

머신러닝을 이용한 온라인 구매전환 예측에 관한 연구

연구배경

예측문제와 머신러닝

- 예측은 마케팅 의사결정의 전제조건임(Politz and Deming, 1953)
- 예측의 정확도가 높아질 수록 ROI가 증가됨(e.g., Lambrecht and Tucker, 2013)
- 머신러닝은 예측 문제 해결에 강점을 가지는 것으로 보고됨
 - ✓ 주식시장의 가격예측(Göçken et al., 2016), 의학 분야의 암 탐지(Cruz et al., 2006), 금융 부문의 신용등급(Tsai and Chen, 2010), 소비자 선호 예측(Huang and Luo, 2016)

머신러닝 적용의 잠재적 영역 “온라인 고객 행동 예측”

- 오프라인과 달리 유입경로 및 다양한 클릭 스트림 데이터를 식별할 수 있음
- 머신러닝을 활용한 온라인 소비자 행동의 예측은 고객생애가치 증가의 새로운 기회가 될 수 있음(Gupta et al., 2002; Libai et al., 2020)
 - ✓ 본 연구는 고객 행동 중, 핵심 성과변수인 “구매예측”에 초점을 맞추고자 함

연구배경

온라인 구매예측 문헌과 연구문제

- 선행연구자는 소비자 구매예측을 위해 *퍼널모델*(funnel model)을 적용해옴 (Hoban and Bucklin, 2015; Howard and Sheth, 1969; Vakratsas and Ambler, 1999)
 - ✓ 온라인 구매행동 예측을 위해, *페이지 뷰*(Bellman et al., 1999), *체류시간*(Lin et al., 2010), *행동의 순서*(Gündüz & Özsü, 2003; Lo et al., 2016), *검색 다양성*(Goldstein et al., 2017; Chan et al., 2014), *스크롤링 행동*(Guo and Agichtein, 2016)을 연구함
- MSI Research Priorities Tier 1(2020-2022): “What is the **most effect way** to conduct **account-based marketing** in the face of **new online technologies**?”
 1. 대부분 *고객이탈(churn)* 예측에 초점을 맞추었을 뿐(e.g., Burez and Van den Poel 2009; Vafeiadis et al., 2015), 구매예측에 대한 연구는 제한적임
 2. 온라인 구매예측에 적합한 *머신러닝 모델*이 무엇인지에 관해 충분히 논의되지 않음
 3. 온라인 구매예측에 적합한 *데이터 샘플링 방법*에 대한 분석이 이루어지지 않음
 4. 실무적 관점에서 머신러닝 모델의 *결과를 해석할 수 있는 방법*에 대한 연구가 충분하지 않음

이론적 배경

온라인 소비자 구매행동과 관련된 연구

- 소비자의 온라인 *클릭 스트림* 데이터와 *구매 전환* 간의 관계를 주로 전통적인 회귀분석을 통해 분석함

예측변수	연구자
체류 시간	• Bellman et al.(1999), Johnson et al.(2004)
페이지 뷰	• Huang(2009)
검색 깊이	• Johnson et al.(2004)
검색 다양성	• Goldstein et al.(2017)
마우스 스크롤	• Guo and Agichtein(2016)
이동 패턴	• Clark et al.(2006), Montgomery et al.(2004), Ting et al.(2007),Ting et al.(2009)

- *머신러닝*을 활용한 온라인 소비자 행동 예측은 잠재력에 비해 충분히 연구되지 않았음(Gupta et al., 2020; Labai et al., 2020)

이론적 배경

마케팅 맥락의 머신러닝 연구

- "명시적으로 프로그램이 작성되지 않아도 컴퓨터가 스스로 학습할 수 있는 능력을 제공하는 학문" (Samuel, 1959)
 - ✓ 머신러닝 활용의 목적과 효과는 예측 정확도의 개선에 있음(Agrawal et al., 2018)
- 연구자들은 머신러닝을 활용하여 마케팅 문제를 해결하고자 함
 - ✓ 마케팅 문제 중 예측(e.g., Cui and Curry, 2005; Vermeer et al., 2019), 특성 추출 (Netzer et al., 2008; Tirunillai and Tellis, 2014), 기술적 해석(Netzer et al., 2012; Trusov et al., 2016), 처방적 분석(Hauser et al., 2009; Schwartz et al., 2017)에 적용함
 - ✓ 모델 측면에서는 SVM(Support-vector machine; Cui and Curry, 2005; Evgeniou et al., 2005), 토픽 모델링(Topic modeling; Tirunillai and Tellis, 2014), 앙상블 트리(Ensemble trees; Guo et al., 2018; Yoganarasimhan, 2018), 딥러닝(Ballestar et al., 2018; Li et al., 2019; Liu et al., 2019; Zhang et al., 2017) 등을 적용함

이론적 배경

마케팅 맥락의 머신러닝 연구

- *SVM* – 마케팅에 적용된 초기 머신러닝 모델
 - ✓ Cui and Curry (2005)는 SVM이 다항 로짓모델 보다 성과가 뛰어나다는 점을 발견함
 - 머신러닝 모델이 예측적인 성과는 매우 우수하지만 설명변수와 예측결과 간의 관계를 설명할 수 없다는 한계점을 지적함
 - ✓ Luo(2016)은 fuzzy SVM (Lin and Wang, 2005) 능동 학습 알고리즘을 개발함
- *딥러닝* - 텍스트 및 이미지 데이터를 분석에 활발히 적용
 - ✓ Liu et al.(2018)은 소비자 리뷰를 분석하여 미학(Aesthetic)과 가격 콘텐츠가 전환에 영향을 미친다는 결과를 보고함
 - 비슷하게, Zhang et al.(2017)은 이미지가 Airbnb의 숙박 수요에 미치는 영향을 분석함, CNN(Convolution neural network)을 사용하여 이미지를 고품질 또는 저품질로 분류했으며 고품질 사진이 숙소의 수요를 증가시킨다는 결과를 확인함
 - ✓ Hartmann et al.(2019)는 텍스트의 감성을 분류하는 방법론으로 머신러닝 모델 5개의 성능을 비교함(SVM, Random forest, Naive Bayes, ANN, KNN)

이론적 배경

마케팅 맥락의 머신러닝 연구

- 앙상블(Ensemble) – 메타 학습으로 정확도가 높음

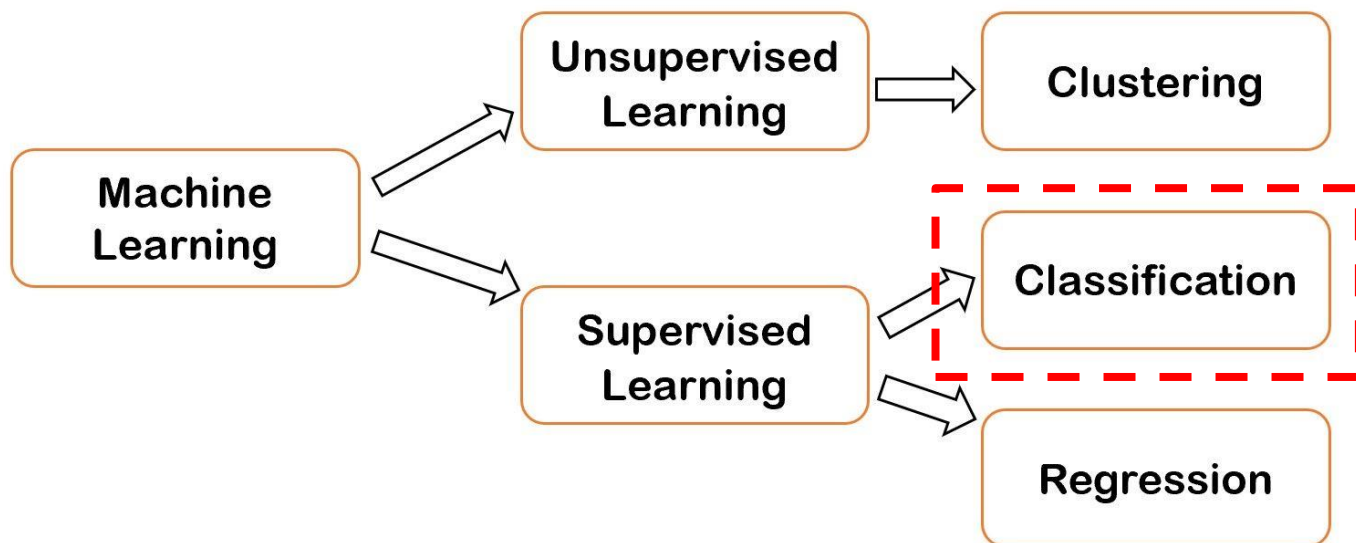
("Ensemble model is one of the most robust and widely used algorithm", Maik and Hussain, 2020)

- ✓ 일반적인 방법으로 스택킹(Stacking), 배깅(Bagging), 부스팅(Boosting)이 활용됨
 - 스택킹은 개별 설명변수의 선형조합을 사용하여 정확도를 높임(Breiman, 1996), 배깅은 각각의 개별 학습 트리가 부트스트랩 샘플을 활용하여 학습하며, 개별 학습 트리의 예측이 집계되어 최종 예측결과를 산출함, 부스팅은 개별 학습 트리가 순차적으로 훈련되고 각각의 정확도에 따라 학습 트리를 생성함
 - 배깅의 경우 병렬로 학습하며 부스팅은 순차 적으로 학습하고 학습이 끝난 후, 가중치를 부여함
- ✓ 랜덤 포레스트(Breiman, 2001)와 Gradient-boosted tree(Fridman, 2001) 대표적임
 - 랜덤 포레스트는 개별 트리가 원본 데이터의 부트스트랩 샘플을 통해 구축되며, 각 분할된 트리는 상관관계를 줄이기 위해 입력변수를 무작위로 할당됨
 - GBM은 여러 트리가 순차 적으로 훈련하며, 각 트리는 이전에 적용된 트리의 오류를 줄여서 정확도를 높임, 희소성을 인식하는 XGBoost는 Kaggle의 데이터 사이언스 대회에서 우승함

연구문제

연구문제1 : 온라인 소비자 구매를 예측하는데 가장 적합한 머신러닝 모델은?

