesting plots which allow us to visualize the relationships between the variables:



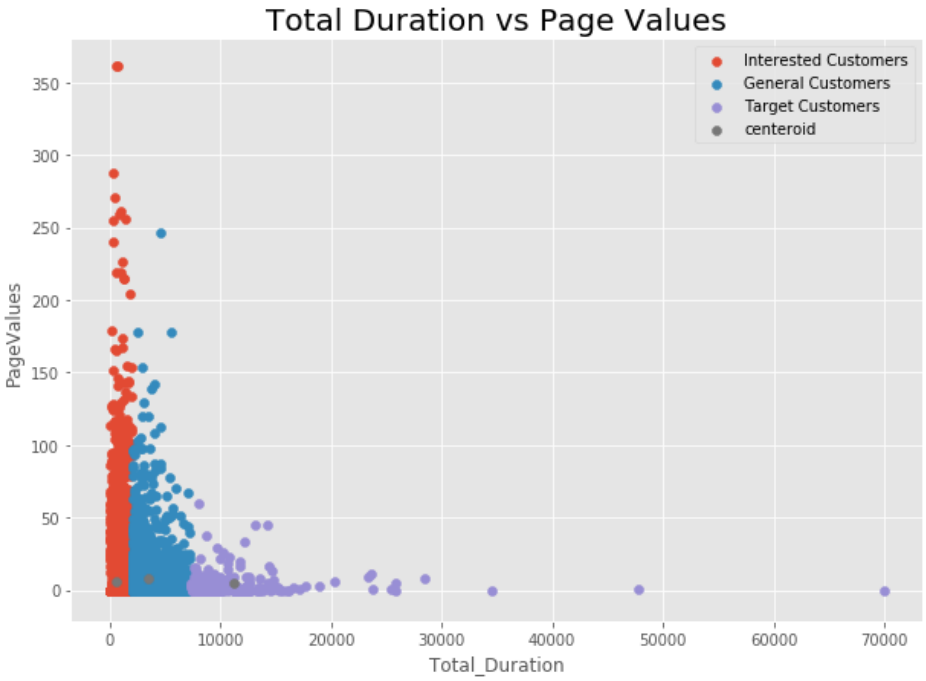




The above three plots show the association of different page view analytics like ExitRates, BounceRates and PageValues with the target variable. Typically, the pages with average high BounceRates and/or ExitRates are corresponding to no revenue generation. The pages with average high PageValues are resulting in revenue generation.



From the above graph, it is evident that the data is not available for all the twelve months. The data for January and April is not present. Further, the SpecialDay variable is not consistent for all the months as it is available for only two months. From other plots, it was observed that SpecialDay with value 0.6 was second most revenue generating after value 0.0 of SpecialDay.



In the above graph, the Total_Duration is calculated by the summation of Administrative_Duration, Informational_Duration and ProductRelated_Duration. The cluserting of different segments of the customers spending the specified total duration in accordance with the PageValues, shows the two extreme scenario. One in red reflect the customers who spend very less duration but visit pages with average high PageValues, these may be the interested customers who might visit with a specific product in mind. The other segment in the purple reflect the customers who spent a lot of time on the website but visit pages with low average PageValues, these might be the visitors who might randomly visit the website to search for a good product and end up spending a lot of time to explore the products but mostly visit pages with low average Page Values. These latter customers are our target to focus on and analyze the pages they visit.

- Logistic Regression Analysis

To analyze the P-values of the various variable with respect to the target variable, we performed Logistic Regression on the data. The resulting co-efficients were log odds (logarithm of odds of success/failure) and hence exponentiating these values would be more interpretable. The P-values are more than 0.05 for many variables, we can try to normalize the variables later on.

Logit Regression Results

Dep. Variable:	Revenue	No. Observations:	12330
Model:	Logit	Df Residuals:	12313
Method:	MLE	Df Model:	16
Date:	Sat, 25 Apr 2020	Pseudo R-squ.:	0.2876
Time:	00:18:49	Log-Likelihood:	-3784.5
converged:	True	LL-Null:	-5312.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Administrative	-0.0168	0.011	-1.532	0.126	-0.038	0.005
Administrative_Duration	-0.0003	0.000	-1.438	0.150	-0.001	0.000
Informational	0.0222	0.026	0.840	0.401	-0.030	0.074
Informational_Duration	9.208e-05	0.000	0.415	0.678	-0.000	0.001
ProductRelated	0.0027	0.001	2.321	0.020	0.000	0.005
ProductRelated_Duration	7.031e-05	2.79e-05	2.521	0.012	1.57e-05	0.000
BounceRates	6.0499	3.196	1.893	0.058	-0.214	12.314
ExitRates	-26.3971	2.395	-11.022	0.000	-31.091	-21.703
PageValues	0.0773	0.002	32.776	0.000	0.073	0.082
SpecialDay	-0.9540	0.219	-4.361	0.000	-1.383	-0.525
Month	0.0077	0.011	0.695	0.487	-0.014	0.029
OperatingSystems	-0.3726	0.035	-10.547	0.000	-0.442	-0.303
Browser	-0.0282	0.019	-1.482	0.138	-0.065	0.009
Region	-0.0687	0.013	-5.439	0.000	-0.093	-0.044
TrafficType	-0.0081	0.008	-0.966	0.334	-0.025	0.008
VisitorType	-0.3932	0.037	-10.733	0.000	-0.465	-0.321
Weekend	-0.0294	0.069	-0.425	0.671	-0.165	0.106

DATA PRE-PROCESSING AND MODELLING

PRE-PROCESSED DATA

Before applying various Machine Learning techniques, we encoded and scaled the data. The 7 categorical features were encoded using *get_dummies* function with the first column dropped for each feature to avoid correlation. This resulted in total of 68 columns including numerical and categorical features. The target variable *Revenue* was label encoded.

The data was further feature scaled using *StandardScalar* function. It was then split into 80% training set and rest testing set.

NAÏVE BAYES

The typical Gaussian Naïve Bayes for the given dataset gave just 29% accuracy which is very low for our scenario. We then balanced the data using SMOTE algorithm and the accuracy increased to 63%.

LOGISTIC REGRESSION

For Logistic Regression also, we performed both on the original dataset and balanced dataset using SMOTE. The accuracy for former method is 88% and with latter strategy is 86%. It is to note that the balanced data has a reduced accuracy.

RANDOM FOREST

For Random Forest, we tried 4 variants. The first is, Random Forest with given dataset which gave 89% accuracy. Next is Random Forest with 40 important extracted features out of 68 and gave the accuracy of 90.51%. The third is, Random Forest with given dataset and balancing using SMOTE which gave 94% accuracy. The last variant we tried is Random Forest with 50 extracted important features and balancing using SMOTE. It gave the accuracy of 90.4%.

XGBOOST

We ran the default XGBoost algorithm which gave us 90.51% accuracy. We then tried to fine tune the parameters and the accuracy was 90.1% suggesting that we need to further hyper tune the model.

- ## AGGREGATED ANALYSIS

The below table shows that Random Forest with extracted features and default XGBoost gave approximately equal accuracies and Random Forest with SMOTE gave highest accuracy of all other models.

Model Variation	Accuracy
1. Naïve Bayes	29%
2. Naïve Bayes with SMOTE	63%
3. Logistic Regression	88%
4. Logistic Regression with SMOTE	86%
5. Random Forest	89%
6. Random Forest with important features	90.51%
7. Random Forest with SMOTE	**94%**
8. Random Forest with important features and SMOTE	90.4%
9. Default XGBoost	90.51%
10. Tuned XGBoost	90.1%