FINAL PROJECT REPORT

Sai Bharath Kumar Bandi, Sachin Dattatraya Hedge, Martin Bautista, Mohmed Hira

CS 418 – Introduction to Data Science

INTRODUCTION:

The objective of this project is to implement all stages of processes in Data Science pipeline by choosing a dataset. We are using the dataset (avacado.csv). This dataset contains the details of the hass avocado price, sales and volumes sold over a period of 3 years. The observations are made weekly. Python 3 is used as a programming language for the given tasks and Jupyter Notebook for implementation.

PROBLEM STATEMENT:

Predict the price of the hass avocado based on sales and volumes. Also classify the avocado into conventional or organic.

DATA COLLECTION:

Data is collected from the below link:

https://www.kaggle.com/neuromusic/avocado-prices

- The dataset contains the details of average price of hass avocado in different counties of the United States from Jan 2015 to March 2018.
- The dataset contains a total of 18249 observations and 14 variables
- The variables info is as follows:
  - Unnamed: 0 - int64
  - Date - object
  - AveragePrice - float64
  - Total Volume - float64
  - 4046 - float64
  - 4225 - float64
  - 4770 - float64
  - Total Bags - float64
  - Small Bags - float64
  - Large Bags - float64
  - XLarge Bags - float64
  - type - object
  - year - int64
  - region - object

DATA PREPARATION:

Irrelevant variables

- The irrelevant and redundant variables are identified.
- In the avocado dataset, the irrelevant variable is ‘Unnamed: 0’
- The irrelevant variable is handled by dropping from the dataset:
    - *data=data.drop(['Unnamed: 0'],axis = 1)*

Transformation of column names

- The name of columns (‘4046’, ‘4225’, ‘4770’) are changed to (‘Small Hass’, ‘Large Hass’, ‘Extra Large Hass’) for convenient analysis

Values mapping

- The values of the variable ‘type’ are [‘conventional’, ‘organic’].
- These are mapped to 1 and 0 as follows:
    - change_values = {'conventional' : 1, 'organic' : 0}
    - data['type'] = data['type'].map(change_values)

Date Format

- The format of the values in the Date variable (YYYY-MM-DD) is converted as (YYYY-MM) for data analysis by month

Data Grouping

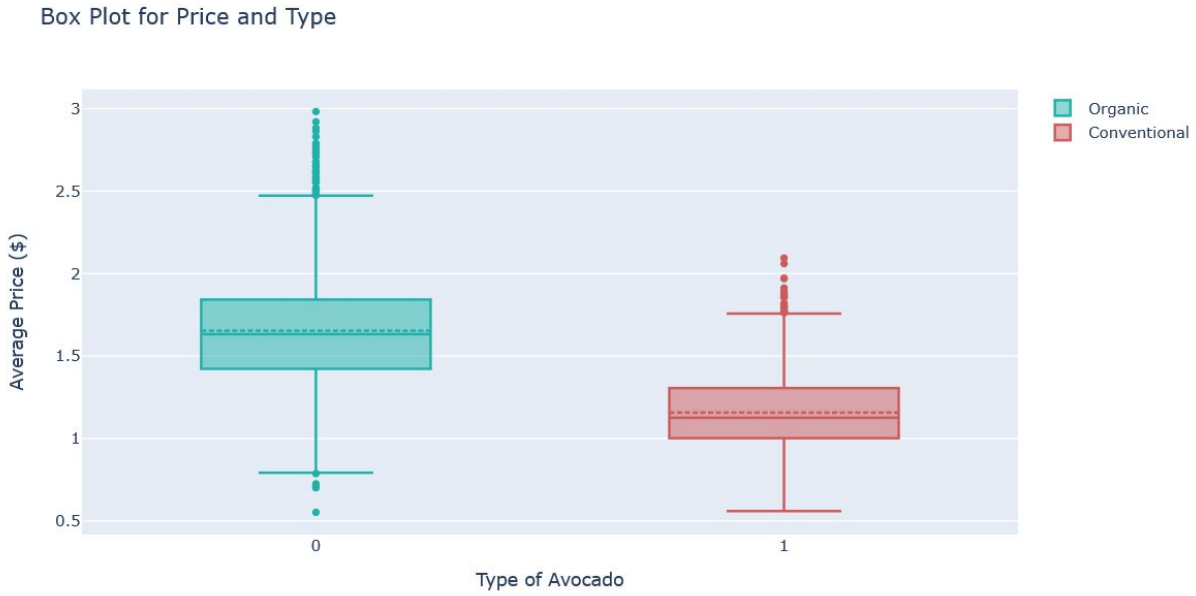
- The given dataset contains details about price and volume of avocado for all weeks in each month.
- The data per each month in every region for both conventional and organic is determined by calculating the mean of all the weeks in each month
- The data is grouped by the variables 'type', 'Date', 'region' for better analysis

