**MACHINE LEARNING**

Avocado Average Price Prediction

Date: 17 April 2021

Author: Pooja Mishra

Contents

[Introduction 1](#_Toc69605529)

[Problem Definition 1](#_Toc69605530)

[Data Analysis 1](#_Toc69605531)

[Workflow 2](#_Toc69605532)

[EDA (Exploratory Data Analysis) 2](#_Toc69605533)

[Data Pre-processing 4](#_Toc69605534)

[Building Machine Learning Models 5](#_Toc69605535)

[Conclusion 7](#_Toc69605536)

Avocado Average Price Prediction

Regression Problem

# Introduction

Estimation of Avocado Average price prediction using Machine Learning Algoriths. With help of Machine Learning (ML) technology, we can predict price problems formulated as Regression analysis, in which a statistical technique used to estimate the relationship between a target variable i.e dependent variable and single or multiple independent variables.

This blog will focus on the ML algorithms performed on Avocado dataset to predict prices using different independent variables.

# Problem Definition

Avocado, is a dark green color botanically large berry containing a single large seed. It is originating from south-central Mexico. There is dozen variety of avocado, but more than 85% of Avocados harvested and sold in world are of Hass one. It heavily consumed by United States’ people.

In this blog, we’ll analyse the avocado price prediction. This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

# Data Analysis

This dataset consists of several columns mentioned below:

* Date - The date of the observation
* AveragePrice - 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 variables of the dataset are present in different types of data types as explained in the following:

* Categorical: ‘region’,’type’
* Date: ‘Date’
* Numerical: ‘Unamed: 0’,’Total Volume’, ‘4046’, ‘4225’, ‘4770’, ‘Total Bags’, ‘Small Bags’,’Large Bags’,’XLarge Bags’,’Year’
* Target: ‘AveragePrice’

The aim is to predict the average price of Avocados which is continuous in nature and using the all independent features present in the dataset. For that purpose, let’s confirm the type of problem, so it is regression type problem on which we’ll be performing several types of regression algorithms to predict avocado price.

# Workflow

To predict the average price of Avocados we’ll be performing following workflow to regularise the dataset followed with all regression algorithms that will help in predicting the prices.

