Data-Driven Strategies for Enhanced Sales Prediction: A Case Study in Confectionery Manufacturing in SRI LANKA

TEAM JARVIS

CONTENTS

Introduction

Company X, a leading confectionery manufacturer in Sri Lanka, is grappling with the challenge of optimizing its sales and distribution strategies in the ever-evolving confectionery industry. With a diverse product portfolio ranging from chocolates to biscuits, the company relies on an extensive network of distributors and outlets spread across the Western Province of Sri Lanka.

The datasets provided on the Kaggle platform, train.csv and test.csv show valuable information regarding week-wise sales data for each outlet.

- ✓ train.csv: Contains historical week-wise sales data.
- ✓ test.csv: Used for testing the developed forecasting model.

The dataset includes the following variables:

• outlet code: Unique code to identify the outlet.

• week start date: Start date of the week in which the transaction occurred.

• expected rainfall: The expected rainfall for the week of the transaction.

• freezer status: Specify if the outlet has a freezer or not.

• outlet region: The geographical regions of the outlets.

• sales quantity: Quantity of items sold.

Problem Statement:

The existing forecasting methodologies, primarily based on ARIMA models, fall short of fully realizing the potential of each outlet. Company X aims to leverage advanced analytics to gain profound insights from the data, allowing for more accurate sales predictions.

Objective:

Optimize distribution plans and maximize sales efficiency.

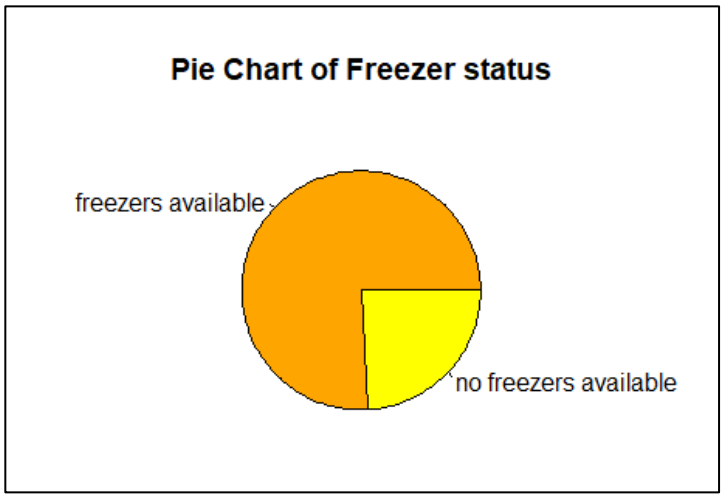
Data Pre-processing

1) No missing values were found in all the columns.
2) No invalid dates, but format of week_start_date in both train.csv and test.csv were changed to (M/D/Y).
3) The variable expected_rainfall initially was in character format. The substr function was employed to remove the last two characters from each string, as they represented units associated with rainfall. This conversion ensured that expected_rainfall is represented as a numeric variable, allowing for more accurate statistical analysis.

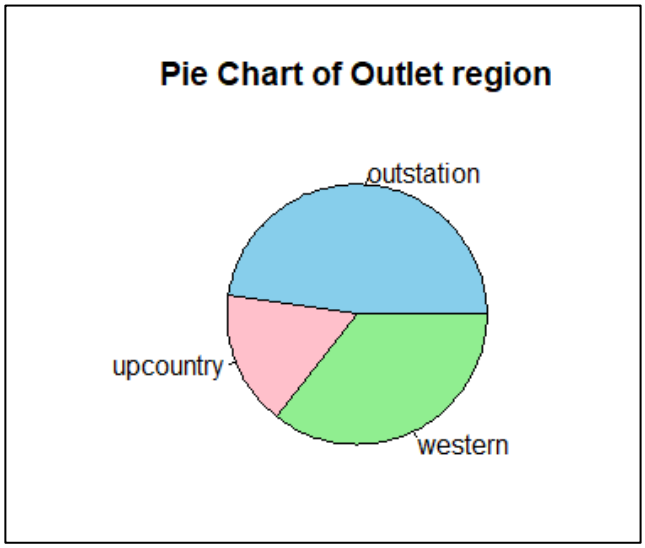
4) The column freezer_status is supposed to have only two values which are ‘freezers available’ and ‘no freezers available’. Originally in the dataset, due to evidence of spaces the variable doesn’t seem like a binary variable which was corrected. Following the correction, the variable was encoded as a factor, ensuring its appropriate representation in subsequent analyses. This data pre-processing step was essential to standardize the freezer_status variable, providing a clear and consistent framework for further exploration and analysis.
5) The variables outlet_region and outlet_code was transformed into factors, allowing for improved representation and interpretation in statistical models.

