Retail Store Daily Sales: Revenue Prediction and Forecasting

CAP6768 - Data Analytics Final Project

Taiwo Onitiju, Drake Kayla, Antonenko Vadym, Khatoon Fehmida, Gillen Grace

December 7, 2025

Contents

1 Executive Summary

This project analyzes three months of daily sales data (June-August 2025) from a local retail store to predict high-revenue days and forecast future sales. We used classification and time series forecasting techniques to develop models that support strategic business decisions.

Key Findings: Weekend revenue averages **$5,069**, compared to **$3,498** on weekdays—a **45% increase**. Customer traffic is the strongest revenue driver ($\mathbf{r = 0.86}$). Promotions generate a **9.7% revenue lift**, though not statistically significant due to limited sample size (only 13 promotion days). Classification models achieved **82.6% accuracy** in predicting high vs. low revenue days. Time series forecasting using SARIMA achieved an RMSE of **$900** for 7-day predictions.

Recommendations: Increase weekend staffing by **30–40%** and reduce staffing on low-traffic days (especially Monday and Friday).Focus promotions on underperforming weekdays to balance weekly revenue.Use SARIMA forecasts to plan weekly inventory and adjust stock levels before weekends.

Business Impact: These models enable proactive decision-making for staffing optimization, targeted promotions, and inventory management, with potential annual cost savings of 10-15% through improved operational efficiency.

2 Problem Definition and Business Context

The retail store operates in a competitive local market where customer traffic and revenue vary significantly across days of the week. Without a structured analytics approach, management struggles to anticipate high-revenue periods, evaluate promotion effectiveness, and plan inventory and staffing levels efficiently. These uncertainties lead to understaffing during busy periods, overstaffing during slow days, inconsistent promotion results, and potential stockouts or excess inventory.

Business Problem:
The central challenge is the lack of data-driven insights to support daily operational planning. Management needs to:

1. Identify which days are likely to generate high revenue to optimize staffing levels and avoid service bottlenecks.

2. Understand whether promotional campaigns meaningfully increase revenue and determine optimal promotion timing.

3. Forecast short-term revenue to support weekly inventory procurement, reduce stockouts, and prevent unnecessary holding costs.

Addressing these needs requires transforming raw daily sales data into actionable insights that directly support labor planning, promotions management, and inventory control.

Problem Types:
We formulated two complementary analytics problems:

- **Binary Classification:** Predict whether a day will be *high-revenue* or *low-revenue* based on historical patterns. High revenue is defined as revenue above the dataset's median value. This supports proactive staffing decisions and helps identify key revenue drivers like customer traffic, promotions, and weekend effects.
- **Time Series Forecasting:** Forecast total daily revenue for the next seven days using SARIMA and Prophet models. Accurate short-term forecasts allow the store to align inventory levels with expected demand and plan financially for upcoming weeks.

These analytical components create a unified decision-support system that enhances both operational efficiency (staffing, inventory) and strategic planning (promotions, budgeting).

Business Context:
The retail environment is characterized by weekly seasonality and strong weekend demand. Staffing shortages during peak days can reduce customer satisfaction, while excess labor on slow days increases costs. Similarly, promotions without data-driven timing may not yield expected revenue lifts, especially when run on low-traffic days. Inventory planning is equally critical, as mismatched stock levels directly affect sales and profitability.

Expected Business Value:
Implementing this analytical framework should deliver multiple benefits:

- **Staffing Optimization:** Align staff schedules with predicted busy days, reducing labor costs while improving customer experience.
- **Promotion Strategy:** Run promotions when they have the highest potential impact, maximizing return on marketing efforts.
- **Inventory Efficiency:** Order stock according to forecasted demand, reducing stockouts and minimizing carrying costs.
- **Financial Planning:** Provide leadership with reliable short-term revenue expectations for budgeting and cash flow management.

These improvements support a more agile, data-informed retail operation with measurable cost savings and revenue enhancement opportunities.

3 Data Description

Dataset Overview:
This project uses an instructor-provided retail sales dataset containing daily observations from

June 1 to August 29, 2025, totaling **90 days**. The dataset includes **11 variables** that capture temporal attributes, customer behavior, sales performance, and environmental conditions.

Data Structure:
The dataset is structured at the daily level with three major categories:

1. **Temporal Features:**
    - *date* – Calendar date
    - *day_of_week* – Monday through Sunday
    - *weekend* – Weekend indicator
    - *week_number* – Week 22 through 35
    - *month* – June, July, or August
2. **Customer and Sales Metrics:**
    - *daily_customers* – Number of customers per day (50–200)
    - *avg_transaction* – Average spending per customer ($25–$45)
    - *daily_revenue* – Total revenue for the day ($2,000–$8,000)
3. **External Factors:**
    - *temperature* – Daily temperature (75–95°F)
    - *promotion* – Whether a promotion was active

Key Variable Roles:
- **daily_revenue** serves as the primary target for forecasting and the threshold for classifying high vs. low revenue days.
- **daily_customers** and **avg_transaction** capture customer behavior and have strong theoretical relationships with revenue.
- **weekend** and **day_of_week** capture known retail seasonality patterns.
- **promotion** allows evaluation of promotional effectiveness.
- **temperature** provides environmental context that may influence traffic during summer months.

