PURCHASE PROPENSITY MODELLING

Bharadwaj Kacharla, Claire Lin, Nikhil Joshi, Nouf Alzamel, Xiaoqing Li, Zoe Yu

PRESENTATION OVERVIEW

- EXECUTIVE SUMMARY
  - Outline background, objective and business use case
- EXPLORATORY DATA ANALYSIS - EDA
  - Overview of the data
- MODELLING AND MODEL VALIDATION
  - Describe model type and success metrics
- FUTURE DIRECTIONS
  - Conclusion and the future work

EXECUTIVE SUMMARY

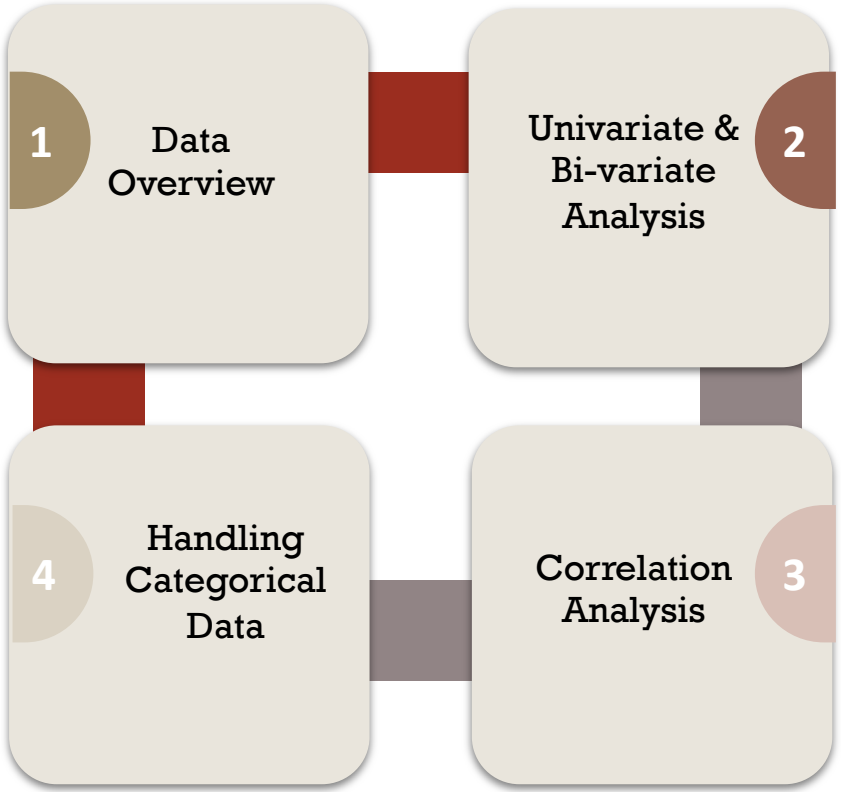
EXECUTIVE SUMMARY

- Background:
  - The increase in e-commerce usage over the past few years has created potential in the market, but the fact that the conversion rates have not increased at the same rate leads to the need for solutions that present customized promotions to the online shoppers
  - In physical retailing, a salesperson can offer a range of customized alternatives to shoppers based on the experience he or she has gained over time. This experience has an important influence on the effective use of time, purchase conversion rates, and sales figures
  - Many e-commerce and information technology companies invest in early detection and behavioral prediction systems which imitate the behavior of a salesperson in virtual shopping environment

- Objective:
  - For this project, we aim to predict the purchase propensity of an online shopper based on the engagement metrics on the website

EXPLORATORY DATA ANALYSIS - EDA

EXPLORATORY DATA ANALYSIS



DATA OVERVIEW

DATA :

- **Online Shoppers Purchasing Intention** UCI dataset
- The data is related to shoppers' behavior as they visit a website
- The dataset consists of 12,330 rows and 18 columns : 10 numerical and 8 categorical
- The rows represent the "session" activity
- The columns represent the dimensions related to the session
- Each session would belong to a different user in a 1-year period to avoid bias
- The 'Revenue' attribute can be used as the class label
- Majority of the dataset is a negative class
- The dataset doesn't have any missing value

DATA OVERVIEW

Data Snapshot :

Numerical (10) | Categorical (8)

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.00	0.10	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.05	0.14	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.02	0.05	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

1 — The number of different types of pages visited by the visitor in that session and total time spent in each of these page categories

2 — Metrics measured by "Google Analytics" for each page in the e-commerce site.

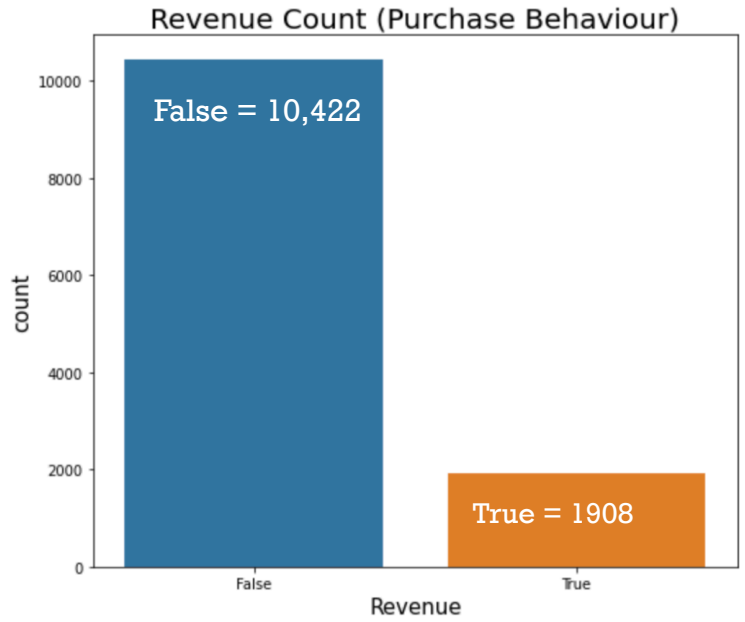
3 — Closeness of the site visiting time to a specific special day

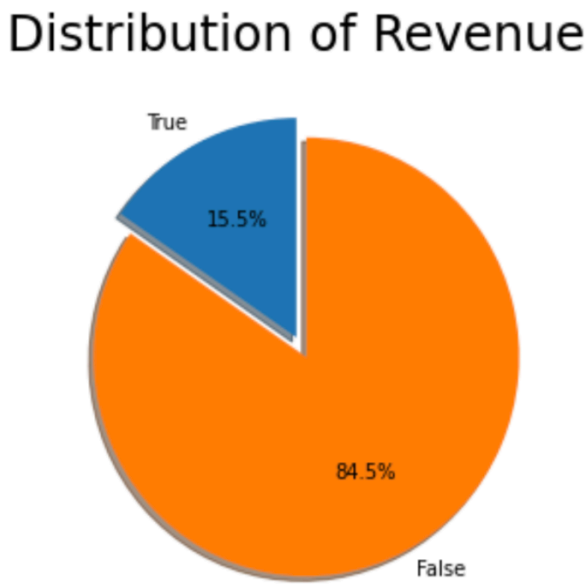
4 — operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

Revenue: Class Label (Target Variable):
has the customer made a purchase ?
('TRUE', 'FALSE')

UNIVARIATE ANALYSIS

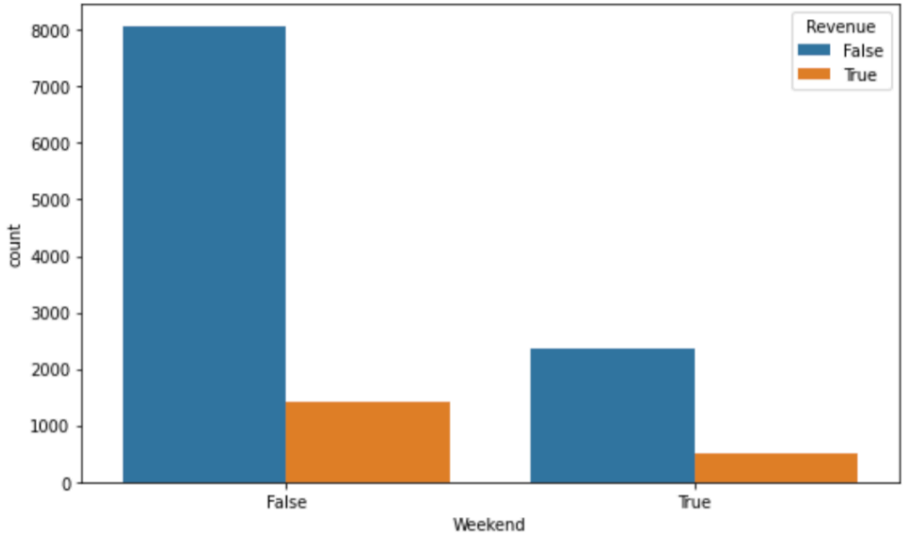
- Check the distribution of the target Variable :
  - As we can infer from the below visuals , the data is imbalanced





BIVARIATE ANALYSIS

- 76.7% of our visitors visit during weekdays (a five-day period)
  - with 14.9% probability of making purchase
- 23.3% of visitors on weekends, (a two-day period)
  - with a 17.4% probability of making purchase



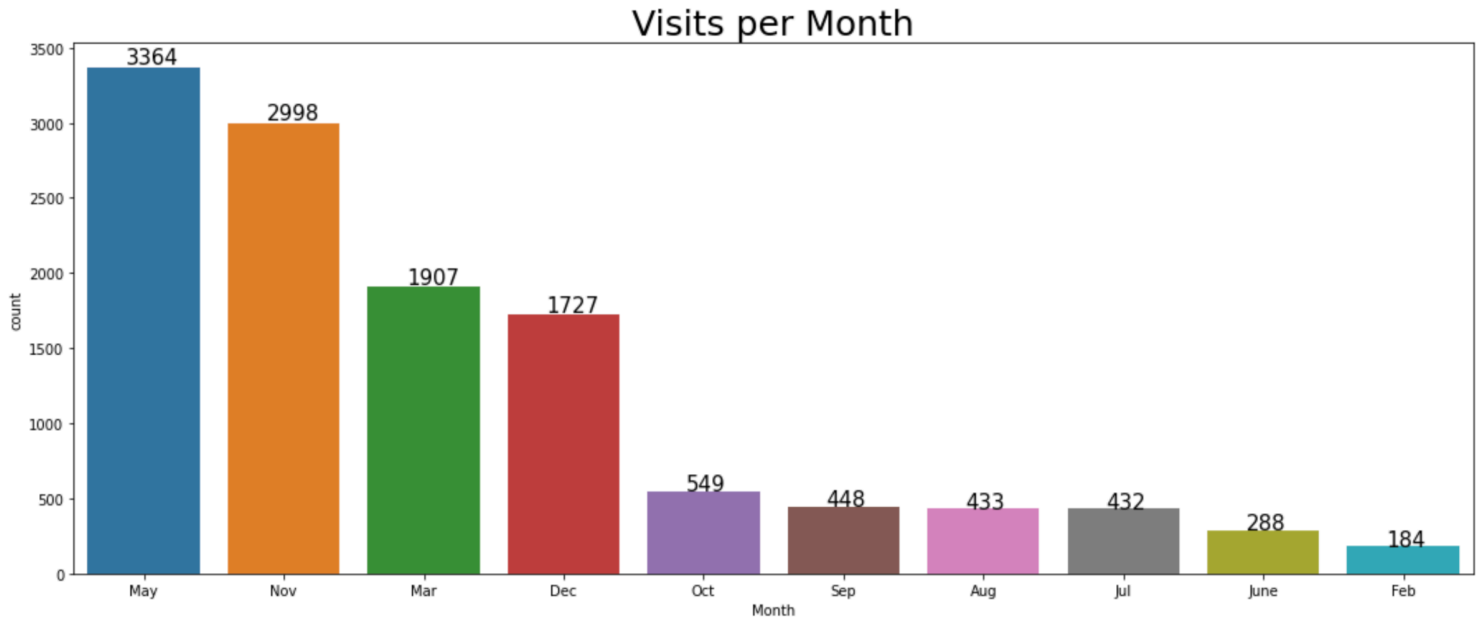
Revenue	False	True	total	visitor_percent	purchase prob
Weekend					
False	8053	1409	9462	0.767397	14.891144
True	2369	499	2868	0.232603	17.398884

