REPORT

Team CodeWave

Mini Hackathon 2023

Preliminary Round

Group Members:

- Nadun Channa.
- Ruvindya Sachinthani.
- Hirusha Adithya.
- Hansaja Amarajith.

Approach and Methods

Data Preprocessing

- Dataset: Utilized datasets include train and test.
- Handling Missing Data: Checked and no missing values.
- Data Encoding: Encoded categorical variables, such as "outlet_code" and "outlet_region," using label encoding and one-hot encoding techniques.
- Feature Extraction: Extracted additional features like year, month, and custom week from the "week_start_date" column.

Model Development

- Algorithm: Utilized the Random Forest Regressor for its predictive capabilities.
- Feature Selection: Identified and selected relevant features based on correlation analysis and domain knowledge.
- Training: Trained the model on the training dataset.
- Validation: Validated the model Mean Absolute Percentage Error (MAPE) as the evaluation metric.

Assumptions

- Data Consistency: Assumed consistent data across training and testing datasets.
- Stationarity: Assumed the time series data to be stationary for model training.

Feature Engineering Techniques

- Datetime Features:
    - Utilized the "week_start_date" column to extract additional datetime features.
    - Converted "week_start_date" to a datetime format (pd.to_datetime).
    - Calculated a custom week number using a function (custom_week_number), considering the first day of the month and ensuring a maximum of 5 weeks.
    - Extracted and added features such as year, month, and custom week to the DataFrame.
    - Drop Year, because it constant for all samples.

- Categorical Encoding:
    - Transformed categorical variables into binary features for enhanced model interpretability and performance.
    - Applied one-hot encoding to categorical variables, such as "outlet_code" and "outlet_region," using pd.get_dummies.

a)

Feature Name	Description	Data Type
months	Extracted month from the "week_start_date"	Integer
custom_week	Calculated custom week number	Integer
outlet_region_outstation	One-hot encoded feature for outlet region outstation	Binary
outlet_region_upcountry	One-hot encoded feature for outlet region upcountry	Binary
outlet_region_western	One-hot encoded feature for outlet region upcountry	Binary
freezer_status_0	One-hot encoded feature for freezer Available	Binary
freezer_status_1	One-hot encoded feature for freezer Not Available.	Binary
rainfall_range_0-70	Extract from expected_rainfall	Integer
rainfall_range_70-150	Extract from expected_rainfall	Integer
rainfall_range_150+	Extract from expected_rainfall	Integer

b)

index	freezer_status_0	freezer_status_1	rainfall_range_0-70	rainfall_range_70-150	rainfall_range_150+
count	113400.0	113400.0	113400.0	113400.0	113400.0
mean	0.7583333333333333	0.24166666666666667	0.6476807760141093	0.2758553791887125	0.07646384479717813
std	0.4280952055172032	0.4280952055172032	0.47769488240354424	0.4469462501484651	0.2657400006914178
min	0.0	0.0	0.0	0.0	0.0
25%	1.0	0.0	0.0	0.0	0.0
50%	1.0	0.0	1.0	0.0	0.0
75%	1.0	0.0	1.0	1.0	0.0
max	1.0	1.0	1.0	1.0	1.0

Table 01 – Summary statistics for feature

index	outlet_region_outstation	outlet_region_upcountry	outlet_region_western
count	113400.0	113400.0	113400.0
mean	0.47619047619047616	0.16666666666666666	0.35714285714285715
std	0.49943498694208266	0.3726796394618745	0.47915953645098136
min	0.0	0.0	0.0
25%	0.0	0.0	0.0
50%	0.0	0.0	0.0
75%	1.0	0.0	1.0
max	1.0	1.0	1.0

Show 25 per page

Table 01 – Summary statistics for feature (Cont.)

	month	custom_week
count	113400.000000	113400.000000
mean	3.592593	3.222222
std	1.831002	1.314690
min	1.000000	1.000000
25%	2.000000	2.000000
50%	4.000000	3.000000
75%	5.000000	4.000000
max	7.000000	5.000000