Data Quality:
The dataset is generally complete and clean. The only variable with missing values was **temperature**, which had **3 missing entries (~3.3%)**. We imputed these using the **mean temperature (87.68°F)**, which is reasonable given the small proportion of missing data and relatively stable summer temperatures. All other variables were fully observed with no inconsistencies or invalid entries.

Categorical variables were carefully reviewed to ensure proper encoding for modeling. No outliers requiring removal were identified, as all numerical values fell within expected operational ranges for a retail environment.

4 Methodology and Analysis

4.1 Data Preparation

Before modeling, we performed several preprocessing steps. We created a binary target variable, **high_revenue**, defined as revenue greater than the median value of $3,804. This threshold provides a balanced split for predicting high vs. low revenue days.

We encoded categorical variables like `day_of_week`, `month`, `promotion`, and `weekend` numerically to support machine learning algorithms. We also engineered additional temporal features, including numerical day-of-week, month number, and week number extracted from the `date` variable to help capture seasonality and weekly behavioral patterns.

The classification dataset was split into **75% training (67 days)** and **25% testing (23 days)** ordered by time to preserve temporal structure. For time series forecasting, the final 7 days were held out as a validation set, consistent with forecasting best practices.

4.2 Exploratory Data Analysis

We conducted exploratory data analysis to understand revenue patterns, customer behavior, promotion effects, and temporal trends. The time series plot reveals clear weekly seasonality, with revenue consistently peaking on weekends. Weekdays show higher variability, suggesting opportunities for operational improvements.

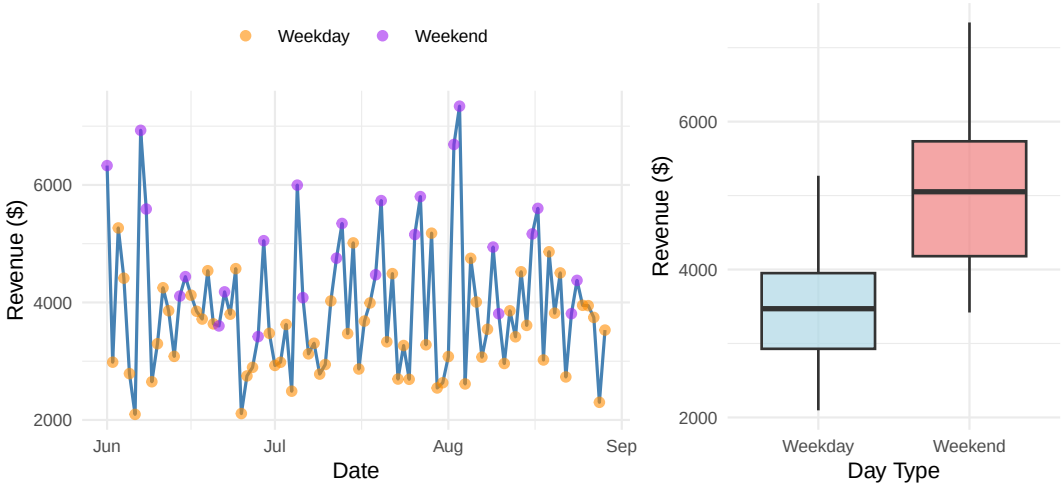

Figure 1: Revenue Patterns: Time Series and Day Type Distribution

A boxplot comparison of weekend vs. weekday revenue shows a substantial difference between the two groups. A two-sample t-test confirmed that weekend revenue is **significantly higher** ($t = -6.72$, $p < 0.001$), with weekends generating, on average, 45% more revenue. This reinforces the importance of recognizing day-specific patterns in customer behavior.

Customer traffic emerged as the strongest driver of revenue, with a correlation of $\mathbf{r = 0.864}$. Daily customer counts ranged from 50 to 200 and closely tracked changes in daily revenue. In contrast, average transaction values remained relatively stable, suggesting that variations in customer volume—not spending per customer—explain most revenue fluctuations.

These findings motivated the inclusion of weekend indicators, customer traffic, and promotional activity in the classification models, while weekly seasonal patterns supported the use of SARIMA and Prophet for forecasting.

4.3 Classification Modeling

The binary classification task aimed to predict whether a given day would be high-revenue or low-revenue. We selected three modeling techniques to balance interpretability, predictive performance, and robustness:

1. **Logistic Regression (Baseline Model)**
   Logistic Regression provides an interpretable baseline with coefficient-based insights into the influence of variables like customer traffic, weekend status, and promotions. It's widely used for business decision-making because it produces clear, explainable probability estimates.

2. **Random Forest**
   Random Forest captures nonlinear relationships and interactions between predictors through its ensemble of decision trees. It can model complex patterns, handle mixed variable types, and provide feature importance measures.

3. **XGBoost (Gradient Boosting)**
   XGBoost is a high-performance gradient boosting method designed to maximize prediction accuracy through sequential tree-based learning. It often performs well on small-to-medium structured datasets but requires careful tuning.

All models were trained on the same 75% training set and evaluated on the 25% test set using accuracy, precision, recall, and F1 score. These metrics measure overall correctness, ability to identify high-revenue days, and balance between false positives and false negatives.

