kaggle x elo

카드사 고객 충성도 예측 모델 분석

Kaggle; Elo Merchant Category Recommendation

혁신성장 청년인재 집중양성 사업 1기
딥러닝 기반 핵심산업별 빅데이터 분석 전문가 과정
Cielo 김하준 강호영 조민정

CONTENT

Cielo

프로젝트 개요

Elo Merchant Category Recommendation

왜 Kaggle 인가?

실질적인 기업데이터를 통해 빅데이터 분석과 머신러닝 경험을 쌓고자 했으나, 기업 데이터는 일반에 공개되지 않을 뿐만아니라 크롤링에도 한계가 있었다. 따라서 **데이터분석 경쟁 플랫폼**인 Kaggle에서 제공하는 실제 기업데이터를 통해 마케팅 관련 데이터 분석 프로젝트를 진행하고자 하였다.

왜 Elo Competition인가?

실제 기업들이 제공하는 데이터분석 미션들을 비교해본 결과, elo 에서 원하는 분석과 본 팀이 지향하는 분석 방향이 일치하였다. elo는 기존의 마케팅 방법을 뛰어넘는 **개인화 알고리즘을 구축하는 것을 목표**로 하고 있다고 설명하고 있어, **마케팅 분야의 머신러닝을 이용한 분석 경험**을 충분히 쌓을 수 있다고 판단했다.

What is Kaggle?

Elo Merchant Category Recommendation

Kaggle Competition 진행 특징

- **기업**이 **빅데이터와 과제**를 캐글에 제공
- 팀이나 개인으로 **문제를 해결하고 상금**을 얻는 방식
- Competition이 종료된 후에도 참여할 수 있음
- Kernel과 Discussion을 통해 **경쟁 중에도 과제에 대한 의견과 코드를 공유** 하는 것을 권장
- 최대한 자유로운 분석을 보장
- 실제 기업에서 요구한 미션을 해결하기 때문에 **실질적 데이터 분석 역량을 증명**할 수 있는 방법으로 부상 중

Kaggle ;세계 최대 데이터 분석 경쟁 플랫폼

2010년에 설립된 예측 모델 및 분석 대회 플랫폼
기업 및 단체에서 데이터와 해결과제를 등록하면
데이터 과학자들이 이를 해결하는 모델을 개발하고 경쟁하는 플랫폼

Kaggle과 실제 기업 데이터

실제 기업의 분석방식

1. 프로젝트의 목표와 Target을 정의
2. 데이터 수집하여 마트를 구축
3. 분석 시행

Kaggle의 분석방식

1. 주어진 데이터 내에서만 분석 운용가능
2. 이미 정해진 타겟값을 가장 가깝게 예측하는 경쟁

》 똑같은 데이터를 이용하더라도 실제 기업의 분석 방식과 Kaggle은 정 반대의 프로세스를 갖고 있음

Elo Competition의 핵심

Elo Merchant Category Recommendation

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

주어진 데이터로 머신러닝을 통해
Loyalty Score를 예측한 후 실제
Target 값에 대한 **RMSE**를 평가

Kaggle에서 제공하는 RAW DATA

historical_transactions

고객별 과거 트랜잭션 데이터

프로모션 시작 날짜로부터의 구매까지 소요 기간
구매 승인 여부. 구매 날짜,
구매 지역 및 구매 상점, 기타 익명 카테고리 제공

new_merchant_transactions

historical transaction에서 고객이 방문하지 않은 상점에 대한 트랜잭션 데이터

historical_transactions와 제공 컬럼 동일하나 고객 당 구매 상점 정보는 일치하지 않음 (historical에서 방문한 매장 정보 삭제)

merchants

상점에 대한 부가적인 정보

상점 고유 id와 카테고리, 상점 지역 정보, 마지막 활성 월의 트랜잭션 량, 3,6,12개월 기준 월 수익과 트랙잭션 데이터, 다수의 익명 카테고리 제공

train

기존 예측 모델 활용 정보
고객 id와 첫 구매날짜
익명화된 평가정보와 target 값

test

기존 예측 모델 활용 정보
고객 id와 첫 구매날짜
익명화된 평가정보, target 없음

MISSION! 유의미한 파생변수를 생성하여 개인 Card id 별 1:1로 대응시키고 하나의 데이터로 통합하는 작업 필요!

