**Avocado problem**

Author- Aditya Kissen Mitra

1. **Problem Definition-**

I have taken a dataset where Avocado consumed by people in United State. Content in the dataset was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

Starting in 2013, the table 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 (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags.

**Predictions**-

* To find the best region to grow Avocado
* To predict the average price of the Avocado

1. **Data Analysis-**

The dataset is containing 13 columns as below-

**Features-**

* Date - The date of the observation
* type - conventional or organic
* year - the year
* 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
* Total Bags- Total number of bags to store avocado
* Small Bags- Total number of small bags to store avocado
* Large Bags- Total number of large bags to store avocado
* XLarge Bags- Total number of extra-large bags to store avocado

**Label**-

* Average Price - the average price of a single avocado
* Region - the city or region of the observation

I have checked for any null values present in the dataset and get to know that now null values present in the dataset. Years mentioned in the dataset ranges from 2015 to 2018 and for region we have different regions across United States.

1. **EDA Concluding Remark-**

I compared the sales between the years and get to conclude that 2018 is the year where less sales has been occurred and 2017 is the most. On 2017 I can see that price increased most. After doing some research find out that due to unsuccessful harvests and rises in demand, avocado prices are doubled on 2017.

Organic type avocado are more costly than convectional type.

1. **Pre-Processing Pipeline-**

Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results.

I have checked the heat-map for correlation between each columns and get to know most of the columns are correlated. Then I checked the skewness and find out most of the columns are highly skewed. To remove skewness I used log1p() function and got a good result and skewness got removed. I used Label encoder to encoded the encode the categorical data.

Then I used box plot to check the outliers present in the dataset and find out outliers are present in many columns. So, I decided to go with outliers removal technique using z\_score technique. After outliers removal I checked the total data loss and find out only 2% of data got lost. I considered it as good to remove so much outliers. Then I checked the heat-map again for correlation and found that correlation also got controlled.

I checked the VIF which according to me was fine. So, I move forward to do the model building.

1. **Building Machine Learning Models-**

I used Standard scaler to scale the dataset and proceed with the model building. I need to predict for average price and region. For Average price prediction I have used different regression techniques and for region prediction used different classification techniques.

* For Prediction of Average Price I have used-

Linear Regression, Ridge, Gradient Boosting Regressor, K-Neighbour Regressor, Random Forest Regressor and Decision Tree Regressor and checked the cross validation score then hyper tune all of them and finally got the best accuracy and cross validation score for Gradient Boosting Regressor with 91% accuracy on test data so I used Gradiant boosting regressor for predicting Average price of Avocado.

* For Prediction of Region I have used-

Logistic Regression, Decision Tree Classifier, Random Forest Classifier, SVC and checked the cross validation score and find out Random Forest gave me the best accuracy and CV score. So, I hyper tuned the Random Forest Classifier and got the accuracy score of 89% on test data and predicted the region.

1. **Concluding Remarks-**

I concluded by saving both the models using pickle function. For predicting Average price used Gradient Boosting Regressor and for region predicting used Random Forest Classifier. Then predicted the results using the best estimators and store it on the data-frame with actual values and predicted values.