Machine Learning Approaches to Predict Avocado Prices

Yiheng Fang

September 16, 2020

Abstract

The avocado fruit is one of the most popular fruits in the world, and its price fluctuation, to a great extent, effects the avocado market and economy. This research applies the basic Machine Leaning operation in analyzing the Hass avocado price since Aug,16th,2015 to Mar,25th,2018, and predict the future prices for avocado. This research combined the knowledge of Python language, basic math knowledge, Statistics knowledge with the machine learning knowledge. According to *Artificial Intelligence and Machine Learning: Policy Paper* (Aritificial Intellegence, Technology, 2017), machine learning is a specific approach to AI and the driving force behind the recent developments. It applies the learning algorithms into the analysis and problem-solving process, with the advantage of data availability, computing power, and algorithmic innovation, to spur more scientific researches and studies, and at the same time inspires the social-service work. In this research, we introduce the utility and performing instructions of three algorithms: Linear Regression, Decision Trees, and Random Forest. The whole analyzing process includes read data, analyze data, data visualization, transform categorical data, split data, and perform algorithms to the data. Comparing the mean squared error and root mean squared error, we find out the random forest algorithm has the best performance, which shows the least error.

Contents

Chapter 1

Introduction

1.1 Overview of the Problem

Avocado has become one of the world's trendiest foods. It is a fruit originating from South-central Mexico. Nowadays, avocado attracts lots of attentions for its abundant beneficial elements and nutrition. At the same time, enormous researches have been done on avocado's market status and situation, with the discovery that it costs more and more in recent years. In 2019, the average domestic price of one single Hass avocado reaches $ 2.1. Forming a strong interest in the avocado market, I decided to conduct scientific analysis on avocado Dataset.

According to *Development of Agriculture* article [?], Arendonk's article, [?] "Agricultural development is essential for overall economic transformation of a country." It points out the important role of agriculture in economic debelopment. According to the arcitle, *Avocado Substainability: What are the Social and Environmental Impacts on Avocado?(Andre, 2018)*, the author suggests that avocados became the latest trend in the Western World's diet. In the United States, there are about 2.4 million tonnes of avocados consumed in 2018. Of the 5,7 million tonnes of avocado produced in 2016 in 564 thousand hectares of land, the Aztec country was accountable for 1,9 million tonnes (33), of which around 37 were sold to the U.S. The U.S. is ranked as the 6th biggest global producer with 0.2 million tones of avocado grown, and 90 percent of which are grown in California. Also, imports from Mexico have increased a lot in order to satisfy the mass demands of avocado. This trend leads to more encouragement to the workers and employment in avocado industry. Therefore, agriculture takes a large part in GDP, and avocado is one of the main products. The avocado industry brings huge benefits and importance for area workers and farmers.

This research contains high values in economic aspect. As the sellers know the future price of avocado will increase, which means the demand will decrease, they will not make too much inventory, preventing the avocado from going bad. As for the farmers know that avocado will turn expensive, they would produce more to reach allocative efficiency. In this way, the avocado market will be more balanced, and there will be less unnecessary surplus or shortage. Also, avocado has high elasticity, so there will be many substitutes. Grasping the price trend of avocado, it will be realistic to predict other substitutes' price, like avocado yogurt and avocado salad. Therefore, it is quite beneficial for accurately predicting the market values of avocado.

Moreover, the area lacks critical investigation. Now although there are many works have been done to analyze the nutritions, values, and the essences of avocado, there are few studies showing trends for future. Also, the avocado market exists instability and uncertainty in the field, like illegal gangs and crime groups such as the Caballeros Templarios (Knights Templar) or the Jalisco New

Generation Cartel. The gangs did lots of crimes like threatening the farmers by killing their families or burning the crops which they live on. Therefore, my work for predicting price and analyzing the avocado market is urgent for stabilize the avocado market.

The Hass Avocado is one of the largest avocado companies in the world, counting vast data about avocado. The dataset of this research is collected in *https://www.kaggle.com/neuromusic/avocado-prices* of the Hass Avocado company, which is reliable and authentic.

1.2 Outline of the dissertation

In this research,the aim is by running different algorithms to find the most suitable set in predicting the future price of avocado. And in the whole process, it goes through several steps including read data, deal with missing or null data, analyze data (which will be detailedly explained in chapter 2), deal with categorical or numerical data, data split, feature scaling, and perform different algorithms (which will be detailedly explained in chapter 3).

