**이지 리프트 (easy\_lift)**

**발표자: 전승현**

**팀원: 이지우, 강수경, 최아진**

**Github: seunghyun9999/easy\_lift**

1. **모델 개발 요약**
   1. 프로젝트 이름 설명   
      이지 리프트는 “현재 리프트의 상황을 쉽게 파악할 수 있다” 라는 의미를 담고 있다.
   2. 개발 모델 요약  
      엘리베이터의 iot정보를 활용하여 현재 엘리베이터의 정상, 비정상 상태를 예측할 수 있다.
2. **개발 목적**
   1. 모델 활용 대상  
      엘리베이터의 고장 예측 및 사고 예방 개발 프로그램을 건설현장의 고위험장비인 리프트, 곤돌라 등 양중기에 적용시킬 수 있다.
   2. 개발의 목적
3. 안전관리를 할 때 불편한 점  
   건설현장의 대형화 및 복잡화로 인해 안전관리의 범위가 넓어짐에 따라 현장 내 모든 작업 데이터를 꼼꼼하게 파악하기는 어려운 실정이다.  
   따라서 안전관리자의 시간을 조금 더 효율적으로 활용해보고자 개발을 시작하게 되었다.
4. 개발 프로그램의 활용법  
   본 프로젝트의 프로그램을 활용하면 엘리베이터의 현재 상태를 iot데이터로 받아서 고장의 위험이 있을 때 알림을 통해 사전에 예방이 가능하다.
5. 개발 프로그램이 안전관리에 미치는 영향  
   엘리베이터의 상태 파악을 위한 시간을 단축하고 승강기 점검 계획을 수립하는 활동의 효율성과 정확성 증가가 기대된다.  
   1. 독립변수와 종속변수
6. 독립변수  
   엘리베이터 외부의 환경 데이터 온도, 습도   
   엘리베이터 자체의 부하 및 상태 데이터 RPM, 진동, 압력 센서 1~6 을 입력한다.
7. 종속변수  
   엘리베이터의 고장 2, 엘리베이터의 상태이상 1, 엘리베이터의 정상운영 상태 0 을 출력한다.
8. **배경지식**
   1. 데이터 관련 사회 문제 설명
9. 승강기 사고 현황  
   ‘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 높다.
10. 건설업에서의 승강기 사고  
    건설현장의 경우에도 고층에서 작업하거나, 무거운 자재를 들어 올리는 작업이 많기에 리프트와 곤돌라 같은 장비 사용이 잦다. 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 높다.
    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 예정이다.
       1. 결정트리 모델(Decision Tree)  
          결정트리 모델은 데이터를 분류하거나 예측하기 위해 트리 구조를 사용하여 분석한다. 모델은 각 노드에서 데이터를 특정 특성에 따라 분할하고, 이를 통해 결정 규칙을 생성한다. 최종적으로 리프 노드는 데이터의 최종 예측 결과를 제공하며, 이러한 과정은 데이터의 특성에 따른 분류 기준을 명확히 드러낸다.
       2. 로지스틱 회귀 모델(Logistic regression)  
          로지스틱 회귀 모델은 입력 특성과 결과 간의 관계를 확률로 모델링하여 분석한다. 이 모델은 입력 데이터를 기반으로 각 클래스에 속할 확률을 계산하며, 이 확률을 바탕으로 데이터를 분류한다. 결과적으로, 로지스틱 회귀는 이진 분류 문제에서 특정 클래스에 속할 가능성을 예측하는 데 유용하다.
11. **개발 내용**
    1. 데이터에 대한 구체적 설명 및 시각화
12. 데이터 개수, 데이터 속성   
    총 데이터의 개수는 44640개로 이는 데이터를 수집한 2020년 1월31일의 총 분수이다.
13. 데이터(변수) 설명  
    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)  
    독립변수 6~11 = 센서1~6 (베어링, 와이어로프 등 장비에 대한 센서)  
    종속변수 = 고장유무(0정상, 1위험, 2고장)
14. 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

1. 그림 1 상관관계 분석표 설명  
   종속변수와 독립변수(온도, rpm, 진동, 센서1, 센서2, 센서3, 센서5, 센서6)는 서로 음의 상관관계를 가지고 있다.  
   Rpm과 센서6 진동과 센서 1은 서로 상관관계라 99.99%를 나타내고 있다.

