

horizontal line

PROJECT INTERIM REPORT

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT Chennai Feb 2022 |
| Team members | 1. Vigneshwar K 2. Tamilselvan M 3. Balaji prasanth S 4. Rahul R 5. Adarsh R |
| Domain of Project | E-Commerce (Supply-Chain) |
| Proposed project title | BackOrder Prediction |
| Group Number | Group 11 |
| Team Leader | Vigneshwar K |
| Mentor Name | Pratik Sonar |

Date: 16-06-2022

Signature of the Mentor Signature of the Team Leader





## TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TOPIC** | **PAGE NO** |
| **1** | **BUSINESS UNDERSTANDING** | **4** |
| 1.1 BUSINESS PROBLEM STATEMENT | 4 |
| 1.2 TOPIC SURVEY | 4 |
| 1.3 CRITICAL ASSESSMENT OF TOPIC SURVEY | 5 |
| **2** | **DATA UNDERSTANDING** | **6** |
| 2.1 DATA DICTIONARY | 6 |
| 2.2 VARIABLE CATEGORIZATION | 11 |
| 2.3 DISTRIBUTION OF VARIABLES | 11 |
| 2.4 REDUNDANT VARIABLES | 13 |
| **3** | **DATA PREPROCESSING** | **14** |
| 3.1 NULL VALUE TREATMENT | 14 |
| 3.2 PRESENCE OF OUTLIERS AND TREATMENT | 15 |
| 3.3 CHECKING FROM MULTICOLLINEARITY TREATMENT | 16 |
| **4** | **EXPLORATORY DATA ANALYSIS** | **18** |





|  |  |  |
| --- | --- | --- |
|  | 4.1 BIVARIANT AND MULTI-VARIANT ANALYSIS | 18 |
| **5** | **BASELINE MODEL BUILDING** | **23** |
| 5.1 LOGISTIC REGRESSION | 23 |
| 5.2 NAIVE BAYES | 24 |
| 5.3 RANDOM FOREST CLASSIFIER | 25 |
| **6** | **REFERENCES** | **27** |





# BUSINESS UNDERSTANDING:

When a customer orders a product, which is not readily available due to lack of unavailability of the product in the store or not in the inventory but can guarantee the delivery of the ordered product a certain date in the future and the customer waits for the same. This scenario is called backorder of that specific product.

Backordering has good and bad faces in the business. So there has to be some balance between the demand and supply. ML Predictive analysis can schedule the production, supply chain and inventory helps to forecast product delivery delay which in return increase the customer satisfaction.

If the back orders are not handled promptly it reflects high impacts on the company’s revenue, share market price, and customer’s trust that leads to losing the customer as well as the order. On the other hand, to satisfy backorders leads to enormous pressure on different stages of the supply chain which may break (exhaust) or can cause extra costs on production, shipment …etc.

In the modern world of E-commerce, a wide range of products are readily available at customers disposal. Consequently, the number of orders handled by the E-commerce business holders in a day has sharply increased. Nevertheless, despite all the cutting-edge business models, machine learning techniques it's hard to cope up with extensive uncertainty in the number of orders and to maintain inventory according to it.

Nowadays, most of the company’s use Machine learning predictive analysis to predict products back orders to overcome the tangible and intangible costs of backorders. We have performed some hypothesis tests considering backorder scenarios. The outcomes of the hypothesis’s tests are helpful to choose the appropriate machine learning model for prediction.

## BUSINESS PROBLEM STATEMENT:

Part backorders is a common supply chain problem, wherein a customer places an order for a product that is temporarily out of stock. The percentage of items backordered and the number of backorder days are important measures of the quality of a company's customer service and the effectiveness of its inventory management. A company can manage its inventory more efficiently using a prediction on the backorder risk for the products. Goal here is to use the past data around the backorders, and provide a prediction on the potential products for backorders.



## TOPIC SURVEY :

* + 1. **Problem understanding:**

Backorders are unavoidable, but by anticipating which things will be backordered, planning can be streamlined at several levels, preventing unexpected strain on production, logistics, and transportation. ERP systems generate a lot of data (mainly structured) and contain a lot of historical data; if this data can be properly utilized, a predictive model to forecast backorders and plan accordingly can be constructed. Based on past data from inventories, supply chain, and sales, classify the products as going into backorder (Yes or No).

## Current solution to the problem:

Despite having a good sales forecasting system sometimes these situations are inevitable because of the factors which can’t be controlled or unpredictable events. If many items are going on Backorders consistently it is a sign that companies’ operations are not properly planned and also there is a very high chance of missing out business on the products.

## Proposed solution to the problem:

The solution here is a Classification based Machine Learning model. It can be implemented by different classification algorithms (like Logistic Regression, Random Forest, Decision tree, XGBoost and so on. Here First we are performing Data pre-processing step, in which Data Profiling, feature engineering, feature selection, feature scaling, PCA steps are performed and then we are going to build model.

## CRITICAL ASSESSMENT OF TOPIC SURVEY:

Part backorders is a common supply chain problem, wherein a customer places an order for a product that is temporarily out of stock. The percentage of items backordered and the number of backorder days are important measures of the quality of a company's customer service and the effectiveness of its inventory management.

## A company can manage its inventory more efficiently using a prediction on the backorder risk for the products. Goal here is to use the past data and metadata around the backorders, and provide a prediction on the potential products for backorders. To accomplish this, we will design a model using sample data, analyze the data, process the data, establish a model, evaluate the output,



and so improve the prediction accuracy. This provides the project with a great deal of flexibility and adaptability.

