
Catch You! (캐치 유)

최종보고서

2022. 12. 05

이 서 연

졸업작품최종보고서

학부/전공			
성명	이서연	졸업정일	2023년 02월
0구분	단독 <input checked="" type="checkbox"/> 공동 <input type="checkbox"/> ()인		
휴대전화	010-2257-5872	E-mail	tjdus5077@naver.com
작품제목	Catch You! (캐치 유)		
작품개요	- 고객과의 지속적인 관계를 유지하고, 고객의 개별적 특성에 따른 관리를 위한 데이터 분석 - 기존 고객 유지 및 신규 고객 창출을 위한 전략 - 고객 구매 데이터 기반 상품 구매 예측 및 추천 - 윈도우 환경에서 아나콘다 주피터 노트북 사용		
위와 같은 내용으로 최종보고서를 제출합니다.			
2022년 12월 05일 제출자 : 이서연(인)			
산업데이터사이언스학부			

※ 아래와 같은 내용의 작성하여 제출하시오.

- 데이터 분석 기획 전 프로세스 내용 포함 (문제 사항 보완 결과 포함)
- 작품 제목, 필요성, 작품 개요(각 구성 요소별 block diagram 수준 기술), 분석 기술(플랫폼, 사용 프로그래밍 언어 및 기타 도구), 필요 장비/소프트웨어, 제작 과정, 작품 해석 등

※ 2인 이상의 공동 작품은 최종보고서를 각자의 담당 부분을 중심으로 기술하여 개별적으로 제출하되, 작품 전체에 대한 요약본을 첨부하시오.

승인여부	가()	부()	가()	부()	가()	부()	가()	부()
지도교수		(인)		(인)		(인)		(인)

보 완 사 항 확 인 서

학 부 / 전 공			
성 명		졸 업 예 정 일	년 월
구 분	단 독 <input type="checkbox"/> 공 동 <input type="checkbox"/> ()인		
휴 대 전 화		E-mail	
작 품 제 목			
요 구 사 항			
최 종 결 과			
<p style="text-align: center;">위와 같이 졸업작품을 최종 보완하였음을 확인합니다.</p> <div style="text-align: right; margin-top: 20px;"> 년 월 일 교 수 : (인) 제 출 자 : (인) </div> <p style="text-align: center; margin-top: 30px;">산 업 데 이 터 사 이 언 스 학 부</p>			

목 차

제 I 장 서론	1
1 목적 및 필요성	0
2 활용 방안	0
제 II 장 분석 기획	0
1 비즈니스 이해 및 범위 설정	0
2 프로젝트 정의 및 계획 설정	0
3 위험계획 수립	0
제 III 장 데이터 준비	0
1 필요 데이터 정의	0
2 데이터 수집 및 전처리	0
3 품질 점검	0
제 IV 장 데이터 분석	0
1 분석용 데이터 준비	0
2 탐색적 분석	0
3 모델링	0
4 모델평가 및 검증	0
제 IV 장 시스템 구현	0
1 설계 및 구현	0
2 시스템 테스트 및 운영	0
제 V 장 평가 및 전개	0
1 모델 발전 계획 수립	0
2 프로젝트 평가 및 향후 계획	0
참고문헌	
별첨 1.	0
별첨 2.	0

I 서론

1. 목적 및 필요성

가. 목적

- ☐ 고객 특성 파악
 - 고객구매데이터 기반 고객 특성에 따른 개별 관리를 위한 데이터 분류
- ☐ 구매 예측 및 추천
 - 상품 구매를 예측하고 추천함으로써 고객과의 관계 유지

나. 필요성

- ☐ 고객과의 지속적인 관계 유지
 - 고객구매데이터에 기반 한 분석으로 고객과의 지속적인 관계유지
- ☐ 신규 고객 창출 및 기존 고객 유지
 - 구매 횟수를 기준으로 데이터 분리/생성

2. 활용 방안

- ☐ 고객 관리 및 마케팅에 활용

II 분석 기획

1. 비즈니스 이해 및 범위 설정

가. 비즈니스 이해

☐ L.POINT 빅데이터 서비스

- 온/오프라인 구매정보, 상품정보, 위치정보, 온라인 행동정보 뿐만 아니라 고객의 정성적인 부분을 이해할 수 있는 리서치 데이터로 구성
- 빅데이터 플랫폼을 통해 데이터 인사이트 확보, 맞춤형 컨설팅 서비스 제공
- 상품 구색, 수요 예측, 부진점/점포 유형화, 신규/이탈/고객가치 비즈니스 분석 이슈 진단 및 개선
- 마케팅 매출 효과 및 수익성 증대
- 통계, 머신 러닝 기반 모델 개발

나. 범위 설정

☐ 고객 관리 및 수요 예측

- 고객과의 관계를 관리하고, 구매데이터에 기반 한 수요 예측

2. 프로젝트 정의 및 계획 설정

가. 프로젝트 정의

- 고객 구매 데이터에 기반 한 상품 구매 예측 및 추천

나. 계획 설정

4주차 : 주제 변경

5주차 : 온라인 쇼핑몰 이용내역 데이터 수집 및 전처리

6주차 : 온라인 쇼핑몰 이용내역 데이터 시각화 탐색, 코로나 확진 데이터 전처리

7주차 : 데이터 변경, 롯데멤버스 고객구매데이터 시각화 탐색, 고객 구매 분석

8주차 : 재구매/신규구매 분석, SVD를 이용한 상품 구매 예측

9주차 : 중간보고서 제출

10주차 : 상품 추천 시뮬레이션 에러 수정, 고객 군집 분석

11주차 : 연관성 분석을 통한 구매 확장 인사이트 도출

12주차 : L.POINT 이용 데이터를 연관하여 포인트 이용 증대 방안 마련

13주차 : 최종 마케팅 전략 제안

14주차 : 최종 발표

3. 위험계획 수립

- 비현실적 일정 : 세부적인 일정 계획 필요
- 실시간 성능의 빈약 : 예측 시뮬레이션 에러 발생 해결 방안 필요
- 대부분의 위험은 수용할 것

III 데이터 준비

1. 필요 데이터 정의

번호	테이블 명	설명	파일명
1	Demo	고객 데모 정보	LPOINT_BIG_COMP_01_DEMO.csv
2	상품 구매 정보	유통사 상품 구매 내역	LPOINT_BIG_COMP_02_PDDE.csv
3	제휴사 이용 정보	제휴사 서비스 이용 내역	LPOINT_BIG_COMP_03_COP_U.csv
4	상품 분류 정보	유통사 상품 카테고리 마스터	LPOINT_BIG_COMP_04_PD_CLAC.csv
5	점포 정보	유통사/제휴사 점포 마스터	LPOINT_BIG_COMP_05_BR.csv
6	엘페이 이용	엘페이 결제 내역	LPOINT_BIG_COMP_06_LPAY.csv

2. 데이터 수집 및 전처리

가. 데이터 수집

- 제 7회 롯데멤버스 빅데이터 경진대회에서 제공받은 데이터로, L.POINT 고객센터에 문의 후 프로젝트에 데이터 사용을 허가 받음.

나. 데이터 전처리

- 영수증 번호에 의미가 들어간 부분은 없어 영수증 번호 열 제거
- 각 데이터별로 결측값 확인 후 제거
- 구매 날짜 열 object 타입에서 date 타입으로 변환
- 월별 분석을 위한 month열 생성
- 상품 구매 데이터에서 상품 코드에 따른 상품 카테고리 분류 열 추가

다. 품질 점검

- 유효성 : 가공된 샘플 데이터로 실제 시장 데이터와 차이가 있을 수 있음.
- 활용성 : 프로젝트 수행을 위해 요구되는 데이터를 충족하며,
21년 1월부터 12월까지 총 12개월의 데이터 포함.
- 보안성 : 롯데 멤버스 엘포인트 고객센터에 문의 후 데이터 사용

IV 데이터 분석

1. 분석용 데이터 준비

번호	테이블 명	설명	파일명
1	Demo	고객 데모 정보	LPOINT_BIG_COMP_01_DEMO.csv
2	상품 구매 정보	유통사 상품 구매 내역	LPOINT_BIG_COMP_02_PDDE.csv
3	제휴사 이용 정보	제휴사 서비스 이용 내역	LPOINT_BIG_COMP_03_COP_U.csv
4	상품 분류 정보	유통사 상품 카테고리 마스터	LPOINT_BIG_COMP_04_PD_CLAC.csv
5	점포 정보	유통사/제휴사 점포 마스터	LPOINT_BIG_COMP_05_BR.csv
6	엘페이 이용	엘페이 결제 내역	LPOINT_BIG_COMP_06_LPAY.csv

가. 분석에 필요한 데이터 범위 지정

- 고객 구매 분석 : 'Demo' 테이블의 29913명의 고객을 대상으로 분석
- 재구매/신규 구매 분석 : '상품 구매 정보' 테이블을 11월 기준으로 데이터 분할
- 구매 예측 모델 : '상품 구매 정보' 테이블을 train/test 데이터로 나누어 분석
- 고객 군집 분석 : 'Demo', '상품 구매 정보' 테이블 이용하여 군집화
- 전체 데이터 중에서 'Demo', '상품 구매 정보', '상품 분류 정보' 데이터만 사용하나, 모든 데이터가 제휴사와도 연관되어 있어 '제휴사 이용 정보'와 '점포 정보' 데이터를 제외할 수는 없음

