**AVOCADO PRICE PREDICTION MODEL**

**Blog by:**

**UJJWAL PRATIK**

## Understanding Increasing and Decreasing in Avocados Price

Avocado has become the world’s trendiest fruit. This superfood is now a mainstay for foodies everywhere and the millennials love it. However, with increasing love for this fruit, the price went up sharply. You should have noticed that your avocado toast seems to be getting more and more expensive. The prices have increased sharply to up to 129%, with an average price of a Hass avocado reaching a price of $2.10 in 2019, almost doubling in one year.

## Problem Definition

Avocado is popular in United states, and it is in high demand for avocado toast. We know that the avocado toast price is skyrocketing because of the popularities among foodies. Seeing this it would be good if some one would be able to predict the price of avocado in various region of the US.

It will help us find where the toast of this fruit would be cheaper so that people should roam around that Using the model area and have their great mean.

Using the model I have developed for this problem, we will be able to predict the price of this fruit with very high accuracy.

## DATA ANALYIS

## About data:

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. It 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.

This data (From 2013) talks about 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.

## Some relevant columns in the dataset:

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.

## Shape:

Before we moved into the details of model I have created, let us understand the data on which the model is created.

The shape and data types of the of the dataset are as follows:

|  |  |
| --- | --- |
| Shape:  Columns: 14  Rows: 18249  Nulls: ZERO | Data types:  Object: 17  Integer: 2  Float: 9 |

## Exploratory data analysis

To get more insight from data let us perform the exploratory analysis.

We will check the Data types of the columns of the dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Column | Data types |  | Column | Data types |  | Column | Data types |
| Unnamed | int64 | Date | object | Average Price | float64 |
| Total Volume | float64 | Total Bags | float64 | region | object |
| PLU4046 | float64 | Small Bags | float64 | year | int64 |
| PLU4770 | float64 | Large Bags | float64 | type | Object |
| PLU4225 | float64 | X Large | float64 |  |  |

These are the columns, and its types present in the dataset.

Upon analysing Year with respect to Average Price and Type of the avocados data with the help of bar plot and count plot, I found that both Year and Type data is evenly distributed. Year 2015, 2016, 2017 and 2018 is almost in equal proportion in year column, but the Year 2017 is bit higher than the other years and type’s data also shows the same pattern here. Conventional and Organic avocados are also nearly equally distributed.

|  |  |
| --- | --- |
|  |  |

The code with which we can achieve the above plot:





The distribution plot of Average Price Column is showing a good distribution and looking like almost bell-shaped curve. This data does not require any correction. The Average Price is also showing the ranges of the distributions in between 0.5 to 3.0.

|  |  |
| --- | --- |
|  |  |

The code:



The distribution plot of PLU4046, PLU4225 and PLU4770 (column is indicating Total number of avocados with PLU, size of the avocado) is showing skewness towards left. This data requires some correction to take itself normal distribution, only if there are no data losses.

|  |  |  |
| --- | --- | --- |
|  |  |  |

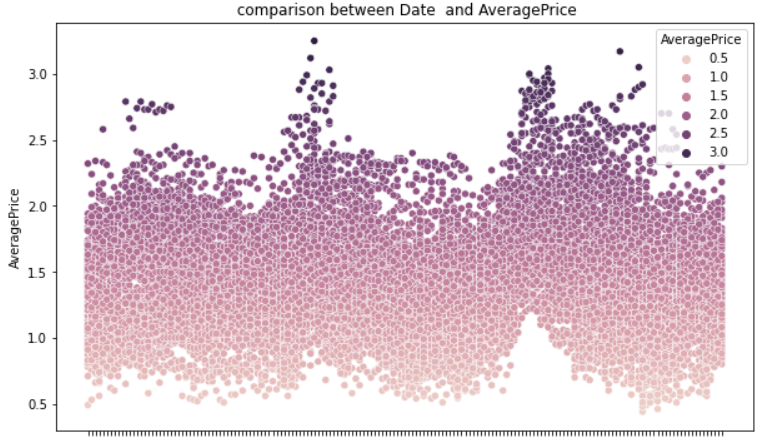
* PLU4046' column is indicating Total number of avocados with PLU (size of the avocado) and ranges in between 0 to .5.
* PLU4225' column is indicating Total number of avocados with PLU (size of the avocado) and ranges in between 0 to approx. 0.4.
* PLU4770' column is indicating Total number of avocados with PLU (size of the avocado) and ranges in between 0 to approx. 0.4.

