Big Mart Sales Prediction Project - Approach Note

Problem Understanding:
The objective of this project was to predict the sales of products at various Big Mart outlets based on historical sales data and store/product attributes. The challenge involved identifying patterns in the data and developing a robust predictive model to provide accurate sales forecasts.

EDA Insights:
During exploratory data analysis (EDA), the following key insights were identified:

- Missing values were found in the Item_Weight and Outlet_Size columns, which were imputed using median and mode values, respectively.
- Outliers in the Item_Visibility feature were capped at the 99th percentile.
- The Outlet_Establishment_Year was transformed into a derived feature, Outlet_Age.
- Sales distribution showed positive skewness, prompting log transformation for better model fit.
- Strong correlations were observed between Item_MRP and the target variable Item_Outlet_Sales.

Feature Engineering:

- Created new features such as Outlet_Age, Item_Type_Combined, and Item_MRP_Band to capture hidden relationships.
- Encoding techniques included one-hot encoding for categorical variables (Outlet_Location_Type, Item_Fat_Content) and label encoding for ordinal features.
- Normalization was applied to numeric features to improve model performance.

Model Experiments:
Several models were experimented with to find the best performer:

1. **Linear Regression:** Provided a baseline model but struggled with complex feature relationships.
2. **Decision Trees:** Improved performance but prone to overfitting.
3. **Random Forest:** Achieved a balance between bias and variance, yielding better accuracy.
4. **XGBoost:** Performed the best after hyperparameter tuning, resulting in the lowest RMSE score.
5. **Grid Search:** Used for hyperparameter optimization, improving the final model's predictive capability.

Challenges Faced:

- Handling missing values without introducing bias.
- Selecting relevant features and minimizing noise during feature engineering.
- Balancing model complexity and interpretability while achieving low RMSE.

Final Model Selection:
The XGBoost model was chosen as the final model due to its superior accuracy and ability to generalize well on unseen data.

Key Achievements:

- Successfully reduced RMSE from the baseline model by over 25%.
- Achieved a competitive rank on the leaderboard.
- Developed a feature-rich and scalable solution for sales prediction.

Future Enhancements:

- Further feature engineering to capture product-seasonality patterns.
- Implementation of advanced ensembling techniques to improve model robustness.