머신러닝을 적용한 해군 정비지원부대 퇴직자 예측 모델

유정민O
O국방대학교, 해군 군참부
e-mail: code_no1@naver.comO

Retirement Prediction Model for ROK Navy's Maintenance Support Unit Based on Machine Learning

Jun-Min YooO
OKorea National Defense University, Dept. of Logistic Branch, ROK Navy

● 요 약 ●

국방 무기체계의 운용유지를 위해서는 숙련자에 의한 신뢰성있는 정비 지원이 필요하다. 특히, 고도의 기술력을 바탕으로 연구/제작된 해군 무기체계를 유지하기 위해서는 이와같은 정비 지원이 무엇보다 중요하다. 해군에서는 효과적인 정비지원을 위해 수개의 정비지원부대를 조직하여 운용하고 있다. 원활한 정비지원부대의 운용을 위해 다년간 기술력을 축적한 정비인원의 중도 이탈을 예방하는 것이 요구되므로, 본 논문에서는 머신러닝을 적용하여 해군 정비지원부대의 퇴직자 예측 모델을 제안하였다. 정비인력의 만족도와 관계가 있을 것으로 예상되는 봉급, 특근율 등을 변수로 사용하였고, F1 Score를 통해 모델의 성능을 평가한 결과 0.7 이상의 높은 성능을 보였다. 이 모델을 통해 조기 퇴직이 예상되는 그룹의 공통 개선소요를 파악하여 사전 조치가 가능할 것으로 판단하였다.

키워드: 머신러닝(Machine Learning)

I. Introduction

무기체계의 정비는 작전 준비도 유지, 수명 연장 및 운용비용 절감, 무기체계 자체의 신뢰도 향상 등 군사 성패를 가를 수 있는 매우 중요한 행위이다. [1] 무기체계 정비를 위해 갖추어야 할 요소에는 조직, 시설, 자재 등 여러 가지가 있겠지만 가장 큰 비중을 차지하는 요소로 정비인력을 꼽을 수 있다. 특정분야 무기체계에 대한 숙련된 정비할 수 있는 인력은 양성 또는 채용을 통해 단기간에 획득할 수 있는 것이 아니다. 다년간의 노력을 통해 경험과 기술의 축적이 숙련된 정비를 위해 요구되는 것이다. 따라서 조직은 이렇게 장기간에 걸쳐 획득된 숙련된 정비인력을 장기간 운용하기 위해 노력할 필요가 있다. 본 연구는 정비조직의 인력 이탈, 즉 퇴직을 예측하는 모델 제안을 통해 이러한 부분에 기여하고자 한다. 어떠한 요소가 퇴직에 영향을 비교적 많이 주고 있는지 식별할 수 있다면, 조직은 그 분야에 대한 개선을 통해 보다 장기적인 관점에서 인력운용이 가능할 것이다.

II. Preliminaries

1. Maintenance Unit of ROK Navy

해군에서 수행하는 정비는 수행 주체에 따라 여러 종류로 분류할 수 있다. 그 중 본 연구에는 전문적인 정비조직인 정비지원부대에서 수행하는 야전정비 이상을 대상으로 하였다. 이러한 야전정비에 대하여 해군은 가능과 목적에 따라 여러 정비지원부대를 운용하고 있는데, 이 정비지원부대 대부분은 군무원 중심의 조직으로, 본 연구에서 다루는 퇴직자 역시 군무원을 대상으로 하였다. 다른 조건이 동일하다고 가정할 때, 해당 조직에서 정비인력이 조기에 이직하는 이유로는 근무 만족도를 거론할 수 있다. 이러한 근무 만족도에는 봉급, 특근율, 근무연수에 따른 대우, 특정 부서의 근무여건 등이 있을 수 있다.

2. Related Works

2.1 Machine Learning

아서 사무엘(Arthur Samuel)은 1959년 게제된 논문에서 머신러닝을 컴퓨터가 명시적으로 프로그래밍되지 않고 스스로 학습하도록 하는 능력으로 정의하였다. [2] 이처럼 머신러닝은 특정 데이터에 대해 스스로 학습하고 개선하는 능력을 갖춘 인공지능의 한 분야이다. 이는 대게 지도학습, 비지도학습, 강화학습으로 구분되는데, 본 연구에서는 이진분류, 즉 퇴직이냐 아니냐를 분류하기 위한 예측모델 학습을 위해 지도학습 방식을 사용하였다. [3] 이를 위해 다음 언급하는 랜덤포레스트와 F1 스코어를 이용하였다.

