자동차 보험 출시 후
효율적인 Lead 확보를 위한 머신러닝

2022. 05. 24. 황인환

어떤 상황인가?

건강보험을 성공적으로 운영하고 있는 우리 회사는 새롭게 **자동차 보험의 출시**를 앞두고 있다.

자동차 보험 출시의 목표

⇒ 시장 확장, 추가 확보를 통한 수익성 강화

어떤 상황인가?

새로운 상품을 출시할 때 중요한 건 무엇인가?

바로 Targeting! Targeting! Targeting!

⇒ 실제로 해당 상품에 관심이 있는 고객에게 더 집중하여,
저희가 가지고 있는 한정된 자원을 아주 효율적으로 사용하는 것

무엇을 풀고 싶은가?

[고객 영업팀의 현재 니즈]

- 불특정 다수를 대상으로 한 영업은 매우 어렵다.
- 특히 출시된지 얼마 안된 상품은 불확실성이 높아 더더욱 어렵다.
- **효율적인 영업 리소스 배분을 통해, 출시 직후 최대한 많은 리드를 끌어오고 싶다.**

무엇을 풀고 싶은가?

[이렇게 한 번 풀어보면 어떨까?]

이미 우리 회사의 건강보험에 가입해 있는 고객 중,

자동차 보험 가입에도 관심이 있는 고객을 찾아내고 그들에게 집중하자!

데이터 개요

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0

(약 38만 건)

[특성]

성별 / 면허 보유 여부 / 지역 / 이미 가입된 자동차 보험이 있는지 / 자동차 연식/ 사고 이력 / 프리미엄 보험료 등

[타겟]

'Response' ⇒ 관심 여부

개요 - 이 발표를 통해 여러분이 얻어갈 수 있는 것

1. 분석 - 데이터 탐색 결과 공유

- 분석한 데이터를 탐색하며 얻은 인사이트를 공유드립니다.
- **여러분의 풍부한 경험을 바탕으로 한 직관에 시너지를 드릴 수 있기를 기대합니다.**

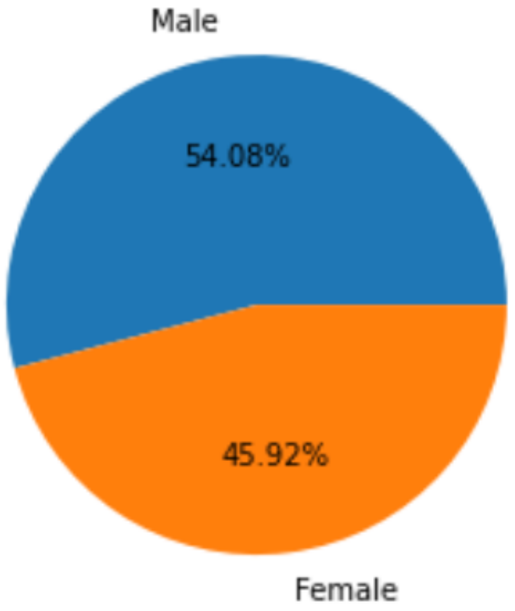
2. 예측 - 머신러닝 모델, 활용성 소개

- 제가 만든 머신러닝 모델을 소개하고, 어떻게 활용할 수 있을지 제안드립니다.
- **어렵지 않아요! 기술적인 부분은 가능한 배제하고 활용성 위주로 설명드리겠습니다.**

