Sales Optimization through Data Driven Insights – Insomnia Cookie

Group 64: Avani Patel, Matthew McDowell, Nicholas Potter, Robert Mallon

Georgia Tech

About Insomnia Cookies

- A popular bakery chain known for its freshly baked cookies and desserts
- Strong brand presence with locations nationwide (300+ stores) and a robust online ordering platform (app & website)
- Relies heavily on promotions to drive sales and customer loyalty

The Challenge

- Sales fluctuate due to promotions, weather, and weekly patterns.
- Promotional decisions lack granular insights into promotion type, timing, weather, and seasonality.
- Insomnia Cookies faces the challenge of optimizing its promotional strategies to maximize revenue and ROI.
- Lack of data-driven insights may lead to ineffective promotions. Suboptimal strategies = Reduced revenue, missed opportunities, inefficient marketing.

Georgia Tech

Opportunity

Data-Driven Promotions
•Use analytics to **target promotions effectively**, maximizing **sales and ROI**.

Uncover Key Sales Drivers
•Analyze **weather, seasonality, and timing** to see if they influence sales.

Smarter Resource Allocation
•Optimize **marketing spend** by understanding how external factors impact sales.

Competitive Advantage
•Move beyond **intuition-based decisions** to **data-backed sales strategies**.

Promote Smarter, Not Harder!

Key Question
How can Insomnia Cookies optimize promotions & sales strategies using data-driven insights to drive sustainable revenue growth?

Georgia Tech

Methodology

Unlocking Cookie Code - Our Data Driven Approach

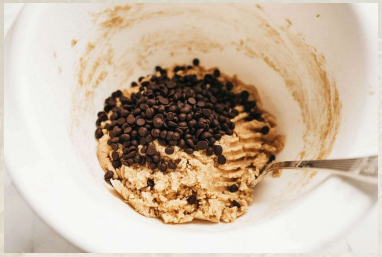
To address the challenge of optimizing sales at Insomnia Cookies, our project will follow a structured analytical approach designed to transform raw data into actionable insights

Gathering the Ingredients

Data Acquisition & Preparation

Collecting and preparing the data 'ingredients' from sales, marketing, and external weather sources.

Mixing the Dough



Exploratory Data Analysis

Combining the ingredients and finding patterns. Understand features affecting cookie sales.

Baking

Baking the model

Training it on historical data, testing its performance, and fine-tuning

Icing/Decorating

Serve with results

Sharing key insights, findings, and recommendations to optimize sales

Georgia Tech

Data Sources

- Insomnia Cookies Transactional Data : Directly from Insomnia Cookies database

Key data points: *Date of transaction, total revenue, store id*
Provides insights into customer behavior, purchasing patterns, and revenue.

- Punchh Marketing Promotion Data : Directly from Insomnia Cookies Punchh instance

Key data points: *Promotion type, promotion start and end dates, promotion details (discount %, points offered)*
Allows correlation of promotional activities with sales performance.

- Weather Data: NOAA Daily Weather Data

Key data points: *Precipitation, min/max temperature, snow (inches) for store locations.*
Allows correlation of weather factor with sales performance.

- Store Information: Directly from Insomnia Cookies database

Key data points: *store IDs, names, locations, store type.*
Used to join with weather data based on location.

Georgia Tech

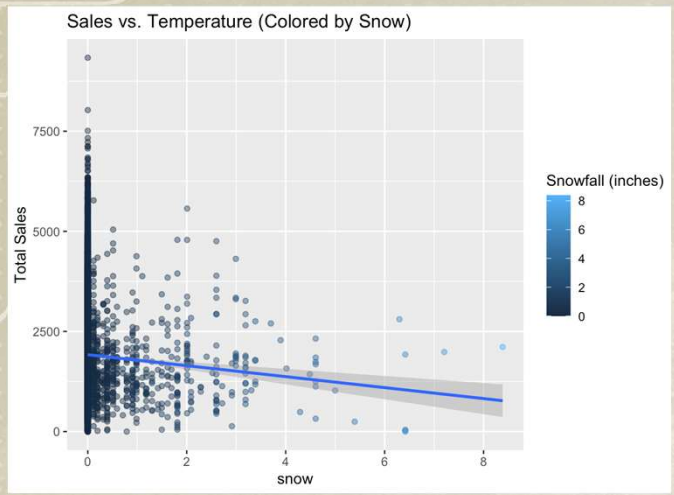
Data Prep

- Handled missing values in promotion and weather data, removing duplicates & outliers from sales data (extremely high amounts and negative amounts)
- Joined datasets on common key columns to create a unified, flat structure for analysis
- Derived columns to capture relevant information like Day of the Week: to analyze weekly sales patterns.
- Calculated Annual/Weekly Seasonality and added in dataset for modelling
- The dataset segmented by store location: New York City, Philadelphia, St. Louis, and a combined "College Towns" group
- Training (80%) and Testing (20%) dataset prepared and used across all 4 groups

