BACK ORDER

Himalay P. Parmar

Modeling in Operation Management

Agenda

Objectives

Benefits

Dataset EDA

Final Model &
Conclusion

Objective

- Material backorder is a common SUPPLY CHAIN PROBLEM, impacting an inventory system se rvice level and effectiveness. Identifying parts with the highest chances of shortage prior its occurrence can present a high opportunity to improve an overall company's performance..
- Investigated in order to propose a predictive model for this imbalanced class problem, wher e the relative frequency of items that go on backorder is rare when compared to items that do not. Specific metrics such as area under the Receiver Operator Characteristic and precisi on-recall curves, sampling techniques and ensemble learning are employed in this particular task. Results are presented and future scope is discussed.

What are 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.

Dataset

- Source of Dataset: Kaggle
- Format: CSV File
- Total num of Rows: 227351
- Any missing data: Yes & Just dropped it’s

Dataset Features	
Sku(Stock Keeping unit	The product id so can be ignored
National_inv	present inventory level of the product
Lead_time	Transit time of product
In_transit_qty	Total product in transit
Forecast_3_month	Forecast of the sales o f the product for comin g 3 , 6 and 9 months re spectively
Forecast_6_month	
Forecast_9_month	
Sales_1_month	Actual sales of the pro duct in last 1 , 3 ,6 and 9 months respectively
Sales_3_month	
Sales_6_month	
Sales_9_month	
Min_bank	Min. amount of stock re commended

Dataset Features	
Potential_issue	Any problem identified in t he product/part
Pieces_past_due	Amount of parts of the product overdue if any
Perf_6_month_avg	Product performance over past 6 and 12 months
Perf_12_month_avg	
Local_bo_qty	Amount of stock overdue
Deck_risk	Different Flags (Yes or No) set for the product
oe_constraint	
ppap_risk	
stop_auto_buy	
rev_stop	
Went_on_backorder	**Our Target Variable**
Total 22 Independence and one Dependent Featur es	

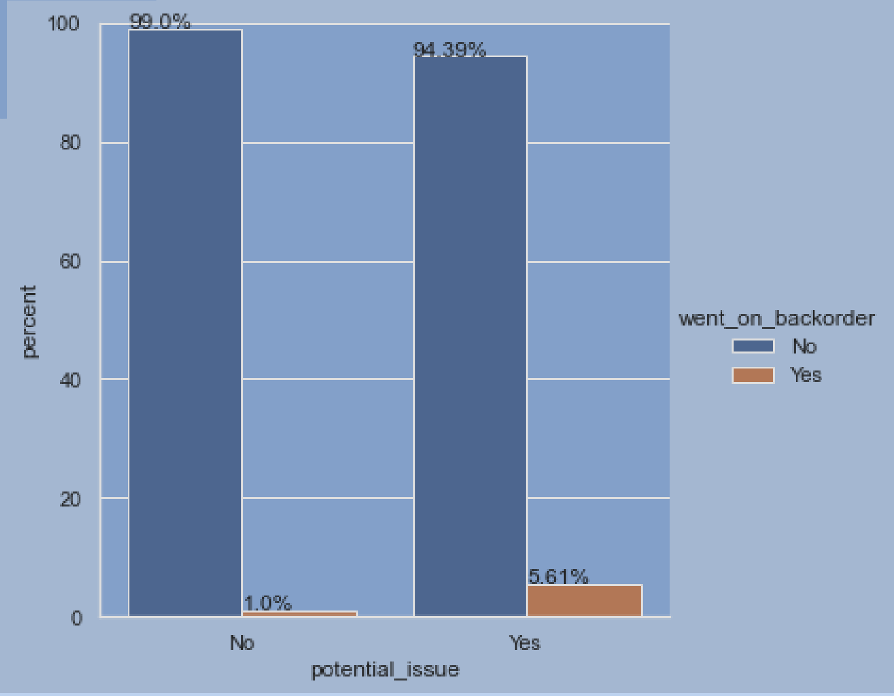
Dataset EDA

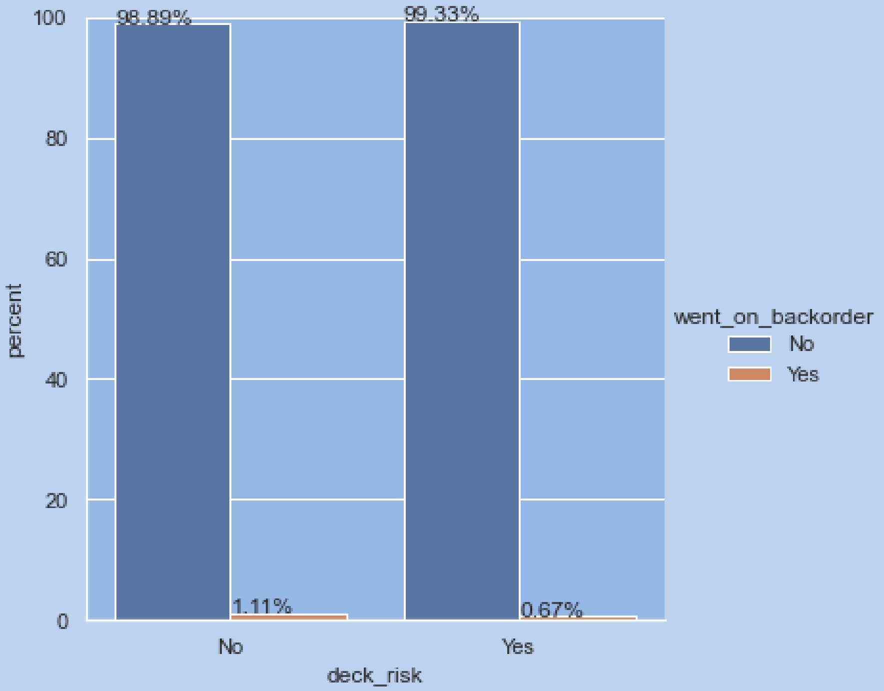
	national_inv	lead_time	in_transit_qty	forecast_3_month	forecast_6_month	forecast_9_month	sales_1_month	sales_3_month	sales_6_month	sales_9_month
count	242075.000000	227351.000000	242075.000000	242075.000000	242075.000000	242075.000000	242075.000000	242075.000000	242075.000000	242075.000000
mean	499.751028	7.923018	36.178213	181.472345	348.807304	508.296301	51.478195	172.139316	340.425414	511.775446
std	29280.390793	7.041410	898.673127	5648.874620	10081.797119	14109.723787	1544.678350	5164.243624	9386.523492	13976.702192
min	-25414.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.000000	4.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	15.000000	8.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	2.000000	4.000000
75%	81.000000	9.000000	0.000000	4.000000	12.000000	20.000000	4.000000	14.000000	30.000000	46.000000
max	12145792.000000	52.000000	265272.000000	1510592.000000	2157024.000000	3162260.000000	349620.000000	1099852.000000	2103389.000000	3195211.000000