# DATA UNDERSTANDING:

## DATA DICTIONARY:

|  |  |  |
| --- | --- | --- |
| **S.No** | **Feature Name** | **Feature Description** |
| **1.** | **national\_inv** | Current inventory level for the part |
| **2.** | **lead\_time** | Transit time for product (if available) |
| **3.** | **in\_transit\_qty** | Amount of product in transit from source |
| **4.** | **forecast\_3\_month** | Forecast sales for the next 3 months |
| **5.** | **forecast\_6\_month** | Forecast sales for the next 6 months |
| **6.** | **forecast\_9\_month** | Forecast sales for the next 9 months |
| **7.** | **sales\_1\_month** | Sales quantity for the prior 1-month time period |
| **8.** | **sales\_3\_month** | Sales quantity for the prior 3-month time period |
| **9.** | **sales\_6\_month** | Sales quantity for the prior 6-month time period |
| **10.** | **sales\_9\_month** | Sales quantity for the prior 9-month time period |
| **11.** | **min\_bank** | Minimum recommend amount to stock |
| **12.** | **potential\_issue** | Source issue for part identified |
| **13.** | **pieces\_past\_due** | Parts overdue from source |
| **14.** | **perf\_6\_month\_avg** | Source performance for prior 6 month period |
| **15.** | **perf\_12\_month\_avg** | Source performance for prior 12-month period |
| **16.** | **local\_bo\_qty** | Amount of stock orders overdue |
| **17.** | **deck\_risk** | Part risk flag |
| **18.** | **oe\_constraint** | Part risk flag |
| **19.** | **ppap\_risk** | Part risk flag |
| **20.** | **stop\_auto\_buy** | Part risk flag |
| **21.** | **rev\_stop** | Part risk flag |