2.2 RandomForest

랜덤포레스트는 의사결정나무(Decision Tree)를 기반으로하는 앙상블(Ensemble) 기법 중 하나로, 머신러닝에 쓰이는 알고리즘이다. 이 알고리즘은 분류와 회귀에 모두 사용되며, 과적합을 줄이는데 사용되는 것으로 알려져 있다. [4] 또한, 특정 변수가 모델 성능에 미치는 영향을 나타내는 특성 중요도(Feature Importance)를 사용할 수 있는데, 특성 중요도의 값이 0.1이상일 때 해당 변수가 모델 예측에 어느정도 기여하는 것으로 해석할 수 있으며, 0.3이상일 때 모델 예측에 매우 중요하게 작용하는 변수로 해석할 수 있다. 본 연구에서는 이를 통해 퇴직에 미치는 영향을 분야별 상대비교하였다.

2.3 F1 Score

F1 스코어는 분류 모델의 성능 평가 지표중의 하나로, 정밀도(precision)와 재현율(recall)의 조화평균을 의미한다. 따라서, 정밀도와 재현율이 하나라도 낮을때에는 상대적으로 낮은 값을 나타내기에 상대적으로 안정적인 성능 지표로 알려져있다. [5] F1 Score는 0에서부터 1사이의 값으로 표현되며, 연구의 목적과 데이터의 상황에 따라 다를 수 있으나, 일반적으로는 0.4 미만일 때 매우 낮은 성능으로, 0.4에서 0.5까지는 상대적으로 낮은 성능으로, 0.5에서 0.6까지는 일반적인 수준의 성능을 나타내며, 0.6이상일 때 상당히 높은 수준의 성능을 나타내는 것으로 해석할 수 있다.

III. The Proposed Scheme

1. Tools and Dataset

본 연구에서는 데이터 수집 및 전처리, 머신러닝을 위해 프로그래밍 언어 파이썬을 사용하였고, pandas, sklearn, matplotlib 등의 머신러닝 라이브러리를 이용하였다.

연구에 사용한 데이터는 해군의 특정 1개 정비지원부대의 정비인력 현황을 사용하였다. 다만, 보안상의 이유로 특정정보가 노출될 수 있는 항목에 대해서는 치환표기로 대체하였다.

구체적인 데이터 획득 범위로, 해당 조직의 운영 또는 발전업무를 하는 부서는 제외하고 실제 정비가 수행되는 정비현장의 인력만을 대상으로 하였으며, 사무직과 현장직을 구분하여 수집한 결과 1,138명의 데이터가 수집되었다. 데이터는 초기부터 특정인을 식별할 수 있는 정보는 제외된 상태로 획득되었다.

퇴직과 관련된 정보는 2021년 이후로 약 2년여간의 데이터가 수집되었으며, 이와 더불어 각 조직별 특근률과 부하율의 자료를 같이 수집하였다. 또한 인사혁신처 공무원보수 규정에 의거 인원별 월급 정보를 같이 수집하였으며, 군무원인사법 시행령에 의거 <표 1>과 같이 직군, 직렬정보를 해당되는 사항에 대해서만 수집하였다.

Table 1. Job Positoin of Maintenance Staff

직군	직렬
행정	행정, 군수
시설	시설
정보통신	전기, 전자, 통신, 사이버
공업	일반기계, 금속, 용접, 물리분석, 화학분석, 유도무기, 탄약, 차량
함정	선체, 선거, 함정기관

2. Preprocessing

분석에 사용할 변수를 준비하기 위해 수집된 데이터들을 통합하여 전처리하였다. 수집된 변수 중 6개 범주형 변수에 대해서는 각각 LableEncoder를 이용하여 수치형으로 변경하여 적용하였으며, 수치형 변수들은 모델의 정확도를 높이기 위해 최대최소 정규화를 실시하였다.

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

이와 같은 과정을 통해 총 24개의 원변수 및 파생변수를 생성하여 적용하였으며 변수의 종류에 따른 세부 내역은 <표 2>와 같다.

Table 2. Variables

구분	변수명
원변수	부서, 과, 팀, 특화직, 군경력, 파견, 직렬, 계급, 근속, 승진연도, 승진종류, 채용구분, 성별, 특근시간, 특근율, 부하율, 나이
파생변수	직종, 직군, 계급차이, 호봉, 봉급, 직렬차이
목표변수	퇴직여부

<표 2>의 변수들 중 부연설명이 필요한 내용은 다음과 같다. 특화직이란 정비업무를 수행하는데 특별한 기술력이 보다 요구되어 별도 관리되는 직위이며, 군경력은 군무원 임용 전 복무여부이다. 파견은 원소속에서 현소속으로 일부 이동근무 중인 인원을 말하며, 승진종류는 일반승진 또는 근속승진을 구분하여 적용하였고, 채용구분은 경채, 공채, 임기제 채용을 구분하여 적용하였다. 직종은 사무직, 현장직을 구분 적용하였고, 계급차이는 해당직위의 편제 대비 현 계급 간 차이를, 직렬차이는 편제 대비 현 직렬 간 차이여부를 반영하였다.

3. Machine Learning

머신러닝 알고리즘으로는 앞서 언급한 바와 같이 랜덤포레스트 알고리즘을 사용하였으며, 성능 평가 지표로는 F1 스코어를 사용하였다. 하이퍼파라미터로 랜덤포레스트의 max_depth는 1~10까지 적용하였으며, 데이터의 70%를 학습용으로, 30%를 검증용으로 사용하되

random_state는 500에서 2000까지 적용하였다.

이 과정에서 특성 중요도(Feature Importance)를 이용하여 차원 축소를 실시하였는데, 과정 중 식별된 변수는 특화직, 군경력, 파견, 성별, 직렬차이, 직종 등 6개로 이 변수들은 특성 중요도 0.01이하의, 상대적으로 모델 성능에 도움이 되지 않는 것으로 판단하였다.

IV. Results

1. Exploratory Data Analysis

먼저 데이터에 대한 이해를 진행하기 위해 탐색적 데이터 분석을 진행하였다. 탐색적 데이터 분석은 통계분석, 상관관계 분석 순으로 진행하였다.

1.1 Statistics Analysis

데이터의 통계적 특성을 확인한 결과, 전 직원의 평균 봉급은 338만원 수준을 기록하였으며, 평균 부하율은 103%로 집계되었다. 이때 부하율의 최대값은 151%, 최소값은 57% 수준으로 차이가 크게 나타났다. 전체 남직원은 1,058명 여직원은 80명 수준으로 성비 또한 큰 차이를 보였다. 퇴직자의 평균 나이는 정년퇴직을 포함한 경우 52세, 정년퇴직을 제외한 경우 48세로 근소한 차이를 보였으며, 본 연구의 목적상 정년퇴직을 한 경우는 분석대상에서 제외하였다. 통계분석의 주요 결과는 <표 3>과 같다.

Table 3. Result of Key Statistical Analysis

구분	봉급	부하율	근속	퇴직자 나이
평균	338만원	103%	17년	48세
최대	607만원	151%	40년	61세
최소	177만원	57%	1년	27세

1.2 Correlation Analysis

특성 중요도와 별개로 데이터 자체에서 퇴직과의 상관관계를 확인하기 위해 상관관계 분석을 수행하였다. 퇴직과 관계된 상관관계는 개별 변수 차원에서는 대체적으로 낮게 나타났으며, 상관계수 0.05이상의 변수들은 계급차이, 근속, 호봉, 봉급, 승진연도, 승진구분, 채용구분 등 7개 변수로 <그림 1>과 같이 Heatmap을 이용하여 나타내었다.

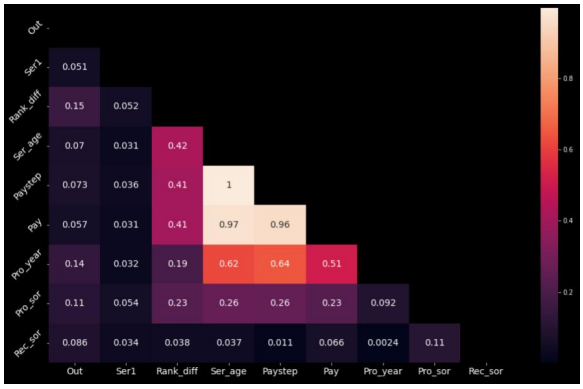

Fig. 1. Heatmap for Correlation

이와는 반대로 소속, 직렬, 계급, 특근율, 부하율, 나이 등의 변수는 퇴직의 상관계수가 낮은 것으로 나타났다. 상관관계가 낮은 변수들의 세부 수치는 생략하였다.

2. Prediction and Evaluation

사용 변수가 비교적 다수인 관계로 랜덤포레스트의 시각화는 생략하였다. 단, 앞서 언급한 바와 같이 데이터의 70%을 사용하여 학습을 수행하고, 나머지 30%에 대한 예측 및 평가를 수행한 결과, F1 스코어는 최소 0.67에서 최대 0.76까지의 변화를 보였고, 대부분 0.70에서 0.72사이의 수치로 확인되어, 상당히 높은 수준의 성능을 보이는 것으로 평가하였다.

모델 성능에 대한 지표들을 정리하여, 최대 성능이 해당하는 경우들을 표현하면 다음과 같다. 전체 데이터 중 맞게 예측한 것의 비율인 정확도는 0.98, 실제 양성인 것을 양성으로 예측한 비율인 정밀도는 0.8, 양성으로 예측한 것중 실제 양성일 비율인 재현율은 0.72, F1 스코어는 0.76을 보였다. 이에 대한 혼동행렬을 시각화한 것은 <그림 2>와 같다.

Fig. 2. Heatmap for Confusionmatrix

3. Feature Importance

변수가 모델의 예측에 얼마나 중요하게 작용하는지 확인하기 위해 특성 중요도를 확인하였다. 특성 중요도는 상위 6개 변수가 다른 변수에 비해 높은 수치를 보였는데 해당 내용은 <그림 3>과 같이 표현하였다.

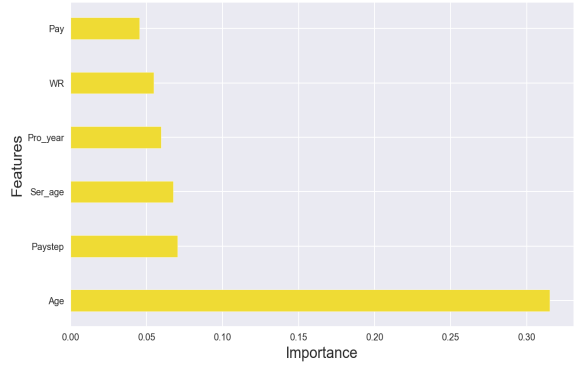

Fig. 3. Feature Importance

위의 그림과 같이 나이가 모델에서 가장 중요한 영향을 미치고 있었으며, 그 뒤로 호봉, 근속, 승진연수, 부하율, 봉급 순으로 영향을 보였다. 특히 나이, 부하율은 상관관계 분석 시에는 낮은 상관계수를 보인 변수들로 나이에 따른 퇴직여부는 선형관계에 있는 것은 아니나, 즉 나이가 많을수록 퇴직확률이 올라가는 것은 아니나, 특정 조건과 결합하여 퇴직 예측에 중요한 작용할 하고 있음을 확인하였다.

V. Conclusions

본 연구에서는 해군 특정 정비지원부대의 정비인력 퇴직을 예측하기 위한 머신러닝 모델을 제안하였다. 연구 결과를 바탕으로, 어떠한 항목이 인력 퇴직에 영향을 많이 주고 있는지 확인할 수 있었으며, 또한 현재의 인원 중 누가 퇴직확률이 높은지 예측할 수 있었다.

모델의 성능은 F1 스코어로 최대 0.76을 기록하여 준수한 성능을 나타내는 것으로 확인하였다. 충분히 활용할 수 있는 수준으로 판단하였다.

향후 보다 현실적인 변수들이 반영된다면 더 정확한 모델을 제안할 수 있을 것으로 예상하였다. 예를 들면, 부서 내 근무만족도 설문조사 결과, 최근 몇 년간 성과연봉 지급현황 등이 그러한 변수가 될 수 있을 것이다.

또한, 본 연구는 최근 2년간의 데이터를 바탕으로 분석을 진행하여 한계점이 있는 바, 보다 다년간의 데이터를 이용하여 보완할 예정이다.

REFERENCES

[1] John Smith, et.al, ""The Importance of Weapon System Maintenance in Ensuring Operational Readiness: A Case Study," International Journal of Military Science and Technology, 2018.

[2] Arthur Samuel, "Some studies in machine learning using the game of checkers," IBM Journal of Research and Development, 1959.

[3] Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," O'Reilly Media, 2019.

[4] Leo Breiman, "Random Forests," Machine Learning, Vol. 45, Issue 1. pp. 5-32, 2001.

[5] C.J. van Rijsbergen, "The relationship between Precision-Recall and ROC curves," Information Processing & Management, 1979.