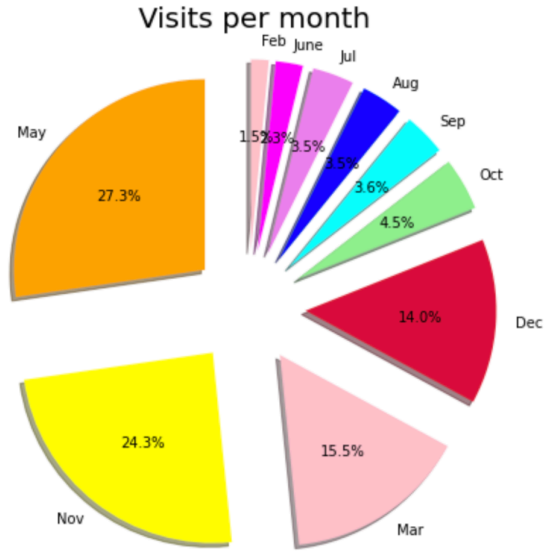
Although Revenue is higher during weekdays,

the percentage of making a purchase is higher during weekends

UNIVARIATE ANALYSIS

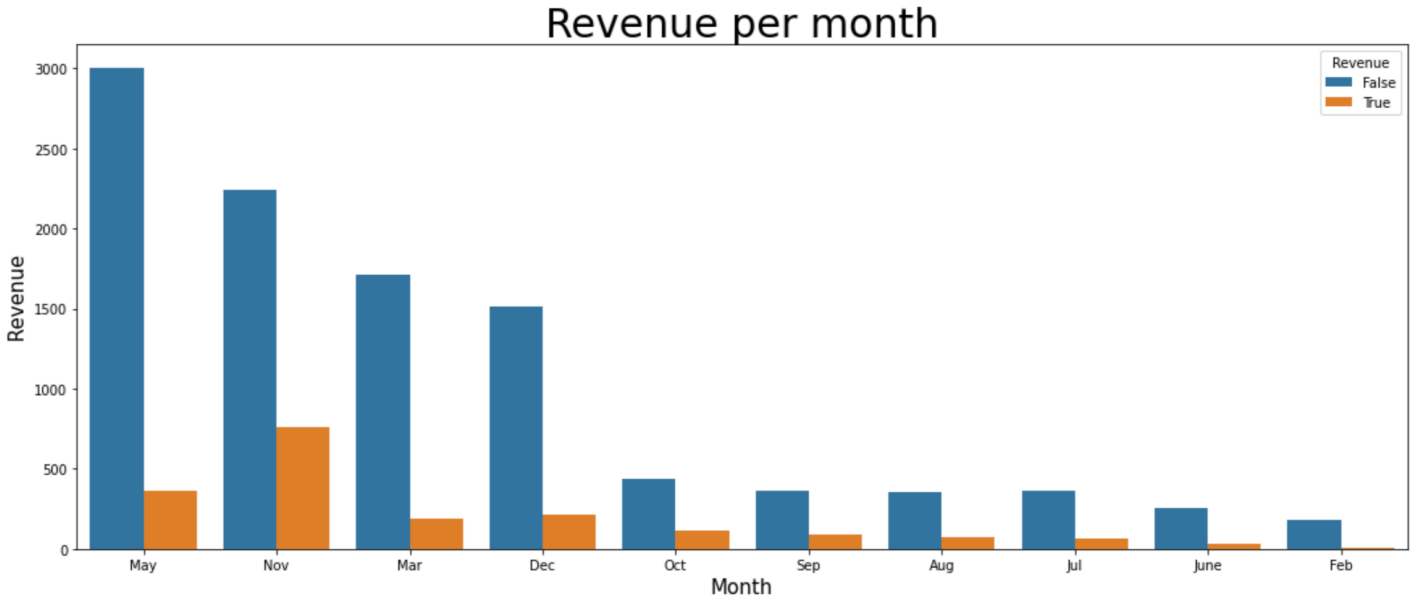
- Customers' visits per month :
  - highest : May with 3364 visits in the year with a (27.3%)
  - Lowest : Feb with 184 visits in the year with a (1.5%)





BIVARIATE ANALYSIS

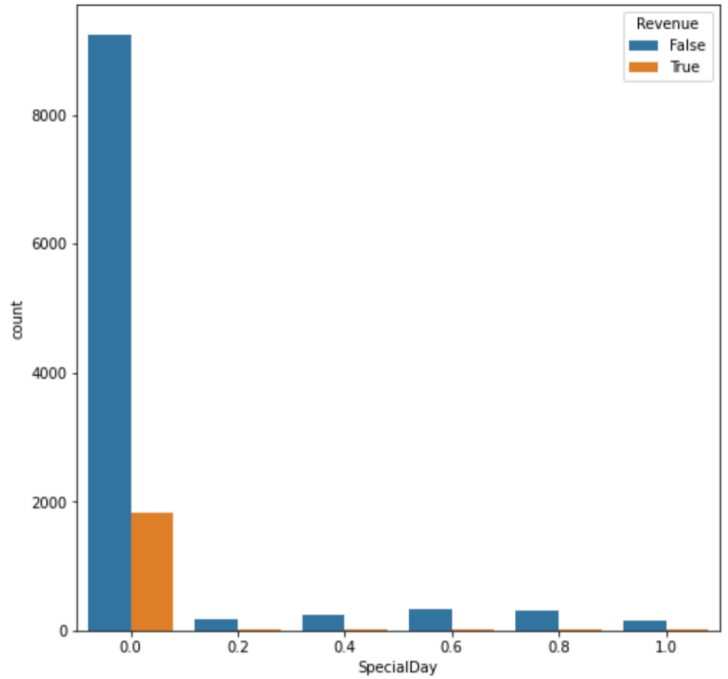
- Customers' Purchase behavior per month :
  - We notice high shopping rates in September, October, and November; and this might be caused by seasonality.



Revenue Month	False	True	total	visitor_percent	purchase percent
Aug	357	76	433	0.035118	17.551963
Dec	1511	216	1727	0.140065	12.507238
Feb	181	3	184	0.014923	1.630435
Jul	366	66	432	0.035036	15.277778
June	259	29	288	0.023358	10.069444
Mar	1715	192	1907	0.154663	10.068170
May	2999	365	3364	0.272830	10.850178
Nov	2238	760	2998	0.243147	25.350233
Oct	434	115	549	0.044526	20.947177
Sep	362	86	448	0.036334	19.196429

BIVARIATE ANALYSIS

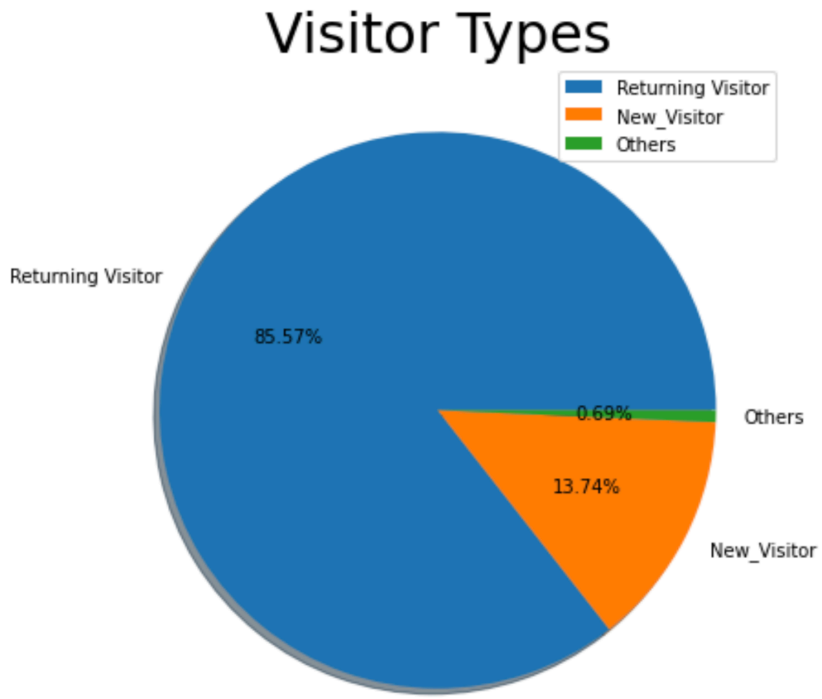
- Special Day VS Revenue:
  - Most purchases has been completed without being close to a special day



Revenue	False	True	total	visitor_percent	purchase percent
SpecialDay					
0.0	9248	1831	11079	0.898540	16.526762
0.2	164	14	178	0.014436	7.865169
0.4	230	13	243	0.019708	5.349794
0.6	322	29	351	0.028467	8.262108
0.8	314	11	325	0.026358	3.384615
1.0	144	10	154	0.012490	6.493506

