Online shoppers Intentions

ML PROJECT

- Venu Gopal Rao Pendyala
Batch Code: l87st2

Problem Statement - Objective

This is the data of an online retailing company where they are trying to find which online shopper will generate revenue by his/her online shoppers' activity on their site.

People often spend lot of time browsing through online shopping websites, but the conversion rate into purchases is low. Determine likelihood of purchase based on the given features in the dataset. The dataset consists of 18 features belonging to 12,330 online transactions.

The Objective of this project is to identify the user behaviour patterns to effectively understand features that influence n create a ML model which predicts shopping intent of website visitors to PURCHASE or NO PURCHASE.

Feature Analysis

Output Variable : Revenue

Datatypes : Int-7, Float – 7, Obj – 2, Bool – 2

No Null Values

No Duplicates

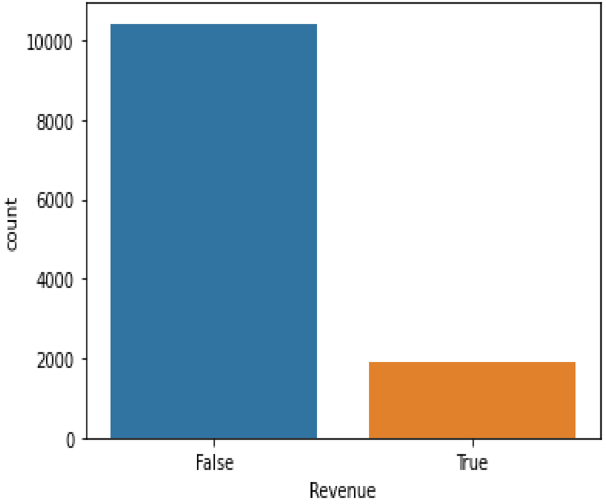
Outliers Treatment to be done

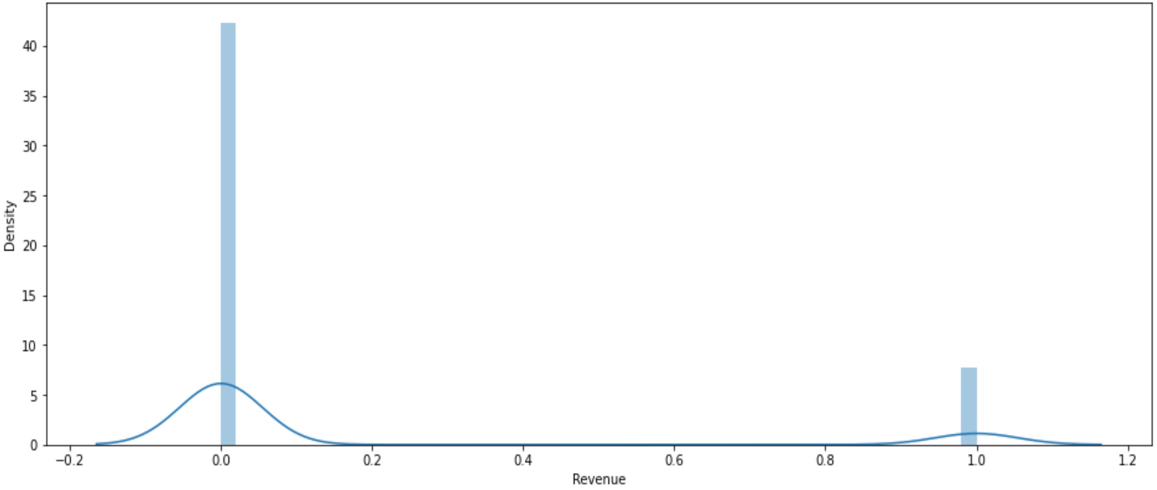
Features	Description	DataType
Administrative	Number of the pages visites by the user for user account management related activities	Discrete values from 0 to 27
Administrative_Duration	Time spent on admn page by user	continuous value of time in seconds
Informational	Number of pages visites by user for information	Discrete values from 0 to 24
Informational_Duration	Time spent on informational page by user	continuous value of time in seconds
ProductRelated	Number of product related visited by the user	Discrete values from 0 to 705
ProductRelated_Duration	Time spent on productrelated pages by user	continuous value of time in seconds
BounceRates in %	Average bounce rate of the pages visited by the user	continuous value
ExitRates in %	Average exit rates of the pages visited by the user	continuous value
PageValues	Average page value of the pages visited by the user	continuous value
SpecialDay (probability)	special event days like mothers day, valentine day etc.,	Discrete values 0.2, 0.4, 0.6, 0.8, 1.0
Month	Month of the visit from Jan to Dec of the year	Categorical
OperatingSystems	Operating systems used by visited users in their systems	Discrete values from 1 to 8
Browser	Browser used by the user to visit the web site/shoppers site	Discrete values from 1 to 13
Region	Region of the user from where they started the session	Discrete values from 1 to 9
TrafficType	Traffic source from where user entered the website	Discrete values from 1 to 20
VisitorType	Visitor type as new visitor or returning visitor	Categorical
Weekend	If the user visited on weekend or not	Boolean
Revenue	If the user revenue generated or not	Boolean

