E-Commerce 구매 및 리뷰 데이터를 활용한 리뷰평점 영향인자 도출

2019년 2월 25일

딥러닝 기반 핵심산업별 빅데이터 분석 전문가 과정
팀: 타이거JK와 아이들
팀원: 권민수, 김이영, 김종인, 문지현

목 차

1. 프로젝트 개요

1.1 프로젝트 주제 및 목표

프로젝트 주제

E-Commerce 의 구매 및 리뷰 텍스트 데이터를 활용한 리뷰 평점에 대한 영향인자 식별

분석 목적

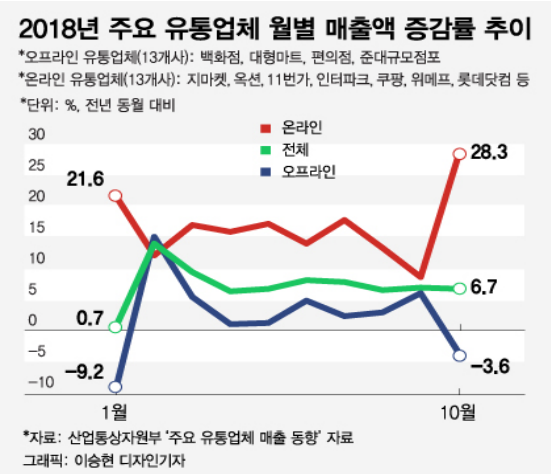
1. 본 프로젝트의 1 차적 목적은 E-Commerce 구매과정에서 축적된 데이터(판매자, 구매자, 상품, 배송 데이터 등의 정형데이터와 리뷰 텍스트 등 비정형 데이터)를 활용하여 리뷰평점에 긍정적/부정적 영향을 미치는 요인을 식별한다.
2. 리뷰 텍스트 데이터 분석을 통해 정형 데이터로는 알 수 없는 새로운 이슈를 인식한다.
3. 식별된 요인에 대한 개선방안을 도출한다.
   - 비즈니스 질문 : "어떻게 리뷰평점을 개선할 수 있는가?"
   - 더 나아가, 본 프로젝트는 식별된 요인을 개선하기 위한 방안을 도출함으로써 E-Commerce 사업자에게 사업 운영에 관한 시사점을 제공한다.
   - 효율/효과적인 분석을 위해 향후 개선할 점 제안
   - 리뷰 작성 시 추가적으로 수집하면 좋은 데이터 제안
   - 데이터 수집/관리 시 개선할 점 제안

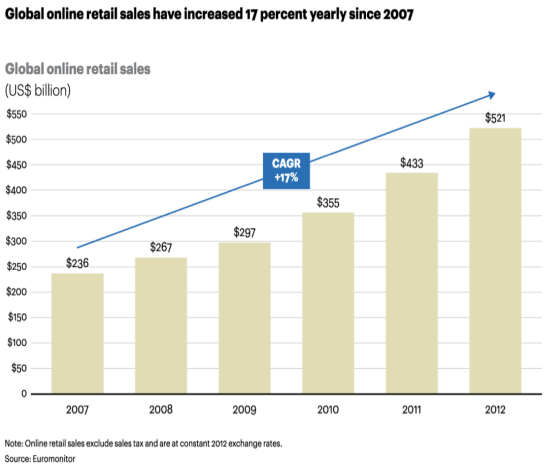
1.2 기획 배경

분석의 주제를 E-Commerce 에서의 리뷰평점 개선으로 정한 이유는

(1) E-Commerce 의 성장과 (2) 리뷰평점이 E-Commerce 에서 미치는 영향력 때문이다.

(1) E-Commerce 시장의 성장





[그림 1] 2018 년 주요 유통업체 월별 매출액 증감률 추이와 [그림 2] 글로벌 온라인 쇼핑 매출 성장률

E-Commerce 시장은 급속도로 성장하고 있다. 국내 기준 2015 년 10 월에 4 조 8222 억원이었던 온라인 쇼핑 월 거래액은 2018 년 10 월 처음으로 10 조원을 돌파했다. 불과 3 년 만에 2 배 이상 늘어난 것으로 연평균 성장률을 계산했을 때는 30.1%수준이다. [그림 1]
또한, 국내 뿐만 아니라 세계적으로도 온라인 쇼핑 시장은 연평균 17%로 성장하고 있다. [그림 2]

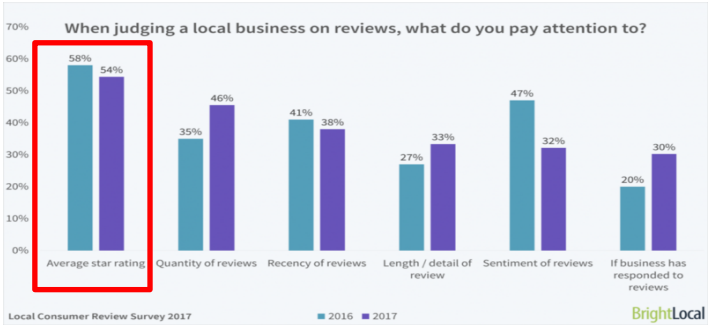
E-Commerce 시장의 성장배경에는 여러가지가 있지만 대표적인 것은 다음과 같다.

- 스마트폰 보급에 따른 모바일 쇼핑 증가
- 당일 배송 가능
- 모바일에 익숙해진 경제력과 시간적 여유를 가진 실버 세대들이 온라인 쇼핑의 새로운 '큰손'으로 등장
- 쉬운 가격 비교
- 언택트(비접촉 쇼핑 심화)
- 로열티 서비스(네이버페이, 포인트, 쿠폰 등)
- 1 인 가구 증가
- 시간 절약
- 24 시간 접근성

(2) 리뷰평점이 E-Commerce 소비자들의 구매 의사결정에 미치는 영향력

E-Commerce 시장이 빠른 속도로 성장하고 전체 유통시장에서 차지하는 비중이 커지는 상황에서 E-Commerce 시장에서 소비자들의 구매 의사결정에 영향을 미치는 요인을 찾아내는 것은 중요하다.

과거 오프라인 유통시장에서는 입소문이나 브랜드 인지도 등이 구매 의사결정 단계에서 큰 영향력을 가졌다. 이에 따라, 유통업체들은 구전 효과를 높이고 브랜드 인지도를 높이기 위해



대규모 마케팅 비용을 들였다. 하지만 E-Commerce 시장에서는 리뷰 평점이 구매 의사결정에 큰 영향을 끼치고 있다. 잠재 소비자들이 사용해보지 않는 브랜드나 제품에 대해 느끼는 불확실성을 기존의 소비자들이 작성한 리뷰 평점을 통해 해소하고 있는 것이다.

[그림 3] 리뷰 평점의 중요성 : 리뷰 정보 중에서도 소비자들이 가장 주목하는 부분

본 프로젝트가 리뷰 평점에 초점을 맞춘 이유는 리뷰 정보 중에서도 리뷰 평점이 가장 중요한 부분을 차지하고 있기 때문이다. Bright Local 에서 시행한 설문 [그림 3]에 따르면, 리뷰

정보 중에서도 소비자들이 가장 주목하는 부분은 리뷰 평점이었다. 실제 하버드에서 미국 Yelp 라는 음식점 리뷰 어플리케이션을 대상으로 연구를 한 결과, 리뷰 평점이 1 점 높을 때 평균적으로 수익이 9% 높은 것으로 나타났다.

리뷰 평점이 E-Commerce 소비자의 구매 의사결정과 매출에 영향을 미치므로 E-Commerce 는 리뷰 평점에 영향은 주는 요인을 파악하고 개선하여 리뷰 평점을 특정 수준 이상으로 관리하는 것이 필요하다.

1.3 Olist 기업 소개 및 데이터 선정 이유

(1) Olist 기업 소개(비즈니스 모델)

Olist 는 2015 년에 설립된 최근 브라질에서 가장 빠르게 성장하고 있는 SaaS 기반 기업이다. 온라인 상점 유무와 관계없이 모든 규모의 소매업체를 대상으로 온라인 시장 판매 증가에 대한 솔루션을 제공하는 업체이다. 소매업체들은 Olist 와의 단일 계약으로 아마존과 같은 온라인 채널에서 제품을 판매할 수 있다.

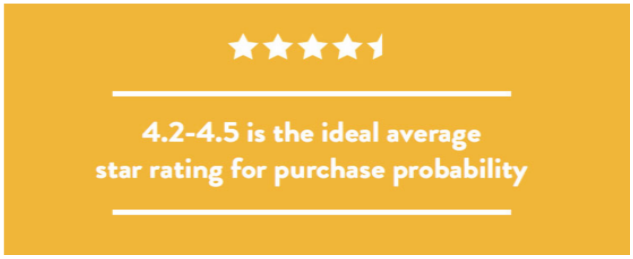
배송 시스템은 Olist 측에서 브라질의 우체국과 계약을 맺고, 판매자가 우체국을 통해 직접 배송하는 시스템으로 배송 관련 문제들의 책임은 대부분 판매자에게 있다.

(2) Olist 데이터셋 선정 이유

2.1. 무엇보다 분석 주제의 초점인 리뷰 평점(5 점 척도) 데이터를 포함하고 있었다.

2.3. 또한, 텍스트 리뷰 (비정형 데이터)를 포함하고 있어 리뷰 평점에 미치는 요인을 다방면으로 분석해 볼 수 있을 것이다.

- Olist 의 리뷰 평점 평균은 3.9 이므로 리뷰 평점 개선이 필요하다.[그림 4]



[그림 4]연구 발표에 따르면 4.2 ~ 4.5 사이의 평점이 가장 이상적인 평점)

2.3. 비정형 데이터에서, Olist 측에서 전문가 3 명이 댓글을 9 개의 이슈로 분류해서 투표를하여 새로운 이슈를 도출하고자 한 노력의 흔적이 보여서 활용 요소가 될 수 있을거라고 판단했다.

1.4 프로젝트 추진 일정 및 구성원 역할

(1) 프로젝트 추진 일정

NO	STEPS	TASK	SUB_TASK	OUTPUT	OWNER
Nov.3-Dec.8	분석기획	1. 비즈니스 이해 및 범위 설정	1.1 Olist 비즈니스 모델 이해	Olist 비즈니스 설명(WORD)	권민수
		2. 프로젝트 정의 및 계획 수립	2.1 분석과제 발굴(하향식 접근 방식)		-
			2.2 분석과제 정의	분석과제정의서(WORD)	김종인
			2.3 과제 수행 일정 계획 수립	과제수행일정계획(EXCEL)	김종인
Dec.8-Dec.22	DATA 정의	1. 데이터 이해 및 필요 데이터 정의	1.1 데이터 이해 (dataset별 데이터/테이블정의서 작성)	데이터/테이블정의서(EXCEL): 8개(+1개)	문지현
			1.2 가설수립(1차)	가설정의서_1차(EXCEL)	김종인 / 권민수
			1.3 필요데이터 선택 및 정의(1차)	필요데이터정의서_1차(EXCEL)	김종인 / 권민수
		2. 포르투갈어 리뷰데이터 처리 방안 논의	2.1 각자 구글링 후 처리 방안 논의 후 대안 선	리뷰데이터 처리방안(PPT)	김이영 / 김종인
Dec.22	DATA 수집	1. 원천 데이터 수집	1.1 데이터 수집(데이터소스에서 데이터 다운로	데이터 SET(CSV): 8개	김이영 / 권민수
Dec.22-Jan.5	DATA 전처리	1. 분석데이터 구성 및 데이터 품질확인	1.1 분석 데이터 구성방안 수립	분석 데이터 구성도(EXCEL or PPT)	김이영 / 권민수
			1.2 분석 데이터 구성(1개 테이블 구성)	분석 데이터 SET(CSV or EXCEL): 1개	김이영 /권민수
			1.3 데이터 품질확인(기초통계 산출) (잡음, 이상치, 결측치 식별 후 처리)	데이터 품질 보고서(EXCEL)	김종인
Jan.5-Jan.19	DATA 탐색	1. dataset 이해(EDA)	1.1 데이터 기술 분석 (= 데이터 이해 = 새로운 인사이트 발견)	데이터 탐색 보고서(EXCEL, PPT) (기술통계 후 인사이트 발굴, 분석방향 결정)	문지현 / 권민수
			1.2 가설수립(2차)	가설정의서_2차(EXCEL)	팀
		2. 변수생성, 선택, 차원 축소 / 분포, 단위변환	2.1 파생 변수 생성	파생변수 정의서(EXCEL)	팀
			2.2 데이터 분포변환, 단위변환	파생변수 정의서(EXCEL)	팀
		3. 학습용 데이터와 검증용 데이터 분리	3.1 데이터 분할	TRAIN DATASET, VALIDATION DATASET	팀
Jan.19-Feb.16	DATA 분석	1. 1차 분석(필수) (리뷰데이터 불포함)	1.1 주요 영향인자 분석	분석결과보고서(PPT)	팀
			1.2 예측모델링		
			1.3 모형 평가 및 검증		팀
		2. 2차 분석(OPTIONAL) (리뷰데이터 포함)	2.1 주요 영향인자 분석	분석결과보고서(PPT)	
			2.2 예측모델링		
			2.3 모형 평가 및 검증		팀
Feb.16-Feb.23	결과 정리	1. 분석결과 활용방안 수립 (VOC 개선방안에 대한 시사점)	1. 분석결과 활용방안 수립	분석결과보고서(PPT)	팀
		2. 분석 제약사항	2. 분석 제약사항		팀
		3. 향후계획	3. 향후계획		팀
Feb.15-Feb25	리포팅	1. 보고서 작성 및 보고	1. 보고서 작성	프로젝트 최종보고서(PPT)	팀
			2. 결과보고	프로젝트 최종보고서(PPT)	팀

(2) 구성원 역할

이름	학교(전공)	역할	구현 부분
권민수	Indiana (MIS)	팀원	Api를 통한 텍스트 번역 결과 해석 및 비즈니스 제안 & 분석
김이영	University of Bristol (Engineering Mathematics)	팀원	데이터 조작, 모형 모델링 & 분석
김종인	중앙대학교 (경영학과, (부) 컴퓨터공학과)	팀장	데이터 조작, 모형 모델링 & 분석
문지현	홍익대학교(경영학과)	팀원	기초통계와 결과 해석 및 비즈니스 제안 & 분석

2. 프로젝트 현황

2.1 시장 조사 [5, 6]

(1) 영화 평점

- 극장을 찾는 관객들은 **영화의 선택에 실패하지 않기 위해, 입소문이든 네티즌들의 평점이든 높은 점수에 따라가고 있다**. 영화 관련 평가를 조사한 결과, 인터넷 포털 평점을 신뢰한다는 응답이 29.7%로 1 위를 차지한다. 영화 관련 커뮤니티 평가(19.1%), 지인의 평가(18.2%), 영화 평론가 평점(15.7%)보다 높은 수치이다.
- 개봉 전 보고 싶었던 작품이 시사회 후 평가가 좋지 않거나 네티즌들의 댓글이 부정적인 뉘앙스를 띠면, 관객들은 의식적이든 무의식적이든 그들의 평점에 큰 영향을 받아 평가 절하하는 경향이 짙다.
- 네티즌들의 평가를 100% 신뢰할 수 없다는 것을 알고 있으면서도 평점과 후기를 따라 선택하려는 마음이 크다 평점이 높을수록 사람들은 영화에 대한 기대감을 갖고 극장으로 향하는 경우가 많다.
- 사람들이 매긴 평점을 무조건 따라가는 군집 행동을 보였다.

