Big-Mart Sales Prediction

1. Objective

The goal of this project is to build a predictive model to forecast the sales of various products across different Big-Mart stores. The process involves comprehensive Exploratory Data Analysis (EDA), Feature Engineering to transform the raw data into a model-ready format, and evaluating multiple Regression Models to find the best sales predictor.

1. Objective

The goal of this project is to build a predictive model to forecast the sales of various products across different Big-Mart stores. The process involves comprehensive **Exploratory Data Analysis (EDA)**, **Feature Engineering** to transform the raw data into a model-ready format, and evaluating multiple **Regression Models** to find the best sales predictor.

Column	Non-Null Count	D-type
Item_Identifier	5966	object
Item_Weight	5966	float64
Item_Fat_Content	5966	object
Item_Visibility	5966	float64
Item_Type	5966	object
Item_MRP	5966	float64
Outlet_Identifier	5966	object
Outlet_Establishment_Year	5966	int64
Outlet_Size	**4276**	object
Outlet_Location_Type	5966	object

Column	Non-Null Count	D-type
Outlet_Type	5966	object

3. Exploratory Data Analysis (EDA)

Numerical Feature Analysis

Numerical features were analyzed for distribution and outliers.

- Item_Weight (eda_item_weight.png): Shows a relatively normal distribution with a wide range, indicating variability in product weights. The missing values were successfully imputed.
- Item_Visibility (eda_item_visibility.png): Highly skewed towards zero, indicating a large number of items with very low visibility. This was addressed by replacing zero values with the median of non-zero visibility in the feature engineering step.
- Item_MRP (eda_item_mrp.png): Exhibits a multimodal distribution, suggesting products are grouped into different price tiers.

Categorical Feature Analysis

Analysis of categorical features was performed to understand data distribution and prepare for encoding.

- Item_Fat_Content (eda_item_fat_content.png): Revealed inconsistent labeling (e.g., 'low fat', 'LF', 'Low Fat') which was corrected to two labels: 'Low Fat' and 'Regular'.
- Outlet_Identifier (eda_outlet_identifier.png): Shows 10 unique outlets, each with a different frequency in the dataset.
- Outlet_Size (eda_outlet_size.png): The missing values were visible as a separate bar before imputation.
- Outlet_Type (eda_outlet_type.png): Shows a clear dominance of 'Supermarket Type1', followed by 'Grocery Store'.

4. Feature Engineering and Preprocessing

The following transformations were applied to prepare the data for modeling:

1. **Imputation of Outlet_Size**: Missing values were imputed with the new category 'Missing'.

2. **Outlet_Age Creation**: A new numerical feature was created by calculating the age of the outlet: 2025 – Outlet_Establishment_Year. The original year column was dropped.
3. **Item_Type_Combined Creation**: A new feature was created by classifying items into three broader categories based on Item_Identifier prefix: 'Food', 'Drinks', and 'Non-Consumable'. The original Item_Identifier and Item_Type were dropped.
4. **Item_Visibility Zero Handling:** Zero values were replaced with the median of the non-zero Item_Visibility values to correct for records where visibility was likely missing or misrecorded.
5. **Encoding:**
    - Item_Fat_Content was transformed using Label Encoding (0 and 1).
    - Remaining categorical columns (Outlet_Identifier, Outlet_Size, Outlet_Location_Type, Outlet_Type, Item_Type_Combined) were transformed using One-Hot Encoding (pd.get_dummies).

Transformation	**Feature(s)**	**Method**
Handling Missing Values	Outlet_Size	**Imputed** with the category **'Missing'**.
New Feature Creation	Outlet_Age	Calculated as **$2025 - \text{Outlet\_Establishment\_Year}$**.
New Feature Creation	Item_Type_Combined	Extracted the prefix from Item_Identifier to group items into **'Food'**, **'Drinks'**, or **'Non-Consumable'**.
Handling Zeros	Item_Visibility	Zero values were replaced with the

Transformation	Feature(s)	Method
		median of all non-zero Item_Visibility values.
Categorical Encoding	Item_Fat_Content	Transformed using **Label Encoding**.
Categorical Encoding	Outlet and Item type features	Transformed using **One-Hot Encoding** (pd.get_dummies).
Feature Dropping	Item_Identifier, Item_Type, Outlet_Establishment_Year	Dropped as their information was captured in new or encoded features.

The final processed training dataset had 5966 rows and 29 columns.

5.Model Building and Evaluation

Three different regression models were trained and evaluated on the test set using Root Mean Squared Error (RMSE) and R-squared (R^2) Score.

Model	RMSE	R2_Score
Random Forest	**1058.3911**	**0.6001**
Decision Tree	1071.7407	0.5899
Linear Regression	1097.8569	0.5697

Key Finding

The Random Forest Regressor demonstrated the best performance among the models successfully executed, achieving the highest R-squared score of 0.6001 and the lowest RMSE of 1058.3911. This indicates that an ensemble, non-linear approach is best suited for modeling sales data.

6. Conclusion

The project successfully executed the sales prediction workflow from data cleanup to model deployment.

1. Data Quality issues (missing weights, missing outlet sizes, inconsistent fat content, and zero visibility) were successfully resolved through careful imputation and cleaning.
2. Feature Engineering effectively created predictive features like Outlet_Age and simplified high-cardinality features into Item_Type_Combined.
3. The Random Forest Regressor is the recommended model, showing an R-squared of 0.6001, indicating it can explain approximately 60% of the variance in sales.