Final Report

Problem Statement

A UK-based online store aims to enhance its operational efficiency by accurately forecasting sales for its range of products. The company operates within the e-commerce sector, catering to customers across various countries.
It is essential for stores to have an idea of how much sales to expect as it will determine their decision on how soon/ how often to restock products they have in their inventory. Thus, stores rely on sales/ demand forecasts to make informed decisions about inventory restocking frequencies. Anticipating demand will enable the company to optimize inventory management, plan marketing strategies, and enhance customer experience, by aligning strategies with anticipated consumer demand.
The primary objective of this project was to develop a sales forecasting model leveraging one year of historical sales data of the online store. The insights gained from the model can be used to optimize inventory levels of the online store and reduce the cost of excess inventory, allowing the store to tailor their marketing and sales strategies.

Data

The raw dataset contains 541909 records and 8 attributes, organized in a long format.
It contains the following:
- Product information: StockCode, Product Description, Unit Price.
- Customer data: Customer ID and Country
- Sales data: Quantity sold, Invoice date and Invoice number
- other attributes: category of the product is not available in the data
- Output variable (desired target): Quantity (sales)

Key Data source:

The dataset is taken from the online platform:
Online Retail. (2015). UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.
Link: (https://archive.ics.uci.edu/dataset/352/online+retail)

Data Wrangling

The raw data required some cleaning, the sales data seems to be mixed with accounting records. Some of the potential issues that were identified are:

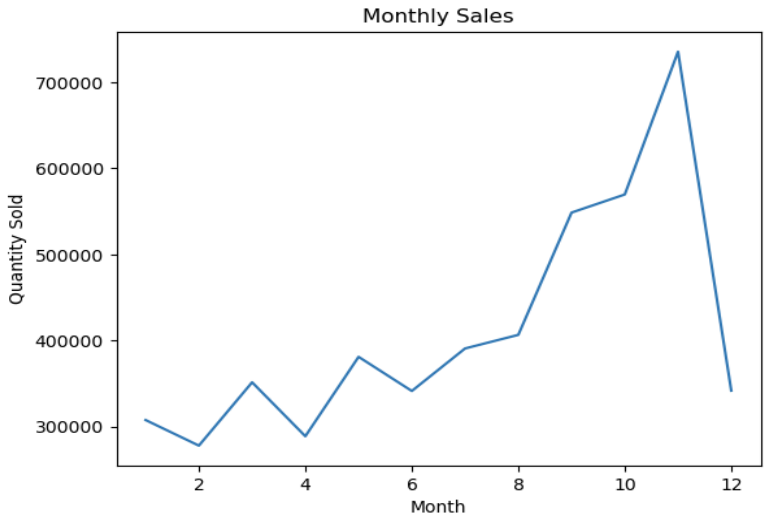
● Missing data: There are records missing CustomerID, and a few missing Descriptions. Since this is a time series forecast CustomerID was not very useful, so it was left as is. However description is an important feature that needed to be used to make product categories so the missing values were removed.
The histogram of InvoiceDate shows a gap in the data as there were no transactions recorded between Dec 24, 2010 and Jan 4, 2011 either due to the website being down or the transactions occurred but were not recorded.

● Out-of-range data/ potential errors: In the ‘Country’ attribute, the records - 'European Community' is represented as a separate record, since it does not represent a country, combining it with the 'unspecified' records was considered.
There are two transaction records with a negative UnitPrice value. The Description of these records implies that transactions are 'adjusting for bad debts' i.e. they may be entered to correct errors in accounting rather than being real transactions. Therefore, they were removed.
There are few records with a negative Quantity which could indicate purchase returns. These values were later removed.
The records are not all sales records, the data includes what could be returns, accounting adjustments, Amazon fees, shipping fees and so on. Therefore, I've filtered the data based on the StockCode to remove the non-transactional/accounting data.