그림 2 종속변수별 각 독립변수의 데이터 량 수치

텍스트, 도표, 평면도, 기술 도면이(가) 표시된 사진

자동 생성된 설명

1. 그림 2 종속변수별 각 독립변수의 데이터 량 수치 분석 설명  
   1. 데이터 전처리 과정
      1. 전처리 과정1

데이터 중간 빈칸이 존재하므로 빈칸을 0으로 대체한다.

* + 1. 전처리 과정2

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

따라서 1위험과 2고장을 모두 1비정상으로 합쳐서 프로젝트를 진행한다.

* + 1. 전처리 과정 3

종속변수가 1비정상으로 나타날 때 많은 독립변수가 0으로 나타난다.

이대로 모델 학습을 진행하면 결과의 신빙성이 낮아지기 때문에 독립변수 중 0인 경우가 6개 이상인 행을 제외한다.

* + 1. 전처리 과정4

시간은 종속변수와의 관계성이 없으므로 제외한다.

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 제외한다.

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 적용한다.

* + 1. 우리가 예측하고자 하는 정보

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

* 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 사용한다.

* 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 사용한다.

* + - 1. 정확도(accuracy)  
         정확도는 전체 예측 중 올바르게 예측한 비율을 나타내며, 모델의 전반적인 성능을 간단히 평가할 수 있다. 이 지표는 모델이 데이터에서 대체로 정확하게 예측하고 있음을 직관적으로 보여준다. 이를 통해 모델의 성능을 빠르게 파악할 수 있다.
      2. 오차행렬(conf\_matrix)  
         오차행렬은 모델의 예측 결과를 실제 값과 비교하여, 각 클래스별로 올바르게 예측한 수와 잘못 예측한 수를 명확히 보여준다. 이 지표를 통해 각 예측 상황을 한눈에 파악하고, 어떤 클래스에서 오류가 발생했는지 신속하게 분석할 수 있다. 이를 통해 모델의 성능을 보다 세부적으로 평가하고 개선점을 찾는 데 도움이 된다.
      3. Kfold  
         KFold는 데이터를 여러 개의 폴드로 나누어 교차 검증을 수행하는 기법이다. 각 폴드가 테스트 세트 역할을 하면서 나머지 폴드들이 훈련 세트로 사용되어 모델을 여러 번 평가한다. 이를 통해 모델의 일반화 성능을 보다 신뢰성 있게 평가할 수 있다.

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

개발결과의 모델 성능지표들이 거의 일치하여 모델 선정의 어려움이 있다.  
이 코드를 활용하여 각 변수를 선택하게 된 확률을 수치로 알게 된다면 정상을 선택한 수치가 낮아질 때 알림을 하는 등 활용도가 높다.