Label Encoding

Data Cleaning

EDA

Identify the Problem

Importing Dataset

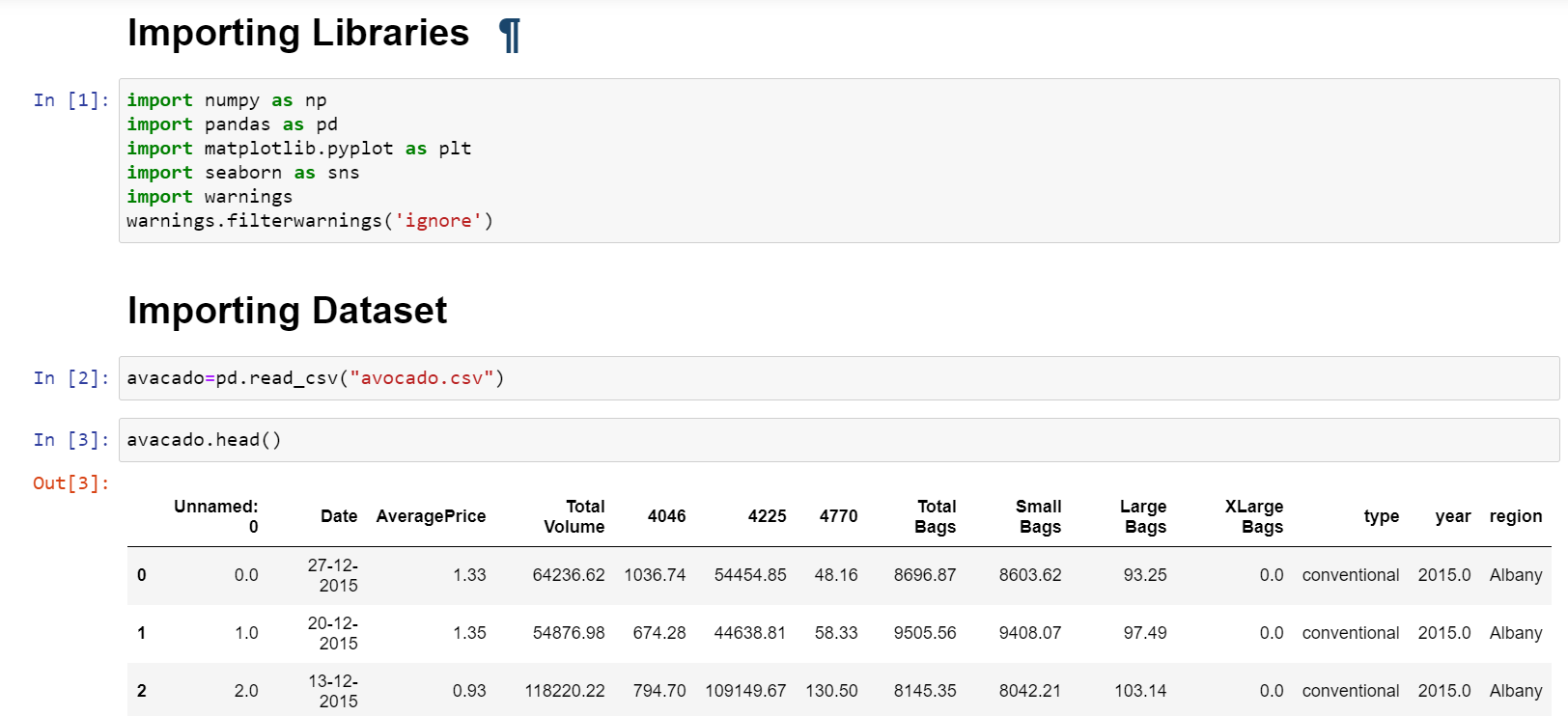
Hyper Parameter Tuning

Performing Algorithms

Find best Random State

Splitting in x & y

Saving the Best Model



In above image we have imported the initial libraries and algorithms libraries are imported later in this programming then imported the dataset so that we can perform further techniques to make the dataset ready prior implementing all algorithms.

# EDA (Exploratory Data Analysis)

Data Preparation:

To visualise the dataset of avocado we’ll perform initial steps to make data standard. We’ll be applying below mentioned activities to make the dataset ready for visualization.

1. Identifying the data types of all columns.
2. Checking total number of rows and columns.
3. Dropping the columns which will not participate in prediction of average price.
4. Checking the null values. Plotting its heatmap.
5. Replacing null values with mean or mode.
6. View the statistical summary of dataset.

After performing all above steps on dataset, we have observed following information about the dataset as well as rows and columns present.

Observations:

1. About the datatypes, we have observed that there are 3 are object type and 10 are float type.
2. Total number of rows present are 16468 entries along with 13 columns.
3. We observed that column ‘Unnamed’, ‘Date’ are not much participating in prediction of price thus dropped it. ‘Region’
4. Filling the null values in columns –