Exploratory Data Analysis: Uncovering Sweet Insights

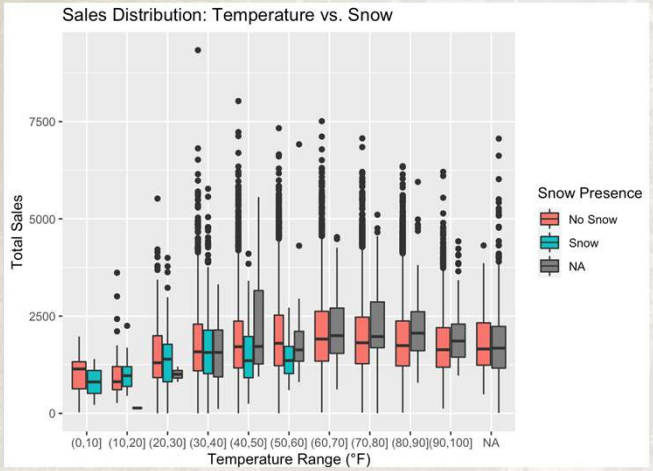
Impact of Snow and Temperature on Sale

Sales vs. Snowfall scatter plot



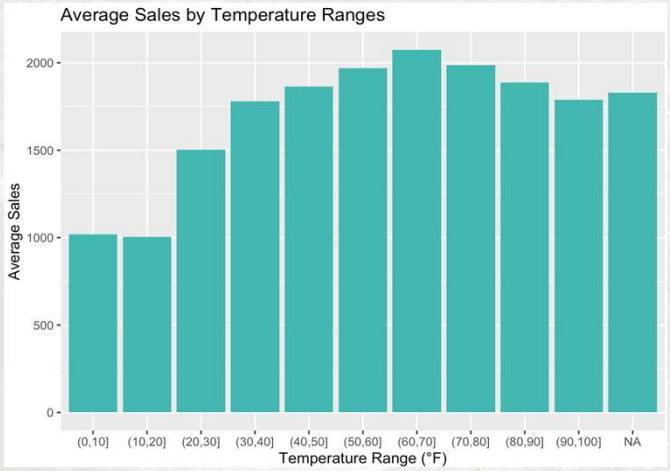
- As snowfall increases, total sales tend to decrease (negative trend)
- Heavy snowfall may deter customers, leading to lower sales.

Sales by Temperature and Snow Presence



- Across nearly all temperature ranges, the median sales value is lower as temperature decreases.
- Additionally, when snow is present (blue boxes) sales are further depressed.

Sales Distribution Across Temperature Ranges



- Sales show an increase within moderate temperatures (40°F-70°F) but are depressed on extreme ends.

Georgia Tech

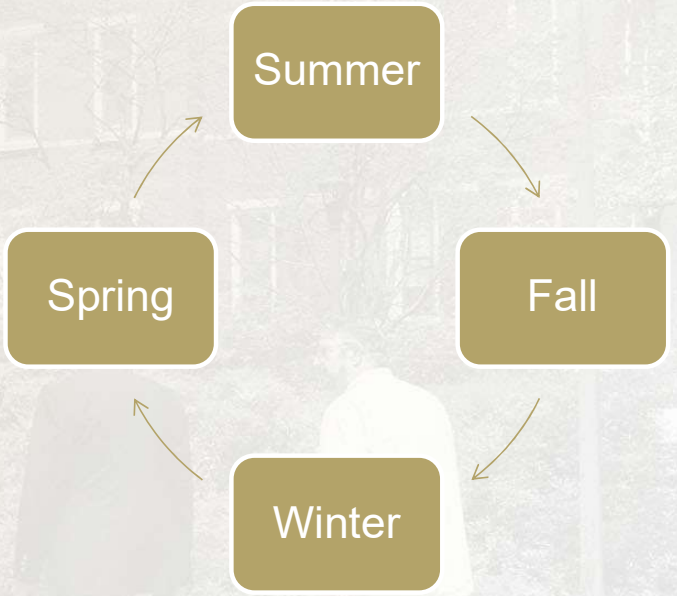
Exploratory Data Analysis: Seasonality

- **Understanding Seasonality:**
  - **Annual seasonality:** Reflects yearly cycles in sales data.
  - **Weekly seasonality:** Reflects Weekly cycles in sales data.

- **General Formula:**
  - $Y_t = T_t * S_t * Rt$
    - **T:** Trend component (long-term movement)
    - **S:** Seasonal Component (Seasonal Component)
    - **R:** Residual

- **Seasonality extraction:**
  - Identifying/isolating cycles from the observed data by removing the long-term trend and residual noise.
  - patterns are calculated by averaging the values for each cycle

- **Model Used: MSTL**
  - **MSTL:** (Multi-Seasonal-Trend decomposition w/LOESS)
    - LOESS (Locally Estimated Scatterplot Smoothing) is a non-parametric regression method used to smooth data

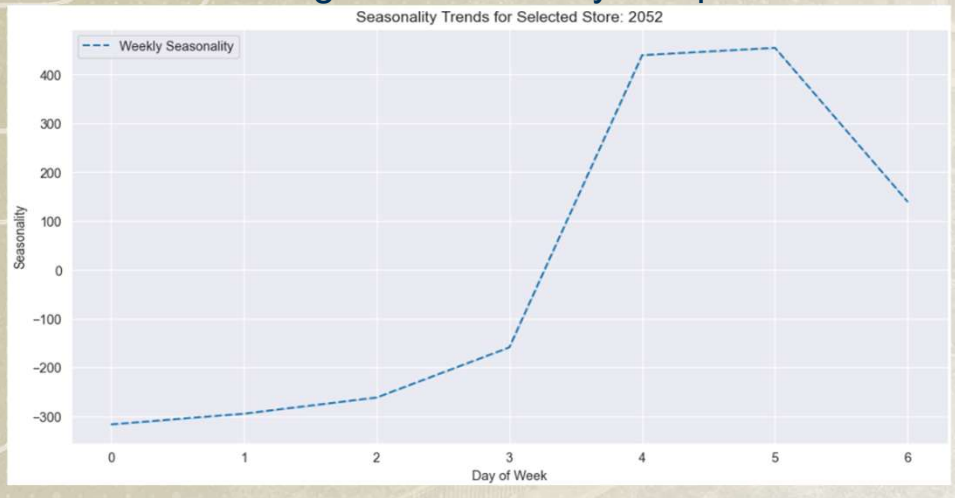


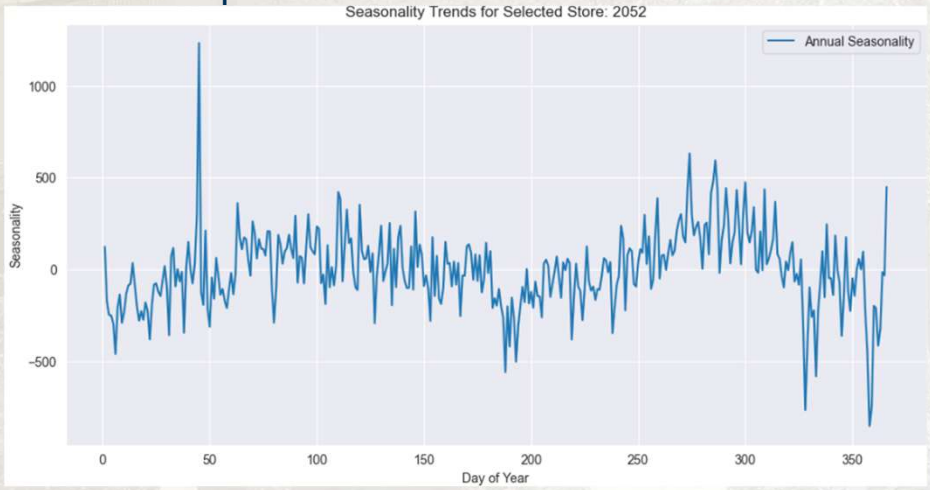
Georgia Tech

Exploratory Data Analysis: Weekly/Annual Seasonality

Weekly and Annual Seasonality:

- Extracting the seasonality component will allow us to build better prediction models.





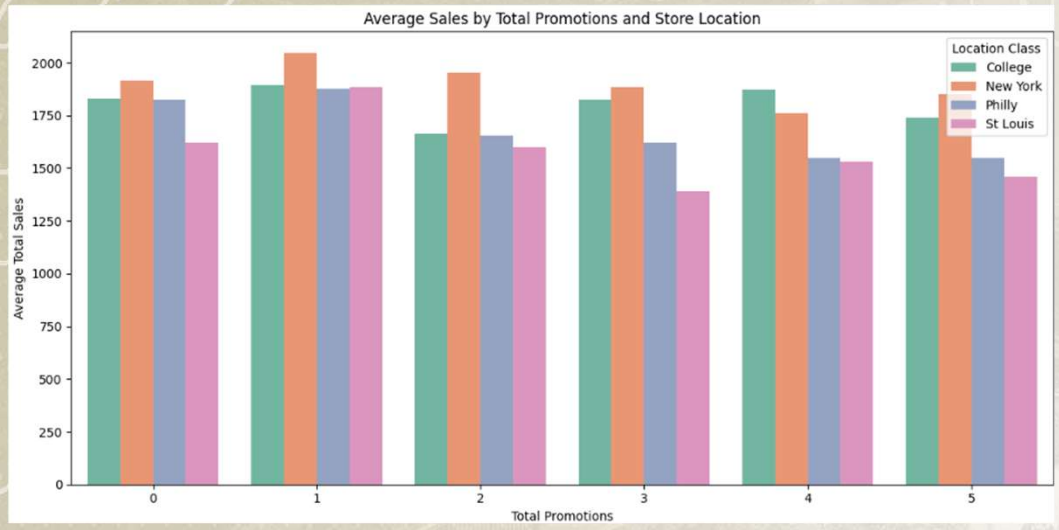
The weekly sales graph reveals distinct seasonality, with higher average sales observed on weekends (Saturday and Sunday), likely driven by increased consumer activity. Sales remain relatively stable but lower during weekdays, with Monday showing the lowest average sales.

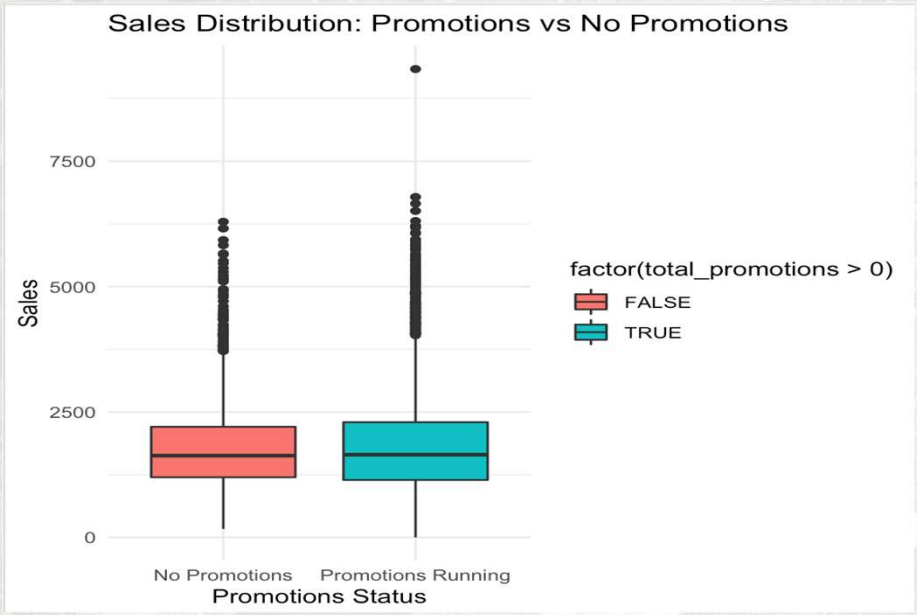
The graph reveals **spring and autumn peaks** (March/April and September/October), a **summer** (June/July), and a **December dip**, which may warrant further analysis to understand the underlying causes, considering other factors affecting the business.

Georgia Tech

Exploratory Data Analysis: Uncovering Sweet Insights

Promotional vs Non-Promotional sales





- Bar plot shows avg daily sales distributed across different numbers of promotions (0-5) and store location.
- Highest average total sales is with 1 promotion.
- For some locations, the overall trend suggests diminishing returns as the number of promotions increases.
- Other factors like type of promotion, channels needs to be taken into consideration.

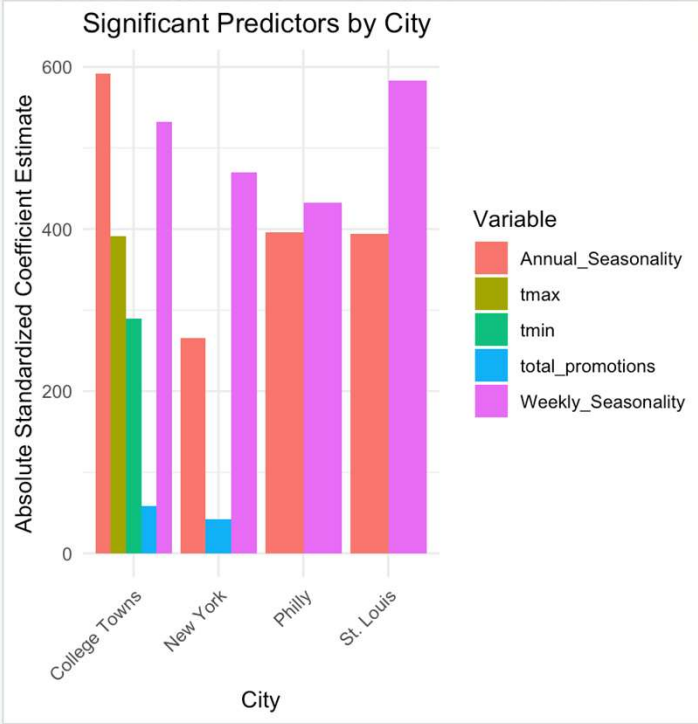
- Promotions occasionally generate high sales spikes (visible as outliers),
- However, promotions don't significantly raise the overal sales median compared to non-promotional periods.
- This suggests the need to consider additional factors in the analysis, such as promotion types, timing, free products, weather

Georgia Tech

Model – Multiple Linear Regression

Predictor Variables: Annual and weekly seasonality, temperature, snow, number of promotions, dayofweek
Response Variable: total.sales
4 models were created based on grouping of stores NY, Philly, St Louis and College towns

- Seasonality strongly impacts sales, Cyclical patterns are key.
- Models struggle to accurately capture sales, suggesting the need for additional variables or more complex modeling techniques.
- Language R, with the dplyr, ggplot2, and tidyr



MLR Model Results				
Store Location	**Average Error Rate**	**Average Error Variance**	**Average Explained Variance**	cross-validation was not prioritized, as baseline model performance was deemed insufficient for practical application
Colleges	$373.01	430.25	79. 56 %	
Philadelphia	$529.89	639.20	48.39 %	
New York City	$662.29	899.69	30.96 %	
St. Louis	$577.10	641.57	55. 24 %	

Next Steps

- Consider random forests and XGBoost to account for non-linear and feature interactions.
- By using the more complex models, we can create more accurate predictions, and better decisions
- Since promotions appear to have some impact for New York stores, identify which promotions are most effective between store locations.

Georgia Tech

Model Selection: Random Forest

•**Captures Nonlinear Relationships**
Effectively models complex interactions between factors like snow, promotions, and weekly seasonality that impact sales.

•**Provides Feature Importance Insights**
It provides feature importance scores, helping to understand the relative influence of variables such as snowfall, promotions, and weekly seasonality on sales.

•**Robust to Missing Data**
Handles missing values seamlessly, reducing the need for extensive data preprocessing.

•**Proven Success in Sales Forecasting**
Frequently used in similar scenarios due to its accuracy and ability to account for diverse external factors.

