BackOrder Prediction

Detail Project Report

Objective:

Development of a predictive model for monitoring Backorder for private supermarket. The model will determine whether the products will go backorder or not.

Benefits:

- Detecting the Backorder of product
- Identifying the indent product based on the model prediction.
- Manual inspection if product backorder is identified.
- Helps in easy flow of product

Data Sharing Agreement:

Length of date stamp (8)

Length of time stamp (6)

Number of columns(23)

Column names

Sku, national_inv, lead_time, in_transit_qty, forecast_3_month, forecast_6_month, forecast_9_month, sales_1_month, sales_3_month, sales_6_month, sales_9_month, min_bank, potential_issue, pieces_past_due, perf_6_month_avg, perf_12_month_avg, local_bo_qty, deck_risk, oe_constraint, ppap_risk, stop_auto_buy, rev_stop, went_on_backorder.

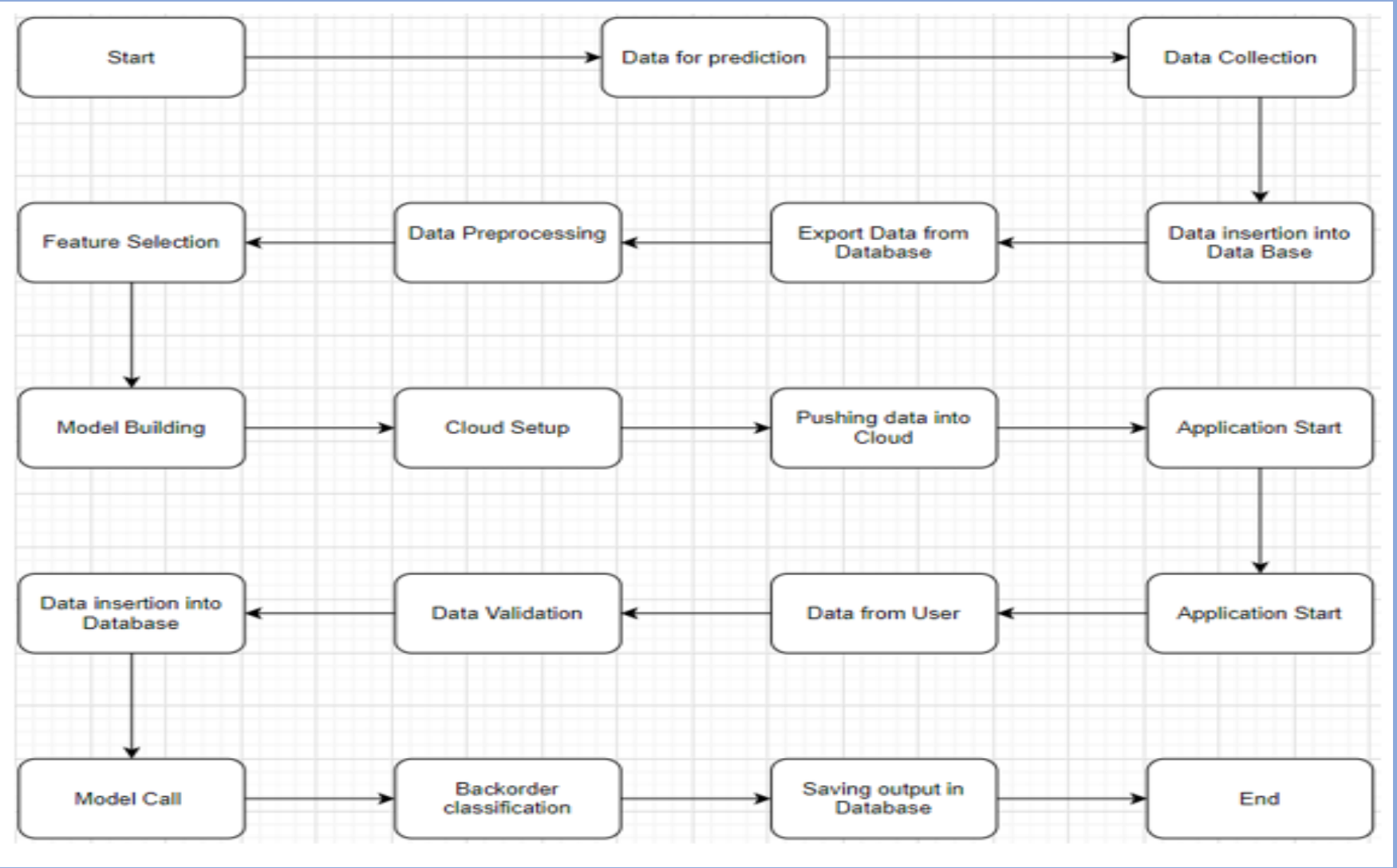
Column data types

Sku : INTEGER, national_inv : FLOAT, lead_time : FLOAT, in_transit_qty : FLOAT,

forecast_3_month : FLOAT, forecast_6_month : FLOAT, forecast_9_month : FLOAT,

sales_1_month : FLOAT, sales_3_month : FLOAT, sales_6_month : FLOAT, sales_9_month : FLOAT, min_bank : FLOAT, potential_issue : VARCHAR, pieces_past_due : FLOAT, perf_6_month_avg : FLOAT, perf_12_month_avg : FLOAT, local_bo_qty : FLOAT , deck_risk : VARCHAR, oe_constraint : VARCHAR, ppap_risk : VARCHAR, stop_auto_buy : VARCHAR, rev_stop : VARCHAR, went_on_backorder : VARCHAR.

ARCHITECTURE:



Data Validation and Data Transformation:

- Name Validation – Validation of file name as per the DSA. We have created a regex pattern for validation. After if checks for date format and time format if these requirements are satisfied, we move such files to “Good_Data_Folder” else “Bad_Data_Folder”.
- Number of Columns – Validation of number of columns present in the files and if it doesn’t match then the file is moved to “Bad_data_Folder”.
- Name of Columns – The name of columns is validated and should be the same as given in the schema file. It not then the file is moved to “Bad_Data_Folder”.
- Data type of columns – The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong , then the file is moved to “Bad_Data_Folder”.
- Null values in columns – If any of the columns in a file have all the values as NULL or missing, I discard such file and move it to “Bad_Data_Folder”.

Data Insertion in Database:

- Table creation – Table name “Good_raw_data” is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- 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 data type in any of the columns, the file is not loaded in the table.

Model Training:

❖ Data Export from Db:

- The accumulated data from DB is exported in csv format for model training

❖ Data Preprocessing

- Performing EDA to get insight of data like identifying distribution, outliers, trend among data.
- Check for null value in the columns. If present remove the null values.
- Encode the categorical values with numeric values.
- Perform Standard Scalar to scale down the values
- Perform Principle component Analysis and reduce the dimension of the data

Model Selection:

After the PCA is performed, we find the best model for data. By using 2 algorithms “Random Forest” and “XG Boost”. For each algorithm hyper parameter tuning is done. We calculate the AUC score for both models and select the model with best score.

If target variable is having only 1 class then we use ACCURACY score for both models and select the model with best score.

Prediction:

- The testing files are shared in batches and we perform the same validation operations, data transformation and data insertion on them.
- We perform data pre-processing techniques on it.
- After performing pre-processing techniques respective model is loaded and is used to predict the data.
- Once the prediction is done for all data, the prediction is saved in CSV format and shared.

Logs:

We are using different logs as per the steps that we follow in validation and modeling like file validation log, Data insertion log, model training log, prediction log and many more.