**Python Term Project**

20101237 배주현

**1. 분석할 데이터에 대해 설명하고, 해당 데이터 분석의 목적을 서술하시오.**

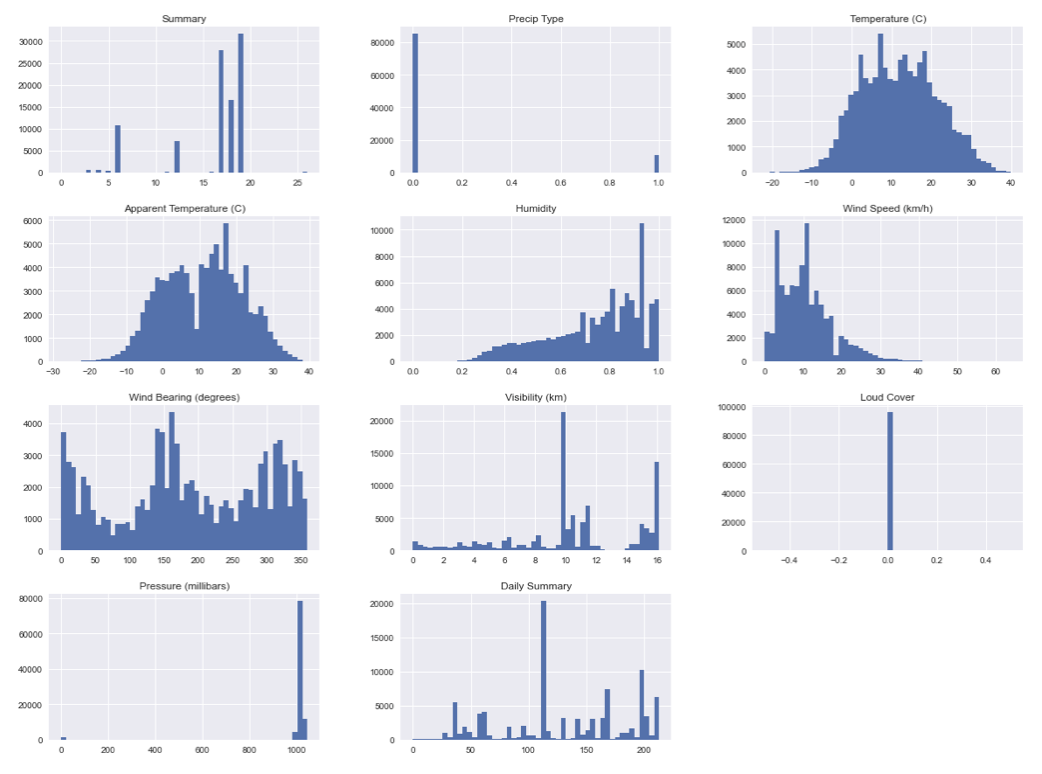
2006년 4월부터 2016년 9월까지의 특정 지역의 날씨를 기록해놓은 데이터를 사용하여 기온, 습도, 풍속 등의 값을 통해 체감­­­­­­ 온도를 예측한다.

**2. 위 데이터를 분석하기 위한 파이썬 코드를 작성하고 해당 코드 전체 동작에 대해 간략히 설명하시오.**

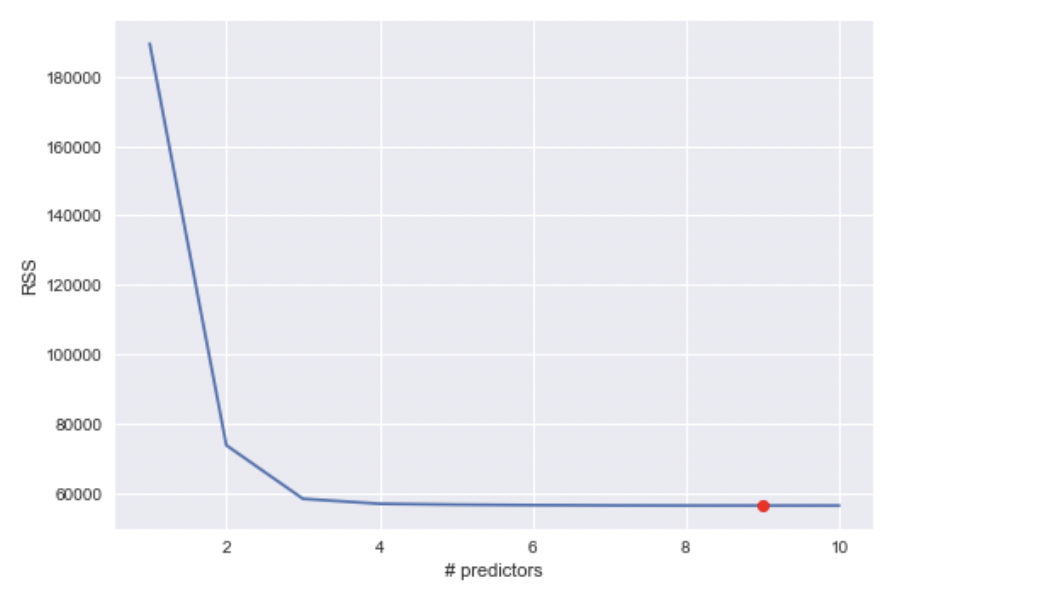
weatherHistory.csv 파일을 읽어오고 숫자가 아닌 feature를 라벨화한 후, validation set approach를 사용한 forward stepwise selection으로 RSS 값을 최소화한는 p개의 predictor를 뽑아낸다. 그 다음 p 개의 predictor subset으로 least squares 기법을 사용하여 모델을 학습시키고 cross validation으로 MSE 값을 알아낸다.

