Advanced Industrial SCIence
Vol. 2, No. 4, pp. 11-16, 2023

ISSN 2951-2476
https://doi.org/10.23153/AI-Science.2023.2.4.011

머신러닝을 이용한 지하철 고장 탐지 및 예측

성국경
한국영상대학교 방송영상미디어과 교수

Detection and Prediction of Subway Failure using Machine Learning

Kuk-Kyung Sung
Professor, Department of Broadcasting and Emerging Media, Korea University of Media Arts

요 약 지하철은 현대 도시의 교통 체계에서 중요한 역할을 하는 대중 교통 수단이다. 하지만, 갑작스런 고장 및 시스템 불통 등의 이유로 혼잡을 야기시키는 경우가 종종 발생하여 불편을 초래하고 있다. 따라서, 본 논문에서는 지하철 시스템의 효율적 운영을 위해 머신러닝을 활용한 고장 예측 및 예방 연구를 진행하였다. UC Irvine의 MetroPT-3 데이터셋을 활용하고, 로지스틱 회귀를 이용하여 지하철 고장 예측 모델을 구축하였다. 모델은 0.991의 높은 정확도로 비고장 상태를 예측하나, 정밀도와 재현율은 상대적으로 낮아 고장 예측에 있어 오류 가능성을 시사하고 있다. ROC_AUC 값이 0.901로, 모델이 무작위 추측보다 뛰어난 분류를 할 수 있다. 구축한 모델은 지하철 시스템의 안정적인 운영 운영에 유용하나, 성능 개선을 위한 추가 연구가 필요하다고 생각한다. 따라서 학습 데이터가 많고 데이터의 정제가 잘 이루어진다면 고장 예측을 통해 사전 점검을 하여 예방할 수 있다.

주제어 : 로지스틱 회귀 분석, 머신러닝, 지하철, 고장 예측, 예방

Abstract The subway is a means of public transportation that plays an important role in the transportation system of modern cities. However, congestion often occurs due to sudden breakdowns and system outages, causing inconvenience. Therefore, in this paper, we conducted a study on failure prediction and prevention using machine learning to efficiently operate the subway system. Using UC Irvine's MetroPT-3 dataset, we built a subway breakdown prediction model using logistic regression. The model predicted the non-failure state with a high accuracy of 0.991. However, precision and recall are relatively low, suggesting the possibility of error in failure prediction. The ROC_AUC value is 0.901, indicating that the model can classify better than random guessing. The constructed model is useful for stable operation of the subway system, but additional research is needed to improve performance. Therefore, in the future, if there is a lot of learning data and the data is well purified, failure can be prevented by pre-inspection through prediction.

Key Words : Logistic Regression Analysis, Machine Learning, Subway, Failure Prediction, Prevention

*Corresponding Author : Kuk-Kyung Sung(kksung@pro.ac.kr)
Received November 14, 2023
Revised November 21, 2023
Accepted December 20, 2023
Published December 30, 2023

1. 서론

지하철은 현대 도시의 교통 체계에서 중요한 역할을 하는 대중 교통 수단이다. 이는 도심과 교외를 연결하며, 매일 수많은 사람들이 직장, 학교, 병원 등의 중요한 목적지로 이동하는 데 이용한다. 지하철 시스템의 효율적이고 안전한 운영은 도시의 교통 흐름을 원활하게 유지하며, 경제적 활동의 지속성을 보장하는 핵심적인 요소로 자리잡고 있다.

하지만 대도시의 교통 혼잡 문제는 도시 계획과 환경 문제에 있어서 지속적으로 중요한 도전 과제로 남아 있다. 교통 혼잡을 줄이기 위한 노력의 일환으로, 지하철 시스템의 효율성과 신뢰성 강화가 강조되고 있다. 이러한 맥락에서 지하철의 고장 예측 및 예방은 단순히 이용객의 안전을 확보하는 것을 넘어, 도시의 지속 가능한 발전을 위해 필수적인 과제로 부상하고 있다[1-3].

