# Group 64 Project Report:

## Sales Optimization - Insomnia Cookies

## ISYE 7406

Robert Mallon (ID:013)
rmallon@gatech.edu

Matt McDowell (ID:156)
mmcdowell31@gatech.edu

Avani Patel (ID:001)
apatel778@gatech.edu

Nicholas Potter (ID:465)
npotter7@gatech.edu

April 18, 2025

# 1 Abstract

Insomnia Cookies is a popular bakery chain known for its freshly baked cookies and desserts with locations nationwide and a robust online ordering platform. Insomnia runs various promotions to drive sales, but the challenge is to properly evaluate these promotions since sales are also affected by seasonal and weather patterns. ++ This paper assesses the effectiveness of ***random forests*** and ***boosting*** to create a prediction model to predict the daily sales of specific store locations. A linear regression model was also created in order to establish a baseline comparison. Additionally, the goal is to look at the feature importance of these models to determine which types of promotions are the most effective. Interaction terms were also created in order to see if the effectiveness of certain promotions varied depending on factors such as day of the week, allowing the models to capture more nuanced relationships between promotional strategies and daily sales performance.

Cross validation was used in order to tune the parameters of these models as well as evaluate the testing error and variance of their predictions. The findings demonstrate that the XGBoost model produced the best predictions and outperformed the other models. It was found that seasonality, weather, and the interaction between the day of the week were the features that impacted the model the most, it was also shown that multiple ongoing promotions do not increase sales. The results highlight the importance of using robust cross validation techniques to assess model performance and tune model parameters, additionally they highlight the learning capability of ensemble methods.

# 2 Introduction/Problem Statement

## 2.1 Introduction

Predicting sales is critical for quick-service restaurants (QSRs). From scheduling labor to preparing ingredients to ordering inventory, an accurate sales forecast is critical, especially when margins in the restaurant industry can be extremely tight. Sales at Insomnia Cookies, a popular bakery chain that specializes in delivering fresh and warm cookies, are heavily influenced by a multitude of factors such as weather and seasonality, making it difficult for their teams to predict sales. Additionally, Insomnia Cookies often offers multiple promotions throughout the week, given a massive percentage of Insomnia's customers are college students who are heavily influenced by discounts and offerings. These promotions can drastically change the sales for a given store.

## 2.2 Problem Statement

Currently, Insomnia uses rolling averages with basic weighting to account for weather and other events. This does not accurately account for the impacts of weather nor does it account for marketing promotions. This results in Insomnia making labor, prep, and inventory

decisions based on often inaccurate forecasts that lead to waste or unmet demand. By building out a more sophisticated machine learning model that can accurately incorporate weather, seasonality, and marketing campaigns into the sales forecast Insomnia will be able to more accurately forecast and therefore make better decisions. Additionally, since there are multiple types of marketing promotions such as free products, free delivery, or double points we decided to evaluate how each type of promotion impacts sales to see if certain promotions work better for certain types of locations or times of year. If there is a significant difference between types of promotions, this could help Insomnia's marketing team determine when to run specific promotions.

# 3 Data Sources and Exploration

In this section, we outline the primary data sources used to build and evaluate our sales prediction models. The goal is to provide a clear understanding of the types of data available, how they are structured, and the role each plays in informing the models. Additionally, we conduct an initial exploratory data analysis (EDA) to examine distributions and potential relationships between variables.

## 3.1 Data Sources

One of our team members is employed at Insomnia Cookies, which allowed us to obtain direct access to their internal data systems. Through this access, we were able to retrieve detailed datasets on daily sales and promotions. While the sales data spans multiple years, the promotions dataset was only available for the past 14 months. To complement this with climate context, we obtained daily weather data using the National Oceanic and Atmospheric Administration (NOAA) API. A custom Python script was developed to scrape weather information from stations located in the same cities as each store. All datasets were then preprocessed to handle missing values and outliers. After being cleaned, they were merged into a single flat table to support modeling and analysis. The final dataset included 1 response variable **total.sales** and 34 predictor variables. For modeling purposes, we focused on the main (**bold**) predictors below.

1. Daily Sales Data: Insomnia Sales Data
    (a) **total.sales**: total daily sales of a given store in $ dollars.
    (b) **location_class**: grouping of stores NY, Philly, St Louis and College towns.
2. Daily Promotions Data : Insomnia Marketing & Loyalty Platform (Punchh) Data
    (a) **total_promotions**: Total number of active promotions on a given day for each store.
    (b) **promotion_types**: type of promotion (expanded to binary columns)

(c) **promotion_channels**: channel to access the promotion, ex. "in-store" (expanded into binary columns)

3. Daily Weather Data : NOAA Weather Service

   (a) **prcp**: Daily precipitation amount (in inches).

   (b) **snow**: Daily snowfall (in inches).

   (c) **tmax**: Maximum temperature recorded for the day.

   (d) **tmin**: Minimum temperature recorded for the day.

4. Seasonality Data:

   (a) **Annual_Seasonality**: A feature that captures recurring patterns or fluctuations in sales observed over the course of a year.

   (b) **Weekly_Seasonality**: A feature that captures consistent sales trends within a week.

## 3.2 Exploratory Analysis

### 3.2.1 Daily Sales Exploration

For the initial exploratory analysis, histograms were created to visualize and compare the distributions of daily sales data across the different locations. (Figure 7) Based on these histograms, the sales data appears to roughly follow a normal distribution, though with a slight right skew. Additionally, we created bar plots to understand how the average sales were impacted by the total number of promotions. (Figure 2) In general it was found that the highest daily sales occurred when there was only one promotion in place. This could indicate that stacking multiple promotions at the same time doesn't yield any significant benefits, but this will need to be confirmed by further analysis.

### 3.2.2 Sales vs. Weather Effects

To gain a comprehensive understanding of factors influencing Insomnia's sales, we extended our analysis to explore the relationship between daily sales and weather conditions, specifically temperature and snowfall. Understanding how these factors influence sales can provide valuable information for forecasting and decision making. Our exploration of sales versus weather data revealed several notable trends. A scatter plot of sales versus snowfall indicates a general negative correlation, with total sales tending to decrease as snowfall increases. This suggests that heavy snowfall may deter customers, leading to lower sales, as depicted in Figure 3. Figure 4 visualizes sales distribution across temperature ranges and snow presence as box plots. The box plots illustrate that across nearly all temperature ranges, the median sales value is lower when snow is present. Moreover, Figure 5 illustrates that sales generally increase within moderate temperatures (40°F-70°F) but are depressed at extreme ends.