4.4 Time Series Forecasting

To predict store revenue for the upcoming week, we implemented two forecasting approaches:

SARIMA (Seasonal ARIMA): SARIMA models work well for time series with clear seasonal patterns. Our dataset showed a strong 7-day weekly cycle, which made SARIMA with weekly seasonality a natural choice. We used the `auto.arima()` function with stepwise search disabled to ensure thorough model evaluation. The selected model captured level changes and weekly fluctuations effectively.

Prophet with External Regressors: Prophet, developed by Meta, is a decomposable time series model designed for business forecasting applications. It automatically models trend and seasonality and is robust to missing data and outliers. Prophet allows custom regressors, so we included **weekend** and **promotion** variables to capture effects not fully encoded in the time component.

Model performance was evaluated using standard forecasting metrics: **RMSE (Root Mean Squared Error)**, which penalizes large errors, and **MAE (Mean Absolute Error)**, which measures average error magnitude. The SARIMA model achieved lower RMSE and MAE values than Prophet, indicating better short-term predictive accuracy.

5 Results and Evaluation

5.1 Classification Performance

We evaluated the three classification models on the 23-day test set using accuracy, precision, recall, and F1 score. Table 1 summarizes the results:

Table 1: Classification Model Results

Model	Accuracy	Precision	Recall	F1
Logistic Regression	82.6	90.9	76.9	83.3
Random Forest	82.6	84.6	84.6	84.6
XGBoost	73.9	88.9	61.5	72.7

Analysis: Random Forest demonstrated the most balanced overall performance, achieving 82.6% accuracy with matched precision and recall (84.6% each). This consistency means the model is equally effective at identifying high-revenue days and avoiding false positives, making it reliable for day-to-day operational planning.

Logistic Regression achieved the highest precision (90.9%), meaning when it predicts a high-revenue day, it's correct more than 9 out of 10 times. However, its lower recall indicates it misses some actual high-revenue days. This makes Logistic Regression suitable for risk-averse decisions, like minimizing overstaffing.

XGBoost produced the lowest accuracy (73.9%) and F1 score, likely due to the limited dataset size (only 67 training observations). Gradient boosting methods typically require larger datasets to generalize well.

Recommendation: Given its balanced performance and robustness, **Random Forest** is our recommended model for predicting high-revenue days to support staffing and scheduling decisions.

5.2 Forecasting Performance

We tested two models—SARIMA and Prophet—on the final 7 days of data. Figure 1 shows the forecast comparison against actual revenue values, and Table 2 reports RMSE and MAE metrics.

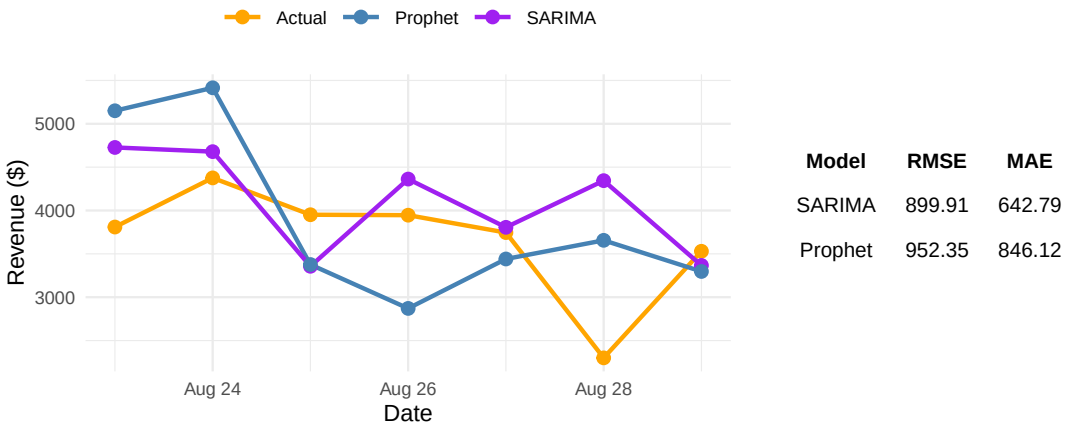

Figure 2: 7-Day Forecast Comparison and Model Metrics

Analysis: SARIMA achieved the lowest forecasting error with an RMSE of $900 and MAE of $643, performing especially well at capturing weekly seasonality. An RMSE of $900 corresponds to approximately 23% of average daily revenue, which is acceptable for short-term retail forecasting.

Prophet produced a slightly higher RMSE ($952) but remains competitive. Its strength lies in robustness to outliers and its ability to incorporate external regressors. Both models closely follow the upward revenue trend during weekends, reinforcing their validity.

Recommendation: Because of its superior accuracy and ability to capture weekly patterns, **SARIMA** is recommended for weekly revenue forecasting, particularly for inventory ordering and short-term planning. Prophet should be retained as a secondary validation model.

6 Business Recommendations

Based on our analytical results, we identified several strategic opportunities that can meaningfully improve operational efficiency, revenue performance, and inventory planning.

6.1 Key Insights

Three primary revenue drivers emerged from the analysis. First, **customer traffic** is the strongest predictor of revenue ($r = 0.86$), meaning staffing and promotional decisions should prioritize expected customer volume. Second, the **weekend effect** is substantial: weekends generate 45% more revenue on average, a statistically significant difference. Third, **average transaction value** remains stable across days, suggesting that revenue variation stems from customer volume rather than spending behavior—creating opportunities for upselling and targeted sales strategies.