Explanatory Data Analysis

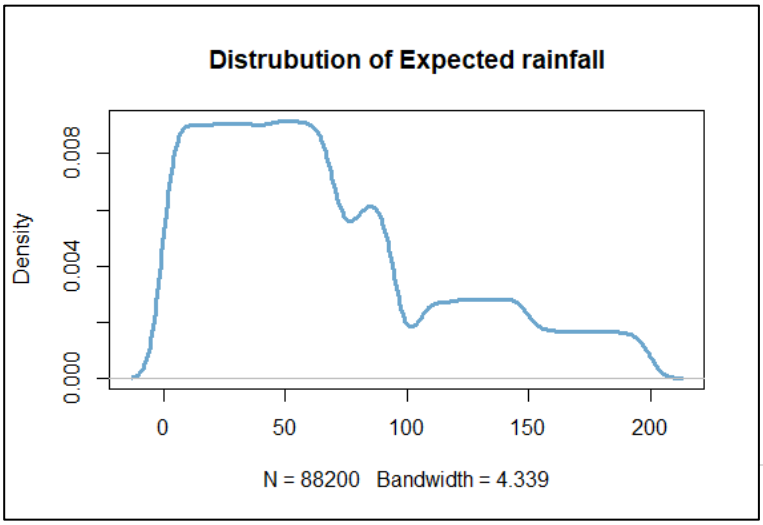
➢ Univariate Analysis



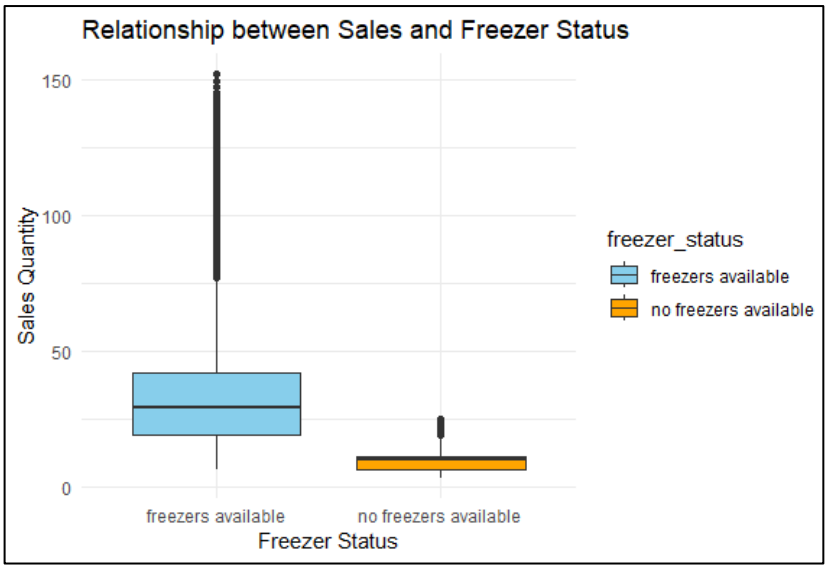
Majority of outlets have freezers.



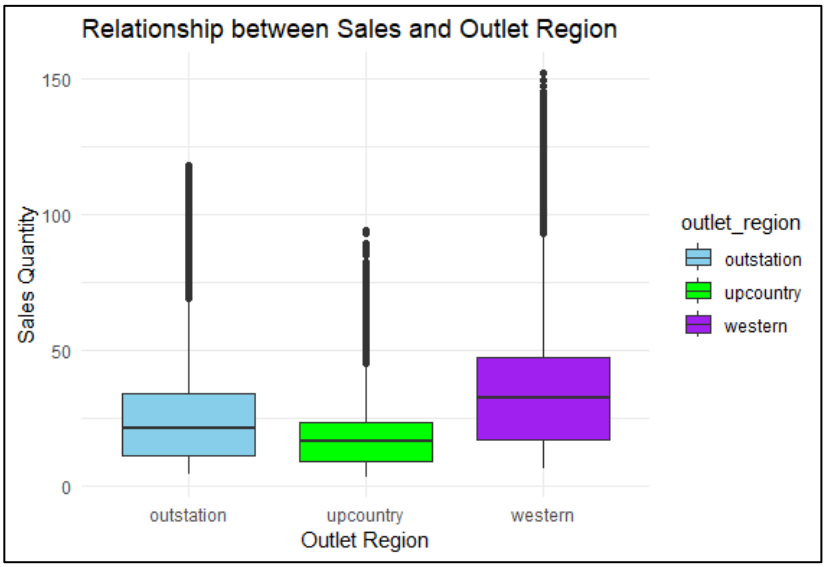
Most of the outlets are found outstation, followed by western province and then upcountry.



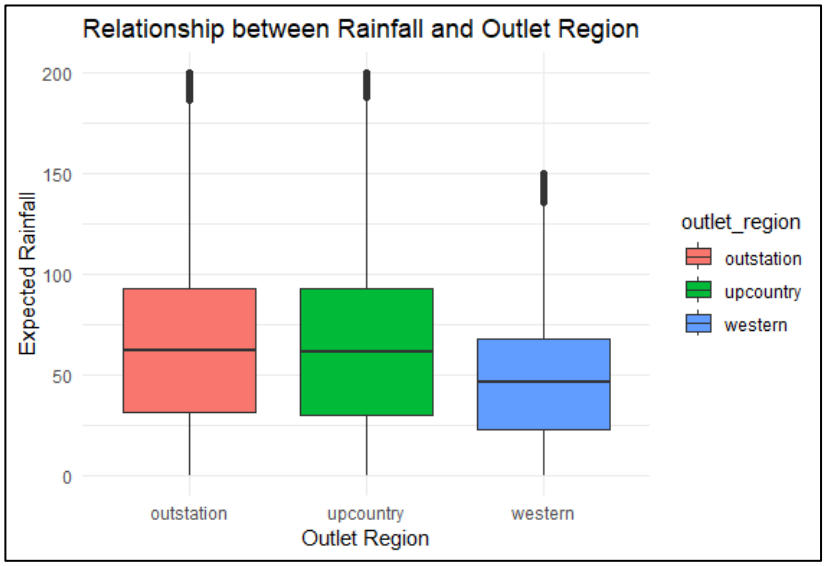
The distribution of expected rainfall is positively skewed indicating that the expected rainfall is dominated in the region between 0mm-100mm.



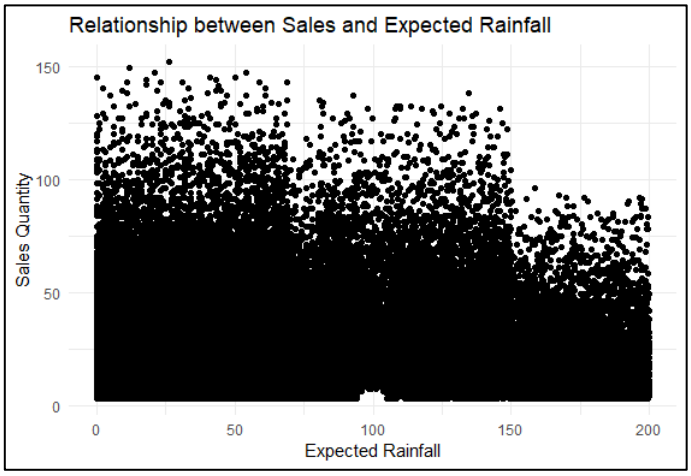
The outlets which have freezers available tend to be having higher number of sales while outlets with no freezers seem to be very poorly in sales.

