Approach Note: Big Data Sales Prediction Hackathon (AV)
Objective: To forecast sales (Item_Outlet_Sales) by using historical retail data.

This involves data preprocessing, feature engineering and applying machine learning models.

1. Data Exploration & Cleaning

Initial Analysis: Analyzed the distribution of categorical variables and assessed sales performance for each category.

Observations:
- Item_Fat_Content had mixed labels like "LF", "low fat", and "reg" which were made consistent.
- Outlet_Size and Item_Weight had missing values that needed to be filled in.

2. Missing Value Imputation

Item_Weight:
- Tested different methods to fill missing values: median, group-wise median, KNN, and MICE.
- Evaluated each method using RMSE from simple ML model.
- The best method was group-wise median based on Item_Type and Item_Fat_Content.

Outlet_Size:
- Imputed using the most common value from combinations of outlet-related features like Outlet_Type, Outlet_Location_Type, and Outlet_Establishment_Year.

3. Feature Engineering

Encoding:
- Compared One-Hot Encoding and Label Encoding.
- Label Encoding worked better as it gave lower RMSE.

Scaling:
- Used StandardScaler before training the Artificial Neural Network (ANN) and other models.

4. Model Building & Evaluation

Models evaluated - Linear Regression, Random Forest, AdaBoost, Gradient Boosting, XGBoost, and ANN.
- Evaluated with RMSE on a validation set.
- Optimal model was ANN with the lowest RMSE score.

5. Hyperparameter Tuning

- Performed grid search with 5-fold cross-validation for the ANN model.
- Tuned parameters like learning rate, number of neurons, activation function, batch size, and number of epochs.

6. Final Model & Submission

- Retrained the ANN model on the complete training data using the best hyperparameters.
- Generated predictions on test data for the final submission.

Future Improvements
- Feature Reduction: Group some Item_Type categories to simplify the model.
- Feature Engineering – Explore deeper feature engineering techniques by merging existing variables or deriving new, more predictive features.
- Performing GridSearchCV for boosting algorithms
- Stacking/Ensembling: Use predictions from multiple models together for better results.