Name – Dinesh Sonawane

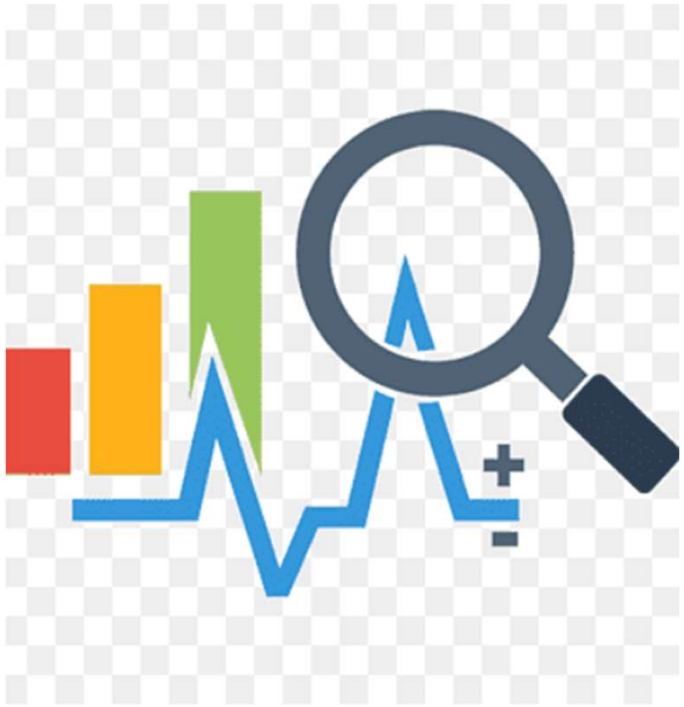
Project – Online Shoppers Intentions

Date – May'23

Project Details

Project Aim:

Online retailing company is trying to find which online shopper will generate revenue by his/her online shoppers' activity on their site



Inputs Provided:

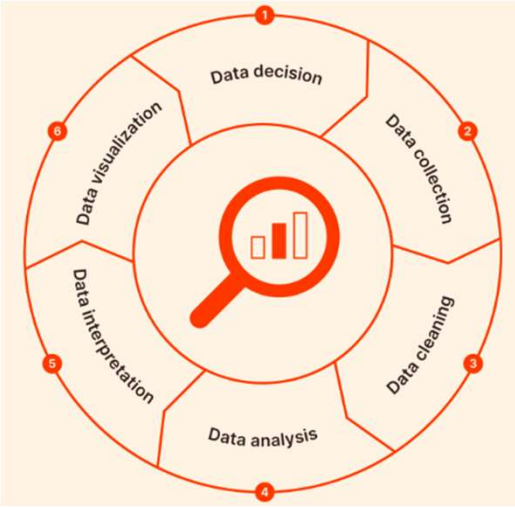
- ➢ Expectations of data analysis (problem statement)
- ➢ Dataset (.csv file) containing the online shoppers' activity details
- ➢ Interpretation of features included in the dataset

Project Expectations:

- ❑ Insights on factors impacting revenue generation
- ❑ Recommendations or suggestions helping revenue generation
- ❑ Prediction of visitors' conversion

High Level Approach

- Data information
  - No. of rows and columns
  - Type of columns (categorical or numerical)
  - Identify and treat the missing data
  - Identify and treat the duplicate data
- Deep dive of data
  - Monthly number of visitors
  - Special days and, visitors count and conversion
  - Variation of regional revenue
  - Monthly number of visitors and associated conversions
- Data preprocessing for ML models
  - Categorical to numerical conversion
  - Outlier identification and treatment
  - Multicollinearity check and feature engineering
- Selection and building of ML models
  - Finalization of ML model (Parameter Hyper tuning)
  - Validation of ML model
  - Identification of important features impacting prediction
- Concluding Summary



Data Information

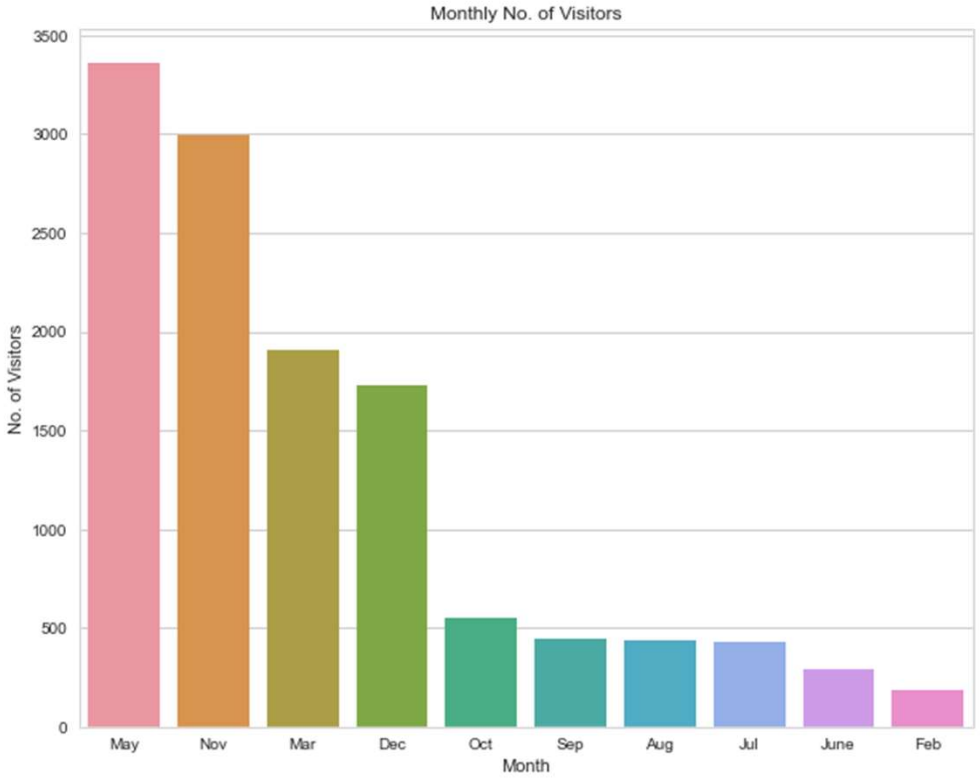
Aim - Which online shopper will generate revenue

- Data information
  - No. of rows and columns
    - Total 18 columns with 12330 entries
  - Type of columns (categorical or numerical)
    - Except Month, weekend and revenue columns, all other are numerical columns
  - Identify and treat the missing data
    - No missing values
  - Identify and treat the duplicate data
    - 125 duplicate entries located
      - There is a possibility that multiple users visiting the same page at same time having same properties such as bounce rates, exit rates etc.
      - We should not neglect such users having same behavior

Deep Dive of Data

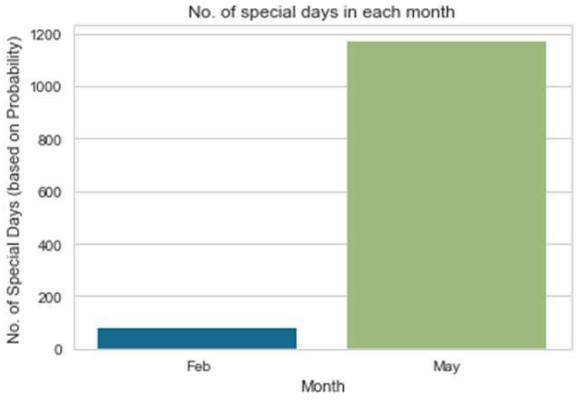
Aim - Which online shopper will generate revenue

- Monthly number of visitors



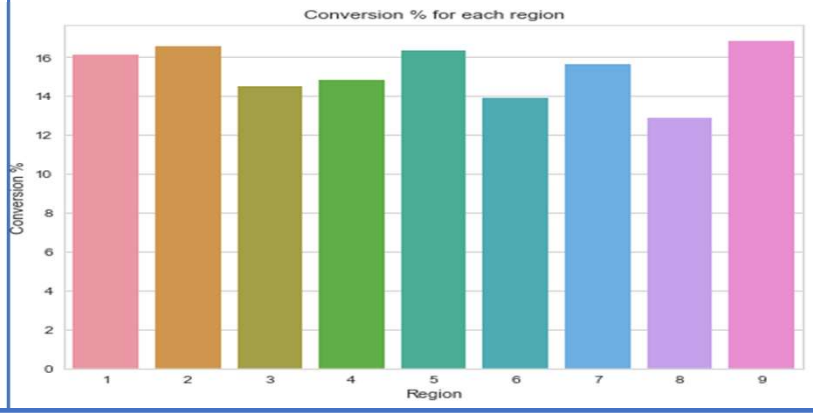
- May has max. no. of visitors followed by Nov and Mar
- Jun to Oct, no. of visitors stayed the same

- Look for relationship between Special days and, visitors count and conversion



Only May and Feb months have probability of having special days, however no. of visitors is min. in Feb, which indicates that special days are not attracting visitors.

