데이터마이닝 프로젝트 최종 보고서

20201634 전민혁

1. 데이터셋 소개

에콰도르의 [Favorita](https://www.corporacionfavorita.com/)사 슈퍼마켓 매장의 판매 데이터입니다.

1) train.csv

train.csv의 열 목록은 id, date, store\_nbr, family, sales, onpromotion입니다.

이중 store\_nbr은 어느 매장인가를 나타냅니다. 매장은 총 54개입니다.\

family는 상품이 어떤 품목인지 나타냅니다. 품목의 개수는 총 33개입니다.\

onpromotion은 해당 품목 중에 세일하는 상품의 개수를 나타냅니다.

train.csv는 2013년 1월 1일부터 2017년 8월 15일까지의 데이터입니다.

2) test.csv

test.csv의 열 목록에서 나머지는 train.csv와 같으나 sales가 빠져 있습니다.=

test.csv는 2017년 8월 16일부터 2017년 8월 31일까지의 데이터입니다.

3) oil.csv

train.csv와 test.csv파일에 기록된 날짜들에 대한 에콰도르의 기름값 데이터입니다.

에콰도르의 경제는 석유 산업에 의존합니다. oil값이 판매량에 영향을 미칠 가능성이 높습니다.

4) stores.csv

stores.csv의 열 목록은 store\_nbr, city, state, type, cluster 입니다.

train.csv, test.csv의 store\_nbr에 해당하는 마켓이 어느 주의 도시에 있는지, 어떤 타입인지 나타냅니다.

유사한 매장들은 같은 cluster로 묶였습니다.

5) holidays\_events.csv

휴일이나 이벤트 등의 특이한 날입니다.

열 목록에서 locale은 어떤 단위(지역인지 국가적인지)에 해당하는지 나타냅니다.

열 목록에서 transferred는 대체공휴일인지 여부를 나타냅니다.

6) transcation.csv

2013년 1월 1일부터 2017년 8월 15일까지 거래량입니다.

**1. 데이터 가공**

- holyday events나 stores 정보가 결과에 영향을 많이 끼치지 않을 것이라 판단하여 train, oil, transection csv파일만 학습에 이용했습니다.

- oil, transection의 행들 중 train.csv파일의 행과 일치하지 않는 자료들이 존재합니다.

만약 이대로 열들을 합치면 NaN(Not a Number) 이 값으로 들어갈 수 있습니다.

oil의 경우 하루정도로 값이 크게 변하지 않을 것이므로 NaN이 나온 값들을 전날의 값으로 채워주었습니다.

transection의 경우 NaN이 나올 때 해당 날짜에는 거래가 발생하지 않았습니다. 따라서 NaN이 나왔을 경우 해당 행은 삭제했습니다.

- 날짜값 처리

2015-05-05처럼 날짜에 해당하는 'Date'열은 (년도 \* 10000 + 월 \* 1000 + 일 \* 10 + 요일정보(일부터 토까지 순서대로 0123456) 로 변환하였습니다.

- 학습 데이터와 테스트 데이터

test.csv파일과 별개로 테스트를 위해 2달정도의 데이터를 분리하였습니다.

**2. 학습**

데이터 학습은 lightgbm으로 진행하였습니다.

(원래 transformer를 변형하여 제작하였으나 모델의 완성도가 처참하여 lightgbm으로 변경하였습니다.)

라인, 직사각형, 스크린샷이(가) 표시된 사진

자동 생성된 설명

학습 이후 가지고 있는 데이터를 바탕으로 테스트를 진행하였습니다.

오차의 평균은 -18.050499311413468 이 나왔고 위 사진은 테스트에서 오차가 얼마나 나왔는지 보여주고 있습니다.

많은 경우 오차가 크지 않으나 몇몇 값들의 경우 오차가 크게 발생하였습니다.

월단위로 계산할 경우 몇몇 오차가 심한 값들을 제외하면 판매량을 예측하는데 크게 문제가 없는 수준으로 보입니다.

**3. 제출**

- prediction을 제출하기 위한 test.csv파일에 transection과 oil이 없는 문제가 발생

이를 해결하기 위해 transection과 oil값을 다시 예측하여 이 값을 test.csv 파일의 데이터에 합쳐서 예측을 진행했습니다.

- 전체 등록된 720명중 613위

스크린샷, 텍스트, 번호, 라인이(가) 표시된 사진

자동 생성된 설명