Date 14951

AveragePrice 14951

Total Volume 14951

4046 14951

4225 14951

4770 14951

Total Bags 14951

Small Bags 14951

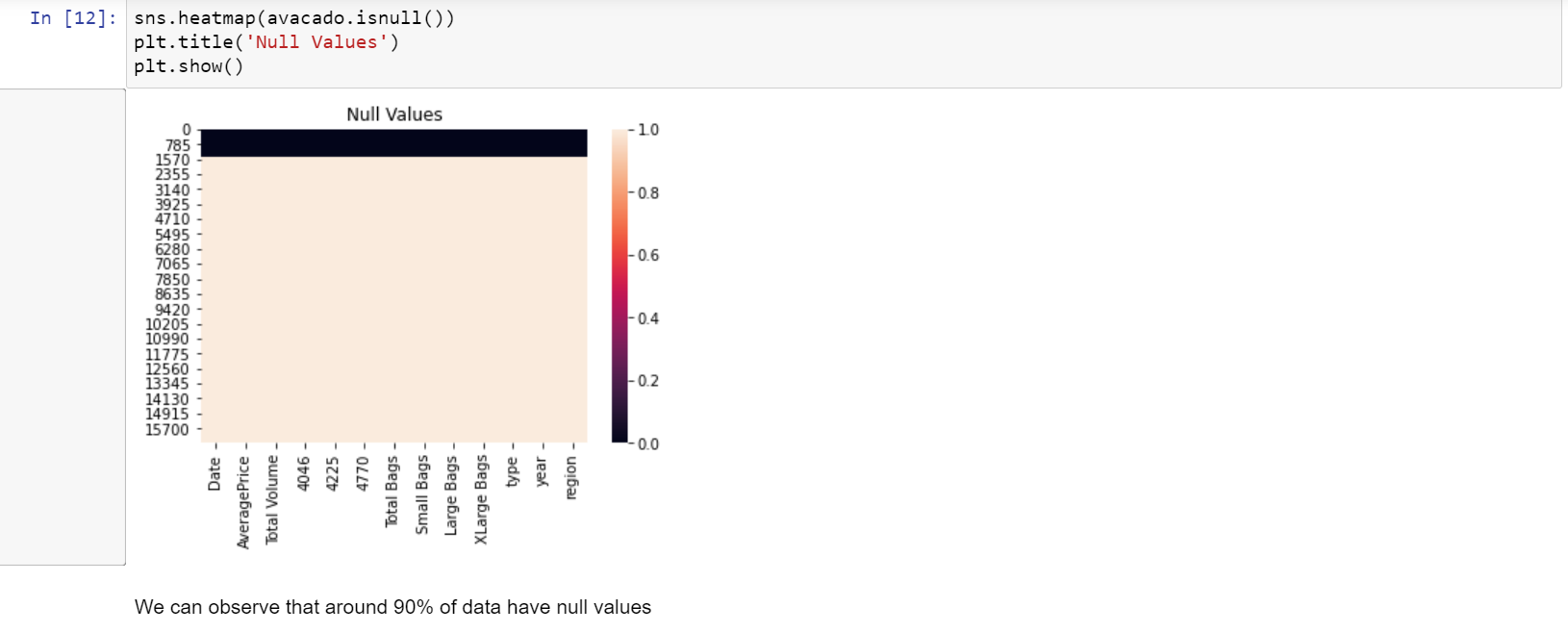
Large Bags 14951

XLarge Bags 14951

type 14951

year 14951

region 14951



From above heatmap we can observe that, around 90% of values are null because the white colour is equally distributed correspond to each other. Almost all the columns have null values present.

Removing Null Values:

Removing of null values is very important to operate algorithms on the dataset to get the predictions. The most common imputations are using of mean, medium, or mode to fill the values of each column. For categorical data, mode value will replace all null values, for non-categorical values use mean, median to fill the null values.

Similarly, in this dataset we have used the mode to fill all rows that have null values and mean for filling non-categorical rows that have null values.