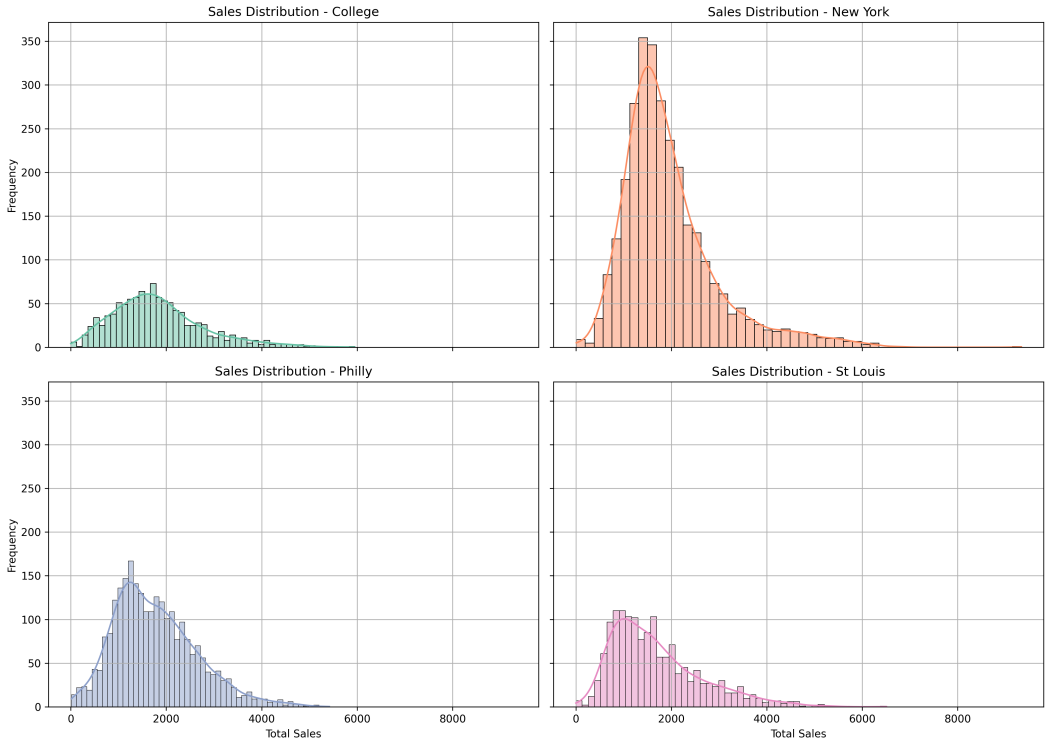


Figure 1: Sales Distribution Histograms by location_class

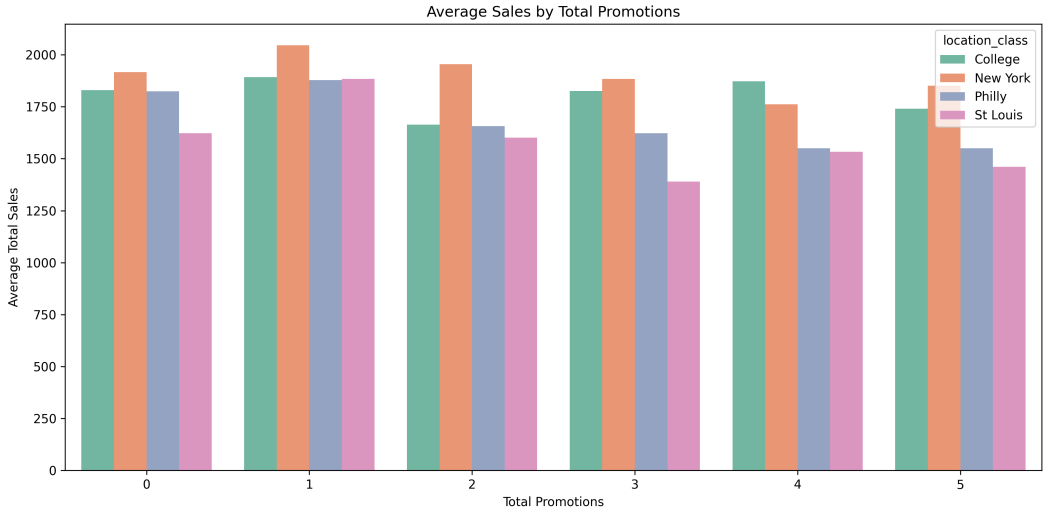


Figure 2: Average Sales by Total Promotions

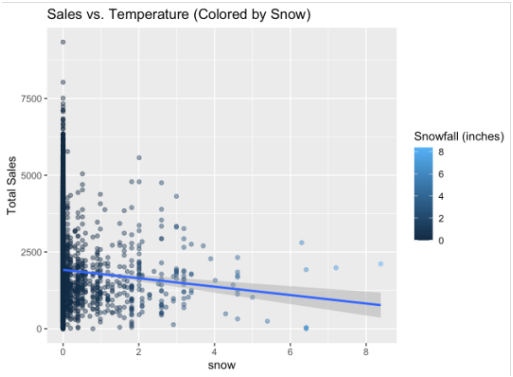


Figure 3: Sales vs. Snowfall

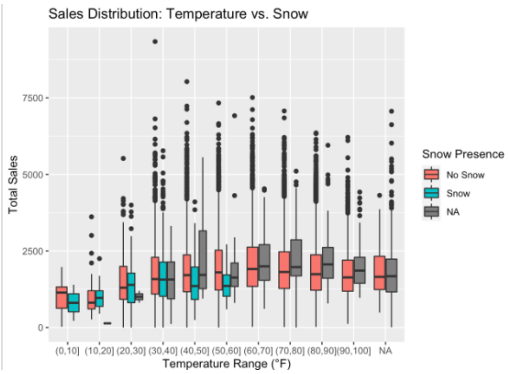


Figure 4: Sales vs. Temp (Snow)

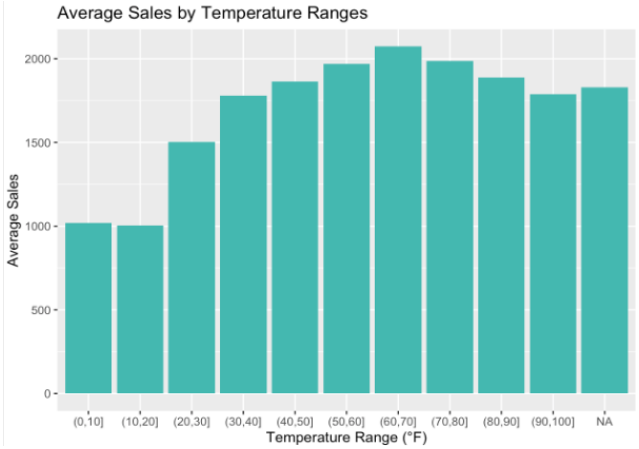


Figure 5: Sales Temp Range

Figure 6: EDA of Sales and Weather Data.

