

ARCHITECTURE DOCUMENT DESIGN (ADD) Backorder Prediction

Written By:- Yogesh Mudalkar

Document Control:**Change Record:**

Version	Date	Author	Comments
0.1	28-July - 2023	Yogesh Mudalkar	Initial Architecture design Draft.
0.2	29-July-2023	Yogesh Mudalkar	Architecture Design appended and updated
0.3	30-July-2023	Yogesh Mudalkar	Final design appended

Reviews:

Version	Date	Reviewer	Comments
1.0	03-Aug-2023	Yogesh Mudalkar	Document Content , Architecture Design Updated

Approval Status:

Version	Date	Reviewed By	Approved By	Comments
1.1	05-Aug-2023	Yogesh Mudalkar		Final

Contents

1.	Introduction —	4
1.1.	What is Architecture Document Design ? -----	4
1 2.	Scope —	5
2.	Backorder Prediction Home Page -----	6
3.	Architecture Description —	7
3 1.	Data Description —	7
3 2.	Data Transformation —	7
3.3.	Exploratory Data Analysis -----	7
3 4.	Data Cleaning —	7
3 5.	Data Preprocessing —	8
3 6.	Feature Engineering —	8
3 7.	Feature Selection —	8
3 8.	Model Building —	8
3 9.	Random Forest —	8
3.10.	DecisionTreeClassifier -----	8
3.11.	XGBoost Model —	9
3 12.	Model Validation —	9
4.	Database —	9
5.	Unit Test Case —	10

1. Introduction:

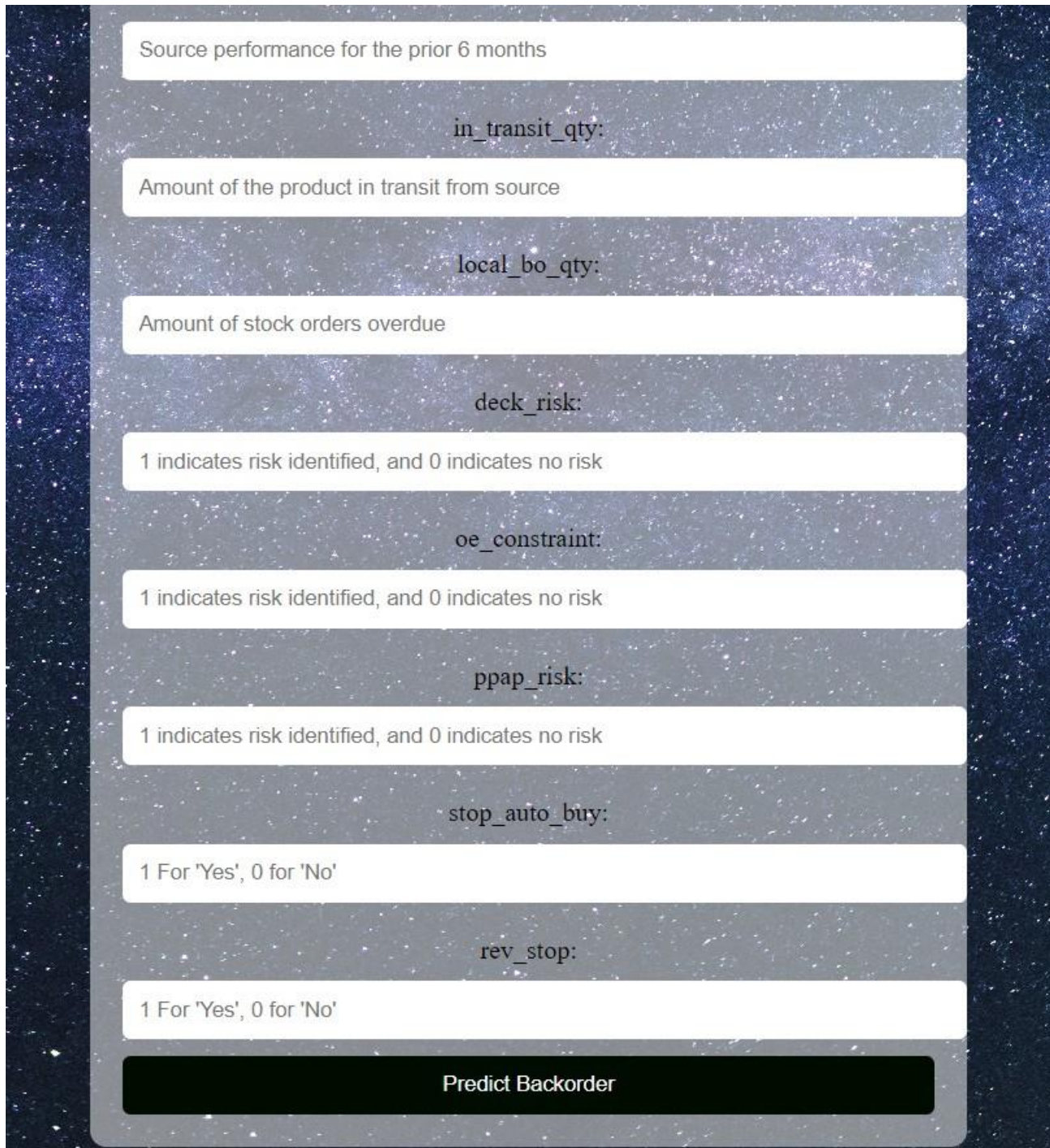
1.1. What is an Architecture design document?

