우리 회사 AI시스템에 설명가능성을 요구한다면?

휴먼지능정보공학과 201910803 박채희

AI면접 시스템 개발 회사의 당/락 결과에 대해, 취준생들이 근거를 요구하는 상황을 가정해보았습니다. 이런 경우에는 feature based approach가 가장 적합하게 근거를 설명할 수 있을 것 같습니다.

당연히 평가이기 때문에 당/락 결과를 결정하는 질문과 대답 등의 요소들을 feature라고 가정해보았을 때, 특별히 중요한 feature가 있을 것입니다. 그 feature에 대해서 XAI의 기술인 feature importance를 사용해 feature 중요도를 근거로 삼을 수 있을 것입니다.

또한, 취준생 지원자들마다 feature의 값이 다를 것인데, 그 다른 값들을 변형시키면서 AI시스템의 결과가 어떻게 변화하는지 관찰하고 그래프로 표현하는 PDP 부분의존도 방법을 사용하여 어떤 부분이 당/락을 결정하였는지 근거로 보여줄 수 있을 것 같습니다.

마지막으로는 취준생 지원자들의 특정 feature값빼고 나머지를 고정시켜 feature의 중요도를 알 수 있는 SHAP방법도 사용하여 근거를 제시할 수 있습니다. 예를 들어, 여러 명의 지원자들을 비교할 때 feature을 하나씩 제거하여 일정 수준 이상이면 합격인 기준으로 당/락에 영향을 미치는 feature을 파악할 수 있을 것 같습니다.