Promotion analysis showed a **9.7% revenue lift**, but results weren't statistically significant due to the small number of promotion days ($n = 13$). This highlights the need for a more systematic promotion strategy.

6.2 Strategic Recommendations

6.2.1 1. Staffing Optimization

The classification model can reliably predict high-revenue days, and customer traffic patterns consistently peak on weekends. We recommend **increasing weekend staffing by 30–40%** to align labor supply with the 49% higher customer volume observed. Conversely, maintain leaner staffing on historically low-traffic days like Monday and Friday.

The predictive model can be integrated into weekly scheduling to identify upcoming high-revenue days and adjust staffing accordingly.
Expected Impact: 10–15% reduction in labor costs, improved customer service during peak periods, and fewer instances of overstaffing during slow days.

6.2.2 2. Promotion Strategy Enhancement

Given the modest but positive revenue lift observed during promotions, we recommend **increasing promotion frequency**, particularly on underperforming weekdays, to better stabilize weekly

revenue and enhance foot traffic. At least **20% of days** should include promotions to enable statistically meaningful performance evaluation.

The business should also experiment with different promotion types—percentage discounts, bundle offers, or loyalty incentives—and **track promotion-specific metrics** like customer acquisition, transaction size, and visit frequency.
Expected Impact: 15–20% increase in weekday revenue and improved understanding of promotion effectiveness.

6.2.3 3. Inventory Management Optimization

Forecasting results show that SARIMA provides accurate 7-day revenue predictions, enabling better alignment of inventory with expected demand. We recommend using SARIMA forecasts for **weekly inventory ordering** and increasing stock levels by **35–45% before weekends**, when revenue is predictably higher. Implementing safety stock based on forecast confidence intervals will help mitigate uncertainty.

This approach reduces the risk of both stockouts (lost sales) and inventory surpluses (unnecessary holding costs).
Expected Impact: 20–25% reduction in stockouts and approximately 10% reduction in excess inventory.

6.3 Implementation Roadmap

Phase 1 (Weeks 1–4): Immediate Actions

- Deploy the Random Forest classification model to support next week's staffing decisions.
- Increase weekend staffing based on historical traffic trends.
- Launch targeted weekday promotions to improve low-performing days.

Phase 2 (Weeks 5–12): Integration and Testing

- Integrate SARIMA forecasts into weekly inventory ordering processes.
- Establish a structured promotion testing framework to determine effective promotion types.
- Train frontline staff on upselling strategies to increase average transaction value.

Phase 3 (Weeks 13–26): Optimization and Scaling

- Retrain models with newly collected data to improve accuracy.
- Conduct deeper analysis of promotion performance using expanded samples.
- Extend forecasting horizon to 14–30 days for enhanced financial planning.

7 Conclusion

This project successfully developed and evaluated predictive models that address key operational challenges for the retail store. Using 90 days of daily sales data, the classification and forecasting models demonstrated strong performance, with Random Forest achieving **82.6% accuracy** in predicting high-revenue days and SARIMA obtaining an **RMSE of $900** for 7-day revenue forecasts.

Several consistent insights emerged from the analysis. Weekend revenue was found to be **45% higher** than weekday revenue, a statistically significant and operationally important trend. Customer traffic showed the strongest correlation with revenue ($\mathbf{r = 0.86}$), confirming its role as the primary revenue driver. Meanwhile, average transaction values remained stable, indicating that demand fluctuations—not spending levels—drive daily revenue variability.

The business impact of these findings is substantial. Implementing the recommended strategies could deliver **10–15% reductions in labor costs**, **15–20% revenue increases on targeted weekdays**, and **20–25% reductions in stockouts** through better inventory alignment. These improvements translate directly into operational efficiency, enhanced customer experience, and stronger financial performance.

Next steps include deploying the staffing optimization strategy guided by the classification model, expanding the promotion testing program to gather more robust insights, and continuing to collect data to refine and enhance forecasting accuracy. As additional data becomes available, the store can move toward richer customer-level and product-level analytics that support more sophisticated decision-making.

Overall, this project demonstrates how data-driven approaches can provide measurable value in retail operations, creating a foundation for more advanced analytics capabilities in the future.

8 References

1. Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.
2. James, G., et al. (2021). *An Introduction to Statistical Learning with R* (2nd ed.). Springer.
3. Taylor, S.J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37-45.
4. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proc. KDD*, 785-794.
5. Wright, M.N., & Ziegler, A. (2017). ranger: Fast Random Forests. *J. Stat. Software*, 77(1), 1-17.