According to the analysis, this dataset has 18249 rows and 67 columns, and this is a time-series data. We use pandas read to read data. About this dataset, we choose regression model to manipulate, since the goal is to predict the exact domestic average price of the avocado, and this is quantity export, which predicts the continuous variable. There are totally three methods including Linear Regression, Decision Trees, and Random Forest of algorithms tried to predict the result, which are all essential regression analyzing tools.

Chapter 2

Machine Learning Algorithms

2.1 Linear Regression

2.1.1 Introduction

Linear regression is a method modeling the linear relationships between observed variables, and its goal is to find the best-fit line. In linear regression analysis, the variable Y is the dependent or target variable, while the variable X is an independent or explanatory variable, which is the regressor. And for the function containing points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the best-fit line can be written as

$$y = mx + b$$

The relationship is deterministic and exact for linear regression analysis.

2.1.2 The Best-Fit Approaches

Consider the pair (x, y), let y_1 be the predicted value of y associated with x if using the linear regression analysis. Then the error of the function is defined as

$$e = y - (y_1)$$

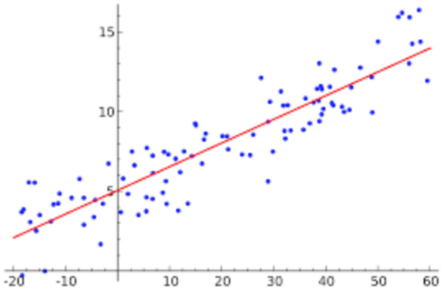


Figure 2.1: The Best-Fitting Linear Relationship Between the Variables x and y

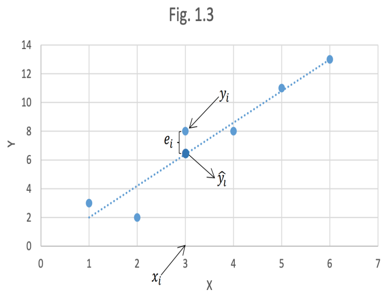


Figure 2.2: The Error occurs in a Linear Regression Analysis

Plus, our criterion must be based on some aggregate measures.

By far, the most common approach to estimating a regression equation is the *least square*, that is, the linear approaches need to minimize the squared errors in the figure above. In the situations that have inevitable and substantial uncertainties, we need errors-in-variables models instead of least squares, which account for measurement errors in the independent variables.

2.1.3 Application

The most common application for linear regression analysis is in data fitting. And in regression analysis, P-values and coefficients in regression analysis work together to show the most statistically significant model and the nature of this relationship. The coefficients describe the relationship between the independent and dependent variables. Sometimes there are curved relationships exist in the independent and dependent variables, and in this curvilinear relationship, the effect of the independent variable is not a constant value.

2.1.4 Advantages and Disadvantages

As one of the essential and fundamental algorithms in machine learning, linear regression has many advantages itself. Technically, it can be easily implemented and applied, and the output results can also be easily interpreted. Especially when we have already known the linear relationship between the variables, it will be most convenient to use linear regression compared to others. Linear regression is indeed susceptible to over-fitting, but it can be avoided using some dimensionality reduction techniques, regularization techniques and cross-validation. However, on the other hand, if there are outliers occur in the dataset graph, it can pose a huge effect on the regression analysis, and the result of analyzing might turn inaccurate.

It is not accurate and rigorous enough to apply linear regression in real-life problems, as it will over-simplies them, assuming the straight-forward linear relationships between variables. And in this way, linear regression cannot explain the complete relationship between variables.

2.2 Decision Trees

2.2.1 Introduction

Decision Tree is a supporting tool to model the possible consequences of the event, like chance event outcome, resource costs, and utility. It is the direct way of visualizing possibilities and display

