AIFFEL 1st HACKATHON

Recruit Restaurant Visitor Forecasting

9해조

Reference

@김준일 일본 휴일/골든위크 데이터 (2016 , 2017)
일본 휴일 제도 (해피먼데이)
Aiffel Exploration 6 : 나의 첫 번째 캐글 경진대회, 무작정 따라해보기
Kaggle Notebook : holiday trick

- 점수를 올리는데 도움이 된다는 내용이 있어서 적용해보았다. 실제로 점수에 미쳤던 영향은 적었다.

@박은영 https://www.kaggle.com/faizalabdkadir/recruit-visual-eda-itadakimasu

- Data Wrangling이 잘 정리되어 있다.
- Feature에서 얻을 수 있는 정보들을 모두 쪼개고 새로운 기준으로 데이터를 구조화 시켜서 다양한 시각으로 데이터를 살펴보는 방법을 배울 수 있었다.
- 다양한 시각화 기법이 소개되어 있다

@황동호 https://www.notion.so/modulabs/5b419c654e12473985402b452538f945#23d7568a36b64a228e905dd4b238fbb2
https://www.notion.so/modulabs/5b419c654e12473985402b452538f945#612bc7140d8d4c7a8ef23e9d626cfaa3
https://www.notion.so/modulabs/5b419c654e12473985402b452538f945#290fa95557484817a6094b05579ba905

- 어떤 데이터를 대상으로 시각화를 하고, 어떻게 해야 할지에 대한 인사이트를 얻을 수 있었다.
- 데이터간 상관관계를 파악할 수 있었다.

EDA

우리가 EDA를 통해 알게된 것은..

- 평일보다 금,토,일(주말)에 방문객이 더 많은 편
- 평일보다 공휴일이 방문객이 더 많은 편
- 골든위크 기간에 그렇지 않은 휴일보다 특히 더 많은 방문자가 방문
- 연휴라고 다른 휴일보다 방문자가 증가하지는 않는 점
- 다음날이 휴일이면 그렇지 않은 날보다 방문자 증가
- 다음날이 주말인 경우 방문자가 증가
- 방문객에게 선호되는 장르가 있는 점
- 어떤 장르의 레스토랑은 평일 방문객은 적으나 주말 방문객 수가 크게 증가한 다는 것
- 예약은 금요일, 토요일 예약이 많고
- 일부 레스토랑이 대부분의 방문객을 커버하는 점처럼 인기있는 레스토랑이 있다는 것

이런 방문에 영향을 주는 변수들을 설명변수로 방문객 수를 예측 해본다!!

Modeling

Feature Engineering

각 상점의 요일별 평균 방문자에 대한 통계들을 Feature로 사용한 것이 점수를 높이는데 큰 영향을 끼쳤다 (score 0.8 → 0.5)

지역(area_name) 데이터와 업종(genre_name) 데이터는 Label인코딩보다 One-Hot 인코딩이 더 좋은 점수를 냈지만 학습 시간이 오래걸렸다. (public score 기준 0.503 → 0.497)

Scailing

StandardScaler 보다 MinMaxScaler가 조금더 좋은 점수를 냈다.
MinMax가 변수간의 범위를 정규화 시켜주며 데이터의 분포를 유지해준다고 한다.
(public score 기준 0.54094 → 0.53461)

Score

Submission and Description	Private Score	Public Score	Use for Final Score
Aiffel-YJ-9_save_me_junil Version 13 : add qoq (version 13/13) 23 minutes ago by kimjunil Notebook Aiffel-YJ-9_save_me_junil | Version 13 : add qoq	0.53429	0.48948	☐

Review

@김준일

학습시킬 Feature들을 선택하거나 가공할 때 EDA를 기반으로 하기 때문에 데이터 분석이 무척 중요하다는 것을 느꼈다.

@박은영

데이터 피쳐링의 중요성에 체감할 수 있던 프로젝트 였다. 무엇을 기준으로 데이터를 다듬는지에 따라 예측결과에 미치는 영향을 볼 수 있었다.
변수끼리 높은 상관관계를 보이는 다중공선성 문제를 확인하고, 변수선택법을 적용해 보려 했는데 연산도중에 환경이 자꾸 재시작 되는 문제로 모델에 적용을 해보지 못해 아쉬웠다.

@황동호

시각화를 통한 전체 데이터에 대한 파악의 중요성과, 데이터를 어떻게 전처리하고 어떤 데이터를 기준으로 모델을 학습시킬지에 대해 고민해볼 수 있었다.