-

매장 단위 판매량 예측 모델 개발

CUAI 8기 **DA 2**팀
김동건(AI학과), 최규원(예술공학부), 안상우(경영학과)

[요약]
본 프로젝트는 Kaggle 대회 'Store Sales - Time Series Forecasting'을 기반으로 매장 내 다양한 품목들의 단위 판매량을 예측하기 위한 모델을 개발하는 것을 목표로 한다. 데이터를 전처리하고 시각화하여 주요 패턴을 분석한 후, LightGBM 모델을 사용하여 예측 모델을 구축하였다. 이 모델은 매장 운영의 효율성을 극대화하고 재고 관리 전략을 개선하는 데 기여할 수 있는 중요한 인사이트를 제공한다.

1. 서 론

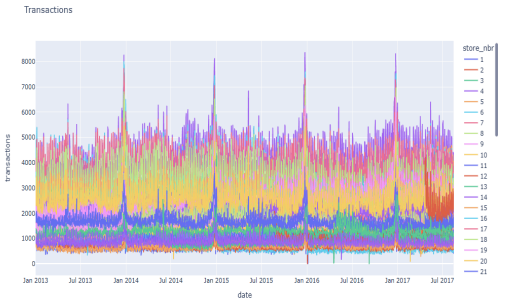
본 프로젝트의 목적은 매장 내 다양한 품목의 판매량을 예측하여 매장 운영의 효율성을 높이고 재고 관리에 있어 정확한 의사결정을 지원하는 데 있다. 이를 위해 데이터 전처리, 시각화 및 머신러닝 모델링 과정을 거쳐 최적의 예측 성능을 달성하고자 하였다.

2. 본 론

1) 데이터 셋 설명 및 시각화

본 프로젝트에서는 다양한 데이터셋을 활용하여 매장 판매량에 영향을 미치는 주요 변수들을 분석하였다. 데이터셋은 매장 정보, 거래 건수, 석유 가격, 공휴일 및 이벤트 정보 등으로 구성되었다.

거래 건수(transactions)와 판매 금액(sales) 간의 상관관계 분석 결과, 두 변수 간에 높은 양의 상관관계가 확인되었으며, 특히 연말 시즌에 거래 건수가 급증하는 패턴이 나타났다. 이는 연말 세일과 블랙프라이데이 같은 이벤트가 주요 원인으로 작용했을 가능성이 클 것으로 보인다.

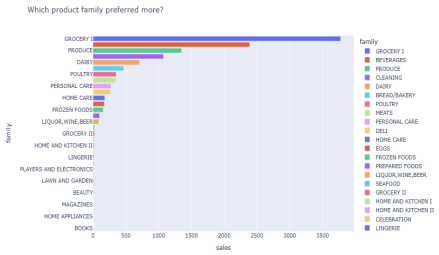


(그래프1) 거래 건수와 판매 금액 간 관계

날짜별 석유 가격 흐름을 분석한 결과, 주말에 결측치가 발생하는 패턴이 확인되었으며, 이는 석유 가격이 에콰도르 경제와 매장 판매량에 미치는 영향을 분석하는 데 중요한 데이터를 제공한다. 매장별 판매량 변화 분석에서는 2016년 4월 특정 매장에서 판매량 급증 현상이 발견되었으며, 이는 당시 발생한 대규모 지진과 관련된 것으로 보인다. 이 결과는 자연재해가 판매량에 미치는 영향을 평가하는 데 중요한 시사점을 제공한다.

(그래프2) 날짜 별 석유 가격 흐름

품목군별 판매량 분석에서는 프로모션 여부에 따라 판매량이 크게 달라지는 패턴이 나타났으며, 특정 품목군에 대한 집중적인 프로모션이 효과적임을 확인하였다. 마지막으로, 도시 및 주별 판매량 분석 결과, 대도시와 경제 중심지의 판매량이 상대적으로 높아, 매장 운영 전략 수립 시 지역별 특성을 고려할 필요가 있음을 시사한다.



(그래프3) 품목군별 판매량 분석

2) 데이터 전처리 과정

시계열 데이터의 특성을 고려하여 다양한 전처리 작업을 수행하였다.

먼저, 여러 CSV 파일에서 데이터를 불러온 후, train과 test 데이터셋을 하나로 결합하고, holidays, stores, oil, transactions 데이터셋을 각각 store_nbr과 date를 기준으로 병합하였다. 이 과정에서 병합된 데이터는 다중 인덱스(store_nbr, date, family)로 설정하여 이후 분석 및 모델 학습 시 데이터 접근성을 높였다. 이후, date 열에서 연도(year), 분기(quarter), 월(month), 일(day), 요일(day_of_week), 연중일(day_of_year) 등의 피처를 추가로 생성하였다. 이러한 피처들은 시계열 데이터의 계절성과 트렌드를 학습하는 데 중요한 역할을 한다.