### 3.2.3 Seasonality: Weekly/Annual

In the context of Insomnia's daily sales data, we wanted to explore whether the data exhibited predictable patterns associated with certain days of the week or certain days of the year. Identifying such patterns can help us better understand sales behavior and ultimately lead to more accurate prediction models. We used **MTSL (Multiple Time Scale Linear) decomposition**, a time series decomposition method, which has the ability to handling multiple seasonalities (weekly, annual) simultaneously. Below is the general form of the MTSL decomposition (Equation 1).

$$Y_t = T_t + S_{annual} + S_{weekly} + \varepsilon_t \tag{1}$$

In order to perform this decomposition we looked at the entire sales dataset (including sales data which preceded our promotions dataset). We then extracted and examined the seasonal components for each store, $S_{annaul}$ and $S_{weekly}$ , to determine if meaningful seasonal patterns existed.

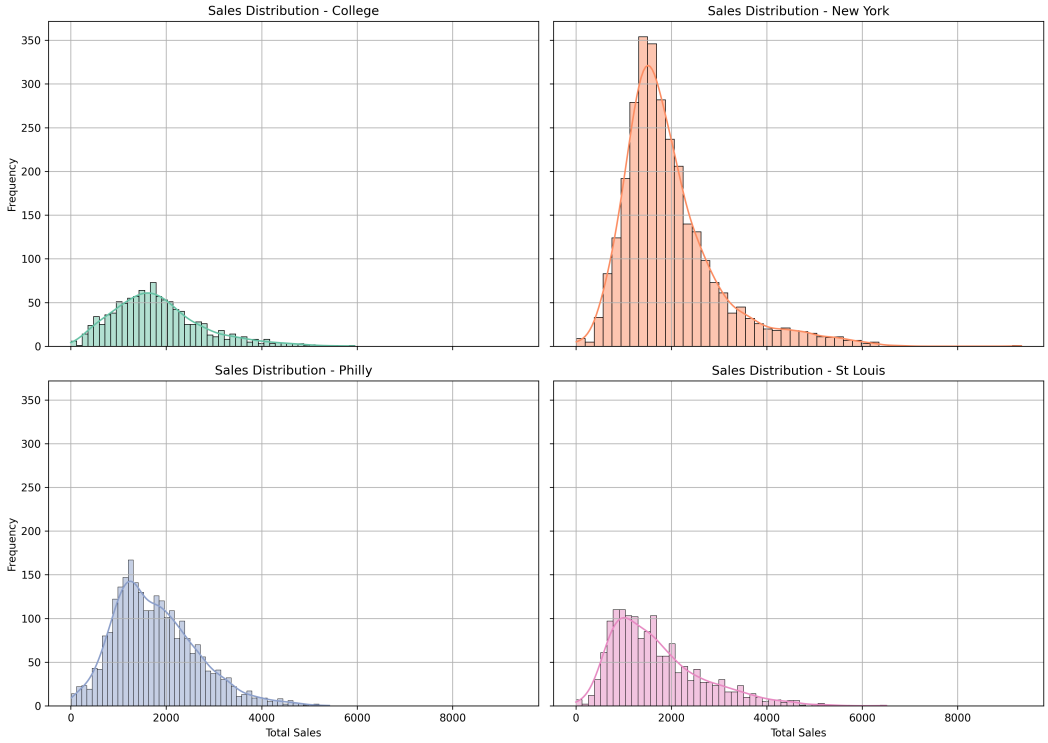


Figure 7: Sales Distribution Histograms by location_class

In general it was found that there were distinct seasonal patterns to the sales data. The weekly sales patterns showed significantly **higher average sales observed on weekends.** While the annual patterns showed reveals **spring and autumn peaks** (March/April and September/October), a **summer dip**(June/July), and a **December dip**. These patterns can be observed for a single store in Figure 8.

Ultimately, we decided to include these seasonal components as additional features in our prediction models, since the daily sales clearly exhibit seasonality that could improve accuracy.

### 3.2.4 Feature Correlations

Finally, to give some initial insight on the predictive potential of this dataset we created a correlation matrix (Figure 9). Based on the correlation matrix, there are many features with "weak" correlations (< 0.5) to but only two (weekly and annual seasonality) that are "strong" (> 0.5) predictors. Therefore this dataset appears to be a good candidate for ensemble models specifically ***boosting*** where multiple weak predictors are combined to make a single strong classifier.

