Product Demand Forecasting for the Sea Island Tennis Shop

Capstone Project - DATA 4200 (Fall 2025)

Gonzalo Tano
B.S. in Data Science (Computational Data Analytics)
College of Coastal Georgia

Faculty Advisor: Renren Zhao

December 2025

Abstract

This capstone project develops a daily product-level demand forecasting system for the Sea Island Tennis Shop, a resort tennis retail store in Georgia, USA. The shop operates in a highly seasonal environment where sales depend on weather, guest traffic, events, and product mix. During two summers working at the shop, I experienced firsthand how difficult it was for managers to make inventory decisions without data-driven forecasts: some days we sold out of key items by noon, while on other days we overstocked products that barely moved.

Using a dataset covering January 2023 to June 2025, I built an end-to-end workflow that includes data cleaning, exploratory data analysis (EDA), feature engineering, predictive modeling, and scenario simulation. The target variable is daily units sold per product. I engineered time-based features (year, month, day of week), weather categories, event flags, lagged sales, and rolling 7-day averages. Three models were compared using a time-based split (train: 2023-2024, test: 2025): Linear Regression, a regularized Random Forest, and XGBoost.

XGBoost achieved the best performance, with an R^2 of approximately 0.76 on the 2025 test set, and lower MAE and RMSE than the other models. Feature importance analysis shows that inventory on hand, price and cost, recent sales behavior, store traffic, and calendar effects are the primary drivers of predicted demand. Scenario simulations illustrate how promotions, tennis events, and warm weather can almost double expected sales relative to a normal day.

The results demonstrate that even a relatively simple machine learning pipeline can significantly improve planning for inventory, staffing, and promotions. The final model and workflow could be integrated into a dashboard to support proactive, data-driven decision-making at the Sea Island Tennis Shop.

Introduction

The Sea Island Tennis Shop is located inside a luxury resort in Georgia and serves a mix of resort guests and club members. Unlike a typical mall store, the shop's demand is highly dynamic and depends on seasonality, weather, guest arrivals, and tennis events. On busy spring and summer weekends, the shop can be full of players buying apparel, rackets, shoes, and accessories; on cold or rainy weekdays, sales slow down significantly.

I worked at the Sea Island Tennis Shop for two summers, which gave me a practical understanding of how the store operates. My manager and I often discussed how difficult it was to make inventory decisions without any forecasting tools. Some days we sold out of specific items early, creating missed sales opportunities. Other days we overstocked products that moved slowly, tying up capital and storage space. Most decisions were based on intuition and past experience rather than on data.

This project grew directly out of those conversations. The goal is not just to experiment with machine learning models, but to build a forecasting system that could realistically support managers in planning inventory, staffing, and promotions. By analyzing historical data and combining it with external factors such as weather and events, the project aims to reduce uncertainty and move the shop toward more proactive, data-driven decision making.

Data Confidentiality Notice

The dataset used in this project is based on real operational data from the Sea Island Tennis Shop.
However, in order to protect internal business information, some numerical values (such as prices, costs, quantities, and traffic counts) were adjusted or anonymized.

These modifications **do not affect the patterns, trends, or conclusions** of the analysis and forecasting models.

Problem Statement and Objectives

The core business problem is that the Sea Island Tennis Shop lacks a systematic way to predict daily demand at the product level. Without reliable forecasts, managers face several challenges:

- Risk of stockouts on high-demand days, especially for high-margin products such as rackets, shoes, and premium apparel.
- Overstocking of slow-moving items, leading to excess inventory and discounted sales.
- Limited ability to plan staffing around busy weekends, holidays, and tennis events.
- Difficulty quantifying the impact of promotions, discounts, and weather conditions.

To address this problem, the project has four main objectives:

1. **Build a clean, analysis-ready dataset** that combines sales, inventory, weather, event, and calendar information for the Sea Island Tennis Shop.
2. **Explore patterns in demand** through EDA, focusing on seasonality, product categories, customer types, and the impact of weather and events.
3. **Develop and compare predictive models** (Linear Regression, Random Forest, and XGBoost) to forecast daily units sold per product using a realistic time-based

evaluation.

4. **Translate the model into business insights and scenarios**, showing how managers can use forecasts to plan inventory, staffing, and promotions more effectively.

The remainder of this report follows a standard data science pipeline. Section 3 describes the dataset and preprocessing steps; Section 4 presents the exploratory analysis; Section 5 explains the feature engineering and modeling approach; Section 6 discusses the results and scenario simulations; and Section 7 summarizes business recommendations and future work.

Data Description

This project uses a realistic synthetic dataset representing daily sales at the Sea Island Tennis Shop from January 2023 to June 2025. Each row corresponds to a product on a given day. The main columns include:

- **date** - Calendar date (daily, 2023-2025).
- **product_id** - Numeric identifier for each product.
- **product_name** - Descriptive product name (e.g., “Pro Staff V14 Racket”).
- **category** - Product category (Rackets, Shoes, Apparel, Balls, Accessories, Beverages).
- **price** - Selling price per unit.
- **cost** - Cost per unit for the shop.
- **on_promo** - Binary flag indicating whether the product is on promotion.
- **discount_pct** - Discount percentage applied.
- **stock_on_hand_end** Inventory on hand at the end of the day.

- **sales_qty** - Units sold during the day (target variable).

- **revenue, cogs, margin** - Financial measures derived from sales.

- **customer_type** - Guest or Member.

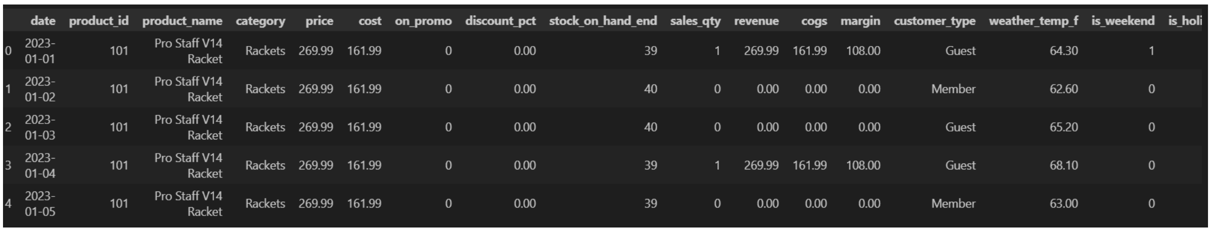
- **weather_temp_f** - Average daily temperature in degrees Fahrenheit.

- **is_weekend, is_holiday** - Calendar flags.

- **event_name** - Name of tennis events (tournaments, camps, etc.) where applicable.

- **store_traffic** - Daily store traffic count.

The dataset contains 14,592 rows and 23 columns, covering multiple product categories and reflecting realistic variation across days, seasons, and events.



	date	product_id	product_name	category	price	cost	on_promo	discount_pct	stock_on_hand_end	sales_qty	revenue	cogs	margin	customer_type	weather_temp_f	is_weekend	is_holi
0	2023-01-01	101	Pro Staff V14 Racket	Rackets	269.99	161.99	0	0.00	39	1	269.99	161.99	108.00	Guest	64.30	1	
1	2023-01-02	101	Pro Staff V14 Racket	Rackets	269.99	161.99	0	0.00	40	0	0.00	0.00	0.00	Member	62.60	0	
2	2023-01-03	101	Pro Staff V14 Racket	Rackets	269.99	161.99	0	0.00	40	0	0.00	0.00	0.00	Guest	65.20	0	
3	2023-01-04	101	Pro Staff V14 Racket	Rackets	269.99	161.99	0	0.00	39	1	269.99	161.99	108.00	Guest	68.10	0	
4	2023-01-05	101	Pro Staff V14 Racket	Rackets	269.99	161.99	0	0.00	39	0	0.00	0.00	0.00	Member	63.00	0	

The table above shows the first few rows of the dataset, illustrating the structure of the data used in this project. Each row represents a single product on a specific day, including key variables such as price, cost, category, inventory, sales quantity, customer type, weather, and calendar indicators. This structure allows the forecasting model to combine internal business variables with external factors such as weather and events.