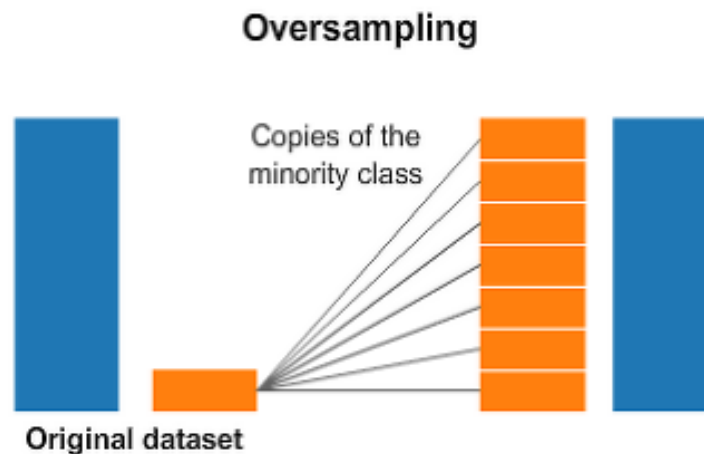
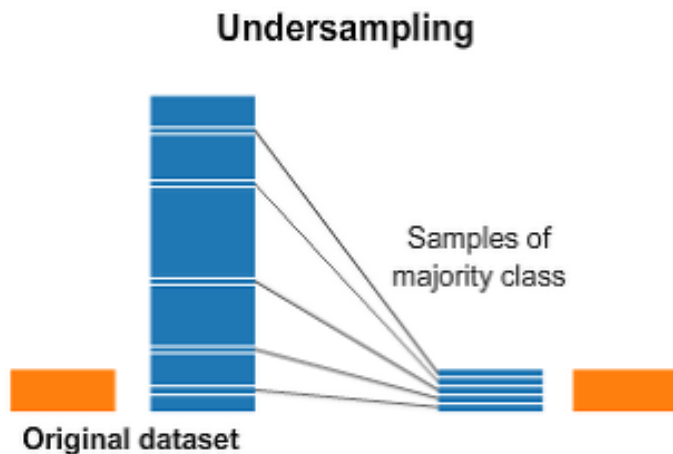
- 선행연구는 *SVM*(Support-vector machine; Cui and Curry, 2005; Evgeniou et al., 2005), *토픽 모델링*(Topic modeling; Tirunillai and Tellis, 2014), *앙상블 트리*(Ensemble trees; Guo et al., 2018; Yoganarasimhan, 2018), *딥러닝*(Ballestar et al., 2018; Li et al., 2019; Liu et al., 2019; Zhang et al., 2017) 등을 적용함



연구문제

연구문제2 : 온라인 소비자 구매를 예측하는데 가장 적합한 샘플링 방법은?

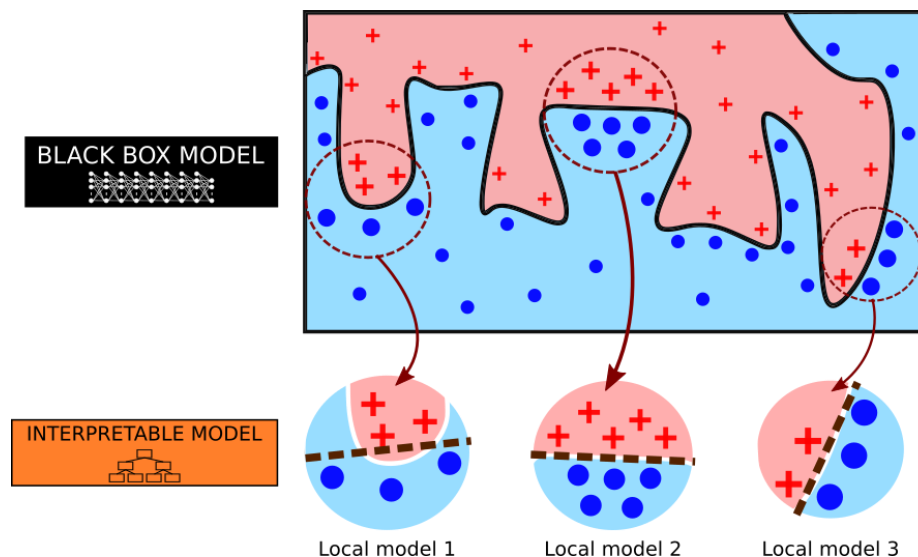
- 머신러닝 분류 알고리즘은 클래스 간에 데이터가 고르게 분포된 것으로 가정하기 때문에, 불균형 데이터의 경우 예측성도가 하락함(Garca-Pedrajas et al., 2010)
 - ✓ 데이터 불균형을 완화하는 방법으로 *내부 알고리즘*과 데이터에 대한 *외부 접근법*이 제시됨(Sun et al., 2007; Garca and Herrera, 2009; Freund and Schapire, 1996)
- 샘플링 방법은 *과대표집*(Over-sampling)과 *과소표집*(Down-sampling)으로 구분됨



연구문제

연구문제3 : 머신러닝 예측결과를 해석할 수 있는가?

- Cui and Curry (2005)이 지적한 것과 같이, 머신러닝 모델은 예측변수와 결과 간의 관계를 해석하는 데 어려움이 있음
 - ✓ 해석 가능성의 문제는 머신러닝에 대한 사용자의 신뢰도에 부정적인 영향을 미치며, 머신러닝 예측결과의 활용에도 장애가 됨(Rai, 2020)
 - ✓ 머신러닝의 알고리즘은 구조 및 학습 메커니즘이 복잡하여 해석하기 어려운 것이 일반적이며, 그 결과 예측결과 활용 가능성이 감소함(Hall and Gill, 2019; Du et al., 2018)



선행연구와의 차별점

마케팅 맥락의 머신러닝 연구

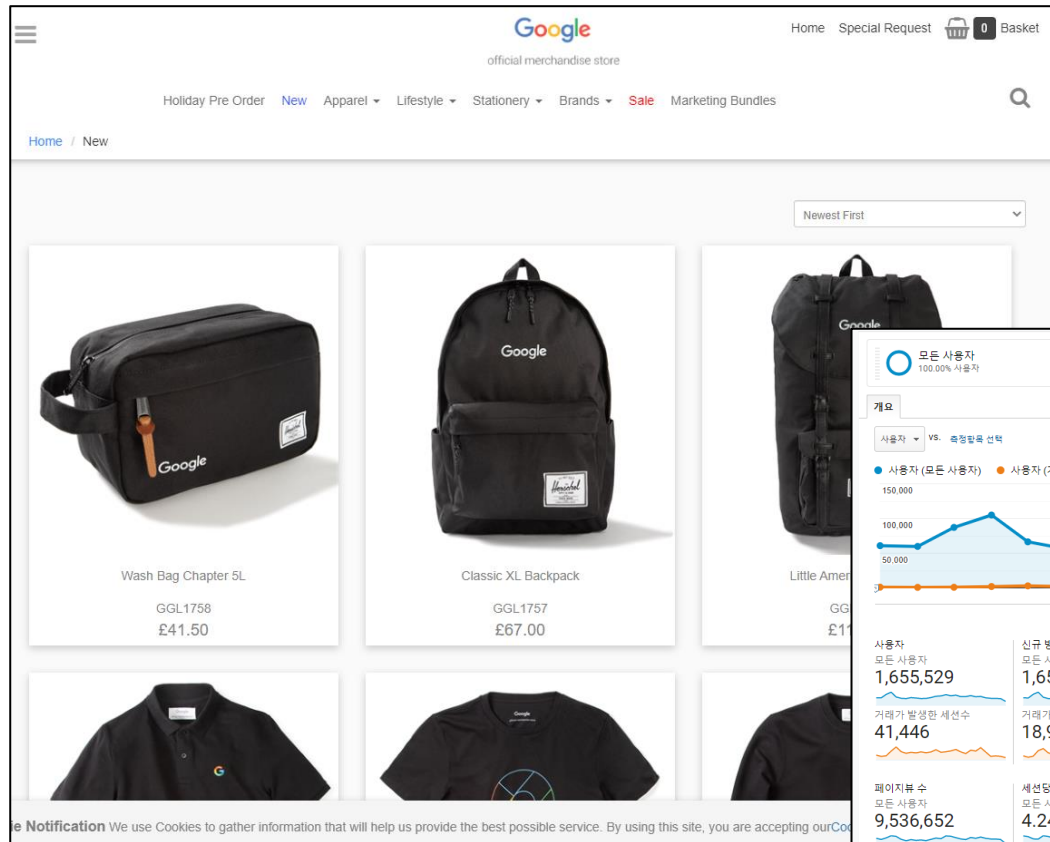
- 마케팅 예측문제에 머신러닝을 적용한 연구 및 본 연구와의 비교

연구자	모델 비교	샘플링 방법 비교	해석 가능성 적용	모델	주요 결과
Cui and Curry (2005)	-	-	-	SVM	• 실험을 통해 다양한 마케팅 예측환경에서 전통적인 예측 모델보다 머신러닝 기반의 SVM이 예측성도가 뛰어나다는 것을 검증함
Huang & Luo (2016)	-	-	-	Fuzzy SVM	• Fuzzy SVM을 통해 복잡한 특성의 제품에 대한 소비자의 선호를 분석할 수 있는 프레임워크를 제안함(adaptive decompositional framework)
Jacobs et al. (2016)	-	-	-	LDA, MDM	• LDA(Latent Dirichlet allocation)와 MDM(Mixtures of Dirichlet-Multinomials)을 활용하여 고객이 이후에 어떤 제품을 구매할지 예측 할 수 있는 방법론을 제시함
Miguéis et al. (2017)	-	Yes	-	Random Forest	• 은행 산업을 대상으로 다이렉트 마케팅에 대한 소비자 반응예측에 대해 샘플링 방법에 대한 성과 비교를 함
Ballestar et al. (2019)	-	-	-	MLP ANN	• Multilayer Perceptron (MLP) Artificial Neural Network (ANN)을 활용하여 E커머스 소셜 네트워크를 대상으로 고객 품질을 예측함
Hartmann et al. (2019)	Yes	-	-	5가지 모델	• 텍스트의 감성을 예측하는 방법론으로 머신러닝 모델 5가지를 비교함(SVM, Random forest, Naive Bayes, ANN, KNN)
본 연구	Yes	Yes	Yes	8가지 모델	• 온라인 구매 예측에 머신러닝을 적용할 때, 가장 적합한 모델, 샘플링 방법, 해석가능성 여부 등을 분석함

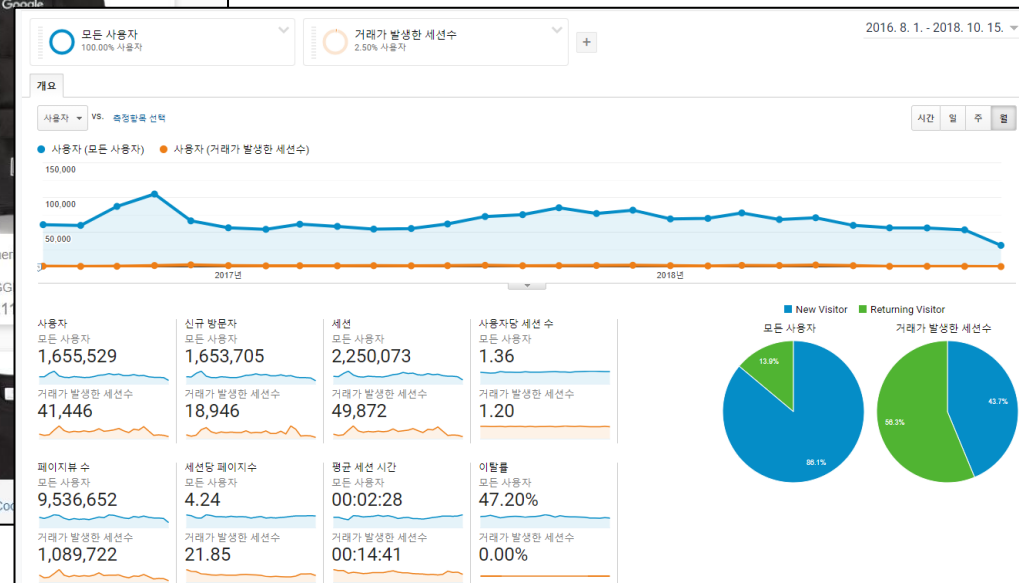
데이터

구글 머천다이즈 스토어 데이터 (2016. 8~2018. 4)

- 구글 머천다이즈 스토어는 전 세계 소비자를 분석하여 일반화 할 수 있는 결과를 도출하는 데 적합함(e.g., Kakalejčík et al., 2019)



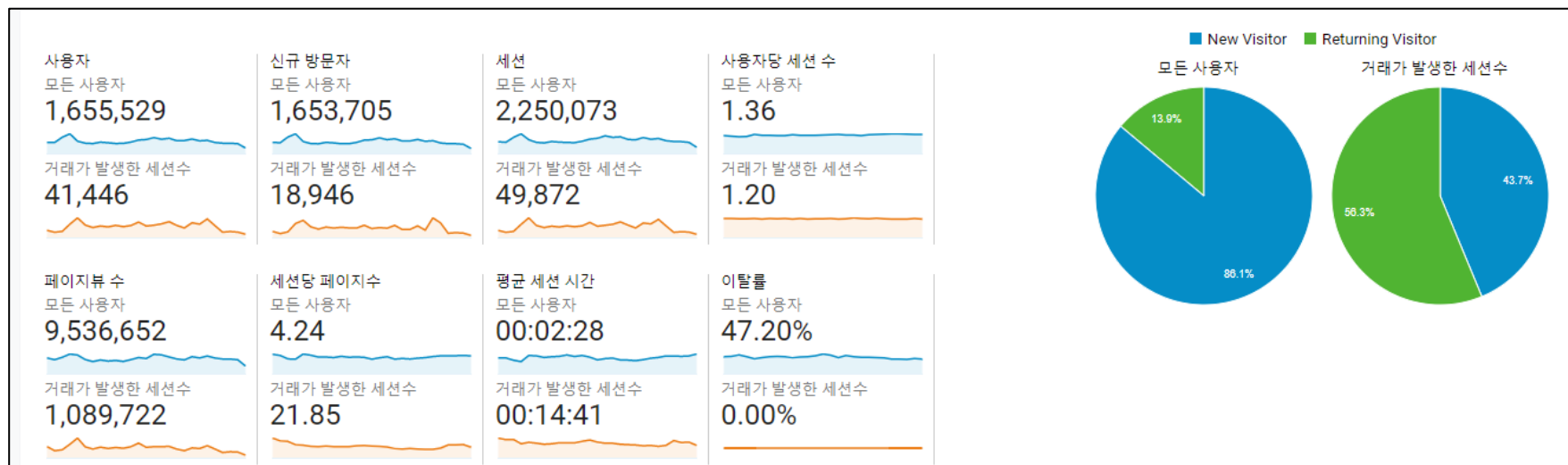
<https://shop.googlemerchandisestore.com>



데이터

데이터 특성

기본 웹사이트 프로파일		유입경로에 따른 사용자		성과	
기간	2016.8~2018.4	자연검색	759,171	수익	\$9,203,274.38
사용자	1,668,539	소셜미디어	363,521	전환율	2.29%
신규 사용자	1,648,685	직접방문	281,996	거래 수	51,498
세션	2,250,073	추천방문	184,878	객단가	\$178.71
사용자당 세션 수	1.35	검색광고	44,078		
세션당 페이지 수	4.22	디스플레이 광고	40,635		
이탈률	47.23	어필리에이트	35,926		
평균 세션 시간	2분 27초	기타	105		
모바일 비중	27.59%	전체	3,448,502		



데이터

데이터 특성

측정 지표	정의
세션	• 특정 기간에 사용자에게서 발생한 상호작용 또는 조회의 집합
사용자	• 특정 기간에 콘텐츠를 조회하거나 이용한 개별 디바이스(사용자)의 수
신규 사용자 수	• 특정 기간에 최초로 웹사이트에 방문한 사용자의 수
사용자당 세션 수	• 사용자당 발생한 평균 세션의 수
세션당 페이지 수	• 세션당 발생한 평균 페이지 조회 수
이탈률	• 페이지의 모든 페이지뷰에 대해 세션의 마지막이었던 비율
평균 세션 시간	• 모든 세션의 전체 시간(초)을 세션수로 나눈 값
모바일 비중	• 사용자 중 모바일 디바이스를 사용하여 접속한 사용자의 비율
자연검색	• 검색엔진에서 유료 광고가 아닌 다른 영역을 클릭하여 접속하는 경우
소셜미디어	• 소셜네트워크 서비스를 통해 유입되는 경우
직접방문	• 사용자가 웹사이트 URL의 이름을 브라우저에 입력하거나 북마크를 통해 사이트를 방문한 세션
추천방문	• 다른 사이트 혹은 특정링크를 통해 유입되는 경우(소셜네트워크 제외)
검색광고	• 검색광고를 클릭하여 접속하는 경우
디스플레이 광고	• 디스플레이 광고를 클릭하여 접속하는 경우
어필리에이트	• 제휴 마케팅을 통해 접속한 경우

데이터

데이터 특성

변수명	비 전환	전환	F
신규방문	.76	.39	14,600**
방문횟수	2.31	4.02	616.1**
페이지뷰	3.44	26.91	281,045**
체류시간(초)	113.99	1,022.44	129,741**
자연 검색	.43	.29	1,398**
소셜미디어	.20	.01	4,531**
직접 방문	.15	.18	58.17**
추천 방문	.11	.46	20,192**
검색 광고	.02	.03	98.06**
디스플레이	.03	.01	230.7**
어필리에이트	.01	.00	334.5
기타	8.048395e-05	5.388221e-05	.162

데이터

소비자 여정 측정 - CDJ(consumer decision journey; Anderl et al., 2016)

변수명	Min	Max	Mean	Standard Deviation
전환 유무	0	1	.05	.31
방문 횟수	1	456	5.71	18.93
페이지 뷰	0	500	5.45	9.30
세션 품질	0	100	3.59	12.59
체류 시간(초)	0	19,017	223.50	520.13
이전 전환 수	0	21	.02	.16
유입 채널	Organic Search, Social Direct, Referral, Paid Search, Display, Affiliates, Other 8개			
브라우저 종류	Chrome, Safari, Edge, Samsung internet, Opera Mini, Muffin, Ye 등 33개			
운영체제 종류	Macintosh, iOS, Chrome OS, Windows, Tizen, Samsung, Xbox, Nintendo Wii 등 19개			
디바이스 종류	Desktop, Mobile, Tablet 등 3개			
접속국가	United States, Canada, Philippines, Mexico, India, Japan, Israel, Austria 등 196개			
접속도시	New York, California, Massachusetts, Gujarat, Taipei City, Tokyo 등 421개			
전체 변수 및 데이터	변수 : 686개, 데이터 수: 374,749개			

분석결과

머신러닝 성능 평가 방법

- 민감도, 특이도, PPV(Positive predictive value), NPV(Negative predictive value), Prevalence 활용(Esuli and Sebastiani, 2010; Vermeer et al., 2019)
 - 실제 양성(전환)을 양성으로 판정한 경우인 TP(True Positive)
 - 실제 음성(비전환)을 양성으로 판정한 경우인 FP(False Positive)
 - 음성을 음성으로 판정한 경우인 FN(False Negative)
 - 양성을 음성으로 판정한 경우인 True Negative(TN)

• 민감도(Sensitivity)	$\frac{TP}{TP + FN} \times 100$
• 특이도(Specificity)	$\frac{TN}{FP + TN} \times 100$
• 정확도(Accuracy)	$\frac{TP + TN}{TP + FP + FN + TN} \times 100$
• PPV(Pose Pred Value) Precision	$\frac{TP}{TP + FP} \times 100$
• NPV(Neg Pred Value) Recall	$\frac{TN}{FN + TN} \times 100$
• Prevalence	$\frac{TP + FN}{TP + FP + FN + TN} \times 100$

분석결과

연구문제1 : 온라인 소비자 구매를 예측하는데 가장 적합한 머신러닝 모델은?

- CARET(classification and regression training) - 복잡한 문제에 대한 모델 훈련(training)과 파라미터 조절(tuning) 과정을 간소화하는 함수

✓ 조율모수 후보 값을 5개로 설정하고 교차검증(K-fold cross validation)은 5로 설정

1	평가할 모형의 모수 집합을 정의
2	for 각 모수 집합에 대해 수행 do
3	for 각 재표본 반복에 대해 수행 do
4	· 특정 표본을 예비표본으로 남김
5	· [선택적] 데이터에 대한 전처리 수행
6	· 남겨진 표본으로 모형 적합
7	· 예비표본에 대해 예측 수행
8	end
9	예비표본 예측에 대한 평균성능 계산
10	End
11	최적의 모수집합 결정
12	최적의 모수집합을 사용하여 모든 훈련 자료에 대한 최종모형을 적합

분석결과

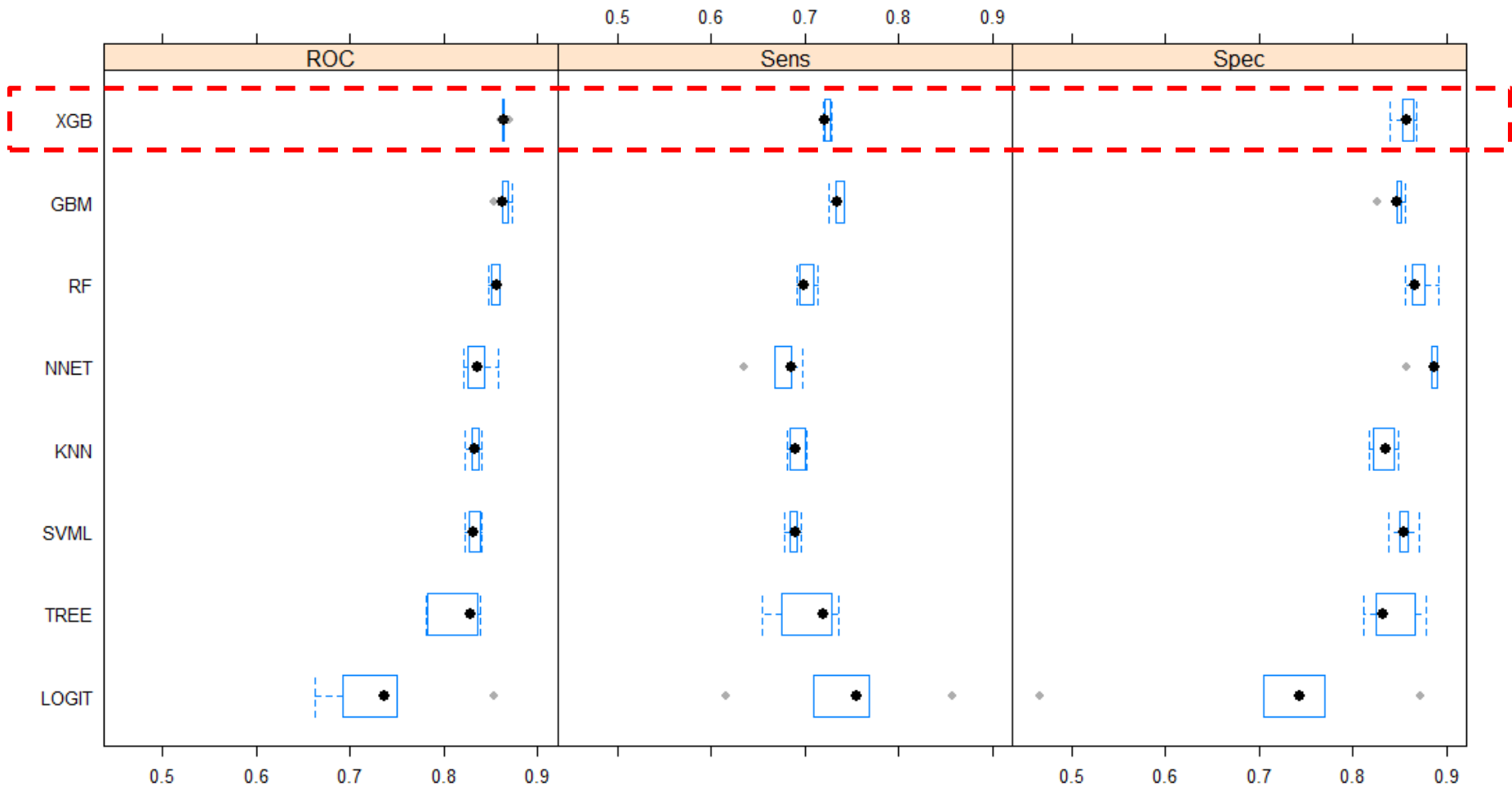
연구문제1 : 온라인 소비자 구매를 예측하는데 가장 적합한 머신러닝 모델은?

번호	머신러닝 알고리즘	ROC	민감도	특이도
1	Classification tree(TREE)	.8133	.7030	.8419
2	Artificial neural network(NNET)	.8367	.6744	.8815
3	k-nearest-neighbor(KNN)	.8323	.6913	.8332
4	Logistic regression(LOGIT)	.7386	.7410	.7113
5	Support vector machine with linear kernel(SVML)	.8318	.6881	.8542
6	Random forest(RF)	.8544	.7017	.8702
7	Gradient Boosting Algorithm(GBM)	.8640	.7356	.8451
8	eXtreme Gradient Boosting(XGB)	.8643	.7235	.8560

ROC : receiver operating characteristic; 신호탐지이론에서 개발된 지표로 오경보 확률 대비 적중 확률

분석결과

연구문제1 : 온라인 소비자 구매를 예측하는데 가장 적합한 머신러닝 모델은?



분석결과

연구문제2 : 온라인 소비자 구매를 예측하는데 가장 적합한 샘플링 방법은?

번호	샘플링 방법	데이터 수		
		전체	전환	비전환
1	원본 데이터	262,324	6,303	256,021
2	다운샘플링	12,606	6,303	6,303
3	오버샘플링(SMOTE)	264,726	132,363	132,363

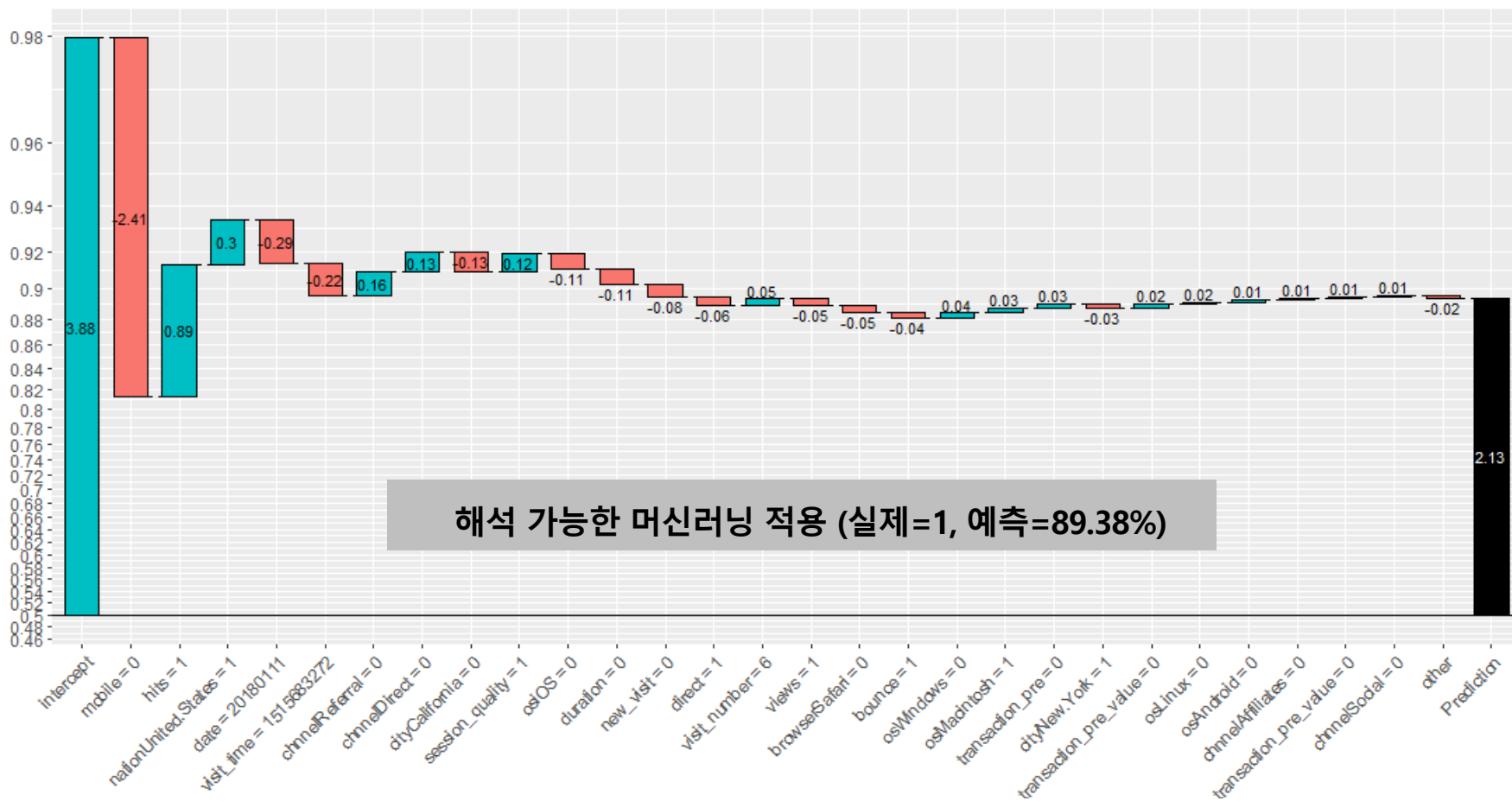
Data	원본데이터	다운 샘플링	오버 샘플링
Accuracy (95% CI)	.7348 (.7322, .7374)	.7324 (.7298, .7350)	.7417 (.7391, .7442)
민감도	.7320	.7296	.7392
특이도	.8501	.8444	.8430
Pose Pred Value	.9950	.9948	.9949
Neg Pred Value	.0714	.0704	.0727
Prevalence	.9763	.9763	.9763
Detection Rate	.7146	.7123	.7217
Detection Prevalence	.7182	.7160	.7254
Balanced Accuracy	.7910	.7870	.7911

분석결과

연구문제3 - 머신러닝 예측결과를 해석할 수 있는가?

- ✓ XGBoost 모델에서 개별 표본에 따라 각각의 설명변수가 어떤 비중으로 예측 값에 작용했는지 로그오즈(log odds)를 분석할 수 있음

$$\log(\text{Odds}) = \log\left(\frac{p}{1-p}\right)$$



분석결과

추가분석 - 리타겟팅(Re-targeting) 광고에 활용 가능성

- Display 광고를 통한 전환예측

Data	SMOTE Sampling
Accuracy (95% CI)	.7147 (.7120, .7173)
민감도	.7143
특이도	.7276
Pose Pred Value	.9908
Neg Pred Value	.0581
Prevalence	.9763

Note: 디스플레이 광고로 유입된 18,340개의 전체 데이터를 7:3 분할 후, Train data set 12,838개에 SMOTE 샘플링을 적용하여 23,885개를 훈련함, Test data set은 5,502개임

결론 및 시사점

이론적 시사점

- 온라인 마케팅 맥락에서 머신러닝을 효과적으로 적용하기 위한 방법에 대해 논의함으로써 온라인 마케팅 및 머신러닝 문헌에 기여함
 - ✓ MSI의 연구요청 "*What is the most effect way to conduct account-based marketing in the face of new online technologies?*" 에 답함
 - ✓ 온라인 예측 분야의 향후 머신러닝 연구에 대해 적절한 연구방법의 가이드를 제공함

실무적 시사점

- 본 연구를 온라인 구매예측을 위한 머신러닝 개발의 가이드로 활용할 수 있음
 - ✓ 온라인 구매예측에는 XGB 모델과 SMOTE 샘플링 방법이 적절하며, 편의적 방법인 과소표집은 성능이 가장 낮은 것으로 분석됨
- 머신러닝 모델을 통해 온라인 마케팅 투자의 ROI를 개선할 수 있음
 - ✓ 리타겟팅 광고 등 선제적 마케팅 기법의 효과예측을 통해 ROI를 높일 수 있음

결론 및 시사점

본 연구의 한계점 및 향후연구방향

- 다양한 산업과 특성의 온라인 쇼핑몰로 연구가 확대되어야 함
 - ✓ 본 연구는 전 세계 소비자가 이용하며, 선행 연구에서 활용된 머천다이즈 스토어 데이터를 활용하였으나, 쇼핑몰 특성에 따라 예측 성과에는 차이가 발생할 수 있음
- 필드실험을 통한 “반사실적 의존관계(counterfactual dependence)” 분석이 필요함
 - ✓ 본 연구는 반사실적 의존관계를 검증하지 않았기 때문에, 필드실험을 통해 머신러닝 예측결과를 실제로 적용함으로써 정확도와 ROI를 검증할 필요가 있음
- 더욱 많은 머신러닝 모델과 조율 모수의 비교가 이루어져야 함
 - ✓ , 머신러닝 모델은 수백개에 달하기 때문에 더 많은 모델 탐색이 필요함
 - ✓ 또한, 컴퓨팅 파워와 분석의 시간제약으로 인해 조율 모수의 후보 값을 5개로 설정하였으나, 로컬 최적화(local optimization)와 글로벌 최적화(global optimization)의 여부에 따라 성과차이가 발생할 수 있음

감사합니다.