ARCHITECTURE

Store Sales Prediction

Revision Number: 1.0
Last Date of Revision: 11/06/2022

Contents:

Document version control 3

Abstract 4

1 Introduction 5
- 1.1 Why this Architecture? 5
- 1.2 Scope 5
- 1.3 Constraints 5

2 Architecture 6
- 2.1 Prediction Data Validation 6
- 2.2 Prediction Data Insertion into Database 7
- 2.3 Prediction 7

Document Version Control:

Date Issued	Version	Description	Author
11/06/2022	1	Initial Architecture-V1.0	Amit Ranjan Sahoo

Abstract:

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

You have to build a solution that should able to predict the sales of the different stores of Big Mart according to the provided dataset.

1 Introduction:

1.1 Why this Architecture:

The purpose of this Architecture is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The Architecture will:

- Present all of the design aspects and define them in a detail
- Describe the user interface being implemented
- Includes design features and architecture of the project
- List and describe the non-functional attribute like:
  - Reliability
  - Maintainability
  - Portability
  - Reusability
  - Application compatibility
  - Serviceability

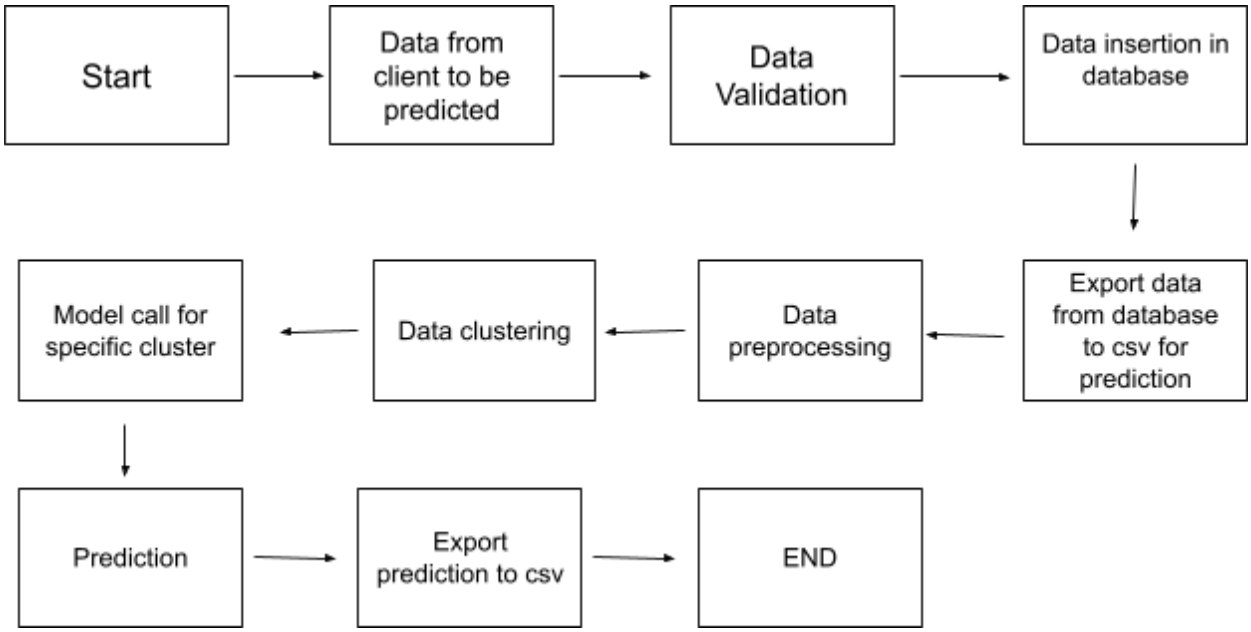
1.2 Scope:

The documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (navigation), and technology architecture.

1.3 Constraints:

The System must be user-friendly, as automated as possible, and the user should notbe required to know any other work.

2 Architecture:



2.1 Prediction Data Validation:

In this step, we perform different sets of validation on the given set of prediction files.

1. Name Validation- We validates the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good Raw" else we move such files to "Bad Raw".

2. Number of Columns - We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad Raw".

4. Name of Columns - The name of the columns is validated and should be the sameas given in the schema file. If not, then the file is moved to "Bad Raw".

5. The datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, thenthe file is moved to "Bad Raw".

6. Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad Raw".

2.2 Prediction Data Insertion in Database:

1) Database Creation and connection - Create database with the given name passed. If the database is already created, open the connection tothe database.

2) Collection creation in the database - Collection with name - "Good_Raw_Data", is created in the database for inserting the files in the"Good Raw" on the basis of given column names and datatype in the schema file. If collection is already present then the collection is deleted and new table is created.

3) Insertions of files in the collection - All the files in the “Good Raw” are inserted in the above-created table. If any file has invalid data type in any ofthe columns, the file is not loaded in the table and is moved to "Bad Raw".

2.3 Prediction:

1) Data Export from Db - The data in the stored database is exported as a CSV file to beused for prediction.
2) Data Preprocessing

ARCHITECTURE

a) Drop columns not useful for training the model. Such columns were selected while doing the EDA.

b) Encode the categorical values

c) Check for null values in the columns. If present, impute the null values using the KNN imputer.

3) Clustering – K-Means model created during training is loaded, and clusters for the preprocessed prediction data is predicted.

4) Prediction - Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster.

5) Once the prediction is made for all the clusters, a zip file containing prediction (csv file) and bad files are sent to client.