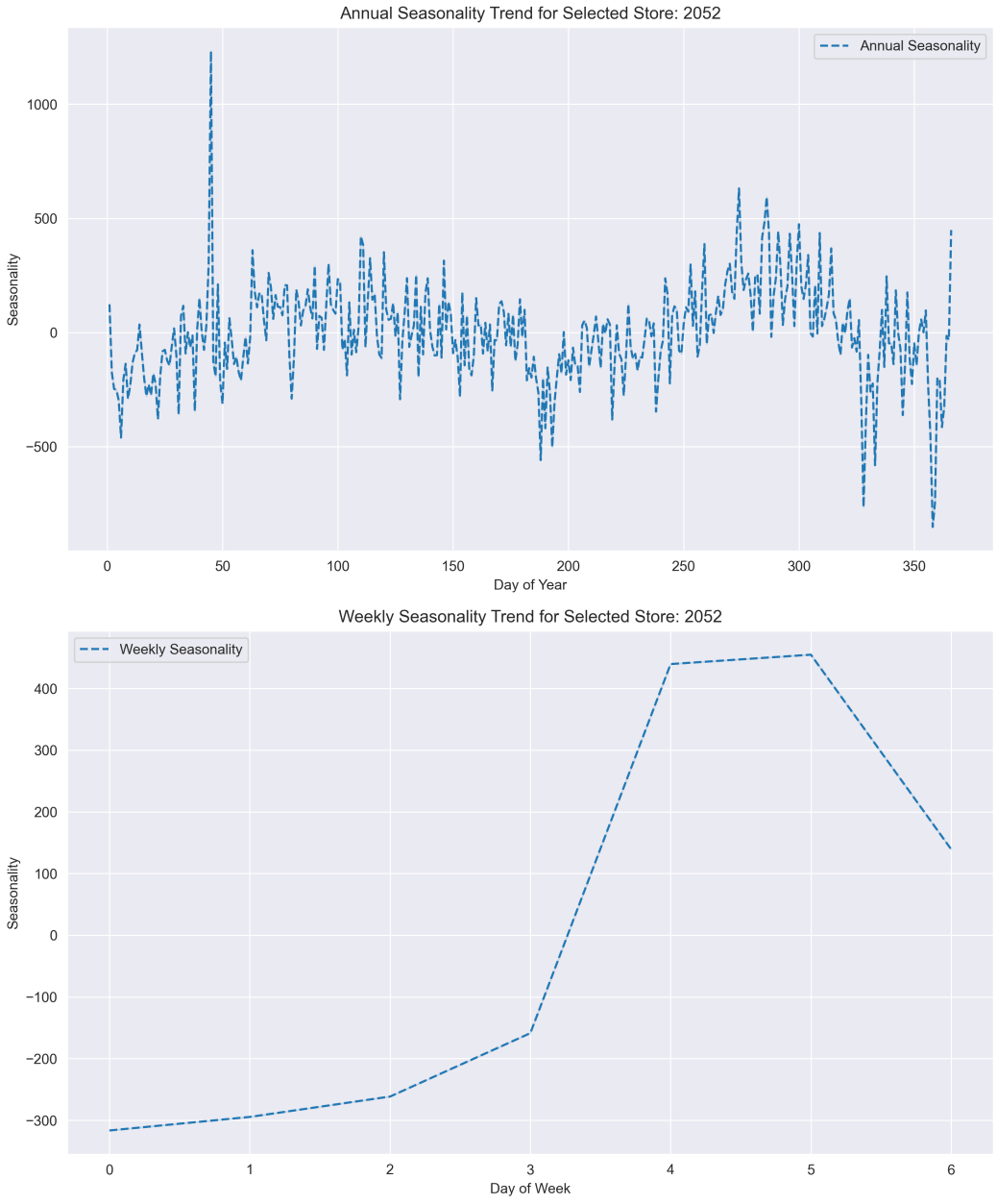


Figure 8: Seasonality Trends for Store 2052

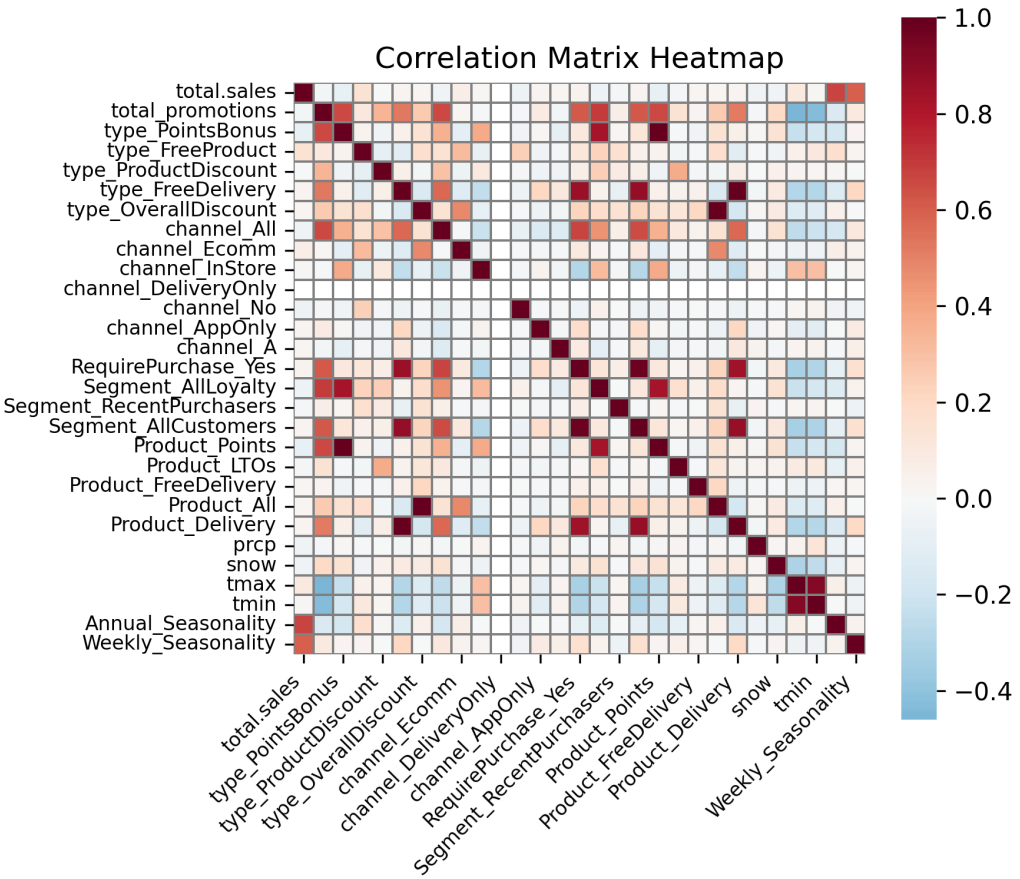


Figure 9: Correlation Matrix: Insomnia Daily Sales

# 4 Methodology

In this section, we will describe the different methods used to create/generate the ensemble sales prediction models as well the baseline model. Additionally, we will briefly summarize their initial performance on the testing data, and where applicable, we will optimize model parameters using RMSE/Explained Variance.

For this section, the data was split (80/20) for training/testing and the same training and testing data was used for each model. When model parameters where optimized k-fold cross validation was used.

## 4.1 Linear Regression

Multiple Linear Regression (MLR) was used as a baseline modeling approach to predict daily total sales using several key predictors, including weekly and annual seasonality, temperature (tmax, tmin), snow, number of promotions, and day of the week. These predictors were selected based on their theoretical and observed relevance to retail sales trends. The dataset was divided into four distinct store location groups: New York City, Philadelphia, St. Louis, and College towns, to account for potential geographic and demographic differences in customer behavior. Separate MLR models were created for each group to evaluate model performance across different regions. Linear regression models were fitted using the lm() function in R, and performance was evaluated using metrics such as mean error rate, error variability, and explained variance (R2). Although the MLR approach provided simplicity, it exhibited limitations in accurately capturing sales patterns. For example, college town stores showed the best performance, with the highest explained variance (79.56%) and the lowest error variance ($430.25). In contrast, the New York City model struggled significantly, with an average error rate of $662.29 and a low explained variance of 30.96%, indicating that linear models were not well-suited to model the complexity in that location.

