SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

삼성 청년 SW 아카데미

데이터분석 이론 및 실습

Confidential

〈알림〉

본 강의는 삼성 청년 SW아카데미의 컨텐츠로
보안서약서에 의거하여
강의 내용을 어떠한 사유로도 임의로 복사, 촬영,
녹음, 복제, 보관, 전송하거나
허가 받지 않은 저장매체를
이용한 보관, 제3자에게 누설, 공개,
또는 사용하는 등의 행위를 금합니다.

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

9월23일 실습 및 과제 알고리즘

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

실습 알고리즘

실습1. 고급 통계 분석

- 실행결과

```
문제 1. 데이터의 구조와 각 열의 데이터 타입:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   date             10000 non-null  object
 1   district         10000 non-null  object
 2   temperature      10000 non-null  float64
 3   precipitation    10000 non-null  float64
 4   num_vehicles     10000 non-null  int64
 5   num_pedestrians  10000 non-null  int64
 6   day_of_week      10000 non-null  int64
 7   is_weekend       10000 non-null  int64
 8   month            10000 non-null  int64
 9   traffic_volume   10000 non-null  int64
 10  avg_speed        10000 non-null  float64
dtypes: float64(3), int64(6), object(2)
memory usage: 859.5+ KB
None
```

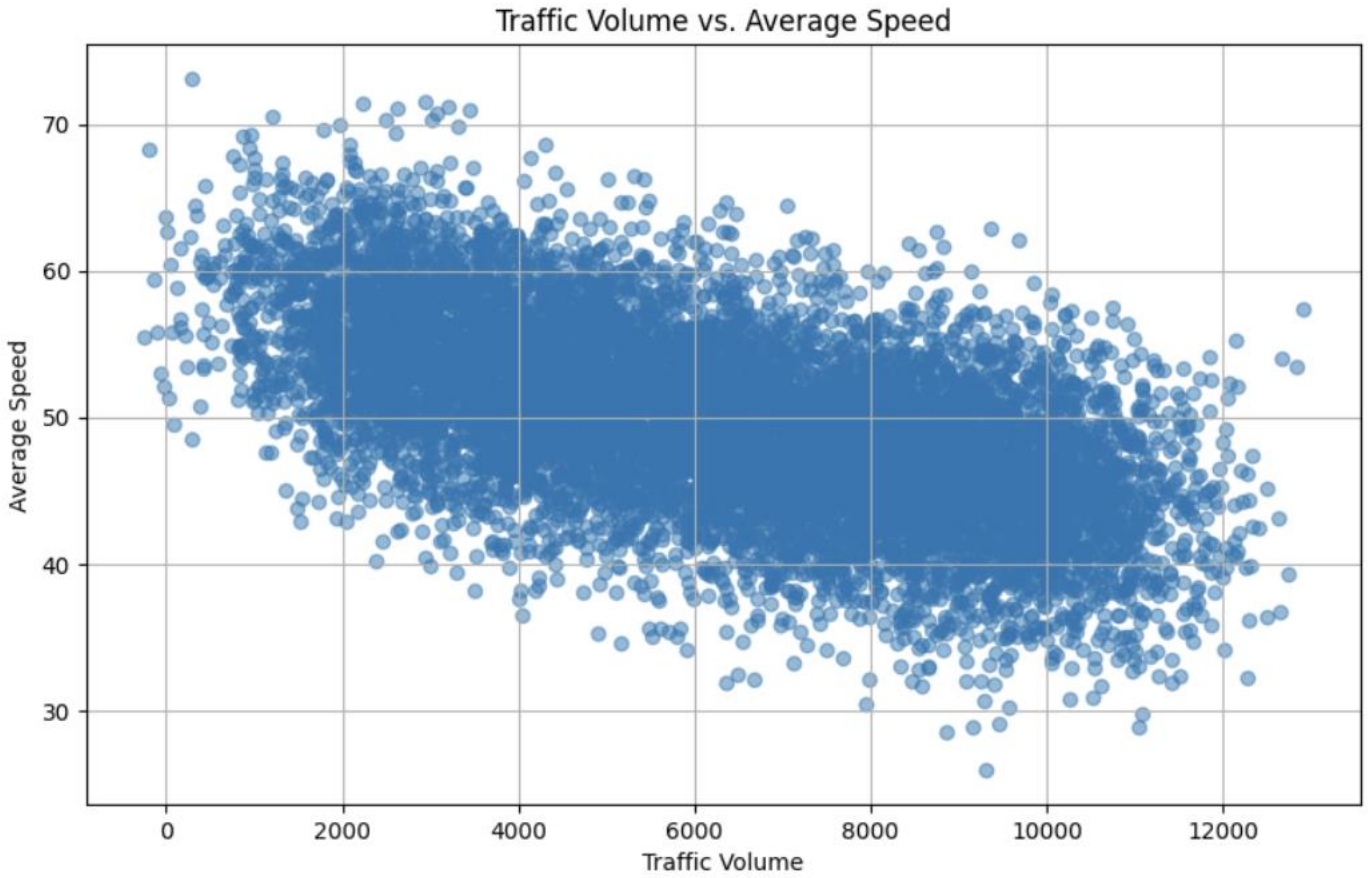
실습1. 고급 통계 분석

- 실행결과

```
문제 1. 데이터의 첫 5행 출력:
         date district  temperature  ...  month  traffic_volume  avg_speed
0  2023-04-13      서초구    -1.971793  ...      4            6528      49.44
1  2023-12-15      강서구    12.788840  ...     12            6287      48.74
2  2023-09-28      마포구    13.508233  ...      9            9597      45.30
3  2023-04-17      강서구    22.650517  ...      4            3450      50.20
4  2023-03-13      강남구     5.461658  ...      3            7867      46.86

[5 rows x 11 columns]
문제 2. 해석: 교통량이 증가할수록 평균 속도가 감소하는 경향이 보입니다.

Process finished with exit code 0
```

실습1. 고급 통계 분석

- 실행결과



실습2. 통계 기법 심화

- 실행결과

```
문제 1. 단순 선형 회귀 모델의 계수(Coefficient): -0.001378102442243744
문제 1. 단순 선형 회귀 모델의 절편(Intercept): 58.448558287440186
문제 2. 모델의 R-squared 값: 0.3577104203386625
문제 3. 해석: 회귀 계수는 -0.001378로, 교통량이 증가할수록 평균 속도가 감소하는 경향을 나타냅니다.
절편은 58.45로, 교통량이 0일 때 예상되는 평균 속도입니다.
