**[ 6월 4주차 프로젝트 수행 일지 ]**

**프로젝트 수행 기간 : 6/17 ~ 6/23**

|  |  |  |  |
| --- | --- | --- | --- |
| **프로젝트 주제** | 서울시 골목상권 구역별(상권코드명) 매출에 영향을 주는 요인과 구역별 매출 금액 예측 | | |
| **프로젝트 팀명** | 이장우조 | **프로젝트 팀원** | 장서영, 장예림, 우성진, 이채린 |

**※**

|  |  |  |
| --- | --- | --- |
| **1. 프로젝트 수행 계획 및 현황** | | |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **장서영** | 1. 매출액 데이터를 가공하고, 매출액 상위 5개와 하위 5개의 그래프를 각각 그리고 시군구별 골목상권 개수를 시각화 함. 2. 각 팀원들이 가공한 데이터들을 취합하고 전처리를 하여 최종 데이터셋을 완성함. 3. 매출액을 기준으로 시군구 군집분석을 실시함. | 1. 군집분석 결과를 보기 쉽게 시각화하고 정리 2. 라쏘, 릿지 회귀분석 검증해 보기 3. 발표 자료 정리 |
| **이채린** | 1) 데이터 수집  2) 기획서 작성  3) 데이터 전처리  - 프렌차이즈 점포 수 추출 및 행 합치기  4) 시각화  - 상관관계 히트맵  - 프렌차이즈 수  - 시군구 별로 트리맵  - 상권구분코드명 파이차트 | 1) 의사결정나무(회귀트리)  2) 피처그래프 시각화  - 의사결정나무 통해서 피처의 중요도 시각화  3) 데이터 분석 시각화  - 프렌차이즈 수 - 상권코드명 top 5 시각화  - top 5 의 시군구 시각화  4) XGBoost / LightGBM/고급회귀 관련 공부 및 수행  5) 발표 자료 정리 |
| **장예림** | 1) 데이터 전처리  - 상권\_소득소비 데이터에서 필요한 열 추출  2) 데이터 분석  - 분야별 소비 지출 top\_5 구역 시각화  3) 회귀분석  - 다중선형회귀  1. 전체 데이터 세트 활용  2. 매출액 상위 25% 데이터 세트 활용 -> 회귀식 완성  3. 매출액 하위 25% 데이터 세트 활용 -> 회귀식 완성  - 릿지 회귀 분석  - 라쏘 회귀 분석  - 다항회귀 분석 | 1) 데이터 분석\_시각화  - 화곡역 4번이 어느 분야에 소비 지출을 많이 했는지 시각화  - 까치산 1번이 어느 분야에 소비 지출을 많이 했는지 시각화  - 까치산 4번이 어느 분야에 소비 지출을 많이 했는지 시각화  - 소비 지출 분야와 나이대는 어떤 상관관계가 있는지 확인  2) 회귀분석 검증  3) 발표 자료 정리 |
| **우성진** | 1) 데이터 전처리  -상권\_소득소비 데이터에서 필요한 열 추출 및 재범주화  2) 데이터 분석  -연령,성별,시간대 서울시 골목상권 구역별 시각화 및 상관관계 분석  3) 랜덤포레스트  -피처별 중요도 시각화 | 1) 데이터 분석\_시각화  -오후 저녁 생활인구와 연령대에 상관성 분석과 시각화  2) XGBoost / LightGBM/고급회귀 관련 공부 및 수행  3) 발표 자료 정리 |

|  |  |
| --- | --- |
| **2. 강사 피드백** | |
| **강사명** | 1) 각자 역할이 골고루 분배 되어있음  2) 머신러닝보다 데이터분석 쪽에 내용을 더 추가하면 좋을듯  3) 더 세세한 과정으로 결과를 도출하는 데이터 분석 방식을 추가하면 좋을듯  4) 회귀분석과 의사결정회귀,랜덤포레스트의 RMSE를 비교  5) 고급분석회귀에서 알고리즘 순서대로 LinearRegression, Ridge, Lasso 한 번에 돌리는 것 참고  6) XGBoost, LightGBM 은 데이터 별로 다르므로 직접 해보고 더 나은 모델을 선택하는 방식 |