Back Order Prediction

Machine Learning Project

Melbin Biju

Objectives

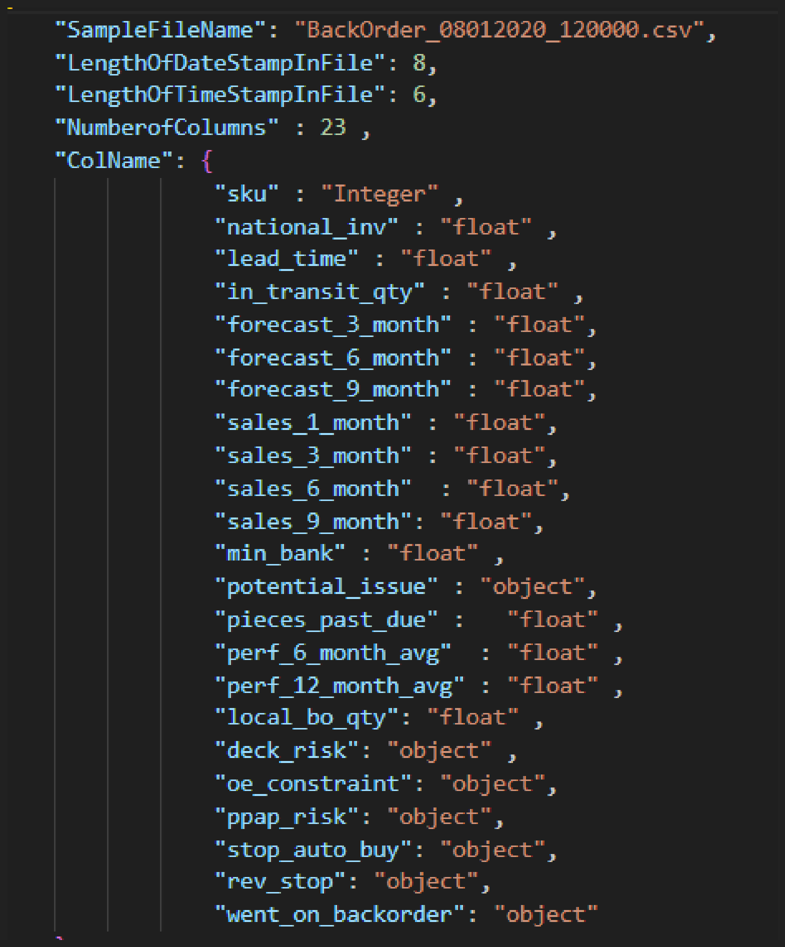
- A Backorder is an order which can’t be fulfilled at the given time due to lack of supply or the product is currently out of stock or not in inventory but can guarantee delivery of the goods or service requested by a certain date in the future because the production of goods or replenishment of inventory is underway. Unlike in the situation of Out-of-stock where the delivery date of the goods can’t be promised , in the Backorder scenario the customers are allowed to shop for the products and order. Simply put Backorder can be thought of as an order with a delayed delivery date.
- The primary goal of all the companies is to increase the demand for the products they offer. Having a poor sales forecast system could one of the reasons for failing to predict the demand. Despite having a good sales forecasting system sometimes these situations are inevitable because of the factors which can’t be controlled or un predictable events.

Benefits :

- Backorders are inevitable but through prediction of the items which may go on backorder planning can be optimized at different levels avoiding unexpected burden on production , logistics and transportation planning.
- ERP systems produce a lot of data (mostly structured) and also would have a lot of historical data , if this data can be leveraged correctly a Predictive model can be developed to forecast the Backorders and plan accordingly.

Data Sharing Agreement File For Training Data Set :

- 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 Details



```
"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"
```

Data Description :

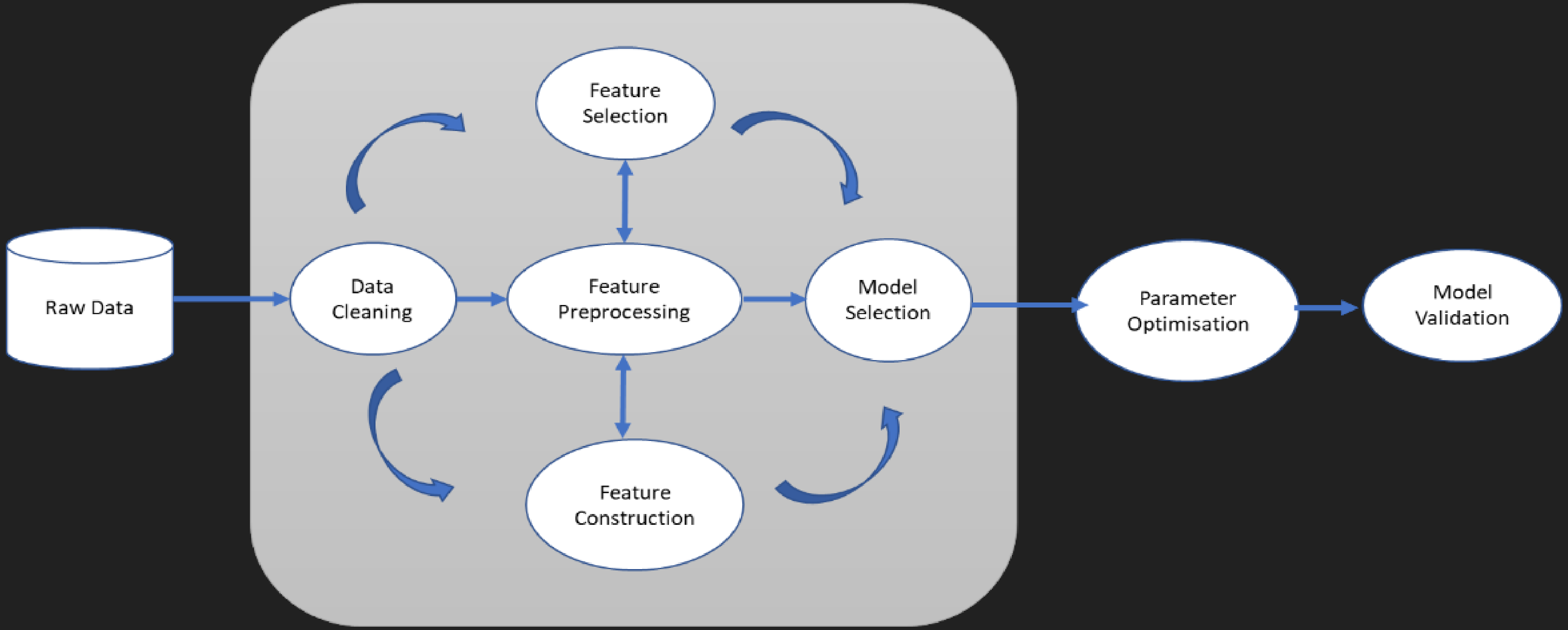
The Client Will Send Data In Multiple Sets Of Files In Batches At A Given Location. Data Will Contain 24 Columns:

- **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_month** – 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.

Application Architecture :



Data Validation & Data Transformation :

- **File Name Validation :** File name validation as per the DSA. We have created regular expression pattern for validation with time stamp format. IF this file satisfy the pattern criteria it will go to valid_data_file folder or else it will go to bad_data file folder.
- **Name and no. of columns :** It will check for number of columns and Name of the columns. If it will pass the validation criteria then it will goto valid_data_file or else bad_data_file .
- **Data types of column :** The datatype of columns is given in the schema file. This is validated when we insert the files into database. If the datatype is incorrect, then the file is moved to "bad_data_folder".
- **Null values in Columns :** If any columns in file contain null values or data ids missing then this file is going to bad_file_folder.

Database :

- **Database creation and connection :** Create the database with the key space name passed (If it's not already present). Connect with the database.
- **Table Creation : we** have to create a table in Database in that key space (If it's not already present) to insert valid data files , if Table is already created then we need to insert new files into database.
- **Insertion :** All valid files are inserted into the tables.

Model Training :

- **Data export from DB**: The data in A stored database is exported as A CSV file to be used for model training.
- **Exploratory Data Analysis & Data Pre-processing:**

  1) Missing value count

  2) No of rows and columns(Shape)

  3) categorical/Numerical columns

  4) Correlation heat map

  5) Null value handling(impute null value)

  6) Outlier detection and remove

  7) Perform standard scalar for scaling down

Model Selection:

- Evaluates classification models using Logistic Regression ,Random forest , Decision Tree , XGBoost through exhaustive Random search, using stratified 5-fold cross-validation, and Area Under Precision-Recall Curve (AUPRC) scorer.
- Compute metrics and generate graphs for model evaluation and importance analysis
- We view AUC values for each model and plot the ROC curves for the top random forest model and down sampled random forest model.
- Finally, we fit the random forest model with optimal tuning parameters on the entire dataset. We then could use this model to predict whether parts will go on backorder.

Prediction :

- The testing files are shared in batches and we perform the same validation operations, data transformation and data insertion on them.
- The accumulated data from database is exported in csv format for prediction.
- We perform data pre-processing techniques in it.

Question & Answers

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.

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 : VSCode

2) Cloud : AWS

3) Data Base : Cassandra

Question 9 : How were you maintain the failure cases?

Answer : If our model is not predicting correctly for data then that dataset goes to database . There will be a report triggered to the support team at the end of the day with all failure scenarios where they can inspect the failure. Once we have a sufficient number of cases we can label and include those data while retraining the model for better performance.

Question 10 : In which technology you are most comfortable?

Answer : I have worked in almost all the fields like machine learning Deep learning and NLP. But personally I prefer deep learning.

Question 11 : 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 12 : What Is Accuracy ?

Answer : Accuracy Is One Metric For Evaluating Classification Models.

Accuracy = Number Of Correct Predictions /Total Number Of Predictions

Question 13 : What will be your expectations?

Answer : I expect to work on different projects to enhance my technical skill and learn new things simultaneously.

Question 14 : What is your future objective?

Answer : My future objective is to learn new things in AI field because it changes continuously, and my aim is to pursue my career as a solution architect near future.

Question 15 : 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 16 : At what frequency are u retraining and updating your model?

Answer : The model gets retrained every 30 days

Question 17 : 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 18 : At what frequency are u retraining and updating your model?

Answer : The model gets retrained every 30 days

Question 19 : Which is more important to you model accuracy or model performance?

Answer : Well, you must know that model accuracy is only a subset of model performance. The accuracy of the model and performance of the model are directly proportional and hence better the performance of the model, more accurate are the predictions.

Question 20 : 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

Question 21 : Explain the Confusion Matrix with Respect to Machine Learning Algorithms ?

Answer : A confusion matrix (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it’s called the matching matrix.

The confusion matrix has two parameters:

- Actual
- Predicted

It also has identical sets of features in both of these dimensions.

Question 22 : How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem??

Answer : While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias