(Essay Summary : Data Science Project in Pharmaceutical R&D)
(DScan을 측정하는 방법 대신 Logistic Regression을 사용하자)

일반적으로 파킨슨 병의 유무를 판단하려면 DScan를 파악 해야 한다. DScan을 사용해 파킨슨 병에 걸렸는지 파악하는 것은 정확하지만 비용이 많이 드는 방법이다. 그렇기 때문에 새로운 대체 biomarker를 찾는 것이 필요하다. 그렇다면 어떻게 파킨슨 병과 관련이 있는 다른 biomarker를 발견할 수 있을까?

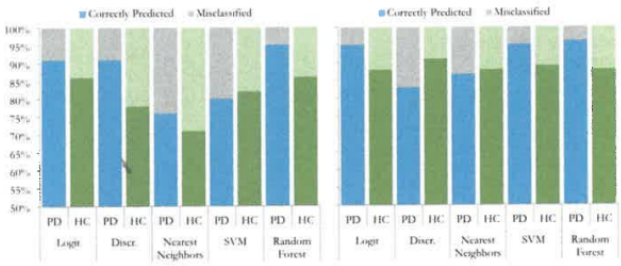
해당 프로젝트는 stepwise regression을 사용해서 데이터를 분석했다. 데이터의 feature를 추가/제거하면서 모델 성능을 판단해 해당 feature를 포함/제거 여부를 결정했다. 데이터의 feature는 76개의 수치형 변수와 17개의 범주형 변수로 구성되어 있다. stepwise regression을 수치형 변수, 범주형 변수 그리고 수치형 변수와 범주형 변수 모두를 포함하는 모델 순서로 필수적인 feature가 무엇인지 분석했다.

먼저 수치형 변수부터 살펴보자. 각각의 feature들과 DScan 및 HC/PD와의 상관관계를 파악해 p값이 0.05보다 작은 feature들을 선택했다. 선택된 feature는 4가지로, the Hoehn and Yahr Motor Score(NHY), Unified Parkinson Disease Rating Score(NUPDRS3), the University of Pennsylvania Smell Identification Test(UPSIT4) 그리고 Tremor Score(TD) 이다. 이 변수들을 이용해 linear regression을 수행한 결과 TD는 HC/PD와 높은 상관관계를 보여주지만 좋은 변수는 아니라고 판단했다. TD의 데이터를 살펴보면 이상치가 존재하는데 이 이상치를 제거하면 상관관계가 많이 떨어지기 때문이다. 따라서 모델이 포함할 변수는 NHY, NUPDRS3, UPIST4와 NHY & NUPDRS3의 cross-term이다.

다음 범주형 변수를 살펴보자. 이전 단계와 마찬가지로 stepwise regression을 통해 총 3가지 변수, REM Sleep Behavior Disorder(RBD), Neurological disorder과 Skin disorder가 선택되었다. 여기서 Skin은 조금 애매한 점이 있는데 PD와 관련이 있어 보이지 않지만 regression 알고리즘이 모델의 성능을 향상시키는 변수라고 판단해 남겨두었다.

최종적으로, 위의 두 분석을 거치면서 선택된 NHY, NUPDRS3, UPIST4, NHY & NUPDRS3의 cross-term, RBD, Neurological, Skin을 모두 포함해 stepwise regression을 수행했다. 그 결과, 낮은 weight와 높은 p-value를 갖는 NHY & NUPDRS3의 cross-term, NUPDRS3, RBD, Neurological을 제외했다. 따라서 NHY와 UPIST4가 가장 좋은 예측 변수라고 할 수 있다.

Linear regression과 비교해서 Logistic Regression을 수행한 결과 5가지 범주형 변수를 사용하지 않는 것이 모델의 성능을 향상시킨다는 것을 알아내었다. 또한 여러 classification model들의 성능을 비교해서 SVM과 Random Forest가 가장 성능이 좋다는 것을 알 수 있었다. 하지만 Logistic Linear Regression과 차이가 거의 나지 않으므로 사용하기 쉬운 Logistic Regression을 사용하는 것을 추천하고 있다.



DScan은 파킨슨 병에 걸렸는지 알아내기 쉬운 biomarker이지만 그 만큼 비용이 많이 들어 DScan의 감소에 영향을 주는 다른 요소를 찾기 위해 해당 프로젝트가 수행되었다. 결과적으로 수치형 변수인 NHY와 UPIST4만을 사용하는 것이 모델의 예측 성능이 좋았고 Logistic Regression 모델을 사용하는 것이 좋다고 판단할 수 있었다.

```
"Written by Lee, Da-In"
```