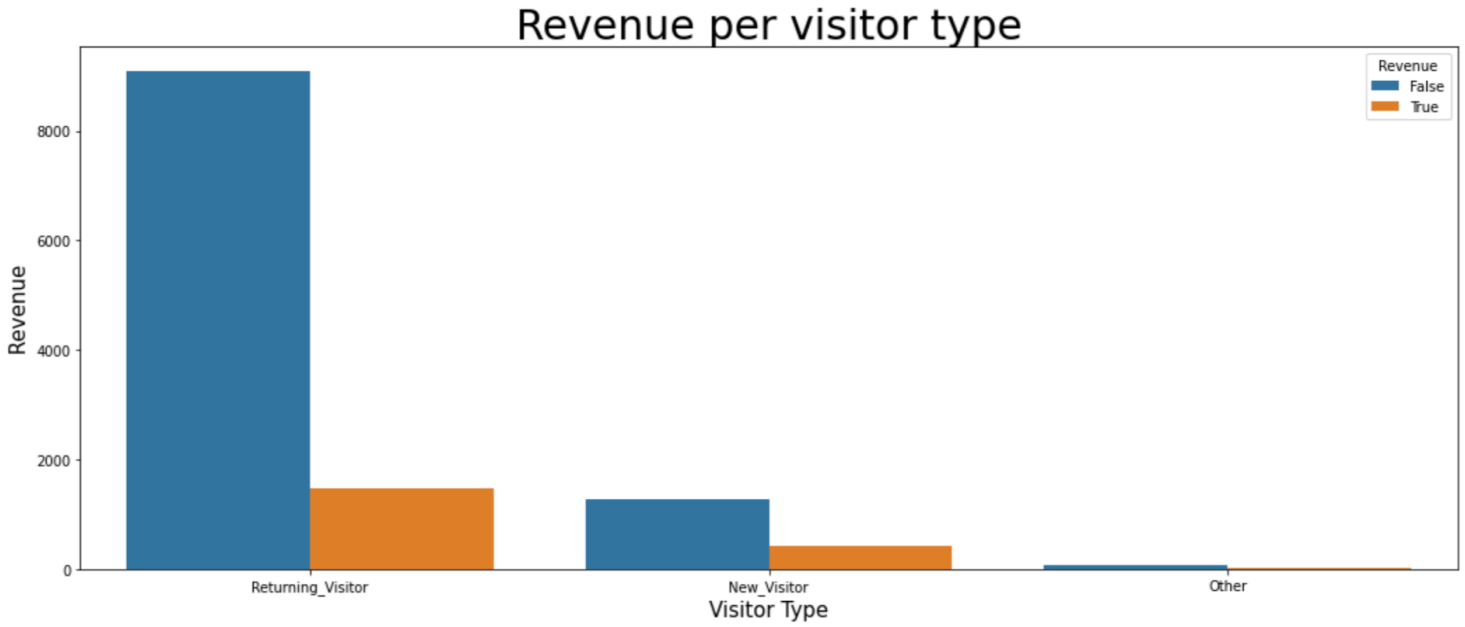
UNIVARIATE ANALYSIS

- Visitor Types :
  - Most visitors are returning visitors (85.6%) compared to (13.7%) new visitors



BIVARIATE ANALYSIS

- However,
  - New visitors have a higher probability of making a purchase with (24.9%) compared to only (13.9%) of returning visitors generating revenue.



VisitorType	Revenue False	True	total	visitor_percent	purchase percent
New_Visitor	1272	422	1694	0.137388	24.911452
Other	69	16	85	0.006894	18.823529
Returning_Visitor	9081	1470	10551	0.855718	13.932329

CORRELATION ANALYSIS

01 Exit rate is highly correlated with BounceRate

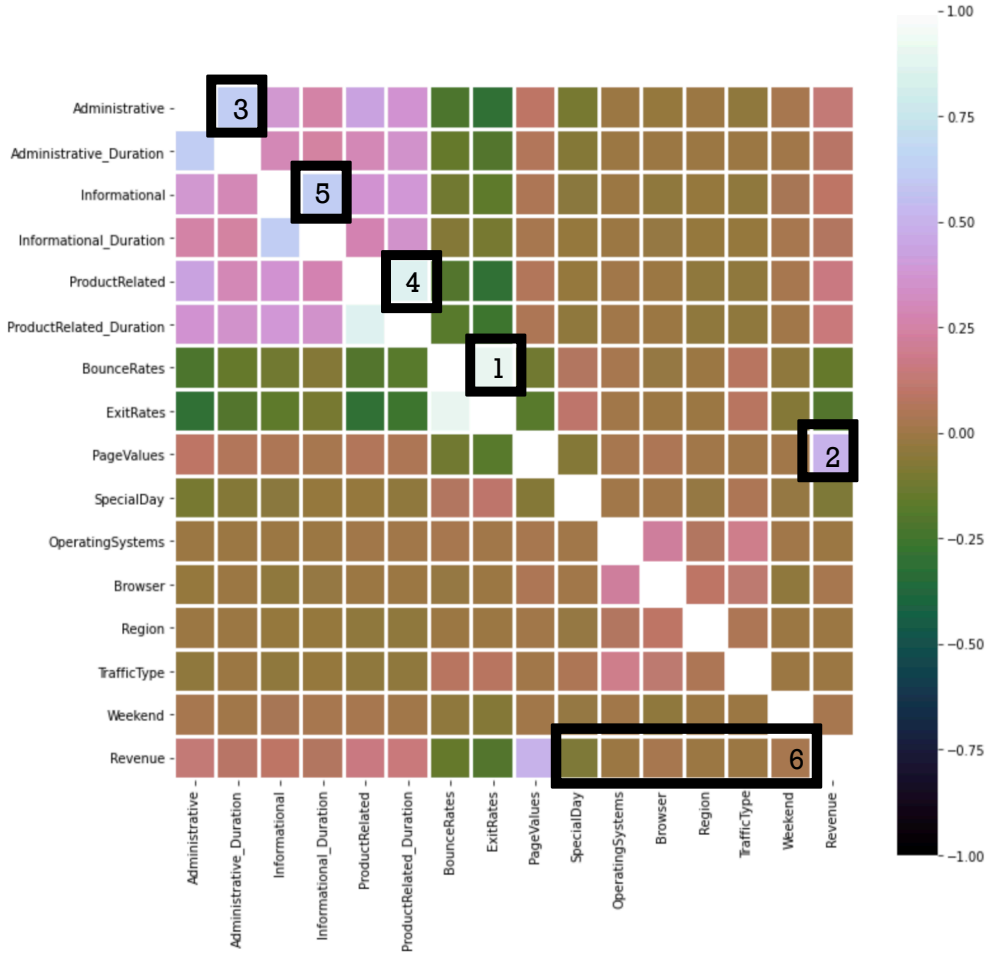
02 Page values and Revenue are highly correlated

03 Administrative and Administrative_Duration are highly correlated

04 ProductRelated & ProductRelated_Duration are highly correlated

05 Informational and Informational_Duration are highly correlated

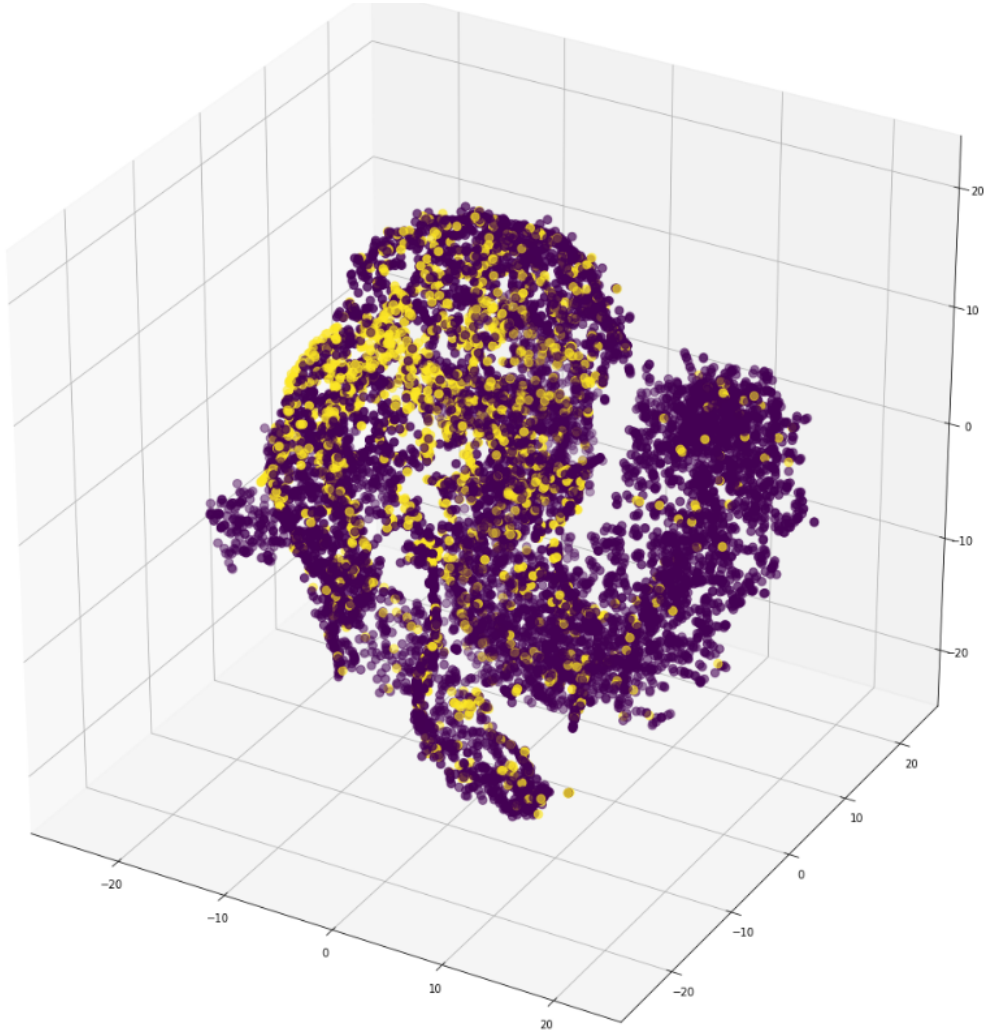
06 SpecialDay, Weekend, and TrafficType has unimportant correlation



HANDLING CATEGORICAL VARIABLES

- To ensure properly feeding our models with the required form.
  - Used **Onehot Encoding** to convert our categorical variables to dummy variables
  - Transformed non-numerical features (Weekend , Revenue) to numerical labels.
- After the encoding, our initial 18 features increased to 75

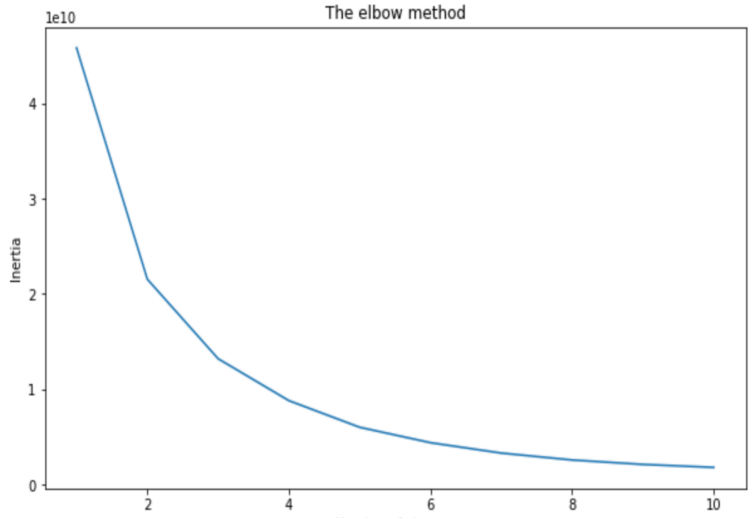
TSNE

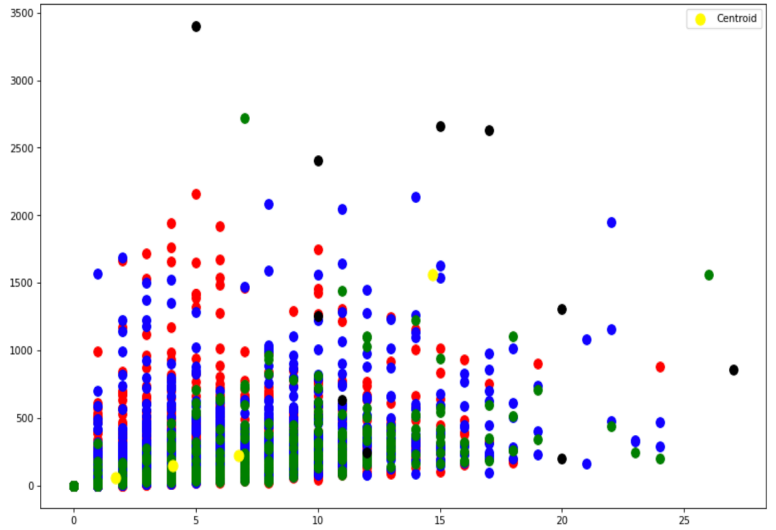


- Overview of dataset by TSNE
- The yellow points are customers who made purchase
- The purple points are customers who did not make purchase