The distribution plot of Small Bags, Large Bags and X-Large Bags is showing skewness towards left. This data requires some correction like PLU columns to take itself normal distribution, only if there are no data losses.

|  |  |  |
| --- | --- | --- |
|  |  |  |

* Small Bags ranges in between 0 to 0.2.
* Large Bags ranges in between 0 to 1.
* X-Large Bags ranges in between 0 to approx. 100000.

I have also analysed the comparison between Date and Average Price with the help of Scatter plot and I found that scatter plot is showing the positivity between the Date and the Average Price. It means that Average Price is increasing Parallelly to Date.



## Pre-processing of the dataset

We will discuss pre-processing pipeline of data in a bit detail here. I have done Label encoding, correlation matrix, statistical summary, skewness analysis and removal, and any outlier detections & removal.

Let me explain the label encoding first.

## *Label Encoding:*

As seen in the data analysis section, most of the data is of object type. We have a limitation in Panda library of Python to process object type of the data. Therefore, we have to convert object data into numerical data. This will also help us process data on outlier, skewness, correlations etc using panda package of Python.

Label encoding will help us achieving this.

Ex- After label encoding, data for Type columns which is object type is converted into numerical data type.



We have applied this label encoding on Date, Type and Region column.

## *Statistical summary:*

Statistics is a pillar of machine learning. You cannot develop a deep understanding and application of machine learning without it.

During Statistical summary I found,

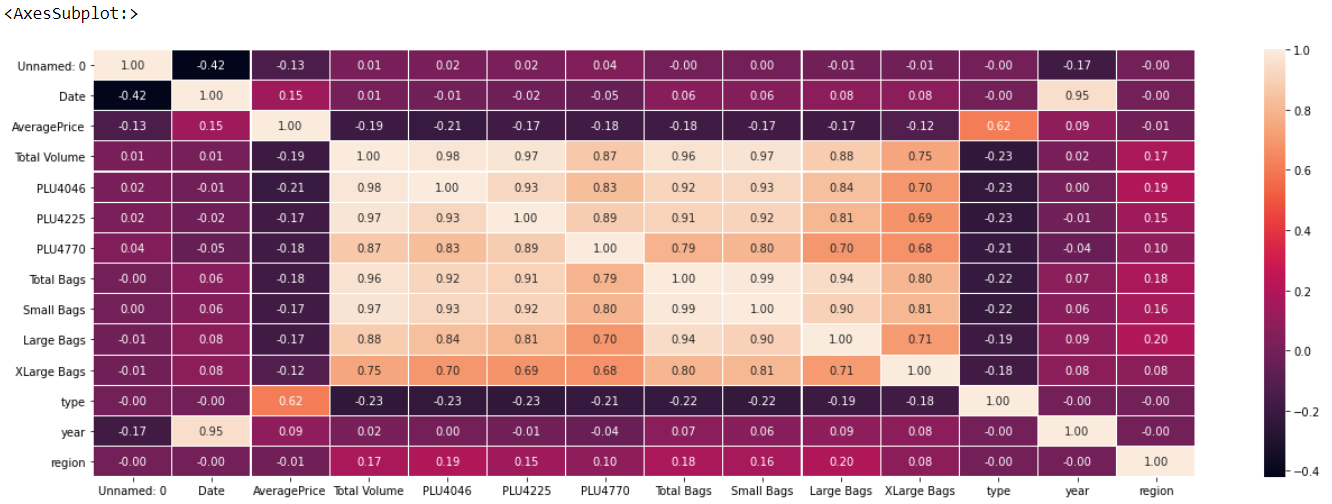
In some of the columns there is difference between mean and median (50%) and in some columns mean and median (50%) are approximately same.

