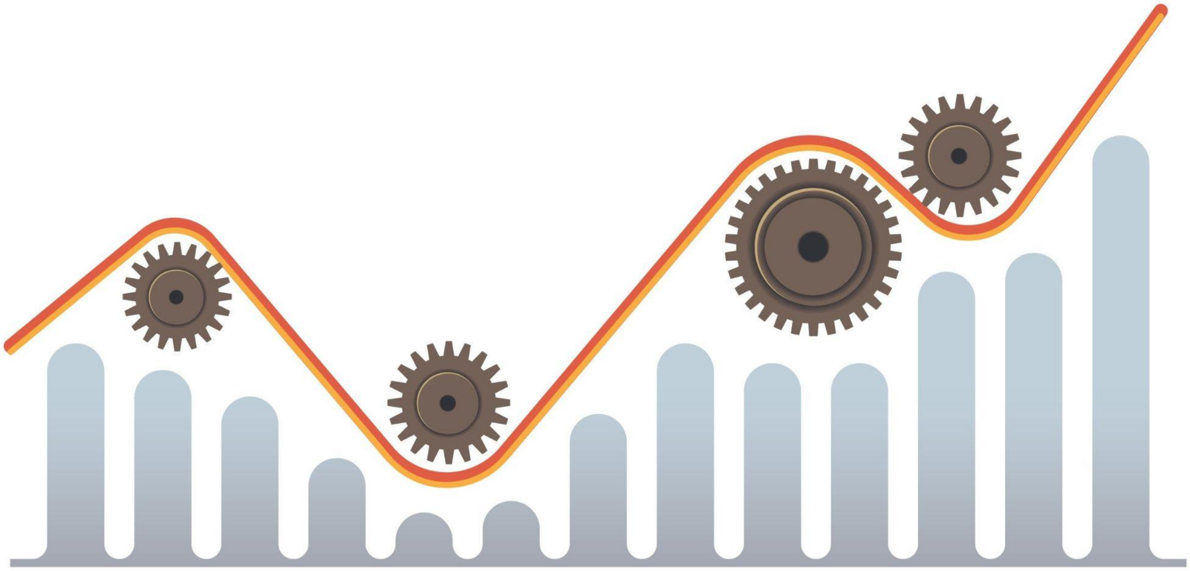
ADIDAS SALES FORECAST MODELING PROJECT REPORT



Ying Li

08/11/2023

TABLE OF CONTENTS

INTRODUCTION

Adidas, a renowned global sportswear brand, entered the US market in 1978, marking its presence in the highly competitive US athletic footwear and apparel industry. Over time, Adidas has established itself as a significant player, securing a considerable market share through its wide array of products. However, challenges like intense competition from rivals such as Nike and Puma, potential brand perception issues, evolving consumer preferences, economic factors, the sway of online retail and e-commerce, and complexity of the global supply chain persist. To tackle these challenges, Adidas needs to stay nimble and innovative; focus on brand differentiation; grasp consumer insights; and enhance online and retail experiences to remain competitive in the US market.

Adidas wants to utilize its historical US sales data to extract valuable insights into the company's performance, market dynamics, and customer behavior. These insights will serve as the foundation for Adidas to drive informed decision-making, optimize short and long term strategies, and ultimately achieve its sales and business objectives.

The objective of this project is to utilize data science tools and techniques to evaluate and identify any noticeable trends and patterns in customer behavior and product preferences, evaluate different segmentation strategies, and construct a robust model capable of accurately forecasting future sales for Adidas products. The expected deliverables include comprehensive Python codes in jupyter notebooks, a final project report and a set of presentation slides.

DATA

The selected dataset for this project is Adidas 2020-2021 US sales data, available at https://data.world/stellabigail/adidas-us-sales-datasets. This dataset, in Excel format, and with a size of 682.28 KB, comprises 9,648 rows and 13 columns. It contains key fields including

- Retailer – name of the retailer
- Retailer ID – unique identification number of the retailer
- Invoice Date – date of the invoice to the retailer
- Region – region the retailer is located in
- State – state the retailer is located in
- City – city the retailer is located in
- Product – product sold
- Price per Unit – unit price of the product
- Units Sold – total number of units sold
- Total Sales – total sales in US dollars
- Operating Profit – operating profit , calculated by subtracting the operating cost from the total sales
- Operating Margin – operating margin, calculated by dividing operating profit by total sales
- Sales Method – sales method, i.e., in-store, outlet and online

Covering Adidas US sales data from 01/01/2020 to 12/31/2021, this dataset provides a comprehensive foundation for analysis.

DATA WRANGLING

Adidas Sales Forecast - Data Wrangling notebook

The dataset is relatively straightforward and doesn’t contain any missing data. However, it includes unnecessary rows and columns due to the Excel format and spreadsheet title. Once these unrelevant elements were removed during the cleaning process, my main focus was on addressing the subsequent data wrangling challenges.

Data Format

Initially, all 13 columns had a data type of “object”. To ensure accurate and consistent data representation, I made specific changes to the data type of certain columns as follows:

- Invoice Date : datetime
- Price per Unit : float
- Units Sold : Integer
- Total Sales : float
- Operating Profit : float
- Operating Margin: float

Data Integrity

1. All numeric columns were evaluated. No obvious outliers were observed.
2. four rows were identified with 0 units sold, 0 total sales, and 0 operating profit, yet displaying a non-zero operating margin. Further investigation revealed that these entries all pertain to Foot Locker, with a retailer ID of 1185732 in Omaha, Nebraska. Two entries were dated 2021-06-05 and the other two 2021-06-11. Considering that the main focus of our analysis centers around total sales and operating profit, and knowing that the inconsistent non-zero operating margin values would not adversely impact the results, I opted to retain these four rows.
3. To validate the accuracy of the values in the "Total Sales" column, I

calculated the total sales for each entry using the formula *Price per Unit x Units Sold = Total Sales*, and compared the calculated values with the original numbers in the dataset. A total of 3886 rows had a mismatch between the calculated and original Total Sales values. In all these instances, the calculated Total Sales was one-tenth of the original Total Sales value. It's noteworthy that all the discrepancies occurred solely at the dataset's outset.

In a real business environment, we'll need to acquire additional sales data or order details to conduct further analysis. Unit Price, Units Sold, or Total Sales values might have been inaccurately input. However, without supplemental data, drawing conclusions on which column might be erroneous is challenging. Acting on unfounded assumptions at this juncture could potentially have adverse implications for our later forecasting or segmentation analysis.

As a result, I proceeded with the assumption that all three original data columns were accurately recorded, and the disparity between calculated and original Total Sales values was caused by factors beyond my understanding.

4. Additionally, I validated the accuracy of the "Operating Margin" values by calculating the operating margin for each entry using the formula *Operating Profit / Total Sales = Operating Margin*, and compared the calculated values with the original numbers in the dataset. This time, I found a total of 848 rows with a mismatch. On closer inspection, the mismatch was notably minor, likely arising from rounding discrepancies and slight decimal variations. Considering the negligible nature of these discrepancies, I deemed it reasonable to disregard these minor inconsistencies and keep the original operating margin values from the dataset.