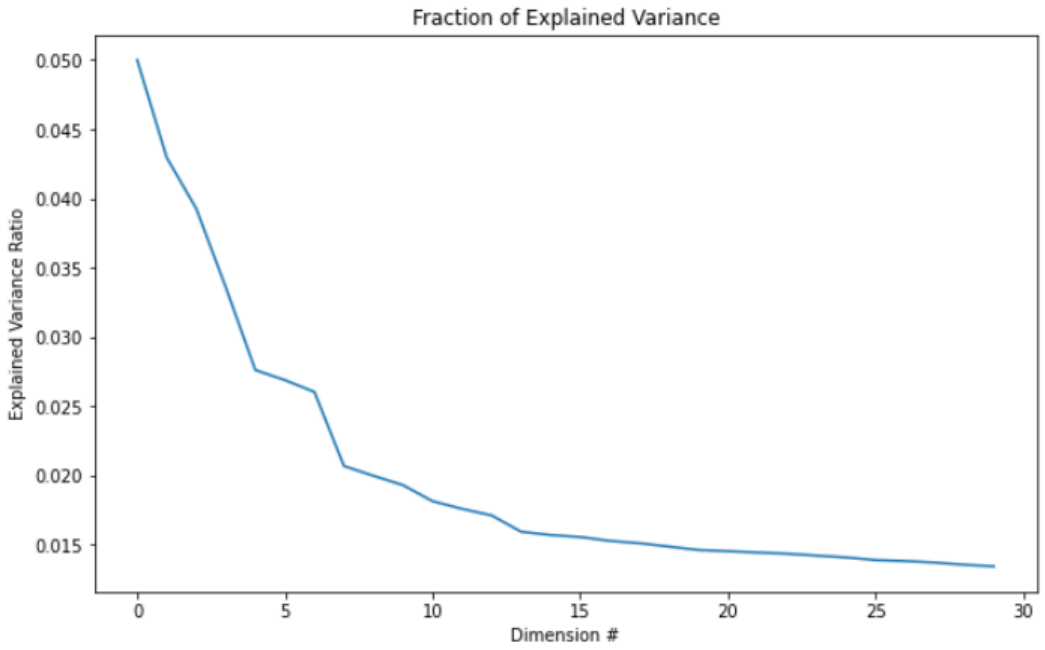
K-MEANS CLUSTERING – PRE PCA

- The k-means clustering on the original dataset has resulted in the following representation of the cluster labels
- Clusters seem to be cluttered and there isn't a clear distinction between the clusters
- Due to this, we opted for clustering post performing PCA

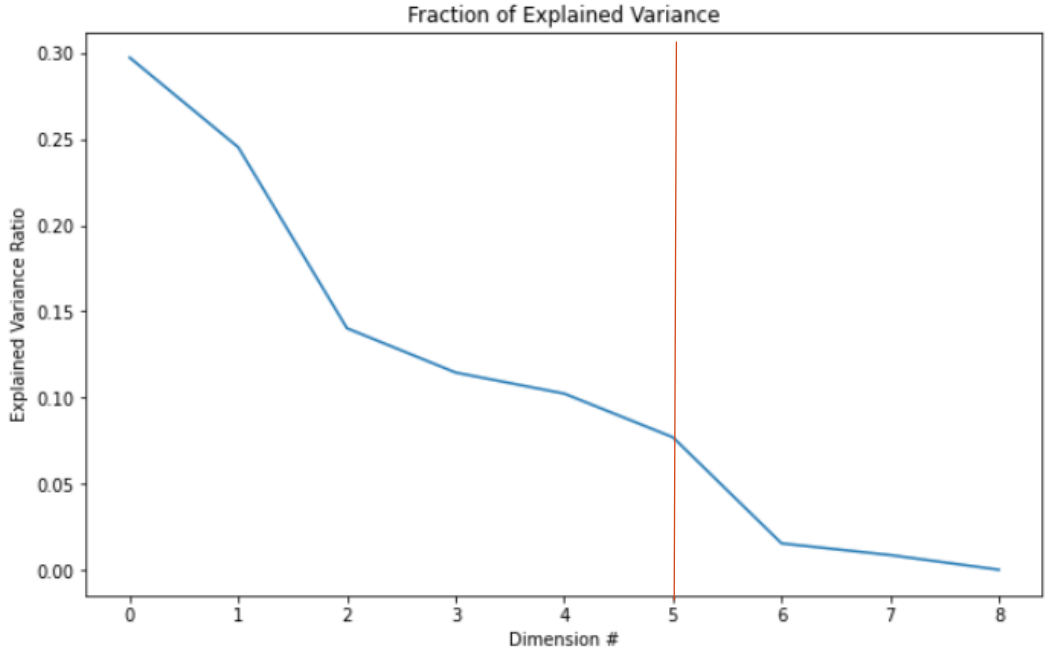




PCA



PCA with all the features

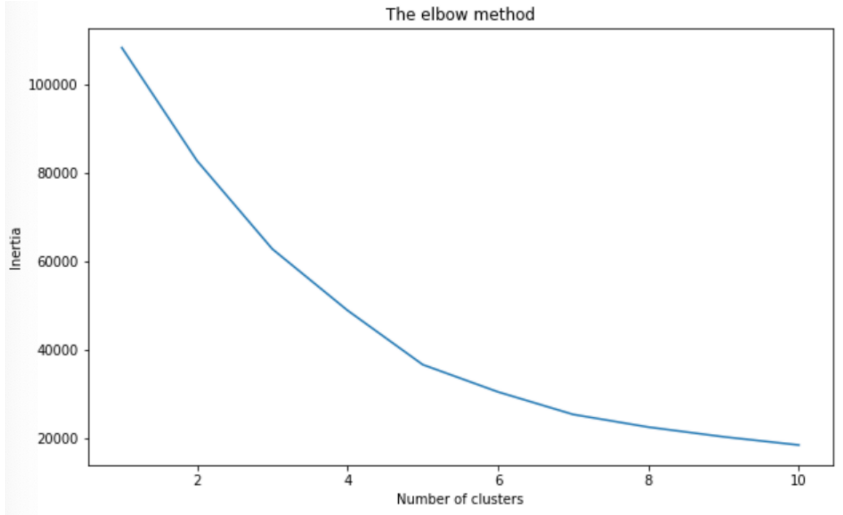


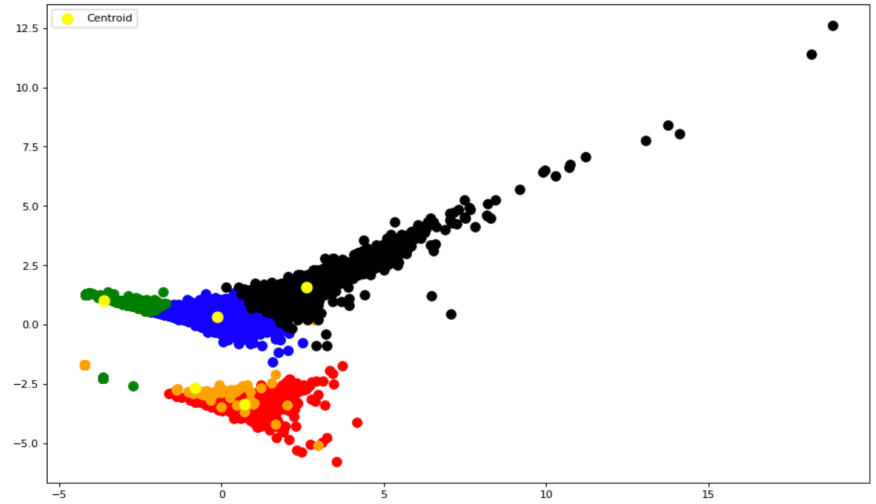
PCA with correlated features

- We will use first 6 components in second graph to do further analysis

K-MEANS CLUSTERING – POST PCA

- Based on the PCA results we've decided on considering the features that explain 95% of the variance
- Clustering of the PCA components has resulted in better segregation of clusters
- Based on the cluster features we've profiled clusters into
  - Direct Shoppers
  - Repeat Shoppers
  - Shallow Shoppers
  - Search/Deliberate Shoppers
  - Hedonic Shoppers





