**Term Project - Predicting Future Sales of Products**

Felipe Rodriguez and Christopher Botts

Bellevue University

DSC 630 Predictive Analytics

Professor Andrew Hua

November 18, 2023

**Introduction**

This project request comes from the need to analyze sales and margin from the electronic division. By exploring the sales data, a model can be created to generate predictions and determine future profitability of the products offered. Growth in the electronics sector is primarily driven by innovation and accelerated by consumer spending on a global level (Beers, 2022). This project’s main stakeholder is the division in charge of these products or the owner of the store, whoever is requesting the analytics for the data. This data was obtained on Kaggle and is useful because it provides transactional sales data that include costs and margins. The margins can be analyzed to create predictive models.

**Methods**

During exploratory analysis, the data was split into the following three subsets to create a more normal distribution of the target variable: low margin products, high margin products, and very high margin products. This division accounted for the wide range of profit margins observed across product categories.

For predictive modeling, a time series analysis was performed in a Python environment. An ARIMA model was chosen for forecasting due to the univariate nature of the business question and a separate model was built for each of the three subsets identified in the exploratory analysis phase. The specific model that will be used is the Autoregressive Moving Average or ARMA. ARMA “uses a combination of past values and white noise in order to predict future values” (Pierre, 2022). ARIMA, or autoregressive integrated moving average, is a flexible predictive model that is effective for sales prediction. There are three main factors involved in ARIMA predictions (p, d, q). The first is autoregression. Auto-regression means that forecasts are built from the dataset’s own prior values. Integrated refers to the use of differencing, which is a method that replaces raw data with the difference between values to make data stationary. Lastly, moving average is a model that utilizes past residuals, meaning forecasts are built from comparing present values with past error values. These three factors are the parameters for the model. Optimal parameters were determined for each of the three models using the Auto ARIMA function from pmdarima. The orders identified by Auto ARIMA were used in the ARIMA model.

Data was split into a 66/34 training set and test set, with the test set being pulled from the most recent 34% of the dataset. The training set was run through the model, the model’s forecasted values were compared against the test set, and the root mean square error was reported (RMSE). Initial results were mediocre, so the model was rebuilt using walk forward validation. Walk forward validation retrains the model as new data becomes available. It was an effective method for improving RMSE with the small sized data set.

**Results**

When evaluated with RMSE, initial results were subpar. The RMSE of the product subsets were reported as:

Low margin model: 1.970

High margin model: 0.446

Very high margin model: 1.344

Walk forward validation was added to the model which significantly improved the fit of the model. After adding this methodology, RMSE reported as follows:

Low margin model: 0.111

High margin model: 0.136

Very high margin model: 0.392

The following chart displays predicted values plotted against the measured values from the test set of the high margin dataset.

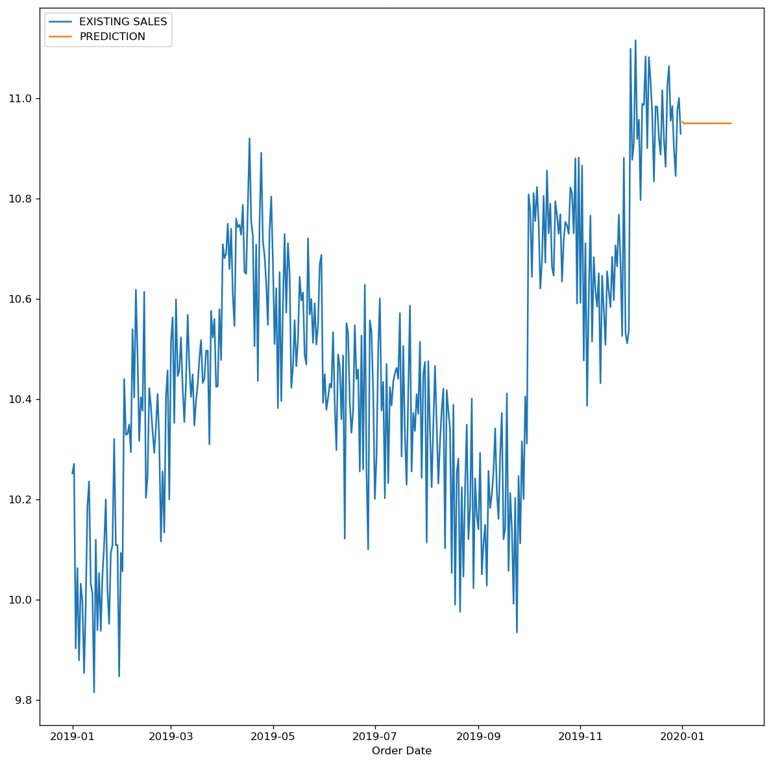
A graph showing different colored lines

Description automatically generated

1.1 Predictions vs Test: High Margin Dataset

With a sufficient fit on the model, data was forecasted for 30 days into the future. All three models suggested that the next month’s profit would be lower than the previous month but remain higher than all other months of the past year. With such strong revenue predictions, we would recommend investing in new products across all categories to increase overall profit.

The weakness of this model is its ability to perform distant future forecasts without additional input data. We found that without further input, predicted profit levels out after 10 days. The following chart displays the predicted profit values for the next 30 days.



1.2 30 Day forecast: High Margin Dataset

**Conclusion**

Our dataset was collected over a relatively small time period (1 year). For this reason, we found that the straightforward ARIMA model was inadequate. Instead, walk forward validation was an effective approach for our dataset. We would recommend retaining this feature during

model maintenance and rebuilding. However, once the collected data grows, this method as coded in this project may be too computationally expensive. At that point we would recommend utilizing a sliding window version of the walk forward validation.

A longer time series also offers the opportunity to introduce an ensemble approach to modeling. With this approach, we believe the predictive power of the model could provide accurate forecasts further into the future. This will allow the company to make more informed financial decisions at the quarterly and annual level.

As it is, this model is most effective at predicting the near future. The company may find it most useful for making staffing decisions for part-time employees and managing cash flow. With a positive outlook across all product groups, we recommend an increased investment in new products, especially those that fall into the very high margin category. When making these investments, the model should be used to monitor cash flow before writing new purchase orders. The model can provide a greater degree of confidence that the revenue flow will support the fulfillment of these purchase orders.

Reference:

Beers, B. (2022, September 6). *Electronics Sector*. Investopedia.

<https://www.investopedia.com/ask/answers/042915/what-electronics-sector.asp>

Cornelius, V. (2023, August 24). Sales Orders. Kaggle. <https://www.kaggle.com/datasets/vincentcornlius/sales-orders>

Pierre, S. (2022, November 4). *A guide to time series forecasting in Python*. Built In. <https://builtin.com/data-science/time-series-forecasting-python>