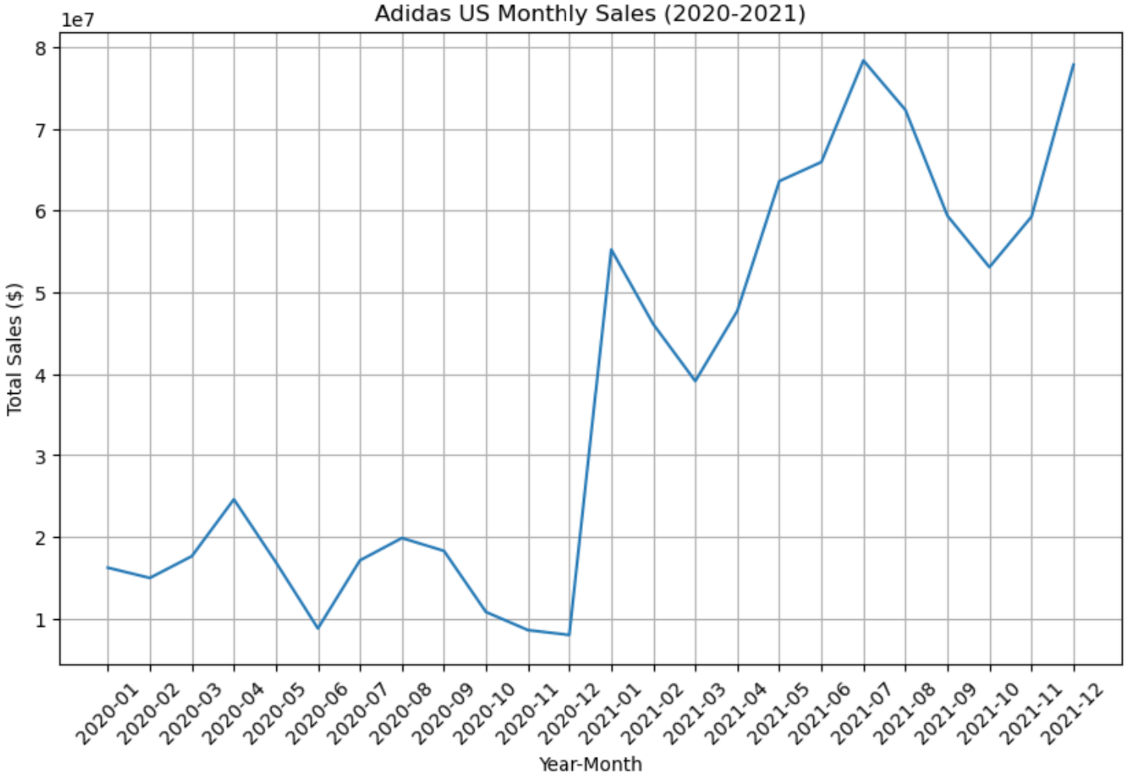
EDA

Adidas Sales Forecast - EDA notebook

Sales Trend

To get an overview of the sales trend, I graphed Adidas daily sales data on a timeline. The chart shows significantly higher and larger sales in 2021 compared to 2020. To enhance visual clarity, I plotted an additional timeline chart (refer to Figure 1) using aggregated monthly sales volume. It's much clearer that Adidas US sales have been consistently increasing from 2020 to 2021, suggesting a positive upward trend.

Figure 1: Adidas 2020-2021 US Sales Trend by Month

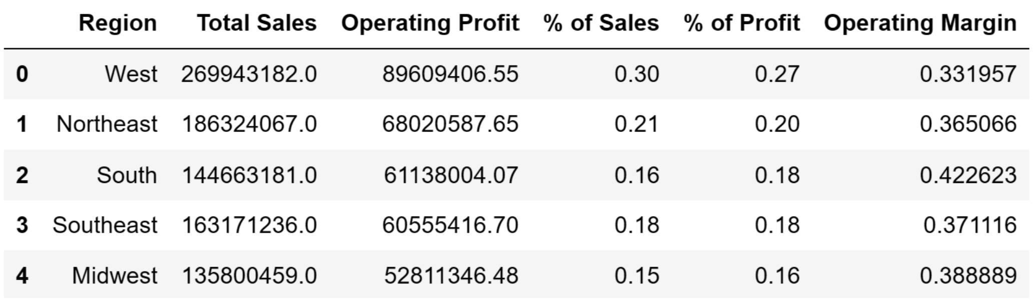


Regional Analysis

Adidas groups its retailers into five regions: Midwest, Northeast, South, Southeast and West.

Among the five regions, the West Region claims the highest sales volume, contributing to 30% of Adidas' total US sales. Following closely, the Northeast Region captures 20.7%. An intriguing finding surfaced when observing the operating margins – the West Region, despite its substantial sales, holds the lowest operating margin, and the Northeast Region follows as the second lowest. Although more data was needed for comprehensive analysis, this observation sparked my curiosity regarding whether elevated sales volumes might come at the expense of profit margins.

Table 1: Adidas 2020-2021 US Sales by Region



	Region	Total Sales	Operating Profit	% of Sales	% of Profit	Operating Margin
0	West	269943182.0	89609406.55	0.30	0.27	0.331957
1	Northeast	186324067.0	68020587.65	0.21	0.20	0.365066
2	South	144663181.0	61138004.07	0.16	0.18	0.422623
3	Southeast	163171236.0	60555416.70	0.18	0.18	0.371116
4	Midwest	135800459.0	52811346.48	0.15	0.16	0.388889

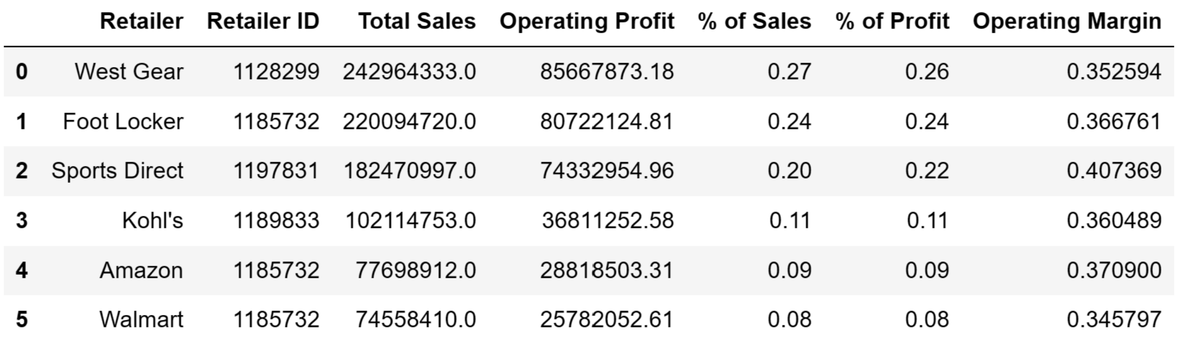
I then examined Adidas' US sales distribution by state. New York, California, Florida, Texas, and South Carolina emerge as the top five states in both sales volume and operating profit. Interestingly, Washington, the sixth top state in sales, exhibits a notably low operating profit of approximately $7M. Similarly, Hawaii records $22.3M in sales but only $5.8M in operating profit. Conversely, some states demonstrate modest sales volumes but yield much more favorable profits – for instance, Alabama with $17.6M sales and $9.1M operating profit, and Tennessee with $18.1M sales and $8.5M operating profit. This observation intensified my curiosity about the trade-off between sales volumes and profit margins.

