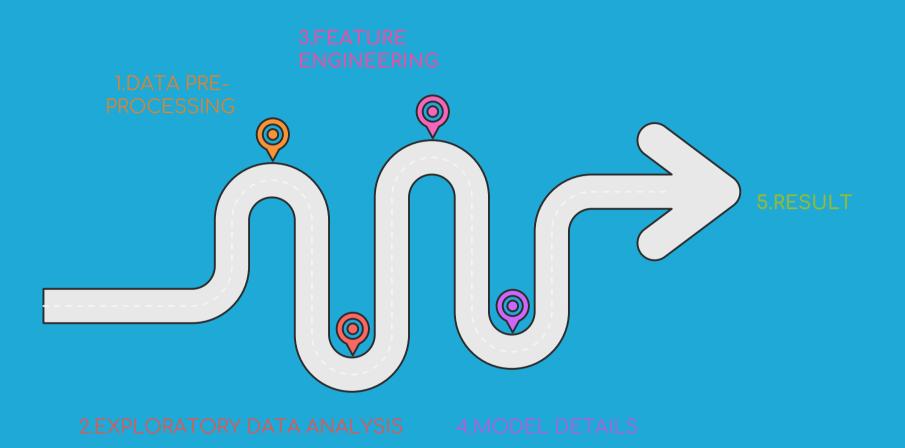
DEEP LEARNING PROJECT ON PRODUCT REORDER PREDICTION

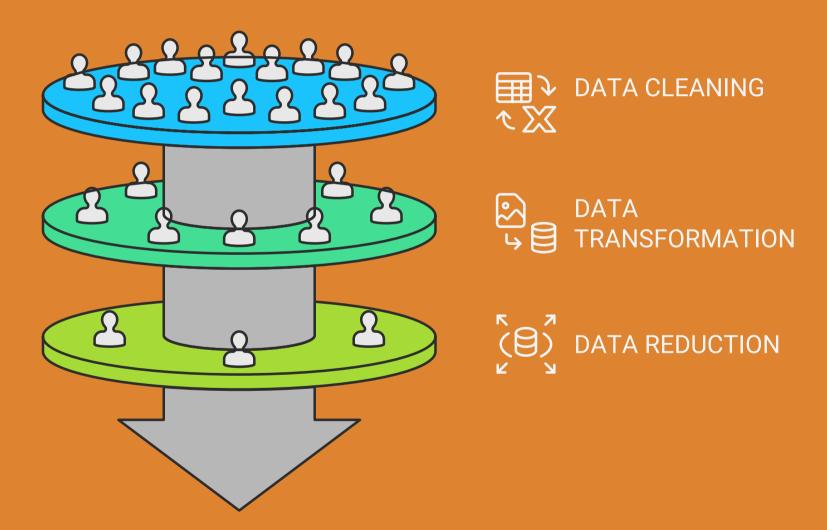
PROBLEM STATEMENT

The project involves building a Deep Learning model to predict product reordering behaviour by users. The aim is to identify patterns in purchasing habits and predict whether a product will be reordered in the next purchase cycle. The project includes data pre-processing, exploratory data analysis, feature engineering, Model building and Model deployment using Streamlit application.

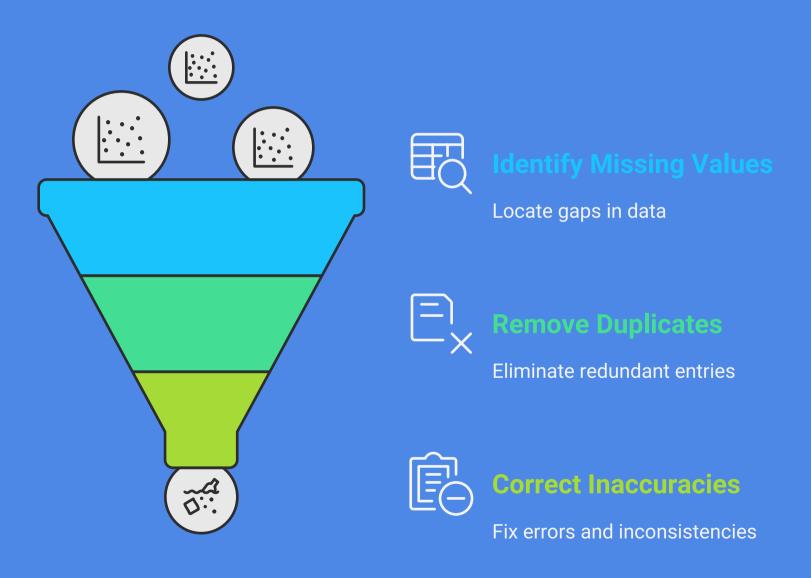
DATA SCIENCE WORK FLOW



DATA PRE-PROCESSING



DATA CLEANING PROCESS



How to handle duplicate data in machine learning?



