**Architecture Design**

**BIG MART SALES PREDICTION**

|  |  |
| --- | --- |
| Written by | Aswini Kumar |
| Version | 1.0 |
| Date | 31-10-2022 |

**Document Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author | Comments |
| 1.0 | 31-10-2022 | Aswini Kumar |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Approval Status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Review date | Reviewed By | Approved By | Comments |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content** | **Page No** |
| Abstract |  |
| 1.Introduction |  |
| 1.1. What is Architecture Design? |  |
| 1.2 Scope |  |
| 1.3 Constraints |  |
| 2. Technical Specification |  |
| 2.1 Dataset |  |
| 2.2 Logging |  |
| 2.3 New Feature Generation |  |
| 2.4 Deployment |  |
| 3. Technical Stack |  |
| 4.Proposed Solution |  |
| 5.Architecture Description |  |
| 6.User Input / Output Workflow |  |

Abstruct

Machine Learning is a category of algorithm that allows software application to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and Employs algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data become available. These models can be applied in different area and trained to match the expectations of management so that accurate steps can be taken to achieve the organization’s target. In this paper, the case of Big Mart, a one-stop shopping-center, has effects of different factors on the items sales. Taking various aspects of a dataset collected of Big Mart, and the methodology followed for building a predictive model, result with high levels of accuracy are generated, and this observation can be employed to make decisions to improve sales.

**1.Introduction**

**1.1 What is Architecture Design ?**

The goal of Architecture Design (AD) or a low-level design document is to give the internal design of the actual program code for the ‘Bike Share Prediction System’. AD describe the class diagram with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

**1.2 Scope**

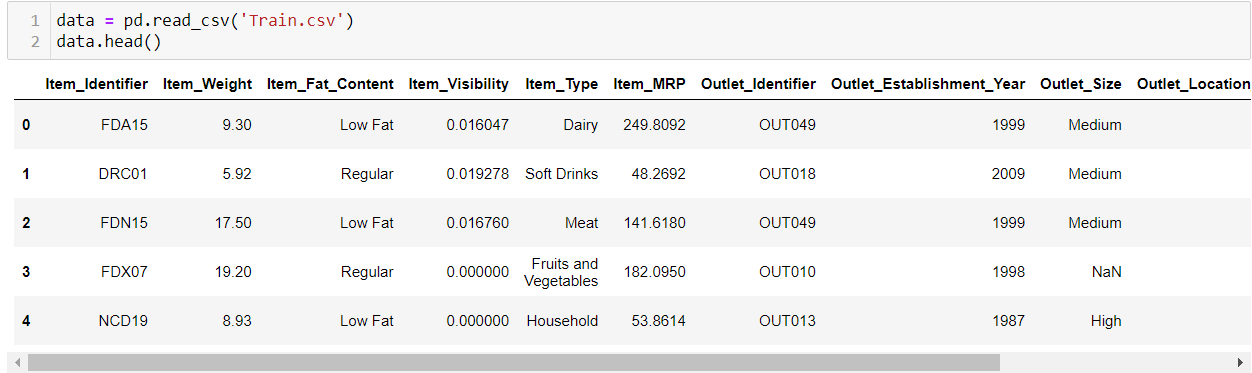
Architecture Design (AD) is a components-level design process that follows a step-by-step refinement process . This process can be used for designing data structure , required software , architecture , source code , and ultimately , performance algorithms. Overall , the data organization may be defined during requirement analysis and then refined during data design work, And the complete workflow

**1.3 Constraints**

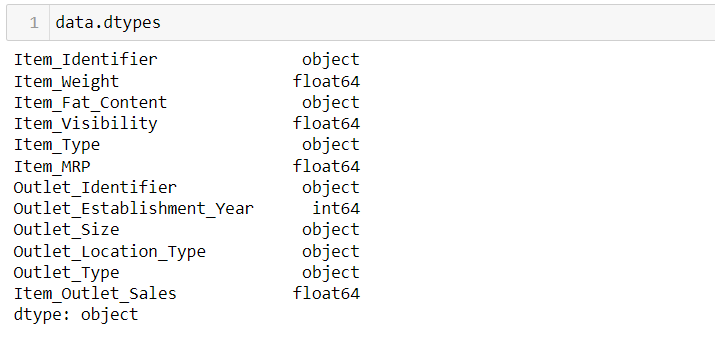
We only predict the expected casual and registered customers based on the weather condition and date information.