**3. 파이썬 코드를 통해 실행한 결과들을 바탕으로 데이터 분석 결과를 서술하시오.**

Predictor들을 라벨화 한 후 값의 분포를 나타낸 그래프는 다음과 같다.



Forward stepwise selection을 수행한 결과, 9개의 predictor를 가진 모델의 estimated test MSE 값이 가장 작게 나왔으나 4, 5, 6, 7, 8, 9 개의 predictor에서의 값이 모두 비슷하다는 것을 알 수 있다. 따라서 해석이 용이하도록 가장 작은 값의 predictor를 가진 모델을 선택하였다.



텍스트이(가) 표시된 사진

자동 생성된 설명

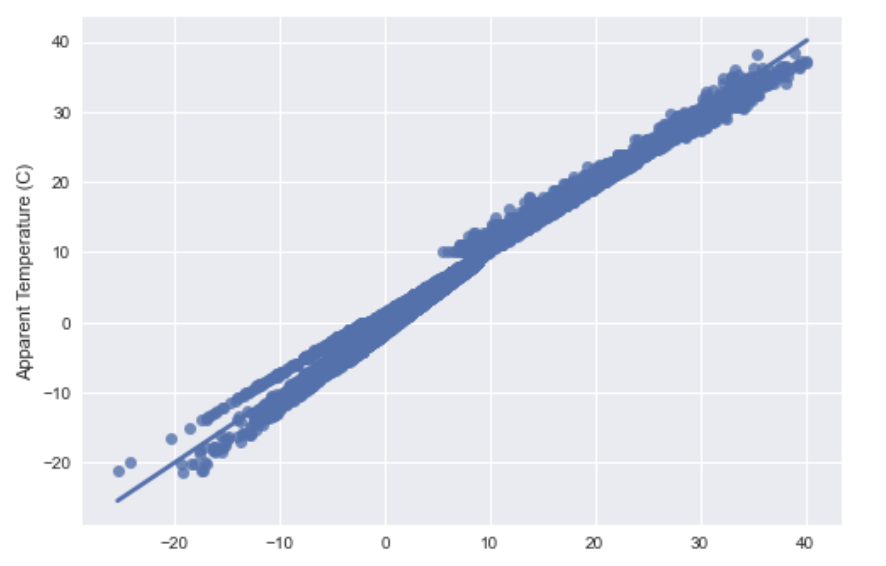
 따라서 10개의 predictor들 중에 'Temperature (C)', 'Wind Speed (km/h)', 'Pressure (millibars)', 'Precip Type' 이렇게 네 가지의 feature를 사용해서 모델을 학습시켰다.

각각의 값의 coefficient 값과 모델의 MSE 값은 다음과 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명

모델의 예측값과 실제값을 그래프로 나타내면 다음과 같다.­­­­­­­­­­­­­­­



**4. 위 파이썬 코드에서, 수업시간에 실습한 library code 및 내용들이 어떻게 활용되고 반영되었는지 설명하시오.**

우선 pandas (Python Data Analysis Library)를 사용하여 csv 파일을 Dataframe으로 처리할 수 있도록 했고 pandas의 dropna, drop\_duplicates 함수를 사용하여 null 값인 열과 중복 값을 처리했다. 그 다음 데이터의 정보를 파악하는 데에 pandas의 columns, info(), shape, describe() 함수를 사용하였고 데이터 정보를 파악한 후 sklearn의 preprocessing의 LabelEncoder 라이브러리를 사용하여 데이터 타입이 Object인 feature 값들을 숫자값으로 처리할 수 있도록 전처리했다. 그리고 모델을 학습시키고, MSE 값을 계산하고 cross validation을 수행하는 데에는 sklearn의 linear\_model, model\_selection, metrics의 LinearRegression, train\_test\_split, mean\_squared\_error 라이브러리를 사용하였다. 마지막으로 그래프를 그릴 때에는 matplotlib의 pyplot 라이브러리와 seaborn을 사용하였다.