Table 01 – Summary statistics for feature (Cont.)

c) The selected target variable for forecasting sales is the "sales_quantity" column in the dataset. This choice is based on the company's objective to predict the quantity of items sold for each item category across stores in the next week. The "sales_quantity" column represents the key metric that aligns with the standard week for operations starting on Monday. The decision to use "sales_quantity" is driven by its significance in gauging consumer demand, optimizing distribution plans, and informing manufacturing and inventory strategies. The inefficiency identified in the existing ARIMA methods motivates the exploration of advanced analytics techniques to enhance forecasting precision and unlock the full potential of the company's outlets. The target variable is engineered by aggregating the quantity of each SKU sold to each outlet within specific weeks, providing a comprehensive dataset for accurate sales forecasting.

d) The forecasting approach for sales employs the RandomForestRegressor algorithm, chosen for its versatility in handling mixed data types and capturing complex patterns. Feature engineering techniques include extracting datetime features and one-hot encoding categorical variables, enhancing the model's ability to discern temporal and categorical nuances. A Mean Absolute Percentage Error (MAPE) is selected as the metric, aligning with the importance of proportional accuracy in sales forecasting. Synthetic features, such as one-hot encoded categorical variables and engineered datetime features, contribute to a comprehensive feature set. The methodology is designed to be iterative, allowing for continuous model refinement through hyperparameter tuning, exploration of alternative algorithms, and adaptive feature engineering. Aligned with business objectives, this approach seeks to provide accurate sales forecasts for distribution planning, inventory management, and strategic decision-making at The Confectionery Company.

e) Average Weekly Sales Volumes for Each Outlet Region:

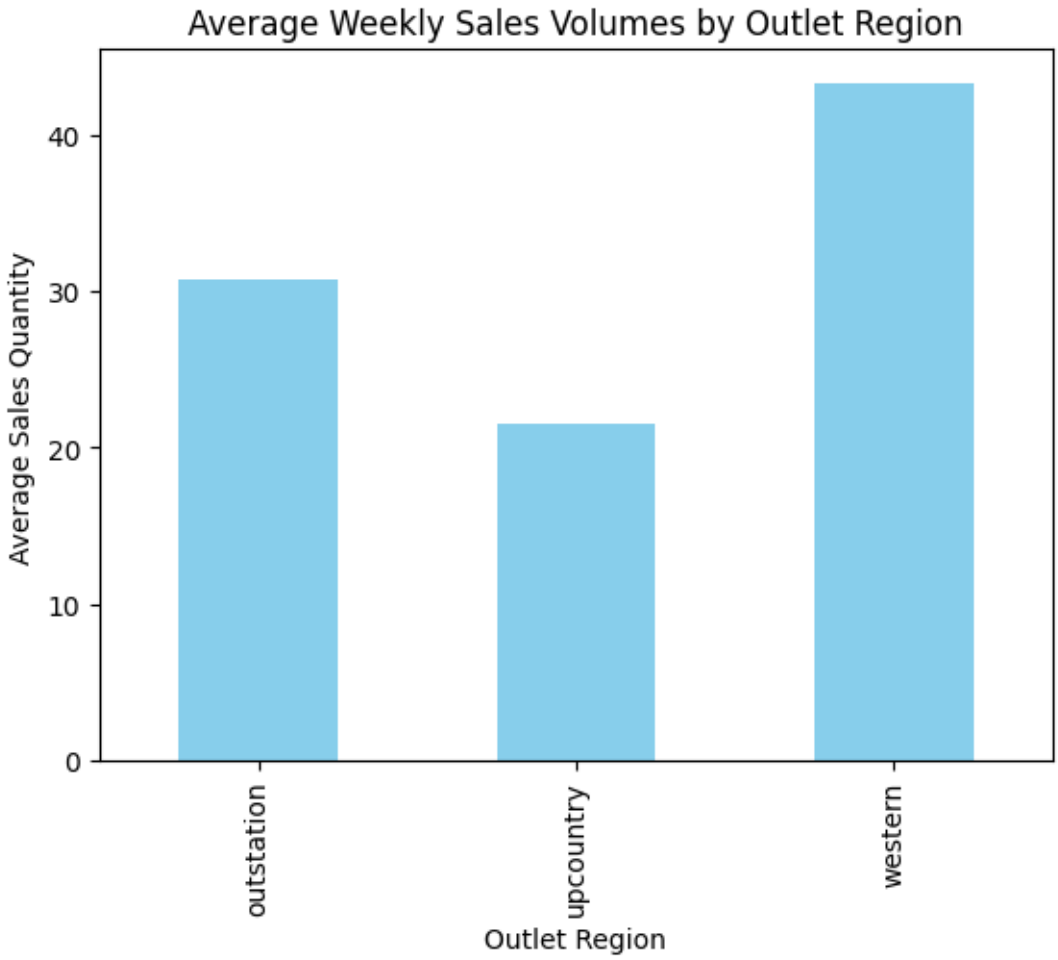

Image 01 - Average weekly sales volumes for each outlet region

Impact of Rainfall on Weekly Sales Volumes:

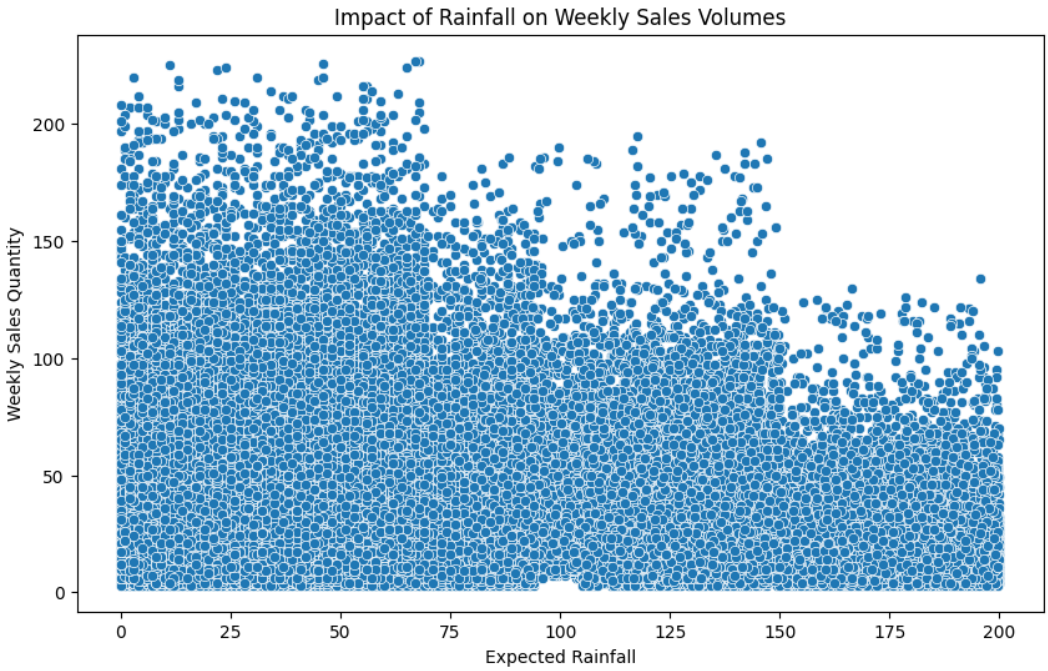

Image 02 - Impact of rainfall on the weekly sales volume

The graph upper shows an impact of rainfall on weekly sales volumes. It shows a weak positive correlation between rainfall and sales. This means that as rainfall increases, sales also increase. But the correlation is not very strong, so there is still a lot of variability in the data. For example, at around 50mm of rainfall, there is a range of sales from 50 to 150 units. This means that even at the same level of rainfall, there can be a big difference in sales.

