테너지

Code States

AI Bootcamp 1기

유병욱

CONTENTS

01. 프로젝트 소개

02. 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

01

01. 프로젝트 소개

02. 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

프로젝트 소개

1. 데이터 소개

1) 서울 시내를 주행 하는 약 9,000 대의 시내 버스로부터의 일일 주행 데이터

2) 메타 데이터 : 차량 번호 / 노선 번호 / 운전자

3) 측정 데이터 : 차량 속도. 가속도 / 엔진 회전 수 / 엔진 토크 / 연료 소모량 / GPS 등의 1sec Modal Data

2. 연료 소모량 Modal Data 의 이상값 처리 기준 제시

1sec 단위로 실시간 데이터를 통신으로 받을 시 차량의 노이즈로 인해 이상값들이 수집 된다.

이상값 판단 알고리즘을 통해 데이터 품질을 높일 수 있는 방안 제시 필요

3. 서울 시내 버스 연비 영향 인자 분석

차량의 연비는, 차량 연식, 운전자 운전 패턴, 도로 조건 등에 따라 차이가 난다.

기존 데이터 들로 부터 차량 연비에 어떠한 인자들이 어느 정도의 영향이 미치는지 분석

02

01. 프로젝트 소개

02. 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

데이터 소개

서울 시내를 주행 하는 약 9,000 대의 시내 버스로부터의 일일 주행 데이터

공항버스

- 당산역 → 개화역광역환승센터 까지의 노선을 가진 버스의 6개월 초단위 데이터

```
1 print(df.shape)
```

```
(8506129, 44)
```

- 위와 같이 6개월의 초단위 데이터를 결합하니 850만개 가량의 샘플

- 44개의 특성

버스	버스	버스	버스	버스	측정장치	측정값	측정값	측정값	측정값	측정값	측정값
버스위치: U=차고지, D=주행	S=Simple : 이전값 사용	버스종류	냉각팬 제어를 위한 버스 상태	CAN 통신상태	장치상태	년	월	일	시	분	초
DrvState	LogType	BusType	BusState	CanState	AuxStatus	Year	Month	Day	Hours	Minute	Sec
측정값	냉각팬제어 장치	측정값	측정값	측정값	측정값	측정값	측정값	측정값	측정값	측정값	측정값
버스냉각수 온도	냉각팬제어 셋팅 값(3= 통신끊김, 4= 셋팅안됨, 1,2 냉각팬 제어 단수)	엔진회전수	선택 속도	차량속도(통신)	차량속도(GPS계산)	차량속도(타코메터)	가속페달위치센서	제동페달	차량가속도	엔진토크	연료소모량
CoolTemp	FanControl	RPM	Speed	TachoSpeed	CanSpeed	GpsSpeed	APS	Brake	ACC	Torque	FuelRate
지시값	측정값	계산값	계산값	계산값	계산값	계산값	계산값	계산값	계산값	계산값	계산값
주전 기어단수	운전자 실 차량 기어	운전자 급출발 점수	운전자 급가속 점수	운전자 제동 점수	운전자 탄력 운전점수	운전자 평균RPM 점수	운전자 가속점수	운전자 변속점수	운전자조기 변속점수	운전자 엑셀조작점수	운전자 경제운전 점수 총점
EcoGear	DriveGear	AccGrade	Acc2Grade	DecGrade	CoastingGrade	AvgRPMGrade	SpeedGrade	SIGrade	PreSIGrade	ShortAccGrade	EcoGrade
운전자 ID	GPS	GPS	GPS	CAN 통신 여유 비트	CAN 통신 여유 비트	CAN 통신 여유 비트	CAN 통신 여유 비트				
운전자 ID	신호상태	위도	경도	No Meaning	No Meaning	No Meaning	No Meaning				
DrvNo	GpsState	Latitude	Longitude	ItemIndex1	ItemIndex2	ItemIndex3	ItemIndex4				

03

01. 프로젝트 소개

02. 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

연료 소모량 Modal Data 의 이상값 처리 기준 제시

IQR Method

- 사분위 값의 편차를 이용
  25%(1/4 지점) : Q1
  50%(2/4 지점) : Q2
  75%(3/4 지점) : Q3
  100%(4/4 지점) : Q4

- IQR은 여기서 Q1 ~ Q3 지점을 뜻한다. 즉, IQR = Q3 - Q1 이다.
IQR * 1.5 를 곱해서 이를 Q3에 더하고,
IQR * 1.5 값을 Q1에서 뺀다.
Q3에 더한 값을 최대값,
Q1에서 뺀 값을 최소값이라고 한다.
그리고 저 최대값, 최소값보다 크거나,
작은 값들을 이상치(outlier)라고 한다.

Standard Deviation Method

- 표준 편차는 분산의 척도 즉, 개별 데이터 포인트가 평균에서 얼마나 분산되어 있는지의 수치.

- 통계에서 데이터 분포가 대략 정규이면 데이터 값의 약 68 %가 평균의 1 표준 편차 내에 있고 약 95 %가 2 표준 편차 내에 있으며 약 99.7 %가 3 표준 편차 내에 있음

Isolation Forest

- 랜덤포레스트가 의사결정나무를 여러번 반복하여 앙상블 하듯이, Isolation Forest는 iTree를 여러번 반복하여 앙상블 합니다.