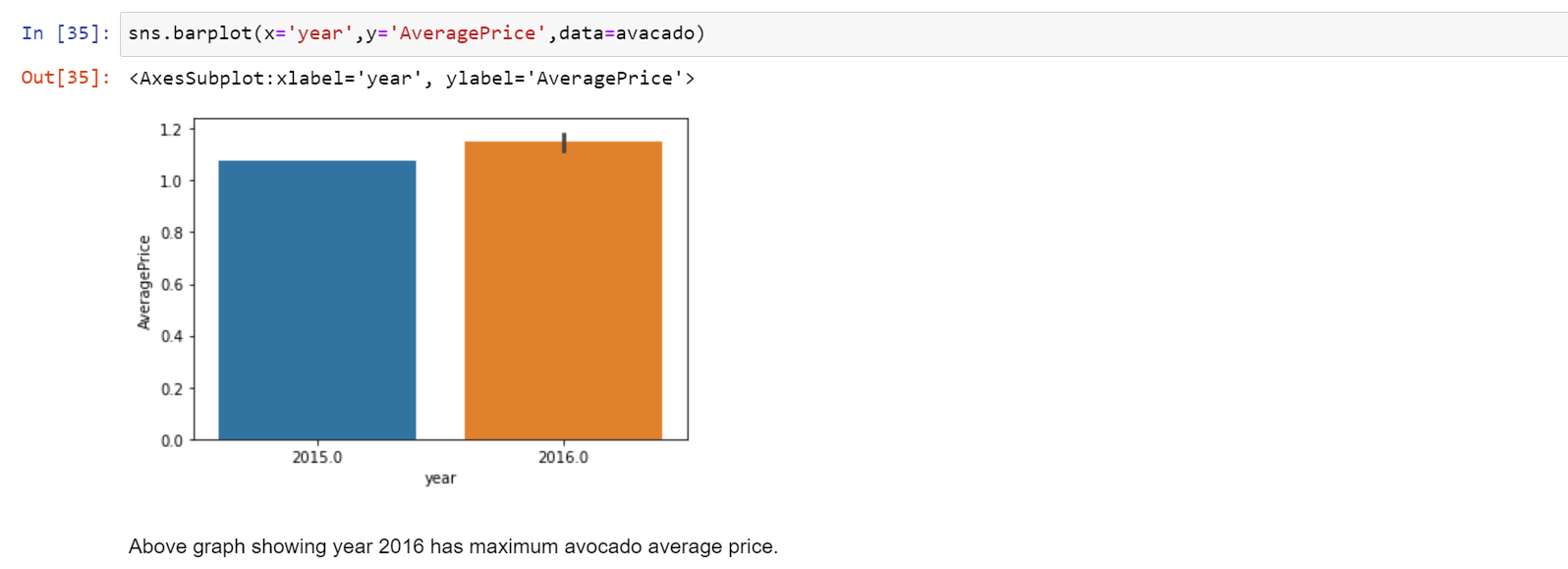
1. After applying mode and mean technique, all null values have been removed without any loss of data.
2. By applying statistical technique, we observed that the mean, 50% of data values and std are same or very close to each other. For few variations in 75% data values and max values, there can be outliers. Outliers present in columns need to treat later to make dataset ready for algorithm performance.

Data Visualization:

In this section we can plot different graph using different columns and try to visualize the data using matplotlib and seaborn library. We used different graphs to visualise the relationship of all features with target variable ‘AveragePrice’. We are univariant and bivariant analysis using barplot.

The histogram analysis performed on all variables showing the relationships with all variables. By histogram analysis we can observe the skewness in the dataset.

Barplot analysis done between ‘year’ and ‘AveragePrice’ showing year 2016 has maximum avocado average price.



