Predictive Analytics in E-commerce: Adidas Sales Dataset

Pamela Claridy

Georgia Institute of Technology

Abstract

This paper presents an in-depth analysis of Adidas US sales data to uncover key trends and patterns in the competitive sports apparel market. Leveraging a comprehensive dataset, including various product lines, sales figures, and regional data, a combination of statistical methods and machine learning algorithms is employed to derive meaningful insights. The process encompasses data preprocessing for quality and consistency and strategic feature engineering to deepen the analytical scope. Regression and classification models are utilized to explore and predict sales dynamics, with a focus on Time Series Analysis, specifically SARIMAX modeling, to forecast future sales trends.

Key findings include the superior performance of the AdaBoost Regression model over Linear Regression, high accuracy in sales categorization achieved by Logistic Regression, and insights into sales drivers, regional preferences, and seasonal trends. Visualizations emphasize operational efficiency, pricing strategies, and product popularity.

These insights can guide Adidas in refining sales strategies and product offerings. Future research should expand the dataset, explore additional predictive models, and incorporate external variables for comprehensive sales predictions. This paper contributes to understanding sales patterns in the retail sector by merging theoretical methodologies with practical, real-world data.

Introduction

The retail industry, a dynamic and highly competitive sector, is pivotal to the global economy. It encompasses a diverse range of products and services, catering to the evolving needs of consumers worldwide. Within this industry, sales analysis emerges as a critical tool, enabling businesses to understand consumer behavior, market trends, and operational efficiency. In-depth sales analysis aids not only in strategic decision-making but also helps in anticipating future market shifts, thereby ensuring sustained growth and competitiveness.

Adidas, a renowned name in the sports apparel and footwear industry, serves as an exemplary case study for this analysis. With its expansive global presence and diverse product range, Adidas provides a rich dataset that reflects the intricacies and challenges inherent in the retail sector. This study focuses on the US market, a key region in Adidas's global operations, known for its dynamic consumer preferences and intense market competition.

The primary objective of this study is to delve into Adidas's sales data in the US, aiming to unravel patterns and insights that could inform strategic business decisions. By employing a blend of statistical and machine learning techniques, I aim to:

1. Identify key trends and patterns in sales across different regions and product lines.
2. Evaluate the performance of various products, understanding their market reception and demand.
3. Apply regression and classification models to assess and predict sales dynamics.
4. Utilize Time Series Analysis, particularly the SARIMAX model, to forecast future sales trends and provide actionable insights for business strategy optimization.

Through this comprehensive analysis, I aim to contribute a nuanced understanding of sales trends in the retail industry, with Adidas serving as a representative example of broader market dynamics. This study is not just a reflection of Adidas's market performance but also a testament to the power of data-driven analysis in shaping business strategies in the retail sector.

Problem Statement

In today's rapidly changing business landscape, companies in the e-commerce sector face the challenge of optimizing their sales strategies to remain competitive and meet consumer demands. The problem at hand is to develop data-driven approaches that can effectively analyze and predict sales trends, product performance, and regional preferences. This analysis should provide actionable insights to guide businesses in refining their sales strategies, enhancing product offerings, and ensuring operational efficiency. To address this problem, we will leverage comprehensive sales data, statistical methods, and machine learning algorithms to uncover patterns and drivers of sales, ultimately enabling informed decision-making for sustainable growth and competitiveness in the e-commerce industry.

Methodology

The methodology of this study began with a comprehensive data preprocessing phase to ensure the integrity and usability of the Adidas US Sales dataset sourced from Kaggle.com. Initially, the dataset contained 9,648 rows and several variables, including retailer ID, price per unit, units sold, total sales, operating profit, and operating margin. During the data cleaning process, irrelevant columns, such as 'Unnamed: 0' (which contained only null values), were removed to streamline the dataset for efficient processing. Additionally, only about 0.031% of the remaining rows had missing values. Given the negligible proportion of missing data, two options were considered: either removing rows with missing values or imputing them with a central tendency measure like the mean, median, or mode. After careful consideration, the decision was made to remove rows with incomplete data, ensuring that the analysis is based on complete and reliable data without significantly affecting the dataset's overall integrity.