Georgia Tech

Random Forest: Model Building/Parameter Tuning

- **Modeling:**
  - A 5-fold cross validation grid search was utilized to determine the optimal number of variables sampled at each split (*mtry*).
    - The R package *caret* was utilized for this grid search & model creation.
  - The resulting RMSE and number of sampled variables tested in the grid search are shown on the right. Variation in RMSE was limited within each model between sampled variable numbers.
  - For each location type (College, St Louis, Philly, New York), 500 tree random forest model was created using the optimal *mtry* parameter

Georgia Tech

Random Forest: Results

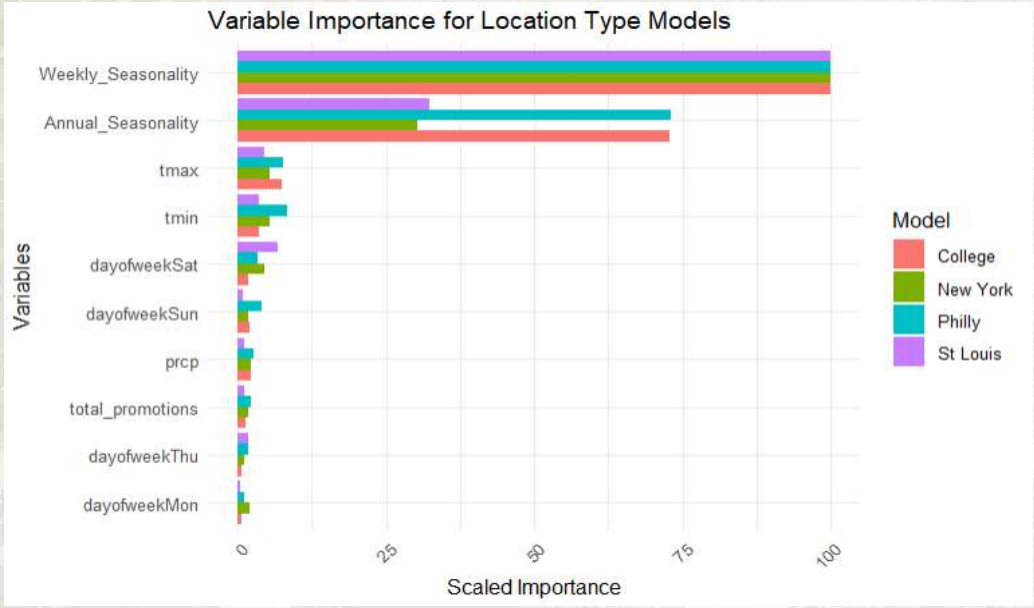
- For assessment of the Random Forest model's predictions for the Test data set, results seen for average error and explained variance are shown in the table on the right.
    - College locations reported the highest Explained Variance (85.41%) with the next lowest density category of locations, St. Louis having the next highest Explained Variance (80.22%).
    - Average Error (MAE) was calculated as the mean of the absolute differences between predicted and actual test data sales values.
    - College and St Louis models reported the lowest Average Error values and highest Average Error Variances.

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 %

Georgia Tech

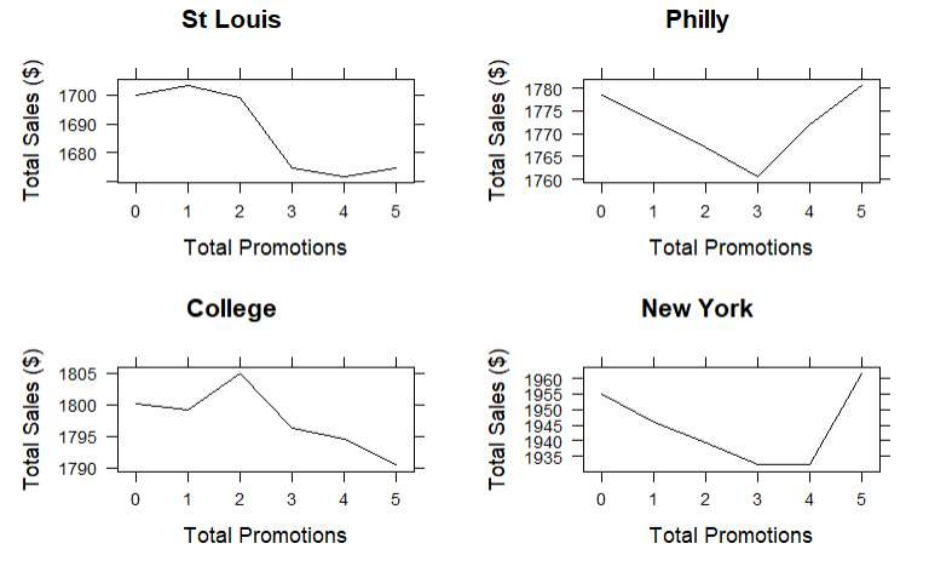
Random Forest: Feature Importance

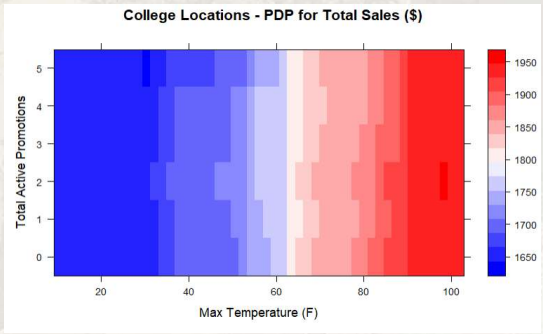
- Variable importance (normalized by standard error) was calculated for each of the models' variables.
    - Weekly & Annual Seasonality are the dominant variables for all models.
        - St Louis and New York are less impacted by annual seasonality than the college and Philly locations.
    - The following are also listed within the top 10 important variables, but relatively weak compared to seasonality:
        - Weather-related variables (Max Temperature, Min Temperature, and Precipitation)
        - Day of the Week
        - Total Number of Promotions