Among predictors, weekly and annual seasonality consistently emerged as important variables, though their impact varied across locations. For example, temperature had a notable effect in college towns, possibly due to student-driven seasonal behavior. However, promotions had limited impact across all regions, suggesting poor targeting or insufficient variation in promotional strategies.

While cross-validation is typically employed to improve model generalization, it was not prioritized here, as the MLR models served mainly to benchmark the effectiveness of more advanced techniques. Given the relatively poor predictive performance in urban areas, these results justified transitioning to nonlinear models like Random Forest and XGBoost, which are better suited to handling feature interactions and complex dependencies.

## 4.2 Random Forest

Random Forest is an ensemble learning method that uses decision trees with a random subset of features and bagging to decrease the variance in predictions. The key idea in

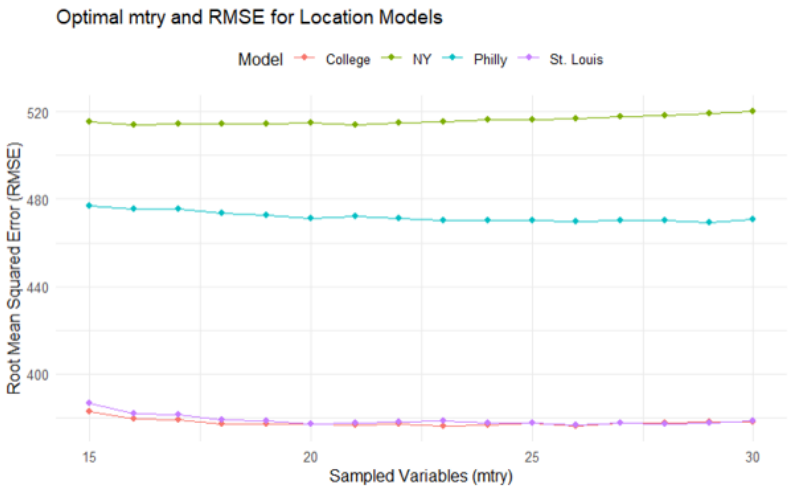


Figure 10: Random Forest Model Tuning

bagging is to use **bootstrapped** (randomly sampled w/replacement) data to make each tree. The final prediction is made by **averaging** the outputs of all individual trees, which helps reduce variance and improve accuracy. The use of a random subset of features introduces diversity among the trees, leading to more robust and generalizable predictions.

A Random Forest model was included with the expected benefits of capturing nonlinear relationships and providing feature importance insights. This model fits with the use case of needing to model complex interactions potentially occurring between factors like promotions, weather, and weekly seasonality as well as providing feature importance scores to help understand the relative influence of variables. Previous research for retail sales forecasting has had success with a modeling approach including Random Forest, with Ma's 2024 analysis of supermarket sales finding a Random Forest model to be the most precise with an R-squared score of 99.27%. [3, 2]

Similar to the baseline linear regression model, a separate Random Forest model was created for each location type (College, St Louis, Philly, New York). Each of these models was created with 500 trees, which was a size that balanced a sufficiently large number of trees for an effective ensemble method and computational efficiency. To determine the optimal *mtry* parameter, which determines the number of variables sampled at each split in each tree, a 5-fold cross validation grid search was completed. The R package *caret* was utilized for this grid search and model creation. The *mtry* parameter per model was chosen based on the lowest corresponding Root Mean Squared Error (RMSE) value. Variation in RMSE was limited within each model between sampled variable numbers. The resulting number of selected predictors for the *mtry* parameter were 26 for the College and St Louis models, 30 for Philly, and 16 for the New York model. See the line graph Figure 10 for more details.

### 4.3 Boosting: XGBoost

**XGBoost (Extreme Gradient Boosting)** is a boosting ensemble method that builds decision trees **sequentially**, with each new tree learning to correct the residual errors of the previous trees. Unlike Random Forest, which builds trees independently, XGBoost focuses on improving performance by giving more weight to observations that were previously predicted poorly. The final prediction is obtained by taking a **weighted average** of the outputs from all trees, effectively minimizing the training error.

XGBoost was chosen due to the presence of many "weak" predictors and its ability to improve the prediction quality of the model using of the complex relationship between these predictors across many trees. Additionally, SHAP Values (Shapley Additive Explanations) can help explain how each feature contributes to a prediction from the XGBoost model. This will be further discussed in the ***results*** section. Prior research in retail sales forecasting has demonstrated the effectiveness of XGBoost. In a 2021 study by Dariu and Shilong, XGBoost significantly outperformed other models, such as Ridge Regression and Multiple Linear Regression, in predicting daily sales for Walmart retail goods. [1]

Similar to the other models, an XGBoost model was developed for each location type (College, St. Louis, Philadelphia, and New York). For each model, the **number of estimators** ,which corresponds to the total number of decision trees built sequentially, was set to the square root of the number of observations in the dataset. The learning rate parameter (**eta**) which controls how much influence each new tree has on the final prediction. In this analysis, we applied K-fold (k = 10) cross-validation to optimize the learning rate.The eta parameter per model was chosen based on the highest corresponding explained variance value. The optimal value for the learning rate was found to be 0.125 for Colleges, 0.145 for St. Louis, 0.185 for Philadelphia, and 0.145 for New York. See the line graph in Figure 11 to see how the different learning rates affect model performance on the testing datasets.

## 5 Results/Analysis

This section contains a brief comparative analysis of the different models and their performance on dataset. Initially, the prediction models were evaluated on one split of of the data while the model parameters were selected using cross validation.

For the baseline model, linear regression, one split of the data was evaluated partly due to it poor performance. For the ensemble models in order to provide a more thorough comparison 100 Monte Carlo cross validation simulations were used to evaluate the testing errors and variances of each model. To evaluate model performance, we calculated three key metrics: Mean Absolute Error (MAE), Error Variance, and Explained Variance. Average Error (MAE) was calculated as the mean of the absolute differences between predicted and actual test data sales values.

