제10회 「 2022 빅콘테스트」 데이터 분석 계획서

* 해당란에 ☑ 표시

참가리그	☑ 데이터분석리그		
세부분야	☐ 이노베이션분야	☑ 데이터분석분야	
세부부문 *해당시 체크	☐ 루키부문 *데이터분석분야에 한함(선택)	☑ 퓨처스부문	☐ 챔피언부문
개인/팀여부	☐ 개인 ☑ 팀(총 3 명)	**개인/팀명**	LstatJ
지도교사명	*루키부문에 한함(선택)		
대표ID	jmss6590@naver.com		

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

분석 주제명	앱 사용성 데이터를 통한 대출신청 예측분석 및 군집분석
분석 배경	핀다는 서류 제출을 하지 않는 비교대출 서비스를 제공하고 있다. 이 서비스는 간단한 몇 가지 정보만 입력하면 빠른 시간 내에 여러 가지 대출 조건을 한 번에 받아 쉽게 비교하여 대출을 받을 수 있다. 이 때 대출 상품별로 고객이 대출 상품을 이용했는지를 예측하는 분석을 진행한다. 핀다 앱의 주된 이용 목적이 대출신청은 맞지만, 모든 고객이 대출신청을 목적으로 이용하고 있는 것은 아니다. 어떤 고객은 자신의 신용도를 확인하는 것이 목적일 수도 있고 또 다른 고객은 기대출정보를 마이데이터로 연결하여 기대출정보를 편리하게 확인하는 것이 목적일 수도 있다. 고객의 특성을 분석하여 특성별로 서비스 메시지를 제공함으로써 이용자들이 더 다양한 기능을 쉽게 이용할 수 있도록 하기 위해 군집분석을 진행한다.
분석 내용 요약	2022년 3월에서 5월까지의 핀다의 홈 화면 진입 고객의 데이터를 통해 2022년 6월 고객의 대출 상품 신청 여부를 예측하고, 유저 정보와 로그 데이터를 이용한 모델 기반 군집분석을 통해 고객의 특성 분석결과를 도출한 후 이를 통해 군집별로 서비스 메시지를 제안
분석방법 및 계획	**안 쓰는 변수 제거** 훈련 데이터셋으로 사용하지 않을 column 제거

loan_result
loanapply_insert_time, bank_id, product_id
user_spec
insert_time, personal_rehabilitation_yn
personal_rehabilitation_complete_yn

근로연월 -> 연차
근로연월을 연차로 변경, ex) 근로연월=20151101.0 -> 연차=7

생년 -> 나이
생년을 나이로 변경, ex) 생년=1985.0 -> 나이=38

미성년자 때부터 경력으로 입력한 데이터 제거
나이와 연차를 비교하여 미성년자 때부터 자신의 경력을 입력한 데이터 제거(만 19세 이상부터 연차를 인정)

대출목적 변수 통일
user_spec 테이블의 대출 목적은 동일한 값이 영어와 한글로 섞여서 입력되어 있음 -> 영어로 통일

대체할 수 없고, 공통 결측치가 존재하는 row 제거
user_spec 테이블에서 나이가 결측치인 row와 성별이 결측치인 row가 동일 -> 제거
근로형태, 고용형태, 주거소유형태, 대출희망금액, 대출목적가 모두 결측치인 85개의 row 제거
위에서 제거하고 남은 연소득이 결측치인 5개의 row 제거

기대출수, 기대출금액이 모두 결측치면 0으로 대체
기대출수와 기대출금액이 모두 결측치이면 기대출이 없어서 공란으로 둔 것으로 판단하여 모두 0으로 대체

연소득이 0인데 연차가 결측치이면 연차를 0으로 대체
연소득이 0인데 연차가 결측치인 데이터는 연차가 0인 것이 일반적이므로 0으로 대체

연소득이 0인데 연차가 존재하는 데이터 제거
연소득이 0인데 연차는 존재하는 데이터 제거(연소득이 0이면 연차는 존재할 수 없다고 판단)