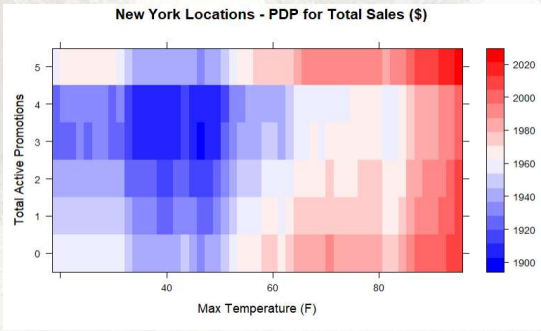


Random Forest: Feature Importance

- Partial Dependence Plots (PDPs) visualize the effects of the Total Promotions variable, while keeping other variables constant.
- Different interactions between promotions and total sales across location models:
  - Philly & New York locations show a negative relationship for 0-3 promotions, but maximum sales values at 5 promotions.
  - St Louis & College locations: maximum sales values at 1 & 2 promotions.
- Temperature has a stronger relationship to sales for College & St Louis locations, relative to New York & Philly.
  - Some visual evidence of promotions providing a bump to sales in college locations in the 30-50 degrees F max temperature range.



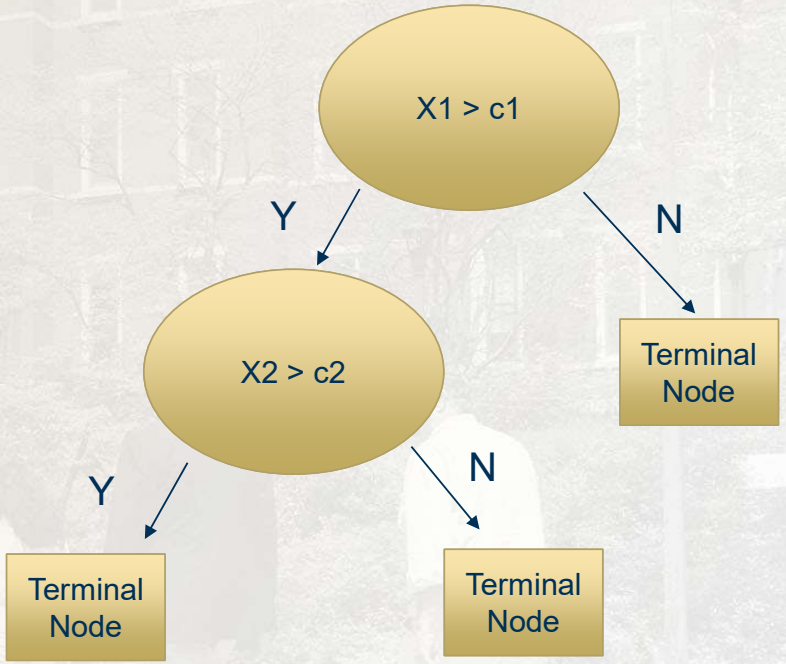




Georgia Tech

Model Selection: XGBoost

- **Extreme Gradient Boosting:** is a *boosting ensemble method* that builds decision trees *sequentially*.
- **Tree Growth:** Trees are grown *independently*, and *greater weight* is assigned to "weaker" learners.
- **Training Speed:** Training generally happens much faster with larger datasets compared to other models. (including Random Forests)
- **Missing Data:** Branch directions for missing values are learned during training.
- **Model Interpretability:** SHAP (Shapley Additive Explanations) values help explain how each feature contributes to a prediction.
- **Tools Used:** Python (pandas) for data handling, XGBoost for modeling, and scikit-learn for cross-validation and model performance evaluation.

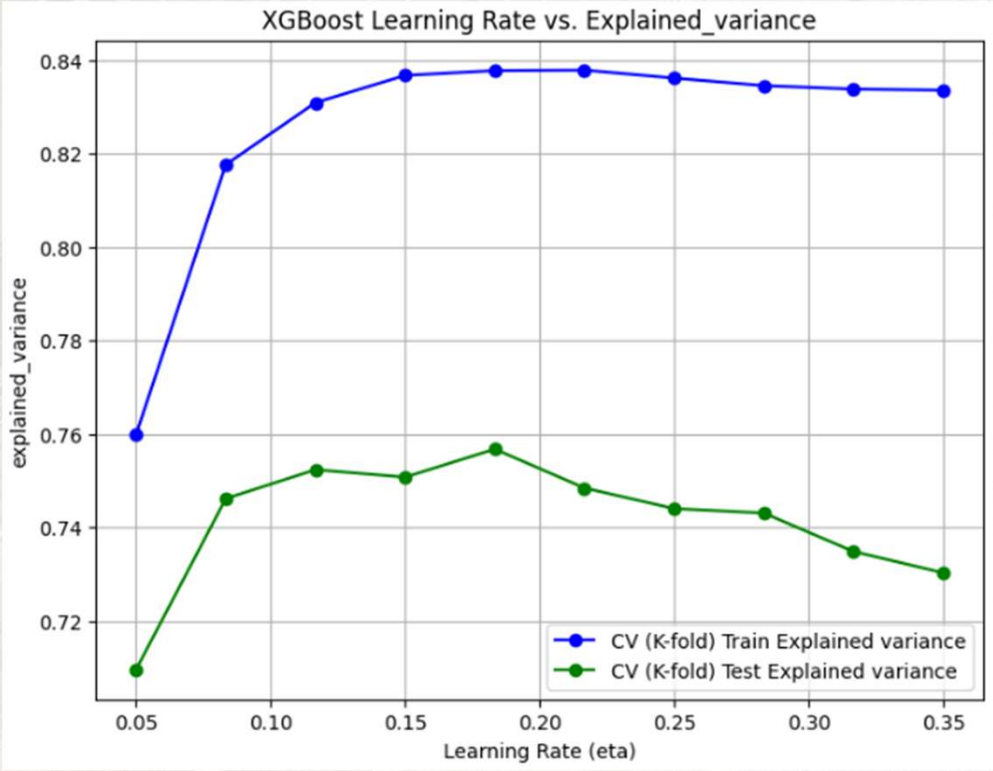


Georgia Tech

XGBoost: Model Building/Parameter Tuning

- **Modeling:** One model was created for each store location
  - College Campuses
  - Philadelphia, PA
  - New York City, New York
  - St. Louis Missouri