Data Description - Stats

Feature	count	mean	std	min	25%	50%	75%	max
Administrative	12330.0	2.315166	3.321784	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12330.0	80.818611	176.779107	0.0	0.000000	7.500000	93.256250	3398.750000
Informational	12330.0	0.503569	1.270156	0.0	0.000000	0.000000	0.000000	24.000000
Informational_Duration	12330.0	34.472398	140.749294	0.0	0.000000	0.000000	0.000000	2549.375000
ProductRelated	12330.0	31.731468	44.475503	0.0	7.000000	18.000000	38.000000	705.000000
ProductRelated_Duration	12330.0	1194.746220	1913.669288	0.0	184.137500	598.936905	1464.157214	63973.522230
BounceRates in %	12330.0	0.022191	0.048488	0.0	0.000000	0.003112	0.016813	0.200000
ExitRates in %	12330.0	0.043073	0.048597	0.0	0.014286	0.025156	0.050000	0.200000
PageValues	12330.0	5.889258	18.568437	0.0	0.000000	0.000000	0.000000	361.763742
SpecialDay (probability)	12330.0	0.061427	0.198917	0.0	0.000000	0.000000	0.000000	1.000000
OperatingSystems	12330.0	2.124006	0.911325	1.0	2.000000	2.000000	3.000000	8.000000
Browser	12330.0	2.357097	1.717277	1.0	2.000000	2.000000	2.000000	13.000000
Region	12330.0	3.147364	2.401591	1.0	1.000000	3.000000	4.000000	9.000000
TrafficType	12330.0	4.069586	4.025169	1.0	2.000000	2.000000	4.000000	20.000000