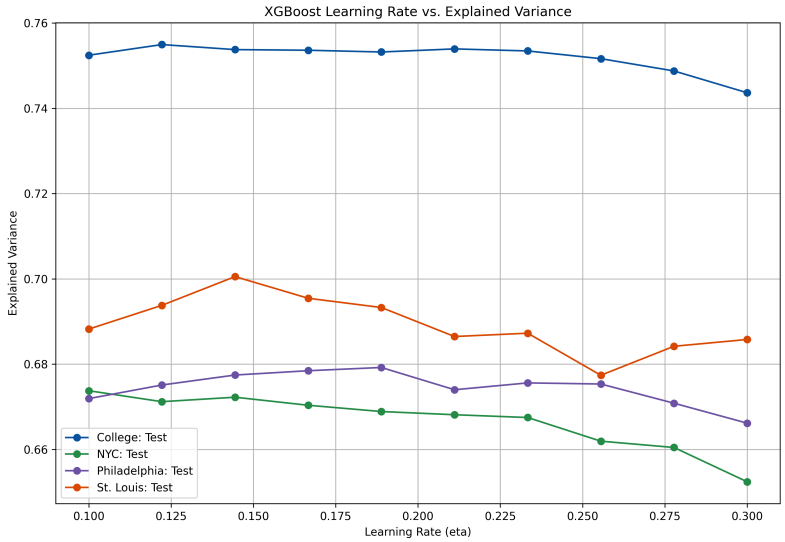


Figure 11: XGBoost Model Learning Rate vs Explained Variance

## 5.1 Linear Regression

### 5.1.1 Model Results

Multiple Linear Regression model was trained using the Training dataset and then used to generate predictions on the Test dataset. College locations showing the highest Average Explained Variance (79.56%) and lowest Average Error Rate ($373.01). New York City exhibits the poorest model fit, indicated by the lowest Average Explained Variance (30.96%) and highest Average Error Rate ($662.29). St. Louis and Philadelphia demonstrate moderate performance, suggesting location-specific factors significantly influence the model's predictive power. Given the insufficient baseline performance, cross-validation was not prioritized. The results are shown in

**Linear Regression Model Results**

| Store Location | Average Error | Average Error Variance | Explained Variance |
|---|---|---|---|
| College | $373.01 | 430.25 | 79.56 % |
| St. Louis | $529.89 | 639.20 | 48.39 % |
| Philadelphia | $662.29 | 899.69 | 30.96 % |
| New York City | $577.10 | 641.57 | 55.24 % |

Table 1: Multiple Linear Regression Model Results

### 5.1.2 Feature Importance

To determine feature importance in the MLR models, we examined the standardized coefficients of the predictor variables for each store location. Standardized coefficients allow for a direct comparison of the relative impact of each predictor, as they account for differences in scale. This approach is conceptually similar to SHAP values in more complex models, as

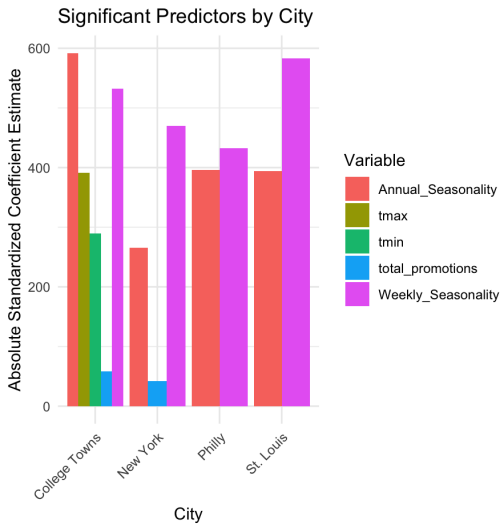


Figure 12: Absolute Standardized Coefficient Estimates of Significant Predictors by City.

it quantifies each feature's contribution to the model's output. Our analysis revealed distinct patterns of importance of features in the four locations.

- College Towns: Annual seasonality, temperature (tmax and tmin), total promotions, and weekly seasonality are the key predictors.
- New York: Annual seasonality, total promotions, and weekly seasonality are key predictors.
- Philadelphia: The model is influenced by Annual and Weekly Seasonality.
- St. Louis: The model is influenced by Annual and Weekly Seasonality.

These patterns emphasize the importance of seasonality and weather effects. The standardized coefficients serve as a measure of the influence of each variable on the sales prediction, as illustrated in Figure 12.

## 5.2 Random Forest

### 5.2.1 Model Results

The Random Forest model was trained using the Training dataset and then used to generate predictions on the Test dataset. College locations reported the highest Explained Variance (85.41%) with St. Louis having the next highest Explained Variance (80.22%). The College and St Louis models also reported the lowest Average Error values and highest Average Error Variances. The results are shown in Table 2.

| Random Forest Model Results (500 Trees) | | | | |
|---|---|---|---|---|
| **Store Location** | **Selected Predictors** | **Average Error** | **Average Error Variance** | **Explained Variance** |
| College | 26 | $309.21 | 795.21 | 85.41 % |
| St. Louis | 26 | $299.20 | 685.47 | 80.22 % |
| Philadelphia | 30 | $366.96 | 415.8 | 75.27 % |
| New York City | 16 | $408.32 | 618.82 | 70.96 % |

Table 2: Random Forest Model Results

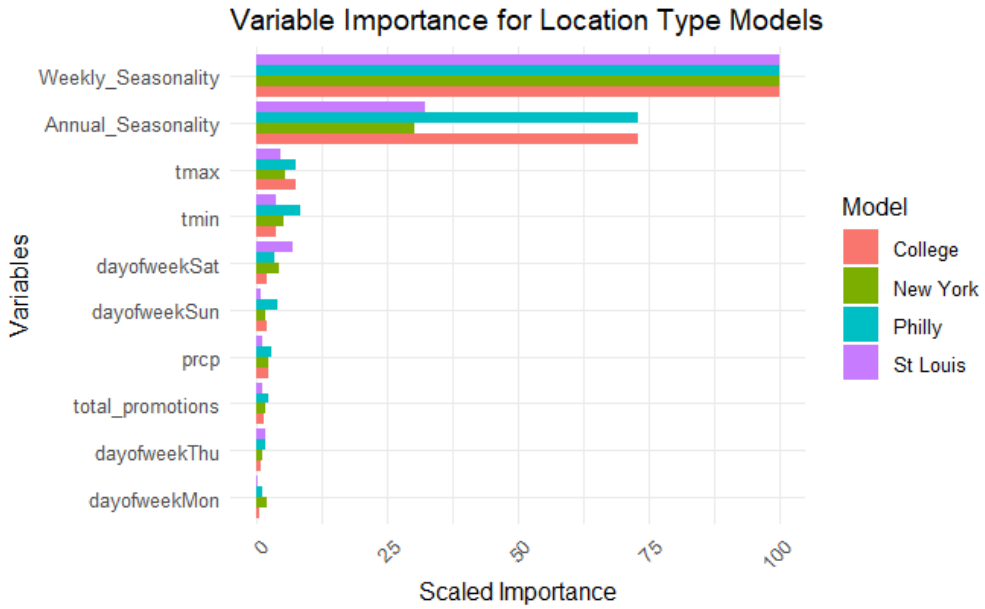


Figure 13: Random Forest Feature Importance

### 5.2.2 Feature Importance

For the Random Forest models, feature importance was assessed by calculating scaled importance based on the reduction in mean squared error associated with each variable when included in decision tree splits. Greater reductions in mean squared error result in higher variable importance values which were normalized by standard error. Weekly and annual seasonality were by far the strongest variables in terms of scaled variable importance. Variation between models was present for annual seasonality importance with the St Louis and New York locations having much lower importance than the College and Philly models. Shown in Figure 13, weather-related variables, day of the week, and total number of promotions were among the most important variables in the models, but were relatively weak in importance compared to seasonality.