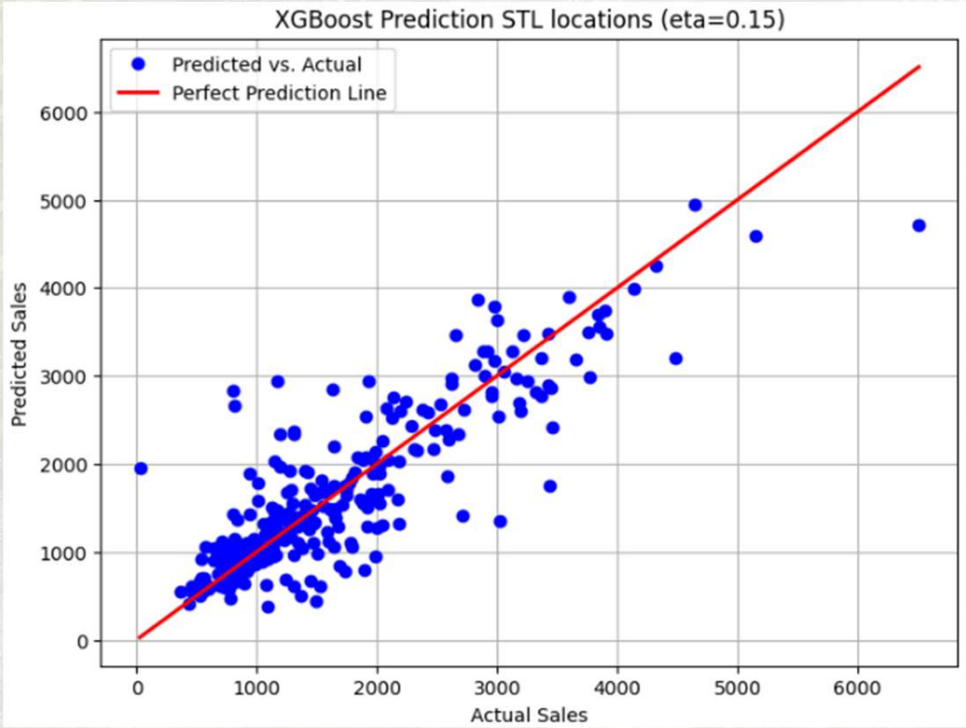
- **Model Parameters:**
  - **Number of Estimators:** Number of *boosted trees* that are added to the model.
    - Estimators = sqrt(length of training data)
  - **Tree Depth:** Controls the complexity of each individual tree.
    - Chosen via K-Fold (k= 10) cross validation
  - **Learning Rate parameter (η)** : Controls how much influence each new tree has on the final prediction.
    - General Equation: $y_i = y_{i-1} + \eta \cdot f(x_i)$
    - Chosen via K-Fold (k= 10) cross validation



Georgia Tech

XGBoost: Results

- **Model Evaluation**:
  - 100 Monte Carlo Cross Validation simulations
  - Metrics: Error Rate, Error Variance, Explained Variance
- **Key Takeaways:**
  - XGBoost models *greatly improved* from their original performance on the testing data set.
  - Most models Improved ~20% from their original performance.
  - Predictions *more often* underestimated actual daily sales.

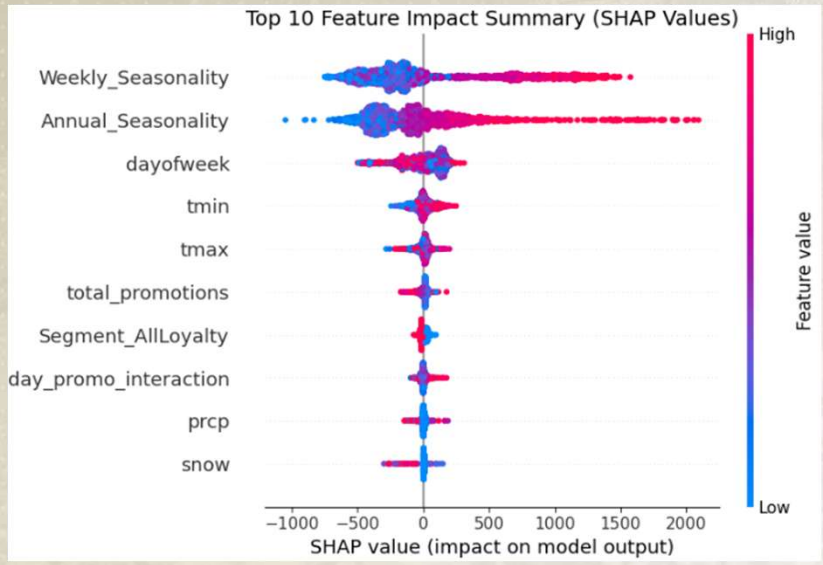


XGBoost Model Results (100 Monte Carlo Simulations)			
Store Location	**Average Error Rate**	**Average Error Variance**	**Average Explained Variance**
Colleges	$71.57	123.39	95.69 %
Philadelphia	$122.55	72.67	92.03 %
New York City	$109. 91	112.94	90.46 %
St. Louis	$72.59	76.38	95.10 %