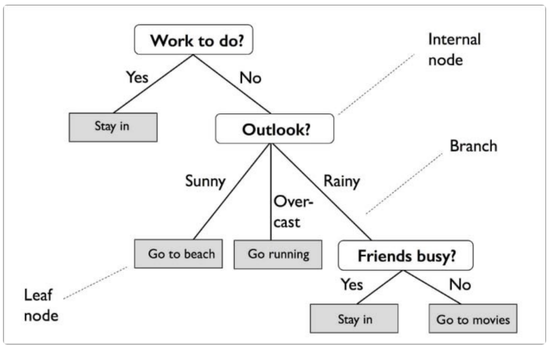


Figure 2.3: Use Decision Tree to Decide Problem in a Particular Way

the algorithms that only contain conditional control systems. Decision Tree is also one of the fundamental, best, and common used analyze methods in Machine Learning, besides its utility in operations research. The solutions in the Decision Tree are shown in the leaves format, containing stems and leaves, and each break-down node represents a point of deliberation and decision. In a decision tree, the target variable is usually categorical. Its two basic functions are:
1. Calculate the probability that a given record belong to each category or,
2. To classify the record by assigning it to the most like class.
Also, decision tree has different types relating with their different target variables. It can be divided into two types:
1. Categorical Variable Decision Tree: where exists categorical potential choose. Eg. In a study of students' performance in school play, the target variable is "Student will participate the play or not", and the result becomes YES or NO.
2. Continuous Variable Decision Tree: Applies when decision tree has continuous target variable. At the beginning, the whole training set is considered as the root. And to recursively record the distributions, every leaf node is a decision needed to make. The internal nodes are the inside attribution node in a decision tree, and some statistical approaches are done to place node of trees orderly.

2.2.2 Advantages and Disadvantages

Since Decision Tree demonstrates the fully phenomenon and chances, it illustrate predictive models bringing high accuracy, stability, and ease of interpretation.
1. Easily being understand, decision tree output does not need any statistical knowledge or analytical background for people. Using the intuitive and vivid graph, readers and users can easily interpret the hypothesis.
2. Decision tree is one of the fastest and most convenient way in visualizing the significant variables in showing their explicit relationships. For the data set who has a large number of variables, the decision tree can help people quickly identify the most significant variable.
3. For decision tree analysis, the steps for data preparation, data organization, and data cleaning are less. Also, it is suitable and applicable with different kinds of relationships, besides pure linear models like what linear-regression analysis could apply to.
On the contrary, the decision tree also contains some flaws compared with other data analysis methods.

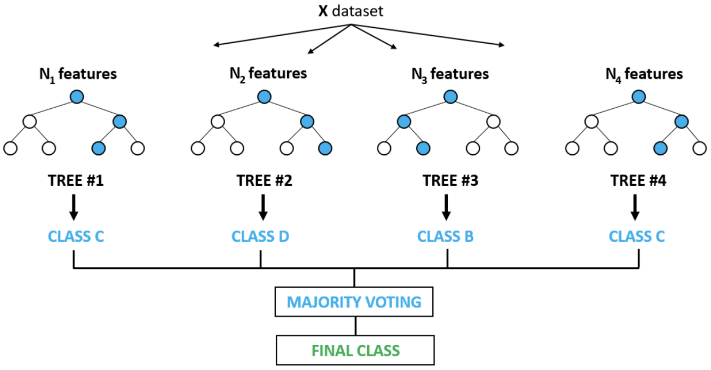


Figure 2.4: Using Random Forest Analysis in Machine Learning

1. **Limited.** Decision tree might be too complex when people designing it with a large data set, and there might occur the over fitting problem, which is one of the most practical difficulty for decision tree models. At the same time, it is also not suitable for analyzing continuous numerical variables, for can only categorize variables in different categories.
2.**Unstable.** The shape of the tree might be effected largely by variance, and methods like bagging and boosting can lower this flaw. Also, the instability can be shown for the bias in different classes where dominance occur. The different attributed categories show the imbalance information collected.

2.2.3 Application

Here are situations that are suitable for applying Decision Tree analysis:
1. When the object of analyzing data set is to achieve the definite goal like optimizing the cost and maximizing the profit.
2. When there are more than one course in the deciding process.
3. The benefits of decision is a countable and calculable measure.
4. When there are environmental factors beyond the control of decision maker
5. Uncertainty concerning which outcome will actually happen.
The decision tree can be applied in many subjects like manufacturing, production, biomedical engineering, astronomy, molecular biology, pharmacology, planning, and medicine. For example, the decision making tree can be applied in the evaluation of

2.3 Random Forest

Roughly, the random forest is a collection of random decision trees. Each tree is built on the random sample of the original data. It can be used in both classification and regression as a useful tool in Machine Learning. Random decision forests correct for decision trees' habit of over fitting to their training set.

Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie et al., "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".

