Low-Level Design

Stores Sales Prediction

|  |  |
| --- | --- |
| Written By | Gaurav.G.Singh |
| Version | 1.0 |
| Date | 28-03-2024 |

**Document Change Control Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### Reviews:

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Comments** |
|  |  |  |  |

### Approval Status:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review**  **Date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content** | **Page No** |
| 1. Introduction | 2 |
| 1.1 What is Low-Level Design Document | 2 |
| 1.2 Scope | 2 |
| Architecture | 3 |
| 2. Architecture Description | 4 |
| 2.1 Data Description | 4 |
| 2.2 Data Gathering | 4 |
| 2.3 Raw Data Validation | 4 |
| 2.4 Data Transformation | 5 |
| 2.5 New Feature Generation | 5 |
| 2.6 Data Preprocessing | 5 |
| 2.7 Feature Engineering | 5 |
| 2.8 Parameter Tuning | 5 |
| 2.9 Model Building | 5 |
| 2.10 Model Saving | 6 |
| 2.11 Django Setup for Data Extraction | 6 |
| 2.12 GitHub | 6 |
| 2.13 Deployment | 6 |
| 3. Unit Test Cases | 6 |

**1. Introduction**

**1.1 What is Low-Level Design Document.**

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code for **‘Stores Sales Prediction’**. 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 code the program from the document.

**1.2 Scope**

Low-level design (LLD) is a component-level design process that follows a step-by-step [refinement](https://en.wikipedia.org/wiki/Refinement_(computing)) 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.

**Architecture :**

**2. Architecture Description**

**2.1 Data Description**

Given is the variable name, variable type, the measurement unit, and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Measurement |
| Item\_Identifier | String | Unique product ID |
| Item\_Weight | Float | Weight of product |
| Item\_Fat\_Content | String | Whether the product is low fat or not |
| Item\_Visibility | Float | The % of a total display area of all products in a store allocated to the particular product |
| Item\_Type | String | The category to which the product belongs |
| Item\_MRP | Float | Maximum Retail Price (list price) of the product |
| Outlet\_Identifier | String | Unique store ID |
| Outlet\_Establishment\_Year | Integer | The year in which the store was established |
| Outlet\_Size | String | The size of the store in terms of ground area covered |
| Outlet\_Location\_Type | String | The type of city in which the store is located |
| Outlet\_Type | String | Whether the outlet is just a grocery store or some sort of supermarket |
| Item\_Outlet\_Sales | Float | Sales of the product in the particular store. This is the outcome variable to be predicted. |

**2.2 Data Gathering**

Data source: [**https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data**](https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data)

Train and Test data are stored in .csv format.

**2.3 Raw Data Validation**

After data is loaded, various types of validation are required before we proceed further with any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because The attributes which contain these are of no use. It will not play role in contributing to the sales of an item from respective outlets.

Like if any attribute is having zero standard deviation, it means that’s all the values are the same, its mean is zero. This indicates that either the sale is increasing or decrease that attribute will remain the same. Similarly, if any attribute is having full missing values, then there is no use in taking that attribute into an account for operation. It’s unnecessary increasing the chances of dimensionality curse.

**2.4 Data Transformation**

Before sending the data into the database, data transformation is required so that data are converted into such form with which it can easily insert into the database. Here, the ‘Item Weight’ and “Outlet Type’ attributes contain the missing values. So they are filled in both the train set as well as the test set with supported appropriate data types.

**2.5 New feature generations**

We can derive new item cateogory from item type and create mrp categories as mrp bin

**2.6 Data Preprocessing**

In data preprocessing all the processes required before sending the data for model building are performed. Like, here the ‘Item Visibility’ attributes are having some values equal to 0, which is not appropriate because if an item is present in the market, then how its visibility can be 0. So, it has been replaced with the average value of the item visibility of the respective ‘Item Identifier’ category. New attributes were added named ‘’Outlet years”, where the given establishment year is subtracted from the current year. A new “Item Type” attribute was added which just takes the first two characters of the Item Identifier which indicates the types of the items. Then mapping of “Fat content” is done based on ‘Low’, ‘Reg’ and ‘Non-edible’.

**2.7 Feature Engineering**

After preprocessing it was found that some of the attributes are not important to the item sales for the particular outlet. So those attributes are removed. Even one hot encoding is also performed to convert the categorical features into numerical features.

**2.8 Parameter Tuning**

Parameters are tuned using Randomized searchCV. Two algorithms are used in this problem, Linear Regressionand Random Forest. The parameters of these 2 algorithms are tunned and passed into the model.

**2.9 Model Building**

After doing all kinds of preprocessing operations mention above and performing scaling and hyperparameter tuning, the data set is passed into 2 models, Linear Regression and Random Forest regressor. It was found that Random Forest Regressor performs best with the smallest RMSE value i.e. 781.64 and the highest R2 score equals 0.55. So ‘Random forest regressor’ performed well in this problem.

**2.13 GitHub**

The whole project directory will be pushed into the GitHub repository.