Data Cleaning and Preprocessing

Before analysis, I performed several basic preprocessing steps in Python:

1. **Standardized column names** by converting them to lowercase and replacing spaces with underscores.

2. **Checked missing values**. All core variables such as sales quantity, price, category, inventory, weather, and traffic had zero missing values. The only column with many missing entries was `event_name`, which is expected because most days do not

have a tennis event.

3. **Verified data types**, converting the `date` column to a proper datetime format and ensuring that numeric variables (price, cost, traffic, etc.) were stored as numeric types.

4. **Checked duplicates** and confirmed that there were no duplicated rows.

5. **Created initial time features**, including year, month, day, and day_of_week extracted from the date.

Overall, the dataset required relatively little cleaning. This allowed me to move quickly into exploratory data analysis and feature engineering.

Exploratory Data Analysis (EDA)

The goal of the exploratory data analysis is to understand the main patterns that drive daily sales at the Sea Island Tennis Shop. In particular, I focused on:

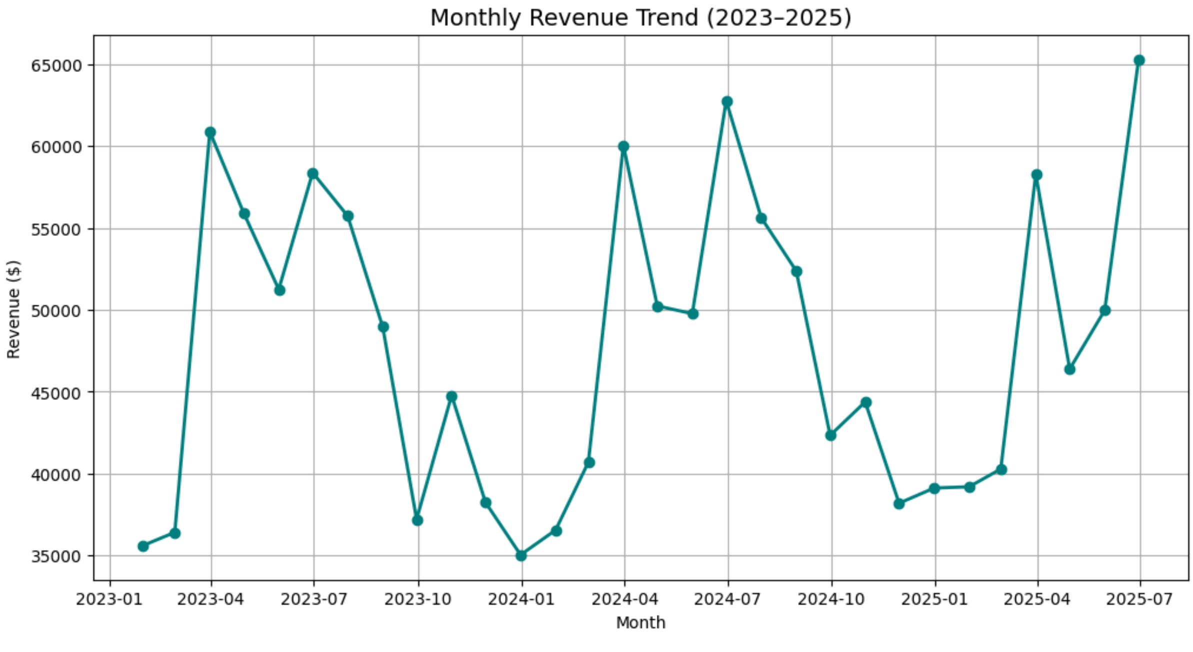
- Overall revenue and margin over time.
- Differences between product categories.
- The contribution of guests vs. members.
- Weekly and seasonal patterns in demand.
- The impact of weather and tennis events.
- Correlations between numeric variables such as traffic, inventory, and sales.

I began by computing simple aggregates such as total revenue, total margin, and average daily revenue. For the full period, the shop generated good numbers in revenue and in margin, with a higher average in daily revenue. These numbers confirm that even relatively small improvements in forecasting can have meaningful financial impact over time.

Seasonality in Monthly Revenue

Figure 2 shows the monthly revenue trend from January 2023 to June 2025. The line plot reveals a strong seasonal pattern: revenue increases sharply during spring and summer, especially between March and July, and declines during fall and winter. This pattern reflects the resort’s peak season, when more guests are on property and tennis activity is higher.

Capturing this seasonality is essential for any forecasting model. If the model cannot learn that March-July months tend to be systematically stronger than November-January, predictions will be biased and inventory will either be overstocked or understocked.

