Project Evaluation-Phase4

In my Data trained internship programme, I worked on several Capstone Projects. I assigned several projects but here I will explain the ‘Avocado Price Analysis’.

**Introduction:**

In May 2018, this information was extracted into a single CSV file from the Hass Avocado Board website.   
  
The weekly 2018 retail scan data for national retail volume (units) and price are shown in the table below. Acquired from actual retail sales of Hass avocados, retail scan data is directly from retailers' cash registers. The table below displays a multi-outlet retail data collection that has been enhanced as of 2013. The channels grocery, mass, club, drug, dollar, and military are aggregated in multi-outlet reporting. Even in cases where many units (avocados) are sold in bags, the Average Price (of avocados) in the table represents a cost per unit (per avocado). Exclusive to Hass avocados are the Product Lookup codes (PLUs) listed in the table. Other avocado kinds, such as green skins, are given as follows:

* Region: the city or area where the observation was made
* Overall 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 Inspiration / Label

The dataset can be viewed from two perspectives to determine the average price and the region.   
One task: one is classification  
Perform both duties simultaneously.ipynb file and send it as a single file.

**Datasets:**

The dataset is available on the Kaggle and I already uploaded it on my GitHub profile.

**Steps for Machine Learning Modelling:**

1. The very basic steps which I followed included downloading important libraries like Pandas, Numpy, Matplotlib, Seaborn
2. Now I use the Pandas for reading the CSV file
3. Now I am Performing the EDA
4. I checked whether any null values are there or not and I found that there are no null values there
5. Now I dropped the column ‘Unnamed’ which is not useful for us
6. I perform various graphical analyses using the Seaborn library for the EDA. I performed the Count plot, Bar plot, Scatterplot, lmplot, regplot for the analysis against the target variable
7. Using the histogram plot I checked whether my data showed right-side skewness or left-side skewness
8. I saw that the data shows skewness so I removed the skewness by using the np. sqrt method
9. Boxplot shows the dataset contains the outliers so we have to remove the outliers by using the Z score method
10. I performed the Pair plot which gave me a detailed analysis of the dataset
11. I converted the categorical dataset into a numerical one
12. I dropped the column which is not useful for us
13. I removed the outliers using the Z-score method which removed the columns which have outliers in it and will ultimately lower the number of columns in our dataset
14. I plotted the correlation using the heatmap seaborn which ultimately showed me which variable is highly correlated positively or negatively correlated
15. Dropping the target variable from the dataset and dividing the dataset into X and Y
16. Performed the standard scaling method for X

**Multicollinearity:**

I checked the multicollinearity in the dataset and I found that the dataset has multicollinearity in the dataset so I removed the multicollinearity using the Variation Inflation factor which removed the multicollinearity from the dataset.

**SMOTE:**

I found that the target variable was unbalanced so I had to balance that target variable and I balanced the target variable using the SMOTE technique.

Modelling:

Now I imported the different machine learning models and I noted that I have to use the Classification machine learning model since my target variable is in categorical form. Since I am proceeding with the classification model I am checking the accuracy score and the random value and I found that the Random Forest Classifier gives the Best accuracy score is 0.9988880652335063 at Random State 52.

Now I run various machine learning models like DecisionTreeClasifier, ExtraTreesClassifier, SVC, KNneighborsClassifier, and GradientBoostingClassifier.

**Cross Validation Score:**

After performing the various machine learning model which gives us accuracy score, confusion matrix, classification report. Since we are following classification I have to consider the accuracy score and the difference between the accuracy score and score value and I found that the RandomForestClassifier gives me the least value hence I selected the RandomForestClassifier as my best model. So now, I proceed with the Hypertuning of my model which is RandomForestClassifier.

**Hypertuning Model:**

I selected GridSeacrhCV for hypertuning the model which gives me the best parameter for my model and I got the following parameters as a best parameter for RandomForestClassifier after hyper tuning the model-

'criterion': 'entropy', 'max\_depth': 20,

'n\_estimators': 300,

'n\_jobs': -2,

'random\_state': 1000

Once I got the best parameter after the GridSearchCV hypertuning now I calculated the accuracy of my model and I found that my model gave me a 99.57% accuracy score which is quite good for me It shows that my model can read the data correctly and correctly answer.

**Saving Model:**

Now I know that my model giving me a 99.57% accuracy score so I have to save my model. I am saving my model using the Joblib method. In this method I am following two processes-

1. The first thing I did was dump my model and give it the name ‘Avocado Price Analysis. pkl’, pkl is an extension.
2. After dumping the model I have to load my saved model for prediction
3. I generated a numpy array for my prediction