- Variation of regional revenue

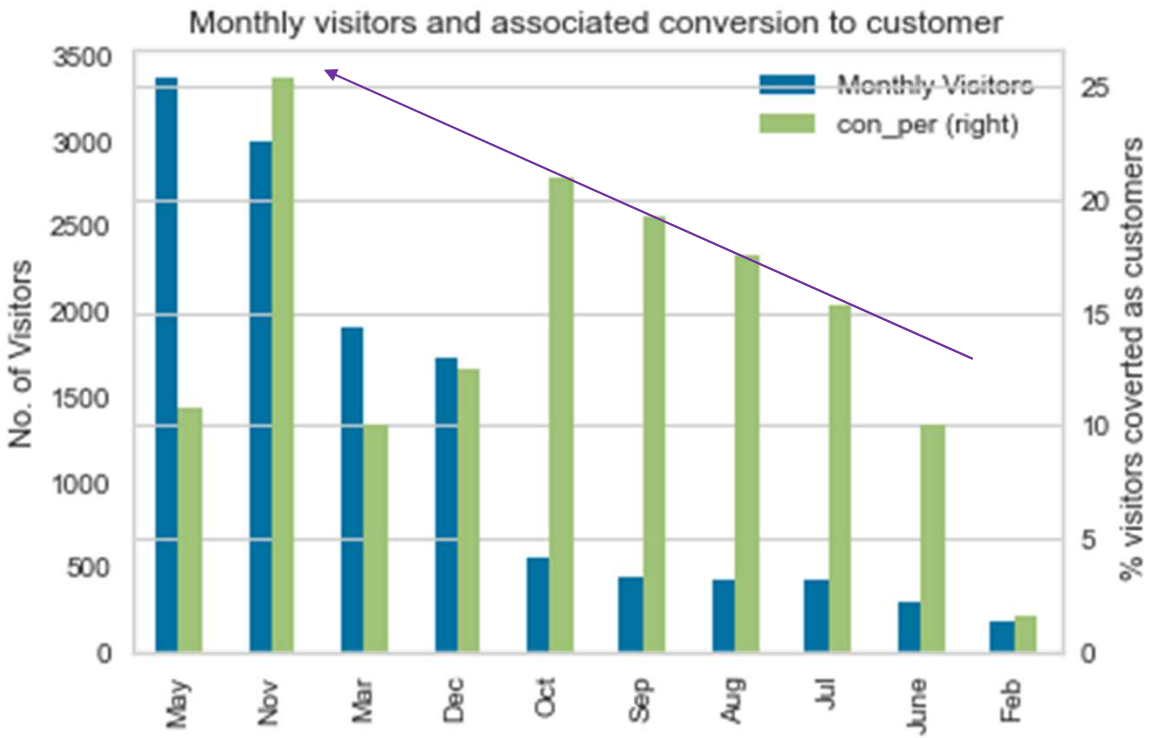


Region of visitors does not impact the conversion

Deep Dive of Data

Aim - Which online shopper will generate revenue

- Monthly number of visitors and associated conversions



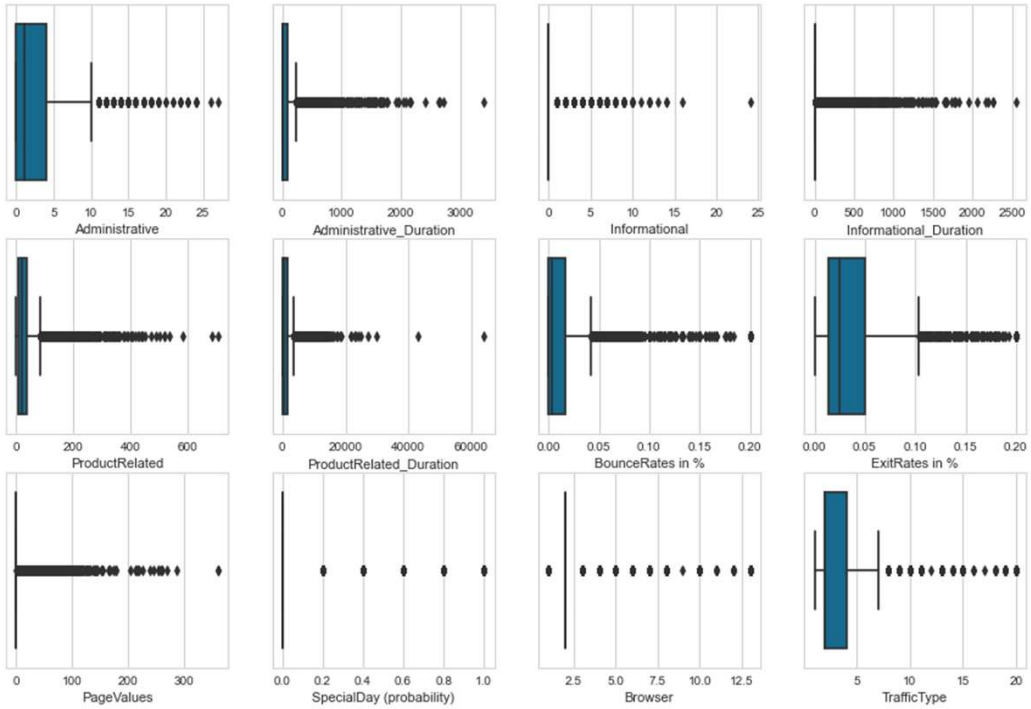
Key Note –
Consistent increase observed in % conversion from June to Nov

1. From the graph, its clear that May and Feb months do not have good conversion ratio of visitors. However these two months only have the probability of having special days. We can conclude that special days need not to be drivers for visitors conversion.
2. Nov month shows the maximum conversion of visitors ~25.3%, so if any marketing strategy adopted in Nov month, recommendation is that same should be expanded for other months too.
3. May month seems to have highest visitors, however conversion is very poor ~11%. Strategies can be deployed for conversion of May month visitors. Recommended to overview product details published on website in May month.

