Data Glacier Internship – Project

Batch: LISUM41 (30 December, 2024 – 30 March, 2025)

Team Member Details:

Name: Mahnoor Farhat

Email: mahnoor.farhat@gmail.com

Country: United Kingdom

College: University of Hertfordshire (Sep 2023 – Oct 2024)

Specialization: Data Science

Business Description:

The large company who is into beverages business in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed forecast of each of products at item level every week in weekly buckets.

Problem Description:

The time series data showed a range of patterns, some with trends, some seasonal, and some with neither. At the time, they were using their own software, written in-house, but it often produced forecasts that did not seem sensible. Company wanted to explore power of AI/ML based forecasting to replace their in-house local solution.

Project Lifecycle:

Week 8: Data Understanding (20 Feb – 26 Feb)

- **Data Collection:**
  - Identify data sources and gather data.
- **Data Exploration:**
  - Assess data types and distributions.
  - Identify missing values (NA values), outliers, skewness, and inconsistencies.
- **Data Quality Issues & Resolution Approaches:**
  - Handle missing values using mean, median, mode, or model-based imputation.

  - Identify and manage outliers using IQR, Z-score, or model-based anomaly detection.
  - Address skewness through transformations.
  - Conduct feature engineering to enhance dataset utility.

Week 9: Data Cleansing & Transformation (27 Feb – 2 March)

- **Techniques Applied:**
  - Handle missing values (e.g., mean imputation vs. model-based imputation).
  - Outlier detection using statistical and machine learning methods.
  - NLP preprocessing (if applicable) using regex, tokenization, and vectorization.
- **Collaboration:**
  - Code review among team members.
  - Maintain documentation and comments in the repository.

Week 10: Exploratory Data Analysis (EDA) & Insights (3 March – 9 March)

- **Visualizations & Trends:**
  - Generate time series plots, correlation heatmaps, and seasonal decomposition.
  - Identify trends, seasonality, and anomalies.
- **Final Recommendations:**
  - Feature selection and engineering decisions for modeling.

Week 11: EDA Presentation for Business Users (10 March – 16 March)

- **Presentation Structure:**
  - Business insights with visualizations.
  - Final slide with technical details and recommended forecasting models.

Week 12: Model Development & Benchmarking (17 March – 23 March)

- **Base Model Selection:**
  - Develop a simple forecasting model as a baseline.
- **Advanced Model Development:**
  - Train at least one model from each family.

- **Performance Evaluation:**
  - Use Weighted MAPE as the key metric.
  - Select the best model based on accuracy and explainability.
- **Parallel Computing Considerations:**
  - Optimize PySpark implementation (if applicable) for faster execution.

Week 13: Final Report & Submission (24 March – 30 March)

- **Consolidation of Work:**
  - Merge code and document findings.
- **Final Deliverables:**
  - Report detailing methodology, challenges, results, and recommendations.
  - PowerPoint presentation for stakeholders.
  - Submission of GitHub repository link with documented code.