Avocado Project Blog

Avocados are highly nutritious and packed with antioxidants. They are great source of Vitamins like Vitamin K, C, B and E also has minerals like copper, phosphorus, magnesium, manganese, iron, and zinc.

There are different types of avocado based on their shape, origin etc. Though we are familiar with a few types of avocado, hundreds of varieties are cultivated around the world. The Avocado fruit are categorized as either A-type or B-type cultivars. A-type cultivars has many varieties like Lula, Reed, Hass, Pinkerton etc.

The Hass avocado is a A-type cultivar of avocado with dark green – coloured, bumpy skin. Today we will have a insight of some dataset downloaded from Avocado Board website in May,2018.

1.Problem Definition

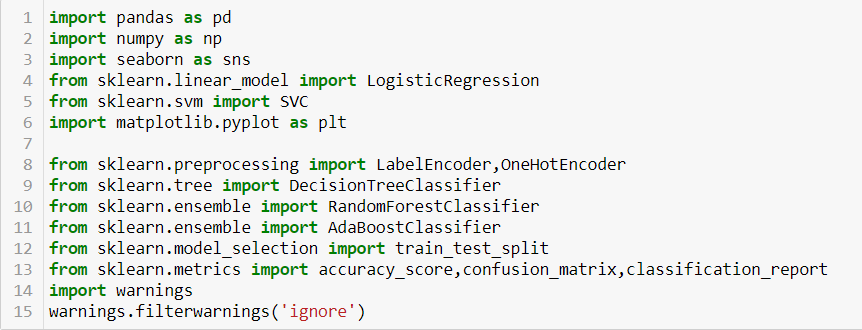
The goal is to predict the city or region and Average price from this dataset. The dataset contains PLU’s also called as Product Lookup Codes, which are a system of numbers that uniquely identify bulk produce sold in grocery and supermarkets which are administered by IFPS (International Federation for Produce Standards).

2.Data Analysis

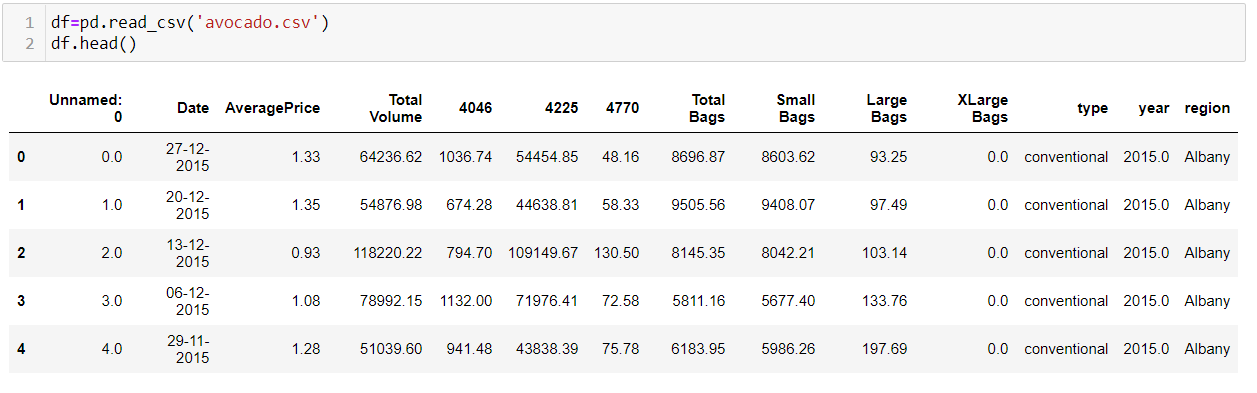
The dataset contains weekly 2018 retail scan data for National retail volume (units) and price, this data comes directly from retailers on actual retail sales of Hass Avocado. This dataset has 16468 rows and 13 columns.

I have 3 PLU’s (Product Lookup Codes) like 4046, 4225, 4770. Dataset also has the size of the bags with Small, Large and X Large. I could analyse the data with Average Price or region as label.

I am starting by importing the libraries which I will require to do the EDA process, prediction, visualization and all the findings from the dataset.



Loading the Dataset into the Variable:



I have loaded the dataset into the variable i.e, “df” and could see that almost all the columns are float data type’s but region, type and year are categorical.

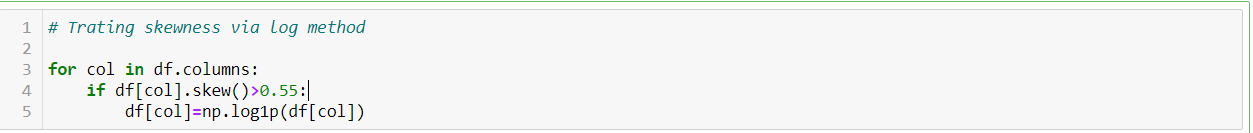
4.Exploratory Data Analysis concluding Remarks:

This dataset has 16468 rows and 13 columns, but most of the columns has NaN values, hence vomiting the NaN columns and removing the unnamed column, I got 1517 rows 12 columns. Though I could also drop the “type” column as it only conventional type in all the dataset but it would neither contribute or harm the prediction.

Since I do not have a null value, I could proceed with checking the co-relation heatmaps, Univariate, Bivariate analysis and could check the skewness of the dataset. The mean of the "Total Volume","4225","Total Bags" and "small Bags" are less than median. There is a large difference in 75% percentile and max in Total Volume,4046, 4225, Total Bags, Small Bags and X Large Bags.

5.Pre-Processing Pipeline:

After passing label encoder for type and region column, removing the skewness via log method and outliers via z-score method I got the dataset ready for splitting and scaling for predictions in the model.



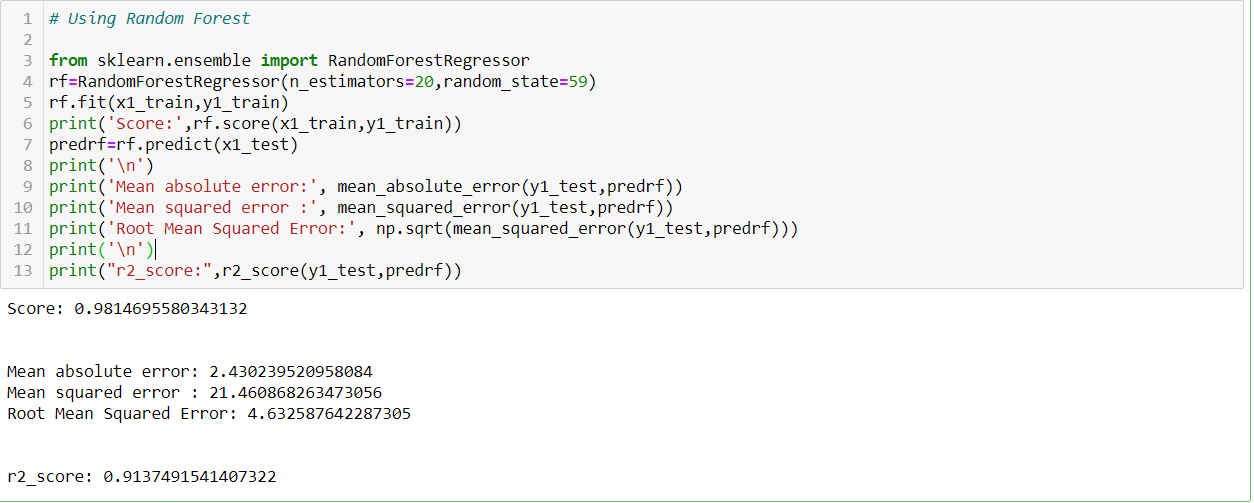
6.Building Machine Learning Models:

Since I have the column Average Price for prediction and next prediction is with region, it has been added to y variable, and rest of the columns are added to x variable. Then passing the parameters for the standard scaler to split the data for test and train accordingly. I have 1183 rows for train and 334 rows for test.





I have used algorithms like LinearRefressions, DecisionTreeRegressor, KNeighborsRegressor, SVR, Lasso, Ridge, ElasticNet and by the R2 score results and cross validation score, we could see that RandomForestRegressor is giving us 97% score (Average Price as Label) and 98% score (Region as Label), hence we could proceed with Random Forest.





7.Concluding Remarks:

Plotting scatter plot between test data and predicted data for Random Forest with Average Price and region as target. Therefore, I could dump our best predicted results in joblib.

