creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:11044-200000491504

석사학위 논문

소비자 리뷰를 통한 이커머스 새벽배송 서비스 요인 분석

-머신 러닝(분류)을 중심으로-

E-commerce early morning delivery service factor analysis through consumer reviews

-Focusing on machine running (Classification)-

2021년 6월

숭실대학교 대학원

IT유통물류학과

전병용

석사학위 논문

소비자 리뷰를 통한 이커머스 새벽배송 서비스 요인 분석

-머신 러닝(분류)을 중심으로-

E-commerce early morning delivery service factor analysis through consumer reviews

-Focusing on machine running (Classification)-

2021년 6월

숭실대학교 대학원

IT유통물류학과

전 병 용

석사학위 논문

소비자 리뷰를 통한 이커머스 새벽배송 서비스 요인 분석

-머신 러닝(분류)을 중심으로-

지도교수 현 병 언

이 논문을 석사학위 논문으로 제출함

2021년 6월

숭실대학교 대학원

IT유통물류학과

전 병 용

전 병 용의 석사학위 논문을 인준함

심 사 위 원 장 최 재 영 인

심 사 위 원 김 인 규 인

심 사 위 원 현 병 언 인

2021년 6월

숭실대학교 대학원

목 차

표 목 차

그 림 목 차

국문초록

소비자 리뷰를 통한 이커머스 새벽배송 서비스 요인 분석

-머신 러닝(분류)을 중심으로-

전병용

IT유통물류학과

숭실대학교 대학원

국내 새벽배송시장이 처음 도입된 시기는 2016년으로 2021년 현재까지 6년 동안 급속도로 성장 중이다. 초창기 새벽배송 시장 규모는 100억이었지만, 2020년에는 1조 5000억으로 초기 대비 150배 성장하였다. 따라서 새벽배송시장이 나아가야할 방향성을 제시하기 위한 연구가 필요하다. 시장의 방향성은 소비자의 요구사항을 반영해야 한다. 그러므로 소비자들이 중요시하는 서비스 요인 분석이 필요하다. 본 연구에서는 급속도로 성장하는 새벽배송시장을 파악하기 위해, 구글, 애플 앱 스토어 리뷰 데이터를 기계학습(분류)을 활용하여 연구모형(애플리케이션 사용요인, 배송요인, 제품요인, 마케팅요인)을 기반으로 새벽배송 서비스에 대한 소비자 만족 요인을 분석하였다.

본 연구는 새벽배송을 처음 선보인 마켓컬리 리뷰 데이터를 중심으로

연구를 진행하였다. 수집된 리뷰 데이터의 기간은 2016년 2월 20일부터 2021년 4월 16일까지이며, 총 10,390개 데이터를 수집하였다. 자연어 처리를 위한 전처리 과정을 진행 후, 최종 6,417개의 데이터로 기계학습(로지스틱 회귀분석)을 진행하였다.

본 논문은 소비자 서비스 요인 중에서 긍정적인 영향을 미치는 요인에 초점을 맞춰 분석을 진행하였다. 단어의 양의 관계와 음의 관계를 분류하기 위해 로지스틱 회귀 분석 모델을 사용하였다. 학습 모델상 양의 관계의 단어의 영향도를 점수로 산출하였고, 상위 10개의 단어를 출력하여 연구모형에 적용하였다.

적용 결과 연구모형 총 점수는 64.7점이며 제품요인은 29(44.8%)점, 애플리케이션 사용요인은 14.8(22.9%)점으로 제품요인이 애플리케이션 사용요인보다 서비스 만족도가 높은 것을 확인할 수 있다. 또한, 포장요인에서 '환경', '친환경' 단어가 군집되어 있다. 이를 보아 재활용 가능한 친환경 포장재 사용은 새벽배송 서비스 경험에 긍정적인 요소라고 판단된다. 향후 지속해서 증가하는 데이터를 기반으로 더 정확하게 고객이 만족에 영향을 미치는 서비스 요인을 파악할 수 있을 것이다. 또한, 고객의 니즈 파악이 가능하므로 새벽배송 서비스 품질 향상을 위한 경영진의 의사결정 지표로 활용될 수 있을 것으로 예상한다.

ABSTRACT

E-commerce early morning delivery service factor analysis through consumer reviews -Focusing on machine running (classification)-

JEON, BYOUNG YONG
Department of IT Distribution Logistics
Graduate School of Soongsil University

The first early morning delivery service in E-commerce market in Korea was introduced in 2016 and has been growing rapidly for last six years. The early morning delivery market has grown in 150 times of its size than that of the beginning; it started with 10 billion won, but it reached 1.5 trillion won in 2020. Accordingly, a well-found research is significantly needed to suggest the direction of the early morning delivery market. The direction of the market should reflect the demand of the consumers, thus the analysis of the service factors is critical in understanding the demand. This study analyzes the factors for customer satisfaction by extracting key words from the

user reviews of Google Play Store and Apple App Store and applying the machine learning (classification) to find the insight about the fast-growing early morning delivery service market.

User reviews of Market Kurly application, which first introduced the early morning delivery service to the market, are the primary data this study develops upon. Collected data date from February 20, 2016 to April 16, 2021, and the total of 10,390 were collected. The data were first pre-processed for the Natural Language Processing and filtered into total 6,417 data. Then, the machine learning (logistic regression) was processed consequently.

This paper focuses on and analyzes the service factors, which influence positively to the consumer experience on the early monring delivery service. A logistic regression model was adopted in order to classify words into positive and negative relationship. The influence of the words in positive relationship within the learning model is calculated and converted into scores, and top 10 words are used for the research model.

As a result, the total score of the research model is 64.7, while other factors score as follow: 'Product' records 29 (44.8%) and 'Application Use' scores 14.8 (22.9%). Thus, it is apparent that the 'Product' factor remarks the higher satisfaction in service than 'Application Use' factor does. In addition, terms such as 'Environment'

and ‘Eco-friendly’ are concentrated in the ‘’Packaging’ factor. This leads to the point which usage of sustainable containers and packaging do have optimistic influence on customers’ service experience. In near future, more precise study on the service factors that affect customer satisfaction is feasible with continuously growing data volume. At last, it is expected that this study, with its insight on customer needs, provides a valuable indication to the management executives for improving the service quality of the early morning delivery.

제 1 장 서 론

1.1 연구 배경 및 목적

새벽배송 서비스란 당일 오후 10시에서 12시 전에 제품을 주문할 경우 다음 날 아침 7시 전까지 집 앞으로 물품을 배송해주는 서비스로 2015년 8월 '마켓컬리'에서 처음 선보인 '샛별배송'이 시초이다. 이 서비스는 바쁜 일정으로 인해 오프라인으로 식품 또는 제품을 구매하기 어려운 현대인의 라이프스타일의 주축으로 자리 잡았다.

새벽배송은 라스트마일의 다양한 분야 중 하나이며, 이커머스가 급속도로 성장함에 따라 그 중요성이 부각되고 있다. 중소벤처기업진흥공단이 조사 및 발간한 '글로벌 이커머스 HOT 리포트'에 따르면 지난해 한국의 이커머스 매출은 1천 41억 달러이며, 세계 5위로 전년보다 19.5% 증가한 것을 알 수 있다[13]. 이와 같은 이커머스 시장의 확대가 새벽배송 서비스의 고속 성장에 기여했음을 유추할 수 있다.

이커머스의 성장으로 인해, 고객들의 소비 형태의 변화가 독립적으로 운영되는 온라인과 오프라인 채널이 경쟁하는 형태에서 서로 보완하는 형태인 옴니채널 단계로 이동하고 있다. 기존 국내의 유통 채널은 오프라인 위주로 운영되고 있었다. 하지만 위와 같은 변화로 오프라인 중심으로 성장했던 국내 물류유통업계는 채널의 전환이 필요해졌고, 현재 온라인 플랫폼 기반의 옴니채널 구축을 위해 지속해서 연구하고 있다.

이커머스의 발전과 함께 소비자들의 니즈가 고도화되고 다양해지면서

라스트마일의 필요성이 주목받기 시작하였다. 그중에서도 새벽배송은 오전, 낮 시간대에 택배를 받기 어려운 직장인, 맞벌이, 1인 가구 중심으로 주요 고객층이 형성되었다. 빠르고 안전하게 주문한 물건을 받을 수 있다는 장점이 현대 트렌드와 맞기에 지속해서 성장할 것으로 예상된다. 게다가 COVID-19의 장기화로 인해 온라인 시장이 더욱 증가하고 있는 상황이므로 다양한 새벽배송 업체들의 경쟁은 심화하고 있다.

특히 새벽배송은 주문 시간의 제약이 없고, 빠르게 물품을 받아볼 수 있다. 이에 따라 점점 더 시장이 확대되고 있으며, 기업들이 주목하고 있는 분야이다. 새벽배송 서비스는 마켓컬리를 선두로 국내 여러 대기업, 중소기업, 스타트업 기업들에게 전파되어 활성화되기 시작했다. 초창기 새벽배송 시장 규모는 100억이었지만 2020년에는 1조 5,000억으로 초기 대비 150배 성장하였다. 현대글로비스 종합 물류연구소의 자료에 따르면 2020년 라스트 마일 시장 규모는 7.5조 원 이었으며, 그 중 새벽배송 서비스는 20%를 차지하고 있다[14].

이처럼 새벽배송 서비스는 이커머스 시장 규모 증가로 인한 옴니채널화 및 라스트마일 서비스 다양화를 발판 삼아 짧은 기간 동안 급성장을 이루었다. 그러나 시장이 성장한 규모에 비해 새벽배송 시장은 아직 정확한 분석이 나오지 않았기에 본 논문에서는 사용자 리뷰를 통해 새벽배송의 현황을 분석하고자 한다.

리뷰 분석에 대한 중요성은 Josh at el.의 논문에서 파악할 수 있다. 제품에 대한 경험이나 지식에 대한 의견이라 할 수 있는 고객리뷰(Customer Review)들은 소비자들이 상품을 구매할 때 많은 영향을 미칠

뿐만 아니라 기업들에도 새로운 마케팅전략을 수립하는데 중요한 자료로 써 활용될 수 있다[1]. 따라서 온라인 애플리케이션 소비자 리뷰를 바탕으로 서비스 사용 만족도와 그에 따른 핵심 키워드를 확인한다면 새벽배송 서비스가 나아가야 할 방향을 제시할 수 있다.

본 연구에서는 온라인상에 존재하는 새벽배송 서비스 리뷰를 기반으로 소비자들이 새벽배송 사용 시 만족 요인 4가지로 분류하고 데이터 수집, 데이터 전처리, 분석, 시각화한 후 로지스틱 회귀 분석을 통해 결과를 도출하고자 한다. 온라인 새벽배송 애플리케이션 배포 플랫폼에서 작성된 만개 이상의 소비자 리뷰를 기반으로 새벽배송 주요 서비스 요인이 무엇인지 핵심 키워드를 통해 파악하고 점수화해, 서비스 품질 개선에 관한 확인이 가능하다. 보완해야 할 서비스를 알 수 있어 새벽배송 서비스 이해에 활용될 수 있을 것이다.

1.2 연구 범위 및 방법

연구를 위해 구글 앱 스토어, 애플 앱 스토어에서 마켓컬리 애플리케이션 사용 리뷰 수집을 진행하였다. 데이터 수집 기간은 2016년 2월 20일부터 2021년 4월 16일까지이며 총 10,390개의 데이터를 수집하였다.