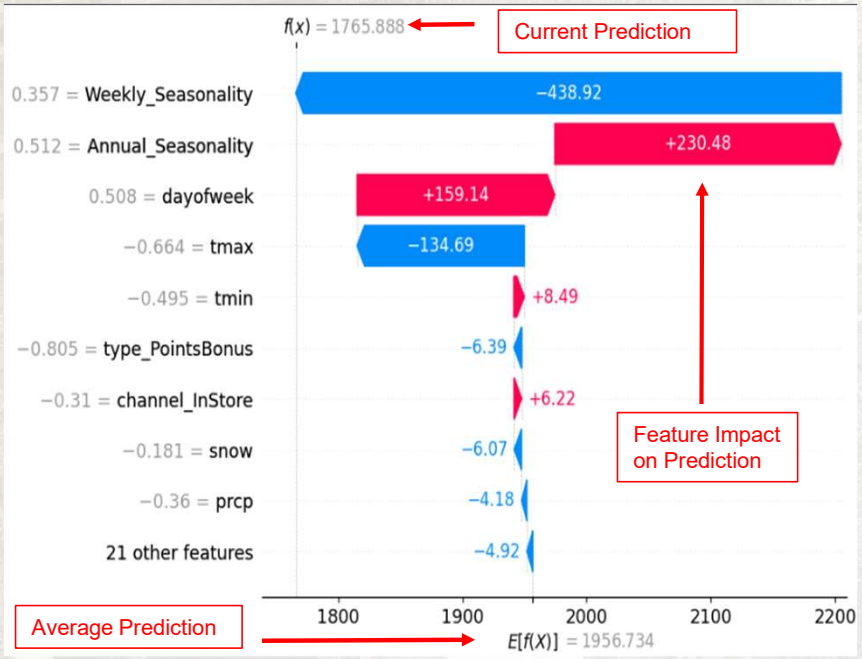
Georgia Tech

XGBoost: Feature Importance

- ***SHAP Values (Shapley Additive Explanations)*** *:*
  - *Explain how each feature contributes to a prediction.*
  - *Uses "fair" allocation results from game theory to allocate credit*
- *Weather effects and seasonality had strongest impact on the model*
- *Other important features:*
  - *Total promotions, Loyalty segment, Day of the week and promo interaction term*



SHAP Waterfall Chart



- ***SHAP Waterfall Chart*** *:*
  - *Chart shows the impact of model features on a single prediction.*
  - *SHAP Values of input features will always sum up to the difference between the current and average predictions.*

Georgia Tech

Conclusions

Model Performance:

- **Regression**
  - High average error rates and modest explained variance (R-squared) a poor overall fit.
- **Random Forests**
  - Stronger explained variance than MLR, but less than XGBoost
- **XGBoost:**
  - Powerful Prediction Capabilities
  - Relatively large error variance

Data-Driven Feature Insights :

- Seasonality Cycles have large impact
- Promotion-Day of the Week Interaction
  - Potential for promotions to drive customers from busy days back on light days.
  - Target favorable days for promotions. Differences in favorable days by cities— e.g. college: Sun and NY: Thurs.
- Weather Impacts:
  - Warmer weather increases sales & Precipitation (Rain/Snow) negatively impacts sales.
  - Potentially focus on delivery-focused promotions on Rain / Snow days to offset decrease in sales.
- Having Multiple Ongoing Promotions doesn't increase sales; one clear promotion at a time should be the focus.

Georgia Tech

Future Directions: Refining the Sales Model

- Develop robust models to robustly predict seasonal patterns in data.
  - Potential to include external timing variables: college academic calendars, sporting events, civic events.
- Time Series Models: Apply time series models (ARIMA, SARIMA) to better forecast/predict temporal dependencies and seasonality.
- Look at subscribing to a quality weather forecasting service.
- Incorporating loyalty programs and customer segmentation can better optimize predictive performance.
- Incorporate more historical data:
  - Could reveal longer-term trends and improve model accuracy.
  - Could help to better identify the true impact of promotions and other critical variables, for informed decisions about sales strategies.
- Use clustering (k-means) to separately group stores with similar features for prediction modeling instead of using just location.
- Evaluate promotions based on type to determine if free delivery for example is more effective in New York on a Thursday than offering a product discount.

Georgia Tech

Challenges

- Difficulty Dealing with/Predicting Time Series Data and availability of Punchh Promotional Data; Punchh data only began being tracked in April 2024.
- Importance of Optimizing Model parameters and using Cross Validation to verify results.
- Limited data makes it difficult to see patterns; more complete historical data will help create a more complete analysis.
- Other factors such as paid media spending and SEO are not currently being considered in the model as that data was not available.
- Time constraints limited the depth of analysis and the ability to explore alternative modeling approaches.

Lessons Learned

- Different team members worked on different models, a consistent evaluation framework and shared dataset are needed to ensure reliable comparisons of model performance across analyses and team members.
- Models may not always yield expected insights; Even when initial models don't fully explain the drivers of sales, particularly regarding promotion effectiveness, we must continue to refine our analysis, explore additional variables, and consider alternative modeling techniques to gain deeper insights and improve sales strategies.

Georgia Tech