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

Felipe Rodriguez and Christopher Botts

Bellevue University

DSC 630 Predictive Analytics

Professor Andrew Hua

October 26, 2023

**Milestone 4 – Finalizing Results**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedPrior to conducting the model, data cleansing and transformations were conducted in Milestone 2. Some of these processes involved removing columns, removing nulls, and formatting the date order column to a datetime. The images below show some examples of these transformations.

While implementing the model, further columns are removed. The columns remaining are order date and margin. These columns are selected for the model since they are the most relevant in the data. Some of the columns removed are strings such as Address and Categorie, that do not provide any value in the model.

To aid with further analysis a dictionary was created to be able to view the data by each individual product. This way, the key (product), can be indexed to create a new data frame without running additional code. If future analysis is required on the individual products, this functionality exists. The screen shot below shows the code for this functionality and an example of showing a data frame. A screenshot of a computer

Description automatically generated

**Models**

A model was built for each of the three subsets of the data identified during data preparation: low margin products, high margin products, and very high margin products.

Before building a model, the series was tested for stationarity with the

Augmented Dickey-Fuller test (ADF). In each case, it was determined that the data was non -stationary. An autocorrelation curve supported the results of the ADF test.

A group of graphs with numbers

Description automatically generated

Figure 1. Autocorrelation curves for the low margin subset of the data

Auto ARIMA was then used to determine the best order (p, q) of the autoregressive component(p) and the moving average component(q) and differencing (d). Auto ARIMA used the Akaike information criterion (AIC) to select the best order for the model.

The results were as follows:

Low margin Products - ARIMA(0,1,1)(0,0,0)[0]

High Margin Products - ARIMA(1,1,1)(0,0,0)[0]

Very High Margin Products - ARIMA(0,1,2)(0,0,0)[0]

The residuals were plotted to check for patterns. The residuals fall along a mean of 0, the histogram and Q-Q plots show a good distribution, and the ACF shows no autocorrelation, so the model with designated parameter orders is ready to forecast.

A collage of graphs and diagrams

Description automatically generated

Figure 2. Residual plots of the low margin subset.

A model was built using the order determined from Auto ARIMA. Our model utilized a rolling forecast to improve the accuracy of predictions. Accuracy was calculated using root mean squared error (RMSE) and plotted against the test set.

A graph of a graph

Description automatically generated with medium confidence

Figure 3. ARIMA forecast plotted against test set for low margin subset.

The RMSE for the models were as follows:

Low margin Products - 0.111

High Margin Products - 0.136

Very High Margin Products - 0.392

**Interpretation**

The auto ARIMA function was effective at finding the best order for model parameters. When second order differencing was manually applied to the model against the recommendation of auto ARIMA, the predictions appeared to be closer to the actual values. However, three evaluation metrics showed that the recommended first order differencing, as recommended by auto ARIMA, was more accurate: the RMSE was closer to zero, the AIC was lower and the BIC was lower.

Overall, predictions closely match the test data set and are particularly strong at predicting margins for the near future. The low RMSE values are an indicator that the models are accurate at predicting the test data set and satisfactory accuracy scores were observed across all subsets of the data. Furthermore, forecasts across all models show a positive trend, indicating an expected rise in profit across all subsets of the data. Because all product categories are expecting positive movement, we will provide recommendations to maximize the increase in profit margin.

**Recommendations**

The models show high accuracy and confidence in the products that are in the current lineup. Since the data shows that there will continue to be rises in margin profits, we recommend adding additional products to the line up to expand the items offered. The models provide reassurance that there will be continued profits for the current products; adding more products will be low risk because the current lineup is expected to perform positively.

The biggest disadvantage to this model is that the predictive capabilities span a small time period. After around 30 days of forecasting, the predictions level out, and the predicted value becomes nearly the same. The next iteration of this model should improve how far out the model is capable of predicting. To achieve this, we might incorporate this model in an ensemble model approach.

We recommend refreshing the model monthly to monitor product performance and identify any opportunities in the future. If in any of the quarters the model predictions show a downward trend, investigation can be done to understand why the group of products is underperforming.

