**[ 2차(6월 1주차) 프로젝트 수행 일지 ]**

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

※ 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) 데이터 수집(크롤링)**  **1.1) 리뷰 데이터 수집**  - 수집 사이트 : 11번가, 옥션  - 수집이유 : 마켓컬리 내 평점 시스템의  부재로 자동 평점 시스템 개발을 위해 타 플랫폼의 평점 데이터를 이용하여 모델 학습 및 학습 결과 크로스 검증에 활용  **1.2) 각각 3만개씩 인당 총 6만개의 데이터**  - 김철기: 밀키트  - 박건하: 가공식품  - 성원선: 냉동식품  - 홍석우: 반찬  - 황진웅: 과일  **2) EDA 작업**  **2.1) 결측치, 이상치 확인**  : 리뷰 텍스트와 평점 누락  **2.2) 변수들 간 상관관계 확인(평점 기준)**  2.2.1) 이미지 포함 여부와 이미지 수  - 두 플랫폼에서 양의 상관관계  - 이미지가 있을수록 평점이 높음  - 긍정적인 리뷰일수록 구매자가 인증샷을  올리는 경우가 많다고 해석 가능  2.2.2) 리뷰 길이  - 음의 상관관계  - 불만이 있을 때 설명이 길어지는 경향  - 부정 리뷰는 길고, 긍정 리뷰는 짧게 쓰는  경우가 많음  2.2.3) 도움돼요 수  - 신뢰도가 낮음  - 매우 약한 음의 상관관계  2.2.4) 작성일자  - 시간 데이터는 자동 평점 생성 시스템에선  유용한 데이터인지 의문(여러 종류가 섞여  있을 가능성도 있음)  **3) 데이터 분포**  **3.1) 평점 분포**  - 대부분 4~5점대 집중되고 1~2점대는 거의  없는 불균형이 심함  **3.2) 리뷰 키워드와 이모지 빈도수**  - 부정적인 표현이 상대적으로 적음  (막대그래프, 워드클라우드로 시각화)  **4) 데이터 정제**  - 이모지만 추출, 텍스트만 추출, 이모지+텍스트  추출 등 여러 버전으로 데이터 정제  - 평점 기준 긍정, 중립, 부정으로 라벨링  - 핵심 키워드를 TF-IDF기반으로 추출하여 감정  별로 맵핑하여 핵심 키워드 분류  **5) 데이터 학습 테스트**  **5.1) 진행**  - 훈련 데이터: 11번가 리뷰 데이터,  테스트 데이터: 옥션 리뷰 데이터  - 훈련 데이터: 옥션 리뷰 데이터,  테스트 데이터: 11번가 리뷰 데이터  - 정확도 비교  **5.2) 감정 분류**  - 샘플링 데이터 사용  - 텍스트 정제(이모지만 제거, 이모지와  특수기호 모두 사용)  - 평점과 감정 매핑  -> 평점 1~2점: 부정  -> 평점 3점: 중립  -> 평점 4~5점: 긍정  - KoBERT 모델로 학습  **6) 금주 핵심 이슈**  [데이터 불균형]  - 평점 데이터 중 90퍼센트 이상이 긍정적인 리뷰로 데이터 불균형 발생 및 부정 데이터 학습 데이터 부족  - 해결 방안 : 부정 부분에 3점(중립)을 추가하여 라벨링  [애매한 문장]  - 맛있지만 배송이 느렸다, 포장 상태가 별로였지만 양이 많았다 등 긍정과 부정이 함께 있는 문장 분류의 모호함  - 해결 방안 : KoBERT 등 앞 뒤 맥락을 파악하는 모델 이용  [라이브커머스 데이터 수집불가]  - 날짜별 음식 선호도 파악 시 대안 필요  - 해결 방안 :  1. 네이버랩, 구글 트렌드에서 해당 날짜와 시기에 사람들이 많이 찾는 음식 조사  2. 