To specifically visualize the effects of the Total Promotions variable while keeping other variables constant, Partial Dependence Plots (PDPs) are shown in Figure 14. Different interactions between promotions and total sales are visible across the location models, with the New York and Philly models having roughly similar patterns of negative relationships with sales for 0-3 promotions and maximum sales values at 5 promotions. The St Louis and

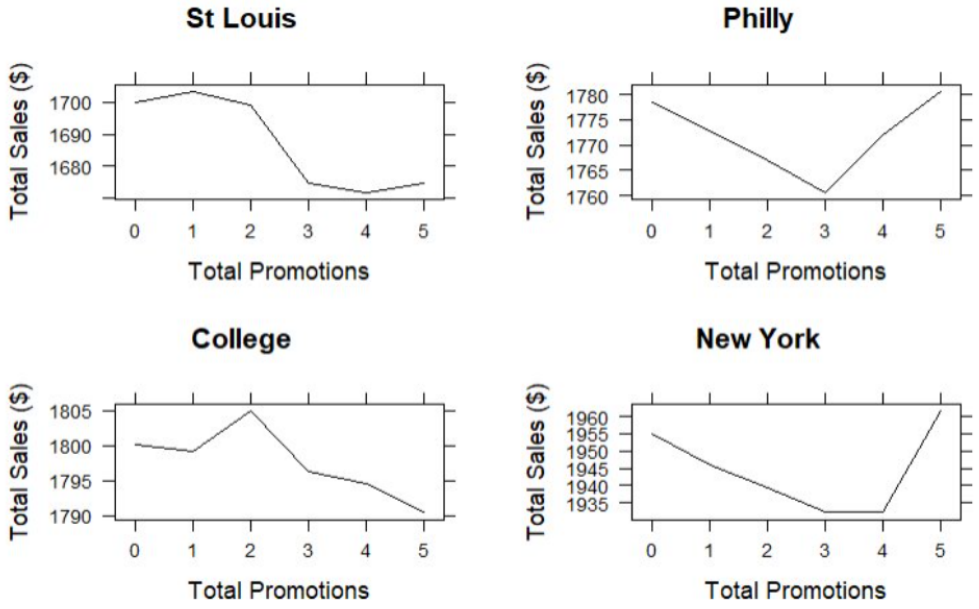


Figure 14: Random Forest Partial Dependence Plots for Promotions

College models exhibit different behavior with maximum sales values at 1 and 2 promotions, respectively.

## 5.3 Boosting: XGBoost

### 5.3.1 Model Results

The XGBoost model was trained using the Training dataset and then used to generate predictions on the Test dataset. College locations reported the highest Explained Variance (95.69%) with St. Louis having the next highest Explained Variance (95.10%). The College and St Louis models also reported the lowest Average Error values and highest Average Error Variances.

**XGBoost Model Results**

| Store Location | Learning Rate (eta) | Average Error | Average Error Variance | Explained Variance |
|---|---|---|---|---|
| College | 0.125 | $71.57 | 123.39 | 95.69 % |
| St. Louis | 0.145 | $72.59 | 76.38 | 95.10 % |
| Philadelphia | 0.185 | $122.55 | 72.67 | 92.03% |
| New York City | 0.145 | $109. 91 | 112.94 | 90.46% |

Table 3: XGBoost Model Results

### 5.3.2 Feature Importance

SHAP (Shapley Additive Explanations) values were used to interpret the XGBoost model by quantifying the contribution of each feature to individual predictions, based on fair allocation principles from game theory. A basic equation of how SHAP values work is shown here Equation 2.

$$\hat{y} = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{2}$$

Where:

- $\hat{y}$ is the model prediction for a given instance.
- $\phi_0$ is the average model prediction (the base value).
- $\phi_i$ is the SHAP value for feature $i$.
- $M$ is the total number of input features.

The analysis revealed that weather effects and seasonality had the strongest impact on the model's output. Other significant drivers included total promotions, customer loyalty segment, day of the week, and the interaction between promotions and weekdays. See Figure 15 for a breakdown of these impacts for each model.

Additionally, SHAP Waterfall Charts (Figure 16) are to shown below to help visualize the impact of model features on single predictions. These charts demonstrate how individual SHAP values contribute to the final prediction, and they always sum to the difference between the current prediction and the model's average prediction.

## 6 Conclusions

In conclusion, this study found that machine learning models, particularly XGBoost, offer significant improvements over traditional regression techniques in predicting Insomnia Cookies sales performance. Although regression analysis exhibited high average error rates and limited explanatory power, ensemble-based approaches such as Random Forests and XGBoost demonstrated stronger predictive capabilities, with XGBoost emerging as the most effective model despite some variability in error. Another significant finding of this analysis is how much XGBoost improved from the model's original performance on the testing data set. It improved from its performance on one split of the data by ~25% after running multiple Monte Carlo simulations. This can be attributed to the fact that by allowing XGBoost to see many partitions of the data it was better able to improve upon its weak learners effectively. In contrast, **Random Forest** showed weaker performance, and its its improvement was more modest compared to XGBoost.