9 Appendix: Additional Visualizations and Code Repository

Code Repository: All analysis code is available at:
https://github.com/T-Oni-01/CAP6768-Data-Analytics-Fall-2025-Group-Project/blob/main/CAP6768%20Final%20Project_Completed.R

9.1 Additional Exploratory Data Analysis

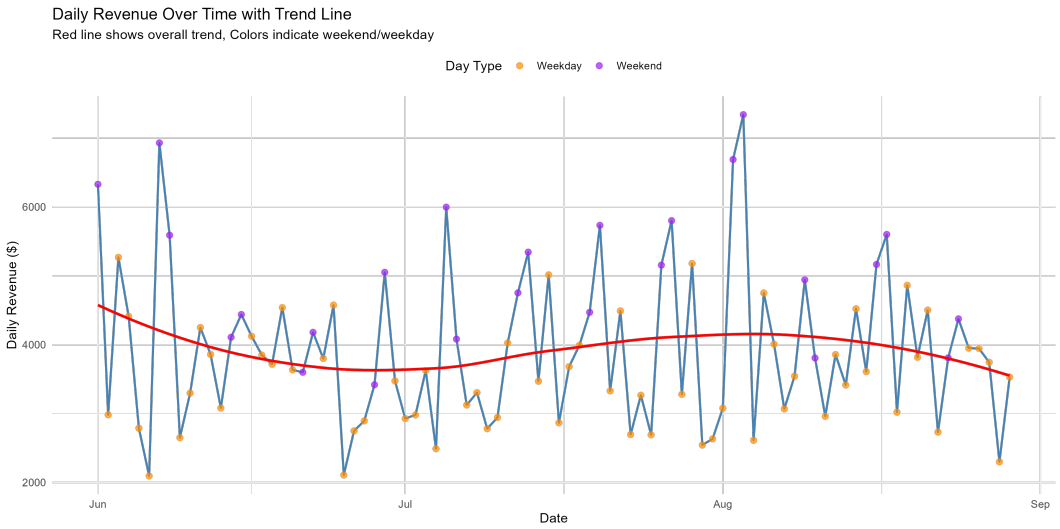

Figure 3: Detailed Revenue Time Series with Trend Analysis