Retailer Segmentation

While conducting sales segmentation by retailer, I encountered a data integrity issue that hadn't been identified during the data wrangling phase. Multiple retailers had been assigned the same retailer, and conversely, the same retailer had been assigned different IDs. While I could understand the same retailer being assigned different IDs due to different geographic store locations, I found it perplexing why different retailers will be assigned the same ID. Without the comprehensive business context to clarify this, I made the assumption that a data entry error in the Retailer ID column might be the reason. With this assumption, I proceeded with the Retailer name for segmentation and subsequent analysis in this project.

Adidas partners with six retailers: Amazon, Foot Locker, Kohl's, Sports Direct, Walmart, and West Gear. Notably, West Gear emerges as the leading retailer in terms of both sales volume and operating profit. West Gear alone contributes to 27% of Adidas' total sales and 26% of its overall operating profit. Foot Locker secures the second spot, making up 24% of both total sales and operating profit. Although Sports Direct claims the third position in total sales, it boasts the highest operating margin among the six retailers, standing at approximately 41%. Conversely, Walmart ranks as the smallest retailer, accounting for merely 8% of Adidas' sales and profit, accompanied by the lowest profit margin of 34%. This retail segmentation provides Adidas with valuable insights for tailoring distinct business strategies for each retailer.

Table 2: Adidas 2020-2021 US Sales by Retailer



	Retailer	Retailer ID	Total Sales	Operating Profit	% of Sales	% of Profit	Operating Margin
0	West Gear	1128299	242964333.0	85667873.18	0.27	0.26	0.352594
1	Foot Locker	1185732	220094720.0	80722124.81	0.24	0.24	0.366761
2	Sports Direct	1197831	182470997.0	74332954.96	0.20	0.22	0.407369
3	Kohl's	1189833	102114753.0	36811252.58	0.11	0.11	0.360489
4	Amazon	1185732	77698912.0	28818503.31	0.09	0.09	0.370900
5	Walmart	1185732	74558410.0	25782052.61	0.08	0.08	0.345797

Product Segmentation

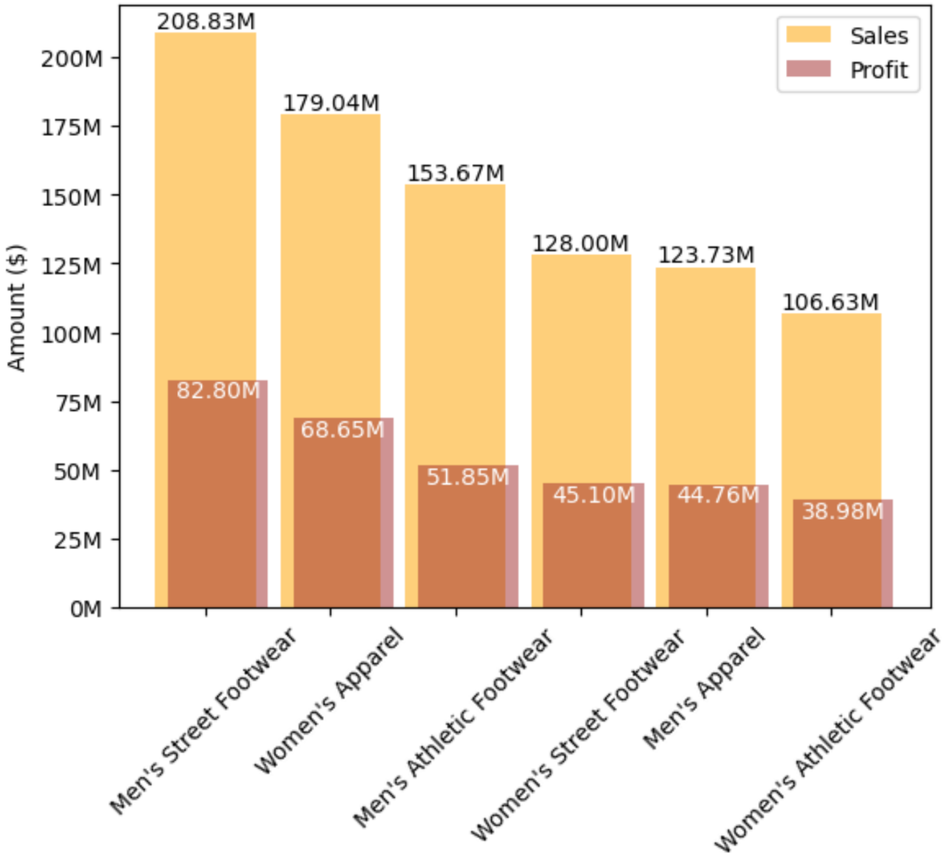
Adidas categorizes its products into six groups:

- Men’s Apparel
- Men’s Athletic Footwear
- Men’s Street Footwear

- Women’s Apparel
- Women’s Athletic Footwear
- Women’s Street Footwear

Out of the six product groups, Men's Street Footwear stands out as the best selling product line, generating over $208M dollar in sales volume and $83M dollar operating profit. Women's Apparel takes the second place, with $179M dollar in sales and $69M dollar in operating profit. Interestingly, these two top selling products also boast the highest operating margin of 40% and 38% respectively. On the other hand, Women's Athletic Footwear emerges as the least successful product line, with only $107M dollar in sales and $39M dollar in operating profit. Nonetheless, even in this lower-ranking position, Women's Athletic Footwear maintains a respectable operating margin of 36.6%, contributing 12% to Adidas' total sales and profitability.

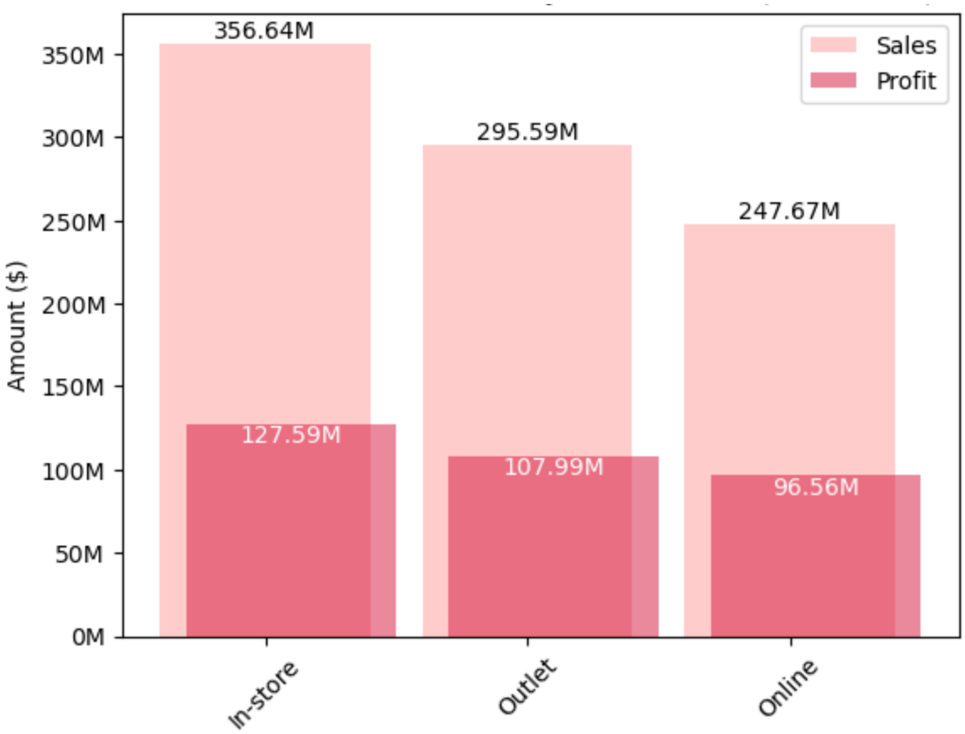
Figure 2: Adidas 2020-2021 US Sales & Profit by Product



Sales Method Analysis

Adidas currently has three established sales channels: in-store, outlet and online. While the in-store channel claims the highest share of Adidas merchandise sales, constituting 40% of the company's total revenue, it has the lowest operating margin at 26%. In contrast, although the online channel accounts for the smallest portion (i.e. 28% of total sales), its profitability is the highest, with an operating margin of 39%. This insight provides Adidas with crucial guidance to refine its sales channel strategy moving forward.

