

Machine Learning Project

GA 데이터를 활용한 고객 수익 예측 및 세그먼트 분류

INDEX

01

프로젝트 소개

- [프로젝트 주제](#)
- [프로젝트 목표](#)

02

데이터 소개 및 문제 해결 EDA

- 데이터 설명 및 주요 문제
- 데이터셋 이슈 해결
- 분석 방법 정의
- 모델 구성 도식

03

모델 개발

- 피처 분석 및 선별 과정
- 유효 피처 요약
- 주요 피처 설명
- 피처 엔지니어링

04

결과

- [프로세스 및 아웃풋](#)
- [분류모델](#)
- [회귀모델](#)
- [RFM 모델](#)

05

결과

- 고객 분류 결과
- 한계 및 아쉬운 점

프로젝트 소개

프로젝트 주제

프로젝트 목표

01

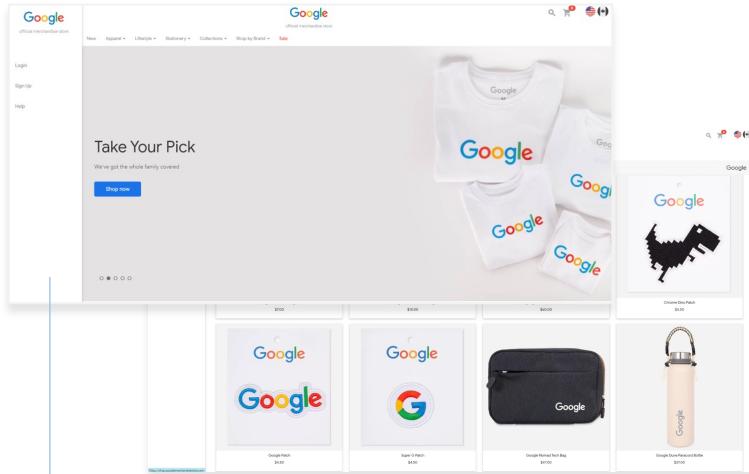
프로젝트 소개

프로젝트 주제

Gstore의 고객 로그 데이터를 활용한 VIP 타겟 분류

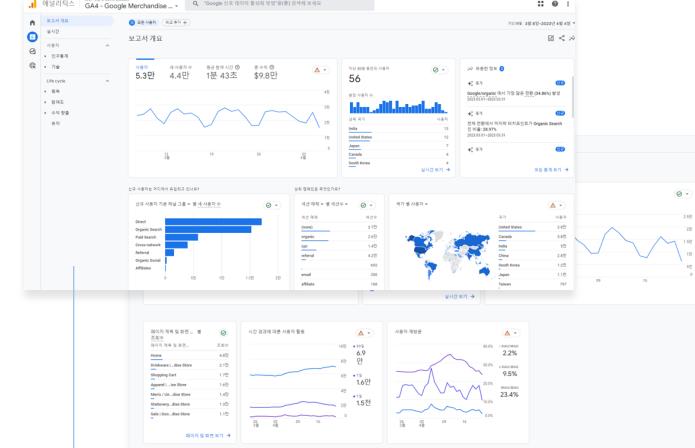
GStore의 마케팅팀에서 타겟 프로모션을 준비중인 상황

고객의 GA로그 데이터로 매출 기여도가 높은 VIP 타겟을 예측하여 전달하라



Gstore(Google Merchandise Store)

Google의 공식 MD 상품을 구매 가능한 온라인 쇼핑몰
<https://shop.googlemerchandise.com/>

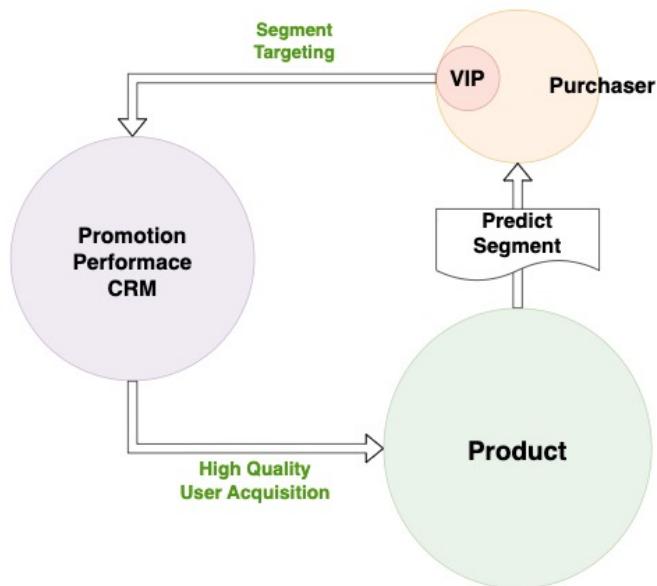


GA(Google Analytics)

구글 애널리틱스 구글에서 제공하는
웹사이트 트래픽을 추적하고 보고하는 분석 서비스

프로젝트 소개

문제 정의 및 목표 설정



20%가 80%의 매출을 발생시키는 파레토법칙에 따라
80%의 매출을 발생시킬 VIP 고객 세그먼트를 예측한다

VIP 세그먼트를 사전에 예측할 수 있다면

- 효과적인 프로모션을 위한 타겟군으로 활용할 수 있다.
- 효과적인 마케팅 수단으로 활용할 수 있다.