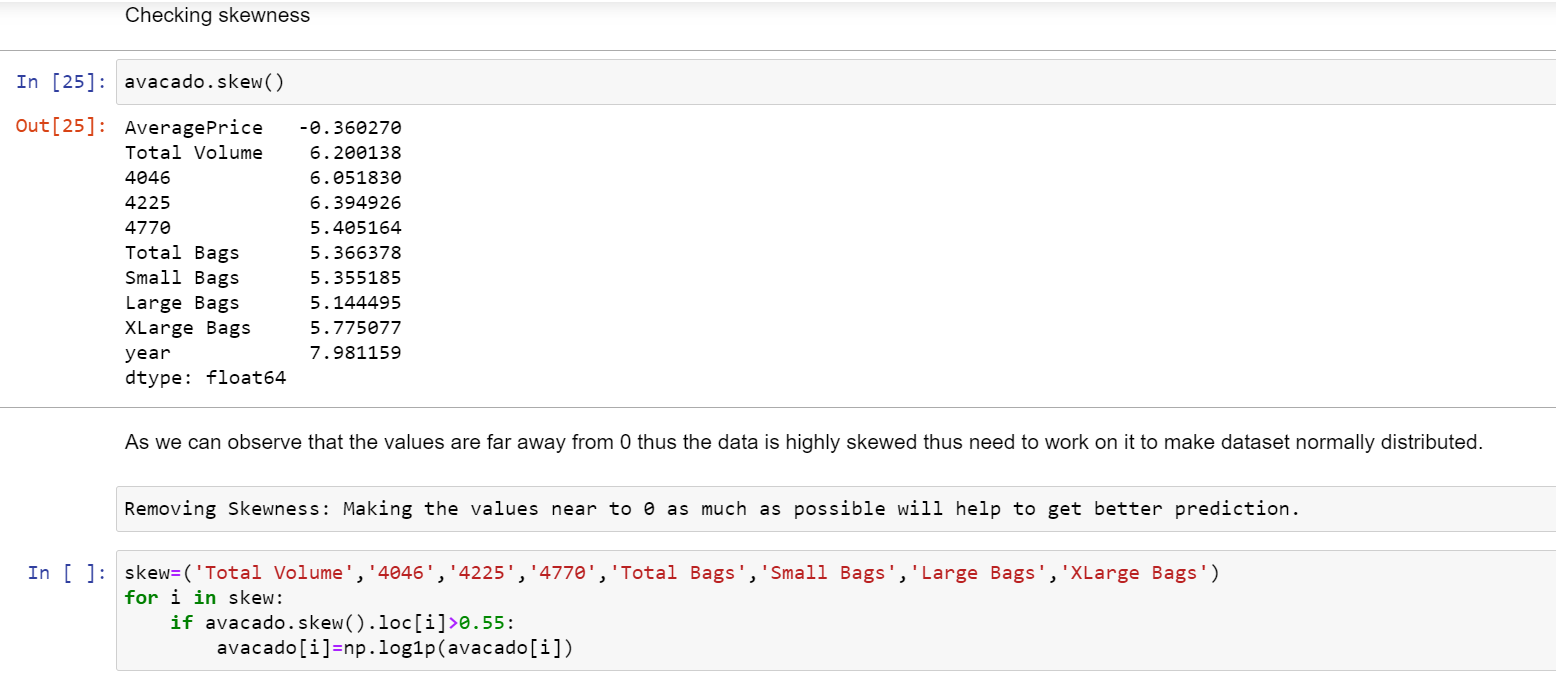
# Data Pre-processing

Data pre-processing is very important to get the dataset into the best format before performing algorithm. This is very important step in Machine Learning that should not be skipped. It involves three stages: Data Cleaning, Data Transformation, and Feature Engineering which converts complicated dataset into quality data.

Data Cleaning:

In Data cleaning, we understand the data, we perform several activities to clean the data. This stage comprises of following activities:

1. Dealing with Null Values – In this dataset, we have found more than 90% values are null so filled it with mean and mode as per the category of values as presented in above workflow step.
2. Dealing with Skewness – In this dataset, few of the columns do not follow the normal distribution that need to treat to convert it into normal distribution. In this step, we checked the skewness and used the NumPy log method to make the skew values into normally distributed.



In above image, you can see the columns which have skewed values, i.e. not normally distributed thus making the values near to 0 using NumPy log.

Outliers:

An **outlier** is a data point in a dataset that is distant from all other observations i.e. a data point that lies outside the overall distribution of the data set. Depending on the frequency of outliers, we’ll perform activities to remove it.

In this Avocado dataset, first we checked the outliers. We found outliers present in the dataset and performed zscore to remove it.

Data Transformation:

In Data transformation we recheck that all present columns are numerical or not? if not then we perform encoding type depending on the type of values present in columns.

Here in this dataset, we performed Label Encoding to convert categorical values in numerical.

Most important thing we need to keep in mind that all data pre-processing steps will be performed according the dataset present.

Correlation Matrix:

**Correlation** Matrix is a covariance matrix. A summary measure called the **correlation** describes the strength of the linear association.

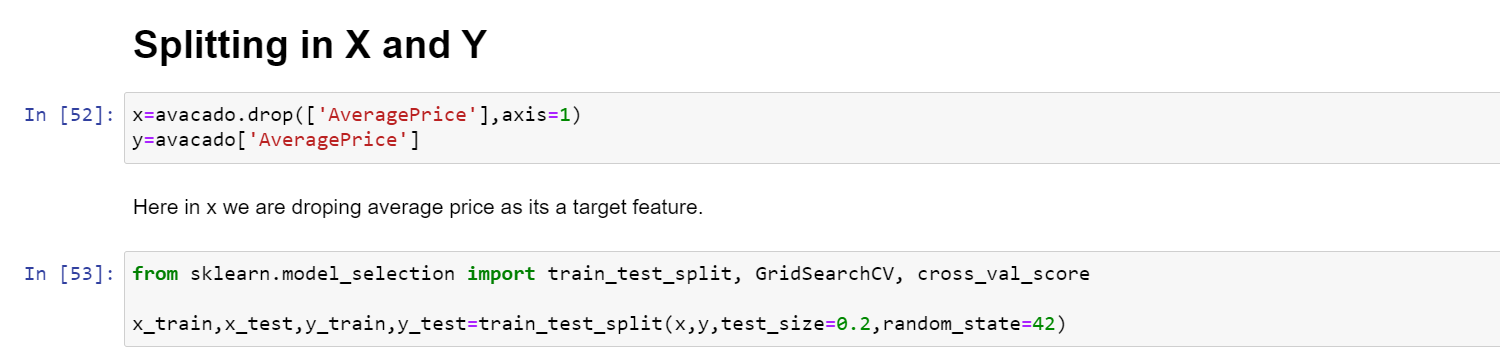
In this dataset, we have described the correlation and relationship present amongst each features with target variable. Here we observed that Total volume, Small Bags, Total Bags, 4225 showing highly positive correlation. With help of heatmap graph we can observe same the positive and negative correlation.

