Approach Note: Big Mart Sales Prediction

Objective

The goal of this project was to develop a robust machine learning model to predict sales for various products across different retail outlets. The approach included a structured pipeline of exploratory data analysis (EDA), preprocessing, feature engineering, model experimentation, and optimization.

Exploratory Data Analysis (EDA)

- Identified numerical and categorical features.
- Standardized categorical values in Item_Fat_Content, handled missing values using Multiple Imputation by Chained Equations (MICE), and imputed zero values.
- Examined sales distribution, outlet characteristics, and pricing trends.
- Used IQR-based filtering to remove anomalies while ensuring test data validity.

Preprocessing & Feature Engineering

- Applied ordinal encoding for categorical features.
- Created new features such as Outlet_Age, Price_per_Unit_Weight, and binned MRP categories.
- Applied RobustScaler to handle skewed distributions.
- Used log transformations on skewed features for better model performance.

Model Experimentation

- Tested RandomForest, XGBoost, LightGBM, and CatBoost.
- Evaluated models using RMSE on a 70-30 train-test split.
- CatBoost performed best, achieving an RMSE of 1135.9065.
- Used Optuna for hyperparameter tuning to optimize model performance.

Results & Final Submission

- Final model achieved the lowest RMSE and was used for predictions.
- Log-transformed predictions were reverted before submission.
- Predictions were saved in submission.csv for final evaluation.

Key Takeaways

- Addressing missing values improved model accuracy.
- Feature engineering significantly enhanced predictive performance.
- CatBoost was the most effective model for handling categorical data.
- Hyperparameter tuning refined the final model for better accuracy.

Conclusion

The final CatBoost model achieved an RMSE of 1135.9065, demonstrating strong predictive accuracy in forecasting sales.