**Customer Retention Case Study Report**

**The main steps that have been undertaken during the analysis are:**

1. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values. Although, there were no null values in the data provided.
2. An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.
3. Removal of skewness- f there are too much skewness in the data, then many statistical model don’t work. So in skewed data, the tail region may act as an outlier for the statistical model and we know that [outliers](https://discuss.analyticsvidhya.com/t/effects-of-outliers-on-regression-model/2403/2) adversely affect the model’s performance especially [regression-based](https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/) models. There are statistical model that are robust to outlier like a [Tree-based](https://www.datacamp.com/community/tutorials/decision-tree-classification-python) models but it will limit the possibility to try other models. So there is a necessity to transform the skewed data to close enough to a Gaussian distribution or Normal distribution. This will allow us to try more number of statistical model.
4. Outlier Removal- Sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data. These are called outliers and often machine learning modeling and model skill in general can be improved by understanding and even removing these outlier values.

1. **Training Process-** The process of training an ML model involves providing an ML algorithm (that is, the *learning algorithm*) with training data to learn from. The term *ML model* refers to the model artifact that is created by the training process.

The training data must contain the correct answer, which is known as a *target* or *target attribute*. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

*To train an ML model, you need to specify the following:*

* *Input training datasource*
* *Name of the data attribute that contains the target to be predicted*
* *Required data transformation instructions*
* *Training parameters to control the learning algorithm*

**Conclusions from the EDA:**

1. From all the factors mentioned for customer retention: best deal, monetary savings and convenience is given the highest priority.
2. The customer base was bifurcated between men and women in the ratio of 1:2.
3. City with the maximum online purchase was Delhi & Noida.
4. The age group doing the maximum online purchase in 31-40 followed by 21-30 group.
5. The most important reason for abandonment of a product in cart is better alternative option available.

6)Customers have maximum perceived trustworthiness on amazon shopping website. This is followed by flip kart and then Myntra. The least votes are given to Paytm online shopping.

7) The most efficient website is amazon according to customer feedback given.

8)Recommendation is also given to amazon website for online shopping.

9) Correlation- Purchase times is negatively correlated with adventure and return & replacement policy of the e-retailer.

10)Skewness- Power transformer was used to remove the skewness in the distribution.

11) Outliers- Z-Score method was used to remove outliers. There was 7.43% data lost on removal of outliers.

12) The performance can be improved using the decision tree classifier as it is showing the highest result.