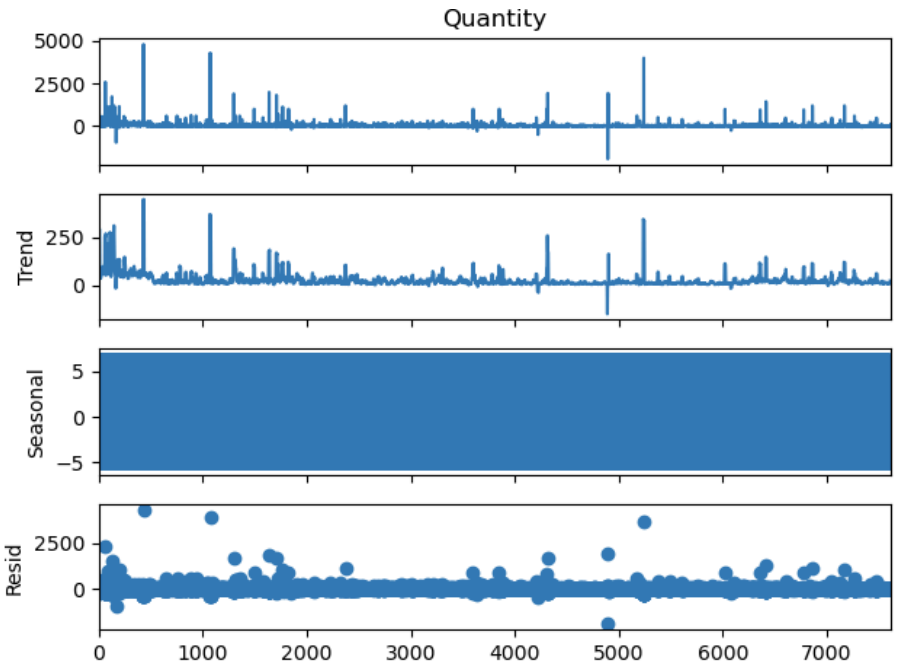
● Skewness - Both Quantity and UnitPrice are highly skewed i.e Quantity has a really high outliers on the upper and lower end and UnitPrice has a really high outliers on the upper end that need to be looked at the following EDA.

Data Exploration

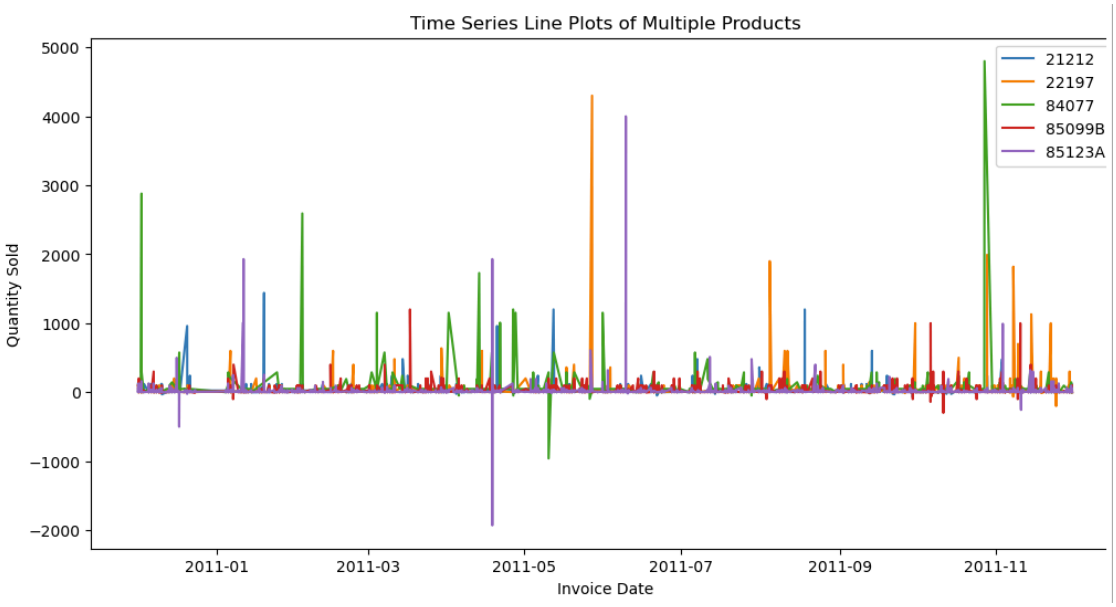
In the initial explorations it’s evident that, throughout the year, the online store recorded a substantial number of transactions, totaling 25,900 purchases. These purchases comprised 5,176,450 individual items sold. On average, around 10 items were sold per invoice, but the data also showed extreme outliers in both the minimum and maximum values that are potentially valid. Among the 4,070 product types sold, the 'POPCORN HOLDER' (stock code = 22197) was the most popular product often sold in bulk, while the 'HANGING HEART T-LIGHT HOLDER' (stock code = 85123A) was frequently sold but not typically in bulk quantities. The store served a customer base of 4,372 individuals across 37 countries plus 1 unspecified category, each making at least one purchase throughout the year.