최근에는 머신러닝 기술이 이러한 문제를 해결하는 데 있어 큰 잠재력을 보이고 있다. 다양한 산업 분야에서 그 효용성이 입증되고 있는 이 기술은, 복잡한 데이터 패턴을 인식하고 분석하여, 장애의 원인을 정확히 파악하고 예방할 수 있는 능력을 갖추고 있다. 본 논문에서는 운행 중인 지하철에서 수집된 데이터를 활용하여, 머신러닝 모델을 개발하고 이를 통해 주된 고장 원인을 식별하였다. 이는 지하철 시스템의 신뢰성 향상뿐만 아니라, 서비스 질의 극대화에 기여할 것으로 기대된다.

본 논문은 머신러닝을 활용하여 지하철 시스템의 신뢰성을 향상시키고, 운영 중 발생할 수 있는 다양한 문제들을 예측하며, 더욱 안전하고 효율적인 대중 교통 환경을 조성하는 데 기여할 수 있는 방안을 제시한다. 연구 결과는 지하철 운영 기관에게 유용한 정보를 제공할 뿐만 아니라, 머신러닝 기술의 교통 분야 적용 가능성을 넓히는 데 기여할 것이다[4-7].

2. 데이터 수집

머신러닝 모델의 학습 및 검증을 위해 UC Irvine Machine Learning Repository에서 MetroPT-3 Dataset인 작동 중인 지하철의 각종 센서 값을 사용하였다. 데이터에는 작동 중인 지하철에서 공기 압축기의 APU(Air Production Unit)의 압력, 온도, 모터 전류 및 흡기 밸브의 판독 값을 수집한 것이다. 속성으로는 TP2, TP3, H1, DV_pressure, Reservoirs, Motor_Current, Oil_Temperature, COMP, DV_electric, TOWERS, MPG, LPS, Pressure_Switch, Oil_Level, Caudal_Impulse로 구성되어 있다.

Table 1. Attribute Name

Attribute Name	Attribute Description
timestamp	Sensor measurement time at 10 second intervals
TP2	the measure of the pressure on the compressor.
TP3	the measure of the pressure generated at the pneumatic panel.
H1	the measure of the pressure generated due to pressure drop when the discharge of the cyclonic separator filter occurs.
DV_pressure	the measure of the pressure drop generated when the towers discharge air dryers; a zero reading indicates that the compressor is operating under load.
Reservoirs	the measure of the downstream pressure of the reservoirs, which should be close to the pneumatic panel pressure (TP3).
Motor_Current	the measure of the current of one phase of the three-phase motor; it presents values close to 0A - when it turns off, 4A - when working offloaded, 7A - when working under load, and 9A - when it starts working.
Oil_Temperature	the measure of the oil temperature on the compressor.
COMP	- the electrical signal of the air intake valve on the compressor; it is active when there is no air intake, indicating that the compressor is either turned off or operating in an offloaded state.
DV_electric	the electrical signal that controls the compressor outlet valve; it is active when the compressor is functioning under load and inactive when the compressor is either off or operating in an offloaded state
TOWERS	the electrical signal that defines the tower responsible for drying the air and the tower responsible for draining the humidity removed from the air; when not active, it indicates that tower one is functioning; when active, it indicates that tower two is in operation

(Continued)
Table 1. Attribute Name

Attribute Name	Attribute Description
MPG	the electrical signal responsible for starting the compressor under load by activating the intake valve when the pressure in the air production unit (APU) falls below 8.2 bar; it activates the COMP sensor, which assumes the same behaviour as the MPG sensor.
LPS	the electrical signal that detects and activates when the pressure drops below 7 bars.
Pressure_Switch	the electrical signal that detects the discharge in the air-drying towers.
Oil_Level	the electrical signal that detects the oil level on the compressor; it is active when the oil is below the expected values.
Caudal_Impulse	the electrical signal that counts the pulse outputs generated by the absolute amount of air flowing from the APU to the reservoirs

종속변수로 사용할 고장 데이터는 데이터 세트에 포함되어 있지 않으며, 따로 주어진다.

Table 2. Attribute Name

Nr.	Start Time	End Time	Failure	Severity	Report
#1	4/18/2020 0:00	4/18/2020 23:59	Air Leak	High stress	
#2	5/29/2020 23:30	5/30/2020 6:00	Air Leak	High stress	...
#3	6/5/2020 10:00	6/7/2020 14:30	Air Leak	High stress	...
#4	7/15/2020 14:30	7/15/2020 19:00	Air Leak	High stress	...

3. 데이터 분석

3.1 로지스틱 회귀 분석

로지스틱 회귀(Logistic Regression)는 독립 변수의 선형 결합을 이용하여 사건 발생 가능성을 예측하는 데에 사용되는 통계 기법이다. 로지스틱 회귀는 일반적인 회귀 분석과 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용한다. 이는 독립 변수의 선형 결합(Linear Regression)으로 종속 변수를 설명한다는 점에서 선형 회귀 분석과 유사하다. 하지만 로지스틱 회귀는 선형 회귀와 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 기법으로 나뉜다[8-10].

로지스틱 회귀는 보통 종속 변수가 이항형 문제를 지칭할 때 사용된다. 이외에도 두 개 이상의 범주를 가지는 문제가 대상인 경우, 다항 로지스틱 회귀 또는 분화 로지스틱 회귀(Polytomous Logistic Regression)라고 하고 복수의 범주이면서 순서가 존재하면 서수 로지스틱 회귀(ordinal logistic regression)라고 한다. 이 프로젝트에서 사용할 로지스틱 회귀는 유효 범주의 개수가 두 개이므로 이항형 로지스틱 회귀(Binomial Logistic Regression)를 사용한다.

로지스틱 회귀는 일반적인 선형 모델의 특수한 경우로 보기 때문에 선형 회귀와 유사하지만 종속 변수와 독립 변수 사이의 관계에 있어서 차이점을 가진다. 이항형인 데이터에 적용하였을 때 종속 변수 Y의 결과가 참과 거짓으로만 제한된다는 것이다. 그래서 선형 함수를 사용하여 추세를 예측하는 선형 회귀와 달리 S자 함수를 사용하여 참과 거짓으로 분류한다. 또, 종속 변수가 이진적이기 때문에 조건부 확률의 분포가 정규 분포 대신 이항 분포를 따른다[11-13].

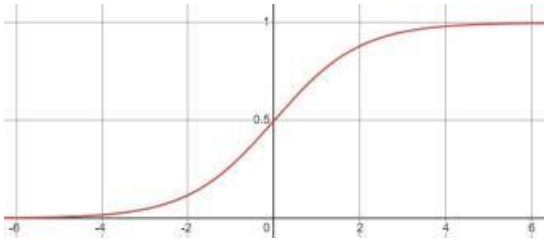

Fig. 1. Logistic Curve

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} \quad (1)$$

3.2 분석

데이터를 분석하기 위해 파이썬을 이용하여 pandas와 numpy 라이브러리를 사용했다. 작동 중인 지하철의 각종 센서 데이터 행의 개수는 1,516,948개가 있다.

```
In [4]: data_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1516948 entries, 0 to 1516947
Data columns (total 17 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   Unnamed: 0       1516948 non-null  int64
 1   timestamp        1516948 non-null  object
 2   TP2              1516948 non-null  float64
 3   TP3              1516948 non-null  float64
 4   H1               1516948 non-null  float64
 5   DV_pressure      1516948 non-null  float64
 6   Reservoirs       1516948 non-null  float64
 7   Oil_temperature  1516948 non-null  float64
 8   Motor_current    1516948 non-null  float64
 9   COMP             1516948 non-null  float64
 10  DV_eletric       1516948 non-null  float64
 11  Towers           1516948 non-null  float64
 12  MPG              1516948 non-null  float64
 13  LPS              1516948 non-null  float64
 14  Pressure_switch  1516948 non-null  float64
 15  Oil_level        1516948 non-null  float64
 16  Caudal_impulses  1516948 non-null  float64
dtypes: float64(15), int64(1), object(1)
memory usage: 196.7+ MB
```