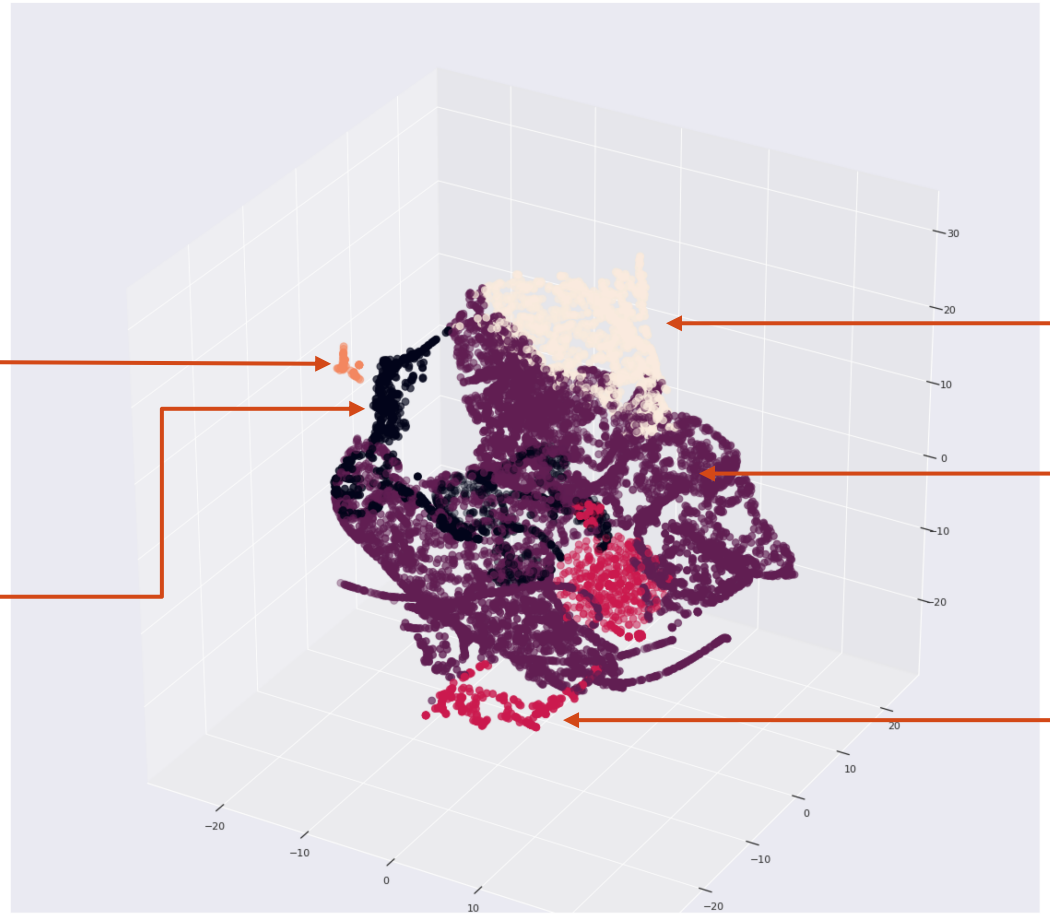
K-MEANS CLUSTERING - PROFILING

Cluster	1	2	3	4	5
PERSONA	DIRECT SHOPPERS	REPEAT SHOPPERS	SHALLOW SHOPPERS	HEDONIC SHOPPERS	SEARCH / DELIBERATE SHOPPERS
DIFFERENTIATING FACTORS	• Don't browse a lot about products, visit the website and make a purchase • Majority of these consumers are new consumers • Most of the purchases happen over the weekend	• Tend to purchase products over the special days like valentine's day, Christmas etc., • Majority of the consumers are return shoppers i.e., tend to purchase two or more times	• Shoppers don't have an intention to purchase a product • They leave the website after only 2 pageviews	• The shoppers view the products but don't tend to purchase any products. These consumers are termed to be hedonic shoppers • hedonic shopping is driven by a shopper's desire to experience satisfaction of shopping	• Research a lot before purchasing a product • The engagement with the website is very high when compared to other shoppers • Majority of the shoppers are repeat shoppers
CLUSTER DISTRIBUTION	13.50%	66.60%	8.22%	1.00%	11.00%

3D TSNE AFTER K-MEANS CLUSTERING - PROFILING

HEDONIC SHOPPERS

SHALLOW SHOPPERS



SEARCH / DELIBERATE SHOPPERS

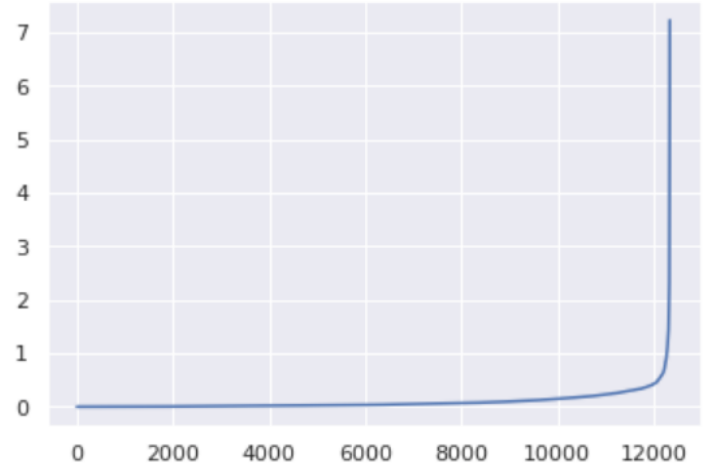
REPEAT SHOPPERS

DIRECT SHOPPERS

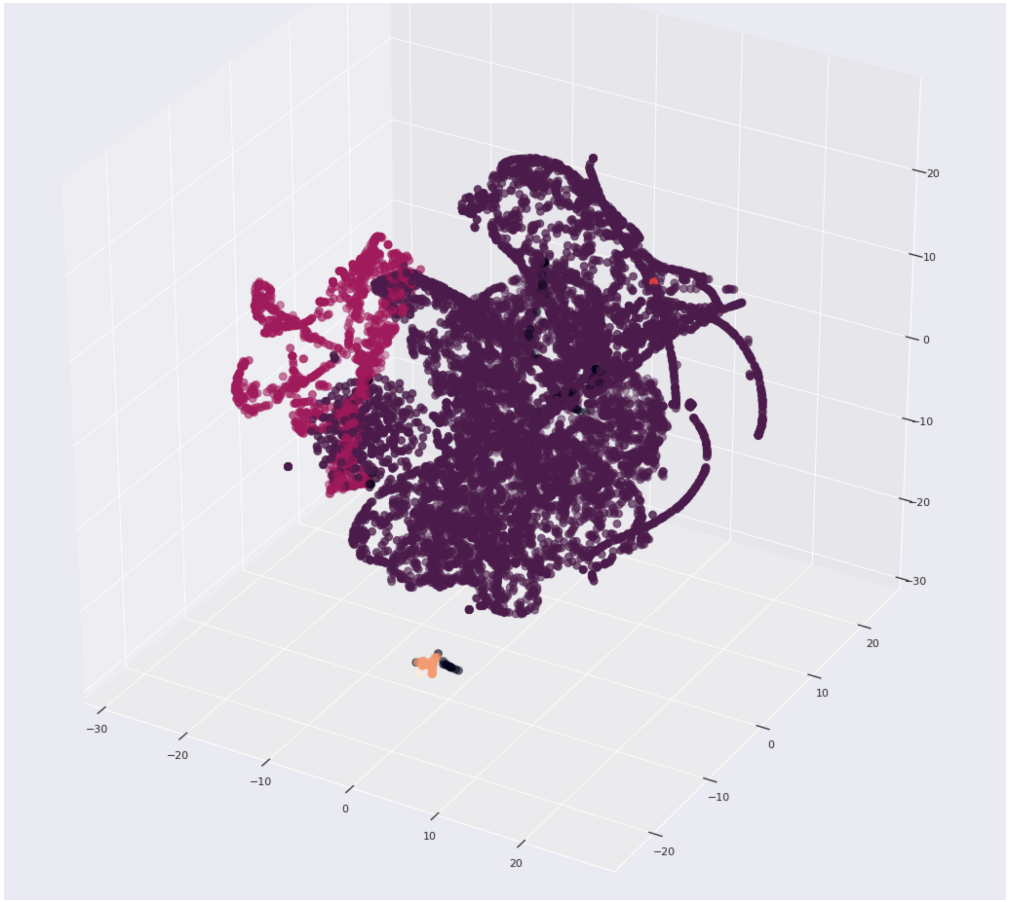
DBSCAN CLUSTERING

- Since DBSCAN clusters data based on the density of the data points. The shoppers that are close to each other. Following is the distribution of the DBSCAN clustering

Cluster	1	2	3	4	5	6
Shopper count	1.16%	84.79%	13.28%	0.23%	0.42%	0.13%



3D TSNE REPRESENTATION OF DBSCAN CLUSTERS



MODELING & VALIDATION

MODEL IMPLEMENTATION

- Logistic Modeling
  - Logistic Regression
  - Logistic RegressionCV
  - SGD Logistic Regression
- SVM Classifier
- Random Forest
- Gradient Boosting
- Extreme Gradient Boosting

LOGISTIC REGRESSION

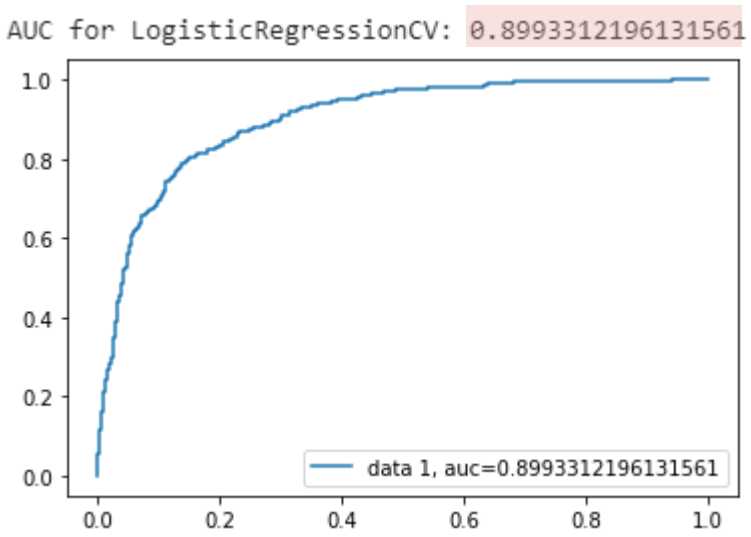
- The outcome or target variable of logistic regression is dichotomous in nature.
- Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.
- LogisticRegressionCV has built-in cross-validation capabilities to automatically select the best hyper-parameters

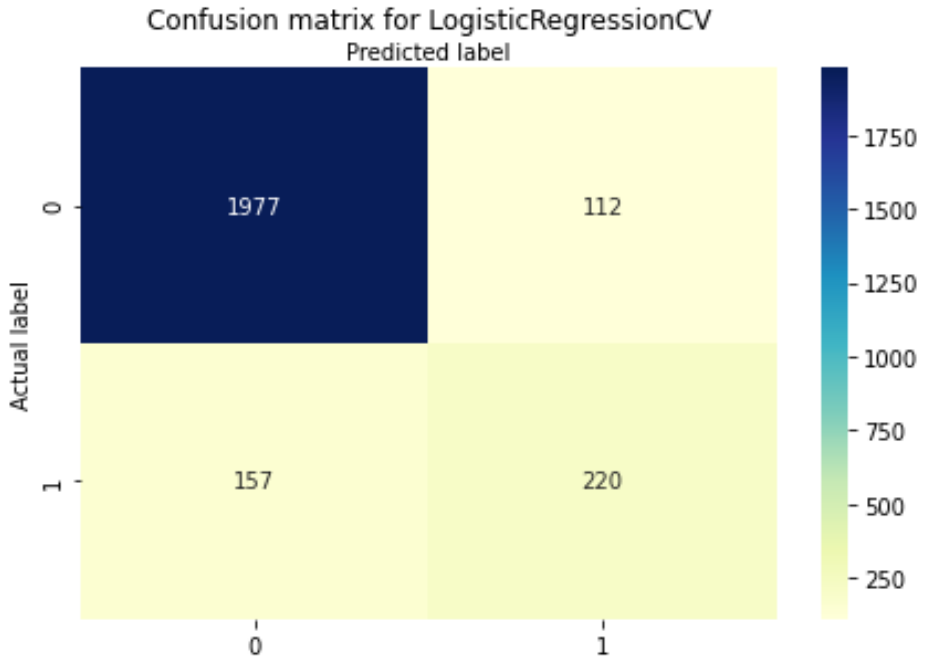
	Initial Model	**Smote Balanced Model**	**PCA Reduced Model**
LogisticRegression	88.74%	89.30%	89.44%
LogisticRegressionCV	88.80%	**89.93%**	89.45%
SGDClassifier(log loss)	69.10%	72.17%	88.86%

Metrics: AUC

LOGISTIC REGRESSION CV – SMOTE BALANCED

	Initial Model	Smote Balanced Model	PCA Reduced Model
Test Accuracy	87.67%	89.09%	88.97%
Train Accuracy	88.72%	83.89%	88.43%





