

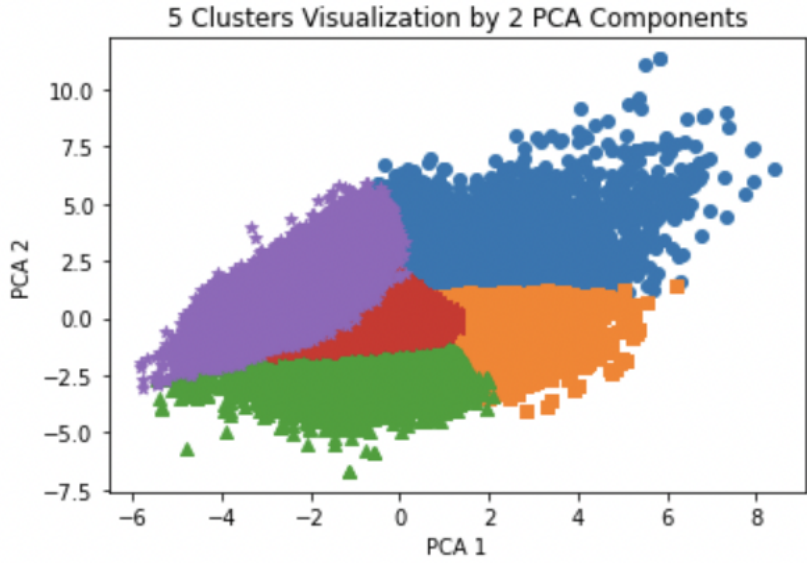
## 제10회 「2022 빅콘테스트」 데이터 분석 계획서

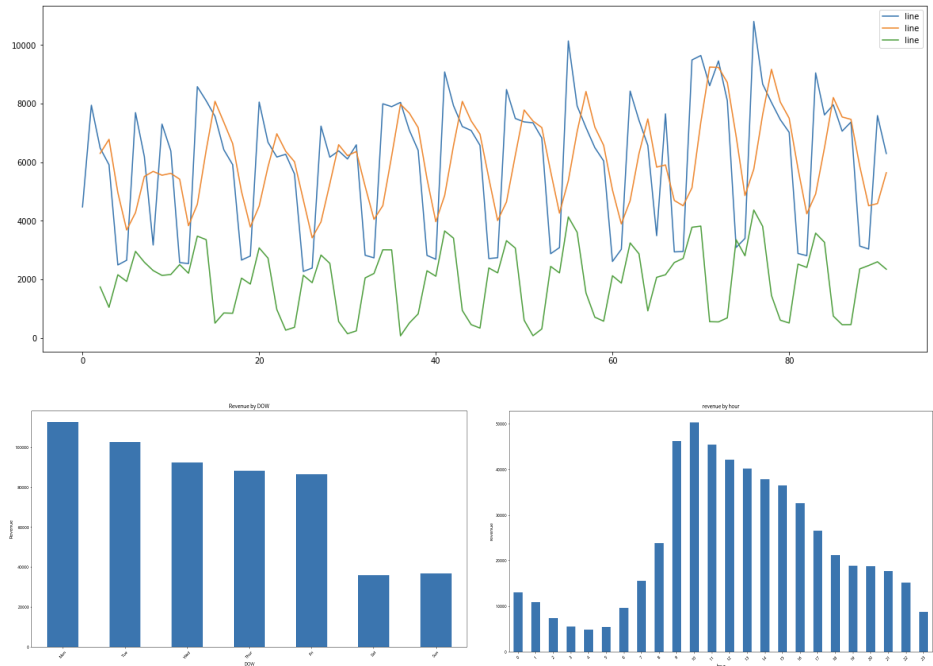
\* 해당란에 ☒ 표시

참가리그	<input checked="" type="checkbox"/> 데이터분석리그		
세부분야	<input type="checkbox"/> 이노베이션분야 <input checked="" type="checkbox"/> 데이터분석분야		
세부부문 <small>*해당시 체크</small>	<input type="checkbox"/> 루키부문 <input checked="" type="checkbox"/> 퓨처스부문 <input type="checkbox"/> 챔피언부문 <small>*데이터분석분야에 한함(선택)</small>		
개인/팀여부	<input type="checkbox"/> 개인 <input checked="" type="checkbox"/> 팀(총 3명)	개인/팀명	우 걱정마세요
지도교사명	<small>*루키부문에 한함(선택)</small>		
대표ID	eric_best@naver.com		