데이터 소개 및 문제 해결

데이터 설명 및 주요 문제

데이터셋 이슈 해결

분석 방법 정의

모델 구성 도식

02

데이터 설명 및 주요 이슈

분석 데이터

Train Data

2016.08.01
~2018.04/30
기간 동안
GA 로그 데이터

Test Data

2018.05.01
~2018.10.15
기간 동안의
GA 로그 데이터

feature

방문자별 페이지뷰, 유입경로, 날짜,
디바이스 등 약 150개 피처

label

방문자별 수익

주요 이슈

① 대용량 데이터셋

Train 데이터 23.67GB, Test 데이터 7.09GB로 용량이 커서
효과적인 데이터 처리와 로드 방법 필요

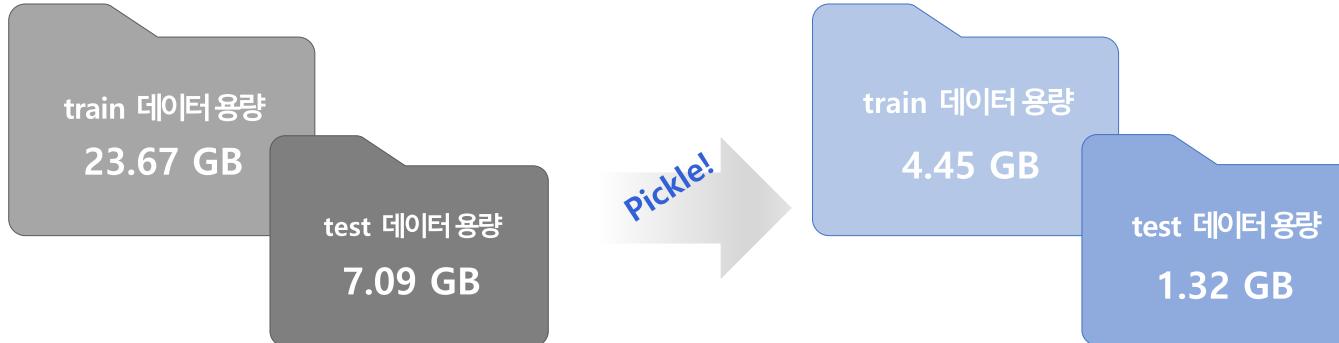
② Json 구조의 컬럼

CSV 파일 13개의 컬럼으로 제공되나, 이중 5개의 컬럼은
Json 구조로 파싱하여 피처를 150개로 만들어야 함

③ 데이터 불균형

Label이 되는 고객의 수익에서 0 값이 약 99%인 상황
이를 타개할 수 있는 모델링 방법 필요

데이터셋 이슈 해결



이슈 1,2

대용량의 데이터셋과 JSON 구조 컬럼

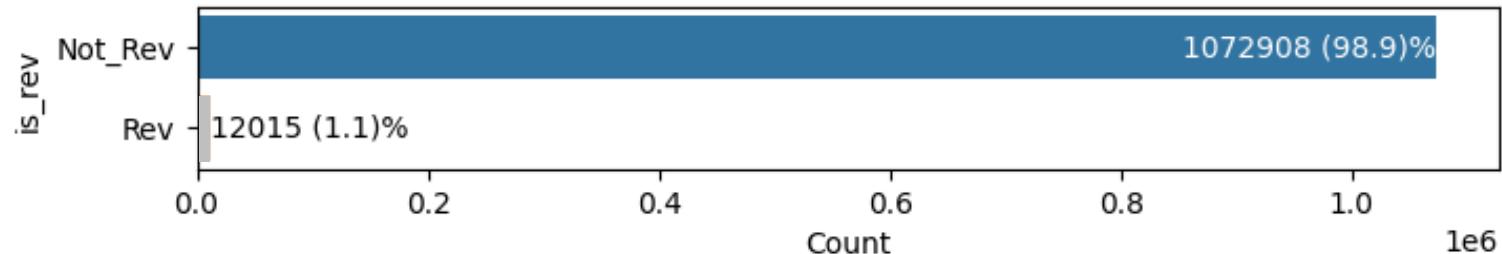
AS-IS

- 본 데이터 내 JSON 컬럼을 파싱하여 피처를 얻어야 함.
- 다만 메모리를 초과하는 용량으로, 쉽게 전처리가 불가능한 상황

TO-BE

- pd.chunksize 옵션을 통해 분할하여 df를 분할 로딩하여 전처리 후 합치는 방식 활용
- 이후, 전처리된 데이터프레임 pickle로 저장하여 메모리 사용량 감축
(데이터를 바이너리 형식으로 압축하여 저장. 따라서 추가적인 메타데이터 저장하지 않음)

데이터셋 이슈 해결



이슈 3

불균형한 데이터셋

AS-IS

- 매출이 0에 가까운 값이 대다수인, 정규분포를 따르지 않은 데이터셋 (커머스 데이터 특성상 매출이 발생한 데이터는 약 1.1%)
- 매출의 분산이 매우 큰 상황이지만, 타겟이기에 아웃라이어를 제거할 수 없는 상황.

TO-BE

- 가치의 상대적인 크기와 순서를 예측하는 것에 유리한 LOG 변환 채택
- 매출이 0인 값을 걸러 예측 모델의 성능을 높이는 프로세스 고안

99%의 0값을 분류하고, 1%의 예상 수익을 도출한 뒤,
RFM 분석기법을 통해 수익의 80%를 견인하는 소수의 고객 세그먼트 예측

① 분류 모델을 통해 구매자와 미구매자 분류

99%의 수익이 0원인 데이터를 걸러내어, 회귀모델에서 최적화된 수익 예측이 가능하도록 함

② 회귀 모델을 통해 구매자의 수익 예측

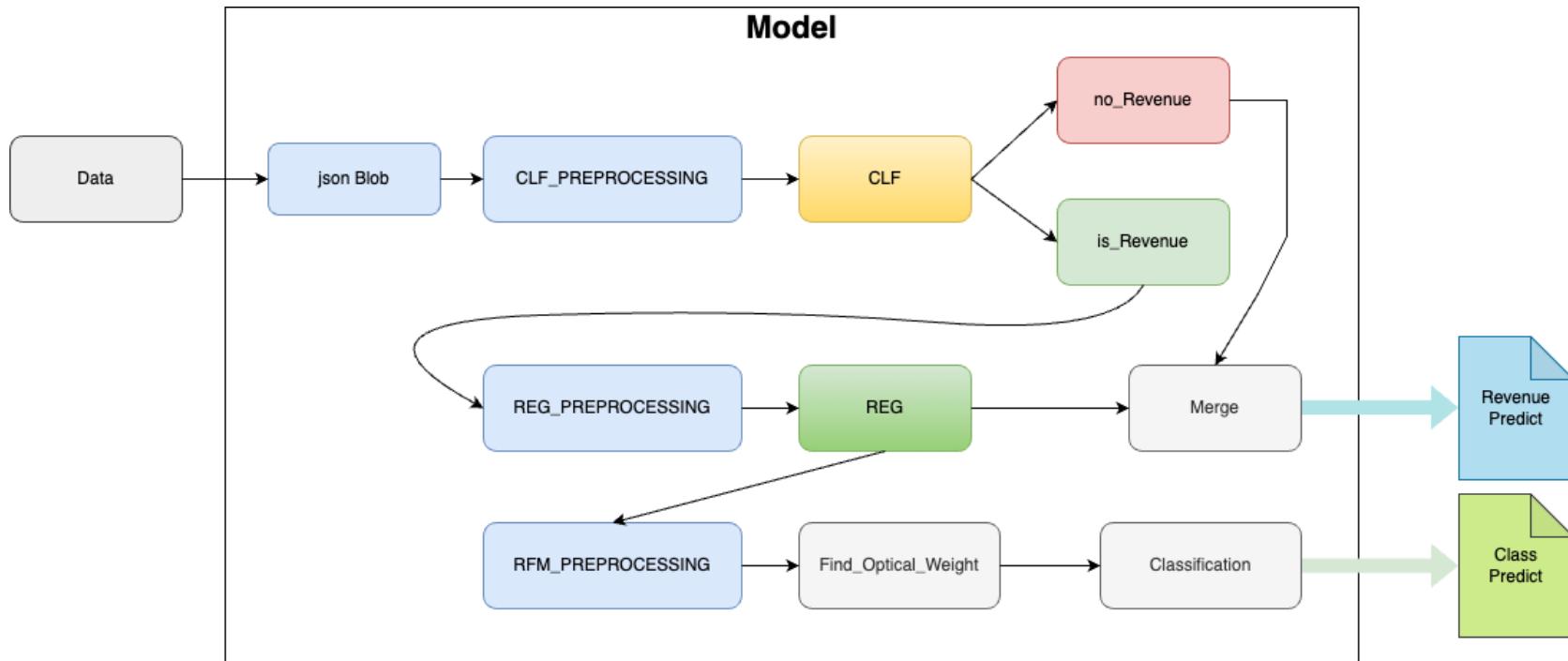
분류 모델을 통해 구매자로 예측된 데이터의 수익 예측

③ RFM 분석을 통해 세그먼트 추출

예측한 구매자의 데이터와 수익에 기반해 최근성, 구매빈도, 구매규모를 점수화하여 세그먼트 도출



모델 구성 도식



EDA

피처 분석 및 선별 과정

유효 피처 요약

주요 피처 설명

피처 엔지니어링

03

모델 적용을 위한 피처 선별 프로세스

유효 피처 선별

127개 → 46개

- 피처의 분포, 수익과의 관계 등을 시각화로 파악하여 유의미한 피처 선별
- 결측치의 개수, 타 피처와 중복 여부, 고윳값 개수 등 고려해 제거할 피처 선별

파생 피처 생성

46 개 → 59개

- 이커머스 데이터 특성을 활용하여 중요한 파생 피처 생성 (상품 관심도, 재방문주기, 총 생애주기, 충성도)
- 시계열 피처(연, 월, 요일, 주) 생성
- 분류모델에서 label이 될 구매, 비구매 여부 label 피처 생성

최종 피처 확정

59 개 → 10개

- 59개의 피처를 통해 분류모델을 돌린 뒤 importance가 높은 상위 10개의 피처 선별
- 해당 피처를 최종으로 분류 모델과 회귀 모델에 적용하며 성능 최적화

