Data Analytics Capstone Topic Approval Form

Student Name: William Townsend

Student ID: 003397146

Capstone Project Name: Time Series Forecasting of United States Hass Avocado Sales

Project Topic: This project will use industry data published by the Hass Avocado Board (www.hassavocadoboard.com) containing price and volume information from 2020 & 2021 to generate an effective time series forecasting model for forecasting Hass avocado sales throughout 2022 and forward into 2023.

X This project does not involve human subjects research and is exempt from WGU IRB review.

Research Question: Can sales of conventional Hass avocado sales in the United States be effectively forecasted based solely on the research data?

Hypothesis: **Null hypothesis**: A predictive time series forecasting model with a mean absolute percentage error of < 20% cannot be generated from the research dataset. **Alternate Hypothesis**- A predictive time series forecasting model with a mean absolute percentage error of < 20% can be generated from the research dataset.

Context: Hass avocados are a popular fruit sold throughout the United States, commonly associated with Tex-Mex food in the western United States. In the United States, the Hass Avocado Board is a trade group that works to "help make avocados America's most popular fruit", and to this end, they publish a variety of supply and market data for both growers and marketers. The Hass Avocado Board releases a small portion of its data to the public, and they encourage the public to work with their data for research purposes, such as this project. With the limited dataset that they provide to the public, data is available regarding both the average sale price of Hass avocados on a per-pound basis and the volume of Hass avocado sales in pounds, broken into different avocado products. With this data, forecasts can be built that suit business interests by giving an idea of the United States' appetite for purchasing Hass avocados year-round.

Data: The data needed to attempt to generate a predictive ARIMA/SARIMA model is published by the Hass Avocado Board at the bottom of their interactive data dashboards. Downloadable volume datasets are only available for 2020, 2021, and 2022. Each year's dataset contains nearly 7,000 observations, and the three will be compiled into one combined dataset. The dataset is has 0% sparseness. These datasets contain the following variables of use:

Field	Data Type
Geographical Area (Market, Region)	Qualitative
Timeframe	Qualitative
Type (Conventional/Organic)	Qualitative
Week Ending (Date)	Quantitative
Average Sale Price	Quantitative
Total Bulk Units (lbs)	Quantitative
Individual Units small Avocados (lbs)	Quantitative
Individual Units large Avocados (lbs)	Quantitative
Individual Units X-large Avocados (lbs)	Quantitative

The primary limitation of this dataset is that it only includes three years of data. A larger dataset would provide more data on which to train a forecasting model, especially one that I expect to have significant annual seasonality. I did attempt to reach out to the Hass Avocado Board to get additional information regarding past years' data, but I did not receive a response. This data also only provides information regarding purchasing and sales of Hass avocados, omitting supplier information which might be impactful. While 90% of Hass avocados sold in the US come from California, which likely minimizes any supplier effects that might exist within the sales and purchasing data, this cannot be verified to be the case without that data.

Some delimitations exist for the intended study. Exploratory data analysis reveals that organic Hass avocados constitute a very small portion of the overall sales of Hass avocados within the United States, at around 3.3%. The dataset distinguishes between both "conventional" (not-organic) and "organic" avocados, so this analysis will omit organic avocados to instead concentrate on the ~97% market share of conventional avocados. This analysis will focus on nationwide trends and forecasting of avocado sales, as this is representative of the 8 regions distinguished by the Hass Avocado Board whose data make up the whole of the total United States data. This analysis is chiefly concerned with forecasting sales volume, so while some exploratory data analysis may involve the pricing of avocados, this is not a variable intended to be forecasted in this study.

Data Gathering: Data containing average sale price and volume of bulk and individual sales in pounds is published by the Hass Avocado Board, aggregated on a weekly basis and organized by market (50 included), region (8 total), or nationwide. Data is downloadable in .CSV format from the Hass Avocado Board for the years 2020, 2021, and 2022. This data will be combined into a complete dataset for these three years. The data is in generally very good shape, being published for public consumption and data analysis by the Hass Avocado Board. The sales price and volume data are complete, while some other columns that are inconsistently used will be omitted from the analysis because they are unnecessary. The most significant problem present in the dataset is that the data for 11 Dec 2022 is missing. This occurs during a low-point in annual Hass avocado sales and production, and the missing values will be generated by taking the average of the values for 4 Dec and 18 Dec 2022 to stand-in for this data.

Data Analytics Tools and Techniques: Exploratory Data Analysis will be performed for a number of views of both the sales price and volume of avocados sold, such as looking at differences between organic vs conventional avocados, price by region, etc. This will give some insight into trends and seasonality of the data. After this exploration is performed, the nationwide time series data for both price and volume will be decomposed to ascertain seasonal and trend effects of both data series. The data will be split into a training set consisting of 67% the observations (2020 & 2021) and a training set consisting of 33% of the observations (2022).

A time series forecasting model(s) will be generated as appropriate and fitted to the training data, which will then be used to perform a forecast for the test set. The mean squared error of the forecast will be calculated to determine the effectiveness of various models amongst each other. The optimal model will then be evaluated for effectiveness by having the mean absolute percentage error (MAPE) calculated for its forecasted predictions relative to the observed data for that time period. A model will be considered to be "effective" for the purposes of accepting or rejecting the null hypothesis by having a MAPE of under 20% for its forecast of 2022 test data. If an effective model is generated, it will then be used to generate a similar forecast for 2023 sales data.

