**BIG MARKET SALES PREDICTION**

**1. Introduction**

**1.1 What is High-Level Design Document?**

The purpose of this high-Level document is to provide a complete description or architecture of the application. The HLD involves system architecture, database design, a description of systems and relationship among modules.

The HLD also includes detailed information about:

• All design aspects and define them in detail.

• The user interface.

• The hardware and software interfaces.

• Describe the performance requirements.

• Portability and compatibility.

* 1. **Scope**

The HLD document presents the structure of the system, such as the database architecture, application architecture, and technology architecture. The HLD uses non-technical to middle-technical terms which should be understandable to the stakeholders.

**2.1 What is Low-Level Design Document?**

The goal of LLD or Low-Level design document (LDD) is to give the internal logical design of the actual program code. Low-Level design is created based on the High-Level design. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly can directly code the program from the document.

* 1. **Scope**

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work

**3. Description of the Dataset**

We have train and test data set, train data set has both input and output variable(s). We need to predict the sales for test data set.

- **Item\_Identifier:** Unique product ID

- **Item\_Weight:** Weight of product

- **Item\_Fat\_Content:** Whether the product is low fat or not

- **Item\_Visibility:** The % of total display area of all products in a store allocated to the particular product

- **Item\_Type:** The category to which the product belongs

- **Item\_MRP:** Maximum Retail Price (list price) of the product

- **Outlet\_Identifier:** Unique store ID

- **Outlet\_Establishment\_Year:** The year in which store was established

- **Outlet\_Size:** The size of the store in terms of ground area covered

- **Outlet\_Location\_Type:** The type of city in which the store is located

- **Outlet\_Type:** Whether the outlet is just a grocery store or some sort of supermarket

- **Item\_Outlet\_Sales:** Sales of the product in the particulat store. This is the outcome variable to be predicted.

**4 General Description**

**4.1 Problem Statement**

Nowadays, shopping malls and Big Marts keep track of individual item sales data in order to forecast client demand and adjust inventory management. In the data warehouse, these data stores hold a significant amount of consumer information and particular item details. By mining the data store from the warehouse, more anomalies and common patterns can be discovered.

**4.2 Proposed Solution**

The proposed solution provides an efficient and simple to use application, that can be used without any prior knowledge.We will use performed EDA to find the important relation between different attributes and will use a machine-learning algorithm to predict the future sales demand. The client will be filled the required feature as input and will get results through the web application. The system will get features and it will be passed into the backend where the features will be validated and preprocessed and then it will be passed to a hyperparameter tuned machine learning model to predict the final outcome.

**4.3 Further Improvements**

As a future scope, features like data visualizations after model training and setting up an auto-mail system to send email to respective domain experts for further analysis.

**4.4 Technical Requirements**

The solution can be a cloud-based or application hosted on an internal server or even be hosted on a local machine. For accessing this application below are the minimum requirements:

• Good internet connection.

• Web Browser.

For training model, the system requirements are as follows:

• +4 GB RAM preferred

• Operation System: Windows, Linux, Mac

• Visual Studio Code / Jupyter notebook

**4.5 Data Requirements**

Data requirements completely depends on out problem statement.

• Comma separated values (CSV) file.

• Input file feature/field names and its sequence should be followed as per decided

**4.6 Tools used**

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Plotly are used to build the whole model.

• Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks.

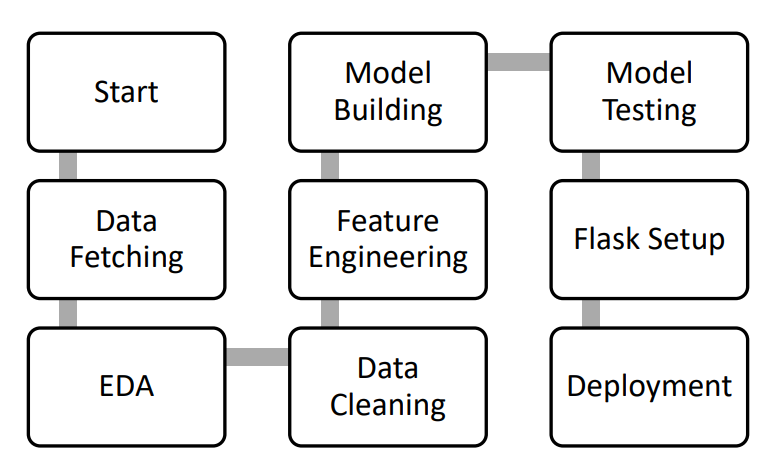
• NumPy is most commonly used package for scientific computing in Python

• Plotly is an open-source data visualization library used to create interactive and quality charts/graphs.

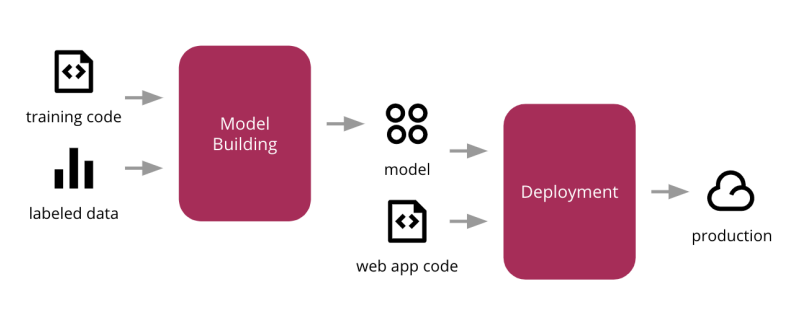
• Scikit-learn is used for a machine learning.

• VS Code is used as IDE (Integrated Development Environment) 5.0 Architecture

**5. Process Flow**



**5.1 Model Training and Evaluation**



**6.0 Architecture Description**

**6.1 Data Collection**

We used Big Mart outlet sales data as a dataset for this project where the dataset consists of 12 attributes. There is another dataset for validation purpose. The information each dataset available in Comma Separated-Values (CSV). Here we use dataset contains 8523 observations and test dataset contains 5681 observations.

**6.2 Exploratory Data Analysis**

Exploring the data by visualizing the distribution of values in some columns of the dataset, and the relationships between ‘Item Outlet Sales’ and other columns.

**6.3 Data Pre-processing**

Data pre-processing is the process of transforming raw data into an understandable format. In data pre-processing all the processes required before sending the data for model building are performed. New attributes were added named ‘’Outlet years”, where the given establishment year is subtracted from the current year. Then mapping of “Fat content” is done based on ‘Low’, ‘Reg’ and ‘Non-edible’.

**6.5 Model Building**

After data pre-processing is done, we will split the dataset into training set and validation set. Then processed data is used to give accurate results by applying multiple algorithms. An effective model can predict accurate results by finding exact insights of data. We will calculate RMSE score for each model and select the model with the best score.

**6.6 Data Validation**

Here Data Validation will be done on the test set.

**7.0 Business Context**

The objective is predicting store sales using historical markdown data. One challenge of modelling sales data is the need to make decisions based on limited history. If Big Market comes so does the chance to see how strategic decisions impacted the bottom line.

**8.0 Conclusion**

This system shows us that the different techniques that are used to forecast sales of the store. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other outlet locations should follow similar patterns for improved sales. Multiple instance parameters and various factors can be used to make this sales prediction more accurate and successful. Accuracy, which plays a key role in prediction-based system. From the results I could see that Random Forest Regressor turned out to be best working model for this problem in terms of the accuracy. These predictions help user to refine their methodologies and strategies which in turn helps them to increase their profit.

**Reference**

<https://github.com/moulimetturu/Big-Market-Sales-Prediction>