THE UNIVERSITY OF TEXAS AT AUSTIN

McCOMBS SCHOOL OF BUSINESS

DEMAND ANALYTICS/PRICING
Spring 23

Sales Forecasting for Grocery Store

By:

Kavya Angara (ka32577)
Rajshree Mishra (rm62528)
Rochan Nehete (rrn479)
Sai Bhargav Tetali (srt2578)
Prathmesh Savale (ps33296)

Table of Contents

1. Introduction

Sales forecasting is an essential aspect of running any business, and the grocery industry is no exception. A grocery store needs to accurately predict future sales to manage inventory, plan promotions, and optimize their operations. However, forecasting sales can be a complex and challenging task due to various factors such as seasonality, changing consumer preferences, and unexpected events.

In this project, we aim to develop a sales forecasting model for a grocery store that leverages historical sales data, external factors such as weather, and other relevant variables. Our goal is to provide the grocery store with accurate predictions of future sales to help them make informed decisions about inventory management, staffing, and promotions.

To achieve this objective, we will use machine learning algorithms and data analysis techniques to analyze the available data and identify patterns and trends. We will also evaluate the performance of the model and fine-tune it to improve its accuracy.

By providing an accurate sales forecasting model, this project aims to help the grocery store optimize its operations, reduce waste, and increase profitability. Ultimately, this project can serve as a valuable tool for any grocery store looking to improve their sales forecasting capabilities and make data-driven decisions.

1.1 Inventory Problems

Lack of popular items, lost revenue, and extra product waste are all common inventory problems that grocery stores can face.

When a grocery store does not have popular items in stock, it can lead to dissatisfied customers who may choose to shop elsewhere. This can result in lost revenue and

decreased customer loyalty, which can have long-term negative effects on the store's profitability. Additionally, a lack of popular items can create an imbalanced inventory, with some products overstocked and others understocked, leading to inefficiencies in the supply chain and a higher risk of excess waste.

On the other hand, overstocking can also lead to waste when products expire before they can be sold, resulting in additional costs for the grocery store. Overstocking can also create storage and handling challenges, leading to damaged goods, increased spoilage, and other losses.

1.2 Forecasting Strategy

In brick and mortar stores, forecasting sales is crucial in determining how much inventory to purchase. Without accurate forecasting, stores risk understocking or overstocking products, both of which can lead to lost revenue and wasted resources.

To develop an effective forecasting strategy for brick and mortar stores, retailers can utilize various data sources, including historical sales data, trends in consumer behavior, and external factors like weather and seasonality. These sources of data can be analyzed using various forecasting techniques, such as time series analysis and machine learning algorithms, to predict future sales accurately.

One effective strategy for forecasting sales is to break down historical data into smaller units, such as hourly or daily sales. This approach can provide more detailed insights into sales patterns, helping retailers understand peak sales periods, popular products, and trends in customer behavior. Retailers can then use this information to optimize their inventory levels, ensuring that they have enough stock to meet demand during busy periods while minimizing waste during slower periods.

Another approach is to use real-time data feeds to monitor sales in real-time and adjust inventory levels accordingly. For example, retailers can use point-of-sale systems to track sales and adjust stock levels automatically, ensuring that popular items are always in stock and that overstocking is avoided.

Ultimately, an effective forecasting strategy for brick and mortar stores requires a combination of data analysis, industry knowledge, and a keen understanding of

customer behavior. By utilizing accurate sales forecasting, brick and mortar stores can optimize inventory levels, reduce waste, and improve profitability.

2. Problem Statement

The problem statement is to develop a sales forecasting model that can accurately predict future daily sales at the store-product level. The objective is to optimize resource utilization and improve operational efficiency by ensuring that the right products are stocked in the right quantities to meet customer demand while minimizing waste and excess inventory. To achieve this, the model needs to leverage historical sales data and other relevant variables to identify patterns and trends, which can then be used to generate accurate sales forecasts. The end goal is to provide decision-makers with actionable insights that can help them make informed decisions about inventory management, staffing, and promotions to maximize revenue and profitability.

3. EDA

3.1 Dataset Description

The time series data set covers the period from 2013 to August 2017 and includes various data categories such as Holiday Events, Oil, Stores, and Promotions.

- The Holiday Events data includes information such as the date, type, locale, location, description, and transferred status of each event. This data can be used to analyze how holiday events impact sales across different locations and product categories.
- The Oil data includes the date and price of oil, which can be used to analyze how changes in oil prices impact the economy and consumer behavior.
- The Stores data includes information about each store, such as the store number (store_nbr), city, state, type, and cluster. This data can be used to analyze store performance and identify patterns and trends in sales across different locations and store types.