(2) YELP

- 2003 년부터 2009 년까지 시애틀 지역에 존재했던 3582 개의 음식점을 전수 조사했을 때, YELP 의 음식점 평점데이터와 미국 국세청의 음식점 수익 데이터를 음식점 이름과 주소를 기준으로 결합하여 분석한 결과 YELP 에서의 평점이 1 점(one-star) 오르면 수익이 5~9% 상승한다는 사실을 발견했다.
- 단 하나의 부정적인 리뷰로 인해 소비자의 22%가 레스토랑을 찾지 않을 수도 있다.
- 리뷰 숫자가 많아지고 Certified Reviewer 가 호의적으로 평가할수록 평점과 수익은 더 강한 상관관계를 보여준다.
- 식당에서는 평점 관리를 하기 위해서 보다 나은 서비스를 제공하고 안 좋은 피드백이 있으면 형식적인 댓글이 아닌 진심이 담긴 댓글을 달아 전화위복의 기회로 삼기도 한다.

2.2 유사 분석 결과 장단점 분석

영화리뷰 감성 분석을 통한 평점 예측 연구

해당 연구는 영화 리뷰 텍스트 데이터를 활용하여 영화 평점을 예측하는 연구이다. 리뷰 텍스트 데이터를 통해 영화 평점을 예측하는 것에 초점을 맞춘 연구로, 평점에 영향을 미치는 요인을 찾으려고 하는 본 프로젝트의 방향성과 다르다. 또한, 감성 분석은 단순히 긍정/중립/부정으로 분류하지만, 본 프로젝트에서는 리뷰 텍스트를 부정을 주제 별로 세분화했다는 점에서 차별점이 있다.

2.3 차별화 핵심 전략 기술

[단순 단어 단위가 아니라, '구' 단위의 딕셔너리로 텍스트 분석 시행]

- 대부분 온라인 쇼핑몰들이 평점이 낮거나 안 좋은 리뷰가 남겨져 있어도 낮은 평점과 안좋은 리뷰의 요인을 분석하여 대응 하기보다는 비용이 많이 드는 마케팅을 강화할 뿐 리뷰와 평점의 중요성을 낮게 평가하는 경향이 있다.
- 이미 온라인 쇼핑몰은 포화 상태이며 새로운 고객을 유치하는 것도 중요하겠지만 기존의 고객들을 지키는 것도 그에 걸맞지 않게 중요하다.

- 리뷰와 평점 관리는 수익 증대의 지름길이고 다른 기업들에서 시도하지 않은 새로운 것을 시도 해보려고 한다.
- 연구결과에 의하면 안 좋은 리뷰에 판매자가 문제를 해결을 할 의지를 보이거나 해결을 했을 경우 안 좋은 리뷰와 평점을 남긴 사용자가 리뷰를 지우거나 평점을 상향 시키는 경향을 보였다.

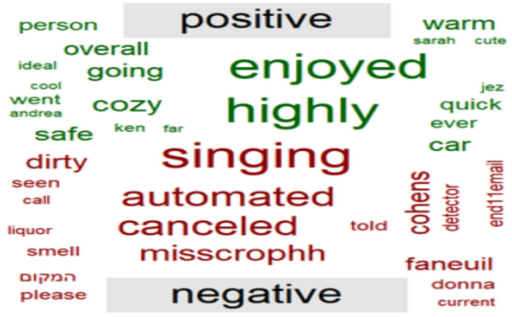
(1) LDA 토픽 모델링

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09
0	love	amazon	home	music	like	one	echo	speaker	great
1	alexa	app	smart	alexa	alexa	bought	dot	sound	works
2	gift	device	lights	weather	dont	got	love	good	product
3	loves	alexa	music	play	doesnt	im	room	bluetooth	easy

[그림 5] 일반적인 LDA 토픽 모델링

다양한 문서에서 추상적인 주제를 발견해내는 통계학적 모델링 기법의 하나.Latent Dirichlet Allocation 의 준말인 LDA는 특정 주제에 맞게 텍스트를 분류하는데 쓰이는 하나의 모델이며, 문서 모델 당 하나의 주제를, 주제 모델 당 '단어'들을 생성.

(2) 감성 분석 (Sentimental Analysis)



[그림 6] 일반적인 감성 분석 결과

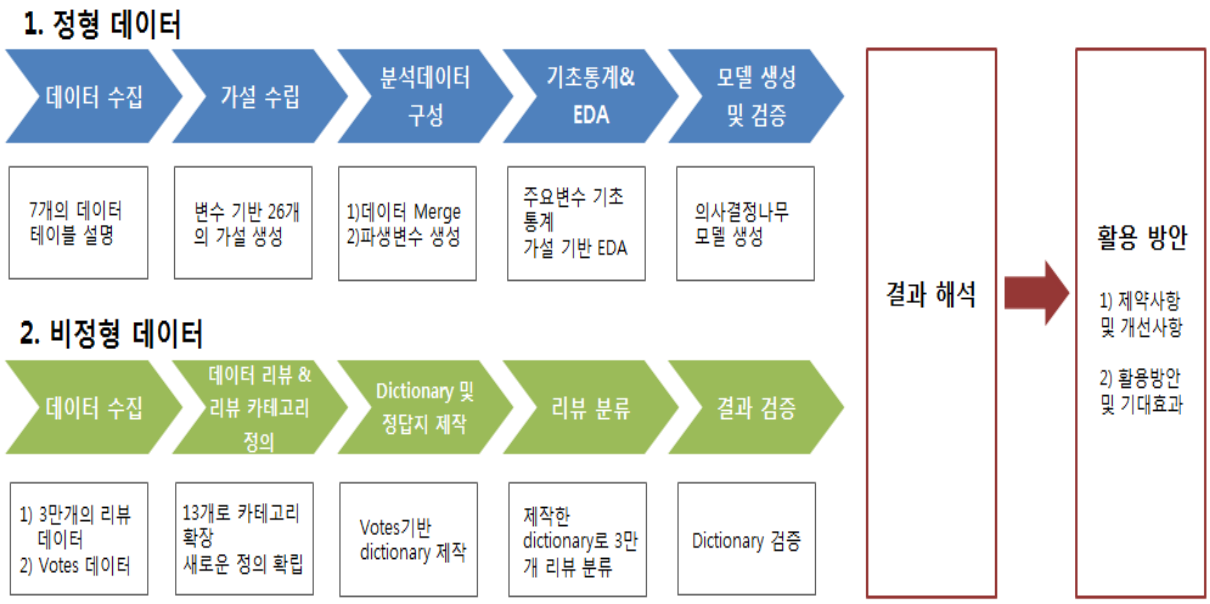
감성 분석(Sentiment Analysis)은 '오피니언 마이닝(Opinion Mining)'으로 도 불리는데, 이는 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적 인 데이터를 분석하는 자연어 처리 기술. 예를 들면 온라인 리뷰에 사용된 단어들을 추출하여 리뷰를 요약하고 온라인 리 뷰에 사용된 단어들의 특징을 이용하여 제품에 대한 소비자의 긍정적 또는 부정적 의도를 예측 하거나, 온라인 리뷰 전체를 한정된 시간에 검토 하기 어려운 점에 착안하여 제품에 대한 구매자 의 태도를 예측하고자 하는 것 등이 진행되었다.

긍정적인 리뷰가 부정적이고 부정적인 리뷰가 긍정적이게 나올 수 있는 치명적인 단점이 있었다.

➔ 감정 분석과 LDA 모델링으로 추출된 단어만으로는 구매 경험의 긍/부정을 파악할 수는 있었지만 실질적인 불만 요인을 파악하는데 한계가 있었음. 그에 반해 우리의 분석은 수천개의 실제 리뷰를 근거로 해서 **문장과 구절로 이루어진 Dictionary 를 만들어서 주제 별로 분류**를 하였다. 그로 인해 보다 정확한 구매자들의 불만 요인을 파악할 수 있게 됐다.

3. 프로젝트 분석 결과

3.1 분석방안 정의

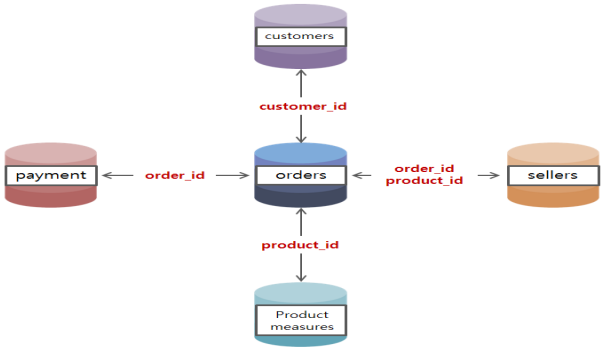


3.2. 정형데이터 분석

데이터 수집 및 변수 설명 [출처]

- 데이터 분석 대회 플랫폼 Kaggle 에서 공개한 Public Dataset
- 7 개의 데이터 파일, 총 100,000 개의 거래 내역 데이터

(1) 데이터 병합 스키마



데이터 긍정적 부정적 평점의 분포 비율 :

Negative : Positive = 0.23 : 0.78

긍정적 평점이 매우 비중이 큰 불균형한 분포를 보여서 비율을 고려한 분할을 실시하였다.

(2) 원본 데이터 변수 설명서

데이터 변수 그룹	테이블 설명	칼럼 속성
orders	주문과 관련된 정보	주문 아이디, 배송상태, 총 제품 가격, 배송료, 주문 내 제품수, 주문 내 판매자 수, 구매 시간, 구매승인 시간, 배송예정 날짜, 고객 물품수령 날짜
customers	고객정보를 관리한다	고객 아이디, 고객 시 주소, 고객 주, 고객 집코드, 고객 unique id,
payments	결제와 관련된 정보	할부 개월, 지불 방법
products	제품과 관련된 정보	제품 아이디, 제품 품목 이름, 제품명 길이, 제품 설명 길이, 제품 사진 개수, 제품 무게, 제품 길이, 제품 높이, 제품 넓이
sellers	판매자와 관련된 정보	판매자 아이디, 판매자 집코드, 판매자 도시, 판매자 주
geolocation	(zipcode기준)지리적 정보	우편번호 앞 3자리, 우편번호의 도시, 우편번호의 주, 우편번호의 위도, 우편번호의 경도
review	리뷰와 관련된 정보	리뷰 아이디, 리뷰 점수, 리뷰 제목, 리뷰 내용, 리뷰 요청 날짜, 리뷰 작성 날짜

[데이터의 한계점]

1. 고객정보가 없는 한정된 데이터셋

- 공개된 데이터를 활용했다보니, E-commerce 거래 내역에서 활용성이 높은 고객에 대한 정보가 부재 돼 있다. 고객에 대한 성별 및 연령 등 없고 지리적 위치 정보만 존재한다. (도시, 주, 우편번호 앞 세자리)
- 따라서, 결과가 배송에 치우쳐졌고 다소 진부한 결과가 도출되었다.

2. 문화적 차이

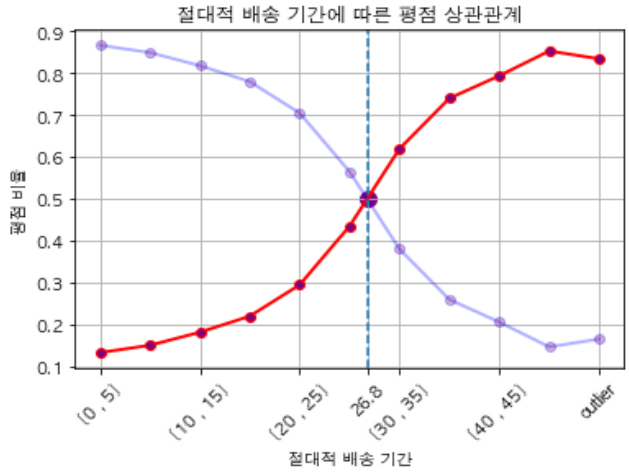
- 한국과 다르게 평균 배송 기한이 이해하기 힘들정도로 매우 길다.
- 구매 수단도 3 가지 다른 카드로 모두 다르게 할부 개월 수를 선택하는 등 한국과 다른 구매 수단 및 방법이 이해하기 힘들다.
- 브라질의 데이터이므로 댓글이 포르투갈어로 작성 돼 있다.
- 브라질의 데이터이므로 불가피하게 지리적, 문화적 이해도가 낮다.

[한계점 해결 방안]

1. 한계된 정보 안에서 유의미한 다른 요인을 도출해 내기 위해서 텍스트 분석을 시행하였다.
2. 포르투갈어를 구글 번역 API를 활용하여 영문으로 변환하여 분석 진행하였다.

