Management Science and Financial Engineering
Vol 19, No 1, May 2013, pp.1-10 http://dx.doi.org/10.7737/MSFE.2013.19.1.001
ISSN 2287-2043 | EISSN 2287-2361 |

Predicting the Performance of Forecasting Strategies for Naval Spare Parts Demand: A Machine Learning Approach

Seongmin Moon*
Integrated Logistics Support Technology Team, Defense Acquisition Program Administration

(Received: May 27, 2012 / Revised: June 1, 2012 / Accepted: June 3, 2012)

ABSTRACT

Hierarchical forecasting strategy does not always outperform direct forecasting strategy. The performance generally depends on demand features. This research guides the use of the alternative forecasting strategies according to demand features. This paper developed and evaluated various classification models such as logistic regression (LR), artificial neural networks (ANN), decision trees (DT), boosted trees (BT), and random forests (RF) for predicting the relative performance of the alternative forecasting strategies for the South Korean navy's spare parts demand which has non-normal characteristics. ANN minimized classification errors and inventory costs, whereas LR minimized the Brier scores and the sum of forecasting errors.

Keywords: Hierarchical Forecasting, Non-Normal Demand, Classification, Machine Learning

* Corresponding Author, E-mail: m1s1m@hotmail.co.uk

1. INTRODUCTION

All Military establishments experience spare parts supply problems caused by inaccurate forecasts of spare parts demand (Seon and U, 2009). In common with many militaries, the South Korean Navy is under immense pressure to maintain adequate stocks of warship's spare parts with budgetary limitations. This requires a careful choice of a forecasting strategy. However, forecasting demand for spare parts is difficult. This is because the spare parts demand exhibits non-normal characteristics (Moon *et al*., 2012). Demand that has infrequent occurrences, low average volumes or highly variable volumes is said to be non-normal (Boylan *et al*., 2008).

A hierarchical structure of time series comprises individual item-level time series and an aggregated group-level time series in which the items are members (Hyndman *et al*., 2007). A hierarchical forecasting strategy (HF) derives forecasts at item level by prorating demand forecasts for the group in which the items are members; whereas a direct forecasting strategy (DF) simply generates a forecast using item-level time series. HF may be divided into two sub-strategies. Top-down forecasting (TDF) models a forecast at the top group level by using the top group-level time series, and then creates lower-level forecasts according to the item's percentage contribution within the group. Combinatorial forecasting (CF) models forecasts at all levels by using all levels of the time series, and then creates lower-level forecasts based on a combination of the forecasts.

When an item-level time series is highly variable and intermittent, a higher group-level time series is usually less variable and less intermittent (Widiarta *et al*., 2009). The lower level of variability and intermittency of a group-level time series can produce a more reliable item-level time series forecast by using HF (Fliedner and Lawrence, 1995).

In practice, the relative performance of the alternative forecasting strategies varied (Moon *et al*., 2012).

Several studies (Dekker *et al*., 2004; Hyndman *et al*., 2007; Kahn, 1998) reported that CF was more accurate than TDF and DF. It is important in practice to identify which of the different forecasting strategies is superior (Fildes *et al*., 2008).

Many authors (Chen and Boylan, 2009; Schwarzkopf *et al*., 1988) have found that the relative performance of HF and DF is conditional on demand features. In order to guide the selection of a forecasting strategy by multivariate demand features, Moon *et al*. (2012b) developed a logistic regression (LR) classification model for predicting the performance of alternative forecasting strategies for the naval spare parts demand, and demonstrated that this model reduced forecasting errors and inventory costs.

Recent developments in the field of machine learning have led to a renewed interest in this classification problem. Given the various machine learning models such as artificial neural networks (ANN), decision trees (DT), boosted trees (BT) (Freund and Schapire, 1999) and random forests (RF) (Breiman, 2001), a question has been raised about which of the different models is suitable for the classification problem of this research. The features of data and the purposes of modeling could be criteria for the choice of a classification model, as claimed by Tu (1996).

The objectives of this research are: (i) to compare the characteristics of the five classification models for predicting the performance of alternative forecasting strategies in forecasting the demand for the naval spare parts; and (ii) to evaluate the performance of the classification models in terms of classification accuracy and inventory costs.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical framework for the relationships between the forecasting strategies and demand features and the characteristics of the machine learning models. Section 3 describes the features of the spare parts demand and the alternative forecasting strategies. Section 4 develops the classification models. This is followed by classification results and analysis in Section 5. Finally, Section 6 presents the conclusions.

2. DEMAND FEATURES AND CLASSIFICATION

This section summarizes research that has investigated the impact of demand features on the relative performance of HF and DF, and describes the characteristics of the machine learning models

2.1 Demand Features

Table 1 compares the results of seven major studies that identified demand features which significantly influenced the performance of the alternative forecasting strategies.

Various demand features were found to have a significant influence on the relative performance of HF and

Table 1. The Influence of Demand Features on the Performance of Hierarchical Forecasting and Direct Forecasting (Moon, 2012)

Demand feature	Impact on relative performance	Comparison	Data	Reference
Correlation	↓TDF	Analytic study		Schwarzkopf *et al*. (1988)
		Simulation	Seasonal models	Chen and Boylan (2009)
	↓DF	Analytic study and Simulation	AR(1)	Widiarta *et al*. (2006)
	↓CF	Empirical study	Warship's spare parts	Moon *et al*. (2012b)
Variability in demand volume	↓CF			
Equipment group	Gun/Radar: ↑CF			
Substitutability (ψ) and variability of proportion (υ)	$\uparrow\psi$ (&$\uparrow\upsilon$) or $\downarrow\upsilon$: ↑TDF, $\downarrow\psi$ and $\uparrow\upsilon$: ↑DF	Simulation	AR(1), MA(1) and ARMA(1, 1)	Widiarta *et al*. (2008b)
Forecasting horizon	↑TDF	Analytic study		Shlifer and Wolff (1979)
Lag-1 auto correlation [$\rho(1)$]	$\rho(1) > 1/3$: ↑DF	Analytic study and Simulation	AR(1)	Widiarta *et al*. (2006)
Grouping criteria	Volume and dollar-volume: ↑TDF	Empirical study	Automotive spare parts	Fliedner and Mabert (1992)

ψ = the portion of the unsatisfied demand for an item that is passed to another item within a group; ↓ (or ↑) = decreasing (or increasing) the value of the demand feature increases the relative performance of the forecasting strategy; TDF = top-down forecasting; DF = direct forecasting; CF = combinatorial forecasting; dollar-volume = demand per period×item price.

DF. Correlation was a contentious feature as the effect across studies was inconsistent. Most studies compared TDF and DF and did not consider a combined influence of multivariate demand features on the performance, with the exception of the study by Moon *et al*. (2012b). The first classification model to predict the performance of CF and DF by the multivariate LR was suggested by Moon *et al*. (2012b). The demand features included in their model were the variability in demand volume (capturing characteristics of non-normal demand), correlation, and the equipment group. However, the classification accuracy of the LR was unsatisfactory (55.6%).

2.2 Classification Models

The machine learning models, for instance ANN, DT, BT, and RF, can be alternatives to the LR model developed by Moon *et al*. (2012b). Table 2 summarizes the characteristics of these models in the literature. Each model has its own advantages and disadvantages. These machine learning models have better capabilities than LR for predicting the performance of forecasting strategies for non-normal demand. ANN can detect more complex nonlinear relationships between outcome and predictor variables than LR. This is because the predictor variables generally go through a nonlinear transformation at each hidden layer and output layer (Tu, 1996). With the tree diagram, DT is easily expressed as a set of rules, and is therefore termed a white-box model (Dreiseitl and Ohno-Machado, 2002). As ensemble methods using DT, BT and RF were also considered. This is because BT (Freund and Schapire, 1999) and RF (Caruana and Niculescu-Mizil, 2006) were claimed to outperform DT.

Despite the better capabilities of the above machine learning models, little attention has been paid to the investigation into machine learning models for such problems. This paper attempts to fill that research gap.

3. CASE STUDY

This section summarizes the results of the previous research (Moon, 2012; Moon *et al*., 2012a; 2012b) on which this paper is based. The features of the demand for spare parts within the South Korean navy and the alternative forecasting strategies are described.

3.1 The Characteristics of the Spare Parts Demand

The time series (2001. 2~2011. 7) for the spare parts demand obtained from the navy were aggregated into monthly time buckets. Table 3 reviews the features of the time series averaged over the 300 items.

The demand features considered were: correlation (Widiarta *et al*., 2008a); the coefficient of variation in demand volume (Cv (vol)) expressed as the standard deviation of demand volume divided by mean demand volume; the number of periods with zero demand (Bauer

Table 2. The Characteristics of the Machine Learning Models

	Advantages	Disadvantages	Performance
ANN	- Ability to detect complex nonlinear relationships among the variables	- Susceptible to over-fitting - A black-box model	- No significant difference between LR and ANN (Dreiseitl and Ohno-Machado, 2002) - Robust performance compared with LR, DT, BT, and RF (Caruana and Niculescu-Mizil, 2006)
DT	- Easy to express as a set of rules (a white-box model) - Ability to detect the structure in data with hierarchical variables	- Susceptible to over-fitting - Discontinuity of the outcome depending on the threshold built in the tree - High variance of outcome due to small perturbations of data	- Superior to LR for large data (larger than 10,000) (Perlich *et al*., 2003)
BT	- Resistance to over-fitting - Few parameters to tune - Ability to identify outliers	- Susceptible to noise - A gray-box model	- Reduces bias and variance compared to DT (Aliev and Aliev, 2007) - Better performance than DT (Freund and Schapire, 1999) - Poor performance with insufficient data or many outliers (Freund and Schapire, 1999)
RF	- Resistance to noise, outliers and over-fitting - Independence of each tree from the other trees when built - Low bias due to random predictor selection - Few parameters to tune	- A gray-box model	- Better performance than DT (Caruana and Niculescu-Mizil, 2006) - Good performance with large data and a large number of input variables (Williams, 2011)

ANN = artificial neural networks; DT = decision trees; BT = boosted trees; RF = random forests.

and Kohavi, 1999); mean demand volume (Mean (vol)) (Fliedner and Lawrence, 1995); and forecasting horizon (Shlifer and Wolff, 1979). In order to standardize the measure in different periods of the time series, the proportion of periods with zero demand (Pr (zero)), defined as the number of periods with zero demand divided by the total period, was used. As the forecasting horizon for the navy was calculated as a procurement lead time (PLT) plus the fixed review cycle (12 months), PLT was simply used as the statistic representing the forecasting horizon.

There were significant correlation, high Cv (vol), high Pr (zero), low Mean (vol), and long PLT. This indicated that the time series were correlated, were non-normal, and required long forecasting horizons. Some relative demand features in the equipment groups were identified. For example, Gun/RD had higher intermittency and shorter forecasting horizon than the others, as indicated by the higher Pr (zero) and the shorter PLT.

Table 3. Statistical Features of the Time Series (Moon, 2012)

	Total	Gun/RD	ME	GE/AC
Correlation	0.77	0.81	0.76	0.79
Cv (vol)	2.18	2.12	2.13	2.37
Pr (zero)	0.49	0.61	0.46	0.50
Mean (vol)	15.82	2.01	22.39	6.60
PLT	9.47	9.20	9.57	9.34

Gun/RD = Gun and Radar (44 items); ME = Main Engine (188 items); GE/AC = Generator and Air Compressor (68 items).

3.2 The Alternative Forecasting Strategies

This paper employed the most robust direct forecasting (DF) and the most robust combinatorial forecasting (CF) among the forecasting strategies tested by Moon *et al*. (2012a) as the alternative forecasting strategies based on simple exponential smoothing for forecasting demand for the spare parts. The forecasts for 300 items were generated in 2001. 5, 2001. 6 and 2001. 7 (i.e. 900 forecasts), and were based on all the available previous records (from 2001. 2).

In order to measure the practical impact of the prediction accuracy on the navy's inventory systems, a simulation exercise was employed by Moon *et al*. (2012a). The simulated inventory system was a periodic review, order-up-to-level system. The total inventory costs were calculated as 'unit variable cost×(0.2×mean inventory per month+0.4×mean stock-out per month).'

DF (the forecast with monthly aggregated data adjusted for linear trend and additive seasonality), which minimizes the mean absolute deviation (MAD) and root mean square error (RMSE) (whereas the forecasting method with monthly aggregated unadjusted data minimizes inventory costs), is called the most robust direct forecast (RDF). SCF (the simple combination (Eq. (1)) between the group-level DF with quarterly aggregated data adjusted for linear trend and the item-level DF with monthly aggregated unadjusted data), which minimizes MAD, RMSE and inventory costs, is called the most robust simple combination forecast (RCF).

$$SCF_{i,t+\tau} = \frac{1}{2}(F_{t+\tau} + \sum_{i=1}^{N} f_{i,t+\tau}) \times \frac{f_{i,t+\tau}}{\sum_{i=1}^{N} f_{i,t+\tau}}. \quad (1)$$

where: $SCF_{i,t+\tau}$ = simple combination forecast of demand i, τ periods ahead, made at time t;
$F_{t+\tau}$ = group-level DF, τ periods ahead, made at time t;
$f_{i,t+\tau}$ = item-level DF of demand i, τ periods ahead, made at time t.

In this section, the features of the spare parts demand and the alternative forecasting strategies were presented. These are used for implementing the classification in the next section.

4. THE DEVELOPMENT OF THE CLASSIFICATION MODELS

This section presents the procedure for building logistic regression (LR) used by Moon *et al*. (2012b) and artificial neural networks (ANN), decision trees (DT), boosted trees (BT), and random forecasts (RF). Other model such as support vector machines (SVM) (Caruana and Niculescu-Mizil, 2006) was also considered. However the error rate of SVM was higher than the error rate of the null model (classifying all the observations into the majority class (RCF)). As SVM in this case was not a sensible model, it was not included in this paper.

4.1 Implementing variables

As the dichotomous outcome variable within the classification models, the performance of each forecast for each item was measured by the absolute deviation divided by the item's monthly mean consumption of data (*AD/M*) in order to eliminate scale dependencies of an item with large consumption, as shown in Eq. (2)

$$AD/M = \frac{|y_{i,t} - \hat{y}_{i,t}|}{\bar{y}_{i,t}}. \quad (2)$$

where: $y_{i,t}$ = the observed demand for item i at time t;
$\hat{y}_{i,t}$ = the forecast demand for item i at time t;
$\bar{y}_{i,t}$ = the monthly mean demand for item i between t-k and t-1;

$$k = \begin{cases} 3, t = 1 \text{ (}i.e.\text{ 05)} \\ 4, t = 2 \text{ (}i.e.\text{ 06)} \\ 5, t = 3 \text{ (}i.e.\text{ 07)} \end{cases}$$

RCF-RDF represents *AD/M* of RCF minus that of RDF. A positive value of RCF-RDF denotes that the RDF is superior to the RCF, and negative value indicates that the RCF is superior.

As predictors, five continuous variables and a categorical variable (Equipment representing the three categories of spare parts used for Gun/RD, ME, and GE/AC) as shown in Table 3 were examined within LR in terms of the log likelihood ratio test. *AD/M* of RCF and RDF at time *t* (as an outcome) and the corresponding demand features between *t*-*k* and *t*-1 (as predictors) were used for fitting the classification models. Cv (vol) and Equipment were within the range of the general threshold for inclusion (P-value, 0.05). In order to examine the impact of the contentious demand feature Correlation (as shown in Table 1), the threshold was expanded to 0.10, as with the criterion used by Ottenbacher *et al*. (2001). Inconsistent with the study of Shlifer and Wolff (1979) as shown in Table 1, PLT (representing the forecasting horizon) was non-significant.

