Data Mining Project Update

The Avocadoers

Evan Lee
University of Colorado - Boulder
CSCI 5502
Evan.N.Lee@colorado.edu

Hector Sanchez
University of Colorado - Boulder
CSCI 4502
Hector.Sanchez@colorado.edu

Madelaine Struwe
University of Colorado - Boulder
CSCI 4502
Madelaine.Struwe@colorado.edu

ABSTRACT

One of the biggest concerns in certain geographical regions of the United States is the relative market price and strength of the avocado, both as a crop and as a market product. The purpose of this project proposal is to determine a relationship between the weather and climate metrics of an avocado-producing region and the subsequent seasonal market share and market price in various geographical regions of the United States in order to produce a predictive model. We will accomplish this by analyzing and mining data sets from government agencies as well as surveys and other data from the private sector in order to determine the effects of individual weather and climate metrics (among which include standard metrics such as temperature, pressure, and precipitation) and using these calculations to produce a weighted model.

KEYWORDS

Avocados, Data Mining, Avocado Prices, Avocado Sales, Weather

ACM Reference format:

Evan Lee, Hector Sanchez, and Madelaine Struwe. 2018. Data Mining Project Update. In *Proceedings of Data Mining, Boulder, CO USA, Fall 2018 (CU BOULDER),* 3 pages.

https://doi.org/10.475/123_4

1 MOTIVATION

The proportion of time the avocado spends in the thoughts of the average modern millennial has increased year after year since the advent and rise in popularity of avocado-based meals and foods. Menu items such as guacamole, avocado toast, and an adaptation on the popular BLT, the BLTA have fully integrated themselves into popular culture.

As such, a general drive exists in the food and crop industries for research into the trends, relationships, and effects of various metrics on the modern market. With our proposed project research, not only can major industry players and consumer resource groups have insight into the predictive and projective capabilities of weather and climate data, but now, with our research and models, so can end-consumers with a vested interest into the market trends of the avocado.

We aim to give users at all levels of the avocado industry the power to predict with a high degree of confidence the future market trends to allow these users to make informed decisions, no matter the scope and impact of such.

1.1 Literature Survey

Research into this particular topic has been mostly top-down, starting at market trends/analysis data and has also been focused on comparative metrics against the competitors of the avocado. The premier source of market data collection and analysis have been consumer groups such as the Hass Avocado Board and the California Avocado Commission. With publications like *the Yearly AvoScore Card* and *State of the Category*, the majority of work by these consumer groups has been driven by increasing market share.

Groups that have performed research and analysis into the avocado industry have tended to be single-focus and small scoped, with each portion isolated and performed by a distinct organization. There has not been research into related trends and the modeling of the relationship as such.

Our team has identified an opportunity to tackle the bigger picture and link these discrete works together by approaching this problem from the bottom-up. Starting at the core of the avocado industry (weather and climate metrics), we can then link the relationship between the base layer metrics to the layer up (avocado production and market volume) and finally to the most visible layer (avocado market trends and analysis).

2 PROPOSED WORK

For the project, we propose to find the correlation between the weather of avocado farms, the crop yields, and the price of avocados in the store.

Generally speaking, we plan to take data correlating to avocado farm yields, avocado retail prices, as well as weather effecting the regions where avocados are grown. We will take this data and come up with a program that will be able to predict avocado prices based on the effects of the weather where avocados are grow.

CU BOULDER, Fall 2018, Boulder, CO USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

For our project, we will most likely use *Python* or *R* to model and analyze the data.

2.1 Data

We will be using data sets from:

- National Oceanic and Atmospheric Administration
  From this we can can get weather data from the area where avocados are grown and distributed to stores across the United States.
- United States Department of Agriculture
  From this we can find data concerning the yield of avocado farms in the United States.
- Hass Avocado Board
  From this we can obtain data on retail prices and sales. This information is divided by conventional and organic avocados, as well as by region in the United States.

2.2 Sub-tasks

The first task that needs to be completed is to obtain data relevant to our problem and then examine those data sets to see what we can mine from them. From there we must connect the different data sets together to further help solve our problem, and make it possible to find a correlation. The best way to tie the data sets together is based on the week, as grocery stores usually change prices based on the week rather than the day. The next steps would be to write the program, train the program, and then test the program. Finally, we will need to analyze the results to check for accuracy. Based on the accuracy of the program, if we need to, we will repeat steps to better solve our problem.

3 WORK ACCOMPLISHED

At this point in the project, we have compiled all data sets that we believe we will be utilizing. Additionally, we have begun rudimentary normalization and standardization of our data sets so that our cross-source data windows will be normalized and easily mined. To this end, we have essentially standardized the format and temporal windows that our project will be focused on. We have also done extensive research into the effects of various weather and climate attributes in order to determine a weighting system.