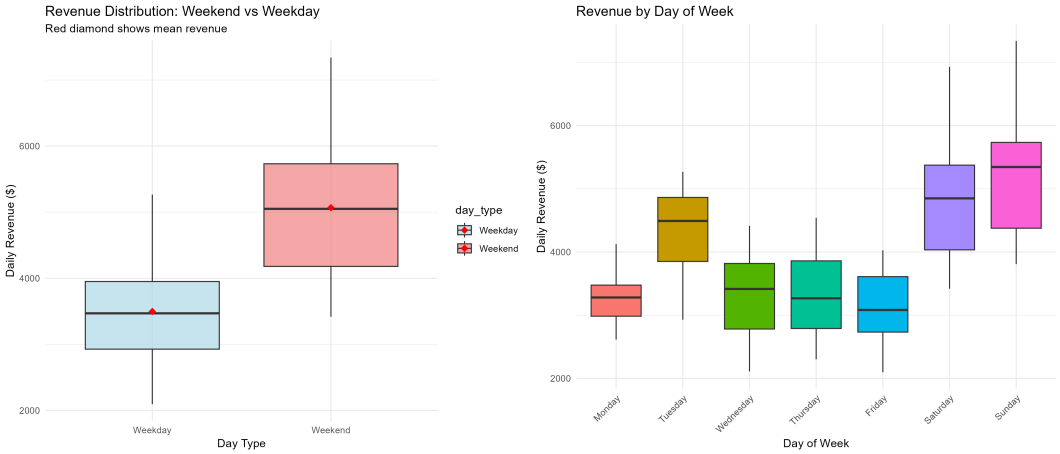

Figure 4: Revenue Distribution by Day Type and Day of Week

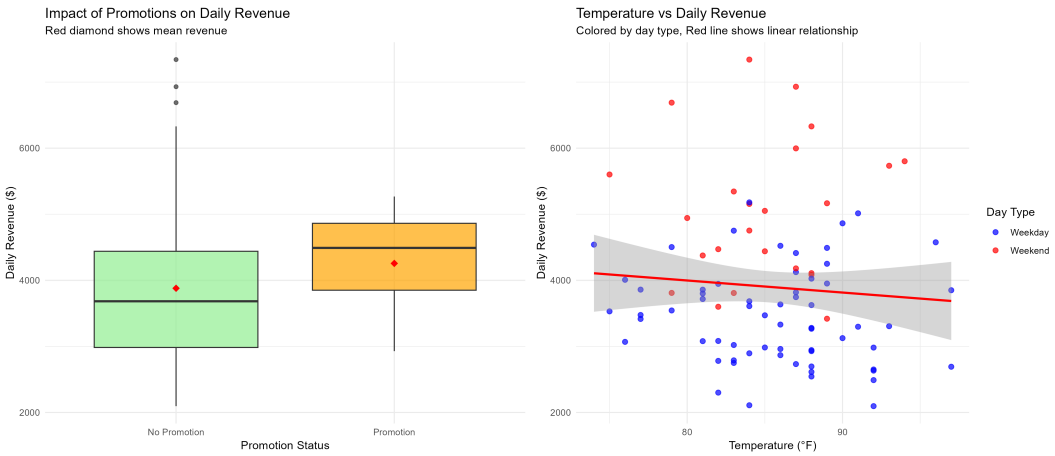

Figure 5: Promotion Impact and Temperature Analysis

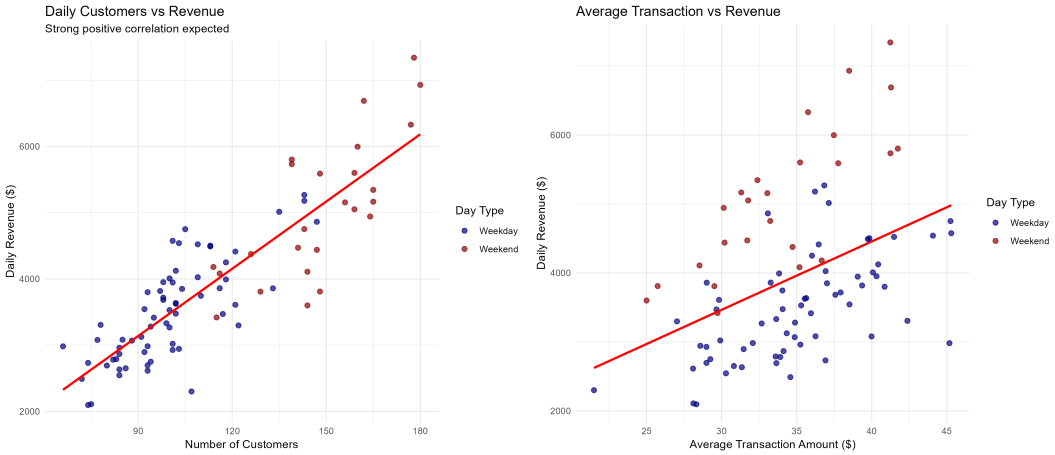

Figure 6: Customer Behavior Analysis

9.2 Additional Classification Model Diagnostics

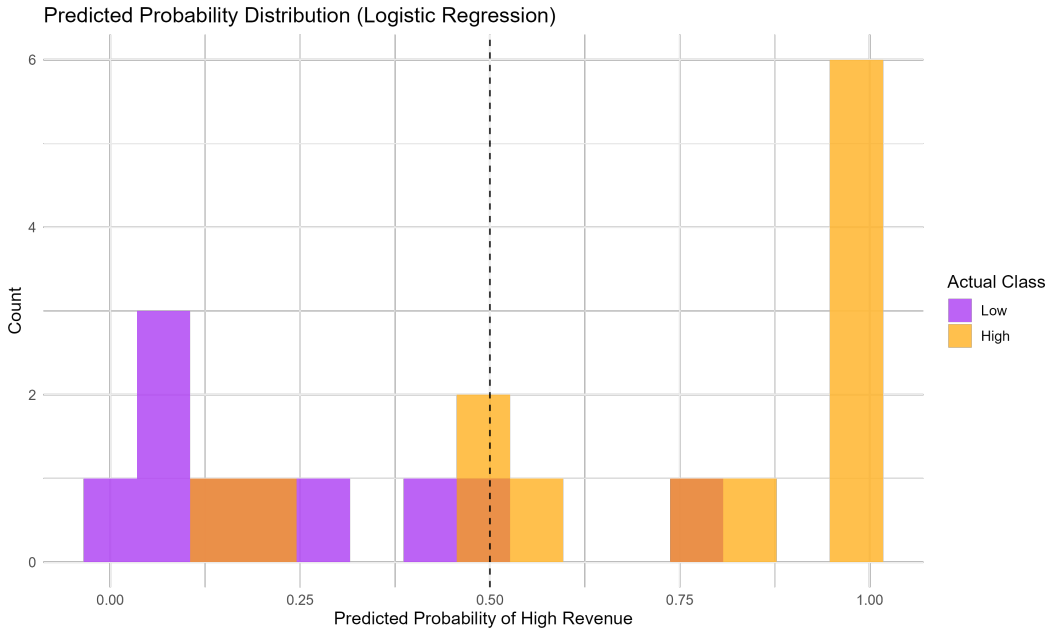