EDA – TARGET ANALYSIS



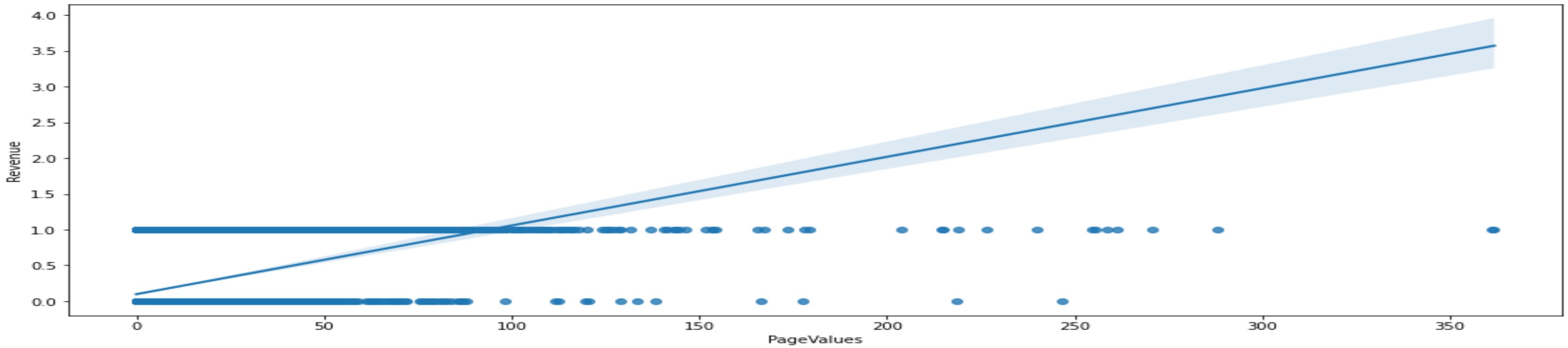


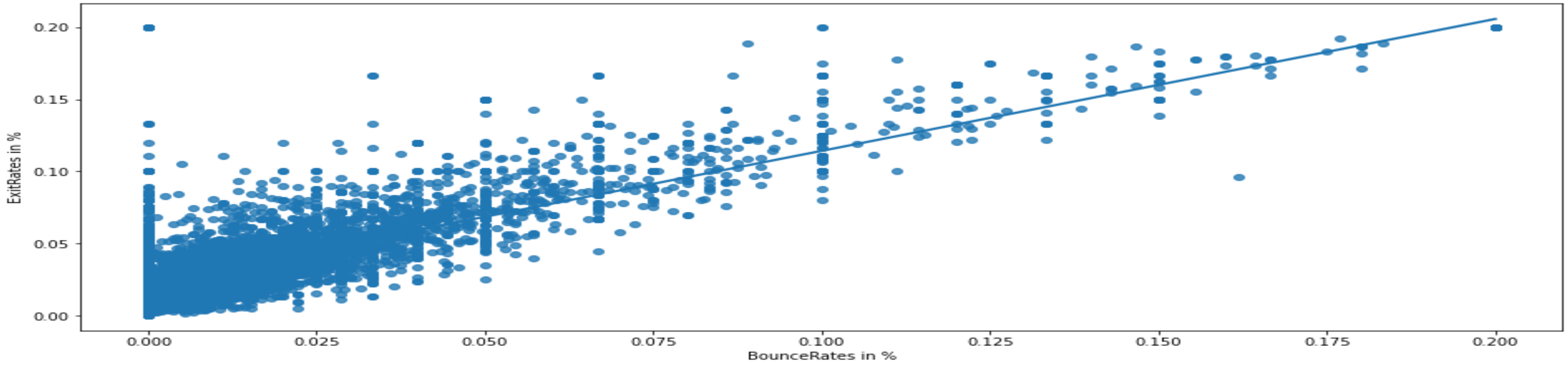
1. Out of total 12330 transactions only 1908 transactions converted to Purchase transactions, which is only 15%.
2. Major chunk of Purchase transactions occurred from the transaction visited Admn and Information pages. Which is around 22%.
3. 50% i.e., transactions directly visited Product page may be visited from out side referal websites but revenue generated transactions are very low at 8% only.

