***AVOCADO PRICE PREDOCTION USING MACHINE LEARNING***

A picture containing fruit, indoor, avocado, green

Description automatically generated

Avocado is the fruit which is so popular among people all over the world. This fruit is priced high due to its high nutrient content and texture and loaded with healthy fats and fibers. we can include this avocado in so many foods like salad, pasta etc.

The most popular variety in avocado is HASS AVOCADO.

It contains wide variety of nutrients and has some 20 minerals and vitamins which is most of the B vitamins and folate, phosphorous, zinc.

One study conducted in US found that people who consume Avocado seems to be healthier than the people who don’t include this fruit in their diet.

But why we are discussing more about Avocado today? . The price of Avocado is high, and we will predict the price of Avocado and what are the factors causing the price of fruit higher?

And how we can predict this Price of fruit as we cannot go to each shop and can’t compare as it is highly Impossible!!

So here comes the *Machine Learning* which has some algorithms and with the help of these algorithms, we can predict the price of the fruit.

Let’s dive into the Machine Learning,

**Data Description-**

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 (units) 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 (of avocados) in the table reflects a 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. Other varieties of avocados (e.g., green skins) are not included in this table.

Date - The date of the observation

Average Price - 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

The Target variable is Price which we need to predict, using ML Techniques.

**Importing the Libraries –**

Graphical user interface, text, application

Description automatically generated

Pandas is open-source library tool which provides high performance data analysis tool by its powerful data structures.

It helps to shorten the procedure of handling the data with extensive set of features.

NumPy is most used package for scientific computing for multi-dimensional array of objects.

Standard Scaler used here to standardize the values to 0 to 1 in order to equalize the range of values as a preprocessing step.

Train Test Split is used to split the train data and test data and we will train our data in the given dataset and use test data to predict the output.

Grid Search CV is used for hyper parameter tuning to increase the model accuracy.

Cross Validation Score is used to check whether the model has been over fit or under fit.

Seaborn and Matplotlib is used here as a visualization library for the stunning plot to understand the data in a better way.

**Loading the Dataset –**

Graphical user interface, application

Description automatically generated

Loading the dataset and head () method will display the first 5 data from the dataset whereas tail () method displays the last 5 data from the dataset.

From the above snap, we can see that we have column – unnamed and we can drop this column as it will not impact our dataset in prediction.

Apart from this, we have date column and need to split day / month and we do have separate column - year already in our dataset.

We have few categorical columns in our dataset, and we need to convert that to numerical columns as well.

Text

Description automatically generated

Luckily, we do not have any null values present in our dataset and total 18249 rows and 14 columns present in this dataset.

Data has 0 values, and we will replace with mean, median values.

Here the main point to be remember is that we can perform mode () to categorical values only but not to numerical values.

So, let’s see how the data has been distributed in the dataset and what are the interpretations we can find from the visualization.

A picture containing shoji, window, crossword puzzle, building

Description automatically generated

So, it is the example of uni-variate analysis, and the above histogram shows the below observations for numerical columns only.

1. The target Variable – Price is distributed normally in the dataset.
2. Rest of the features has outliers/ skewness in the given dataset, and we need to treat it.

The below plot shows the ratio between types of avocado / year, and we can see that 2017 year is having high records.

A picture containing chart

Description automatically generated

Here we have two types of Avocados – Organic and Conventional.