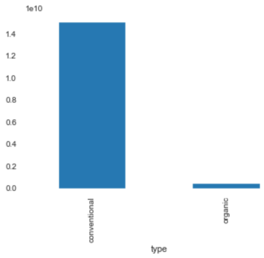
Chapter 3

Data Analysis

3.1 Data Visualization

The figure 3.1 shows the first graph of data visualization. We gathered the overall volumes of conventional and organic avocado separated. From the statistic result, we can see that the conventional avocado is much more than organic-type avocado, and it takes a large part in the whole dataset. With a fair and random dataset chosen, it is convincing to say that conventional avocado takes a larger role than organic ones in the avocado market.

Figure 3.1: The Total Amount of Organic and Conventional Avocado



The Figure 3.2 shows that for each price, the amount of avocados existed in our dataset. Using the Histogram to visualize the dataset, we found out that the selling peak falls in the price of 1.0-1.5 dollars.

The Figure 3.3 shows each month's average price. Because the ripeness of avocado is related to seasons, according to our hypothesis, so the variable of months might effect the price of avocado. In the graph, it shows that the price peak occurs in March, April, August, October, and November. In February, May, June, July, and December, the average price is relatively low.

The Figure 3.4 shows the selling amount of three varieties 4046,4225,4770 during Jan.2015 to Jan.2018. From the graph, we can see that the selling amounts of 4225 and 4046 avocado are generally similiar, keeping the data around 30,0000-40,0000, and the selling amount of 4770 is generally speaking much lower than the other two, fluctuating under 5,0000. In around every February in each year, the selling amount of 4225 and 4046 is experiencing an obvious peak, while there is also a relatively less obvious peak in every year's May. In February of 2016 and 2017, there are peaks occur for 4770 avocado, although the integral trend of 4770 avocado is stable and approximately a

Figure 3.2: The Selling Amount of avocado for each price

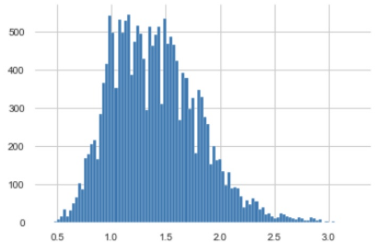
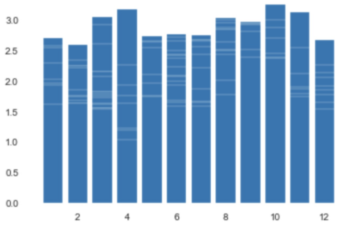
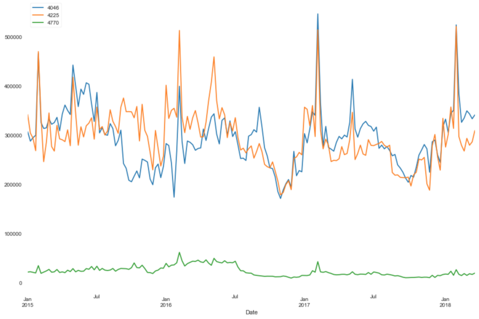


Figure 3.3: The Average Price for Each Month



horizontal line.

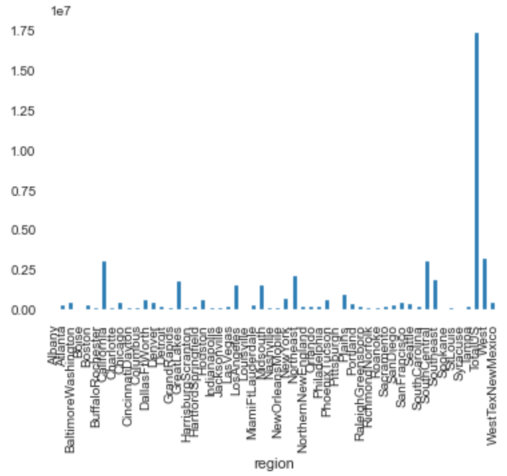
Figure 3.4: The Selling Amount of avocado for Three Varieties



After data visualization, we also need to transform categorical variables "Type" and "Region". For categorical variable is the data type that groups the information with similar characteristics, so we should transform the data type into features which regression can be used.

Also, we need to take the step of spliting data. Spliting data is the process of filtrating and separating train and test data.

Figure 3.5: The Average Price for Different Regions



Chapter 4

Performances of Different Algorithms

In this chapter, we use three machine learning algorithms mentioned in the previous chapters to show the analysis, and included the conducting graph of three algorithms. As we know, in classification, there is accuracy that can be compared to test out the best-performance algorithm. However, in regression data set like this, we can only judge the winner by looking for the least error. We calculate three algorithms' mean squared error and root mean squared error.

