DPR

STORE SALE PREDICTION

Revision Number – 1.0

Last Date of Revision – 04/09/2024

Swapnil Shinde

Document Version Control

Date	Version	Description	Author
01-09-2024	1.0	Abstract, Introduction	Swapnil Shinde
02-09-2024	1.1	Deployment	Swapnil Shinde
03-09-2024	1.2	Q and A	Swapnil Shinde

Contents

Abstract

In the contemporary retail ecosystem, shopping malls and Big Marts meticulously collect and archive individual item sales data, yielding invaluable insights into consumer behavior and product specifics. These data repositories, securely stored within a data warehouse, serve as reservoirs of opportunity. Our system embarks on a transformative journey through the realms of data, expertly harnessing its power. The journey unfolds through meticulously orchestrated steps: data ingestion from Kaggle datasets, data transformation for cleanliness and relevance, model building to extract meaningful patterns, and the establishment of an efficient batch prediction pipeline. We don't stop there; we extend this journey to the end-users with a well-crafted, user-friendly interface, bridging the gap between data and actionable insights. This document, a testament to our technical prowess, delves deep into the modular architecture, interfaces, algorithms, and visualizations that underpin this transformative solution, setting the stage for a future where anomalies and common patterns emerge as strategic assets in the world of retail..

INTRODUCTION

Why this DPR Documentation?

The main purpose of this DPR documentation is to add the necessary details of the project andprovide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points:

Describes the design flow
Implementations
Software requirements
Architecture of the project
Non-functional attributes like:
Reusability
Portability
Resource utilization

1 Description

1.1 Problem Perspective

Utilizing data warehousing techniques to analyze individual item sales data from shopping malls and Big Marts presents a promising prospect. This approach enables us to unearth valuable insights, including anomalies and recurrent patterns, within vast repositories of consumer information and product specifics. By harnessing advanced data mining and analytics, we can enhance demand forecasting accuracy and refine inventory management strategies, ultimately optimizing operations and customer satisfaction. This data-driven approach has the potential to revolutionize how retailers adapt to ever-evolving market dynamics and consumer preferences

1.2 Problem Statement

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..

1.3 Proposed Solution

Developing a comprehensive solution for leveraging data warehousing in shopping malls and Big Marts is essential. First, we must establish robust data pipelines to ingest and store individual item sales data in the data warehouse. This data repository should be designed for efficient retrieval and analysis, considering both consumer demographics and product attributes.
Next, employing advanced data mining and machine learning techniques, we can extract valuable insights from this data store. This includes identifying anomalies that might indicate theft or data entry errors and discovering common patterns that offer invaluable information for demand forecasting and inventory management..

1.4 Solution Improvements

Improvements the solution for leveraging data warehousing in shopping malls and Big Marts involves ensuring data quality, scalability, predictive maintenance, personalization, real-time analytics, security, and compliance, fostering cross-functional collaboration, integrating external data sources, continuous monitoring, machine learning models, and customer feedback. By implementing these enhancements, the solution becomes more comprehensive, adaptable, and capable of addressing evolving challenges, enabling retailers to stay competitive and responsive to customer demands while optimizing operations and customer satisfaction through data-driven strategies

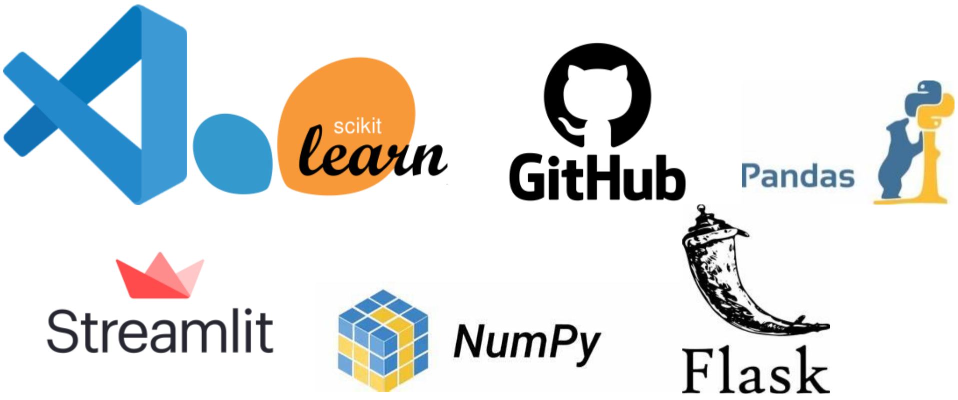
2 Technical Requirements

There are not any hardware needs needed for victimization this application, the user should have an interactive device that has access to the web and should have the fundamental understanding of providing the input. And for the backend half the server should run all the package that's needed for the process and provided information to show the results.