EDA – COLLINEARITY

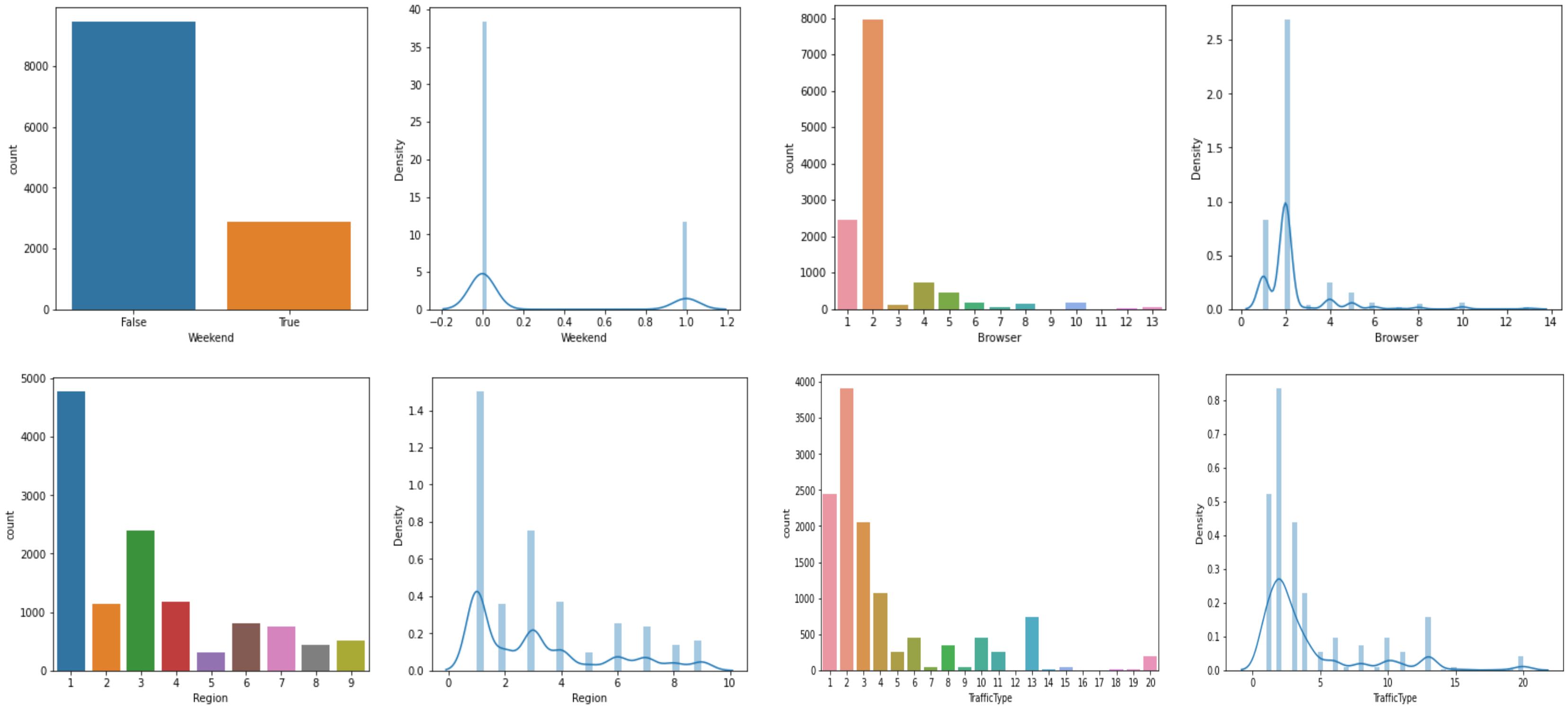
Observations :

1. Based on correlation and heat map observations it is found that only Page values have linear relation with Revenue
2. There is also multicollinearity observed between Bouncerates and Exitrates