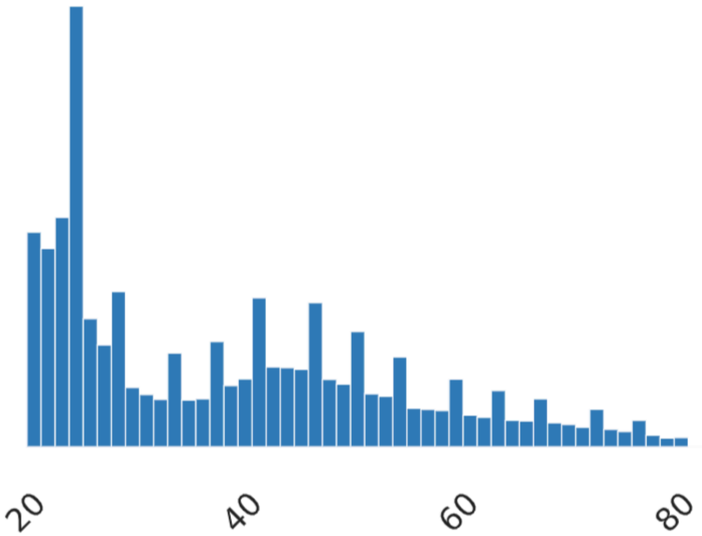
Part 1. 분석 - 데이터 탐색 결과 공유

Part 1. 분석 데이터 탐색 결과 공유

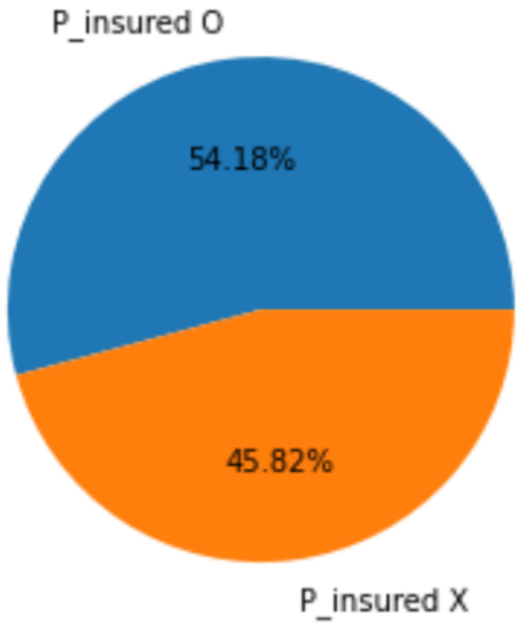
1. 특성별 분포 소개 (약 38만 건의 샘플에 대해)



성별



연령대

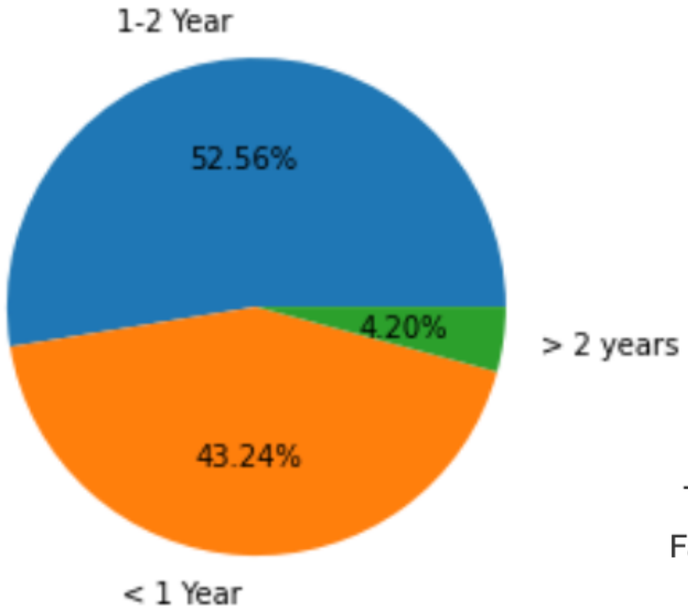


이미 가입된 자동차 보험이 있는지

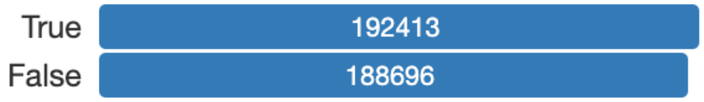
참고로 데이터 전처리(ex. 결측치 관리 등)은 데이터 탐색 결과 불필요하여 이 단계에서는 진행하지 않았고, 모델을 만들 때 인코딩만 해주었습니다.

Part 1. 분석 데이터 탐색 결과 공유

1. 특성별 분포 소개 (약 38만 건의 샘플에 대해)



자동차 연식



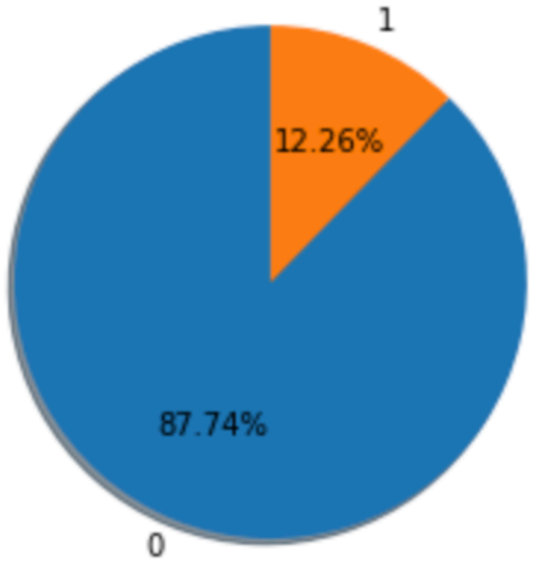
과거 사고이력

면허 보유 여부(!)

참고로 데이터 전처리(ex. 결측치 관리 등)은 데이터 탐색 결과 불필요하여 이 단계에서는 진행하지 않았고, 모델을 만들 때 인코딩만 해주었습니다.

Part 1. 분석 데이터 탐색 결과 공유

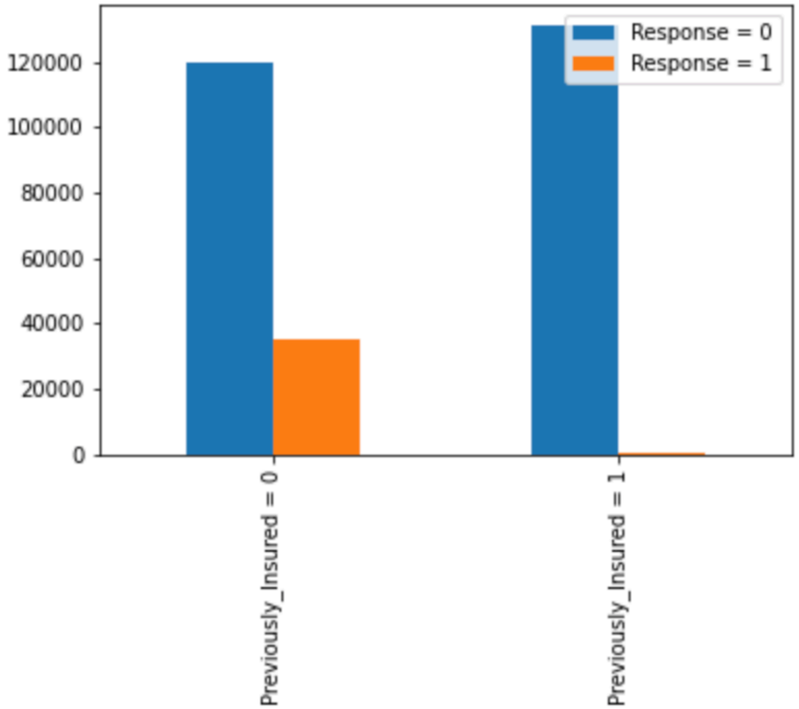
1. 특성별 분포 소개 (약 20만 건의 훈련 샘플에 대해)



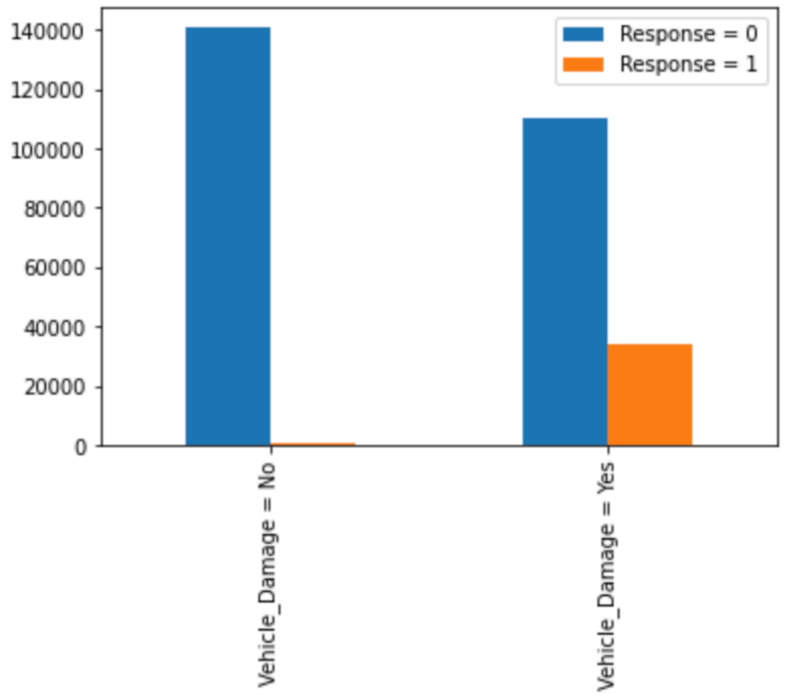
타겟 - 자동차 보험에 관심이 있는가?

Part 1. 분석 데이터 탐색 결과 공유

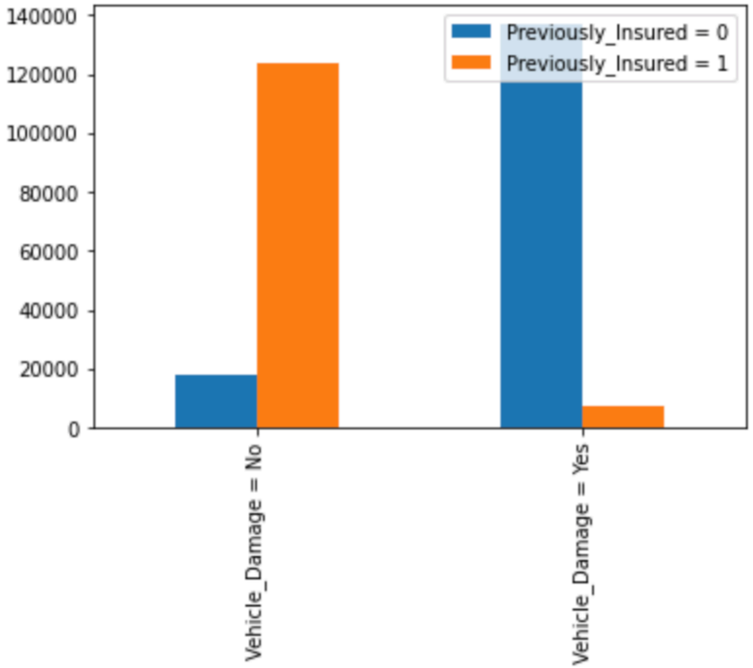
2. 특성과 타겟(관심 여부)간의 관계 소개 (약 20만 건의 샘플에 대해)



이미 자동차 보험에 가입한 여부 & 관심도



과거 사고이력 & 관심도



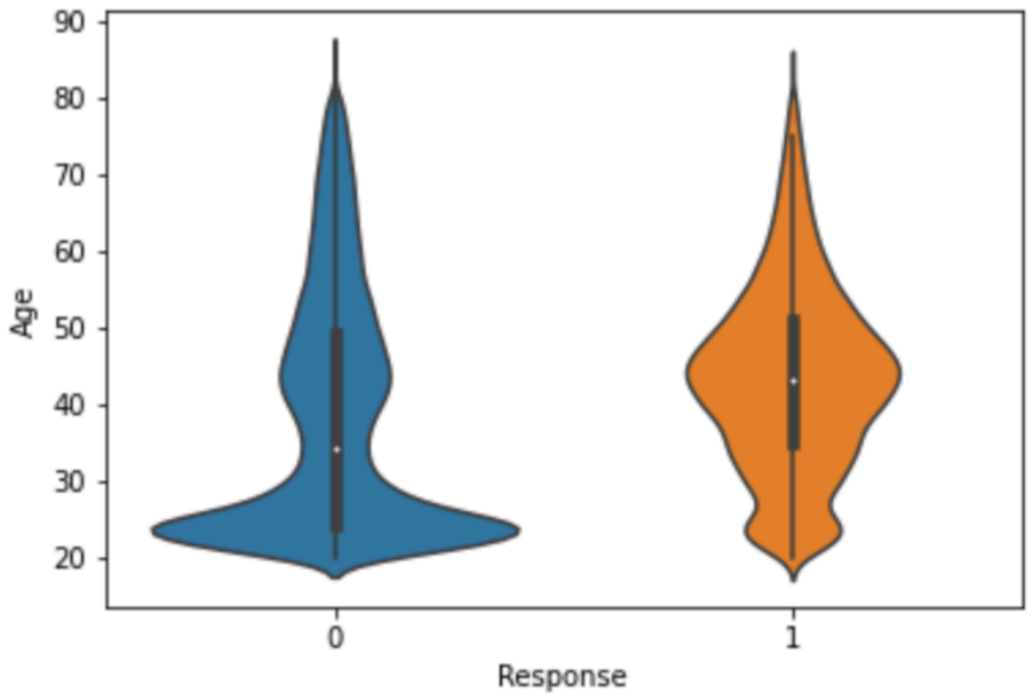
과거 사고이력 & 이미 자동차 보험에 가입한 여부

1)사고이력 있고, 2)이미 가입된 보험 없는 사람을 노리는게 좋겠다는 가설을 세워볼 수 있겠다.

Part 1. 분석 데이터 탐색 결과 공유

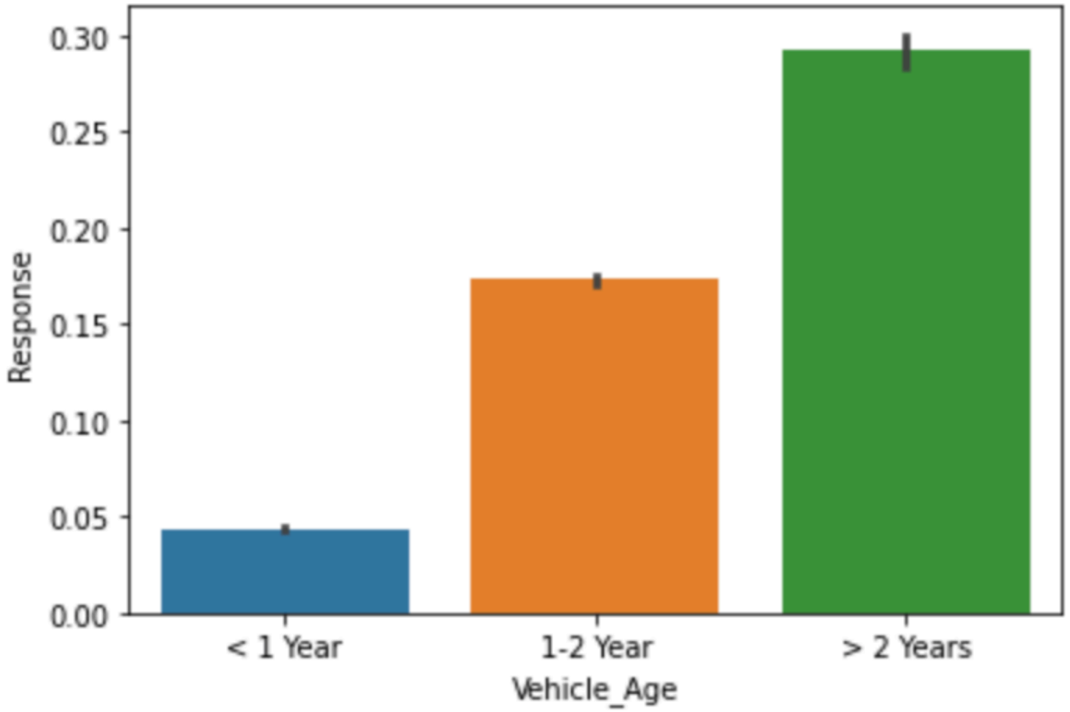
2. 특성과 타겟(관심 여부)간의 관계 가설

(약 20만 건의 샘플에 대해)



나이가 어릴수록 자동차 보험 가입에
관심이 없을 것이다.

⇒참일 가능성이 높아보임



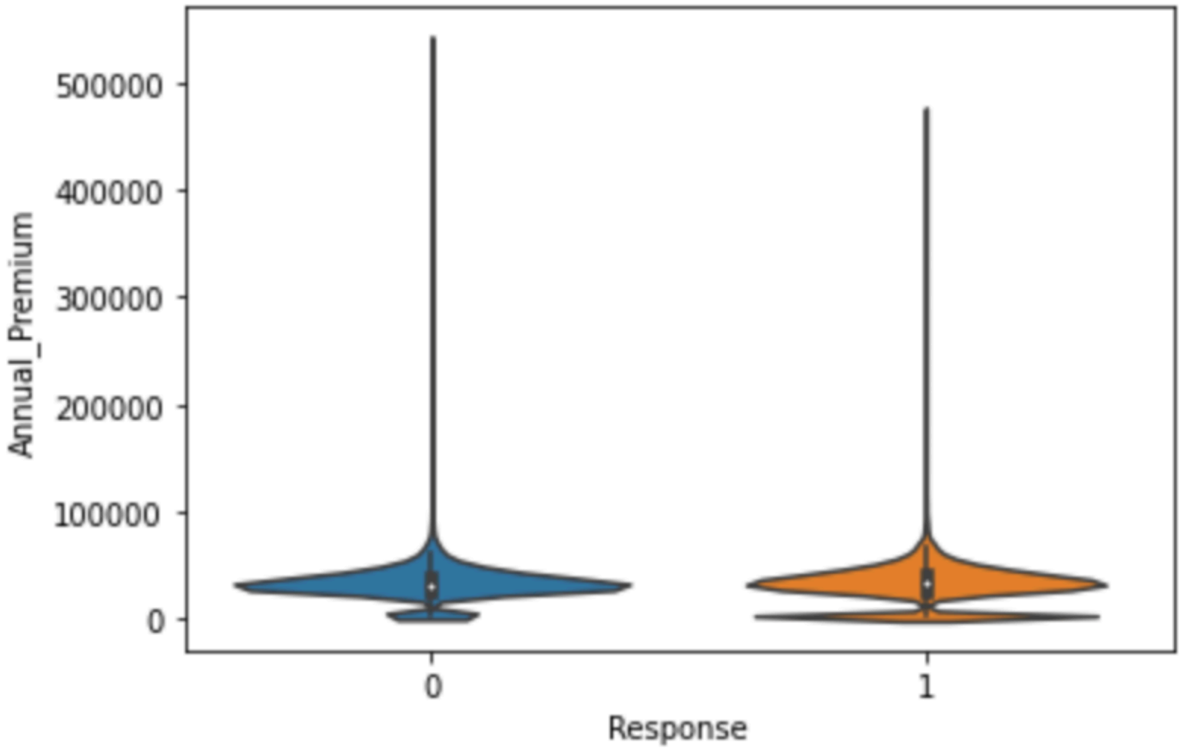
자동차 연식이 오래될 수록
자동차보험에 관심이 많을 것이다.

⇒참일 가능성이 높아보임

Part 1. 분석 데이터 탐색 결과 공유

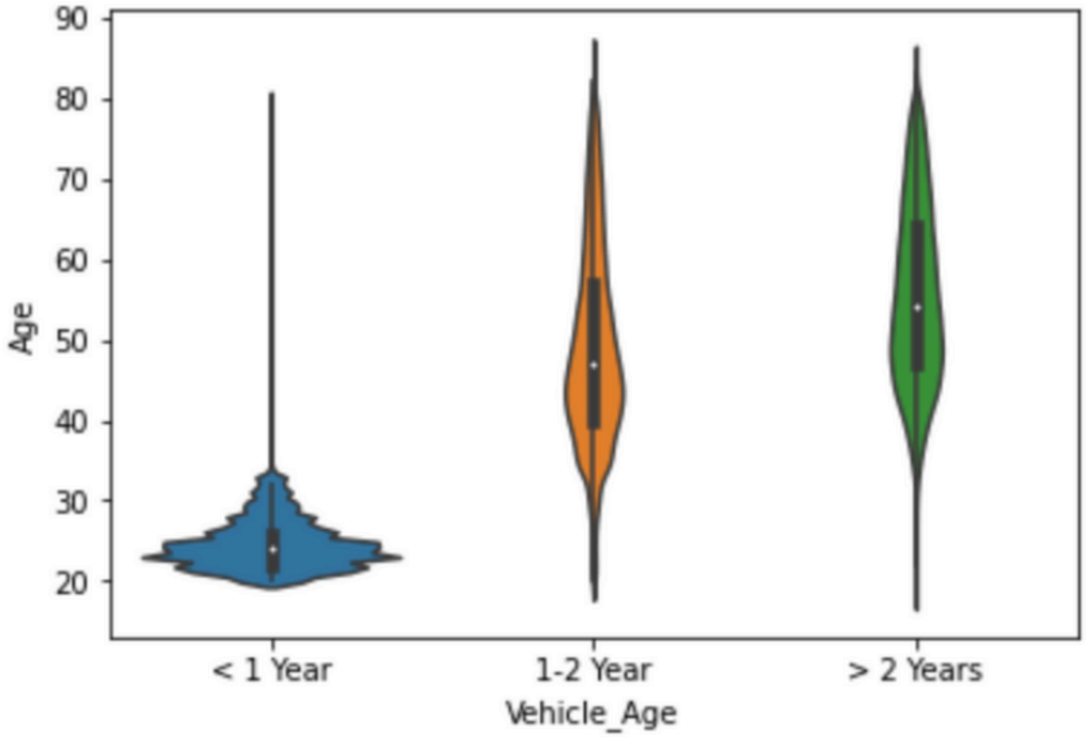
2. 특성과 타겟(관심 여부)간의 관계 가설

(약 20만 건의 샘플에 대해)



연간 보험료가 낮을수록
자동차 보험에 관심이 많을 것이다.

⇒ 경향이 아예 없진 않으나 대체적으로 뚜렷한 차이가 없어보임



참고 - 나이와 자동차 연식 관계

Part 2. 예측 - 머신러닝 모델, 활용성 소개

(!!) 우리는 고객 정보에 따라
자동차 보험 가입에 관심이 있는지 없는지를
예측하는 분류 예측 모델을 만드는 것을 목표로 합니다.

Part 2. 예측 - 머신러닝 모델, 활용성 소개

0. 머신러닝이란? (기본 프로세스)

머신러닝은 데이터와 정답을 넣으면 규칙을 찾아주는 도구입니다.

따라서 저도 주어진 데이터를 **훈련, 검증 데이터**로 나누어,

- 훈련 데이터를 통해 데이터와 정답(Response)를 학습하는 모델을 만들고,
- 검증 데이터를 통해 모델의 성능을 측정하는 프로세스를 밟았습니다.

(이 과정에서 훈련, 검증 데이터 중 필요한 특성에 대해 인코딩 및 기본적인 전처리를 진행함)

Part 2. 예측 - 머신러닝 모델, 활용성 소개

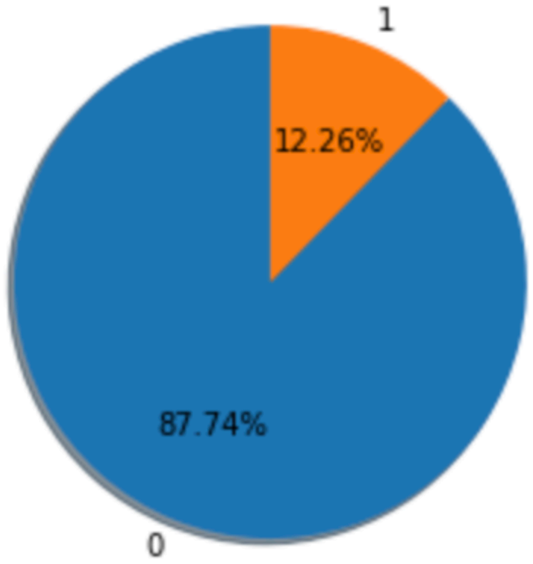
1. 기준 모델 - 가장 간단한 형태의 모델로 성능의 최저 기준이 됨.

- 평가 지표 : f1_score (타겟 데이터의 불균형으로 '정확도'는 사용이 제한됨)
- 기준 모델: 랜덤 포레스트 기본형 사용 (f1_score를 통해 이후 모델과 비교를 위함)

	precision	recall	f1-score	support
0	0.88	0.99	0.93	62700
1	0.41	0.06	0.10	8758
accuracy			0.87	71458
macro avg	0.64	0.52	0.52	71458
weighted avg	0.82	0.87	0.83	71458

Part 2. 예측 - 머신러닝 모델, 활용성 소개

2. 본격적인 모델링 …..



타겟 - 자동차 보험에 관심이 있는가?

타겟 분포가 불균형한 경우 예측 모델을 만들 때 주의해야 합니다.

따라서 여러 모델을 만들어 보고 가장 성능이 좋은 모델을 찾아야 합니다.

RandomForest.. RandomSearchCV, GridSearchCV
XGboost, Downsampling …..

Part 2. 예측 - 머신러닝 모델, 활용성 소개

3. 채택한 모델 소개

1. DownSampling (타겟 분포 불균형 해소) ⇒ RandomSearchCV를 통한 하이퍼 파라미터 최적화
2. 훈련 데이터에서 교차 검증을 해본 결과 최고의 성능 지표는 0.82(f1_score)가 나왔습니다.
3. 테스트 데이터에서는 다음과 같은 결과를 최종적으로 얻었습니다.

	precision	recall	f1-score	support
0	0.99	0.67	0.80	83600
1	0.28	0.93	0.43	11678
accuracy			0.70	95278
macro avg	0.63	0.80	0.61	95278
weighted avg	0.90	0.70	0.75	95278

Part 2. 예측 - 머신러닝 모델, 활용성 소개

3. 채택한 모델 성능 소개

	precision	recall	f1-score	support
0	0.99	0.67	0.80	83600
1	0.28	0.93	0.43	11678
accuracy			0.70	95278
macro avg	0.63	0.80	0.61	95278
weighted avg	0.90	0.70	0.75	95278

Recall이 0.93이라는 건 자동차 보험에 관심이 있지만, 모델이 잘못 예측할 확률이 0.07밖에 되지 않는다는 것.

- 저는 precision (정밀도)보다 이 Recall 수치를 좀 더 중요하다고 생각했는데,
  진짜 원하는 고객을 놓치지 않고 저희 리소스를 배분하는 것이, 한정된 자원에서 최대한의 퍼포먼스를 내는 것이 중요하다고 보기 때문입니다. (+ 출시 직후 폭발적인 유입을 늘리는 것이 차후 유지에도 유리함)

Part 2. 예측 - 머신러닝 모델, 활용성 소개

3. 채택한 모델 소개 - Permutation Importances

Gender_Male	-0.000912
Gender_Female	-0.000803
Vintage	-0.000249
Driving_License	0.000461
Region_Code	0.002135
Policy_Sales_Channel	0.010614
Vehicle_Age	0.015894
Age	0.019057
Previously_Insured	0.078041
Vehicle_Damage	0.088413

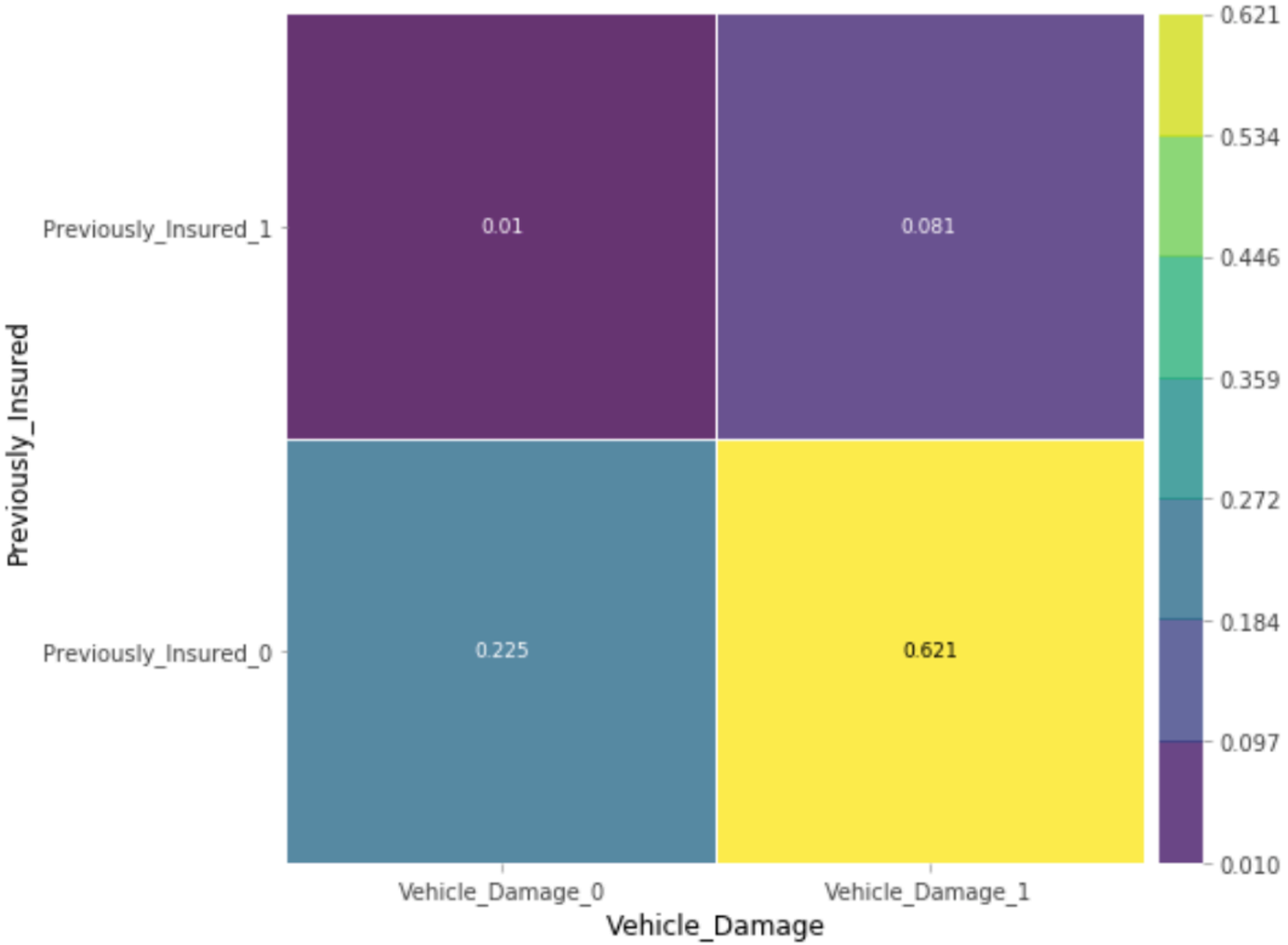
Permutation Importances란 모델이 잘 예측하는데 있어 각각의 특성이 얼마나 중요한지를 측정해줍니다.

위 수치를 보면 제가 만든 모델이

과거 사고이력 여부, 이미 가입된 자동차 보험이 있는지 여부를 가장 중요한 특성으로 판단한 것을 알 수 있습니다.

Part 2. 예측 - 머신러닝 모델, 활용성 소개

3. 채택한 모델 소개 - PDP



위 두 가지 특성을 가져와서 타겟과의 관계를 보면,

사고이력이 있고, 가입된 자동차 보험이 없는 경우

자동차 보험 가입에 관심도가 올라간다는 걸 알수 있습니다.

데이터 탐색 당시 세웠던 가설이 맞을 가능성이 높다는 걸 모델링을 통해서도 확인해볼 수 있었습니다.

"1)사고이력 있고, 2)이미 가입된 보험 없는 사람을 노리는게 좋겠다는 가설을 세워볼 수 있겠다."

Part 2. 예측 - 머신러닝 모델, 활용성 소개

4. 채택한 모델의 활용 제안

1.
데이터셋에 없는 나머지 기존 고객에게도 모델을 적용하여 효율적인 출시 초기 영업 리소스 분배하는데 활용.

2.
앞으로 건강보험에 가입하려는 신규 잠재 고객들을 대상으로 관련 정보도 함께 받아, 자동차 보험 가입 가능성이 높을시 동시에 가입 유도하는데 활용.

끝.

감사합니다.