4.1 Linear Regression

Importing the train test, we draw the linear regression graph that shows points related with a linear relationship. The mean squared error for Linear Regression Analysis is approximately 0.066 and the root mean squared error is 0.257.

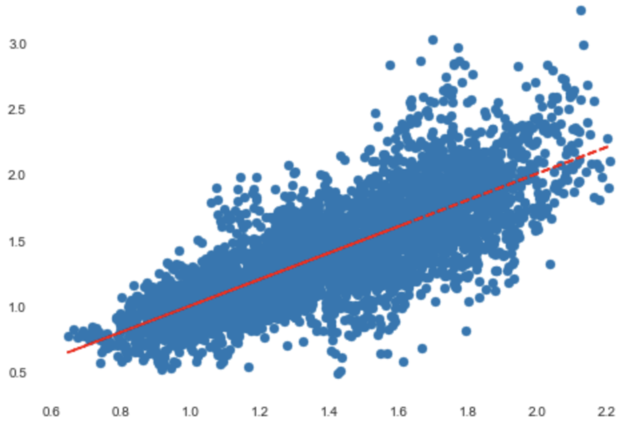


Figure 4.1: The Linear Regression Analysis in Avocado Dataset

4.2 Decision Tree

Importing the train set, we draw the decision tree graph. The mean squared error for Decision Tree Analysis is approximately 0.027 and the root mean squared error is 0.165.

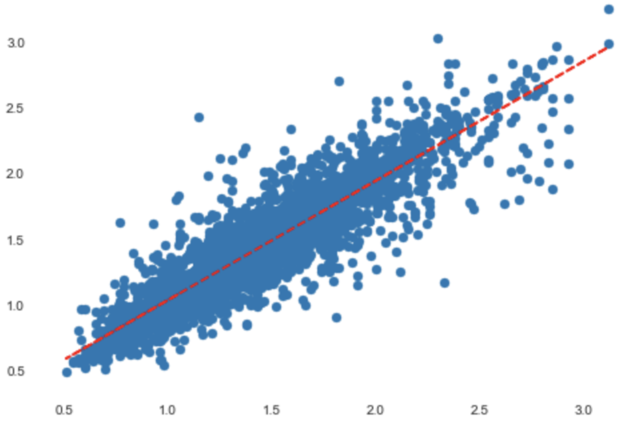


Figure 4.2: The Decision Tree Analysis in Avocado Dataset

4.3 Random Forest

Importing the train set, we draw the random forest graph. The mean squared error for Random Forest Analysis is approximately 0.013 and the root mean squared error is 0.115.

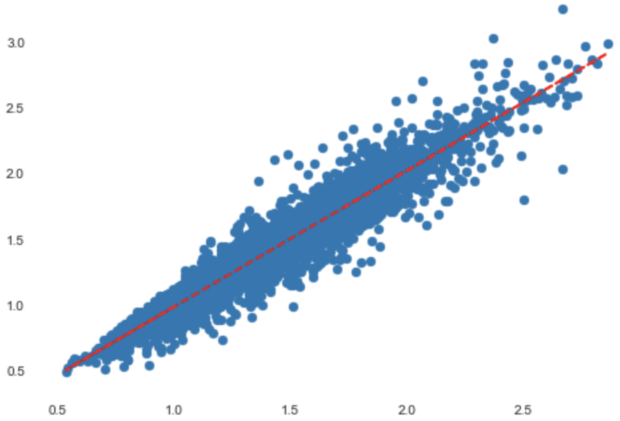


Figure 4.3: The Random Forest Analysis in Avocado Dataset

Chapter 5

Conclusion

In this research relating with the avocado market, we use three fundamental machine learning algorithms to analyze the price trend of Hass avocado, and calculate their mean squared error and root mean square error. In the future, some other more advanced and complicated algorithms are considered to be applied in the project, like Logistic Regression, Support Vector Machines, Lasso Regression, Maltivariate Regression algorithm.

5.1 Reference

https://www.kaggle.com/neuromusic/avocado-prices
https://youmatter.world/en/benefits-avocados-production-bad-people-planet-27107/
https://www.researchgate.net/publication/329156107ArtificialIntelligenceandMachineLearning
https://en.m.wikipedia.org/wiki/LinearRegression
https://en.m.wikipedia.org/wiki/DecisionTree
https://en.m.wikipedia.org/wiki/RandomForest
https://www.researchgate.net/publication/297736556AvocadoCharacteristicshealthbenefitsanduses