# Building Machine Learning Models

As our data are ready now! Let's now begin to train out regression model! We will need to first split up our data into an x array that contains the features to train on, and a y array with the target variable.

Prediction on Target variable i.e. ‘Average Price’

Here we are making two variable x and y where x is having all column except Average Price and Date, and y is having only Average Price column as shown in below image.



Creating and Training the model:

We had done this prediction by taking Average price as an output variable which is continuity in nature so that why using the regression techniques.

Now using multiple regression algorithms to find the best fit model for predicting the average price.

In this dataset, we have performed all regression type algorithms as mentioned below –

i)DecisionTreeRegressor

Using Decision Tree Regressor, we are getting accuracy of 68%

ii)KNeighborsRegressor

Using Kneighbors Regressor, we are getting accuracy of 73%

iii)AdaBoostRegressor

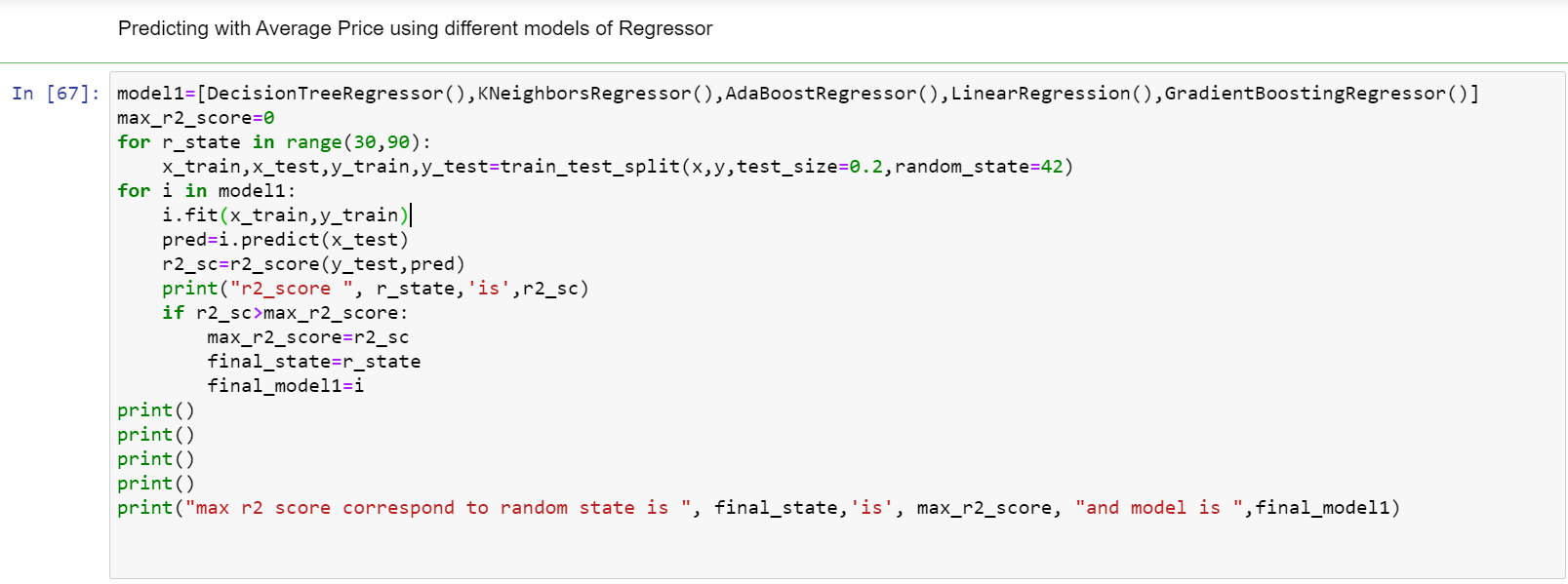
Using Ada Boost Regressor, we are getting accuracy of -117%