Figure 3: Adidas 2020-2021 US Sales & Profit by Sales Method



MODELING

Adidas Sales Forecast - Modeling notebook

Before progressing to the modeling phase, I conducted thorough data preprocessing. The initial timeline chart shows distinct variations in sales levels and patterns between 2020 and 2021. Given the fact that 2020 marked the onset of the Covid pandemic and the subsequent lockdowns mandated by the US government, the drastic shift between the two years is not surprising. Due to the significant influence of Covid and the lockdown measures, the sales data from 2020 doesn't provide a reliable foundation for sales forecasting. Including 2020 sales data in our timeline might introduce bias and potentially lead to an underestimation in our forecasts, given that Covid's impact has diminished over time. Therefore, I made the decision to exclude the 2020 data due to its inherent bias and chose to utilize only the 2021 data for constructing, training and testing the forecasting model.

The original 2021 data was not stationary, evident from the fluctuating mean and variance in the timeline chart. To address this, I applied a logarithmic transformation to stabilize the variance. Following a single application of the log method, the adfuller results confirmed the achievement of stationarity. Subsequently, I used a 70/30 ratio to determine the split point and separated the data into the training and testing sets.

To capture the dynamic patterns exhibited in the sales data, I chose to test out three widely used time series models:

1. ARIMA (AutoRegressive Integrated Moving Average)
2. SARIMA (Seasonal AutoRegressive Integrated Moving Average)
3. Facebook Prophet

By evaluating and comparing their performance metrics, I was able to determine the most robust and suitable model for our purpose.

ARIMA

For the ARIMA model, I began by plotting and evaluating the ACF (Autocorrelation Function) and the PACF (Partial Autocorrelation Function) charts. Neither chart exhibited a clear tail, posing challenges in determining suitable orders (p and q) for the model. I established a baseline model with p=0 and q=0, and then conducted a grid search to identify the best model using the lowest Mean Absolute Percentage Error (MAPE) value..

Based on the grid search result, the optimal ARIMA order for the best-performing model with the lowest MAPE is (8, 0, 6). I used this order to build, train and test the ARIMA model. Model diagnostics, including Standard residual for "T" chart, normal Q-Q chart, and Correlogram chart, were reasonably satisfactory. Based upon the observation, I concluded that the model was sufficiently reliable for forecasting purposes.

Then, I calculated performance metrics for the model, including MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAPE. Furthermore, I performed a 5-fold cross validation, affirming the stability of the ARIMA model based on the results obtained.

SARIMA

SARIMA, similar to ARIMA, incorporates an additional seasonal component. To enhance the visibility of data seasonality, I decomposed the data and plotted corresponding charts. The seasonal chart clearly shows a consistent monthly pattern inherent within the sales data. This was further confirmed by the ACF chart (refer to Figure 5).

Figure 4: Adidas 2021 US Sales Seasonal Chart after Decomposition

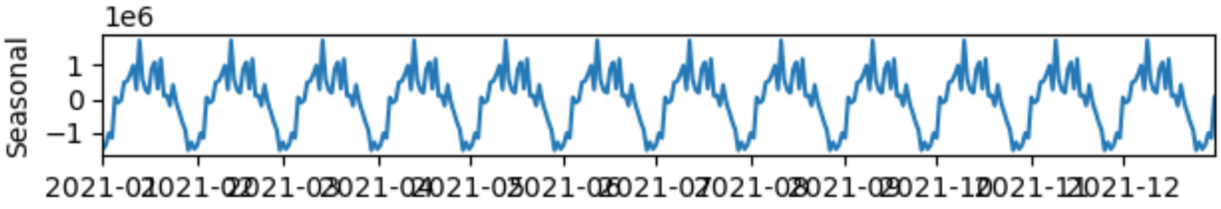
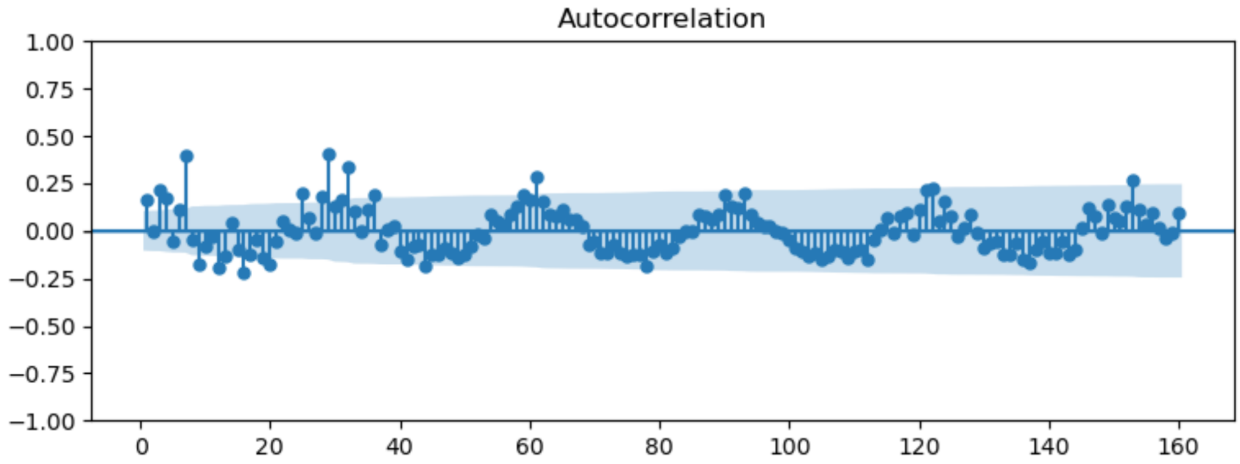


