**Hackathon Submission Documentation**

**1. Final Script / GitHub Repository**

**Deliverable:** A script or GitHub repository containing all relevant code used in this project.

* File: hackathon\_final\_1.py
* Description: This script includes the following components:
  + Data loading and initial exploration
  + Handling missing data and imputations
  + Feature engineering and transformation
  + Model experimentation and final selection
  + Predictions and submission generation

**Note:** The complete script can be found in the provided file hackathon\_final\_1.py, accessible via the Colab notebook link: [Hackathon Script Notebook](https://colab.research.google.com/drive/1BhMtbhooumrcz-9dtzThPU8Wdjegy7zB?usp=sharing). This includes all preprocessing, feature engineering, and modeling steps.

**2. Notebook with EDA and Feature Engineering**

**Deliverable:** A Jupyter notebook (or similar format) documenting the exploratory data analysis (EDA) and feature engineering steps.

**Key Components:**

1. **EDA:**
   * **Target Variable Exploration:**
     + Visualized the distribution of Item\_Outlet\_Sales using histograms and boxplots to understand skewness and central tendency.
     + Created boxplots grouped by Outlet\_Type to compare sales distributions.
     + Examined correlations between numerical features and Item\_Outlet\_Sales using a heatmap.
   * **Categorical Analysis:**
     + Countplots for Outlet\_Size, Item\_Type, and Price\_Segment to understand their distributions.
     + Boxplots of Item\_Category against Item\_Outlet\_Sales to explore group-level differences.
2. **Handling Missing Data:**
   * Imputed missing Item\_Weight values using the mean of corresponding Item\_Type.
   * Filled missing Outlet\_Size values based on the mode of Outlet\_Type.
3. **Feature Engineering:**
   * Derived Outlet\_Age as the difference between the current year (2024) and Outlet\_Establishment\_Year.
   * Created Price\_Per\_Visibility as Item\_MRP / Item\_Visibility to capture price visibility dynamics.
   * Applied log transformations to Item\_Visibility to reduce skewness.
   * Categorized items into broader Item\_Category (e.g., Food, Beverages, Non-Food).
   * Segmented Item\_MRP into Price\_Segment using quantiles (Budget, Economy, Mid\_Range, Premium).

**Objective:** Showcase the thought process and rationale behind these transformations, supported by visualizations and summaries. The full notebook, including visualizations and EDA steps, is accessible here: [Big Mart Sales Prediction Notebook](https://colab.research.google.com/drive/1BhMtbhooumrcz-9dtzThPU8Wdjegy7zB?usp=sharing).

**3. Modeling Scripts/Notebooks**

**Deliverable:** A collection of scripts or notebooks documenting the modeling phase.

**Key Components:**

1. **Model Experimentation:**
   * **Models Evaluated:**
     + Linear Regression
     + Decision Tree Regressor
     + Random Forest Regressor
     + XGBoost Regressor
     + Neural Network
   * Evaluated using K-Fold Cross-Validation with RMSE as the primary metric.
2. **Hyperparameter Tuning:**
   * Applied GridSearchCV for XGBoost:
     + Parameters tuned: n\_estimators, learning\_rate, max\_depth.
     + Optimal parameters identified through cross-validation.
3. **Model Insights:**
   * Linear Regression struggled with feature complexity.
   * Random Forest provided better generalization but required more computation.
   * XGBoost performed best with minimal overfitting, achieving the lowest RMSE on validation data.
4. **Final Model:**
   * Selected XGBoost with optimal parameters (n\_estimators, learning\_rate, and max\_depth).
   * Trained on the entire dataset for final predictions.
5. **Results:**
   * Performance metrics for all models (mean RMSE, standard deviation).
   * Saved final predictions in final\_submissionDay0.csv.

**4. Approach Note**

**Deliverable:** A concise 1-page document explaining the entire thought process and steps undertaken.

**Key Sections:**

1. **Problem Understanding:**
   * The objective was to predict Item\_Outlet\_Sales based on retail data, with challenges like missing data, skewness, and complex feature relationships.
2. **EDA Insights:**
   * Analyzed patterns and distributions to identify key trends.
   * Observed significant variance in sales by Outlet\_Type and Item\_Type.
3. **Feature Engineering:**
   * Derived new features like Outlet\_Age, Price\_Per\_Visibility, and categorized MRP into segments.
   * Standardized categorical features (Item\_Fat\_Content).
4. **Modeling:**
   * Experimented with multiple regression models and advanced techniques (e.g., XGBoost).
   * Tuned hyperparameters for optimal performance.
5. **Outcomes and Reflections:**
   * Final RMSE achieved with XGBoost: 20.54 (calculated from the XGBoost model during validation).
   * Highlighted learnings and potential improvements (e.g., ensembling models).

**Note:** This document references key results and visualizations from the shared Colab notebook: [Big Mart Sales Prediction Notebook](https://colab.research.google.com/drive/1BhMtbhooumrcz-9dtzThPU8Wdjegy7zB?usp=sharing). The approach combines structured scripts and detailed exploratory insights.

**Summary of Deliverables:**

1. **Code/Repository:** Full script with detailed comments.
2. **EDA/Feature Engineering Notebook:** Documentation of transformations and insights.
3. **Modeling Notebooks:** Scripts for model experimentation, tuning, and final selection.
4. **Approach Note:** High-level summary of the entire process.

By consolidating these deliverables, the submission will provide a comprehensive overview of the problem-solving journey and outcomes.