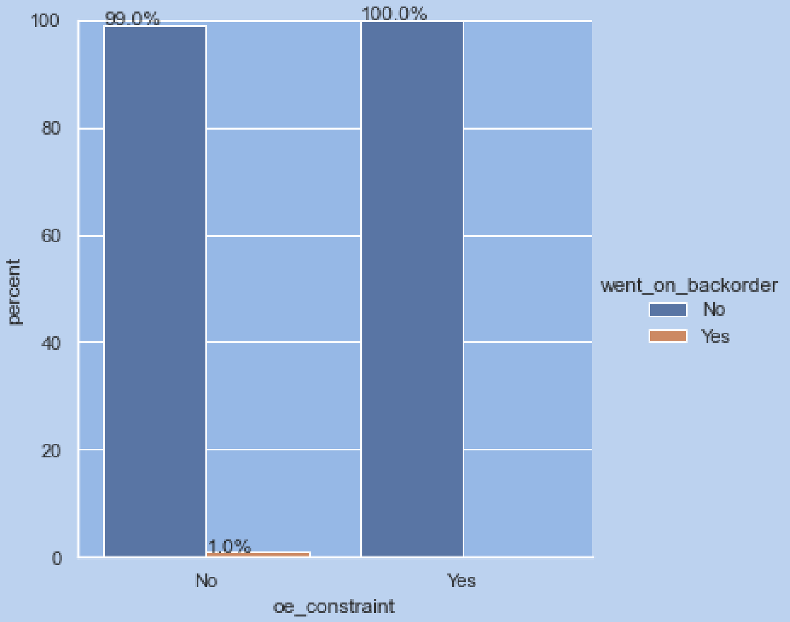
min_bank	pieces_past_due	perf_6_month_avg	perf_12_month_avg	local_bo_qty
242075.000000	242075.000000	242075.000000	242075.000000	242075.000000
52.804693	1.824236	-7.093779	-6.632445	0.843726
1278.591177	178.679263	26.900636	26.160720	45.606626
0.000000	0.000000	-99.000000	-99.000000	0.000000
0.000000	0.000000	0.630000	0.660000	0.000000
0.000000	0.000000	0.820000	0.810000	0.000000
3.000000	0.000000	0.960000	0.950000	0.000000
303713.000000	79964.000000	1.000000	1.000000	6232.000000

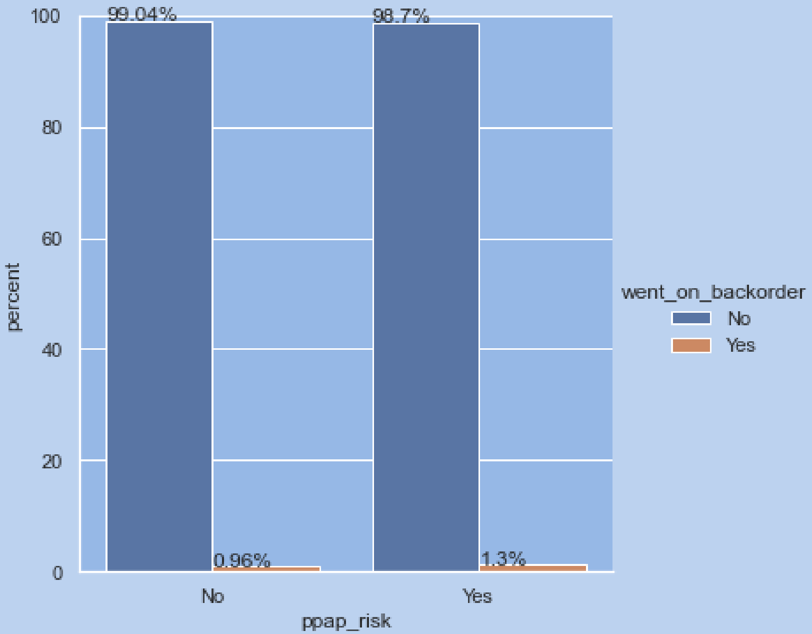
- **Most of the feature mean value is greater than 75 percentile so it is extreamly postive skewed**
- **Most of the features max value is greater than the 75% so they have outliers**
- **The features perf_6_month_avg and perf_12_month_avg has max value 1 and min value -99 so most of the missing value is replace with -99.**