Looking at the sales data, grouped by month, shows a gradual rise in sales after a slow start in the earlier months, followed by a substantial surge in September, peaking significantly in November before a sharp decline. November was identified as the month with the highest sales, while February being the lowest. Regarding the day of the month, there's no distinctive pattern observed, yet a slight decrease in sales towards month-end is noticeable. Saturdays indicate no recorded sales, while Thursdays show a slight increase in sales compared to other weekdays. Furthermore, the sales gap observed between December 24, 2010, and January 4, 2011, as identified during data wrangling, is shown in the sales data.



The time series sales data doesn't exhibit a clear upward or downward trend, it also does not have a repeating seasonal pattern but mainly consists of random fluctuations or noise. After performing an Augmented Dickey-Fuller test, resulting in a p-value of 0.01, I've concluded that this time series data is stationary.



The sales trend for the best selling products also seem to be stationary, with a few spikes from time to time.

Feature Engineering

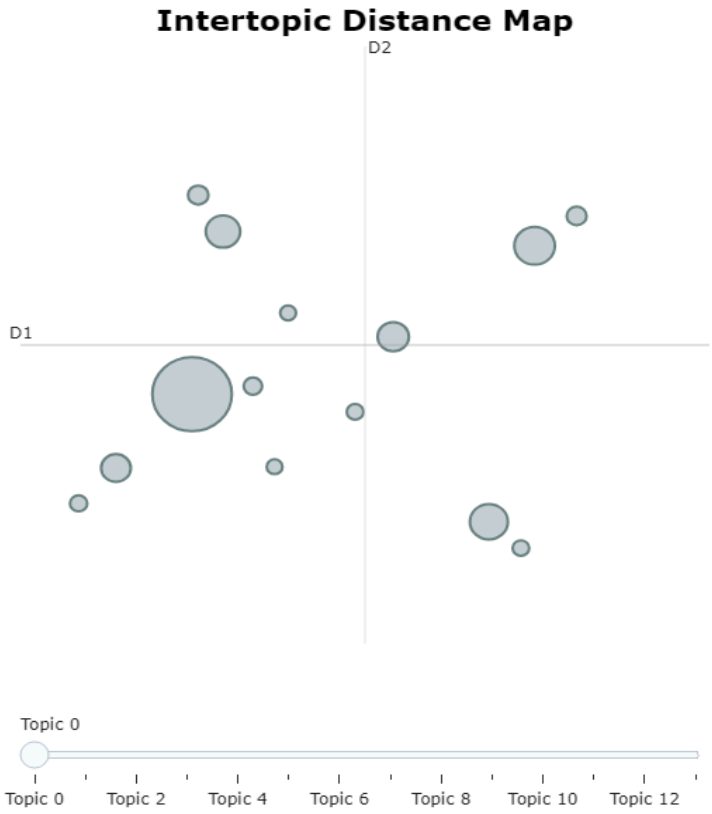
Creating Product Categories

Since there are over 4000 products of mostly household items to forecast, to make the task a little easier, I've used an NLP model-- BERTopic to categorize products into distinct groups using the product description attribute. Subsequently, I constructed a comprehensive pipeline to preprocess the time-series data, ensuring it is well-structured and ready for integration into the forecasting model.

To train the BERTopic model, I used a semi-supervised approach where I input a small amount of manually categorized data along with the abundance of unlabeled/uncategorised data to improve the model's performance.

Topic	Count
0	1400
-1	855
1	364
2	319
3	262
4	217
5	195
6	92
7	87
8	76
9	66
10	62
11	61
12	56
13	55

The model identified 14 categories of products - Home decorations, kitchen items, light holders, jewelry, bags, signs, vintage items, children's toys, stationary items, gifts, arts & crafts and a collection of miscellaneous products(i.e. Category -1).



