배달 데이터를 활용한 Regression/Clustering Modeling

2019312259, 김유진

Introduction

1) Describe your problem with some background information.

: 우리나라 사람이라면 누구나 코로나바이러스 이후 배달 음식 이용이 급증하게 되었음을 실감할 수 있을 것이다. 한국농촌경제연구원의 '2020 식품소비행태조사'에 따르면, 2020 년 기준 전년 대비 음식 배달이나 포장에 지출이 늘었다는 답변이 총 33.4%였다. 실제로 재택 근무가 늘어나면서 4050 세대도 점심을 배달 음식으로 해결하는 경우가 많아졌으며, 치킨을 제치고 한식 주문 비중이 높아지기도 하는 등 대한민국의 배달 음식 문화에 많은 변화가 찾아왔다.

2) Describe why the problem is important and explaining that you are working on something significant.

: 위와 같은 추세를 따라 국내 여러 가지 배달 음식 플랫폼들이 활성화 되고 있으며 이 과정에서 사용자들의 큰 만족을 얻는 서비스나 음식점이 있는 반면, 불만이 쏟아져 나오는 곳도 있었다. 따라서 이번 머신러닝 프로젝트를 통해 '배달 음식' 관련 데이터를 분석하고, 소비자들에게 도움이 되는, 유익한 결론 혹은 서비스를 도출해내고자 한다.

3) What is your final goal? What do you want to achieve in the project?

: 우리나라의 주요 배달음식 플랫폼(요기요, 배달의 민족, 쿠팡 이츠) 등의 크롤링 데이터 혹은 검색을 통해 얻을 수 있는 데이터 분석을 통해 지역 별로 어느 시간에, 어느 장소에서 어느 음식에 배달 주문이 몰리는지 등을 예측할 수 있는 모델을 만드는 것으로 목표를 잡았다.

Data

1) What data do you plan to use?

: 지역별, 음식 별 주문 건수 데이터, 배달 소요시간 데이터

2) Where are you going to get the data (download it from a website or collect by yourself)?

: KT 빅데이터 플랫폼을 통해 배달 관련 데이터 수집

3) Data description (if available)

업종-지역별 평균배달소요시간 데이터(2019-07-17 ~ 2020-09-30)

- 출처: 경기대학교 빅데이터 센터
- Column: 날짜, 시간대별 시간, 배달 상점 업종명, 배달상점 광역시도명 / 시군구명, 주문건수

Methods

1. **EDA**

   Kdeplot, lineplot 등을 이용해 데이터의 분포를 시각적으로 확인해본다.

2. **데이터 정제**

   초기 데이터는 무려 1940382 개의 row 로 이루어져 있었기 때문에 특정 조건에 따라 row 의 개수를 줄여보고자 했다. 먼저 count plot 에서 확인한 메뉴 개수 중 하위 네 가지인 (심부름, 도시락, 배달전문업체, 회)를 제거하였고, 현재와 비슷한 시기를 비교하기 위해 2020 년 5 월의 데이터로 축소하였으며, 우리 대학이 속해있는 서울특별시로 제한시켜 우리에게 더욱 와 닿는 결론을 도출하고자 하였다.
   최종적으로 완성된 데이터프레임은 20977 개의 row 와 5 개의 column 으로 구성되었다.
   완성된 서울의 데이터 프레임으로 다시 lineplot 등을 이용해 전체와 비교하며 분포를 살펴보았다.

3. **데이터 전처리**

   먼저 categorial feature 들을 숫자로 encoding 해주는 과정을 거쳤다. 특히 date column 의 경우, 월요일은 1, 화요일은 2 과 같이 요일 별로 데이터를 구분해주었다.
   또한 필요에 따라 scaling 과 outlier 제거 등의 과정을 거쳤는데, 이는 뒤에 regression 을 적용하면서 scaling 이나 outlier 실행 전/후를 비교하는 데에도 사용했다.