유효 피처 및 파생 피처 요약

유효 피처

- 127개의 피처 중 1차로 유효하다고 판단한 피처
- 지리정보, 유입정보, 상품 정보 등 총 46개 피처

1 totals_hits	24 hits_hitNumber
2 totals_pageviews	25 hits_hour
3 totals_sessionQualityDim	26 hits_appInfo.exitScreenName
4 totals_timeOnSite	27 hits_appInfo.landingScreenName
5 visitNumber	28 hits_dataSource
6 channelGrouping	29 hits_eCommerceAction.action_type
7 totals_bounces	30 hits_eCommerceAction.option
8 totals_newVisits	31 hits_eCommerceAction.step
9 channelGrouping	32 hits_eventInfo.eventAction
10 device_browser	33 hits_eventInfo.eventCategory
11 device_deviceCategory	34 hits_eventInfo.eventLabel
12 device_operatingSystem	35 hits_item.currencyCode
13 geoNetwork_continent	36 hits_page.hostname
14 geoNetwork_subContinent	37 hits_page.pagePath
15 geoNetwork_country	38 hits_page.searchCategory
16 trafficSource_adContent	39 hits_page.searchKeyword
17 trafficSource_campaign	40 hits_referer
18 trafficSource_keyword	41 hits_isEntrance
19 trafficSource_medium	42 hits_isExit
20 trafficSource_referralPath	43 hits_isInteraction
21 trafficSource_source	44 hits_promotionActionInfo.promolsView
22 trafficSource_isTrueDirect	45 hits_promotionActionInfo.promolsClick
23 trafficSource_adwordsClickInfo.isVideoAd	46 hits_exceptionInfo.isFatal

파생 피처

- 이커머스 특성을 활용한 중요 지표 생성

- 상품 관심도

상품 클릭 엔트로피 피쳐 생성

- 재방문주기

고객 별 최초 방문일로부터 재방문일 까지 걸린 기간

- 총 생애 주기

총 방문일 간 쇼핑몰 방문 주기

- 충성도

재방문율을 계산한 충성도

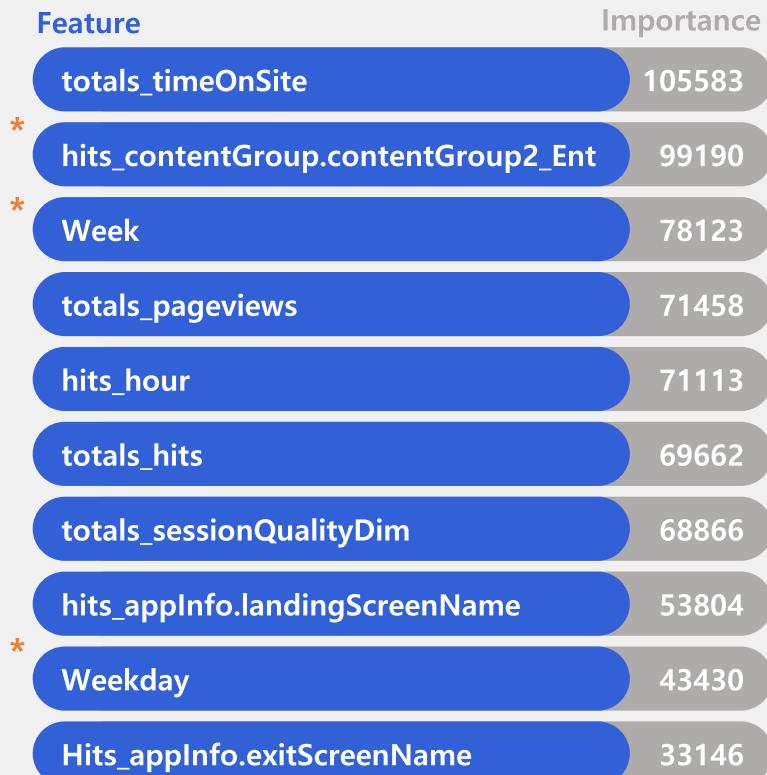
- 연, 월, 요일, 주의 시계열 피처 생성

1 agg_total_visit_count
2 total_life_time
3 monthly_visit
4 weekly_visit
5 weekdaily_visit
6 revisit_dur_time
7 month
8 week
9 weekday
10 stickiness
11 hits_contentGroup.contentGroup2_Ent
12 trafficSource_adwordsClickInfo.gcld_Ent

위 59개 피처 중 중요도 높은 피처를 찾기 위해 분류모델에 테스트로 돌려 feature importance 상위 10개의 피처를 최종으로 선정

최종 확정 피처 설명

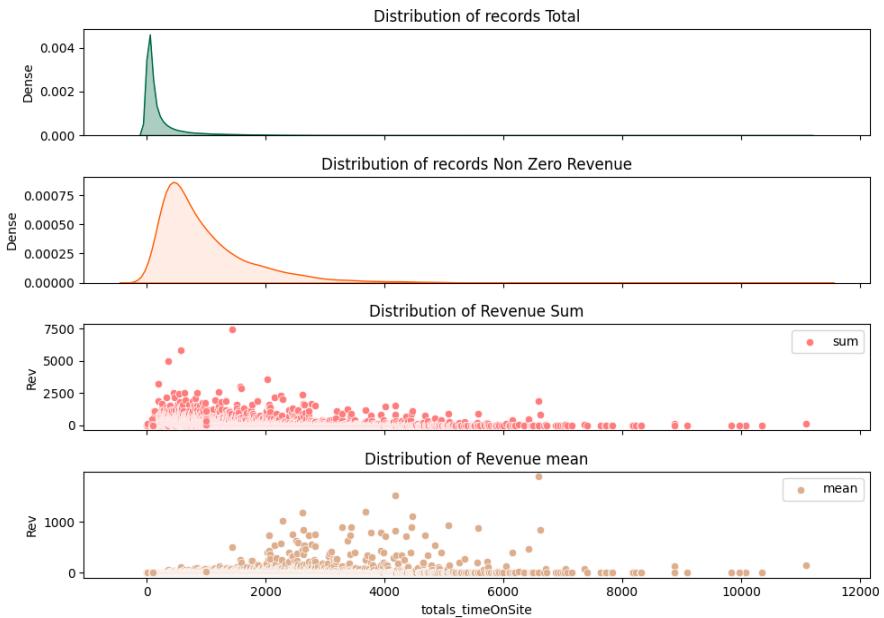
분류모델을 통해 도출한 Feature Importance 상위 10개 피처 (LGBM Classifier 적용)



- 새롭게 생성한 파생변수*가 3개 포함되어 중요한 역할을 함
- 위 상위 10개의 피처를 분류와 회귀모델에 적용

totals_timeOnSite

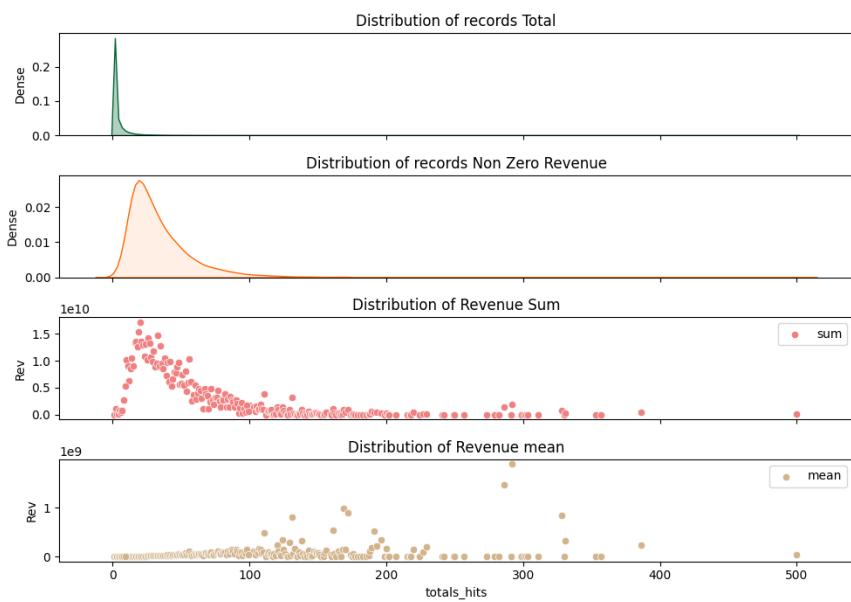
- 고객의 사이트 during time
- 시간이 지속됨에 따라 평균 수익이 늘어나는 경향
- 구매자의 오른쪽 꼬리가 다른 피처 대비 두꺼움



