Sales Forecast

By Lina Abdullahi

The Problem

- 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.

The solution

- A **sales forecasting model** leveraging one year of historical sales data of the online store can be used to generate forecasts that can optimize inventory levels of the online store and reduce the cost of excess inventory, allowing the store to tailor their marketing and sales strategies.

The Data

The raw dataset contains **541909 records** and **8 attributes**

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

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, and had potential issues that were identified:

- **Missing data:** There are records missing CustomerID, and a few missing Descriptions. 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. All missing data were removed before model building.

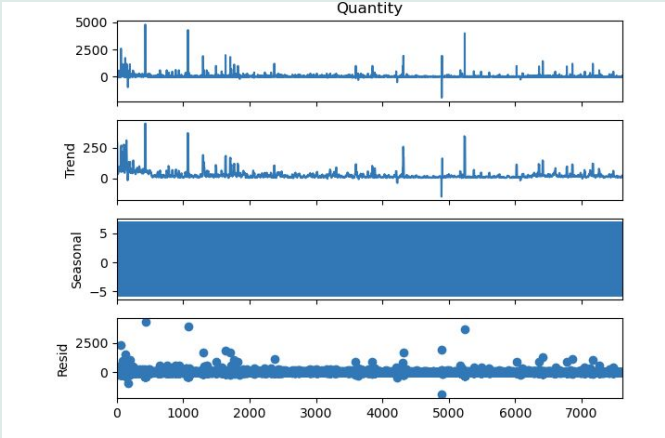
- **Out-of-range data/ potential errors:** In the ‘Country’ attribute, the records - 'European Community' is represented as a separate record, since it does not represent a country, combining it with the 'unspecified' records was considered.There are transaction records with a negative UnitPrice value that may have been entered to correct errors in accounting rather than being real transactions. There are few records with a negative Quantity which could indicate purchase returns. These values along with non-transactional/accounting data were later removed.

- **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.

The Analysis

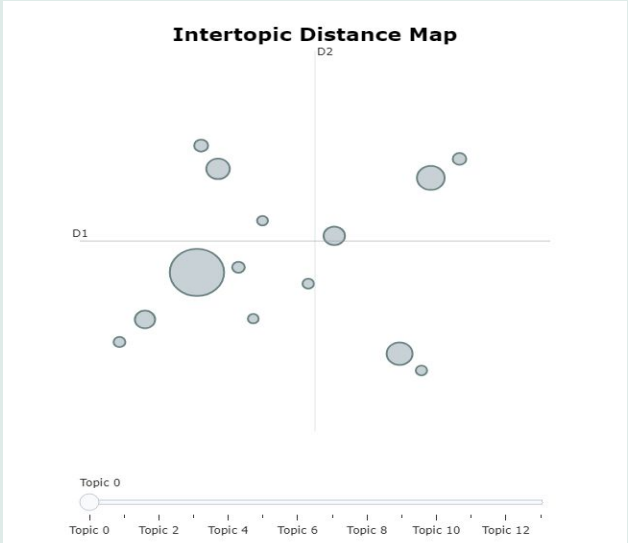
The initial data exploration revealed that:

- Throughout the year, the online store recorded a substantial number of transactions, totaling 25,900 purchases. On average, around 10 items were sold per invoice. And 4070 product types were sold.
- Looking at the sales pattern in time, the data 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.
- 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. Augmented Dickey-Fuller test, resulting in a p-value of 0.01, indicated that this time series data is stationary.



Feature Engineering

- An NLP model- BERTopic was used a semi-supervised approach to categorize products into distinct groups using the product description attribute.
- 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).



Forecasting Model

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

- **The ARIMA model** yielded a Mean Absolute Percentage Error (MAPE) of 17.62%, which is acceptable, but has room for improvement.
- **The Prophet model,** among the three models tested, proved to be the most effective and adaptable to the dataset, achieving a **MAPE** of **14.41%.**
- **The LSTM model** demonstrated lower performance, primarily due to the limited dataset of 48 weeks and the complexity of tuning numerous hyperparameters.

Model Training

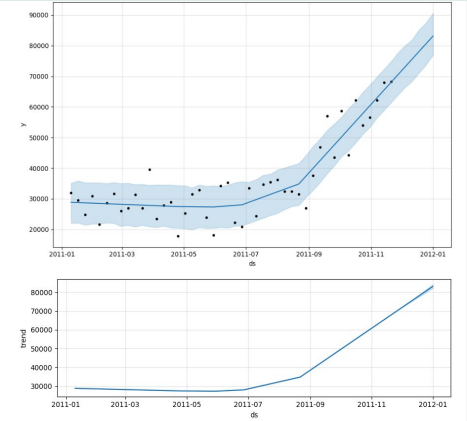
Hyperparameter tuning was performed to identify the combination of parameters that minimized the Mean Absolute Percentage Error (MAPE), for each category. The tuning involved adjusting various parameters of the model, such as:

- changepoint prior scale,
- seasonality prior scale,
- holidays prior scale and
- seasonality mode.

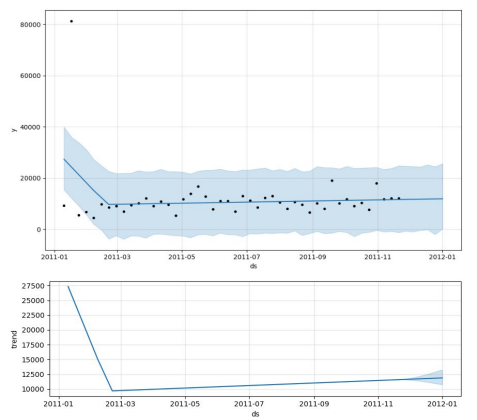
After the rounds of hyperparameter tuning, the final model has demonstrated a good performance by accurately forecasting sales for 9 out of the 14 categories for which the Mean Absolute Percentage Error (MAPE) consistently remained below the threshold of 18%.

Results

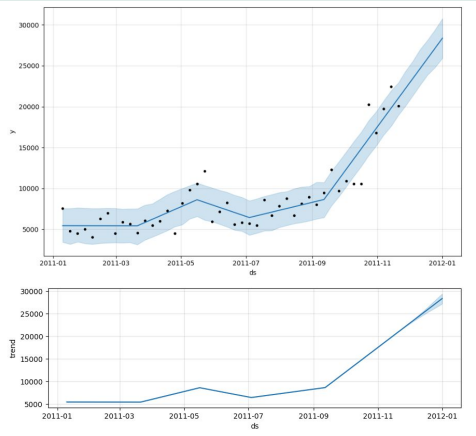
- **Miscellaneous (Category -1):**
  Sales have been consistently low, around 30,000, but are showing an increasing trend since September. The forecast predicts a continued rise until the end of December.



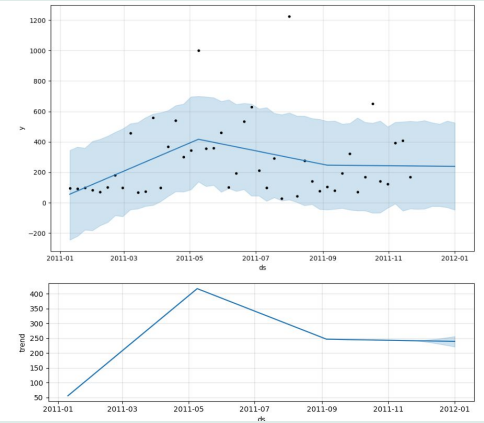
- **Home Decorations (Category 0):**
  Starting high in January, sales surge in mid-February, stabilizing at around 10,000. The forecast indicates that these levels will be maintained throughout December.



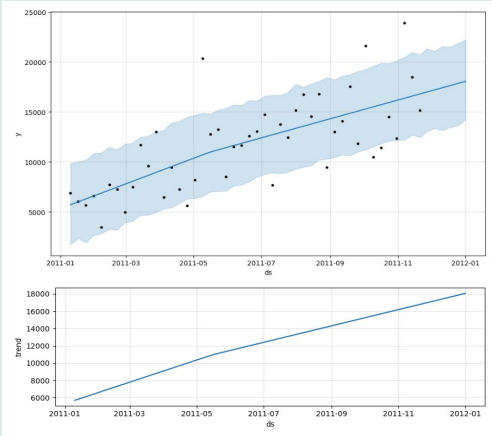
- **Kitchen Items (Category 1):**
  Sales increased in May, dipped in July, and rose again after September. The forecast anticipates a further increase, reaching upwards of 25,000 in sales.



