# **DATATRAINED ACADEMY**

# **Blog Article – 2**

Submitted by: Sindhu Shree N

Batch Number: 1832

# **Avocado Project**

**Problem Statement:**

**Avocado is a fruit consumed by people heavily in the United States.**

### Content

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

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
  1. **Problem Definition:**

Avocado price data includes observations from 2015 to 2018 and was originally extracted from Avocado project and downloaded the dataset covers the .csv files average prices, types (conventional or organic), and cities and regions where avocados were sold. The goal is to predict the average price which is continuous in nature of the different type of avocado and using the region that in which region they are lying.

* 1. **Data Analysis:**

The first stage of this analysis is to describe the dataset (where we can find mean, median, standard deviation, minimum and maximum values set in a table), understanding the meaning of each variable, detecting possible patterns and performing the necessary adjustments to ensure that the data will be proceeded correctly during the machine learning process.

**Data Preparation and Cleaning:**

* Reading the CSV file and doing initial statistical analysis (shape, values etc)
* Data Pre-processing: Reading the unique values for each column and removing those which won’t be significant in the analysis further.
* Create a new data frame to proceed with the analysis further.

Historical data that was selected for solving the problem must be transformed into a format suitable for machine learning. Since model performance and therefore the quality of received insights depend on the quality of data, the primarily aim is to make sure all the data points are presented using the same logic, and the overall dataset is free of inconsistencies.

Dataset contains:

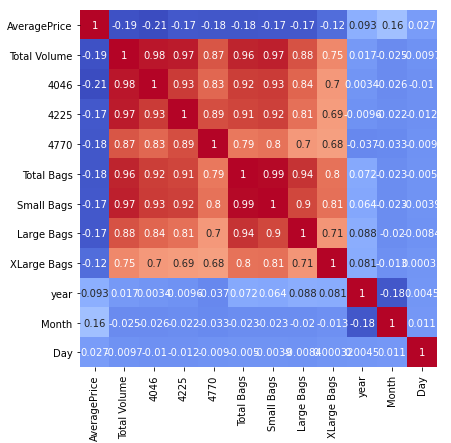
* Date — The date of the observation
* Average Price — the average price of a single avocado
* type — conventional or organic
* year — the year
* Region — the city or region of the observation
* Total Volume — Total number of avocados sold
* 4046 — Total number of avocados with PLU 4046 sold
* 4225 — Total number of avocados with PLU 4225 sold
* 4770 — Total number of avocados with PLU 4770 sold

This project is based on a hypothetical dataset downloaded from Avocado project it has

18,249 data points and 13 features (dependent and independent variables) describing each fruit type and region they are consumed in high level; and labelled (supervised learning)

I had a hypothesis of demand and supply, which means, in other words, that if the consumed volumes are higher, then the prices would be lower. The scatter plot I plotted using matplotlib in Python libraries displays, that it seems there is a trend for that direction. The Pearson correlation coefficient showed a small negative correlation between the average price and average volume consumption. Thus, there is an association between demand and supply, but that cannot explain everything about how the prices are structured. Having some outliers on the right side of the plot where some cities had the highest prices while the consumed volumes are limited.

The correlation between different features of the dataset using heat map is shown below:



The correlation matrix does not indicate any high degree of correlation with the dependent variable. However, it does provide us with a holistic view off all the factors.

* As we can from the heatmap above, all the features are not correlated with the target feature **Average Price**, instead most of them are correlated with each other. So now I am bit worried because that will not help us get a good model. Let’s try and see.
* First, we have to do some Feature Engineering on the **categorical Features: region and type**
  1. **EDA concluding Remark:**
* Finding the patterns of data through visualization and reveal the hidden trends from data.
* Using both matplotlib and seaborn library to visualize the data.
* Finding relationships between features using bar graphs, histograms, box plots, heatmap.
* Analysing both the numerical and the categorical columns separately.
  1. **Pre-processing Pipeline:**
* Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.
* This is required to do since the machine learning algorithms only work on the numerical data. That is why there is a need to convert the categorical column into a numerical one.
* For each of the transformations in Python that has a fit\_transform() method , we can wrap them up in an actual pipeline that executes them all in order and even go back and view attributes of each of the transformations. Additionally, you can set parameters for each transformation and the syntax for that is in the link I just shared. If anything, using this pipeline has just cleaned up my code a lot.
* Sklearn provides a very efficient tool for **encoding** the levels of categorical features into numeric values. **Label Encoder encode labels** with a value between 0 and n\_classes-1 where n is the number of distinct **labels**. If a **label** repeats it assigns the same value to as assigned earlier .Convert Region and Type into numeric value by using encoder.
* Data has to be pre-processed as machine learning models are better at reading numbers than words. Using label encoding, categorical data can be replaced with numbers. Below code is to display all categorical data.

# .**5. Building Machine Learning Models:**

The main goal of this project is to develop a model which predicts average price of Avocado fruit. Before applying the machine learning algorithms we need to split the data in the way that 80% of data should go the training phase and rest of the data should go to the testing phase. This is done by using “train\_test\_split” function in sklearn model selection.

Later we use cross validation in training the models, and each baseline model performance can be tabulated.

The model will be cross-validated using a 5-fold cross validation returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the train\_test\_split.

Classic machine learning models are commonly used for predicting averageprice, for example, linear regression, decision tree, random forest and others. Using Random forest as a baseline model. We generally use a baseline model’s performance.

* Let’s apply our model which is going to be the **Regression algorithms because our Target variable 'AveragePrice' which is continuous.**
* Let's now begin to train our regression model. We will need to first split up our data into an **X array that contains the target variable**, and a **y array with the target variable**.

**Regression models:** Avocado fruit average price prediction can be formulated as a regression task. Regression analysis is a statistical technique to estimate the relationship between a target variable and other data values that influence the target variables, expressed in continuous values. If that’s too hard- the result of regression is always some number, In addition, regression analysis allows for estimating how many different variables in data influence a target variable. With regression, businesses can forecast in what period of time a specific region has a highest (or lowest) AveragePrice of the fruit.

Here we are using some of the regression models and they are:

1. Linear Regression:
2. Decision Tree Regressor
3. Gradient Boosting Regressor
4. Random Forest Regressor

* I had done this prediction by taking Average price as an target variable which is continuity in nature so that why I’m using the regression technique
* While calculating the best random state the 42 is best state which providing the highest R2 score value for this model.
* After using the GridSearchCV, I can find the best parameters and then I used these parameters for that model.
* There are following matrices which I find, and which are providing the best score.
* I have also plotted the scatter plot graph and the actual value and predicted values are very close to each other, so the line is best fit line.

Now I am finding the score by taking region as an target value, I am using classification method because the region data is categorical in nature, so I am importing the classification models and their matrices.

The final model of the dataset: So, utilizing data science on AveragePrice provided significant benefit to the business as we can tag each fruit with the averages price and come up with customized Avocado retention strategy for each group.

According to the Regressor report the accuracy of the model is 85% however its recall is lower at 15% of positive cases. The Random Forest Regressor model is providing excellent results, however the purpose of the problem is to identify the average price of the fruits. This is the reason that accuracy then becomes a very important measure. Accuracy measures the fraction of values that are identified correctly.

Also Random Forest Classifier has emerged as the final winning model with 90% and highest. This could be the highest possible score achieved with the inherent limitations in the dataset. Here Random Forest of both regression and classification is the best model.

Machine learning models are as good as the data to feed it, and more data would strengthen the model. For example, in this dataset.

**6. Concluding Remarks.**

* With the help of notebook I learnt how **EDA** can be carried out using **Pandas and other plotting libraries**.
* Also, I have seen making use of packages like **matplotlib and seaborn** to develop better insights about the data.
* I have also seen how **preprocessing** helps in dealing with **missing values and irregularities** present in the data. I also learnt **how to create new features** which will in turn help us to better predict the survival.
* I also make use of **pandas profiling** feature to generate an html report containing all the information of the various features present in the dataset.
* I have seen the impact of columns like **type, year/date** on the **Average price increase/decrease rate**.
* The most important inference drawn from all this analysis is, I get to know what are the **features on which average price is highly positively and negatively correlated with.**
* This project helped me to gain insights and how I should go with flow, which model to choose first and go step by step to attain results with good accuracy. Also get to know **where to use Linear, Decision Tree and other applicable and required models to fine tune the predictions**.
* Random Forest Regressor model predicts the average price more accurately than other regression model.
* In this project the trend and periodity of avocado price and sales volume time series and also analysed their association .we extracted monthly and annual patterns from the spectrum density analysis ,and also determined the trend of price variation from the spectrum decomposition ,which is not constantly increasing but shows a decreasing trend in recent years.in addition,we applied a regression on the price and sales volume time series and discovered a negative correlation between the two time series. Which is consistent with our empirical knowledge.
* The visualisation we were dreaming of at the beginning of this project has now become a reality.