Missing Values:

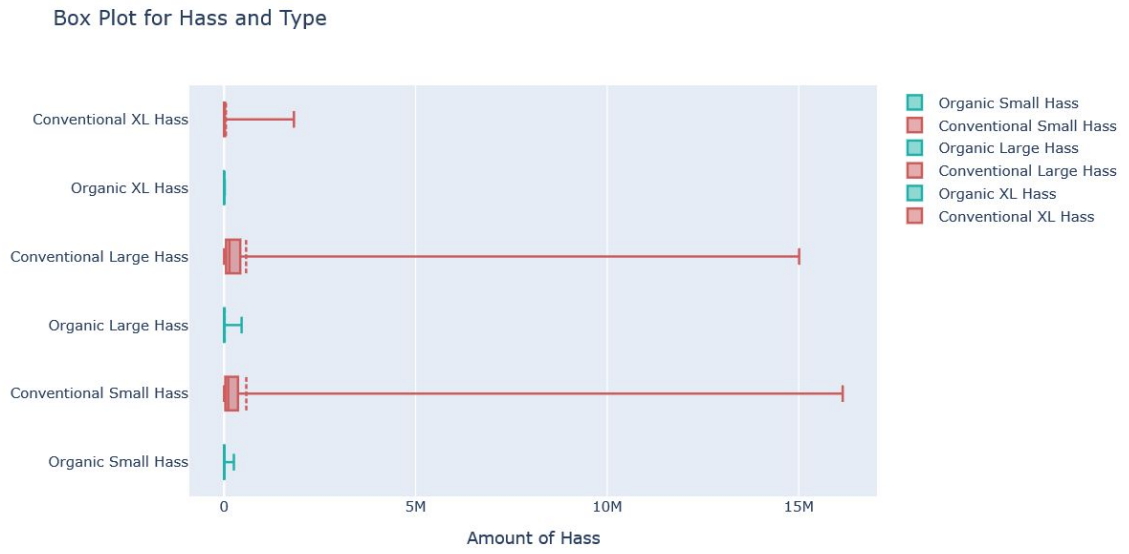
- The missing values in this dataset are assigned with '0'
- The variable 'XLarge Bags' has more than 50% missing values
- If we group the data separately by the variable 'type' and calculate the missing values in the column 'XLarge Bags' then there are 2053 missing values for organic and 420 missing values for conventional.
- We can drop the column 'XLarge Bags' for further analysis
- The variable is dropped as follows:
    - *groupdata = groupdata.drop(['XLarge Bags'], axis = 1)*

DATA EXPLORATION:

- The descriptive statistics are calculated by using describe() function on the respective columns.
- Descriptive statistics is calculated on volumes sold for each size and types of bags sold.
- The results are present in final_project.pdf
- If we consider the results of descriptive statistics for both conventional and organic, The conventional avocados are sold more than organic avocado
- Data visualisation is done by comparing different variables in conventional group and organic group