Figure 5: ACF Chart on Adidas 2021 US Sales Data



Then, I used the "auto_arima" function to find out the optimal hyperparameters that would yield the best model based upon the MAPE value. Consequently, I built the SARIMA model using the auto_arima outcome, conducted training and testing of the model, calculated performance metrics, and performed a cross-validation to confirm the stability of the SARIMA model.

PROPHET

Prophet is an open-source forecasting tool developed by Facebook. It provides a user-friendly platform for time series forecasting. To meet the specific requirements of Prophet, I adjusted the dataset, renaming the time column to "ds" and the value column to "y" before splitting it into the training and testing sets.

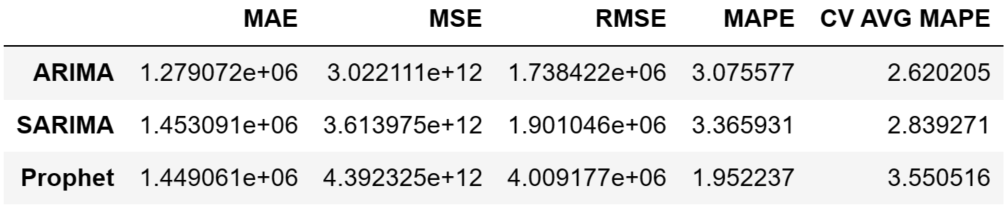
Subsequently, I replicated the same approach used for the ARIMA model to:

- build a baseline model for training and testing;
- identify the most favorable parameters for the optimal model as determined by the MAPE score;
- build and fit the best Prophet model based upon the grid search outcome;
- evaluate the performance metrics of the preferred Prophet model; and
- conduct a cross-validation to affirm the stability of the model.

Model Evaluation & Comparison

Initially, my plan was to select the model with the lowest MAPE score, which would lead to choosing the Prophet model with a MAPE score of 1.95%. However, despite the lowest MAPE score, the Prophet model has the highest MSE and second highest MAE scores among three models. Additionally, the Prophet model has the highest average MAPE score from cross-validation with a much bigger variance in cross-validation MAPE scores compared to the ARIMA and SARIMA models. In light of these findings, I decided that the Prophet model is not the optimal choice. Between the ARIMA and SARIMA models, the ARIMA model has consistently better performance scores compared to SARIMA. Therefore, I came to the final conclusion that the ARIMA model is the most suitable choice for our forecasting purposes.

Table 3: Performance Metrics of Three Models



	MAE	MSE	RMSE	MAPE	CV AVG MAPE
ARIMA	1.279072e+06	3.022111e+12	1.738422e+06	3.075577	2.620205
SARIMA	1.453091e+06	3.613975e+12	1.901046e+06	3.365931	2.839271
Prophet	1.449061e+06	4.392325e+12	4.009177e+06	1.952237	3.550516