The architecture design document is a technical document describing the components and specifications required to support the solution and ensure that the specific business and technical requirements of the design are satisfied. The architecture design document includes the conceptual, logical, and physical designs.

1.2. Scope:

The scope of this document is specific to the implementation of the Predictive Machine Learning Model to deploy on to the cloud. The supporting Libraries are defined in the Architecture Design. These python libraries are used to mining the data to understand the anomalies and frequent patterns from the row data. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques.

1.1.2 Backorder Prediction Home Page :

The image shows a web form for backorder prediction. It has a dark blue, starry background. The form is a light gray rounded rectangle with several white input fields and a black button. The fields are labeled with variable names and their descriptions. The labels are: 'Source performance for the prior 6 months', 'in_transit_qty:', 'Amount of the product in transit from source', 'local_bo_qty:', 'Amount of stock orders overdue', 'deck_risk:', '1 indicates risk identified, and 0 indicates no risk', 'oe_constraint:', '1 indicates risk identified, and 0 indicates no risk', 'ppap_risk:', '1 indicates risk identified, and 0 indicates no risk', 'stop_auto_buy:', '1 For 'Yes', 0 for 'No'', 'rev_stop:', '1 For 'Yes', 0 for 'No''. The button is black with white text that says 'Predict Backorder'.

Source performance for the prior 6 months

in_transit_qty:

Amount of the product in transit from source

local_bo_qty:

Amount of stock orders overdue

deck_risk:

1 indicates risk identified, and 0 indicates no risk

oe_constraint:

1 indicates risk identified, and 0 indicates no risk

ppap_risk:

1 indicates risk identified, and 0 indicates no risk

stop_auto_buy:

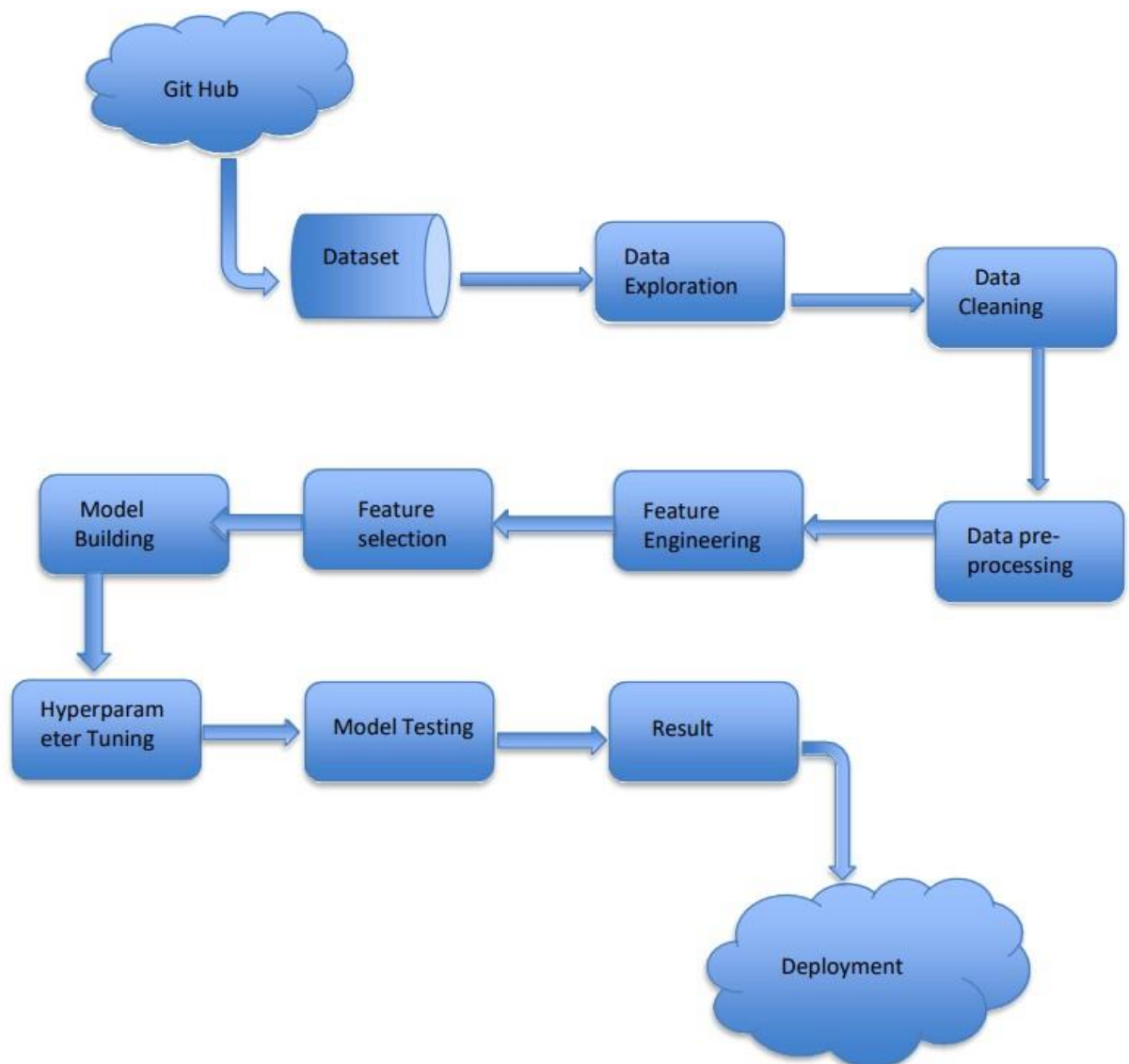
1 For 'Yes', 0 for 'No'