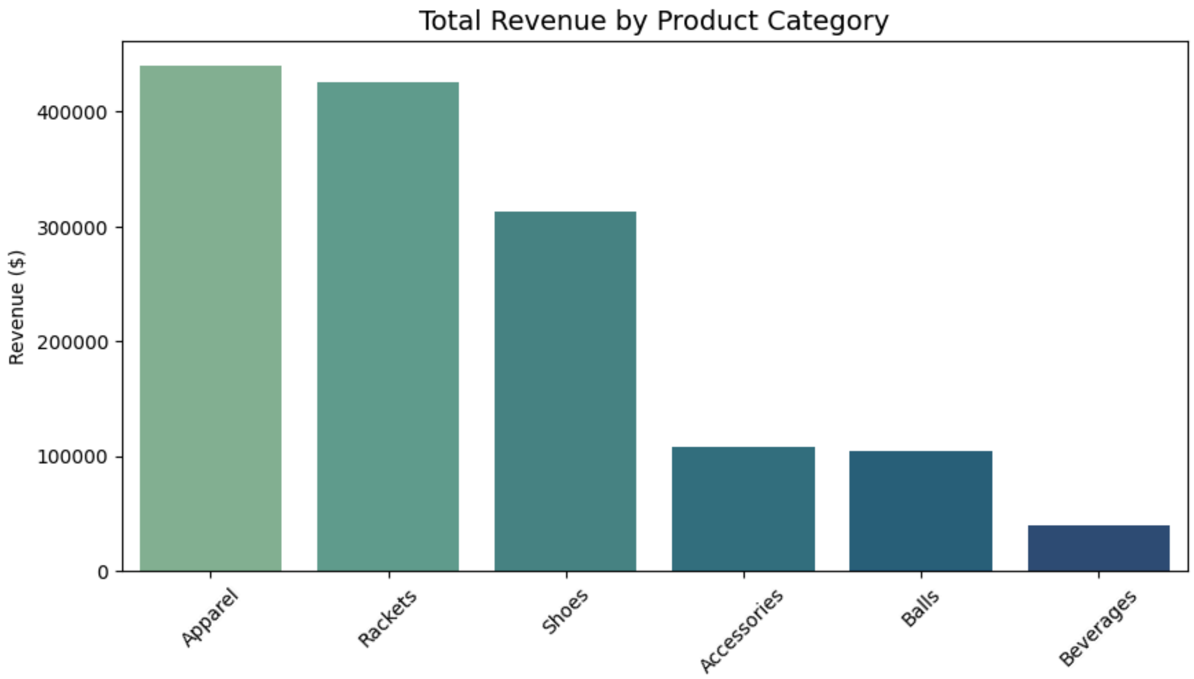


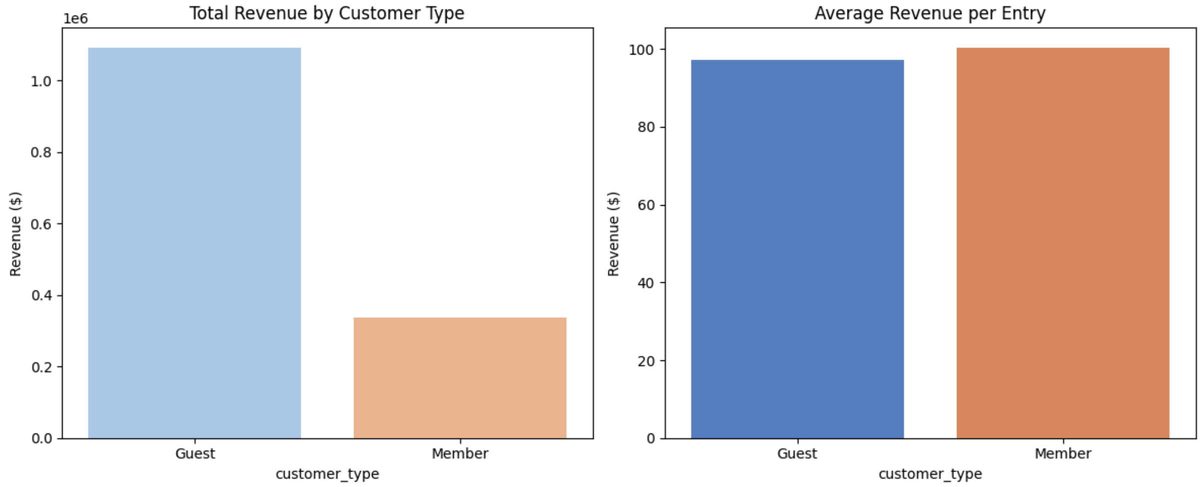
Product Categories and Customer Types

Next, I examined how revenue is distributed across product categories. Figure 3 shows total revenue by category. Apparel, rackets, and shoes dominate the shop's performance, accounting for the majority of revenue. Accessories, balls, and beverages contribute smaller but still meaningful portions.

Understanding which categories drive revenue is important for both forecasting and management. From a modeling perspective, it suggests that `category` is a key feature. From a business perspective, it highlights which segments deserve more attention in pricing, inventory planning, and merchandising.

I also compared revenue between guests and members. Guests generate higher total revenue, which is expected given the volume of resort visitors, while members contribute more stable, recurring purchases over time. This split reinforces the idea that demand is tied not only to local membership but also to resort occupancy and events.

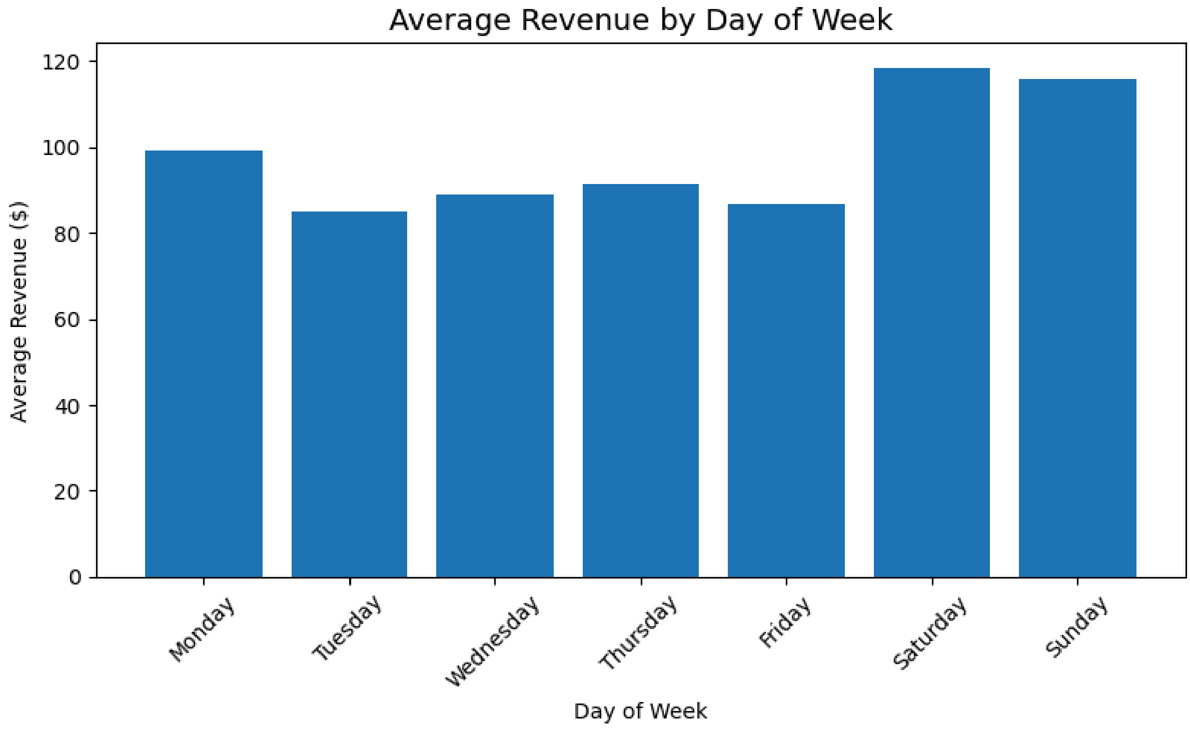


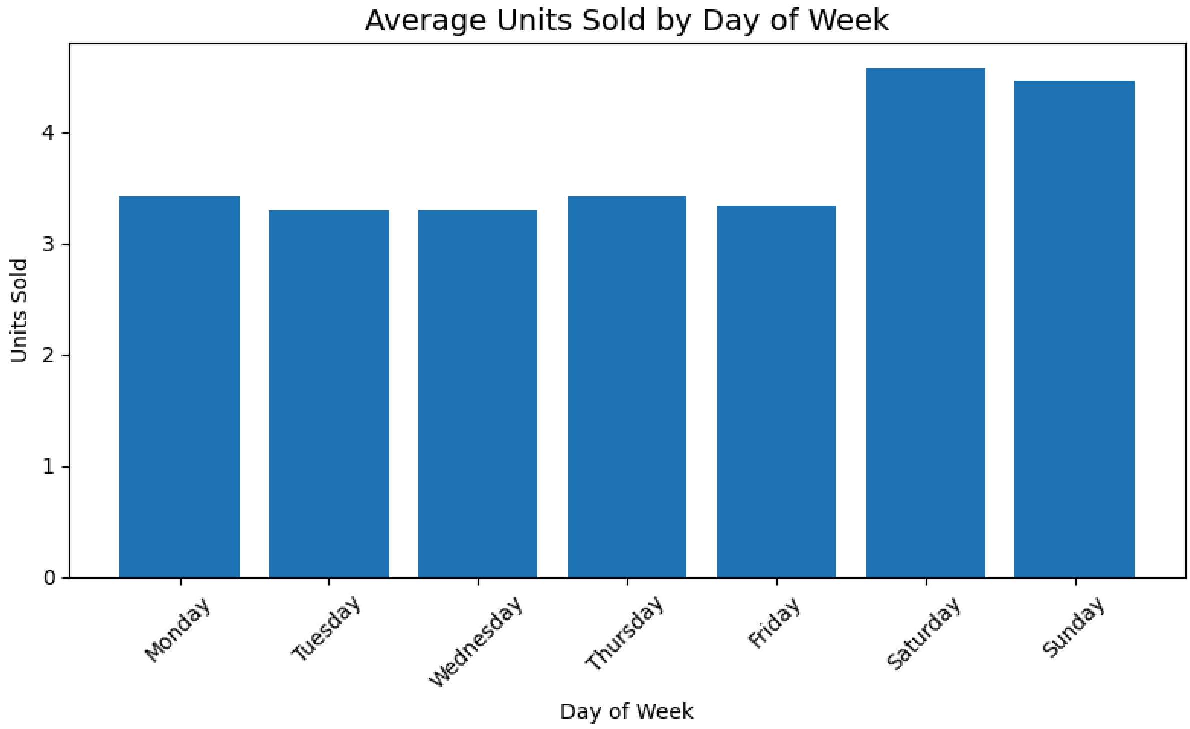


Day-of-Week and Weekly Patterns

Daily sales are not uniform across the week. Figure 5 shows average revenue or units sold by day of the week. The pattern is clear: weekends (especially Saturday and Sunday) exhibit significantly higher sales than weekdays. This matches the resort’s guest traffic, which tends to peak on weekends.

