**Big Mart Sales Prediction: Summary of Approaches Considered**

**1. Objective**

The goal of this project is to build a predictive model that accurately estimates the sales of products across various Big Mart outlets using historical sales and product data.

**2. Data Understanding and Preprocessing (EDA)**

* Data Read:
  + Read train and test data csv files provided
  + Performed transformation on both train and test data sets to create the MBT data for modeling
* Missing Value Treatment:
  + Imputed missing values in Item\_Weight using the mean of similar items.
  + Filled missing Outlet\_Size values based on mode per Outlet\_Type
* Outlier detection:
  + Outlier detection in numerical columns and cap the values of the outlier columns within the IQR range
* Data Cleaning:
  + Corrected inconsistent entries in Item\_Fat\_Content (e.g., combining “LF” and “low fat”).
  + Addressed zero values in Item\_Visibility by replacing with the average visibility for that item.
  + Relook into remaining columns to have overall summary and visualize the data
  + Bring the characteristics of 'Item\_Outlet\_Sales' with respect of Item\_identifier and Outlet\_Identifier
* Categorical Encoding:
  + Applied Label Encoding for ordinal variables like Outlet\_Size.
  + Used One-Hot Encoding for nominal variables such as Item\_Type, Outlet\_Type, and Outlet\_Location\_Type

3. Feature Engineering

* Created new features to enhance model accuracy:
  + Outlet\_Years = 2013 - Outlet\_Establishment\_Year
  + Item\_Visibility\_MeanRatio = Item\_Visibility / mean visibility for that item
  + Binned Item\_MRP into tiers to reduce variance

4. Modeling Approaches

* + Split train and test within main train data as 80-20 split and evaluated RMSE, R2, MSE on all chosen algorithms
* Baseline Model:
  + Linear Regression to set a performance benchmark
* Advanced Models Tested:
  + Decision Tree Regressor
  + Random Forest Regressor
  + XGBoost Regressor
  + LightGBM Regressor
* Model Optimization:
  + Performed hyperparameter tuning using GridSearchCV and RandomizedSearchCV where applicable

**5. Evaluation Metrics**

* Models evaluated based on:
  + Root Mean Squared Error (RMSE)
  + R-squared (R² Score)
  + Mean Absolute Error (MAE)
* Performance Summary:
  + Random Forest with hyper parameter tuning performed significantly better than linear models

Bottom of Form