**Avocado**

This is the dataset repository consist of the information of the weekly sales of Hass Avocado and reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the channels such as: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the data reflects as per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados.

**Problem Definition**

**Our goal or objective** consists of two tasks i.e. to find the average price of the avocados and the region. So we can say that this is both a classification and regression problem. We have to predict the values with the help of other dependent features given to us such as the date of the observation, type of farming of avocado, year information, city or region of observation which we also have to predict in one of the scenario, total no. of sales and few other important features.

**Data Analysis and EDA Conclusions**

When we started analysis of Dataset given to us we found that out of our 13 independent features from total of 14 features 3 of them were categorical features which were “Date”, “type” and “region”, which were needed to be treated and also that there are a total of more than 16k observations in the dataset but out of them only approx. 1500 have information and rest of them have null values, so we also had to drop them.

When I started for the EDA I found out that the first feature have no heading and all of the different values in the dataset it was like an indexing for the observations and thus won’t making any impact on the prediction, so it was no harm in dropping the column.

Now I had to treat the rows with missing information which was starting from row 1517 so I also drop the rows after that and thus removing all the rows with missing information and now the shape of dataset which we I had to work was of 1517 rows and 13 columns.

Next step was to handle the categorical columns which were “Date”, “type” and “region”. For Date I separated the year, month and date and store into separate columns so that they won’t lose any their impact in prediction and I did it by using the panda’s function “to\_datetime”, and then later convert them into integer so they can be now accepted by the ML model. After the Date feature was treated and further new features were added in place of that I dropped the original Date column thus treating the first categorical feature. For second feature type we found that there are only one type of information was present in the entire dataset so we can drop this feature as presence of this feature won’t make any impact on the predictions, so we drop the column “type”. Now we are only left with one categorical column “region” which we still need to convert into model acceptable form.

Before converting last categorical column into acceptable form I perform some visual analytics using Seaborn library, first I tried to draw correlation using heat map of Seaborn library and tried to figure out the correlation between different features of the dataset. Then I tried to check the normal distribution using the plot function from pandas library itself after plotting I found out that there is little skewness in the dataset features because of few outliers. For feature “region”, we are using One Hot Encoding which we will be using by the help of pandas “get\_dummies” function and then storing the new encoded features in new dataframe, after treating the remaining categorical feature it’s time to concatenate the new encoded dataframe with the original dataframe after concatenation we have to drop the initial Region feature containing the categorical values and now the final size of the dataset is 1517 rows and 112 columns.

After conversion of the categorical values there are no other scenarios such as of missing values or curse of dimensionality and our dataset is ready for prediction.

We have to perform two types of prediction i.e. predicting the price of avocado which is a regression problem and predicting the region which is a classification problem.

I started with the regression problem i.e. predicting the Price of avocado, for that I segregated the dependent and independent variables from each other and store them in variable x and y respectively, after segregating. I also tried to find the best random state for the prediction so I use Linear regression for initial prediction (As this is a regression problem) and tried to compare the “**r2\_score**” of the train and predicted train with the “**r2\_score**” of the test and predicted test up to 1000 times, but sadly there was no match so I defined random state by my own.

After splitting our dataset into train and test dataset I started applying multiple algorithms starting with Linear Regression along with the regularisation technique Lasso Regression from library “sklearn.linear\_model” module “Lasso” for better results I tried to optimise the parameters of the algorithm (alpha, random\_state) by using Grid Search CV and then using the best parameters for the dataset generated by Grid Search CV while fitting the model with best parameters I also tried to find out the best cross validation for the dataset using some mathematical concepts and fortunately I found some values which can be used, after fitting the model I find the R2 score of 71% generated by model which was fine at initial stages but not good enough to pass the model so we move ahead with using another technique of Decision Tree for our model building available from package “sklearn.tree” and receives the score of 100% which clearly shows that our model was over fitted at some stages of model building (one of the major problem of Decision tree algo.) so we can’t use this model in this condition so we proceed with parameter optimisation (“criterion”, “splitter”) using Grid Search CV, then I received the best parameters of the decision tree technique for the given algorithm and I fit the model again with this technique and received R2 score of 77% which was little better than the Lasso regression, but still not god enough for model deployment.

Later I started with ensemble techniques and used Random Forest which is initially based on Decision tree technique but overcome the flaws of decision tree after using the Random forest model after hyper parameter tuning I received the score of 88% which is very good and acceptable for model building. So we use the Random forest method with tuned parameters for the prediction of Avocado’s price.

Now for the prediction of region, as our prediction variable is different we have to segregate our columns again and as we can’t use many columns (columns generated by one hot encoding) I have to encode them using different technique.

For encoding “region” this time we are using Label Encoding from library sklearn package preprocessing.

Now I segregated all the columns into training dataset except the label column region and stored them in a variable x and the encoded column in variable y.

Now I repeat all the steps from splitting the train test data to applying different algorithms but as this is a classification problem so we used different algorithms starting with Decision Tree Classifier which after hyper parameter tuning gave the accuracy score of 92% which is very good accuracy score and shows that the model can be used for further predictions but as far we have seen that using ensemble techniques gives higher accuracy over the normal Decision tree so I also used Random forest which gave us the accuracy of 98% which is a very good score for any model prediction and better than decision tree classifier.

So for both regression and classification problem of the dataset we are using Random Forest Algorithm along with some parameter tuning as they are giving the finest result.

I have saved both the models separately with different names into a binary file using pickle library.