**리프트가 위험해~**

**발표자: 전승현**

**팀원: 이지우, 강수경, 최아진**

**Github: seunghyun9999/~**

1. **안전 관련 머신러닝 모델 개발 관련 요약**
   1. 엘리베이터를 건설 현장의 리프트와 곤돌라 등 양중기에 접목하여IoT 센서가 한 달 동안 1분 단위로 캡처한 데이터를 분석해 실시간 모니터링을 통해 엘리베이터의 잠재적인 고장 가능성 및 고장 유무를 예측할 수 있는 모델 개발
2. **개발 목적**
   1. 모델 활용 대상:

엘리베이터의 고장 예측 및 사고 예방 개발 프로그램을 건설현장의 고위험장비인 리프트, 곤돌라 등 양중기에 적용시켜 안전관리의 자동화

* 1. 개발의 의의:

건설현장의 대형화 및 복잡화로 인해 안전관리의 범위가 넓어짐에 따라 현장 내 모든 작업 데이터를 꼼꼼하게 파악하기는 어려운 실정임. 잠재적인 위험을 사전에 예측하기 위해서 장비의 상태와 작업 환경 데이터를 실시간으로 추적 및 모니터링해 사고를 예방하는 것이 필요함. 엘리베이터의 고장 예측 및 사고 예방 개발 프로그램을 활용 시 실시간으로 센서 데이터를 분석하여 승강 장비의 작동 패턴, 이상 징후, 고장 전조 등 식별이 가능하며, 잠재적인 고장 가능성이 있을 경우 사전 알림을 통해 파악 가능함. 따라서 승강 장비의 상태 파악으로 적절한 유지보수 시점 결정, 고장 예측 및 조기 경고 시스템으로 승강 장비의 안전성 확보 및 안전관리 효율성과 정확성 기대 가능.

* 1. 독립변수와 종속변수

엘리베이터 외부의 환경 데이터 온도, 습도

엘리베이터 자체의 부하 및 상태 데이터 RPM, 진동, 압력 센서 1~6 을 입력하여 엘리베이터의 고장유무를 출력

1. **배경지식**
   1. 데이터 관련 사회 문제 설명:

‘승강기 사고 및 기술자 사고 통계'에 따르면 2019년부터 약 5년간 승강기 사고로 인한 사상자는 335명. 이 중 사망자는 27명, 부상자는 308명 수준으로 사고 발생률이 높음. 건설현장의 경우에도 고층에서 작업하거나, 무거운 자재를 들어 올리는 작업이 많기에 리프트와 곤돌라 같은 장비 사용이 잦음. 이러한 승강 장비 사용으로 인한 주요 사고사례로는 권과방지장치 미흡으로 과상승한 운반구로 인한 추락, 운행 전 점검 미흡으로 리프트 운반구 추락 등 추락 사고가 다수 발생하므로 사고 위험도가 매우 높음.

* 1. 머신러닝 모델 관련 설명 등

본 프로젝트의 종속변수 즉 결과값이 정상 혹은 고장을 나타내는 카테고리 형식이기 때문에 머신러닝 모델로 결정트리(Decision Tree)와 로지스틱 회귀(Logistic regression)를 사용할 예정

* + 1. 결정트리 모델

독립변수를 두개로 계속 구분하여 학습하고 종속변수를 분석할 때 yes or no를 반복하여 의사결정을 내리는 모델

분석을 도식화하면 나무모양

* + 1. 로지스틱 회귀 모델

종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내어 향후 예측모델에 사용

1. **개발 내용**
   1. 데이터에 대한 구체적 설명 및 시각화
      1. 데이터 개수, 데이터 속성

총 데이터의 개수는 44640개로 이는 데이터를 수집한 2020년 1월31일의 총 분수

* + 1. 데이터 간 상관관계

데이터 상관관계 분석 표는 아래 그림1, 2와 같다

그림 1 각 변수의 상관관계 분석표

텍스트, 스크린샷, 사각형, 패턴이(가) 표시된 사진

자동 생성된 설명

상관관계 수치가 0에 가까울수록 관계가 적음을 의미

그림 2 종속변수별 각 독립변수의 데이터 량 수치

텍스트, 도표, 평면도, 기술 도면이(가) 표시된 사진

자동 생성된 설명

* 1. 데이터(변수) 설명
     1. 독립변수, 종속변수 설정 (각 독립변수가 종속변수에 미칠 수 있는 영향)

독립변수 1 = 온도 (장비의 과열, 외부 온도에 의한 인장)

독립변수 2 = 습도 (습도, 장비의 부식, 무게 증량)

독립변수 3 = RPM (장비의 과부화, 오작동 상태)

독립변수 4 = 진동 (장비의 과부화, 장비의 부하)

독립변수 5 = 압력 (장비의 부하, 하중)

센서1~6 (베어링, 와이어로프 등 장비에 대한 센서)

종속변수 = 고장유무(0정상, 1위험, 2고장)

* 1. 데이터 전처리 과정
     1. 전처리 과정1

데이터 중간 빈칸이 존재하므로 빈칸을 0으로 대체

* + 1. 전처리 과정2

44640의 종속변수 중 1위험을 나타내는 데이터는 단 2개이기 때문에 데이터 로서의 학습이 어려워 가치가 낮음

따라서 1위험과 2고장을 모두 1비정상으로 합쳐서 프로젝트를 진행

* + 1. 전처리 과정 3

종속변수가 1비정상으로 나타날 때 많은 독립변수가 0으로 나타남

이대로 모델 학습을 진행하면 결과의 신빈성이 낮아지기 때문에 독립변수 중 0인 경우가 6개 이상인 행을 제외

* + 1. 전처리 과정4

시간은 종속변수와의 관계성이 없으므로 제외

센서1과 진동(Vibrations)의 상관계수: 0.9999

센서6과 RPM의 상관계수: 0.9980

매우 높은 상관관계를 나타내고 각 센서에 대한 출처가 분명하지 않은 4개의 독립변수가 종속 변수와의 관계마저 동일 하기 때문에 센서1, 6을 제외

총 독립변수 12->9

* + 1. 전처리 과정 5

종속변수 0정상의 개수 = 약 40000

종속변수 1 비정상의 개수 = 약 4000

데이터 불균형의 정도가 매우 심하기 때문에 모델학습을 할 때 종속변수가 1인경우 가중치를 3배 적용

* + 1. 우리가 예측하고자 하는 정보

엘리베이터의 현재 상태(독립변수)를 대입하여 종속변수 0정상 혹은 1비정상인지 판단

* 1. 머신러닝 모델 선정 이유
     1. 설명한 데이터를 기반으로 머신러닝 모델 선정 이유

종속변수가 0 정상 1비정상으로 이루어져 있기 때문에 카테고리형으로 판단 카테고리형을 잘 분석해주는 대표 모델인 결정트리(Decision Tree)와 로지스틱 회귀(Logistic regression)를 사용

* + 1. 성능 비교를 위한 머신러닝 모델 선정 이유

??????????????????????????????

* 1. 사용할 성능 지표
     1. 머신러닝 모델의 성능을 평가하기 위해 사용하는 성능 지표에 관한 설명

사용하는 머신러닝 모델인 결정트리(Decision Tree)와 로지스틱 회귀(Logistic regression)의 대표적 성능 지표인 정확도(accuracy)와 오차행렬(conf\_matrix)을 사용

* + - 1. 정확도(accuracy)

전체 예측 중에서 정확히 예측된 비율을 수치화 시켜 줌

* + - 1. 오차행렬(conf\_matrix)

모델이 예측으로 맞춘 수치와 맞추지 못한 수치를 열을 원래의 종속변수 값 행을 예측한 종속변수 값으로 아래 표1처럼 나타내어 줌

표 1

|  |  |  |
| --- | --- | --- |
| 실제 종속변수 0 | 100 | 1 |
| 실제 종속변수1 | 2 | 50 |
|  | 예측 종속변수 0 | 예측 종속변수 1 |

* + 1. 성능 지표 선정 이유 등
       1. 정확도(accuracy)

전체 예측의 전반적인 성능을 간단하게 파악 가능

* + - 1. 오차행렬(conf\_matrix)

모델 예측 각각의 상황을 한눈에 파악하기 편함

어떠한 종속변수의 예측을 실패했는지 파악이 빠름

예외) Model.predict (각 종속변수에 따른 예측하게 된 확률)

개발결과의 모델 성능지표들이 거의 일치하여 모델 선정의 어려움이 있음

이 코드를 활용하여 각 변수를 선택하게 된 확률을 수치로 알게 된다면 정상을 선택한 수치가 낮아질 때 알림을 하는 등 활용도가 높음

1. **개발 결과**
   1. 성능 지표에 따른 머신러닝 모델 성능 평가
      1. Accuracy(정확도)

결정트리(Decision Tree) = 정확도 1.0

로지스틱 회귀(Logistic regression) = 정확도 1.0

* + 1. 오차 행렬

결정트리(Decision Tree) = [[8150, 0]

[ 0, 359]]

로지스틱 회귀(Logistic regression) = [[8150, 0]

[ 0, 359]]

* + 1. KFold 결과

결정트리(Decision Tree) = K-Fold 교차검증 정확도: 0.999 ± 0.001

로지스틱 회귀(Logistic regression) = K-Fold 교차검증 정확도: 0.999 ± 0.001

* + 1. 시각화 자료

그림 3 각 모델 별 실제 값과 예측 값을 시각화한 자료

텍스트, 스크린샷, 라인, 폰트이(가) 표시된 사진

자동 생성된 설명

* + 1. Model.predict(각 종속변수에 따른 예측하게 된 확률)

결정트리(Decision Tree) =

[[1. 0.], [1. 0.], [1. 0.] . . .[1. 0.], [1. 0.], [1. 0.]]

로지스틱 회귀(Logistic regression) =

[[9.9997e-01 3.0000e-05], [9.9998e-01 2.0000e-05],

[9.9952e-01 4.8000e-04] . . . [9.9995e-01 5.0000e-05],

[9.9998e-01 2.0000e-05], [9.9997e-01 3.0000e-05]]

* 1. 머신러닝 모델의 성능 결과에 대한 해석
     1. 성능지표 정확도, 오차행렬, kfold의 수치가 일치
     2. 정확도가 비정상적으로 높다

데이터 자체를 분석해본 결과 종속변수가 1비정상인 경우 독립변수가 0이 되는 경우가 많아서 이러한 현상이 있는 것으로 보인다. 해당 현상을 줄이기 위해 여러 전처리 과정을 거쳤지만 눈에 띄는 결과는 없음

* + 1. 따라서 Model.predict (각 종속변수에 따른 예측하게 된 확률) 자료를 활용

결정트리의 경우 각각의 종속변수를 결정함에 확률이 100%지만 로지스틱 회귀의 경우 각각의 종속변수를 결정함에 다른 변수의 고려도 하는 모습을 보여 더 의미가 있는 자료를 보여주는 모델이라고 판단함

1. **결론**
   1. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명
   2. 개발 의의 등
   3. 머신러닝 모델의 한계