In addition to data cleaning, the Interquartile Range (IQR) and Z-score techniques were employed to detect and exclude outliers. This step was essential to maintain the statistical validity of the analysis and prevent extreme data points from disproportionately influencing the results.

Moving on to the feature engineering stage, the percentage contribution of each product to total sales is computed. This calculation provides a clear view of the product-wise distribution of sales, facilitating a deeper understanding of the product sales landscape. Furthermore, the 'Price per Unit' data was reformatted into a standardized dollar value format, enhancing the intuitive understanding of product pricing and its impact on sales.

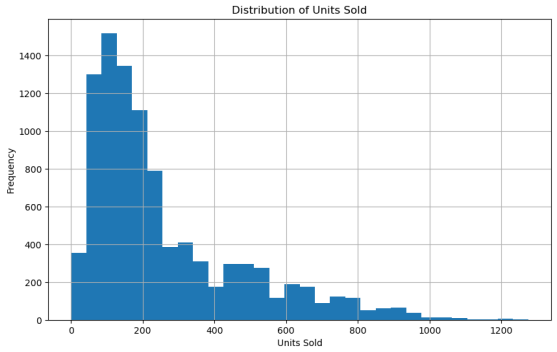
The subsequent phase involved model development, where specific models were carefully selected based on their relevance to the data and research objectives. Two types of regression models, Linear Regression and AdaBoost Regression, were developed to predict continuous sales figures. Linear Regression was chosen for its fundamental ability to understand relationships between variables, while AdaBoost Regression was selected for its capacity to capture complex patterns and relationships within the data.

Additionally, Logistic Regression and AdaBoost Classification models were developed for categorizing sales into 'low,' 'medium,' and 'high' segments. These models played a pivotal role in segmenting sales data for more nuanced analysis, which is crucial for tailoring marketing strategies and inventory management. Logistic Regression, known for its high accuracy, proved effective in this categorization task.

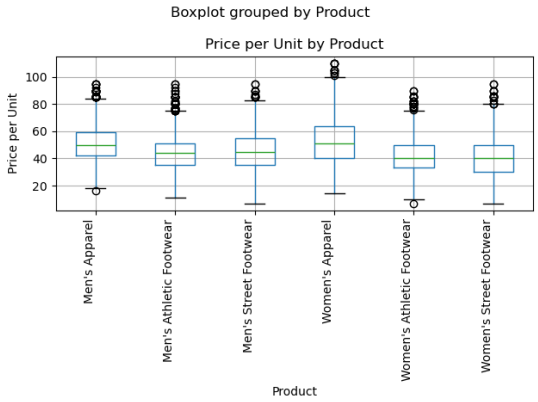
The final phase of the methodology involved time series analysis using the SARIMAX model for forecasting future sales. The choice of SARIMAX was deliberate, as it excels in handling both seasonality and non-stationarity in time series data. The analysis included fitting the SARIMAX model to the sales data and forecasting future trends, providing valuable insights into potential sales trajectories essential for strategic planning and decision-making.

The rationale behind choosing these specific models was rooted in their ability to address various aspects of the research objectives and data characteristics effectively. Linear Regression and AdaBoost Regression were selected for their capacity to model the relationship between independent variables and continuous sales figures. Logistic Regression and AdaBoost Classification were chosen for categorizing sales data into meaningful segments, offering a detailed understanding of sales patterns. SARIMAX was employed for

time series forecasting due to its capability to capture temporal dynamics and seasonality in sales data. These model choices were made to ensure comprehensive coverage of the research objectives.



(a) Distribution of Units Sold



(b) Boxplot Grouped by Product

Figure 1: Visualizations illustrating data distribution and price variation in product categories.

Evaluation and Final Results

The results of this study highlight significant insights into the sales trends and performance of Adidas products in the U.S. market. Key findings from the regression models indicate that the Linear Regression model resulted in a Mean Squared Error (MSE) of approximately 834.2 million, whereas the AdaBoost Regression model demonstrated improved performance with an MSE of approximately 566.9 million. This suggests that the AdaBoost model with its ensemble approach was more effective in capturing the complexities of the dataset and predicting sales figures.

