-

|  |  |  |
| --- | --- | --- |
| 실습제목 | 신용카드 사용자 연체 예측 AI 경진대회 | 일련번호4-0011 |
| 실습목적 | 신용카드 사용자들의 개인 신상정보 데이터로 사용자의 신용카드 대금 연체 정도를 예측 | |

|  |
| --- |
| 21.11.01일자 |
| 팀원:신문혁,정기호,송승한,최지호 |
| <프로젝트 개요>   1. 신용카드 사용자 데이터 EDA 2. 신용카드 사용자 데이터 전처리 3. 모델을 통해 사용자 Credit 평가/분류 4. Credit 분류 값 결과분석 |
| <프로젝트 목적>  신용카드 신청자가 제출한 개인정보 데이터와 대금 연체 정도를 예측할 수 있는 인공지능 알고리즘 활용해 신용 점수를 산정함. 이 신용 점수를 활용해 신청자의 향후 채무 불이행과 신용카드 대급 연체 가능성을 예측 가능할 것으로 예상 중. |
| <실습 내용> |
| 신문혁: Encoding |
| 정기호: Encoding |
| 송승한: Encoding |
| 최지호: 일지작성 |
|  |
| <회의내용> |
| 정: 소득은 ‌income\_scale해야한다  Id 중복은 이미 다 처리가 되었으니 drop 진행해야 한다.  정: 1등 코드는 drop 안해서 인코딩 했으니 drop하면 된다. 날자 관련 된 건 선택적 drop 해야함  정: encoding 진행해야 하나.  신: encoding 나눠서 ordinal object만 돌리면 된다.  변수 (남자 여자 저/고학년 대학원 이상) 각자 번호 부여하는데, 항목당 컬럼으로 변환시키는 것(숫자 컬럼화) . catboosting은 이를 이용한다. 하지만 진행할 필요 있음.  Training set만 세 개 이므로 없을 가능성이 더 높다. beginmonth업데이트되면서 직업변경/무직으로 변질 되었을 가능성 있음. 모델링을 어느 쪽을 채택할건지 (왜 떨궜는지) 정해야 한다  정: 0 과 1 사이에 1000이 나오는데 차이가 너무 많아서 그 차이를 보려 outliner을 보려 진행함.  min-max scaler/standard scaler/ robust scaler 등 여러가지 방식이 있는 중 income에서 어떤 scaler 방식을 적용해야 하는지 고민 중 robust 시도함.  (0.1에서 099퍼가(그래프의 한쪽에 극단적으로 분포) 비정상적 분포되어 있을 시 조정하는 방법이 robust scaler)  1. 길은 어디며,  2. robust 진행 시 outliner 진행해야 하는지,  위 2가지를 상대로 실험중임  Log scaler 사용시 이상치가 확 감소(1위 코드)함. Robust 진행 시 분포에 변환이 있는지 확인이 필요함.  어느 쪽 방식이 더 좋은지 비교 필요함.  Robust는 사용 안함-> data 범위가 넓어짐. 이가 오히려 모델링을 해치는 듯 함.  결론적으로 log scaler 진행해야 할 듯. |
| <대회 진행 일정> |

|  |  |  |
| --- | --- | --- |
| 기록자:최지호 | 점검자 | 점검자 |
| 일자 2021년 11월1일 | 일자 2021년 11월1일 | 일자 2021년 11월1일 |

|  |  |
| --- | --- |
| 실습 후기 | * 수상 코드 리뷰 및 분석 후 적용 시 유용할지 확인 필요. * 모르는 코드 있을 시 검색(구글)및 학습/복습 필요함 |
| 향후 계획 | * XGboost, Catboost, S.Emsemble공부 후 설명 및 자료공유 |