- iTree
  1. Sub-sampling : 비복원 추출로 데이터 중 일부를 샘플링
  2. 변수 선택 : 데이터 X의 변수 중 q를 랜덤 선택
  3. split point 설정 : 변수 q의 범위(max~min) 중 uniform하게 split point를 선택
  4. 1~3번 과정을 모든 관측치가 split 되거나, 임의의 split 횟수까지 반복(=재귀 나무)하며, 경로길이를 모두 저장

I- solation Forest
  5. 1~4번 과정(iTree)을 여러번 반복

03

01. 프로젝트 소개

02. 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

연료 소모량 Modal Data 의 이상값 처리 기준 제시

IQR Method

```
1 # IQR 이상치 확인
2 outlier_iqr(df, 'FuelRate')
```

```
IQR은 6.499999999999999 이다.
lower bound 값은 -7.799999999999998 이다.
upper bound 값은 18.199999999999996 이다.
총 이상치 개수는 1292 이다.
```

Standard Deviation Method(SDM)

```
1 # SDM Method 로 이상치 확인
2 out_sdm(df, 'FuelRate')
```

```
lower bound 값은 -5.762916826825103
upper bound 값은 15.853423407603445
총 이상치 개수는 2798 이다
```

Isolation Forest

```
1 # 해당 이상치 index 확인
2 df_iso_anomaly.index
```

```
Int64Index([    0,     1,     2,     3,
                9,
            ...
            50259, 50265, 50266, 50267,
            50289],
           dtype='int64', length=503)
```

```
1 # 3가지의 방법으로 뽑아낸 이상치들중 중복된값 보기
2 df_out_index = df_out_iqr.index & df_out_std.index & df_out_iso.index
3 len(df_out_index)
```

```
70
```

04

01. 프로젝트 소개

02. 데이터 소개

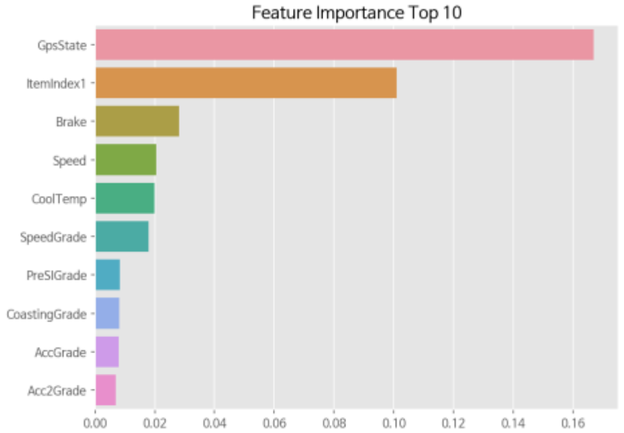
03. 과제 1

04. 과제 2

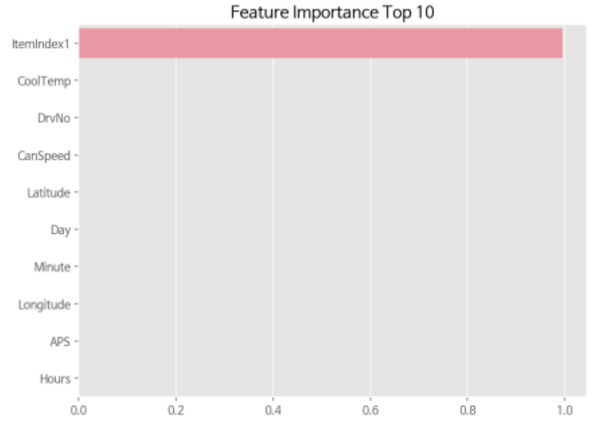
05. 고찰

서울 시내 버스 연비 영향 인자 분석

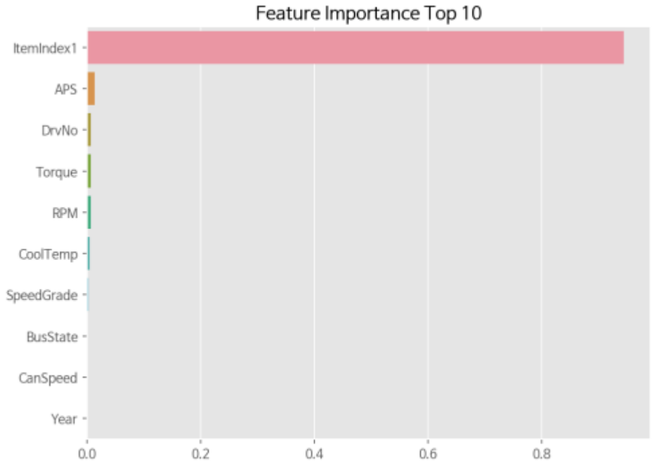
Linear Regression



RandomForest



Xgboost



05

01. 프로젝트 소개

02 데이터 소개

03. 과제 1

04. 과제 2

05. 고찰

고찰

데이터

프로젝트

감사합니다

AI Bootcamp 1기

유병욱