Data preprocessing for ML models

Aim - Which online shopper will generate revenue

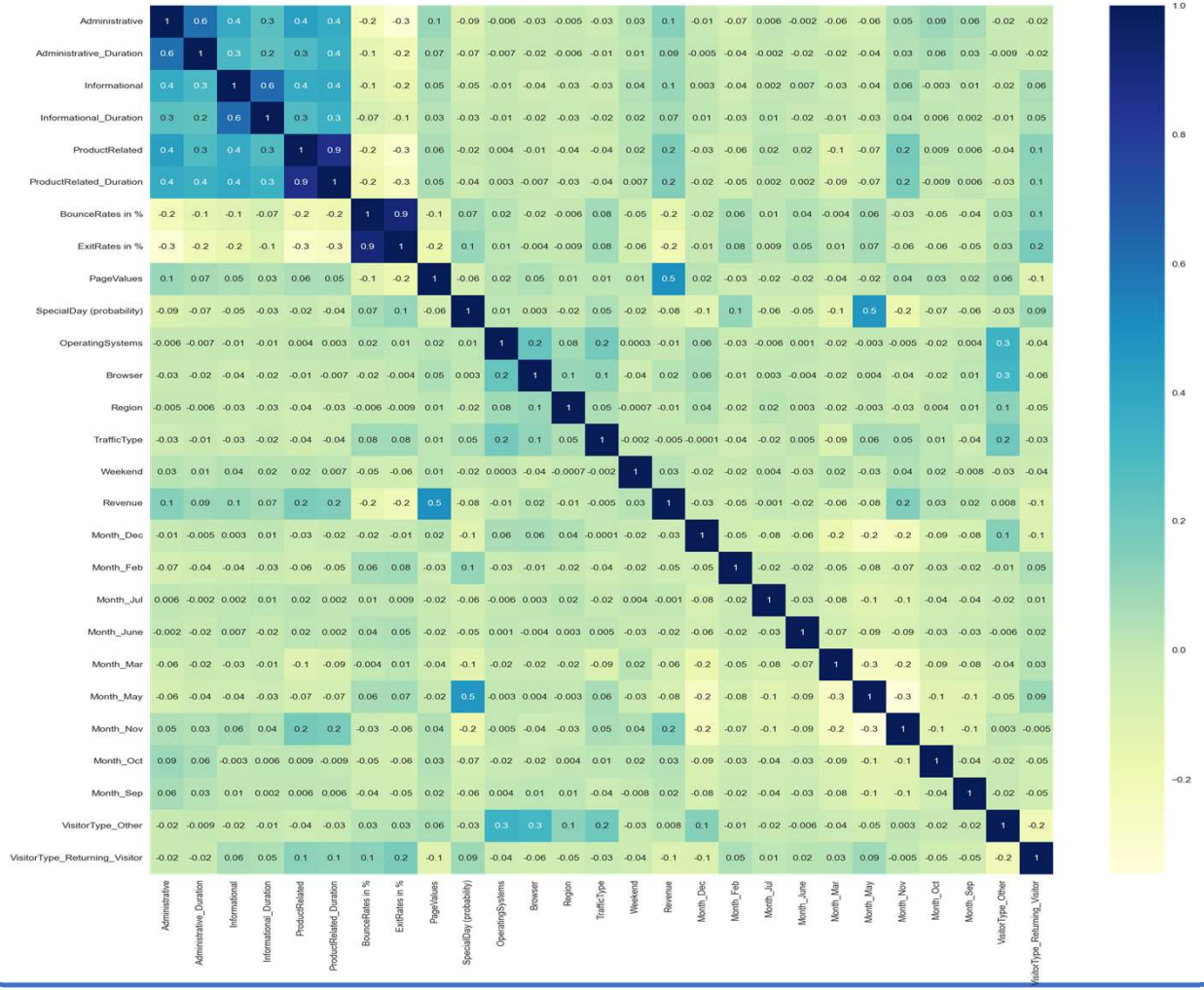
- Categorical to numerical conversion
  - Month and visitor Type columns converted to numerical using OneHotEncoder technique
  - Weekend and Revenue columns converted to numerical using label encoder technique
- Outlier identification and treatment
  - Only two columns found to be normally distributed, OperatingSystems and Region . Rest all are skewed.
  - However most of this columns represents the duration that customer has spent on the website. Which means that if higher duration spent, it indicates outliers but such visitors are important as they are showing