Student name: Avagyan

Date of submission: June 27, 2022

Revised abstract

Avocado price prediction

The high fat and smooth texture of avocados make it a useful and diverse food in different cuisines. In 2018, the [US Department of Agriculture](https://en.wikipedia.org/wiki/US_Department_of_Agriculture) estimated that 231,028 hectares (570,880 acres) in total were under cultivation for avocado production in

Mexico, a 6% increase over the previous year, and those 2 million tonnes would be exported. Seventy-six percent of Mexico's avocado exports go to the United States.

Most Mexican growers produce the [Hass variety](https://en.wikipedia.org/wiki/Hass_avocado) due to its longer [shelf life](https://en.wikipedia.org/wiki/Shelf_life) for shipping and high demand among consumers.

The traditional consumers who buy it for health reasons have contributed to the rise in demand over the years, the intensifying awareness of its nutritional value and numerous health benefits among the growing middle class with their rising income and dormant lifestyles has been significantly responsible for driving the interest in the superfood.

The goal of this project is to predict the trend of avocado prices, organic and non-organic for US cities for next year.

Another goal of this project is to identify which cites in the US have the cheapest avocado.

Calculate the predicted price per person for the next year (2022) with a consumption rate of minimum 1 avocado per day for the 5 chippiest cities in US.

Historical data on avocado prices and sales volume in multiple US markets

The data consists from 3 csv files for 2019, 2020 and 2021 years. has combined 17 936 rows and 14 variables, the dataset is from The Hass Avocado Board.

The techniques and the tools are proposing to use to solve the problem:

\_ data exploration, cleaning and manipulating

\_ exploratory analyses and predicting the avocado price

\_ multiple liner regression

\_ decision tree regression

\_ random forest regression

\_ Python

\_ R

Note: The reason to revise the submitted abstract is using the current datasets which is more actual. Instead of 2016, 2017, 2018 years’ data it is preferable to use data for 2019, 2020 and 2021 periods.

Link to this project in github to see the codes:

<https://github.com/aavagian/aavagian>

Literature Review, Data Description, and Approach

The data set was downloaded from Hass Avocado Board (HAB). HAB’s strategic priority is to support fresh Hass avocados becoming the most popular and desired fruit by 2025 in US. The price prediction is important to compare the avocado users’ income changes with avocado’s average price changes.

<https://hassavocadoboard.com/category-data/>

The dataset consists of three yearly based different csv formatted datasets for 2019, 2020 and 2021 that summarize the weekly sales of Hass Avocados at different regional levels and cities for each year.

Data description and importing

Text, application

Description automatically generated

Index

Geography \_ Locations are selected cities and Regions

Timeframe\_ Weekly bases

Current Year Week Ending\_ Current Year Week Ending the date of the observation

Type\_ Nonorganic (conventional) and organic

ASP Current Year\_ Average Sales Price for a single avocado

Total Bulk and Bags Units\_ Sales Volume including Bulk and Bags units (total number of avocados sold)

4046 Units\_ Sales Volume of avocados with PLU 4046 (Small/Medium Hass Avocado, non-organic), The most commonly sold sizes of fresh Hass avocado can be identified by their Price Look Up code or PLU

4225 Units\_ Sales Volume of avocados with PLU 4225 (Large Hass Avocado, non-organic)

4770 Units\_ Sales Volume of avocados with PLU 4770 (Extra Large Hass Avocado, non-organic)

TotalBagged Units\_ Sales Volume of bagged avocado, Bags indicates a pre-packaged container consisting of a variable number of avocados of mixed PLU type

SmlBagged Units\_ The amount of sold non-organic Hass avocados (they can be a mix of PLUs) sold in small bags

LrgBagged Units\_ The amount of sold non-organic Hass avocados (they can be a mix of PLUs) sold in large bags

X-LrgBagged Units\_ Sales Volume of XLarge Bags

Bulk GTIN\_ Bulk typically means avocados sold as individual pieces

Data exploration

The output shows that dataset has 17936 entries with 14 column, no null values. "Current Year Week Ending" columns’ data type should be changed from object to data.

The result of generating count statistics of duplicate entries shows “No duplicated rows”.

Text

Description automatically generated

Text

Description automatically generated

Categorical variables

Monitoring the messy “Geography” column

The “Geography” column from the dataset is a mix of cities, states and generic location such as “West” and “Northeast”. I will create a new column located before the “Geography” column and having it to group cities, stats and generic location if necessary.

Besides cities dataset has 2 states \_ California and South Carolina, 6 regions\_ Midsouth, Northeast, NorthernNewEngland, SouthCentral, Southeast, West and also Total U.S. .

To plot data to see the ASP variation between cities and regions. The plot shows, that the average price is mainly between 1.25 and 1.50.

Dataset has 2 types of avocado\_ non organic(conventional) and organic. The first boxplot shows the average price difference between 2 types of avocado, the second boxplot shows the volume difference between soled organic and soled non organic avocados, the non organic avocado has significantly big demand if compare with organic. So, this is an other sign, that the ASP prediction is important since it has significant influence on users' behavior.

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

The column "Timeframe" is good to have in original dataset to understand the frequency of data compiling, but it is extra for the future since has only 1 unique value.

To delete the column "Timeframe".

The dataset for 2021 has less rows. It doesn't include data for December 2021, the last week for 2021 is 2021/11/28. The project’s target is predicting ASP per cities for next 2022 year. I think the missing data for December 2021 (1 only month) is not critical since dataset has 35 months' information. So tis is not an issue to have influence on accuracy.

Numerical variables

To rename columns’ names to make it clearer and shorter.

Text

Description automatically generated

The “Total Volume” is the combination of the columns “4046”, “4225”, “4770”, "TotalBagged" and "Bulk GTIN". See the numbers of first row:

3541.90+95713.80+149.31+10904.73+967.73 = 111277.469

The "TotalBagged" doesn't equal to the sum of "SmlBagged", "LrgBagged" and "X-LrgBagged". It is preferable to remove these columns since the data is not full.

The cleared dataset has below view:

Table

Description automatically generated

Comparison with previously done projects

There are already completed projects about Avocado price prediction. The accessible for review projects have been completed based on old datasets (for years 2016, 2017, 2018). The target of published projects and this project is almost the same to predict the future price of avocado’s unit since it became more popular in North America and the significant average price change can change the users’ behavior.

There are some differences between columns if compare the old data based published projects and this project, but it has no influence on the target of projects for both cases.

The approach to prepare this project’s database and the data preparation approach used in completed projects are almost the same.

Taking account, the unpredictable high level of inflation during 2021, it is probable to have significantly different outcome for this project from the published projects because of timeframe.

Some sources for published Avocado price prediction projects:

<https://rpubs.com/kar_ng/786249>

<https://github.com/abhijitshow07/Avocado-Price-Prediction-in-USA/tree/5a982dbc83eff885de581061fe87cd33e9ef0f19>

Link to this project in github to see the codes:

<https://github.com/aavagian/aavagian>

Initial Results and the Code

Create new columns with ‘year’ and 'month\_year' based on column "cy\_week".

Average Price distribution shows that for most cases price of avocado is between 1.1, 1.7.

Chart, histogram

Description automatically generated

Average Price distribution for organic avocado shows that for most cases price of avocado is between 1.5, 1.7.

Chart, histogram

Description automatically generated

Average Price distribution for nonorganic avocado shows that for most cases price of avocado is between 1.0, 1.15.

Chart, histogram

Description automatically generated

Plotting the Avocado's Average Price tendency during 3 years.

Chart, line chart

Description automatically generated

Plotting the Avocado's Average Price through the weekly period for 3 years to see the fluctuation. Seasonal changes can have an influence on prices

Graphical user interface, chart, line chart

Description automatically generated

The list of 5 cities where prices are low

1. New Orleans/Mobile, 2. Tampa, 3. Phoenix/Tucson, 4. Dallas/Ft. Worth, 5. Orlando

Chart, bar chart

Description automatically generated

The list of 5 cities where the usage is very high

1. Los Angeles, 2. New York, 3. Dallas/Ft. Worth, 4.Phoenix/Tucson, 5. Houston

Chart, histogram

Description automatically generated

The plot shows the price change tendency during 2019, 2020, 2021 for each city.

A picture containing timeline

Description automatically generated

The pair plot clearly shows strong correlation between Total Volume and other parameters so Linear Regrssion will be used, but the correlation matrix shows that "Average\_price\_per\_avocado" has no correlation with other components of dataset. The Liner Regression model will help to understand more.

Chart

Description automatically generated

A picture containing icon

Description automatically generated

Application, PowerPoint

Description automatically generated with medium confidence

Linear Regression, creating and training the model.

The objective of Linear Regression is to find a line that minimizes the prediction error of all the data points.

Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset.

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set

Root Mean Squared Error is the square root of Mean Squared error

The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model.

MAE: 0.21208523891188677

MSE: 0.07206664623864492

RMSE: 0.2684523165082487

The RMSE is low so the model is good but this is not enough to have a final opinion. The test vs pred plot shows not a straight line.

Chart, scatter chart

Description automatically generated

To use Decision Tree Regression model

Chart, scatter chart

Description automatically generated

MAE: 0.14799338399804907

MSE: 0.04872486725693925

RMSE: 0.22073709986529055

The results (both plot and MAE, MSE, RMSE) show better correlation, so the Decision Tree Regression model is more accurate than Linear Regression model.

To use Random Forest Regressor model

MAE: 0.11122494643897993

MSE: 0.02298790552987485

RMSE: 0.1516176293505305

The RMSE is lower than the two previous models. So Random Forest Regressor model is the preferable one.

The residuals looked to be normally distributed which means that our model can be a correct choice for the data.

Chart, histogram

Description automatically generated

Actual Vs Predicted sample

Chart, scatter chart

Description automatically generated

Link to this project in github to see the codes:

<https://github.com/aavagian/aavagian>