This weekly cycle is crucial for forecasting. A model that predicts the same sales for a Monday and a Saturday will systematically misallocate inventory and staffing. Including variables such as `day_of_week` and `is_weekend` helps the model learn that different days of the week have different expected demand.



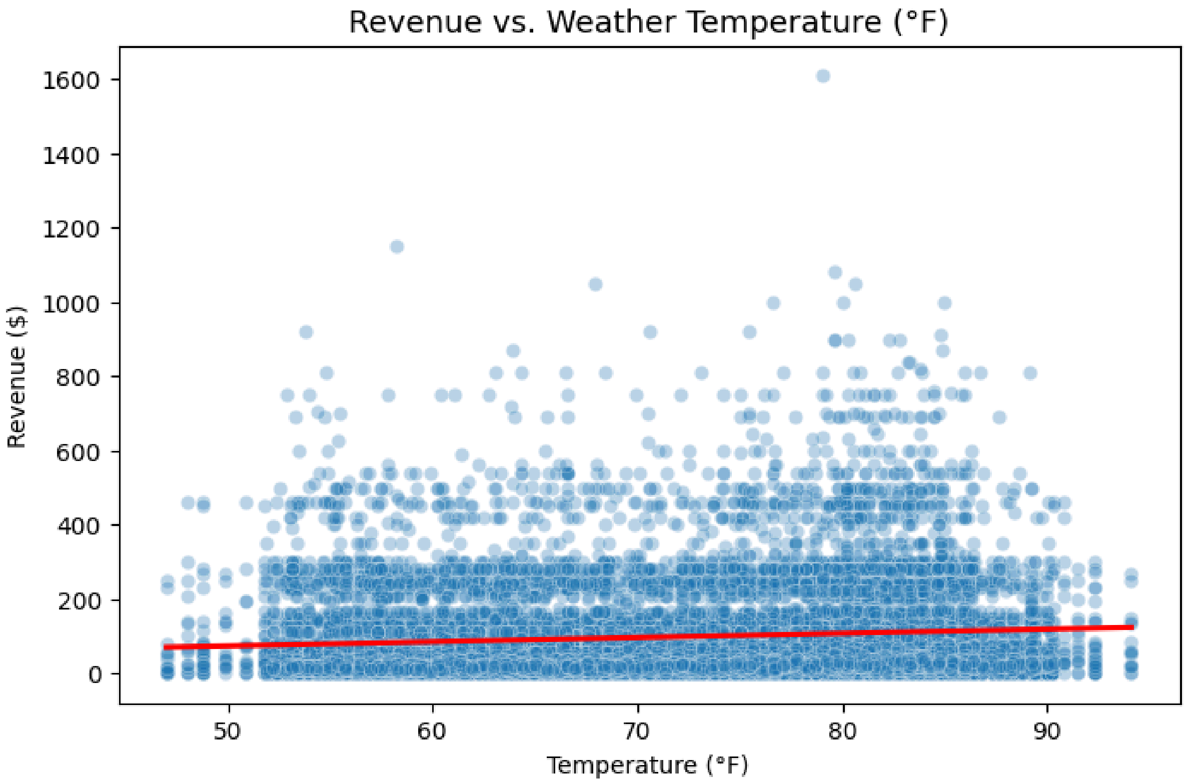


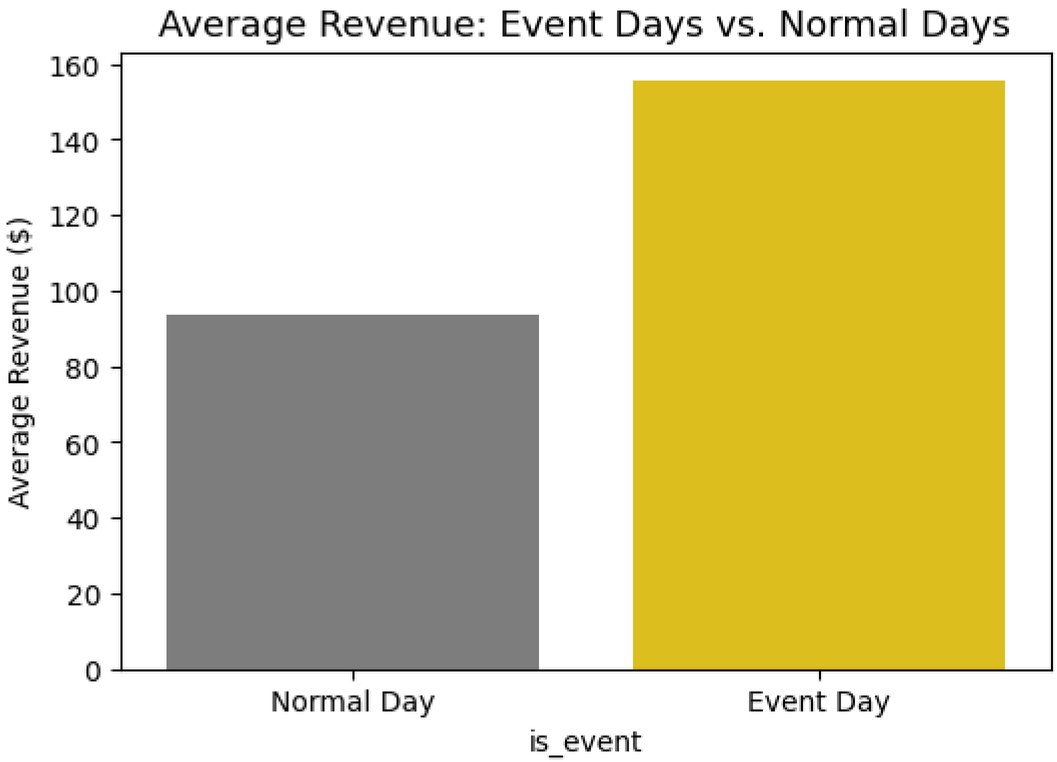
Impact of Weather and Tennis Events

Because tennis is an outdoor sport, weather conditions strongly influence both court usage and shop traffic. Figure 6 plots daily revenue against average temperature. The scatter plot, together with a fitted regression line, shows a positive relationship: warmer days are associated with higher revenue, while cooler days tend to generate lower sales.

I also evaluated the impact of tennis events such as tournaments, junior camps, and member-guest weekends. By creating a simple `is_event` flag, I compared average revenue on event days vs. normal days. Figure 7 shows that event days generate substantially higher revenue, roughly 50-60% more than non-event days. This makes intuitive sense, as events bring many players and families to the courts and the shop.

Both weather and events are therefore important features for the forecasting model. They represent external drivers of demand that complement internal factors like price and inventory.