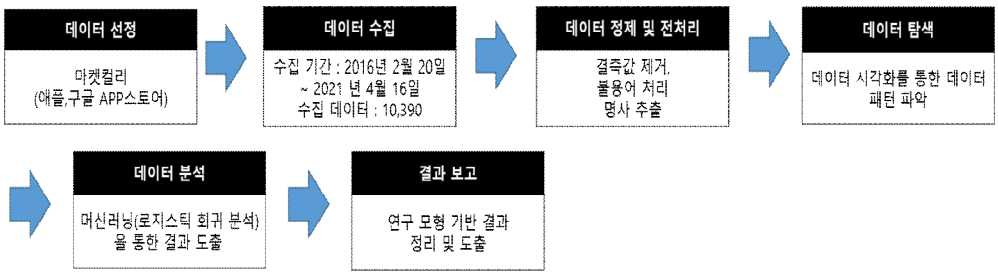
수집 및 분석 프로그램으로는 파이선(Python)을 사용하였다. 수집 라이브러리는 Selenium과 Beautiful Soup을 사용하였으며, 분석 라이브러리는 Pandas, Matplotlib, Konlpy, Soynlp, Sklearn를 사용하여 연구를 수행하였다. 수집한 데이터는 비정형 텍스트 데이터로 text cleaning을 통해 텍스트 데이터를 정제할 필요가 있어, 불용어 사전을 제작하여 불용

어를 제거하였다. 그 후, Soynlp, Konlpy을 이용해 명사 단어들만 추출하여 컴퓨터가 이해 할 수 있게 One-Hot-Encoding으로 변환시켰다.

[표 1-1] 수집, 분석 프로그램 및 라이브러리

파이썬					
수집		분석			
브라우저 자동화	HTML, XML 문서 파싱	행렬화	시각화	명사, 형용사 추출	머신러닝 활용
Selenium	BeautifulSoup	Pandas	Matplotlib	Konlpy, Soynlp	Sklearn

소비자 리뷰 긍정, 부정 분류는 소비자들이 매긴 평점으로 진행하였다. 평점은 소비자가 서비스를 사용했을 때 느낀 감정을 숫자 1부터 5 사이로 표현하는 기능이다. 따라서 긍정, 부정 분류에서 label(종속변수)로 사용할 수 있다고 판단하여, 전체 데이터에서 학습데이터를 70%, 테스트 데이터를 30%로 나누어 분류 모형 중 하나인 로지스틱 회귀 분류로 학습시켰다. 분류된 단어들을 설계한 연구모형에 맞게 분류하고 단어별로 결과에 미치는 영향도를 합쳐 연구 모형별 랭킹화 시켰다.



[그림 1-1] 연구 방법 및 진행 순서

제 2 장 이론적 배경 및 선행 연구

2.1 새벽배송 서비스 이론

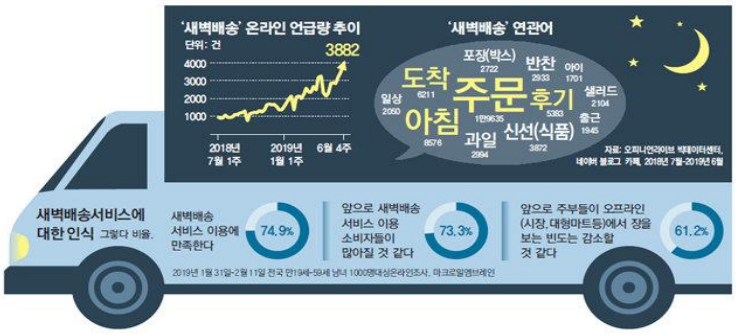
2.1.1 새벽배송 서비스의 개념

기업별로 선보이는 새벽배송 서비스는 다양하지만, 공통으로 오후 10에서 12시 전까지 원하는 제품을 주문하면 다음 날 새벽 7시 전에 주문한 제품을 받을 수 있다. 현재 1인 가구, 맞벌이 가족이 증가하는 추세인데, 이러한 가구 형태의 구성원들은 오프라인으로 장을 보고, 식사를 준비하는데 시간이 모자란 경향이 있다. 하지만 새벽배송 서비스를 사용할 경우, 애플리케이션에 접속해 원하는 식품을 결제하면 새벽에 신선한 상태로 제품이 오기 때문에 물품 구매에 사용하는 번거로움과 시간 낭비가 줄어든다. 즉, 새벽배송은 현재 가구 형태와 라이프 스타일 트렌드에서 불편함을 해결해줄 수 있기에 COVID-19와 맞물려 점점 더 시장이 증가하는 추세라고 볼 수 있다.

오피니언 라이브 빅데이터센터에서 2018년 7월부터 2019년 6월까지 네이버 블로그, 카페에서 새벽배송 관련 키워드를 수집한 결과 상위에 '주문', '아침', '도착'이 나온다. 세 단어는 새벽배송을 핵심적으로 설명하는 단어이다. 상품 관련으로는 '신선(식품)', '과일', '이유식', '우유', '샐러드', '반찬', '도시락', '야채', '유기농', '달걀' 등의 상품이 함께 거론되고 있다.

키워드를 기반으로 분석하자면, 새벽에 배송받아 아침 식사로 먹을 수 있는 음식이 새벽배송의 주력 물품이라는 것을 알 수 있다[15]. 이렇듯 새벽배송 서비스는 주로 신선식품을 다루고 다른 배송 서비스보다 더 늦

게 주문을 해도 더 일찍 도착한다는 차별성을 가지고 있다.



[그림 2-1] 새벽배송 키워드

그러나 새벽배송 모든 과정에서의 만족도를 분석하기 위해 본 연구에서는 단순히 배송에 한정하지 않는다. 애플리케이션 접속부터 제품을 받는 과정을 포함해 데이터 분석을 진행했다.

2.1.2 새벽배송 서비스 현황

새벽배송이 나오기 전까지는 소기업이 새벽에 아침밥, 이유식, 와이셔츠 등 소량 제품을 직접 갖다주는 시장이 블루 오션이었다. 하지만 마켓컬리의 샛별배송을 시작으로 새벽 배송 분야가 라스트마일 시장에서 성장하기 시작했다. 처음 마켓컬리는 새벽 배송을 시작하며 '강남 엄마들의 장보기 필수 애플리케이션'이라는 이미지를 형성하였다. 그 후, 새벽배송이 안정적으로 자리 잡으면서 마켓컬리는 2018년에 월 매출 100억 원을 달성하였다. 이와 같은 마켓컬리의 급성장에 국내 유통 대기업들과 식품 기업들도 새벽배송 서비스를 도입하기 시작하였다.

2017년 GS리테일은 GS프레시를, 한국야쿠르트는 가정간편식 브랜드 잇츠온을 도입하는 등 대기업에서 차별성 있는 새벽배송 서비스를 만들었다. 또한, 새벽배송 관련 스타트업도 등장하면서 2015년 100억이었던 새벽배송 시장은 2021년 1조 5,000억으로 급속 성장하였다[16].

[표 2-1] 대기업, 중견기업 새벽배송 서비스

	서비스명	서비스 시작	배송지역	무료배송 기준	주문마감시간	특징
이마트 SSG 닷컴	쓱새벽배송	2016년 6월	서울 10개구	4만원 이상	자정	이마트, 노브랜드등의 1만여개 배송상품
현대백화점	현대 e슈퍼마켓 새벽배송	2019년 7월	서울, 경기, 인천(일부지역)	5만원 이상	오후 5시	수입치즈, 수입소스 등 고급 식자재 판매
롯데프레시	롯데 프레시 새벽배송	2018년 2월	서울, 경기, 광주, 부산, 천안, 청주 등 일부지역	3만원 이상	오후 10시	롯데프레시센터 거점지역 중심으로 운영
쿠팡	로켓프레시	2018년 10월	전국	1만 5천원 이상	자정	식품 외 200만 개의 상품구성
헬로네이처	새벽배송	2018년 6월	서울, 경기, 인천 (일부지역)	4만원이상	자정(지역별 마감시간 상이)	중간 유통단계 생략, 산지배송도 하루
GS프레시	GS 프레시 새벽배송	2017년 7월	서울, 경기 일부	3만원 이상	오후 11시	GS25 편의점 상품도 구매
동원 F&B	밴드프레시	2018년 2월	서울, 경기, 인천 일부	3만원 이상	오후 5시	동원몰 기반 서비스

[표 2-2] 스타트업 새벽배송 서비스

브랜드	취급 제품, 특징	배송
마켓컬리	신선식품, 유명 베이커리 제품, 가공식품 등 제품 7000종	오후 11시 주문 -> 다음 날 오전 7시 전 도착
정육각	도축 4일 이내 돼지고기, 당일 도계 닭고기, 당일 산란 달걀 , 당일 착유 우유	주문 1시간 이내 도착, 오전 10~ 밤 10시까지 배송
오늘회	회, 고급 수산물 등 200종	오후 3시 이전 주문 -> 당일 오후 7시전 도착
삼삼해물	생물, 내동 등 해산물 200종	자사몰, 마켓컬리, 이마트, 쿠팡 등 판매 채널 이용
나물투데이	시금치, 참나물, 방풍 나물 등 데친 나물 80종	오전 9시 이전 주문 -> 다음 날 도착
팀프레시	신생 이커머스 업체에 유통망 제공	새벽배송, 냉장배송 전문

2019년, 2020년 엠브레인에서 새벽배송 서비스 인식 조사를 진행하였다. 조사 결과를 보자면, 1년 사이 새벽배송 서비스를 인지하는 소비자 비율

은 72.7%에서 95.7%로 증가하였다. 이 중 실제 새벽배송을 경험한 소비자는 53.1에서 70%로 변화하였다[17]. 그 밖에도 새벽배송의 다양한 부가서비스들이 등장하면서 시장의 지속적인 성장이 예상된다.

2.2 관련 선행연구

2.2.1 새벽배송 관련 선행연구

김민정[2]의 연구에 따르면 새벽배송 서비스의 이용 지속성은 소비자의 피해 경험과 감정 경험, 서비스 만족도에 따라 결정된다. 약 12%의 소비자가 새벽배송 서비스 중 상품 훼손을 경험하였으며, 문제를 제기한 약 69%의 소비자 중 절반 정도만 만족스러운 해결이 되었다. 또한, 부정적인 감정 경험과 긍정적인 감정 경험이 이용 지속성에 영향을 미치며, 그 중에서도 긍정적인 감정 경험이 더 큰 영향력을 나타냈다. 그러므로 소비자의 새벽배송 사용에 대한 감정 경험을 정확하게 파악하는 것이 중요하다는 결론을 도출할 수 있다.

윤덕환[3]의 연구인 새벽배송 서비스 관련 인식 조사에서 새벽배송 서비스인지도 는 72.7%로 매우 높게 나타났으며 실제 이용 경험도 절반 이상이었다. 주요 이용 서비스 순위는 마켓 컬리, 쿠팡, 이마트 순으로 나타났다. 새벽배송 서비스의 만족도 및 만족 이유로는 신속도 와 상품의 신선도로 드러났으며, 여성과 50대 고연령층의 만족도가 가장 높게 평가되었다. 위 조사를 참고하였을 때, 배송의 신속성과 상품의 신선도를 중점으로 새벽배송에 관한 소비자 리뷰의 텍스트 분석이 필요하다는 사실을 알 수 있다.

추진기[4]의 연구에 따르면 새벽배송은 온라인과 오프라인 물류가 융합한 새로운 판매 형태에 인공지능, 인터넷, 빅데이터 등의 기술을 이용하여 제품 생산에서 판매까지 전반적 과정을 업그레이드한 신유통의 한 흐름을 반영하고 있다. 이런 신유통에 조금 더 접근할 필요가 있는 요소는 고객과의 커뮤니케이션 요소라고 연구에서 밝히고 있는데, 이러한 커뮤니케이션 요소의 일부로 판단할 수 있는 것이 소비자 리뷰이다. 따라서 소비자 리뷰를 분석할 경우 새벽배송의 브랜드 아이덴티티 방향성을 찾아낼 수 있다고 사료된다.

정지희[5]의 연구에서 새벽배송 기업의 경쟁력은 물류서비스의 품질에 따라 결정된다. 식품 구매가 급증하면서 활성화 된 새벽배송은 '속도'와 '안전'의 두 가지 형태를 중점으로 배송이 이루어지고 있다. 그 중에서도 신속성과 정확성에 대한 품질 수준은 평준화되면서 안전배송, 친절배송의 서비스 속성을 고객들이 중점으로 판단하고 있다. 전자상거래 플랫폼 물류의 서비스 품질 요인은 배송, 반품, 사후, 고객대응의 4개 차원으로 구성되어 있으며, 이러한 다각도의 차원에서 고객의 반응과 서비스의 방향성에 대한 대응이 필요하다. 따라서 본 연구에서는 배송에만 한정하지 않고 넓은 범위로 확장하여 고객의 만족도에 대한 키워드 분석을 실시하고자 한다.

2.2.2 로지스틱 회귀 관련 선행연구

데이터 마이닝 관련 연구는 소셜 네트워크 서비스의 확산으로 인해 다양한 방향으로 진행되고 있다. 특히, 다양한 이용자들의 빅데이터가 비정형 데이터인 텍스트 형식으로 활발하게 수집되고 있다. 리뷰 데이터 역

시 비정형 데이터로 분석을 위해서는 text cleaning 등의 전처리 과정이 필요하지만, 유의미한 정보를 도출해낼 수 있기에 다양한 분야에서 연구가 진행되고 있다. 따라서 본 연구에서도 고객 리뷰 빅데이터를 기반으로 데이터 마이닝과 로지스틱 회귀를 사용하고자 한다. 이와 관련된 상세내용은 '3장 연구 방법 및 분석'에서 설명하도록 한다.

윤지선[6]의 연구에 따르면 로지스틱 회귀분석은 예측 민감도와 분류정확도가 높아 노인의 우울 예측모형을 구축하는 데 유용하게 사용될 수 있다. 소비자 리뷰 분석에서도 분류 정확도가 중요하므로 로지스틱 회귀분석이 유용하게 사용될 수 있다고 판단된다.

최자영[7]의 연구에서 리뷰를 제목과 내용으로 구분해 분석하는 방식을 통해 영향력을 살펴보았으며, 해당 분석에 회귀분석을 사용하였다. 회귀분석을 통해 리뷰가 매출에 미치는 영향력을 분석할 수 있었으며, 각 변수의 영향력을 명확하게 파악하기 위해 단계적으로 실시하는 방법을 사용하였다. 따라서 회귀분석을 통해 리뷰의 영향력을 정확하게 판단할 수 있다고 사료된다.

2.2.3 본 연구의 차별성

선행 연구를 살펴보았을 때, 사용자들의 새벽배송 사용 현황과 미래 사용 만족에 대해 예측하기 위해서는 만족도에 대한 파악이 필요하다. 만족도 파악을 위해서는 특히 소비자 리뷰를 통한 핵심 키워드 파악과 부정점수, 긍정점수 산출이 진행되어야 한다.

그러나 이전의 연구들에서는 애플리케이션 사용 리뷰 데이터를 바탕으로 긍·부정 점수 산출을 진행한 바가 없다. 현재 대다수의 사용자는 애플리케이션을 통해 배송을 주문하고, 수령하고 있기 때문에 사용 플랫폼으로써 배송 애플리케이션의 중요도는 상당하다. 그러므로 구글 플레이스토어와 애플 앱 스토어의 새벽배송 애플리케이션 사용자 리뷰를 바탕으로 분석을 진행하여, 소비자 만족도를 측정하는 차별성을 가지는 데에 본 논문의 의의가 있다.

제 3 장 연구 방법 및 분석

3.1 자료의 구성

3.1.1 연구모형

새벽배송 서비스를 작은 의미로 봤을 때, 고객이 주문한 제품을 새벽 7시 전까지 배송한다는 의미로 볼 수 있다. 하지만 큰 의미로 접근하자면, 새벽배송 서비스 이용 시 구매할 수 있는 제품과 배송 서비스가 기존과 다르기 때문에 고객이 애플리케이션 혹은 온라인으로 접속해 물건을 주문하고 새벽 7시 전에 받는 전 과정을 의미한다고 볼 수도 있다.

[그림 3-1] 새벽배송 서비스 프로세스

따라서 본 연구에서는 큰 의미의 새벽배송 서비스를 기반으로 결제까지 프로세스를 애플리케이션 사용 요인과 배송 프로세스를 배송요인, 홍보활동을 마케팅 요인, 제품을 제품요인 나누어 연구모형을 설계하였다. 각

요인의 의미는 다음과 같다.

첫째, 애플리케이션 사용요인은 애플리케이션 UI, UX 편리성 관련 요인이다. 리뷰 데이터에서 편리, 간편 등의 단어들을 군집한다.

둘째, 배송요인은 마켓컬리에서 고객의 결제 제품을 확인하고 고객이 주문한 제품을 정확하게 포장해 새벽 7시까지 고객이 입력한 주소에 전달하는 과정을 의미하며 배송요인에는 시간, 포장, 정확도가 포함되어 있다.

시간은 고객이 주문한 제품이 새벽 7시 전에 입력한 주소에 도착했는지를 의미하며 빠름, 당일 등의 단어들을 군집한다.

포장은 제품의 포장상태를 의미하며 포장, 재질 등의 단어들을 군집한다.

정확도는 고객이 주문한 제품을 정확하게 받았는지를 의미하며 정확, 맞게 등의 단어들을 군집한다.

셋째, 마케팅요인은 마켓컬리에서 진행하고 있는 행사 및 홍보 요인이다. 리뷰데이터에서 행사, 할인, 쿠폰 등의 단어가 나왔을 때 군집 된다.

넷째, 제품요인은 마켓컬리에서 판매하는 제품과 관련된 요인으로 제품요인에는 다양성, 제품 품질이 포함되어있다.

다양성은 제품의 종류를 의미하며 많다, 다양 등의 단어들을 군집한다.

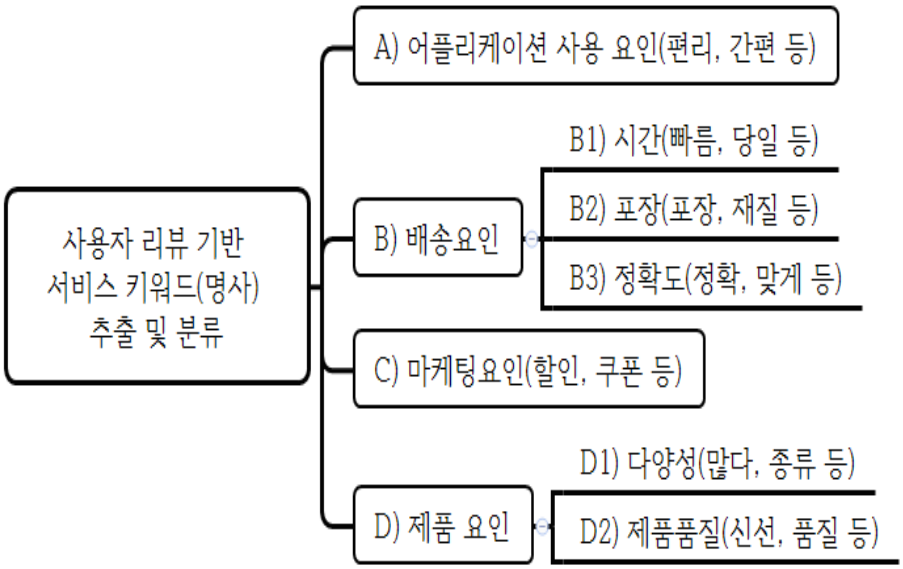
제품 품질은 제품의 상태를 의미하며 선선, 품질 등의 단어들을 군집한다.

분석 결과를 구체적으로 나타내기 위해 오즈비(Odds Ratio)를 사용하도록 한다. 오즈(Odds)는 처리그룹에서 사건이 발생할 확률이 발생하지 않을 확률의 몇 배가 되는지 나타내는 값이다. 오즈비(Odds Ratio)는 특정한 처리를 한 그룹과 아닌 그룹의 오즈 비율이다. 오즈비는 연구하고자 하는 변수를 범주화하여 각 그룹에 대한 특성을 설명하는데 용이하다.

따라서 [표 3-1]과 같이 단어라는 변수들을 범주화하여 긍정 그룹과 부정 그룹에 대한 특성을 설명하도록 한다[8].

[표 3-1] 예상 오즈비

	단어 1	단어 2	단어3	…	단어 N-1	단어 N
긍정 1	긍정값1	긍정값2	긍정값3	…	긍정값N-1	긍정값N
부정 0	부정값1	부정값2	부정값3	…	부정값N-1	부정값N



[그림 3-2] 연구 모형

3.1.2 데이터 선정

구글 앱 스토어, 애플 앱 스토어는 아이폰, 안드로이드 스마트폰 사용자가 필요한 애플리케이션을 설치 시 사용해야 하는 필수적인 애플리케이션이다. 태블릿, 스마트폰 등 애플리케이션 사용 기기가 증가하면서 애플리케이션 플랫폼 스토어는 가파르게 성장하였다.

한국모바일산업연합회에 따르면 2019년 구글 플레이스토어(구글 앱 스토어)와 애플 앱 스토어 매출은 각각 5조 9천억 원, 2조 3천억 원으로 작년보다 11.1%, 9.6% 증가하였다[18].

앱마켓별 매출액 현황(단위: 억 원)

구분	2018년				2019년(P)			
	구글 플레이	애플 앱스토어	원 스토어	기타	구글 플레이	애플 앱스토어	원 스토어	기타
매출액 (커머스 제외)	53,999	21,062	9,403	1,144	59,996	23,086	10,561	932
증가율	10.6%	8.7%	9.3%	5.2%	11.1%	9.6%	12.3%	-18.5%
비중	63.1%	24.6%	11.0%	1.3%	63.4%	24.4%	11.2%	1.0%

주) 기타 앱스토어는 독자적인 앱스토어(삼성전자·LG전자 앱마켓, 해외 각국 로컬마켓, 중국의 경우 텐센트 마이앱, 360 모바일 어시스턴트, 바이두 모바일 어시스턴트 등)

[그림 3-3] 국내 앱 마켓별 매출액

이처럼 1인 1 스마트폰 보유 시대인 현재 구글 앱 스토어와 애플 앱 스토어 사용은 점점 증가할 것이고, 앱 스토어에 애플리케이션 리뷰를 남기는 행위도 증가할 것으로 전망된다. 또한, 애플리케이션 내부에서도 리뷰와 평점을 남기도록 유도하는 경우가 많음으로 자주 사용하는 고객들의 솔직한 리뷰 또는 애플리케이션이나 서비스에 대한 불편으로 남기는 리뷰 등 다양한 방면의 리뷰가 존재하게 된다.

온라인 고객 리뷰 데이터는 고객의 소리를 획득할 수 있는 주요 채널 중 하나로 서비스 품질 평가를 위하여 유용하게 활용될 수 있다[9]. 인터넷 쇼핑 플랫폼에서는 각 제품에 대해 애플리케이션 내부에서 직접 리뷰를 남기는 기능을 탑재하여, 해당 제품에 대한 리뷰를 통해 정보를 얻기도 한다. 대부분의 애플리케이션에서는 앱 스토어에 고객의 리뷰 작성을 유도하는 기능을 탑재하고 있는데, 그만큼 애플리케이션 전반적인 사용성과 세부 서비스에 대한 리뷰가 동시에 기재될 수 있는 것이 앱 스토어 리뷰로 중요하기 때문이다.

즉, 사용자가 해당 애플리케이션을 사용 후 남긴 리뷰는 애플리케이션의 서비스 품질을 평가하는 데 도움이 되는 데이터이다. 그러므로 사용자가 원하는 애플리케이션을 다운받을 수 있는 앱 스토어 리뷰 데이터가 본 연구에 적합한 데이터라고 판단하였으며, 해당 데이터 수집을 진행하였다.

3.1.3 데이터 수집 및 EDA

3.1.3.1 데이터 크롤링

본 연구에서는 마켓컬리 리뷰를 중심으로 연구를 진행했다. 마켓컬리는 새벽배송 서비스 선도주자이며 친환경 포장재, 다양한 서비스를 제공하며 소비자, 기업들에 많은 주목을 받았다. 따라서 마켓컬리 리뷰 분석을 통해 다양한 서비스 만족 요인을 파악할 수 있다고 판단하였다.

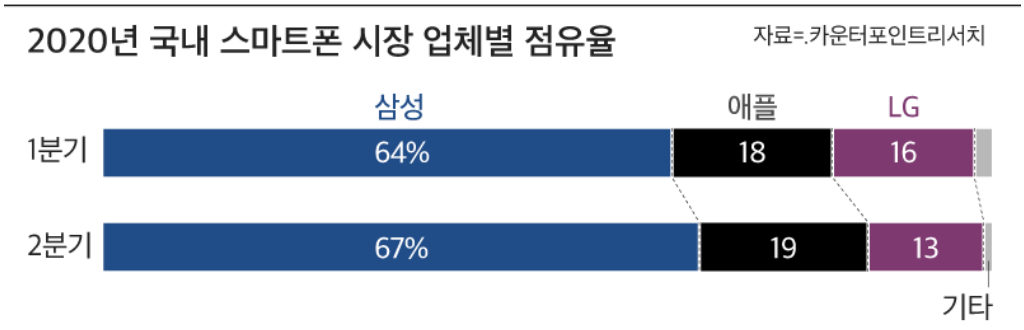
앞서 데이터의 수집 플랫폼은 구글 앱스토어, 애플 앱스토어이다. 두 앱스토어에서 파이선 라이브러리를 사용하여 마켓컬리 애플리케이션 사용자 리뷰를 수집하였다. 구글 앱 스토어에서는 2016년 3월 1일부터 2021년 4월 16일까지의 리뷰를 크롤링하여 7,395개 데이터를 수집하였고, 애플 앱 스토어에서는 2016년 2월 20일부터 2021년 4월 15일까지의 리뷰를 크롤링하여 2,995개 데이터를 수집하였다. 두 개의 앱 스토어의 리뷰는 약 5년 정도의 분량이며 총 10,390개이다.

[표 3-2] 마켓컬리(구글, 애플 앱스토어) 리뷰 수집 기간 및 리뷰 수

	마켓컬리 구글 앱 스토어	마켓컬리 애플 앱 스토어
기간	2016년 3월 1일 ~ 2021년 4월16일	2016년 2월 20일 ~ 2021년 4월 15일
리뷰(개)	7,395	2,995

데이터 수를 봤을 때 구글 앱 스토어 리뷰 수가 애플 앱 스토어 리뷰 수보다 약 2배 더 많다. 원인은 국내 스마트폰 점유율에 있다고 추측된다. 국내 사용자 중 다수는 구글 운영체제 기반의 스마트폰을 사용하는데 이는 스마트폰 점유율 조사 결과를 보면 알 수 있다.

카운터포인트리서치에 따르면 2020년 1분기 안드로이드폰(삼성, LG) 점유율은 80% 정도이고, 애플 점유율은 18%였다. 2분기 때는 안드로이드폰(삼성, LG) 80%, 애플 점유율은 19%이다[19]. 즉 국내 대다수 스마트폰 사용자가 안드로이드 폰을 사용하기 때문에 점유율과 같은 비율로 두 가지 앱 스토어의 리뷰 비율이 나왔다고 판단된다.

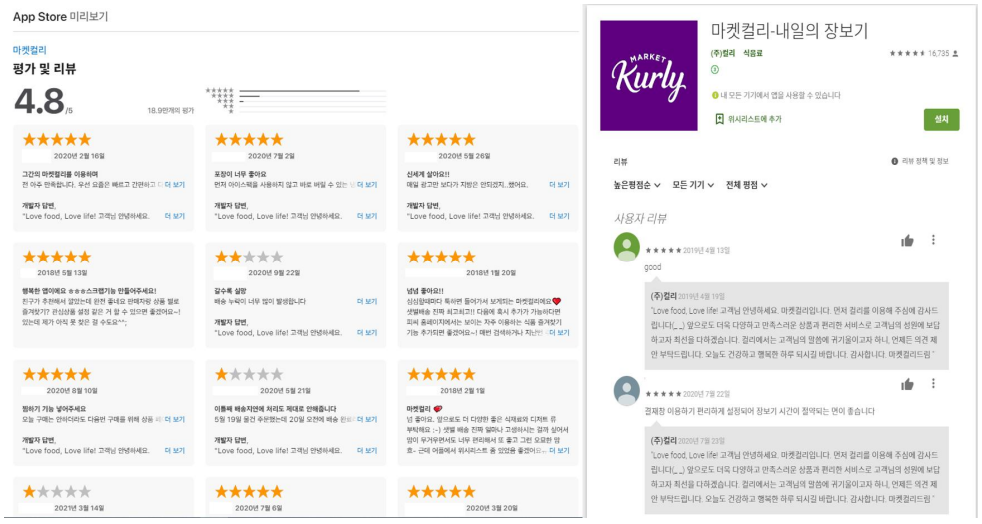


[그림 3-4] 스마트폰 점유율

데이터 수집 프로그램으로는 오픈 소스인 파이선(Python)을 사용하였다.

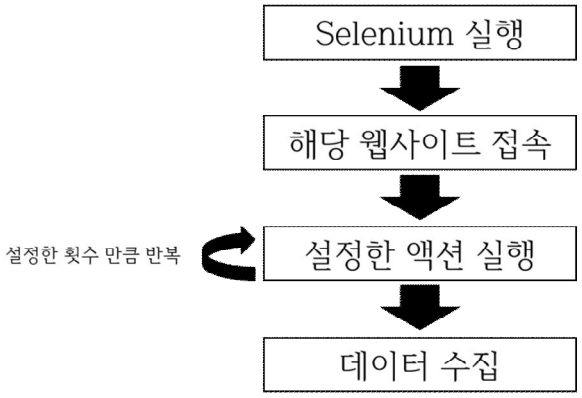
특히 구글, 애플 앱 스토어 페이지는 동적 웹페이지이므로 Selenium을 활용하여 웹 크롤링을 진행하였다. Selenium은 웹앱을 테스트하는 웹 프레임워크로 Web driver의 API를 통해 브라우저를 제어하기 때문에 동적 웹페이지의 데이터를 크롤링할 때 유용하게 사용되는 스크래핑 라이브러리이다.

동적 웹페이지는 사용자에 액션에 의해 페이지에 변화가 일어나는 페이지로 구글, 애플 앱 스토어도 전체 리뷰를 보기 위해서는 하단으로 마우스를 드래그해야 한다. 이를 자동으로 수행할 수 있는 라이브러리가 바로 Selenium이다.



[그림 3-5] 애플(좌), 구글(우) 앱 스토어 웹 사이트

Selenium으로 브라우저를 제어해 구글, 아이폰 앱스토어 전체 리뷰를 표시하고 그다음 데이터 수집을 진행하는 과정을 통해 데이터를 수집하였다.



[그림 3-6] 데이터 수집 프로세스

3.1.3.2 탐색적 데이터 분석(Exploratory Data Analysis, EDA)

탐색적 데이터 분석은 raw 데이터를 이해하기 위해 다양한 각도로 데이터를 관찰하는 행위이다. 주로 데이터를 분석하기 전, 그래프와 통계적인 방법으로 자료를 이해하는 과정이다. 탐색적 데이터 분석이 필요한 이유는 데이터의 값과 분포를 다양한 시각으로 보면서 전체 데이터의 양상과 보이지 않던 문제점을 발견 혹은 이해할 수 있도록 해주기 때문이다.

즉, EDA 과정에서 데이터 수집 후 미처 파악하지 못한 문제들을 발견할 수 있다. 또한, 데이터를 다양한 관점으로 관찰하면 기존의 가설을 수정하거나 새로운 가설을 만들 수 있다.

그렇기에 구글 앱 스토어, 애플 앱 스토어 기준으로 년 도별 리뷰 수, 평점 변화 단어 출현 수 등을 다양한 관점으로 수집한 데이터를 분석하였다.

구글 앱 스토어에서 데이터 수집 데이터의 양식은 name(작성자 아이디), rating(평점), date(작성 날짜), helpful(작성자 외 다른 고객이 내용에 만족하였을 때 누른 횟수), comment(작성자 리뷰 내용), developer_comment(관리자 댓글)로 구성하여 수집하였다. 애플 앱 스토어에서의 양식은 name(작성자 아이디), rating(평점), date(작성 날짜), title(제목), comment(작성자 내용)로 구성되어 있다.

필요하다고 판단되는 컬럼은 date, comment, ratings로 나머지 불필요한 컬럼은 제거하였다.

	date	comment	ratings
0	2021-04-16	너무편리해요	5
1	2021-04-16	깔끔한 앱	5
2	2021-04-16	가격대별로 볼수있는 기능있음 좋겠습니다 딱 5000원만 딱3000원 더 올르면 더있...	1
3	2021-04-16	빠른배송감사합니다	5
4	2021-04-16	품품다양 그룹리티 참르곳 배송곳 서비스곳	5
...	...	...	...
7390	2016-03-02	아이를 키우며 좋은 식자재에 대한 갈증이 있는데 이렇게 건강하고 신선한 먹거리를 ...	5
7391	2016-03-02	아는동생소개로알게된마켓컬리 늦게안거아쉬울정도ㅠㅠ제품퀄리티완전좋고,포장배송완벽합니는...	5
7392	2016-03-02	샛별배송 정말 좋은듭니다. 컬리라면 무조건 믿고 주문~ ^^	5
7393	2016-03-01	마이상 빔늦게 슈퍼를 헤매고 다니지않아도됨	5
7394	2016-03-01	Good	5

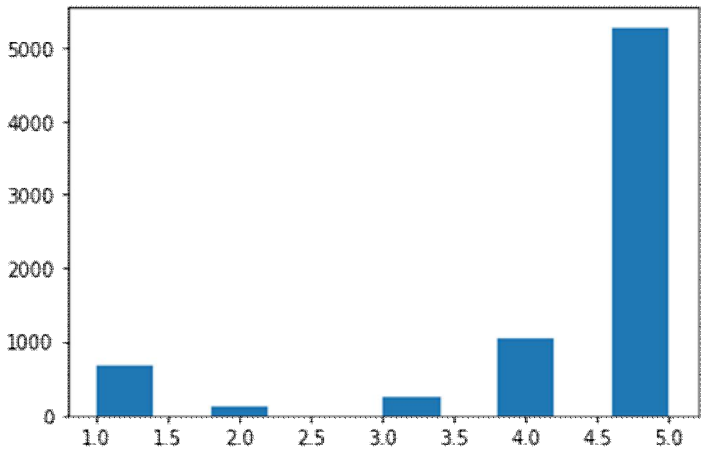
7395 rows × 3 columns

	date	comment	ratings
0	2020-02-16	전 아주 만족합니다. 우선 요즘은 빠르고 간편하고 다양한 음식을 접하려고 하잖아요....	5
1	2020-07-02	먼저 아이스팩을 사용하지 않고 바로 버릴 수 있는 냉각을 사용하는 것과 종이 테이프...	5
2	2020-05-26	매일 광고만 보다가 지방은 안되겠지..됐어요.	5
3	2018-05-13	친구가 추천해서 같았는데 완전 좋네요 판매자랑 상품 별로 물어봤기? 관심상품 설정 ...	5
4	2020-09-22	배송 누락이 너무 많이 발생합니다	2
...	...	...	...
2990	2016-03-02	매일 출근하면 컬리 사이트 보는게 일과였던 요즘~ㅠㅠ 앱 출시로 더 가까이 접하게...	5
2991	2016-03-02	시원찮아기 있을듯은 하나,	1
2992	2016-03-02	드뎌 컬리업이 나왔네요	5
2993	2016-03-02	처음 인정했을 때가 신혼 초였는데요 7개월 된 아기 엄마가 된 지금까지 잘새 방앗...	5
2994	2021-04-15	밤에 시켜서 아침에 받아볼 수 있어서 너무 좋습니다.	5

2995 rows × 3 columns

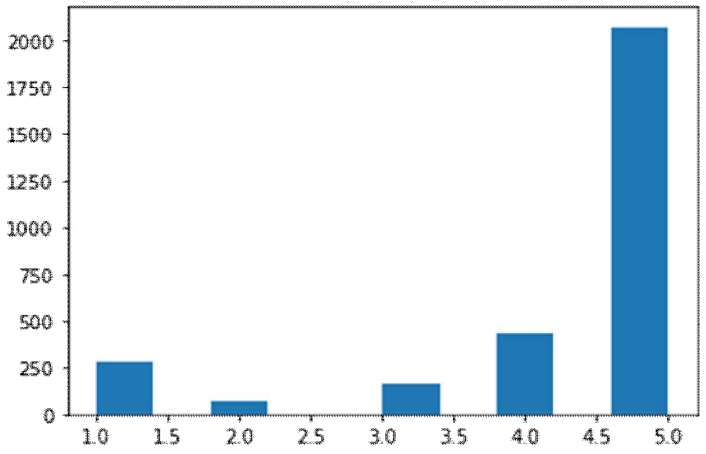
[그림 3-7] 구글(좌), 애플(우) 불필요 컬럼 제거 후 형태

[그림 3-8]과 [그림 3-9]를 보았을 때, 리뷰 데이터 평점의 평균 및 편차를 확인할 수 있다. 구글 앱 스토어의 리뷰 평점은 평균 4.37, 편차 1.23 이며, 애플 앱 스토어의 리뷰 평점은 평균 4.32. 편차 1.25 이다. 이를 통해 구글 앱 스토어의 리뷰와 애플 앱 스토어의 리뷰 평균은 0.05차이, 편차는 0.02차로 두 데이터의 분포가 비슷하다는 것을 확인할 수 있다.



ratings	
count	7395.00
mean	4.37
std	1.23
min	1.00
25%	4.00
50%	5.00
75%	5.00
max	5.00

[그림 3-8] 구글 평점 기술통계 및 히스토그램



ratings	
count	2995.00
mean	4.32
std	1.25
min	1.00
25%	4.00
50%	5.00
75%	5.00
max	5.00

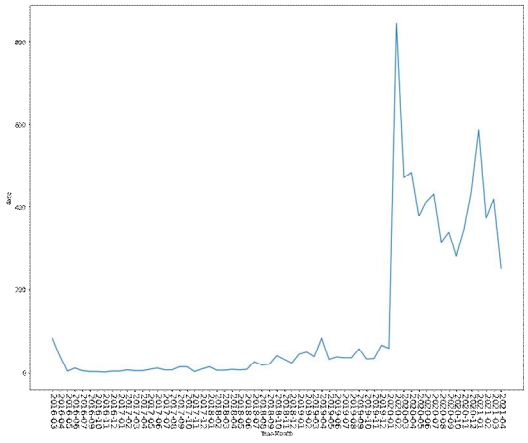
[그림 3-9] 애플 평점 기술통계 및 히스토그램

다음은 두 데이터의 사분위(Inter quartile range)에 대한 파악 내용이다. 사분위는 전체 데이터의 분포를 나타내는 것으로 25%, 50%, 75%에 있는 값을 확인한다.

구글 앱 스토어와 애플 앱 스토어의 리뷰 데이터 사분위가 비슷하게 나

타나고 있다. 특히 25%부터 4점이 나오는걸 보아 전체 데이터에서 3점 이하의 평점보다 4점 이상의 평점 데이터가 더 많이 있다는 것을 확인할 수 있다. 즉 대다수의 소비자가 마켓컬리 서비스에 대한 사용 만족도가 높으나, 불편함을 느꼈을 때는 1점을 주는 극단적인 경우가 많다는 것을 알 수 있다.

다음은 구글, 애플 앱 스토어의 월별 리뷰 수 변화 추세를 그래프 및 표로 표현하였다.



2016											
1	2	3	4	5	6	7	8	9	10	11	12
		80	39	3	10	4	0	2	2	1	3

2017											
1	2	3	4	5	6	7	8	9	10	11	12
3	6	4	4	7	10	6	6	13	13	2	8

2018											
1	2	3	4	5	6	7	8	9	10	11	12
13	5	5	7	6	7	25	17	19	40	31	22

2019											
1	2	3	4	5	6	7	8	9	10	11	12
44	49	38	81	31	37	35	35	57	32	33	65

2020											
1	2	3	4	5	6	7	8	9	10	11	12
58	847	472	484	378	414	431	315	339	283	344	436

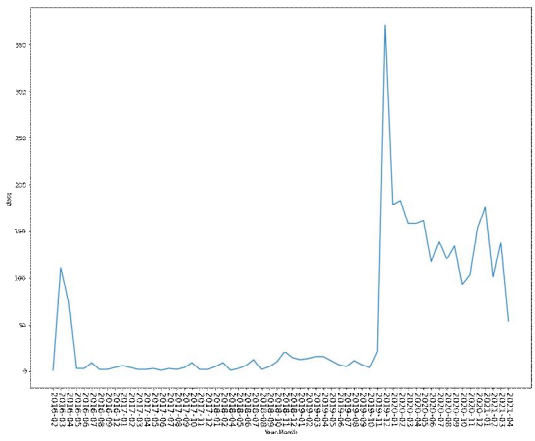
2021											
1	2	3	4	5	6	7	8	9	10	11	12
588	373	420	253								

[그림 3-10] 구글 앱 스토어 월별 리뷰수 변화 및 시각화

구글 앱 스토어 월별 리뷰 수를 확인한 결과는 다음과 같다. 2020년 1월까지 100개 이하의 리뷰가 올라왔지만, 2020년 2월부터 847개로 전월 대비 1,360% 증가하였다는 사실을 알 수 있다.

리뷰 수의 가파른 증가 원인으로는 2월에 발생한 COVID-19 1차 유행을 예상할 수 있다. COVID-19 1차 유행으로 외출이 제한되면서 애플리

케이션으로 제품을 구입하는 신규 고객들이 대거 유입되었다. 이로 인해 마켓컬리 사용량이 증가하면서 리뷰 수가 폭발적으로 증가한 것으로 판단할 수 있다.



2016											
1	2	3	4	5	6	7	8	9	10	11	12
	1	111	75	3	3	9	2	2	9	9	4

2017											
1	2	3	4	5	6	7	8	9	10	11	12
6	4	2	2	3	1	3	2	4	9	2	2

2018											
1	2	3	4	5	6	7	8	9	10	11	12
5	3	0	1	3	6	12	2	5	10	20	14

2019											
1	2	3	4	5	6	7	8	9	10	11	12
12	13	15	15	11	7	5	11	7	4	20	371

2020											
1	2	3	4	5	6	7	8	9	10	11	12
179	183	158	158	161	117	139	120	134	93	103	153

2021											
1	2	3	4	5	6	7	8	9	10	11	12
176	101	138	54								

[그림 3-11] 애플 앱 스토어 월별 리뷰수 변화 및 시각화

애플 앱 스토어는 구글 앱 스토어와 조금 다른 양상을 보였다. 애플 앱 스토어는 2016년 3, 4월을 제외하고 2019년 11월까지 월별로 리뷰 개수가 20개 이하였다. 하지만 2019년 12월부터 371개로 전월 대비 1,755% 증가하였음을 볼 수 있다. 그 이후에는 월별 리뷰 수가 약 세 자릿수로 유지되고 있다.

2019년 12월에 리뷰 수가 폭발적으로 증가한 것은 2019년 11월 25일부터 12월 31일까지 마켓컬리에서 진행한 연말 기획 '컬리랜드' 행사가 원인이 된 것으로 판단된다.

서로 리뷰가 증가하기 시작한 월이 다르다는 것을 바탕으로 구글 앱 스토어와 애플 앱 스토어의 리뷰 수 증가 원인은 달랐을 것으로 예측할 수

있다. 애플 앱 스토어는 마켓컬리에서 2019년 11월 말부터 12월까지 진행한 행사로 인해 리뷰 수가 증가하였으며, 구글 앱 스토어는 2020년 2월 COVID-19 1차 유행이 시작되면서 리뷰 수가 급속도로 증가하였다고 판단된다.

각 스토어별 단어별 출현 빈도 및 평점을 확인하기 위해 간단한 전처리를 진행하였고, 전처리 방식에 대해서는 3.2.2 데이터 전처리 및 학습에서 설명하도록 한다.

구글 앱 스토어 단어별 출현 빈도 및 평점은 아래와 같다.

[표 3-3] 구글 앱 스토어 단어 출현 빈도 및 평점

words	satisfaction	count
배송	4.16	1726
상품	4.14	877
컬리	4.28	687
사용	4.34	571
제품	4.42	571
마켓	4.24	498
이용	4.28	422
새벽	4.40	365
포장	4.48	353
구매	3.91	316
쿠폰	4.00	310
자주	4.56	269
가격	3.99	265
정말	4.50	252
물건	4.02	249
결제	2.79	226
가입	1.83	211
품질	4.55	186
샛별	4.18	183
최고	4.88	182

구글 앱 스토어에서 나온 단어 중 출현 빈도수를 기준으로 상위 20개를 출력하였다. 대다수의 단어가 3.9 이상이었으며, 특히 배송이란 단어는 1,726개로 상품이란 단어보다 2배 이상 많고 평점도 4.16으로 높았다. 그 외에도 사용, 상품, 물건 등 여러 서비스도 4점 이상 받고 있었으며, 이 분석을 통해 고객들이 배송, 사용, 상품 등 마켓컬리의 전반적인 서비스를 만족하며 사용하고 있다는 것을 알 수 있다.

하지만 결제 2.79, 가입 1.83으로 두 단어는 다른 서비스에 비해 매우 낮은 평점을 받고 있었다. 이를 보아 마켓컬리의 결제와 가입 부분에서 안드로이드 사용 고객들이 불편함을 겪고 있다는 걸 알 수 있다.

애플 앱 스토어 단어 출현 빈도 및 평점은 아래와 같다.

[표 3-4] 애플 앱 스토어 단어 출현 빈도 및 평점

words	satisfaction	count
배송	4.19	670
컬리	4.51	573
마켓	4.49	443
상품	4.42	331
제품	4.43	248
이용	4.50	204
구매	4.11	196
포장	4.54	165
사용	4.25	152
새벽	4.61	143
정말	4.48	127
아침	4.71	118
자주	4.61	114
물건	4.21	108
쿠폰	4.21	102
항상	4.81	93
어플	3.44	93
재료	4.84	90
결제	3.42	89
가격	4.51	86

애플 앱 스토어 역시 구글 앱 스토어와 같은 방식으로 단어 출현 빈도

수 기준으로 상위 20개를 출력하였다.

애플 앱 스토어도 구글 앱 스토어와 비슷하게 출력되었지만, 최저 점수가 3.42로 구글 앱 스토어에 비해 높은 점수를 받았다. 특히 구글 앱 스토어에서 결제 단어는 2.79로 낮은 점수를 받았지만, 애플 앱스토어에서는 3.42로 0.63 더 높았다.

이 수치를 통해 아이폰에서 마켓컬리를 사용 하는 사용자들은 마켓컬리 서비스를 높은 만족도로 사용하고 있으며, 결제 서비스 부분도 안드로이드 마켓컬리보다 더 낫다고 판단할 수 있다.

3.2 로지스틱 분류 모형

3.2.1 로지스틱 회귀 분류 개념

로지스틱 회귀모형은 이분형 구조를 가지는 종속변수와 하나 이상의 독립변수 사이의 관계를 나타낸다. 독립변수와 종속변수 값의 분포가 같은 모양의 종속변수를 따라야 하는 등의 다양한 제약이 존재하는 일반적인 회귀분석에 비해 분석이 용이하다는 장점을 가지고 있다. 특히 본 연구에서는 서비스에 대한 만족 여부(긍정, 부정)을 종속변수로 하며, 다양한 단어를 독립변수로 사용하기 때문에 로지스틱 회귀분석으로 분석이 가능하다[10].

일반적인 선형 회귀분석은 종속변수의 평균이 독립변수에 대한 선형 결합이 되므로 식 (1)과 같이 나타난다.

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \cdots\cdots (1)$$

본 연구에서처럼 종속변수가 리뷰의 긍정 또는 부정처럼 이분형일 경우는 $E(y|x)$가 독립변수 x가 주어졌을 때, 사건 Y가 발생할 확률로 0에서 1사이의 값을 가지게 된다. 리뷰의 긍정 또는 부정의 확률은 식 (2)와 같다.

$$E(y|x) = probability = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \cdots\cdots (2)$$

위의 식 (2)를 로지스틱 선형관계로 변형할 경우는 식 (3)과 같이 표현된다.

$$Y = \log \frac{P(x)}{1 - P(x)} = \alpha + \beta x_i \quad \cdots\cdots (3)$$

즉, 로지스틱 회귀모형을 구축하였을 때 베타 계수를 통해 양과 음의 영향을 알 수 있다. 사용된 독립변수들의 Odds Ratio 값이 1 증가하면 사건 발생 확률이 비교 대상과 대비하여 Odds Ratio 배라고 해석할 수 있음으로 Odds Ratio를 사용한 랭킹화를 통해 결과를 분석하도록 한다. 오즈와 오즈비는 각각 식 (4)와 식 (5)로 나타낸다.

$$Odds = \frac{p}{1-p} = \begin{cases} \exp(\alpha + \beta), & x = 1 \\ \exp(\alpha), & x = 0 \end{cases} \quad \cdots\cdots (4)$$

$$Odds\ ratio = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_i (x_i + 1) + \cdots + \beta_p x_p)}{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_i (x_i) + \cdots + \beta_p x_p)} = \exp(\beta_i) \quad \cdots\cdots (5)$$

3.2.2 데이터 전처리 및 학습 방법

전처리는 데이터 분석, 처리 및 모델 학습에 도움이 되는 형태로 만드는 일련의 과정으로 데이터 분석 및 처리 과정에서 필수적인 단계이다.

데이터를 기반으로 한 기계학습은 학습 데이터의 양, 학습 모델, 그리고 데이터의 특징의 수 등 학습 환경에 따라 모델의 성능이 크게 좌우된다[11].

또한 데이터의 처리 방법에 따라서도 상이한 결과를 보인다. 따라서 보다 정확하고 효율적인 기계학습 결과를 얻기 위해서는 정확한 데이터 전처리 과정이 요구된다. 여기서 데이터의 전처리란 학습 이전에 학습에 불필요한 데이터와, 학습에 반드시 필요한 데이터를 구분하여 제거 및 분류하는 과정이다[11]. 이러한 데이터의 전처리 과정은 기계학습 모델의 처리 속도 및 정확도 등 다양한 성능을 향상시킬 수 있다[11].

이와 같이 의미 있는 결과를 도출하기 위해 구글, 애플 앱 스토어에서 수집된 데이터 전처리를 진행하기 위해 먼저 결측값을 확인하였다. 결측값이란 데이터의 값이 누락된 것을 의미한다. 결측값이 존재하면 데이터 분석 및 모델 학습에 치명적인 문제를 야기할 수 있다. 그러므로 결측값이 존재하는 데이터 row를 제거하거나, 주위 데이터와 비슷한 값으로 대체하는 방법으로 결측값을 처리하는데 각 방법은 장, 단점이 존재한다.

먼저, 구글 앱 스토어 전체 데이터 중 5개를 확인하였다.

	name	ratings	date	helpful	comment	developer_comment
0	박혜성	5	2021-04-16	0	너무편리해요	NaN
1	정윤우	5	2021-04-16	0	깔끔한 앱	NaN
2	김한울	1	2021-04-16	0	가격대별로 볼수있는 기능있음 좋겠습니다 딱 5000원만 딱3000원 더 골르면 더있...	NaN
3	이미옥	5	2021-04-16	0	빠른배송감사합니다	NaN
4	박민	5	2021-04-16	0	물품다양 고퀄리티 원료곳 배송곳 서비스곳	NaN

[그림 3-12] 구글 앱 스토어 데이터(5개)

그림 3-12을 보면 developer_comment에는 데이터가 존재하지 않아 NaN으로 표시되는 것을 확인할 수 있다.

결측값이 몇 개 존재하는지를 확인하기 위해 Info 명령어를 통해 각 컬럼별 데이터 수를 확인하였다.

[표 3-5] 구글 앱 스토어 Info

RangeIndex: 7395 entries, 0 to 7394			
Data columns (total 6 columns):			
	Column	Non-Null Count	Dtype
0	name	7395 non-null	object
1	ratings	7395 non-null	int64
2	date	7395 non-null	object
3	helpful	7395 non-null	int64
4	comment	7395 non-null	object
5	developer_comment	7028 non-null	object

구글 앱 스토어 데이터는 총 7,395개이다. developer_comment 컬럼을 제외한 모든 컬럼은 동일하게 7,395개의 데이터가 존재하고, developer_comments는 7,028개로 전체 데이터에서 367개의 데이터가 NaN(결측값)

인 걸 확인 할 수 있다.

다음은 애플 앱 스토어 전체 데이터 중 5개를 확인 하였다.

	name	ratings	date	title	comment
0	NaN	5	2020-02-16	그간의 마켓컬리를 이용하며	전 아주 만족합니다. 우선 요즘은 빠르고 간편하고 다양한 음식을 접하려고 하잖아요....
1	NaN	5	2020-07-02	포장이 너무 좋아요	먼저 아이스팩을 사용하지 않고 바로 버릴 수 있는 냉각을 사용하는 것과 종이 테이프...
2	NaN	5	2020-05-26	신세계 살아요!!	매일 광고만 보다가 지방은 안되겠지..했어요.
3	NaN	5	2018-05-13	행복한 앱이에요 ㅎㅎㅎ스크랩기능 만들어주세요!	친구가 추천해서 깔았는데 완전 좋네요 판매자랑 상품 별로 골겨찾기? 관심상품 설정 ...
4	NaN	2	2020-09-22	갈수록 실망	배송 누락이 너무 많이 발생합니다

[그림 3-13] 애플 앱 스토어 데이터(5개)

name을 제외한 컬럼에는 내용이 존재하고 name컬럼은NaN 값이 존재하는 것을 확인할 수 있다. Info 명령어를 통해 애플 앱 스토어 데이터 수를 확인하였다.

[표 3-6] 애플 앱 스토어 Info

RangeIndex: 2995 entries, 0 to 2994			
Data columns (total 6 Data columns (total 5 columns):			
	Column	Non-Null Count	Dtype
0	name	0 non-null	float64
1	ratings	2995 non-null	int64
2	date	2995 non-null	object
3	title	2995 non-null	object
4	comment	2995 non-null	object

애플 앱 스토어의 데이터는 총 2995개이다. name을 제외한 컬럼은 2,995개 데이터가 온전히 있는 것을 확인할 수 있다. 하지만 name컬럼은

NaN값으로 채워져 있는 것을 확인할 수 있었고, 데이터를 수집하는 과정에서 문제가 생겼음을 알 수 있다.

본 연구의 데이터 분석에는 'date, comment, ratings' 컬럼이 필요하다. 구글, 애플 앱 스토어 데이터에서 'date, comment, ratings' 컬럼을 확인 시 결측값은 존재하지 않기 때문에 별도의 결측값 수정 작업은 진행하지 않았다. 수집한 데이터프레임에서 date, comment, ratings 컬럼을 제외한 다른 컬럼들을 삭제하여 기본적인 데이터 형태를 정제하였다.

결측값 처리가 끝난 후에는 자연어 데이터를 분석에 사용하기 위해 단어 정제가 필요하다. 따라서 리뷰 데이터 단어 정제를 위해 'Konlpy, Soynlp, py-hanspell' 파이선 라이브러리와 Ranks nl에서 제공하는 'Korean Stopwords(불용어 사전)'파일을 사용하였다.

먼저 soynlp 라이브러리로 리뷰 데이터에 등장하는 반복되는 이모티콘, 영어, 숫자 제거와 띄어쓰기 교정을 진행하였다.

soynlp는 한국어 NLP(Natural Language Processing)를 위한 비지도 학습 기반 오픈소스 라이브러리이다. 기존의 형태소 분석기는 말뭉치를 기반으로 학습이 이루어지므로 학습되지 않은 단어를 제대로 인식하지 못하는 미등록 문제(Out Of Vocabulary)가 발생한다. 하지만 사람은 새로운 단어가 포함된 문장을 여러 개 읽게 되면, 이를 새로운 명사로 인식하게 된다. 이렇게 사람이 단어를 인식하는 방법으로 새로운 명사를 인식하는 방법을 사용하는 것이 soynlp이다[12].

py-hanspell은 네이버 맞춤법 검사기를 이용한 파이선용 한글 맞춤법 검사 라이브러리이다. 인터넷 리뷰 특성상 오타 및 적절하지 못한 맞춤법 등으로 특정 단어가 여러 개의 형태로 나뉠 수 있다. 따라서 리뷰 데이터의 원활한 분석을 위해 py-hanspell 라이브러리를 이용하여 리뷰 데이터 내의 맞춤법을 교정하였다.

soynlp는 명사와 형용사를 동시에 추출하기 때문에 명사만 따로 추출하기 위해서는 Konlpy 라이브러리로 추출된 명사, 형용사 단어에서 명사만 다시 추출해야 한다. Konlpy는 한국어 정보처리를 위한 파이선 패키지로 'Hannanum, KKMA, Komoran, Mecab, Okt(Twitter)'가 존재한다. 본 연구에서는 수집된 리뷰 데이터가 SNS 데이터와 형태가 비슷하다고 판단하여 Twitter 분석 시 사용하는 Okt 라이브러리를 사용하였다.

수집된 리뷰 데이터의 Label(ratings)은 5점 척도로 1점(매우 안 좋음), 2점(안 좋음), 3점(중립), 4(좋음), 5점(매우 좋음)으로 구성되어있다. 본 연구에서는 소비자가 어떤 서비스 요인을 만족하고 ratings 4, 5점을 부여했는지 확인하는 연구로 4점, 5점을 1로 군집 하였고 나머지 점수(1점, 2점, 3점)는 0으로 이진 분류하였다.

3점을 0으로 분류한 이유는 3점 리뷰에서 10개 랜덤 표본을 뽑아보았을 때 긍정적인 내용이 다소 있었으나 대부분 부정적인 리뷰로 구성되어 있었기 때문이다.

다음은 랜덤 추출한 3점 리뷰의 내용이다.

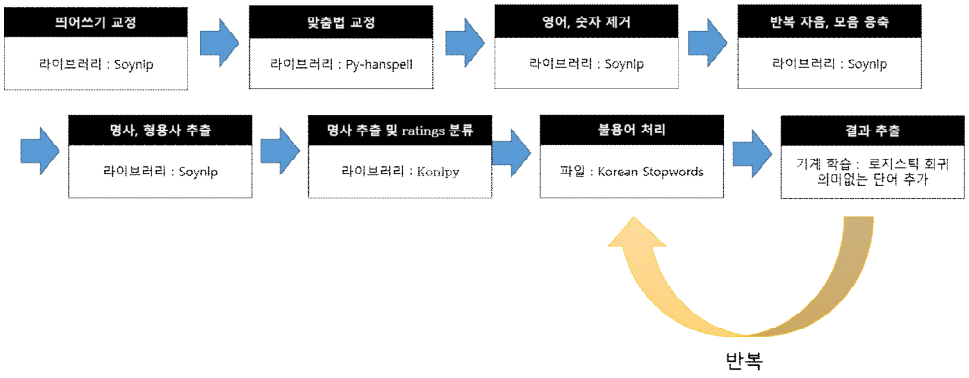
	date	comment	ratings
1156	2020-11-27	저렴한 상품도 준비해 주셨으면 좋겠어요.	3
5481	2019-12-13	베스트는	3
3420	2020-01-10	울산은 일반 택배도 불가능하다네요~ 쿠팡은 전국 다 된다길래 깔아봅니다. 그럼 안녕히~	3
2778	2020-03-24	관심상품 선택할 수 있도록 위시리스트 만들어주시고요! 과대포장 하지 말아주세요! 마...	3
1434	2020-10-06	다 좋은데..배송추적이 안됩니다. 추가로 pc버전에서는 배송지 수정을 할 수가 없습...	3
3885	2018-11-27	회원가입하려고하는데 자꾸 올바르지않은 휴대폰번호라고 뜹네요ㅠㅠ 며칠째 다시 해봐도 ...	3
3414	2020-01-13	어느 페이지에 있든 장바구니 아이콘이 떠있으면 좋겠습니다.	3
1937	2020-07-10	다 좋은데 키워드 검색이 정확하게 안되네요. 광고 넣은데 먼저 뜨는 느낌.	3
103	2021-04-06	결제 수단 추가및 확인 메뉴 찾기가 힘드네요	3
398	2021-02-23	주문하면 배송까지 오니 간편. ㅡ	3

[그림 3-14] 3점 리뷰 랜덤 추출

[표 3-7] 단어 전처리 라이브러리 및 불용어 사전

	이름	용도
파이썬 라이브러리	Soynlp	명사, 형용사 추출 영어, 숫자 제거 반복 이모티콘 제거
	Konlpy(Okt)	명사 추출
	Py-hanspell	맞춤법 교정
파일	Ranks Nl	불용어 처리

정제된 단어들은 Ranks nl에서 제공하는 불용어 사전 바탕으로 무의미한 단어들을 제거하였다. 기계학습(로지스틱 회귀)으로 나온 결과를 기반으로 연구에 불필요한 단어들을 불용어 사전에 추가하였고, 의미 있는 결과가 도출될 때까지 이 과정을 반복하였다.



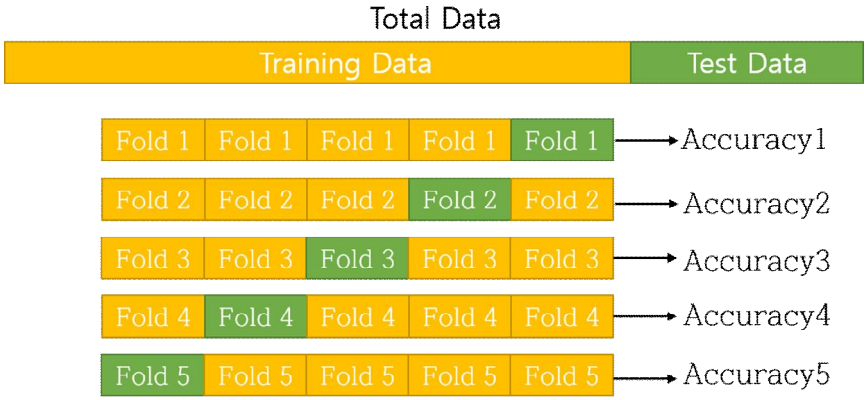
[그림 3-15] 전처리 프로세스

3.3 데이터 분류 결과 및 수치화

전처리한 데이터를 기반으로 로지스틱 회귀 모델(학습모델)에 학습시켰다. 전처리 후 리뷰 데이터양은 8,253개로 전체 데이터 10,390개에서 2,137개 데이터가 삭제되었다. 이렇게 전처리한 데이터를 훈련데이터(Traing data) 5,777개, 테스트데이터(Test date) 2,476개로 데이터를 7:3으로 나누었다.

모델 연구 평가는 K-fold 교차검증과 혼동행렬(confusion matrix)을 통해 평가를 진행하였다. 보통 기계학습은 하나의 테스트 데이터 세트를 가지고 모델의 성능을 확인하고 결과를 낸다. 이 경우 하나의 테스트 데이터 세트로 학습을 진행하기 때문에 과적합(Overfitting)이 발생할 수 있다. 따라서 과적합을 방지하기 위해 교차 검증으로 학습을 진행하였다. 교차 검증 기법은 전체 데이터에서 고정된 테스트 데이터로만 학습하지 않고, 모든 데이터를 사용하여 교차 검증하는 과정을 통해 모델의 성능을 검증한다.

본 연구에서는 K-fold 교차 검증을 통해 모델성능을 검증하였다. K-fold 교차 검증 프로세스는 전체 데이터를 K등분으로 분할하고, K-1개의 데이터 집합을 학습 데이터 나머지 1개의 데이터 집합을 테스트 데이터 세트를 할당한다. 교차 검증을 총 K번만큼 반복한다.



[그림 3-16] K-fold 프로세스

추후 K-fold 교차검증을 통해 나온 정확도 합한 후 K번만큼 반복한 횟수를 나누어서 평균을 구한다.

$$Average\ Accuracy = \frac{Accuracy1 + Accuracy2 + \cdots + AccuracyN}{N} \quad \cdots\cdots\cdots\cdots (6)$$

혼동행렬을 통해 정확도 외에 다른 지표를 함께 사용하여 모델의 성능을 평가하였다. 혼동행렬은 분류 모델의 성능을 평가하는 지표로 모델이 예측한 값과 실제 값을 배열해 행렬로 표현한 것이다. 본 연구에서는 실제 값과 예측값을 긍정(1)/부정(0)으로 분류하여 혼동행렬을 구성하였다.

[표 3-8] 혼동행렬 표

		실제 값	
		긍정(1)	부정(0)
예측값	긍정(1)	True Positive(TP)	False Positive(FP)
	부정(0)	False Negative(FN)	True Negative(TN)

[표 3-8]과 같이 혼동행렬은 TP, TN, FP, FN 4가지 요소로 구성되어 있다. 4가지 요소의 의미는 [그림 3-17]과 같다.

True Positive(TP): 학습모델의 예측 값이 참, 실제 값이 참
True Negative(TN): 학습모델의 예측 값이 거짓, 실제 값이 거짓
False Positive (FP): 학습모델의 예측 값이 참, 실제 값이 거짓
False Negative(FN): 학습모델의 예측 값이 거짓, 실제 값이 참

[그림 3-17] 혼동행렬 4가지 구성요소 의미

혼동행렬의 구성요소 4가지로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1-Score)를 구해 분류 모델 평가지표를 구할 수 있고 점수는 1에 가까울수록 좋은 성능을 낸다고 판단 할 수 있다.

정확도는 학습모델에서 정확하게 예측한 값의 비율을 나타내는 지표이다.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad \cdots\cdots (7)$$

정밀도는 학습 모델이 참으로 예측한 값에서 실제 참값의 비율을 나타내는 지표이다.

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \quad \cdots\cdots (8)$$

재현율은 실제 참값 중 학습모델이 참으로 예측한 비율을 나타내는 지표이다.

$$Recall = \frac{TP}{TP+FN} \quad \cdots\cdots (9)$$

F1점수는 재현율과 정밀도의 조화평균이다.

$$F1-Score = 2 \times \frac{1}{\frac{1}{\mathrm{Pr}ecision} + \frac{1}{Recall}} = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall} \quad \cdots\cdots (10)$$

K-fold 교차검증으로 전체 데이터 품질을 혼동행렬을 통해 정확도, 정밀도, 재현율, F1 점수를 산출하여 학습 모델의 성능을 평가하였다.

본 연구에서는 K-fold를 교차검증을 15번 반복하였고, 그 결과 전체 데이터 평균 정확도는 87.4%였다. 80% 이상의 정확도를 보이기 때문에 전체 데이터의 품질 및 모델의 정확도는 준수한 성능을 내고 있다고 판단할 수 있다.

[표 3-9] K-fold 교차 검증 정확도 및 전체 데이터 정확도 평균

K-fold 교차 검증 15번 실행				
86.6%	88.4%	87.3%	86.5%	88.1%
87.6%	89.3%	90.9%	87.3%	85.8%
85.3%	88.5%	85%	87.3%	86.4%

전체 데이터 평균 정확도
87.4%

학습모델의 혼동행렬 결과는 True Positive(TP) 1,989개, True Negative(TN) 194개, False Positive(FP) 254개, False Negative(FN) 39개이다.

[표 3-10] 학습모델결과의 혼동행렬

		실제 값	
		긍정(1)	부정(0)
예측값	긍정(1)	1989 (TP)	254(FP)
	부정(0)	39(FN)	194(TN)

혼동행렬의 4가지 구성요소로, 정확도, 정밀도, 재현율, F1 점수를 확인하였다. 정확도 0.88 정밀도 부정(0) 0.83, 긍정(1) 0.89 재현율 부정(0) 0.43, 긍정(1) 0.98, F1 점수 부정(0) 0.57, 긍정(1) 0.93인 것을 확인하였다.

[표 3-11] 정밀도, 재현율, F1-스코어, 정확도

	Precision	Recall	F1-Score	Accuracy	Support
0	0.83	0.43	0.57		448
1	0.89	0.98	0.93		2,028
				0.88	2,476

현재 긍정(1)에서 성능 평가는 좋지만, 상대적으로 부정(0)에서 성능 평가는 낮은 편이다. 부정 데이터의 양이 매우 낮아 학습하기에 부족했기 때문이라고 판단된다. 하지만 본 연구에서는 긍정(1)일 때 서비스 요인을 확인하는 연구이기 때문에 부정 데이터에서의 성능 문제는 연구 진행에 지장이 없을 것으로 판단하여 계속 연구를 진행하였다.

학습모델이 긍정(1)이라고 판단할 때 각 단어가 미치는 점수를 상위 10개 출력하였다.

[표 3-12] 학습모델 상위 10개 단어

단어	Coef
신선	2.49
만족	2.35
최고	2.33
편리	2.25
간편	1.98
항상	1.8
애용	1.71
아주	1.66
재료	1.58
요즘	1.57

단어의 분류 기준은 단어가 가지는 의미를 이해도와 연구모델에 적합성으로 설정한 후 연구를 진행하였다.

처음 결괏값으로 '최고, 항상, 애용, 아주, 재료, 요즘'이라는 단어가 출력되었다. 그러나 해당 단어들은 연구모델에서 의미가 불명확하기 때문에 위 단어들을 불용어 사전에 추가하여 제거한 후 다시 학습을 진행하였다. 상위 10개 단어 모두가 단어 분류 기준을 만족하였을 때를 기준으로 Ranks Nl에서 제공해준 기본 불용어 사전 단어를 제외하고 1,025개 단어가 제거되었다.

최종 전처리 후 리뷰 데이터양은 6,417개로 전체 데이터 10,390개에서 3,973개 데이터가 삭제되었다. 이를 훈련데이터(Traing data) 4,491개, 테스트데이터(Test date) 1,926개 7:3 비율로 데이터를 나누었다. 최종 데이터를 기반으로 K-fold 교차 검증한 결과 전체 데이터 정확도 평균은 84%로 초기 전체 데이터 정확도 평균보다 3.4% 떨어졌다.

[표 3-13] 최종 전처리 후 K-fold 교차 검증 정확도 및 전체 데이터 정확도 평균

최종 전처리 후 K-fold 교차 검증 15번 실행				
82.2%	85.5%	83.4%	85.3%	84.8%
85.5%	86.4%	87.4%	81.5%	83.4%
81.3%	85.3%	81.7%	83.6%	82%

전체 데이터 평균 정확도
84%

학습모델의 혼동행렬 결과는 True Positive(TP) 1,436개, True Negative

(TN) 187개, False Positive(FP) 261개, False Negative(FN) 42개이다.

[표 3-14] 최종 전처리 후 학습모델결과의 혼동행렬

		실제 값	
		긍정(1)	부정(0)
예측값	긍정(1)	1436 (TP)	261(FP)
	부정(0)	42(FN)	187(TN)

혼동행렬의 4가지 구성요소로, 정확도, 정밀도, 재현율, F1 점수를 확인하였다. 정확도 0.84 정밀도 부정(0) 0.82, 긍정(1) 0.85 재현율 부정(0) 0.42, 긍정(1) 0.97, F1 점수 부정(0) 0.55, 긍정(1) 0.90인 것을 확인하였다.

[표 3-15] 최종 전처리 후 정밀도, 재현율, F1-스코어, 정확도

	Precision	Recall	F1-Score	Accuracy	Support
0	0.82	0.42	0.55		448
1	0.85	0.97	0.90		1,478
				0.84	1,926

최종 전처리 후 학습모델 성능이 전 학습모델보다 다소 하락한 것을 확인하였다. 학습모델이 긍정(1)이라고 판단할 때 각 단어가 미치는 점수 기준으로 상위 10개 단어를 출력하였다.

[표 3-16] 최종 전처리 후 학습모델
상위 10개 단어

단어	Coef
신선	3
편리	2.15
간편	1.82
품질	1.69
포장	1.56
할인	1.46
혜택	1.4
환경	1.29
퀄리티	1.25
친환경	1.24

새벽배송 서비스 만족 요인을 오즈비를 통해 결과를 분석하였다. 따라서 설계한 연구 모형을 기반으로 학습 모델 상위 10개 단어를 군집시켰고, 학습모델에서 나온 단어 점수를 오즈비로 변경하여 계산을 진행하였다.

[표 3-17] 학습모델 단어 오즈비 변환

단어	Coef
신선	20.1
편리	8.6
간편	6.2
품질	5.4
포장	4.8
할인	4.3
혜택	4.1
환경	3.6
퀄리티	3.5
친환경	3.5

오즈비 변환 후 '신선'이란 단어가 다른 단어들에 비해 높은 점수가 매겨진 것을 확인할 수 있다. 오즈비로 변환한 점수를 연구 모형에 적용하고 합산한 결과는 다음과 같다.

[표 3-18] 연구모형 점수

A) 어플리케이션 사용요인(14.8)		편리(8.6) , 간편(6.2)
B) 배송요인(12.5)	B1) 시간	
	B2) 포장(12.5)	포장(5.4), 환경(3.6), 친환경(3.5)
	B3) 정확도	
C) 마케팅요인(8.4)		할인(4.3), 혜택(4.1)
D) 제품요인(29)	D1) 다양성	
	D2) 제품품질(29)	신선(20.1), 품질(5.4), 퀄리티(3.5)

제 4 장 결론

4.1 요약 및 시사점

본 연구의 결과를 기반으로 한 요약은 다음과 같다.

첫째, 연구모형의 총 점수는 64.7이다. 64.7점 중 제품요인 29점(44.8%), 애플리케이션 사용요인 14.8(22.9%)점, 배송요인 12.5(19.3%)점, 마케팅요인 8.4(13%)점으로 제품요인이 다른 요인들보다 새벽배송 서비스 만족도에 미치는 영향력이 높다는 사실을 확인할 수 있다.

둘째, 새벽배송 서비스 중 배송요인에서 제품 품질 요인이 중요하다. 당일 배송이 다수 서비스되고 있는 현재 배송 시장에서 시간과 정확도는 차별화된 서비스가 아닌 것으로 판단된다. 소비자들은 시간과 정확도 외에도 추가적인 요인을 고려하는 것으로 드러났다. 바로 제품 품질 요인이다. 새벽배송은 신선 제품을 주 제품으로 판매하고 있기 때문에, 일반 배송보다 좋은 품질로 배송이 가능하다. 따라서 신선 제품을 구매하는 고객들에게 긍정적인 서비스라고 판단된다.

셋째, 포장요인에서 '환경', '친환경' 단어가 군집되어 있는 것을 확인할 수 있다. 마켓컬리는 재활용 가능한 친환경 포장재를 제품 배송에 사용한다. 현재 배송 시장 성장과 더불어 배송에서 발생하는 재활용이 불가능한 쓰레기가 증가하였다. 그로 인해 배송 서비스를 이용하는 고객들이 환경에 대한 중요성을 인식하게 되었다. 이에 마켓컬리의 친환경 패키징 서비스가 새벽배송 서비스를 이용하는데 긍정적인 영향을 미쳤다고 판단된다.

넷째, 애플리케이션 사용요인은 고객 만족에 영향을 미친다. 모바일로 제품을 구매하는 고객들이 증가하면서 기업 대다수가 애플리케이션 개발에 집중하게 되었다. 이로 인해 애플리케이션 UI, UX는 급속도로 성장하였다. 고객들은 다양한 UI, UX를 접하게 되었으며, 이제는 UI, UX로부터 비롯된 애플리케이션 사용요인이 만족도에까지 영향을 미치게 된 것이다. 특히 개인추천 서비스가 고도화되면서 고객의 패턴을 파악해 제품을 추천해주는 UX가 필수적인 요소로 자리 잡았다고 판단된다. 앞으로도 고객을 만족시킬 수 있는 UI, UX 분야의 발전이 요구된다.

정리하자면 새벽배송 서비스에서 고객만족이 제일 높은 요인은 제품요인이었고 다음은 차례대로 애플리케이션 사용요인, 배송요인, 마케팅요인이었다. 특히 배송요인에서는 친환경 서비스 요인의 비중이 크다는 것을 확인할 수 있었다.

새벽배송 시장은 2016년에 시작하여 2021년까지 폭발적으로 성장하였다. 모바일, 온라인 쇼핑의 급속적인 성장으로 새벽배송 시장의 성장은 지속할 것이라고 판단된다. 그러므로 규모의 확장에 비례하여 시장에 대한 더욱 상세한 연구가 필요하다. 본 연구에서는 새벽배송 서비스 요인 파악을 위해 소비자와 가장 가까이 접해있는 애플리케이션 리뷰 데이터를 분석하였다. 또한 본 논문은 보다 상세한 분석을 위해 새벽배송 서비스 요인의 연구 범위를 단순히 배송으로만 국한하지 않았다. 고객이 애플리케이션에 접속해서 주문한 제품을 받는 과정과 마케팅, 제품구성을 연구모형(애플리케이션 사용요인, 배송요인, 마케팅요인, 제품요인)으로 구축해 소비자들이 어떤 서비스 요인으로 만족에 영향을 받는지 분석하

였다.

본 연구를 통해 알아본 새벽배송 서비스의 사용자 요인은 배송요인과 애플리케이션 사용요인으로 나뉘었다. 배송 요인은 현재 상향 평준화로 신속성과 정확성은 소비자가 기본적으로 요구하는 요인이 되어버린 경향이 있기에 제품 품질에 대한 차별화가 필요하다. 친환경적 요인의 경우는 이전의 친환경적 요인에 관한 연구와 비슷한 시사점을 나타내었는데, 서비스에 긍정적인 영향을 미치는 것을 알 수 있었다. 그러나 해당 요인이 영향을 미치는 범주에 대해서는 차후 해당 범주에 연관된 데이터만을 사용하여 보다 세분화된 연구가 필요하다. 마지막으로 애플리케이션의 사용성이 소비자 만족에 유의미한 영향을 미치는 것을 알 수 있었다. 그러므로 기업은 소비자 만족을 위해 보다 애플리케이션 사용성 향상에 대한 집중적인 연구가 필요하다. 해당 연구를 통해 알아본 결과들을 발전시키면 앞으로 새벽 배송 관련 실무진 및 경영진의 의사결정 기반 자료로 사용될 수 있을 것으로 판단된다.

4.2 향후 연구 과제

본 연구를 진행하면서 발생한 몇 가지 한계점 및 추후 연구에 대한 제언을 제시한다.

첫째, 부정 리뷰데이터 부족이다. 마켓컬리의 대다수의 데이터가 긍정 데이터라서 긍정적인 요인 결과는 쉽게 도출할 수 있었다. 그러나 부정 리뷰 데이터 부족으로 요인의 결과에 대한 신뢰도가 낮게 산출되었다. 그로 인해 고객에게 부정적 영향을 미치는 서비스 요인을 찾을 수 없었

다는 한계점이 있다. 추후 연구에서는 부정 데이터를 더 확보해 기존 학습모델보다 좋은 모델을 구축하여 부정적인 서비스 요인 연구를 수행할 예정이다.

둘째, 마켓컬리 외에도 새벽배송 서비스를 하는 대기업, 스타트업이 등장하였다. 다양한 기업의 리뷰데이터를 바탕으로 각 기업의 서비스 요인 특징을 분석할 필요가 있어 보인다. 이를 통해 거시적인 새벽배송 서비스를 연구하고 더 나아가서는 각 기업의 서비스 강점을 분석할 계획이다. 분석된 강점을 바탕으로 각 기업이 나아가야 할 방향을 제시할 수 있을 것으로 기대한다.

셋째, 본 연구에서 진행한 학습 모델은 단순히 새벽배송 서비스 요인 분석에만 국한되는 모델이 아니다. 그러므로 추후 다른 서비스에도 접목하여 해당 서비스 사용 시 만족 및 불만족 서비스 요인을 확인해 볼 예정이다.

넷째, 본 연구에서 사용한 로지스틱 회귀 분석뿐만 아니라 다양한 딥러닝 모델에도 해당 연구 주제를 적용하면 보다 넓은 범주에서의 연구가 가능할 것으로 예상된다. 더 많은 데이터를 기반으로 해당 분야로의 연구로 범위를 확장할 예정이다.

참고문헌

[1] 김근형 (2011. 10), 고정적 차수의 관계형 테이블을 기반으로 한 온라인 고객리뷰의 분석 기법 「한국정보과학회」

[2] 김민정, 곽민주 (2020. 12), “새벽배송 서비스의 이용지속성에 관한 연구 : 소비자피해경험, 감정경험 및 서비스만족도의 영향” 「소비자정책교육연구」

[3] 윤덕환, 채선애, 송으뜸 (2019. 02), “새벽배송 서비스 관련 인식 조사” 「리서치보고서」

[4] 추진기(2019. 11), “신유통 트렌드 시대 브랜드 확장 아이덴티티(B.I) 전략 연구 -새벽배송 택배회사의 사용자 경험(UX) 요인을 중심으로- 「한국디자인트렌드학회」

[5] 정지희, 신재익(2020. 10), 새벽배송의 물류 서비스 품질이 고객만족에 미치는 영향: 친환경 태도의 조절효과 「한국컴퓨터정보학회논문지」

[6] 윤지선(2020. 12), 의사결정나무분석과 로지스틱 회귀분석을 이용한 우울 예측요인 비교연구 「한국사회복지경영학회」

[7]최자영, 김현아, 김용범(2020.07), 온라인 리뷰가 매출에 미치는 영향력 분석: 텍스트기반 감성지수를 중심으로 「한국유통학회」

[8]장정아, 이현미, 박형원(2019. 08), 로지스틱회귀모형과 오즈비분석을 이용한 KNCAP제도의 효과평가방법 연구 「한국자동차공학회논문집」

[9] 송보미, 윤병운, 박용태 (2011. 11), 고객 리뷰 분석을 통한 서비스 품질 진단 : 감성 분석 및 갭 분석 접근 「대한산업공학회」

[10] 손희영, 강만수, 박상규(2014. 09), 내부마케팅이 직무만족, 애호도, 기업성과에 미치는 영향 -로지스틱회귀분석 방법을 이용. 「한국경영과학회지」

[11]김동현, 유승언, 이병준, 김경태, 윤희용(2019. 01), 효율적인 기계학습을 위한 데이터 전처리 「한국컴퓨터정보학회」

[12]김정욱, 정지완, 차미영, (2020. 02), 포털 뉴스 기사를 이용한 신조어 목록 자동 추출 「한국HCI학회」

[13] 한국 전자상거래 시장, 지난해 세계 5위…중국 1위 . (2021). https://www.yna.co.kr/view/AKR20210209060000030

[14] Part1. 지속되는 '라스트마일' 시장의 성장 . (2021). https://www.klnews.co.kr/news/articleView.html?idxno=122698

[15]아침에 바로 먹을 수 있게…유통계 새벽배송 전쟁 시작됐나 . (2019). https://www.donga.com/news/Opinion/article/all/20190714/96470925/1

[16] 박찬석의 물류時論(72) / 새벽배송시장 현황과 전망 . (2019).
http://www.ulogistics.co.kr/test/board.php?board=column2&command=body&no=520

[17] [트렌드모니터] '새벽 배송' 서비스, 소비자의 라이프스타일에도 영향을 줄까? .
(2020).
https://www.madtimes.org/news/articleView.html?idxno=3849

[18] 구글, '앱 수수료 30%' 움직임…업계 "너무 높다" 촉각 . (2020).
https://www.edaily.co.kr/news/read?newsId=03824486625833208&mediaCodeNo=257

[19] 국내 2분기 스마트폰 시장, 삼성 67% 점유율로 견고한 1위 .
(2020).
https://news.zum.com/articles/11020200911162715610