R-squared 값이 0.36이라는 것은 이 모델이 약 36%의 데이터를 설명할 수 있다는 뜻입니다. 이는 모델이 교통량과 속도 간의 관계를 어느 정도 설명하지만, 완벽하지 않다는 것을 보여줍니다.

Process finished with exit code 0
```

실습2. 통계 기법 심화

- 실행결과

실습3. 복잡 데이터 분석

Confidential

- 실행결과

```
문제 1. 데이터의 각 열의 데이터 타입:
date                object
district            object
temperature        float64
precipitation      float64
num_vehicles         int64
num_pedestrians      int64
day_of_week          int64
is_weekend           int64
month                int64
traffic_volume       int64
avg_speed          float64
dtype: object
문제 1. 날짜 컬럼을 변환하여 새로운 연, 월, 일 정보가 추가된 데이터 프레임:
  district  temperature  precipitation  ...  avg_speed  year  day
0      서초구    -1.971793      23.419412  ...      49.44  2023   13
1      강서구    12.788840       1.572606  ...      48.74  2023   15
2      마포구    13.508233       7.186694  ...      45.30  2023   28
3      강서구    22.650517       0.305907  ...      50.20  2023   17
4      강남구     5.461658       3.150569  ...      46.86  2023   13

[5 rows x 12 columns]
```

실습3. 복잡 데이터 분석

- 실행결과

```
문제 1. 다중 선형 회귀 모델의 각 변수의 계수:
           Variable   Coefficient
0       temperature  3.196737e+00
1     precipitation -4.783435e+01
2      num_vehicles  9.765077e-01
3   num_pedestrians  1.246103e-01
4       day_of_week -2.482618e+00
5        is_weekend  9.831388e+02
6             month -8.159089e-01
7         avg_speed -2.120044e+01
8              year -1.705303e-13
9               day  2.732239e-01
10     district_강서구  1.413146e+00
11     district_마포구 -1.253215e+01
12     district_서초구  2.791298e+01
13     district_종로구  1.228269e+01
문제 1. 모델의 R-squared 값: 0.9427111715638055
문제 2. 가장 영향력 있는 변수 3개:
        Variable  Coefficient
5     is_weekend   983.138845
12  district_서초구    27.912981
13  district_종로구    12.282687
```

실습3. 복잡 데이터 분석

- 실행결과

```
문제 3. 성능 비교 및 해석:
실습 2의 단순 회귀 모델 R-squared 값: 0.3577104203386625
실습 3의 다중 회귀 모델 R-squared 값: 0.9427111715638055
해석: 단순 회귀 모델의 R-squared 값은 약 36%였지만, 다중 회귀 모델에서는 약 94%로 성능이 크게 향상되었습니다.
이는 다중 회귀 모델이 더 많은 변수들을 포함하여, 더 많은 데이터의 변화를 설명할 수 있기 때문입니다.


Process finished with exit code 0
```

실습4. 고급 분석 기법

Confidential

- 실행결과

```
문제 1. 데이터의 각 열의 데이터 타입:
date                object
district            object
temperature        float64
precipitation      float64
num_vehicles         int64
num_pedestrians      int64
day_of_week          int64
is_weekend           int64
month                int64
traffic_volume       int64
avg_speed          float64
dtype: object
문제 2. 날짜 컬럼을 변환하여 새로운 연, 월, 일 정보가 추가된 데이터 프레임:
  district  temperature  precipitation  ...  avg_speed  year  day
0      서초구    -1.971793      23.419412  ...      49.44  2023   13
1      강서구    12.788840       1.572606  ...      48.74  2023   15
2      마포구    13.508233       7.186694  ...      45.30  2023   28
3      강서구    22.650517       0.305907  ...      50.20  2023   17
4      강남구     5.461658       3.150569  ...      46.86  2023   13

[5 rows x 12 columns]
```

실습4. 고급 분석 기법

- 실행결과

```
문제 4. 데이터 분할:
훈련 데이터셋 크기: 7000개, 테스트 데이터셋 크기: 3000개

문제 5. 모델 훈련 완료

문제 6. 모델 평가 지표:
R-squared: 0.9423360734779507
MSE: 449884.12329080456
MAE: 521.9258165642168

문제 7. 5-fold 교차 검증 결과:
Average R-squared from 5-fold CV: 0.9424993876050699

문제 7. 모델의 성능 종합 평가 및 결과 해석:
R-squared 값이 약 94%라는 것은 모델이 대부분의 데이터를 잘 설명하고 있다는 것을 의미합니다.
MSE와 MAE 값은 모델이 예측하는 값이 실제 값과 얼마나 차이가 있는지를 보여주며, 값이 작을수록 예측이 정확합니다.
5-fold 교차 검증 결과에서 평균 R-squared 값이 높게 나온 것은 모델이 데이터의 패턴을 일관되게 잘 설명한다는 것을 나타냅니다.
전반적으로, 이 모델은 교통량 예측에 매우 유용하며, 대부분의 변수들이 모델 성능을 향상시키는 데 기여하고 있습니다.

Process finished with exit code 0
```

실습5. 통계적 심화 분석

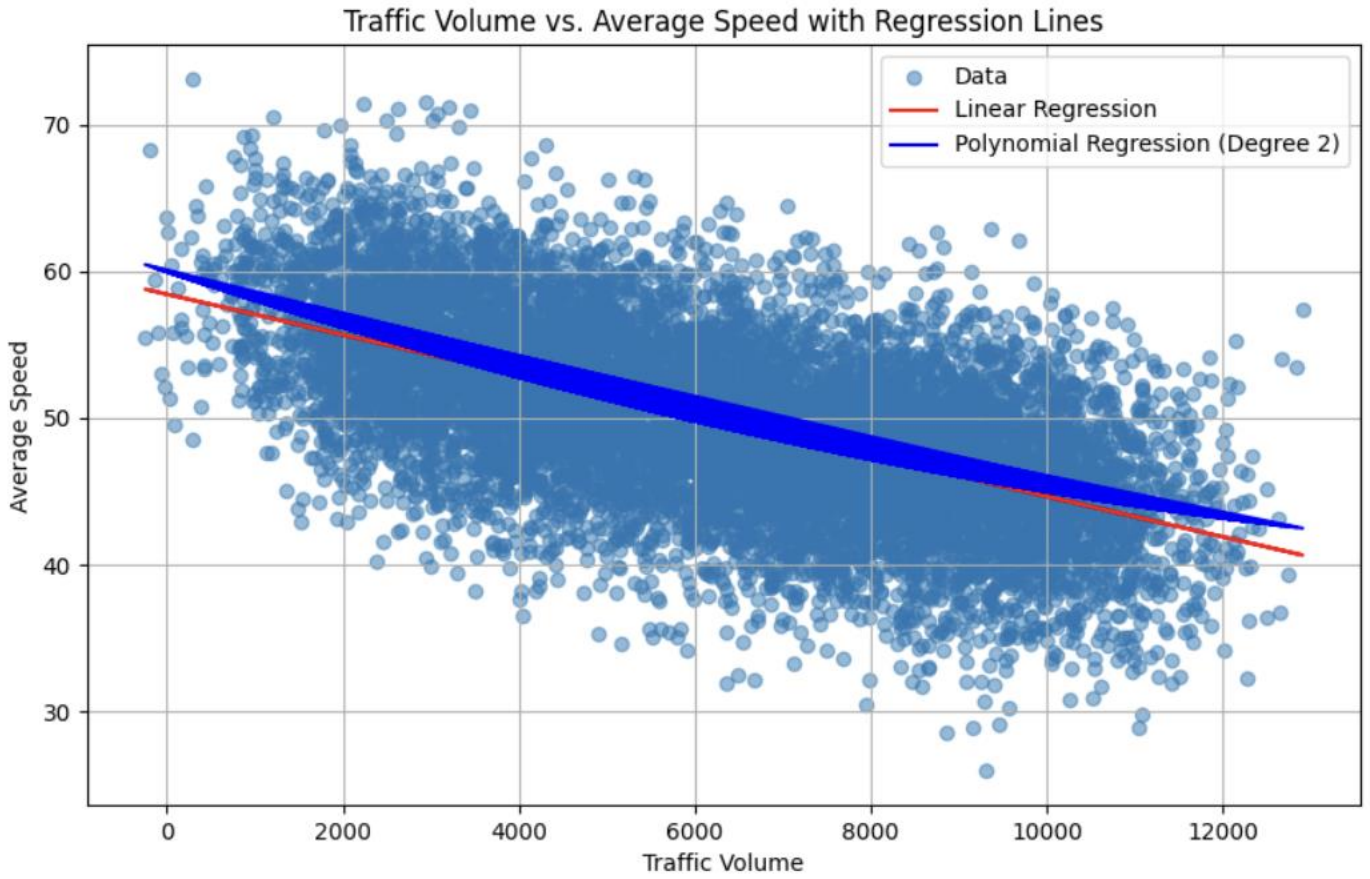
- 실행결과

```
문제 2. Linear Model R-squared: 0.3577104203386625
문제 2. Polynomial Model (Degree 2) R-squared: 0.36118194238184487

문제 4. 해석:
선형 회귀 모델과 다항 회귀 모델을 비교한 결과, 다항 회귀 모델의 R-squared 값이 더 높게 나왔습니다.
이는 다항 회귀 모델이 선형 모델보다 데이터의 비선형 관계를 더 잘 설명한다는 것을 의미합니다.

Process finished with exit code 0
```

실습5. 통계적 심화 분석

Confidential

- 실행결과



SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

과제 알고리즘

과제1. 데이터 처리 및 분석 기법

Confidential

- 실행결과

```
문제 1. 월별 평균 교통량:
    month  traffic_volume
0       1     6290.400466
1       2     6230.382082
2       3     6261.591019
3       4     6333.330636
4       5     6320.668981
5       6     5996.664685
6       7     6198.705251
7       8     6236.859977
8       9     6051.539846
9      10     6396.167264
10     11     5932.342482
11     12     6250.742857
문제 2. 주말 고온 데이터:
          date district  temperature  ...  month  traffic_volume  avg_speed
19  2023-12-10      마포구    27.292771  ...     12            9394      42.74
26  2023-09-10      강남구    32.079485  ...      9            7209      51.45
34  2023-07-09      마포구    26.286728  ...      7            9079      45.81
50  2023-01-14      강남구    33.272187  ...      1            4827      53.03
92  2023-01-15      마포구    28.847027  ...      1            3820      58.42

[5 rows x 11 columns]
```

과제1. 데이터 처리 및 분석 기법

- 실행결과

```
문제 3. 평균 속도가 40 이하인 데이터 수: 580
문제 4. 교통량과 기온의 상관관계: 0.019089624296374612
문제 5. 새로운 변수 traffic_density 생성:
    traffic_volume  num_vehicles  traffic_density
0             6528          8230         0.793196
1             6287          5362         1.172510
2             9597          7176         1.337375
3             3450          2832         1.218220
4             7867          7183         1.095225

Process finished with exit code 0
```

과제2. 고급 데이터 분석 및 응용

Confidential

- 실행결과

```
문제 1. 이상치 제거 후 데이터:
    temperature  precipitation  ...  district_서초구  district_종로구
0     -1.971793      23.419412  ...          True         False
1     12.788840       1.572606  ...         False         False
2     13.508233       7.186694  ...         False         False
3     22.650517       0.305907  ...         False         False
4      5.461658       3.150569  ...         False         False

[5 rows x 15 columns]
문제 2. 원본 데이터 R-squared: 0.9427111715638055
문제 2. 이상치 제거 데이터 R-squared: 0.9412901231840743
해석: 이상치를 제거했지만 R-squared 값이 하락하거나 거의 변화가 없었습니다. 이는 모든 이상치가 모델 성능에 부정적인 영향을 주지 않으며, 이상치 제거가 항상 모델 성능을 개선하지 않는다는 점을 보여줍니다.
문제 3. 평균 R-squared from 5-fold CV: 0.9410749963762098

Process finished with exit code 0
```

SAMSUNG SW ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

내일 방송에서 만나요!

삼성 청년 SW 아카데미