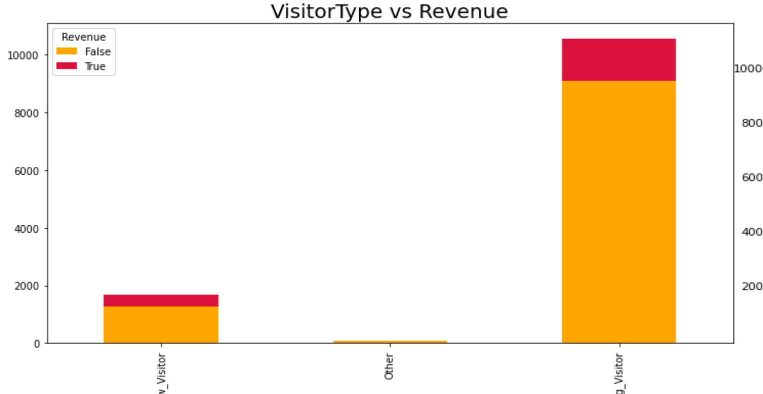
EDA – UNIVARIATE ANALYSIS

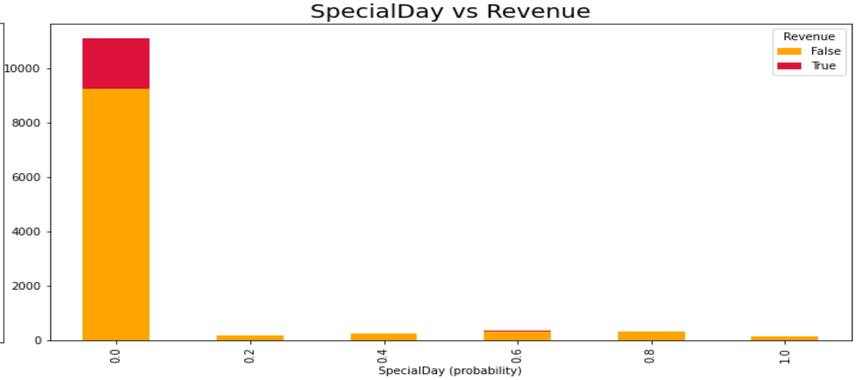


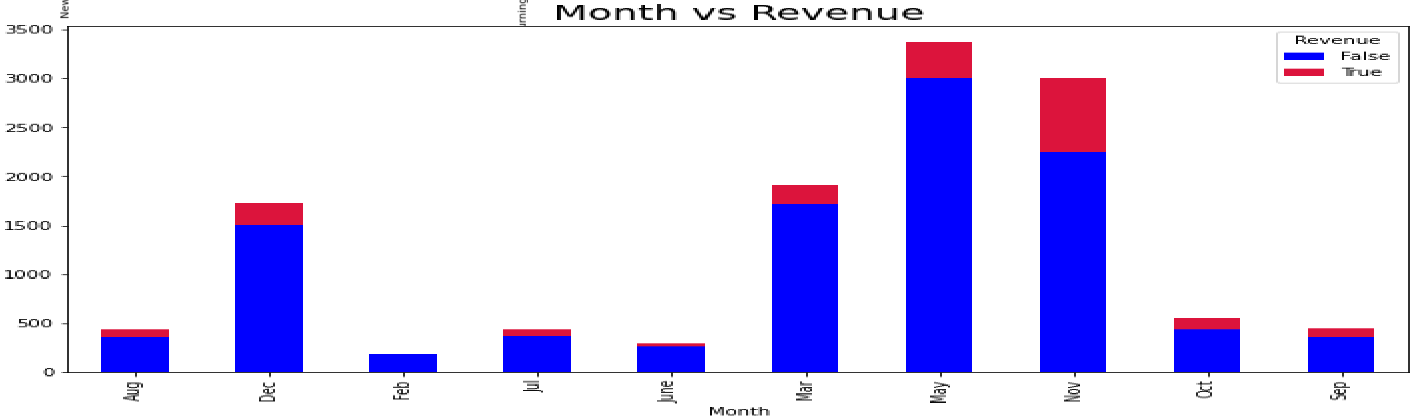
1. Only 23%(around 2900) page visits happened in weekends, rest of 77% visits happened in other days.
2. Highest page visits and Purchase page visits observed from Browser 2 Followed by page visits of Browsers 1, 4,5.
3. It is observed that Region 1 leading with 39% page visits followed by 3 with 20%
4. It is observed that Traffic type 2 leading with 31% page visits followed by 1 with 20% Next level observed with Traffic type 3 and 4 with 16% and 8% page visits.

EDA – BI VARIATE ANALYSIS – TARGET VS FEATURE

1. Highest Purchase page visits happened in Nov month and then in May month.
2. Highest page visits and Purchase page visits observed from Returning Visitor, Followed by page visits of New visitors.
3. Special day impact on page visits or Revenue generated page visits is not visible hence there is no impact