3.1 Processing & Cleaning

The first task was clearly to identify and compile data at the scope that we wished to accomplish. At this stage, we ran into our first issue of mismatches in availability and granularity of data. For example, the data sets from the National Oceanic and Atmospheric Administration (NOAA) are very extensive, ranging as far back as 1945, depending on the operational date of a particular land station. NOAA data sets are also extremely granular, with certain metrics recorded at a fifteen minute interval, others at an hour interval, and yet others only having daily summaries, averages, or ranges. In contrast, a large amount of the market volume/price data is aggregated at the weekly level. A great deal of work and time thus far has been put into normalization and standardization across each data set, especially into cleaning of each weather attribute set. The use of *python* for some simple scripting as well as basic *Microsoft Excel* operations have been put into use in this section.

3.2 Delays

Unfortunately, it has been identified thus far that the intended scope of the project is too large to accomplish in the remaining time allotted. We intended to accomplish the mining of the relationships of avocado market share with one or more of the following weather attributes:

- Sunshine & Visibility, Cloud Cover
- Precipitation
- Temperature
- Humidity

We intended to assign a weighting system to these attributes, either accomplished by applying a regression model to the test data or by researching a more naive approach via research and educated arbitrary assignment. However, due to the mere volume of data and the time taken to process such, we have scaled our goal back to only the relationship of one metric, which we have selected as **temperature** which we felt was most appropriate. This would eliminate the extra time and overhead required to mine or assign a weighting system and also cuts down on the amount of processing time of extra data sets.

Our normalization scheme for the temperature metric is described below.

3.3 Scale/Rate The Weather Data for Avocado Growth

We will implement 0-3 scale that will help us to interpret our data. Avocados have a wide range in temperatures that they can handle, however, the peek condition for avocados to grow is between 75-85$^{\circ}F$ weather. But they can handle extreme heat and extreme cold pretty well (only for a certain amount of time). We will use a grading or scaling system, in which 0 is peek weather conditions and 3 is not good weather conditions, for weather conditions that are both hot and cold.

3.4 Data to be Used in The Project

Due to the vast amounts of data available, we have decided to reduce the size of the data we will be using for the mining process. Because of the sheer amount of weather data available, we are opting to analyze one year of data, at most two years. As stated above, instead of using all the weather data, such as "Dew Point" and "Sea Level Pressure", we are electing to narrow down the weather attributes.

4 HOW TO EVALUATE

We have many years worth of data on past avocado prices and the weather. However, not all of this data will be used in our training set. With the data sets that are not used in our training data, we will use those to determine the accuracy of our project.

For instance, if we do not use 2017 year data as training data, we can use that to test our accuracy. We would put in the weather patterns into our program and get a prediction of avocado prices. With that prediction we can compare it to the actual data of avocado prices. If the predicted price is no where near the actual price, we know that our model is not accurate and more work needs to be done.

However, the converse is also true. If the predicted price is close to the actual price, we know that our model is accurate. As such, we can assume that it will be accurate for predicting "what if" scenarios.

4.1 Milestones

For our project, we have milestones to help keep us on track and moving forward.

- **Select a topic for the project** *September 25* The group met beforehand to brainstorm project ideas, and we needed to settle on one topic to get research started.
- **Obtain data** *October 11* Searching for relevant data sets for the problem we want to solve.
- **Examine data** *October 16* Examine the data sets to see what data can be extracted from them. Also find out what other projects have been done with the data sets (if any).
- **Connect the data** *October 21* Connect the data sets to each other. Link the weather to the data, price of avocados to the date, number of avocados harvested to the date, number of avocados exported to the date.
- **Write the program** *November 6* Write the program to do the stuff
- **Train the program** *November 13* Use previous years data sets to train the program
- **Test the program** *November 15* Use a years worth of data that was not used in the training set to test the accuracy of the program.
- **Analyze results** *November 16* Determine if the results from the test the accuracy.
- **Repeat Write-Analyze if necessary** *December 16* If the results were inaccurate or abnormal, rework the program.

5 BRIEF SUMMARY OF PROJECT DISCUSSION

For the project, we plan to use data sets from Hass Avocado Board and California Avocado Growers for data on avocado consumption, and pricing. We plan to use the National Operational Model Archive and Distribution System to look at archived weather patterns where avocados are grown. We will also be using National Agricultural Statistics Service to determine the yield of crops of avocados. Using this data, we plan to look into how weather patterns can effect avocado harvesting, exports, and retail pricing.

We plan to do this by creating a program, possibly using python, that will correlates historical weather patterns to subsequent market prices and shares. The program will provide an interface that will allow prediction of avocado prices based on theoretical or projected inputs.

We will use data from previous years as our training data for our program. After our program has been "trained" we will use another previous year of data that was not used in the training set to test the accuracy of our program. If the test data shows accurate results, we know our program is also accurate. However, if the results are inaccurate with the actual prices, our program is inaccurate and more work must be done.

ACKNOWLEDGMENTS

The authors would like to thank avocados everywhere.

6 REFERENCES

[1] Hass Avocado Board. 2018. Retrieved from http://www.hassavocadoboard.com/retail/volume-and-price-data.

[2] NOAA. National Centers for environmental information. Retrieved from https://www.ncdc.noaa.gov/cdo-web/datatools/lcd.

[3] United States Department of Agriculture. National Agricultural Statistics Service. Retrieved from https://quickstats.nass.usda.gov/.