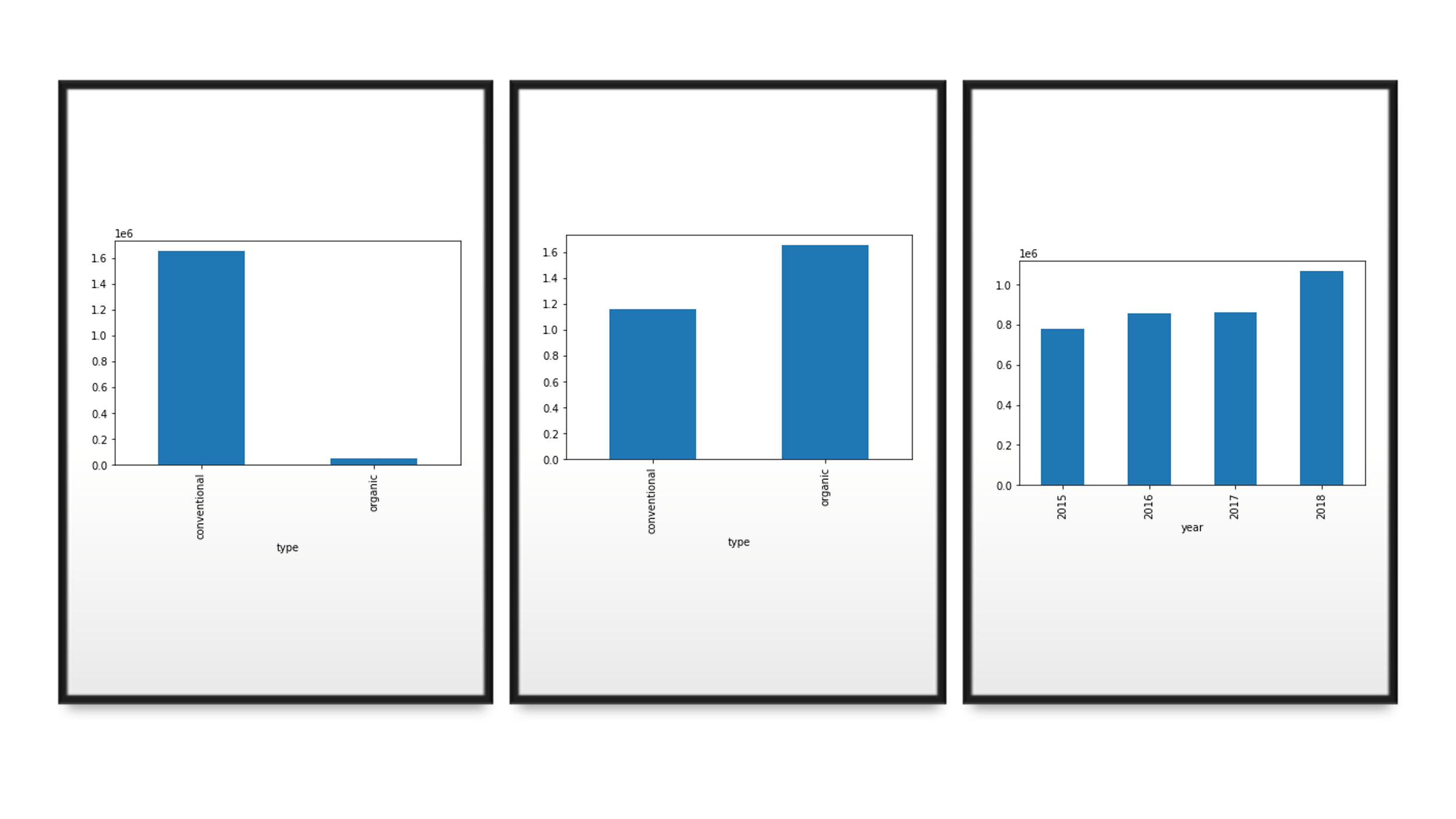
Icon

Description automatically generated with medium confidence

Chart

Description automatically generated with medium confidence

The above plots shows that September / October months has high price for Avocados in terms of price and the price of Avocados are extremely high specifically in 2017.



The above plot explains that Organic avocados are more than conventional in terms of price and year 2018 recorded high in Sales of avocados.