**Milestone 3 – Preliminary Analysis**

**Data**

The dataset contains the necessary information to address the business questions asked for this analysis: which product decision should be made to maximize profit margins. Therefore, the driving question does not need to be adjusted. Features in this dataset include sales price, product cost, margin per sale, and datetime stamps. The sales price and datetime features enable us to model revenue predictions. The product cost feature allows us to calculate gross profit from revenue predictions. Lastly, profit margin per sale as a feature can be used to forecast gross profit margins under different scenarios with different product offerings.

The limiting factor of the data is the time period over which data was collected. This dataset contains one year of sales information. If the information were provided over a longer period, we could generate better models and potentially identify seasonal trends.

The data does require modification before it can be used in predictive modeling. The first adjustment is the conversion of the date and time feature from string to datetime. As a datetime object, this feature can be most effectively used in a time series model.

After initial data processing, it was also determined that the dataset should be split into three separate datasets with different product offerings. There was a broad range across the product portfolio in the feature ‘margin,’ which caused a skewed distribution for our target variable, gross profit. The products were split into three separate datasets to create more appropriate dispersions for feeding into a predictive model. The new datasets contain the following sets of products:

**Data Set 1 – Low Margin Products**

AAA Batteries (4-pack)

AA Batteries (4-pack)

Wired Headphones

USB-C Charging Cable

20in Monitor

Lightning Charging Cable

27in FHD Monitor

Bose SoundSport Headphones

Apple AirPods Headphones

**Data Set 2 – High Margin Products**

LG Dryer

LG Washing Machine

Vareebadd Phone

ThinkPad Laptop

Flatscreen TV

Google Phone

34in Ultrawide Monitor

27in 4K Gaming Monitor

iPhone

**Data Set 3 – Very High Margin Products**

MacBook Pro Laptop

Below are a few descriptive statistics that show the dispersion of the feature ‘margin’ across the three datasets.

**Dispersion of 'margin' in split datasets**

|  |  |  |  |
| --- | --- | --- | --- |
|  | low margin | high margin | very high margin |
| mean | 17.96 | 362.79 | 1139.25 |
| 25% | 3.84 | 254.59 | 1139.00 |
| 75% | 49.99 | 469.00 | 1139.00 |

Table 1. Average margin and spread vary widely by dataset. Splitting the initial dataset into three datasets created feature distributions better suited for use in predictive models.

Even after this modification, the distribution of our target variable was still positively skewed. To resolve this problem, the feature was log transformed.

**Visualizations**

There are a few key visualizations that were used to understand and will be used to explain the data. Histograms were important for identifying nonnormal distributions in the dataset. Bar charts were useful for understanding the data and will also be used to explain the results of this project. This type of visualization excels at displaying comparison in an intuitive format. It will be used to compare the contributions of various products to gross revenue, gross profit, and sales counts.

A graph of sales and sales

Description automatically generated

Chart 1. Comparison of individual product revenues. This type of chart shows the degree to which certain products contribute to gross revenue compared to other products.

Line and area charts will also be used to explain the data. These visualizations are good for displaying value change over time. Current revenue over time and the results of the time series model will be well displayed with these two types of visualizations.

**Conclusion**

With the preliminary analysis conducted, the original expectation can be adjusted slightly. The original plan was to uncover which products are most profitable or carry the most interest among our consumers. While the original plan will still be conducted, another expectation is to understand the least profitable items and provide recommendations with the information provided by the model. Additionally, the products will be explored by low margin, high margin, and MacBooks. The division data will allow a more thorough analysis and additional accuracy when conducting the models.

The original model, time series forecast (ARMA), will still be pursued for this analysis. However, it will be conducted with three subsets, low margin, high margin items, and MacBooks, instead of looking at all products in the dataset. In the preliminary analysis, since the data was skewed to high margin items, the division of the data will allow for exploration of the items that still are sold with great frequency, but the margins are not as high due to the original value of the product. Separating the MacBooks from the data set also allows for sales forecasting of that product.

In conclusion, after initial exploration, the dataset selected provides useful information that can continue to be explored by a model. While the model selected and data remains the same, there is a slight adjustment to how the data will be used. The goal of the term project is within scope, if there are additional challenges encountered those will be discussed and documented, and the contingency plan will be conducted instead.

