Predicting Product Backorder Using Machine Learning Algorithms

Visanu Chumongkhon
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, USA.
vzc15@psu.edu

Rutvik Sharedalal
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, USA.
rns26@psu.edu

***Abstract*—** One of the major problem faced by Supply Chain of any industry is delivery performance. Quantifying the uncertainties to improve delivery efficiency is a major challenge. Multiple criteria should be considered for optimizing service level (Percentage of customers whose demand is met). Various key performance indices like yield from manufacturing, quality of supplies from suppliers, lead time, etc., are to be quantified for inventory planning. Along with aforementioned factors, the cost is an inevitable constraint in formulating the objective function for capacity planning. The aim of all parameters mentioned above is to make sure that the organization doesn't incur backorder during the lead time for a given service level. This paper elaborates upon exploring inventory management data to find patterns that could explain instances of backorder and applying machine learning to develop models that can predict backorders given various features identified from the dataset. Sampling techniques, various classification algorithms and validation metric such as area under ROC curve are employed in this particular task. Accurate predictions obtained by these models can then be used to plan accordingly to avoid instances of backorder and improve or maintain service level.

Keywords—Supply chain, Imbalanced class problem, Sampling methods, Stratified cross-validation

I. INTRODUCTION

If a company acquires a lot of product backorders(shortage), it might lose its customers over a long run. It is always important from a business point of view to decide the optimal number of order quantity as well as the time of ordering it. Most of the approaches proposed so far in industries try to formulate this problem as a multi-objective optimization: minimizing ordering and storage cost and simultaneously keeping service level as high as possible. In this context, machine learning proposes a different approach for managing the inventory more efficiently. Machine learning can be used to build predictive models which can identify products that are at risk of stockout before it actually occurs. This eventually helps an enterprise gain ample time to react to this adversary situation. This paper proposes the application of various classification techniques to predict the product backorder based on the various predictor attributes. The dataset contains the information of 8 weeks prior to the week for which the prediction is performed. Attributes are defined below:

TABLE I – DATASET SUMMARY

No. of Attributes	No. of Positive class samples	No. of Negative class samples	Total Samples
22	13,981	1,915,954	1,929,936

X_1 = Current inventory level of a component
X_2 = Registered transit time for the product
X_3 = In transit quantity
$X_{4,5,6}$ = Forecast for next 3,6 and 9 months
$X_{7,8,9,10}$ = Sales quantity for the prior 1,3,6,9 months
X_{11} = Minimum recommended amount to stock
X_{12} = Parts overdue from source
$X_{13,14}$ = Source performance in last 6 and 8 months
X_{15} = Amount of stock orders overdue
$X_{16\text{-}21}$ = General risk flags
y = Product went on backorder

One of the complications related to the data is that the positive class(backorder) observations are very inferior compared to negative class (no backorder) observations. Hence, the dataset is dealing with class imbalance problem. This paper elaborates upon using sampling techniques to make data balanced. The performance measure and the validation technique that have been used in this paper are area under the ROC curve and stratified k-fold cross-validation. Grid search is used to select the best parameters among the ones specified for some of the classification techniques.

The paper is divided into 5 sections: Following the introduction, the section 2 provides a brief literature review on the approach used to perform the classification on a similar dataset, section-3 gives brief overview of sampling techniques as well as classification algorithms used in this paper to perform analysis, section-4 exhibits the results and discussions obtained from the analysis. Finally, section-5 proposes possible future scope pertaining to the data.

II. Literature Review

De Santis et al. applied different classification models on the similar data and tried to compare their accuracies [1]. The sampling techniques used in their paper are Random Under Sampling and Synthetic Minority Over-sampling Technique. The classifiers used by them are Logistic Regression, Classification Tree and ensemble learners such as Random Forest, Gradient Tree Boosting and Blagging. They used 5-fold stratified cross-validation after dividing the data into training and testing sets with 0.85:0.15 ratio and calculated AUC score of the ROC curve to compare the accuracies of these models. Their conclusion reflected gradient boosting as the most efficient algorithm to predict backorders. Besides this, it is mentioned in their future work that Support Vector Machines and Neural Network algorithms can be tried to identify potential performance improvement. Dai et al. incorporated Random Under Sampling in their research to identify rare genetic disease called Hereditary Angioedema(HAE) [2]. RUS was combined with Bagging Random Forest and Bagging Lasso and the traditional logistic regression model was set as a benchmark for the purpose of model comparison. AUC score and AUPR score were used as the performance evaluation metrics and 5-fold cross-validation was used as a model validation technique. Bagging RF showed 21% performance improvement when AUPR score was compared to that of the logistic regression model. However, Bagging Lasso proved to be inferior compared to the logistic regression model.