interest in our website. So we will not treat these outliers, but we cannot use logistic regression for prediction.



Data preprocessing for ML models

Aim - Which online shopper will generate revenue

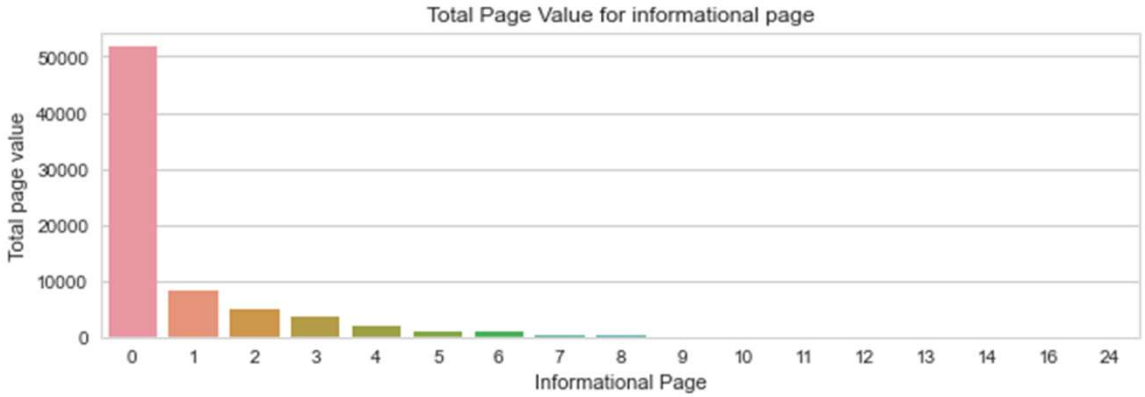
- Multicollinearity check and feature engineering
  - Bounce rate and exits rates are highly positively correlated, And productRelated_duration and productRelated columns are also highly positively correlated. We can use any one of these columns before building ML model. Apart from this, most of the other columns are non-correalted or observed to have weak correlation.
  - pageValues has high correlation with revenue compared to all other features