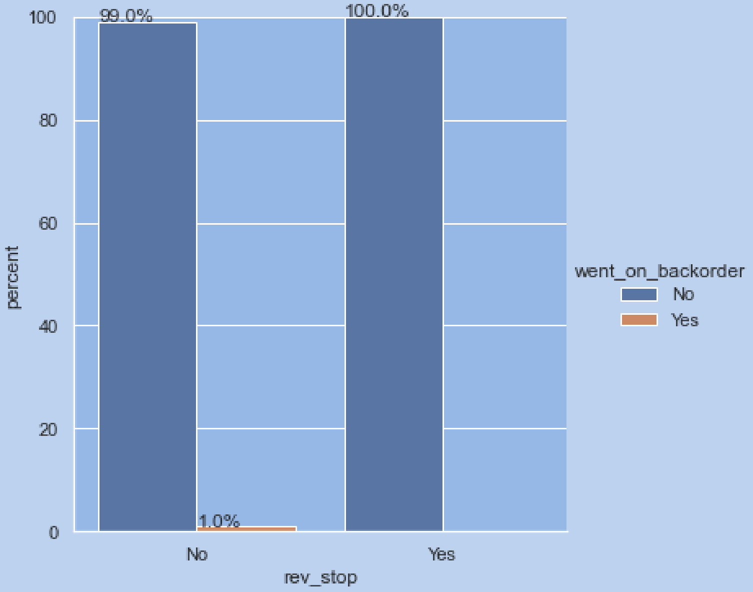
Target Var Against Categorical Features

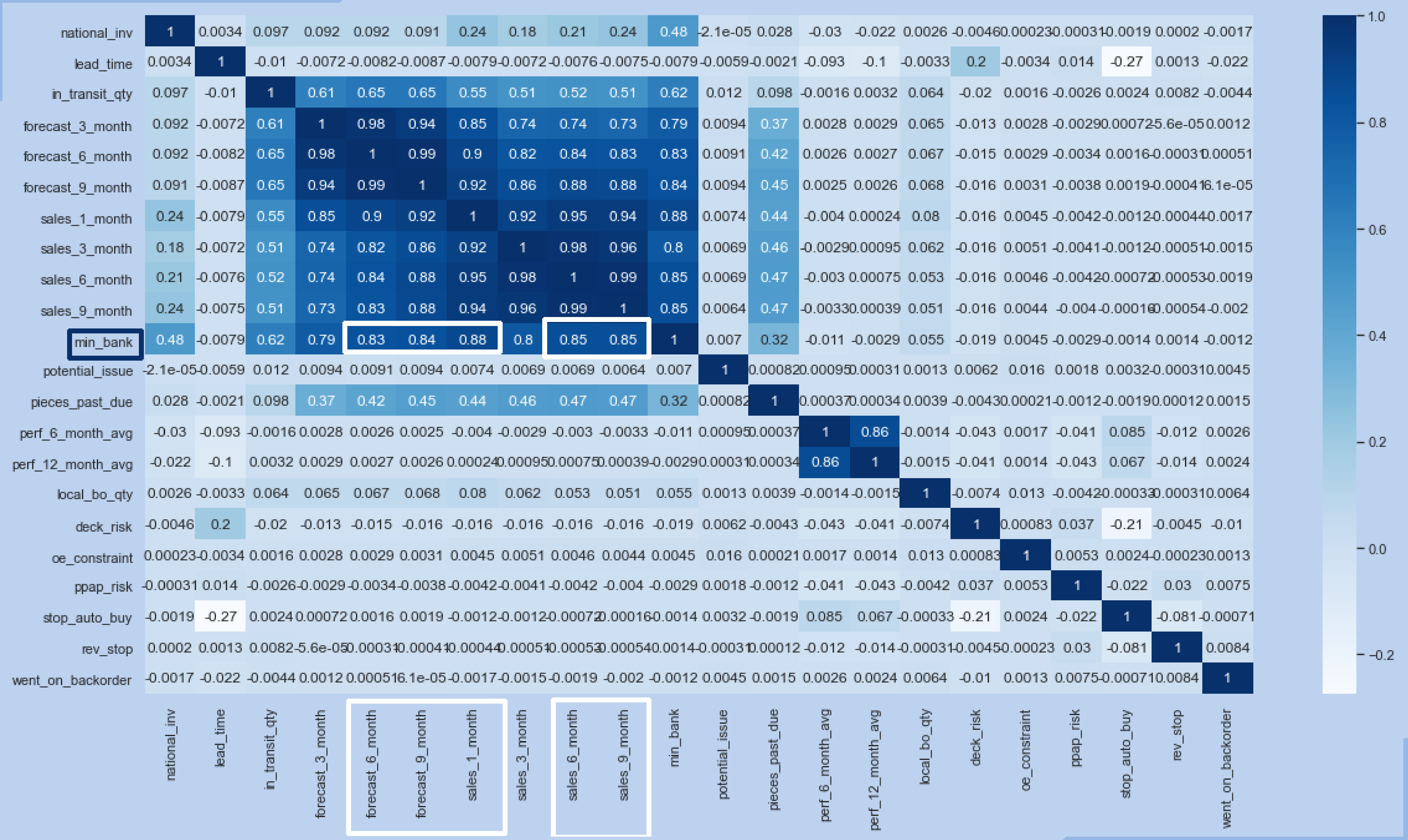






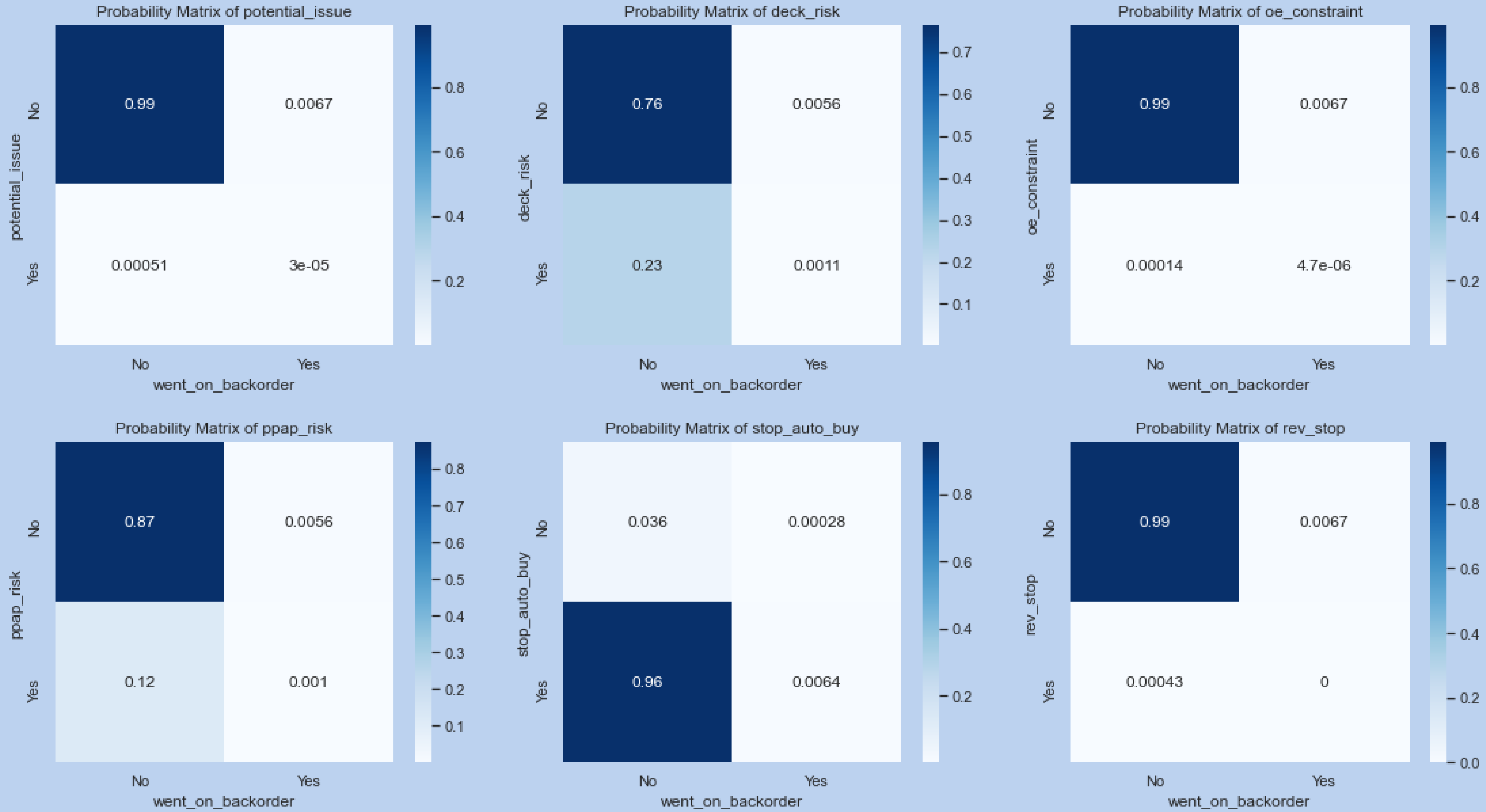





national_inv
lead_time
in_transit_qty
forecast_3_month
forecast_6_month
forecast_9_month
sales_1_month
sales_3_month
sales_6_month
sales_9_month
min_bank
potential_issue
pieces_past_due
perf_6_month_avg
perf_12_month_avg
local_bo_qty
deck_risk
oe_constraint
ppap_risk
stop_auto_buy
rev_stop
went_on_backorder
1.0
0.8
0.6
0.4
0.2
0.0
-0.2

- All the significant correlations observed are positive.
- forecast_3_month , forecast_6_month and forecast_9_month are very strogly correlated with each other to a degree of 0.99.
- sales_1_month, sales_3_month, sales_6_month and sales_9_month are strongly correlated with each other with a degree varying from 0.82 to 0.98.
- forecast and sale columns are correlated with each other with a minimum degree of 0.62 var ying upto 0.88. It is obvious that when the sales for a certain products is high in the past sale s the forecast for the same in the coming months will be higher and viceversa.
- perf_6_month_avg and perf_12_month_avg are very highly correlated with each other to a d egree of 0.97.
- min_bank (minimum amount of stock recommended) is highly correlated with sales and fore cast columns as stock in inventory is directly proportional to sales.
- in_transit_qty is highly correlated with sales, forecast and min_bank columns. This is obvious because high sales of a product => more of that product in transport for inventory replenishin g high sales of a product => high forecast.
- pieces_past_due is meekly correlated with sales and forecast columns.
- national_inv is meekly correlated with min_bank and weekly correlated with sale columns.
- As many features are correlated the linear models like logistic regression, Linear SVM and ot her linear models may not perform well as the coeffiecients of separating plane change.
- By checking VIF(Variance Inflation Factor) value between the correlated features the redund ant features can be removed if needed or using PCA we can reduce dimensions if feature int repretability of model is not important

Probability Matrix for categorical features



From the above set of probaility matrices for all the categorical features we see that most of these categorical features have a very high probability of having a negetive flag when the product did not go into backorder. Therefore, I can say that when a product does not go into backorder, most of the general risk flag are negative.

Test: Realtionship with deck_risk, ppap_risk and stop_auto_buy with the outcome variable went_backorder

Inorder to further understand the relationship between the categorical variable with the outcome variable, we can start using the crosstabulation and chi-square test.

- Ho : Feature are independent, no association between the variables exists
- H1 : Feature are not independent; there is an association between the variable exists.

Deck_risk

	No	Yes
No	150598	1686
Yes	47395	321

Chi-Square Critical value	3.84145
chi_deck_risk	68.580
p_val_deck_risk	1.21829e-16

PPAP Risk

	No	Yes
No	173809	1688
Yes	24184	319

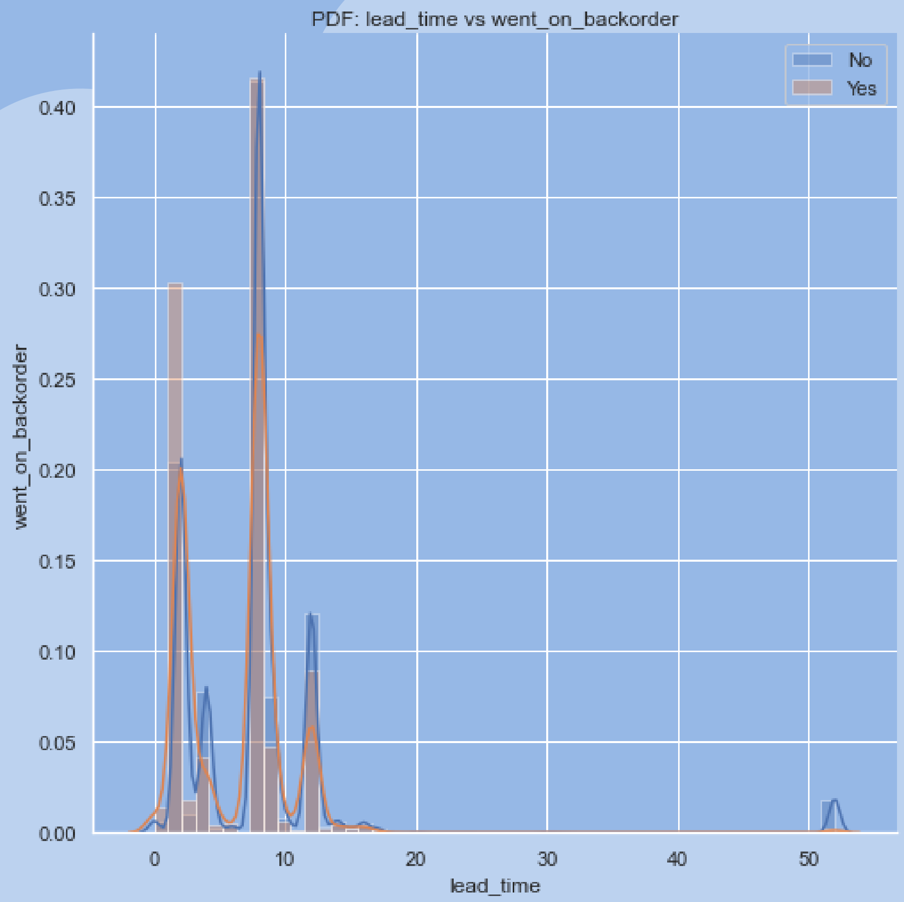
Chi-Square Critical value	3.8414588
chi_deck_risk	24.68455
p_val_deck_risk	6.752305335491265e-07