According to results we can say that a light drizzle might not have much of an impact on sales, but a heavy rainfall could keep people away from the store. Another possibility is that the impact of rainfall varies depending on the time of year. For example, rainfall might be an issue during the new year season or any other festival seasons. Overall, the graph suggests that there is a weak positive correlation between rainfall and sales. However, there is still a lot of variability in the data.

Weekly Sales Trends Across Different Outlet Regions:

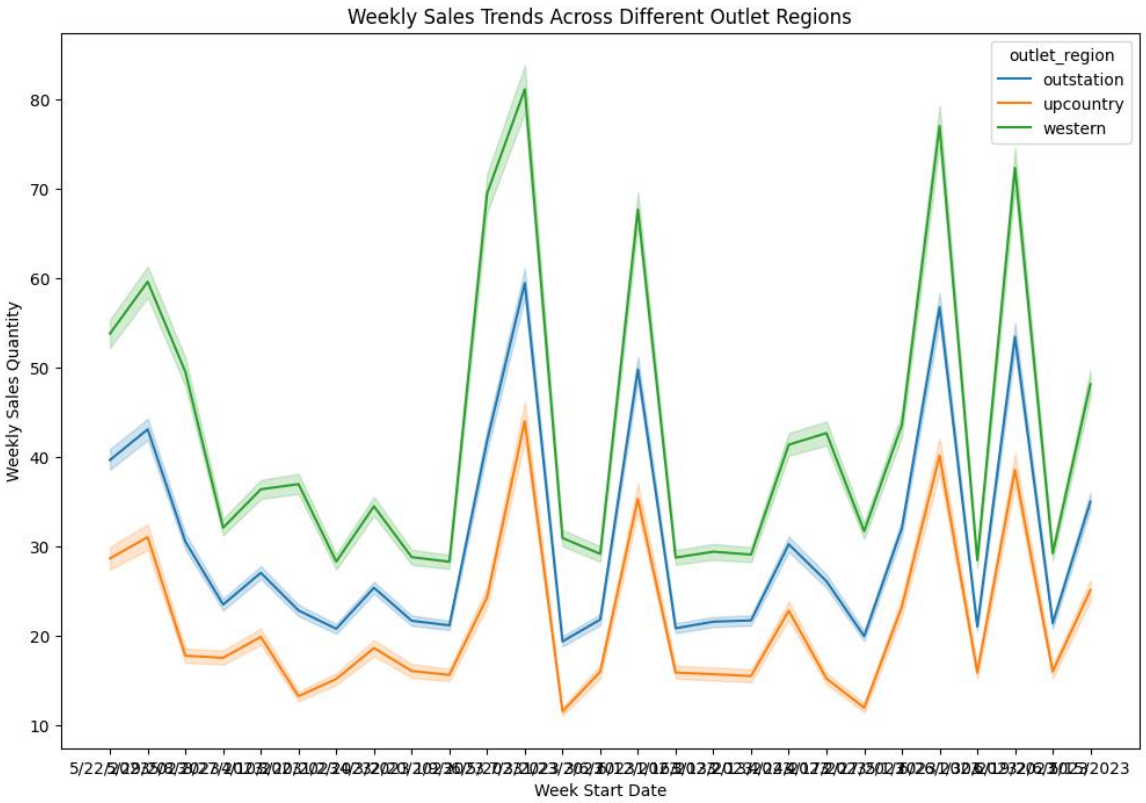

Image 03 - Weekly sales trends across different outlet regions

Outstation: Sales in the outstation region tend to be the highest, followed by upcountry and then western.

Upcountry: Sales in the upcountry region show a slight upward trend over time.

Western: Sales in the western region show a more volatile pattern, with some sharp peaks and dips.

Correlation between rainfall and sales. For example, in the week starting on 5/22, there was a relatively high amount of rainfall (around 70mm), and sales in all three regions were also relatively high.