4. **Regression**

   주문 시간(hour)와 배달 소요시간(time)의 scatterplot 을 통해 회귀의 가능성을 생각하게 되었다. IQR 을 활용해 outlier 를 제거하기 전과 후의 회귀 결과를 살펴보았다.
   확실히 해당 분포에는 linear 나 quadratic 보다 cubic 의 곡선이 더 잘 들어맞았고, outlier 를 제거해주었을 때 보다 나은 결과를 얻을 수 있었다.

5. **Clustering**

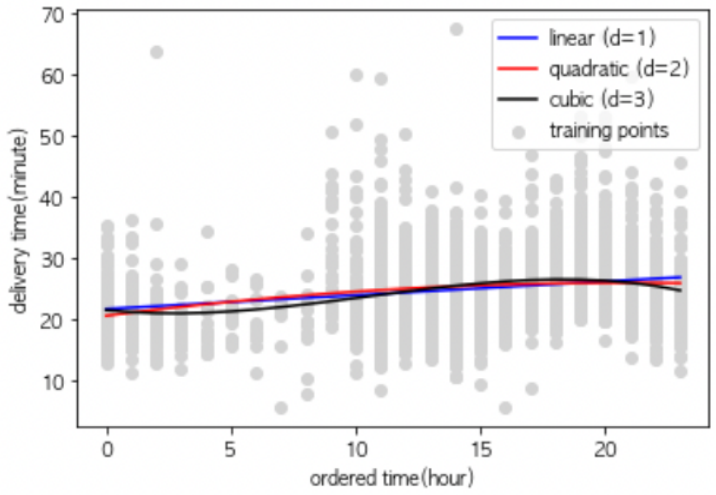
   또한 k-means clustering 기법을 적용시켜보았다. 우선 그래프를 통해 k 의 값에 따라 군집 내 거리가 어떻게 변하는지 확인하였다. 이후 k 의 값을 바꿔가면서 스케일링 전/후 데이터에 클러스터링을 적용시키고 결과를 확인했다.

Experimental Results

1. Regression(chicken 메뉴 예시)

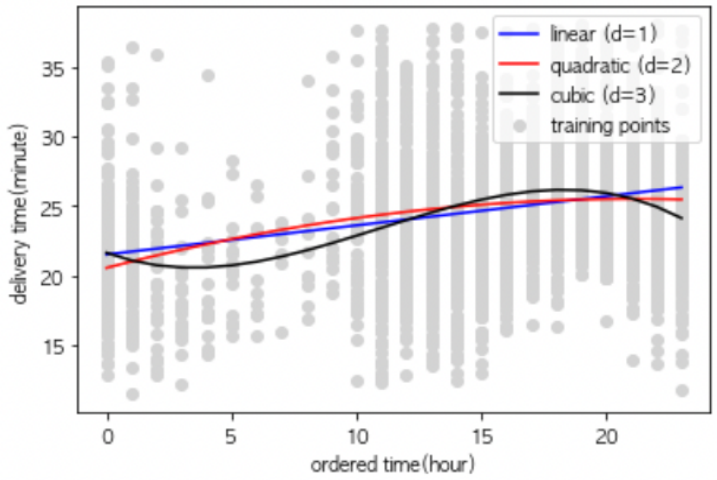
(outlier 제거 전)

```
MSE       Linear: 35.20,          Quadratic: 34.88,        Cubic: 34.49
R2        Linear: 0.05,   Quadratic: 0.06,         Cubic: 0.07
```



(outlier 제거 후)