최종 확정 피처 설명

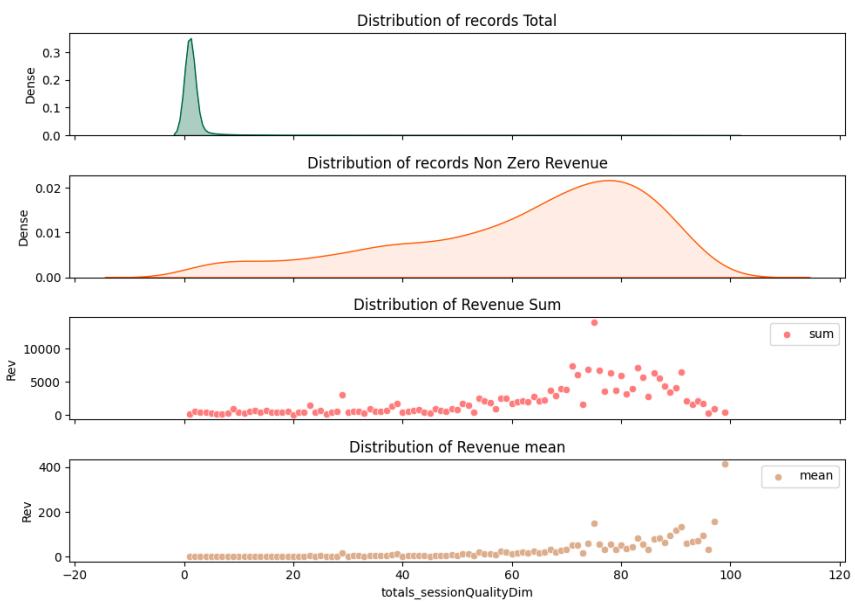
totals_hits

- 고객이 페이지 내에서 hit 한 수
- **Totals_pageviews** 피처와 0.97의 강한 상관관계
- 구매자의 경향을 확인할 수 있음



Totals_sessionQualityDim

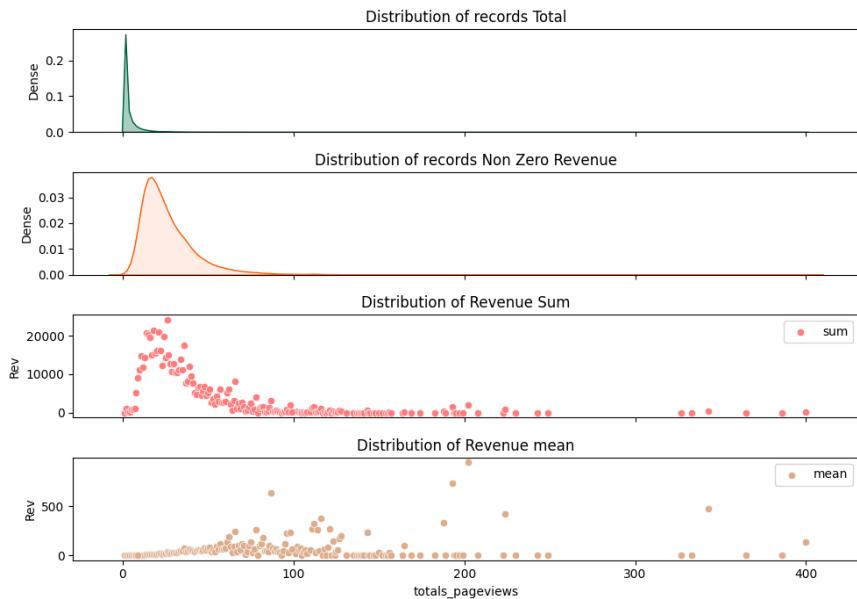
- GA에서 자체적으로 세션을 평가하는 지표
- Session과 상품 조회 횟수 등을 평가하는 것으로 추정
- 매출과 점수의 양의 상관관계가 명확히 드러남



최종 확정 피처 설명

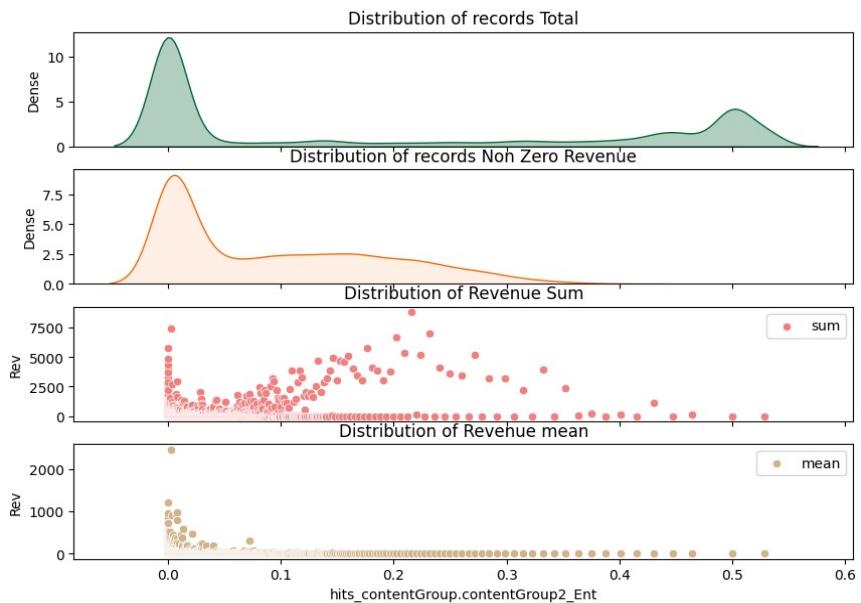
Totals_pageviews

- 고객이 페이지를 본 수
- 구매자와 비구매자의 분포가 비슷하나,
밀집되어 있는 위치가 다르므로 유의미함
- 수익의 격차가 매우 큰 점을 고려, 유의미한 차이가 있음



hits_contentGroup.ContentGroup2_Ent *

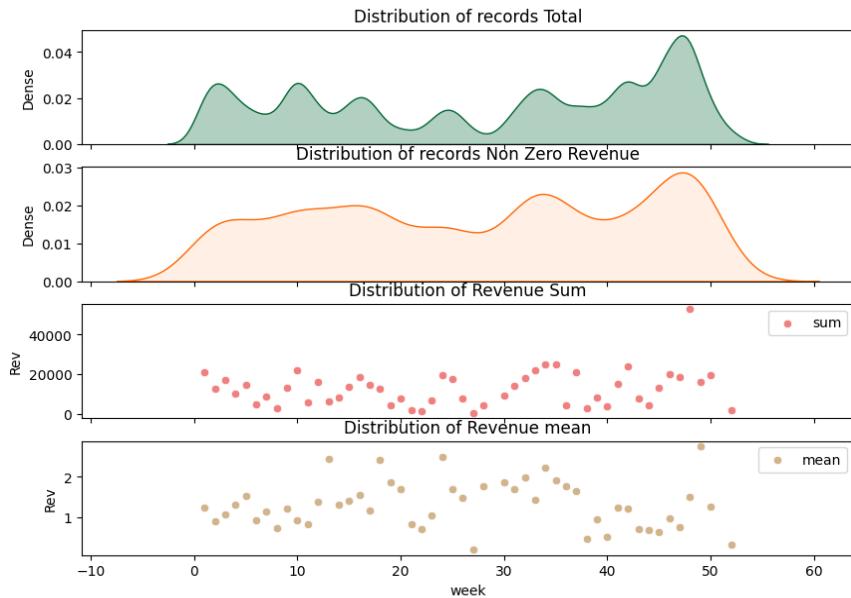
- Hit_contentGroup.contentGroup2의 엔트로피
- 세션 내 클릭한 상품들의 종류를 히트로
엔트로피를 계산한 파생변수
- 비구매자를 명확히 구분하는 경향을 보임



