**Approach Note: Sales Prediction Hackathon**

* Problem Understanding:  
  Understood that it’s a supervised regression problem. The goal was to predict outlet-wise item sales using item and outlet features.
* Data Investigation and Cleaning:
  + Started by looking at summary statistics, missing values, and data types.
  + Found missing values, outliers, skewness, and inconsistent labels.
  + Used correlations to decide how to fill missing values.
    - Example- Estimated missing Item Weight using Item Identifier.
  + Standardized inconsistent labels in Item\_Fat\_Content.
  + Checked for outliers but kept them because tree-based models handle outliers well.
* Exploratory Data Analysis (EDA):
  + Discovered a non-linear relationship between item visibility and sales.
  + Visualized skewness, presence of zeros, and outliers in Item\_Visibility.
  + Decided to convert Item\_Visibility into a categorical feature based on these insights.
  + Found Item\_MRP to be strongly correlated with sales.
* Feature Engineering:
  + Binned Item\_Visibility into Zero, Low, Medium, High to better capture patterns.
  + Created additional features like zero visibility flag and outlet age.
  + Used Label Encoding for categorical variables for effective model training.
* Scaling and Preprocessing:
  + Applied StandardScaler on numerical features (Item\_MRP, Item\_Weight), fitting on train data and applying same transformation on test data.
  + Encoders and bin edges were fit on train dataset and reused for test data to avoid leakage.
* Model Experimentation and Selection:
  + Tested Linear Regression as baseline model.
  + Tried XGBoost and Random Forest; both showed strong performance.
  + Chose Random Forest for its accuracy.
* Hyperparameter Tuning:
  + Used RandomizedSearchCV to fine-tune key hyperparameters.
  + Tuning improved performance and prevented overfitting.
  + Cross-validation ensured model stability.
* Final Model:
  + Trained Random Forest with best hyperparameters on full train data.
  + Generated predictions for test data.
  + Achieved a top rank of 668 in the hackathon.