**Milestone 2 – Project Plan**

Growth in the electronics sector is primarily driven by innovation and accelerated by consumer spending on a global level (Beers, 2022). This growth has induced market entry by new retailers. The rise of online shopping has also fragmented the market and created a highly competitive market (Beers, 2022). An analysis of this market could prove beneficial and will uncover the possibilities for investments and areas of concern.

**The Data**

This data is from a retail company that sells a variety of electronic products in the United States. This data was obtained from Kaggle and contains sales data by consumer for the year 2019. Products include small ticket items such as batteries and more expensive products like MacBooks.

The history of sales and consumer information collected by this company provide the opportunity to analyze the sales of specific product categories in the electronics market. Furthermore, this dataset includes information about product costs that can be used to find profit margin. This enables us to determine not only which products bring in the most revenue, but also which products are the most profitable. The data also includes the quantity of items sold and the purchasing address.

**What we hope to learn**

By analyzing the data, we hope to uncover which products are most profitable or carry the most interest among our consumers. The products with the leading profits will be further analyzed to forecast sales and see how the performance will do over time. This project will also observe the items with the least profits and interest. This analysis will show where improvements can be made and determine if products need to be discontinued or if additional investment is needed. The models explored will allow us to forecast profitability over time. With the products identified and studied, a recommendation can be provided to the stakeholders. This analysis would be most useful to stakeholders and will allow them to make decisions that can allocate company resources and future investments to the products analyzed. Insights can be used to drive R&D investments to make informed innovation decisions. Product managers would also be able to utilize this information to make marketing decisions. By using the results, product managers can determine what the key marketing differences are for products that perform better than others. If there are notable areas that can be improved, these adjustments can be made, and the new results can be analyzed.

**The Models – Analysis and Evaluation**

The plan for research is to conduct a Time Series Forecast. 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). Prior to implementing and running the model, the individual sales of each product will be analyzed and the highest selling versus least selling will be selected and used in the model. This model will allow us to see which products will perform better or worse in the future. To confirm that an ARMA model will be effective for this data, an augmented Dickey-Fuller test (ADF) will be performed to determine the stationarity of the time series. Assuming the time series is stationary, we will proceed with an ARMA model. Otherwise, we will implement an ARIMA model, so that trends can be removed, and the time series can be made stationary.

When building the ARMA modeling, we will determine the order (p, q) of the autoregressive component(p) and the moving average component(q). This can be accomplished by plotting the autocorrelation and partial autocorrelation functions of the time series. Another option for finding the (p, q) values is to use the Auto ARIMA function. The dataset will be split into training, validation and testing sets. To avoid the look ahead bias, the split will be based on time, with the testing set being the most recent part of the data. The performance of the model will be evaluated and compared using Bayesian Information Criterion (BIC); we are looking for the model with the lowest BIC value. The goal of utilizing BIC is to ensure the model is not overfitted but includes enough parameters to capture patterns in product sales.  After deciding the order for ARMA model, we will use maximum likelihood estimation (MLE) to determine the coefficients. This will provide the maximum likelihood estimate – the point at which the observed data in our dataset is the most probable*.*

**Risks and Ethical Concerns**

This analysis has ethical concerns we want to note. This data only contains the electronic division data, which is not all encompassing of the entire company. As we explore this data and publish it, it does not contain or reflect any changes in cost or sale price, meaning that no seasonal discounts or promotions will be included in the data analysis. When consumers look at our data, they will not be able to identify target dates where there are promotions and price reductions.

**Contingency Plan**

When conducting research, if there are any issues or roadblocks that we cannot overcome, there is a contingency plan in place. Since the data includes purchasing address, the alternative that can be explored is identifying the most profitable locations, allowing for a geographical analysis. By studying the data geographically, we can make marking assumptions and determine which areas to focus marketing in. This discovered information can be relayed to the marketing team for further analysis and exploration to understand the differences between the analyzed areas. The same models can be used for this contingency plan; however, the layout will be by geographical location as opposed to by product.

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