This project will use the following tools in the course of its analysis:

- Jupyter Notebook
- Python programming language in the Anaconda environment
- The following Python libraries:
    - pandas
    - NumPy
    - MatPlotLib
    - SciKit-Learn
    - StatsModels
    - itertools
    - pmdarima
    - Prophet (formerly FBProphet)

Justification of Tools/Techniques:

Mean average percentage error will be used as an evaluation metric to determine if the optimized forecasting model is effective. Mean squared error (MSE) and root mean squared error (RMSE) are potential alternatives, but the magnitude of these is dependent upon the units involved in the study. The volume of avocado sales nationwide by weight is in the hundreds of thousands of pounds on a weekly basis, meaning that an MSE/RMSE are likely to be very large in their absolute value as well. A more standard and easily intuited measurement of error for this situation would instead be the mean average percentage error (MAPE), because this is standardized on a traditional 0 – 100% scale, representing the percentage error between the observed and predicted values. Evaluation of a forecast's MAPE still requires some scenario-specific considerations, but a common rule of thumb in business forecasting is that a MAPE of 20% or less is a "good" model, while a MAPE of under 10% is a very good model. As such, the threshold for whether or not

a model is considered "effective" for the purposes of evaluating the null hypothesis will be if the MAPE is under that 20% threshold for a "good" forecasting model.

Python provides a useful and robust programming language to perform this data analysis, and the Anaconda environment provides an all-in-one solution that allows this to be done in one convenient package. Jupyter Notebooks, as a part of that environment, allows for a "show-your-work" approach that easily moves back and forth between Python script and narrative output all in one place, while allowing for a quickly iterative approach to working with the data.

The pandas library will allow me to place all of the data into a dataframe, which is essentially a Python table that is capable of sophisticated operations and manipulation upon the data within it. The NumPy library is often used by pandas, as well as other packages, because of its mathematical functions. The MatPlotLib library allows for the generation of highly customizable data visualizations.

Scikit-learn is useful for a number of other functions, including splitting between training and test groups, quickly calculating the mean squared error of a forecast, or several other tasks necessary as a part of the process of this analysis. StatsModels will be used similarly, for some elements of decomposing time series data and plotting data related to the time series.
The pmdarima library provides a useful 'auto_arima() function for time series analysis. Rather than building 'for' loops that may be poorly optimized to iterate through different ARIMA/SARIMA models, auto_arima handles this in an optimized all-in-one function that is flexible and handles a variety of different parameters.

Prophet (formerly FBProphet) is an open source automated time series forecasting library released by Facebook's core data science team. Prophet is capable of performing complex forecasting beyond the ARIMA/SARIMA examples used in the MSDA program, it uses the SciKit-Learn model API (fit, predict, etc.) that we were trained on throughout the MSDA program, and it is well documented. Prophet also has hyperparameter tuning capabilities, which the itertools library will be used to interact with.

Project Outcomes: The project will generate a model that is capable of forecasting the volume of conventional avocado sales through CY 2022 with effectiveness demonstrated by a mean absolute percentage error (MAPE) of under 20% in order to reject the null hypothesis. If an effective forecasting model is generated, a forecast will also be performed to forecast the volume of conventional avocado sales through CY 2023.

Projected Project End Date: 3/31/2023

Sources:

Hass Avocado Board: Category Data is the source for this dataset. When applying filtering to isolate 2020, 2021, or 2022 data, at the bottom right of several windows within the dashboard, an option will display to download the "Weekly Retail Volume & Price Report".

Colorado State University - Food Source Information was used for observations and information regarding where avocados come from and how this could impact prices and availability.

Stephen Allwright: What is a good MAPE score?, Cross Validated: Is there any standard of a good forecast measured by SMAPE?, and C.D. Lewis: Industrial and Business Forecasting Methods (1982) were all used for justifying evaluation of the final model based on its MAPE score.

Prophet was used for gathering information regarding the capabilities of the package as an alternative to basic ARIMA/SARIMA/SARIMAX forecasting methods.

Course Instructor Signature/Date: 3/8/23

Daniel J. Smith, PhD, MBA

Institutional Review Board Quiz and Approval

Have you read and understood the "Human Subjects FAQ" page and completed the "Human Subjects FAQ Quiz" at the WGU Institutional Review Board (IRB) website? (https://irb.wgu.edu/info/Pages/Home.aspx)

☐ Yes, I have read and understood the "Human Subjects FAQ" and have provided email proof of my completed quiz in appendix A. (https://irb.wgu.edu/info/Pages/Human-Subjects-FAQ-Quiz.aspx)

☒ No, I have not completed the Human Subjects FAQ quiz.

Assess whether your capstone proposal complies with WGU's IRB standards for exemption status. Explain why you believe the proposed project complies with the standards for exemption status. If it does not, make arrangements with a course mentor and the IRB for approval.

☒ The research complies with WGU's IRB exemption status because:

- Research involving the collection or study of freely available de-identified existing data
- Research that does not employ methodology on human subjects.

☐ The research requires approval from WGU's IRB because:

☐ Yes, I would like to schedule a conference to discuss my project.

To be filled out by a course mentor:

☐ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Mentor's Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.