

브라질 이커머스 플랫폼 고객 만족 분석

고객 만족 데이터 분석을 통한 사업 개선 방향 제시 및 만족도 점수 예측 모델 개발

0. Summary

- (Kaggle)브라질 이커머스 기업 데이터셋
<https://www.kaggle.com/olistbr/brazilian-ecommerce>
- 실제 기업 데이터를 활용하여 비즈니스 문제 정의 및 인사이트 도출을 위한
- 프로젝트 기간 : 21.02.01 - 21.03.25(8주)
- 프로젝트 참여 인원 : 3명
- 언어 : Python
- 기술스택 : Pandas, Numpy, Seaborn, Matplotlib, Sklearn, lightGBM

1. Goal

- 프로젝트 배경 및 목표 선정
 - Olist 수익 구조 ➔ 판매자를 통한 수익 창출
 - 판매자가 잔존하여 지속적인 판매를 해야 기업 수익이 증가
 - 판매자의 지속 성장이 중요한데, 만족도 점수가 낮을수록 성장에 악영향을 미칠 것으로 예측
- 프로젝트 목표
 - 1) 데이터 분석을 통한 비즈니스 현황 파악 및 사업 개선 방향 제시
 - 2) 실제 Olist DB에서 만족도 데이터가 없는 구매 건의 만족도 점수 예측 모델 개발

2. Preprocessing

➤ EDA

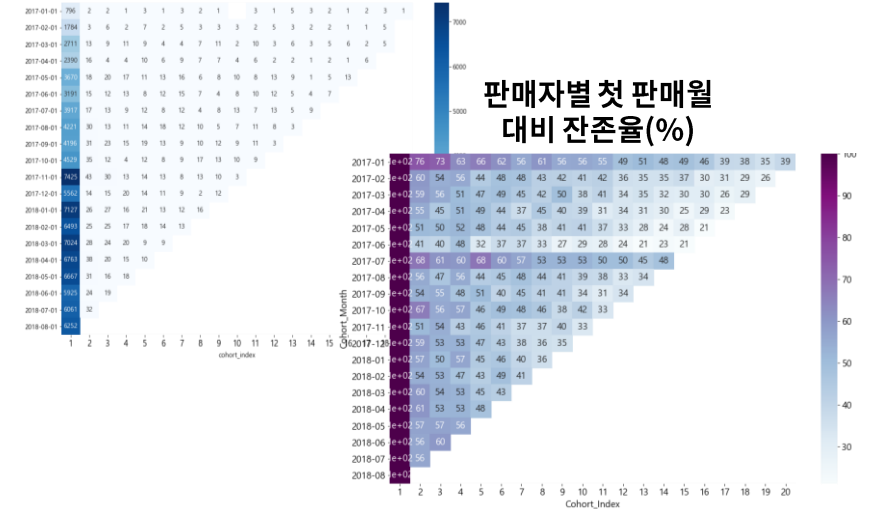
1) 고객 & 고객 데이터

- 고객은 약 97% 이상이 1회만 구매
- 반면, 판매자는 지속적 상품 판매 (상대적)

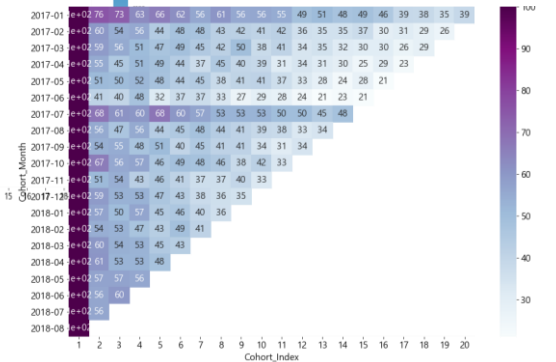
2) 리뷰 데이터

- 전체 리뷰 중 불만족 비율이 약 23%(1~3점)
- 리뷰 커멘트에서 전반적으로 Product, delivery(receive, came 등)이 언급이 많이 된 것으로 보아 배송 관련 사항이 많을 것으로 판단

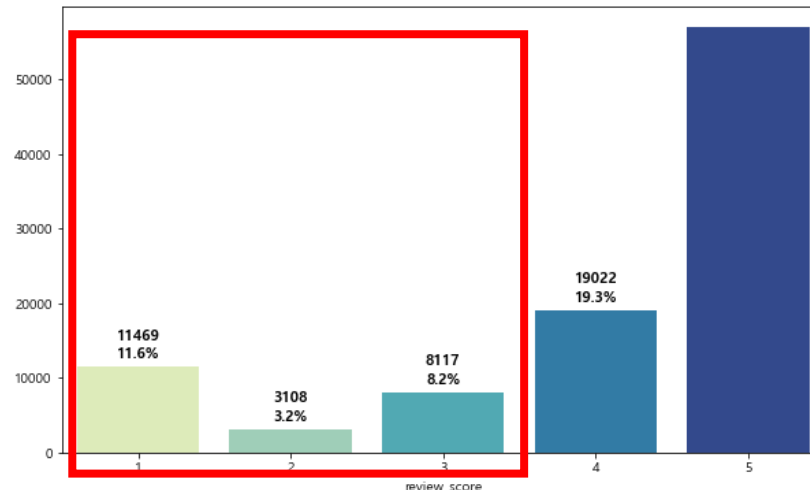
고객별 첫 구매월 대비 재구매 건수



판매자별 첫 판매월 대비 잔존율(%)



리뷰 스코어 비율



만족

불만족



브라질 이커머스 플랫폼 고객 만족 분석

고객 만족 데이터 분석을 통한 사업 개선 방향 제시 및 만족도 점수 예측 모델 개발

2. Preprocessing

➤ Feature Engineering

기준 : Olist측에서 리뷰가 있는 데이터만 추출하여 제공

➔ 리뷰가 없는 데이터에 적용 가능한 변수 생성이 목표

1) 목표 변수 선정

- review score 1,0 (1 = 1~3점 / 0 = 4~5점)

➔ 불만 사항 개선으로 판매자 수익성 및 잔존율 향상을 위해
1~3점의 만족도 점수를 주 관심 대상으로 선정

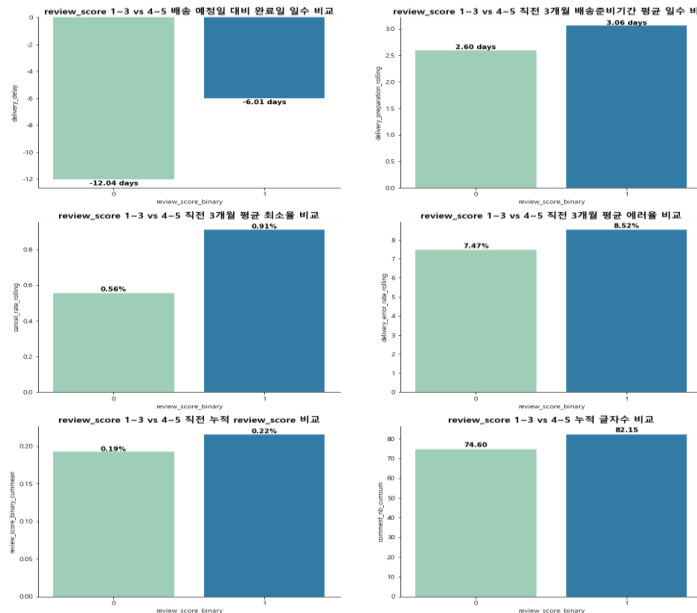
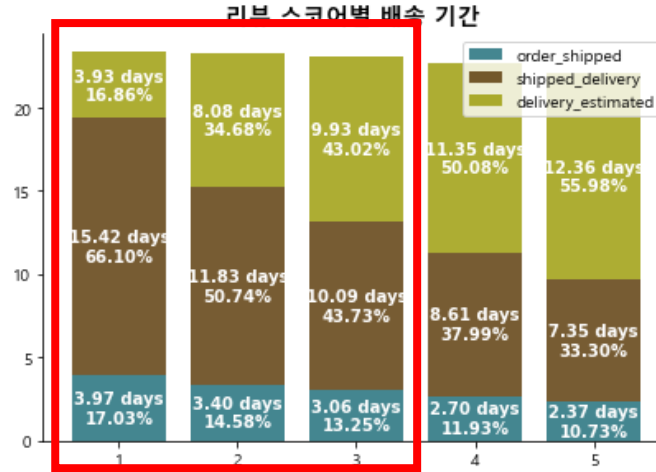
2) 리뷰 코멘트 분석 결과 관련 변수 생성

- Product 기준)
상품의 설명 글의 길이, 상품 사진 개수 등

- 배송 기준)
- 구매 요청 시점 기준 특정 소요 일수
(배송 완료, 상품 준비 등)

- 과거의 배송 내역 반영

➔ 구매 요청 시점 기준 직전 3개월의 배송 특징 변수



3. Modeling

LightGBM Model 사용

➤ 기본 입력 변수 시 ➔ recall : 0.01

➤ Feature Eng. 변수 추가 시 ➔ recall : 0.27

➤ 오버샘플링 + 하이퍼 파라미터 조정 시 ➔ recall : 0.48
(1 : 0.25 / 0 : 0.75 ➔ 1 : 0.5 / 0 : 0.5)

Base model

	precision	recall	f1-score	support
0	0.78	1.00	0.87	37212
1	0.50	0.01	0.01	10820

Feature Eng 변수 추가

	precision	recall	f1-score	support
0	0.82	0.97	0.89	37212
1	0.72	0.27	0.39	10820

오버샘플링 및 하이퍼 파라미터 조정

	precision	recall	f1-score	support
0	0.84	0.82	0.83	14896
1	0.43	0.48	0.45	4317

브라질 이커머스 플랫폼 고객 만족 분석

고객 만족 데이터 분석을 통한 사업 개선 방향 제시 및 만족도 점수 예측 모델 개발

4. Suggestion

➤ Logistics 기준

- 물류 관련 변수의 경향성

i) 제품 배송 기간과 고객 만족도 반비례

ii) 배송비 금액과 고객 만족도 반비례

➔ 제안 : 물류 파트너 및 인프라 확보를 통한

배송 기간 단축

➤ 판매자 care

- 판매자 관련 변수의 경향성

i) 판매자의 제품 준비 기간과 고객 만족도 반비례

ii) 판매자가 받았던 리뷰 코멘트 길이와 반비례

➔ 제안 : 판매자별 과거 판매 데이터 중 만족도에 영향을

미치는 지표를 추적 관리하여 서비스 개선이

필요한 셀러를 조기에 탐지 및 관리

5. Restrictions

➤ 판매자 분포

- 판매 건수가 1건인 셀러는 전체 3,095명 중 563명(18.2%)

- 데이터 부족으로 판매자 특징을 구분할 클러스터링 불가

➤ 데이터 구성

- 다양한 제품을 한 번에 구매 시에도 리뷰가 1건으로

집계되어 어떤 대상에 관한 리뷰인지 구분이 모호함

- 리뷰가 제품 단위로 기록되어 있다면 보다 정확한 분석이

가능할 것으로 보임

6. Improvement

➤ 새로운 모델 생성 및 앙상블

- 추가적인 Tree_Based Model을 생성 후 앙상블을 통한 예측 성능 개선 가능

➤ Feature Engineering

- 리뷰 스코어와 상관성이 높은 추가 변수를 확보 시 예측 성능 개선 가능