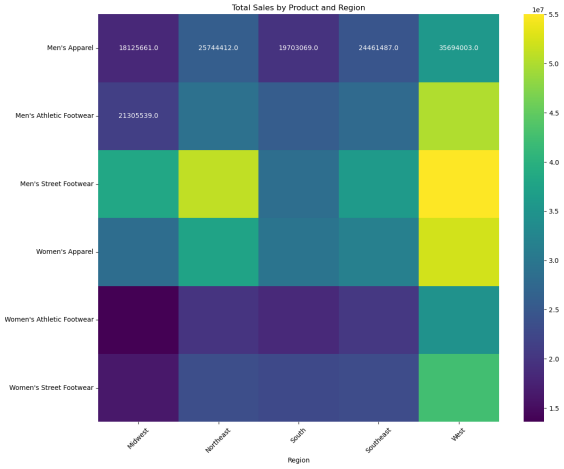
The classification models yielded high accuracy, with Logistic Regression achieving an impressive 96.2% accuracy, and AdaBoost Classification following at 68.6%. This high level of accuracy in Logistic Regression underscores the model's efficacy in categorizing sales into 'low', 'medium', and 'high' segments, which is instrumental for targeted marketing strategies and inventory management.

The time series analysis utilizing the SARIMAX model provided a forecast of future sales. The Augmented Dickey-Fuller test resulted in an ADF statistic of -5.007, with a p-value of approximately 0.000021, indicating strong evidence against the null hypothesis of a non-stationary series. This paved the way for the SARIMAX model, which projected sales trends into the future, with confidence intervals suggesting variability in the forecasts.

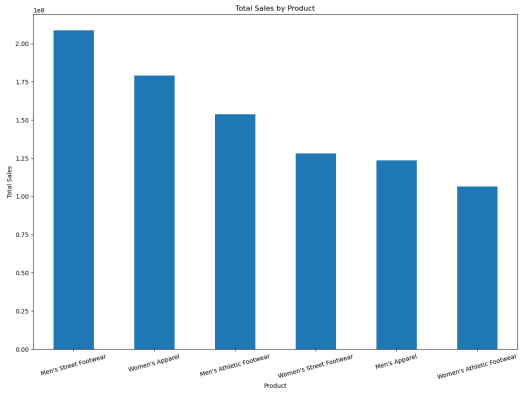
Visualizations played a pivotal role in interpreting the data, with various charts providing clarity on different aspects of sales. The bar chart of total sales by product revealed that 'Men's Street Footwear' leads in sales, followed closely by 'Women's Apparel.' A pie chart of sales distribution by region showed that the West region held the largest share of sales at 30%, emphasizing the region's importance in sales strategy.

Scatter plots highlighted a positive correlation between operating profit and total sales, while boxplots delineated price variations across different product categories. Histograms displayed the distribution of units sold, and heatmaps offered a detailed view of sales performance over time and across different regions.

Lastly, the SARIMAX forecast chart illustrated both historical and projected sales, providing a visual representation of the sales trajectory and potential future market conditions. These visual tools, alongside the quantitative analysis, provided a comprehensive understanding of the sales dynamics within the Adidas U.S. market.