III. Methodology

A. Sampling Techniques

1) Random Under Sampling

According to general framework given in [2], Random Under-sampling creates a subset of the original training set by randomly sampling observations from each class with or without replacement. The rate at which the sampling is performed is given by N_c/N_i where N_c stands for desired sample size in the subset and N_i stands for original sample size of class i. The main drawback of this technique is that it might remove potentially useful information by removing some of the majority class observations and the power of algorithm might decrease as fewer data points are used to train the model.

2) Synthetic Minority Over-Sampling Technique

SMOTE is a sampling technique which artificially generates data points of minority class to make the dataset balanced [5]. In this technique, a minority class sample is considered and synthetic points are generated on a line segment joining this particular sample and its k nearest neighbors in a feature space. The algorithm working behind this logic is taking a difference between feature vector of the sample considered and that of its nearest neighbor and multiplying this difference with a random number between [0,1] to generate a point along the line segment joining them.

2.Classification Algorithms

1) Gradient Boosting

Boosting is a popular ensemble technique in which many weak classifiers are combined to create a stronger learner. Boosting sequentially build estimators that try to get higher accuracy by focusing or placing more weight on the samples which were incorrectly classified by previous learner [1]. Gradient boosting is a variant of boosting ensemble techniques. Gradient tree boosting uses decision trees as weak learners and try to minimize the loss function using gradient descent algorithm. Each iteration tries to minimize the loss incurred in the previous iteration. According to [1], gradient tree is robust to outliers and hence efficiently deal with the class imbalance problem.

2) Support Vector Machine

If the data points are linearly inseparable, there can exist infinite many hyperplanes to separate the data. Support vector classifier tries to find the optimal hyperplane which is the farthest from the nearest training observations of each class. However, if the data points are not linearly separable, non-linear class boundaries need to be incorporated. Support Vector Machines try to address this problem by enlarging the feature space using quadratic, cubic or higher-order polynomial functions of the predictors with the help of kernels.

3) Neural Networks

Neural networks are powerful nonlinear regression and classification technique inspired by theories about how the brain functions [6]. In Neural networks, outcome is modeled by an intermediary set of hidden units and each hidden unit has some weight parameter associated with it. Each hidden unit is a linear combination of some or all of the predictor variables. To introduce non-linearity, this linear combination is passed through a non-linear function such as the logistic function [6]. Back-propagation algorithm is used to find the optimal value of the parameters. In this algorithm, error is calculated at the output layer in each iteration and the weight parameters are optimized by taking derivative of the error loss function with respect to the parameters.

4) Logistic Regression

Logistic Regression is a popular and simple approach for binary classification [1,3]. In logistic regression, the response variable is a log function of odds of an event happening. This

log function can take any value between -∞ and +∞. The shape of the end function obtained is no longer linear [3]. The difference between logistic and linear regression is in the choice of model parameters and the error function. Contrary to least square, the approach used in the logistic regression is minimizing log-likelihood function. Logistic shows improved sensitivity to class imbalance problem because of this [1].

3.Model Validation

1) Stratified k-fold cross-validation

Stratified cross-validation is the most appropriate technique to use when the data is unbalanced. In stratified k-fold cross-validation, each fold approximately has the same mean response value i.e. each fold has the same proportion of observations from each class. The main advantage of this technique of validation is that the model is fitted on a training set that is not biased towards a specific class.

2) AUC score of the ROC curve

Area under the ROC curve is a standard approach used to evaluate the performance of a classifier in the unbalanced class environment [1]. In ROC space, False positive rate is plotted on X-axis and True positive rate is plotted on Y-axis to form ROC curve. The FPR provides the fraction of negative examples that are incorrectly classified as positive and TPR provides the fraction of positive examples that are correctly labeled [2]. As mentioned in [1], any classifier cannot increase the number of true positives without also increasing the number of false positives. Hence, accuracy of any classifier is measured by the area under the curve. The greater the area, the more the TPR can be increased without significantly increasing the FPR.

TABLE II – CONFUSION MATRIX

	Real = 1	Real = 0
Predicted = 1	True Positive(TP)	False Positive(FP)
Predicted = 0	False Negative(FN)	True Negative(TN)

TABLE III – MODEL PERFORMANCE MEASURES

Measure	Definition
AUC	Area under ROC curve
True Positive Rate(TPR)	TP/(TP+FN)
False Positive Rate(FPR)	FP/(FP+TN)

IV. RESULTS & DISCUSSION

Table-I represents the AUC score achieved by proposed algorithms obtained by applying 5-fold stratified cross-validation. The logistic regression model performed the lowest score. The AUC score of the SMOTE technique showed better results compared to the Random Under Sampling when combined with logistic regression. Support Vector Machine performed slightly better than Logistic Regression when combined with Random Under Sampling. Neural Network with Random Under Sampling showed promising results giving accuracy of 0.9224. The best accuracy was shown by Gradient Boosting with grid search even when applied on the unbalanced dataset. This was justified as tree-based algorithms are usually robust to class imbalance problem.