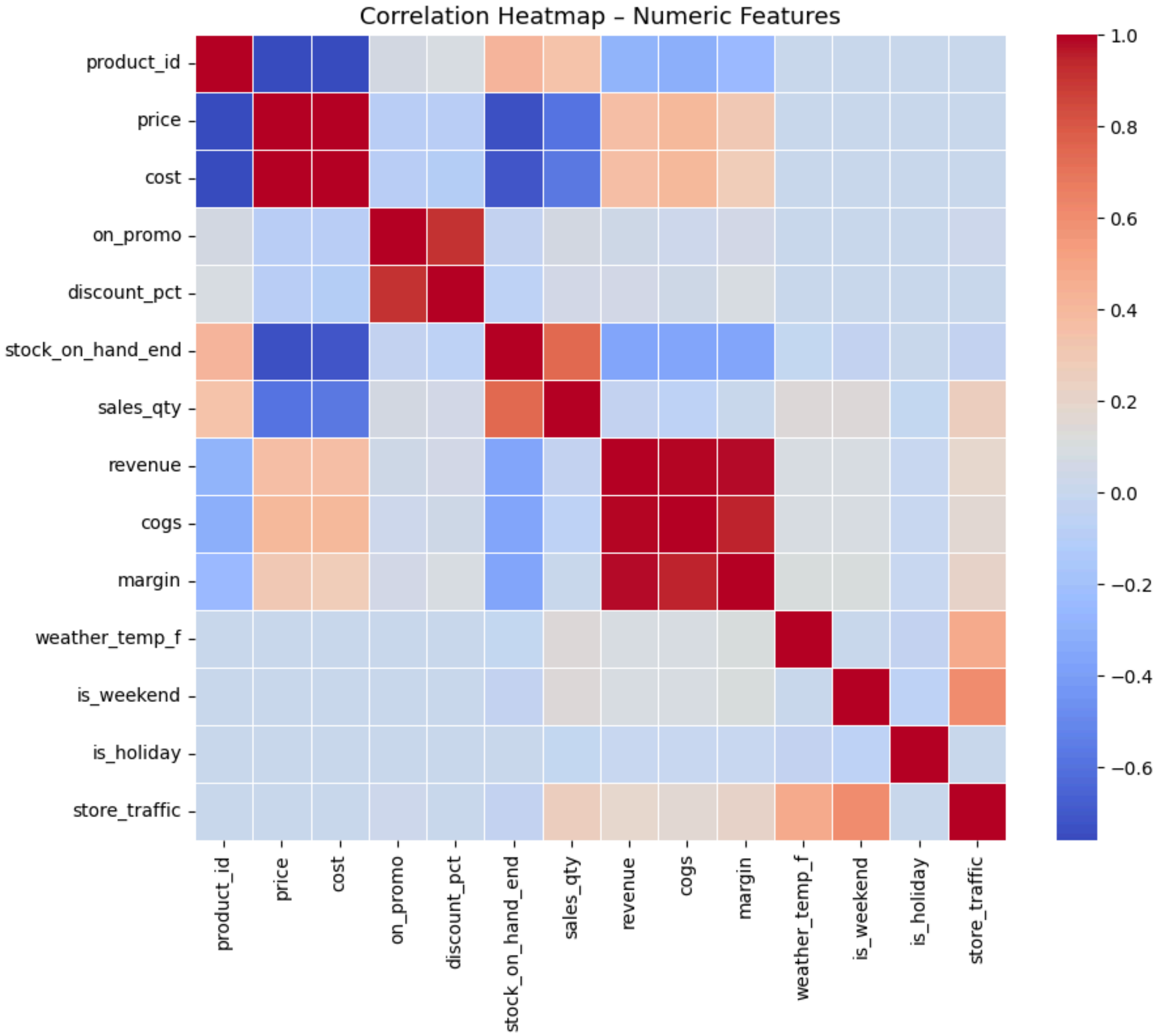


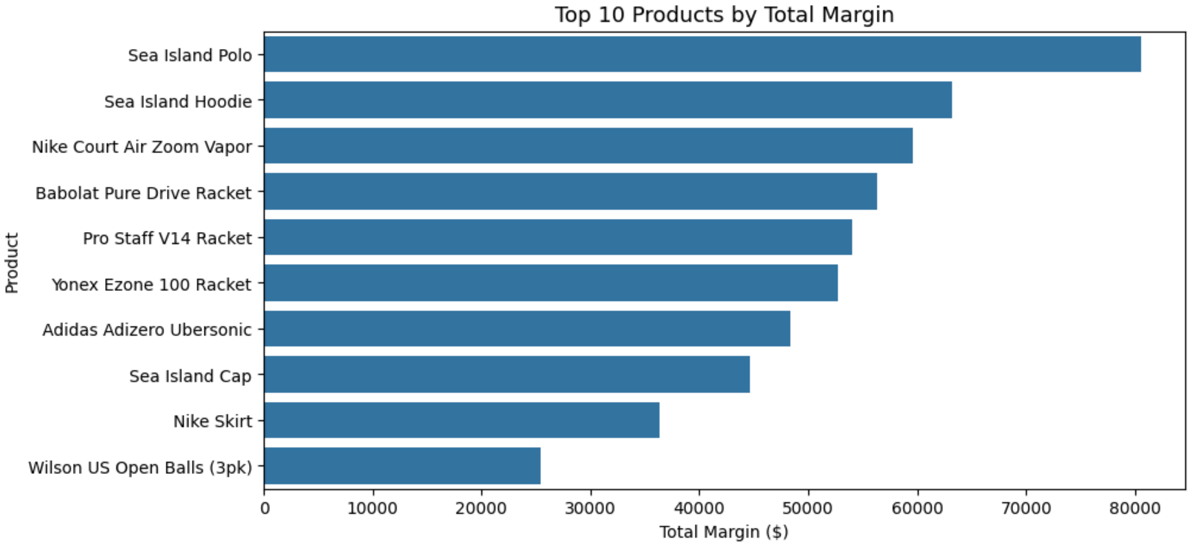


Correlations and Top Products

To better understand the relationships among numeric variables, I computed a correlation matrix and visualized it using a heatmap (Figure 8). The heatmap shows that store traffic, inventory on hand, and price are among the variables most strongly associated with sales quantity and revenue. These correlations do not prove causality, but they provide a useful guide for feature selection and highlight which drivers are likely to be important for forecasting.

I also examined which individual products contribute most to total margin. Figure 9 ranks the top 10 products by cumulative margin. A small number of premium items-such as high-end apparel, performance shoes, and advanced rackets-account for a disproportionate share of profit. This concentration suggests that accurate forecasts for these key products can have an outsized impact on overall profitability.





Overall, the exploratory analysis reveals strong seasonal and weekly patterns, clear differences across categories, and meaningful effects of weather, events, and traffic. These insights guided the design of the feature engineering and modeling steps described next.

Feature Engineering and Model Mathematics

This section describes the feature engineering steps used to prepare the dataset for modeling, followed by the mathematical formulation of each model and the evaluation metrics used to compare performance.

Feature Engineering

Based on the exploratory data analysis, several engineered features were created to help the models capture seasonality, weather effects, product behavior, and sales momentum. These include:

1. Time-Based Features

- Year
- Month
- Day of month
- Day of week
- Weekend flag (0/1)

These features help the model understand monthly seasonality and weekly sales patterns.

2. Weather Features

- Daily average temperature
- Temperature category (cold, cool, warm, hot)

This allows the model to capture how warm weather increases tennis activity and store traffic.

3. Event Indicator

- A binary flag indicating whether a tennis event (tournament, camp, or member-guest weekend) occurred on that day.

4. Lag Features (Sales Momentum)

Lagged sales for each product were created to capture short-term sales patterns.