**2.Technical Specification**

Big Mart’s data scientists collected sales data of their 10 stores situated at different locations with each store having 1559 different products as per data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their sales. The dataset looks like as follow

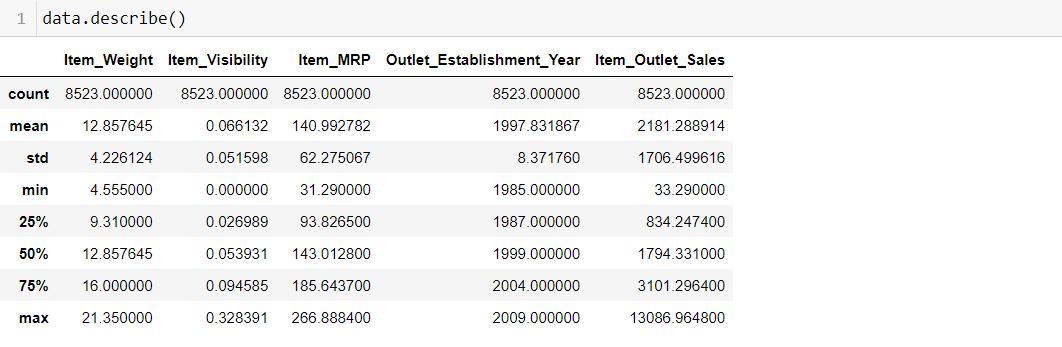


The data set consists of various data types from integer to floating to object as show in Fig.



In the raw data there can be various types of underlying patterns which also give an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed with respect to data as it may contain null value , or redundant values , or various types of ambiguity , which also demands pre-processing of data. The dataset should therefore be explored as much as possible.

Various factors important by statistical mean , standard deviation , median , count and values and maximum value etc. are show below for numerical attributes.



Preprocessing of this dataset includes doing analysis on the independent variable like checking for null values in each column and then replacing or filtering them with supported appropriate data types so that analysis and model fitting is not hindered from their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tell about the variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns and along with their percentile values for median, play an important factor in deciding which value to be chosen at priority for further exploration task and analysis. Data types of different columns are used further in label processing and a one-hot encoding scheme during the model building.

**2.2 Logging**

We should be able to log every activity done by the user

* The system identifies at which step logging require
* The system, should be able to log each and every system flow.
* Developers can choose logging methods. Also, can choose database logging
* The System Should be not be hung even after using so much logging. logging just because we can easily debug issuing so logging is mandatory to do.

**2.3 Database:**

The system needs to store every request into the database and we need to store it in such a way that it is easy to retain and look into the records.

The system should capture every data that any user gave and the prediction that has been made by that input.

**2.4 Deployment**

For the hosting of the project, we will use Heroku

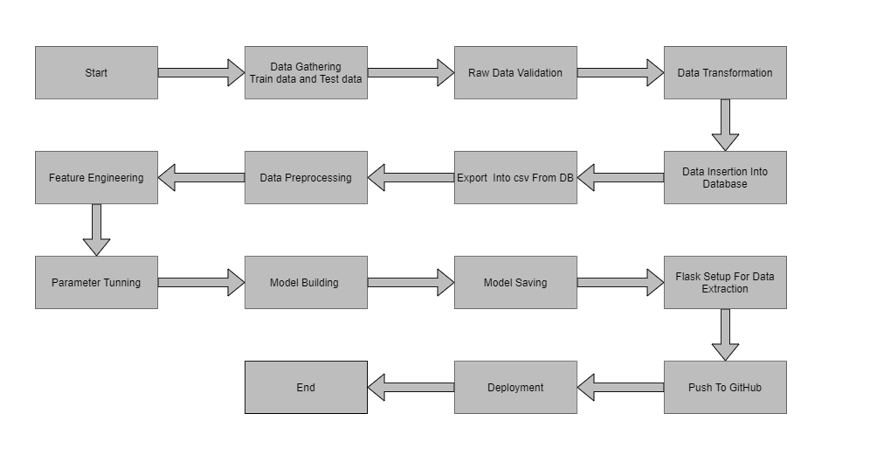
**3.Technology Stack**

|  |  |
| --- | --- |
| Front End | HTML/JavaScript |
| Backend | Python |
| Deployment | Heroku |

**4. Proposed Solution**

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.

**5. Architecture detail:**

****

**5.1Data Gathering**

Data source: <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data>

Train and Test data are stored in .csv format.

**5.2 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.

**5.3 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.

**5.4 New Feature Generation**

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

**5.5 Data Pre-processing**

In data pre-processing 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’.

**5.6 Feature Engineering**

After pre-processing 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.

**5.7 Parameter Tuning**

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

**5.8 Model Building**

After doing all kinds of pre-processing operations mention above and performing scaling and hyperparameter tuning, the data set is passed into 2 models, Linear Regression and Random Forest. It was found that Radom 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.

**5.9 Model Saving**

Model is saved using pickle library in `.sav` format.

**5.10 GitHub**

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

**5.11 Deployment**

The cloud environment was set up and the project was deployed from GitHub into the Heroku cloud platform.

App link :

**6. User Input / Output Workflow.**