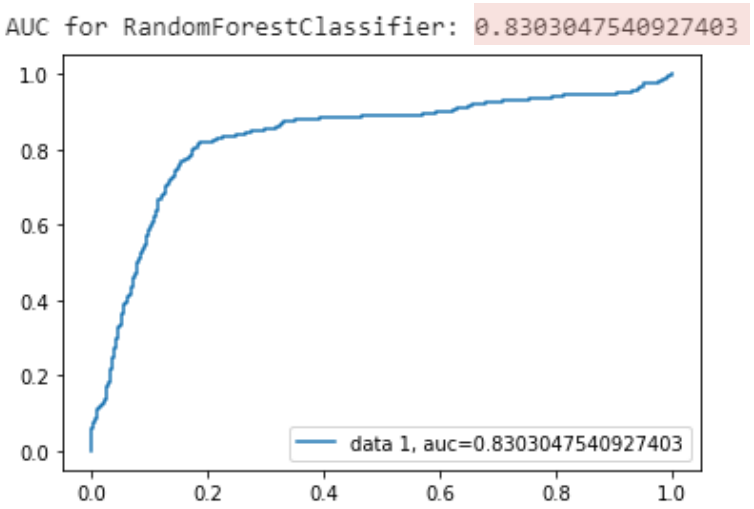
SUPPORT VECTOR CLASSIFIER

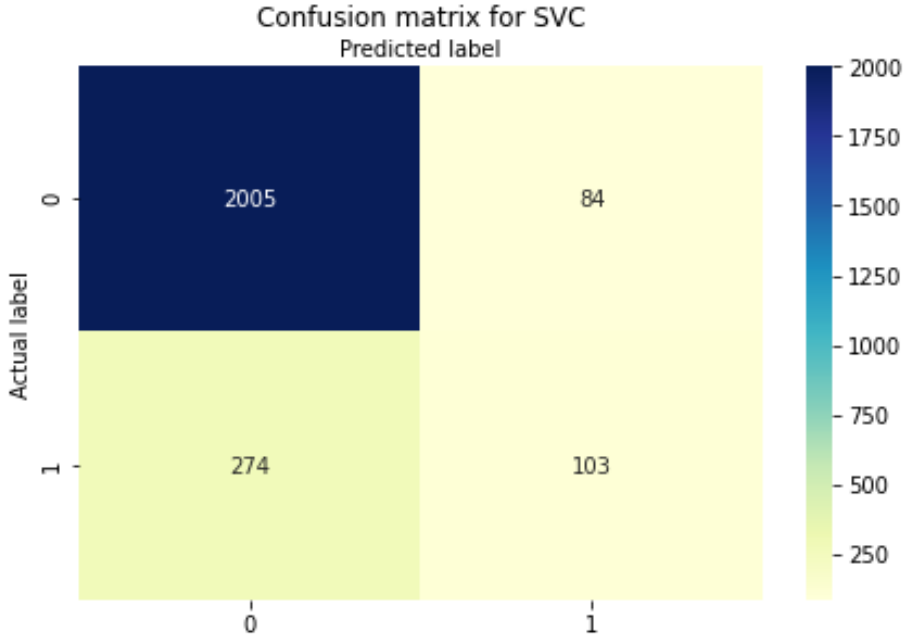
- Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.
- Kernel choice: rbf (radial basis function kernel)
- Does not directly provide probability estimates.

	Initial Model	Smote Balanced Model	PCA Reduced Model
AUC	80.41%	**83.03%**	82.40%
Precision	100.00%	55.08%	72.96%
Recall	1.86%	27.32%	48.43%

SUPPORT VECTOR CLASSIFIER – SMOTE BALANCED

	Initial Model	Smote Balanced Model	PCA Reduced Model
Test Accuracy	85.00%	85.28%	90.11%
Train Accuracy	84.71%	82.57%	89.66%
CV Accuracy	84.71%	84.71%	84.50%





RANDOM FOREST CLASSIFIER

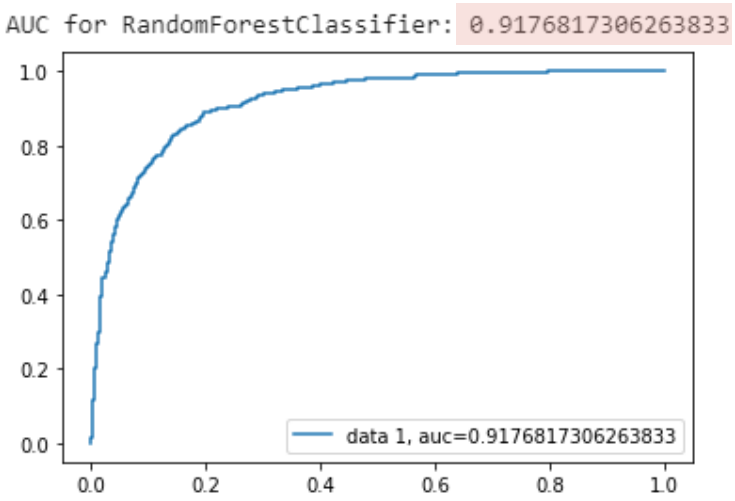
- Random Forest is a tree-based algorithm. It chooses features randomly during the training process.
- Utilize GridSearchCV to find the best params:
  - max_depth=12 and n_estimators=300

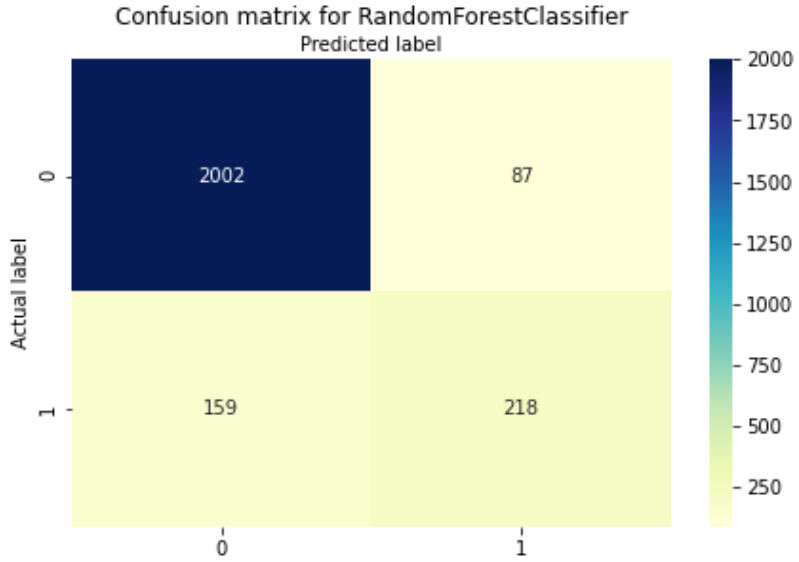
	Initial Model	Smote Balanced Model	PCA Reduced Model
AUC	91.51%	**91.77%**	90.53%
Precision	78.44%	68.57%	67.60%
Recall	45.35%	57.29%	55.27%

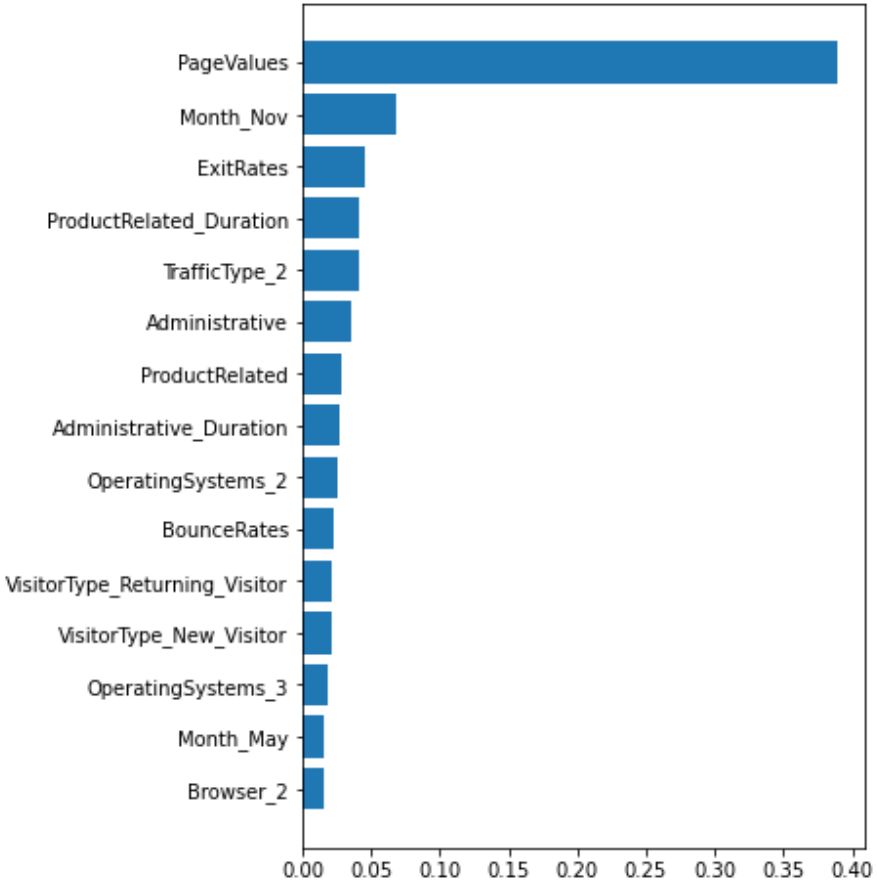
	max_depth	n_estimators	mean_test_score
18	12	300	0.920877
19	12	400	0.920077
17	12	200	0.919997
20	12	500	0.919917
16	12	100	0.918957
11	10	100	0.917437
13	10	300	0.916157
14	10	400	0.916077
15	10	500	0.916077
12	10	200	0.915357
6	7	100	0.910636
8	7	300	0.909677
9	7	400	0.909436
7	7	200	0.908637
10	7	500	0.908156
2	5	200	0.898076
5	5	500	0.896716
3	5	300	0.896636
1	5	100	0.895836
4	5	400	0.895355

RANDOM FOREST CLASSIFIER – SMOTE BALANCED

	Initial Model	Smote Balanced Model	PCA Reduced Model
Test Accuracy	89.74%	90.24%	89.86%
Train Accuracy	95.60%	96.33%	94.47%
CV Accuracy	90.83%	90.79%	89.91%