또한, 석유 가격(oil_price) 데이터는 에콰도르와 같은 석유 의존도가 높은 국가에서 매출에 중요한 영향을 미칠 수 있는 변수로, 주말에 석유 가격이 책정되지 않는다는 점을 고려하여 결측치를 처리하고, 이를 매출 예측 모델에 반영하였다. 마지막으로, 매장 번호(store_nbr), 상품군(family) 등 범주형 변수는 라벨 인코딩(Label Encoding)을 통해 수치형 데이터로 변환하였다. 이는 모델이 범주형 데이터의 고유 특성을 학습할 수 있도록 돕는데, 예를 들어, 특정 매장이나 상품군의 고유한 판매 패턴을 모델이 인식할 수 있도록 한다.

3)모델 학습 및 평가

전처리된 데이터를 바탕으로 LightGBM 모델을 활용하여 매장 판매량 예측을 수행하였다. LightGBM은 그레이디언트 부스팅 알고리즘을 기반으로 한 빠르고 효율적인 모델로, 특히 대규모 데이터셋에서 우수한 성능을 보인다.

모델 학습을 위해 전체 데이터셋을 학습 데이터(X_train, y_train), 검증 데이터(X_val, y_val), 테스트 데이터(X_test)로 분리하였다. 테스트 데이터는 최종 모델의 성능 평가에 사용되며, 검증 데이터는 학습 도중 모델의 성능을 모니터링하고 과적합을 방지하기 위해 사용된다. 학습 과정에서는 데이터의 열 이름에서 특수문자를 제거하는 전처리 과정을 거쳤으며, 이는 모델 학습 중 발생할 수 있는 오류를 방지하고 모든 피처가 올바르게 학습되도록 하기 위함이다.

LightGBM 모델 학습에는 learning_rate, max_depth, num_leaves, feature_fraction, bagging_fraction 등의 하이퍼파라미터가 사용되었다. 학습 과정에서 검증 데이터를 통해 실시간으로 모델의 성능을 평가하였으며, 이를 통해 모델의 일반화 성능을 높였다. 모델의 예측 성능은 RMSE(Root Mean Squared Error) 지표를 사용하여 평가되었으며, 이 지표는 예측 값과 실제 값 간의 오차를 평가하는 데 유용하다. 또한, K-Fold 교차 검증을 통해 모델의 일반화 성능을 추가로 검증하였으며, 이를 통해 데이터셋의 특정 부분에 치우치지 않고 다양한 데이터 분포에 대해 안정적인 성능을 보이는 모델을 구축할 수 있었다.

하이퍼파라미터 튜닝 과정에서는 GridSearchCV를 사용하여 다양한 하이퍼파라미터 조합을 탐색하고 최적의 모델을 도출하였다. 이 과정에서 학습률(learning_rate), 최대 깊이(max_depth), 리프 노드 수(num_leaves) 등의 하이퍼파라미터를 조정하여 모델의 예측 성능을 극대화하였다. 이는 모델이 특정 데이터에 과적합되지 않도록 조정하며 성능을 향상시키는 중요한 단계이다.

```
LGBMRegressor
LGBMRegressor(bagging_fraction=0.7, bagging_freq=10, feature_fraction=0.9,
              max_bin=512, max_depth=50, metric=['l1', 'l2'], num_leaves=128,
              objective='regression', task='train', verbose=0)
```

(그림1) 하이퍼파라미터

결과는 실제 운영 환경에서 활용될 수 있도록 적절한 형식으로 저장되었다. 이 예측 결과는 매장의 위치, 프로모션 여부, 기념일 등이 판매량에 중요한 영향을 미친다는 것을 잘 반영하고 있으며, 이는 매장 운영 전략 수립에 유용한 인사이트를 제공할 수 있다.

id	sales
3000888	3.056196
3000889	1.445302
3000890	13.310976
3000891	2216.884255
3000892	1.445302
...	...
3029395	447.752389
3029396	136.339060
3029397	1984.648576
3029398	147.236049
3029399	18.000321

28512 rows × 1 columns

(표1) 예측 결과

3. 결 론

본 프로젝트는 매장 단위의 판매량 예측을 통해 매장 운영의 효율성을 극대화하고 재고 관리 전략을 개선하는 것을 목표로 하였다. 다양한 시각화와 전처리 과정을 통해 데이터를 깊이 있게 분석하였으며, LightGBM 모델을 활용하여 최적의 예측 성능을 도출하였다. 향후 연구에서는 추가적인 피처 개발이나 딥러닝 모델을 적용하여 성능을 향상시키는 방안을 모색할 수 있을 것이다.

참고 문헌 (앞 뒤 두줄 삽입, 9p, 진하게, 왼쪽 정렬)

권철민, “파이썬 머신러닝 완벽 가이드”, 2022