## VARIABLE CATEGORIZATION :

### Independent variables:

Numerical column: 15

Categorical column: 7

### Target variable:

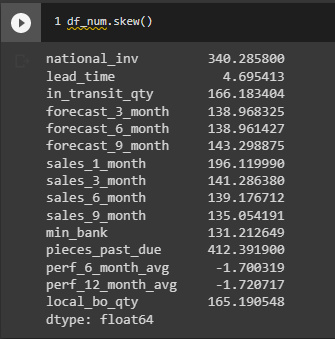
Categorical - 1

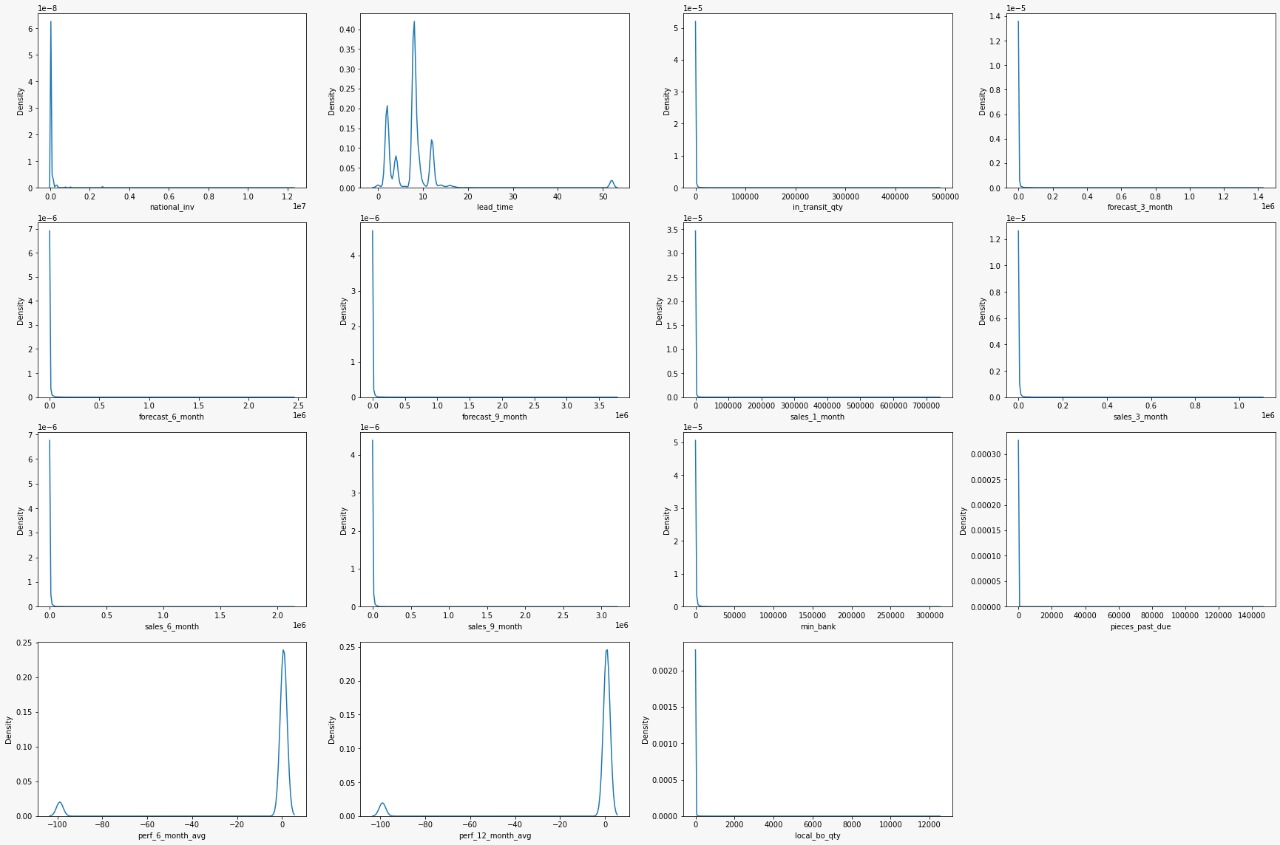
Total columns: 23

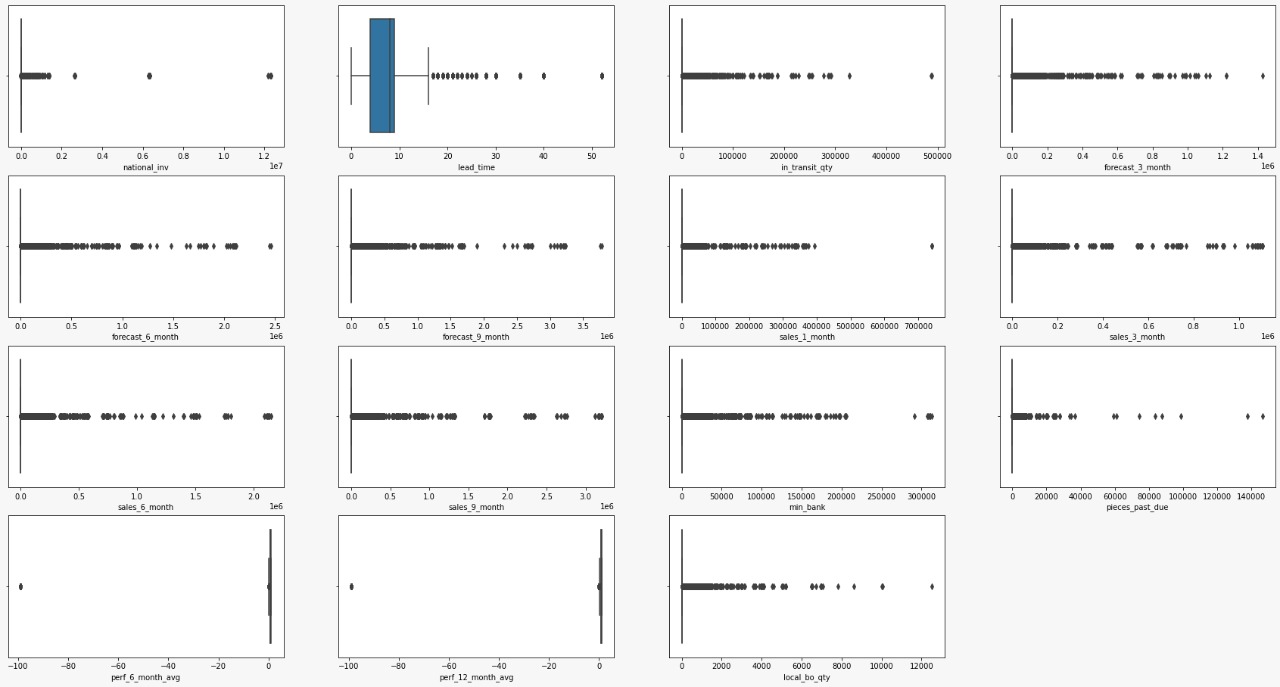
## DISTRIBUTION OF VARIABLES:

The data given to us is in shape a 1687860 \* 23 data frame. The data consists of Numerical and Categorical data. While further analyzing the data, we realize that the categorical data is mostly Binary flagged as 0 or 1. The sku column has a unique value for each row, so it is the index column and should be dropped. The numerical features have different scales, which may be a problem for some machine learning algorithms. The features should be rescaled to have similar scale

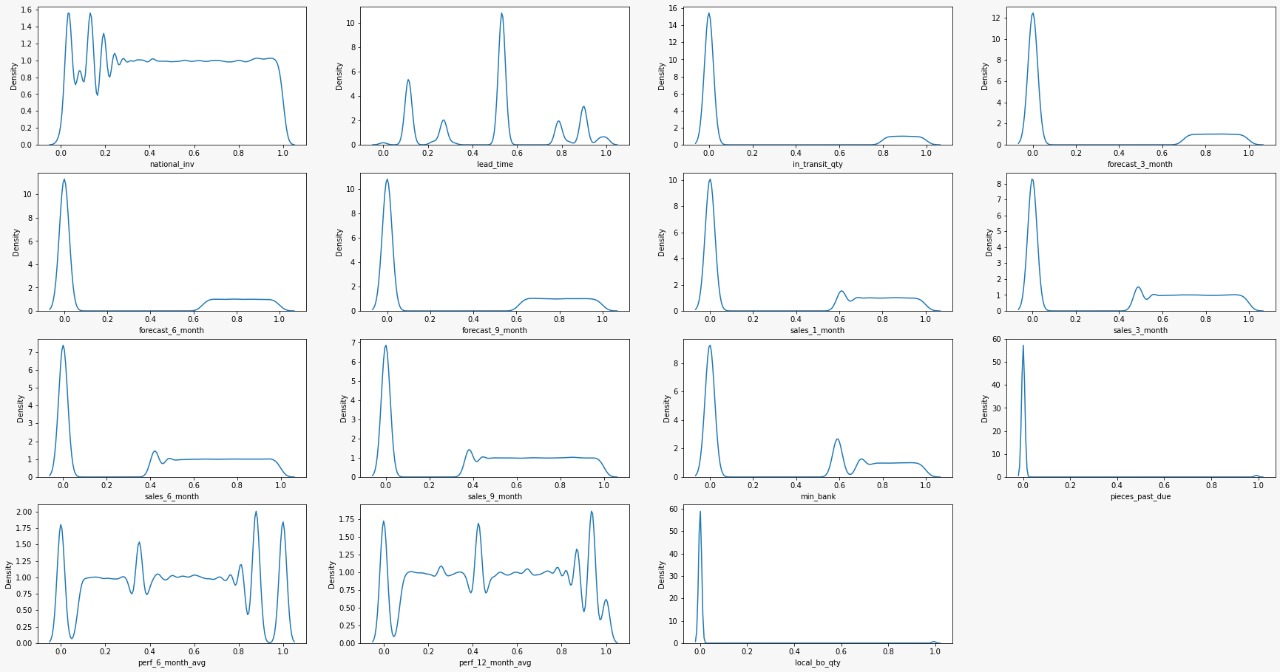
**Skewness of Original & Scaled Numeric Data:**



