Study Problem & Hypothesis

Using a research dataset from the Hass Avocado Board of US Hass avocado sales by volume (in pounds) from 2020-2022, could an effective forecasting model be developed?

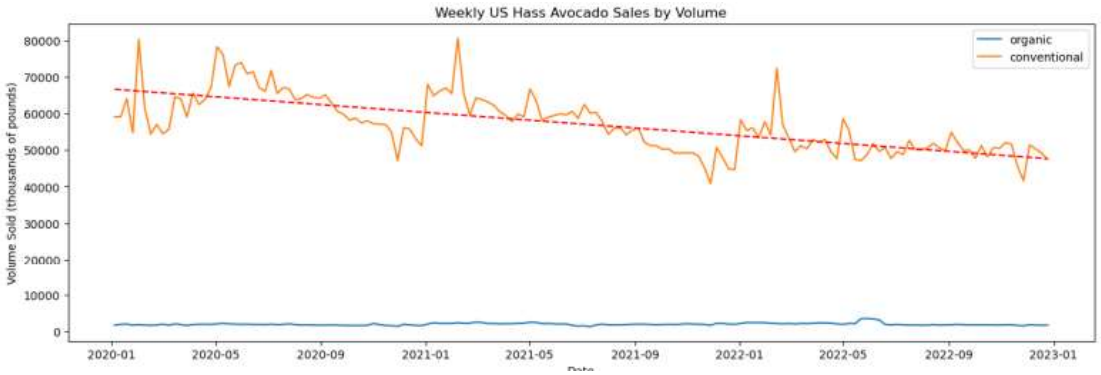
Null hypothesis: An effective predictive time series forecasting model with a mean absolute percentage error of < 20% **cannot** be generated from the research dataset.

Alternative hypothesis: An effective predictive time series forecasting model with a mean absolute percentage error of < 20% **can** be generated from the research dataset.

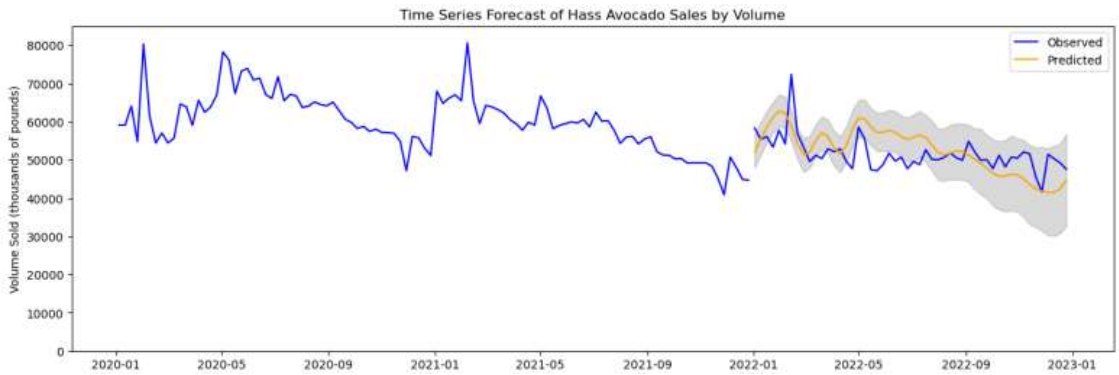
Historical data for Hass avocado sales is easily generated in the process of growing, shipping, and selling avocados across the United States. Generating a forecasting model for future avocado sales improves efficiency for every layer of the supply chain by enhancing planning abilities for those entities.

Data Analysis Process:

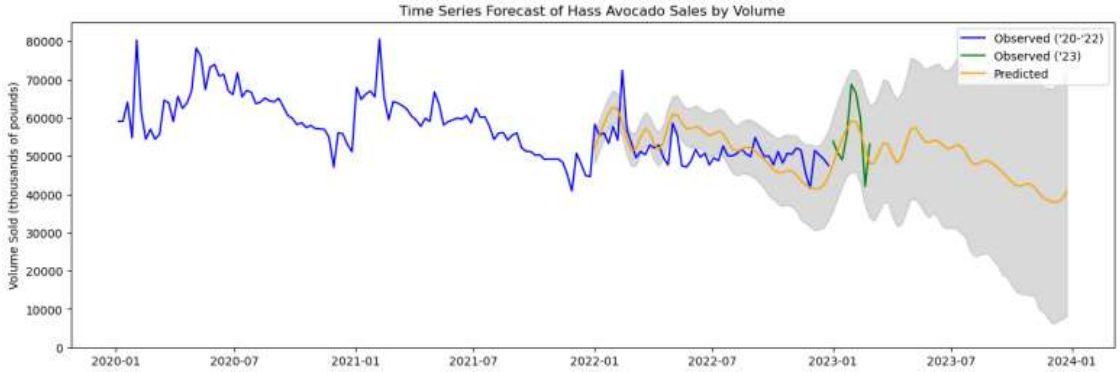
Data was compiled into weekly nationwide totals of sales in pounds, summing all packaging types (bulk and individual) into a total volume. Organic avocados (constituting only 3.3% of nationwide sales) were omitted to focus on conventional avocados. Exploratory data analysis immediately revealed a consistent negative trend in the sales over the study's timeframe.