In some columns there is very big difference between 75% and maximum indicating outliers are present in some columns.

Once statistical summary is done, I have moved towards correlation matrix.

## *Correlation matrix:*

This technique is very crucial to find out the correlation among the numerical variables.



By carefully observing the matrix I found that:

* Light shades are indicating Positive correlations and Dark Shades are indicating negative correlations.
* Most of the columns are showing negative showing correlations with target column Average Price.
* Type column is positively correlated with Average Price (target column).
* Most of the column is showing positive correlation between each other.

## *Dropping Unnecessary Column:*

I have dropped the Date column here because the Year is already present in the dataset.

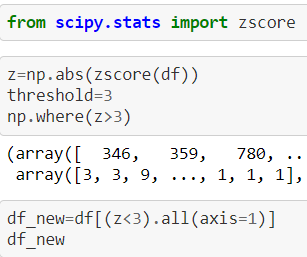
The code:



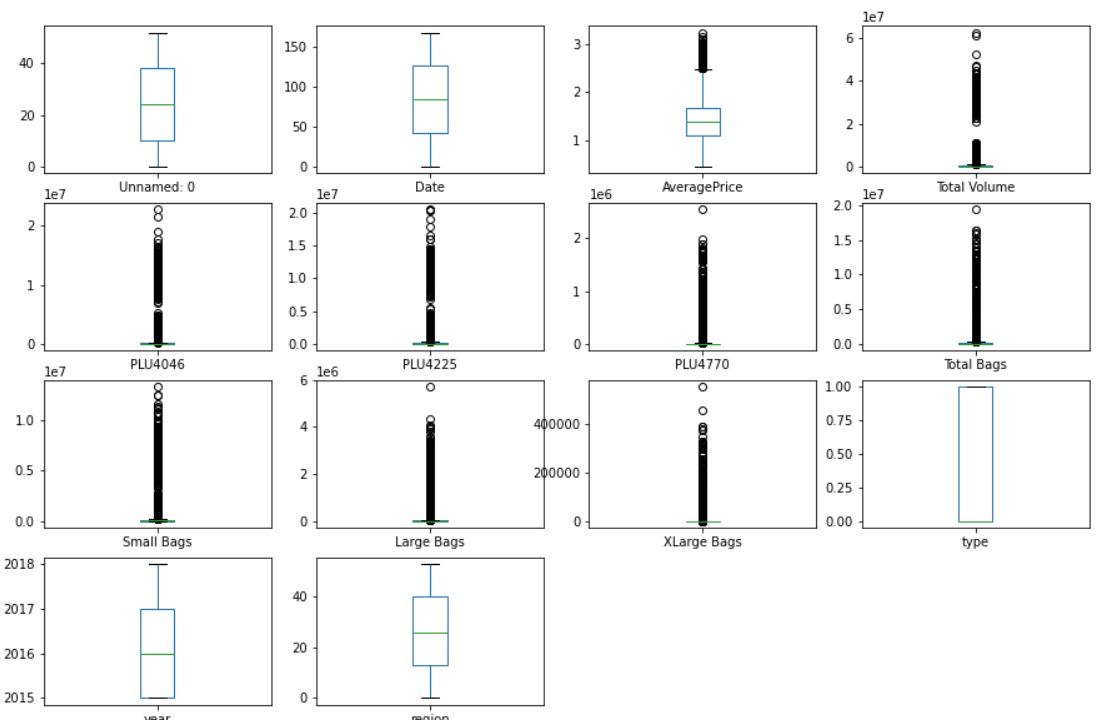
## *Outliers:*

Outliers are data points that are far from other data points. In other words, they are unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

The code:



During removal of outlier, Data was becoming Nan and I was losing the important data of the dataset, so outliers’ removal was not considered in the dataset.