Fig. 2. Data Information

현재 데이터에는 고장에 대한 값이 들어 있지 않아 Table 1 속성에 failure라는 속성을 추가했다. 초기 값은 0으로 초기화를 했다. Table 2에서 고장이 난 기간의 시작 시각과 종료 시각으로 데이터의 failure 값을 1로 변경하였다.

```
In [7]: print(data_df[(data_df['timestamp'] >= start_time_list[0])& (data_df['timestamp'] <= end_time_list[0])])

        Unnamed: 0            timestamp     TP2     TP3      H1  DV_pressure  \
562564      562560  2020-04-18 00:00:01 -0.018   8.248   8.238      -0.024
562565      562561  2020-04-18 00:00:13 -0.018   8.248   8.238      -0.024
562566      562562  2020-04-18 00:00:24 -0.018   8.248   8.238      -0.024
562567      562563  2020-04-18 00:00:36 -0.018   8.248   8.239      -0.024
562568      562564  2020-04-18 00:00:49 -0.018   8.248   8.239      -0.024
...            ...                  ...    ...     ...     ...         ...
571222      571220  2020-04-18 23:59:16   9.024   8.866  -0.008       2.002
571223      571221  2020-04-18 23:59:25   9.026   8.878  -0.006       2.002
571224      571240  2020-04-18 23:59:36   9.032   8.888  -0.006       2.006
571225      571250  2020-04-18 23:59:45   8.806   8.990  -0.006       1.354
571226      571260  2020-04-18 23:59:56   8.434   8.974  -0.006       1.690

        Reservoirs  Oil_temperature  Motor_current  COMP  DV_eletric  Towers  \
562564       8.248           49.450         0.0400   1.0         0.0     1.0
562565       8.248           49.450         0.0400   1.0         0.0     1.0
562566       8.248           49.450         0.0400   1.0         0.0     1.0
562567       8.248           49.450         0.0400   0.0         0.0     0.0
562568       8.248           49.450         0.0400   1.0         0.0     1.0
...            ...              ...            ...   ...         ...     ...
571222       8.864           73.500         5.7025   0.0         1.0     0.0
571223       8.876           73.625         5.8250   0.0         1.0     0.0
571224       8.884           73.650         5.8325   0.0         1.0     0.0
571225       8.896           73.425         5.1925   0.0         1.0     1.0
571226       8.870           73.425         5.7050   0.0         1.0     1.0

        MPG  LPS  Pressure_switch  Oil_level  Caudal_impulses  failure
562564  1.0  0.0             1.0        1.0              1.0        0
562565  1.0  0.0             1.0        1.0              1.0        0
562566  1.0  0.0             1.0        1.0              1.0        0
562567  0.0  0.0             0.0        0.0              0.0        0
562568  1.0  0.0             1.0        1.0              1.0        0
...     ...  ...             ...        ...              ...      ...
571222  0.0  0.0             1.0        1.0              1.0        0
571223  0.0  0.0             1.0        1.0              1.0        0
571224  0.0  0.0             1.0        1.0              1.0        0
571225  0.0  0.0             1.0        1.0              1.0        0
571226  0.0  0.0             1.0        1.0              1.0        0

[8663 rows x 18 columns]
```

Fig. 3. Data Result

로지스틱 회귀 분석 모델은 scikit-learn의 선형 분석 모델 패키지의 Logistic-Regression을 불러 모델 생성을 진행했다. 또한, 성능 확인을 위해 train_test_split을 사용하여 데이터를 학습 데이터와 테스트 데이터로 분류하여 성능을 확인했다.

4. 예측 및 결과

failure 속성을 제외한 나머지 속성 값을 데이터 프레임을 X에 저장을 했고, failure에 대한 데이터를 Y에 저장했다. X는 독립변수, Y는 종속변수에 해당한다.

```
feature_column = data_df.columns.difference(['failure'])
X=data_df[feature_column]
Y=data_df['failure']
```

Fig. 4. Data Frame Split

데이터 프레임 X와 Y를 넣고 테스트 사이즈를 0.3으로 지정하여 학습 데이터와 테스트 데이터를 분류하였다. 로지스틱 회귀 분석 모델을 생성하여 데이터의 0.7인 학습 데이터로 모델 학습을 수행하였다.