Remove Duplicates

Ensures unique data points for accuracy



Keep Duplicates

May skew model results and predictions

Data Preprocessing and Feature Engineering Sequence



MICE Imputation It is an advanced statistical technique used to fill in (impute) missing data in a dataset by generating plausible values based on other available information.

Identify and remove duplicate

Duplicate Removal





Outlier checking is the process of identifying data points that significantly differ from the rest of the dataset, as they may indicate errors, variability, or rare events. Detecting outliers helps improve data quality and model accuracy by deciding whether to remove, transform, or analyze these unusual values separately. Z Score and IQR Method

Collinearity checking is the process of identifying whether two or more independent (predictor) variables are highly correlated with each other.

Collinearity Check





Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics using statistics and visualizations.

Feature engineering is the process of creating and transforming input variables to make data more useful for machine learning models. It helps improve model accuracy by providing better





representations of the data. Sampling is the process of selecting a subset of data from a larger population to analyze & make inferences about the whole. Random sampling means selecting samples entirely by chance, where every member of the population has an equal probability of being chosen, ensuring unbiased representation.



Scaling is the process of resizing data features to a common range or distribution to improve model performance. MinMax Scaling, scales features to a range between 0 and 1, useful when features need to be normalized.



Encoding is the process of

converting categorical variables into numerical values for use in machine learning models.OHE, Creates binary columns for each category, suitable for non-ordinal data.



One Hot Encoding

OUTLIER CHECKING



Z Score Method

Z-score detects
outliers using mean
and standard
deviation,
best for normally
distributed data.





IQR Method

IQR method uses the interquartile range and is more robust for skewed or nonnormal distributions.



Cycle of Exploratory Data Analysis

Univariate Analysis Analyze one variable at a time. Histogram and KDE plot

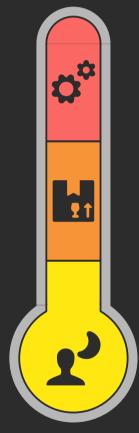


Statistical Analysis
Use summary stats
like mean, median,
variance, and
correlation,
Skewness and
Kurtosis

Graphical Analysis Histograms, scatter plots, and box plots.

FEATURE ENGINEERING

Combined



User-product

Focuses on specific user-product relationship.(how many times user reordered the product)

Product-based

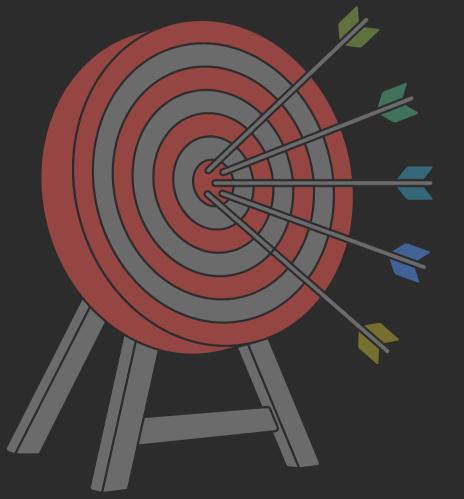
Focuses on inherent product characteristics. (reorder ratio, number of times reordered)

User-based

Focuses on individual purchasing behavior (average days between orders, total orders)

Individual

NEURAL NETWORK ARCHITECTURE





EVALUATION METRICS-F1 Score, Accuracy, Precision, Recall

Assess model performance



OPTIMIZER-Adam

Adjust model parameters



LOSS FUNCTION -Categorical Cross-Entropy

Measure prediction error



OUTPUT LAYER -¹∞ Softmax

Generate final predictions



INNER LAYERS -Relu

Process input data

PRODUCT REORDER PREDICTION

