**Online Shoppers Intention**

E-commerce is a booming industry with lots of potential for applying Data Science. Websites have their customer activity data that can be used to improve customer performance. There are various applications such as creating a Recommendation System, Inventory management and Warranty Management etc.

**Problem Statement:**

The boom of internet has given e-commerce websites accessibility to a large audience. In fact, an increase in website visitors can be observed. But, this increase in e-commerce usage over the past few years has created potential in the market but the conversion rate is not as high. Aim is to investigate the conversion rates, and understand what is happening.

**Data:**

Data was collected from Kaggle, linked it here (https://www.kaggle.com/roshansharma/online-shoppers-intention).

The dataset consists of feature vectors belonging to 12,330 sessions. Each session belongs to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

**Data Wrangling:**

1. Identifying categorical and continuous variables.
2. Handling NaN values: Session details of customers were essentially missing. In such a case all the columns corresponding to these sessions were NaN values, hence these session details were dropped.
3. Dealing with duplicate data: There were almost 117 rows that were duplicated and hence were dropped.
4. Outlier Detection: On further analysis, we can observe that ‘ProductRelated\_Duration’ column has outliers. Hence, used tukey's method on the column to remove outliers.

**EDA:**

<pie chart>

In the model, Revenue is the target variable. In the pie chart, it can be observed that data is highly imbalanced.

<Special Day vs Revenue>

There is hardly any shopping on or the day's approaching any special day

<Month>

May, November, March and December have peak. There are 2 opportunity windows, March - May and November - December

<Weekend vs Revenue>

Even if on surface it looks like there is more sale during Weekdays, but during Weekends customers seem to buy more wrt visiting the site.

<No. of pages and Duration>

From the above plots we can see that, the time spent by visitors’ pages are related to number of pages. Eg. in above figure a visitor’s duration on ProductRelated pages increases as number of pages do. It seems like Visitors who don't convert stay ideal and don't interact with the website.

<Correlation>

Exit Rate and Bounce Rate are highly corelated.

**Algorithms and Machine Learning**

This can be treated as a classification problem, with target variable as ‘Revenue’.

In this case our governing metric would be ‘F1 Score’ of minority class.

As the target variable is skewed. Tried using 3 methods to deal with the imbalance:

1. Class Weight: One of the simplest ways to address the class imbalance is to simply provide a weight for each class which h places more emphasis on the minority classes such that the end result is a classifier which can learn equally from all classes.
2. Under Sampling – Near miss: The general idea behind near miss is to only the sample the points from the majority class necessary to distinguish between other classes.
3. Over Sampling – SMOTE:  Synthetic Minority Over-sampling Technique (SMOTE) is a technique that generates new observations by interpolating between observations in the original dataset.

Classification Algorithms used were: Logistic Regression, Decision Tree, Random Forest, XG Boost and Random Forest.

There were 2 models that performed well, Class Weight in SVM and Random Forest with SMOTE and Feature selection.

After hyper parameter tuning of the Random forest gave the highest F1 score of 71.

<cumulative importance>

**Future Improvement**

A added feature could be to add Customer detail and Customer segmentation, to analyse more about the customer demography.