STORE SALES PREDICTION

MANU VATS

OBJECTIVE & BENEFITS

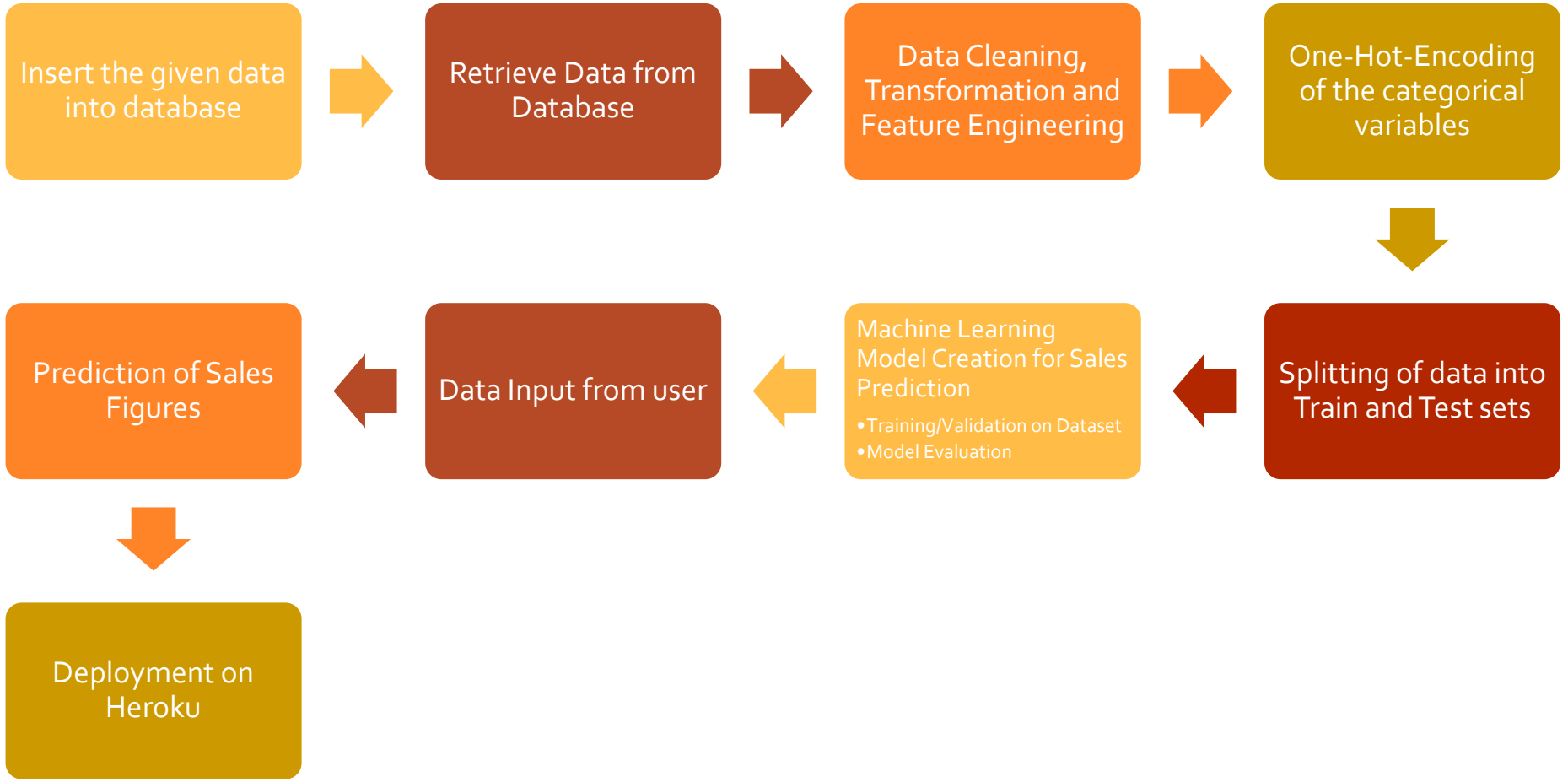
- Objective
  - Development of a predictive model for sales of a item in a store . The model will predict the sales of a product whose details like item type, weight, outlet, etc. will be provided.

- BENEFITS
  - PREDICTION OF SALES OF AN ITEM.
  - HELPS IN INVENTORY PLANNING AND LOGISTICS
  - GIVES INSIGHT INTO WHICH PRODUCT SHOULD BE MARKETED OR NOT
  - HELPS TO DESIGN VARIOUS PROMOTIONAL OFFERS FOR THE CUSTOMERS

Data Sharing Agreement

- Sample file name (ex Big mart store sales)
- Number of Columns
- Column names
- Column data type

ARCHITECTURE



Data Insertion in Database

- Table creation :- Table name “big_mart" is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the “data_given" are inserted in the above-created table. There are two files namely Train.csv and Test.csv. The data obtained from them is combined and inserted into one table.

EDA, Retrieve data, Pre-Processing and Splitting

- Exploratory data analysis was performed on the raw data to understand the dataset and its features. EDA was performed in the following steps-:
  - Removing unwanted attributes
  - Visualizing relation of independent variables with each other and output variables
  - Checking and changing Distribution of continuous values
  - Removing outliers
- The data is fetched from Cassandra to perform data cleaning, transformation and feature engineering.
- Feature engineering was performed on some columns of the dataset
- Missing values were imputed with the help of knn imputer.
- After that one-hot-encoding was done on categorical variables so that our Machine Learning Model can treat them as numerical variables to predict the outcome.
- The data was then split into Train and test datasets

Model Training and Prediction

- After all the pre-processing steps, model training was performed. Different models were trained and tested. Following models were tried-:
  - Linear Regression
  - Ridge Regression
  - Decision Tree Regression
  - Random Forest Regression
- We calculate the RMSE, MAE and R-squared scores for all the models and select the Decision Tree regression model as it had the best scores.
- The model trained was stored in a file to perform prediction
- The saved model was used to perform predictions on the test data and the input data.
- The input data had to be transformed according to the test data before performing prediction.

THANK YOU