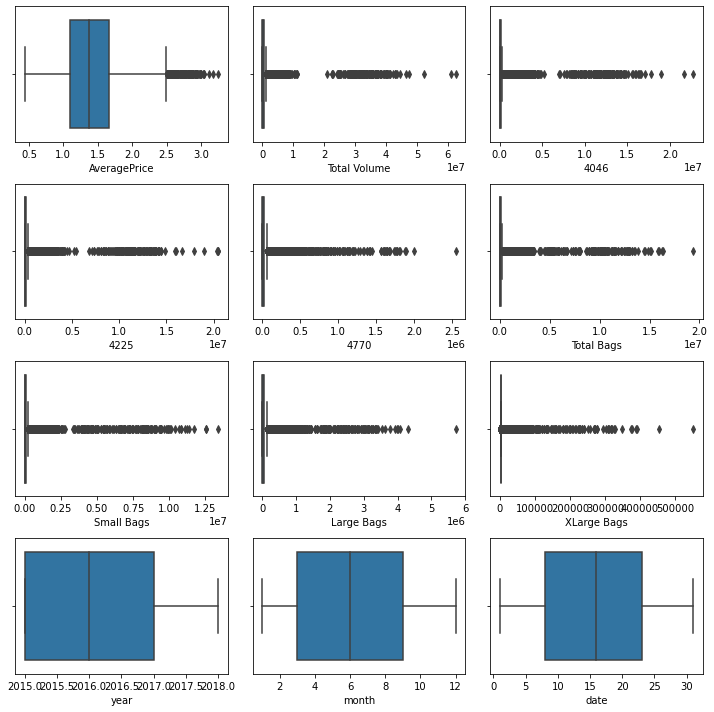
Conventional one is easily available than organic in terms of total volume and that's the reason, Organic avocados are selling high in price.

As we have “Type, Date, Region” as categorical data, we should encode it to Numerical columns.

1. Dropping the region Columns as it may increase the more column and it won’t cause any impact for our prediction.
2. Splitted the Date column values into day and month separately and dropping the actual column – Date.
3. We have one categorical column now which is Type and it has two values – Conventional, Organic.
4. Replacing as 1 and 0.

Here Our Feature Engineering part is over!!

Let’s check the outliers/ skewness removal part as we already discussed that we have outliers some columns.



So, here I am using Z- Score method for treating the outlier and loss of data % should not go beyond 8 -10 % and it is 3 % here.

Text, letter

Description automatically generated

After Outlier treatment we can check if there is any skewness exist and treat it with any transformation methods like power transformation, box cox, sqrt, cbrt, log function.

Then checking the correlation matrix if any features are correlated with each other,

Correlation Matrix is a statistical method of showing the relationship between two or more variables and the interrelation in their movements etc.

In short, it helps in defining the relationship and dependence among the variables.

A correlation matrix is “square”, with the same variables shown in the rows and columns.

Calendar

Description automatically generated

So, everything looks good and perfect, we will proceed for next step for splitting target and feature variables and perform scaling for the features variables in order to equalize the range of values to 0 and 1.

Next step is very important and interesting step, which is building a model,

Here I builded the model with the below Algorithms and our problem statement is to predict the price of the fruit and the data is continuous, it’s a Regression problem.

1. **K Neighbors Regressor-**

The KNN algorithm uses ‘**feature similarity**’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

Chart, scatter chart

Description automatically generated

As it’s a Regression problem, we will use r2 score, MSE, RMSE as metrics to calculate the accuracy score.

The above plot shows that line is linear, and we have data points which is nearer to that line and some data points are far from the line.

1. **Decision Tree Regressor –**

Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

**Chart, scatter chart

Description automatically generated**

The above plot shows that line is linear, and we have data points which is nearer to that line and most data points are far from the line as well.

1. **Linear Regression –**

Linear regression is an algorithm used to predict, or visualize, a [relationship between two different features/variables](https://www.statisticssolutions.com/what-is-linear-regression/).

Chart, scatter chart

Description automatically generated

We can see that line is not linear and data points are far as well and that shows the model score is less.

1. **Ada Boost Regression –**

An AdaBoost Regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

A picture containing text, electronics, screenshot, display

Description automatically generated

We can see here that line is not linear and can see more data points far as well and that shows that model score is less.

Out of all models, Decision Tree Regressor model is the best in terms of Cross validation Score.

Cross validation Score is nothing but the score which will check the model overfit or underfit.

Next, we will try to Apply hyper parameter tune for accuracy to increase the model score and depends on parameter and values we can increase our model score.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two Strategies are,

1. Grid Search CV
2. Randomized Search CV

**GridSearchCV**

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values.

This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

**RandomizedSearchCV**

RandomizedSearchCV solves the drawbacks of GridSearchCV, as it goes through only a fixed number of hyperparameter settings.

It moves within the grid in random fashion to find the best set hyperparameters. This approach reduces unnecessary computation.

It is fast than Grid Search CV.

We can check for actual and Predicted values and save it in a new data frame for to check whether our Predictions are same or not.

Final step is to Save the model pkl file.

Graphical user interface, text

Description automatically generated