※ 5장 내외로 목차는 준수하여 자유롭게 작성

분석 주제명	앱 사용성 데이터를 통한 대출신청 예측 분석
분석 배경	<p>❖ 선택한 주제(부문)의 데이터 분석 제안에 대한 배경을 기술해주십시오</p> <p>대한민국의 급진적인 경제적인 발전과 동시에 저금리시대, 주택가격 상승기대 등의 수요측면과 금융회사의 대출 경쟁등 공급측면이 동시에 작용하여, 대출을 받는 사람은 점점 늘어나고 있는 추세이다. 한국은행이 발간한 2022년 9월 금융안정 상황 보고서에 따르면 올해 6월 말 기준 국내 총생산(GDP) 대비 민간신용 비율은 221.2%로 최고치를 기록했다.</p> <p>이러한 상황속에서 개인의 상황 및 데이터를 고려하여 가장 유리한 조건의 상품을 추천하는 '핀다'의 대출비교서비스는 고객에게 필연적인 역할을 한다.</p> <p>대출총량규제나 총부채원리금상환비율(DSR)적용, 금리 인상등과 같은 변동요인이 늘어날때, 소비자들은 여러가지 대출 상품을 한번의 조회만으로 신용등급에 영향을 주지 않고 결과를 알아볼수 있는 '핀다'앱을 많이 찾는다. 핀다의 대출비교서비스는 신용도 향상을 위한 중-저신용자뿐만 아니라 고신용자에게도 유리한 서비스이다.</p> <p>'앱 사용성 데이터를 통한 대출신청 예측'이라는 분석을 이행하면서, 기존 고객의 데이터를 통해 신규고객이 현재의 상황이나 변동요인을 감안하여 어떠한 상품을 이용할지를 군집화를 이용해 알아보고, 더 나아가 이 고객의 대출신청 여부를 머신러닝 기법의 모델로 예측하기로한다. 이를 이용하여 '핀다'는 예산을 올바르게 구축하여, 기업자체의 수익예측과 새로운 고객유입이라는 두가지의 파생효과를 가져올것을 기대한다,</p>

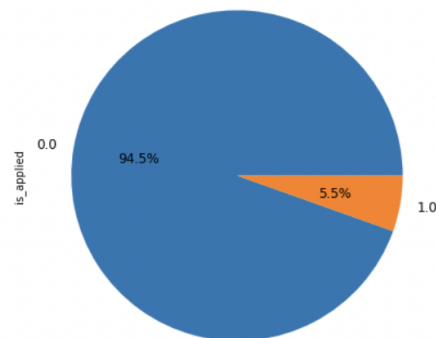
분석 내용 요약	<p>※ 분석 내용을 200자 내외로 간략하게 요약 기술해 주십시오</p> <p>2022/03 ~ 2022/05까지의 ‘핀다’ 앱 유저들의 스펙 정보, 앱 사용자들의 행동 패턴을 담은 log데이터, 사용자들이 신청한 금융사/상품 별 승인결과 등을 활용하여, 2022/06의 대출 상품 당 신청 여부 예측을 목표로 한다. ‘핀다’ 앱을 이용하는 고객들의 특성(스펙 정보, 앱 행동 패턴) 과 금융사/상품들의 특성을 결합한 후, 고객 유형 구분을 통해 본 팀은 고객 유형 별 대출 예측 모형을 구축하고자 한다.</p>
분석방법 및 계획	<p>※ 대회기간동안 선택한 주제(부문)에 대한 분석 방향에 대하여 기술해 주십시오</p> <ul style="list-style-type: none"> <li>- 분석에 활용되는 추가데이터(출처 기재)</li> <li>- 분석에 적용·활용할 통계·분석 기법, 방법론</li> <li>- 분석 결과에 대한 시각화 방법 등</li> </ul> <p><b>[ 분석 방향 및 시각화 ]</b></p> <p>1) 고객 유형에 따른 군집 시각화 그래프</p>  <p>위 ‘고객 유형에 따른 군집 시각화 그래프’는 고객을 5개의 유형으로 분류해 본 결과이다. ‘고객 스펙 정보 테이블’에는 연소득, 근로형태, 대출목적 등 다양한 고객들의 정보가 존재한다. 이 정보들을 결합하여 군집분석을 통해 고객의 유형을 분류해보고 군집화 된 고객 스펙 정보와 함께 ‘사용자가 신청한 대출/금융사별 승인결과 테이블’과 결합하여 대출 상품 신청여부에 어떠한 연관성을 띄는지 확인해보며 최종적으로 대출 신청 여부를 예측해보고자 한다.</p> <p>-&gt; 고객 스펙 정보 데이터는 군집분석 결과로 대체한다.</p> <p>2) 2022년 3~5월 요일/시간별 대출량 추이 그래프</p>