rev_stop:

1 For 'Yes', 0 for 'No'

Predict Backorder

2. Architecture:



3. Architecture Description:

3.1. Data Description:

The dataset of this research has been published in Kaggle. It is divided into training and testing datasets. Data Source:

https://github.com/rodrigasantis1/backorder_prediction/blob/master/dataset.rar

The dataset is highly imbalanced which should be addressed for accurate predictions by the model; each dataset contains 23 attributes with 1,687,862 and 242,077 observations for the training and testing sets, respectively.

3.2. Data Transformation:

In the Transformation Process, we will convert our original dataset which is in CSV Format to a pandas data frame. After reading the dataset we have two pandas data frame train and test concatenated into one data frame.

3.3. Exploratory Data Analysis:

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

3.4 Data Insertion into Database:

- a. Database Creation and connection - Create a database with name passed. If the database is already created, open the connection to the database.
- b. Table creation in the database.
- c. Insertion of files in the table.

3.5. ExportData from Database:

The data in a stored database is exported as a CSV file to be used for Data Preprocessing and Model Training.

3.6. Data Cleaning:

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

3.7 Data Pre-processing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always the case that we come across clean and formatted data.

3.8 Feature Engineering:

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process.

3.9 Feature Selection:

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimises the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets.

3.10 Model Building:

A machine learning model is built by learning and generalising from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfil its purpose. Lack of data will prevent you from building the model, and access to data isn't enough.

3.11 Random Forest:

A random forest classifier, a random forest is a Meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

3.12 DecisionTreeClassifier:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

3.13 XGBoost Model:

XGBoost Model for Classification XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm. The stochastic gradient boosting algorithm, also called gradient boosting machines or tree boosting, is a powerful machine learning technique that performs well or even best on a wide range of challenging machine learning problems.

3.16. Model Validation:

For machine learning systems, we should be running model evaluation and model tests in parallel. Model evaluation covers metrics and plots which summarise performance on a validation or test dataset. Model testing involves explicit checks for behaviours that we expect our model to follow.

3.17 Flask:

Flask is a micro web framework written in python. Here we use flask to create an API that allows us to send data, and receive a prediction as a response. Flask supports extensions that can add application features as if they were implemented in Flask itself.

4. Database:

Cassandra DB is used to retrieve, insert, delete and update the customer input database. Flask is a micro web framework written in python. Here we use the flask to create an API that allows sending data, and receives a prediction as a response. Flask API's act as an interface between the user, ML- Model and DB. Users interact with the model and the result will be returned to the user by API. The Flask API will collect all the customer input data stored in Cassandra DB.

Unit test Cases:

Test Case Description	Pre-requisite	Expected Result
Verify the user should able to see their input data	1. Application is accessible 2. Application is responsive	Users can see their data in the form.
Verify the user can edit their information in the form	1.Application is responsive	User should be able to edit all input field
Verify whether the user gets Submit button to submit the inputs.	1.Application is accessible 2.Application is responsive	User should able to submit values
Verify whether the wrong information should not be submitted by the user	Application is secured	Correct information should be submitted