**<응용머신러닝 final 발표 대본>**

1. **Introduction**

저번 중간 프로젝트 발표에서도 말씀 드렸다시피, 코로나 속에서 우리나라는 현재 그야말로 ‘배달의 민족’이 되었습니다. 시장의 규모가 갑자기 확 불어난 만큼 생겨나는 여러 가지 문제점을 지난 발표에서도 살펴봤었는데요, 대표 적으로 배달 음식의 퀄리티 문제나, 심각한 배달 지연 문제 등이 있겠습니다.

따라서 저는, 소비자들이 어느 음식을 어느 시간에 시켰을 때 주문량이 몰리는지를 미리 파악할 수 있으면 편리 하겠다는 생각을 하게 되어 해당 주제를 선정하게 되었습니다.

1. **Data**

주제를 정하고 나서, 데이터를 살펴보는 과정을 먼저 거쳤습니다. 제가 가져왔던 데이터는 KT 빅데이터 플랫폼 이라는 사이트에서 “업종-지역별 평균배달소요시간”데이터 였습니다. 해당 데이터를 파이썬 라이브러리를 이용해 EDA 하는 과정을 거쳤습니다.

먼저 데이터프레임은 다음과 같이 형성이 되어 있었습니다. 컬럼에는 주문 날짜, 주문 시간, 배달 상점 업종명, 광역시도명과, 시군구명, 그리고 마지막으로 평균 배달 소요시간으로 구성이 되어있습니다.

하지만 이 데이터는 row의 개수가 column의 개수에 비해 너무 많았습니다. 저는 row를 먼저 줄이는 과정을 거쳤습니다. 메뉴 별로 countplot을 그려보았을 때, 하위 네 가지의 메뉴인 “회”, “배달전문업체”, “도시락”, “심부름”을 제거하였습니다.

하지만 여전히 row가 많은 것을 볼 수 있었고, 머신러닝 할 데이터의 범위를 더욱 좁혀야 겠다는 생각을 했습니다. 따라서, 지금이 5월인만큼 2020년 5월의 데이터로 제한하였고, 또한 성균관대학교 학생들이 주로 거주하고 있는 서울특별시로 범위를 좁혀보았습니다.

그 결과 다음과 같이 약 2만개의 row로 된 데이터프레임을 완성하였습니다.

이 데이터로 여러가지 시각화를 해보았는데요, 먼저 주문 시간 대, 즉 hour column에 따른 KDE plot을 그려보았습니다. 보시다시피 녹색 선은, 데이터 정제를 통해 만든 서울 데이터프레임의 plot이며, 검정색 선은 전체 데이터의 분포입니다. 서울은 특히 15시 경의 주문 밀도가 높았으며 오전 5-7에도 전체 분포에 비해 약간 높은 것을 알 수 있었습니다.

그 다음, x축은 주문 시간으로, 그리고 y축은 배달 소요시간으로 lineplot을 그려보았는데요, 여기서 직선은 서울과 전체 각각의 배달소요시간 평균을 나타냅니다.

서울이 전체적인 배달 소요시간이 높았으며, 특히 9-10시 오전과 19시 저녁 시간대에 배달 소요시간이 두드러지게 높았습니다.

분포를 확인 한 후에, 직접 모델에 데이터를 학습 시키기 위해 categorial value 들을 숫자로 encoding 해주는 과정을 거쳤습니다. 특히 날짜 컬럼의 경우 요일 별로 숫자를 나누어 값을 매칭시켜주었습니다.

이 데이터를 이용해 저는 유의미한 결론을 얻기 위해서 두 가지 머신러닝 기법을 활용해보기로 하였습니다.

먼저 배달 소요 시간 예측을 위해 regression 기법을, 그리고 주문 분포를 묶어서 확인해보기 위해 clustering을 적용시켜보았습니다.

1. **ML method**

먼저 regression입니다. 해당 데이터에서 x축을 주문시간, y축을 배달 소요시간으로 두었을 때가 가장 어느 정도의 상관성을 확인할 수 있어, 이 두개의 컬럼을 이용해 regression을 진행했는데요,

그 전에 먼저 time 컬럼의 outlier를 IQR을 이용한 boxplot으로 확인해보았더니, 다음과 같았습니다. 여기서 꽤 많은 outlier들이 존재했기 때문에, outlier 제거 전/후 regression 결과를 비교해보게 되었습니다.

다음과 같이 outlier를 제거했을 때 조금 더 적합한 분포가 나타났으며, linear, quadratic, cubic 중에는 cubic이 가장 적합해보였습니다.

다음은 clustering 파트인데요, 저는 그 중에서 K-means 클러스터링을 적용시켜보았습니다. K값을 결정하기 위해 k값에 따른 군집 내 거리를 그래프를 이용해 확인을 먼저 해보았고, 다음과 같이 scaling 전과 후로 나누어 군집화를 시켜보았습니다.

1. **Conclusion**

이번 프로젝트를 하면서 데이터를 다루는 전반적인 과정이나 머신러닝 알고리즘을 직접 만져볼 수 있어서 정말 유익한 시간이 되었습니다. 다만 제가 선정했던 데이터가 컬럼의 수가 충분하지 않고, 카테고리얼한 값들이 많아서 명확한 상관관계를 도출하는데 어려움을 느껴 크게 아쉬움으로 남았습니다.

앞으로 머신러닝에 대해 더 열심히 공부를 해나가면서 보다 심화된 프로젝트를 꼭 한번 해보고 싶습니다.