ONLINE SHOPPERS INTENTION

This project is aimed to analyze online shoppers' intention dataset stored on UCI Machine Learning Repository. Data set consists of more than 12,000 instances and 18 attributes that include customers 'duration' data at various category including other numerical and categorical features to predict whether a customer will be ended up with shopping (revenue True) or not (revenue False).

Data Visualization

```
df[['Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated','BounceRat
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	BounceRates	ExitRates	PageValues	SpecialDay	Month
0	0.0	0.0	0.0	0.0	1.0	0.20	0.20	0.0	0.0	Feb
1	0.0	0.0	0.0	0.0	2.0	0.00	0.10	0.0	0.0	Feb
2	0.0	-1.0	0.0	-1.0	1.0	0.20	0.20	0.0	0.0	Feb
3	0.0	0.0	0.0	0.0	2.0	0.05	0.14	0.0	0.0	Feb
4	0.0	0.0	0.0	0.0	10.0	0.02	0.05	0.0	0.0	Feb

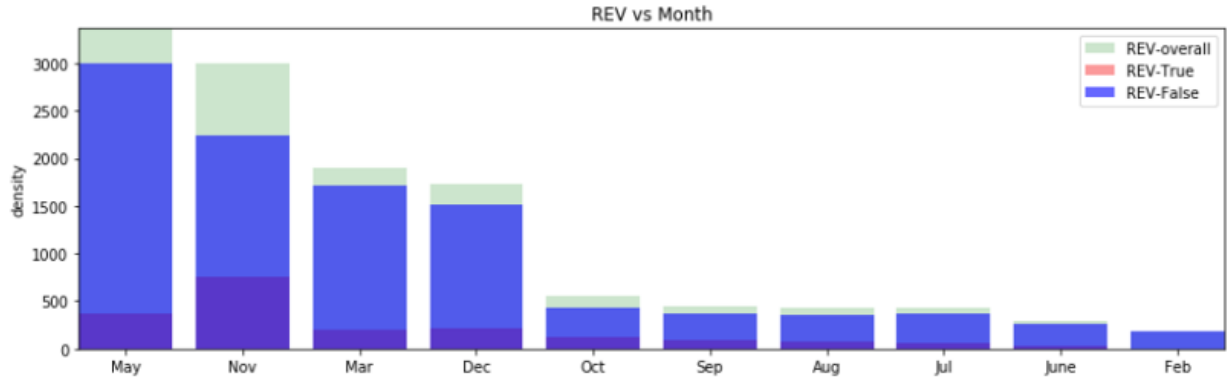
```
df [ ['OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType','Weekend', 'Revenue'] ].head()
```

	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	1	1	1	1	Returning_Visitor	False	False
1	2	2	1	2	Returning_Visitor	False	False
2	4	1	9	3	Returning_Visitor	False	False
3	3	2	2	4	Returning_Visitor	False	False
4	3	3	1	4	Returning_Visitor	True	False

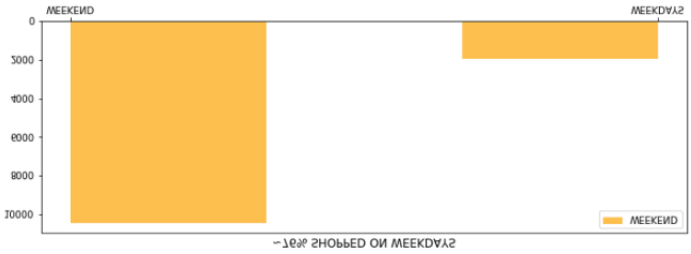
Data set consists of 15% positive class sample, i.e.; 85% did not ended up in shopping.

Among the eighteen attributes as listed here ['Administrative', 'Administrative_Duration', 'Informational','Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month', 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType', 'Weekend', 'Revenue'], ten of which are numerical and eight categorical.