학습이 된 모델에 대해 테스트 데이터를 가지고 예측을 진행하여 예측값을 구한다. 생성된 모델의 성능 확인을 위해 scikit-learn에서 제공하는 모듈인 confusion_matrix, accuracy_score, precision_score, recall_score, 1_score, roc_auc_score를 사용하였다. 평가를 위해 7:3으로 분할한 455,085개의 테스트 데이터에 대해 이진 분류의 성능 평가 기본이 되는 오차행렬을 구한다. 실행 결과를 보면 TN 443,685개, FP 2,342개, FN 1,753개, TP 7,305개인 오차행렬이 나온다.

```
In [21]: import seaborn as sns
         import matplotlib.pyplot as plt

In [22]: sns.heatmap(confusion_matrix(Y_test, Y_predict), annot=True, fmt='d', cmap='BuGn')
         plt.title('confusion matrix')
         plt.xlabel('true')
         plt.ylabel('pred')
         plt.xticks([0.5,1.5], ['Negative(0)', 'Positive(1)'])
         plt.yticks([0.5,1.5], ['Negative(0)', 'Positive(1)'])
         plt.show()
```

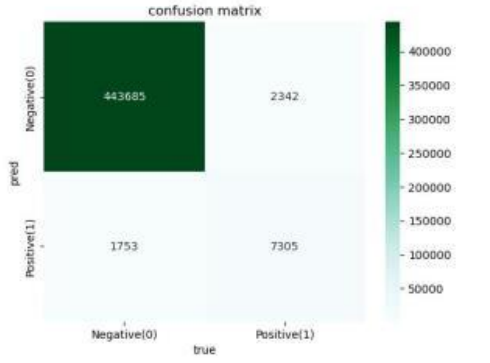

Fig. 5. Confusion Matrix Heat-Map

Scikit-learn에서 제공된 성능 평가 함수를 사용하여 정확도, 정밀도, 재현율을 구할 수 있다[14,15].

Fig. 6. Evaluation Metrics

5. 결론

본 연구에서는 머신러닝 기법 중 하나인 로지스틱 회귀 모델을 적용하여 지하철 시스템의 고장 탐지 및 예방에 관한 연구를 수행하였다. 오차 행렬 분석 결과, TN 비율이 높은 것을 보면 모델은 고장이 없는 상태를 정확하게 식별하는 데 매우 효과적이었음을 보여주었다. 이는 시스템이 안정적으로 작동 중임을 확신할 수 있는 높은 신뢰성을 의미한다.

모델의 정확도는 0.991로 매우 높게 나타나, 전반적으로 고장 상태와 비고장 상태를 잘 구분함을 나타낸다. 그러나 정밀도는 0.757, 재현율은 0.806, F1 스코어는 0.781로 상대적으로 낮게 나타났다. 이는 모델이 고장 상태를 예측하는데 있어서 일부 오류를 포함할 수 있음을 시사한다. 특히, 낮은 정밀도는 모델이 고장이라고 예측했지만 실제로는 고장이 아닌 경우가 다수 있음을 의미한다. 반면, 재현율은 실제 고장 상태를 어느 정도 잘 포착하고 있음을 보여준다. F1 스코어는 정밀도와 재현율의 조화 평균으로, 이 두 메트릭 간의 균형을 반영한다.

ROC_AUC 값은 0.901로, 모델이 무작위 추측보다 훨씬 뛰어난 분류 능력을 나타낸다. 이는 모델이 고장을 예측하는 데 있어서 높은 진단 능력을 가지고 있음을 의미한다.

이러한 결과를 바탕으로, 본 모델은 지하철 시스템의 안정적인 운영을 지원하는 데 있어서 유용한 도구가 될 수 있다. 하지만, 정밀도와 재현율을 개선하기 위한 추가적인 연구가 필요하다. 더 많은 데이터와 다양한 변수, 개선된 특징 선택 기법을 통해 모델의 성능을 높일 수 있을 것이다.