“2022년 3~5월 요일별/시간별 대출량 추이 그래프”는 주말과 평일의 대출량이 매우 다름을 보여준다. 특히 매주 월요일, 09시에 대출량이 높아지는 추세를 띈다. 대출량은 요일과 시간에 따라 크게 변할 것이라 생각한다.

-> 예측하는 기간에 대한 변수는 요일별/시간별 추세를 활용하여 대체할 것이다.

### 3) 타겟 변수 'is\_applied' 분포 시각화



최종 예측하고자 하는 'is\_applied' 타겟 변수에는 데이터 불균형이 존재한다. 따라서 데이터 불균형 문제를 해결하고자 다양한 over sampling 기법과 under sampling 기법을 비교하며 예측을 진행해 '재현율'을 높이려고 한다.

### 4) 분석 목표

앱 사용 로그 데이터에서 추출한 다양한 인사이트와 고객유형을 군집분석 한 결과를 활용하여 '사용자가 신청한 대출/금융사별 승인여부 테이블'을 중심으로 결합해, 어떠한 특성을 띄는 고객이 대출 신청을 하게 되는지 파악하고 예측하는 것을 주요 목표로 한다.

## [ 데이터 활용 ]

-> “고객의 유형”이 대출에 미치는 영향을 알아보기 위해 유저 스펙 정보를 군집화하여 ‘사용자가 신청한 대출/금융사별 승인결과 테이블’의 새로운 파생 변수를 생성한다.

-> 팀 내 아이디어 회의를 통해 신용점수가 유저 스펙의 정보를 대표한다고 판단하였다. 따라서 결측값을 채운 다른 유저 스펙 데이터를 사용해 신용점수를 예측하여 결측값을 채우는 모델을 구축하고자 하였다. 본 팀에서는 여러 머신러닝 알고리즘들 중 XGBoost에서 가장 좋은 성능을 띄어 XGBoost 모델을 사용하고자 한다.

-> 로그데이터 파생변수 생성:

- 1) 각 유저별로 하루에 행동(‘event’)을 얼마나 했는지
- 2) 유저가 했던 행동의 수를 총합
- 3) 유저가 기간동안 총 몇일 접속하였는지
- 4) 한도조회, 신용 정보 조회, 대출 관리 서비스 이용, 여윳돈 계산기 서비스 이용 중 어떤 것을 먼저 이용한 유저인지

이처럼 앱 로그 데이터에서 얻을 수 있는 인사이트들을 통해 파생변수들을 생성한다. 위 파생변수들을 이용하여 로그데이터를 군집화하고, 군집별로 대출 신청여부를 파악하여 새로운 고객 데이터가 들어왔을때 고객의 행동패턴을 기준으로 군집을 부여한후, 더 나은 모델링 방안을 제시한다.

## [ 분석 기법 ]

본 팀은 머신러닝에서 트리 기반의 XGBoost, LightGBM을 사용하고, Random Search와 Bayesian Optimization을 이용해 하이퍼 파라미터를 최적화 할 것이다. 또한, 머신러닝 비지도 학습에 속하는 K-Means와 K-Prototypes을 사용하여 군집화를 진행할 것이다.