GRADIENT BOOSTING

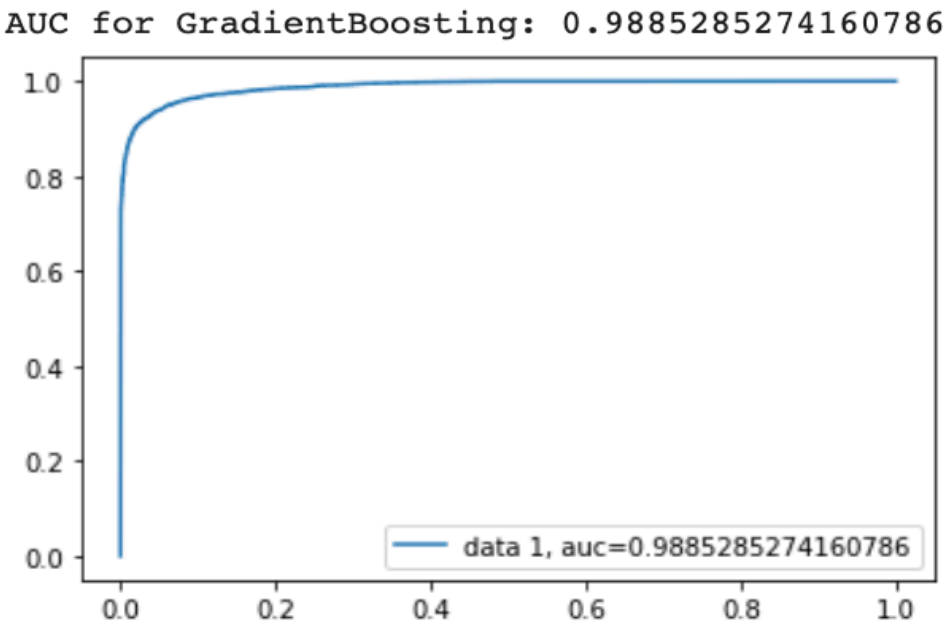
- Gradient Boosting is a boosting algorithm that tries to identify the optimal function that best approximates the data. The model creates a strong learner based on an ensemble of weak learners
- Initially trained the model on the original dataset with parameters
  - max_depth =4 , estimators = 200
- Further improvements were possible by balancing the target variable using SMOTE

	Initial Model	Smote Balanced Model
Train Accuracy	94.77%	95.41%
Test Accuracy	90.55%	90.26%
CV Accuracy	90.1%	92.0%

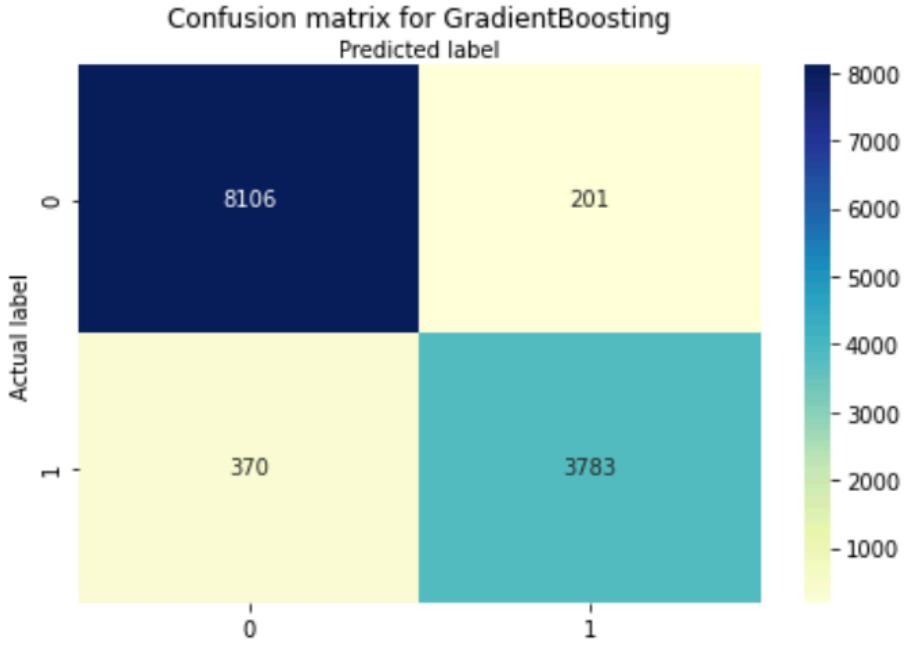
	Initial Model	Smote Balanced Model
Precision	88.88%	94.95%
Recall	76.49%	91.09%
AUC	97.33%	98.88%

GRADIENT BOOSTING

ROC plot



Confusion Matrix



EXTREME GRADIENT BOOSTING - XGBOOST

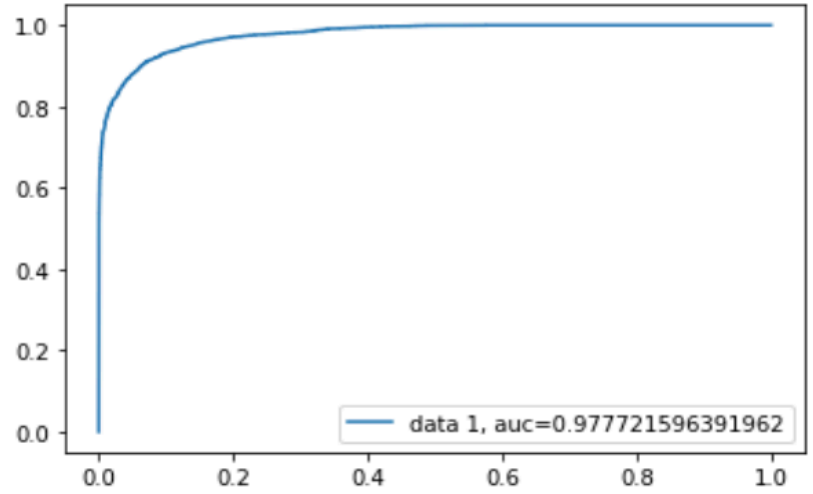
- A specific implementation of gradient boosting model. The most important differentiating factor is computation of second order partial derivative of the loss function
  - This provides more information about the direction of the gradient and how to get to the minimum of our loss function
- Trained the model with default parameters
- Further improvements were possible by balancing the target variable using SMOTE.

	Initial Model	Smote Balanced Model
Train Accuracy	91.70%	92.61%
Test Accuracy	90.79%	90.51%
CV Accuracy	90.1%	91.00%

	Initial Model	Smote Balanced Model
Precision	77.92%	89.68%
Recall	66.21%	87.96%
AUC	94.77%	97.77%

EXTREME GRADIENT BOOSTING - XGBOOST

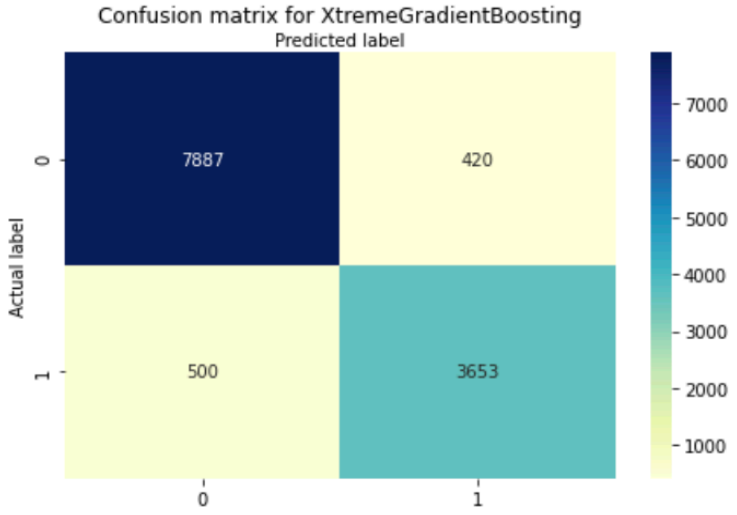
```
AUC for XtremeGradientBoosting: 0.977721596391962
```

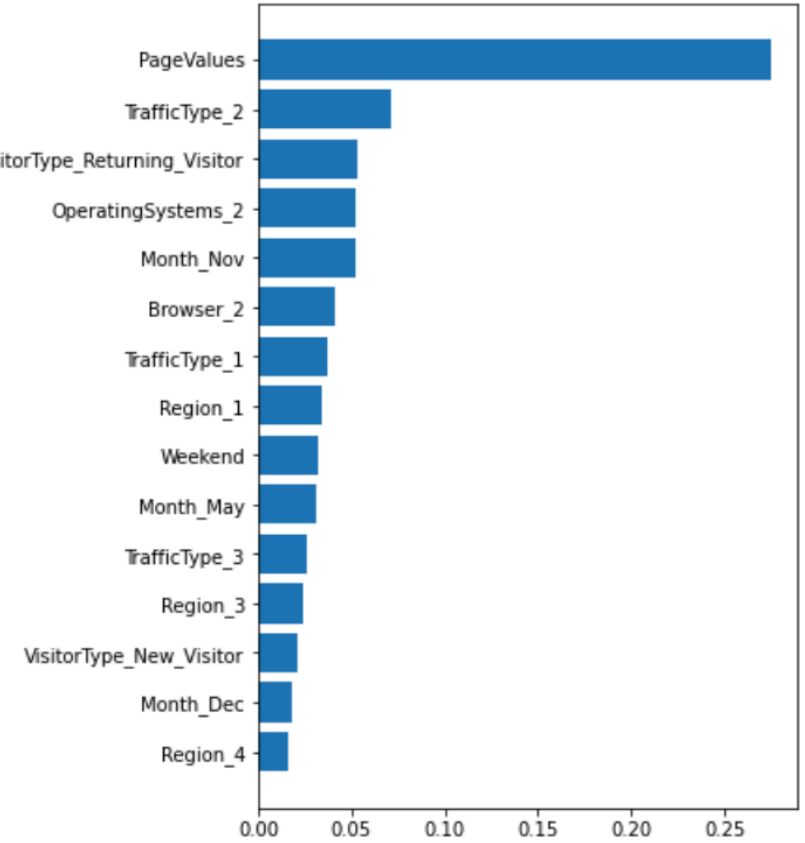


ROC plot

Confusion Matrix

```
Text(0.5, 257.44, 'Predicted label')
```





MODEL COMPARISON

Logistic Regression CV

AUC: 89.93%
Recall: 58.36%

- Doesn't require high computation power, easy to implement, easily interpretable.
- Provides a probability score for observations

Support Vector Classifier

AUC: 83.03%
Recall: 27.32%

- Effective in high dimensional spaces.
- Doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

Random Forest

AUC: 91.77%
Recall: 57.29%

- Effective in high dimensional dataset
- Can handle thousands of input variables without variable deletion

Gradient Boosting

AUC: 98.88%
Recall: 91.09%

- Lots of flexibility
- Often provides predictive accuracy that cannot be beat

Extreme Gradient Boosting

- AUC: 97.77%
- Recall: 87.98%

- Highly scalable/ parallelizable
- Quick to execute
- Regularization

CONCLUSION

CONCLUSION

Limitations :

- Lack of extensive engagement metrics and demographics constrained the scope of profiling the shopper personas
- Lack of available information for few categorical variables has constrained our understanding of the features
- Currently the web interface is designed to handle a single instance of data

Improvements and Future work :

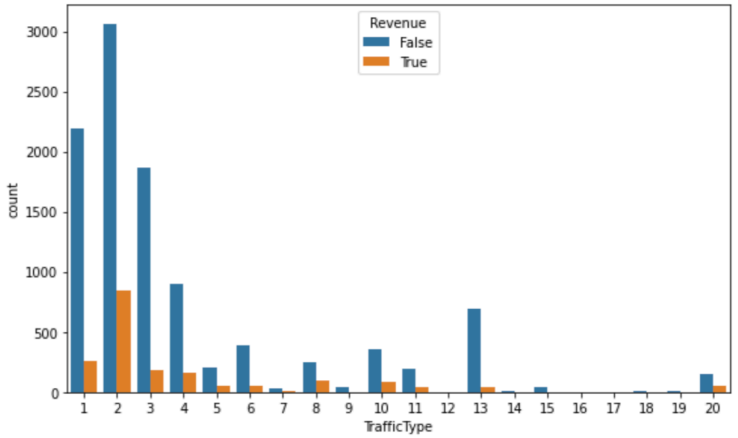
- Build an estimator to predict the probability of making a purchase instead of just classifying whether a consumer will purchase or not
- Using the probability scores build a recommendation engine to target the consumers using personalized emails
- The Web Interface can be improved to handle multiple to observations instead of just one
- Deploying web application on cloud