PROBLEM! 데이터 파악 중 발견된 ELO DATASET의 문제점과 해결 해야 할 미션

▷ Target과 feature의 상관성 부족

제공된 feature와 target이 correlation이 부족해 변수의 중요도 파악이 힘듦

▷ Normalize formula에 대한 설명 부족

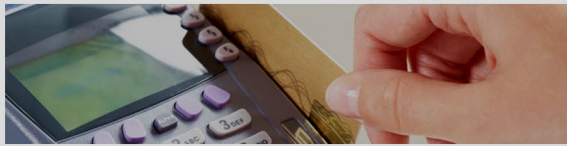
주요 피쳐인 구매량과 target, 익명의 수치형 컬럼의 정규화 공식에 대한 정보가 부족하여 데이터 파악 어려움

▷ Target에 대한 명확한 정의 불명

의도적으로 loyalty score, target 값에 대한 명확한 정의를 제공하지 않아 Target 값이 어떤 의미인지 명확하게 파악하는 것이 필요함

▷ 익명화된 Category의 불확실성

각 dataset별로 의도적으로 익명처리 된 컬럼들이 많아, feature engineering에서 새로운 파생변수를 만들기가 어려움



분석 Process

Elo Merchant Category Recommendation

Raw data 파악

주어진 데이터 파악을
통해 분석 방향 결정

Data Quality Check

제공된 데이터셋의 특성 파악
결측치 및 이상치 제거
활용 가능한 데이터 정리
익명화된 데이터와
Loyalty Score의 의미 파악

EDA

각 컬럼의 통계적 특성 파악
각 변수와 Loyalty Score의 관계 파악

Data Mart 구축

분석 요건에 맞게, 필요할
때마다 접근이 용이하도록
데이터 웨어하우스 구성,

Feature 생성

주어진 데이터를 통해
활용 가능한 파생변수 생성

Algorithm 생성

Loyalty Score 예측을 위한
회귀 머신러닝 모델 생성

Feature Selection

주요 영향 feature를 선정해
모델 효율성 증가 및 정확도 상승

Validation

Feature값 조정,
Parameter tuning 및
Ensemble 기법을 통해 예측률 조정
Root Mean Squared Error (RMSE)
점수 산정

모델선정

최종 평가를
위한 모델 선정

데이터 분석 결과; 모델링

Elo Merchant Category Recommendation

Gradient Boosting Decision Tree : XGBoost, Light GBM, Catboost

여러 개의 결정 트리를 묶어 강력한 모델을 만드는 다른 앙상블 방법. Kaggle Competition에 가장 많이 이용되는 모델로, 무작위성이 연관성이 약한 피쳐를 가진 모델들을 연결하고 앙상블하여 예측률을 높임. 기본적인 변수들과 타겟 변수인 loyalty score의 상관도를 확인해 본 결과 강한 상관을 가진 변수가 존재하지 않아 약한 변수들을 많이 연결하여 성능을 향상시키는 그래디언트 부스팅 모델을 사용하기로 결정하였고, 대표적 모델인 XGBoost와 Light GBM을 사용

	XGboost	Light GBM	Catboost
모델 특징	병렬처리 사용, 학습과 분류가 빠름 다양한 커스텀 최적화 옵션 설정 가능 자동가지치기를 통해 과적합 방지 앙상블 학습에 용이	대용량 데이터 처리에 용이 속도가 빠르고 예측정확도가 가장 높음 가지치기 시 loss를 줄일 수 있음 과적합에 민감하여 가장 많이 쓰임	Level-wise Tree 구조 사용 범주형 변수 처리를 모델 훈련과 동시 진행 모델 학습 시간 단축 및 과적합 방지 메모리 효율이 높음
Parameters Used	learning late = 0.01 max_depth = 6 gamma = 0.1 seed = 1000 nthread = 4 min_child_weight = 30	learning_rate = 0.01 num_leaves = 30 min_child_weight = 50 bagging_fraction = 0.7 feature_fraction = 0.7 bagging_freq = 5	learning_rate = 0.015, iterations = 1000 bagging_temperature = 0.3 ls_leaf_leg=10 depth = 10, one_hot_max_size=500
Training Score	3.6806	3.65204	3.8035
Leaderboard Score	3.767	3.687	3.882

데이터 분석 결과; 최종 모델 선정

Elo Merchant Category Recommendation

```
def run_lgb(train_X, train_y, val_X, val_y, test_X):
    params = {
        "objective" : "regression",
        "metric" : "rmse",
        "num_leaves" : 30,
        "min_child_weight" : 50,
        "learning_rate" : 0.01,
        "bagging_fraction" : 0.7,
        "feature_fraction" : 0.7,
        "bagging_freq" : 5,
        "bagging_seed" : 2019,
        "verbosity" : -1
    }

    lgtrain = lgb.Dataset(train_X, label=train_y)
    lgval = lgb.Dataset(val_X, label=val_y)
    evals_result = {}
    model = lgb.train(params, lgtrain, 3000, valid_sets=[lgval], early_stopping_rounds=200, ver
bose_eval=100, evals_result=evals_result)

    pred_test_y = model.predict(test_X, num_iteration=model.best_iteration)
    return pred_test_y, model, evals_result

train_X = train_df[Features]
test_X = test_df[Features]
train_y = train_df['target'].values

pred_test = 0
kf = model_selection.KFold(n_splits=6, random_state=1992, shuffle=False)
for folds, (dev_index, val_index) in enumerate(kf.split(train_df)):
    dev_X, val_X = train_X.loc[dev_index,:], train_X.loc[val_index,:]
    dev_y, val_y = train_y[dev_index], train_y[val_index]

    pred_test_tmp, model, evals_result = run_lgb(dev_X, dev_y, val_X, val_y, test_X)
    pred_test += model.predict(test_X, num_iteration = model.best_iteration) / 6
# pred_test /= 6
```

Our Champion Model; Light GBM

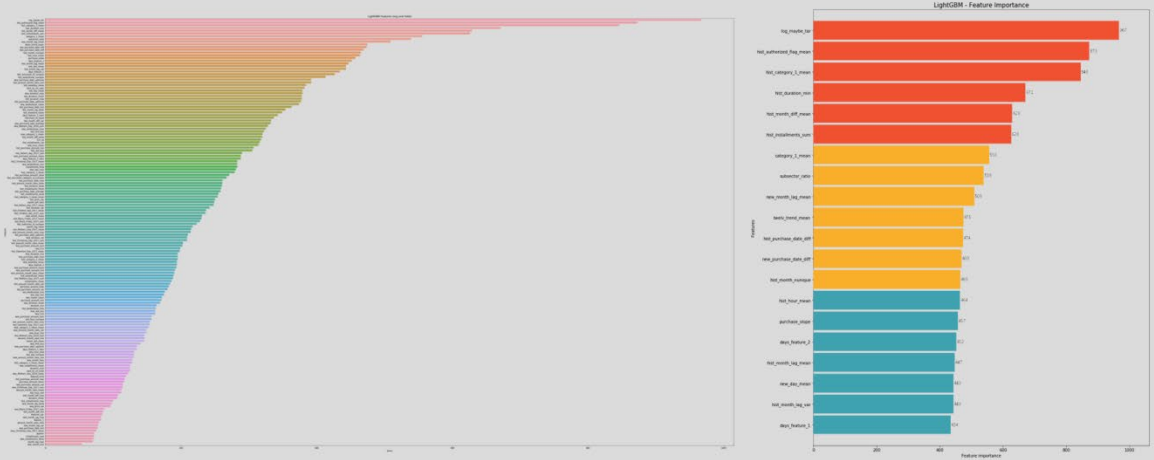
parameters= { "objective" : "regression",
"metric" : "rmse",
"num_leaves" : 30,
"min_child_weight" : 50,
"learning_rate" : 0.01,
"bagging_fraction" : 0.7,
"feature_fraction" : 0.7,
"bagging_freq" : 5,
"bagging_seed" : 2019,
"verbosity" : −1 }

- 동일 Tree-Based Model 중 Score가 가장 높게 나옴
- 가장 Train 속도가 빨라 효율적이면서도 정확도가 높음
- 단일모델 이용 시 가장 효과적이라는 것이 이미 검증
  * Kaggle Elo Competition 단일 모델중 가장 효과적이라고 이미 커뮤니티 내에서 충분히 디스커션이 이루어져 있음
- Tree-Based가 아닌 모델과 앙상블이 용이하고 평가함수 조정을 통해 최적화 모델을 쉽게 찾을 수 있음
- 모델 자체에서 과적합을 방지할 수 있어 과적합 문제가 발생 하기 쉬운 Elo Competition에 가장 적합한 모델

Ciel

데이터 분석 결과; 주요 영향 피쳐

Elo Merchant Category Recommendation



Loyalty Score에 영향도가 높았던 피쳐

LGBM모델은 Tree-Based model이기 때문에 모델 기여도 평가가 가능, Cross Validatrion이 줄어드는 정도로 영향도 판정

LGBM regency 관련된 피쳐들

month_diff, purchase date elapsed days(최근으로부터 마지막 구매 일까지의 차이)

Frequency관련 피쳐들

month 별 거래 변화량
과거 특정 기간 purchase_amount 와 예측된 미래의 purchae amount 의 관계 지

데이터 분석 결과; Score와 등수 변화

Elo Merchant Category Recommendation

Score 및 등수 급상승 Point 1 ; 주요 피쳐 생성 및 변경
Score 및 등수 급상승 Point 2 ; Cross Validation 분포 재조정

아직 최종 Competition이 종료되지 않았지만(2/27일 종료)
교육과정과 프로젝트를 통해 데이터 분석 경쟁에서 유의미한 점수와 등수 획득

*상위 10%(동메달)권 안착, 상위 5%(은메달)권 목표

Ciel

데이터 분석 결과

Elo Merchant Category Recommendation

What was LOYALTY SCORE?