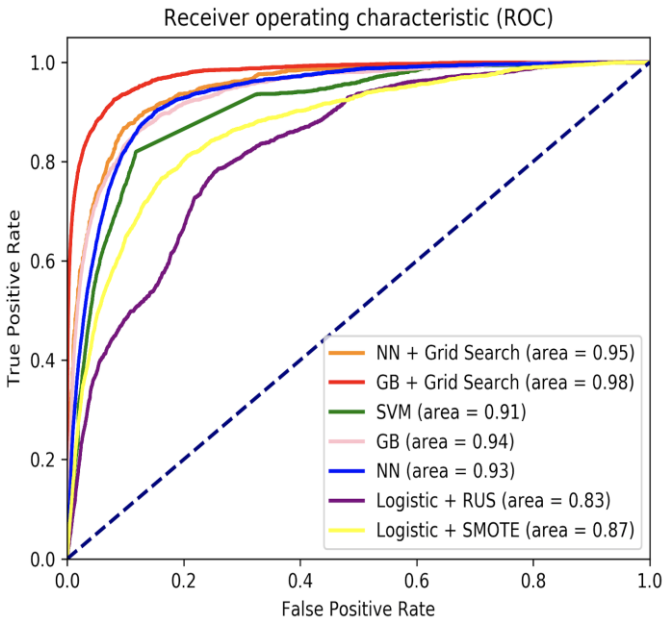

FIGURE I – COMBINED ROC CURVE

TABLE - IV – PARAMETERS ADOPTED FOR CLASSIFIERS

METHOD	PARAMETERS	VALUE
Gradient Boosting	Learning Rate	0.1
	Loss function	Deviance
	No. of estimators	10
	Maximum Depth	9
	Minimum Sample Leaf	5
Neural Network with RUS	Activation	Rectified Linear Unit Function
	Solver	lbfgs (Quasi-Newton)
	Learning Rate	0.001(Constant)
	Hidden Layer Structure	{50,100,200,100,10}
Support Vector Machine with RUS	Penalty Term of error	C=1.00
	Kernel	rbf
Logistic Regression with RUS and SMOTE	Penalty	L2
	Regularization	C=1.00
Neural Network with RUS and Grid Search	Activation	Tanh,
	Solver	'adam'
	Learning Rate	0.001(Constant)
	Hidden Layer Structure	{100,50,25,5}
Gradient Boosting with Grid Search	Learning Rate	0.1
	Loss function	Deviance
	No. of estimators	500
	Maximum Depth	9
	Minimum Sample Leaf	9

TABLE V – COMPARISON OF AUC OF VARIOUS MODELS

MODEL	AUC
Gradient Boosting	0.9324
Gradient Boosting with Grid Search	**0.9759**
Logistic Regression with RUS	0.8378
Logistic Regression with SMOTE	0.8805
SVM with RUS	0.8991
Neural Network with RUS	0.9224
Neural Network + RUS + Grid Search	0.9392

It can be seen from the table – V that when grid search was applied to Neural Network and Gradient Boosting, they performed better than their normal versions. Table – IV shows the parameters used for the proposed classification techniques both for with and without grid search. Figure - I shows the combined ROC curve for all of the proposed techniques with their AUC scores. Python 3.6.4 was used for programming purpose. The computer specifications used for grid search are: 2.2 GHz Intel Xeon Processor, 24 CPU/server, 128 GB RAM, 40 Gbps Ethernet.

V. Future Work

Future work includes incorporating cost-based classification for Support Vector Machine and Neural Network in which minority class observations are given more weight than those from the other class. The accuracy of this cost-based classification can be compared with the accuracy of the same algorithms combined with various sampling techniques.

References

[1] de Santis, Rodrigo Barbosa, Eduardo Pestana de Aguiar, and Leonardo Goliatt. “Predicting material backorders in inventory management using machine learning.” *Computational Intelligence (LA-CCI), 2017 IEEE Latin American Conference on*. IEEE, 2017.

[2] Dai, Dong, and Shaowen Hua. "Random Under-Sampling Ensemble Methods for Highly Imbalanced Rare Disease Classification." *Proceeding of the International Conference on Data Mining (DMIN)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016.

[3] D. W. Hosmer Jr., S. Lemeshow and R. X. Sturdivant, “Applied logistic regression,” John Wiley & Sons, vol. 398, 2013.

[4] http://www.businessscience.io/business/2017/10/16/sales_backorders_prediction.html

[5] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*16 (2002): 321-357.

[6] Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. Vol. 26. New York: Springer, 2013.