Figure 7: Logistic Regression Probability Distribution

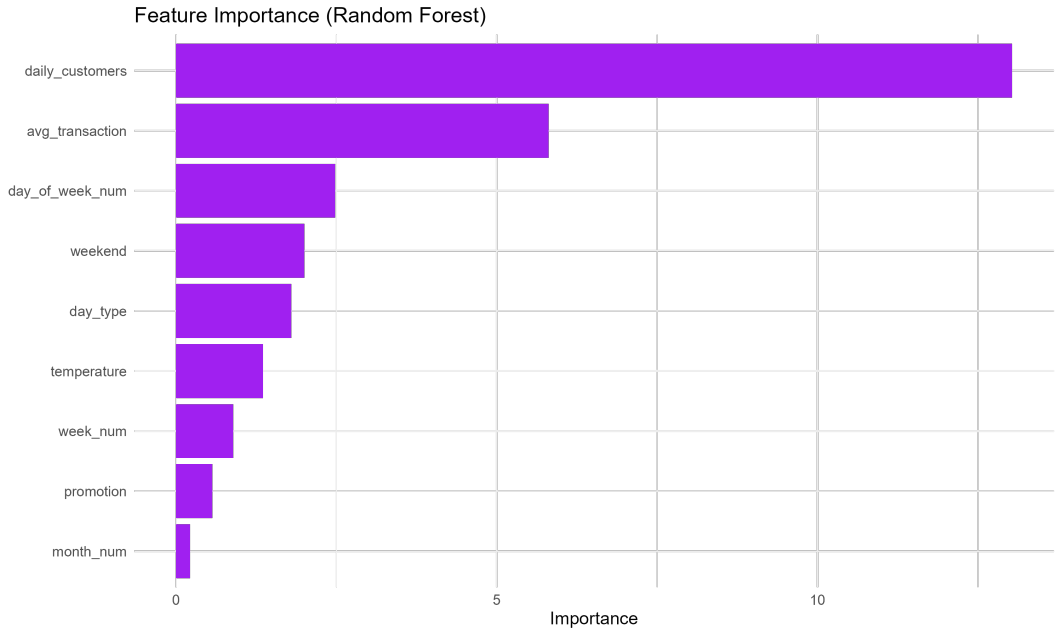

Figure 8: Random Forest Feature Importance Rankings

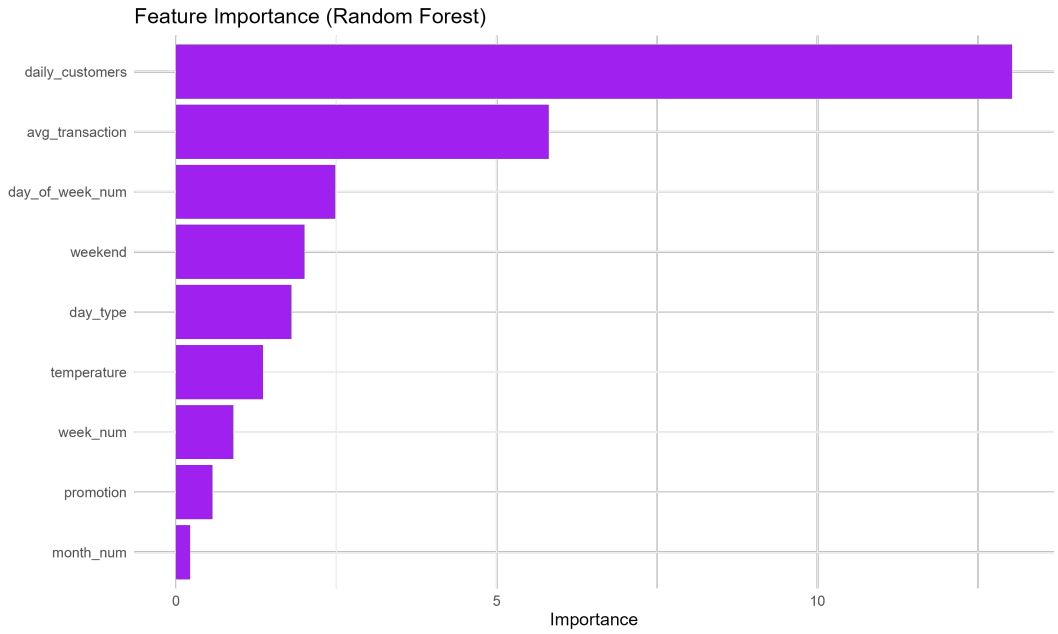

Figure 9: XGBoost ROC Curve and AUC Score

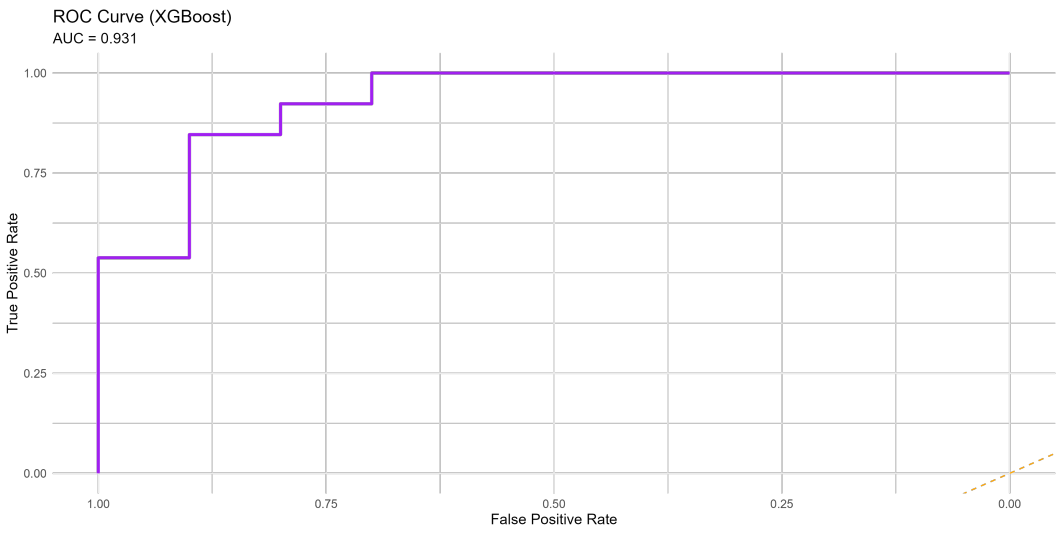

Figure 10: Classification Model Performance Comparison

9.3 Additional Forecasting Model Diagnostics