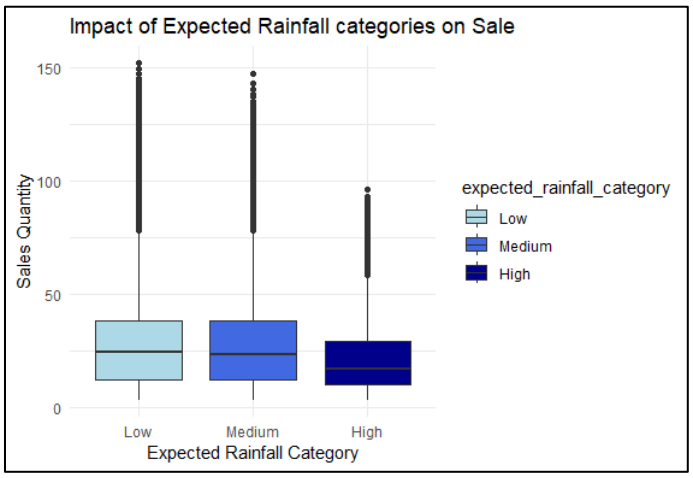


Wester province outlets show the highest sales whereas outstation outlets do better than upcountry outlets.



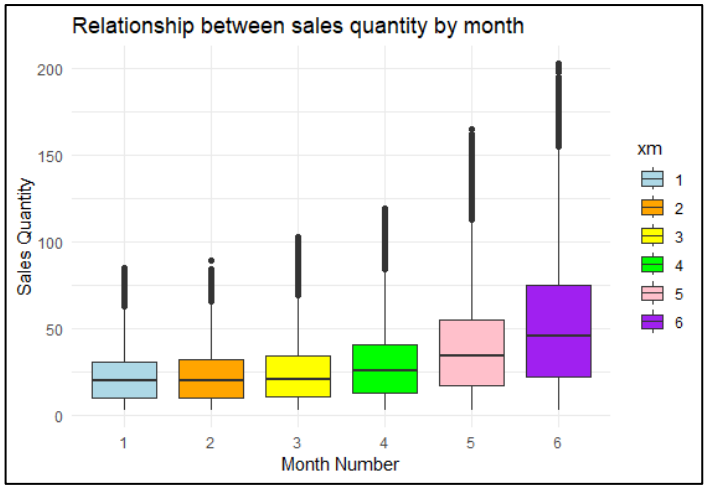
When observing the relationship between expected rainfall and outlet region, outstation, and upcountry experience more rainfall which might have caused them for lower sales as seen in the above boxplot.

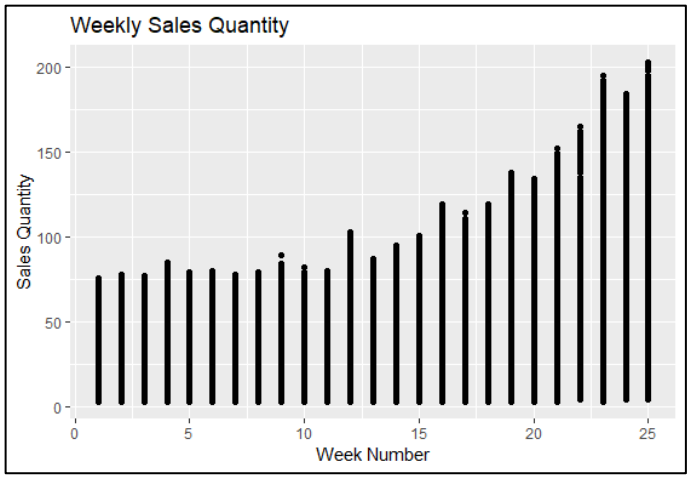




A weak negative correlation of -0.06182801 between sales and expected rainfall can be observed.

Rainfall was transformed into categories, low(0-50mm), medium(50mm,150mm) and high(150mm,200mm). The above box plot indicates that when the expected rainfall is high the sales is low.





When months go by, the sales seem to increase and the 6th Month(June) the sales is the highest.

Overall increasing trend can be observed between sales and week numbers.

Feature Engineering

❖ Feature Engineering for Time Components:

To enrich the temporal information within the dataset, feature engineering was conducted on the week_start_date column. The objective was to derive additional variables representing the month and week of each transaction.

❖ Feature Engineering to transform to categorical variable.

In the process of refining the dataset for analysis, specific attention was given to the expected_rainfall variable. This variable, originally formatted as a character string, required transformation to a numeric type to facilitate meaningful quantitative analysis. Thereafter, for better insights a categorical variable 'expected_rainfall_category' was created indicating low(0-50mm), medium(50-100mm) or high rainfall(100mm-200mm).

