**Machine Learning & Data Mining, Spring 2020**

**중간대체 리포트**

사이버보안학과

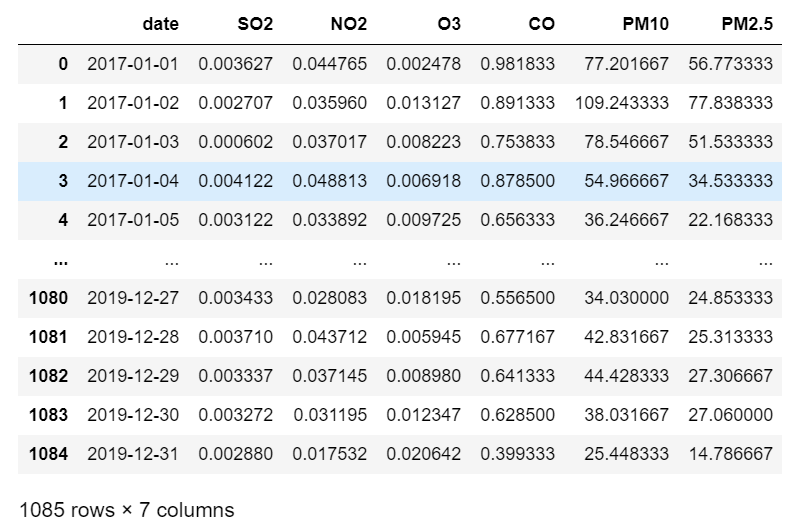
201620641

유 상 정

**과제 목표**

* 프로젝트주제로 이용할 만한 data를 EDA후 분석
* A4 3~5page 분량

1. 서울시 대기오염 dataset

이 dataset은 2017년 1월 1일부터 2019년 12월 31일까지 서울시의 일 평균 이산화황, 이산화질소, 오존, 일산화탄소, 미세먼지, 초미세먼지 수치를 나타낸 data이다. 이 data를 통해서 서울시의 지난 3년간의 대기오염 수치 변화를 파악할 수 있다.

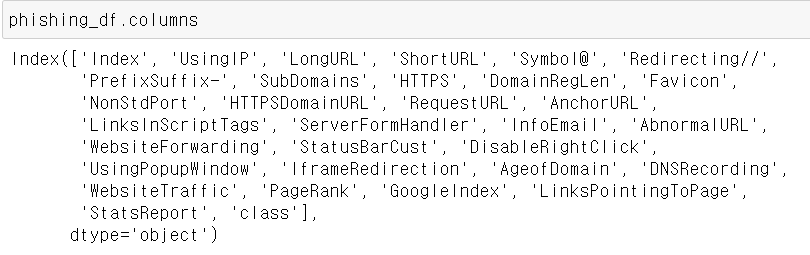
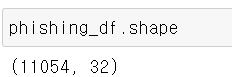


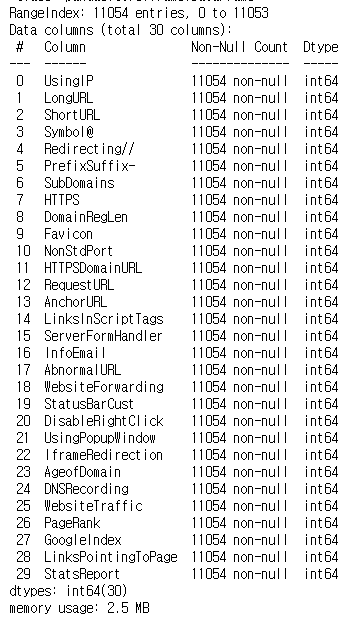
옆의 heatmap은 각 feature간의 상관관계를 알 수 있다. 이 heatmap을 통해서 각 오염 물질 사이의 상관관계를 파악하여, 관계를 feature만을 학습 후, 오염물질의 오염정도를 파악하는 것을 행할 수 있다.

본인은 이 data를 통해서 요즘 이슈가 되는 초 미세먼지를 예측하는 것을 만들 수 있다고 생각된다. 옆의 heatmap을 보면 초미세먼지와 가장 큰 상관관계를 보이는 것은 PM10(미세먼지)와 CO(일산화탄소)와 강한 양적 상관관계를 보인다. 본인은 class를 옆의 기준을 통해서 4가지로 나눈 후, 학습시키고 초미세먼지의 나쁨치를 예측할 수 있을 것이라고 기대한다.

이를 통해서 환경부가 초미세먼지를 줄이고자 한다면, 일산화질소를 함께 줄여야 함을 느끼고 미세먼지와 일산화탄소를 함께 줄이는 정책을 펼칠 수 있을 것이라고 판단하여 유의미한 dataset이라고 생각된다. 또한 이 data를 학습시키고 초미세먼지를 예측하게 된다면 그날, 그날의 미세먼지 농도를 바로바로 알려줄 수 있게 된다. 물론 위의 데이터로 미세먼지뿐만 아니라 이산화황과 이산화질소, 오존과의 관계를 통해서 이산화황의 증가가 오존파괴현상을 예측하거나 대기 중의 이산화질소의 농도 등을 예측할 수도 있다.

1. phishing사이트 dataset

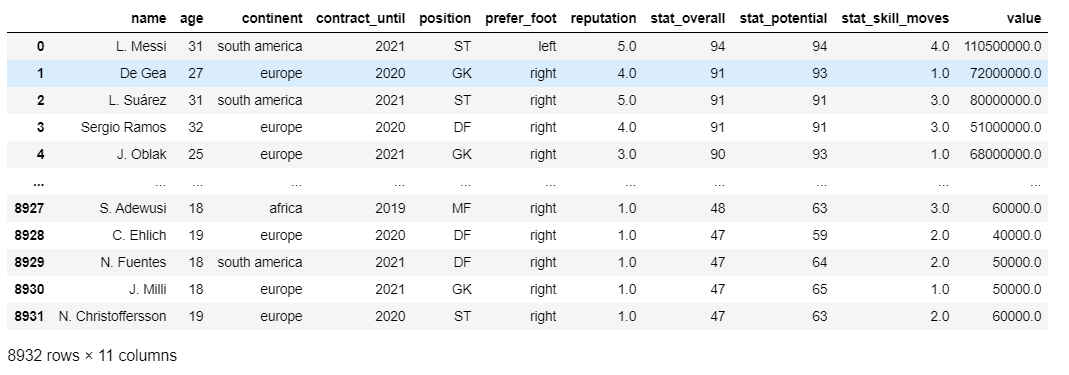
이 dataset은 sample수 11054개이며 feature수가 32개인 피싱사이트 판별 dataset이다. Feature중 index와 class를 빼면 총 30개의 feature이다. 이 feature들을 통해서 피싱사이트를 판별하고, class = 1이면 정상사이

트이고 -1이면 피싱사이트임을 판별한다. 그래서 dataset의 target을 class로 정하고 학습을 통하여, 피싱사이트를 예측할 수 있다.

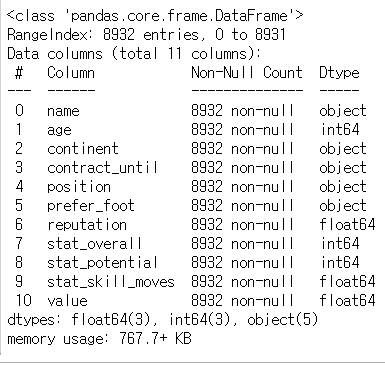
Feature의 특징을 보면 null값이 존재하지 않고, 모두 int type으로서 모두 classification적인 feature들이다. 따라서 이러한 feature일때는 correlation coefficient의 관계로 feature를 drop하기는 어렵다고 판단했다. 그래서 feature를 버리기를 판단하기에는 직접 feature를 버렸을 때의 평가지표와 모든 feature를 사용했을 때의 평가지표를 비교, 분석하여서 feature를 버릴지 말지를 결정해야한다고 생각된다.

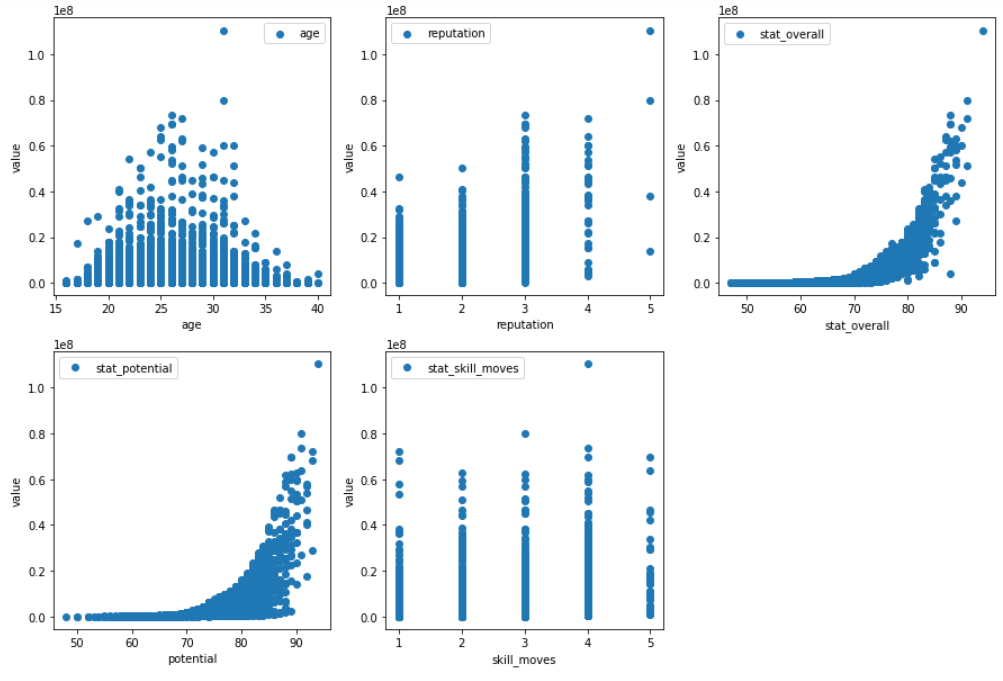
이러한 특징을 가진 이 dataset을 통해서 좋은 성능을 가진 피싱사이트를 판별하는 classifier를 만들게 되면 실제 인터넷 사용자들이 잘못된 사이트를 들어가는 것을 막을 수 있다고 생각된다. 즉, 이 dataset을 통해서 피싱사이트 예측기를 만들 수 있다.

1. 해외 축구 선수 몸값 예측 dataset



이 dataset은 해외축구 선수의 몸값을 나타낸 dataset이다. 각 feature에는 선수의 남은 계약기간, 국가, 포지션, 선호발, 현재 스텟,명성 ,잠재력, 감각적 움직임과 target이 될value(몸값) 이 있다.

data.info()를 확인해보면 각 data type이 나타나고, NULL값이 없음을 확인할 수 있다. 그리고 축구선수의 몸값측정에는 이름과 대륙과 관련이 없으므로 name feature, continent feature는 버린다. 또한 prefer\_Foot도 몸값과 관련이 없는 feature이므로 버린다. 그리고 position 또한 몸값과는 관련 없다고 할 수 있다. 마지막으로 남은 계약기간도 몸값과는 관련 없다고 판단하여 버린다.



실제로 위에서 판단한 feature들을 버리고 각 feature와 선수 가치의 상관관계를 그래프로 나타냈다. 확인해본 결과 예상과는 달리 age와 stat\_overall, stat\_potential만이 value와 관련 있는 feature라고 판단된다. Reputation은 높을수록 몸값에 영향이 있을 줄 알았지만 데이터 분포를 보니 그렇지 않다. 또한 skill적인 움직임 또한 값에 상관없이 4일때의 outlier값을 제외하면 1~5의 value 분포가 비슷한다. 따라서 age와 stat\_overall, stat\_potential만을 이용하여 몸값예측을 할 수 있을 것이라고 판단된다. 이 dataset은 regression dataset으로서 linear regression을 이용하여 학습하고 예측하면 좋을 것이라고 생각한다. 이 data의 활용은 실제로 구단이 어떠한 선수를 데리고 오고 싶어서 스카우터에게 제시할 금액을 예측할 때 활용할 수 있다.