Model Selection

After finalizing the categories, I explored three different time series forecasting models for a specific product sales category.

When analyzing the data trend for the first category of products(Category 0 - home decorations), it was observed to be stationary, and no differencing was necessary before fitting the ARIMA model.

- The ARIMA model yielded a Mean Absolute Percentage Error (MAPE) of 17.62%, which is acceptable, but has room for improvement.

Among the three models tested, the Prophet model proved to be the most effective and adaptable to the dataset, achieving a MAPE of 14.41%. On the other hand, the LSTM model demonstrated lower performance, primarily due to the limited dataset of 48 weeks and the complexity of tuning numerous hyperparameters.

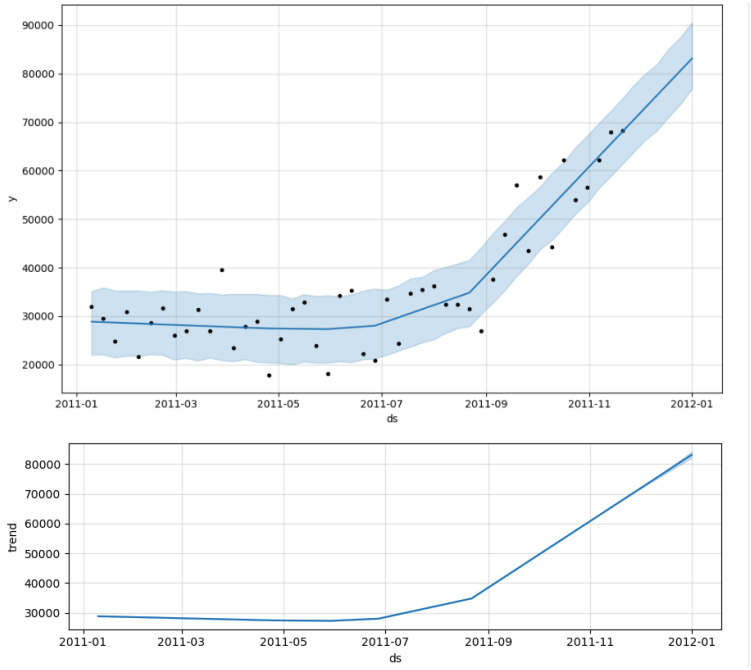
Taking all these factors into account, the Prophet model was ultimately chosen for the modeling phase to be implemented on all 14 categories.

The Final Model

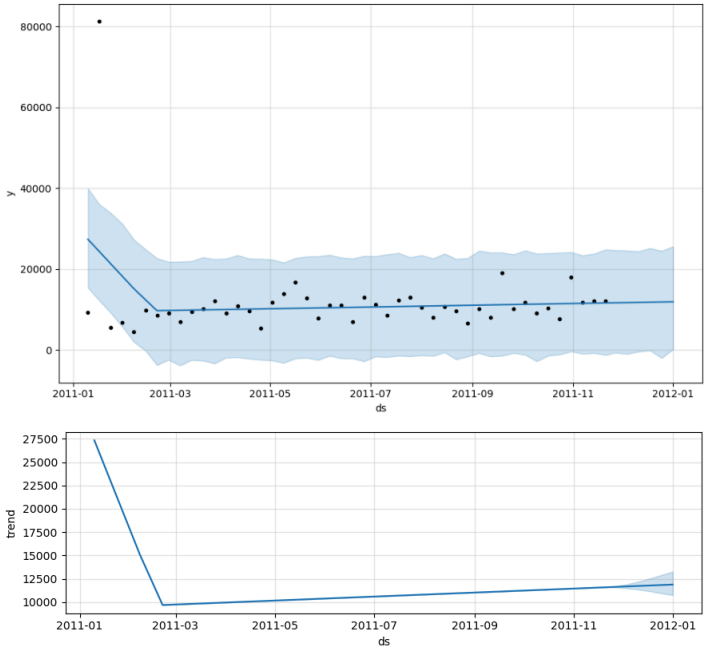
In this phase of Final Modeling, I performed an iterative process of hyperparameter tuning for each category to optimize the model's performance and improve the accuracy of the forecasting model. Mean Absolute Percentage Error (MAPE) was used as the model evaluation metric for the hyperparameter tuning as it is a reliable measure for gauging the accuracy of predictions especially in the context of time series forecasting.
The hyperparameter tuning involved adjusting various parameters of the model, such as changepoint prior scale, seasonality prior scale, holidays prior scale and seasonality mode. These adjustments aimed to identify the combination of parameters that minimized the MAPE, resulting in a model that could provide more accurate predictions for each category.
After the rounds of hyperparameter tuning, the final model has demonstrated a good performance by accurately forecasting sales for 9 out of the 14 categories for which the Mean Absolute Percentage Error (MAPE) consistently remained below the threshold of 18%.

