

데이터마이닝을 이용한 공군 F-15K 수리부속 예측모형에 관한 연구

송민규, 마정목*
국방대학교 국방과학학과

A study on predictive model for forecasting Airforce F-15K spare parts Using Data Mining

Min-Gyu Song, Jung-Mok Ma*
Department of Defense Science, Korea National Defense University

요약 군에서 사용하고 있는 수리부속은 약 72만 종으로, 국방 예산의 약 3%(약 1.3조원)을 차지하고 있으며, 적시적기의 수리부속 보급은 무기체계 전투준비태세에 직접적인 영향을 미치고 있어 수리부속 수요예측은 예산 운영과 장비 가동률 측면에서 매우 중요한 요소가 되고 있다. 그래서 현재 공군에서 적용 중인 시계열 모형의 예측한계를 개선하기 위해 다수의 연구들이 데이터마이닝 중 지도학습을 통한 새로운 외부변수를 추가하여 예측률을 향상시키는데 집중하였고, 비지도학습과 피쳐 엔지니어링을 통한 예측 정확도 향상 연구는 한정되어 있었다. 따라서 본 연구에서는 수리부속 수요예측 정확도 제고를 위해 공군에서 운용중인 F-15K 전투기에 대한 2014 ~ 2019년까지의 5,752개 수리부속 품목의 소모개수와 표준단가 데이터 수집하였고, 데이터마이닝 모형(Random forest, GBM, XGBoost, LightGBM, Ridge, Lasso, Elastic Net regression)을 활용하여 시계열모형과 예측률을 비교하였다. 예측력척도는 RMSE (평균제곱근오차)를 사용하였다. 연구결과 데이터마이닝 모형이 현재 공군에서 적용중인 시계열 예측 모형에 비해 예측정확도가 높았으며, 이어서 비지도학습의 차원축소와 피쳐 엔지니어링을 통해 예측정확도를 더욱 높일 수 있었다.

Abstract The military uses approximately 720,000 kinds of repair parts, accounting for approximately 3% (~ 1.3 trillion won) of the defense budget. The timely supply of repair parts has a direct impact on weapons systems. Therefore, a demand forecast for repair parts has become a very important factor in terms of budget and equipment utilization. To improve the prediction limits with the time series model currently being applied in the Airforce, several studies focused on improving the prediction rates by adding new external variables through supervised learning. Research on improving the predictive accuracy through unsupervised learning and feature engineering has been limited. In this study, the consumption and standard unit price data of 5,752 repair accessory items from 2014 to 2019 were collected for F-15K fighter jets operated by the Airforce. In this study, the data-mining model (Randomforest, GBM, XGBoost, LightGBM, Ridge, Lasso, ElasticNet regression) was used to compare the time series models with the prediction rate. The predictive power scale was RMSE (mean square root error). This study showed that data mining models have higher prediction accuracy than the time series prediction model currently applied in the Airforce. In addition, dimensionality reduction and feature engineering improve the prediction accuracy further.

Keywords : Data Mining, Demand Forecasting, Unsupervised Learning, Feature Engineering, F-15K

*Corresponding Author : Jungmok Ma(Korea National Defense Univ.)

email: jxm1023@gmail.com

Received April 16, 2021

Accepted August 5, 2021

Revised May 11, 2021

Published August 31, 2021

1. 서론

1.1 문제 제기

현재 군에서 운영되는 수리부속은 다음과 같은 특징으로 수리부속 예측에 어려움을 겪고 있다. 첫째, 전군 수리부속은 약 72만여 종에 이르며 다양한 종류와 크기를 가지고 있는 다품종 군수품이다. 둘째, 소요발생이 일정하지 않고 불규칙적이라 수요예측 분석이 어렵다. 끝으로, 수리부속의 원가산정이 지연되거나, 생산업체와의 계약 지연이 빈번하게 발생되어 해외 도입장비는 부품이 단종되는 경우와 같이 확보의 어려움이 발생하기 때문에 조달기간이 장기간 소요된다.

특히 항공기 수리부속은 수만 종으로 품목이 많아 공급망(Supply Chain)이 길고, 수요예측이 정확하지 않아, 공급망의 불확실성으로 인해 비효율성이 존재하며, 정확한 수요예측이 이루어지지 않을 경우, 물류체계의 비효율성이 증가하게 되는데 수요예측 미흡으로 재고가 고갈되면 긴급하게 수리부속을 조달해야 하고, 이로 인한 조달비용 증가와 신규 발주에 따른 조달 기간이 증가하게 된다. 이 경우 재고 고갈에 대비하여 더 많은 재고를 보유해야 하고, 재고관리 목표도 증가하여 재고와 재고비용 증가를 초래한다[1]. 이로 인해 최근 2020년 국정감사에서 항공기의 장기미사용 수리부속을 지적받기도 하였다[2].

Table 1. Airforce demand forecasting models

Methodology	Optimum forecasting
Weighted moving average method	Minimum absolute deviation
Linear moving average method	
Simple exponential smoothing	
Linear exponential smoothing	
Least squares method	

현재 공군은 Table 1과 같이 5가지 시계열 예측기법을 사용하여 연간 수요를 예측하고 있다. 5가지 시계열 예측 기법은 데이터의 추세변동(Trend variation)과 불규칙 변동(Irregular variation)을 반영하며, 각 시계열 기법의 장·단점은 다음과 같다.

첫째, 가중이동평균법은 일정한 추세변동이 있고, 수요변동이 적은 데이터에 적합하지만 수요변동이 큰 경우에는 적합하지 않다. 둘째, 선형이동평균법은 가중이동평균법 보다 추세변동이 큰 데이터에 적합하지만 수요변동이 큰 경우에는 적합하지 않다. 셋째, 단순지수평활법은 불규칙적인 변동이 있는 데이터에 적합하지만 추세변

동이 큰 데이터에는 적합하지 않다. 넷째, 선형지수 평활법은 단순지수평활법 보다 추세변동이 큰 데이터에 적합하다. 다섯째, 최소자승법은 추세변동이 일정한 경우에 적합하다.

5가지 시계열 예측 기법이 데이터의 특성의 일부 반영하여 수요를 예측하고 있으나, 예측오차가 최소인 예측 기법을 선택하여 연간 수요를 선정하기 때문에 예측오차가 커도 5가지 시계열 예측기법 내에서 선택해야 하는 제한이 있다[3].

구체적인 방법은 Fig. 1과 같이 과거 17개 분기의 소모실적 자료를 사용하여 5가지 기법으로 최근 3개 분기의 수요를 예측 한 후, 3개 분기별로 각 기법의 예측치와 실제 소모량과의 차이가 가장 작은 기법으로 미래 분기의 수요를 예측하는 방법이지만 예측 정확도에 한계를 보이고 있다.

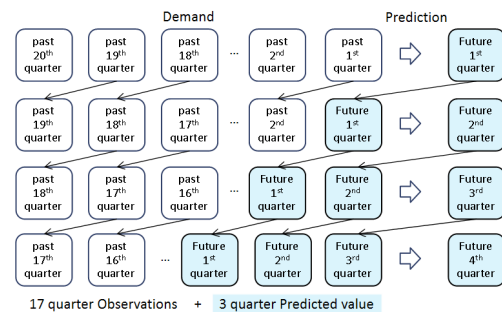


Fig. 1. Demand forecasting process in airforce

1.2 기존연구 고찰

1.2.1 군 분야 시계열 기법 수요예측

군에서는 수리부속 정확도 제고를 위해 많은 연구를 수행해 왔다. 군 분야에서 시계열 기법을 적용한 수리부속 수요예측에 대한 선행연구는 다음과 같다. K. Y. Lee는 차량 수리부속의 5년간 자료를 이용하여 품목별 최적기법을 선정하여 수요예측하였으며, Y. B. Ko는 해군의 함정 엔진의 수리부속 자료를 바탕으로 ARIMA 기법을 활용하여 수요예측하였다. G. H. Nam은 최근린법을 이용하여 A무기체계의 6년간 자료를 수요량 변화와 고·저단가 품목의 특성을 반영하여 수리부속 수요를 예측하였으며, S. R. Yoon은 K-9 자주포의 6년간 자료를 바탕으로 월별 수요량 분포추적을 활용한 수요예측을 진행하였다[4].

1.2.2 군 분야 데이터마이닝 수요예측

군 분야에서 데이터마이닝을 적용한 수리부속 수요예

측에 대한 선행연구는 다음과 같다.

J. D. Kim은 공군 F-XX 항공기의 수리부속 수요예측을 위해 수리부속의 소모율, 소모개수 등을 변수로 하여 데이터 마이닝을 활용하였다[5]. J. S. Kim은 공군 B 항공기의 수리부속 수요를 장단기 기억 메모리(LSTM: Long-Short Term Memory, 이하 LSTM)를 활용하여 예측하였다[6]. M. K. Park은 공군 KF-16C의 수리부속을 주기성을 갖는 수리부속과 간헐적 수요를 갖는 수리부속으로 구분하여 수요예측 방법을 제안하였다. 이 때, 공군에서 사용하는 5가지 시계열 기법을 활용하는 다수결 투표 및 hybrid 방법을 제시하였다[7]. B. H. Oh는 함정 수리부속의 간헐적 수요예측을 위해 순환 신경망(RNN: Recurrent Neural Network), LSTM, 다층 퍼셉트론(MP: MultiLayer Perceptron), 게이트된 순환 유닛(GRU: Gated Recurrent Unit)과 같은 딥러닝을 활용하였다[8]. T. G. Kim은 공군 E-737 수리부속 중 간헐적 수요를 갖는 수리부속에 대해 수리부속 소모개수와 더불어 비행시간/횡수당 소모율, 분기별 온도/습도의 외부 요인을 고려하여 데이터 마이닝으로 수요예측을 수행하였다. 또한, 4년간의 자료를 1년부터 4년까지의 기간으로 세분화하여 시계열모형과 데이터 마이닝 모형의 예측 정확도를 비교하여 제시하였다[9]. H. T. Kim은 육군의 ATCMS, K-9 자주포, ARTHUR-K에 대해 수리부속 소모개수와 표준단가를 설명변수로 선정하여 데이터 마이닝으로 수요예측을 수행하였으며, 군집분석을 통해 예측 정확도를 더욱 높이는 방안을 제안하였다[10]. 최근에는 C. K. Park은 공군 패트리엇 수리부속을 간헐적 수요를 갖는 수리부속으로 구분 후 외부요인인 기온과 장비운영시간을 고려하여 데이터마이닝으로 수요예측을 실시하였다[11].

지금까지 살펴본 데이터마이닝 수요예측 관련 기존 연구들과 수요예측 관련 기존 연구들과 본 연구의 차이점은 다음과 같다.

첫째, Table 2와 같이 대부분 목표변수가 있는 지도 학습 유형의 데이터마이닝 기법을 활용하였으며, 비지도 학습 기법과 연계된 수리부속 예측 연구가 많지 않은 실정이다.

둘째, 군에서 사용하는 실제 데이터를 사용함에 있어 Table 3과 같이 피쳐 엔지니어링(Feature engineering)을 사용한 예측률 향상 연구가 없고, 일부 연구에서 변수 간 스케일(scale) 조정을 위한 정규화만 진행하였다. 따라서 다양한 스케일 조정과 데이터의 로그, 제곱근 변환을 통한 예측률을 향상시키는 연구는 없는 실정이다.

따라서 본 연구에서는 기존 연구들과의 차이점을 바탕으로 수요예측 정확도 제고를 위해 다음과 같은 연구를 수행하고자 한다.

첫째, 지금까지 연구되지 않았던 공군 F-15K를 대상으로 실제 공군 군수사에서 비용분석에 사용되는 14 ~ 19년도 6년간의 분기별 소모개수와 표준단가 수리부속 자료를 수집하여 데이터 마이닝 예측모형과 기존 시계열 모형의 예측값을 비교 분석한다. 둘째, 비교 분석 후에 비지도 학습 기법인 주성분분석(Principal Component Analysis, 이하 PCA), 선형판별분석(Linear Discriminant Analysis, 이하 LDA)을 적용하여 예측정확도에 어떠한 차이가 있는지 확인한다. 셋째, 차원축소 후 피쳐 엔지니어링을 적용하여 예측정확도에 어떠한 차이가 있는지 확인한다.

Table 2. Relevant prior research for military data mining

dissertation author	Subject	Unsupervised learning (Methodology)	Feature engineering (Type)
J. D. Kim [5]	F-XX (Airforce)	no	no
J. S. Kim [6]	B aircraft (Airforce)	no	yes (Minmax scaling)
M. K. Park [7]	KF-16C (Airforce)	no	no
B. H. Oh [8]	Vessels (Navy)	no	no
T. G. Kim [9]	E-737 (Airforce)	no	no
H. T. Kim [10]	ATCMS, K-9, ARTHUR-K (Army)	yes (k-means)	yes (Standard scaling)
C. K. Kim [11]	Patriot (Airforce)	no	no
This paper	F-15K (Airforce)	yes (PCA, LCA)	yes (Feature engineering)

Table 3. Feature engineering type

Scaling type	Scaling characteristics
Robustscaler	Use medians and quaitles instead of means and variances
Minmaxscaler	All attributes are between 0 and 1
Sqrt conversion	Transforming the distribution of data using the sqrt function
Log conversion	Transforming the distribution of data using the log function

2. 본론

2.1 자료수집 및 모형구축

연구절차는 Fig. 2와 같이 진행하였다.

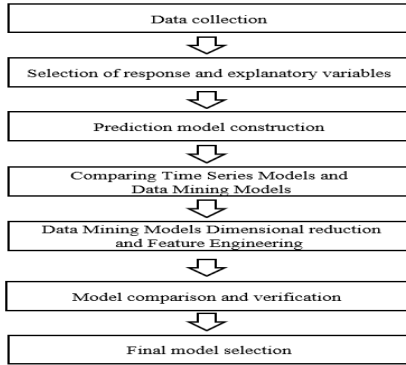


Fig. 2. Research procedure

수집자료는 Table 4와 같이 공군 군수사 분석평가실에서 관리 중인 14 ~ 19년도 F-15K의 수리부속 실제 자료를 수집하였으며, 수요대상은 국방부 성과지원 성과관리 훈령에 따라 5년간 수요가 1회 이상 발생한 5,752개 품목으로 한정하였다.

Table 4. Data collection

Category	Content
Subject	F-15K
Collection period	2014 ~ 2019
Number of parts	5,752

비교 데이터는 정확한 비교를 위해 Table 5와 같이 14 ~ 16년, 14 ~ 17년, 14 ~ 18년으로 3개의 데이터 세트로 구분하였다.

변수선정에서 설명변수는 각 데이터 세트의 분기별 소모개수를 사용하였다. 종속변수는 17 ~ 19년도 1분기 소모개수를 사용하였다. 데이터 마이닝 모형은 표준단가를 추가하였다.

Table 5. Separating data sets

Explanatory variable		Dependent variable
1	Q1401 ~ Q1604	Q171 (1Q of 2017)
2	Q1401 ~ Q1704	Q181 (1Q of 2018)
3	Q1401 ~ Q1804	Q191 (1Q of 2019)

* Data mining : Quarter data + Standard unit price

시계열 모형 구축은 현재 공군에서 사용하는 5가지 시계열 예측기법을 사용하여 구축하였다. 시계열 모형은 평균(Average)과 평활(smoothing)을 기반으로 하는 단순한 모형이라 Table 6과 같이 엑셀 함수와 데이터 분석 기능을 사용하여 모형을 구축하였다.

Table 6. Time series models and excel functions

Methodology	Excel Functions
Weighted moving average	Sumproduct
Linear moving average	Stat data analysis tool
Simple exponential smoothing	Stat data analysis tool
Linear exponential smoothing	Stat data analysis tool
Least squares method	Linest

첫 번째 기법으로 가중이동평균법을 사용하였다. 수요 변동이 적은 경우 적용하며, 최근 분기수요에 높은 가중치를 부여하여 수요량을 예측하는 방법이다. 공군은 0.2, 0.3, 0.5로 가중치를 부여하여 사용하고 있으며, Eq. (1)과 같이 정의된다.

$$F_{t+i} = A_t \times w_1 + A_{t-1} \times w_2 + \dots + A_{t-i+1} \times w_i \quad (1)$$

Where, F_{t+i} denotes predicted value at t+i, A_t denotes real value at t, w_i denotes weight at I

두 번째 기법으로 선형이동평균법을 사용하였다. 이동평균법(Moving average)에 의한 예측결과가 추세 요인으로 인한 과대/과소가 예측될 경우 적용하며, 이동평균 결과를 다시 이동평균하여 추세 요인을 고려하여 미래수요를 예측하는 방법이다. Eq. (2)와 같이 정의된다.

$$F_{t+i} = MA + (MA - MA') = 2MA - MA' \quad (2)$$

Where, F_{t+i} denotes predicted value at t+i, MA denotes first moving average value, MA' denotes second moving average value

세 번째 기법으로 단순지수평활법을 사용하였다. 불규칙적인 수요변동의 경우 적용하며, 관측치에는 평활상수(Smoothing constant) 0.1부터 0.9까지 0.1씩 간격으로 α 값을 사용하여 수요값을 예측한다. 예측오차의 평균 제곱합을 최소화 하는 α 값을 선택하며, Eq. (3)과 같이 정의된다.

$$F_{t+i} = \alpha A_t + (1 - \alpha)F_t, \quad 0 \leq \alpha \leq 1 \quad (3)$$

Where, F_{t+i} denotes predicted value at t+i, A_t

denotes real value at t, α denotes smoothing constant

네 번째 기법으로 선형지수평활법을 사용하였다. 추세를 갖는 불규칙적인 수요변동의 경우 적용하며, 지수평활한 결과(1차 예측)와 이중 지수평활한 결과(2차예측)를 활용하는 방법으로 Eq. (4)와 같이 정의된다.

$$F_{t+i} = F_t + (F_t - F'_t) = 2F_t - F'_t \quad (4)$$

Where, F_{t+i} denotes predicted value at t+i, F_t denotes exponential smoothing value, F'_t denotes double exponential smoothing value

다섯 번째 기법으로 최소자승법을 사용하였다. 시계열 자료가 일정한 추세를 가지고 있을 경우 적용하는 방법으로 관측치와 예측치의 편차제곱합이 최소가 되는 추세를 찾아 예측을 수행하며 Eq. (5)와 같이 정의된다.

$$Y = \beta_0 + \beta_1 X \quad (5)$$

Where, β_0 denotes y intercept, β_1 denotes slope

Table 7. Data mining models

Assortment	Methodology
Ensemble	Random forest
	GBM
	XGBoost
	LightGBM
Regularization	Ridge regression
	Lasso regression
	Elastic Net regression

데이터 마이닝 모형은 Table 7과 같이 종속변수가 연속형이기 때문에 회귀분석 시 예측률이 뛰어나며, 다중공선성(Multicollinearity)과 과적합(Overfitting)에 강한 앙상블(Ensemble) 모형과 규제(Regularization)모형을 파이썬을 사용하여 구축하였으며, 기본 파라미터 값을 적용하였다.

2.2 검증 / 평가 방법

시계열 모형과 데이터마이닝 모형 비교 시 정확한 검증을 위해 시계열 데이터 세트를 3개로 나누었으며, 3개 데이터 세트의 RMSE값 평균을 최종적으로 비교하였다.

평가방법은 회귀지표의 대표적인 척도인 평균제곱근 오차(Root Mean Squared Error, 이하 RMSE)을 사용

하였다. RMSE값은 관측값과 예측값의 차이이므로 적을 수록 성능이 우수하며, 수식은 Eq. (6)과 같이 정의된다.

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (Y_t - \hat{Y}_t)^2}{n}} \quad (6)$$

비교 이후에는 데이터마이닝 모형을 차원축소와 피쳐 엔지니어링을 실시하여 전과 후의 예측값을 비교하였고, 예측값 비교 시 정확한 검증을 위해 10-fold(k-fold cross validation, 이하 10-fold)검증을 실시하였다.

시계열 모형과 데이터마이닝 모형 구축 시 사용 환경은 Table 8과 같다.

Table 8. Usage environment for model deployment

Category	Value
CPU	Intel(R) core(TM) i7-1065G7
Memory	16GB
OS	Window 10 64bit
S/W	Data mining modeling : Phythion(ver. 3.9.2) Time series modeling : Excel(ver 2020)

2.3 결과분석

2.3.1 모형 비교

시계열 모형은 Table 9와 같이 5가지 모형 모두 RMSE값이 높아 예측률이 낮았다.

Table 9. Prediction results by model

Assortment	Methodology	RMSE			
		Q1701	Q1801	Q1901	Average
Time series models	Weighted moving average method	52.85	48.63	55.47	52.31
	Linear moving average method	58.83	54.07	63.27	58.72
	Simple exponential smoothing	49.10	46.10	52.70	49.3
	Linear exponential smoothing	53.43	50.49	58.67	54.19
	Least squares method	69.85	63.69	67.66	67.06
	Random forest	31.61	29.78	36.09	32.49
Data mining models	GBM	29.36	29.53	41.74	33.54
	XGBoost	38.12	37.73	50.43	42.09
	LightGBM	31.50	28.32	38.61	32.81
	Ridge regression	40.22	71.30	71.30	60.94
	Lasso regression	33.26	57.58	57.58	49.47
	Elastic Net regression	33.29	57.67	57.67	49.54

불규칙 변동(Irregular variation)에 적합한 단순지수 평활법($\alpha=0.6$)의 RMSE값이 49.3으로 상대적으로 예측률이 높았다. 반면 추세변동(Trend variation)을 반영하기 위해 이중평균을 하는 선형이동평균법과 이중평활하는 선형이동평균법, 그리고 최소가 되는 추세를 찾기 위해 기울기와 절편을 사용하는 최소자승법 모두 예측률이 높지 않았다. 이는 데이터의 전반적인 형태가 불규칙변동이 크고 추세변동이 적다고 추정할 수 있다.

데이터마이닝 모형은 과적합을 방지하기 위해 10-fold 검증을 실시 하였다. 데이터마이닝 모형은 기본 파라미터값을 적용했음에도 불구하고, Table 9와 같이 4개의 앙상블 모형 모두 시계열 모형 보다 예측률이 우수하였다. 이는 앙상블 모형이 불규칙 변동이 큰 자료 예측에 적합하다고 추정할 수 있다.

3개의 규제모형은 단순지수평활법을 제외한 나머지 시계열 모형보다는 RMSE값이 낮아 예측률이 우수하였지만 앙상블 모형들보다는 예측률이 모두 낮았다. 이는 규제모형 모두 최소자승법 기반 모형이기 때문에 최소자승법의 영향을 받았기 때문으로 추정할 수 있다. 최소자승법의 RMSE값은 67.06으로 전체 모형 중에서 예측률이 가장 낮았다.

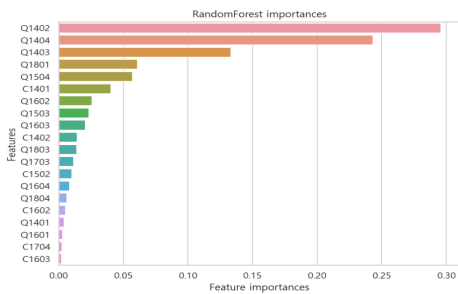


Fig. 3. Feature importances

세부적으로 랜덤포레스트 모형이 RMSE값이 32.63으로 예측률이 가장 높았다. Fig. 3과 같이 랜덤포레스트 모형(Q1901)의 설명변수들의 중요도(상위 3)를 확인해 보면 Q1402 (14년 2분기 소모) > Q1404 (14년 4분기 소모) > Q1403 (14년 3분기 소모)로서 소요 수량 예측에 있어 과거 자료가 최근자료 보다 예측에 영향을 많이 미친다는 것을 알 수 있었다.

2.3.2 차원축소 / 피쳐 엔지니어링 전·후 비교

결과적으로 각 모형에서 가장 우수한 모형의 RMSE값 차이는 16.67로 예측률에 많은 차이가 나며, 이는 예측

률에 있어 데이터마이닝 모형이 시계열 모형보다 우수함을 보여주었다. 특히 데이터 형태가 불규칙 변동이 클 경우 시계열 모형 보다는 데이터마이닝 모형의 앙상블 모형이 더 적합함을 알 수 있었다.

이어서 본 연구에서는 수리부속 예측정확도를 높이는 방법을 찾기 위해 데이터마이닝 모형에 차원축소를 실시 하였다. 설명변수가 40개로 가장 많은 데이터 세트 3번(Q1901) 자료를 사용하였으며, 정확한 검증을 위해 10-fold 검증을 실시 후 RMSE값을 비교하였다.

적용 모형은 차원축소하는데 대표적으로 사용되는 PCA와 LDA모형을 사용하였다. 주성분 수 선정방법은 Fig. 4와 같이 variance_ratio함수를 사용 하여 적용하였다.

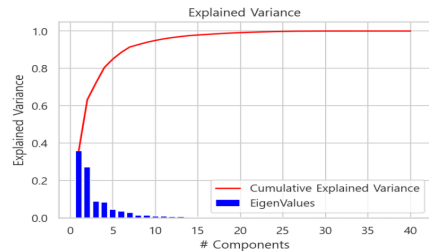


Fig. 4. PCA explained variance

적용 결과 Table 10과 같이 7개 데이터마이닝 모형 모두 차원축소 후 RMSE값이 감소하였다. 이는 차원축소가 예측률 향상에 도움이 됨을 보여준다. 세부적으로 Ridge모형에서 RMSE값이 최대 46.62 감소하였다. 차원축소 전에는 랜덤포레스트 모형이 예측률이 가장 우수하였으나, 차원축소 후에는 Ridge모형이 가장 예측률이 우수하였다. PCA모형은 RMSE값이 평균 22.93, LDA모형은 평균 14.49 감소하였다. 7개 모형 모두 LDA보다 PCA모형 적용 시 예측률이 더 향상되었다.

Table 10. Dimensionality reduction comparison results

Methodology	Before	RMSE	
		After (Gap)	
		PCA	LDA
Random forest	36.09	28.33(-7.76)	31.25(-4.84)
GBM	41.74	26.73(-15.01)	29.40(-12.34)
XGBoost	50.43	27.39(-23.04)	33.11(-17.32)
LightGBM	38.61	32.53(-6.08)	35.87(-2.74)
Ridge regression	71.30	24.68(-46.62)	41.93(-29.37)
Lasso regression	57.58	25.52(-32.06)	40.60(-16.98)
ElasticNet regression	57.67	27.70(-29.97)	39.80(-17.87)

예측률을 더 높이기 위해 Table 11과 같이 설명변수 간의 다중공선성을 제거하여 PCA모형으로 다시 차원축소를 실시하였다. variance_inflation_factor함수를 사용하였으며, 분산팽창지수(VIF: Variance inflation factor, 이하 VIF) 10 이상의 설명변수를 제거한 결과는 Table 12와 같다. 설명변수가 40개에서 11개로 감소하였다.

적용결과 Table 11과 같이 7개 데이터마이닝 모형 모두 설명변수간의 다중공선성을 제거 후 차원축소 시 RMSE값이 감소하였다. 이는 설명변수간의 다중공선성을 제거하는 것이 예측률 향상에 도움이 됨을 보여준다. 세부적으로 Lasso모형에서 RMSE값이 최대 6 감소하였다. 다중공선성 제거 전에는 Ridge모형이 예측률이 가장 우수하였으나, 제거 후에는 Lasso모형이 가장 예측률이 우수하였다.

Table 11. Dimensionality reduction re-comparison results

Methodology	PCA RMSE	
	Before	After (Gap)
Random forest	28.33	26.33(-2)
GBM	26.73	24.68(-2.05)
XGBoost	27.39	23.52(-3.87)
LightGBM	32.53	28.95(-3.58)
Ridge regression	24.68	20.67(-4.01)
Lasso regression	25.52	19.52(-6)
ElasticNet regression	27.70	22.13(-5.57)

예측률을 한층 더 높이기 위해 Table 13과 같이 4항목의 피쳐 엔지니어링을 적용하였다. 피쳐 엔지니어링은 일부 왜곡된 분포를 정규분포에 가깝게 만들어 주거나 종속변수와 설명변수와의 관계를 비선형관계에서 선형관계에 가깝게 데이터를 변환 해주기 때문에 예측률 향상에 도움이 될 수 있다.

Table 12. Variancevariance inflation factor results

Explanatory variable		VIF
1	Q1403	4.1570
2	Q1402	3.3709
3	Q1702	2.8783
4	Q1704	2.8783
5	Q1701	2.8192
6	Q1703	2.8027
7	Q1401	2.7929
8	Q1501	2.1381
9	Q1802	2.1209
10	Q1601	2.1018
11	C1704	1

적용결과 설명변수간의 다중공선성을 제거 후 차원축소 한 결과보다 일부 모형들은 RMSE값이 감소하였지만, 7개 모형 모두 감소 되지는 않았다. 하지만 각 모형에 적합한 피쳐 엔지니어링 적용 시 RMSE값이 감소 되었다. 규제모형에 Roubuster 스케일 적용 시 모두 RMSE값이 감소하였으며, 앙상블모형에 제곱근변환 적용 시 모두 RMSE값이 감소하였다. 피쳐 엔지니어링 후에는 Lasso와 ElasticNet모형이 가장 예측률이 우수하였다.

Table 13. Feature engineering comparison results

Methodology	Feature engineering RMSE				
	Before	After (Gap)			
		Roubuster scale	Minmax scale	Sqrt	Log
Random forest	26.33	24.80	28.69	22.00	28.66
GBM	24.68	22.85	28.75	19.93	33.78
XGBoost	23.52	24.43	25.37	21.15	48.30
LightGBM	28.95	30.64	34.18	26.17	28.08
Ridge regression	20.67	19.30	20.54	26.20	32.86
Lasso regression	19.52	19.29	34.82	25.95	32.08
ElasticNet regression	22.13	19.29	34.87	24.97	30.94

결과적으로 본 연구를 통해 차원축소와 피쳐 엔지니어링 모두 실제 항공기 데이터 자료에서 예측률 향상에 도움이 됨을 알 수 있었다.

3. 결론

본 연구는 수리부속 수요예측 정확도 제고를 위해 현재 공군에서 운용중인 F15K 14년 ~ 19년까지의 5,752개 수리부속 데이터 수집을 통해 시계열 모형과 데이터마이닝 모형을 비교 분석하였다.

연구결과 현재 공군에서 운용중인 시계열 모형에 비해 소모개수, 표준단가 등 다양한 변수를 활용하는 데이터마이닝 모형의 성능이 더욱 우수함을 알 수 있었다. 그 중에서도 예측함에 있어서 Lasso와 ElasticNet 모형의 성능이 가장 우수하였다. 이어서 더욱 예측 정확도를 높이기 위한 노력으로 차원축소와 피쳐 엔지니어링을 진행했으며, 전과 비교해서 예측정확도가 더욱 향상됨을 알 수 있었다.

따라서 실제 군 자료 형태의 다양성을 고려한다면 앞으로 우리 군이 수리부속 수요예측에 시계열모형 뿐만 아니라 데이터마이닝 모형도 활용하되 비지도학습의 차원축소와 피쳐 엔지니어링도 같이 활용한다면, 예측정확도

향상으로 경제적인 군 운영에 크게 기여할 것으로 판단된다.

끝으로, 본 연구에서는 현재 공군에서 운용중인 F-15K 장비에 대해서만 연구를 진행했기 때문에 공군 전체 수리 부속에 적용하기에는 한계가 있다고 판단되며, 향후에는 추가 변수와 다른 비지도학습 기법과 피쳐 엔지니어링 기법을 적용하여 다양성과 신뢰성을 높일 필요가 있다.

References

- [1] J. S. Kim, J. S. Hwang, J.W.Jung "A New LSTM Method Using Data Decomposition of Time Series for Forecasting the Demand of Aircraft Spare Parts", *korean Management Science Review*, , Vol.37, No.2, pp.2, June 2020
DOI : <https://doi.org/10.7737/KMSR.2020.37.2.001>
- [2] S. C. Park, "Military, 300 billion won worth of unused repair accessories in the warehouse"[Internet], World Daily, c2020 [cited 2020 Oct. 15], Available From: <http://www.segye.com/newsView/20201015522160> (accessed Oct. 15 2020)
- [3] Airforce Headquarters, Maintenance of Repair parts, p.69, Airforce Headquarters, 2019, pp.12.
- [4] H. T. Kim, S. H. Kim, "Data mining based army -repair parts demand forecast", *journal of the Korean Data & information Science Society*, Vol.30, No.2, pp.430-431, March 2019
DOI : <https://doi.org/10.7465/jkdi.2019.30.2.429>
- [5] J. D. Kim, H. J. Lee "A Study on Forecasting Spare Parts Demand based on Data-Mining", *journal of the internet Computing and Services*, Vol.18, No.1, pp.121-129, February 2017
DOI : <https://dx.doi.org/10.7472/jksii.2017.18.1.121>
- [6] J. S. Kim, J. S. Hwang, J.W.Jung "A New LSTM Method Using Data Decomposition of Time Series for Forecasting the Demand of Aircraft Spare Parts", *korean Management Science Review*, , Vol.37, No.2, pp.1-18, June 2020
DOI : <https://doi.org/10.7737/KMSR.2020.37.2.001>
- [7] M. K. Park, J. G. Baek, "Demand Forecast of Spare Parts for Low Consumption with Unclear Pattern", *journal of the KIMST*, , Vol.21, No.4, pp.529-540, August 2018
DOI : <https://doi.org/10.9766/KIMST.2018.21.4.529>
- [8] B. H. Oh, *A study on weapon system spare parts intermittent demand forecasting using deep learning*, Master's thesis, Korea University of information and Communication, Seoul, Korea, pp.2-10, 2017
- [9] T. G. Kim, J. M. Ma, "A Data Mining Approach for intermittent Demand Forecasting of Aircraft Spare Parts - Focusing on the E-737(AEW&C: Airborne Early Warning & Control) Spare Parts -" *journal of the AMSOK*, , Vol.16, No.4, pp.155-164, August 2018

DOI : <https://doi.org/10.30529/amsok.2018.16.4.008>

- [10] H. T. Kim, S. H. Kim, "Data mining based army repair parts demand forecast", *journal of the Korean Data & information Science Society*, Vol.30, No.2, pp.429-444, March 2019
DOI : <https://doi.org/10.7465/jkdi.2019.30.2.429>
- [11] C. K. Park, J. M. Ma "A Study on Intermittent Demand Forecasting of Patriot Spare Parts Using Data Mining", *Korea Academy Industrial Cooperation Society*, Vol.22, No.3, pp.234-241, February 2021
DOI : <http://doi.org/10.5762/KAIS.2021.22.3.234>

송 민 규(Min-Gyu Song)

[준회원]



- 2010년 2월 : 한국성서대학교 성서학과 (성서학 학사)
- 2020년 2월 ~ 현재 : 국방대학교 국방과학학과 (석사과정)

<관심분야>

수요예측, 데이터마이닝

마 정 목(Jung-Mok Ma)

[정회원]



- 2002년 2월 : 육군사관학교 운영분석학과 (운영분석 석사)
- 2008년 8월 : 미국 펜실베이니아 주립대(PSU) (산업공학 석사)
- 2015년 5월 : 미국 일리노이대(UIUC) (산업공학 박사)
- 2015년 9월 ~ 현재 : 국방대학교 국방과학학과 부교수

<관심분야>

국방 모델링 및 데이터 분석학, 무기체계 획득관리