iv)LinearRegression

Using Linear Regression, we are getting accuracy of 30%

v)GradientBoostingRegressor()

Using Gradient Boosting Regressor we are getting accuracy of 65%



Thus, in this dataset we are getting highest score of 73% for model Kneighbors Regressor.

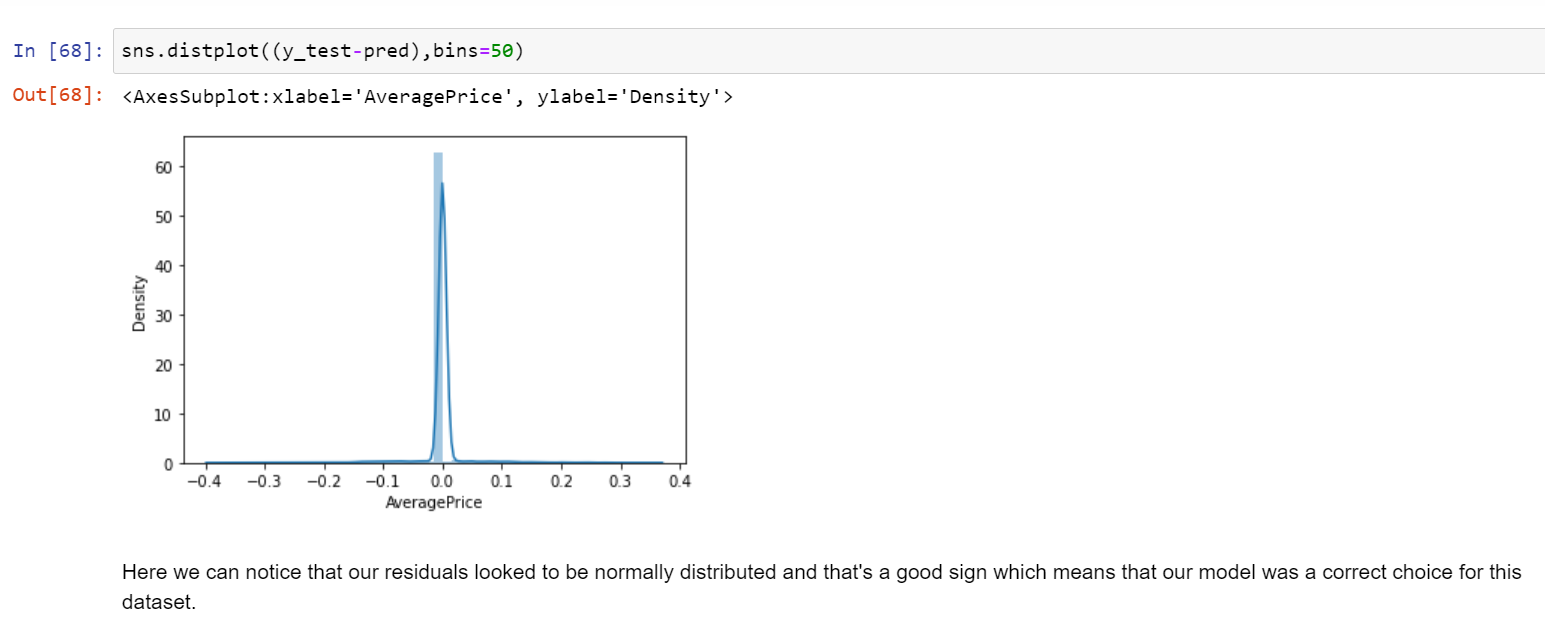
Hyper Parameter Tuning:

* After identifying the best model as Kneighbors Regressor, we used the GridSeachCV, so we can find the best param and then we used these param for that model.