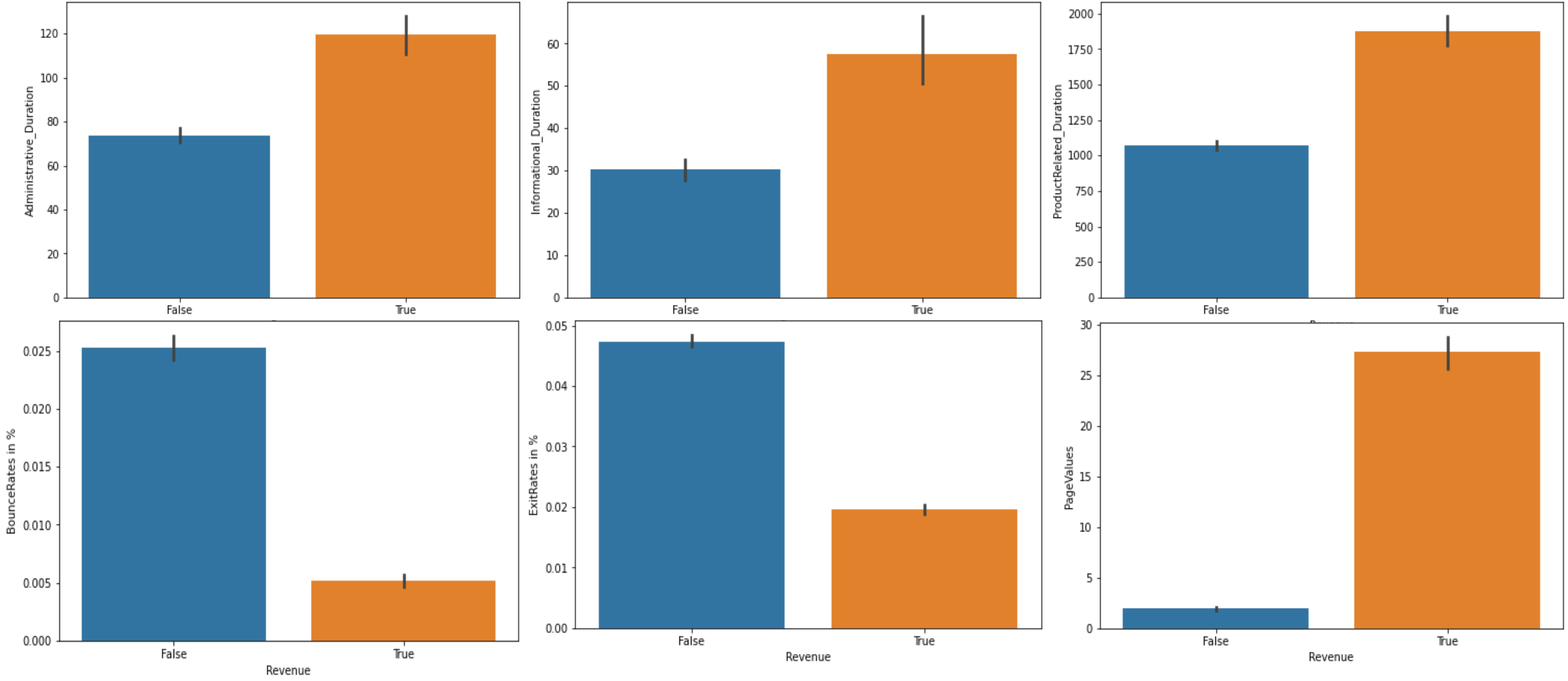






EDA – BI VARIATE ANALYSIS – TARGET VS FEATURE

1. Administrative Duration, Informational duration, product related duration and Page values has positive impact on Revenue.
2. Where as Bounce Rates, Exit Rates have negative impact on Revenue generated page visits.