A training/test split was generated using 2020/21 data for training and 2022 data for testing. ARIMA and SARIMA forecasting models were attempted with marginal success. An automated approach using the Facebook prophet time series forecasting package was attempted with improved results. Hyperparameter tuning yielded a final optimized model.



This forecasting model achieved a mean absolute percentage error of 9.2%, leading to rejection of the null hypothesis in favor of the alternative. With an effective model trained on the 2020/21 data, this same model was used to make a 2-year forecast into 2023, being compared to data harvested directly from the Hass Avocado Board's data dashboards.



The data published to the dashboards was found to be slightly inflated relative to the data published to the downloadable dataset, being approximately 7% larger than it "should" be, without any explanation for this discrepancy. Despite this discrepancy, for the period of Jan/Feb 2023 – a forecast made 13-14 months out, based on only 2 years of data – the model had a mean absolute percentage error of 9.9%. This increase in error was of similar magnitude to the unexplained increase in the harvested sales data from the Hass Avocado Board's dashboards.

Study Findings:

As mentioned, the optimized model had a mean absolute percentage error of 9.2% on the test (2022) set. This easily cleared the established threshold of 20% or less mean absolute percentage error needed to establish an "effective" forecasting model and reject the null hypothesis. The model's forecasting accuracy was much better than what I'd expected to generate from such a small timeframe.

Study Limitations:

The primary limitations of this analysis relate to the timeframe of the analysis. Only a three year period is examined, of which only two years were used for training. As a result, the model can only anticipate patterns based on those two years of data. A larger dataset would be an improvement, but in this regard, I was bound by the limitations of the data provided by the Hass Avocado Board.

A larger concern with the time period examined in this study is that it runs from 2020 – 2022, being almost entirely composed of data from within the Covid-19 pandemic beginning in Mar 2020. This had countless ripple effects, impacting shipping and logistics, inflation of goods, changes in the workforce, consumer behaviors, etc. The study's dataset beginning in 2020 might make it more accurate in its predictions of 2022, 2023, and beyond because it is not swayed by pre-pandemic observations. Alternatively, it might only be accurate in the pandemic period if the future reflects behavior closer to the pre-pandemic period. A worst case is that this model might also become less accurate if things get worse, such as through climate change damages to avocado crops or a new/more destructive wave of

Covid-19. A larger dataset would help provide some of that context, but this uncertainty is a reality of living in unusual times.

Aside from these limitations of the study itself, some minor issues with the data must also be acknowledged. The data published by the Hass Avocado Board was generally in good shape, with only some minor issues, except for one issue where the 11 Dec 2022 observations were completely missing. This was filled by generating observations for this date using the midpoint for each group and variable of the 4 Dec and 18 Dec 2022 datapoints. As this occurred at the nadir of annual avocado sales, the impact of this inference is likely minimal.

Recommended Action:

The visible downward trend in Hass avocado sales over the three-year period in the study is concerning for all involved, from farm to store. This is a serious concern for the business, meriting a rigorous causal analysis to determine why sales are trending downwards. This includes looking at historical data (pre-2020) to determine if the observed trend is part of a bigger picture or a recent phenomenon, analysis of supplier data to determine if there are fewer avocados being produced for sale, and possibly even analysis of retailer data to determine if there are changes with regard to rates of sale (sold avocados vs unsold “waste” avocados). Being a project with a very wide scope, this might be best organized into a series of smaller scoped analyses, some of which can be completed in parallel by different teams, before being synthesized into a final causal analysis.

Expected Benefits:

Expected benefits of this forecasting model relate primarily to increasing efficiency at various stages of the Hass avocado supply chain through enhancing planning. For example, growers of Hass avocados can avoid creating excess avocado crops that are likely to go to waste by reducing their intended yield in proportion with the reduced expected sales compared to prior years. Some growers may even wish to consider alternative crops, which can provide increased opportunities for those growers who remain in the market. Similarly, shippers can use this forecast to both anticipate a lower overall logistic burden from avocado shipment in proportion with the reduced demand, as well as anticipating increased or decreased need based on the seasonality of avocado demand. Retailers can make similar adjustments, whether through their purchasing decisions or their pricing of the avocados they have purchased. This forecasting model allows for efficiency adjustments throughout the supply chain.