REFERENCES

[1] Lee, S. Y, Seo, B. W., & Park, S. M. (2023). Conv-LSTM-based Range Modeling and Traffic Congestion Prediction Algorithm for the Efficient Transportation System. *The Journal of the Korea institute of electronic communication sciences, 18(2),* 321-327.
DOI : 10.13067/JKIECS.2023.18.2.321

[2] Kim, J. Y, Lim, S. Y, Choo, S. H., & Park, I. K. (2015). Analysis of Transit Ridership Patterns and Influencing Factors in Seoul. *The Korea Spatial Planning Review.* 49-65.
DOI : 10.15793/KSPR.2015.87..004

[3] Kim, J. I. (2013). The Determinants of Subway Riderships at AM-peak in Daegu Metropolitan City: Focusing on the Land Use of Station Neighborhood Areas. *Journal of Transport Research, 20(1),* 15-25.
DOI : 10.34143/JTR.2013.20.1.15

[4] Ki, T. S., & Lee, S. H. (2017). A Prediction Scheme for Power Apparatus using Artificial Neural Networks. *Journal of Convergence for Information Technology, 7(6),* 201-207.
DOI : 10.22156/CS4SMB.2017.7.6.201

[5] Lee, H. W. (2011). Development of Supervised Machine Learning based Catalog Entry Classification and Recommendation System. *Journal of Internet Computing and Services, 20(1),* 57-65.
DOI : 10.7472/JKSII.2019.20.1.57

[6] Park, Y. K., & Youn, J. H. (2021). A Study on Detection of a Keyboard Trigger Based on Machine Learning. *Proceedings of the Korea Information Processing Society Conference,* 179-180.
DOI : 10.3745/PKIPS.Y2021M05A.179

[7] Paik, G. O., Kang, M. C., Soul, M. W. & Lim, S. J. (2020). ARIMA, Machine Learning Approach to Forecasting Empty Container Volumes. *Proceedings of the Korea Information Processing Society Conference*, 953-955.
DOI : 10.3745/PKIPS.Y2020M11A.953

[8] Kim, M. Y. (2017). Analysis for Factors of Predicting Problem Drinking by Logistic Regression Analysis. *Journal of Digital Convergence, 15(5),* 487-494.
DOI : 10.14400/JDC.2017.15.5.487

[9] Vasanth, R. et al. (2018). Identification of Environmental Factors in Fruit Disease by Logistic Regression. *Journal of Knowledge*

Information Technology and Systems, 13(5), 521-532.
DOI : 10.34163/JKITS.2018.13.5.002

[10] Baek, S. A. et al. (2016). Assessment of Slope Failures Potential in Forest Roads using a Logistic Regression Model. *Journal of Korean Forest Society, 105(4),* 429-434.
DOI : 10.14578/JKFS.2016.105.4.429

[11] Chun, J. A, Lee, H. J, Im, S. H, Kim, D. H. & Baek, S. S. (2021). Comparative assessment of frost event prediction models using logistic regression, random forest, and LSTM networks. *Korea Water Resources Association, 54(9),* 667-680.
DOI : 10.3741/JKWRA.2021.54.9.667

[12] Song, J. H, Shin, J. W. & Han, H. S. (2023). Multimetric Measurement Data Monitoring System Using Sigmoid Function. *The Journal of Engineering Geology, 33(1),* 137-149.
DOI : 10.9720/KSEG.2023.1.137

[13] Park, I. S. & Han, J. T. (2019). Study of effect on the obesity status using multilevel logistic regression analysis. *Journal of the Korean Data And Information Science Sociaty, 30(1).* 205-217.
DOI : 10.7465/jkdi.2019.30.1.205

[14] Kim, C. J., Jeong, J. H., Jo, C. W., & Yoo, J. K. (2019). A performance evaluation analysis of product recommendation techniques. *Journal of Knowledge Information Technology and System, 14(5),* 515-525.
DOI : 10.34163/jkits.2019.14.5.008

[15] Lee, S. M. & Kwon, H. Y. (n. d.). A Performance Evaluation of Deep Learning Methods for Anomaly Detection and Distributed Learning Model. *Korean Institute of Information Scientists and Engineers, 48(1),* 864-866.

성 국 경(Kuk-Kyung Sung) **[정회원]**

· 1991년 2월 : 한남대학교 전산학과 (이학석사)
· 1994년 3월 ~ 현재 : 한국영상대학교 방송영상미디어과 교수
· 2016년 1월 ~ 2023년 8월 : 한국영상대학교 평생교육원장
· 관심분야 : 빅데이터, 소프트웨어공학
· E-Mail : kksung@pro.ac.kr