**Flow of the Project**

**Problem Statement**

To build a model which will be able to predict whether an order for a given product can go on backorder or not.

A backorder is the order which could not be fulfilled by the company. Due to high demand of a product, the company was not able to keep up with the delivery of the order.

**Architecture**



**Data Description**

The client will send data in multiple sets of files in batches at a given location. Data will contain 24 columns:

**sku** – Random ID for the product

**national**\_**inv** – Current inventory level for the part

**lead\_time** – Transit time for product (if available)

**in\_transit\_qty** – Amount of product in transit from source

**forecast\_3\_mont**h – Forecast sales for the next 3 months

**forecast\_6\_month** – Forecast sales for the next 6 months

**forecast\_9\_month** – Forecast sales for the next 9 months

**sales\_1\_month** – Sales quantity for the prior 1 month time period

**sales\_3\_month** – Sales quantity for the prior 3 month time period

**sales\_6\_month** – Sales quantity for the prior 6 month time period

**sales\_9\_month** – Sales quantity for the prior 9 month time period

**min\_bank** – Minimum recommend amount to stock

**potential\_issue** – Source issue for part identified

**pieces\_past\_due** – Parts overdue from source

**perf\_6\_month\_avg** – Source performance for prior 6 month period

**perf\_12\_month\_avg** – Source performance for prior 12 month period

**local\_bo\_qty** – Amount of stock orders overdue

**deck\_risk** – Part risk flag

**oe\_constraint** – Part risk flag

**ppap\_risk** – Part risk flag

**stop\_auto\_buy** – Part risk flag

**rev\_stop** – Part risk flag

**went\_on\_backorder** – Product actually went on backorder. This is the target value.

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

**Data Validation**

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

1. Name Validation- We validate 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\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."
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\_Data\_Folder."
3. Name of Columns - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".
4. 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, then the file is moved to "Bad\_Data\_Folder".
5. 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\_Data\_Folder".

**Data Insertion in Database**

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

2) Table creation in the database - Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file. If the table is already present, then the new table is not created and new files are inserted in the already present table as we want training to be done on new as well as old training files.

3) Insertion of files in the table - All the files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

**Model Training**

1) Data Export from Db - The data in a stored database is exported as a CSV file to be used for model training.

2) Data Preprocessing

a) Check for null values in the columns. If present, drop the null values.

b) Check if any column has zero standard deviation, remove such columns as they don't give any information during model training.

c) Apply scaling and PCA in the columns , remove multi collinearity.

3) Model Selection - After clusters are created, we find the best model for each cluster. We are using two algorithms, "Random Forest" and "XGBoost". For each cluster, both the algorithms are passed with the best parameters derived from GridSearch. We calculate the AUC scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

**Prediction Data Description**

Client will send the data in multiple set of files in batches at a given location.Data will contain 23 columns and we have to predict whether an order will become back order or not.

Apart from prediction files, we also require a "schema" file from client which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns and their datatype.

**Data Validation**

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

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

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

3) Name of Columns - The name of the columns is validated and should be same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".

4) Datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If dataype is wrong then the file is moved to "Bad\_Data\_Folder".

5) Null values in columns - If any of the columns in a file has all the values as NULL or missing, we discard such file and move it to "Bad\_Data\_Folder".

**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 to the database.

2) Table creation in the database - Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" on the basis of given column names and datatype in the schema file. If table is already present then new table is not created, and new files are inserted the already present table as we want training to be done on new as well old training files.

3) Insertion of files in the table - All the files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

**Prediction**

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

2) Data Preprocessing

a) Check for null values in the columns. If present, drop the null values.

b) Check if any column has zero standard deviation, remove such columns as we did in training.

3) Prediction - model is loaded and is used to prediction.

4) Once the prediction is made, the predictions are saved in a CSV file at a given location and the location is returned to the client.