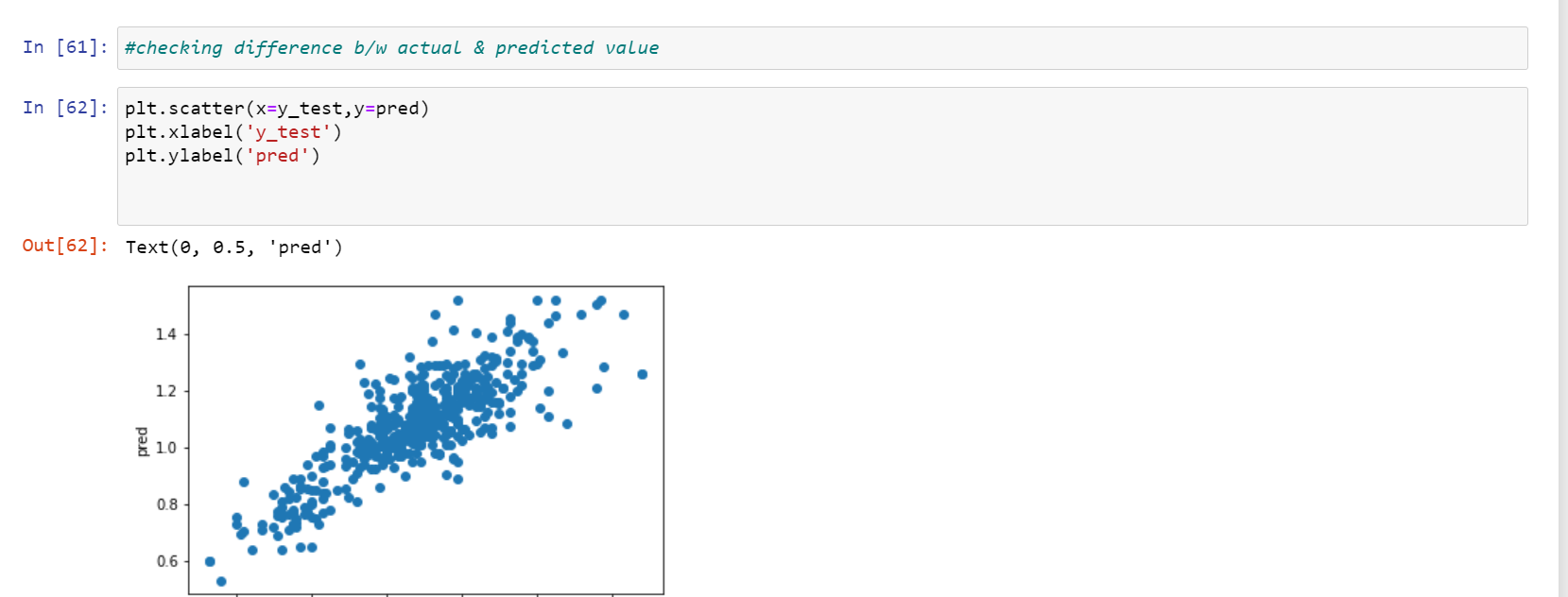
**Let’s see final Actual Vs Predicted sample.**

To confirm that we have chosen best model, lets plot normal distribution to confirm the model.

Here you can see the normally distributed model:



Now checking the difference between actual and predicted value, we plotted the scatter graph which shows positive correlation in actual and predicted values thus this model proves to be a best fit mode for this dataset.



Using pickle, we can save the best model KNN and can be useful to calculate the accuracy score and predict outcomes on new data.

# Conclusion

Thus, we can conclude that by performing all regression models, we came to know about the best model that predict the average price of Avocados. The Kneighbors Regressor model with 73% amongst all models proves that it’s the best model to go on. The normal distribution and positive correlation in graphs of actual and predicted value confirms the model is best fit.

Check out the ‘Avocado Average Price Prediction’ entire code on my Github profile:

<https://github.com/github-pooja/Data-Trained-Practice-Projects/blob/main/Avocado%20Price%20Prediction%20%20(1).ipynb>

Thank You.