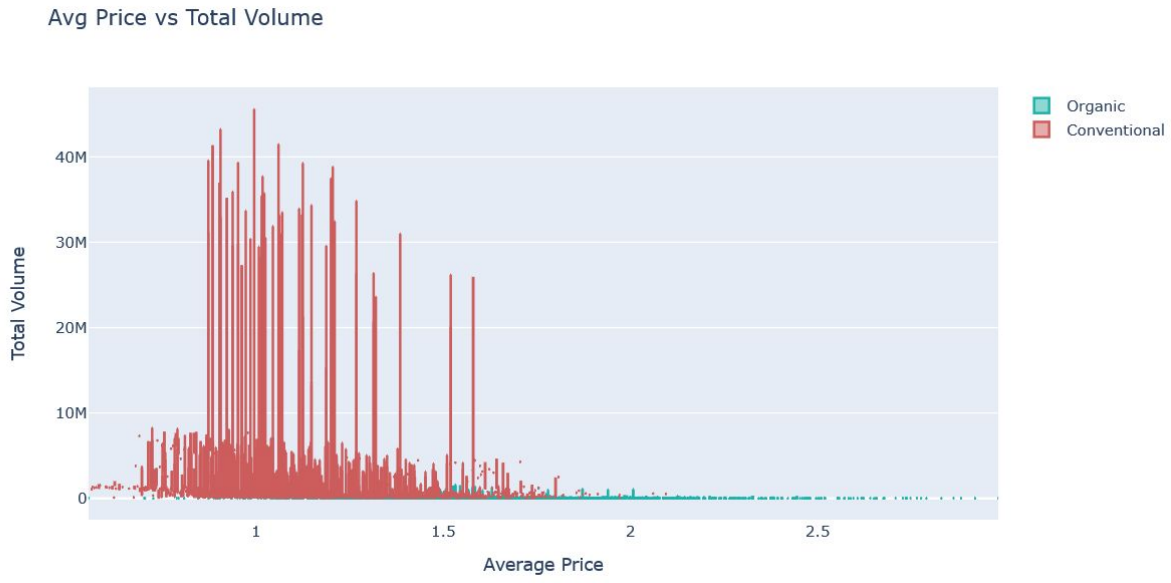


- From the above graph, we can say that the average price for organic avocado (type 0) is greater than the average price of conventional avocado (type 1)



- From the above graph,
  - Volumes of organic avocados are sold very less than conventional avocados irrespective of size of avocado.
  - 25% of conventional small hass avocados has values varying from 1M to 15M, which means a wide variation is present in the quantities of avocados sold.

- The average price of each county is calculated over the years for organic and conventional
- From both graphs, we can see that the pattern in which the average price changes over the years for both organic and conventional are almost the same. So, 'year' might not be important variable for classification



- From the above graph, we can say the total volume play an important role for both conventional and organic in determining the average praise.

From the above analysis, we say that size and volume of hass avocado plays an important role.

Therefore the important in this dataset are '**Total Volume**', '**Small Hass**', '**Large Hass**', '**Extra Large Hass**'

DATA MODELING:

Data Partitioning:

- The given dataset is split into train and test sets by using Hold-out method:
    - Data - train set (80%), test set (20%) and assigned random_state = 0
- The train set is split into training and validation sets by using Hold-out method for evaluating the performance of the models:
    - Train - training set (80%), valid set (20%) and assigned random_state = 0

Data Standardization:

- The train set, valid set and test set are standardized by removing the variables - 'Date', 'AveragePrice', 'year' and 'region'

1. Regression:

- Different regression models are considered for predicting the Average Price of Avocado - Simple Linear regression model, Multi-linear regression model, LASSO regression model, Ridge regression model and Elastic Net regression model
- Different combinations of variables are used for better performance
- The performance of the regression model is calculated based on the Adjusted R Squared value and Root Mean Square Error value.
- The model with high adjusted R squared value and low RMSE is considered as the best.

- The best regression model is the Multi-linear regression model with the variables - 'Total Volume', 'Small Hass', 'Extra Large Hass'

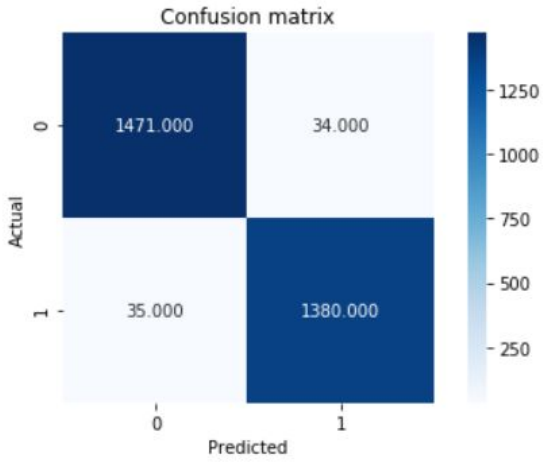
  The performance of the model is:
  R square - 0.05239386909627955
  Adjusted R square - 0.04815085656984486
  RMSE - 0.39717080143144706

  We have selected different combinations of variables:
  1. All variables
  2. ['Total Volume', 'Small Hass', 'Large Hass', 'Extra Large Hass', 'Small Bags']
  3. ['Total Volume', 'Small Hass', 'Large Hass', 'Extra Large Hass', 'Large Bags']
  4. ['Total Volume', 'Small Hass', 'Extra Large Hass']

