**[ 7차(7월 2주차) 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 판매 데이터 분석을 통한 수요 예측 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 3조 - 데이터 맛집 | 프로젝트 팀원 | 황진웅(팀장), 홍석우(퇴소)  김철기, 박건하, 성원선 |

8※ 프로젝트 수행 일지는 매 주 금요일 17시까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
|  | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
|  | **1) 자동 평점 산출 시스템**  **1.1) 자동 평점 시스템 :** 작업완료    **1.2) 키워드 자동생성 시스템**  [해시태그 모델 생성]   * Okt, Mecab 등 형태소 분석기를 활용하여 형태소별 빈도수 집계 및 상위 주요 키워드 작성 * mlxtend, apriori, association\_rules를 사용해 리뷰에서 자주 함께 등장하는 단어 조합과 이들의 연관 규칙 분석 * 밀키트용 해시태그 사전 확장과 리뷰 길이 및 카테고리 기반 해시태그 자동화 로직을 정리, 과다출현 중복 태그 조정 및 랜덤성 도입 등으로 태깅의 다양성 강화   **2) 맞춤형 상품 추천 시스템**  **2.1) 해외 레시피 데이터셋**  **2.1.1) 데이터 가공**  - [Food.com](http://food.com) Recipes and Interactions  - 챗봇에서 사용할 사용자의 문장을 생성하기  위한 특징 테이블 생성  - 추천 시스템에서 문장을 입력으로 들어오면  상황(태그)를 분류하기 위한 데이터셋 생성 및  가공  ▶ 특징: category(음식 종류), ingredient(식재료)  ▶ 특징: emotion(감정), tags(상황)  - 입력 데이터: 영어 문장, 타겟값: tags(상황)  ▶ 주제가 사용자의 상황에 따른 음식 추천 시스템이기 때문에 타겟값을 tags로 설정하였음  - 마켓컬리와 해외 레시피 데이터와 연결하기 위해 카테고리(음식 종류) 매핑을 위한 데이터 가공 시도        **2.1.2) 모델**  - BERT 모델(bert-base-uncased)  ▶ 문장을 통해서 상황(태그) 분류하기 때문에  자연어 처리하는 BERT 모델 선택  **2.1.3) 결과 및 분석**  - 문장 생성    - 상황 분류  -> loss: 0.1162 - accuracy: 0.9747    -> 분석  ▶ snacks이 자주 출력이 되어서 조정이 필요함  ▶ 상황은 괜찮게 분류되므로 마켓컬리와 어떻게  연결할지를 생각해봐야함  - 마켓컬리와 해외 레시피 데이터셋 연결      -> 분석  ▶ 데이터 분석을 통해서 카테고리(음식 종류)를  매핑한 것이 아니라 유사도를 통해 하였기  때문에 신뢰도와 정확도가 낮다.  ▶ 결론  해외 데이터와 국내 데이터를 연결하기가 매우 어렵다. 특히 식품에서는 환경과 문화가 달라서 차이가 더욱 극심하다. 어느 정도 유용한 국내 데이터를 찾아서 추천 시스템을 구현하면 신뢰도와 정확도가 높을 것이다.  **2.2) KurlyPool**  **2.2.1) EDA**  배달음식판매건수 vs 간편식판매량  상관계수: 0.6917    연령대별 판매금액 vs 간편식 판매량  상관계수: 0.7054    지역별 배달매출 간편식 판매량간의 비교    -> 높은 상관관계를 보이므로 배달음식데이터와 간편식 데이터 높은 연관성 근거  간편식 판매량 비교(가구원수별, 연령별, 성별, 지역별)    평균 판매건수 비율(연령, 음식카테고리)      재구매율 (연령,성별)      -> 가구원수에서 주로 만두/피자류/면류가 높게나옴  1인가구에서 즉석식품, 밥류가 높게나옴  가구원수가 많을수록 육류쪽 구매를 많이함  -> 구매율은 40대가 제일 높음  -> 여성이 남성보다 재구매율 높음  7~8월은 재구매율이낮음  평균객단가    40대가 가장 높다,    월별 판매건수와 재구매율을 추이를 비교했더니  ,7,8월은 판매건수가 높은데 재구매율이 가장 적다  -> 판매량과 재구매율이 정비례하지 않음을 보여주며, 단순 판매량 지표만으로 충성 고객의 구매 성향을 파악하는 데에는 한계가 있음을 시사한다.  따라서 재구매율을 판매건수의 근거 자료로 사용하는 것은 타당하지 않으며, 고객 유지율 분석에는 별도의 지표 활용이 필요하다  지역,연령,판매건수 히트맵    지역,연령,판매금액 히트맵    -> 수도권 3040세대에서 구매비중이 높다.  가구주 연령별 히트맵    면류,만두,밥류가 높게나오고있고  2030세대에서 특히 즉석밥류를 많이구매한다.  연령대가높을수록 밥류는 적게찾고  만두/면류를 많이 구매한다.  연도별 품목 선호 추세    2018~2020년: 면류/밥류 수요가 높았음  2021년 이후: 면·밥류 수요 급감, 반면 즉석식품 수요 급증  이 현상으로부터 알 수 있는 것  소비자의 조리 부담을 줄이고자 하는 수요 변화가  기초식품(면·밥)에서 완조리제품(즉석식품)으로의 구조적 소비 트렌드 전환을 촉진시켰다.  지역별 판매량과 누적비율    지역별 판매량은 전체적으로 고르게 분포되어 있지만, 수도권과 일부 상위 지역이 누적 판매량의 과반 이상을 차지한다.  전국 단위 마케팅보다는 핵심 지역(Top 3~4)에 집중된 전략이 더욱 효율적일 수 있음을 시사한다.  연령대별 모든 밀키트 평균 판매량    밀키트 소비는 50대와 30대에서 가장 활발하게 나타나며, 중·장년층의 간편식 선호 증가를 반영한다.  연도별 밀키트 평균 판매량    한식 밀키트는 전통적인 수요를 기반으로 꾸준히 성장하며, 2023년에 최고점을 기록했다.  반면, 동양식은 일시적인 유행의 영향을 받아 단기 상승 후 하락했고,  서양식은 전체적으로 낮은 수요를 유지하며 상대적으로 비인기군에 속한다.  **가구원수별 거주지역별 밀키트 판매량**        시사점.  충청권·대경권은 밀키트 중심 소비지이며,  수도권은 외식·신선식품 경쟁으로 상대적으로 밀키트 수요가 낮다.  → 지역별 선호도와 라이프스타일에 따른 맞춤 전략이 필요하다.    연령대별 밀키트 유형 선호도는 명확히 구분된다.  30대는 동양식에 대한 선호가 높아 트렌디한 메뉴 추천에 적합하며,  50대는 한식 밀키트에 대한 압도적 수요를 보여 전통식 중심의 추천이 효과적이다.  서양식은 전 연령대에서 수요가 낮아, 이를 찾는 특정 사용자에게만 한정적으로 제공하는 전략이 적절하다.  따라서 챗봇은 사용자의 연령 정보를 바탕으로 선호도 기반 메뉴 필터링 및 맞춤형 메시지 제공 기능을 포함해야 하며, 추천의 정확도와 사용자의 만족도 향상에 핵심적인 역할을 할 것이다.  **연도별 간편식 구입주기 추이**      최근년도 음식 품목군별 구입주기    간편식 구입 주기는 ‘고빈도에서 중빈도로 이동하는 추세’를 보이며,  특히 ‘2주일에 1회’ 구입이 가장 일반적인 소비 패턴으로 자리잡았다.  2020년 팬데믹 시기 일시적으로 급증했던 ‘주 1회 이상’ 소비는 2022년 이후 급격히 감소하며 시장 포화 또는 식사 루틴 변화 가능성을 시사한다.  **챗봇 대화에 구입 주기 반영**  이용자 : 내가 2주전에 이 간편식을 구매했는데 추천해줄만한 간편식 있어?  챗봇 : "마지막 구매 이후 2주 정도 지났네요.  이 시점에 다시 찾으시는 분들이 많아요.  