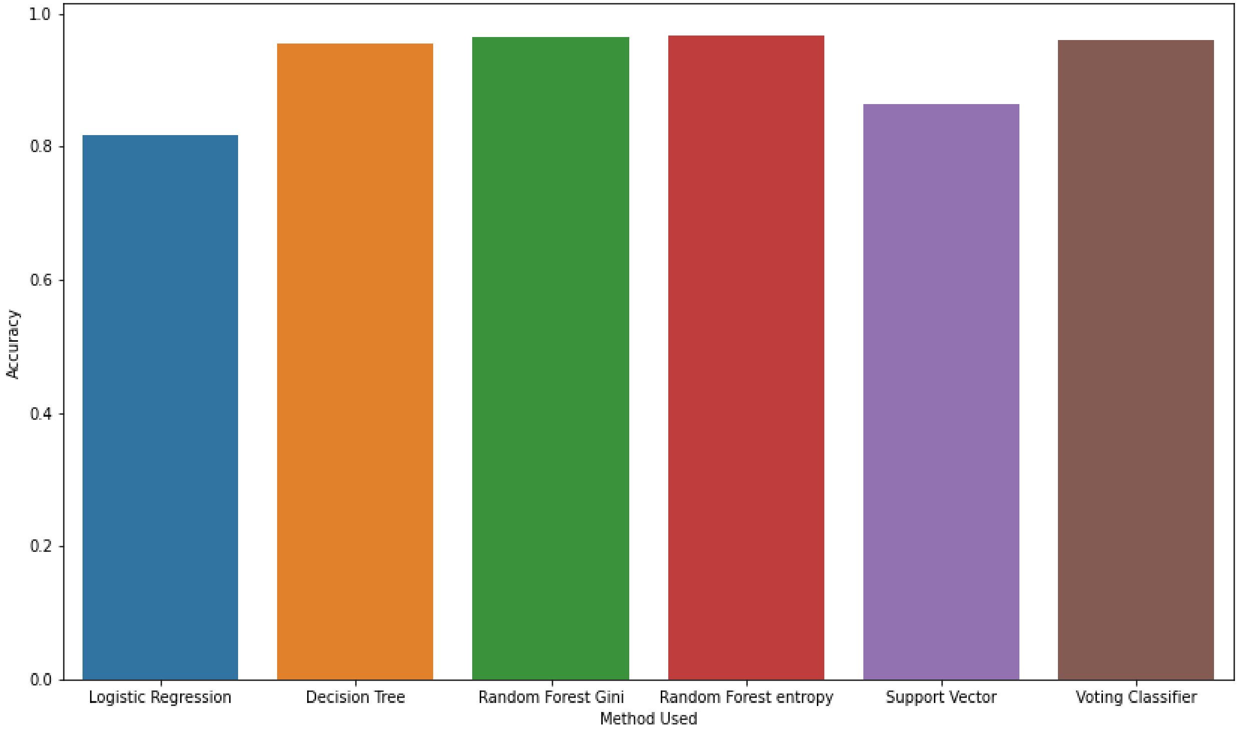


DATA PREPROCESSING

1. No null values and no duplicates.
2. ONE HOT ENCODING done on two categorical variables i.e., Month and Visitor type.
3. Label Encoding done on Bool variables i.e., Weekend and Revenue.
4. Scaling done on numerical variables using Standard scaler.
5. Outliers treatment done thru capping and flooring by taking <1% for flooring and >99% for capping.
6. Data set found to be imbalanced as major > 2 times minor. To handle this Random oversampling done after applying K fold cross validation.