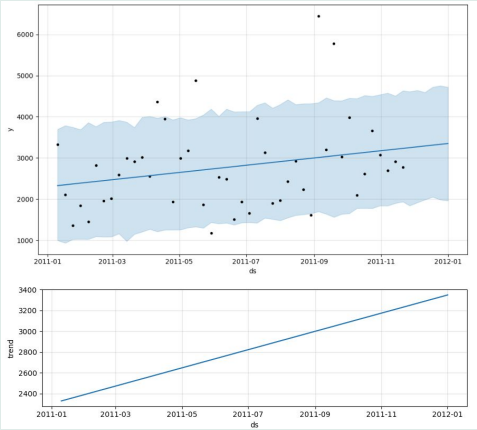
- **Light Holders (Category 2):**
  Experiencing a slight sales increase in mid-May, light holders have stabilized around 200 units. The forecast suggests these levels will be maintained.



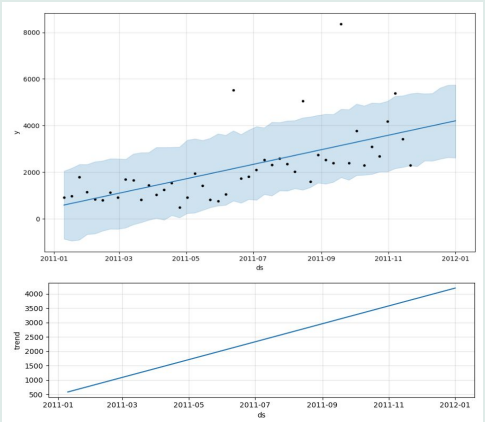
- **Bathroom Items (Category 4):**
  Sales have shown a consistent increase, growing from 5,000 to nearly 18,000 by November. The forecast predicts a continued upward trajectory in December.



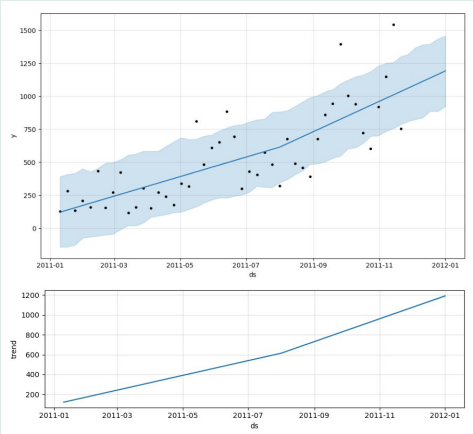
- **Bags (Category 5):**
  Steadily increasing since January, starting at 2,000 sales and reaching 3,500 by November. The forecast suggests a continued steady rise.



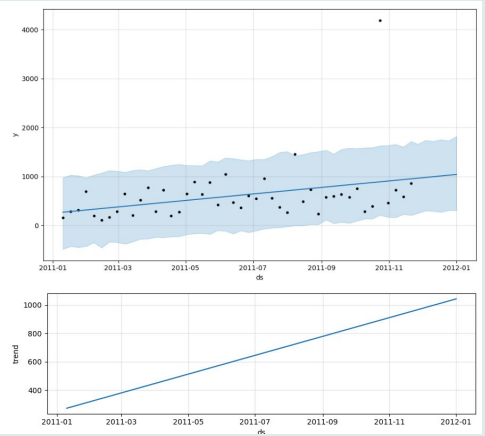
- **Signs (Category 6):**
  Similar to bags, signs have shown a consistent increase, going from 1,200 to over 4,000 in sales. The forecast anticipates this upward trend continuing.



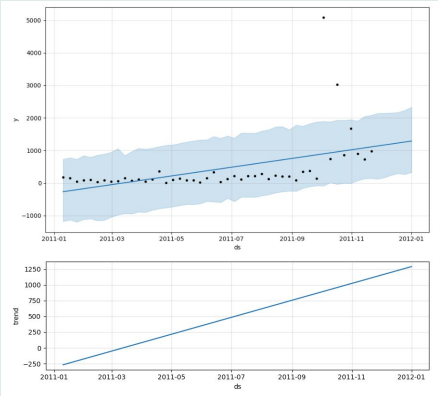
- **Children's Toys (Category 8):**
  Sales have been increasing, with a brief stability in late August, followed by continued growth. The forecast predicts an increase in December, reaching up to 1,250 sales.



- **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

In conclusion, 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%.

The model can be used to forecast the next level of sales for the 9 product categories on a bi-monthly basis. However, using the current model for forecasting Category 7 items as well as Categories 11, 13, and 9 is not recommended, as it may result in suboptimal inventory management, inefficient resource allocation, and other potential business decision inaccuracies.

Recommendation

- 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.

Questions?