2.1 Tools Used

- Python 3.9 is employed because the programming language and frame works like numpy, pandas, sklearn,flask, streamlit and alternative modules for building the model.

- Visual Studio Code is employed as IDE.
- Front end development is completed victimization HTML/CSS
- Flask is employed for each information and backend readying
- GitHub is employed for version management
- Streamlit Cloud and localhost is used for Deployment



3 Data Requirements

The info demand is totally supported the matter statement. and also, the information set is accessible on the Kaggle within the type of standout sheet(.xlsx), because the main theme of the project is to induce the expertise of real time issues, we have a tendency to once more mercantilism {the information into the prophetess data base and commerce it into csv format.

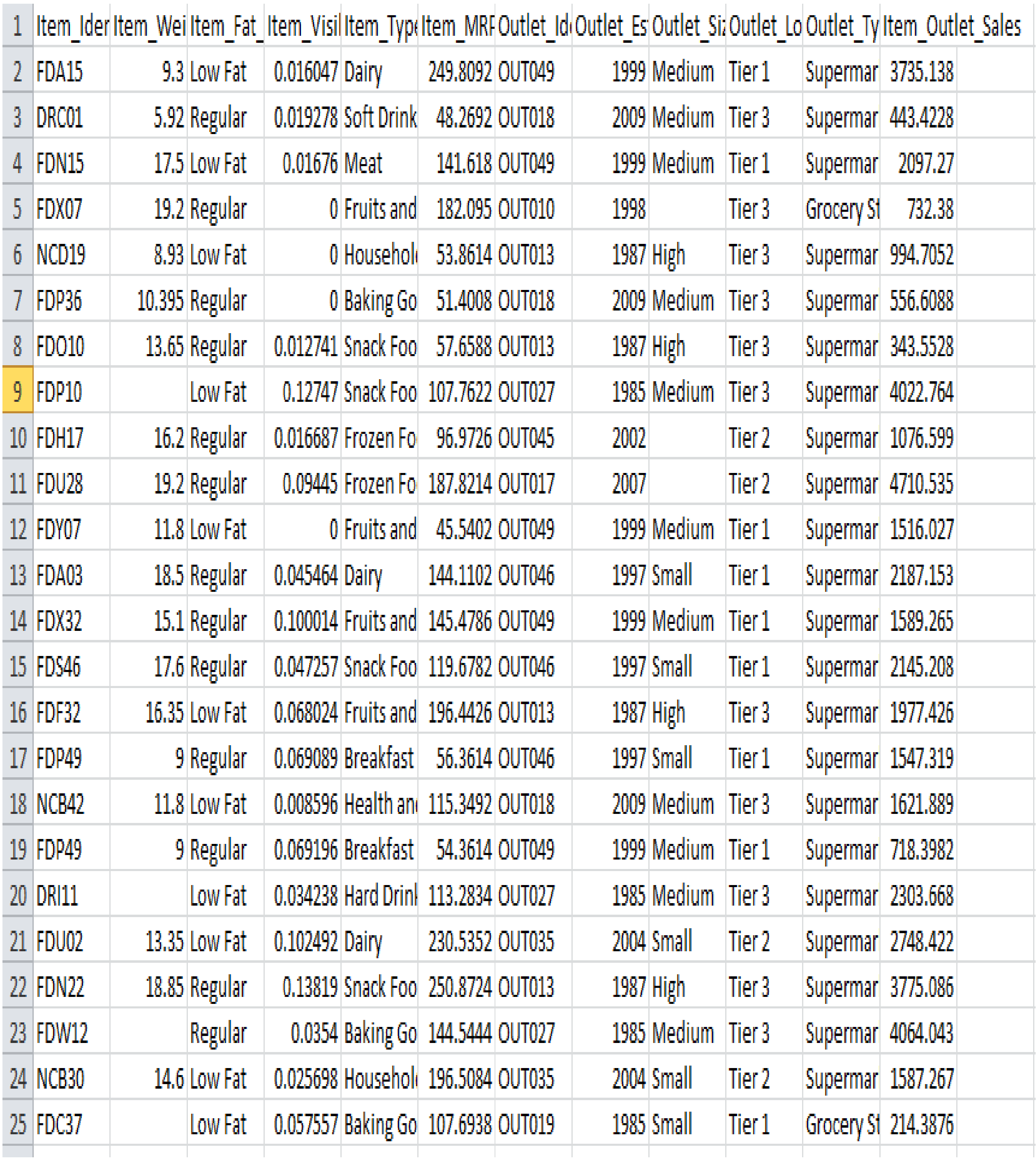
3.1 Data Gathering from Main Source

The data for the current project is being gathered from Kaggle dataset, the linkto the data is: BigMart Sales Data | Kaggle

3.2 Data Description

We have train (8523) and test (5681) data set, train data set has both input and output

Columns Are :
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.



1	Item_Ider	Item_Wei	Item_Fat_	Item_Visil	Item_Typ	Item_MRF	Outlet_Id	Outlet_Es	Outlet_Si	Outlet_Lo	Outlet_Ty	Item_Outlet_Sales
2	FDA15	9.3	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermar	3735.138
3	DRC01	5.92	Regular	0.019278	Soft Drink	48.2692	OUT018	2009	Medium	Tier 3	Supermar	443.4228
4	FDN15	17.5	Low Fat	0.01676	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermar	2097.27
5	FDX07	19.2	Regular	0	Fruits and	182.095	OUT010	1998		Tier 3	Grocery St	732.38
6	NCD19	8.93	Low Fat	0	Househol	53.8614	OUT013	1987	High	Tier 3	Supermar	994.7052
7	FDP36	10.395	Regular	0	Baking Go	51.4008	OUT018	2009	Medium	Tier 3	Supermar	556.6088
8	FDO10	13.65	Regular	0.012741	Snack Foo	57.6588	OUT013	1987	High	Tier 3	Supermar	343.5528
9	FDP10		Low Fat	0.12747	Snack Foo	107.7622	OUT027	1985	Medium	Tier 3	Supermar	4022.764
10	FDH17	16.2	Regular	0.016687	Frozen Fo	96.9726	OUT045	2002		Tier 2	Supermar	1076.599
11	FDU28	19.2	Regular	0.09445	Frozen Fo	187.8214	OUT017	2007		Tier 2	Supermar	4710.535
12	FDY07	11.8	Low Fat	0	Fruits and	45.5402	OUT049	1999	Medium	Tier 1	Supermar	1516.027
13	FDA03	18.5	Regular	0.045464	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermar	2187.153
14	FDX32	15.1	Regular	0.100014	Fruits and	145.4786	OUT049	1999	Medium	Tier 1	Supermar	1589.265
15	FDS46	17.6	Regular	0.047257	Snack Foo	119.6782	OUT046	1997	Small	Tier 1	Supermar	2145.208
16	FDF32	16.35	Low Fat	0.068024	Fruits and	196.4426	OUT013	1987	High	Tier 3	Supermar	1977.426
17	FDP49	9	Regular	0.069089	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermar	1547.319
18	NCB42	11.8	Low Fat	0.008596	Health an	115.3492	OUT018	2009	Medium	Tier 3	Supermar	1621.889
19	FDP49	9	Regular	0.069196	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermar	718.3982
20	DRI11		Low Fat	0.034238	Hard Drin	113.2834	OUT027	1985	Medium	Tier 3	Supermar	2303.668
21	FDU02	13.35	Low Fat	0.102492	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermar	2748.422
22	FDN22	18.85	Regular	0.13819	Snack Foo	250.8724	OUT013	1987	High	Tier 3	Supermar	3775.086
23	FDW12		Regular	0.0354	Baking Go	144.5444	OUT027	1985	Medium	Tier 3	Supermar	4064.043
24	NCB30	14.6	Low Fat	0.025698	Househol	196.5084	OUT035	2004	Small	Tier 2	Supermar	1587.267
25	FDC37		Low Fat	0.057557	Baking Go	107.6938	OUT019	1985	Small	Tier 1	Grocery St	214.3876

