Initial Draft

Data Science Project Lifecycle

Group B11

Sivaraj Krishnadayal
Sandew Mahawatta
Roshen Liayara
Yasiru Witharana

Initial Draft

Project Primary Goal

To develop an analytical solution that accurately classifies new **KJ Marketing** customers into predefined customer segments based on their purchasing behavior, enabling personalized marketing strategies.

Company Overview

KJ Marketing is a leading retail supermarket chain in Sri Lanka which runs 22 supermarket outlets across both urban and suburban regions of the country, offering a diverse product range including:

- **Dry Goods (**non-perishable items like grains, canned goods, beverages, and snacks**)**
- **Fresh Items** (perishable goods like fruits, vegetables, dairy, meat, and seafood)
- **Luxury Products** (premium, high-end consumer goods)

The company aims to improve its marketing strategies by shifting from traditional approach to a more personalized, data-driven strategy.

Business Objectives

1) Personalized Marketing Strategy
   - Enhance marketing effectiveness by tailoring promotions and engagement strategies based on customer purchasing behavior.

2) Customer Segmentation
   - Develop an analytical model to classify new customers into relevant segments for targeted marketing.

3) Sales Optimization
   - Identify high-value customer groups and improve the sales of luxury, fresh, and dry items by targeting them more effectively.

4) Customer Retention & Loyalty
   - Improve customer satisfaction and engagement by offering relevant products and promotions based on purchasing patterns.

5) Data-Driven Decision
   - Use insights from customer segments to optimize product offerings and expand into new markets effectively.

Methodology

The dataset provided by KJ Marketing consists of historical sales data, including average monthly sales per customer across different product categories and outlets.

This dataset has two sets which are the test set and the train set.

The **train data set** contains **774,156 rows** and **6 columns**, while the **test set** contains **40,750 rows** and **5 columns**. The test set does not include the cluster_category column.

The train set is used for model training and testing, while the test set is used to evaluate the model.

Data Preprocessing Strategy

In our initial analysis, we found that each variable has less than 1% null values. Therefore, we will apply mean imputation for numerical variables and mode imputation for categorical variables. Additionally, we will remove any rows that contain missing values in the customer_id column.

Outliers will be identified using the interquartile range method. Depending on the context, we may cap extreme upper and lower values or choose to retain them.

We will check each column to determine whether it has any typographical errors and correct them.

After cleaning the dataset, we will conduct exploratory data analysis (EDA) to identify patterns, trends, and insights about the data. For this, we will use univariate, multivariate analysis.

Modelling Technique

The goal of this project is to classify new customers into predefined segments based on purchasing behavior. For this purpose, we have chosen classification approach, because we typically use classification model to predict categorical variables. However, we still haven't figured out which type of classification model to use.

Tools & Technologies

We will use the following tools for data preprocessing, modelling, and visualizations:

Component	Tools & Libraries
Data Wrangling	Python (Pandas, Numpy)
Data Visualization	Python (Matplotlib, Seaborn), Power BI
Model Development	Python (Scikit-learn)
Modal Evaluation	Precision, Recall, F1- score
Version Control & Collaboration	Git, GitHub