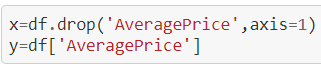
Outliers are present in the dataset, but it is near to the path that is why during removal of outliers the dataset was becoming Nan.

## Building Machine Learning Models for our dataset

With all the analysis and pre-processing we have done so far, it’s time to use the understanding of the data and build the data model.

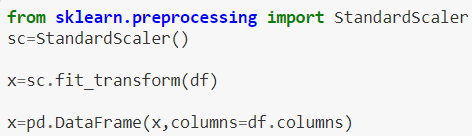
For Building Machine Learnings Model, we have to split our dataset into features and target variables that is x and y.

The code:



After that I am using Standard Scaling here for bringing all features to a common scale for getting better prediction.

The code:

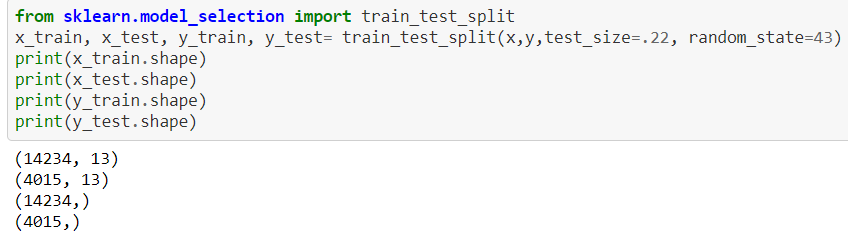


## *It’s time to split data into training and test data:*

Train test split is a function in Scikit-learn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually.

By default, Scikit-learn train test split will make random partitions for the two subsets.

The code:

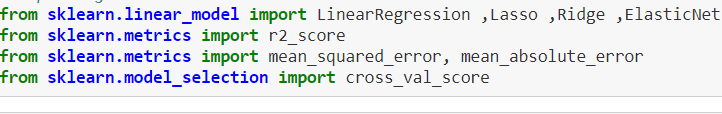


With the given code, I am splitting by taking test size 22 and random state 47.

## *Importing Libraries:*

In python, we require some libraries to be imported in our code in order to perform specified functions. The libraries we needed for our model building are:

The code:

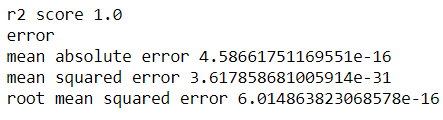


We have done all required things until now. Its time to build the data model now. We will test multiple algorithms to find our best model which has capability to predict accurately. We have taken all supervised machine learning regression algorithm here because we have to predict the Average Price.

## *Linear Regression:*

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear Regression gives 100% r2 score (accuracy). With cross validation applied on it by cross- validation-score to avoid overfitting, it remains same.



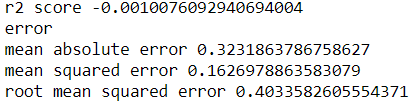
The code to generate the Linear regression model.

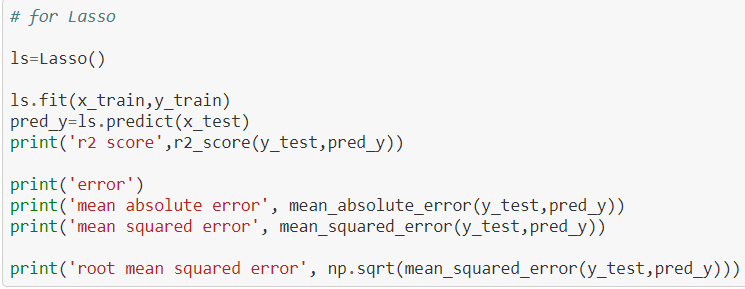
## 

## *Lasso Regularization:*

LASSO stands for Least Absolute Shrinkable and Selection Operator. As mentioned in the regularization definition, it is the process of adding information to prevent the over-fitting problem, it’s also helps into selection of features.

This model gives very low r2 score and very low cross validation score.