The formulas used were:

- lag_1 = sales from the previous day
- lag_7 = sales from the same weekday one week earlier

A 7-day rolling average was also created:

- rolling_7 = average of the previous 7 days of sales

These features help the model understand trends such as increasing or decreasing demand.

5. Promotion-Traffic Interaction

- promo_traffic = on_promo * store_traffic

This feature captures how promotions work differently depending on traffic levels.

6. One-Hot Encoding of Categorical Variables
Categorical variables such as product category and temperature category were converted into dummy variables for modeling.

Linear Regression Model

Linear Regression was used as a baseline model.
It predicts daily sales using a weighted sum of the input features.

Mathematical formulation:

$\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_p \cdot x_p$

Where:

- y_hat = predicted daily sales
- β_0 = intercept
- $\beta_1 \dots \beta_p$ = feature coefficients
- $x_1 \dots x_p$ = input features (price, inventory, lag features, traffic, etc.)

Linear Regression is easy to interpret but cannot capture non-linear relationships.

Random Forest Model

A Random Forest is an ensemble of many decision trees.
Each tree makes a prediction, and the final output is the average of all trees.

Mathematical formulation:

y_hat = (1 / K) * (f1(x) + f2(x) + … + fK(x))

Where:

- K = number of trees
- fk(x) = prediction of the k-th tree

Random Forest captures non-linear patterns well.

XGBoost Model (Final Model)

XGBoost builds trees **sequentially**, where each new tree tries to correct the errors of the previous trees.

Mathematical formulation:

y_hat = f1(x) + f2(x) + ... + fK(x)

The model optimizes:

Loss = (sum of squared errors) + regularization

Regularization helps prevent overfitting and improves generalization.

XGBoost is known for high accuracy in forecasting tasks.

Evaluation Metrics

Three evaluation metrics were used to compare the models.

1. Mean Absolute Error (MAE)

Measures average absolute difference between predicted and actual sales.

MAE = average of | y - y_hat |

(Write exactly like this in Docs-no symbols break.)

2. Root Mean Squared Error (RMSE)

Penalizes large errors more heavily.

RMSE = square root of the average of (y - y_hat)^2

3. R-Squared (R^2)

Measures how much of the variation in sales the model explains.

R2 = 1 - (sum of squared errors / total variance)

Train-Test Split

A time-based split was used:

- Training: January 2023 - December 2024
- Testing: January 2025 - June 2025

This ensures that the model predicts **future** sales using **past** data, just like in real operations.

Model Training and Comparison

This section describes how the models were trained, how the data was split for evaluation, and how performance was compared across Linear Regression, Random Forest, and XGBoost.

Time-Based Train-Test Split

Because this project involves forecasting future daily sales, a time-based split was applied instead of a random split. This ensures the model is always predicting **future data** rather than memorizing shuffled values.

The split was:

- **Training period:** January 2023 to December 2024
- **Testing period:** January 2025 to June 2025

This approach mimics real-world use of the model, where managers generate predictions for upcoming days based on historical data.

Training Procedure

All models were trained using the engineered features described earlier. The training pipeline included:

1. **Feature selection and encoding**
   Numerical features were scaled when necessary, and categorical variables were one-hot encoded.
2. **Handling lagged features**
   Lag values were shifted forward in time to ensure no future information leaked into the model.
3. **Hyperparameter tuning**
   - For Random Forest, parameters such as max_depth, min_samples_leaf, and the number of trees were adjusted.
   - For XGBoost, learning_rate, max_depth, subsample, and number of boosting rounds were tuned.
4. **Model fitting**
   Models were trained on the 2023-2024 period using scikit-learn and the XGBoost library.
5. **Prediction on 2025 test data**
   All models generated predictions on unseen 2025 data, allowing for a fair performance comparison.

Evaluation Metrics

To compare the models, three standard regression metrics were used:

1. Mean Absolute Error (MAE)

Measures the average size of errors.

MAE = average of absolute differences between actual and predicted values.

2. Root Mean Squared Error (RMSE)

Penalizes larger errors more strongly.

RMSE = square root of the average of squared errors.

3. R-Squared (R2)

Represents the proportion of variance explained by the model.

R2 = 1 minus the ratio of squared prediction errors to total variance.

These metrics provide a balanced perspective, capturing both accuracy and stability.

Model Performance Comparison

The table below summarizes performance on the 2025 test data:

(You will insert this as a table in Docs)

Table 2. Model performance on the 2025 test set

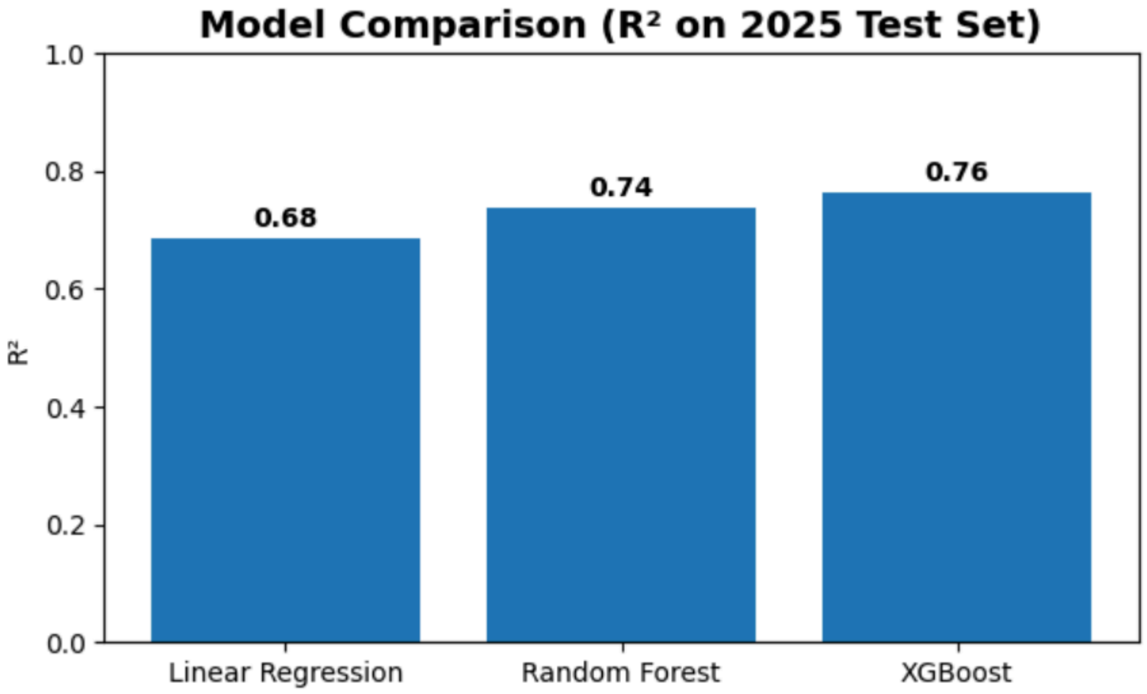
Model	R2	MAE	RMSE
Linear Regression	0.68	1.59	2.20
Random Forest Regressor	0.74	1.42	2.00
XGBoost Regressor	0.76	1.32	1.90

This comparison highlights several important observations:

- Linear Regression provides a useful baseline but struggles with non-linear relationships.

- Random Forest improves performance by capturing more complex patterns.
- **XGBoost clearly performs best**, achieving the highest R2 and the lowest MAE and RMSE.

Based on these results, XGBoost was selected as the **final forecasting model** for this project.



Feature Importance Analysis

After selecting XGBoost as the final forecasting model, I analyzed feature importance to understand which variables have the strongest influence on predicted daily sales. Feature importance helps managers see which factors matter most for planning promotions, inventory, and staffing.

The model's importance scores show that both **operational features** and **external factors** contribute to accurate predictions.

Top Predictive Features

According to the XGBoost model, the most important features include:

1. Inventory on hand (stock_on_hand_end)

Inventory is the strongest driver of sales. If stock is low or zero, the model cannot predict high sales regardless of demand. This validates a common retail principle: “No stock, no sales.”

2. Store traffic

Traffic represents the number of customers entering the store. Higher traffic increases the likelihood of purchases, especially for apparel, accessories, and impulse items.

3. Lagged sales (lag_1 and lag_7)

These features capture short-term sales momentum.
Products that performed well yesterday or last week often continue to perform well, especially during peak season.

4. Rolling 7-day average (rolling_7)

This smooths out day-to-day noise and gives the model a stable trend indicator.

5. Weather temperature

Warm days increase tennis activity and store visits, improving the chances of sales. This aligns with patterns observed in the exploratory analysis.

6. Calendar variables (month, day_of_week)

Seasonality is extremely important.
Sales are highest from March to July and much lower in winter months.
Weekends also consistently outperform weekdays.

7. Price and product-specific factors

Product ID and price help the model differentiate between items and adjust expected demand accordingly. Higher-priced items tend to sell fewer units but contribute more profit per sale.

Business Interpretation

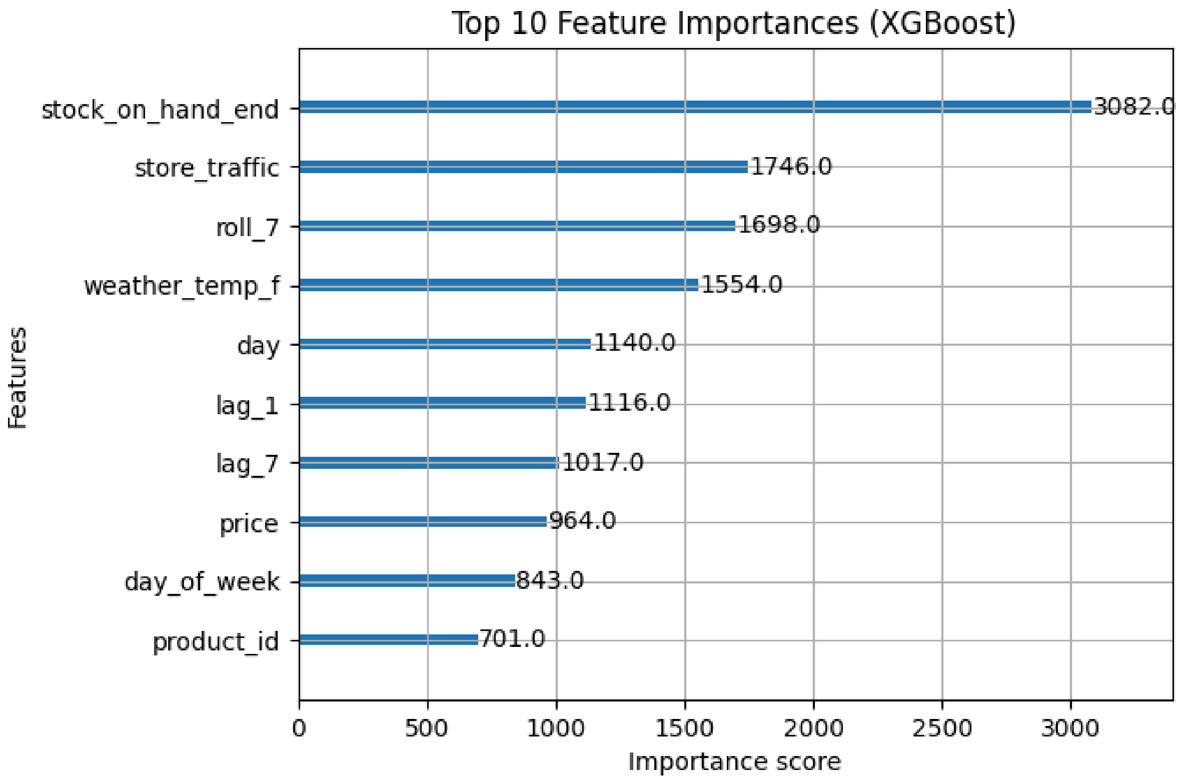
The feature importance results align with operational intuition from the tennis shop:

- Products with strong recent sales should be monitored closely and restocked proactively.

- Weather and traffic should be considered when planning staffing and promotions.

- Key products (especially rackets and apparel) must maintain stronger inventory levels to avoid missed sales opportunities.

These insights reinforce the value of combining internal and external variables in the forecasting process.



Scenario Simulations

Beyond forecasting individual daily sales, one of the goals of this project was to help the Sea Island Tennis Shop make better operational decisions. To demonstrate the practical use of the model, I created three simulation scenarios that show how predicted sales change under different conditions.

These scenarios are realistic, based on patterns observed in the data, and can be used by store managers to plan inventory, promotions, and staffing ahead of busy days or events.

Scenario 1: Normal Day (Baseline)

This scenario represents a typical weekday with average weather and no events or promotions.

Assumptions:

- Normal temperature (around seasonal average)
- No tennis events
- No promotions
- Average store traffic

Predicted daily sales: approximately **1.4 units per product**

This serves as the baseline against which other scenarios are compared.

Scenario 2: Promotional Weekend

This scenario simulates a weekend with higher-than-average traffic and a moderate promotion on selected products.

Assumptions:

- Weekend (Saturday or Sunday)
- Higher store traffic
- Small discount (10 to 15 percent)
- Warm weather

Predicted daily sales: approximately **1.7 units per product**

Sales increase due to both the promotion and the naturally higher weekend traffic.
This indicates that promotions are more effective when combined with high-traffic days.

Scenario 3: Tennis Event + Promotion + Warm Weather

This is the most favorable scenario and represents the type of day when the store experiences peak sales.

Assumptions:

- Member-guest tournament or junior event
- High store traffic

- Small promotion
- Warm or hot weather

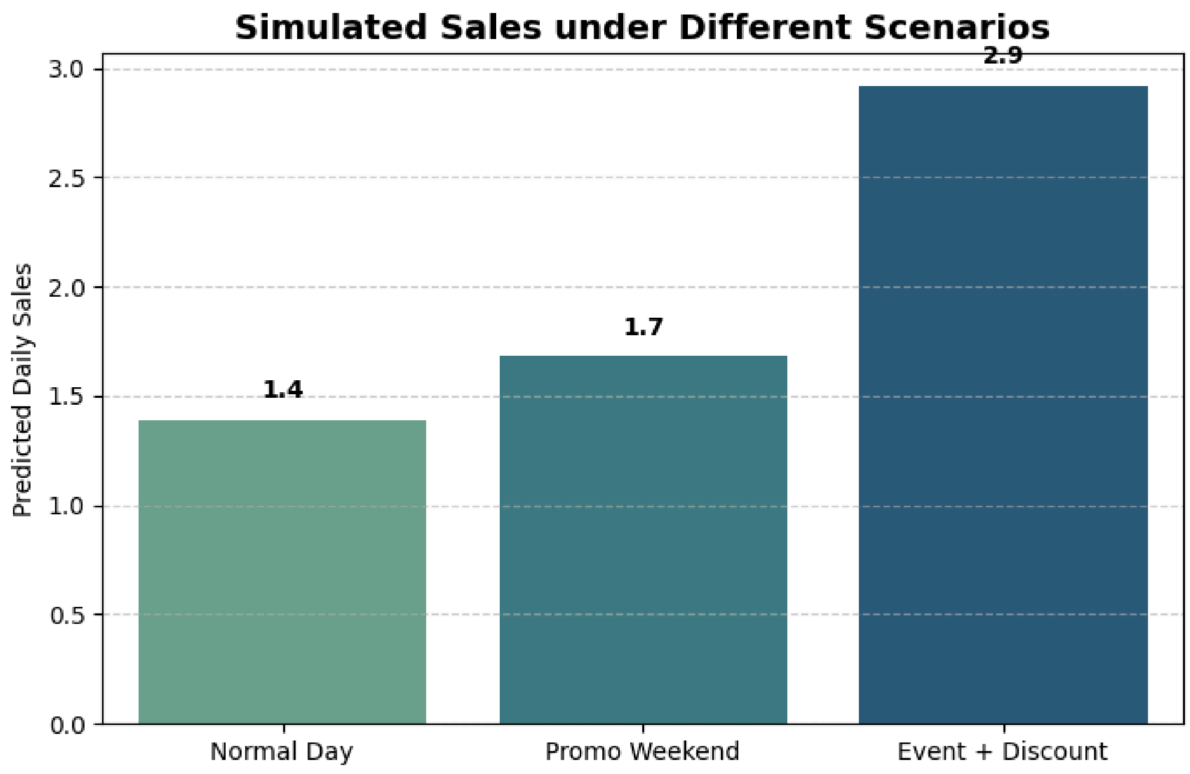
Predicted daily sales: approximately **2.9 units per product**

Under these combined conditions, demand nearly doubles compared to a normal day. This scenario emphasizes how weather, events, and promotions work together to create stronger demand.

Business Use Cases

These simulations provide practical guidance for the tennis shop:

1. **Inventory Planning**
   - Ensure key products (apparel, shoes, rackets) are stocked ahead of tournaments.
   - Avoid stockouts during predictable high-demand periods.
2. **Promotion Strategy**
   - Promotions are more effective when paired with high-traffic or event days.
   - Avoid using discounts during slow periods unless necessary.
3. **Staffing Decisions**
   - Managers can schedule more staff during warm weekends and tournament days.
4. **Pricing and Display**
   - High-margin products can be promoted strategically when demand is expected to spike.



Business Recommendations

Based on the forecasting results, feature importance analysis, and scenario simulations, several actionable recommendations were developed to help the Sea Island Tennis Shop improve its operations. These recommendations are grounded in the data and directly address the challenges the store faces with inventory planning, demand uncertainty, and staffing during peak periods.

Inventory Optimization

1. Maintain higher inventory levels for top-margin products.
Items such as premium apparel, high-end rackets, and performance tennis shoes drive a large portion of total margin. Stockouts in these categories result in significant lost revenue. Ensuring that these items remain available during peak months is essential.

2. Prepare inventory ahead of events.
Tennis events (tournaments, camps, member-guest weekends) consistently generate higher sales. Inventory orders should be placed earlier, especially for sizes and products that commonly sell out.

3. Use forecasts to reduce overstock.
During slower months (November-January), the model can help prevent unnecessary overstock by predicting lower demand ahead of time.

Promotion Strategy

4. Concentrate promotions during higher-traffic days.
The scenario analysis shows that discounts are more effective on weekends and event days.
Avoid spreading promotions evenly throughout the month, which dilutes impact.

5. Use targeted discounts for low-velocity products.
Rather than broad promotions, targeted markdowns can help move slow-moving products without reducing margin unnecessarily.

Staffing and Scheduling

6. Increase staffing during warm weekends and events.
The model shows clear sales spikes under these conditions. Adding additional staff can improve customer service, reduce wait times, and increase upselling opportunities.

7. Reduce staffing during historically slow periods.
Forecasting can help managers schedule more efficiently during cold-weather weekdays or post-holiday periods.

Weather and Traffic Integration

8. Combine weather forecasts with model predictions.
Because warm weather correlates with higher sales, managers can use weather forecasts to anticipate busy days even before the weekend arrives.

9. Monitor store traffic daily.
Traffic is one of the strongest predictors of sales. Tracking this metric allows managers to react quickly when demand increases unexpectedly.

Long-Term Improvements

10. Create a simple dashboard for weekly planning.
A dashboard combining sales forecasts, weather, traffic, and event information would allow staff to make proactive decisions each week.

11. Autom ate the forecasting workflow.
Running the model weekly or daily would ensure managers always have updated predictions for smarter planning.

12. Integrate model outputs into ordering decisions.
Linking forecasts to purchasing thresholds could reduce manual effort and improve inventory efficiency.

Summary of Recommendations

To summarize the most impactful actions:

- Stock up before events and warm weekends
- Use promotions strategically during high-traffic periods
- Keep premium, high-margin items in strong supply
- Use forecasts to avoid over- or under-ordering
- Schedule staff based on predicted demand patterns
- Build a simple operational dashboard

These recommendations translate the forecasting results into practical actions the shop can implement immediately.

Conclusion

This project developed a complete end-to-end demand forecasting system for the Sea Island Tennis Shop, addressing the real business challenges of inventory planning, seasonality, and demand uncertainty. Using daily product-level data from 2023 to 2025, the forecasting pipeline included data cleaning, exploratory analysis, feature engineering, machine learning modeling, and scenario simulations.

Key Findings

Several important insights emerged from the analysis:

1. Strong seasonality patterns
Sales peak between March and July, while winter months show consistent slowdowns.

2. Weekends significantly outperform weekdays
Traffic increases during weekends, leading to higher revenue and unit sales.

3. Weather matters
Warm days drive tennis activity and increase store traffic, producing higher sales.

4. Events generate substantial demand spikes
Tournaments and member-guest weekends consistently result in elevated sales compared to non-event days.

5. A small number of products drive most of the margin
Premium apparel, shoes, and rackets account for a large share of profitability.

Model Performance

Among the three machine learning models tested-Linear Regression, Random Forest, and XGBoost-**XGBoost delivered the highest accuracy** with:

- Highest R2
- Lowest MAE
- Lowest RMSE

Its ability to capture non-linear relationships, seasonality, and interactions between demand drivers made it the strongest forecasting model for this project.

Practical Applications

The forecasting system is not only accurate but also actionable. Store managers can use model predictions to:

- Plan inventory more efficiently
- Reduce stockouts during peak demand
- Adjust staffing according to expected traffic
- Schedule promotions on impactful days
- Prepare for events and seasonal fluctuations

Scenario simulations demonstrated how demand changes under different conditions, providing a clear decision-making tool for the shop.

Limitations and Future Improvements

While the model performed well, a few limitations remain:

1. No real-time POS data
Predictions would be even stronger with minute-by-minute transaction data.

2. Limited promotional history
More detailed price elasticity information would allow for stronger pricing optimization.

3. Weather conditions simplified
Future models could integrate humidity, rainfall, or UV index for more precise analysis.

4. No product attributes
Including product-level metadata (brand, sport type, material) could refine forecasts.

5. Lack of foot-traffic segmentation
Segmenting members, resort guests, juniors, and event participants would add accuracy.

Final Remarks

Overall, this project demonstrates how data science and machine learning can significantly improve retail operations inside a dynamic environment like the Sea Island Tennis Shop. The final forecasting model provides a reliable and scalable foundation that the store can use to make better decisions throughout the year.

By applying the recommendations and integrating these tools into weekly and monthly planning, the tennis shop can operate more proactively, avoid unnecessary costs, and increase both customer satisfaction and profitability.

References

[1] Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830.

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[3] McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.

[4] Harris, C. R. et al. (2020). *Array programming with NumPy*. Nature, 585, 357-362.

[5] Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90-95.

[6] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

[7] Sea Island Tennis Shop (2023-2025). *Retail Sales and Inventory Dataset*. Internal dataset used for academic purposes.

[8] College of Coastal Georgia - DATA 3371 Course Materials (2025).