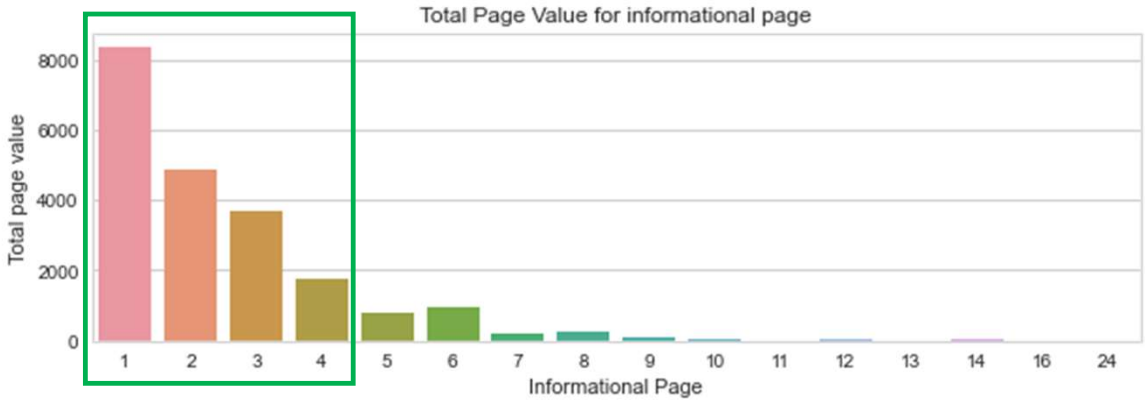


Data preprocessing for ML models

Aim - Which online shopper will generate revenue

- Why PageValues column highly corelated to revenue?
  - Page value is a metric that measures the average value of a webpage to a website's business. It is calculated by taking the total revenue generated by a webpage and dividing it by the number of views the page received.
- Which are the webpages that have impact on pageValues?
  - Informational page 0 has highest page values. However it seems that page 0 can be home page leading to all other pages. Pages 9 onwards does not impact revenue.

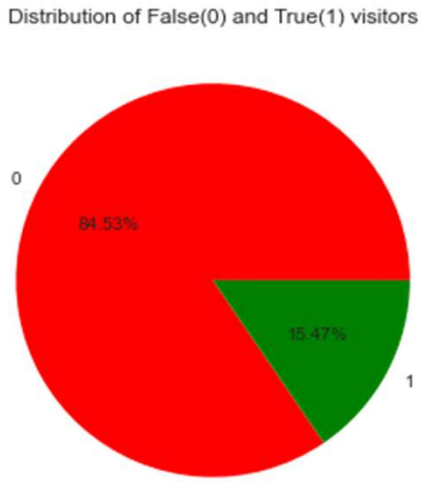




Selection and Building of ML Models

Aim - Which online shopper will generate revenue

- Target variable 'Revenue' found to be highly imbalanced
  - False or 0 → Indicates visitors failed to turn as customers
  - True or 1 → Indicates visitors turned as customers



- Model Selection
  - Preliminary accuracies determined using pycaret
  - Best model selected considering overall accuracy and recall accuracy

	Model	Accuracy	AUC	Recall
rf	Random Forest Classifier	0.9012	0.9217	0.5538
gbc	Gradient Boosting Classifier	0.9005	0.9299	0.6077
lightgbm	Light Gradient Boosting Machine	0.8992	0.9271	0.5935
ada	Ada Boost Classifier	0.8906	0.9131	0.5575
et	Extra Trees Classifier	0.8888	0.9094	0.4333
lr	Logistic Regression	0.8843	0.8873	0.3824
lda	Linear Discriminant Analysis	0.8796	0.9002	0.3322
ridge	Ridge Classifier	0.8730	0.0000	0.2425
knn	K Neighbors Classifier	0.8615	0.7658	0.3009
dt	Decision Tree Classifier	0.8547	0.7291	0.5471
svm	SVM - Linear Kernel	0.8498	0.0000	0.3111
dummy	Dummy Classifier	0.8452	0.5000	0.0000
nb	Naive Bayes	0.8124	0.8245	0.5906
qda	Quadratic Discriminant Analysis	0.7709	0.8268	0.7208

