**Avacado Project**

**Business Problem Framing/Problem Definition:** The given dataset talks about Avacado which is a fruit consumed in USA. The source of data is Hass Avocado Board website (May 2018). In the given dataset, it represents weekly 2018 retail scan data for National retail volume (units) and price. The dataset contains The Product Lookup codes (PLU’s) which are only for Hass avocados. The codes are 4046, 4225 and 4770. Based on various parameters provided, we have to build a model to predict the Average Price

**Data Analysis:**

Data Analysis is a key part of model building and it requires various parameters that needs to be analysed before arriving at an accuracy to avoid attrition. So, in this case, the target variable will be AveragePrice and various input parameters are listed below

* Date
* Total Volume
* 4046
* 4225
* 4770
* Total Bags
* Small Bags
* Large Bags
* XLarge Bags
* type
* year
* region

**So, to perform analysis, we have to follow the below steps**

* Import all the libraries in Jupyter like Numpy, Pandas, Matplotlib, seaborn etc.
* Load the data into the file. This will help us to read the data and start performing analysis
* Find out the number of rows and columns in the dataset
* Identify the types of columns in the dataset
* Describe the data to find out mean, median, count, min, max and the percentiles
* Find out information about the data
* Identify datatypes from the dataset
* Identify if there are any null values
* Plot histogram
* Find out unique values
* Find out the count from the dataset

**EDA (Exploratory Data Analysis):**

EDA is a key part in Machine Learning as it helps us find out information from the data, find out insights, infer from them and perform visualization of the data to make better data driven decisions. This involves Univariate (single variable), bivariate (comparing 2 variables), multivariate analysis (finding correlation and plotting heatmap)

As part of Univariate Analysis, we perform the below

* Use countplot to plot the target variable to find out the class imbalance
* We use distplot to find out the distribution of curves for the given columns or variables (Bell shaped curve). From this, we can infer if the curve is normally distributed or not
* We use box plot to identify if there are any outliers in the dataset

As part of Univariate Analysis, we perform the below

* Here we can use two related variables to compare the data, so that we can infer insights from 2 different variables which are used for plotting. Here we have used 4046 and 4225 along with 4046 and 4770. Here we have used scatterplot for the same
* We can also use “hue” to differentiate target variable from input variables

As part of Multivariate Analysis, we perform the below

* Here we can use cor command to find out the correlation of different variables in the dataset
* Plot the heatmap for visualization

**Encoding**

Since the given dataset has categorical variables, it has to be converted to numerical values before processing the data. Here we are applying Ordinal Encoding. Label Encoding can also be applied

After the data is encoded, we have to remove the outliers. We can apply Zscore or IQR for the same. For our dataset, we have applied Zscore

After the outliers are removed, we have to find out the new shape to find out the number of rows and columns and find out the skewness

**Pre-processing Pipeline**

* Before processing the data, we have the import MinMaxScaler, train test split and import r2 score
* After importing, we have to identify the best random state for train and test data
* Then, we have to identify the best accuracy on the Random State

**Building Machine Learning Models**

* Once we have pre processed the data and identified the problem, we have to decide the algorithm. Since we have to find AveragePrice in this case which is a target variable and since these values that are continuous, it is a regression algorithm
* Here we have to import precision\_score, recall\_score, f1\_score, roc\_auc\_score, accuracy\_score, classification\_report, confusion\_matrix
* Here we have tested and applied Random Forest Regressor algorithm while building model
* Once we have obtained accuracy using the algorithm, we have to use cross validation to improve the accuracy of the model

**Concluding Remarks**

Once we have found the improved accuracy using CV (87 accuracy in this case), we will be deploying the final model on pickle and comparing original V/s predicted values