Results

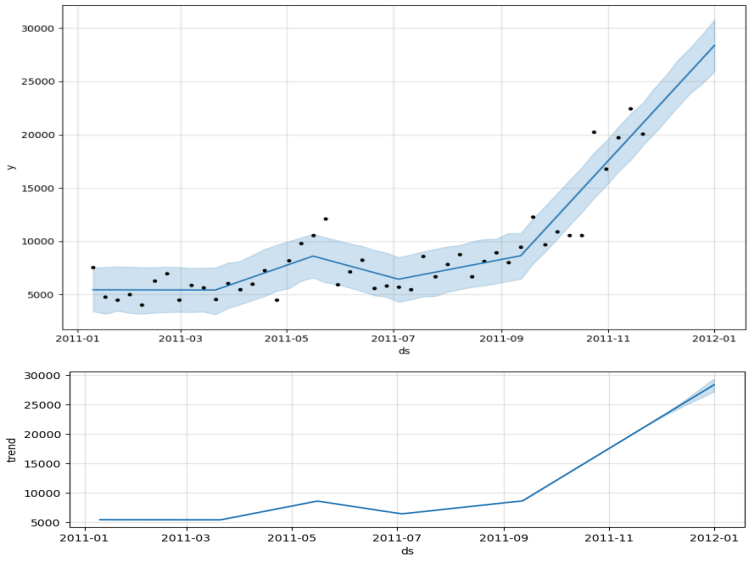
- Miscellaneous (Category -1):
  Sales have been consistently low, around 30,000, but are showing an increasing trend since September. The forecast predicts a continued rise until the end of December.



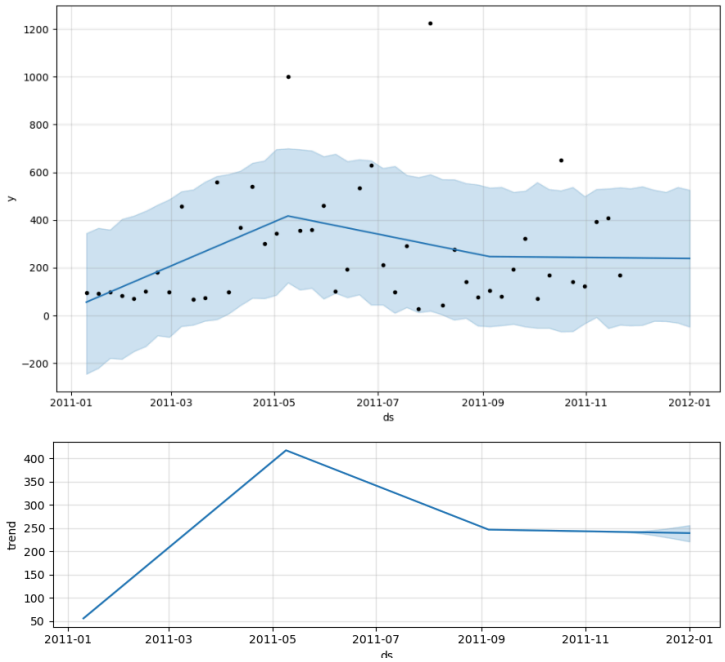
- Home Decorations (Category 0):
  Starting high in January, sales surge in mid-February, stabilizing at around 10,000. The forecast indicates that these levels will be maintained throughout December.