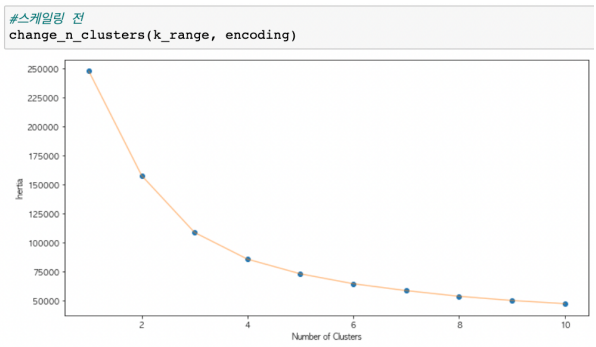
```
MSE       Linear: 22.81,          Quadratic: 22.56,        Cubic: 22.01
R2        Linear: 0.07,   Quadratic: 0.08,         Cubic: 0.10
```

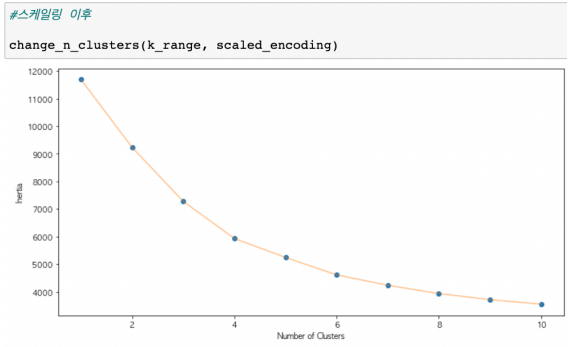


두 개의 figure 은 x 축을 주문 시간(hour) y 축을 배달소요시간(time)으로 하여 적용시킨 regression 의 결과이다. 해당 분포에는 3 차 함수가 가장 들어맞는 것을 알 수 있었으며, 또한 Outlier 제거를 통해 더 나은 결과를 얻을 수 있었다.

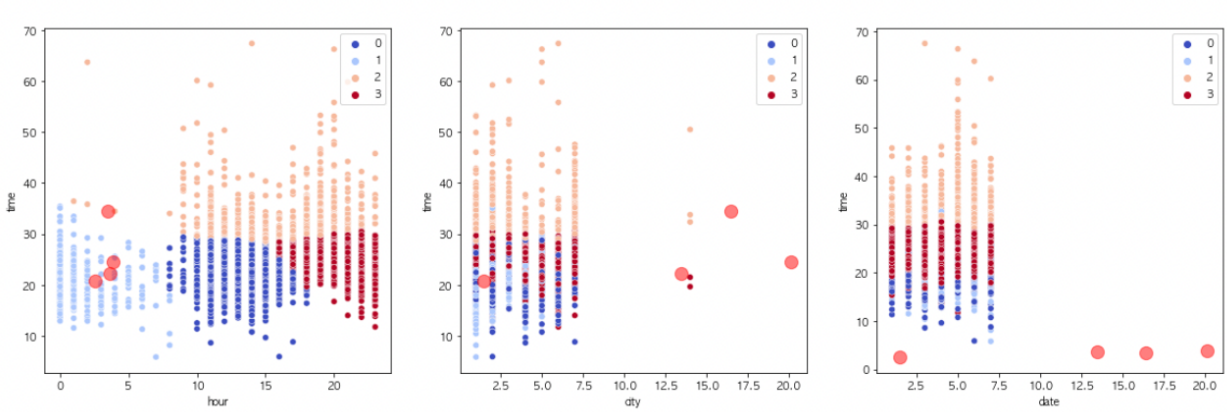
2. K-means Clustering(k=4) (chicken 메뉴 예시)

(k 값에 따른 군집 내 거리 확인)