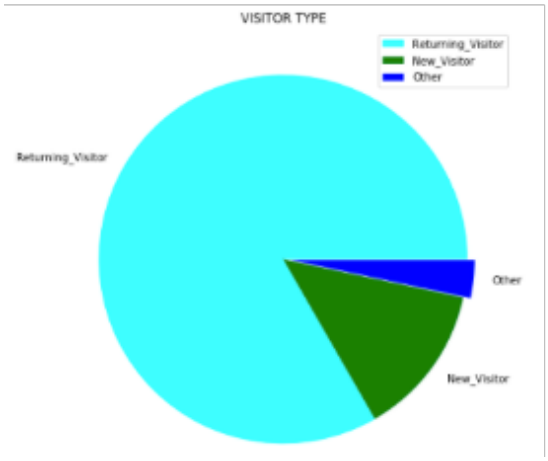
Data visualization shows that May is the month when maximum shoppers attempted shopping, however they really did shop max in the month of Nov.

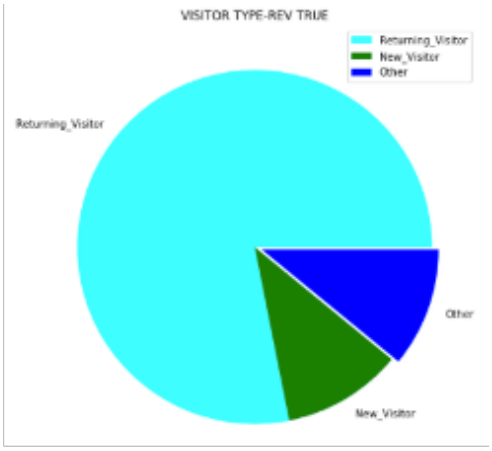


The ratio of week-day vs week-end revenue is 3:1 as shown here



In terms of visitor types ‘returning-visitor’ is the largest group irrespective of whether they ended up in shopping or not as visualized in pie-charts here.

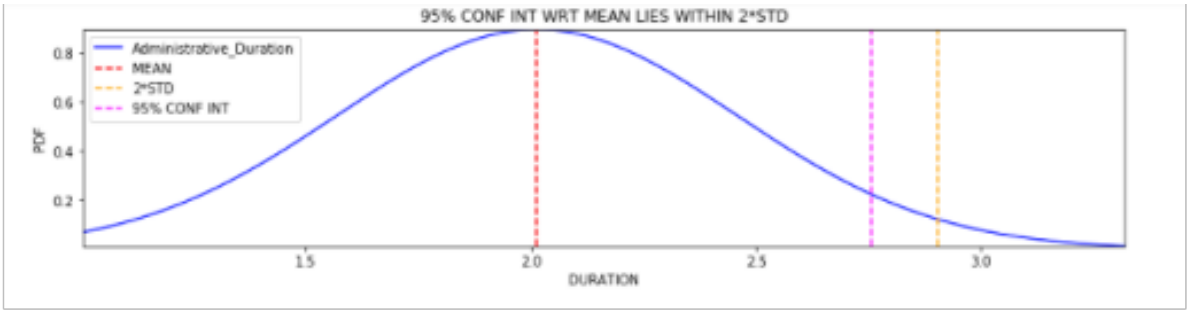


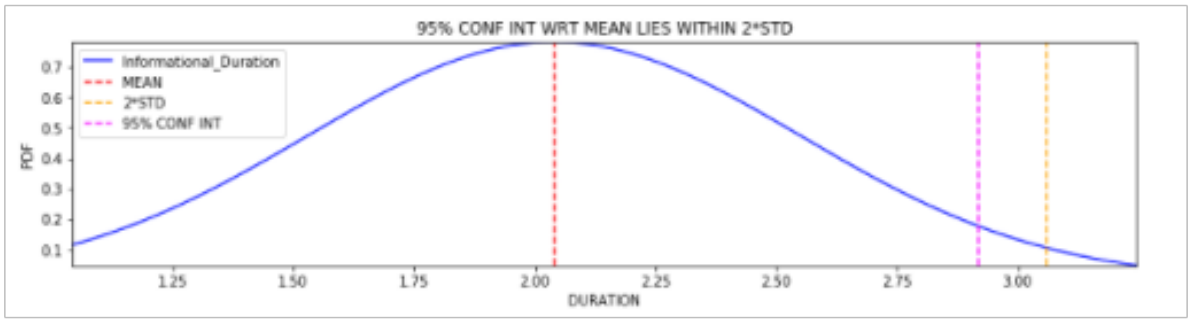


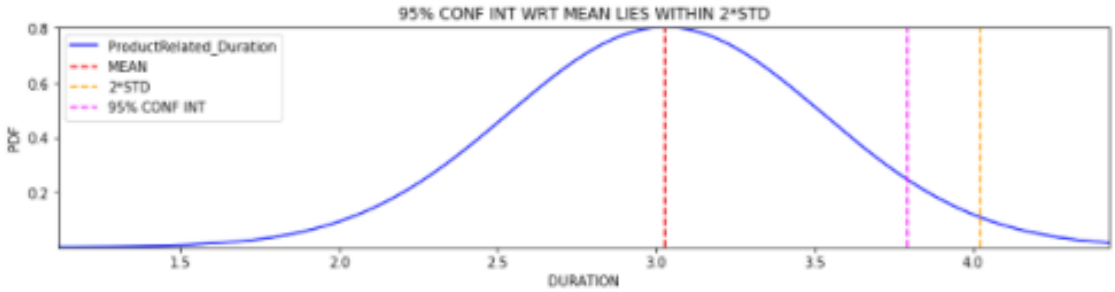
Statistical Analysis

While analyzing the shoppers duration in different category such as ['Administrative_Duration', 'Informational_Duration', 'ProductRelated_Duration'], we found 95% confidence interval with respect to the mean within 2 STD and this holds pretty much for overall shoppers or shoppers categorized in terms Rev-True or Rev-False as shown here, though those groups are not statistically identical in terms ttest.