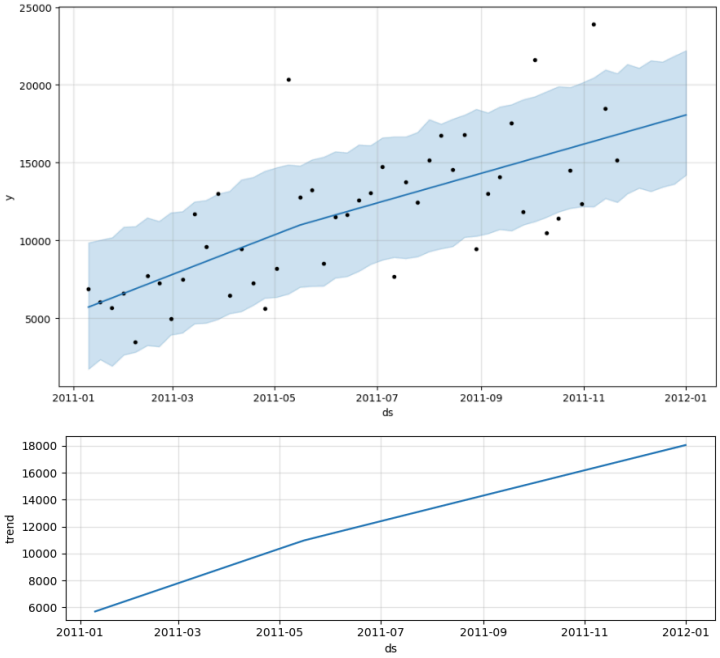
- Kitchen Items (Category 1):
  Sales increased in May, dipped in July, and rose again after September. The forecast anticipates a further increase, reaching upwards of 25,000 in sales.



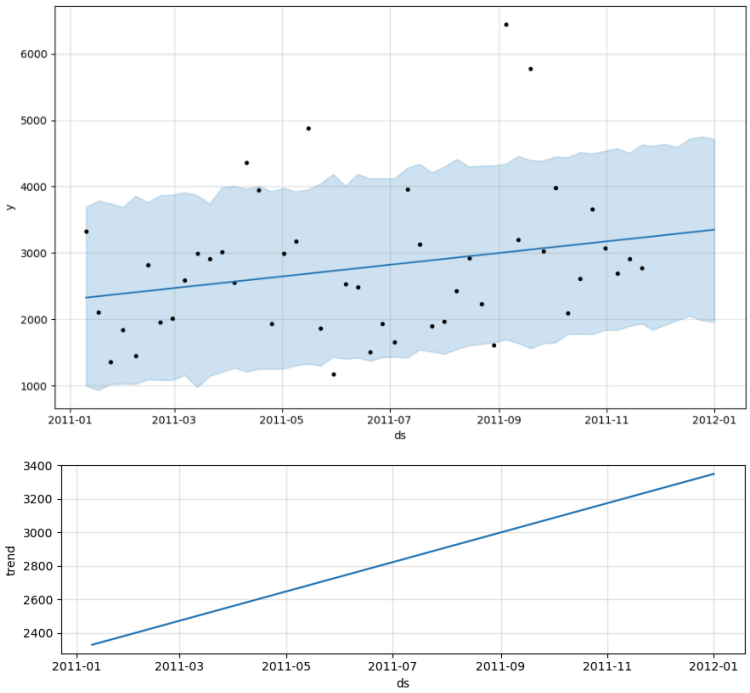
- Light Holders (Category 2):
  Experiencing a slight sales increase in mid-May, light holders have stabilized around 200 units. The forecast suggests these levels will be maintained.