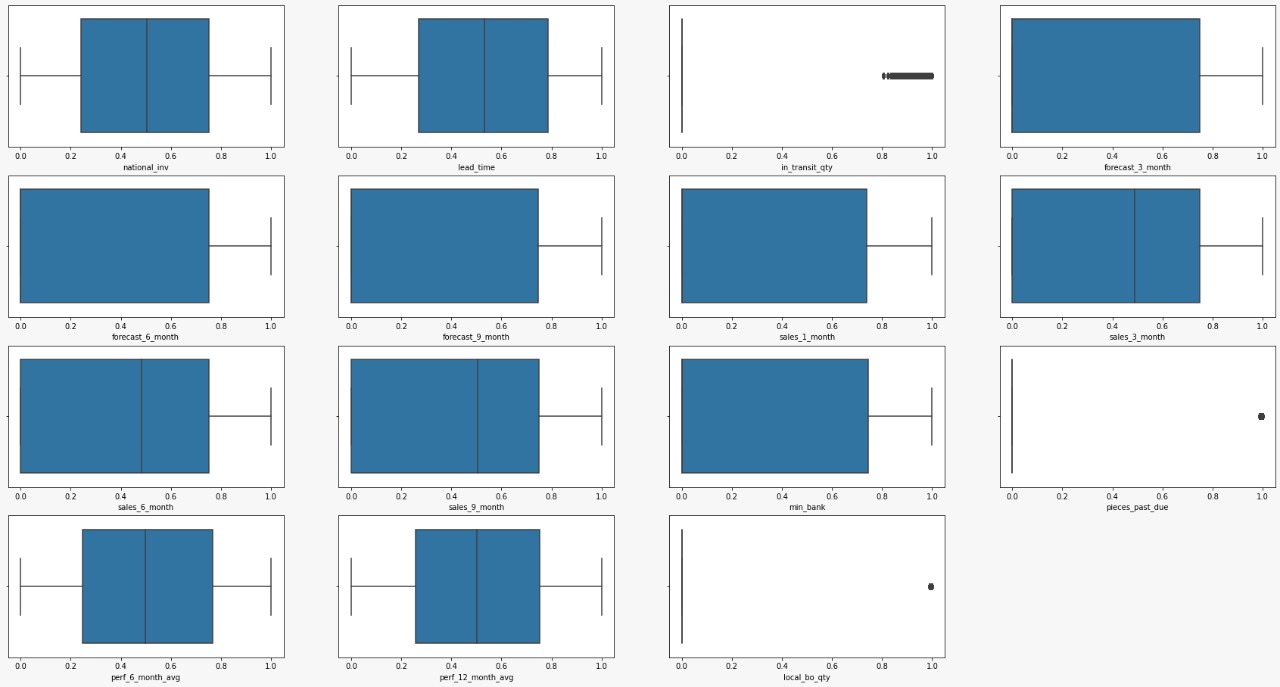
**Distribution of Numeric Variables Original Data:**