Random forest is further fine tuned for consideration of imbalanced target feature and hypertuning of parameters

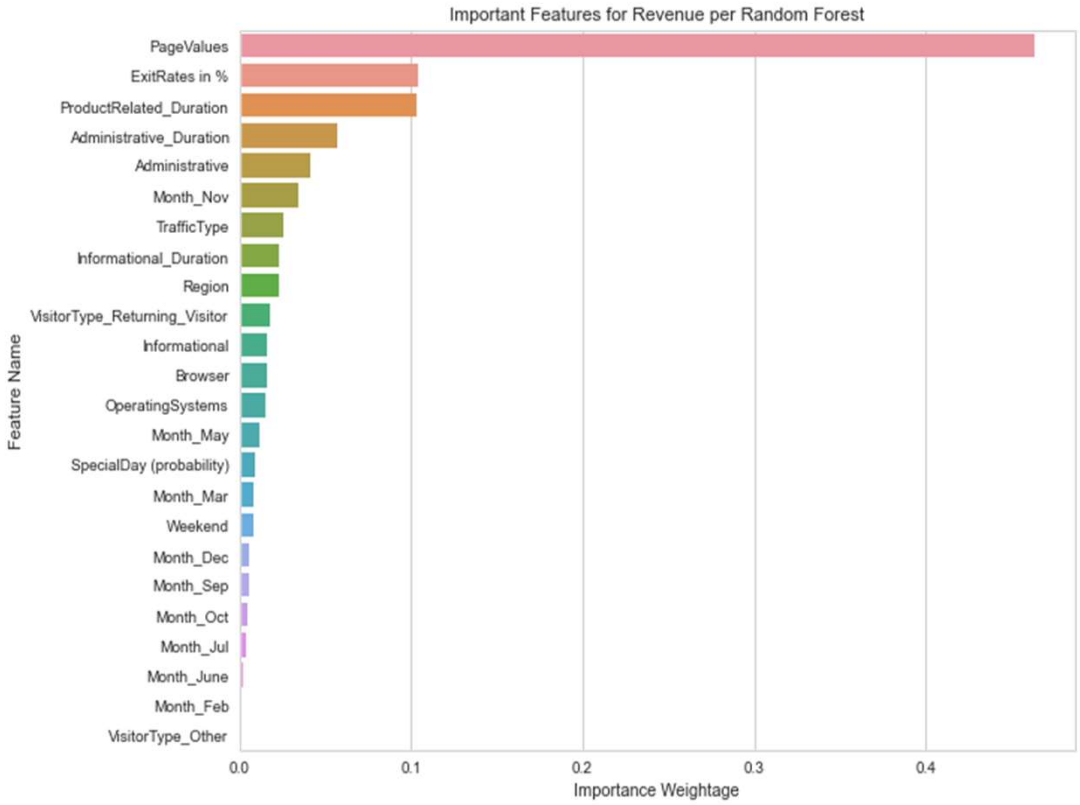
Selection and Building of ML Models

Aim - Which online shopper will generate revenue

- Finalization of ML model (Parameter Hyper tuning)
  - Random Forest is finalized ML model
  - Imbalanced data is taken care by following ways:
    - ✓ Target variable is stratified while splitting train and test data
    - ✓ Class_weight is considered as tuning parameter
- Validation of ML model
  - Probability of predicting the TRUE (1) class is important here, so Racall is prioritized for model validation over overall accuracy
  - No underfitting and overfitting issue observed with tuned model

	precision	recall	f1-score
0	0.96	0.91	0.94
1	0.62	0.80	0.70
accuracy			0.89

- Identification of important features impacting prediction



Concluding Summary

Aim - Which online shopper will generate revenue

- Data Analysis Observations:
  - May month has seen max number of visitors, however conversion rate was poor ~11%
  - June to Oct number of visitor were almost same, however conversion rate consistently increased from 10% to 25%
  - PageValues is highly positively correlated to revenue. And pageValues are seen to substantially improved by informational pages 0 to 4

- Recommendations:
  - Opportunity to improve pageValues of informational pages 4 onwards by improving the web content. Contents on pages 0 to 4 can be referenced for improvement.
  - May has max number of visitors with very poor conversion %. Validate the content of pages in May, and avoid similar content to be published in future
  - Administrative pages not to be neglected as those have significant impact on customer conversion