Project: Identifying Items With The Highest Chances of Shortage Prior Its Occurrence For Better Inventory Management

Abstract

Material backorder in a supply chain environment is a frequent issue which impacts the service level and effectiveness of the enterprise at inventory level. Out of many products maintained at any inventory or WH's, there are always few products which attract higher demand compared to other products and chances for their stock to run out is very common due to the unforeseen high level demand arising out of any situation like festivals, recent marketing update, discounts etc. The problem here is to identify those products and maintain sufficient inventory stock so that we don't face any backorder issue.

1. Business problem

1.1 Introduction:

At business level frequent backorders indicate mismanagement of an inventory system wherein they have to maintain sufficient space and stock for each product which not only affects the optimum utilization of the facility but also impacts the revenue and brings dissatisfaction among consumers/businesses over regular shortages and non-supply. Frequent backorders or low supply compared to demand could also lead the consumers/businesses to opt in for other available alternatives leading to lower demands in future. Hence, Identifying parts with the highest chances of shortage prior to its occurrence can present a high opportunity to improve an overall company's performance.

1.2 Business/Real-world impact of solving this problem

.

Solving the material backorder issues could help in

i.) healthy utilization of the facilities undertaken
ii.) maintaining optimum supply level compared to demand
iii.) lower chances consumer/business dissatisfaction over supply problems
iv.) Improved business performance

2. Dataset

2.1 Source of the dataset:

For this project we are using a real world imbalanced dataset available on Kaggle's competition and can be found on link provided below.
backorder_prediction/dataset.rar at master · rodrigosantis1/backorder_prediction · GitHub

Project: Identifying Items With The Highest Chances of Shortage Prior Its Occurrence For Better Inventory Management

2.2 Explanation of each feature and datapoint available

The dataset contains the historical data for the 8 weeks prior to the week we are trying to predict, taken as weekly snapshot at the start of the week.Attributes of the dataset are provided below:

1. x1: Current inventory level of component;
2. x2: Registered transit time;
3. x3: In transit quantity;
4. x4;5;6: Forecast sales for the next 3, 6 and 9 months;
5. x7;8;9;10: Sales quantity for the prior 1, 3, 6, 9 months;
6. x11: Minimum recommended amount in stock;
7. x12: Parts overdue from source;
8. x13;14: Source performance in last 6 and 12 months;
9. x15: Amount of stock orders overdue;
10. x16-21: General risk flags;
11. y: Product went on backorder.

2.2 Data Size and any challenges:

We have a huge dataset consisting of 23 columns/features with more than 10.5 Lac rows of data per feature. There are extra data columns “national_inv” & “potential_isssue” compared to attributes. Also, basic transformations are required such as binaries feature encoding, quantity related features normalization and missing values imputation.

2.3 Tools (Pandas, SQL, Spark etc) used to process this data

The project will be implemented in Python 3.6.1 programming language, using Scikit-learn and Imbalanced learn machine learning libraries.

2.4 Data Acquisition:

Data is a real world gathered dataset provided on Kaggle competition website as open for all. More data may be acquired based upon request to original authors of the paper as provided in the reference section.

Project: Identifying Items With The Highest Chances of Shortage Prior Its Occurrence For Better Inventory Management

3. Key Metric (KPI)

3.1 Business Metric definition:

Selecting the right evaluation metrics is a key determinant for guiding the construction of a predictive model. The accuracy rate has been the most usually applied empirical measure in classification, but in our case we will be using “AUC” as our classifier evaluation metrics.

A standard approach used to evaluate classification models in imbalanced problems is to use the Receiver Operating Characteristic (ROC). The Area Under the ROC Curve (AUC) corresponds to the probability of correctly identifying which one of the two stimuli is noise and which one is signal plus noise. AUC provides a single measure of a classifier’ capability of evaluating which model is better on average and can be computed by:

$$AUC = \frac{1 + P - F}{2} \qquad -(3)$$

3.2 Why is this metric used:

In the framework of imbalanced datasets, accuracy is no longer a suitable metric, since it doesnot distinguish between the number of correctly classified examples of different classes. Many machine learning models are designed around the assumption of balanced class distribution, and often learn simple rules like always predicting the majority class, causing them to achieve an accuracy of 99 percent, although in practice performing no better than an unskilled majority class classifier.A typical example is an estimator which classifies all examples as negatives, leading to equivocated conclusions.

3.3 Alternative metrics that can be used? Why are they not preferred in this case?

Several specific metrics that can be used within imbalanced problems domain in order is to take into account the class distribution: **“Precision”, defined by (2),** express the accuracy of an estimator when predicting the positive class, while **“Recall” (3)**, also known as true positive rate or sensitivity, indicates its ability of finding all the positive samples.

$$Acc=\frac{Tp + Tn}{Tp + Fn + Fp + Tn} \qquad -(1)$$

$$P= \frac{Tp}{Tp + Fp} \qquad -(2)$$

$$R = \frac{Tp}{Tp + Fn} \qquad -(3)$$

Precision-recall curves represent the conflict existing between both metrics and are commonly used in binary classification to understand the output of a classifier and aid the choice of the decision function threshold.”Precision-Recall” curve graphic allows the visualization of the trade-off between the precision and recall, as it evidences that any classifier cannot increase the number of true positives without also increasing the false positives.

Other metrics can also be used in classifiers evaluation, although AUC has been one of the most applied in literature for assessment and benchmark reference.

3.4 Pros and cons of the metric used:

The biggest benefit of using the metrics is that it solves the major challenge of correctly evaluating our model for the problems where our dataset is hugely skewed like identifying a fraud to happen, or detect a smoke etc which are very real world cases where occurrence of 1 class is very less compared to the major class but the impact is huge.

Since this is a special case metric used in certain scenarios no cons can be provided as such.

3.5 Where does this metric fail? Where should it not be used?

This metric will not be very suitable for balanced class problems.

4. References

1) (PDF) Predicting Material Backorders in Inventory Management using Machine Learning (researchgate.net)
2) https://www.bigcommerce.com/blog/inventory-management/#coqmmon-inventory-management-questions
3) https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba
4) https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/
5) https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/