- The Promotions data includes information about sales promotions, such as the date, duration, and type of promotion. This data can be used to analyze the effectiveness of promotions and their impact on sales.

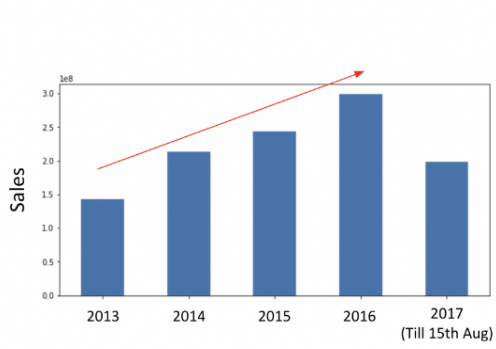
Overall, the time series data set covers a broad range of variables that can be used to analyze sales patterns and trends across different product categories, locations, and time periods. It provides a rich source of data for forecasting, trend analysis, and identifying opportunities for improving sales and operational efficiency.

For Model evaluation, we used Mean absolute percentage error (MAPE), which is a widely used metric for evaluating the accuracy of a forecasting model. It measures the average percentage difference between actual and predicted values, with lower MAPE scores indicating better forecasting accuracy. MAPE is calculated by taking the absolute difference between actual and predicted values, dividing this value by the actual value, and then multiplying the result by 100. The MAPE score represents the average percentage difference between actual and predicted values across all observations in the dataset. MAPE is a useful metric for evaluating forecasting models as it provides a simple, intuitive measure of accuracy that can be easily compared across different models and datasets.

3.2 Data Exploration

To gain insights into the performance of Favorita, several factors can be analyzed.

- The overall sales trend can be evaluated to understand the company's growth trajectory over time.



Top 4 Stores by Sales

Store No.	Sales	City	State
44	63.3K	Quito	Pichincha
45	55.6K	Quito	Pichincha
47	52.0K	Quito	Pichincha
3	51.5K	Quito	Pichincha

Last 5 Stores by Sales

Store No.	Sales	City	State
35	7.8	Playas	Guayas
30	7.5	Guayaquil	Guayas
32	6.1	Guayaquil	Guayas
22	4.2	Puyo	Pastaza
52	2.7	Manta	Manabi

Figure 1

- Time series analysis can be conducted for each store and product type to identify trends, patterns, and potential areas for improvement.

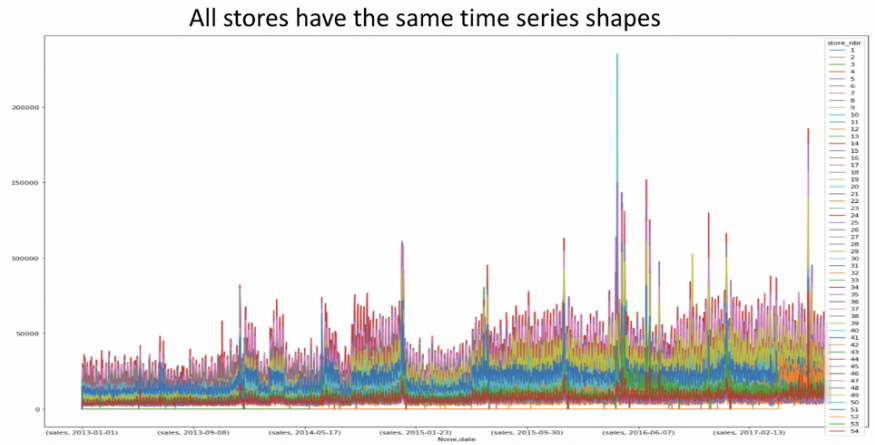


Figure 2

- Time series analysis can also be performed for product families to identify which product categories are driving sales growth.

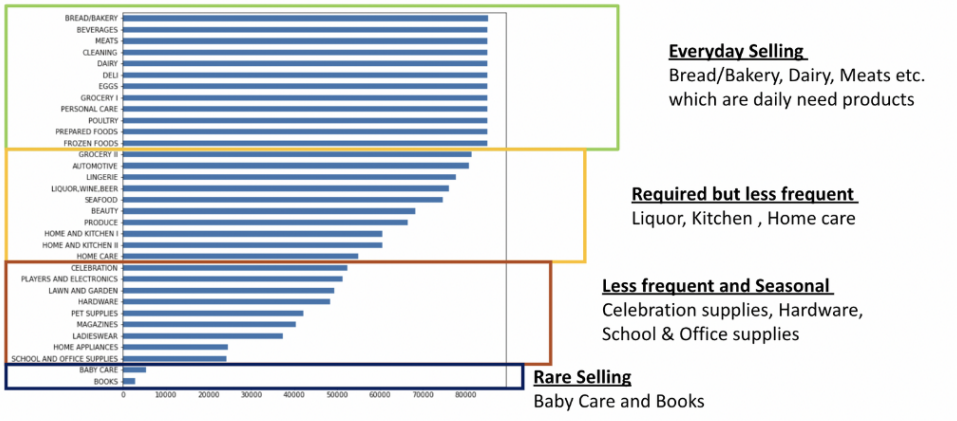


Figure 3

- External factors such as holidays, promotions, and oil prices can have a significant impact on sales.

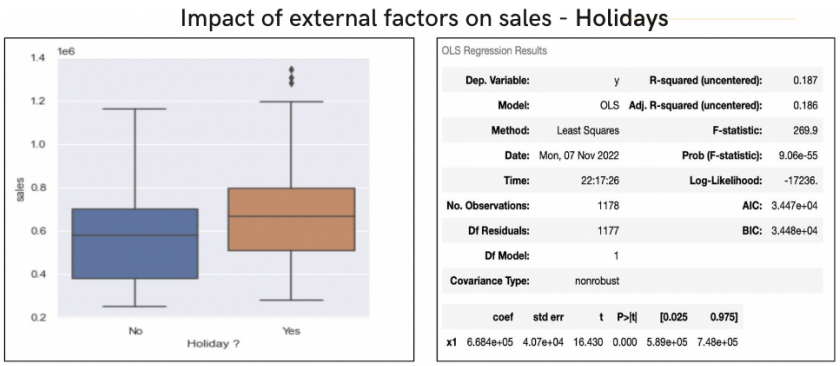


Figure 4

This information can be used to optimize inventory levels, pricing strategies, and marketing campaigns to drive sales growth and improve overall performance. By analyzing these factors, Favorita can gain a better understanding of their business and identify opportunities for growth and optimization.

Additionally, Time series decomposition can be used to break down a time series into its underlying components, including trend, seasonality, and residual. By decomposing the time series, we can identify the different patterns and trends that are driving the data and gain insights into the underlying factors that are impacting sales.

In the context of the above data, time series decomposition can be used to identify trends and seasonality patterns in overall sales, store-level sales, and sales by product type and family. By analyzing these patterns, we can identify areas for improvement and make data-driven decisions to optimize sales growth.

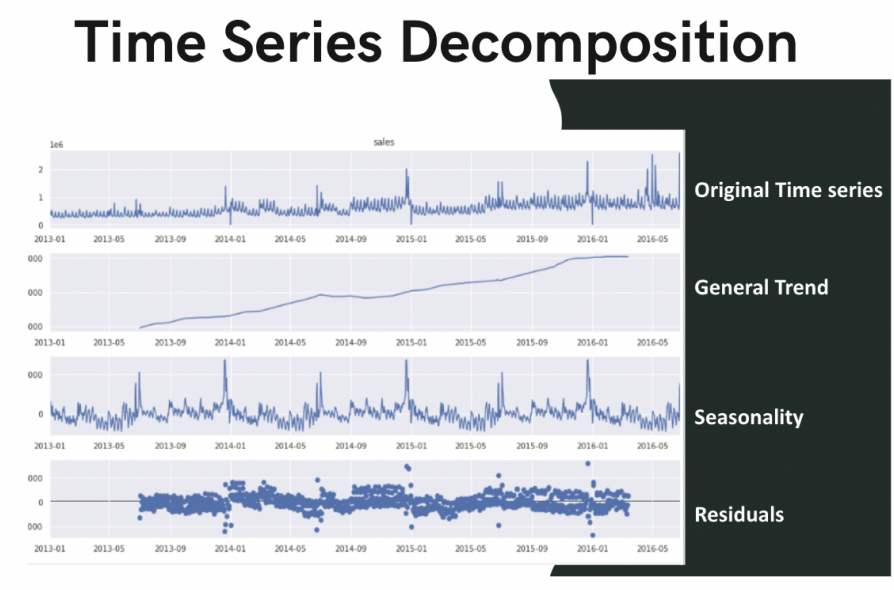


Figure 5

For example, if the decomposition reveals a strong seasonal pattern in sales for a particular product type or family, the company can adjust its inventory levels and marketing campaigns to capitalize on this trend. Alternatively, if the decomposition reveals a declining trend in overall sales, the company can investigate the underlying factors that are driving this trend and take corrective action.

