1. 문제 정의

카드회사에 사례를 들면서 문제 정의 및 프로잭트 이유 설명

1. 데이터 수집

카드 데이터를 받았다~

1. 데이터 EDA& 전처리
2. 이 데이터의 변수들이 무엇을 의미하나 설명
3. 직접 어떤 분포를 가지나 시각화해서 각각 설명
4. 이거를 통해서 ‘데이터를 정규화 해야한다’ 설명
5. 오버샘플링도 해야한다고 설명
6. 좀 더 자세히 보기 위해 상관관계 분석을 했다고 설명
7. 봤더니 상관관계가 높게 나온 변수들이 있어서 상관관계가 높은 변수들을 편차로 합친 새로운 파생변수를 만들었다고 설명(대조군1)
8. 모델링
9. 모델 설명
10. 위에서 상관관계가 높게 나온 변수들이 실제로 얼마나 모델에 영향을 줄지 궁금하여 상관관계에 영향을 많이 받는 ‘나이브 베이즈’모델을 선택
11. 가장 대표적인 로지스틱 회귀를 사용하여 다른 모델들을 객관적으로 비교하기 위해 선택하였다고 설명
12. 상대적으로 중요한 변수를 선정 및 순위 매기기 가능한 랜덤 포레스트 모델을 선택하였다고 설명
13. 랜덤 포레스트 모델
14. 하기전 변수 중요도를 보기위해 FEATURE IMPORTANTCES를 봄
15. 보았더니 최근 소비패턴이 중요하고 나머지 변수들은 중요도가 없기 때문에 전처리과정을 한번 더 한다고 설명
16. 데이터 재 전처리
17. 최근 소비패턴 관련 변수만 남긴 대조군2와 최근소비패턴만 남기고 편차를 합친 대조군3 데이터를 새로 생성
18. 데이터 재 모델링
19. ‘랜덤 포레스트’부터 4가지 데이터를 오버 샘플링한 것 안한 것(2종류)으로 나눠(총 8가지) F1 스코어로 모델의 성능 평가함
20. ‘나이브 베이즈’ 모델을 통해 8가지 데이터를 분석한 결과 실제로 변수들의 상관관계가 모델링에 영향을 많이 끼친다고 결과 도출
21. ‘로지스틱 회귀’ 모델을 통해 객관적으로 두가지 모델과 비교를 해봤다고 설명
22. 랜덤 포레스트 모델이 다른 모델에 비해 제일 좋았다고 결과 도출
23. 모델링 추가
24. 랜덤 포레스트가 제일 적합한 모델로 보였기 때문에 랜덤 포레스트 모델에 상위 모델인 XGBOOST를 이용하면 더 좋은 F1스코어를 얻을 것이라고 가설 세움
25. XGBOOST를 실제로 돌려봤더니 랜덤 포레스트 모델과 생각보다 다르지 않았다고 결과가 나왔음
26. 실험 계획법
27. 실제로 이 데이터를 예측하는데 데이터 전처리와 모델 설정이 영향을 주는지 알아보기 위해 이때까지 나왔던 F1 스코어로 ‘분산 분석’을 진행한다고 설명
28. 분산 분석을 돌려본 결과 ‘변수 제거’와 ‘편차 파생 변수 생성’의 P-VALUE가 0.829로 높게 나왔고 이 두개를 동시에 한 것을 제외한 것들이 예측력 향상에 효과가 있었다고 결과 도출