Product Revenue Prediction

TEAM MEMBERS

1. Saichand Thota – A02394497 (Grad Student)
2. Preethi Reddy Tera – A02394358 (Grad Student)
3. Yagnashree Velanki – A02395651 (Grad Student)

DESCRIPTION OF THE PROJECT

The sales data for 1559 products across 10 stores in different cities has been collected by the data scientists at Big Mart. It has become common for malls and marts worldwide to keep track of their product sales data in order to predict future demand and manage their inventory. The primary goal of this project is to gain insights into consumer behaviour and to optimise product placement and promotions. Based on transactional data, identifying the items that are frequently purchased together and examining the relationships between different products, compare customers' preferences between the outlets and scope of introducing the new products in outlets.

DATASET

The "Product Revenue Prediction" dataset, sourced from Kaggle, contains information on the products sold at a mart, as well as outlet details and sales values. The dataset comprises two CSV files: train.csv and test.csv, with 8523 and 5681 records, respectively. The dataset encompasses product attributes such as unique ID, weight, fat content, product visibility, category, maximum retail price, outlet ID, establishment year, store size, city type, and store type. The target variable is the sales of the product in a particular outlet. This dataset is suitable for analysing sales and identifying the factors that influence sales at a particular outlet. The dataset is available at the following URL: **https://www.kaggle.com/datasets/akashdeepkuila/big-mart-sales.**

IMPLEMENTATION PLAN

I. ***Data Collection*:** Data is collected from Kaggle website. Link of the dataset is present in the Dataset section.

II. ***Pre-processing Techniques*:** The dataset has two columns with null values, Item_Weight and Outlet_Size, which will be addressed either by eliminating or by substituting them with mean or random values. For example, for the Item_Weight column, null values were replaced with the mean value of all other values in the column, while for the Outlet_Size column, the mode value of all other values in the column was used. When analysing the dataset in deeper, irrelevant columns will be removed or merged with other columns as needed.

   a. ***Feature selection*:** Based on the correlation value, we will be selecting the Features as per the requirement of the analysis. Features which are not required will be removed.

   b. ***Removing Null values*:** Null values will be replaced either by mean or mode of the column as per the requirements. Such as if we need to replace the null values from column which represents weight, we will replace them with the Mean values and if null values are in outlet details, then the most repeated value i.e., mode will be used to replace the null values

III. ***Models*:** To gain insights into the sales volume of outlets based on their location and established year, as well as insights into sales data regardless of outlet details, we wanted

to use machine learning algorithms such as linear regression, DecisionTreeRegressor, and random forest regressor to analyze the dataset. We also want to analyze various factors that affect product sales, such as nutritional value, cost, packaging, and fat content. Lastly, compare the outlets which have similar products and suggest the product – introducing of new product (from the available dataset), into the outlet which might increase the outlet sales.

ROLE OF MEMBERS

Methods	**Done by**
Data Collection	Preethi
Pre-processing	Yagnashree
Analysis and Implementing Models	Saichand, Preethi, Yagnashree
Accuracy testing	Saichand