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

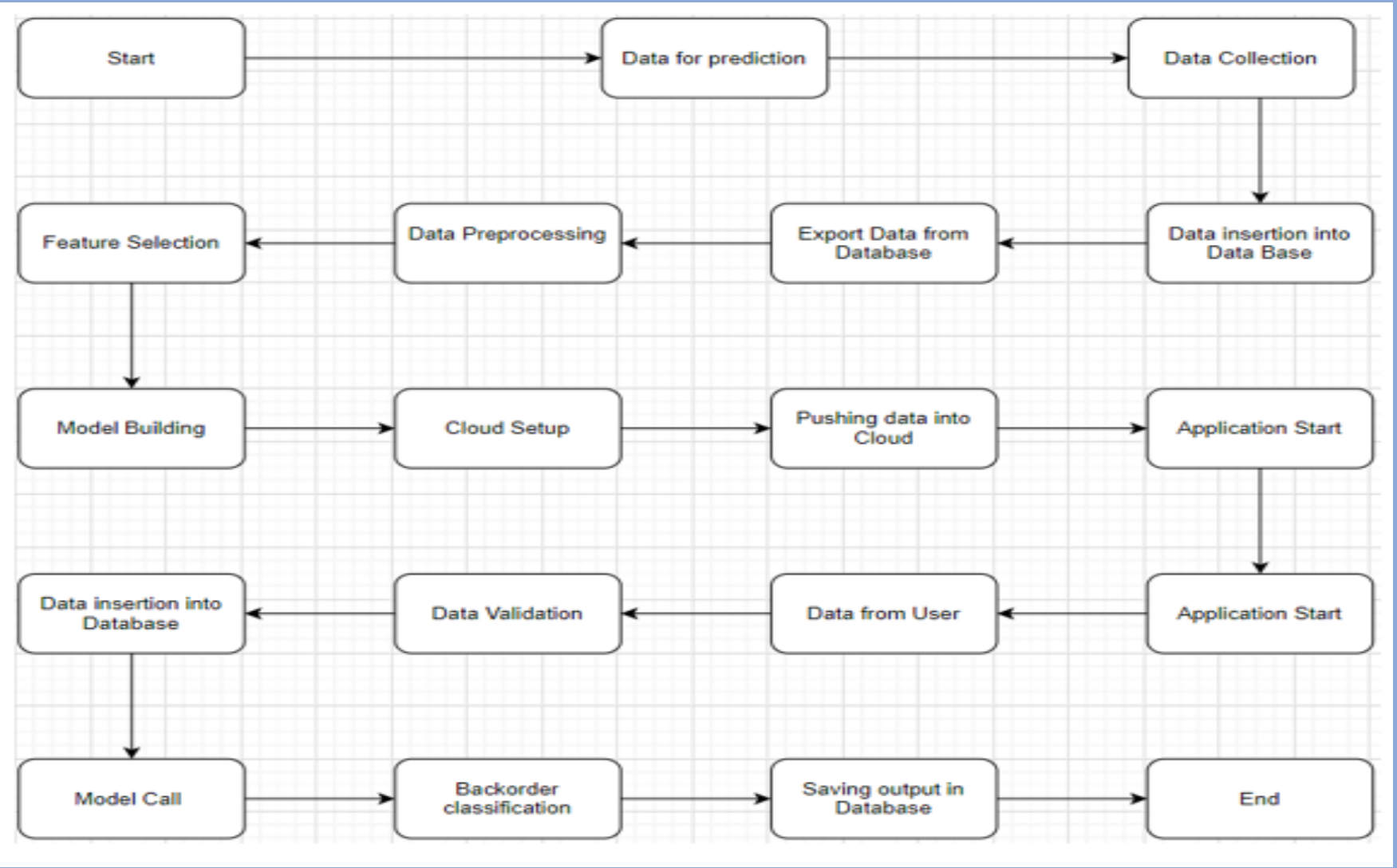
Sku : INTEGER, national_inv : FLOAT, lead_time : FLOAT, in_transit_qty : FLOAT,

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.

```
"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": "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.

FREQUENTLY ASKED QUESTIONS :-

Question 1: Explain about the Project and your day to day task :

Answer : Product backorder may be the result of strong sales performance (e.g. the product is in such high demand that production cannot keep up with sales). However, backorders can upset consumers, lead to canceled orders and decreased customer loyalty. Companies want to avoid backorders, but also avoid overstocking every product (leading to higher inventory costs). As a data scientist I am involving in each an every phase of the project. My responsibility consisted of gathering the dataset ,labelling the data for the model, training the model on the prepared dataset , deploying the training model to the cloud, monitoring the deployed model for any issues. Mixed in are calls, stand ups and the attending Scrum meeting.

Question 2 : How Logs Are Managed?

Answer : 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.

Question 2 : What is the source and size of data ?

Answer : The data for train is provided by client in batches . Size of the data usually in MB.

Question 3 : How Prediction Was Done?

Answer : The Testing Files Are Shared By The Client .We Perform The Same Life Cycle Till The Data Is Clustered. Then On The Basis Of Cluster Number Model Is Loaded And Perform Prediction. In The End We Get The Accumulated Data Of Predictions.

Question 4 : What is AUC Curve ?

Answer : AUC stands for "Area under the ROC Curve" .AUC measures the entire 2D area underneath the entire ROC curve.

Question 5 : What Is The Type Of Data?

Answer : The Data Is The Combination Of Numerical And Categorical Values.

Question 6 : What techniques r you using for data pre-processing ?

Answer : 1) Removing unwanted attributes.

2) Visualizing relation of independent variables with each other and with dependent variable.

3) Removing Outliers.

4) Cleaning data and imputing if null values are present.

5) Convert Categorical data to numerical data.

6) Scaling the data.

7) Dimensionality reduction

Question 7 : Does Your Dataset Show Normally Distributed Or Not? If Not Then Which Techniques You Will Use To Make It Normal?

Answer : No, These Data Set Does Not Show Normal Distribution Behavior. I Used Reciprocal, Square, Log, Exponential Techniques To Make It Normally Distributes.

Question 8 : Which Tool You Are Used For Implementation This Model?

Answer :

1) Ide : Pycharm

2) Cloud : AWS

3) Data Base : Cassandra

Question 9 : What Kind of challenges have u faced during the project?

Answer : The biggest challenge I face in project is in obtaining good dataset , cleaning it to be fit for model and then labeling prepared dataset. Labeling is a time consuming task and it takes lots of our. Then comes the task of finding the correct algorithm to be used for business case.

Question 10 : What Is Accuracy ?

Answer : Accuracy Is One Metric For Evaluating Classification Models. Accuracy = Number Of Correct Predictions /Total Number Of Predictions

Question 11 : How did you optimize your solution?

Answer : 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.

4) Try and use multithreaded approaches.

Question 12 : At what frequency are u retraining and updating your model?

Answer : The model gets retrained every 30 days

Question 13: How did you optimize your solution?

Answer : 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.

4) Try and use multithreaded approaches

Question 14 : What is Overfitting, and How Can You Avoid It?

Answer : Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.

When a model is given the training data, it shows 100 percent accuracy technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.

There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used

• If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters