BACKORDER PREDICTION DETAILED PROJECT REPORT

End-to-End Machine Learning Project by:-

Amar Hulamani
B.Tech in Production Engineering, Nit Trichy

Objectives

- A backorder occurs when a product is temporarily out of stock or not available in inventory but can still be ordered with a guaranteed delivery date in the future. This typically happens when production is ongoing, or inventory replenishment is in progress. Unlike a typical out-of-stock scenario—where the availability of the product remains uncertain—backorders allow customers to place orders in advance, ensuring fulfillment once stock is replenished.
- For businesses, the primary goal is to maintain a balance between supply and demand to maximize sales and customer satisfaction. However, even with an efficient sales forecasting system, backorders may still occur due to unexpected demand spikes, supply chain disruptions, or production delays. When backorders become frequent, it signals inefficiencies in demand forecasting, inventory planning, or supplier performance, potentially leading to lost revenue and dissatisfied customers.
- A well-managed backorder system can help businesses retain customers, prevent lost sales, and optimize inventory management, ensuring a more agile and responsive supply chain. However, excessive backorders may indicate poor operational planning, causing logistical strain, increased lead times, and potential reputational damage. This is why predictive modeling plays a crucial role in helping businesses anticipate and prevent excessive backorders, improving overall supply chain efficiency.

Benefits :

- Backorders are inevitable in supply chain operations, but by accurately predicting which products are likely to go on backorder, businesses can optimize planning at multiple levels. This helps in reducing unexpected strain on production, logistics, and transportation, ensuring smoother operations and improved customer satisfaction.
- Enterprise Resource Planning (ERP) systems generate vast amounts of structured data, including historical sales, inventory levels, supplier performance, and demand trends. By leveraging this data effectively, businesses can develop a predictive model to forecast backorders in advance. This proactive approach allows companies to streamline inventory management, enhance demand forecasting, and minimize supply chain disruptions, ultimately leading to better operational efficiency and profitability.

Data Sharing Agreement File For Training Data Set :

- Sample file name(Ex:backorder.csv)
- Length of datestamp(8digits)
- Length of timestamp(6 Digits)
- Number of columns
- Column names
- Columns Data type

```
 #   Column             Dtype
---  ------             -----
 0   sku                object
 1   national_inv       float64
 2   lead_time          float64
 3   in_transit_qty     float64
 4   forecast_3_month   float64
 5   forecast_6_month   float64
 6   forecast_9_month   float64
 7   sales_1_month      float64
 8   sales_3_month      float64
 9   sales_6_month      float64
 10  sales_9_month      float64
 11  min_bank           float64
 12  potential_issue    object
 13  pieces_past_due    float64
 14  perf_6_month_avg   float64
 15  perf_12_month_avg  float64
 16  local_bo_qty       float64
 17  deck_risk          object
 18  oe_constraint      object
 19  ppap_risk          object
 20  stop_auto_buy      object
 21  rev_stop           object
 22  went_on_backorder  object
dtypes: float64(15), object(8)
memory usage: 353.4+ MB
```

Data Description :

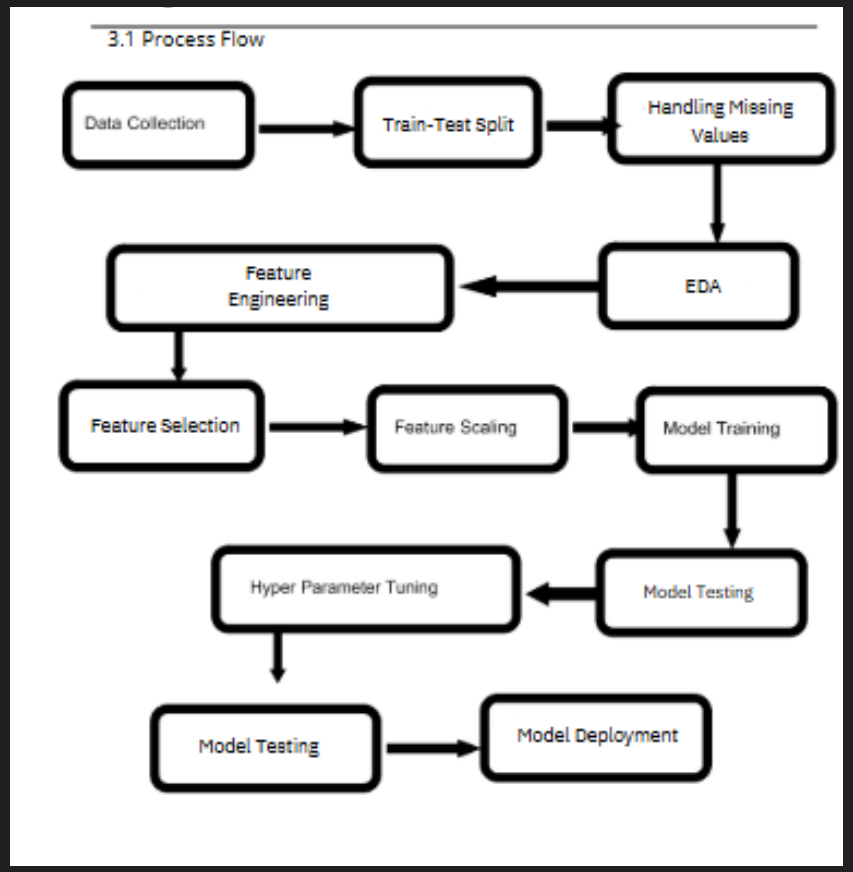
THE CLIENT WILL SEND DATA IN MULTIPLE SETS OF FILES IN BATCHES AT A GIVEN LOCATION. DATA WILL CONTAIN 23 COLUMNS:

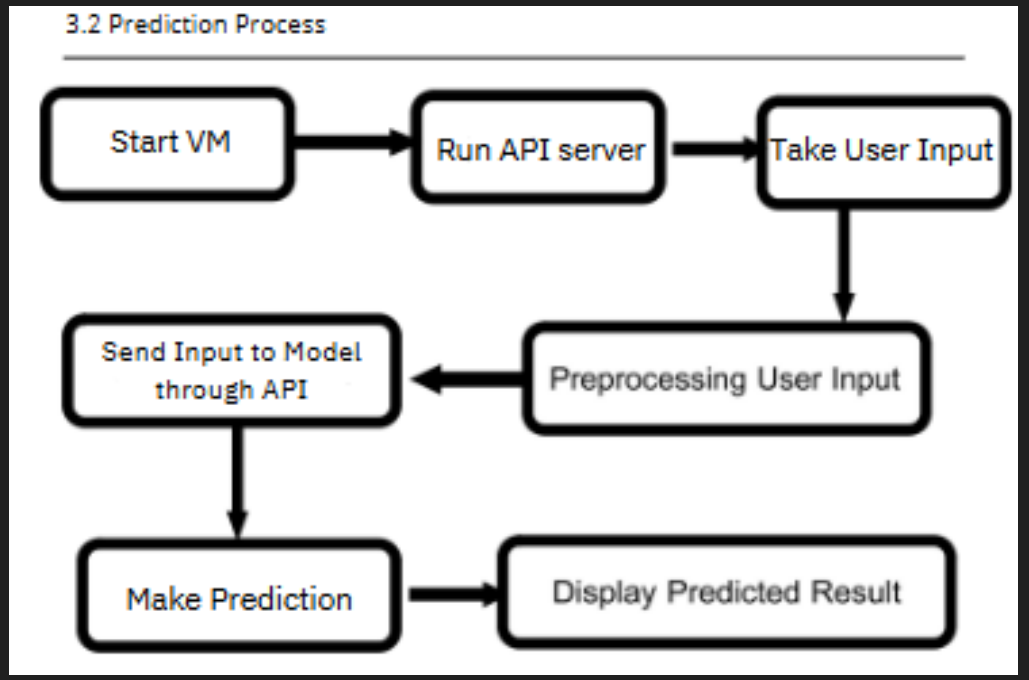
- **sku (Stock Keeping Unit) :** Unique Alphanumeric code assigned by the retailer to each product.
- n**ational_inv :** Current National Inventory of the product
- **lead_time** : Lead Time of the product
- **in_transit_qty :** Total quantity of products in transit
- **forecast_3_month :** Forecasted Sales of the product for next 3 months
- **forecast_6_month :** Forecasted Sales of the product for next 6 months
- **forecast_9_month :** Forecasted Sales of the product for next 9 months
- **sales_1_month :** Previous month Sales of the product

- **sales_3_month :** Previous 3 months Sales of the product
- **sales_6_month :** Previous 6 months Sales of the product
- **sales_9_month :** Previous 9 months Sales of the product
- **min_bank :** Minimum Quantity of Products required to keep in stock
- **potential_issue :** Is there a potential issue with the product?
- **pieces_past_due :** No of pieces/products overdued
- **perf_6_month_avg :** Average Source Performance over the last 6 month
- **perf_12_month_avg :** Average Source Performance over the last 12 month
- **local_bo_qty :** Current Unmet demand

- **deck_risk :** Is there a risk of mismatch between forecasted and actual demand of the product?
- **oe_constraint :** Did the product experience operational constraints?
- **ppap_risk :** Is there a Production Part Approval Process risk associated with the product?
- **stop_auto_buy :** Is automatic purchasing for this product halted?
- **rev_stop :** Is review process for this product paused?
- **went_on_backorder :** If the product will go to backorder or not. (Target Variable)

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".

Model Training & Selection:

- Data export from DB: The data in Casandra database is exported as ACSV file to be used for model training.
- Exploratory Data Analysis (EDA) & 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) Normalization of skewed features
    7) OneHotEncoding for categorical feature
    8) Feature selection

- Models like LogisticRegression, SVM, Randomforest, DecisionTree, Gradient Boosting, XGBoost, LightGBM with sampling techniques like RUS, ROS, SMOTE and imblearn ensemble techniques like Balanced Random Forest and Easy Ensemble were fitted by performing exhaustive Gridsearch or stratified5-foldcross-validation to get the Train and Test score simultaneously.
- All the models were evaluated using metrics like AUCROC, AUCPR, & recall.
- Finally Balanced Random Forest was choosen as the best performing model in terms of recall (false negatives) and this model was trained again on the whole dataset. This same model was then used for predicting the backorder situation of the products.