3.3 Data Ingestion

The cornerstone of our data-driven project was established through a systematic process of data acquisition and ingestion. Utilizing Kaggle, a reputable platform renowned for its high-quality datasets, we identified and acquired the crucial data required for our price prediction project. This dataset, integral to our goal of accurate price forecasting, was meticulously downloaded and securely stored within our local system infrastructure. Subsequently, we initiated the data ingestion phase, where the dataset seamlessly integrated into our project's data pipeline. This meticulous approach ensures that our project is built upon a solid foundation, setting the stage for robust and precise price prediction models and analysis

4 Data Transformation

Steps performed in pre-processing are:

- First read data from Artifact folder
- Checking unnecessary columns
- One column has product id which is unique for every product so I deleted that column.
- Checked for null values
- there are too many null values are present in two columns that’s why I deleted them
- Performed one-hot encoder on categorical columns.
- Perform Ordinal Encoder on Ordinal Columns.
- Scaling is performed for needed information.
- And, the info is prepared for passing to the machine learning formula

5 Design Flow

5.1 Modelling

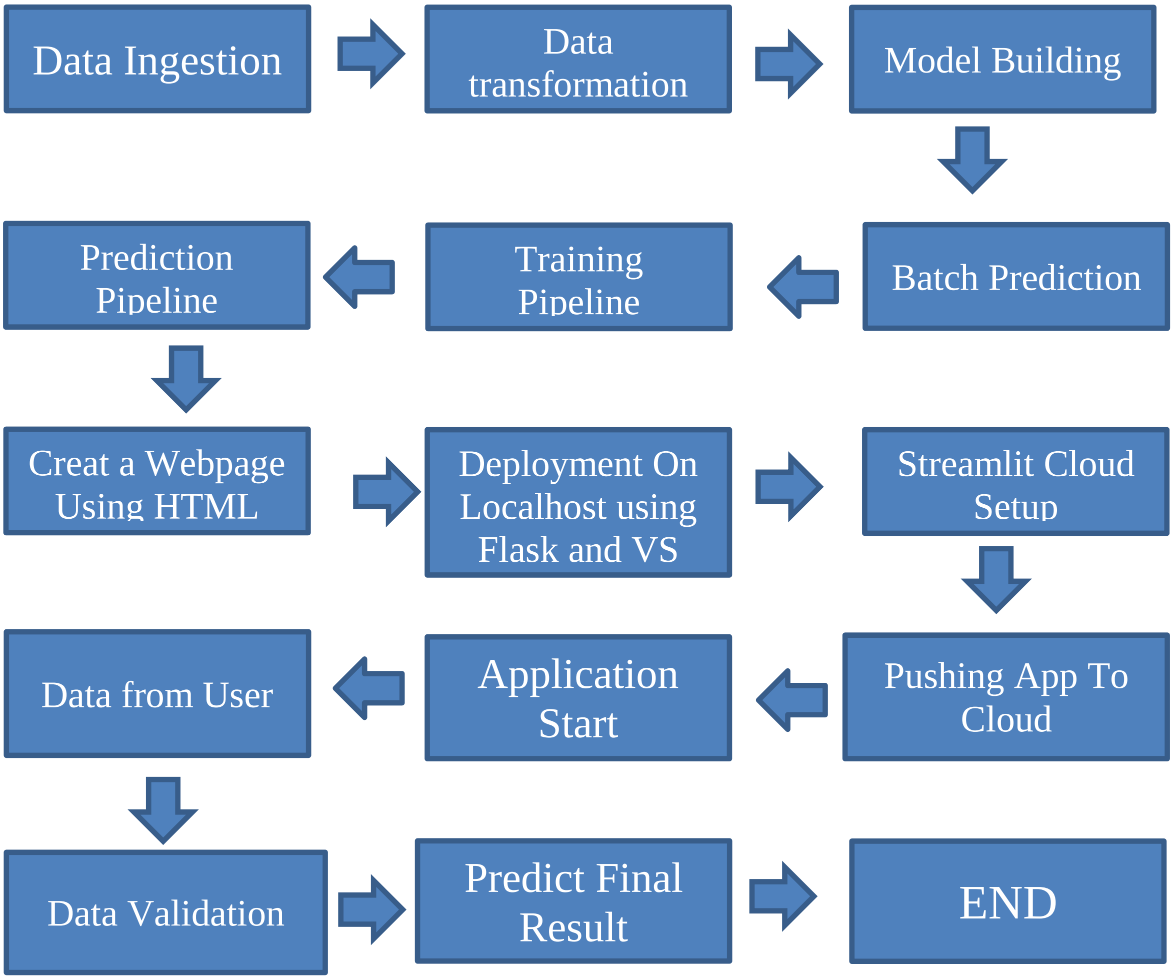
The pre-processed information is then envisioned and every one the specified insights are being drawn. though from the drawn insights, the info is at randomunfold however still modelling is performed with completely different machinelearning algorithms to form positive we tend to cowl all the chances. and eventually, Gradient Boosting performed well .