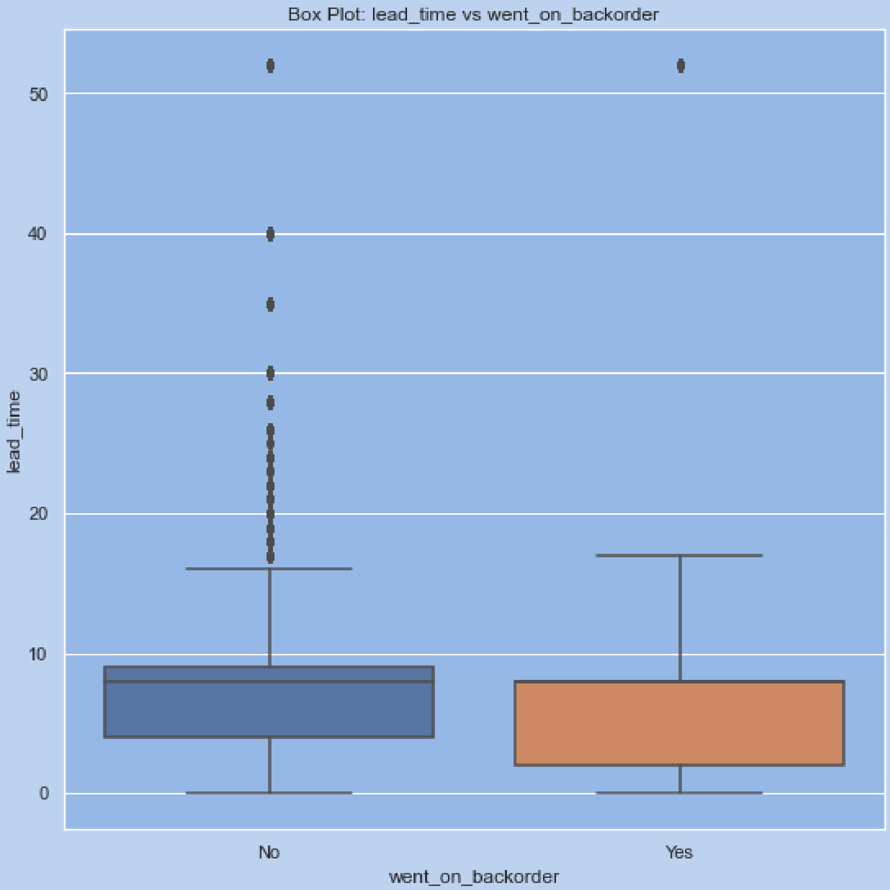
Auto Buy

	No	Yes
No	6959	81
Yes	191034	1926

Chi-Square Critical value	3.84145
chi_deck_risk	1.4389400
p_val_deck_risk	0.00178555963

CrossTab and chi-square to find the relation between target variable with other categorical variables. All the relations has p-values is less than 0.05 and we also have chi-square calculated value is greater than the chi-square critical value. Based on these two evidence we can reject the null hypothesis and can go with the alternate hypothesis.
Here we can say that went_on_backorder is related to deck_risk,ppap_risk and stop_auto_buy, so we will keep all these features for modeling

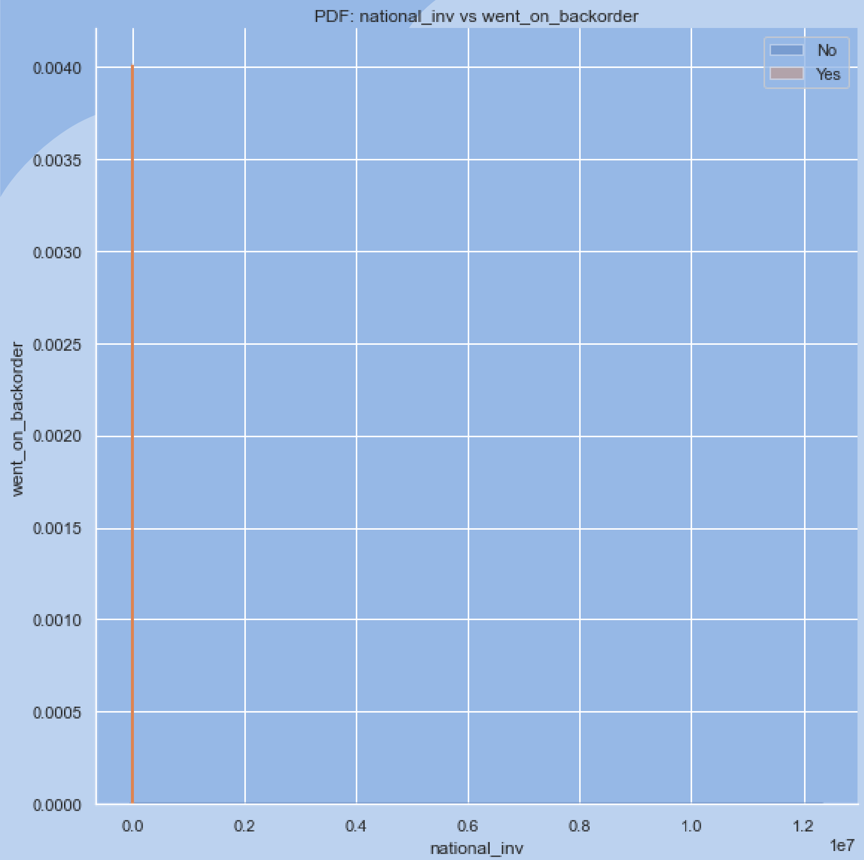


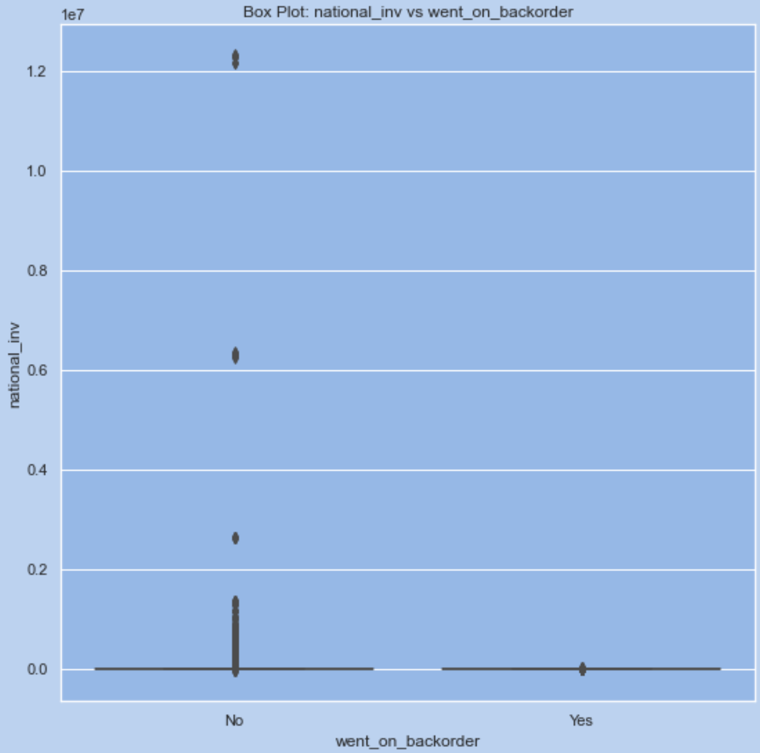


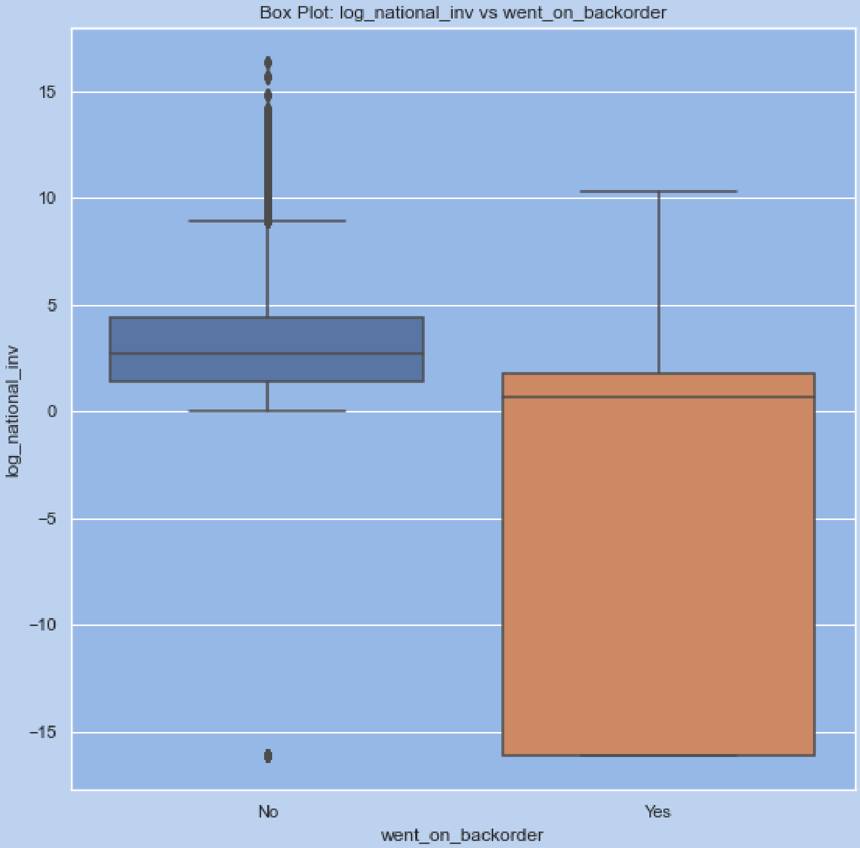
- feature is not normally distributed as per the first pdf plot.
- There is a lot of overlap and we see that the are a lot of datapoints spread towards the right side of the graph which means skewness. The feature 'lead_time' is extremely skewed towards the positive side.

When we look at the box plot, we see that there is no distinct median for the positive class. The median seems to have been m erged into the Q1 value. Therefore, we can say that most of the datapoints in the feature is that one value at Q1 for the positiv e class. Howver, for the negative class we see the median but it is closer to the Q3 value. Here as well, we see a skewness bu t due to outliers.

The minumum for both the classes seem to be similar. We also see many outlier here, especially for the negative class.

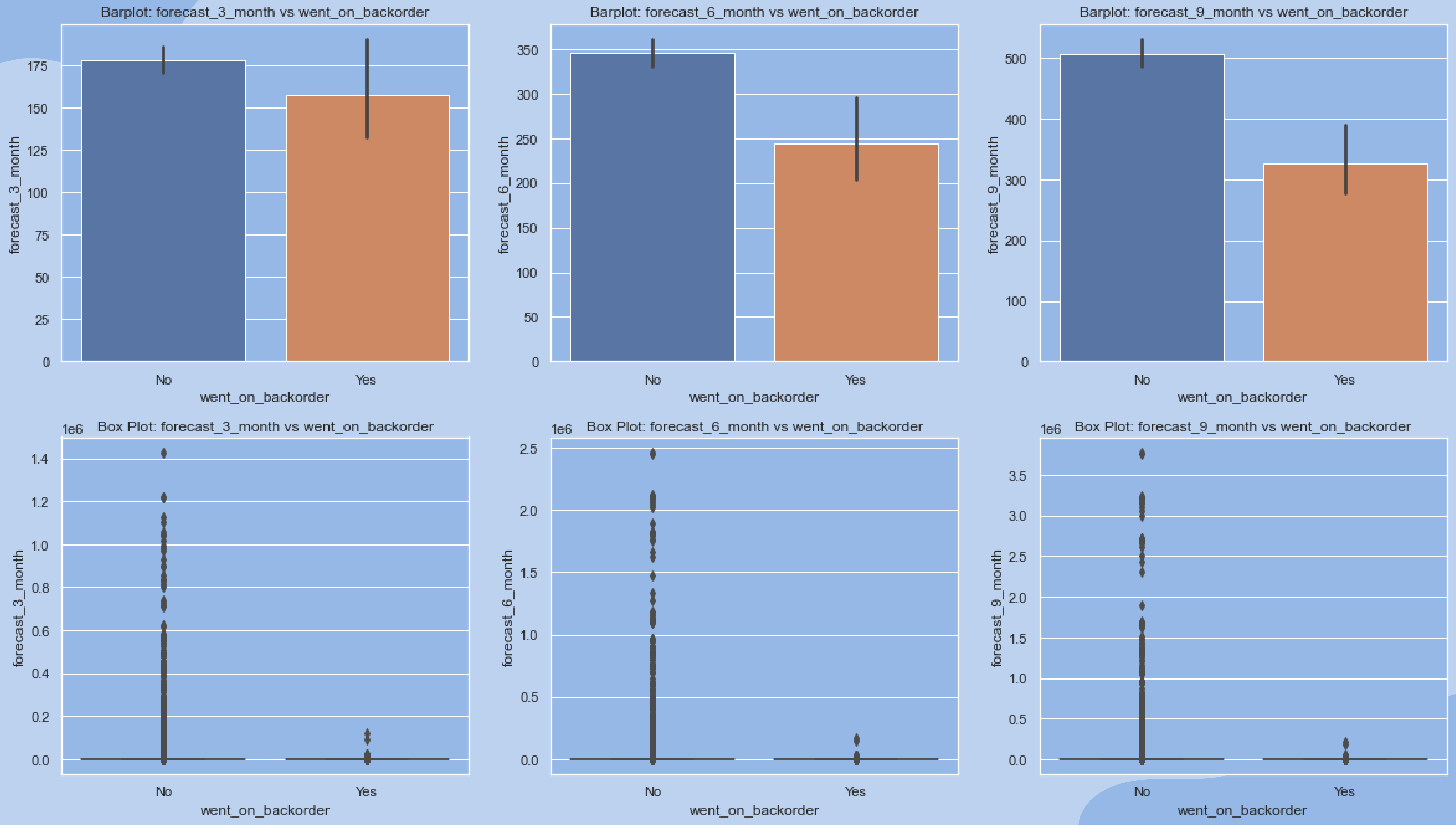






- From the initial plots, it is evident that there are a lot of outliers and the distritution is extremely skewed towards the positiv e side. However, we are unable to properly see that the Inter Quartile Range (IQR) for both the box plots. Therefore, we ha ve modified the national_inv to show it's log values. And since there are zero values in the feature, we have added a small value 'epsilon' which is 1e-7, to avoid infinity.
- From the box plot of the logarithm of national_inv, we see that the IQRs are now visible. The median and the maximums fo r both the classes seems to be similar but the IQRs themselves vary a lot. We still do see outliers for the feature, especially for the negative class label.
- With regard to the positive class, we quickly observe that there is no seperate minimum. The minimum seems to be same as the 25th percentile. And the number of points lying between the 25th percentile and the median is quite large compared to the median and the 75th percentile.

forecast_3_month, forecast_6_month and forecast_9_month vs went_on_backorder



sales_1_month, sales_3_month, sales_6_month and sales_9_month vs went_on_backorder

From the above all barplots, we understand that the mean number of orders that went in to backorder over a span of a few months decreases as the number of orders increase.

Logistic Regression Model Build with RStudio

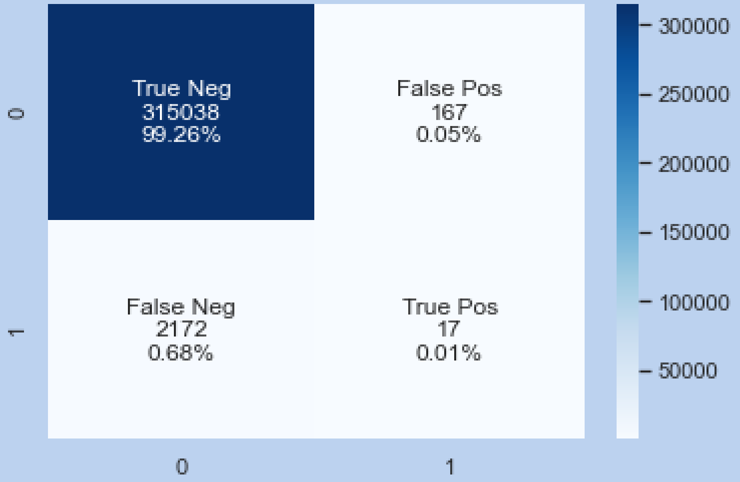
```
Deviance Residuals:
    Min       1Q    Median       3Q      Max
-8.4904  -0.1263  -0.1212  -0.1017   8.4904

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.721e+00  6.579e-02 -56.561  < 2e-16 ***
national_inv     -2.182e-03  1.168e-04 -18.674  < 2e-16 ***
lead_time        -6.913e-02  2.978e-03 -23.217  < 2e-16 ***
in_transit_qty   -6.751e-03  6.762e-04  -9.983  < 2e-16 ***
sales_1_month    -5.293e-05  1.338e-04  -0.396  0.69236
min_bank          7.060e-05  1.691e-05   4.174 3.00e-05 ***
potential_issue   2.190e+00  1.653e-01  13.251  < 2e-16 ***
pieces_past_due   1.236e-05  1.933e-05   0.639  0.52270
perf_6_month_avg  3.010e-03  1.049e-03   2.869  0.00412 **
local_bo_qty      2.030e-04  3.219e-04   0.631  0.52828
deck_risk        -4.267e-01  2.924e-02 -14.593  < 2e-16 ***
oe_constraint     2.171e+00  3.697e-01   5.873 4.28e-09 ***
ppap_risk         2.979e-01  2.984e-02   9.983  < 2e-16 ***
stop_auto_buy    -5.536e-01  6.454e-02  -8.578  < 2e-16 ***
rev_stop         -1.234e+01  1.283e+02  -0.096  0.92339
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105070  on 1269572  degrees of freedom
Residual deviance: 102046  on 1269558  degrees of freedom
AIC: 102076

Number of Fisher Scoring iterations: 15
```

```
    in_transit_qty + min_bank + potential_issue + perf_6_month_av
    deck_risk + oe_constraint + ppap_risk + stop_auto_buy, family
al",
    data = df)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-8.4904  -0.1263  -0.1212  -0.1017   8.4904

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.738e+00  6.573e-02 -56.871  < 2e-16 ***
national_inv     -2.213e-03  9.672e-05 -22.880  < 2e-16 ***
lead_time        -6.927e-02  2.980e-03 -23.246  < 2e-16 ***
in_transit_qty   -6.673e-03  5.724e-04 -11.658  < 2e-16 ***
min_bank          7.177e-05  1.664e-05   4.313 1.61e-05 ***
potential_issue   2.190e+00  1.652e-01  13.251  < 2e-16 ***
perf_6_month_avg  3.013e-03  1.049e-03   2.873  0.00406 **
deck_risk        -4.259e-01  2.923e-02 -14.571  < 2e-16 ***
oe_constraint     2.170e+00  3.698e-01   5.870 4.36e-09 ***
ppap_risk         2.952e-01  2.985e-02   9.890  < 2e-16 ***
stop_auto_buy    -5.352e-01  6.448e-02  -8.300  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105070  on 1269572  degrees of freedom
Residual deviance: 102030  on 1269562  degrees of freedom
AIC: 102052

Number of Fisher Scoring iterations: 14
```

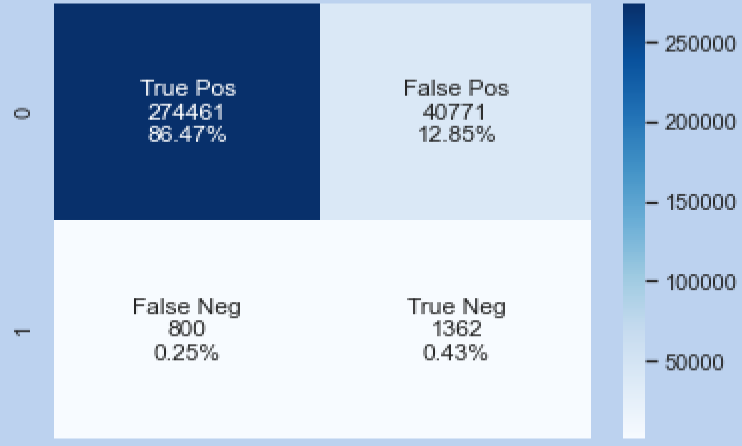
Final Model

WentToBackorder= -3.738e+00 -2.213e-03 (national_inv) -6.927e-02(lead_time) i-6.673e-03 (in_transit_qty) + 7.177e-05 (min_bank) + 2.190e+00 (potential_issue) +3.013e-03(perf_6_month_avg) -4.259e-01 (deck_risk) + 2.170e+00 (oe_constraint) +2.952e-01 (ppap_risk) -5.352e-01(stop_auto_buy)