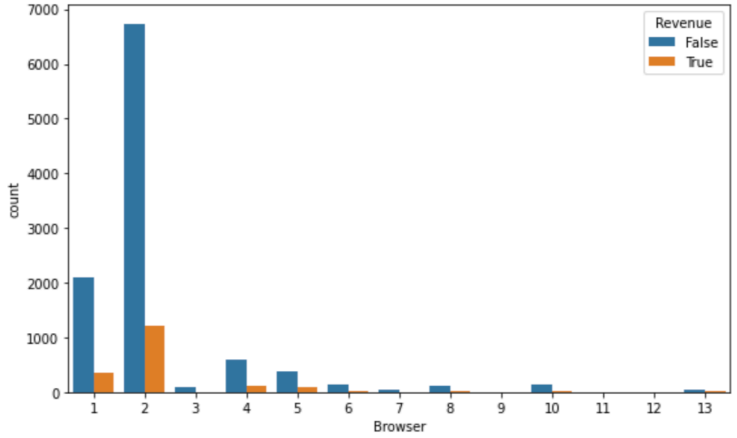
THANK YOU ☺

APPENDIX

BIVARIATE ANALYSIS

Lots of variation in revenue with the type of traffic the website is getting.

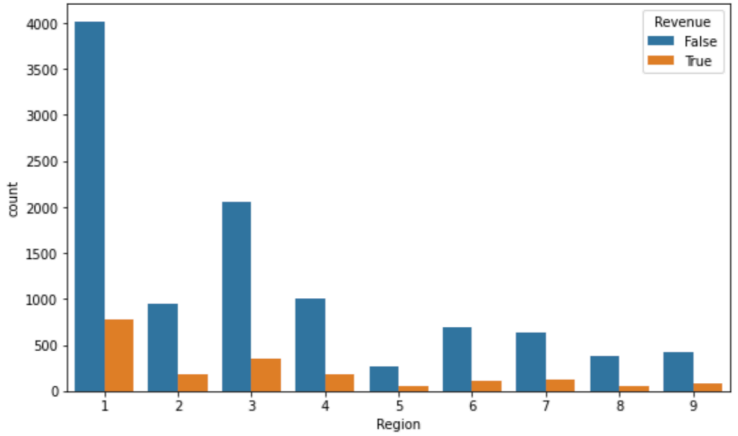


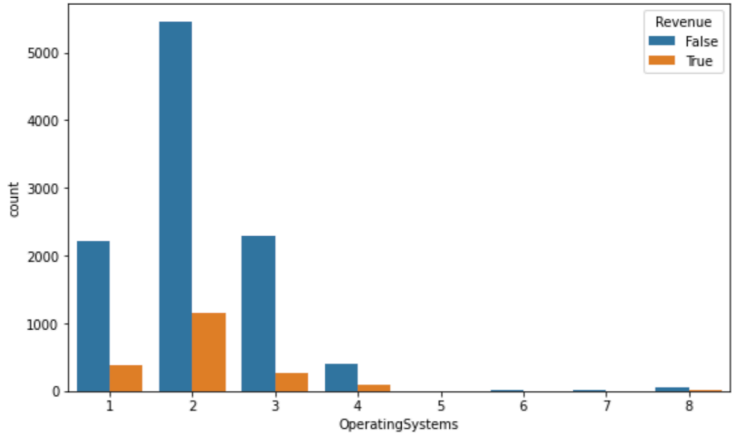


Browser 12 and 13 show high conversions

Every other browser preforms almost similarly

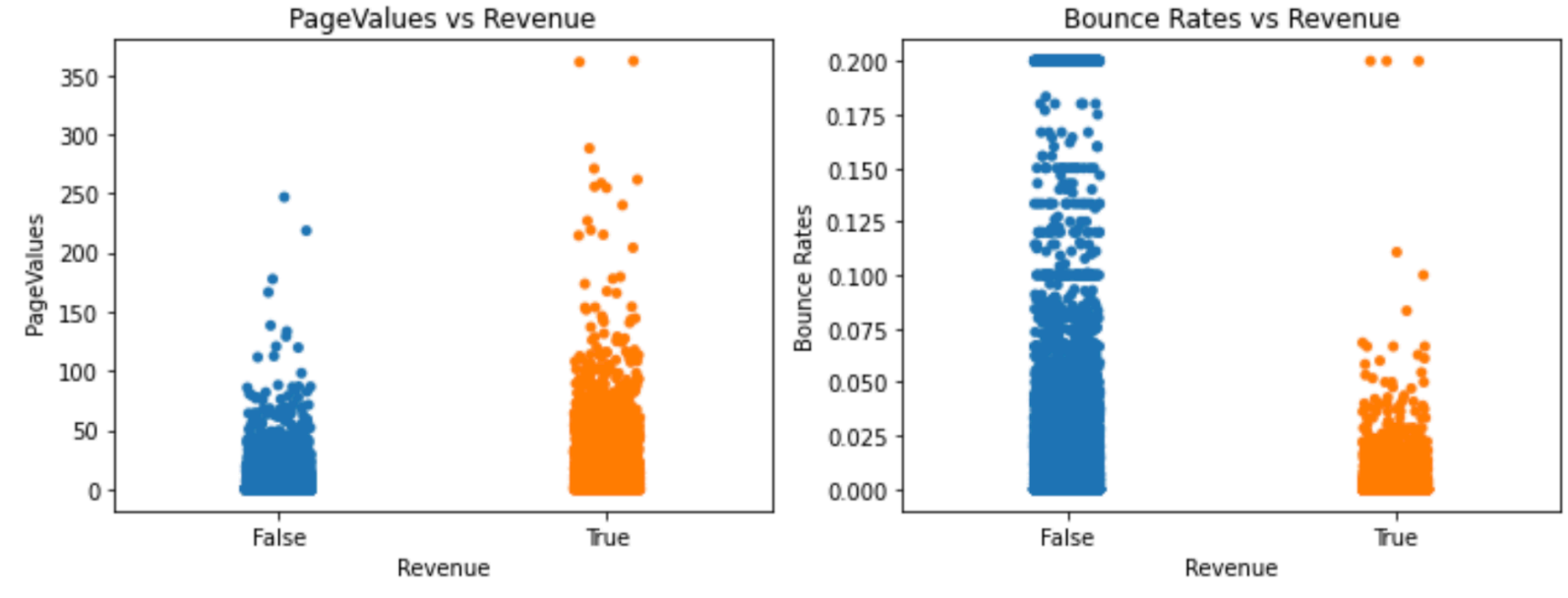
Very little variation by region

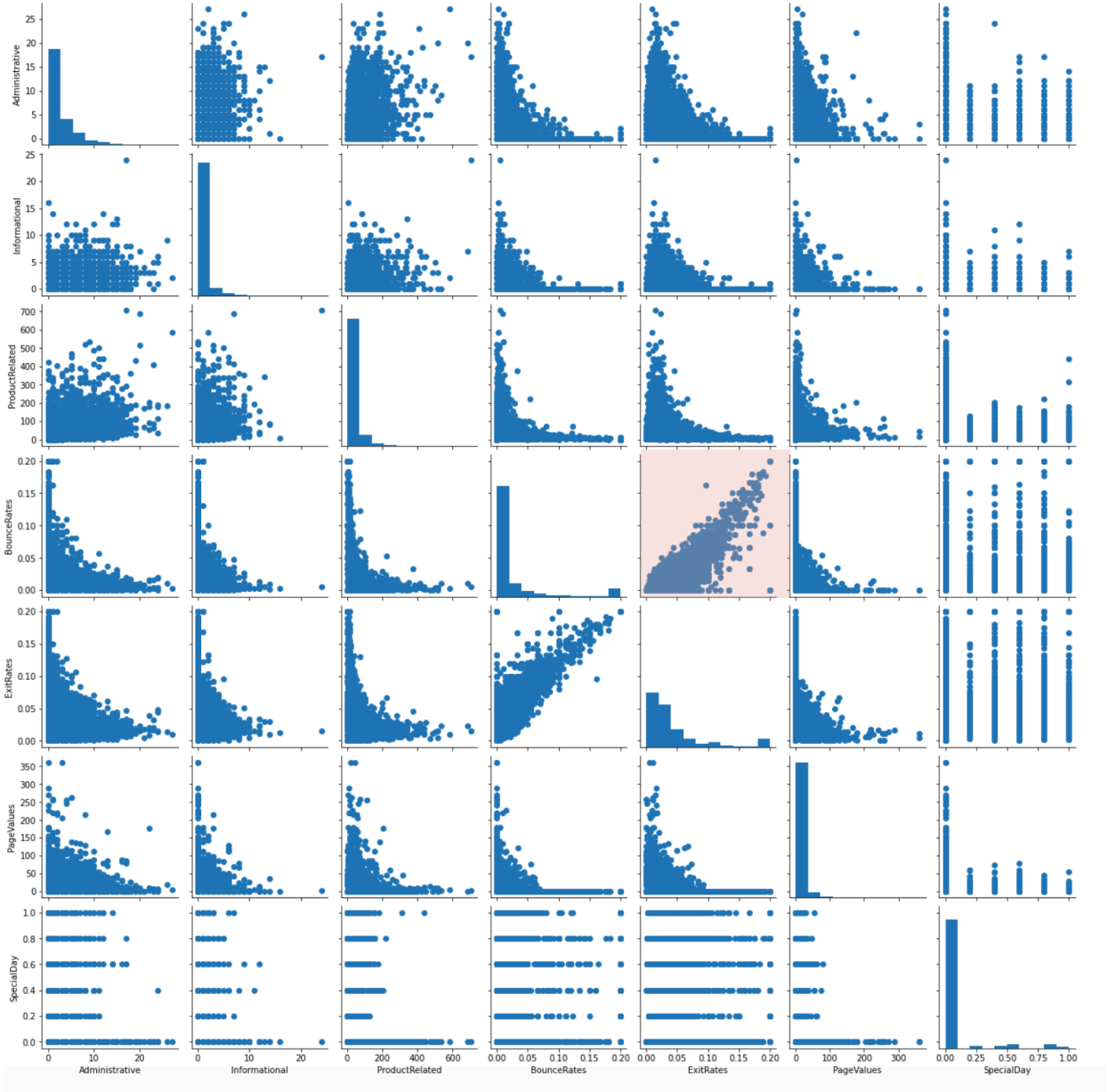




OS type 8 stands out although it doesn't comprise the majority of customers

BIVARIATE ANALYSIS




Administrative
Informational
ProductRelated
BounceRates
ExitRates
PageValues
SpecialDay