### 1. XGBoost

: 트리 기반의 앙상블 학습에서 가장 각광받고 있는 알고리즘 중 하나로, 일반적으로 머신러닝 알고리즘 중 뛰어난 예측 성능을 발휘한다. 병렬 CPU환경에서 병렬학습이 가능하다는 점에서 기존 GBM에 비해 빠른 수행 성능을 보장한다. 또한 자체에 과적합 규제 기능이 있으며 tree pruning 기능으로 더 이상 긍정 이득이 없는 분할을 가지치기를 통해 분할 수를 더 줄이는 장점도 있다. 반복 수행 시마다 내부적으로 학습 데이터 세트와 평가 데이터 세트에 대한 교차 검증을 수행해 과적합을 방지할 수 있다. 지정된 반복 횟수가 아니라 교차 검증을 통해 평가 데이터 세트의 평가 값이 최적화 되면 반복을 중간에 멈추는 조기 중단 기능과 결손 값을 자체적으로 처리할 수 있는 기능이 있다. 추가로, 성능 평가와 피쳐 중요도를 시각화로 나타낼 수 있는 기능도 있다.

### 2. LightGBM

: XGBoost와 함께 부스팅 계열 알고리즘에서 가장 각광받고 있는 알고리즘으로, XGBoost보다 예측 수행시간이 빠르고 메모리 사용량이 적다는 점 등에서 XGBoost의 단점을 보완한 알고리즘이다. Root Node와 가까운 Node를 우선으로 대칭 분할하는 다른 트리 기반 알고리즘과는 다르게 Loss 변화가 가장 큰 Node를 지속해서 비대칭 분할한다. 이를 통해 깊은 비대칭 트리를 생성하며 대칭 분할보다 예측 오류의 손실을 감소할 수 있다. 또한, 카테고리형 피처의 자동 변환과 최적 분할이 가능하다.

### 3. K-Means

: 연속형을 가지는 객체들을 K개의 군집으로 군집화 하는 방법이며, 군집의 중심값은 군집의 평균으로 정한다. 몇 개의 군집(K)으로 군집화할지 결정하는 것이 중요한 알고리즘이다.

### 4. K-Prototypes

: 연속형과 범주형 속성이 혼합된 데이터를 군집분석하는 방법으로 연속형 속성을 군집화 하는 K-means 방법과 범주형 속성을 군집화 하는 K-modes 방법이 결합된 형태이다. 군집의 중심값은 군집의 연속형 속성의 평균값과 범주형 속성의 최빈값으로 정하며, 군집의 중심값과 객체간의 거리 계산은 범주형 속성 거리에 가중치를 주어 연속형 속성 거리와 합한 식을 따른다.

## [ Hyper parameter 최적화 방법 ]

### 1. Random Search

: Grid Search에 비해 불필요한 반복 수행 횟수를 대폭 줄이면서, 동시에 정해진 간격(grid) 사이에 위치한 값들에 대해서도 확률적으로 탐색이 가능하므로, 최적 하이퍼 파라미터 값을 더 빨리 찾을 수 있다.

### 2. Bayesian Optimization

: 불필요한 하이퍼 파라미터 반복 탐색을 줄여 보다 빠르게 최적 하이퍼 파라미터를 찾을 수 있는 최적화 방법으로 알려지지 않은 목적 함수를 최대화 하는 최적해를 찾는 기법이다.

## [ 과적합 방지 및 성능 향상 ]

### 1. Ensemble

: 다양한 종류의 여러 estimator를 결합하여 더 좋은 estimator를 만드는 것이다. 종류는 estimator들을 어떻게 결합할 것인지에 의해 결정된다. 대표적으로 배깅, 부스팅, 보팅, 스택킹이 있다.

**분석결과 활용 및  
시사점**

❖ 분석 결과에 대한 활용 방안, 적용대상, 결과 적용시 기대효과 및 시사점 등에 대하여 기술해 주십시오

‘핀다’는 주로 앱 기반으로 사용되는 서비스이다. 따라서 모바일 광고를 많이 사용하는데, 위에서 진행한 분석을 바탕으로 광고비용을 효율화하고, 앱 중도 이탈 사용자나 새로운 고객을 전략적으로 공략하여 광고를 개편하는 작업을 통해 ‘핀다’ 앱을 사용하는 고객과 ‘핀다’ 앱에서 대출을 신청하는 고객을 늘리며, 기업의 효율적인 수익창출을 이끌 수 있다.

또한 ‘무료 신용보험 서비스’ 나 ‘대출 관리 서비스’를 통하여 사용자의 대출위험성을 낮추어 최근 가계부채의 증가로 인한 민간소비 약화 문제를 완화 시키고, 더 넓게는 대한민국의 성장 잠재력을 키울 수 있다.

※ 제출자료는 평가에 반영 예정