(3) 탐색적 분석

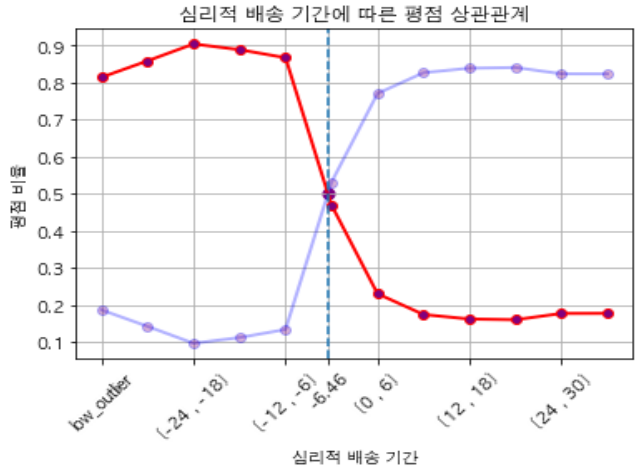
가설 1-1 : 절대적 배송기간이 길수록 평점은 낮다



절대적 배송 기간 의미
= '배송 소요 기간'

길어질수록 평점이 부정적인 비율이 높아진다.
가설 수립

가설 1-2 : 심리적 배송기간이 길수록 평점은 낮다

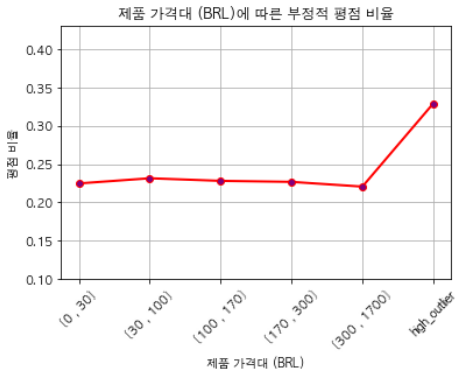


심리적 배송 기간 의미
= '예상 배송일 대비 배송 기간'
음수일수록, 예상 배송일보다 **늦음**
양수일수록, 예상 배송일보다 **빠름**

예상배송일보다 늦게 왔을 수록 부정적 평점의 비율이 높아진다.

가설 수립

가설 1-3 : 지불한 가격이 클수록 민감하게 반응한다.



지불 가격대 (BRL) 가격대
KRW 환원하면,
약 [만원 - 3 만원 - 5 만원 – 10 만원 – 50 만원 – 50 만원 이상] 으로 구간화하여 살펴본 결과,
50 만원 이상의 VIP 고객들이 더 부정적 비율이 높다.
가설 수립

*** 기타 통계들은 Appendix 참고**

(4) 모델 생성 및 검증

[모형 1]. 의사 결정 나무에 들어갈 변수 선정 : 로지스틱 회귀

데이터셋에 존재했던 변수들과 새롭게 생성한 여러가지 파생변수들 중 EDA에 유의미 했던 변수들, 유의미한 변수들을 투입하여 선형 분류 모델로 유의 변수들을 선정하였다.

- Train – 검증 : 0.7 :0.3 의 비율로 분할했고, 긍정 부정의 비율을 맞춰서 분할함
- 로지스틱회귀분석 실시(일반화 선형모형: Generalized Linear Model)
- 패키지와 함수 : glm 함수

1.1. 로지스틱 회귀 분석 결과 :

주제	변수명	한글 변수명	Estimate	Pr(>|z|)
배송	del_period_deadline_ynTrue	배송 기한 준수 여부 (약속 이행)	1.303(+)	0.00
배송	del_period	절대적 배송 기간	-0.048(-)	0.00
거리	distance	판매자 - 고객 직선 거리(km)	0.000(-)	0.00
비용	order_product_value	제품 가격대	0.000(-)	0.00
비용	freight_value_proportion	배송료 비중	-0.150(-)	0.00
배송	del_period_psy	심리적 배송 기간	0.004(+)	0.02
주문시기	order_day.L	주문 요일	-0.059(-)	0.03
비용	installments_ynTrue	할부 여부 (True)	-0.060(-)	0.04

[표1] 로지스틱 회귀 모형의 결과로 유의한 변수로 선정된 8개 변수들

TRAIN	NEGATIVE	POSITIVE		TEST	NEGATIVE	POSITIVE	
0	3085	1292	4377	0	1309	579	1888
1	10191	47900	58091	1	4379	20476	24855
SUM	13276	49192	62668	**SUM**	5688	21055	26743
정확도			81.62%	**정확도**			81.46%

[표2] 훈련, 테스트 데이터의 혼동표, 정확도

1.2. 변수 선택을 위한 결과 해석 :

모델의 정확도는 훈련, 테스트 모두 약 81.62%가 측정되었고, 총 8개의 변수들이 p-value 값이 0.05보다 작아서 유의미한 변수로 선정되었다.

(1) 첫번째로, 배송과 관련된 변수들로는 '배송 기한 준수 여부', '절대적 배송기간', '심리적 배송 기간' 3 가지가 있다.

배송 기한은 약속을 이행 했때 Estimate 가 1.303(+)으로 평점이 긍정적으로 매겨진다.

절대적 배송 기간은 길어질수록 평점이 부정적인 방향으로 간다는 것을 알 수 있다.

심리적 배송 기간은 양수일수록 예상 배송일보다 빨리 배송이 온 것이므로, Estimate 가 양수로 긍정적인 반응으로 간다는 것을 알 수 있다.

(2) 두번째로, 금액과 관련된 변수들로 '제품 가격대', '배송료 비중', '할부 여부' 3 가지가 도출됐다.

배송료 비중은 커질수록 부정적인 반응으로 흘러가고, 할부로 결제한 경우에 더 부정적인 반응이 나온것으로 측정되었다.

(1) 그외로 '거리'와 '주문시기' 변수가 유의하게 선정되었다.

즉, 고객 평점의 긍정적/부정적 반응은 배송 시간과 제품 및 운송 비용에 민감하게 반응한다는 것을 알게됐다.

[모형 2]. 모형선정 : 의사 결정 나무 → 요인 파악 & 변수 조합 보기

종속변수 : 리뷰평점 긍/부 (12,3 : 부정, 4,5 : 긍정)

독립변수 [로지스틱 회귀 분석으로 선택된 변수들]

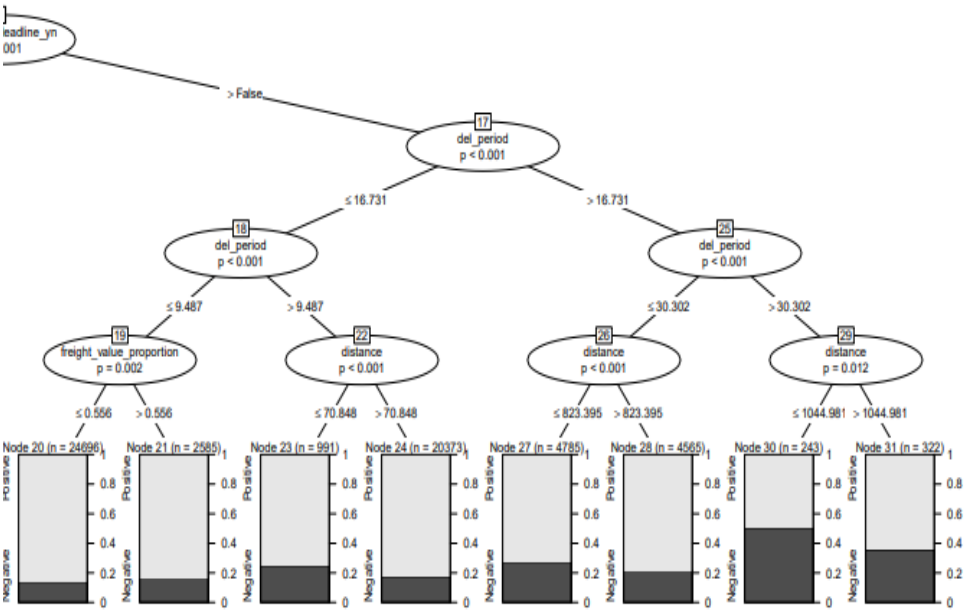
: 배송 기한 준수 여부, 절대적 배송기간, 심리적 배송기간, 제품 가격대, 배송료의 비중, 할부 여부, (판매자와 구매자 간) 거리, 주문 요일

모델 성능 평가

```
Confusion Matrix and Statistics

          Reference
Prediction Positive Negative
  Positive    21072     4599
  Negative      445     1783

               Accuracy : 0.8192
                 95% CI : (0.8146, 0.8237)
    No Information Rate : 0.7712
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3355
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9793
            Specificity : 0.2794
         Pos Pred Value : 0.8208
         Neg Pred Value : 0.8003
             Prevalence : 0.7712
         Detection Rate : 0.7553
   Detection Prevalence : 0.9201
      Balanced Accuracy : 0.6293

       'Positive' Class : Positive
```

결과 해석

[오른쪽 분기 나무]

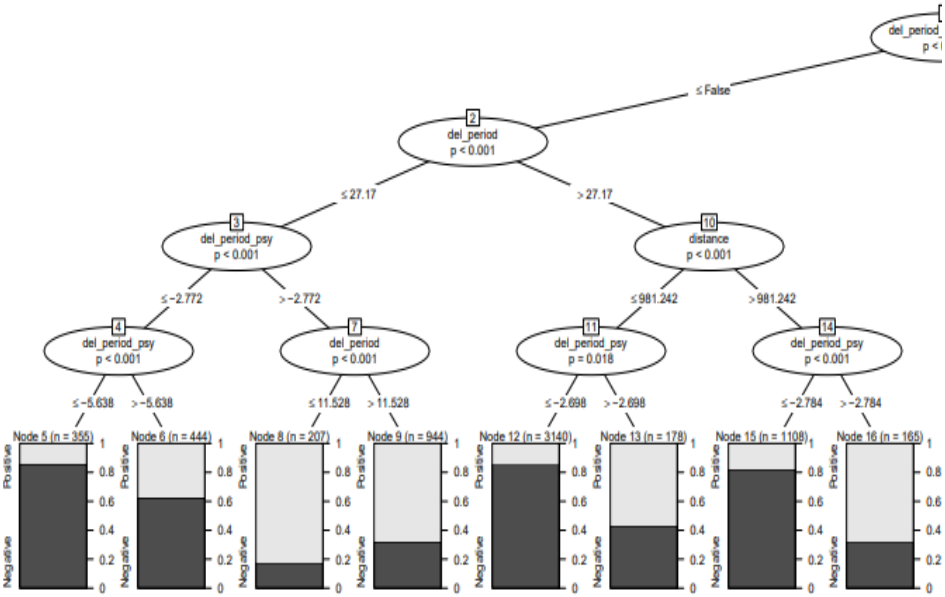


첫번째 분기 = 배송 이행 여부 -→ (배송시간 불이행 – 배송시간 이행)

[예상 배송일 시간을 이행한 경우 = 오른쪽 나무]

두번째 분기인 절대적 배송기간이 길어지면 더 부정적인 비중이 커지는 것을 볼 수 있다. 그리고 마지막으로 분기되는 것이 거의 대부분 '판매자와 고객간의 거리' 변수로 인해 분기되는데, 모두 판매자와 고객간의 거리가 짧을 때, 길때 보다 더 부정적인 비율이 높다. 이는 거리가 짧을 수록 배송이 더 빨리 올 것이라는 기대 심리가 작용 했을 것이라고 가정할 수 있다. 예를 들어, 고객이 북동부나 북부 지역에 거주 중 이라면, olist의 판매자의 대부분 상파울루 주에 거주 중이므로, 심리적으로 더 배송 기간에 기대치가 낮아지고 민감하지 않게 반응할 수 있다.

[왼쪽 분기 나무]



첫번째 분기 = 배송 이행 여부 → (배송시간 불이행 – 배송시간 이행)

[예상 배송일 시간을 이행하지 못한 경우 = 왼쪽 나무]

왼쪽의 가지 또한 '절대적 배송 기간', 27일 기준으로 분기되었다. 그 다음 분기들은 오른쪽과 다르게 대부분 '심리적 배송 기간'을 기준으로 분기 되었다. 심리적 배송기간은 음수일수록 예상 배송일보다 늦게 온 것이므로 모두 심리적 배송기간이 더 작아질수록 부정의 비율이 높은 것을 볼 수 있다. 3 번 분기점을 보면, 배송 기한이 지켜지지 않았음에도 불구하고 오른쪽으로 나뉘어진 예상 배송일보다 2.7일 정도 늦은 경우는 그렇지 않은 경우들보다 부정 평점의 비율이 현저하게 낮다. 이를 통해 배송 예상일에 따른 사람들의 민감도가 매우 크다는 것을 알 수 있다.

[결과 요약]

결과에서 볼 수 있는 영향인자를 정리하면 아래의 [표] 와 같다.

[표] 리뷰평점 긍/부에 영향을 주는 주요 변수와 영향의 방향성

변수	영향
배송 기한 준수여부	배송 기한 준수하지 않을 때 부정적 평점
절대적 배송기간	배송기간이 길수록 부정적 평점
(판매자-구매자 간) 거리	거리가 짧을수록 부정적 평점
제품 가격대	제품가격이 높을수록 부정적 평점

한정된 변수를 활용 했지만, 구매 평점에 영향을 끼치는 주요 요인로 크게 배송, 비용, 배송에 영향을 끼치는 운송 거리, 주문 시기가 유의하게 도출되었다.

의사결정 나무 모델의 결과로는, 배송과 관련된 것 중에서도 심리적인 배송 기간, 판매자와 고객 간의 거리가 중요하게 도출되었다. 이는 소비자가 배송 시기에 갖는 기대치와 연관이 높고 구매 평점에 민감하게 반응한다는 것을 알 수 있다. 현재 Olist는 거리를 기반으로한 정밀한 예상 배송일 측정이 이루어지지 않고 판매자가 배송일을 지키지 않았을 때에 대한 엄격한 관리가 이루지지 않고 있음을 가정할 수 있다.

따라서,

[1] 예상 배송일을 판매자와 고객간의 거리기반으로 측정할 필요성이 있다.

[2] 판매자의 예상 배송일 이행 여부에 대한 엄격한 관리 조치가 필요하다.

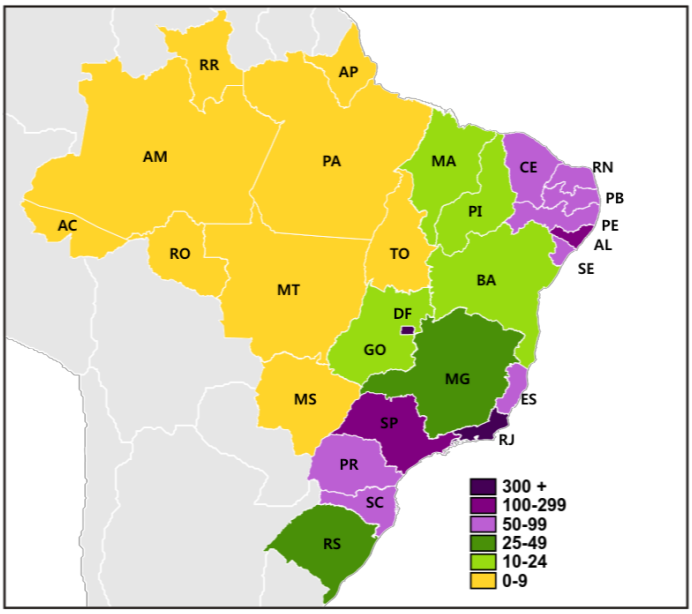
3.3. 정형 데이터 분석 결과 기반으로 비즈니스적 해결 방안 제시

[1] 거리 기반 배송 기간 제안

모델의 결과, 배송에 관련된 변수들이 평점에 영향을 끼치는 것으로 도출됐고, 배송 기간의 지연 원인을 도출해보고자 한다.

1. 브라질 지형과 주의 특징

브라질의 국토면적은 약 8,56,000km² (29%는 농지, 66%는 산림지)로 러시아, 캐나다, 중국, 미국에 이어 세계에서 5위로 큰 면적을 가지고 있다. 지역적 차이에 따라 크게 북부 지역(45%), 북동부지역(18%), 남동부지역(11%), 남부지역(7%), 중서부지역(19%)의 5개 지역으로 분류되고 있으며, 극남과 극북의 직선거리는 약 4,400km로 전 세계를 통틀어 남북으로 가장 긴 나라이다. [1]



순위	약어	인구	비율
1	SP	44,035,304	21.7%
2	MG	20,734,097	10.2%
3	RJ	16,461,173	8.1%
4	BA	15,126,371	7.5%
5	RS	11,207,274	5.5%
6	PR	11,081,692	5.5%
7	PE	9,277,727	4.6%
8	CE	8,842,791	4.4%
9	PA	8,073,924	4.0%
10	MA	6,850,884	3.4%

[그림 1] 브라질 각 주의 인구 밀도[2] [표 1] 2014 년 기준 인구가 많은 상위 10 개 States [4]

1.1. 북부 지역 (면적 45%, 인구 7%) – AC, AP, AM, PA, RO, RR, TO

전 국토 면적의 45%, 총인구의 7%를 차지하며 Acre(AC), Amapá(AP), Amazonas(AM), Pará(PA), Rondônia(RO), Roraima(RR), Tocantins(TO) 등으로 구성되어 있다. 아마존 강이 흐르는 열대 우림 지역으로, [그림 1]에 따르면, 북부의 인구 밀도는 모두 0-9로 매우 낮은 인구 밀도를 가지고 있다.

1.2. 북동부 지역 (면적 18%, 인구 29%) – MA, PI, CE, RN, PE, AL, SE, PB, BA

전 국토 면적의 18%, 총인구의 29%를 차지하며 Maranhão(MA), Piaúi(PI), Ceará(CE), Rio Grande do Norte(RN), Pernambuco(PE), Alagoas(AL), Sergipe(SE), Paraíba(PB), Bahia(BA)주 등으로 구성되어 있다. 해안선을 따라 전개된 비옥한 토양지대와 반사막 평원의 내륙지역으로 구성되어 있다. [3]

1.3. 남동부지역 (면적 11%, 인구 43%) – ES, MG, RJ, SP

전 국토 면적의 11%, 총인구의 42.5%를 차지하며 Espírito Santo(ES), Minas Gerais(MG), Rio de Janeiro(RJ), São Paulo(SP) 주 등으로 구성되어 있다.

상파울루주의 수도인 상파울루는 인구가 약 1,900만명인 남미 최대 도시로, 각종 현대적 공업지대가 밀집되어 있으며 국내 제조업 총생산의 약 55%를 차지한다. 특히, **상파울루시는 브라질 최대의 소비도시이자 공업중심지**이다.

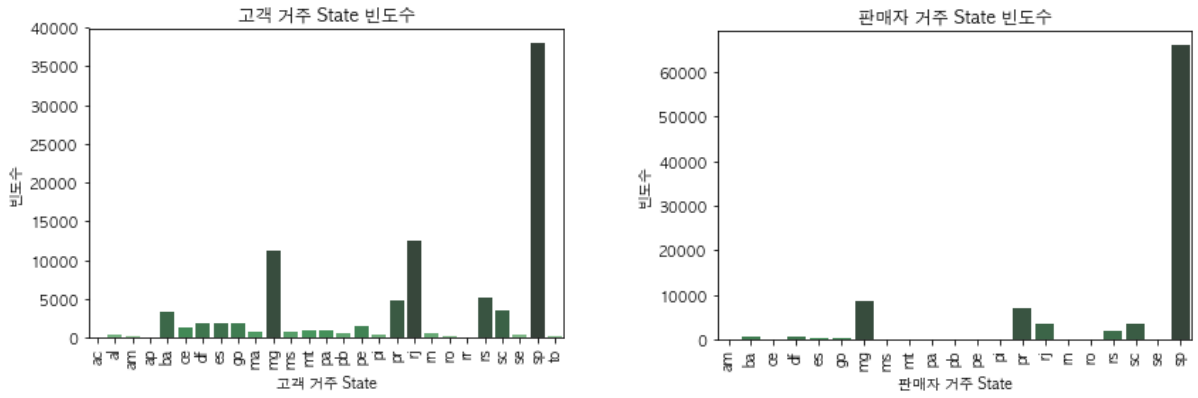
1.4. 남부지역 (면적 7%, 인구 15%) – PR, SC, RS

전 국토 면적의 7%, 총인구의 15%를 차지하며, Paraná(PR), Santa Catarina(SC), Rio Grande do Sul(RS)주 등으로 구성되어 있다.

1.5. 중서부지역 (면적 19% 인구 7%) – GO, MS, MT, DF

전 국토 면적의 19%, 총인구 7%를 차지하며 Goiás(GO), Mato Grosso do Sul(MS), Mato Grosso(MT), 연방특별구(Distrito Federal; DF)로 구성되어 있다. 고온다습한 열대성 기후지대로 축산업, 농업, 광업 등이 발달하였으며, **수도 브라질리아**가 있는 중서부 지역은 중앙고원을 비롯하여 완만한 기복이 있는 고지대로 대부분 미개발 상태이다. 또한, 브라질리아의 인구는 2017년 기준 약 303만명이며 연방특별구(Distrito Federal; DF)에 위치해 있다.

2. Olist 데이터에서 판매자와 고객의 거주 지역 분포



[그림2] 총 거래 내역 중 고객의 거주 State 분포와 판매자의 거주 State 분포

[표1]의 인구와 대부분 비율이 유사해보이지만,마찬가지로 상파울루 주(SP)에 거주하는 사람들의 거래 비율이 높다. [표2]는 전체 데이터 93000개의 거래 중에서 고객과 판매자의 거주 지역 빈도를 보여주고 있다. 전체 거래 중 약 41%가 소비자가 상파울루에 거주 했던 경우였고, 상위 3개의 주; SP, RJ, MG 에서의 거래가 약 67%로, 대부분 남동부에 위치한 3개의 주에서 소비자의 주문이 이루어졌다.

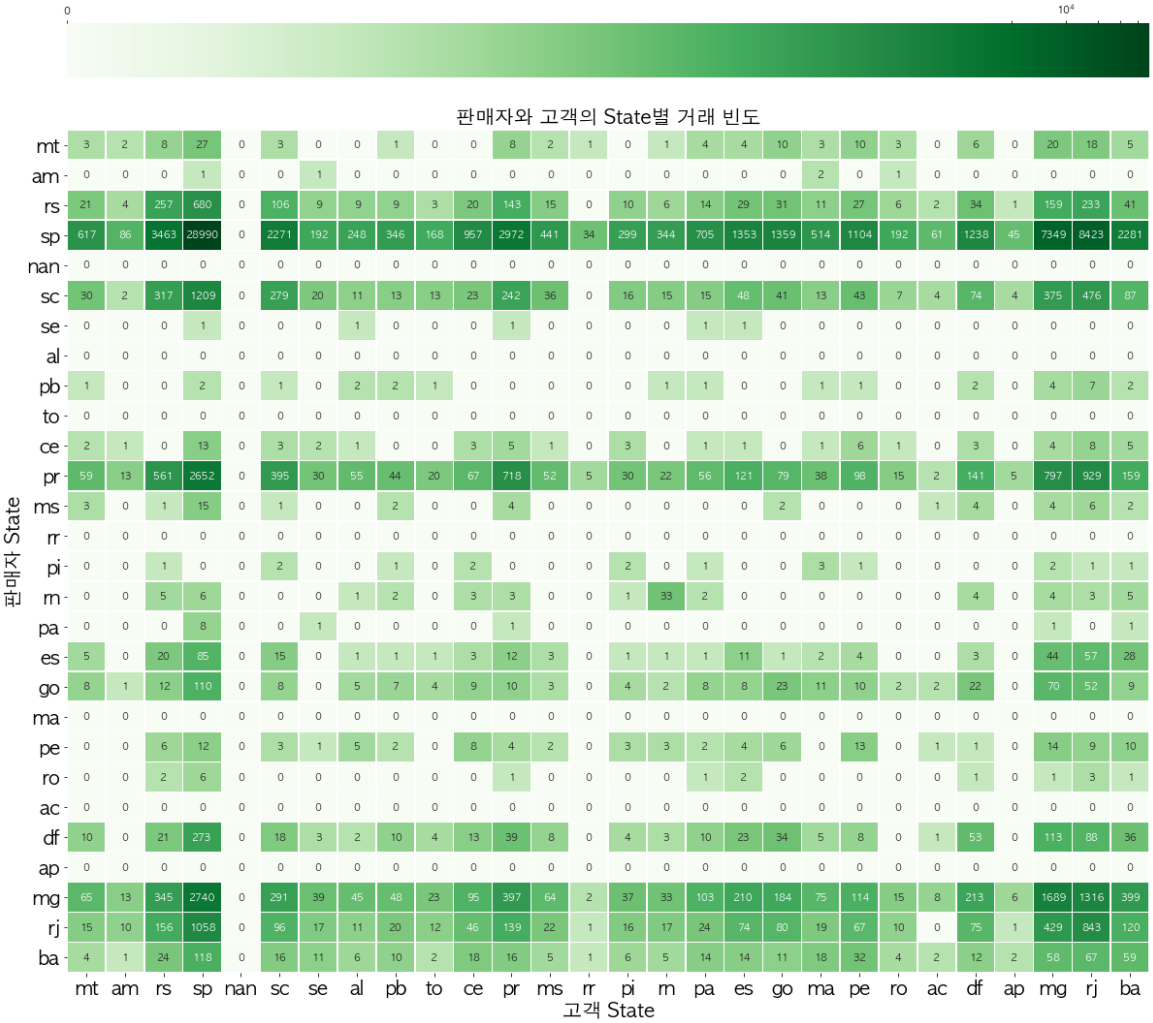
순위	약어	고객 빈도수	비율	약어	판매자 빈도수	비율
1	SP	38013	40.8%	SP	66052	71.0%
2	RJ	12541	13.5%	MG	8569	9.2%
3	MG	11138	12.0%	PR	7163	7.7%
4	RS	5201	5.6%	SC	3413	3.7%
5	PR	4715	5.1%	RJ	3378	3.6%

[표2] States별 고객, 판매자의 수 : 전체 거래 중 SP인 상파울루에서 고객은 약 41%, 판매자는 약 71% 의 거래 빈도가 있다.

판매자의 거주 State는 더 편향된 분포를 보이는데, 전체 거래의 상파울루(SP)에 거주하는 판매자의 비율이 약 71%로 대부분을 차지하고 있다. 따라서 상위 3개의 주; SP, MG, PR에서의 거래가 88%로 고객과 마찬가지로 판매자의 대부분은 3개의 주에 거주 중이다. 하지만, RJ 주 대신 PR 주가 상위에 존재함으로서 단순히 남동쪽이 아닌 남부 주가 상위권에 위치한 것을 볼 수 있다.

3. 배송 루트 탐색적 분석

1.1.(판매자 State – 고객 State) 별 거래 빈도



[그림3] 판매자와 고객 State 조합별 거래 빈도수를 나타낸 Heatmap : 매우 편향된 빈도수를 가지고 있었으므로 컬러맵을 로그스케일링 해주었다.

(판매자State – 고객 State) 별로 허용 가능한 배송 기간의 임계치(특정 배송 기간 기준으로 평점의 '부정' 비율이 '긍정' 비율보다 커지는 배송기간)가 다를 것

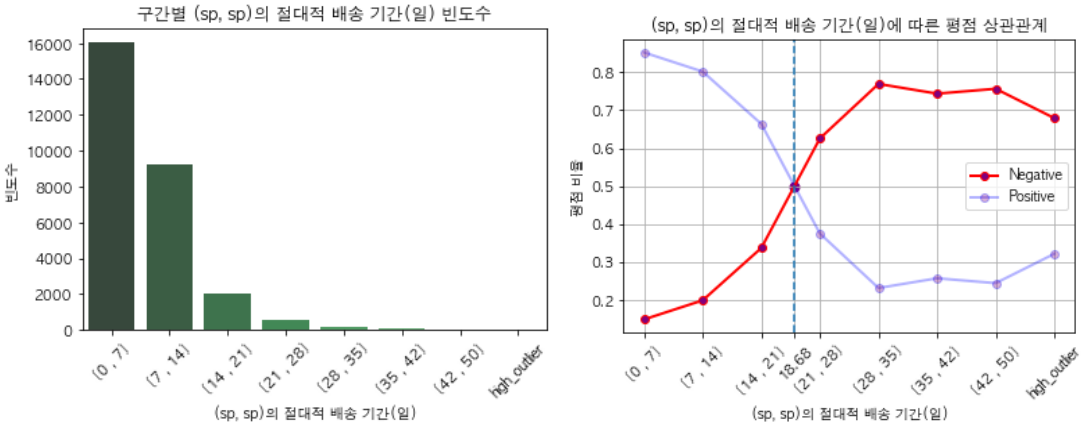
1. 거래 빈도수가 높은 (판매자 State – 고객 State) 별 허용 가능한 배송 기간 임계치 탐색
2. 지연된 배송의 비율이 높았던 (판매자 State – 고객 State) 별 허용 가능한 배송 기간 임계치 탐색

을 진행 하였다.

[탐색 1] 거래 빈도수가 높은 (판매자State – 고객 State) 별 허용 가능한 배송 기간 임계치 탐색

순위	조합	빈도 (비율)	임계 배송기간 (일)
1	SP - SP	28990 (31%)	18. 7일
2	SP - RJ	8423 (9%)	22.3일
3	SP - MG	7349 (8%)	22.8일
4	SP - RS	3463 (4%)	28.1 일
5	SP - PR	2972 (3%)	26.9일

[표3] 거래 빈도가 높은 상위 5개 조합 : 판매자와 고객 모두 남동쪽과 남부지역의 인접한 지역들의 조합이 빈도가 높다.



[그림4] 판매자, 고객 모두 SP에 거주하는 거래들의 배송 기간별 분포와 리뷰 평점 비율

[표3]를 보면, 예상대로 모든 조합 모두 남동쪽과 남부지방의 인접한 지역들이다. 하지만 판매자와 고객 모두 SP에 사는 조합의 경우가 전체의 약 31%로 매우 크고 그 중 일주일 이내에 배송이 온 경우가 매우 많은 것을 알 수 있다. 배송 기간의 일주일 단위마다 평점 기간의 비율을 살펴봤을 때, 긍정비율보다 부정 비율이 더 높아지는 배송 기간은 약 18.7일로 임계 배송 기간이 가장 짧게 도출되었다.

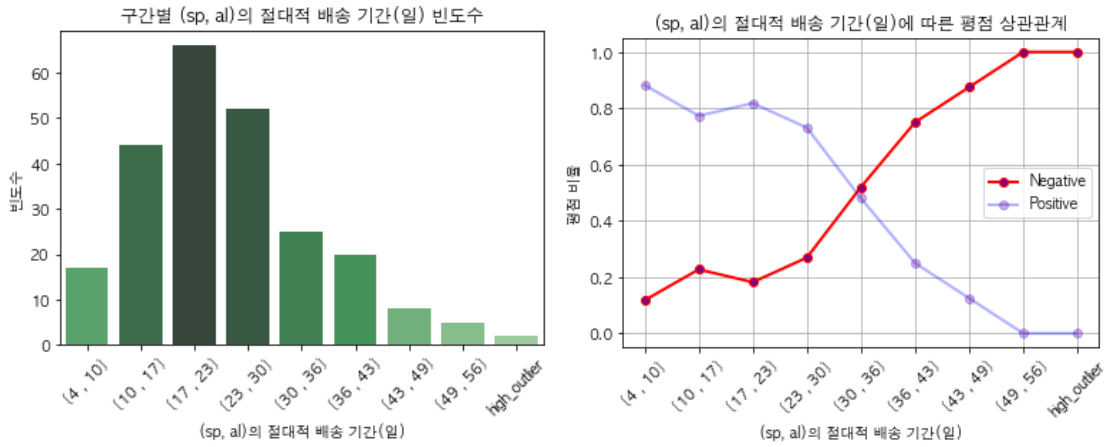
하지만, [그림 1]을 따르면, 상위 5개의 조합 중 판매자와 고객의 거주 State의 거리가 가장 먼 조합은 (SP – RS)이고 이 조합의 임계 배송 기간은 28.1일로 가장 크게 도출되었다. 또한, 다른 조합들도 마찬가지로 임계 배송 기간과 거리가 어느정도 비례하는 것으로 보인다. 같은 남동부지방인 SP, RJ, MG 간의 조합보다, 남부로 보내는 경우 임계 배송 기간이 더 커지는 것을 알 수 있다. 이것으로 판매자와 고객의 거주 State가 배송 기간 뿐만 아니라 평점에도 영향을 끼쳐 임계 배송기간과의 상관성이 있다는 것을 알 수 있게 되었다.

[탐색 2] 지연된 배송의 비율이 높은 (판매자State – 고객 State) 별 허용 가능한 배송 기간 임계치 탐색

순위	조합	지연된 배송 비율	임계 배송기간 (일)
1	PR(남부) - AL(북동부)	36.4%	24.8일
2	RS(남부) - PE(북동부)	25.9%	- (27.1)
3	SP(남동부) - AL (북동부)	25.4%	29.8일
4	BA(북동부) - SC (남부)	25.0%	- (27.1)
5	SP(남동부) - MA(북동부)	24.1%	26.7일

[표4] 지연된 배송의 비율이 높은 상위 5개 조합 : 모두 남부/남동부 – 북동부의 조합이 배송 이행을 지키지 못한 경우가 많게 도출되었다.

- 제약 : 조합별로 10 건 이상의 거래건 수(전체의 거래 건 수의 50%)가 있어야한다.
- 전체 조합의 지연된 배송 비율의 평균 : 10.5%



[그림4] 판매자는 SP, 고객은 AL에 거주하는 거래들의 배송 기간별 분포와 리뷰 평점 비율

우선, 지연된 배송의 의미는 주문의 예상 배송일 내에 도착을 못한 배송이고, 각 조합별로 그런 배송의 비중이 큰 순서대로 나열해 보았다. 모든 조합들의 지연된 배송의 비율의 평균은 약 10.5%이고, 상위 5개는 모두 평균보다 2.4배 이상의 지연된 배송의 비율이 높은 경우들이다. [표4]에서 보이듯이, 모든 조합이 (남부/남동부 – 북동부) 사이의 조합이며, 1절에서 언급했듯이 어느 정도 이상의 인구가 거주하고 있고 가장 거리가 먼 조합들이 도출되었다고 볼 수 있다. 이는 예상 배송일이 측정이 거리를 기반으로 되지 않고 있음을 보여준다.

하지만, 2번 조합인 (RS – PE)은 22건, (BA – SC)는 12건으로 빈도수가 너무 작아 임계 배송 기간 설정이 유의미하다 판단하기 힘들었으므로 1,3,5위 조합들의 평균으로 제시한다.

[[모델 결과 해석을 바탕으로 한 정리]]

앞서 의사 결정 나무 모형의 결과로 Olist는

[1]. 거리를 기반으로한 적절한 예상 배송일 선정

[2]. 판매자에 대한 엄격한 관리 필요성

이 결과로 도출 됐었다. 하지만, 정확한 수치적 제안을 위해 더 깊이 있는 탐색 과정을 거쳤고, 1차적으로는

[탐색1] Olist의 (판매자 – 고객) 거주 State조합 별로 서로 다른 임계 배송 기간을 가지고 있고 그것이 거리와 비례한다는 것,
[탐색 2] 그리고 배송 불이행의 비율이 높을수록 (남부/남동부 – 북동부)를 오가는 운송이라는 통일성을 발견함으로써 더 깊이 있는 검증 과정을 거쳤다.
2차적으로는 각 States **조합별 거리를 기반으로한 수치적 임계 배송 기간을 제시**하게 되었다.

3.4. 비정형 데이터 분석

- 리뷰 텍스트 분석을 통한 요인 파악

(1) 데이터 수집

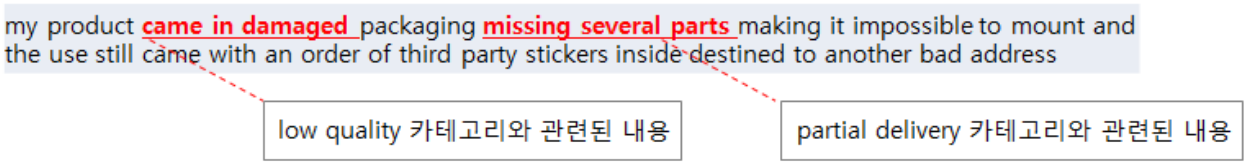
- 3,584개의 데이터
- 이 데이터셋은 기존의 리뷰 메세지를 3명의 분석가가 본인들의 판단 하에 class에 따라 분류 해 놓은 데이터셋이다. 가장 투표를 많이 받은 class로 리뷰 메시지가 분류된다.

데이터 변수 그룹	테이블 설명	칼럼 속성	
review	리뷰와 관련된 정보	리뷰 점수, 리뷰 제목, 리뷰 메시지, 리뷰 작성 날짜, 리뷰 요청 날짜	
votes	3명의 분석가의 판단에 의해 받은 득표수	votes_before estimate	제품을 배송 예정일 보다 빨리 받음
		votes_delayed	늦은 배송에 따른 불만사항
		votes_low quality	낮은 제품의 질에 따른 불만사항
		votes_return	셀러에게 다시 반송하려는 불만사항
		votes_not as anounced	공지한것과 다른 내용에 따른 불만사항
		votes_partial delivery	부분 배송에 따른 불만사항
		votes_other delivery	기타 배송에 관한 불만사항
		votes_other order	기타 주문내용과 관련된 불만사항
		votes_satisfied	주문에 대해 만족한다는 내용
most voted	가장 득표수가 높은 class	most_voted_class	가장 득표수가 높아 분류된 카테고리
		most_voted_subclass	subclass를 3개로 묶은 class (satisfied, delivery complaints, quality complaints)

(2) 데이터 리뷰 및 리뷰 카테고리 정의

- 리뷰 데이터를 팀원들이 직접 읽으면서 리뷰 하던 중 내용과 카테고리가 일치하지 않는 부분이 발견됨.

예시) return 카테고리에 분류되었던 리뷰 메세지



➔ 사람의 의견이 들어가기 때문에 **제품이상 -> 반품** 이라고 판단하여 return 카테고리에 들어갔다. 하지만 damaged는 low quality에, missing several parts는 partial delivery 카테고리에도 해당되는 구문.

- 기존의 9개의 카테고리가 리뷰 메세지를 제대로 분류하지 못한다는 의견이 나옴.
- 너무 많은 문제들이 하나의 카테고리에 포괄적으로 들어 있었다.

기존 9개의 카테고리		
low quality	not as announced	other delivery
delayed	return	other order
before estimate	partial delivery	satisfied

팀원들이 3,584개의 리뷰를 다 읽은 후 토의 후 재정의

새롭게 재정의한 14개의 카테고리		
주제	**카테고리**	**조작적 정의**
제품관련	**제품 불량**	제품의 질이 나쁨, 제품이 생산된 이후에 생긴 문제들
	제품 훼손	제품이 훼손 된 경우, 제품에 스크레치, 찌그러짐
	not as announced	받은 제품이 웹사이트 설명 혹은 사진과 다른 경우
배송관련	**배송부주의**	확실히 배송으로 인해 패키징에 생긴 문제들
	배송누락 및 부분배송	배송누락: 1개 제품을 주문했는데 해당 제품의 부분이 배송되지 않은 경우 부분배송: 복수의 제품을 주문했는데 모든 개수가 배송되지 않은 경우
	잘못된 배송	주문한 제품과 상이한 제품이 온 경우
	delayed	배송예정일자보다 배송이 늦은 경우
	before estimate	배송예정일자보다 배송이 일찍 온 경우
	other delivery	배송관련 문제이나 위의 분류에 포함되지 않는 경우
주문 관련	**응답 문제**	문제가 발생하여 고객이 문의했는데 응답이 느리거나 없는 경우
	return	'명시적으로' 교환, 환불을 원한다는 표현이 있는 경우
	other order	주문 관련 문제이나 위의 분류에 포함되지 않는 경우
기타	satisfied	제품과 서비스에 대해 만족한 모든 표현
	기타	배송 관련 문제 혹은 주문 관련 문제로 특정하기도 어려우며 매우 특이한 문제

(3) dictionary 및 정답지 제작

'구/문장' 단위로 딕셔너리 생성 :

기존 LDA나 토픽분석, 단순 워드 카운팅의 문제점은 단어들의 빈도를 기준으로 결과가 나오고 대부분 단어 단위로 결과가 나오게되어 유의성이 떨어지게 된다. 특히, '토픽'으로서 도출된 단어들간에 연관성을 '인간'으로서는 유추 가능하지만 말 그대로 유추일 뿐, 어떠한 문맥이 작용했는지 정확한 근거를 잃게된다. 따라서, 우리는 문맥을 고려하여 단어가 아닌 '구/문장' 단위로 딕셔너리 작성하였다.

정답지 생성 : 팀원 두명이 직접 모든 리뷰를 읽으면서 새롭게 Binary로 해당 카테고리에 표시했다.

(4) 리뷰 분류

카테고리별 카운트 기준 :

1. 카테고리별 딕셔너리를 전체 표현 기준 빈도를 센다.

2. 주의사항 : '만족' 카테고리

짧고 단순한 형용사 및 동사의 표현이 많다. (예) 'good' 'recommend'

→ 'not good' 'do not recommend'라는 표현과 구분이 필요하다.

따라서, 정규표현식을 활용하여 앞에 'not'이 붙이 않아야한다는 조건을 달아서 오분류율을 낮췄다.

(5) 결과 : 혼동표

```
만족 Confusion Matrix :
[[1401  153]
 [ 420  812]]
만족 Accuracy Score : 0.7943287867910983
만족 Report :
              precision    recall  f1-score   support

         0.0       0.77      0.90      0.83      1554
         1.0       0.84      0.66      0.74      1232

   micro avg       0.79      0.79      0.79      2786
   macro avg       0.81      0.78      0.78      2786
weighted avg       0.80      0.79      0.79      2786
```

```
제품불량 Confusion Matrix :
[[2600   59]
 [  29   98]]
제품불량 Accuracy Score : 0.968413496051687
제품불량 Report :
              precision    recall  f1-score   support

         0.0       0.99      0.98      0.98      2659
         1.0       0.62      0.77      0.69       127

   micro avg       0.97      0.97      0.97      2786
   macro avg       0.81      0.87      0.84      2786
weighted avg       0.97      0.97      0.97      2786
```

```
제품훼손 Confusion Matrix :
[[2710   27]
 [   4   45]]
제품훼손 Accuracy Score : 0.988872936109117
제품훼손 Report :
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99      2737
         1.0       0.62      0.92      0.74        49

   micro avg       0.99      0.99      0.99      2786
   macro avg       0.81      0.95      0.87      2786
weighted avg       0.99      0.99      0.99      2786
```

```
포장문제 Confusion Matrix :
[[2745   13]
 [   4   24]]
포장문제 Accuracy Score : 0.9938980617372577
포장문제 Report :
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      2758
         1.0       0.65      0.86      0.74        28

   micro avg       0.99      0.99      0.99      2786
   macro avg       0.82      0.93      0.87      2786
weighted avg       1.00      0.99      0.99      2786
```

```
과대광고 Confusion Matrix :
[[2696   31]
 [  17   42]]
과대광고 Accuracy Score : 0.9827709978463748
과대광고 Report :
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99      2727
         1.0       0.58      0.71      0.64        59

   micro avg       0.98      0.98      0.98      2786
   macro avg       0.78      0.85      0.81      2786
weighted avg       0.98      0.98      0.98      2786
```

```
빨리도착 Confusion Matrix :
[[1561  403]
 [  39  783]]
빨리도착 Accuracy Score : 0.8413496051687006
빨리도착 Report :
              precision    recall  f1-score   support

         0.0       0.98      0.79      0.88      1964
         1.0       0.66      0.95      0.78       822

   micro avg       0.84      0.84      0.84      2786
   macro avg       0.82      0.87      0.83      2786
weighted avg       0.88      0.84      0.85      2786
```

```
배송지연 Confusion Matrix :
[[2124  251]
 [  27  384]]
배송지연 Accuracy Score : 0.9002153625269204
배송지연 Report :
              precision    recall  f1-score   support

         0.0       0.99      0.89      0.94      2375
         1.0       0.60      0.93      0.73       411

   micro avg       0.90      0.90      0.90      2786
   macro avg       0.80      0.91      0.84      2786
weighted avg       0.93      0.90      0.91      2786
```

```
부분배송&누락 Confusion Matrix :
[[2461   52]
 [  67  206]]
부분배송&누락 Accuracy Score : 0.957286432160804
부분배송&누락 Report :
              precision    recall  f1-score   support

         0.0       0.97      0.98      0.98      2513
         1.0       0.80      0.75      0.78       273

   micro avg       0.96      0.96      0.96      2786
   macro avg       0.89      0.87      0.88      2786
weighted avg       0.96      0.96      0.96      2786
```

```
잘못배송 Confusion Matrix :
[[2661   29]
 [  33   63]]
잘못배송 Accuracy Score : 0.9777458722182341
잘못배송 Report :
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99      2690
         1.0       0.68      0.66      0.67        96

   micro avg       0.98      0.98      0.98      2786
   macro avg       0.84      0.82      0.83      2786
weighted avg       0.98      0.98      0.98      2786
```

```
기타배송 Confusion Matrix :
[[2585   64]
 [  41   96]]
기타배송 Accuracy Score : 0.9623115577889447
기타배송 Report :
              precision    recall  f1-score   support

         0.0       0.98      0.98      0.98      2649
         1.0       0.60      0.70      0.65       137

   micro avg       0.96      0.96      0.96      2786
   macro avg       0.79      0.84      0.81      2786
weighted avg       0.97      0.96      0.96      2786
```

```
서비스 Confusion Matrix :
[[2575   51]
 [  76   84]]
서비스 Accuracy Score : 0.9544149318018664
서비스 Report :
              precision    recall  f1-score   support

         0.0       0.97      0.98      0.98      2626
         1.0       0.62      0.53      0.57       160

   micro avg       0.95      0.95      0.95      2786
   macro avg       0.80      0.75      0.77      2786
weighted avg       0.95      0.95      0.95      2786
```

```
기타 Confusion Matrix :
[[2481   94]
 [ 145   66]]
기타 Accuracy Score : 0.9142139267767408
기타 Report :
              precision    recall  f1-score   support

         0.0       0.94      0.96      0.95      2575
         1.0       0.41      0.31      0.36       211

   micro avg       0.91      0.91      0.91      2786
   macro avg       0.68      0.64      0.65      2786
weighted avg       0.90      0.91      0.91      2786
```

```
Return Confusion Matrix :
[[2609   50]
 [   8  119]]
Return Accuracy Score : 0.9791816223977028
Return Report :
              precision    recall  f1-score   support

         0.0       1.00      0.98      0.99      2659
         1.0       0.70      0.94      0.80       127

   micro avg       0.98      0.98      0.98      2786
   macro avg       0.85      0.96      0.90      2786
weighted avg       0.98      0.98      0.98      2786
```

[결과 해석]

F-1 점수가 0.65이하인 카테고리 : '서비스', '기타배송' , '기타', '과대광고'

1. 카테고리 생성때부터 어렵고 애매한 표현이 많았던 '서비스' '기타배송' '기타' 부분의 카테고리들의 f1-score가 낮은 것을 알 수 있다.

- [예상된 원인] 각 카테고리에 대한 정답지 생성자의 이해도 차이

2. '과대광고' 카테고리는 표현이 대부분 연속된 단어의 구로 딕셔너리를 생성하기 어려운 경우들이 많았다. 정교한 정규표현을 활용한 분류 방안을 마련해야한다.

대부분 90% 이상의 정확도 : '빨리도착' 을 제외한 모든 카테고리의 정확도 90% 이상이 나왔다.

[딕셔너리 결과 예시]

Hello.

I just received two pendants, but I bought three

and I sent an e-mail complaining to notify that it is missing an item for you so far.

I have not received an answer.

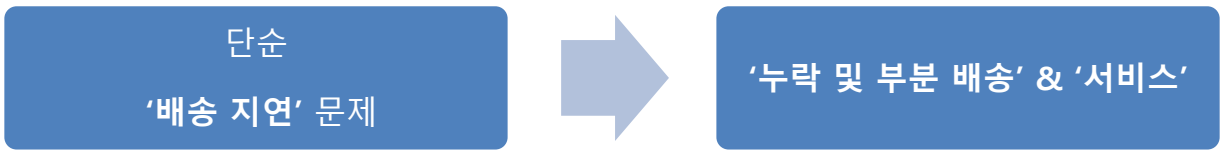
I do not know about the fault of the mail or the shop request.

[한글 내용] 3개를 주문했지만 2개만 받았다, 그리고 불만을 이메일로 보냈지만 아직까지 나는 물건을 받지 못했다. 그리고 답장도 받지 못했다. **→ 가장 큰 문제 = 판매자의 무응답**

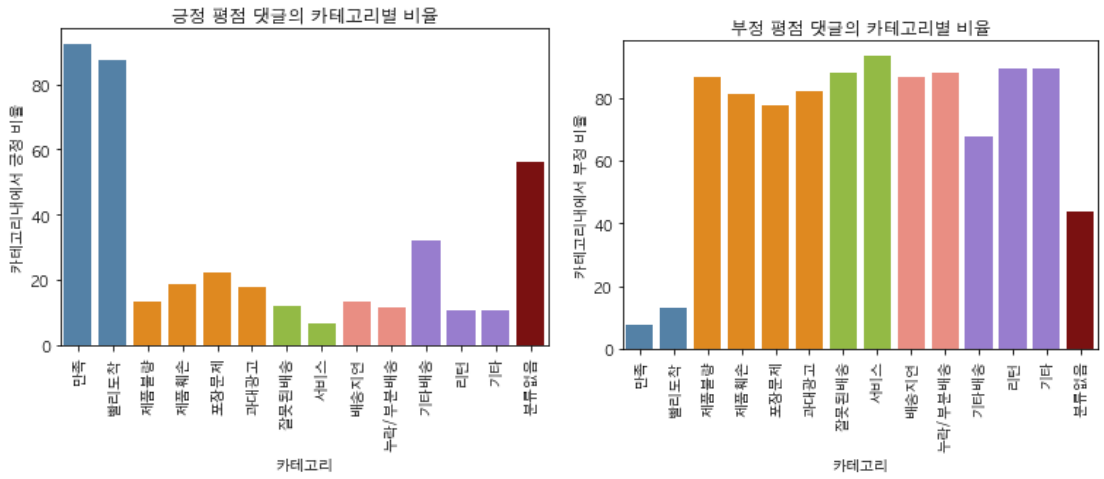
3 개의 카테고리에 중복 투표된다.

	배송지연	배송누락 및 부분배송	서비스
투표수	4	2	5
딕셔너리 표현	so far, have not received, not received, so far i have not received	missing an item, just received	sent an e-mail complaining, have not received an answer, complaining to notify, not received an answer, sent an e-mail

단순 **'배송 지연'** 문제에서 **'누락' & '서비스'** 문제까지 도출 가능해졌다.



(6) 검증



[한계점] :

제한된 데이터이기 때문에, 3500 개의 데이터셋에서는 우리가 생성한 파생변수 생성이 어려워서 같은 데이터셋에 분류 검증하기가 불가능하다.

[대체 검증 과정] :

1. 동일 딕셔너리를 93000 개 중 리뷰데이터가 있는 약 '38892' 개의 데이터에 카테고리별 빈도를 측정
2. '가장 많이 득표받은 카테고리'를 기준으로 각 리뷰를 분류를 하였다.

(예) [카테고리별 특표수] = [0,0,0,4,0,2,0,0,0,1,0,0,0,1] 인 경우 4 표를 받은 카테고리명으로 분류

- 이때, 아무런 득표를 받지 못한 텍스트 리뷰들은 '분류없음'으로 분류했다.

3. 각 분류된 카테고리를 검증하기 위해서
(1) 리뷰 점수가 긍정(4점 / 5점)인 리뷰에서 카테고리별 비율과
(2) 리뷰 점수가 부정(1점 /2점 / 3점)인 리뷰에서 카테고리별 비율을 보았다.[그림]

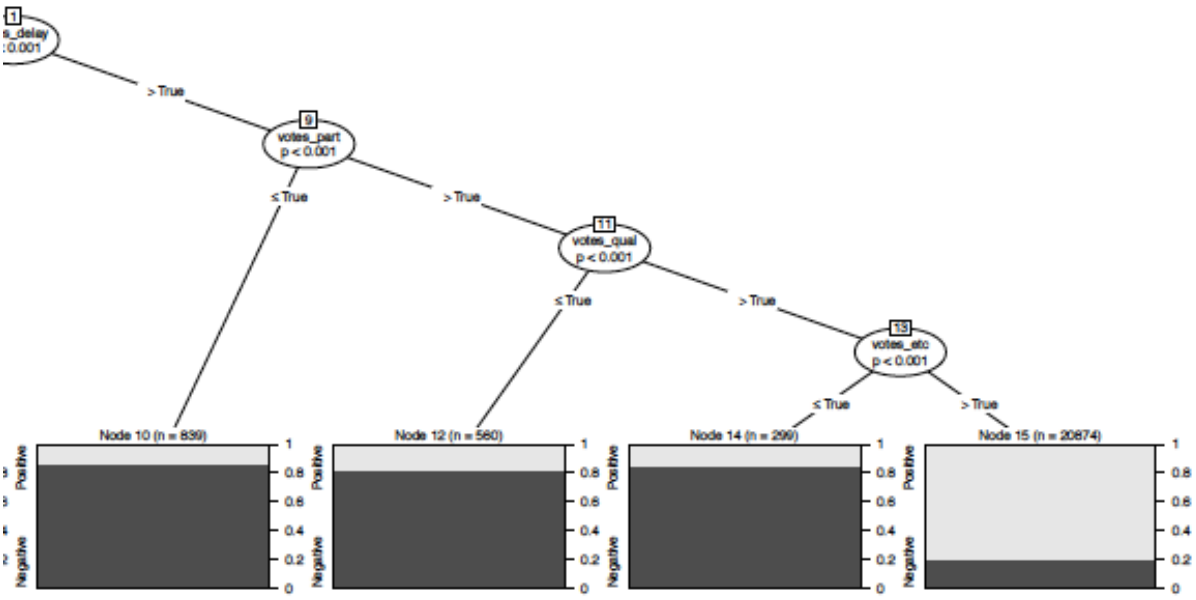
(7) 결과 해석

검증 결과 :

1. 왼쪽인 긍정점수를 가진 리뷰들은 확실히 긍정적 카테고리인 '만족'과 '빨리도착'의 비율이 압도적으로 큰 것을 볼 수 있다.
2. '분류 없음'은 부정, 긍정에 거의 동일한 비율로 분포함으로서, 어느쪽으로도 분류하기 애매한 카테고리라는 것을 검증한 것이라 가정할 수 있다.
3. 특히, '포장문제'와 '잘못된 배송', '누락 및 부분배송' 이라는 명백한 판매자의 실수로 부터 야기된 문제들이 부정적 리뷰에 많이 분포되어 있는데, 이것은 판매자에 대한 엄격한 관리의 필요성을 상기시킨다.

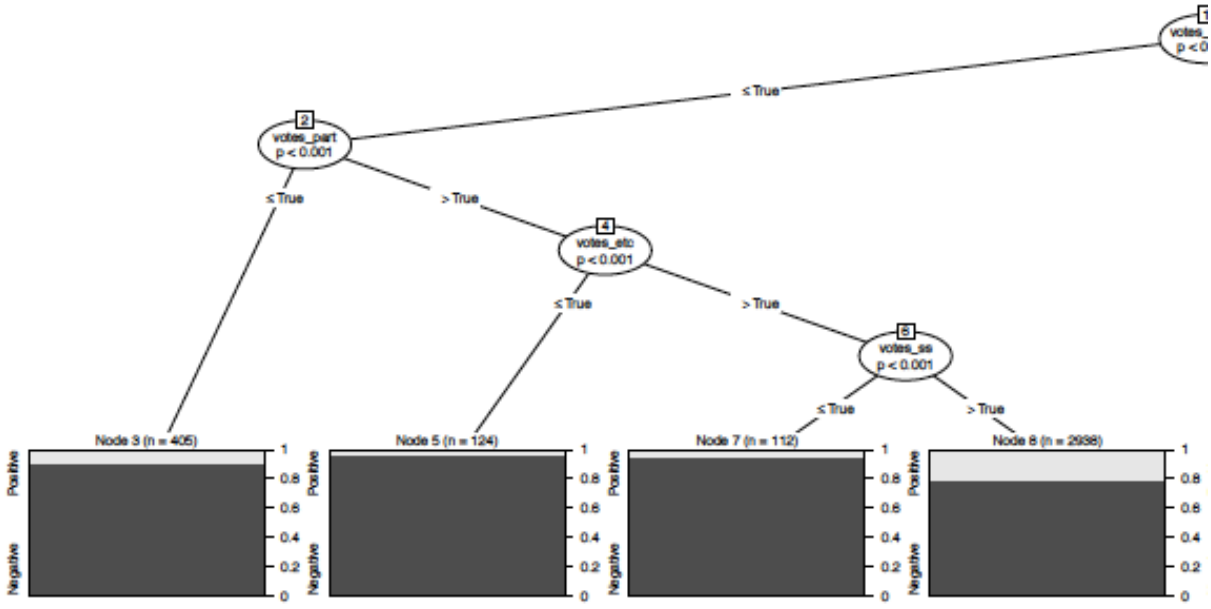
3.5. 정형과 비정형 데이터 결과 결합

3500개의 비정형 데이터의 훈련세트가 아닌 정형 데이터셋에 우리가 딕셔너리로 투표한 카테고리별 투표수 빈도 칼럼 14개를 추가하여, 다시 의사 결정 나무를 실행시켜 보았다.



[오른쪽 분기 해석]

여전히 첫번째 가장 중요한 변수로는 '배송 지연' 변수가 도출됐다. 하지만 제품 불량이어도 기타 카테고리 부분이 중요 변수로 도출되었다.



중요 변수로 부분 배송, 기타, 서비스 부분이 도출 되었다.
이로서, 배송 관련 변수가 결합한 모델에서도 중요 변수로 도출되었지만, 우리는 새로운 요인 인자를 도출해 낼 수 있게 됐다.

```
Confusion Matrix and Statistics

          Reference
Prediction Positive Negative
  Positive     7151     1731
  Negative      414     1911

               Accuracy : 0.8086
                 95% CI : (0.8012, 0.8158)
    No Information Rate : 0.675
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5186
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9453
            Specificity : 0.5247
         Pos Pred Value : 0.8051
         Neg Pred Value : 0.8219
             Prevalence : 0.6750
         Detection Rate : 0.6381
   Detection Prevalence : 0.7925
      Balanced Accuracy : 0.7350

       'Positive' Class : Positive
```

[결과 정확도] : 결과 정확도가 80%로 이전 정형 데이터로 돌린 결과보다 매우 향상되었고, 특이도도 0.2에서 0.5로 많은 성능이 개선됐다.

4. 기대 효과

4.1 향후 개선 사항

[정형데이터]

다른 알고리즘을 사용하여 정확도 개선

[비정형데이터]

정규 표현식을 활용하여 더 복잡한 표현을 읽어낼 수 있도록 개선

4.2 기대 효과

1. 고객 불만족에 대한 원인 파악

- 리뷰 평점은 고객 만족도의 척도이므로 리뷰 평점 영향인자를 식별하는 것은 고객 불만족에 대한 원인을 파악할 수 있음을 의미한다.

2. 고객 만족도 향상

- E-Commerce 사업자는 식별된 요인뿐만 아니라 함께 제시하는 개선방안을 활용하여 실제 고객 만족도를 향상 시킬 수 있을 것으로 기대된다.

3. 리뷰 평점 개선으로 인한 판매량 증가

- E-commerce 에서 구매 결정 고려에 있어 가장 중요한 기준인 리뷰 평점 자체를 개선함으로서 판매량이 증가될 것으로 기대된다. (평점이 1 점 오를 때 평균적으로 매출이 5-9% 증가한다는 연구결과가 있음)

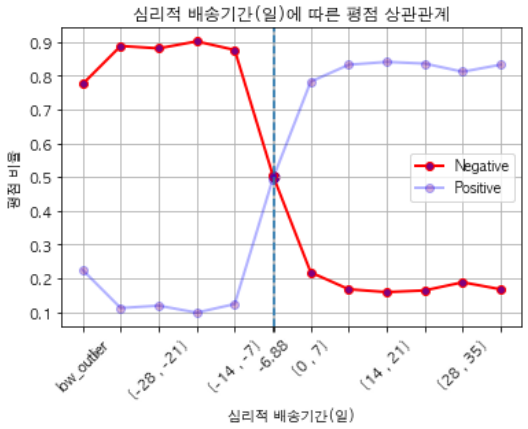
4. 리뷰 코멘트를 활용한 정확한 요인 파악

- 전체 리뷰 코멘트의 샘플을 추출하여 불만사항 주제별로 분류한 후, 각 분류마다 딕셔너리 작성할 계획이다. 이를 활용하면 전체 리뷰 코멘트 혹은 새로운 리뷰 코멘트에 대해 불만 사항의 종류를 자동으로 집계 가능하다. 텍스트로 더 정확한 요인 파악이 가능해진다.

[비즈니스적 Olist 제안 사항]

1. 앞선 정형 데이터의 결과 = 거리 기반 임계 배송일

2.

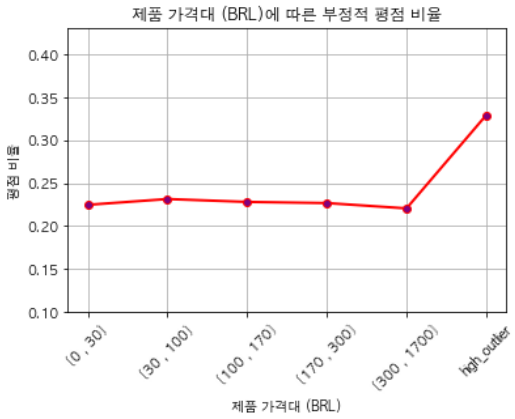


예상 배송일보다 약 7 일 이상 늦게 도착한 배송에서의 부정적 평점 비율의 매우 높게 나타났다.

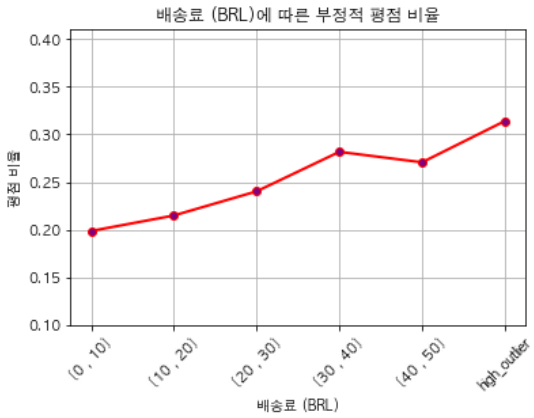
2.1. 예상 배송일보다 7 일 이상 늦게 보내는 판매자들에 대해서 패널티 부여

2.2. 판매자의 거주지역 표시 및 평균 배송일 표시로 경각심 부여

3. 가격대 + 배송 거리 :가격대 50 만원 이상 주문 고객은 VIP 로 최 우선 고려



4. 배송료의 비중 & 적절한 배송료 산정 필요



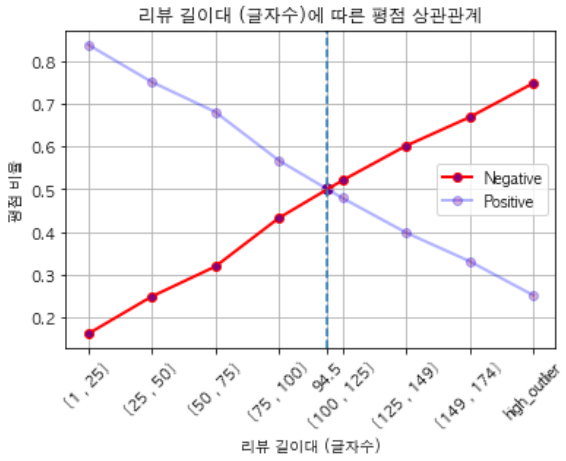
3. 거주 지역 조합별로 임계 배송일 제안

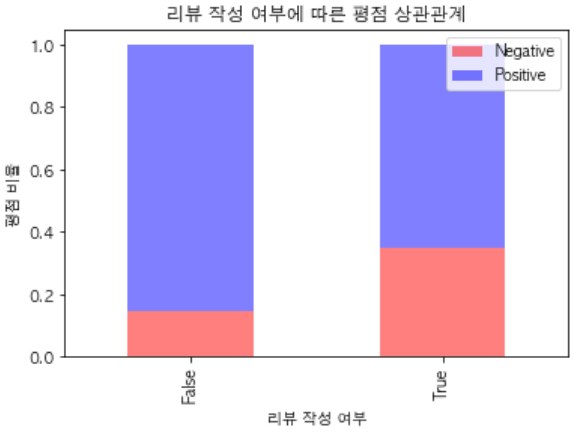
4. **글자수** 제한 & 리뷰 작성 유도

긍정적 리뷰의 글자수는 대부분 100 자 이상이 넘어가지 않는다. 따라서 양질의 긍적적 후기를 작성하도록 유도하도록

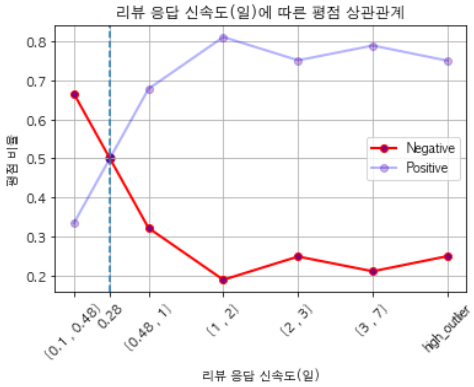
4.1. 베스트 후기를 남겨준 고객에게 리워드 지급

4.2. 리뷰 작성 길이에 최소한의 제한(90 자)을 주어 리뷰의 질을 높인다.

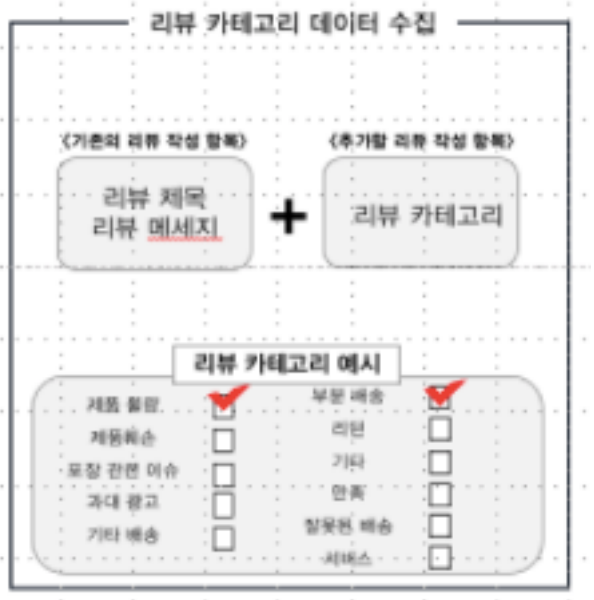




5. 빨리 리뷰를 쓰는 사람들(0.28 일 = 7 시간 이내) → 부정적 → 받자마자 부정적이었다는 것은 우선적이고 신속한 조치가 필요하다.



6. 비정형 데이터에서 14 개의 카테고리별로 주제를 선정하였으므로, 앞으로 리뷰 쓰는 곳에 14 개의 카테고리를 작성자가 체크하도록 시스템을 개선한다.



5. 분석 후기

팀 사진 첨부

성명	후기
권민수	빅데이터 분석이란 세상에 넘쳐나는 무의미한 데이터들 사이에서 다이아몬드를 채굴하는 것이라는 지문을 읽은적이 있습니다. 이해도 하기 어렵고 무의미 해보였던 olist 데이터에서 유를 찾은 것 같아 뿌듯합니다 JK 타이거조..사랑합니다
김이영	어려운 점도 많았고 힘든 점도 많았지만 좋은 경험이었습니다. 감사합니다.

김종인	비록 배운지 얼마 안된 지식을 활용하여 프로젝트를 진행하느라 힘든 점이 많았지만, 프로젝트 중간중간 분석을 통해 새로운 발견을 할 때마다 그러한 수고들을 잊을 수 있었습니다. 무엇보다 끝까지 자신의 임무를 수행해준 팀원들에게 고맙다는 말을 전하고 싶습니다.
문지현	분석 프로젝트라는 것을 처음 접하면서 새로운 특징의 팀플을 할 수 있어서 좋았고 현직에서 어떤식의 일을 하는지 간접적으로 배울 수 있어서 좋았습니다.앞으로 현업에 나가서도 이러한 경험이 크게 도움이 될 것 같습니다. 사랑합니다 JK 조

Reference

[1] https://ko.wikipedia.org/wiki/%EB%B8%8C%EB%9D%BC%EC%A7%88#cite_note-2

[2]https://ko.wikipedia.org/wiki/%EB%B8%8C%EB%9D%BC%EC%A7%88%EC%9D%98_%EC%A3%BC

[3] 주 브라질 대한민국 대사관[09112017] : http://overseas.mofa.go.kr/br-ko/brd/m_6115/view.do?seq=1159945&srchFr=&srchTo=&srchWord=&srchTp=&multi_itm_seq=0&itm_seq_1=0&itm_seq_2=0&company_cd=&company_nm=&page=1

[4] "2014 IBGE Estimates - Estimates of Resident Population in Brazil, Federative Units and Municipalities" (PDF) (in Portuguese). IBGE.gov.br. Retrieved 12 September 2014.

[5] 영화평점 출처 - 한겨레 '누리꾼 영화 평점이 흥행 흔든다'

[6] Yelp 출처 Reviews,Reputation,and Revenue: The Case of Yelp.com by Michael Luca. Harvard Business School Working Journal

[7] 데이터 링크 : https://www.kaggle.com/olistbr/Brazilian-ecommerce/version/5#geolocation_olist_public_dataset.csv

Appendix

가설과 그에 따른 파생변수

	가설	변수명	자료형	파생변수 로직
1	(판매자-고객 거리를 고려하지 않은) 절대적 배송기간이 짧을수록 평점은 높다.	절대적 배송기간	FLOAT	고객물품수령 날짜 – 구매시간
2	심리적 배송기간이 짧을수록(예상 배송일자보다 실제 배송일자가 빠를수록) 평점은 높다.	심리적 배송기간	FLOAT	배송예정 날짜 – 고객물품수령 날짜
3	FALSE인 경우, 즉 예상일을 초과하여 배송한 경우 평점은 낮다.	배송기한 준수 여부	BOOL	예상일을 초과해서 배송한 경우 FALSE 예상일 내에 배송한 경우 TRUE
4	길이가 적당하게 긴 경우 평점이 높다.	제품묘사 길이	INT	-
5	제품에 대한 사진 개수가 많을 수록 평점이 높다.	제품사진 개수	INT	-
6	쿠폰을 사용해서 구매한 경우 평점이 높다.	쿠폰사용 여부	BOOL	지불방식 중 voucher가 포함되어 있는 경우 TRUE, 아닌 경우 FALSE (참고: order_id는 전체 데이터셋에서 primaty key로서 unique하다)
7	주문의 배송료가 작을수록 평점이 높다	배송료	FLOAT	-
	상품가격 대비 배송료 비중이 낮을수록 평점이 높다.	배송료 비중	FLOAT	배송료 / 제품가격
8	주문시간대와 리뷰평점은 관련이 있을 것이다.	주문 시간대	범주형	24시간을 6시간 단위로 나누어 범주형 변수를 생성 *1시간 단위, 4시간 단위, 6시간 단위로 나눌 수 있음
9	구매요일이 월요일, 화요일인 경우에 리뷰 평점이 높을 것이다.	주문 요일	범주형	날짜를 요일로 변환
10	평일일 경우에 리뷰 평점이 높을 것이다.	주문 평일주말	범주형	평일인 경우 1, 주말일 경우 0
11	리뷰작성시간대와 리뷰평점은 관련이 있을 것이다.	리뷰작성 시간대	범주형	24시간을 6시간 단위로 나누어 범주형 변수를 생성 *1시간 단위, 4시간 단위, 6시간 단위로 나눌 수 있음
12	리뷰작성요일과 리뷰평점은 관련이 있을 것이다.	리뷰작성 요일	범주형	날짜를 요일로 변환
13	리뷰작성 요일이 평일인지 주말인지가 리뷰평점에 영향을 미칠 것이다.	리유작성 평일주말	범주형	평일인 경우 1, 주말일 경우 0
14	할부를 한 경우 리뷰평점이 높을 것이다.	할부여부	BOOL	할부를 한 경우 1, 하지 않은 경우 0
	평균 할부개월수가 높을수록 리뷰평점이 높을 것이다	평균 할부개월수	FLOAT	할부개월수 / 결제수단 개수
	결제금액을 가중치로 계산한 가중평균 할부개월이 높을수록 리뷰평점이 높을 것이다	가중평균 할부개월수	FLOAT	(할부개월수 * 결제금액)의 합 / (결제금액)의 합
15	리뷰응답 신속도와 리뷰평점과 관계가 있을 것이다.	리뷰응답 신속도		리뷰 작성 날짜 – 리뷰 요청 날짜
16	리뷰작성유무와 리뷰평점과 상관관계가 있다	리뷰작성 여부	BOOL	리뷰를 작성한 경우 TRUE 리뷰 작성하지 않은 경우 FALSE

17	"고객이 상품을 구매결정 ~ 판매자의 승인"이 짧을수록 리뷰평점이 높을 것이다.	판매자 응답속도	INT/FLOAT	구매승인 시간 – 구매시간
18	지역별로 리뷰평점이 다를 것이다.	고객 거주도시	범주형	-
19	가격은 리뷰평점과 관련이 있을 것이다.	가격	FLOAT	총 제품가격 / 주문 내 제품 수
20	가격대와 리뷰평점은 관련이 있을 것이다.	가격대	범주형	파생변수 총 제품가격을 구간화 하여 파생변수 생성
21	리뷰의 길이가 짧은 경우 평점이 높다.	리뷰길이	INT	리뷰 내용 길이
22	거리가 길수록 리뷰 평점이 낮을 것이다.	판매자와 고객 직선거리	FLOAT	우편번호 앞 3자리를 기준 평균 위도 경도로 고객과 판매자 거리 계산(km)

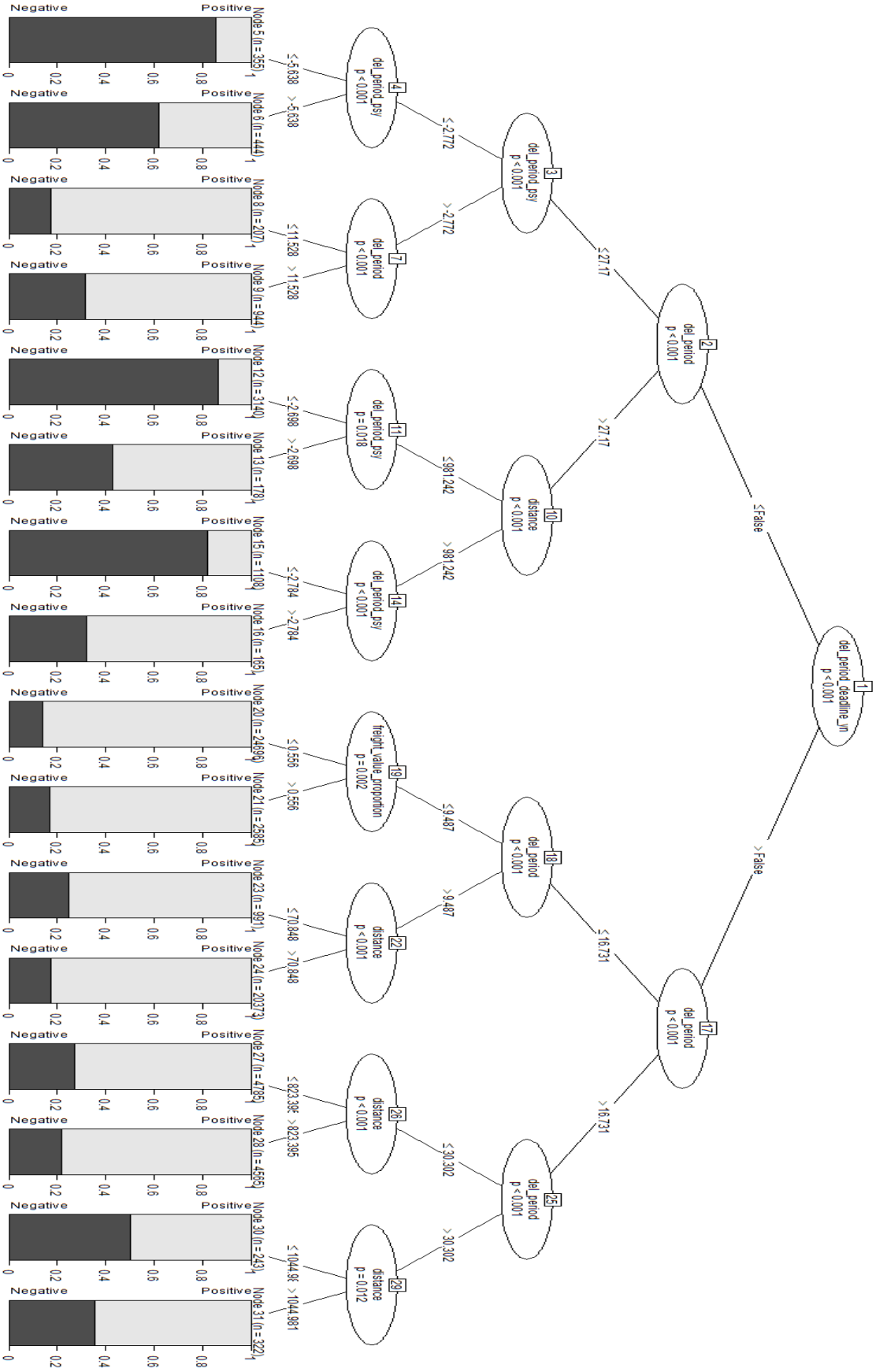
기초 통계

Table 1 : 변수 빈도수 분포

Feature	Unique Count	Top Value	Top Freq Count	Top Freq %
order_id	93000	79d2bfad385ea...	1	0.00108
order_status	7	delivered	90700	97.5269
order_purchase_timestamp	92673	2017-12-10 22:51	3	0.00323
order_aproved_at	92976	2017-12-27 14:03	3	0.00323
order_estimated_delivery_date	476	2017-12-20 0:00	493	0.53011
product_id	23175	aca2eb7d00ea1...	612	0.65806
product_category_name_english	71	bed_bath_table	8980	9.65591
customer_unique_id	90063	ff4ea78481e00...	10	0.01075
customer_state	27	sp	38013	40.8742
customer_city	4087	sao paulo	14003	15.057
customer_zip_code_prefix	851	130	1226	1.31828
seller_id	2091	6560211a19b47...	2598	2.79406
seller_state	21	sp	66052	71.0366
seller_city	475	sao paulo	21368	22.9805
seller_zip_code_prefix	427	149	6511	7.00236
review_score	5	5	53535	57.5645
review_creation_date	666	2017-12-19 0:00	437	0.46989
pn_review_score	2	Positive	71724	77.1226
del_period_deadline_yn	2	TRUE	83691	89.9903
payment_voucher_yn	2	FALSE	89323	96.0462
order_time_6	4	time_3	35968	38.6753
order_day	7	Tuesday	15024	16.1548
order_week_day_end	2	TRUE	71384	76.757
review_ans_time_6	4	time_1	31642	34.0237
review_ans_day	7	Friday	16158	17.3742
review_ans_w_d_e	2	TRUE	68843	74.0247
installments_yn	2	TRUE	48878	52.557
review_comment_yn	2	FALSE	54108	58.1806
distance				
price_range				

Table 2 변수 기술통계

Feature	Mean	Std	Min	25%	50%	75%	Max
order_products_value	131.94	198.42	2.00	48.90	85.00	149.00	13440.00
order_freight_value	21.88	20.12	0.00	13.64	16.79	23.07	1562.10
order_items_qty	1.10	0.46	1.00	1.00	1.00	1.00	20.00
order_sellers_qty	1.03	0.23	1.00	1.00	1.00	1.00	12.00
product_name_lenght	48.79	10.15	5.00	42.00	52.00	57.00	72.00
product_description_lenght	791.82	672.89	8.00	341.00	600.00	1001.00	3992.00
product_photos_qty	2.30	1.75	1.00	1.00	2.00	3.00	20.00
product_weight_g	2153.28	3895.27	50.00	250.00	700.00	1800.00	30000.00
product_length_cm	30.32	16.45	2.00	17.00	25.00	40.00	105.00
product_height_cm	15.84	13.61	1.00	7.00	12.00	20.00	105.00
product_width_cm	22.95	12.02	7.00	14.00	20.00	30.00	105.00
sequential	1.05	0.41	1.00	1.00	1.00	1.00	29.00
review_score	4.09	1.33	1.00	4.00	5.00	5.00	5.00
del_period	12.84	9.59	0.53	7.01	10.65	15.95	209.63
del_period_psy	11.38	10.21	-188.98	7.06	12.11	16.29	146.02
freight_value_proportion	0.30	0.29	0.00	0.13	0.22	0.36	12.24
sim_installments_mean	2.96	2.72	0.00	1.00	2.00	4.00	24.00
wgt_installments_mean	3.01	2.74	0.00	1.00	2.00	4.00	24.00
review_ans_period	3.34	11.09	0.17	1.13	1.80	3.24	532.83
seller_response_time	0.42	0.89	0.00	0.01	0.01	0.58	60.45
cus_lat	-21.15	5.64	-32.32	-23.60	-22.91	-20.02	2.72
cus_lng	-46.14	4.08	-69.15	-48.12	-46.62	-43.48	-33.77
sel_lat	-22.84	2.35	-31.75	-23.62	-23.31	-21.62	-3.08
sel_lng	-47.32	2.25	-63.69	-48.93	-46.77	-46.52	-34.85
order_product_value	123.61	179.57	2.00	45.00	79.90	139.90	9798.00
review_length	66.84	52.75	1.00	27.00	51.00	92.00	204.00
distance							
price_range							


del_period_deadline_ym
p < 0.001
≤False
>False
del_period
p < 0.001
del_period_psy
p < 0.001
≤27.17
>27.17
del_period_psy
p = 0.018
distance
p < 0.001
≤2.772
>2.772
≤11.528
>11.528
≤5.638
>5.638
≤2.698
>2.698
≤981.242
>981.242
del_period_psy
p < 0.001
del_period_psy
p < 0.001
≤2.784
>2.784
Node 5 (n = 355)
Node 6 (n = 444)
Node 8 (n = 207)
Node 9 (n = 944)
Node 12 (n = 3140)
Node 13 (n = 178)
Node 15 (n = 1108)
Node 16 (n = 165)
del_period
p < 0.001
freight_value_proportion
p = 0.002
≤16.731
>16.731
≤9.487
>9.487
distance
p < 0.001
distance
p < 0.001
≤70.848
>70.848
≤0.556
>0.556
Node 20 (n = 24696)
Node 21 (n = 2585)
Node 23 (n = 991)
Node 24 (n = 20373)
del_period
p < 0.001
distance
p < 0.001
distance
p = 0.012
≤30.302
>30.302
≤823.395
>823.395
≤1044.981
>1044.981
Node 27 (n = 47785)
Node 28 (n = 4565)
Node 30 (n = 243)
Node 31 (n = 327)
Negative
Positive
0
0.2
0.4
0.6
0.8
1