Data Insights for Company X: Answers to the Questions

(a) Synthetic Features Table:

Name	Description	Data Type
expected_rainfall_category	Indicating low(0-50mm), medium(50-100mm) or high rainfall(100mm-200mm)	Categorical
month	Numeric representation of the month extracted from date	Categorical
Week_No	Numeric representation of the week number extracted from date	Ordinal

(b) The summary statistics for the variables:

```
> summary(df)
 week_start_date      expected_rainfall              freezer_status
 Min.   :2023-01-02   Min.   :  0.00    freezers available   :79625
 1st Qu.:2023-02-13   1st Qu.: 27.00    no freezers available:25375
 Median :2023-03-27   Median : 55.00
 Mean   :2023-03-27   Mean   : 65.74
 3rd Qu.:2023-05-08   3rd Qu.: 90.00
 Max.   :2023-06-19   Max.   :199.99

   outlet_region             outlet_code     sales_quantity      month
outstation:50000   outlet_code_1   :    25   Min.   :  3.0   Min.   :1.00
upcountry :17500   outlet_code_10  :    25   1st Qu.: 13.0   1st Qu.:2.00
western   :37500   outlet_code_100 :    25   Median : 25.0   Median :3.00
                   outlet_code_1000:    25   Mean   : 31.3   Mean   :3.36
                   outlet_code_1001:    25   3rd Qu.: 42.0   3rd Qu.:5.00
                   outlet_code_1002:    25   Max.   :203.0   Max.   :6.00
                   (Other)         :104850
     Week          Week_No   expected_rainfall_category
 Min.   : 1.00   Min.   : 1   Low   :48319
 1st Qu.: 8.00   1st Qu.: 7   Medium:48011
 Median :15.00   Median :13   High  : 8670
 Mean   :15.08   Mean   :13
 3rd Qu.:22.00   3rd Qu.:19
 Max.   :30.00   Max.   :25
```

Some key observations based on the EDA:

- Outlets equipped with freezers demonstrate higher sales, while those without freezers exhibit lower sales performance.
- Sales in the Western province outperform other regions, with outstation outlets generally performing better than upcountry outlets.
- High expected rainfall appears to correlate with lower sales figures.
- Over the course of the months, there is a noticeable upward trend in sales. Particularly, sales peak in the 6th month (June).

(c) The selected target variable for forecasting sales is sales_quantity. In the process of feature engineering, the logarithm of sales_quantity was taken, i.e., log(sales_quantity), to address potential issues related to the distribution of the response variable as it was positively skewed. Additionally, during the model fitting process, the log-transformed values were used as the target variable.

(d)

Forecasting Methodology :

The forecasting methodology involved employing multiple models to predict sales quantity. The models used are:

1. Multiple Regression:
   The dataset was divided into training and testing sets. Feature engineering involved transforming the expected_rainfall variable and cleaning categorical variables. Logarithmic transformation of the sales quantity was applied to address distribution issues.
2. Regression Trees:
   Regression trees were fitted, and the model was developed to optimize its performance, and the mean squared error (MSE) was used for model evaluation.
3. Gradient Boosting (XGBoost):
   Hyperparameters such as the number of trees, interaction depth, and shrinkage were tuned. The model was trained on the logarithmic transformation of sales quantity.
4. Bayesian Additive Regression Trees (BART):
   The BART model was fitted where it utilized Bayesian methods for tree-based regression.

Model Evaluation:

Model performance was assessed using the Mean Absolute Percentage Error (MAPE). MAPE is calculated as follows:

$$\textbf{MAPE} = \textbf{Sum}(|\textbf{Predicted Sales (I)} - \textbf{Actual Sales(I)}|) \,/\, \textbf{Sum}(\textbf{Actual Sales(I)})$$

- ❖ The choice of models was driven by the need to explore various approaches, considering the strengths and weaknesses of each. The models were trained and evaluated on a log-transformed sales quantity to enhance their predictive accuracy. The final selection of the forecasting model was chosen to be XGBoost with the logarithm transformation on the response variable. The final model resulted in the lowest MAPE, indicating the model with the least prediction error.
- ❖ While the normality assumption is less critical for tree-based models like XGBoost, log transformation is a useful preprocessing step for addressing specific characteristics of the data distribution. It aids in stabilizing variance and mitigating the impact of extreme values on error metrics like MAPE. The choice of log transformation and the selection of XGBoost as the final model indicates a thoughtful approach to handling data characteristics and improving the accuracy of predictions.

