오존예측모델

2018010
김영동
Github address
https://github.com/2018010-youngdong/Ozone_Predictor

1. 오존예측모델 개발의 목적

오존예측모델은 유아 혹은 폐 기능이 약한 노인분들이 주 대상으로, 온도와 바람의 수치를 독립 변수로 사용하여 오존의 양을 예측하는 모델이다.

고농도의 오존은 기관지염 등 조직을 손상시키거나 폐 기능을 약화시키는데 오존은 눈에 보이지 않아 오존이 많은지 적은지 알 방법이 없다. 오존예측모델로 사람들은 오존이 많은 날을 그날의 바람과 온도로 예측하여 바깥활동을 줄여 몸에 해로운 영향을 막을 수 있다.

2. 오존예측모델의 네이밍의 의미

한번 보면 뜻을 바로 알 수 있도록 오존을 예측하는 머신러닝 모델이라는 의미로 간결하게 오존예측모델이라고 지었다.

3. 개발 계획

데이터 파일이름은 airquality로 아래 사진과 같은 데이터를 갖고 있다.

	Ozone	Solar.R	Wind	Temp	Month	Day
0	41.0	190.0	7.4	67	5	1
1	36.0	118.0	8.0	72	5	2
2	12.0	149.0	12.6	74	5	3
3	18.0	313.0	11.5	62	5	4
6	23.0	299.0	8.6	65	5	7

	Ozone	Solar.R	Wind	Temp	Month	Day
count	111.000000	111.000000	111.000000	111.000000	111.000000	111.000000
mean	42.099099	184.801802	9.939640	77.792793	7.216216	15.945946
std	33.275969	91.152302	3.557713	9.529969	1.473434	8.707194
min	1.000000	7.000000	2.300000	57.000000	5.000000	1.000000
25%	18.000000	113.500000	7.400000	71.000000	6.000000	9.000000

데이터에 결측치가 있는지 확인하여 삭제하고, scikit-learn의 선형회

귀모델을 사용할 것이다.
여기서 선형회귀모델은 예측값과 실제값 간의 오차를 최소화하는 선형 함수의 계수를 찾는 최소제곱법을 사용하여 선형회귀를 구현한 것으로 데이터셋의 크기가 작거나 특성이 많지 않을 때 효과적이다.
이 머신러닝 모델이 성공적으로 예측한다면 그래프는 우상향 대각선으로 나타날 것이다.
사용할 성능 지표는 평균제곱오차 mse (mean squared error)로 모델의 예측값과 실제값 간의 차이를 제곱한 후에 평균을 구한 값이다.
mse가 작을수록 모델의 예측이 실제값과 가깝고 이를 확인해볼 것이다.

4. 개발 과정

```
filename = './data/airquality.csv'
data=pd.read_csv(filename)

print(data.isnull().sum())
data = data.dropna(subset=['Ozone', 'Solar.R'])
```

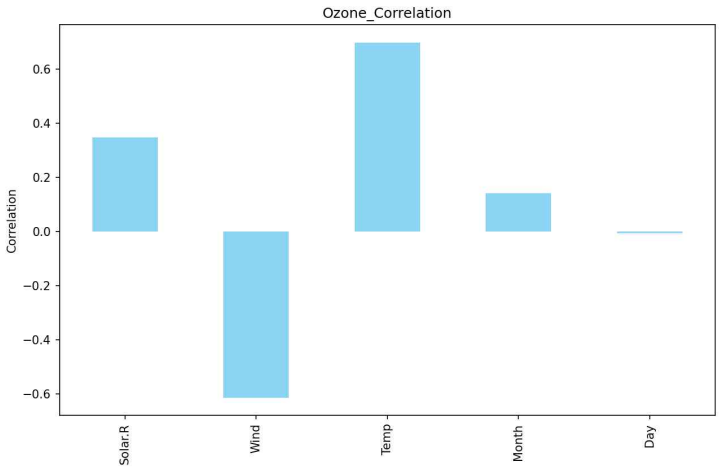
```
Ozone      37
Solar.R     7
Wind        0
Temp        0
Month       0
Day         0
dtype: int64
```

airquality 파일을 불러오고 data라는 변수로 저장했다. 이후 오류발생 원인에서 다룰 결측치를 오른쪽 사진과 같이 확인하고 제거했다.

```
correlation_matrix = data.corr()
correlation_ozone = correlation_matrix['Ozone']
plt.figure(figsize=(10, 6))
correlation_ozone.drop('Ozone').plot(kind='bar', color='skyblue')
plt.title("Ozone_Correlation")
plt.xlabel("x")
plt.ylabel("Correlation")
plt.show()
```

corr함수를 통해 오존과 각 변수간의 양과 음의 상관관계를 알도록 했고 그래프의 가로 세로 크기를 지정하고 파란색의 바 형태로 지정하였다.
아래 그래프는 그 결과로 가장 큰 양의 상관관계는 Temp로 고온에서 오존의 양이 증가함을 알 수 있고 가장 작은 음의 상관관계는 Wind로

바람이 많이 불 때 오존의 양이 감소함을 알 수 있다.

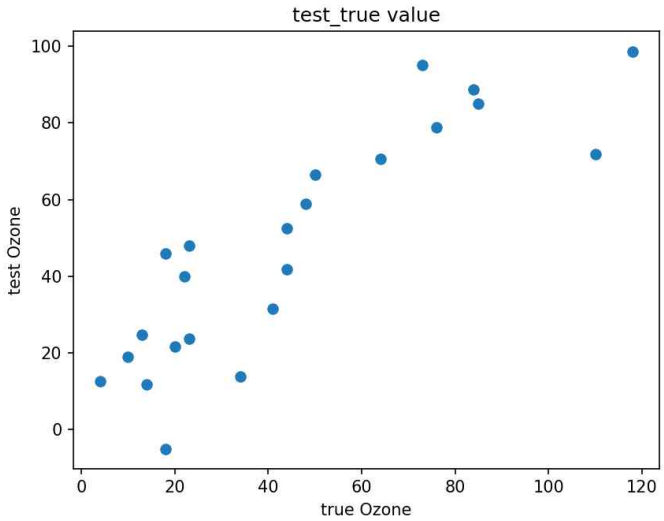



```
X = data[['Temp', 'Wind']]
y = data['Ozone']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
print(mse)
```

온도,바람 독립변수와 오존 종속변수를 지정하고 학습용과 테스트용 데이터로 나눠 선형회귀모델을 생성하고 학습시켰다. 이때 다운받은 데이터에 결측치가 있는 행이 존재하여 모델 학습에서 오류가 발생하였다. 따라서 isnull로 Ozone과 Solar.R에 결측치를 확인하고 dropna로 제거해주었다. 이후 테스트 데이터로 예측을 해 평균제곱오차로 실제값과 비교하였고 그 값은 257.22로 데이터의 수가 적어 다소 큰 값이 나왔다.

```
plt.scatter(y_test, y_pred)
plt.xlabel("true Ozone")
plt.ylabel("test Ozone")
plt.title("test_true value")
plt.show()
```

코드에서 예측값과 실제값을 시각화하여 아래 그래프처럼 산점도로 표시했다.



5. 개발 후기

실제 데이터는 깨끗한 형태가 아닌 결측치, 이상치, 범주형 데이터 등이 존재할 수 있으므로 전처리 단계가 중요하다는 점을 느꼈고 많은 양의 숫자 데이터를 시각화하여 그래프로 만들면 데이터를 이해가 용이해 시간을 절약할 수 있음을 깨달았다.