**Outliers of Numeric Variables Original Data** :

**Distribution of Numeric Variables Skewed Data:**



**Outliers of Numeric Variables Skewed Data:**





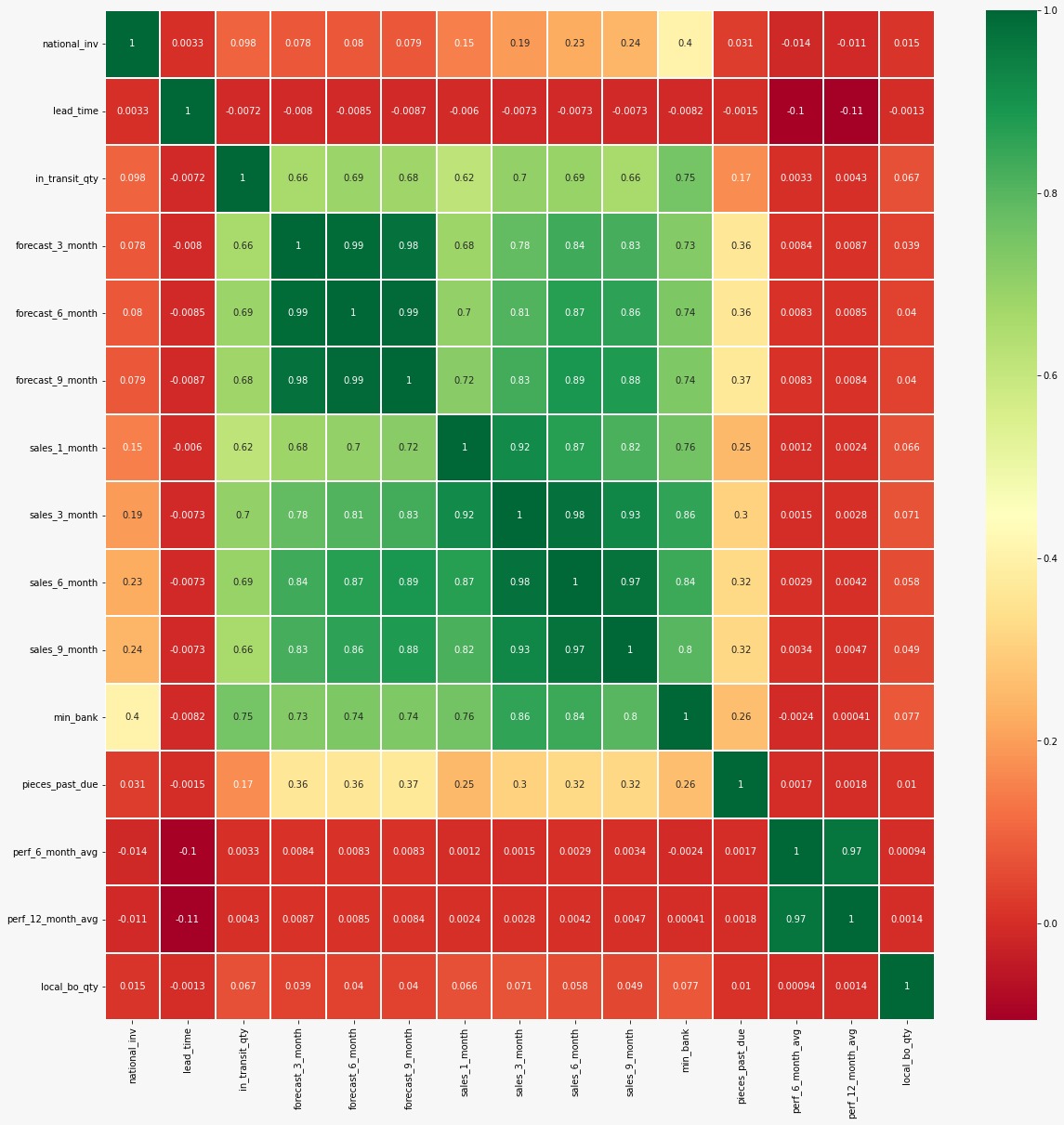




## REDUNDANT FEATURES :

**h1\_diasbp\_max,h1\_diasbp\_min,h1\_diasbp\_noninvasive\_max,h1\_diasbp\_noninvasive\_min, h1\_heartrate\_max,h1\_heartrate\_min,h1\_mbp\_max,h1\_mbp\_min,h1\_mbp\_noninvasive\_m ax,h1\_mbp\_noninvasive\_min, h1\_resprate\_max, h1\_resprate\_min, h1\_spo2\_max, h1\_spo2\_min,h1\_sysbp\_max,h1\_sysbp\_min,h1\_sysbp\_noninvasive\_max,h1\_sysbp\_noninva sive\_min** -- Since all these features are having first 1 hour of diagnosis records , we are dropping it because the other features with first 24 hours of diagnosis records will have more impact than 1 hour of diagnosis

**encounter\_id, patient\_id, hospital\_id, icu\_id -** These are the unique IDs of the respective features, which would not be considered for analysis, hence it can be dropped.



**icu\_admit\_source, icu\_stay\_type, icu\_type, pre\_icu\_los\_days** – Since these features have no effects on patients’ survivability we are dropping them.

**d1\_diabp\_noninvasive\_max , d1\_diabp\_noninvasive\_min, d1\_sysbp\_noninvasive\_max, d1\_sysbp\_noninvasive\_max –** These shows similar range of blood pressure which has been observed in 'd1\_diasbp\_max', 'd1\_diasbp\_min','d1\_sysbp\_max', 'd1\_sysbp\_min'.

**gcs\_eyes\_apache, gcs\_motor\_apache, gcs\_unable\_apache, gcs\_verbal\_apache –** This shows the person’s level of consciousness which is already used to calculate the apache 3j diagnosis scoring

**map\_apache** is having multi collinearity effect with **'d1\_sysbp\_max', 'd1\_sysbp\_min'** and

**'d1\_diasbp\_min' -** So instead of dropping 3 columns we are dropping 'map\_apache'

**Ethnicity, Region** of the patient are not required to predict the patient’s survivability, hence these variables can be removed



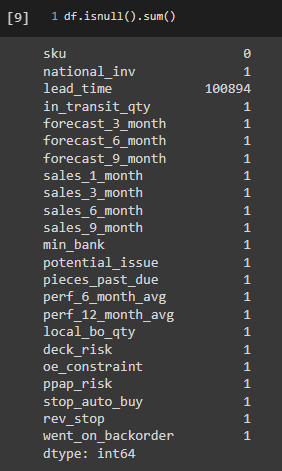


**Height and weight-** since there is a column for BMI which is calculated using height and weight these columns will have severe multicollinearity, so we drop them

# DATA PREPROCESSING:

## NULL VALUE TREATMENT:

Null value treatment is essential to building most of the commonly used machine learning classification models such as logistic regression, decision tree, KNN, and others. To infer that we have used isnull() function the null values from the dataset.