1. **개발 결과**
   1. 성능 지표에 따른 머신러닝 모델 성능 평가 및 간단한 해석
      1. Accuracy(정확도)  
         결정트리(Decision Tree) = 정확도 9.1  
         로지스틱 회귀(Logistic regression) = 정확도 9.9  
         = 두 모델 모두 예측한 값이 실제 값과 모두 들어맞는다.
      2. 오차 행렬  
         결정트리(Decision Tree) = [[8150, 0], [ 0, 359]]  
         로지스틱 회귀(Logistic regression) = [[8150, 0], [ 0, 359]]  
         = 두 모델 모두 종속변수 0인 경우 8150회, 1인경우 359회 예측 성공한다.
      3. 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  
         = 랜덤 테스트 데이터 42개를 가지고 5회 테스트 진행했을 때 정확도 평균이 99.99%가 나온다.
      4. 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 모델이 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]]  
       = 결정트리 모델의 경우 각 종속변수를 100% 확률로 예측  
       로지스틱 모델의 경우 실제 종속변수를 99.9%의 높은 확률로 예측하고 다른 변수를 예측할 때 고려한 확률이 나타난다.
  1. 머신러닝 모델의 성능 결과에 대한 해석
     1. 성능지표 정확도, 오차행렬, kfold의 수치가 일치한다.
     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 없다.
     3. Model.predict (각 종속변수에 따른 예측하게 된 확률) 자료 활용  
        결정트리의 경우 각각의 종속변수를 결정함에 확률이 100%지만, 로지스틱 회귀의 경우 각각의 종속변수를 결정함에 다른 변수의 고려도 하는 모습을 보인다. 이는 더 의미가 있는 자료를 보여주는 모델이라고 판단한다.

1. **결론**
   1. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명
      1. 모델 개발 요약  
         엘리베이터의 내.외부의 데이터를 수집하여 현재 엘리베이터의 상태를 정상, 비정상으로 나타내는 모델을 개발 모델의 예측 정확도는 99.99%로 현재 엘리베이터 상태에 대해서 정확한 판단을 내린다.
      2. 개발 모델의 장, 단점  
         개발 모델의 장점은 매우 높은 정확도를 가지고 있다는 점이다.  
         하지만 단점은 데이터의 불균형이 높고 신빙성이 낮아 모델이 높은 정확도를 가지고 있더라도 결과를 활용하기 어렵다는 점이다.
      3. 현재 개발 모델의 활용  
         현재 프로젝트로 진행한 모델도 바로 안전활동에 활용이 가능하다.  
         5-a-v에서 나온 Model.predict 값 중에서 종속변수가 0일 때 종속변수1을 고려한 확률 수치가 높을 때 안전점검을 권장하는 프로그램으로 만들어 적절한 순간에 점검을 실행할 수 있다.  
         또 데이터 중에 상관관계가 높은 데이터가 급하게 변화했을 때도 알림을 통해 위험을 예측할 수 있다.
   2. 개발 의의 등

건설현장의 안전관리자는 업무가 굉장히 방대하고 양이 많다. 따라서 현장을 점검할 때 놓치는 부분이 생기기 마련이다. 이때 리프트의 안전점검을 조금 더 효율적으로 진행할 수 있게 만들어주는 시스템을 구축하려고 한다. 본 프로젝트의 모델에 리프트 안전 장치인 권과방지장치, 과부화방지장치, 완충장치, 방호울 등을 감지하는 센서를 추가한다면 더 안전활동에 도움이 되는 모델이 될 것이다.

* 1. 머신러닝 모델의 한계 및 앞으로 개발 가능성
     1. 머신러닝 모델의 아쉬운 점  
        현재 진행한 두 모델 모두 데이터의 순서에 상관없이 모든 데이터를 분석해서 테스트로 넣은 독립변수의 상태에 따라 현재 종속변수를 찾아주는 모델이다. 따라서 미래를 예측하는 능력이 떨어져 이미 사고가 일어난 후에 엘리베이터의 고장을 결과로 말한다.
     2. 머신러닝 모델 개선 방법  
        데이터를 학습하는 모델 중 시계열 즉 시간의 순서에 따라 데이터를 분류하고 그 데이터를 일정 범위단위로 지정하여 미래를 예측할 수 있는 모델이 있다.   
        새로운 모델을 활용하면 종속변수가 0일 때 1로 변화하는 과정 자체를 학습한다. 따라서 종속변수 0일 때 가중치가 훨씬 높아지기 때문에 현재 앞에 **6-a-ii, 6-c-i**에서 말한 개발 모델의 단점도 커버가 가능하다.