수리비 예측 모델 제작 1차 보고

2023/01/26 박성완

1. **데이터셋 제작**

총 12만4천개의 수리비 예측 데이터셋 중 5만개를 사용해서 데이터셋 제작.

테이블이(가) 표시된 사진

자동 생성된 설명

(기존 수리비 견적서 형식)

우리의 파손 모델의 output과 input으로 예측할 수 있는 feature 정리

* + - COMPANY 🡪 제작사 : BMW
    - MODELTYPE 🡪 차량 세부 모델 : 520d
    - MILEAGE 🡪 주행거리 : 100000km
    - FIRSTDAY 🡪 최초등록일 : 20150101
    - CARNAME 🡪 차량 이름 : BMW 5시리즈
    - PART 🡪 작업 항목 및 부품 명 : 도장, 앞범퍼, 헤드램프 …
    - SEVERITY 🡪 작업 내용 : 교환, 탈착, 판금 …
    - PRICE 🡪 공임비 : 100000원

50000개의 기존 데이터셋에서부터 445000개의 데이터 ROW 생성

텍스트이(가) 표시된 사진

자동 생성된 설명

(CSV 형식으로 새로 생성된 데이터셋)

1. **데이터셋 가공**

**2-1.** 899개의 MODELTYPE 존재, feature를 압축 하려고 노력해 보았지만, 데이터에 규칙성이 없어 오히려 정확도가 떨어지는 결과가 나와 MODELTYPE feature 삭제

**2-2.** 18000개의 PART 존재, 18000개의 PART를 우리의 모델이 해석할 수 있는 11개의 PART로 압축함.

텍스트이(가) 표시된 사진

자동 생성된 설명

우리의 모델이 해석할 수 없는 PART는 제외.

**2-3.** Dataset SEVERITY의 경우 18개가 존재.

대표적으로는 오버홀, 교환, 판금, 수리, 조정, 도장, OH, 탈착이 존재.

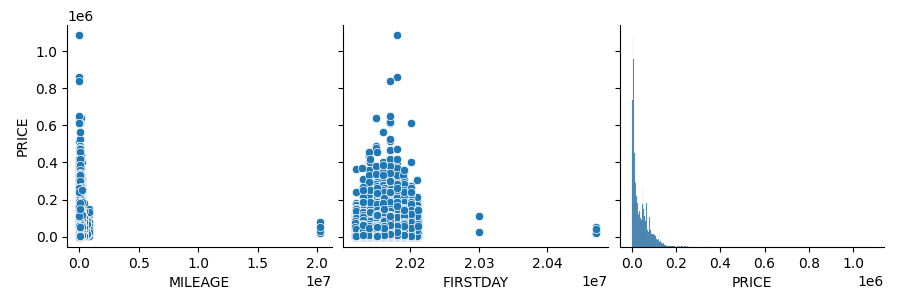
위와 같은 feature들을 우리 모델이 해석할 수 있는 4단계의 심각도와 matching 시킴.

텍스트이(가) 표시된 사진

자동 생성된 설명

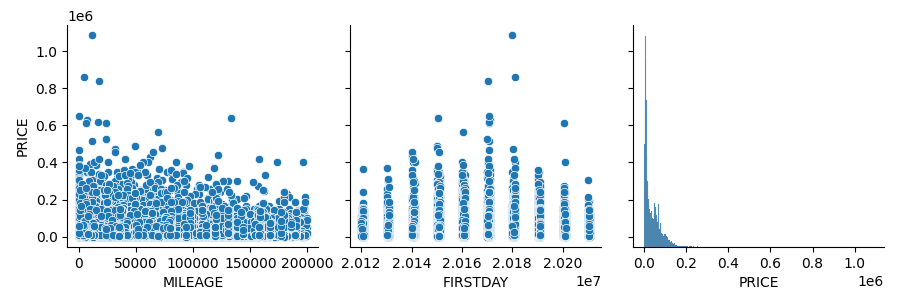
2-4. 결측값 제거와 위와 같은 과정을 진행 후 445000 row에서 68000 row가 남음.

1. **Integer features들 Outlier 제거**



(제거 전 데이터 분포)

MILEAGE와 FIRSTDAY에서 outlier 제거 후 데이터 분포



(제거 후 데이터 분포)

1. **Linear regression 모델 학습**

**4-1**. categorical features 들은 one-hot-encoding으로 변환

CARNAME : 323개

COMPANY : 28개

SEVERITY : 4개

PART : 11개

총 366개의 column 존재

**4-2**. Integer features들은 MinMaxScaler를 사용하여 0~1 값으로 정규화

Train, test 데이터셋은 85 : 15 비율로 설정.

먼저 linear regression을 사용하여 학습을 진행해봄.

텍스트이(가) 표시된 사진

자동 생성된 설명

성능이 상당히 좋지 않은 것을 알 수 있음.

1. **Ensemble 모델 사용**

**5-1.** Gradient Boost 모델 사용

텍스트이(가) 표시된 사진

자동 생성된 설명

조금 나아졌지만 여전히 좋지 않다.

**5-2.** Random Forest 모델 사용

텍스트이(가) 표시된 사진

자동 생성된 설명

1. **SVM 모델 사용**

Sklearn SVR, kernel로는 rbf 사용

**텍스트이(가) 표시된 사진

자동 생성된 설명**

1. **향후 방안**

여러 모델들을 사용해 보았지만 사용할 수 없는 성능이 나오는 것을 확인 할 수 있음.

데이터 feature들의 질이 좋지 않아 발생하는 문제로, 문제의 이유를 추측해보면

1. MODELTYPE field 삭제,
2. 18개의 SEVERITY를 4단계로 부적절하게 대응
3. One-hot-encoding값까지 scaling 해버림
4. Data Column(feature)이 너무 많음 366개 (one-hot-encoding을 사용해서 sparse함)

정도로 파악할 수 있고 해결 방안으로는

1. feature들을 결합해서 의미 있는 feature 생성
2. MODELTYPE field 적절하게 사용
3. SEVERITY field 4단계로 적절하게 대응
4. 더 좋은 모델 사용 및 하이퍼 파라미터 튜닝
5. One-hot-encoding 방식말고 다른 방식으로 categorical features 처리 등을 시도할 예정이다.