|  |
| --- |
| **제10회「2022 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 ☑ 표시 | | |
| **참가리그** | 🗹 데이터분석리그 | | |
| **세부분야** | □ 이노베이션분야 🗹 데이터분석분야 | | |
| **세부부문**  \*해당시 체크 | □ 루키부문 🗹 퓨처스부문 □ 챔피언부문  \*데이터분석분야에 한함(선택) | | |
| **개인/팀여부** | □ 개인 🗹 팀(총 4 명) | **개인/팀명** | CriticalPoint |
| **지도교사명** | \*루키부문에 한함(선택) | | |
| **대표ID** | rmsmsgood@naver.com | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

|  |  |
| --- | --- |
| **분석 주제명** | 핀다 어플을 이용하는 고객의 대출신청여부 예측,  모델 기반 고객 군집분석 및 군집 별 최적의 서비스 메세지 제안 |
| **분석 배경** | 핀다 어플은 일반인들에게도 금융정보 접근을 쉽게 해주는 대출중계 어플이다. 이러한 특성으로 인해 대출신청이 목적이 아닌 신용등급확인, 대출상품 확인 등의 목적으로 이용되는 경우도 있다. 따라서 어떤 사람,군집이 대출을 신청하지 않는지 등을 조사하고, 여러 원인들을 분석해 더 많은 고객을 대출신청과정까지 유도할 필요가 있다. |
| **분석 내용**  **요약** | 주어진 3개의 데이터는 각각 고객의 인적 정보, 대출상품관련 정보,  고객의 핀다 어플 사용 정보이다. 이 중 공통된 변수를 바탕으로 데이터를 종합하고, 핀다 어플을 설치해서 이용해본 결과 , 한 고객의 기본 인적 정보는 동일하고 여러 대출신청서를 여러 은행에 대출하는 과정을 확인하였다.  세개의 데이터의 통합을 위해 log\_data에서 user\_id별 이벤트들의 횟수를 나타내는 데이터를 새로 만들고, 이 데이터와 나머지 두개의 raw데이터를 통합하였다. 이 데이터를 이용해 변수들간의 관계를 시각화를 통해 파악하고, 이것들을 토대로 예측모델을 만들 예정이다. 그리고 군집분석을 통해 어떤 유형의 고객이 대출을 신청하지 않는지 분석하고 적절한 메시지를 고안할 예정이다. |
| **분석방법 및**  **계획** | **※ 대회기간동안 선택한 주제(부문)에 대한 분석 방향에 대하여 기술해 주십시오**  **- 분석에 활용되는 추가데이터(출처 기재)**  **- 분석에 적용·활용할 통계·분석 기법, 방법론**  **- 분석 결과에 대한 시각화 방법 등**  로지스틱회귀분석  로지스틱회귀분석은 예측할 변수가 범주형 변수인 경우 사용할 수 있는 회귀분석기법이다., 여러 통계량을 통해 변수의 유의성, 모델의 설명력을 알 수 있고, 예측력보단 설명력에 중점을 둔 기법이다.    Is\_applied가 0,1일 확률을 출력해주고, 임계값을 통해 대출신청여부를 결정할 수 있다.  랜덤포레스트  랜덤포레스트는 다수의 의사 결정 트리를 기반으로 분류하는 기법이다.  대용량 데이터 처리에 효과적이고, 상대적으로 중요한 변수들을 선정할 수 있는 기능이 있다. 모델 훈련시간은 길지만, 설명력이 좋은 머신러닝 기법중 하나이다.    그림과 같이 변수의 중요도를 출력할 수 있고, 이 그래프를 기반으로 모델에 사용될 변수를 선정할 수 있다.  FFNN  **FFNN**은 층 간의 모든 노드끼리 연결시키는 덴스 레이어(Dense Layer)만으로 이루어진 신경망이다. 입력값은 은닉층을 통해 한 방향으로만 이동하고, 순환이나 루프가 없다. 퍼포먼스는 가장 베이스라인이 되는 수준이다  XGBoost  XGBoost란 순차적으로 약한 학습자를 추가 결합하여 하나의 강한 모델을 만드는 기법이다. 새로운 약한 학습자를 훈련시킬 때 바로 전 약한 학습자의 학습결과에서 잘못 예측된 샘플에 대한 가중치는 증가시키고 올바르게 예측된 샘플에 대한 가중치는 감소시켜서 사용한다.    모델의 훈련속도가 빠른편이고, 뛰어난 예측 성능이 발휘되는 기법 중 하나이다.  SVM  SVM이란 분류를 위한 결정 경계선을 정의하는 모델이다. 결정 경계선을 기준으로 데이터가 어느 카테고리에 속할지 판단하는 분류 모델이다.    예측에 이용되는 변수의 개수가 많아질수록 결정경계의 차원이 늘어나는 데 이것을 초평면이라 부른다. 서포트 벡터는 결정 경계에 가장 가까운 각 군집의 점들을 뜻하고, 서포트 벡터와 결정경계 사이의 거리를 마진(margin)이라 한다. SVM은 마진을 최대로 만들어서 분류하는 기법이다. 모델 훈련시간은 길지만, 예측력이 좋은 머신러닝 기법중 하나이다. |
| **분석결과 활용 및 시사점** | 고객의 대출신청 여부 및 군집분석을 함으로써 고객이 대출신청여부의 원인을 알 수 있고, 이것을 토대로 고객이 대출하기까지 더 쉽게 할 수 있다. |

※ 제출자료는 평가에 반영 예정