5.2 UI Integration

Both CSS and HTML files are being created and are being integrated with the created machine learning model. All the required files are then integrated to the app.py file and tested locally

.

5.3 Modelling Process&5.4 Deployment Process



6 Data from User

The data from the user is retrieved from the created HTML web page.

7 Data Validation

The data provided by the user is then being processed by app.py or application.py file and validated. The validated datais then sent for the prediction.

8 Rendering the Results

The data sent for the prediction is then rendered to the web page.

9 Deployment

The tested model is then deployed to Streamlit Cloud. So, users can access the project from anyinternet devices.

Conclusion

In conclusion, implementing a data warehousing solution in shopping malls and Big Marts holds immense potential for optimizing operations, enhancing customer experiences, and staying competitive in the dynamic retail landscape. By effectively harnessing data from various store operations and ensuring compliance with data privacy regulations, retailers can gain valuable insights into consumer behaviors and preferences. While constraints such as high infrastructure costs and cybersecurity risks are challenges to navigate, they can be mitigated with prudent planning and investment. With a commitment to data quality and ongoing adaptation to changing market dynamics, the future for these retail giants looks promising as they leverage data-driven strategies to meet customer demands and drive business success.

Q & A:

Q1) What’s the source of data?

Ans-The data for training is provided by the client in multiple batches and each batch contain multiplefiles.

Q 2) What was the type of data?

Ans-The data was the combination of numerical and Categorical values.

Q 3) What’s the complete flow you followed in this Project?
Ans-Refer Page no 11 for better Understanding.

Q 4) After the File validation what you do with incompatible file or files which didn’t pass the validation?
Ans-Files like these are moved to the Achieve Folder and a list of these files has beenshared with the client and we removed the bad data folder.

Q 5) How logs are managed?
Ans- We are using different logs as per the steps that we follow in validation and modeling like File validation log, Data Insertion, Model Training log, prediction log etc.

Q 6) What techniques were you using for data pre-processing?
Ans-Removing unwanted attributes Cleaning data and imputing if null values are present. Converting categorical data into numeric values.

Q 7) How training was done or what models were used?
Ans-Before dividing the data in training and validation set, we performed pre-processing over the data set and made the final dataset. As per the dataset training and validation data were divided. Algorithms like Linear regression, Decision Tree, Random Forest, Gradient Boosting were usedbased on the recall, final model was used on the dataset and we saved that model.

Q 8) How Prediction was done?
Ans-The testing files are shared by the client. We Performed the same life cycle on the provided dataset. Then, on the basis of dataset, model is loaded and prediction is performed. In the end we get the accumulated data of predictions.

Q 9) What are the different stages of deployment?
Ans-First, the scripts are stored on GitHub as a storage interface.
The model is first tested in the local environment.
After successful testing, it is deployed on Streamlit Cloud.