The analysis also highlighted the critical role of external factors, including seasonality, day-of-week promotional timing, and weather, in shaping sales results. Targeted promotions

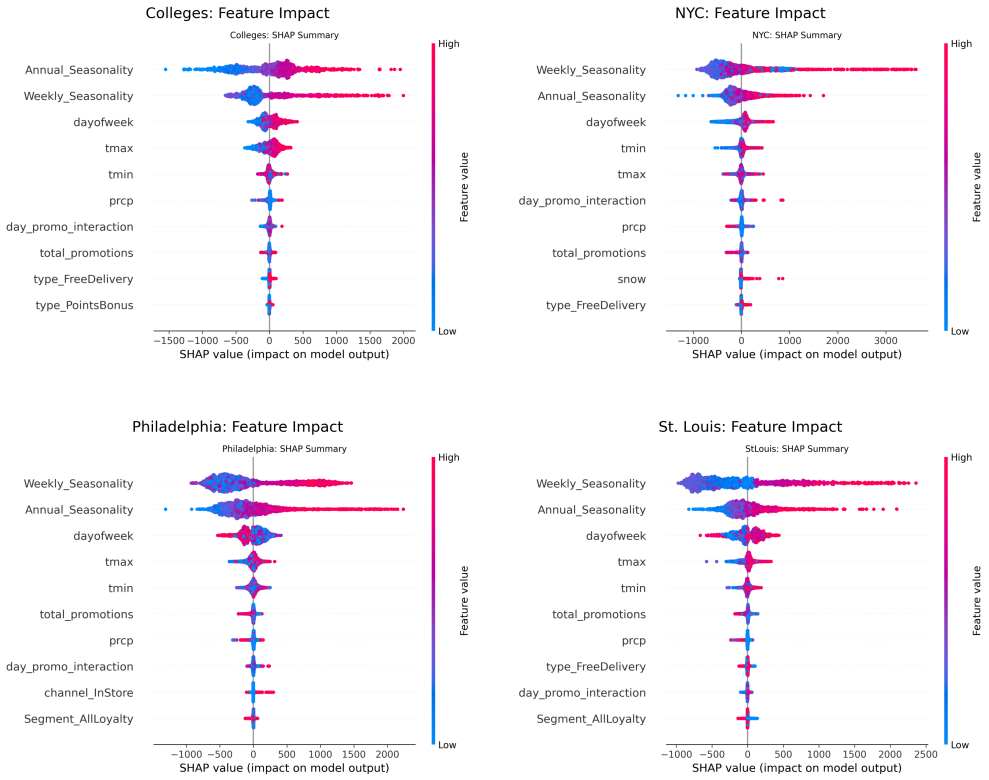


Figure 15: SHAP Feature Plots

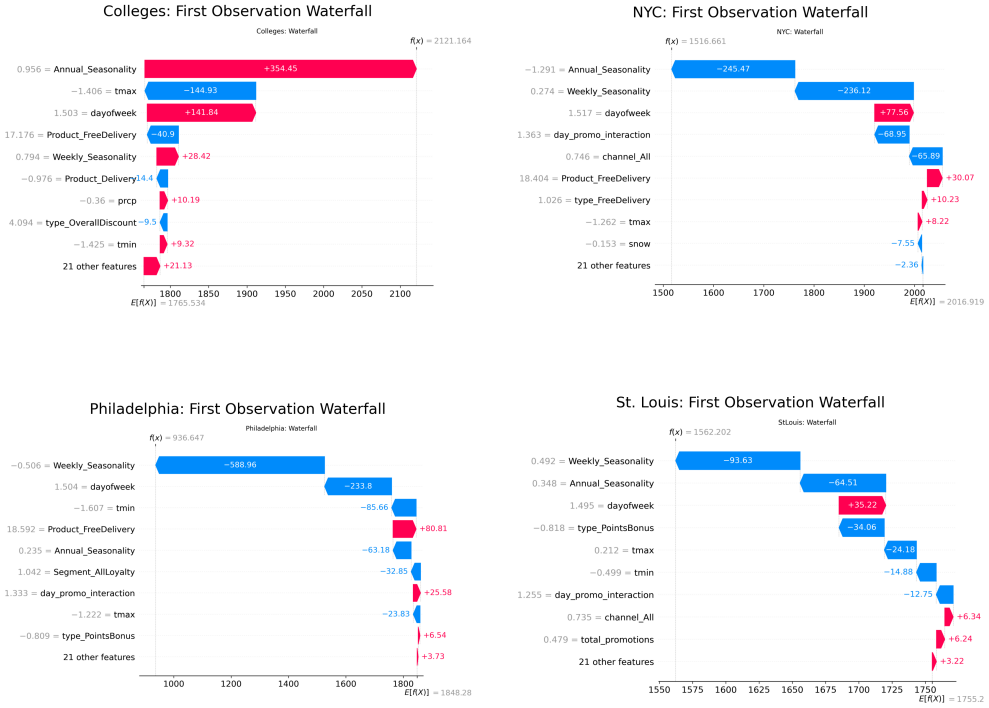


Figure 16: SHAP Waterfall Plots

were shown to shift customer demand from peak days to lighter days, and the optimal timing of these promotions varied between markets, with distinct preferences observed in college towns and urban centers like New York. In addition, weather patterns were found to significantly influence sales, reinforcing the value of weather-adaptive marketing strategies, particularly on days with rain or snow. Interestingly, the data suggested that offering multiple simultaneous promotions diluted effectiveness, highlighting the importance of clear, singular promotional messaging. These findings underscore the value of using advanced predictive models and data-driven insights to guide promotional planning, allowing more precise, efficient, and ROI-focused marketing strategies.

This initial study provides our team with confidence that with additional time, data, and resources, our team could continue to improve the XGBoost model and help the Insomnia marketing team optimize their promotions and obtain a significantly greater Return on Investment.

## 6.1 Lessons Learned

Working with limited and newly available data, particularly time series and promotional data, posed challenges in identifying clear patterns and building predictive models. The lack of key variables like paid media and SEO, along with time constraints, limited the depth of analysis and ability to explore alternative modeling techniques. Additionally, differences in modeling approaches among team members highlighted the need for a consistent evaluation framework and shared datasets to ensure reliable model comparisons. We learned that initial models may not always yield expected insights, particularly regarding promotion effectiveness, highlighting the need for iterative refinement, incorporation of additional variables, and experimentation with diverse modeling strategies. Employing cross-validation techniques proved essential for both model selection and performance evaluation. While the XGBoost model had substantially improved predictive accuracy, there remains room for enhancement. Future work could focus on fine-tuning model parameters and exploring alternative ensemble methods to further reduce error rates and strengthen overall model performance.

## References


[1] Xie dairu and Zhang Shilong. “Machine Learning Model for Sales Forecasting by Using XGBoost”. In: *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. 2021, pp. 480–483. DOI: 10.1109/ICCECE51280.2021.9342304.

[2] Ruiyun Kang. “Sales Prediction of Big Mart based on Linear Regression, Random Forest, and Gradient Boosting”. In: *Advances in Economics, Management and Political Sciences,17,200-207* (2023).

[3] Yining Ma. “Utilizing machine learning for sales forecasting in urban supermarkets”. In: *Applied and Computational Engineering,45,278-285* (2024).

# A Appendix A: Code

```
See Supplemental Project Code folder.
```

Listing 1: Code