기타소득이 아니고 연차가 결측치인 데이터 제거
근로형태가 기타소득이면 연차는 정의되지 않는 것이 맞다고 판단하여

	기타소득이고 연차가 결측치인 데이터는 보류, 기타소득이 아니고 연차가 결측치인 데이터는 제거 **기대출수 이상치 제거 (Q3+3*IQR 이상인 값을 이상치로 판단)** 기대출수가 0부터 200이 넘는 값까지 존재하는 것을 확인했고 기대출수가 많은 데이터들은 아주 적은 것도 확인했음. 따라서 기대출수의 Q3+3*IQR(28.5) 이상인 값을 이상치로 판단하여 해당 row 제거 **기대출금액이 0인데 기대출수가 0이 아닌 데이터 0으로 변환** 기대출금액은 0이지만 기대출수가 0이 아닌 데이터가 존재하여 기대출수를 0으로 대체(기대출금액이 0이면 기대출수가 0이 아닌 다른 값일 수 없다고 판단) **기대출금액 이상치 제거 (평균+5*표준편차를 초과하는 값을 이상치로 판단)** 기대출금액에서 매우 큰 이상치들이 존재하는 것을 확인했음 따라서 평균+5*표준편차를 초과하는 값을 이상치로 판단하여 제거 **신용점수, 기대출금액 결측치 처리** 신용점수와 기대출금액의 결측치는 모두 10만개 이상이고 결측치를 모두 제거하고 분석할 지, 결측치를 대표값으로 대체하고 분석할 지 판단하기 쉽지 않음. 따라서 결측치를 제거한 데이터셋과 대체한 데이터셋으로 각각 모델을 만들 예정 **- 결측치를 제거하는 경우:** 신용점수와 기대출금액이 하나라도 결측치인 row는 모두 제거 **- 결측치를 대체하는 경우:** 1) 신용점수는 이상치가 존재하지 않고 분포가 정규분포에 가까운 형태를 보이기 때문에 결측치를 전체 신용점수 평균으로 대체 2) 기대출금액이 결측치인 row의 기대출수는 모두 1인 것을 확인했고, 기대출수가 1인 기대출금액의 분포를 보았을 때 오른쪽으로 꼬리가 긴 형태를 나타내고 있고, 이상치를 제거했음에도 boxplot에서 기준으로 제시하는 이상치가 아직 상당 수 존재하여 중앙값으로 결측치를 대체 **승인한도, 승인금리 결측치 제거** 승인한도와 승인금리가 결측치인 데이터는 모두 제거 **중복되지 않는 신청서 번호를 포함하는 row 제거** user_spec 테이블과 loan_result 테이블에서 신청서 번호는 모두 unique한 값인 것을 확인했고, 훈련 데이터셋에서 두 테이블을 모두 활용하기 위해 신청서 번호가 중복되지 않는 row는 모두 제거

모델용 데이터
훈련 데이터셋을 만드는 것이기 때문에 loan_result에서 is_applied가 NaN인 데이터(예측해야하는 데이터) 제거

전처리 된 user_spec 테이블과 loan_result 테이블 merge
전처리 된 각각의 테이블의 신청서 번호로 inner join하여 merge함

더미변수 생성
명목형 변수를 더미변수를 이용하여 가변화

근로형태가 기타소득인 데이터는 연차를 빼고 분석하기 위해 데이터프레임을 나눔
근로형태가 기타소득인 경우, 소득이 불규칙하고 연차를 정의할 수 없는 소득이라고 판단하여 모델 훈련 데이터셋을 연차를 포함하고 근로형태가 기타소득이 아닌 데이터셋과 연차를 포함하지 않고 근로형태가 기타소득인 데이터셋으로 나눔

근로형태가 OTHERINCOME인 데이터는 연차를 빼고 분석하기 위해 데이터프레임을 나눔
근로형태가 기타소득인 경우, 소득이 불규칙한 경우이고 연차를 정의할 수 없는 소득이라고 판단하여 모델 훈련 데이터셋을 연차를 포함하고 근로형태가 기타소득이 아닌 데이터셋과 연차를 포함하지 않고 근로형태가 기타소득인 데이터셋으로 나눔

분석 방법
대출 여부를 예측하는 것이 목표이고, 타겟이 0 또는 1로 나타나는 이진 분류의 형태이다. 이에 따라 로지스틱 회귀분석, 랜덤포레스트, 인공신경망 등을 분석 방법으로 선정

안 쓰는 변수 제거
사용자 기기의 종류와 앱 버전은 분석에 큰 도움이 되지 않는다고 판단하여 mp_os, mp_app_version 변수를 log_data에서 제거

log_data와 user_spec를 통해서 분석을 진행하므로 log_data와 user_spec에 공통적으로 있는 user_id 변수를 통해서 전처리
log_data에는 있지만 user_spec에는 없는 사용자 데이터와

user_spec에는 있지만 log_data에는 없는 사용자 데이터를 제거
군집분석의 해석의 용이성을 위해서 모든 변수를 사용하는 것이 아닌 군집이 중요도가 떨어진다고 생각되는 변수를 user_spec에서 제거

	중요도가 떨어진 변수를 제거한 데이터를 통하여 클러스터링을 진행 **군집 분석 방법** 연속형 변수와 명목형 변수가 모두 포함된 데이터로 군집분석을 할 예정이고 그 방법에는 더미변수 이용, Gower의 방법, Eskin의 방법 등이 있다. 이 방법들로 모델링을 하고 Silhouette 지수, Dunn 지수, 수정된 Rand 지수를 이용하여 모델을 평가하고 선택
분석결과 활용 및 시사점	고객들의 정보에 따른 대출 상품별 대출 여부를 예측함으로써 대출한 경우의 상품 정보를 파악하고 이를 통해 고객들에게 보다 적합한 대출 상품정보를 제공할 수 있다. 현재 핀다는 다양한 기능이 있다. 핀다 이용 고객들을 군집화하여 각 군집의 주된 이용 목적을 파악하고 다른 군집에서 주로 이용하고 있는 기능을 제안하는 서비스 메시지를 통해 고객들에게 핀다의 다양한 서비스 이용을 유도하여 앱 사용성을 높일 수 있다.

※ 제출자료는 평가에 반영 예정