**Approach Note: Big Mart Sales Prediction Project**

Objective: The primary objective of this project was to develop a predictive model to forecast `Item\_Outlet\_Sales` for various products across different Big Mart stores, based on the provided training and testing datasets. This task was approached as a supervised regression problem.

**1. Project Initiation and Data Understanding:**

The project commenced with loading the provided `bm\_Train.csv` and `bm\_Test.csv` datasets using pandas. Initial steps involved examining the dimensions (`.shape`), data types (`.dtypes`), and summary statistics (`.describe()`) of both datasets to gain a preliminary understanding of their structure and content. Copies of the original dataframes were made to preserve the initial state. For consistent preprocessing, the training and testing datasets were concatenated into a single dataframe.

Prior to deep diving into the data, a hypothesis generation phase was undertaken. This involved brainstorming potential factors influencing sales at both the store level (e.g., location, size, type, operational tenure) and the product level (e.g., type, weight, visibility, fat content). These hypotheses served as guiding principles for subsequent data exploration and feature engineering.

**2. Exploratory Data Analysis (EDA):**

Comprehensive exploratory data analysis was performed to understand the distribution of variables and identify relationships.

\* **Univariate Analysis**: The distribution of the target variable, `Item\_Outlet\_Sales`, was visualized using a histogram, revealing a skewed distribution. Categorical features such as `Item\_Fat\_Content`, `Item\_Type`, `Outlet\_Identifier`, `Outlet\_Size`, `Outlet\_Location\_Type`, and `Outlet\_Type` were analyzed using value counts and bar plots to understand the frequency and distribution of their respective categories. This analysis highlighted inconsistencies in categorical entries (e.g., `Item\_Fat\_Content`) and provided insights into the composition of items and outlets.

\* **Bivariate Analysis**: Relationships between key categorical features were explored using cross-tabulations and stacked bar plots. Visualizations depicting the relationship between `Item\_Fat\_Content` and `Outlet\_Identifier`, and `Item\_Type` and `Item\_Fat\_Content` were generated to identify potential interactions or dependencies between these variables.

3. **Data Preprocessing and Cleaning:**

Addressing data quality issues was a critical step.

\* **Missing Value Imputation**: Missing values were identified using `.isnull().sum()`. `Item\_Weight` contained missing values, which were imputed using the mean of the column. Rows with a value of 0 in `Item\_Weight` were initially treated as missing (`NaN`) before imputation. `Outlet\_Size` also had missing values, which were imputed using the mode of the column, a suitable strategy for this categorical/ordinal feature. Additionally, 0 values observed in the `Item\_Outlet\_Sales` column were imputed using the mode of the column.

\* **Categorical Standardization**: Based on EDA findings, inconsistent entries in `Item\_Fat\_Content` ('LF', 'low fat', 'reg') were standardized to 'Low Fat' and 'Regular' to ensure uniformity.

4. **Feature Engineering**:

New features were created and existing ones transformed to potentially enhance model performance.

\* A new categorical feature was derived from `Item\_Identifier` by extracting the first two characters and mapping them to broader categories: 'Food' (FD), 'Non\_Consumable' (NC), and 'Drinks' (DR).

\* The operational tenure of each outlet, `Outlet\_Years`, was calculated by subtracting `Outlet\_Establishment\_Year` from the year 2013 (as per the code's logic).

5. **Data Preparation for Modeling**:

To prepare the data for machine learning algorithms:

\* Categorical features were converted into a numerical format using one-hot encoding (`pd.get\_dummies`). This created binary columns for each category, expanding the feature space.

\* The combined dataframe was split back into the original training and testing sets based on the initial data dimensions.

\* The training data was separated into features (`x`) and the target variable (`y`, `Item\_Outlet\_Sales`).

\* Finally, the training data (`x`, `y`) was partitioned into training (`x\_train`, `y\_train`) and validation (`x\_test`, `y\_test`) sets using `train\_test\_split` with a test size of 30% (0.3) for model training and evaluation.

6. **Model Selection and Experimentation**:

A range of regression models were experimented with to identify a suitable approach for predicting sales. The models trained and evaluated included:

\* Linear Regression

\* AdaBoost Regressor

\* Random Forest Regressor

\* Decision Tree Regressor

\* Support Vector Regressor (SVR)

\* A simple Neural Network implemented using TensorFlow, comprising multiple dense layers with ReLU activation and a linear output layer, trained using Gradient Descent.

Each model was trained on the `x\_train`, `y\_train` split and evaluated on the `x\_test`, `y\_test` validation set.

7. **Evaluation Methodology:**

Model performance was primarily evaluated using the Root Mean Squared Error (RMSE) calculated on the predictions made on the validation set (`y\_test`). RMSE provides a measure of the average magnitude of the errors. The R2 score, indicating the proportion of variance in the dependent variable predictable from the independent variables, was also calculated for some models. For the Neural Network, the training process involved monitoring the Mean Squared Error (MSE) (the cost function) over epochs for both the training and validation sets, visualized through cost plots.

**Conclusion**:

This approach involved a systematic workflow, starting with understanding the problem and data through hypothesis generation and initial exploration. Rigorous data cleaning and preprocessing steps, including imputation and standardization, were performed. Feature engineering was applied to derive potentially informative variables. The data was appropriately prepared for modeling through encoding and splitting. A diverse set of regression models were experimented with, and their performance was evaluated using standard metrics, primarily RMSE, to assess their predictive capability on unseen data.