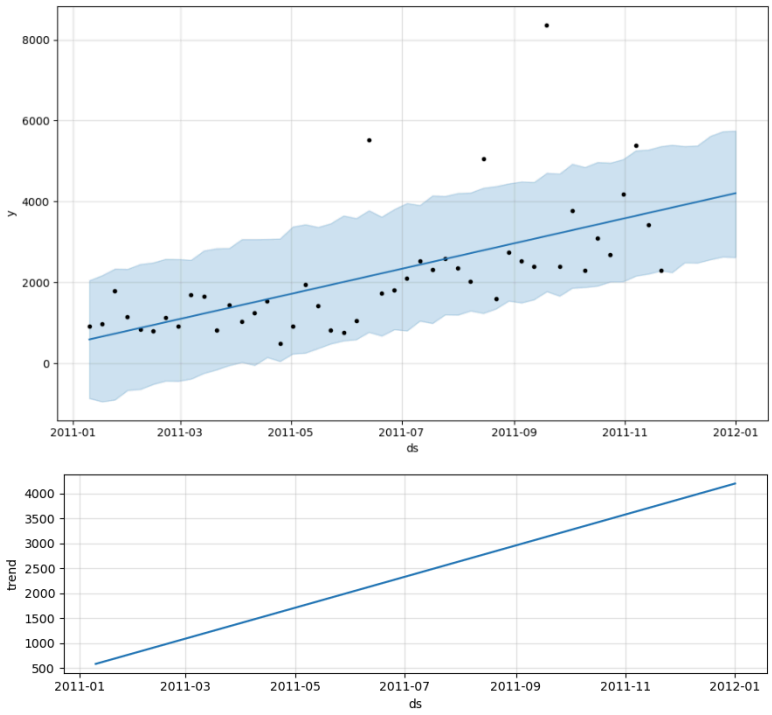
- Bathroom Items (Category 4):
  Sales have shown a consistent increase, growing from 5,000 to nearly 18,000 by November. The forecast predicts a continued upward trajectory in December.



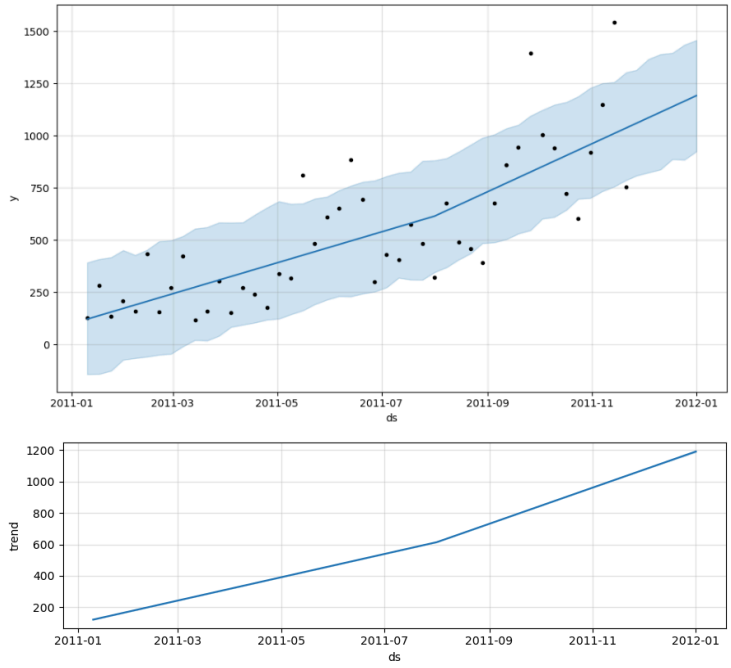
- Bags (Category 5):
  Steadily increasing since January, starting at 2,000 sales and reaching 3,500 by November. The forecast suggests a continued steady rise.