The distributions of Cv (vol) and Correlation had a number of outliers and were significantly skewed (0.90 for Cv (vol); -1.46 for Correlation). When data deviate from a normal distribution, it is sensible to consider transforming the data (Miller, 1986). However, skewness is difficult to eliminate with a transformation (Chat-field, 2004). The linear transformation using a quadratic, cubic, log, or inverse function makes a case, which is close to being an outlier before transformation, an extreme case after transformation (Miles and Shevlin, 2001). Ro-bust estimators (e.g. trimming and winsorizing), which could convert outliers into proximity with the rest of the data, are based on the assumption that underlying distribution is symmetric about its median (Miller, 1986). Asymmetric distribution such as Cv (vol) and Correlation could not be handled by the robust estimators. Therefore, transformation was not conducted in the research.

The scatter plots of the demand features compared with forecasting performance with the least-square line are presented in Figure 1.

Although there are no strong relationships between Correlation (or Cv (vol)) and RCF-RDF, the least-square line suggests that RCF was superior; however, when Correlation (or Cv (vol)) increased, the performance of RCF became moderate compared to that of RDF.

Figure 2 presents a mosaic plot for Equipment by the number of the superior forecasting strategies in terms of Eq. (2). RCF was roughly three times better than RDF in the case of Gun/RD and it was therefore used as the reference group with its demand features in comparison to the other two groups.

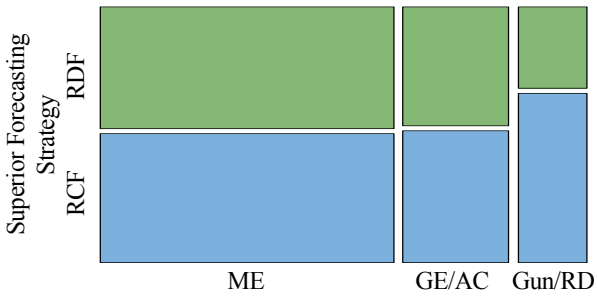

Figure 2. Equipment by the Superior Forecasting Strategies

4.2 Implementing the Classification Models

For the purpose of comparison, the identical variables and data used for LR were entered into the machine learning models. LR was fitted with the 900 observations (forecasts), with the predictors (Correlation, Cv (vol) and Equipment) by using *R Commander* 1.5-5 (Fox, 2012). Each machine learning model was implemented using *Rattle* (Williams, 2009) embedded with several other R software packages used in constructing ANN, DT, BT, and RF.

If *AD/M* for RCF was greater than that for RDF, the outcome was encoded as 0 (RDF); otherwise it was encoded as 1 (RCF). This classification result of the

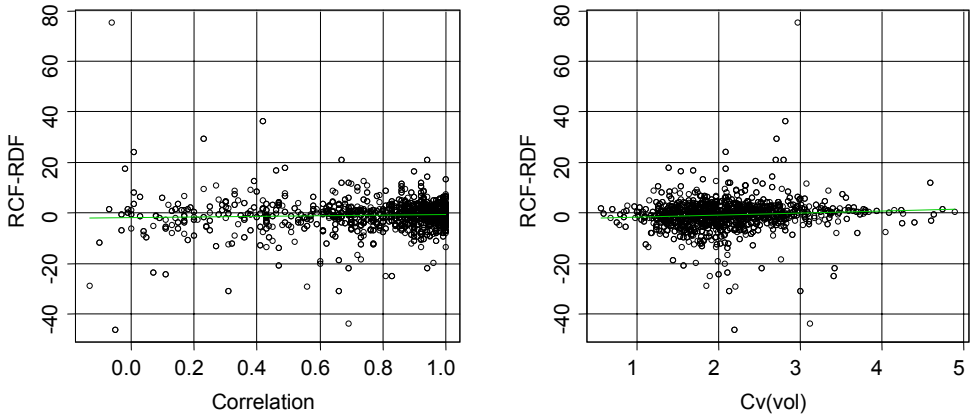

Figure 1. Demand Features vs. Forecasting Performance (Moon *et al*., 2012b)

forecasting methods is called the observed classification. If the outcome was smaller than 0.5 it was classified as RDF; otherwise it was classified as RCF.

In order to establish internal validity for a predictive model, this paper tested 10 sets for the 10% cross-validation. Many parameter settings for each model were tested and the settings which provided the best performance were used for generating results.

4.2.1 Logistic Regression (LR)

The probability of RCF being superior to RDF (Pr (*RCF*)) can be defined in Eq. (3).

$$\Pr(RCF) = \frac{e^{(\beta_0 + \beta_1 Corr + \beta_2 Cv(vol) + \beta_{3-4} Equipment)}}{1 + e^{(\beta_0 + \beta_1 Corr + \beta_2 Cv(vol) + \beta_{3-4} Equipment)}}. \quad (3)$$

LR requires continuous predictors; however, it can include a categorical predictor by using dummy coding. The categorical variable, Equipment was encoded as two dummy variables (Equipment (ME), Equipment (GE/AC)) with values of 00 for Gun/RD (as a reference category), 10 for ME, and 01 for GE/AC respectively.

4.2.2 Artificial Neural Networks (ANN)

Feed-forward neural networks with a single hidden layer (Venables and Ripley, 2002) were built using *Rattle* embedded with R software package *nnet* (Ripley, 2012). Dreiseitl and Ohno-Machado (Dreiseitl and Ohno-Machado, 2002) pointed out that one layer of hidden neurons is generally sufficient for classifying most data sets. One of the disadvantages of ANN is over-fitting. Limiting the number of hidden nodes may prevent over-fitting, whereas no theory exists for predetermining the optimal number of hidden nodes (Tu, 1996). After a series of examinations this paper selected the settings of 3 hidden nodes.

4.2.3 Decision Trees (DT)

A classification and regression tree (CART) (Breiman *et al*., 1984) was structured using R software package *rpart* (Thereau and Atkinson, 2012). An information gain measure was employed for deciding between alternative splits. In order to avoid over-fitting, several options are available. The maximum depth of a tree, 3, and complexity parameter, 0.01, were chosen.

4.2.4 Boosted Trees (BT)

Adaptive Boosting (AdaBoost) was built using R software package *ada* (Culp *et al*., 2012), following the algorithms listed in Friedman *et al*. (2000). The individual decision trees within AdaBoost were built using *rpart*. AdaBoost has a parameters to tune (i.e. the number of iterations to boost) (Freund and Schapire, 1999). An examination of the error rates suggested that there was very little reduction gained by adding more than 50 trees. This led to the choice of 50 iterations.

4.2.5 Random Forests (RF)

A random forest algorithm (Breiman, 2001) was implemented using R software package *random Forest* (Liaw and Wiener, 2002). RF has only two parameters (the number of trees in the forest and the number of variables at each node) (Liaw and Wiener, 2002). An examination of the error rates suggested that there was very little change achieved by adding more than 1,000 trees to the forest. This resulted in 1,000 as the number of trees. The output of RF depends primarily on the number of variables to be chosen randomly at each tree node (Prasad *et al*., 2006). The general default value is one-third the number of variables (Prasad *et al*., 2006) or the square root of the total number of variables available for classification tasks (Williams, 2011). Liaw and Wiener (2002) stipulated that one variable can give a very good performance for some data. Based on these recommendations, the number of variables (to be chosen randomly at each node) for this research was selected to be one.

5. CLASSIFICATION RESULTS AND ANALYSIS

This section analyses the results of the five classification models and evaluates the performance of the models.

5.1 The Interpretation of the Results

There was no significant difference in the estimated coefficients for LR across the 10 cross-validation training sets and overall data set. Table 4 presents the predictors in LR built with the overall data. The significance of the predictors was estimated using a *z*-statistic (defined as the estimated coefficient divided by the standard error).

Table 4. Predictors in the Logistic Regression (Moon *et al*., 2012b)

Predictor	β	Std error	z	P-value	e^{β}
Correlation	-.366	.266	-1.38	1.68-e01	.694
Cv (vol)	-.367	.109	-3.37	7.43-e04	.693
Equipment (ME)	-.702	.206	-3.41	6.41e-04	.496
Equipment (GE/AC)	-.560	.236	-2.37	1.76-e02	.571
Intercept	1.792	.346	5.18	2.27e-07	

β = the estimated regression coefficient.

Cv (vol) had the most significant effect on the relative performance of alternative forecasting methods, followed by Equipment. Correlation had a marginal effect on the performance.

The odds ratio (defined as e^{β}) indicates that, when Correlation or Cv (vol) increased with other predictors

held constant, the log odds of the probability of RCF being superior decreased. This is consistent with the effects of Correlation and Cv (vol) as shown in Figure 1.

The effect of Correlation is consistent with the results of Schwarzkopf *et al*. (1988) and Chen and Boylan (2007), as shown in Table 1, if CF is considered to be a variant of TDF.

With regard to Equipment, RCF for ME is 0.496 times as likely as RCF for Gun/RD to be superior to RDF, and RCF for GE/AC is 0.571 times as likely as RCF for Gun/RD to be superior to RDF.

As Cv (vol), Equipment, and Correlation were presented respectively in all sets, nine sets, and two sets of the 10 cross-validation training sets for DT, Cv (vol) was the most important variable, followed by Equipment. With the overall data, three rules were generated, as shown in Table 5.

Table 5. Decision Trees Rules Built with Overall Observations

#	Rules
1	If (Cv (vol) ≥ 2.065) and (Equipment = Gun/RD), then RSF is superior.
2	If (Cv (vol) < 2.065), then RSF is superior.
3	If (Cv (vol) ≥ 2.065) and (Equipment = GE/AC or ME), then RDF is superior.

$Rule_1$ indicates a 65% chance of RCF being superior and covers 6% (54) of the data set; $Rule_2$ denotes a 60% chance of RCF being superior and covers 53% (474) of the data set; $Rule_3$ represents a 55% chance of RDF being superior and covers 41% (372) of the data set. $Rule_1$ is consistent with the results of LR in that RCF performed better for Gun/RD. $Rule_2$ also corroborated the results of LR in that low Cv (vol) is a favorable condition for RCF.

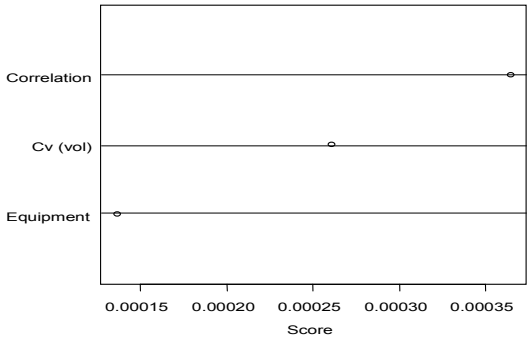

Figure 3. Variable Importance Plot for AdaBoost

There was no significant difference in the rank of the predictors for AdaBoost across the 10 cross-validation training sets and overall data set, as Cv (vol) was used in the greatest number of iterations, followed by Correlation and Equipment. Cv (vol), Correlation, and Equipment were employed in 49, 41, and 35 iterations respectively, out of 50 iterations using the overall data set. The variable importance was calculated as the average improvement in accuracy for the variable selected to split the data over all trees in the ensemble. Figure 3 presents the variable importance plot for AdaBoost using the overall data set.

It was interesting that Correlation (which was the least significant variable in other models) was highest ranked for AdaBoost.

No significant difference was identified in the rank of the predictors for RF across the 10 cross-validation training sets and overall data set. The mean decrease accuracy is calculated as a scaled average of the prediction accuracy of each variable. The mean decrease Gini indicates the total decrease of the impurity in a DT node when splitting on a variable using the Gini index. A higher value of these measures indicates that the variable is more important. The variable importance plots for RF built with the overall data are presented in Figure 4. Both measures identified Cv(vol) as the most important predictor, with the other two predictors following in different orders.

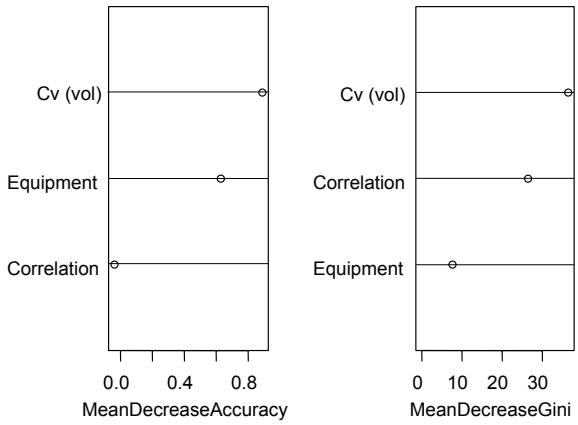

Figure 4. Variable Importance Plots for RF

The impact of the variables in LR, DT and RF (in terms of *z*-statistics of β as shown in Table 4, the number of employments in the 10 cross-validation sets, and the mean decrease accuracy as shown in Figure 4 respectively) was consistent. The most important variable was Cv (vol) followed by Equipment and Correlation. AdaBoost focused on the most difficult variable (albeit the least significant variable in other models) and produced the highest score. This might lead to the inconsistent result in AdaBoost that Correlation took the highest rank in terms of the variable importance (as shown in Figure 3).

5.2 The Performance of the Classification Models

Table 6 compares the prediction results of the models in the 10 cross-validation test sets. 'Predicted' indicates these prediction results; 'observed' indicates the classification results from the observed classification. All the models produced marginally smaller error rates

than the null model classifying all the observations into the majority class (RCF). ANN minimized the classification errors, followed by LR.

Table 6. Error Matrix for the Models

	Observed	Predicted		Error (%)
		RDF	RSF	
RSF	RDF	0	414	100
	RSF	0	486	0
	Overall			46.0
LR	RDF	164	250	60.4
	RSF	150	336	30.9
	Overall			44.4
ANN	RDF	166	248	59.9
	RSF	130	350	26.7
	Overall			42.0
DT	RDF	167	247	59.7
	RSF	156	330	32.1
	Overall			44.8
BT	RDF	143	271	65.5
	RSF	138	348	28.4
	Overall			45.4
RF	RDF	119	295	71.3
	RSF	113	373	23.3
	Overall			45.3

Table 7 compares the performance of the models in terms of the sum of forecasting errors (calculated by Eq. (2)), the Brier score (Wilks, 2011), and the inventory costs (Moon *et al*., 2012) in the 10 cross-validation test sets.

Table 7. Performance of the Models

	Sum of errors	Brier	Inventory costs
RSF	7,155		\$597,572
LR	**7,062**	**0.2435**	**\$584,195**
ANN	**7,084**	**0.2475**	**\$571,920**
DT	7,209	**0.2499**	\$617,824
BT	7,284	**0.2497**	\$618,264
RF	7,268	0.2835	\$616,261

The Brier score of a sensible model ranges from 0 (perfect) to 0.25 (Steyerberg *et al*., 2001). The sum of errors and the inventory costs that are smaller than the null model and the sensible Brier scores are shown in bold.

LR minimized the sum of forecasting errors and the Brier score, whereas ANN minimized the inventory costs. All models except RF were within the range of a sensible model in terms of the Brier score. ANN and LR were superior to the null model in terms of all the measures. These results are consistent with the argument of authors (Caruana and Niculescu-Mizil, 2006) that ANN was robust, and the claim of researchers (Dreiseitl and Ohno-Machado, 2002) that the performances of ANN and LR were similar. On the basis of these results, it might be claimed that the internal validity of LR and ANN has been established. The performance of ANN and LR was marginally higher than the null model. This might be because the low reliability of data (due to the recent stabilization of the naval logistical data base) introduced extremely non-normal characteristics.

As shown in Table 2, DT performed worse than LR with small amounts of data (e.g. smaller than 10,000 observations) (Perlich *et al*., 2003); BT performed poorly with insufficient data and many outliers (Freund and Schapire, 1999) and was susceptible to noise (Aliev and Aliev, 2007); RF is generally suitable for large data and a large number of input variables (Williams, 2011). The unsatisfactory performance of DT, BT, and RF could be explained by the outliers and noise originating from the low reliability of data, the small size of the data sets (i.e. 900), or the small number of input variables (i.e. 3).

In general, ANN or LR might be the recommended choice for the problem facing this research. However, when the inventory manager wants to understand the causal relationships between the performance of forecasting strategies and the demand features, LR can provide a clear advantage due to the black-box characteristics of ANN (Tu, 1996). If the primary goal is the interpretation of the results, DT still remains one of the choices for that problem.

6. CONCLUSIONS

This paper compared the five classification models for predicting the performance of the alternative forecasting strategies (most robust simple combination forecast and most robust direct forecast) for forecasting non-normal demand associated with spare parts demand in the South Korean navy. The five classification models were constructed using logistic regression (LR), artificial neural networks (ANN), decision trees (DT), boosted trees (BT), and random forests (RF), with three predictors (Correlation, variability in demand volume (Cv (vol)), and Equipment).

The most important predictor in LR (in terms of z-statistics of β), DT (in terms of the number of employments in the 10 cross-validation sets) and RF (in terms of the mean decrease accuracy) was Cv (vol), followed by Equipment and Correlation. The feature of BT that concentrates on the most difficult variable might be the reason that Correlation took the highest rank in terms of the variable importance for BT. Consistent impacts of Cv (vol) and Equipment in LR and DT were identified. When Cv (vol) within LR increased, the log odds of the probability of RCF being superior decreased. This was presented as a rule by DT; that is, if Cv (vol) is less than 2.065, then RCF is superior. LR identified that RCF

performed better for Gun/RD, and DT also generated a rule including this effect.

The performance of the five models was evaluated by several measures. In terms of the error matrix (as shown in Table 6), all the five models performed better than the null model, and the best model was ANN, followed by LR. All models except RF were sensible in terms of the Brier score. Only ANN and LR were superior to the null model in terms of all the measures such as the sum of forecasting errors, the Brier score, and the inventory costs, as shown in Table 7. LR minimized the sum of forecasting errors and the Brier score; ANN minimized the classification errors and the inventory costs. These points lead to the conclusion that the reliability and internal validity of ANN and RF have been established.

The eventual purpose of the classification model was to minimize the inventory costs through guiding the use of the alternative forecasting strategies. Therefore ANN might be the best choice for the inventory manager.

Although ANN performed nicely, it is difficult to interpret the weights. Sometimes the primary goal of the inventory manager might be an interpretation of the results. In this case, LR (or even DT) can be recommended.

This paper identified the characteristics of the five classification models in order to predict the performance of the alternative strategies for forecasting the spare parts demand and to evaluate the performance of the classification models in terms of classification accuracy and inventory costs. Therefore the research gap might be claimed to be filled.

The performance of ANN and LR was found to be marginally superior to the null model. This may be caused by the low reliability of data obtained from the unsettled database. The low reliability of data might lead to the poor performance of BT. As DT, BT, and RF are generally used with sufficient data (as shown in Table 2), the data set for this research with only 900 items might be an unfavorable setting for these models. Further investigation into the performance of the classification with more reliable and/or larger datasets is strongly recommended.

REFERENCES

Bauer, E. and R. Kohavi, "An empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning* 36, 1/2 (1999), 105-139.

Boylan, J. E., A. A. Syntetos, and G. C. Karakostas, "Classification for forecasting and stock control: a case study," *Journal of the Operational Research Society* 59 (2008), 473- 481.

Breiman, L., "Random forests," *Machine Learning* 45, 1 (2001), 5-32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Chapman and Hall, NY, USA, 1984.

Caruana, R. and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms using different performance metrics*, Pittsburgh, USA: 2006.

Chatfield, C., *The Analysis of Time Series: an Introduction*, *6th edn*., Chapman and Hall/CRC, London, UK, 2004.

Chen, H. and J. E. Boylan, The effect of correlation between demands on hierarchical forecasting, In: Lawrence, K. D. Klimberg, R. K. (Eds.), *Advances in business and management forecasting*, Bingley, UK: Emerald Group Publishing Limited, (2009), 173-188.

Culp, M., K. Johnson, G. Michailidis, *ada: an R package for stochastic boosting*, Available at: http://cran.ma.imperial.ac.uk/, Accessed on 4 June 2012.

Dekker, M., K. V. Donselaar, and A. P. Ouwehand, "How to use aggregation and combined forecasting to improve seasonal demand forecasts," *International Journal of Production Economic* 90, 2 (2004), 151-167.

Dreiseitl, S. and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics* 35, 5/6 (2002), 352-359.

Fildes, R., K. Nikolopoulos, S. F. Crone, and A. A. Syntetos, "Forecasting and operational research: a review," *Journal of the Operational Research Society* 59 (2008), 1150-1172.

Fliedner, E. B. and B. Lawrence, "Forecasting system parent group formation: An empirical application of cluster analysis," *Journal of Operations Management* 12, 2 (1995), 119-130.

Fliedner, E. B. and Mabert, "Constrained forecasting: some implementation guidelines," *Decision Sciences* 23, 5 (1992), 1143-1161.

Fox, J., *R Commander 1.5-5*, Available at: http://cran.ma.imperial.ac.uk/, Accessed on 4 June 2012.

Freund, Y. and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence* 14, 5 (1999), 771-780.

Friedman, J., T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics* 28, 2 (2000), 337-407.

Hyndman, R. J., R. A. Ahmed, and G. Athanasopoulos, "Optimal combination forecasts for hierarchical time series," *Monash University* (2007), 1-21.

Kahn, K. B., "Revisiting top-down versus bottom-up forecasting," *The Journal of Business Forecasting* (1998), 14-19.

Liaw, A. A. and M. Wiener, "Classification and regression by random forest," *R news* 2, 3 (2002), 18-22.

Miles, J. and M. Shevlin, *Applying Regression and Correlation: a Guide for Students and Researchers*,

Sage publications, London, UK, 2001.

Miller Jr., R. G., *Beyond ANOVA, Basics of Applied Statistics*, John Wiley and Sons, Inc., NY, USA, 1986.

Moon, S., "The impact of demand features on the performance of hierarchical forecasting: case study for spare parts in the navy," *Korean Management Science Review* 29, 1 (2012), 101-114.

Moon, S., A. Simpson, and, C. Hicks, "The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean Navy-A case study," *International Journal of Production Economics* 140 (2012a), 794-802.

Moon, S., A. Simpson, and C. Hicks, "The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand," *International Journal of Production Economics* (2012b), 10.1016/j.ijpe.2012.02.016.

Ottenbacher, K. J., P. M. Smith, S. B. Illig, R. T. Linn, R. C. Fiedler, and C. V. Granger, "Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke," *Journal of Clinical Epidemiology* 54, 11 (2001), 1159-1165.

Perlich, C., F. Provost, and J. S. Simonoff, "Tree induction vs. logistic regression: a learning-curve analysis," *The Journal of Machine Learning Research* 4, (2003), 211-255.

Prasad, A. M., L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems* 9, 2 (2006), 181-199.

Ripley, B., *nnet: feed-forward neural networks and multinomial log-linear models*, Available at: http://cran.ma.imperial.ac.uk/, Accessed on 4 June 2012.

Schwarzkopf, A. B., R. J. Tersine, and J. S. Morris, "Top-down versus Bottom-up Forecasting Strategies," *International Journal of Production Research* 26, 11 (1988), 1833-1843.

Seon, M.-S. and U, J.-U., "A study on forecasting of repair part demands of Korean Military: focused on Navy," *The Korean Journal of Defense Analysis* 85 (2009), 201-234.

Shlifer, E. and R. W. Wolff, "Aggregation and proration in forecasting," *Management Science* 25, 6 (1979), 594-603.

Steyerberg, E. W., F. E. Harrell, G. J. J. M. Borsboom, M. J. C. R. Eijkemans, Y. V. Vergouwe, and J. D. F. Habbema, "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," *Journal of Clinical Epidemiology* 54, 8 (2001), 774-781.

Thereau, T. M. and B. Atkinson, *rpart: recursive partitioning*, Available at: http://CRAN.R-project.org/package = rpart, Accessed on 23 January 2012.

Tu, J. V., "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology* 49, 11 (1996), 1225-1231.

Venables, W. N., B. D. Ripley, *Modern Applied Statistics with S*, *4th edn*., Springer, NY, 2002.

Widiarta, H., S. Viswanathan, and R. Piplani, "On the effectiveness of top-down strategy for forecasting autoregressive demands," *Naval Research Logistics* 54, 2 (2006), 176-188.

Widiarta, H., S. Viswanathan, and R. Piplani, "Forecasting item-level demands: an analytical evaluation of top-down versus bottom-up forecasting in a production-planning framework," *IMA Journal of Management Mathematics* 19 (2008a), 207-218.

Widiarta, H., S. Viswanthan, and R. Piplani, "Evaluation of hierarchical forecasting for substitutable products," *International Journal of Services and Operations Management* 4, 3 (2008b), 277-295.

Widiarta, H., S. Viswanathan, and R. Piplani, "Forecasting aggregate demand: an analytical evaluation of top-down versus bottom-up forecasting in a production planning framework," *International Journal of Production Economics* 118, 1 (2009), 87-94.

Wilks, D. S., *Statistical Methods in the Atmospheric Science*, *3rd edn*., Elsevier, Oxford, UK, 2011.

Williams, G., "Rattle: a data mining GUI for R," *The R Journal* 1, 2 (2009), 45-55.

Williams, G., *Data Mining with Rattle and R: the Art of Excavating Data for Knowledge Discovery*, Springer, London, UK, 2011.