**Avacado Project**

**Problem Definition:**

***Avocado is a fruit consumed by people heavily in the United States.***

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. The table below represents weekly 2018 retail scan data for National retail volume and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price in the table reflects a per unit (per avocado) cost, even when multiple units are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (greenskins) are not included in this table.

**Columns in the dataset:**

* + Date - The date of the observation
  + AveragePrice - the average price of a single avocado
  + type - conventional or organic
  + year - the year
  + Region - the city or region of the observation
  + Total Volume - Total number of avocados sold
  + 4046 - Total number of avocados with PLU 4046 sold
  + 4225 - Total number of avocados with PLU 4225 sold
  + 4770 - Total number of avocados with PLU 4770 sold

**Target**

The dataset can be seen in two angles to find the region and find the average price .

Task: One of Classification and other of Regression.

**EDA**

First we need to do EDA process in which we analyse, clean and preprocess the data. In this dataset—

* There are 13 features in data.
* Region and Type has object datatype.
* Date by default data type is object which need to change to datetime datatype.
* Other variable are float & int datatypes.

For this I have used df.columns, df.info and df.shape commands. Now checked null and duplicate values in dataset but thankfully there is no null and duplicates in dataset.

**Statistical Summary:**

* Value of standard deviation is greater than means data is spread.
* Some of the features contain outliers after looking at 50% and max columns.
* For most of columns have minimum value of zero.
* Almost in all columns value of mean is greater than median. Data is right skewed.
* We will check skewness with df.skew commands.

**Visualisation:**

With different types of graphs we can analyse relation and dependence between features. Before this I have extracted days and months from date and drop Date and Unwanted: 0 columns as both are not for any use.

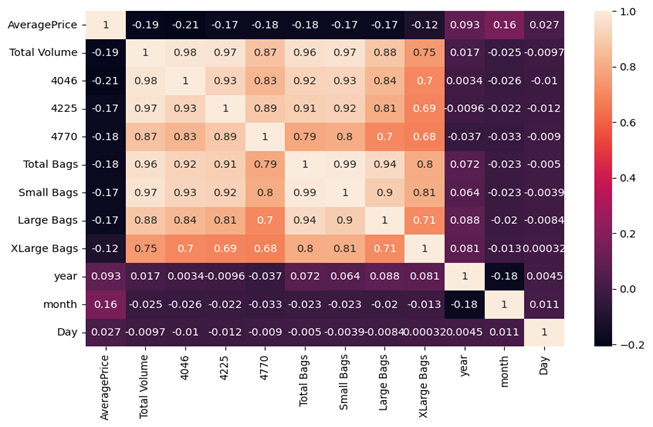
Using scatter plot I have checked relation between region and average price and got that highest price for Avacado lies in Senfransisco.

A blue and white lines with black text

Description automatically generated

Price are high for conventional and demand is high for organic.

* Using heatmap checked for multicollinearity and correlation between variables. Small bags are highly correlated with Total Bags correlation coeffient 0.978.
* Total volume is highly correleated. 4046,4225,4770 are shows multicollinearity with Total Volume. These feature are nothing but Total avacado sold under particular grade, which are already counted in Total Volume. As we do not have any price data according each grade of avacado. These feature doesnot have any meaning in ML model.So I will drop them.
* Same goes with Total bags. Total bags is sum of all other type of bags. I will drop Total bags and going to keep other differnt size bags counts features.
* Next is region feature, we also going drop this as it very poorly correlated with other features.



With more visuals and graphs we can get more information about data.

Before removing outliers we have to encode the data using label encoding so that all features can be in integer form. For this I have used below command-

**from** sklearn.preprocessing **import** LabelEncoder

lb**=** LabelEncoder()

df['type']**=**lb**.**fit\_transform(df['type'])

As only ‘Type we have to change. After this removed outliers using zscore and removed skewness using boxcox method.

z**=**np**.**abs(zscore(df)**.**values)

threshold**=**3

print(np**.**where(z**>**3))

skew\_list**=**['Total Volume','4046','4225','4770','Total Bags','Small Bags','Large Bags','XLarge Bags']

**for** i **in** skew\_list:

df[i]**=**boxcox(df[i],0.5)

We need to clean our data by removing outliers,skewness and unnecessary columns so that our model can predict more accurately. Now our pre processing is complete but many more techniques are there to be used according to nature of data and goals of analysis. Scaling is important for put all features on same scale to get more accuracy for our prediction.

Building ML Model

Our first target is to predict average price which is a Regression problem so I have imported some Regressors.

Linear Regression

Logistic Regression

Random Forest Regressor

Decision Tree Regressor

SVR

AdaBoost Regressor

K Neighbors Regressor

Imported Regressors and then trained our model with fit command and test with predict command. Random Forest Regressor among all gave best score 78% so done hypermeter tuning to get best parameters.

A graph with blue dots

Description automatically generated

Second target is ‘Region’ which is categorical. W have imported some classifiers for this task.

SVC

Decision Tree Classifier

Random Forest Classifier

Again we got Random Forest Classifier as best accuracy giving model with 90% accuracy.

In conclusion, the Random Forest Regressor outperforms the other models with the lowest errors (MAE, MSE, RMSE) and the highest accuracy (r2). It proves to be the most effective in predicting avocado sales data compared to the Decision Tree Regressor and Linear Regression models.

Now imported joblib and saved the model.