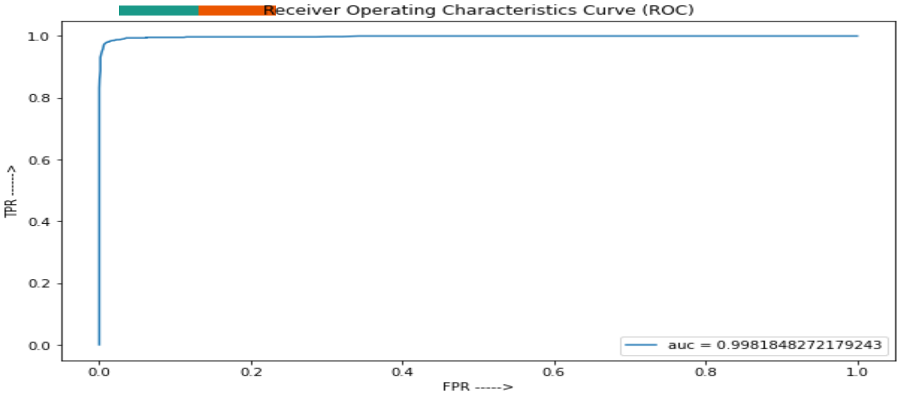
MODEL FITTING & PERFORMANCE

1. Total 5 models used in the process to observe which model is delivering better results, then all models given input to voting classifier to combine the output of all models to get refined output.
2. But observed Random Forest model has done farely good job and given better out put.



```
       Method Used            Accuracy
0   Logistic Regression       0.824898
1   Decision Tree             0.957544
2   Random Forest Gini        0.968098
3   Random Forest entropy     0.966659
4   Support Vector            0.866395
5   Voting Classifier         0.960902
```

AUC & METRICS



S.No.	model_name	metrics	Is_Revenue (False)	Is_Revenue (True)
0	Random Forest(entropy)	precision	1.00	0.94
1	Support Vector Machine	precision	0.85	0.88
2	Voting Classifier	precision	1.00	0.93
3	Logistic Regression	precision	0.78	0.86
4	Decision Tree	precision	1.00	0.92
5	Random Forest(entropy)	recall	0.94	1.00
6	Support Vector Machine	recall	0.88	0.84
7	Voting Classifier	recall	0.92	1.00
8	Logistic Regression	recall	0.87	0.76
9	Decision Tree	recall	0.91	1.00
10	Random Forest(entropy)	f1_score	0.97	0.97
11	Support Vector Machine	f1_score	0.87	0.86
12	Voting Classifier	f1_score	0.96	0.96
13	Logistic Regression	f1_score	0.82	0.80
14	Decision Tree	f1_score	0.95	0.95

NOTE:

Random Forest model has delivered accurate predictions both in False as well True scenarios with 97% f1_score. Hence it is recommended to use this model.

DATASET OBSERVATIONS

1. Out of total 12330 transactions 6562 visited Admin page, out of these 6537 i.e., 99% visited product page, 2168 visited information pages. Out of 1908 Revenue generated transactions 1394 Transactions occurred from these pages. Which is 21%.
2. Out of total 12330 transactions 2631 visited Info page, out of these 2623 i.e., 99% visited product page, 2168 came to info from Admn pages. Revenue generated transactions is 23% from information pages
3. 5299 transactions directly visited Product page. But Revenue generated transactions are only 427 from these 5299 transactions which is just 8%.
4. Overall out of 12330, 12292 visited Product pages which is 99%.
5. Feature importance table appended below with Top 10 features, as per which Page values feature is at top with 38% followed by producted related duration and Exit rates.

S.No.	Feature Name	importance
1	PageValues	0.381310
2	ProductRelated_Duration	0.091788
3	ExitRates in %	0.089217
4	ProductRelated	0.067117
5	BounceRates in %	0.057214
6	Administrative_Duration	0.046151
7	Administrative	0.039228
8	Month_Nov	0.029544
9	TrafficType	0.028132
10	Region	0.026786

Best feature	Name	NO. OF VISITS	% OF REVENUE GEN. VISITS
Month	November	760	39.83%
Browser	2	1223	64.10%
Visitor type	Returning_Visitor	1470	77%
Traffic Type	2	847	44.39%
Region Tye	1	771	40.41%
Browser Type	2	1223	64.10%
OS type	2	1155	60.53%

RECOMMENDATIONS

As per Feature Importances, PageValue is the most important feature and pages with high average value contribute more to revenue generation. Replicate the characteristics of these pages in other pages to enhance page value

From EDA, and survey the month of November and May has always been great for ecommerce and revenue may be due to Festival season and Summer. We can leverage this and produce better turnouts by offering customers with holiday gifts, discounts sale, goodies etc.,

Identify the pages with high bounce rate, exit rates and divert traffic to other pages. Also enhance webpage content so that these parameters reduced.

Focus on Bests in features like returning visitor, browser-2, region-2 etc., which are generating more transactions and revenue.

Deploy the model if possible to measure time to predict behaviour in realtime

THANK YOU