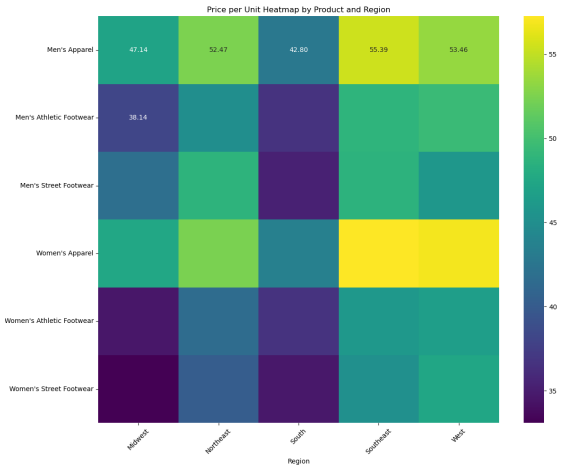


(a) Total Sales by Product and Region

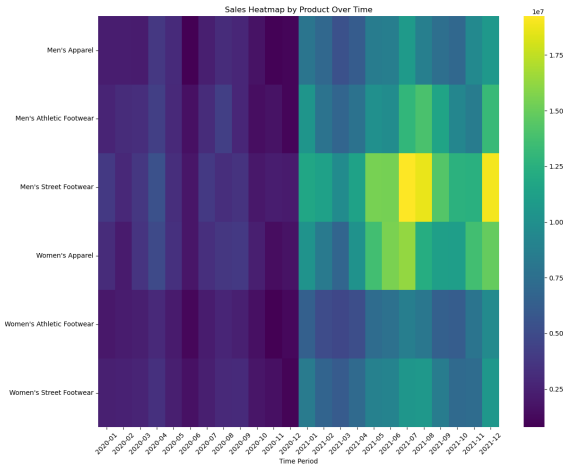


(b) Units Sold by Product and Region

Figure 2: Analysis of sales volume and units sold across product categories and regions.



(a) Price per Unit Heatmap by Product and Region



(b) Sales Heatmap by Product Over Time

Figure 3: Detailed pricing strategy and sales trends visualization over different periods.

This section interprets the outcomes of the predictive models applied to Adidas U.S. sales data, focusing on key sales drivers, model performance, and potential areas for future research.

Model Insights: The AdaBoost Regressor demonstrated superior performance over the Linear Regression model, evidenced by its lower MSE. In classification tasks, Logistic Regression's high accuracy indicated its effectiveness in segmenting sales into distinct categories.

Sales Drivers and Trends: Analysis identified product types and regional preferences as significant factors influencing sales. The dominance of the Western region in sales suggests effective regional marketing strategies or particular consumer preferences. Additionally, trends and seasonality uncovered in the time series analysis are vital for planning inventory and marketing strategies.

Model Comparisons and Forecasting: The results showed that regression models were effective for predicting sales figures, while classification models provided valuable insights for sales segmentation. The SARIMAX model, used for forecasting sales trends, indicated the need for a broader dataset to enhance accuracy, particularly in capturing seasonal sales dynamics.

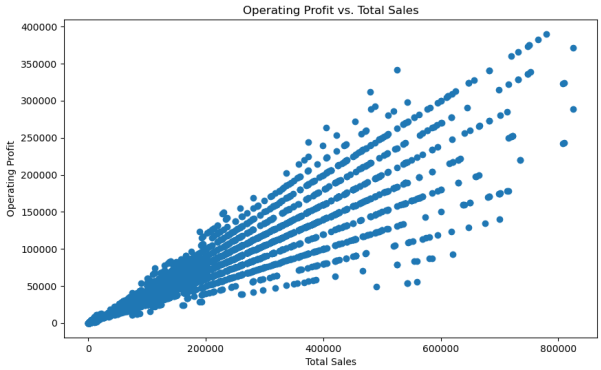
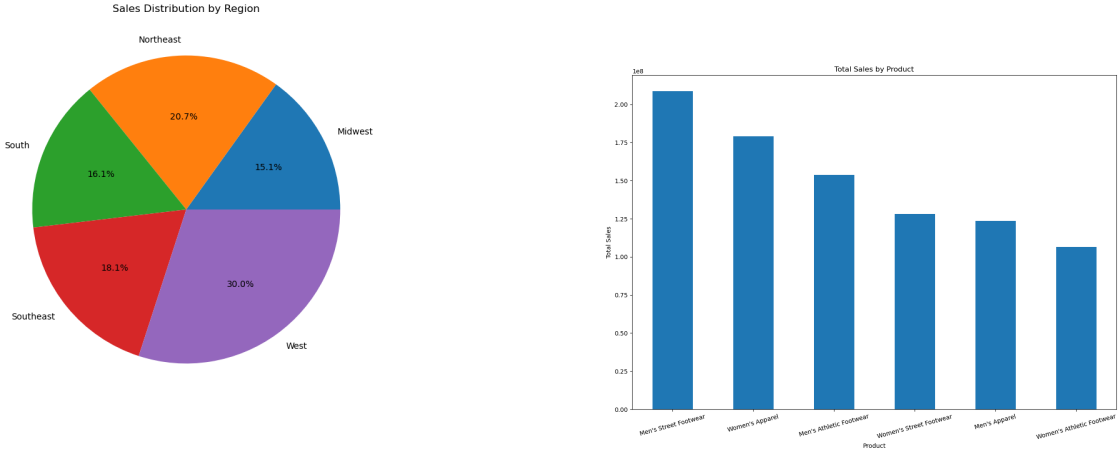