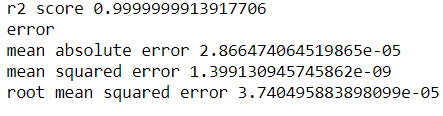


The code to generate the Lasso Regularization model.

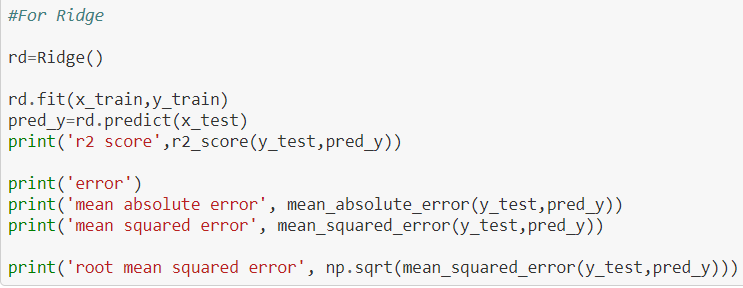
## *Ridge Regularization:*

Ridge regularization is another variation for LASSO as the term added to the cost function is as shown below. Cost Function of Ridge Regression Model in Ridge regularization, the penalty term can approach zero but will not be zero as it squares the coefficient (slope).

Ridge Regularization gives 99% r2 score (accuracy). With cross validation applied on it by cross- validation-score to avoid overfitting, it remains same.



The code to generate the Ridge Regularization model



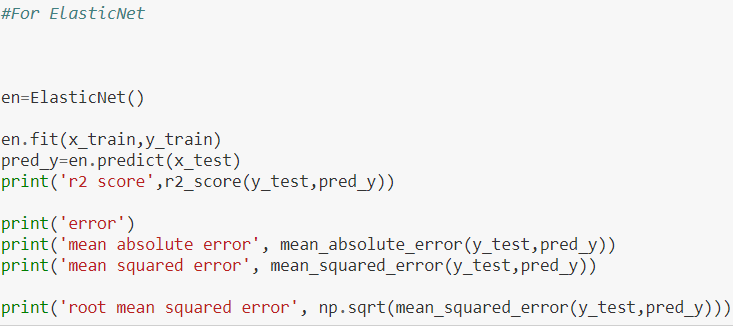
## *Elastic-Net:*

Elastic Net Regularization is a regularization technique that uses both L1 and L2 regularizations to produce most optimized output. This is one of the best regularization techniques as it takes the best parts of other techniques.

This model gives very low r2 score and very low cross validation score.

## 

The code to generate Elastic-Net.



## Hyper parameter tuning

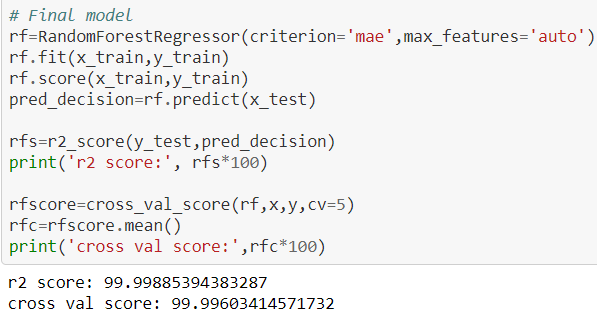
Finally, I have done the hyper parameter tuning which is being done for fine tuning of the model’s accuracy score. Hyperparameter tuning is one of the most essential knowledge for machine learning engineers and data scientists. I have taken Random Forest Regression as an algorithm and taken help of Grid Search CV for finding the best prediction.

After this fine tuning of the Random Forest regressor comes out very accurate and it gives us the r2 score (accuracy) of 99% and cross validation score of 99%.



The Process for Hyper Parameter Tuning of the model.

|  |  |
| --- | --- |
|  |  |



## Conclusion

With all analysis, pre-processing, model building and hyperparameter tuning, I recommend Random Forest Regressor as the best model of prediction with highest accuracy of 99%.

This Regression model will help people finding their favourite toast cheaper at various regions of United States. People will be able to predict very accurately with 99% and would be able to roam without worrying about their pocket.