캐글 데이터로 연령대, 성별, 날짜에 따라 구매하는 음식(식품) 조사  **7) 회의 및 피드백**  [회의주제]  - 옥션, 11번가 데이터 통합 사용 여부  - 강수량 데이터를 추가 여부  - 임의의 단어와 리뷰를 분석하여 유사도로 표현  [피드백]  - 날씨 데이터 : 위치 데이터도 필요함  - 날씨에 따른 음식 추천할 때 근거가 필요  - 리뷰 텍스트 기반으로 분석하며 진행 제안  - 감정 표현 패턴은 정제 방식 선택 권유  **8) 인원별 세부 수행 내역**  - 김철기  : 텍스트 뿐 아니라 이모지,특수기호까지 여러 다양한 방면에서의 상관관계를 해석하고 분석  - 박건하  : 수집한 데이터들을 취합 후 세부 카테고리 별로 정리하고 데이터 전처리  - 성원선  : 옥션, 11번가 뿐만 아니라 다른 플랫폼을 찾아보고 필요한 데이터 추가 크롤링  - 홍석우  : 데이터 이해를 위한 서칭과 정제된 외부 데이터(아마존 리뷰 데이터)로 필요한 데이터 분석 및 모델 테스트  - 황진웅  : 지금까지 수집한 데이터 중 일부 데이터로 다양한 모델 테스트와 정확도를 올리기 위한 방법 서칭 | **1) 데이터 추가 수집**  - 데이터 불균형으로 추가 데이터 수집과 상관 관계가 높은 다른 데이터 수집  - 카카오맵 API이용(라이브커머스 대체)  지역별 날씨 데이터를 활용하기 위해서 지역별 음식점 데이터수집  **2) EDA**  [필요성 및 작업방향]  - 변수 간 관계, 분포 등을 파악해 분석 전략을 수립하여 데이터 이해도 향상  - 어떤 피처가 타깃에 영향을 줄 수 있을지 가설 설정 기반 마련  -데이터 오류나 편향, 왜도 등의 문제를 사전에 탐지  [처리방식]  히트맵, 꺾은선 그래프(시계열), 히스토그램    **3) 텍스트 정제 및 평점분류**  **3.1) 텍스트 정제**  - 오탈자, 특수문자, HTML 태그 등 노이즈 제거  - 같은 의미의 단어를 통일된 형태로 바꾸는 정규화 작업으로 통일성 확보  - 노이즈 제거 범위 정하기(정확도 비교)  **3.2) 평점 분류**  - 샘플링 데이터 사용  - 평점과 감정 데이터 수정  -> 평점: 1 (0~4)  -> 감정: 부정: -1, 중립: 0, 긍정: 1  - 앙상블 + Light GBM 모델로 학습  (XGBoost + Light GBM + CatBoost)  - 마켓컬리 리뷰 데이터로 예측  **3.3) 작업방향**  - 데이터 노이즈 제거를 통한 정확한 분석 기반 확보  - 일관된 형식으로 정리된 데이터를 이용하여 모델 학습 효율 향상  - 분석 목표에 맞는 변수 재구성을 통해 유의미한 피처 구성  **4) 데이터 불균형 해결 방법 찾기**  - 언더샘플링 (긍정 줄이기) - 오버샘플링 (부정/중립 늘리기)  - 가중치 적용해보기(가장 효율적인 가중치 수치)  **5) 모델 설계 준비**  - EDA를 바탕으로 감정 분류 모델, 평점 예측 모델을 설계하기 위한 준비  - Kaggle 데이터와 한국어 리뷰 데이터를 함께 사용할 수 있는 모델 찾기(Hugging Face 등) |

**2. 강사님 피드백**

| 모델 성능 평가 기준이 명확하지 않습니다. 실제 리뷰와 예측 평점의 차이 분석을 진행하여 정량적 평가 방식을 명시하면 좋을 듯 합니다.  모델 설계 부분에서 KoBERT, 앙상블(XGB + LGBM + CatBoost) 등의 언급은 좋지만, 어떤 기준으로 모델을 선택했고 어떤 성능 차이가 있었는지 설명이 부족합니다. 근거가 추가되면 좋을 것 같습니다.  결측치, 이상치 제거는 잘 이루어졌으나 데이터 편향이나 노이즈와 관련된 분석, 근거들도 확보하시면 좋을 것 같습니다.  마지막으로 상품 추천을 하기 위한 구체적인 계획과 진행 상황도 궁금합니다 ㅎ |
| --- |