CONCLUSION

After evaluating and comparing model metrics, I have determined that the ARIMA model performs the best. The optimal hyperparameters for the ARIMA model are as follows:

p = 8

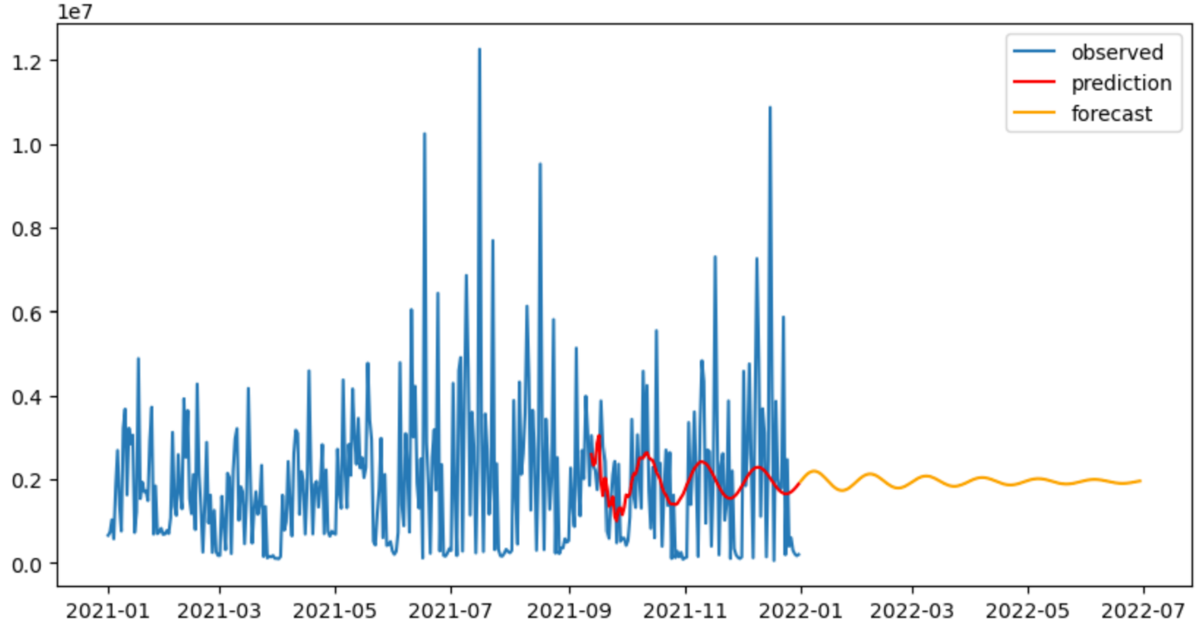
d = 0

q = 6

The ARIMA model stands out with the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) scores, along with a relatively low Mean Absolute Percentage Error (MAPE) score, as presented in Table 3.

Utilizing the ARIMA model, I forecasted future sales and visualized the results in a chart as illustrated in Figure 6.

Figure 6: Adidas 2021 US Sales Data with Prediction on Test Data & Future Forecast for 6 Months



FUTURE IMPROVEMENTS

To improve the accuracy and reliability of the model, we'll need more data.

1. The dataset used in this study covers the years of 2020 through 2021, which is outdated and particularly influenced by the significant impact of the Covid pandemic and subsequent lockdowns in 2020. Acquiring more recent data for the years of 2022 and 2023 year-to-date would be very valuable in providing more insight into Adidas' post-pandemic sales trend, customer behaviors and seasonality patterns. The inclusion of such up-to-date information has the potential to considerably enhance the forecasting accuracy of the model and may even lead to a re-evaluation of our choice of the optimal model, if the newly acquired data unveil previously unseen seasonal patterns that were not observed in the 2021 dataset.
2. Incorporating historical data from years prior to 2020 is equally important. This supplementary information will offer deeper insights into Adidas' segmentation, year-round seasonality, and evolving sales patterns. Such understanding will play a vital role in formulating effective sales and segmentation strategies for Adidas moving forward.
3. More detailed and multidimensional data is also necessary. For instance, data that further breaks down products could offer additional insights into the sales trend of established product lines versus new innovative ones. Gaining this knowledge and identifying trends will enable Adidas to make informed decisions regarding innovation, new product introductions, and the phasing out of obsolete products. Furthermore, the enriched data will enhance the overall accuracy and effectiveness of the forecasting model and segmentation analysis.
4. We would also like to incorporate data concerning external factors that may impact sales, such as economic indicators, consumer sentiment, sales promotions and competitor activities, etc. Integrating these external variables into the forecasting model will greatly enhance its predictive power by accounting for the influence of broader market dynamics on

sales.

We could consider developing and implementing an automated process to continually update the forecasting model when new data becomes available, ensuring the model's accuracy and relevance.

Furthermore, we could explore an interactive dashboard that allows users to explore and visualize sales data and forecasting results across various dimensions like time, product, retailer and/or sales channel. This will greatly enhance the usability and accessibility of insights for different stakeholders.

By delving into these areas, we have the potential to improve the outcomes of this project, achieving better accuracy, reliability, and effectiveness in Adidas' sales forecasting. This enhanced insight could enable Adidas to optimize supply chain planning, resource allocation , and ultimately achieve better sales outcomes and consequently improve its bottom line.