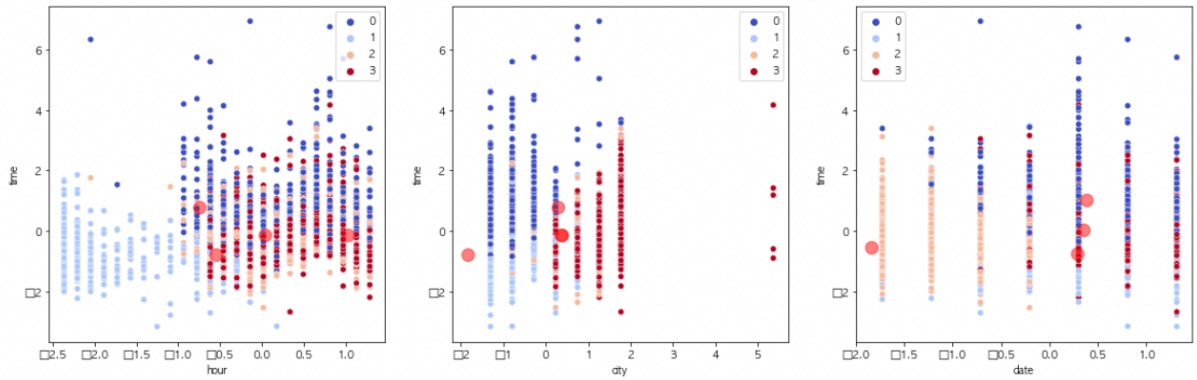




(scaling 전)



(scaling 후)



먼저 k 값에 따른 군집 내 거리 변화를 그래프를 통해 확인한 뒤, 모델에 k=2, k=4, k=6 을 적용시켜보았다. 위의 그림은 그 중에서도 가장 데이터에 적합해보이는 k=4 를 적용시켜 나타난 figure 들이다. y 축은 모두 배달 소요시간(time)이고 x 축을 각각 주문 시간(hour),

시군구명(city), 주문 날짜(date)로 맞추었다. Categorial 값들 때문인지 clustering 은 scaling 전이 군집의 구분이 뚜렷해 보였다. Scaling 전의 첫번째 subplot(x=hour, y=time)을 보면, 가장 뚜렷하게 네 개의 그룹으로 구분이 되었는데, 0: 주문이 몰리는 점심시간 대에 주문 하였지만, 비교적 음식을 빨리 받은 그룹 / 1: 새벽 시간대에 주문 한 그룹 / 2: 주문시간대에 상관 없이 항상 오래 걸린 그룹 / 3: 저녁 시간대에 주문했지만, 비교적 음식을 빨리 받은 그룹으로 나눠볼 수 있었다.

Conclusions

1) A summary of what you have achieved.
   한학기 동안 배워왔던 머신러닝을 직접 데이터를 찾아서 적용시켜본 경험이 처음이었는데, 데이터 상태에 따라 모델이 반응하는 과정이 다르게 나타나는 것이 인상깊었고 신기했다. 뿐만 아니라 해당 모델의 성능을 평가하는데 사용되는 여러가지 통계적 지표들을 알게 되었고, 내가 아직 많이 부족함을 느꼈다. 아쉽게도 이번에 내가 선정했던 데이터들은 categorial 한 값들이 많아서 상관관계가 두드러지지 않은 것 같다. 앞으로는 이번 학기를 통해 배운 내용을 바탕으로 공모전이나 대회 등을 참여해보면서 실전적인 경험을 많이 쌓아야 겠다는 생각을 했다.
2) 서울은 아침(오전 9-10 시)에도 배달소요시간이 높음 / 서울의 배달 데이터에서 주문 시간, 배달소요시간에 따라 4 개의 뚜렷한 군집화가 가능함(0: 주문이 몰리는 점심시간 대에 주문 하였지만, 비교적 음식을 빨리 받은 그룹 / 1: 새벽 시간대에 주문 한 그룹 / 2: 주문시간대에 상관 없이 항상 오래 걸린 그룹 / 3: 저녁 시간대에 주문했지만, 비교적 음식을 빨리 받은 그룹)

References

https://www.bigdata-telecom.kr/invoke/SOKBP2603/?goodsCode=KGUTIMEDLVR
https://newsis.com/view/?id=NISX20201218_0001275733
https://www.yna.co.kr/view/AKR20210111023700030