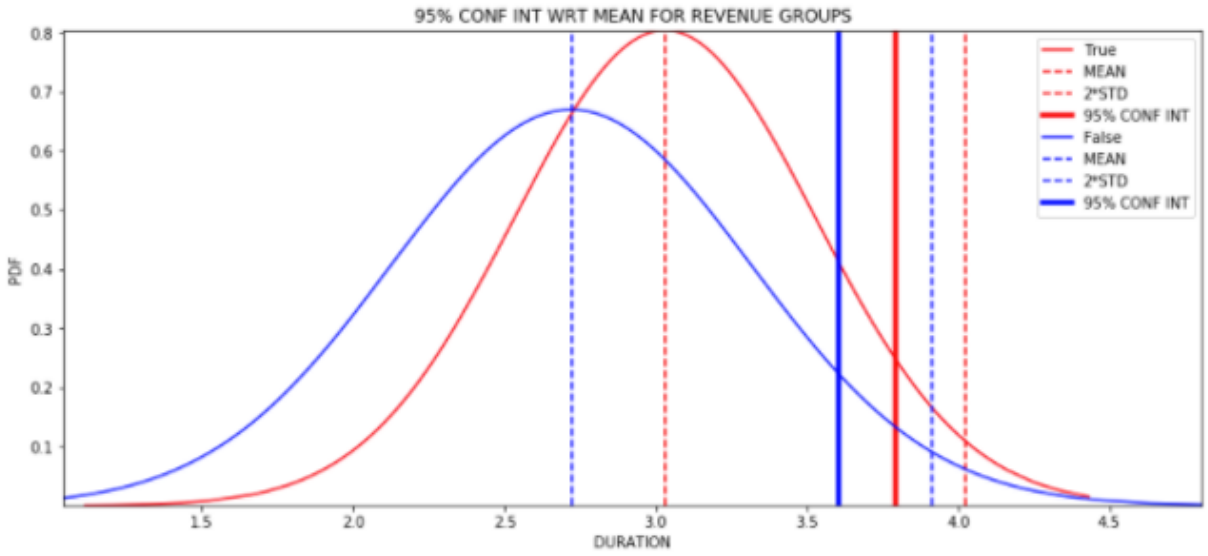
ProductRelated_Duration







We also analyzed the 95% confidence interval of 'ProductRelated_Duration' with respect to the mean of two groups (Rev-True and Rev-False) and found similar results, however hypothesis testing via ttest shows Rev-True and Rev-False groups are not identical.



ML Predictions

Several ML algorithms such as Random Forest Classifier, Extra Trees Classifier, Logistic Regression, ensemble Gradient, GaussianNB have been implemented to predict whether the revenue is true or false. Accuracy of the model was checked by grid search and cross-validation. Via cross-validation, we found Gradient Boosting Classifier predicts with highest accuracy. Based on feature selection and prediction, certain offers can be made to customer to increase the revenue.