최종 확정 피처 설명

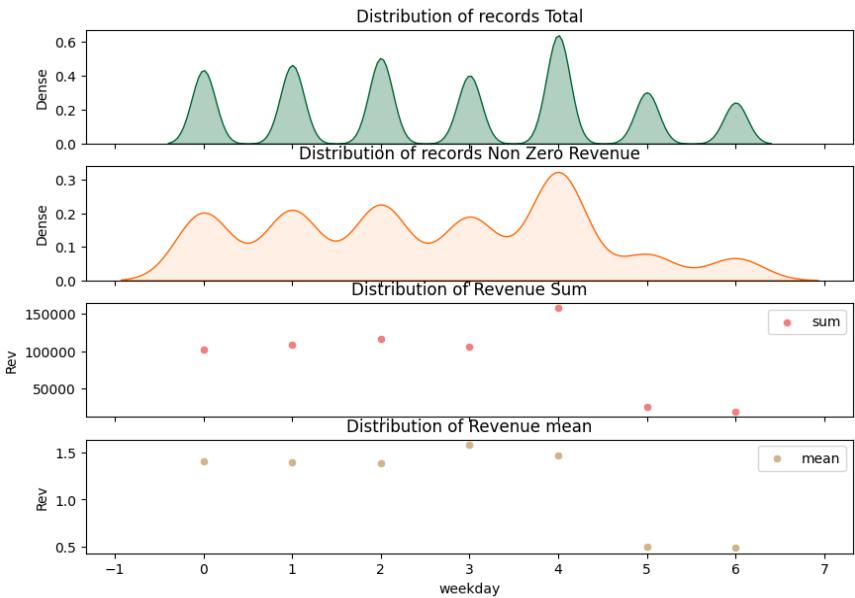
Week *

- 고객 방문 ISO WEEK
- 데이터셋의 계절성을 파악할 수 있음.



Weekday *

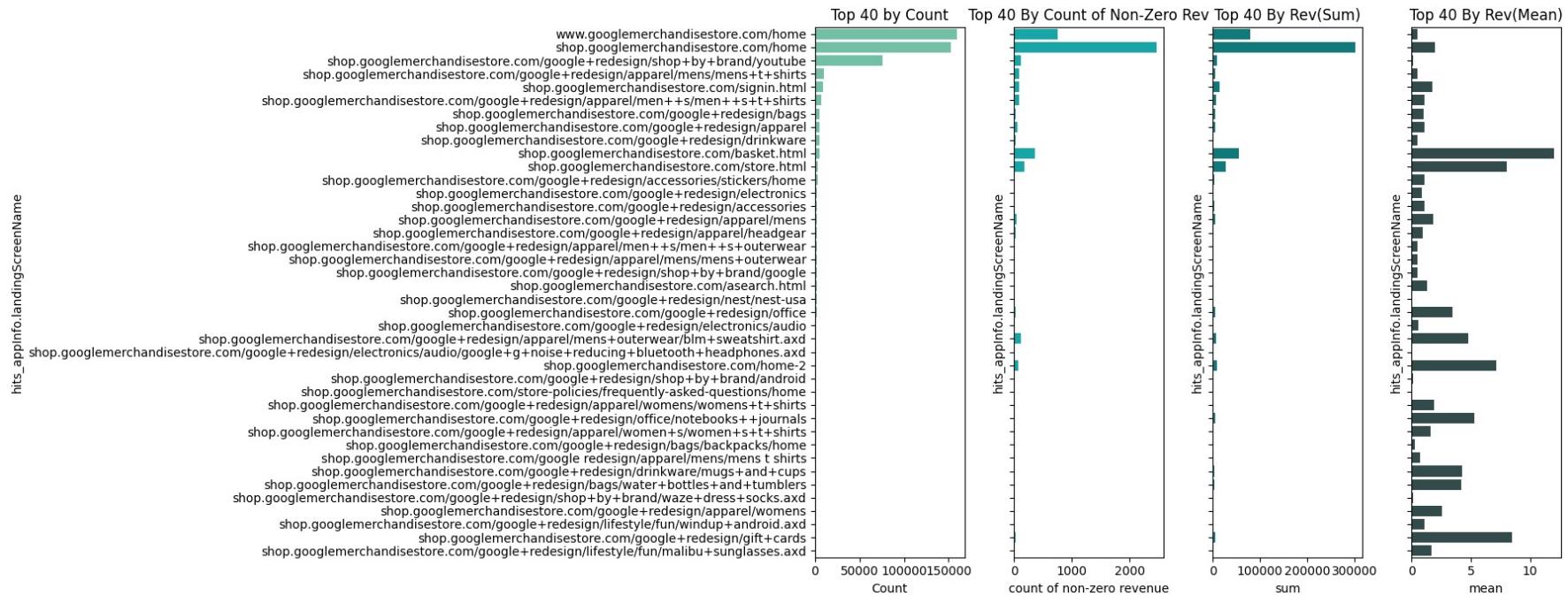
- 고객 방문 요일
- 데이터셋의 주 수익 창출 요일을 파악할 수 있음.



최종 확정 피처 설명

Hits_appInfo.landingScreenName

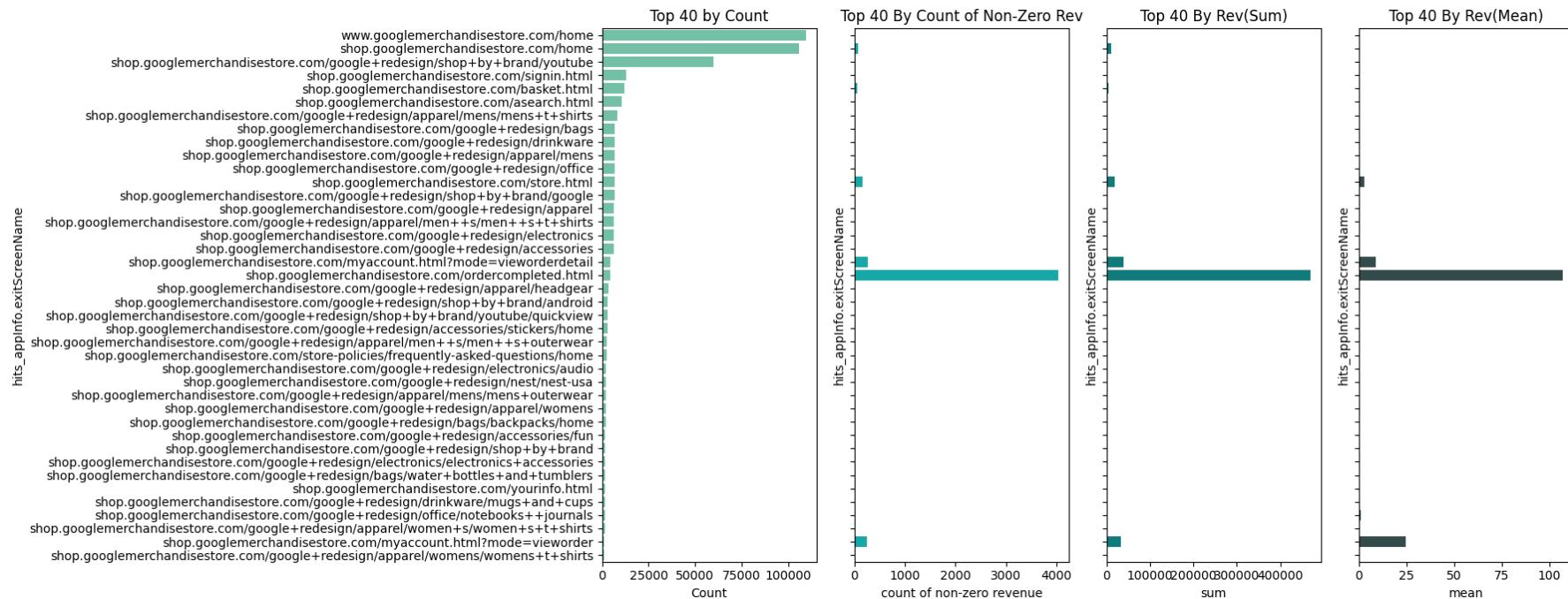
- 고객의 랜딩 화면 이름
- 재방문자의 유입경로와 신규의 유입경로가 다를 수 있음 (URL 자동 완성)
- 따라서 구매자와 비구매자를 어느 정도 구분할 수 있는 피처



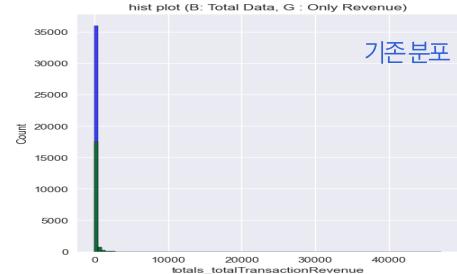
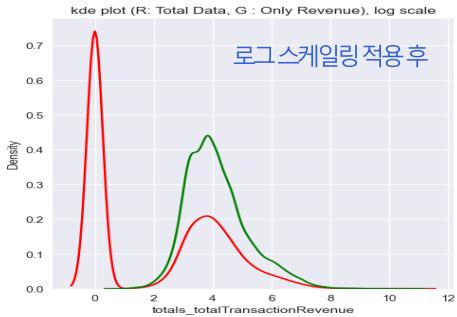
최종 확정 피처 설명

Hits_appInfo.exitScreenName

- 고객의 종료 화면 이름
- 1 세션 내 종료하는 비정상 유저, 비구매자를 어느정도 구분할 수 있는 피처
- 따라서 구매자와 비구매자의 명확한 차이를 가지고 있음



피처 엔지니어링

결측치 처리	명목형 피처 인코딩	수익 label 로그 스케일링
<p>개인정보보호 등의 이유로 결측처리된 값들은 각 피쳐에 맞는 결측처리 진행</p> <p>ex01/ hits_appInfo.landingScreenName → 0.1%가 결측치 결측치에(Not Recored) 값 배정</p> <p>ex02/ totals_sessionQualityDim → 48%가 결측치 1부터 시작하는 연속형 타입으로 결측치에 0배정</p>	<p>모델 적용을 위해 명목형 피처 인코딩 전략 필요, 라벨 인코딩 적용</p> <ul style="list-style-type: none"> Label 인코딩이 One-Hot 인코딩에 비해 인코딩 후 데이터 용량 부담이 적으며, Tree 계열 분류/회귀모델에서는 label과 One-Hot 인코딩의 성능차이 미비 실제 lgbmRegressor 에서 두 인코딩에 따른 성능을 비교했을 때 label 인코딩의 성능이 다소 높았음 <ul style="list-style-type: none"> 라벨 인코딩 시 RMSE : 1.22446 원핫 인코딩 시 RMSE : 1.30397 	<p>회귀 모델에 적용할 수익 label이 낮은 값에 몰려 있어 로그 스케일링 적용</p> <ul style="list-style-type: none"> 낮은 수익 값에 몰려 있어(극단적인 Right Skewed) 모델 성능 저하가 우려되는 상황  <p>기존 분포</p> <ul style="list-style-type: none"> 왜도 완화에 효과적인 로그 스케일링을 적용  <p>로그 스케일링 적용 후</p>

모델 개발

프로세스 및 아웃풋

분류모델

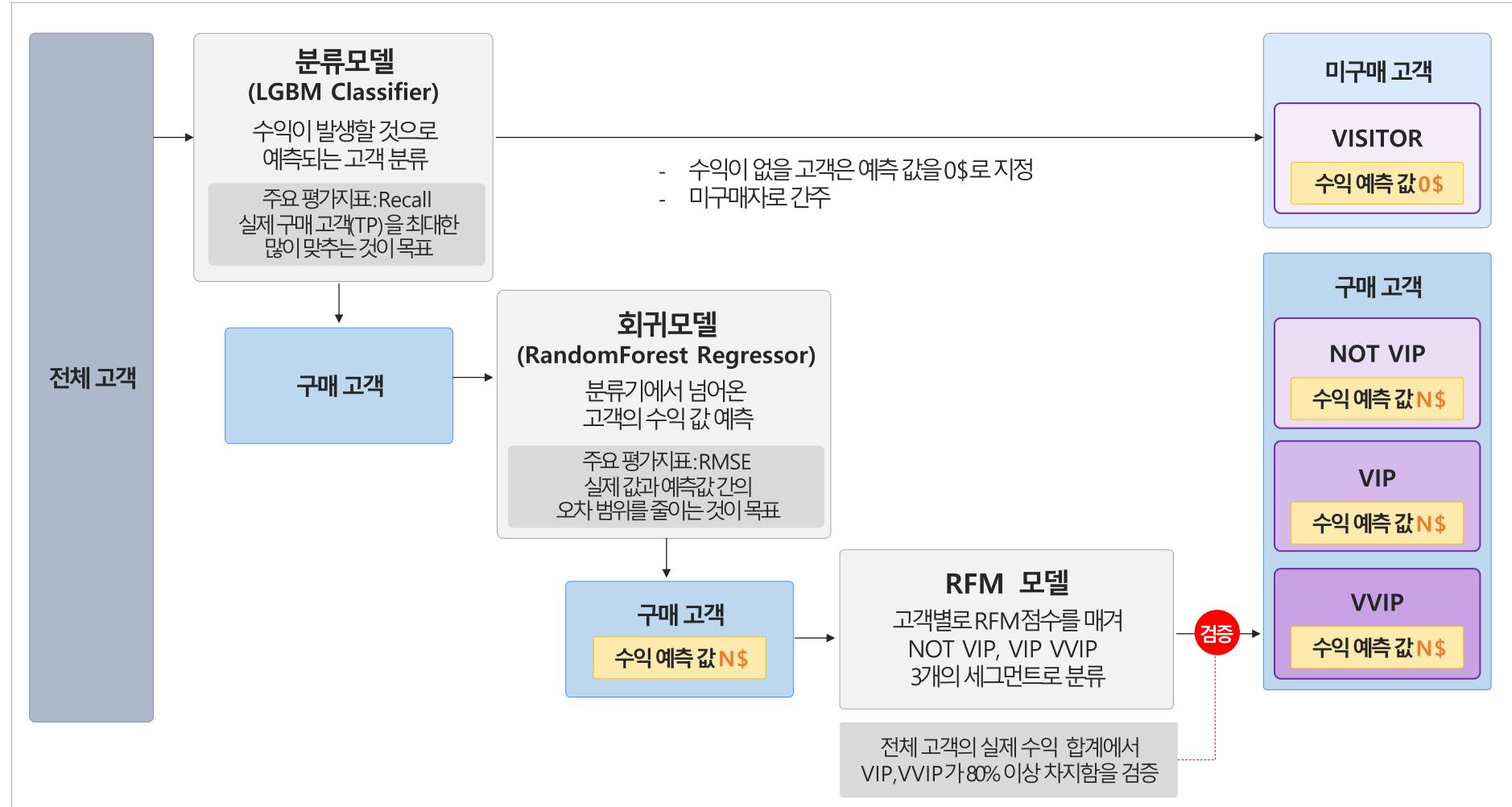
회귀모델

RFM 모델

04

모델 프로세스 및 아웃풋

분류, 회귀, RFM 3개의 모델을 통해 고객의 개별 수익을 예측하여 3개의 세그먼트(NOT VIP, VIP, VVIP)로 분류



분류 모델

**1%의 구매 고객과
99%의 미구매 고객을
분류하는 단계**

- 본 데이터에서는 구매자가 1%대로 매우 적기 때문에, 이를 최대한 잘 찾아내는 것이 중요한 Key
- 분류 평가 지표 중 Recall은 모델이 실제 양성 샘플을 얼마나 잘 감지하는지를 나타냄
- 소수의 구매자(양성)를 잘 분류해내어 Recall값을 개선시키는 것을 모델 최적화의 목표로 삼음
- 최종적으로 가장 향상된 Recall 지표 0.86를 갖는 모델을 선정하였음

데이터

피처

totals_timeOnSite, week,
hits_hour, totals_hits,
totals_pageviews 등 10개

라벨

Is_revenue(수익 발생 여부)

모델

LightGBM Classifier

하이퍼 파라미터 셋팅하지 않을 때 가장 성능 좋았음

다양한 분류 모델 실험 결과 속도와 Recall에서 우수하여 채택

모델	Time	Recall
Logistic Reg	15.886	0.71
RandomForest Clf	8.579	0.82
GradientBoost Clf	68.754	0.81
XGBoost Clf	16.436	0.81
LightGBM Clf	2.349	0.86

성능

Recall

Train Data : 0.96

Test Data : 0.86

Test Data의 Classification Report

	Precision	Recall	F1-Score	Support
0	0.998	0.995	0.997	396995
1	0.668	0.859	0.752	4594
accuracy			0.994	401589
Macro avg	0.833	0.927	0.874	401589
Weighted avg	0.995	0.994	0.994	401589

ROC-AUC score : 0.9267907841595555

Threshold : 0.5

회귀 모델

구매 고객의 수익을 예측하는 단계

- 분류 모델을 통해 넘어온, 구매 고객의 수익을 실제와의 오차(error)를 최소화하며 예측하는 것이 목표
- 회귀모델의 평가 지표 중 예측값과 실제값 사이의 평균 차이를 측정하는 RMSE가 가장 좋은 모델로 최종 선정
- 로그 스케일링을 통해 대다수의 매출이 0에 가까이 분포한 이슈 해소 (p.nn 참고)
- 목적 달성 이후, 지수변환을 통해 달러 스케일 복구

데이터

모델

성능

피처

totals_timeOnSite, week, hits_hour, totals_hits, totals_pageviews 등 10개

라벨

total_transactionRevenue
log1p 스케일링(고객의 수익 값)

RandomForest Regressor

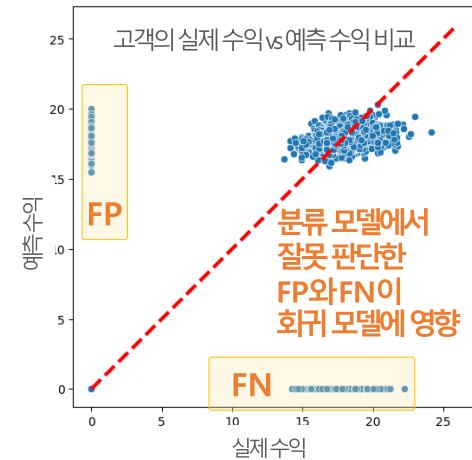
하이퍼파라미터 설정하지 않을 때 가장 성능 좋았음

다양한 회귀모델 실험결과 속도와 RMSE에서 우수하여 채택

모델	Time	RMSE
Linear_Reg	2.007	1.918
Quadratic Reg	34.259	1.755
RandomForest Reg	57.899	1.428
GradientBoost Reg	74.648	1.511
XGBoost Reg	18.566	1.514
LightGBM Reg	3.339	1.455

RMSE

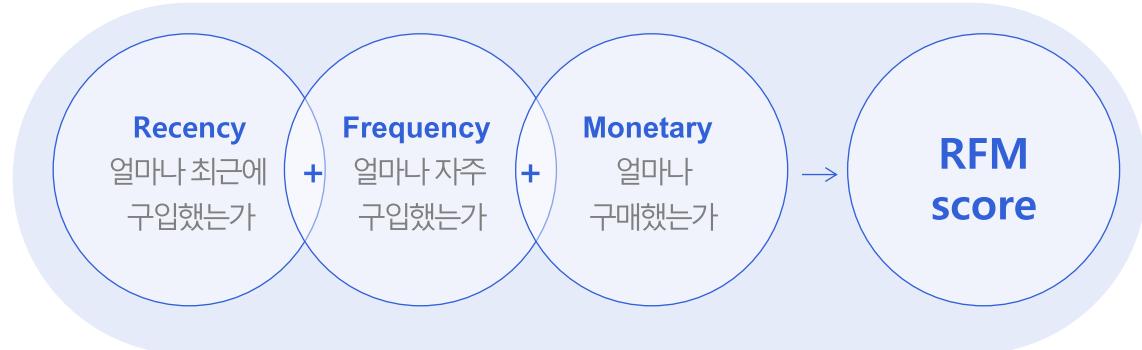
Train Data : 0.06043792898194

Test Data : 1.42843443308592

RFM 모델

구매고객의 예측 수익과 정보에
기반해 RFM 점수를 도출하고
고객 세그먼트를 분류하는 단계

RFM 분석의 이해



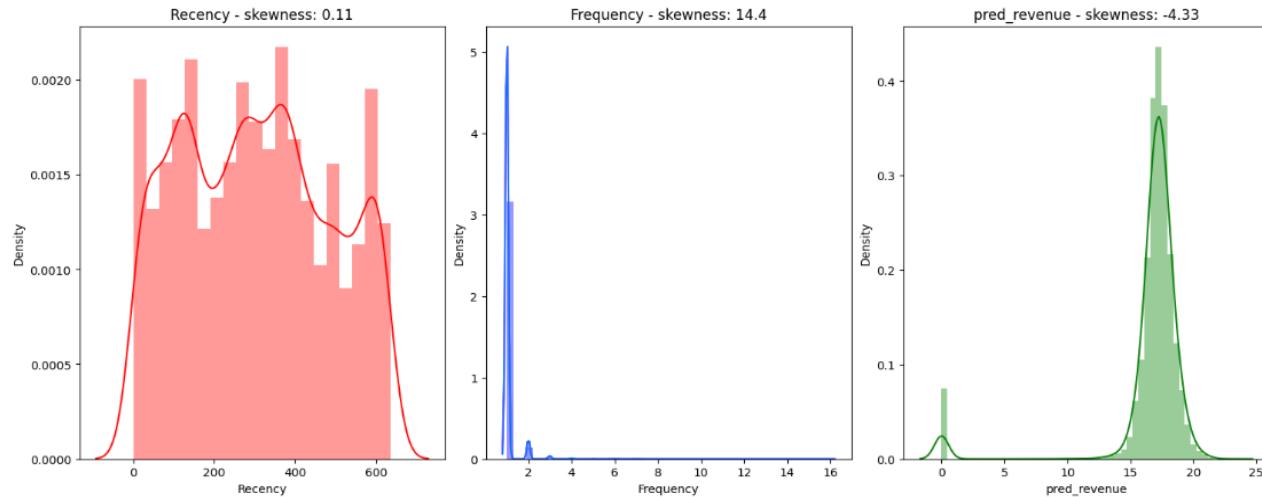
- RFM은 CRM(Customer Relation Management, 고객관계관리)의 방법론 중 하나로 점수부여방식(Scoring system)의 분석 기법이며 고객 세분화 작업 시 자주 활용
- 고객별 Recency, Frequency, Monetary 점수를 산정한 뒤 목적에 따라 각 점수별 최적 가중치를 적용 후 합산하여 RFM 점수 도출

RFM 모델 설계

STEP 01 분류-회귀 모델을 거친 구매고객의 예측 수익과 data를 기반으로 R, F, M 계산

- Recency : 데이터의 기준일(2018.05.01)과 고객의 최근 구매일자 간의 차이
- Frequency : 고객의 예측 구매 빈도
- Monetary : 고객의 예측 수익의 총합

구매 고객의 Recency, Frequency, Monetary의 분포



RFM 모델 설계

STEP 02 R, F, M 각각의 값을 3 Grade의 점수로 변환 후, 최적의 가중치 선정

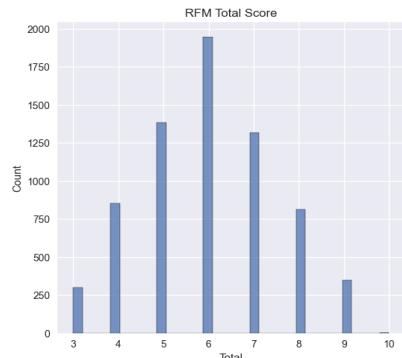
- 실험결과, 3, 4, 5 Grade 중 高가치 고객을 보다 잘 분류하는 3 Grade 채택
(3 Grade는 총 수익의 81%를 견인하는 고객을 분류함. 4 Grade는 67%, 5 Grade는 75%)
- 최적의 RFM 점수 구간을 선정하기 위한 가중치 탐색 (80% 목표)

Recency, Frequency, Monetary 3grade 스코어링

> Recency, Frequency, Monetary 각 가중치 선정

> 최적 가중치

- Recency, Frequency, Monetary 각 수치 별 적절한 구간으로 나눠 1~3점까지 스코어링
- RFM 점수를 단순 합산할 경우 아래와 같이 균등한 분포



RFM 점수 상위 20%의 고객이
전체 수익 중 80%를 차지할 수 있는 가중치 도출

Recency : 0.0
Frequency : 0.462
Monetary : 0.538

Optical Weight 선정 절차

- ① 가능한 모든 경우의 수(각 가중치의 합 = 1)에 대한 가중치를 선택한다.
- ② ①의 가중치를 이용하여 각 가중치별 RFM 점수 상위 20%의 전체 매출 비중을 산출한다.
- ③ 3RFM 점수 상위 20%의 전체 매출 비중이 80%와 가장 가까운 가중치를 선택한다

RFM 모델

고객 분류 및 RFM 모델 검증

- 가중치를 적용한 RFM 점수 구간별로 고객 세그먼트 분류
- 회귀모델을 통해 예측한 수익에 기반한 RFM 모델의 성능을 검증하기 위해
고가치 세그먼트 고객이 실제로도 높은 수익 비중을 차지하는지 확인

① 고객별 가중치 적용한 RFM 점수 계산

$$0 * \text{Recency} + 0.462 * \text{Frequency} + 0.538 * \text{Monetary} = \text{RFM Score}$$

② RFM 점수 구간에 따른 3단계 고객 분류

RFM 점수 상위 20% : VVIP / RFM 점수 상위 21~60% : VIP / RFM 점수 상위 61~100% : NOT VIP

③ 고객 세그먼트별 실제 수익 확인을 통한 검증

Train	고객 수	예측 수익	실제 수익
Not VIP	584369 (99.2%)	52973 (9.4%)	53887 (5.9%)
VIP	2345 (0.4%)	113556 (20.1%)	147598 (16.3%)
VVIP	2412 (0.4%)	399829 (70.6%)	705032 (77.8%)

Train	고객 수	예측 수익	실제 수익
Not VIP	292761 (98.7%)	49811 (14.3%)	105981 (19.0%)
VIP	1857 (0.6%)	91326 (26.2%)	120459 (21.6%)
VVIP	1912 (0.6%)	207798 (59.6%)	331906 (59.4%)

결과

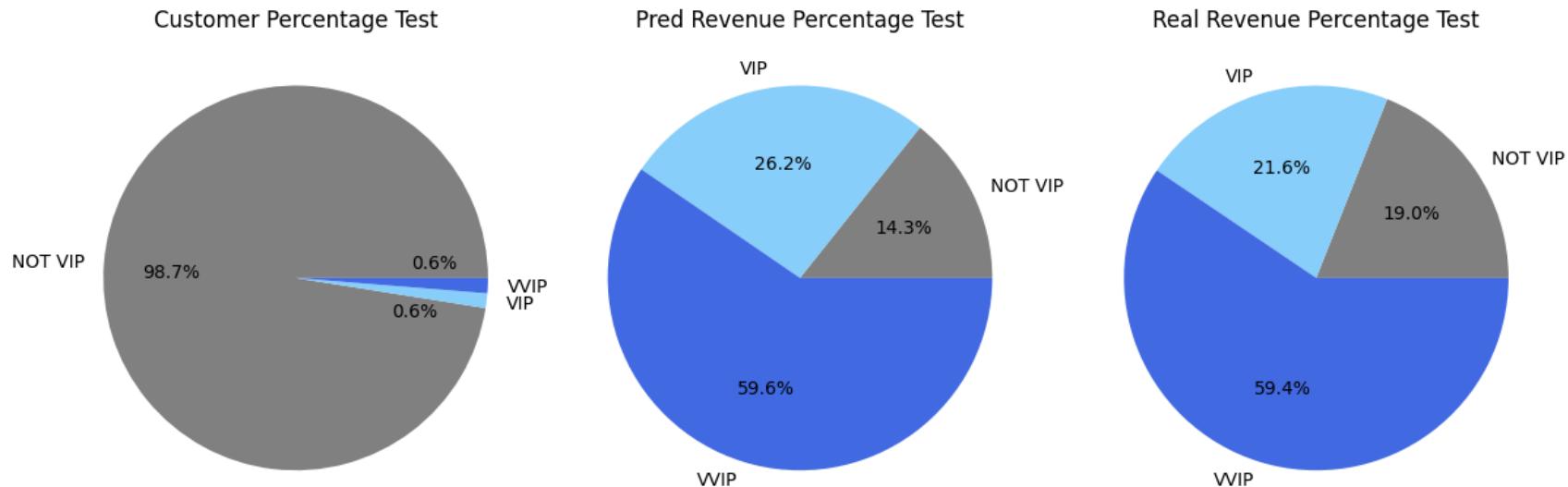
고객 분류 결과

한계 및 아쉬운 점

05

결과

최종 고객 분류 결과



Test 데이터의 고객 296,530명 중 1.3%인 약 3769명을 VIP(VVIP + VIP)고객으로 분류

VIP 고객의 수익 기여 VIP 고객은 회귀모델을 통해 예측한 수익의 85.8%를 발생시키며, 실제로는 81.0%의 수익을 발생시키는 것을 검증

한계 및 아쉬운 점

- 원본데이터의 큰 데이터로 인해 전체 훈련데이터를 사용하지 못하고 40%의 데이터로 프로젝트를 수행하였다
- 메모리 부족으로 인해 분류, 회귀 모델의 최적 파라미터 값을 찾지 못함
하이퍼파라미터 튜닝 시, 성능 개선의 여지 존재
- 전체 데이터의 크기는 충분히 크나, 수익이 있는 세션만으로 분석하기에는 적은 세션 수로 인한 분석의 신뢰도 하락
- 분류모델에서의 FP(False Positive), FN(False Negative)이
회귀분석, RFM분석에서 노이즈로 발생

Reference

- 주영혁, 한상만. 수익성 있는 고객의 웹사이트 방문행동특성에 관한 연구: 수익모델간 비교를 중심으로. *마케팅연구*, 2001, 16.2: 69-91.
- 김동석, et al. RFM 모형의 가중치 선택에 관한 연구. 2021. PhD Thesis. 제주대학교 대학원.
- 김규곤, et al. 고객 세분화를 위한 최적 RFM 모형 구축에 관한 연구. *Journal of the Korean Data Analysis Society*, 2004, 6.6: 1829-1840.
- 이영진, et al. CRFM 모형을 이용한 고객세분화와 모형평가. *Journal of The Korean Data Analysis Society*, 2010, 12.6: 3283-3293.

E.O.D