Main Question: 대체 무엇을 예측해야 했는가?

1. 타겟값; 충성도의 정의

카드사에서 특정 상점에 대한 프로모션을 진행하기 일정 기간(3개월로 추정) 이전의
그 상점에 대한 트랜잭션 양과 프로모션 진행 이후 트랜잭션 양의 변화를 나타내는 지표

2. historical과 new merchant를 분리해 제공한 이유

프로모션이 진행된 후 historical 상점에 대한 트랜잭션이 target에 직접적 영향을 미치고,
프로모션 후 new merchant 소비 형태와 프로모션 전 historical 소비 분석을 통해
historical 데이터의 상점에 대한 거래량이 loyalty score일 것이라고 추측

3. 정규화된 target의 비밀

극단적인 outlier를 제외하고는 정규분포의 형태를 보이는 target값은 대다수의 feature와
상관성이 너무나 낮았음. 이는 자체 공식을 통해 일차적으로 산정한 충성도 점수를
로그처리 한 값이기 때문. 정규화를 풀고 로그처리 이전의 원본 값으로 변환하면
일반적으로 나타나는 불규칙 분포임을 확인

Ciel

데이터 분석 결과

Elo Merchant Category Recommendation

유의미한 파생변수 생성의 중요성

충성도 점수에 대한 이해를 바탕으로 고객의 소비 성향과 패턴에 관련한 피쳐들을 생성

대표 피쳐 1 프로모션 시작 기준일로부터 3개월간의 구매횟수와 프로모션 시작 이후의 구매 횟수를 비교한 변수

대표 피쳐 2 소비패턴의 변화와 관련해 구매량의 변화를 전체적으로 나타낼 수 있는 기울기를 반영

대표 피쳐 3 실제 마케팅과 고객 충성도 판단에서 사용하는 CLV(고객 생애 가치)지수나 RFM(최근성, 빈도, 금액)지수

» **kaggle 분석의 유의미한 파생변수 생성은**
도메인 지식과 Target값에 대한 명확한 이해가 필수적

데이터 분석 결과

Elo Merchant Category Recommendation

Feature Scaling Issue ; 데이터 왜곡 해결

정규화 된 연속형 feature들이 다수 존재. 이는 분석 결과를 왜곡을 발생시킬 가능성이 높아 각 컬럼별로 연산이 가능한 수로 스케일링하여 사용 *세부적인 방법은 보고서 참조

Target Uniformed ; 이상치 분포 조정

K-Fold 방식의 한계. 과적합은 방지하지만 각 fold별로 이상치의 영향력을 조정할 수 없어 모델의 불균형을 방생시키므로 타겟값의 분포를 균일하게 조정 후 학습 *세부적인 방법은 보고서 참조

Cross Validation or Leader Board? ; 과적합과 점수 신뢰도 파악

대회의 특성 상 교차검증 점수와 리더보드 점수 차이가 크기 때문에 신뢰도 문제가 존재.
교차검증 점수를 기반으로 모델 성능을 실험하면서 리더보드 점수를 향상시키는 방향 설정

* Kaggle의 스코어의 특징 : 테스트셋의 30%만 평가하는 Pubic Leaderboard와 나머지 70%의 테스트셋을 평가하는 Private Leaderboard로 구분하여 부분 과적합 위험이 있음

Lessons learned

Elo Merchant Category Recommendation

팀장 김하준

“ 데이터 분석에 있어 엔지니어링적 측면 뿐만 아니다 도메인 산업과 사람의 행동에 대한 이해가 바탕이 된 인문학적, 창의적 사고로 데이터를 대하는 것이 결과를 내는데 결정적이라는 사실을 배웠다.”

팀원 강호영

“ 완전히 가공되지 않은 데이터를 파악하고 어떻게 사용할지 설계하는 과정을 경험하며 많은 공부가 되었다. 상관성이 낮고, 실제 관련 없어 보이는 데이터를 활용하여 예측모델을 만들 수 있다는 것이 놀라웠다.”

팀원 조민정

“최종 과제로 관심있는 마케팅 분야에서 많이 사용하는 회귀분석과 머신러닝 기법을 직접적으로 사용하고 구축해보며 부족한 부분을 채울 수 있었고, 분석가로서 성장하기 위해 실질적으로 필요한 것들을 알아볼 수 있는 기회였다.”


[Q&A]

👋

감사합니다