2. 탐색적 분석

가. 데이터 탐색

- Demo (29913 rows × 4 columns)

칼럼명	칼럼 설명	자료 설명
cust	고객 고유 코드	고객 별 고유한 값
ma_fem_dv	성별	여성/남성으로 구성
ages	나이 대	6가지 나이대로 구성
zon_hlv	지역 코드	17가지 코드로 구성

- 상품 구매 정보 (4381743 rows × 10 columns)

칼럼명	칼럼 설명	자료 설명
cust	고객 고유 코드	고객 별 고유한 값
cop_c	제휴사 코드	
br_c	점포 코드	
pd_c	상품 코드	
de_dt	거래 날짜	날짜(yyyy-mm-dd)
de_hr	거래 소요 시간	
buy_am	거래 금액	
buy_ct	거래 수량	
de_month	거래 월	

- 제휴사 이용 정보 (248304 rows × 9 columns)

칼럼명	칼럼 설명	자료 설명
cust	고객 고유 코드	고객 별 고유한 값
cop_c	제휴사 코드	
br_c	점포 코드	
de_dt	거래 날짜	날짜(yyyy-mm-dd)
vst_dt	방문 날짜	날짜(yyyy-mm-dd)
de_hr	거래 소요 시간	
buy_am	거래 금액	
de_month	거래 월	
vst_month	방문 월	

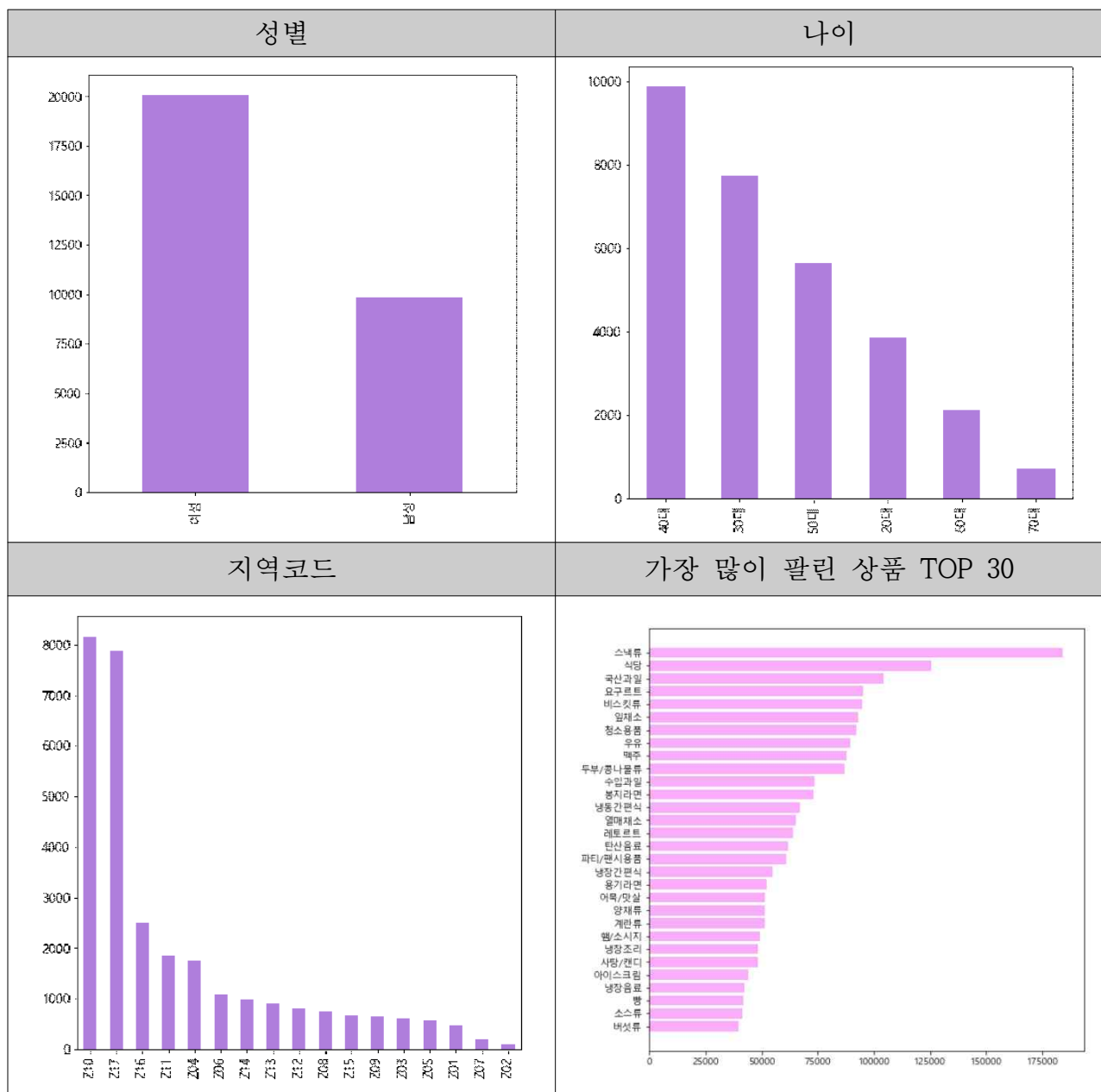
- 상품 분류 정보 (1933 rows × 4 columns)

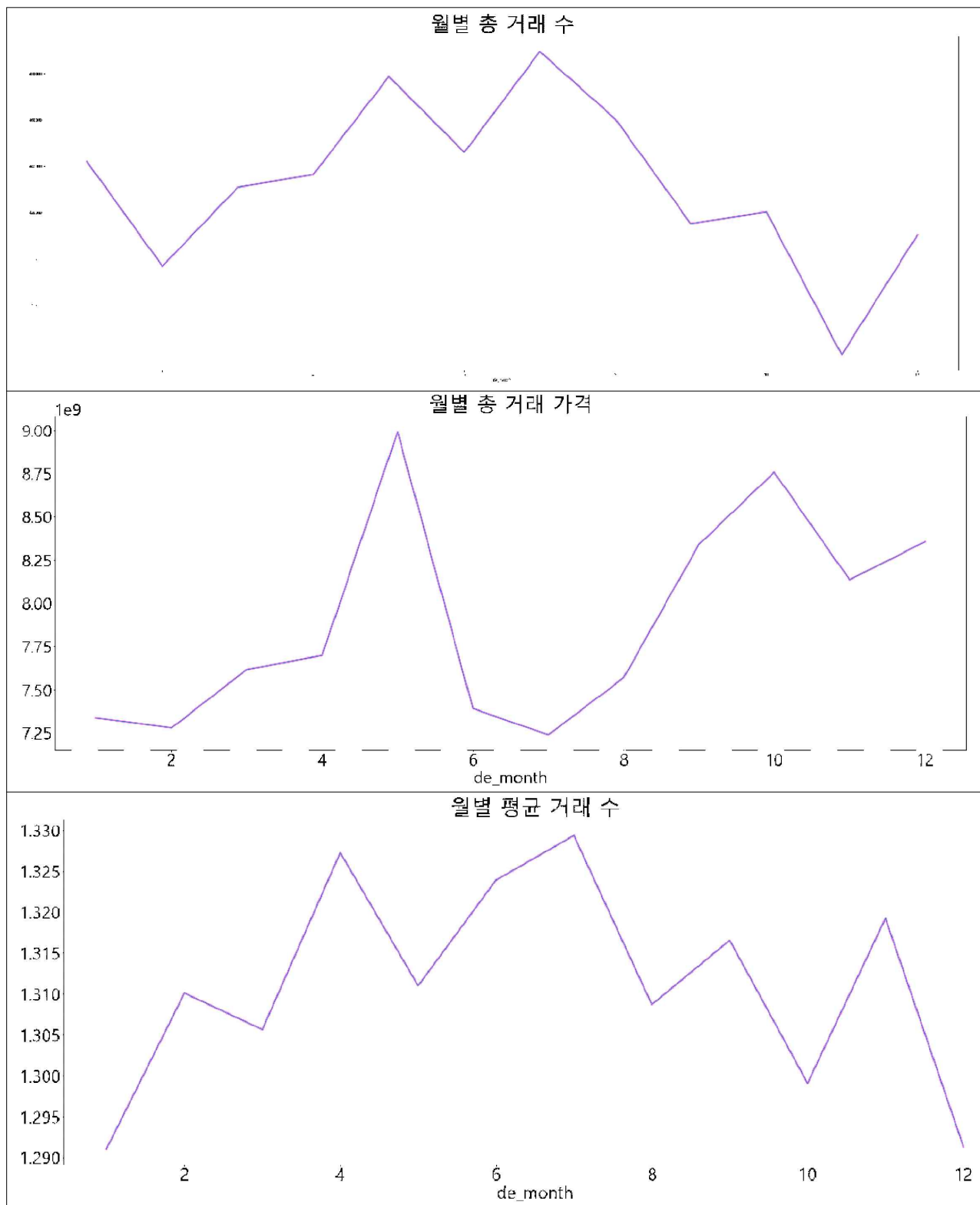
칼럼명	칼럼 설명	자료 설명
pd_c	상품 코드	
pd_nm	상품 상세 분류	ex) 소파
clac_hlv_nm	카테고리 분류	ex) 가구
clac_mcls_nm	상품 분류	ex) 거실가구

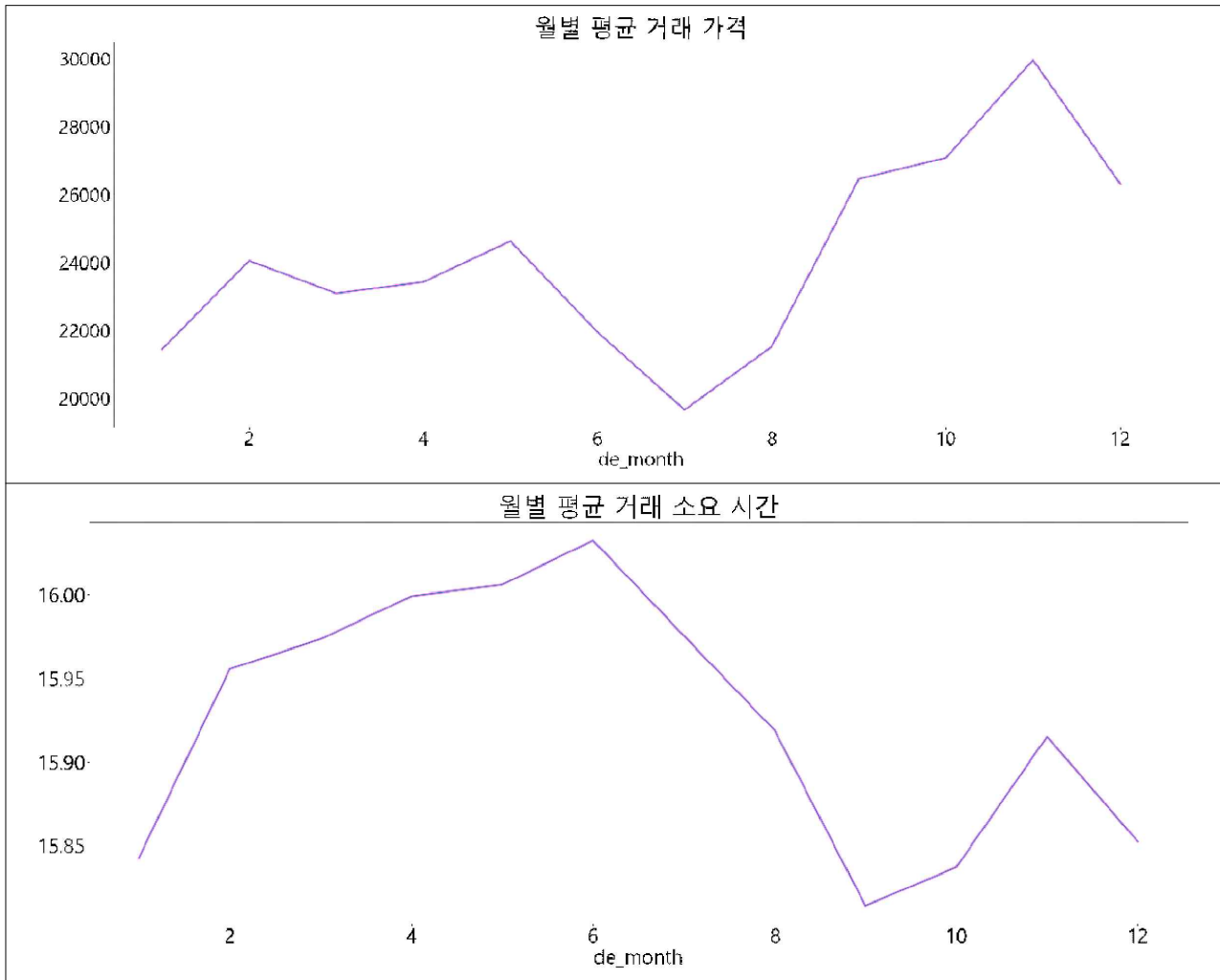
- 점포 정보 (8808 rows × 4 columns)

칼럼명	칼럼 설명	자료 설명
br_c	점포 코드	
cop_c	제휴사 코드	
zon_hlv	지역 코드	
zon_mcls	지역 상세 번호	

나. 데이터 분포 시각화







다. 고객 구매 패턴

구매 횟수 탐색	상품 구매 종류 탐색
<p>1</p>	<p>1</p>
<p>평균 4회 구매, 대부분 1~5회 정도 구매하는 것으로 나타남.</p>	<p>평균 57.5개 종류를 구매하는 것으로 나타남. 그러나 편차가 매우 큼.</p>

라. 재구매/신규 구매 분석

- 11월을 기준으로 데이터를 분할하여, 11월 이전/이후 고객별 구매했던 상품의 집합 추출하여 고객별로 기록하기 위한 딕셔너리 생성

11월 이전에 구매한 상품은 'old', 11월 이후에만 구매한 상품은 'new', 11월 이전과 이후 모두 구매한 상품은 'both'로 데이터 프레임 생성

- 생성된 데이터 프레임

	cust	old	new	both
0	M000034966	25	3	1
1	M000136117	36	2	4
2	M000201112	14	1	1
3	M000225114	28	2	15
4	M000261625	15	1	1

- 재구매/신규 구매 분석 결과

- 전체 고객의 수 26027명

- 11월 이후, 기존에 구매한 적 없는 새로운 상품을 구매한 고객의 수는 17573명

- 11월 이후, 재구매한 상품이 있는 고객의 수는 18047명

- 총 상품의 수는 1933개

- 신규 구매한 상품이 있는 고객들은 평균적으로 약 385개 종류의 신규 상품을 구매하는 것으로 나타남. 그러나 편차가 매우 큰 것으로 보아 신규 상품을 구매하는 고객들은 일반적으로 많은 종류의 상품을 구매하지 않을 것으로 예상할 수 있음.

3. 모델링

가. SVD 모델을 이용한 상품 구매 예측

- 특정 시점 이전의 데이터에 SVD 모델을 학습하고, 특정 시점 이후의 고객과 상품의 선호도 점수 예측

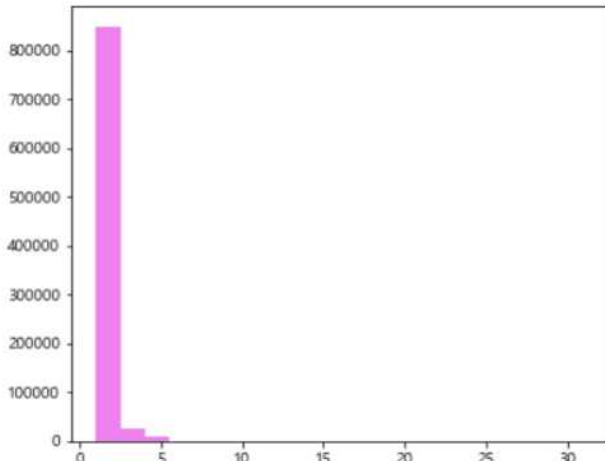
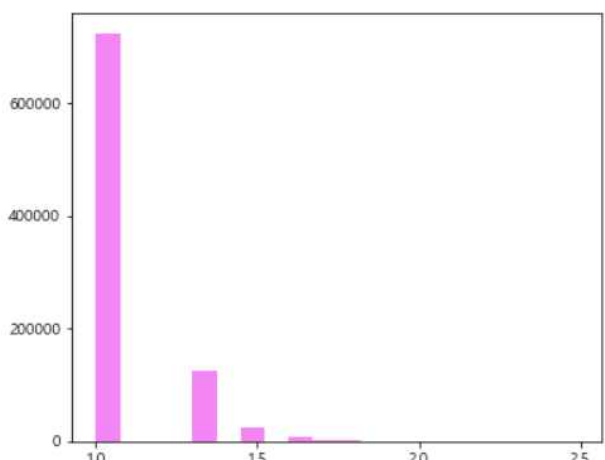
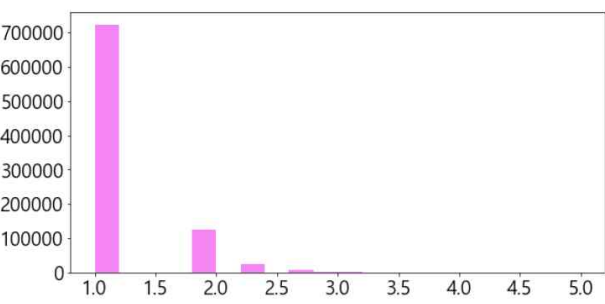
- 추천 대상 고객의 수는 255517명, 추천 대상 상품의 수는 1838개

- 본 데이터를 가지고 SVD 모델을 학습함에 있어서 고객과 상품의 선호도 점수를 가지고 있지 않아 피처엔지니어링을 통해 점수 생성

- '고객별 구매 횟수는 일반적으로 1~5사이에 분포되어 있다' 는 정보 이용

- 이 정보를 이용하여 고객-상품 간 구매 횟수가 Rating으로 사용하기에 적절한지 탐색

- 다음은 Rating 생성 과정

	<ul style="list-style-type: none"> - Rating이라고 가정한 고객-상품 간의 구매 횟수 분포를 나타내는 그래프 생성 - 그래프를 살펴보면 대부분의 점수가 1~5사이에 위치하기는 하지만 점수가 낮은 쪽으로 많이 치우쳐져 있음. - 이러한 분포를 가진 Rating으로 SVD 모델을 학습한다면 제대로 완성하지 못할 확률이 높음
	<ul style="list-style-type: none"> - 이러한 상황에 적용할 수 있는 피쳐 엔지니어링 기법으로 로그를 통한 피쳐 정규화 실시 - 여전히 왜도가 높지만 적용 이전에 비해서는 피쳐를 Rating으로 사용하기에 적합한 모습을 보임
	<ul style="list-style-type: none"> - 로그를 통한 정규화 적용 후, 다시 피쳐 스케일링을 적용하여 1~5사이의 값으로 변환 - 최대-최소 스케일링 방법 사용

나. SVD 모델 성능 테스트

- 11월 이전 데이터로 생성한 학습 데이터를 또 다시 학습 데이터와 테스트 데이터로 분리하여 모델을 학습하고 평가
- 성능 평가 결과
 - training time of model : 182.07 seconds
 - RMSE : 0.3640

다. 상품 추천 시뮬레이션

- 11월 데이터를 모두 학습 데이터로 사용하여 모델을 학습
- 추천대상은 11월 이전 데이터에 존재하는 모든 고객
- 추천의 대상이 되는 상품은 3가지로 분류
 - 분류 기준은 재구매/신규 구매에 대한 탐색을 기반으로 한 것
 - 1) 이전에 구매한 적 없던 상품 추천 : 신규 구매 타겟
 - 2) 이전에 구매했던 상품 추천 : 재구매 타겟
 - 3) 모든 상품 대상 추천 : 모든 고객-상품 점수 고려하여 추천
- 1) 이전에 구매한 적 없던 상품 추천 : 신규 구매 타겟
 - 이전에 구매하지 않았던 상품을 예측 대상으로 선정하고, 구매 예측 결과를 딕셔너리 형태로 변환
- 2) 이전에 구매했던 상품 추천 : 재구매 타겟
 - 이전에 구매했었던 상품을 예측의 대상으로 선정하고, 구매 예측 결과를 딕셔너리 형태로 변환
- 3) 모든 상품 대상 추천 : 모든 고객-상품 점수 고려하여 추천
 - 위에서 생성한 두 개의 딕셔너리를 하나의 딕셔너리로 통합

4. 모델평가 및 검증

- 11월 이후 데이터를 테스트 데이터로 활용하기 위해 각 고객들이 11월 이후에 실제로 구매한 상품의 리스트를 데이터 프레임 형태로 정리

	cust	RealOrdered
0	M000034966	{애견용품, 스낵류, 남아완구, 비스킷류}
1	M000136117	{여아의류전신, 여성속옷, 유아의류전신, 기타아웃도어/레저, 패션액세서리, 남성의류세트}
2	M000201112	{과채음료, 비스킷류}
3	M000225114	{스낵류, 초콜릿, 사탕/캔디, 기능성음료, 잎채소, 남성일반스포츠의류, 열매채소,...}
4	M000261625	{푸드코트, 스킨케어}

- 시뮬레이션 테스트용 데이터 프레임에 3개의 딕셔너리 시뮬레이션 결과 추가

	cust	RealOrdered	PredictedOrder(New)	PredictedOrder(Reorder)	PredictedOrder(Total)
0	M000034966	{애견용품, 스낵류, 남아완구, 비스킷류}	{여아의류특수목적의류, 용기보충금, 일반담배, 냉장음료, 소주, 아이스크림, 남아의...}	{맥주, 기능성음료, 수입과일, 열매채소, 사탕/캔디, 청소용품, 과채음료, 여성속...}	{맥주, 여아의류특수목적의류, 용기보충금, 일반담배, 냉장음료, 소주, 아이스크림,...}
1	M000136117	{여아의류전신, 여성속옷, 유아의류전신, 기타아웃도어/레저, 패션액세서리, 남성의류세트}	{남아특수소재의류, 동물병원, 여아의류특수목적의류, 용기보충금, 파티/팬시용품, 남...}	{여성속옷, 식당, 스킨케어, 맥주, 빵, 차음료, 남성의류상의, 열매채소, 탄산음...}	{남아특수소재의류, 동물병원, 여아의류특수목적의류, 용기보충금, 파티/팬시용품, 남...}
2	M000201112	{과채음료, 비스킷류}	{남아특수소재의류, 파티/팬시용품, 여아의류특수목적의류, 동물병원, 용기보충금, 잡...}	{생수, 스낵류, 소주, 용기라면, 비스킷류, 맥주, 냉장간편식, 탄산음료, 기타,...}	{남아특수소재의류, 파티/팬시용품, 여아의류특수목적의류, 동물병원, 용기보충금, 잡...}
3	M000225114	{스낵류, 초콜릿, 사탕/캔디, 기능성음료, 잎채소, 남성일반스포츠의류, 열매채소,...}	{동물병원, 아이스크림, 맥주, 냉장음료, 용기보충금, 용기라면, 잡화균일가, 생수...}	{스낵류, 비스킷류, 초콜릿, 사탕/캔디, 과채음료, 기능성음료, 가공유, 봉지라면...}	{동물병원, 스낵류, 아이스크림, 맥주, 냉장음료, 비스킷류, 용기보충금, 용기라면...}
4	M000261625	{푸드코트, 스킨케어}	{남아특수소재의류, 파티/팬시용품, 여아의류특수목적의류, 잡화균일가, 용기보충금, ...}	{스킨케어, 스낵류, 식당, 여성의류상의, 베이커리, 와인, 그릇/식기, 바디케어,...}	{남아특수소재의류, 파티/팬시용품, 여아의류특수목적의류, 잡화균일가, 용기보충금, ...}

- 상품 추천 시뮬레이션 평가

- 고객별로 예측된 상품의 점수 순으로 상위 k개의 상품을 추천 대상으로 정의
- 추천한 k개의 상품 중, 실제 구매로 얼마만큼 이어졌는지 평가

- 평가 결과

재구매 타겟	0.33495469960011504
신규 구매 타겟	0.01864655565116105
전체 상품 고려하여 추천	0.13074859543452988

- 이미 한 번 구매했던 상품을 대상으로 하여 추천해주었을 때, 평균 재현도는 약 33%
- 신규 구매를 대상으로 추천해주었을 때, 평균 재현도는 약 2%
- 전체 상품을 대상으로 추천해주었을 때, 평균 재현도는 약 13%
- 이를 통해 재구매 할 만한 상품을 추천해 주는 것이 새로운 상품을 추천해 주는 것보다 더 좋은 결과를 낼 것이라고 예상

V 평가 및 전개

1. 모델 발전계획 수립

가. 고객 특성을 고려한 군집 분석

현재 진행한 분석 과정은 상품 구매 데이터의 구매 횟수에 기반 한 것으로 고객의 개별적 특성을 고려한 분석이라고 하기에 부족함. 따라서 고객 개별적 특성인 나이, 성별, 지역, 재구매 했다면 재구매한 기간 등에 따른 분석이 필요한 것으로 보임.

나. 모델 평가

모델 평가하고 검증하는 부분이 부실하므로 모델을 선택한 수치적 근거와 시뮬레이션 결과에 대한 구체적 평가를 보완할 필요가 있음.

2. 프로젝트 평가 및 향후 계획

가. 프로젝트 평가

- 연관성 분석도 실행하였지만 이를 해석하는 능력이 부족하여 상품 연관 관계의 인사이트를 도출하지 못했음. 수치적인 결과뿐 만 아니라 결과에 대한 인사이트를 도출하여 흥미로운 결과를 내지 못한 것에 대해 아쉬움.
- 제공받은 데이터에 비해 실제 분석으로 이어진 데이터는 절반 이하. 데이터들의 관계를 파악하여 넓게 보고 이용할 줄 알아야 함.
- 초기 주제 선정에 ‘마케팅 전략 제안’을 포함하였으나, 결과적으로는 제외했으며 마케팅에 대한 지식이 필요함.

나. 향후 계획

- 데이터를 분석하는 것도 중요하지만, 분석 결과를 보고 인사이트를 도출하는 것 또한 중요함. 결과에 대한 해석을 할 수 있도록 해야 함.
- 기한이 제한 적인 프로젝트의 경우 일자별, 주차별 구체적인 계획을 세우는 것이 중요하다는 것을 느낌.

참 고 문 헌

[별첨 1]

- heehee's study note, “빅데이터 분석 방법론“, 2022-10-26,
<https://heehee-ds.tistory.com/entry/ADsP-2-1-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%84%EC%84%9D-%EA%B8%B0%ED%9A%8D%EC%9D%98-%EC%9D%B4%ED%95%B4-3-%EB%B9%85%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%84%EC%84%9D-%EB%B0%A9%EB%B2%95%EB%A1%A0>
- 위키백과, “CRM“, 2022-10-26,
https://ko.wikipedia.org/wiki/%EA%B3%A0%EA%B0%9D_%EA%B4%80%EA%B3%84_%EA%B4%80%EB%A6%AC
- 남기백, 박상원, (2017), 머신러닝 기반 고객 재구매 상품 예측, 2017년 추계학술발표대회 논문집, 제 24권(제 2호), 421-423
- Data Makes Our Future, “파이썬 그룹 연산“, 2022-11,
<https://data-make.tistory.com/139>
- Everly's Data&Life, “쇼핑몰 데이터 분석“, 2022-11, <https://suy379.tistory.com/21>
- 내공남남 기술 BLOF, “파이썬 연관성 분석“, 2022-11, <https://hezzong.tistory.com/23>
- 어쩐지 오늘은, “비즈니스 이해“, 2022-11,
<https://zzsza.github.io/diary/2020/08/02/how-to-study-business/>

[별첨 2]