DrivenData 캐글처럼 데이터분석과제 올라오는 사이트!

[Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines](https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/)

코로나가 백신개발중인게 전세계적으로 가장 주목받는 핫이슈인 상황이니까

이전에 가장 유행했던 신종플루 백신과 계절독감 백신을 맞은 사람들의 설문 데이터를 가지고 이사람이 두 백신을 맞았는가 / 하나만 맞았는가 / 둘다 안맞았는가를 예측하는 문제

**장점 :** 시의적절한 주제! EDA할것도 많아보이고 모델링도 다양하게 진행할 수 있을듯.

그리고 범주형 변수도 많고 binary 분류문제인만큼 범주하다는 장점

**걱정되는 점:** 작년 범주팀에서도 설문조사 데이터를 다루고 binary 분류 문제 다룸

**그래서 받은 피드백**:

코로나나 코로나 백신과 관련된 데이터들을 추가해서 주제를 조금더 확장시키는것도 좋을 것 같다! -> 근데 주제 확장은 좀 어렵지 않을까? 데이터만으로도 꽤나 귀찮을것같다

작년 범주랑 겹치는 걱정은 안해도 될 것같다 데이터 특성보다는 분석의 가치가 있는지, 어디에서 이전 범주랑 다른 분석방향(논리, 새로운 분석방법 등) 을 잡아낼수있는지가 중요한것같다!

- 단순예측 이상의 해석을 이끌어낼 수 있을까? – 작년 범주도 SHAP value로 보완하려고 했던 부분인데 복잡한 신러닝 모델이 어떤 알고리즘으로 왜 이런 예측을 냈는지 알고자 할 수 있는 방법들이 있는 것 같다! 이런 부분들을 알아가면 도움 될 것 같다는 생각

[*https://blogsaskorea.com/78*](https://blogsaskorea.com/78) *:* ***사후 모델링(post-modeling) 해석력(사후 해석력)*** *랜덤 포레스트의 변수 중요도 출력(variable importance output)과 같은 특정 모델에 특화된 기법 외에도 부분 의존성(PD; Partial Dependence) 플롯, 개별 조건부 기대치(ICE; Individual Conditional Expectation) 플롯, LIME(Local Interpretable Model-Agnostic Explanations) 등 특정 모델에 구애 받지 않는 기법 등이 있습니다.*

[*https://blogs.sas.com/content/saskorea/2018/07/27/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%ED%95%B4%EC%84%9D%EB%A0%A5-%EC%8B%9C%EB%A6%AC%EC%A6%88-3%ED%83%84-%EB%B6%80%EB%B6%84%EC%9D%98%EC%A1%B4%EC%84%B1pd-%EA%B0%9C%EB%B3%84%EC%A1%B0%EA%B1%B4/*](https://blogs.sas.com/content/saskorea/2018/07/27/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%ED%95%B4%EC%84%9D%EB%A0%A5-%EC%8B%9C%EB%A6%AC%EC%A6%88-3%ED%83%84-%EB%B6%80%EB%B6%84%EC%9D%98%EC%A1%B4%EC%84%B1pd-%EA%B0%9C%EB%B3%84%EC%A1%B0%EA%B1%B4/)

-분석방향을 어떻게 잡을 것인가?

두가지 모형을 만들어야함(계절독감 백신과 신종플루 백신)

=> 백신 접종 여부에 대해서는 “건강신념모형” 이라는 심리학적 개념 있음

건강 신념 모델(HBM,health belief model)은 특히 건강 서비스의 채택과 관련하여 건강 관련 행동을 설명하고 예측하기 위해 개발된 사회적 심리적 건강 행동 변화 모델이다. HBM은 1950 년대 미국 공중 보건국의 사회 심리학자에 의해 개발되었으며 건강 행동 연구에서 가장 잘 알려져 있고 가장 널리 사용되는 이론 중 하나이다. HBM은 건강 문제, 행동의 이점과 행동의 장벽에 대한 사람들의 신념, 건강 증진 행동에서의 참여 (또는 참여의 결여)를 설명하는 자기 효능을 제안한다. 건강을 증진시키는 행동을 유발하기 위해서는 자극, 즉 행동의 단서가 있어야 한다

이런 도메인적 지식들 공부해가면서 **파생변수도 생성**

클러스터링 가능할까? - EDA용으로 사람들 유형 나눠보는 것도 가능하지 않을까?

나이브베이즈분류모형 스태킹앙상블모형 등등 다양하게 머신러닝 모델 사용해보자

전처리 관련

결측치 처리 어떻게 할 것인가?

범주형 변수 인코딩 어떻게 할 것인지?

클래스 불균형 문제 해결하기 위한 샘플링 방법도 적용 가능할 듯

\*multi label problem

Y변수가 두개이기 때문에 각각 진행해도 문제 없지만 두 변수가 의존적인 경우 정보를 무시하는 것이 되기 때문에 이걸 보완할 방법론들이 필요함

<https://tonnykwon.github.io/blog/machine/2020-01-16-Multi-Label%20Classification/> : 관련되어 여러 방법들 소개하는 듯

전 범주팀은 결측치 해결이 되게 크리티컬한 문제여서 거기에 힘을 많이 주었다면, 우린 어디에 포인트를 잡고 진행 할 수 있을까??? 고민고민