  These variable combinations are chosen based on adjusted r-square value and RMSE values

2. Classification:

- We used four classification techniques we used were Decision Tree and Naive Bayes and K nearest Neighbour and Support Vector Machine. The best of the four classification models proved to be Decision Trees as we got better values for the metrics. The performance of the model was (0.977, 0.975). The parameters we chose for said model included the minimum number of samples required to be a leaf node at 2 because that improved the results for the f1-score. The variables that we chose were Large Bags, Total Bags, Total Volume, Small Hass, Large Hass, Extra Large Hass and we chose these because we received a better f1-score with them.
- Using Variables 'Large Bags', 'Total Bags', 'Total Volume', 'Small Hass', 'Large Hass', 'Extra Large Hass'
- This model (BEST) accurately classifies 2875 observations of the for the party. The variable were chosen based on comparing the F1 score.



Different combinations of variables and parameters are chosen and mentioned below:

Decision Tree

1. Variables: All variables
    a. F1 score of: 0.817, and 0.155
    b. Parameters: Default

2. Variables : Total Bags, Total Volume
    a. F1 score of: 0.916, and 0.910
    b. Parameters: criterion = entropy, min_sample_leaf = 2, splitter = best

Naive Bayes

1. Variables : All Variables
    a. F1 score of: 0.851, and 0.784
    b. Parameters: Default
2. Variables : Total Bags, Total Volume
    a. F1 score of: 072, and 0.601
    b. Parameters: Default
3. Variables : Large Hass , Small Hass , Extra Large Hass
    a. F1 score of: 0.875, and 0.830
    b. Parameters : var_smoothing = 3e-09

K - Nearest Neighbours

1. Variables : All the variables
    a. F1 score of: 0.972, and 0.970
    b. parameters : n_neighbors = 3
2. Variables : Large Bags,Total Bags,Total Volume,Large Hass,Small Hass,Extra Large Hass
    a. F1 score of: 0.973, and 0.972
    b. parameters : n_neighbors = 3
3. Variables : Total Bags,Total Volume,Large Hass,Small Hass,Extra Large Hass
    a. F1 score of: 0.969, and 0.968
    b. parameters : n_neighbors = 5,algorithm = kd_tree ,weights = distance

Support Vector Machines

1. variables : All the variables
    a. F1 score of: 0.903, and 0.885
    b. parameters : kernel = rbf
2. variables :Total Bags, Large Bags, Small Bags
    a. F1 score of: 0.772, and 0.329
    b. parameters : kernel = poly,probability=True,shrinking=True
3. variables : Total Bags,Large Hass,Small Hass,Extra Large Hass
    a. F1 score of: 0.727, and 0.336
    b. parameters : kernel = rbf, decision_function_shape = ovo

Different combinations of variables are chosen based on the values of accuracy, error, precision, recall and F1- score

Results:

- The best regression model is the Multi-linear regression for calculating the average price Variables: ['Total Volume', 'Small Hass', 'Extra Large Hass']

The best classification model is Decision Tree with the parameters - (criterion = entropy, min_sample_leaf = 2, splitter = best) and the variables are - [‘Large Bags’, ‘Total Bags’, ‘Total Volume’, ‘Small Hass’, ‘Large Hass’, ‘Extra Large Hass’]

- Standardize the x_test using x_train
- Predict and classify the data using the above models
- Export the data into **test_output.csv**

Conclusion:

- From Data Exploration, we can conclude that the important variables for determining average price of an avocado in both the types are ‘**Total Volume**’, ‘**Small Hass**’, ‘**Large Hass**’, ‘**Extra Large Hass**’
- The best regression model to calculate the average price of hass avocado is the multi-linear regression model with the variables - ‘**Total Volume**’, ‘**Small Hass**’, ‘**Extra Large Hass**’
- The performance of the regression models are poor. So we can conclude that the data provided is insufficient to determine average price more effectively
- The best classification model is Decision tree with the parameters - (criterion = entropy, min_sample_leaf = 2, splitter = best) and the variables - [**‘Large Bags’, ‘Total Bags’, ‘Total Volume’, ‘Small Hass’, ‘Large Hass’, ‘Extra Large Hass’**]
- The F1-Score is high for classification model. So, we can classify the avocado effectively with the given data