데이터 사이언스 입문

ITS\_EATS팀 보고서

<목차>

1. 프로젝트 소개
2. 데이터 처리
3. 서비스 제안
4. 제언

**1. 프로젝트 소개**

배달 주문량 데이터를 통해 배달 수요를 예측하여새로운 배달 앱에 서비스를 서비스를 제안하는 것이 프로젝트의 목적이다.

**2-1. 데이터 전처리**

1) 날씨 변수 추가

배달을 예측하는 데에 있어 상식적으로 가장 중요할 거 같기도 하고.

실제로 조사를 해봤을 때 날씨와 배달 간의 상관관계에 대한 연구가 많았다.

2) 한강 변수 추가

한강 근처 지역에서는 배달량이 많을 것으로 예상되어 법정동별 한강 인접 여부를 조사하여 변수로 추가했다.

3) 공휴일 추가

날짜를 기준으로 데이터를 시각화한 결과 공휴일이 중요한 요인임을 확인하였다.

4) 스포츠 경기 관련 변수 추가

공휴일이 아닌 경우 스포츠 경기가 있는 경우 눈에 띄게 배달량이 증가함을 확인하였다.

**2-2. 데이터 정제**

문제점: 데이터 사이즈가 너무 크기에 R로 분석하기가 불가능하였다. 데이터를 필수적으로 줄이고 정제해야 한다.

해결방향:

1) 표본추출

데이터가 요약된 정보를 가지고 있기에 대표성을 잃을 수 있어 표본 추출은 불가능하다.

2) 구별로 나눠서 분석을 진행

행정구 간에 상관관계를 파악할 수 없다.

3)연령대 카테고리 병합

시각화를 통해 볼 수 있듯이 30 40대, 50 60대 배달량 비슷하기에 합쳐 데이터를 줄여도 크게 손실이 없을 것으로 판단하였다.

4)동을 연속형 변수로 변환

동이 300여개로서 one hot encoding이 진행되기에 가장 데이터를 많이 차지한다. 연속형으로 변환하여 column수를 줄일 수 있다.

**3. 모델링**

1) 시계열

시간의 흐름이랑 관련있는 데이터이기에 시간 흐름에 따라 변하는 변수들만 가지고 시계열 분석 진행

2) 디시전 트리

디시전 트리를 이용하여 결과 예측 및 변수 중요도 확인

3) XGBOOSTING

일반적으로 성능이 가장 높은 Xgboosting을 통하여 결과 예측

**4. 예측 모델 이용한 서비스 제안**

xgboost 모델을 통해 서울시 구별 지점당 평균 주문건수를 예측할 수 있다.

1) 서비스 1: 데이터를 이용하여 미래 예상 배달주문 건수 예측을 통한 수요 대비.

예를 들어 2019년 3월 16일 대치동의 치킨 배달 수요 예측. 우리에게 이런 데이터가 주어졌을 때, 예측 모델을 이용하면 수요를 예측할 수 있다.

2) 서비스 2: 고객 맞춤 팝업 광고

개인이 치킨, 중국집, 피자 등을 주문할 확률이 어느 정도 되는지 예측하여 고객 맞춤 광고를 설정할 수 있다.

예를 들어 우리 모델을 이용하면 (합정동, 20대, 여자) 의 경우 (11월 11일)에 치킨을 시킬 확률이 00%, 중국음식을 주문할 확률이 00%, 피자를 주문할 확률이 00% 임을 알 수 있다.

이를 이용하여 개인의 특성에 맞는 팝업 광고를 설정할 수 있는데, 예를 들어 이러한 특성을 가진 고객의 경우 치킨 메뉴를 위주로 팝업 광고를 띄우는 것이 도움이 된다.

**5. 제언**

1) 파생 변수 생성 과정:

우리 배달량 데이터는 법정동을 기준으로 데이터가 구분되어 있었는데, 서울시에서 동별 정보는 행정동 기준으로 데이터가 나와 있어서 추가하지 못했다. (변환 불가능)

우리동네 상권 분석 서비스에서 제공하는 소득 분위, 1인가구 비율 등의 추가정보를 사용했으면 더 예측력을 높일 수 있었을 것이다.

2) 비용의 한계

지역에 중국음식점, 치킨전문점, 피자전문점이 몇 곳 위치해 있는지 등에 관한 상권 정보를 파생 변수로 추가하고 싶었지만 비용이 많이 들어서 분석에 활용하지 못했다 (NICE 지니빅데이터). 실제로 의뢰를 받아서 예산이 있다면 충분히 가능한 문제.

3) 위도 경도

동 변수를 위도 경도로 바꿨을 때 트리 모델에서는 문제가 생길 수 있다. 지리적으로 먼 곳에 위치한 지역의 경우 동으로 넣고 돌렸을 때는 한 그룹으로 묶일 수 있지만 위도 , 경도로 나뉘는 경우에는 불가능하다.

감사합니다.