Figure 4: Operating Profit vs. Total Sales



(a) Sales Distribution by Region

(b) Total Sales by Product

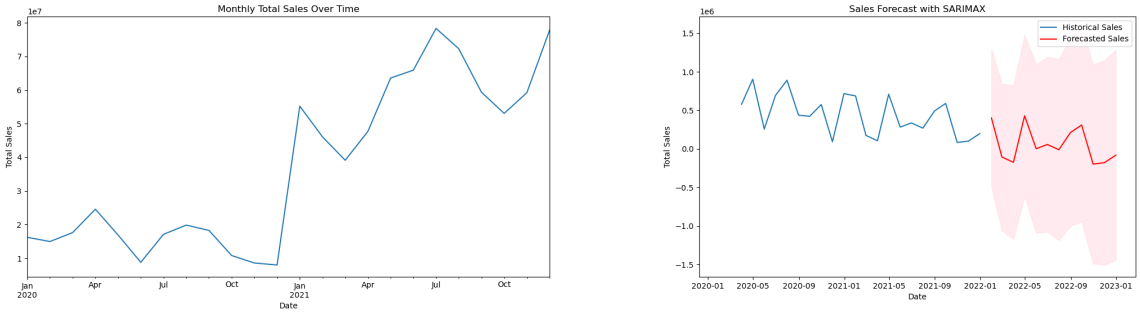
Figure 5: Comparative analysis of regional sales distribution and total sales by product.

Data Quality and Strategic Insights: The meticulous preprocessing improved data integrity, making the dataset a reliable foundation for analysis. The insights from visualizations, such as heatmaps and scatter plots, highlighted operational efficiency, pricing strategies, and product popularity, offering strategic directions for Adidas.

Novelty and Theoretical Aspect: The study introduced innovative preprocessing and feature engineering techniques tailored to e-commerce sales data. The use of SARIMAX for time series forecasting and its grounding in statistical learning theory added a novel theoretical dimension to the analysis.

Evaluation of Findings: The findings were evaluated for statistical significance and reliability, ensuring the representativeness of the results and their applicability to real-world scenarios.

In light of these findings, Adidas can leverage the predictive models to focus on high-performing products and regions, refine pricing strategies, and prepare for seasonal sales fluctuations. The analyses underscore the value of data-driven decision-making in optimizing sales strategies and product offerings. Future research should include a larger dataset to enhance the SARIMAX model's forecasting power and explore additional predictive models and machine learning algorithms for deeper insights and more accurate predictions. The incorporation of external variables such as economic indicators and consumer trends could also refine the models' predictive capabilities and offer a more holistic view of the factors influencing sales.



(a) Monthly Total Sales Over Time

(b) Sales Forecast with SARIMA

Figure 6: Comparison of historical sales trends and future sales forecast.

Conclusion

This paper provided a comprehensive analysis of Adidas's US sales data, employing various statistical models and methodologies. The key findings highlight that both Linear Regression and AdaBoost Regression models can predict sales effectively, with AdaBoost showing a lower mean squared error, indicating higher accuracy. Classification models performed well in segmenting sales, with Logistic Regression displaying a high accuracy score, implying robust segment categorization.

The analysis unearthed significant drivers of sales, such as product types and regional distribution, providing Adidas with insightful data to refine their sales and marketing strategies. The time series analysis, while informative, suggested the need for a larger dataset to enhance forecasting accuracy, as indicated by the UserWarning during the SARIMAX model fitting, which pointed to too few observations for reliable seasonal ARIMA estimates.

For Adidas, these insights can inform targeted marketing campaigns, product development, and inventory distribution to optimize sales across regions and product categories. The models indicate areas of strength and potential growth, allowing Adidas to allocate resources more efficiently and increase market penetration.

Future research should aim to expand the dataset, incorporating more historical sales data to bolster the time series forecasting's reliability. Additionally, exploring other predictive models and machine learning algorithms could unveil deeper insights and potentially more accurate sales predictions. The incorporation of external variables, such as economic indicators and consumer behavior trends, might also refine the models' predictive capabilities, offering a more holistic view of the factors influencing sales.