BackOrder Prediction

MACHINE LEARNING PROJECT

Detailed Project Report
On
BACKORDER PREDICTION

Made by: Gaurav Singh
Domain: E-commerce

Objective

- 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.

Benifits

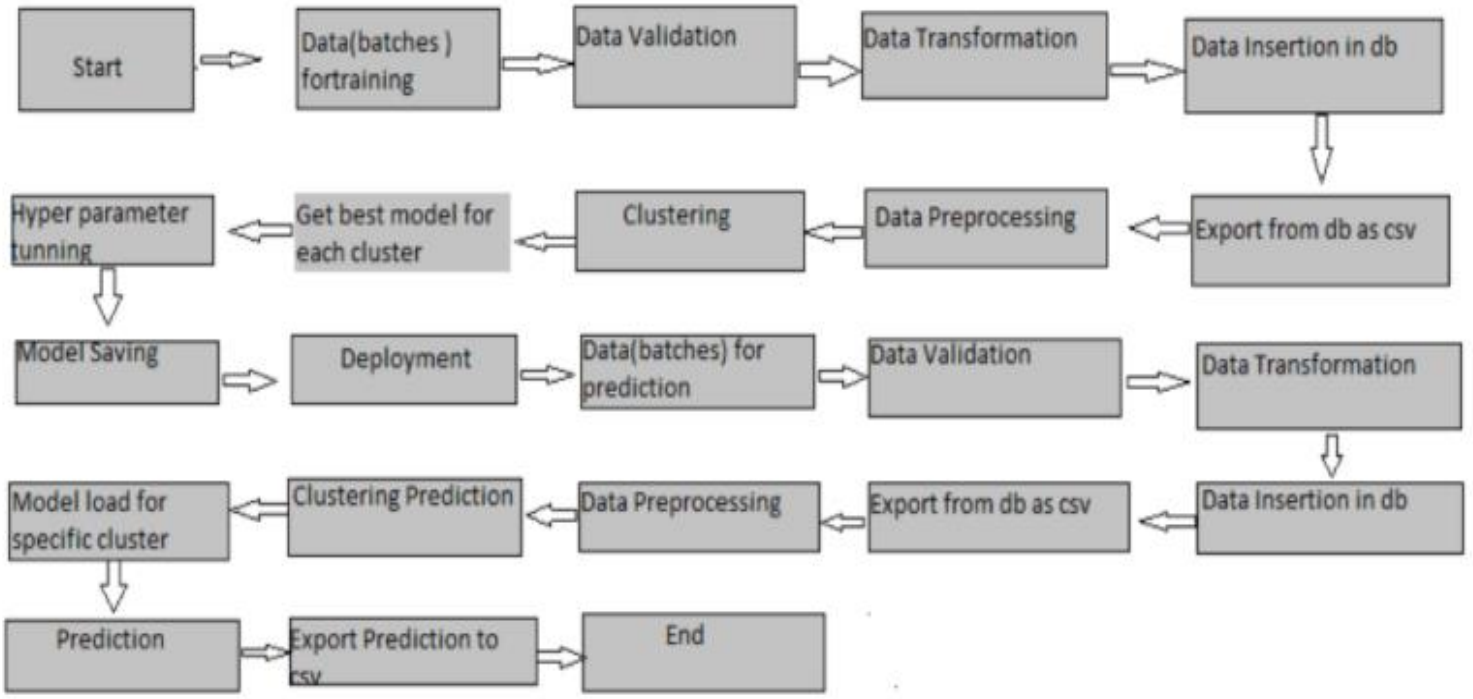
- ***True positive***: *Benefit by predicting correctly the backorder SKUs. Profit generated from such items is benefit.*
- ***True negative***: *Benefit by predicting SKUs not in backorder list correctly. Though benefit from this is zero as customer has not bought item, storage cost saved in addition to opportunity cost for not manufacturing such items is also benefit.*

Data Sharing Agreement

- Sample file name(Ex:backorder_02082021_010101.csv)
- Length of date stamp(8 digits)
- Length of time stamp(6 Digits)
- Number of columns
- Name of column names
- Columns Data type
- Column Detail

```
    "SampleFileName": "BackOrder_08012020_120000.csv",
                        "LengthOfDateStampInFile": 8,
                        "LengthOfTimeStampInFile": 6,
                        "NumberofColumns" : 23 ,
                        "ColName": {
            "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" : "object",
            "pieces_past_due" :   "float" ,
            "perf_6_month_avg"  : "float" ,
            "perf_12_month_avg" : "float" ,
            "local_bo_qty": "float" ,
            "deck_risk": "object" ,
            "oe_constraint": "object",
            "ppap_risk": "object",
            "stop_auto_buy": "object",
            "rev_stop": "object",
            "went_on_backorder": "object"
                        }
    }
```

Architecture



Data Validation & Transformation

- Name Validation - Validation of files name as per the DSA. We have created a regex pattern for validation. After it 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 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".
- 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, we discard such a file and move it to "Bad_Data_Folder".

Data Insertion in Database:

- Table creation :- Table name “good_training_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 invalid 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 etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Perform Standard Scalar to scale down the values.

Model Selection

- In this machine learning classifiers are investigated in order to propose a predictive model for this imbalanced class problem, where the relative frequency of items that goes into backorder is rare when compared to items that do not. Specific metrics such as area under the Receiver Operator Characteristic and precision-recall curves, sampling techniques and ensemble learning are employed in this particular task.

Prediction

- The testing files are shared in the batches and we perform the same Validation operations ,data transformation and data insertion on them.
  - The accumulated data from db is exported in csv format for prediction
  - We perform data pre-processing techniques on it.
  - KMeans model created during training is loaded and clusters for the preprocessed data is predicted
  - Based on the cluster number respective model is loaded and is used to predict the data for that cluster.
- Once the prediction is done for all the clusters. The predictions are saved in csv format and shared

Questions & Answers

- Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

- Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

- Q 3) What's the complete flow you followed in this Project?

Refer slide 5th for better Understanding

- Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

- Q 5) How logs are managed?

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?
  - Removing unwanted attributes
  - Visualizing relation of independent variables with each other and output variables
  - Checking and changing Distribution of continuous values
  - Removing outliers
  - Cleaning data and imputing if null values are present.
  - Converting categorical data into numeric values.
  - Scaling the data

- Q7 How did you optimize your solution?

1) Model optimization depends on various factors

2) Train with better data or do data pre-processing in efficient way.

3) Increase the quantity of training data etc.

- Q8 At what frequency are u retraining and updating your model?

The model gets retrained every 30 days