이번엔 이런 간편식은 어떠세요?"  (해당 구입주기 높은 품목의 음식 추천)  **2.2.2) 문장 생성**  **2.2.2.1) 목표**  **- 챗봇 로직**  -> Chatbot 흐름구상도  [1단계] 사용자 입력  ↓  [2단계] 의도 분류 (Intent Classification)  - 질문 → ex) '상품추천', ''트렌드\_월별' 등으로 분류(BERT/ROBERTA 트랜스포머모델 사용)  ↓  [3단계] 조건 추출 (Slot/Condition Extraction)  - 키워드 매칭  ↓  [4단계] 조건 확인 및 보완 질문 (선택형 or 자유 입력)  ↓  [5단계] 응답 템플릿 생성 or 추천 리스트 출력  ↓  [6단계] 후속 질문 or 선택 유도  **2.2.3) 마켓컬리 상품정보 데이터 가공**  -> 국내 데이터와 마켓컬리 상품정보 데이터 연결하기 위한 가공이 필요함  -> 브랜드 추출  ▶ [스윗밸런스] 오늘의 샐러드 10종 (리뉴얼) (택1)  ▶ 스윗밸런스  -> 음식을 수월하게 추천하기 위해 분류 작업(필터링)  ▶ 한식, 동양식, 서양식  -> 사용자의 취향을 고려하기 위해 음식 종류를 세분화  -> 결과    **3) 금주 핵심 이슈**  [챗봇]  - 챗봇 로직 정의 및 초기 모델 생성  **4) 회의 및 피드백**  [회의주제]  (1) 7월 5일 : 진행 상황 발표 - 멘토링  1) 자동 평점 시스템  -> 최종 결과 발표  2) 해시 태그  -> 비슷한 태그들만 출력되는 문제 발생  3) 맞춤형 추천 시스템  -> 현재까지 진행한 내용 발표  (2) 7월 9일 : 진행 상황  1) 해시 태그  -> 사전의 단어량을 증가하고 진행  -> 자동 평점 시스템에 사용된 리뷰 데이터에서 해시  태그를 달기  2) 맞춤형 추천 시스템  -> 추천 시스템 사용할 간편식 데이터 EDA  ● 간편식에 대한 데이터  ● 가구원수별 음식 카테고리 간편식 선호도  ● 연령대별 총 판매건수  ● 품목 평균 비중  ● 연령대별 평균 재구매율 비율과 분포  ● 성별 평균 재구매율 비율과 분포  ● 월별 평균 재구매율 추이 - 일반, 연령대별, 성별  ● 월별 판매금액 및 판매건수 추이 - 일반, 연령대별,  성별  ● 판매건수 - 지역, 연령대, 성별 기준  ● 지역별 품목 선호  ● 연도별 품목 선호 추세  -> Chatbot 로직설계 및 시나리오  -> Chatbot 문장생성(의도,질문,답변)  [피드백]  (1) 7월 5일 - 멘토링  1) 자동 평점 시스템  -> 최종적으로 평점 1~5점으로 분류된 것이기  때문에 내용이 분류 vs 회귀가 안 되도록 명시를  제대로 했으면 좋겠음  -> 분류 모델 vs 회귀 모델을 혼동 행렬로 보여주는  것은 조금 어색한 것 같다. 처음 볼 때는 오해를  할 수 있기 때문에 ‘분류 vs 회귀’라는 느낌이 안  들도록 정리하면 좋겠음  -> 전체를 5점으로 분류하는 경우와 R2score를  비교해보면 좋을 것 같음  2) 해시 태그  -> 사전의 단어량 늘리면 성능이 좋아질 수 있다.  3) 맞춤형 추천 시스템  -> 외국은 국내와 환경과 문화에서 차이가 크기  때문에 마켓컬리 데이터와 해외 데이터를 연결할  때 주의를 기울여야 한다.  -> 마켓컬리에서 없는 기능을 구현하면 좋음  -> 마켓컬리의 검색 기능하고 비교 하면 좋을 것  같음  -> 일반적인 추천 시스템은 감정에 관한 것은 없기  때문에 감정 기반 추천 시스템 구현하면 좋을 것  같음  (2) 7월 9일  1) 해시 태그  -> 수작업으로 해시 태그의 종류를 늘려서 만든  것이라면, 실제 결과가 적절하게 나왔는지 확인이  필요하다.  2) 맞춤형 추천 시스템  -> 시간에 대한 데이터는 필요없을 것 같다.  ▶ 컬리에는 상관없기 때문이다.  -> 롯데카드 배달음식에 대한 판매금액과 판매건수  ▶ 데이터의 근본적인 해석이 부족한 것 같다.  ▶ 데이터 내부에 어떤 피처가 있고 어떤 관계가  있는지 확인 필요  ▶ 각각의 데이터에는 어떤 정보가 있는지 확인  필요  **5) 인원별 세부 수행 내역**  - 김철기  : 마켓컬리 자동평점 시스템 Streamlit으로 예측 평점 결과를 시각화 ★⯨☆ 형태로 구현  추천시스템(챗봇) EDA작업 및 시각화  간편식/배달음식 데이터를 활용한 트렌드 분석 등  챗봇 로직 및 시나리오 의견 제시  챗봇 답변 문장 생성  - 박건하  : 자동평점예측 결과데이터에 형태소 분석을 적용해 주요 키워드 선별, 확장된 해시태그 사전 기반 해시태깅 로직 구축, 카테고리 분산 및 랜덤성을 반영하여 다양성 확보, 빈도 및 연관 분석을 통해 과다출현 및 중복 해시태그 조정  - 성원선  : 식품 관련 내용에 카테고리 정보를 연결하여 추천할 수 있게 도와주기위해 식품에 상품명과 상세 카테고리를 맵핑하는 작업 / 사용자에게 필요한 내용을 제공해주는 웹챗봇 구현 (Streamlit을 통해)  챗봇 구현을 위한 작업   * FAQ 데이터 수집 (SSG, 마켓컬리 등) * 챗봇이미지 생성   - 황진웅  : 레시피 데이터 상황(태그) 분류 테스트, 최종 출력으로 사용할 마켓컬리 데이터를 추천 시스템과 연결할 수 있도록 데이터 가공, 챗봇에 사용할 사용자의 문장 생성 | **1) 자동 평점 산출 시스템**  **1.0) 작주 진행 완료사항**  [자동평점 시스템]  스트림릿 이용해서 별점표시 나타내봄  [해시태그 자동생성 시스템]  - EDA 및 초기구축  - 구조 전환 및 기능 구현 테스트  - 사전 기반 자동 태깅 및 성능 개선 반복  **1.1) 해시태그/키워드 자동생성 시스템**  - 해시태그 결과검증처리 EDA 및 세부조정  - 실제 시스템 접목방법 고안  - 키워드 파일 정제 및 기능구현  **2) 맞춤형 추천 시스템**  **2.0) 작주 진행완료 사항**  - 마켓컬리와의 연결을 위한 마켓컬리 상품정보 데이터 가공  **2.1) 문장 생성**  - EDA한 데이터 기반으로 사용자의 질문과 챗봇 답변 문장 데이터셋 완성  **2.2) 챗봇 로직 설계**  키워드 기반 BIO 태깅 로직 시도해봄.  챗봇 응답 패턴 시나리오 구축  정규식 및 유사어 매칭 룰 최적화  **2.3) 테스트**  - 모델 테스트 및 선정  - 챗봇 동작  **2.4) 모델 선정**  - KoBERT  - 앙상블(XGBoost, Light GBM, CatBoost)  - BERT/ROBERTA 트랜스포머  **3) 전체**  - UI, 대시보드 꾸미기 |

**2. 강사님 피드백**

| EDA 과정에서 특히 판매 데이터의 경우 가구수와 같은 추가적인 요소들을 고려해야 합니다. EDA 분석 과정에서의 해석을 바탕으로 추가적인 진행방향을 만들어낼 수 있도록 진행하는 것이 좋습니다.  배달음식을 줄이자, 다이어트를 하자, 배달음식보단 간편식을 먹자 와 같은 이벤트성 아이템을 기반으로 생각을 하면 좀 더 아이디어를 구체화하는 데 도움이 될 수 있으며, Feature 들을 제대로 활용할 수 있도록 진행해보시는 것을 추천드립니다.  마켓컬리 방문을 유도하겠다, 유저들의 쇼핑을 도와주겠다 등의 방향성을 잡고 마켓컬리 챗봇 캐릭터를 만드는 등의 특정 컨셉을 잡아 진행하는 것이 구체화하기 편할 것 같습니다. 다양한 관점에서 도전해보세요.  해시태그는 평점 시스템 모델을 연결하여 진행한 부분이 좋았으며 결과 검증, 후처리 그리고 결과물에 어떻게 접목시킬지 등을 활용방안에 대해 고민해보세요. |
| --- |