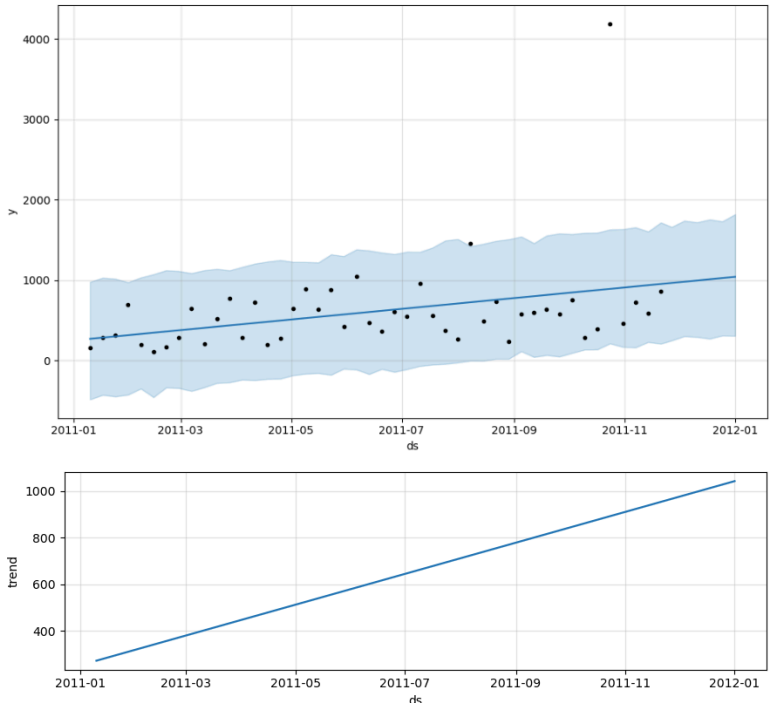
- Signs (Category 6):
  Similar to bags, signs have shown a consistent increase, going from 1,200 to over 4,000 in sales. The forecast anticipates this upward trend continuing.



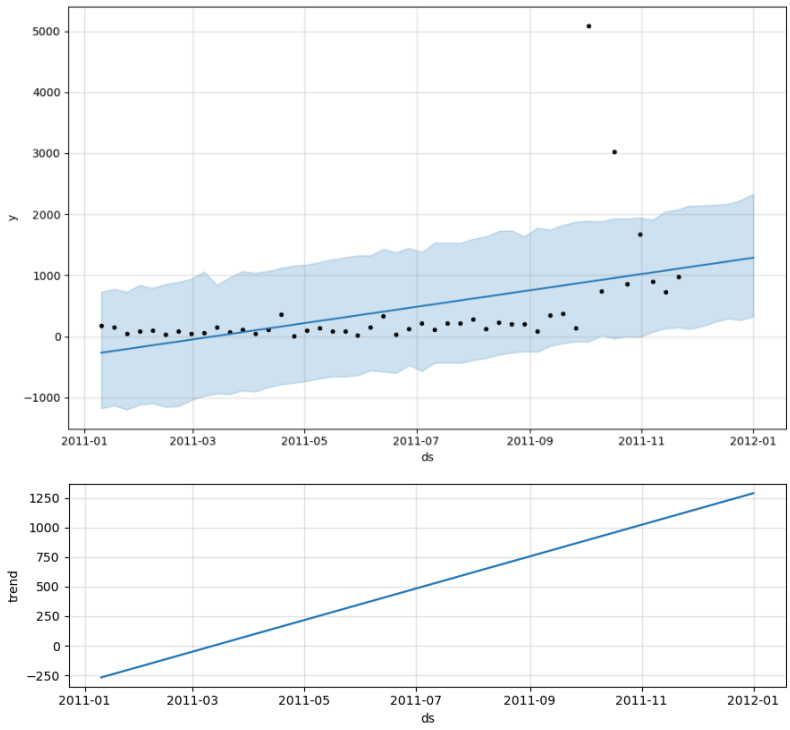
- Children's Toys (Category 8):
  Sales have been increasing, with a brief stability in late August, followed by continued growth. The forecast predicts an increase in December, reaching up to 1,250 sales.



- Stationary Items (Category 10):
  Sales have remained relatively stable, reaching almost 1,000 items in November. The forecast indicates a slight increase to just over 1,000 in December.



- Gifts (Category 12):
  Sales have been consistently below 1,000 items each month but exhibit a notable increase starting mid-November. The forecast suggests this upward trend will persist in December.



Conclusion and Recommendation

While the model was able to accurately forecast sales of the 9 categories, it did not achieve a good level of accuracy for some categories. Category 7 in particular has the highest MAPE, indicating a significant error of 46%. Similarly, Categories 11, 13, and 9 also have high MAPE values, exceeding 30%.
Therefore, using the current model for forecasting Category 7 items is not recommended, as it may result in suboptimal inventory management, inefficient resource allocation, and other potential business decision inaccuracies.

In the next steps, to enhance the accuracy of predictions for these categories, it is advisable to delve deeper into their characteristics. This could involve acquiring more data, extending the observation period, or exploring alternative models in future research.