Prediction :

- The user input is taken through Streamlit Web application where all the necessary input fields values are send to a backend api which acts as a midman between the user interface and the model.
- The API preprocesses the data before sending it to the model for prediction. The model takes the input data and gives the appropriate prediction of the backorder situation. This prediction is sent back to the UI through API and is displayed to the user.

Question & Answers

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

Backorders occur when a product is temporarily unavailable due to high demand, supply chain delays, or insufficient inventory levels. While strong sales performance can contribute to backorders, excessive backorders may lead to customer dissatisfaction, canceled orders, and reduced brand loyalty. Businesses aim to minimize backorders while avoiding excessive stock, which increases inventory costs.

As a data scientist working on the Backorder Prediction System, I am involved in every phase of the project, ensuring a robust machine learning pipeline. My responsibilities include:

- Data Collection & Preprocessing: Gathering the ERP dataset, handling missing values, and performing feature engineering.
- Model Training & Selection: Experimenting with various machine learning algorithms such as Random Forest, XGBoost, and Balanced Random Forest to optimize prediction accuracy.
- Hyperparameter Tuning & Model Evaluation: Ensuring that the model effectively distinguishes between products likely to go on backorder and those that will not.
- Model Deployment: Deploying the trained model using FastAPI as the backend and Streamlit as the frontend on an Azure Linux VM.
- Collaboration & Agile Workflow: Attending daily stand-ups, Scrum meetings, and sprint planning to coordinate with team members and ensure smooth project execution.

This end-to-end workflow helps businesses anticipate and mitigate backorders, improving inventory management and customer satisfaction while optimizing supply chain operations.

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

The data for train is provided by client in batches . Size of the data is 350 MB.

Question 3 : How Prediction Was Done?

- Receiving Data – The client provides the testing dataset, which undergoes the same preprocessing steps as training data.
- Data Preprocessing – Handling missing values, feature scaling, and encoding to maintain consistency.
- Model Selection – If clustering is used, data is assigned to a predefined cluster, and the corresponding model is loaded.
- Prediction – The trained model (Balanced Random Forest, XGBoost, etc.) predicts whether a product will go on backorder (Yes/No).
- Results Compilation – Predictions are shared with the client to help in inventory management and supply chain planning.

Question 4 : What is AUC Curve ?

AUC stands for "**Area under the ROC Curve**". AUC measures the entire 2D area underneath the entire ROC curve (Receiver Operating Charateristics).

Question 5 : What Is The Type Of Data?

The data is a structured labeled data which contains both numerical and categorical features.

Question 6 : What Kind of challenges have u faced during the project?

The biggest challenge I face in project is in obtaining good dataset , cleaning it to be fit for model and then comes the task of finding the correct algorithm to be used for business case.

Question 7 : What Is Recall?

Recall is one of the metric used for classification problems, particularly in situations where the cost of missing a positive class is high.
Recall is the proportion of actual positive instances that were correctly identified by the model. In other words, it measures the ability of the model to identify all relevant positive cases.
Where:

- True Positives (TP): The number of positive instances that were correctly predicted as positive.
- False Negatives (FN): The number of positive instances that were incorrectly predicted as negative.

Question 8 : What will be your expectations?

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

simultaneously.

Question 9 : What is your future objective?

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 data scientist or data analyst in the near future.

Question 10 : How did you optimize your solution?

Model optimization depends on various factors:-

1) Train with better data or do data pre-processing in efficient way.
2) Increase the quantity of training data etc.
3) Try and use multithreaded approaches

Question 11 : At what frequency are u retraining and updating your model?

The model gets retrained every 30 days

Question12 : What is Overfitting, and How Can You Avoid It?

Overfitting occurs when a model learns the training data too well, including noise and random fluctuations, which hurts its ability to generalize to new data. It results in high accuracy on training data but poor performance on test data.

Ways to Avoid Overfitting:

1. Regularization: Adds a penalty to the loss function (e.g., L1/L2 regularization) to prevent the model from becoming too complex.
2. Simplify the Model: Reduce the number of parameters or features, or use dimensionality reduction techniques.
3. Cross-Validation: Use techniques like k-fold cross-validation to ensure the model generalizes well.
4. Pruning: In decision trees, remove unnecessary branches to prevent overfitting.
5. Early Stopping: Stop training when performance on the validation set starts to degrade.
6. Increase Training Data: More data helps the model generalize better.
7. Ensemble Methods: Combine multiple models to improve performance and reduce overfitting.
8. Dropout: In neural networks, randomly drop neurons during training to prevent over-reliance on any single neuron.

These methods help strike a balance between fitting the data well and maintaining the model’s ability to generalize to new data.