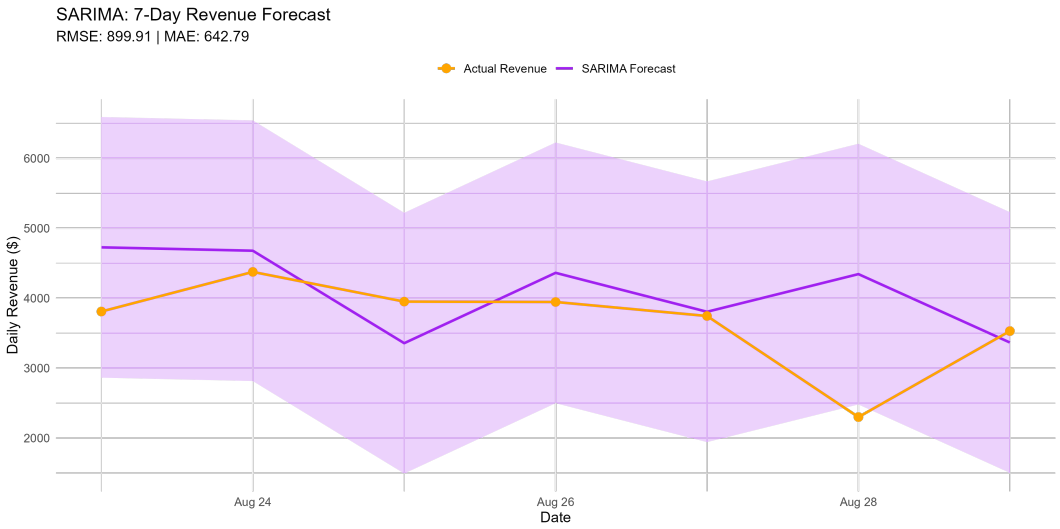

Figure 11: SARIMA Forecast with 95% Confidence Intervals

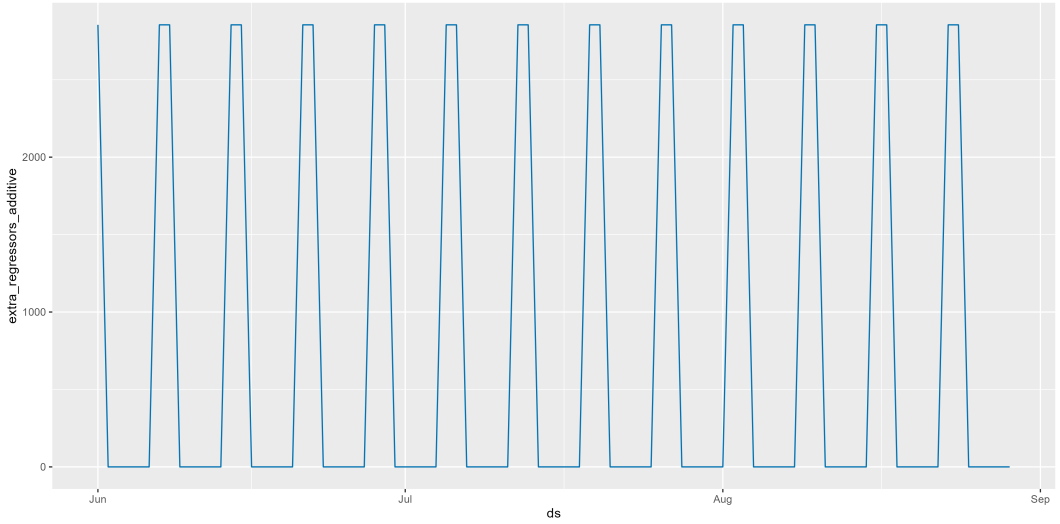

Figure 12: Prophet Forecast with 95% Confidence Intervals

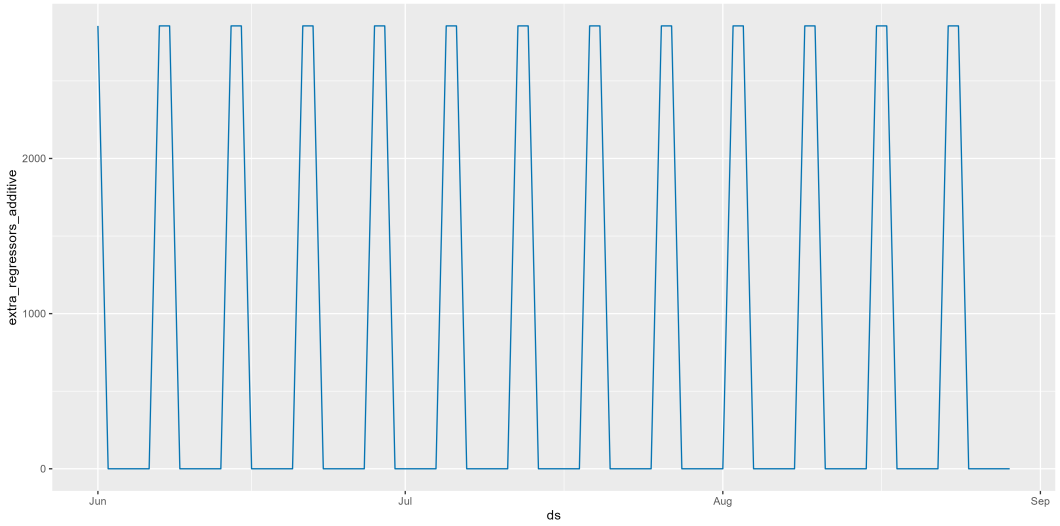

Figure 13: Forecast Model Performance Metrics Comparison

Figure 14: Combined Forecast Comparison: SARIMA vs Prophet