Project Planning

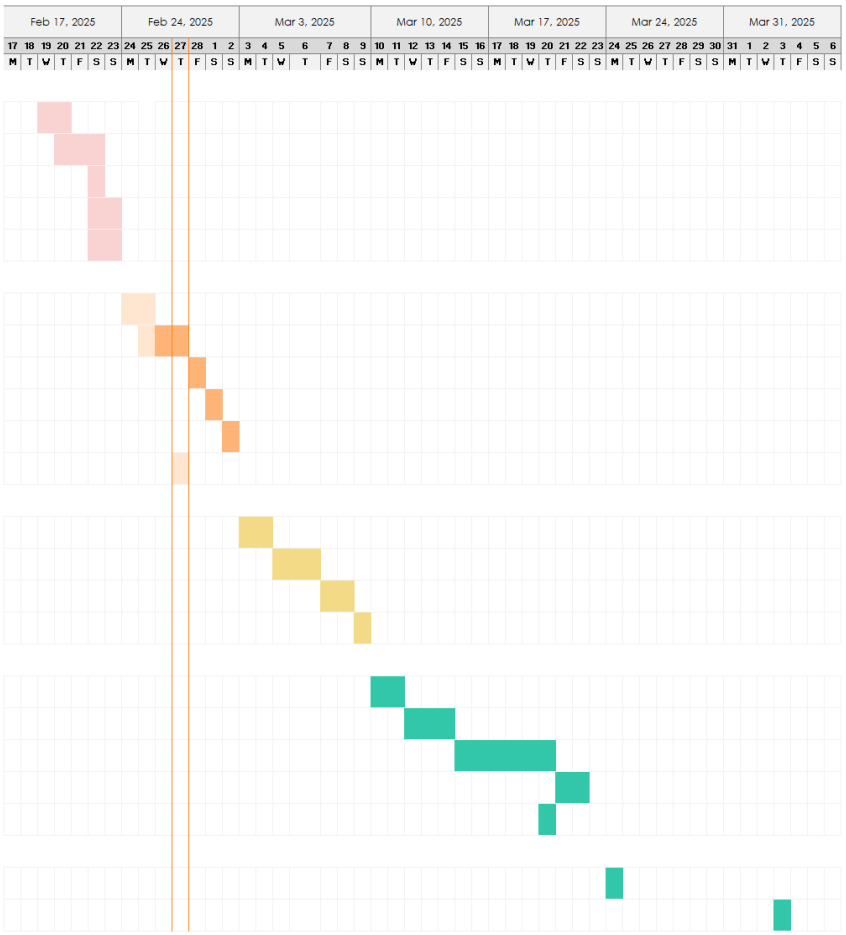
KJ Marketing Project

Group B11 Project lead: Krishnadayal

Project start: 19 February 2025

Display week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Project Definition & Planning				
Understanding the business problem	All members	100%	19-02-25	20-02-25
Researching existing solution	Sadivya	100%	20-02-25	22-02-25
Defining project scope & objectives	Yasiru	100%	22-02-25	22-02-25
Assigning team roles & project schedule	Krishna	100%	22-02-25	23-02-25
Risk assessment & mitigation strategies	Roshen	100%	22-02-25	23-02-25
Data Acquisition & Preprocessing				
Data exploration	Yasiru	100%	24-02-25	25-02-25
Data cleaning	Roshen	40%	25-02-25	27-02-25
Data intergration	Sadivya		28-02-25	28-02-25
Data reduction	Sadivya		01-03-25	01-03-25
Data transformation	Krishna		02-03-25	02-03-25
Initial Draft Project Analysis Submission	Sadivya	100%	27-02-25	27-02-25
Exploratory Data Analysis				
Descriptive statistics	Roshen		03-03-25	04-03-25
Univariate anlaysis	Sadivya		05-03-25	06-03-25
Multivariate analysis	Roshen		07-03-25	08-03-25
Documentation of key insights	Sadivya		09-03-25	09-03-25
Model Development & Deployment				
Select appropriate modelling technique	Krishna		10-03-25	11-03-25
Generate test design	Yasiru		12-03-25	14-03-25
Build model	Krishna & Yasiru		15-03-25	20-03-25
Evaluate model	Krishna & Yasiru		21-03-25	22-03-25
Draft Project Report for Client Submissio	Roshen		20-03-25	20-03-25
Communication				
Presentation	All members		24-03-25	24-03-25
Final Project Report Submission	All members		03-04-25	03-04-25



Potential Challenges and Mitigation Strategies

- The team could go through Business Understanding Challenges due to lack of domain knowledge and understanding of the retail industry. so at first, we will have do to enough research for a better understanding.
- We might face Modeling challenges when selecting the most appropriate model, in this case, we would start with simple models and gradually increase complexity
- Project management challenges could occur while dealing with team members with different skill sets and levels, we can overcome this by holding weekly team meetings with clear agendas and clear roles for each member.