4. Methodology

Below are the methods we considered for our problem statement :

- **Rolling exponential smoothing:** This method applies a simple exponential smoothing model to a rolling window of historical data to make short-term predictions. This approach allows for changes in the data pattern over time to be captured and used in the forecasting model.
- **Holt's (double exponential smoothing):** This method extends simple exponential smoothing to capture trends in the data. It uses two smoothing parameters, one for the level of the series and another for the trend component. This model can be useful when the data exhibits a trend that is not seasonal.
- **Holt's (double exponential smoothing) - rolling:** This is a variant of Holt's model that applies a rolling window to the historical data. By using a rolling window, the model can capture changes in the trend over time and make short-term forecasts.
- **Holt's Winter (triple exponential smoothing):** This method extends Holt's model to include seasonal components in the data. It uses three smoothing parameters,

one for the level of the series, one for the trend component, and another for the seasonal component. This model is useful when the data exhibits both trend and seasonal patterns.

- **Holt's Winter (triple exponential smoothing) - rolling:** This is a variant of the Holt's Winter model that applies a rolling window to the historical data. By using a rolling window, the model can capture changes in the trend and seasonal patterns over time and make short-term forecasts.
- **Propagate trend and seasonality:** This method uses a linear regression model to estimate the trend and seasonal components of the time series. The estimated components are then used to make forecasts. This model can be useful when the data exhibits a strong trend and seasonal patterns.
- **Propagate trend and seasonality and adjust residue:** This method extends the previous model to also include an adjustment for the residual component of the time series. This adjustment can help improve the accuracy of the forecasts by accounting for any unusual or unexpected changes in the data that are not captured by the trend and seasonal components.

5. Results

Method	MAPE
Simple Exp smoothing	37%
Rolling exponential smoothing	37%
Holts (double exp smoothing)	46.14%
Holts (double exp smoothing) - rolling	35.17%
Holts Winter (triple exp smoothing)	48.45%
Holts Winter (triple exp smoothing) - rolling	45.99%
Propagate trend and seasonality	35.09%
Propagate trend and seasonality and adjust residue	35.08%

We calculate MAPE for all of our 8 methods and obtain the results presented in the above table. We see that a simplistic method such a Simple or rolling exponential smoothing yields a lower MAPE than double or triple exponential smoothing with or without rolling. Only Holts double exponential smoothing with rolling time period yields better results. The best(lowest) MAPE is obtained by propagating the trend and seasonality and adjusting for residues.

6. Recommendation/Next Steps

From the results we see that by propagating trend and seasonality and adjusting for residue, we get the lowest MAPE. But the improvement is not much when you compare it with a simple model such as exponential smoothing. In a real world scenario, there would be a tradeoff between a simple model such as exponential smoothing and choosing to propagate trend and seasonality which gives an improvement of only 2% in MAPE, on the cost of having lesser explanability and ease of understanding.

Overall, we have found that using the forecast data can greatly benefit Favorita stores in planning their inventory for the upcoming months. By analyzing the historical sales data, we can identify the patterns and trends in customer demand, and use that information to make informed decisions about what products to stock and in what quantities.

Another important factor that we need to consider is the fluctuation of oil prices, as they have a significant impact on sales. Therefore, keeping track of oil prices and incorporating this information into our forecasting models can help us better predict the demand for certain products and adjust our inventory accordingly.

Furthermore, by including adjustments for holidays and promotions in our models, we can accurately anticipate the surge in sales during these times and plan our inventory accordingly. This can help us avoid stockouts or overstocking, both of which can result in lost sales and revenue.

Finally, at a store level, it is important for Favorita to focus on the stores that are performing well and have the most potential for growth. Based on our analysis, stores 3, 44, 45, 47, and 49 have shown promising sales trends and should be the main focus for inventory planning and optimization efforts. By using the insights gained from our analysis, Favorita can make data-driven decisions to optimize their inventory management and ultimately drive growth and profitability.

For our next steps, we can perform forecasting on individual product-family group and on a store level using PySpark. In a real world scenario, we can do this by setting up a pipeline that automatically updates the models and forecasts daily as new data becomes available, ensuring that the inventory planning is always based on the most up-to-date information. We would also need to implement a monitoring system to alert us of any anomalies or issues with the forecasts so that we can quickly address them.