Summary of Approach for BigMart Sales Prediction

The **BigMart Sales Prediction** project aimed to build a predictive model that forecasts the sales of various products across different retail stores. The solution was developed through a **structured, step-by-step approach**, incorporating **data cleaning, feature engineering, model selection, hyperparameter tuning, and final predictions**. Below is a detailed summary of the key steps undertaken in this project.

📌 1. Understanding the Problem Statement
The dataset consisted of **sales data from BigMart outlets**, containing information about **products and store attributes**. The primary objective was to **predict Item_Outlet_Sales**, which is the total sales of a particular product at a specific store.

Key Challenges:
✅ Handling missing values
✅ Feature engineering to extract meaningful insights
✅ Encoding categorical variables
✅ Selecting the best model for sales prediction

📌 2. Data Cleaning & Preprocessing

🔹 Handling Missing Values
- **Item_Weight**: Filled missing values with the **mean weight** of the respective **Item_Identifier**.
- **Outlet_Size**: Imputed missing values using the **mode** of the corresponding **Outlet_Type**.

🔹 Standardizing Categorical Variables
- **Item_Fat_Content** was normalized (e.g., ‘LF’, ‘low fat’ → ‘Low Fat’).
- **Outlet_Age** was calculated as `2025 - Outlet_Establishment_Year`.

🔹 Addressing Data Skewness
- **Item_Visibility** had zeros, which were replaced with the **median visibility** of the respective category.

📌 3. Exploratory Data Analysis (EDA)
EDA was performed to understand the relationships between variables and sales.

🔹 Key Visualizations & Insights
✅ **Sales Distribution** → Right-skewed, indicating a few high-selling products contribute to most revenue.
✅ **Sales vs. Outlet Type** → Supermarket Type 3 had the highest median sales, whereas grocery stores had the lowest.
✅ **Item Type vs. Sales** → Food products dominated sales compared to drinks and non-consumables.
✅ **Item MRP vs. Sales** → A **positive correlation** indicated that higher-priced items tend to sell more.

📌 4. Feature Engineering
To improve model accuracy, **new features** were introduced:

- **Price_per_Unit_Weight** → Identified pricing efficiency.
- **Item_Visibility_Log** → Applied log transformation to normalize skewed data.
- **Outlet_Age_Category** → Grouped outlets into "Young," "Mid," and "Old"

categories.
- **Item_Category** → Extracted Food, Drinks, and Non-Consumables from **Item_Identifier**.
- **Non_Consumable Flag** → Created a binary indicator for non-food items.

📌 5. Model Training & Evaluation
Multiple machine learning models were trained and evaluated based on **Root Mean Squared Error (RMSE)**.

🔹 Baseline Model Results
Model	Train RMSE	Validation RMSE
Linear Regression	1141.31	1068.91
Decision Tree	0.00	1499.10
Random Forest	434.18	1091.85
Gradient Boosting	**1035.67**	**1040.11**

✔ **Gradient Boosting had the best RMSE**, making it the **best-performing model** at this stage.

📌 6. Hyperparameter Tuning
To further optimize the **Gradient Boosting Model**, **RandomizedSearchCV** was used.

✅ **Best Hyperparameters Found**:
- **n_estimators**: 300
- **learning_rate**: 0.01
- **max_depth**: 5
- **subsample**: 0.9

📊 **Optimized RMSE on Validation Set**: **1030.01**

📌 7. Advanced Models (XGBoost & LightGBM)
To improve performance, **XGBoost** and **LightGBM** were tested.

Model	Train RMSE	Validation RMSE	R² Score
XGBoost	890.15	1061.97	0.5851
LightGBM	944.81	**1045.83**	**0.5976**
Gradient Boosting (Prev. Best)	1035.67	**1040.11**	**0.6097**

🔹 **Gradient Boosting remained the best model**, achieving the highest **R² score** and lowest **RMSE**.

📌 8. Final Prediction on Test Data
After training the best model, **final sales predictions** were made on the test dataset.

📂 **Final Submission File**: `BigMartSales_Final_Predictions.csv`

✅ Predictions saved successfully!

📌 9. Performance Metrics
To evaluate the **final model**, the following metrics were computed:

Metric	Score
MAE	727.33
MSE	1,060,921.07
RMSE	1030.01
R² Score	0.6097

✅ **~61% of the variance in sales is explained by the model**, showing **reasonable performance** with room for improvement.

📌 10. Next Steps & Business Recommendations
✔️ **Optimize Pricing Strategy** → Introduce more products in **100-150 MRP** range.
✔️ **Improve Inventory Planning** → Reduce stockouts for high-selling items.
✔️ **Invest in High-Performing Outlets** → Expand **Supermarket Type 3** format.
✔️ **Category-Specific Promotions** → Boost marketing for **high-margin items**.

🚀 Conclusion
This project successfully developed a **robust machine learning model** to predict **BigMart sales** using **feature engineering, advanced models, and hyperparameter tuning**. The **Gradient Boosting Model** emerged as the best performer, achieving a **RMSE of 1030.01** and an **R² score of 0.6097**.

Future improvements could include **deep learning approaches, time-series forecasting, and additional feature engineering** to further enhance prediction accuracy. 🚀

🎯 Final Deliverables
📄 **Cleaned & Processed Data**
📊 **EDA & Visualizations**
🤖 **Trained Machine Learning Models**
📂 **Final Predictions File**
📈 **Performance Metrics Report**

📌 **Author:** Shivam Namdeo, Data Scientist
🔗 **GitHub Repository:**
https://github.com/sivm22/BigMartSalesPrediction_Shivam_Namdeo_Data_Scientist.git 🚀