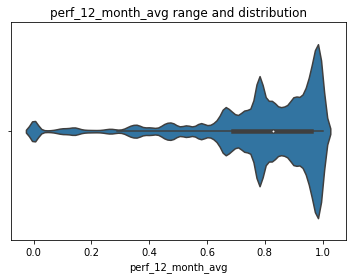
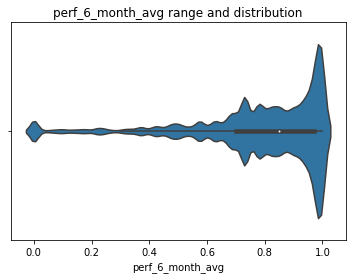
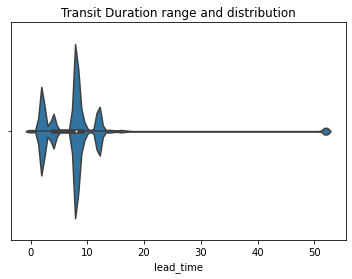






From the above figure, it is evident that the maximum of missing value is **100894** which is observed only in lead\_time column. This means that **median imputation** is applicable as an imputation technique to impute the missing values safely without drastically impacting the distribution of the variables.

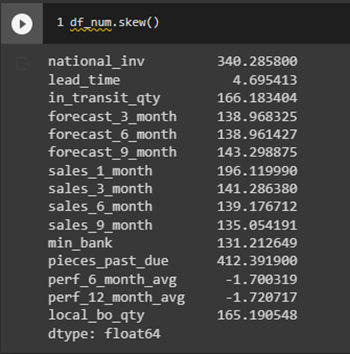
Missing values in columns **source\_performance\_6\_months** and source**\_performance\_12\_months** were represented as -99. We had replaced -99 with NaN for the ease of processing. **source\_performance\_6\_months** had 129478 and **source\_performance\_12\_months** had 122050 missing values.



It’s clearly visible from violin plot that data is not distributed normally. So picked up median to fill remaining values.

## PRESENCE OF OUTLIERS AND TREATMENT:

There are about 15 numerical variables on which the presence of outliers is to be determined. A distribution was assumed to be skewed when the skewness is outside the range of **±0.5** as it is quite impossible to have a real dataset with skewness of each variable or at least one of the variables with a perfect zero skewness. It is also understood that as the value of skewness increases, the farther the outlier is. Out of the 15 numerical variables, it was found that all numerical variables were outside the range of acceptable skewness. The skewness values for those 15 numerical variables are shown below.



Once there are outliers outside the acceptable range, it has to be treated. Dropping the rows with outliers or capping outliers are not recommended as the dataset consists of inventory data with unique product detail which can mislead a model when it comes to prediction. So, transformations are the only means to treat outliers. Out of all the transformation techniques, most of them don’t deal with negative values and zero values. This dataset doesn’t contain negative values however, it



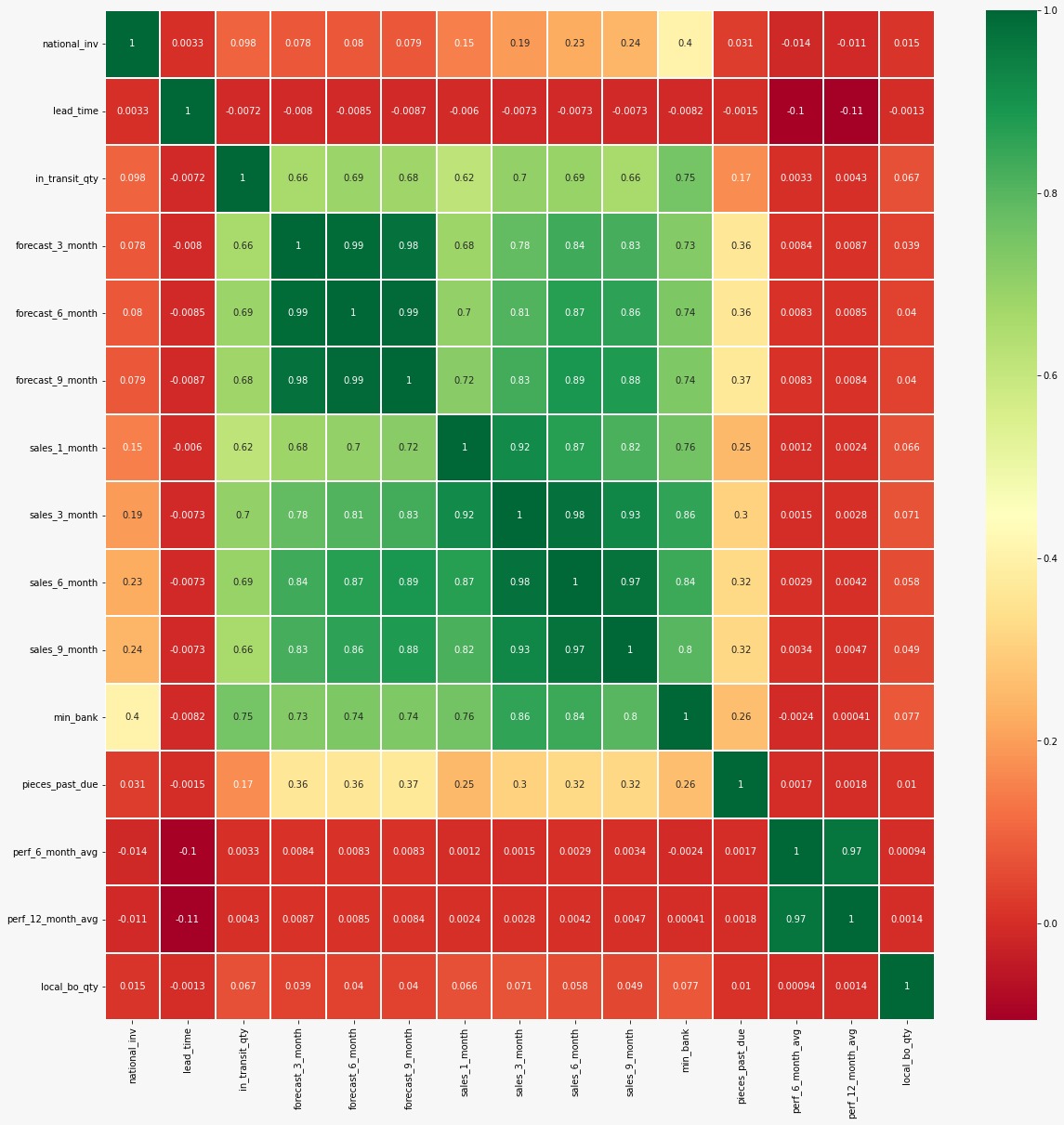


contains values that are zero and very close to zero which will cause problems when applying the

transformation. It was found that the Quantile transformation technique handles both negative and zero values appropriately thereby reducing skewness considerably as shown below

## CHECKING FOR MULTI-COLLINEARITY AND TREATMENT:

When a pair of independent variables exhibit high correlation (that is when a pair of independent variables can explain one another with strong linear relation either positively or negatively) with each other it is termed a collinear effect. When more than one pair of independent variables exhibit a high correlation with each other it is termed a multi-collinear effect.



There is a threshold that is to be set to the correlation value to categorize it between the collinear effect and non-collinear effect. For this project, the threshold is set to be **± 0.5**. The dataset taken into consideration for this project has **more than 10 pairs** of independent variables which exhibit a multi-collinearity effect. This effect is visualized using a heatmap from the python seaborn library. Based on the heatmap we have observed multi-collinearity effect which will be treated post building the base model based on the domain knowledge.

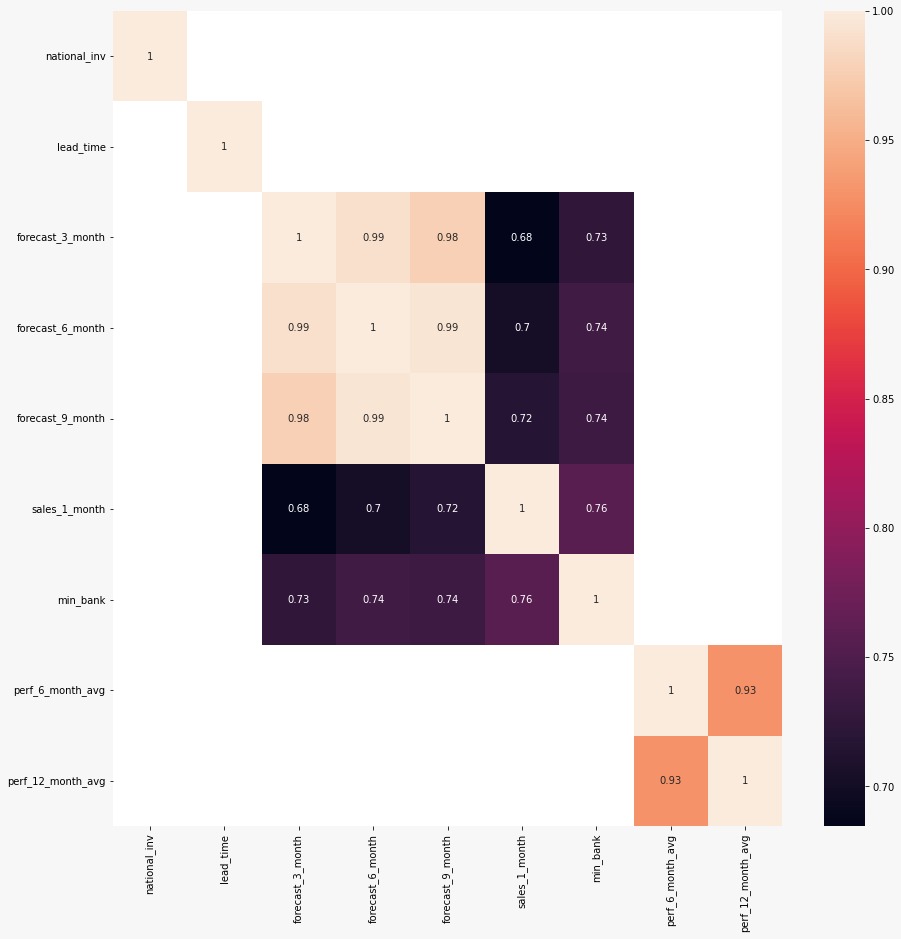




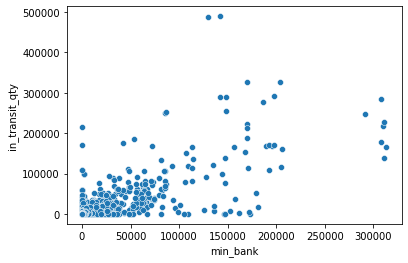


Figure 7: Heatmap of the correlation matrix(treated)