Logistic Model Comparison
Imbalance Original Model vs SMOTE Oversample Model



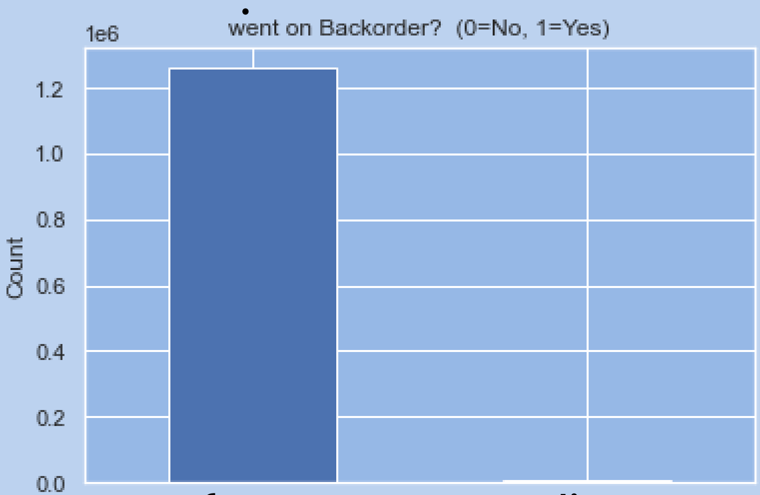
Original model build with i
mbalance data



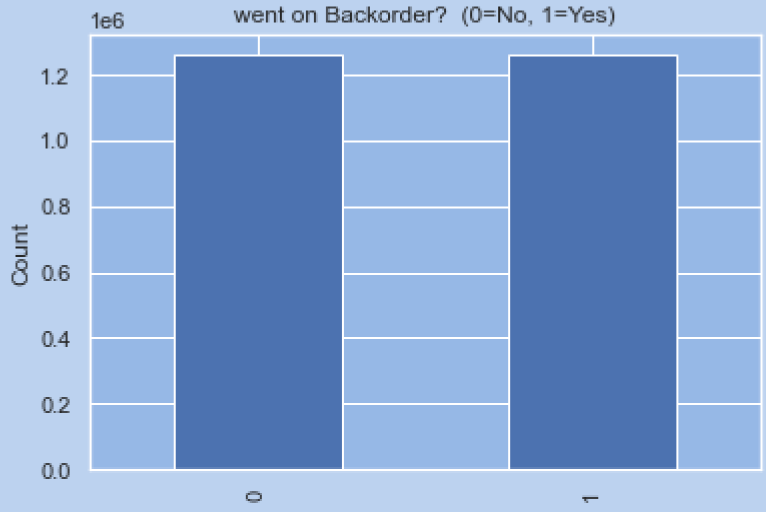
Sensitivity : 0.5406399096538422
Specificity : 0.8385753931544866

Model build with SMOTE B
alance data

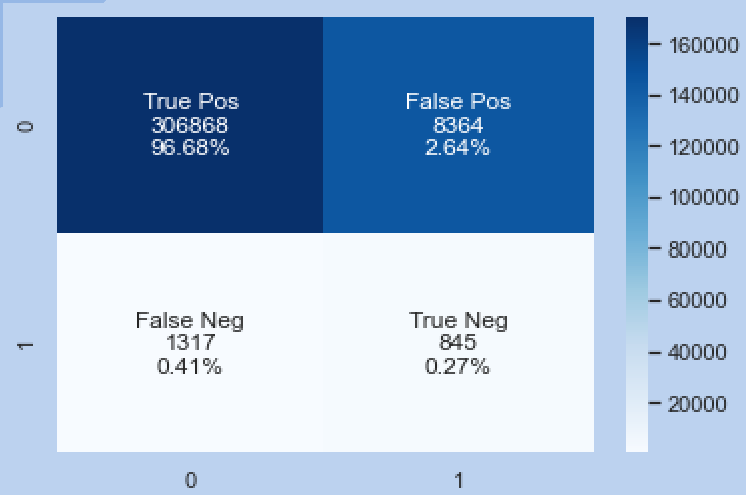
SMOTE - Synthetic Minority Oversampling Technique

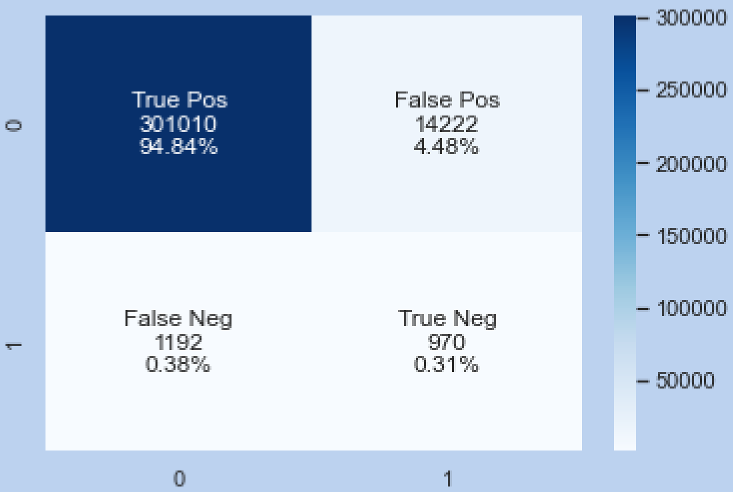


Before Oversampling



After oversampling





Sensitivity : 0.9734671606943458
Specificity : 0.3908418131359852

Model	Accuracy	Precision	Recall	AUC	Sensitivity	Specificity
Logistic Regression	0.8690	0.8077	0.87	0.750 3178	0.870663	0.629972247
Decision Tree	0.9094984	0.543742 327	0.6821 5448	0.6821 5448 69	0.9734671	0.39084
Random Forest	0.9533741 6	0.771026 99	0.5777 2583	0.577 7258	0.948840 21	0.44865864 9

Final Conlusion

- It is important for us to note that the trained dateset is highly imbalanced. Total: 1693050 Positive: 10914 (0.64% of total)
- Therefore the task of classification is difficult. However, the model can be trained to classify minority classes well by :
- Oversampling the minority class (here, the products that went on backorder).
- Undersampling the majority class (here, the products that didn't go on backorder).

- Help with SMOTE: Synthetic Minority Oversampling Technique: To oversample the minority class. This is because undersampling would result in loss of data and important information about the class distribution.
- It is important to understand that accuracy is not a measure that can be used to explain our model performance as the problem that we are dealing with is of imbalanced classification and hence, metrics like ROC score and Recall score are of importance. Here, we have achieved Recall score if 0.87 on the test set. This is a decent classifier as this suggests that 87% of our minority class is correctly classified.

Business Insights

- Important Features that are responsible for affecting the model's performance are:
  - National Inventory
  - Intransit Quantity
  - Lead Time
  - Forecast for 6 Months
  - Sales for 1 and 3 Months

Thank You