(e)

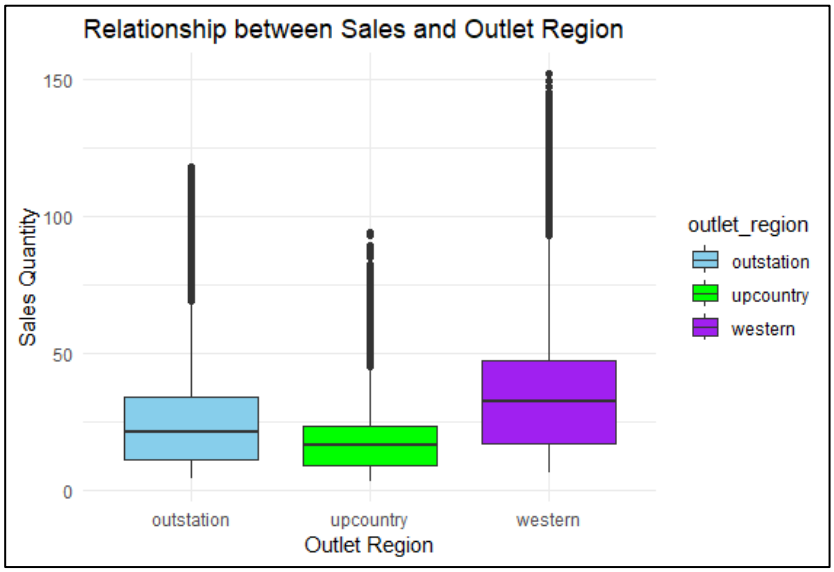
1) The table below shows the average weekly sales for each outlet region.

Week No	Outstation	Upcountry	Western
1	20.79250	15.17571	28.29600
2	21.17250	15.62857	28.28667
3	20.84750	15.89143	28.75600
4	19.37950	11.55000	30.93400
5	21.04550	15.90714	28.45733
6	21.40650	16.01286	29.25000
7	21.72200	15.50000	29.08533
8	21.68600	16.07857	28.80000
9	19.96900	11.95714	31.70067
10	21.81550	15.97143	29.16667
11	21.58400	15.71571	29.40733
12	22.83650	13.25286	36.96867
13	23.50300	17.54571	32.09133
14	25.37950	18.63571	34.48667
15	27.04850	19.88571	36.38667
16	26.11050	15.23286	42.67000
17	30.24050	22.81429	41.37467
18	32.04200	23.25143	43.58067
19	30.52900	17.77429	49.50200
20	35.00000	25.11714	48.14867
21	39.66400	28.65714	53.79733

22	43.09100	31.04000	59.60200
23	41.79650	24.30714	69.45267
24	49.78800	35.30714	67.66400
25	53.43550	38.55000	72.33133

2) A weak negative correlation can be observed between sales and rainfall. Correlation coefficient: -0.06182801. The scatterplot and the box plot in the EDA above also indicate the same.

3) As mentioned in the EDA above Western province outlets show the highest sales whereas outstation outlets do better than upcountry outlets. Consider the box plot below:



According to the boxplot, approximately 75% of sales is above 30 at the western outlets, whereas 50% of sales is below 30 at the outstation outlets.

Conclusion

By delving into the sales data, valuable insights into the driving forces behind sales quantity were discovered. Key observations uncovered through exploratory data analysis provided a roadmap for decision-making and model selection.

The choice of models was driven by a thorough understanding of the dataset's characteristics. Various modeling techniques, including multiple regression, regression trees, XGBoost, and Bayesian Additive Regression Trees (BART), were explored. The log transformation applied to

the sales quantity variable aimed to enhance the models' predictive accuracy, particularly benefiting models that assume a normal distribution of the response variable. Synthetic features, such as Week_No and month, were engineered to capture temporal patterns and seasonal variations, contributing to a more comprehensive representation of the data.

After rigorous training and evaluation, XGBoost emerged as the preferred forecasting model, demonstrating the lowest Mean Absolute Percentage Error (MAPE) and, consequently, the least prediction error.

Overall, this comprehensive approach, involving feature engineering, exploratory data analysis, and the application of diverse modeling techniques, enhances the reliability of the forecasting models. The insights gained from this analysis provide a foundation for strategic decision-making and optimization of sales strategies in the future.