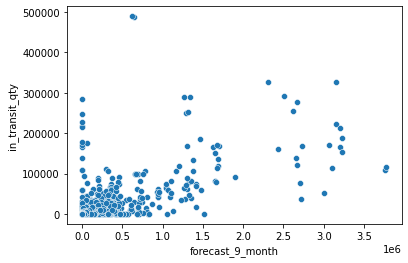
# EXPLORATORY DATA ANALYSIS:

## BIVARIANT AND MULTI-VARIANT ANALYSIS:

From the below bi-variate analysis we could infer that there is slightly linear relationship observed between min\_bank and intransit\_quantity features



The importance of blood pressure is visibly seen in the graph, the patients with low blood pressure both systolic and diastolic were deceased, low blood pressure may lead to various serious health effects, in worst cases If low blood pressure causes a lack of blood flow to the organs of the body, then those organs will start to fail. This may result in stroke, heart attack, kidney failure, and bowel ischemia



When the heart beats too fast, it may not pump enough blood to the rest of the body. As a result, the organs and tissues may not get enough oxygen. This results, that the patients with high rates of





being deceased compared to patients with normal heart rates. The normal heart rate for a healthy person is 60 to 100 beats per minute.

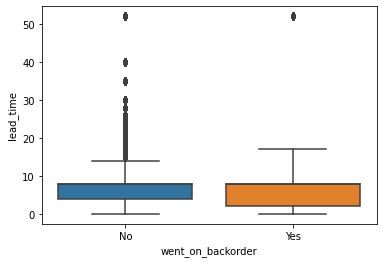
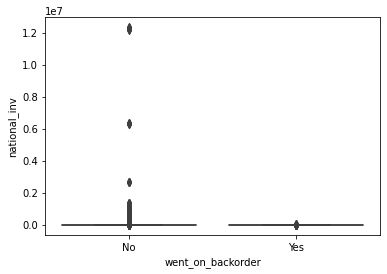


Figure 10: hospital death vs heart rate

It is observed that the patients with a respiratory rate of more than 24 BPM are said to be an abnormal or serious issue, from the below graph it clearly shows that the persons with high respiratory rates tend to have more severe conditions and are mostly deceased, compared to the patients with normal respiratory rate.











# BASELINE MODEL BUILDING:

We have used Logistic Regression classification models as base model and the metrics that we used to validate our model is

* Classification Report
* Accuracy score

For binary classification, the AUC score and kappa score are the best validation metrics. Hence, we try to increase the AUC score rather than the accuracy score.

## LOGISTIC REGRESSION:

We have considered Logistic Regression as a base model. The dependent variable has only two classes (product will go in backorder or not), so we are using binary logistic regression because the dataset contains 15 numerical features out of 22 total features which are favorable for building a logistic regression model.

The following assumptions should be considered while building the Logistic Regression model:

1. Observations must be independent of each other, i.e., they should not come from repeated or paired data
2. Linearity in the log odds for all continuous independent variables
3. There should be no multicollinearity between the independent variables as it reduces the precision of the estimated coefficients
4. Outliers need to be treated before building the model

The model can be built after satisfying all the assumptions above. But the target variable is highly imbalanced. Classification algorithms work well when the data is sufficiently balanced. Hence, we can balance the data using any of the sampling techniques. Here we are using SMOTE to deal with





class imbalance as it generates data based on the available data. Since oversampling generates duplicate data, it automatically violates the assumptions of logistic regression that there should not be any duplicate responses. Since it reduces the data of the majority class, the data obtained does not represent the mean of the original population. After creating the Logistic Regression model after meeting all the above assumptions, the **AUC\_SCORE** obtained for the data is **0.5**, which states that the predictor makes random guesses and there is no need of calculating the accuracy score. Hence, we opt for another model for better prediction.

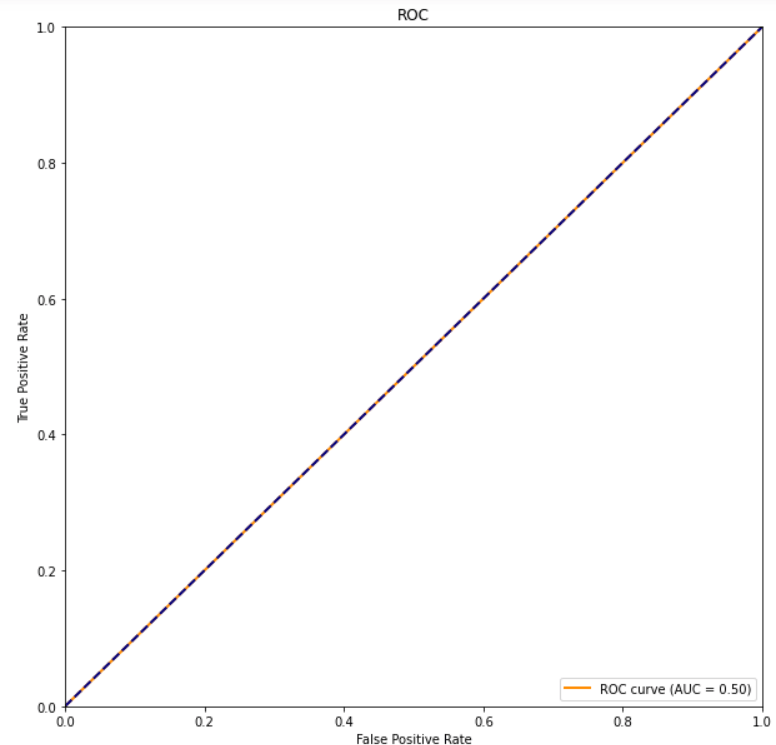


Figure 13: AUC plot for logistic regression





