데이터 증강 및 통계적 가설 검정을 통한 불균형 보험 데이터 예측

TEAM | 일등(최)해원

목 차

팀원 소개

01

01 팀원 소개

역할 분담

17 김보경

- 평가지표 비교 및 분석
- **Fisher Kim**

20 김현우

- 진행 총괄
- 실험 코드 작성
- **Drunken Coder**

18 이준희

- 데이터 전처리 및 보고서 작성
- **Emperor of MT**

20 임예림

- 평가지표 비교 및 PPT 제작
- **Mom of Stat, UP!**

20 최해원

- 데이터 전처리 및 PPT 제작
- **The Kidnapped**

주제 선정 이유

02

02 주제 선정 이유

문제 배경

Paradox of Accuracy

Accuracy는 불균형 데이터에 대해 단순히 다수 범주로 통일하여 예측하여도 상당히 높은 수치 기록

데이터 품질이 예측의 성능을 좌우하는 경향성을 불균형 데이터에서는 찾아보기 어려움

신용카드 사기 거래와 같이 개인에게 심각한 악영향을 끼치는 실생활 사례의 경우 소수 경우를 정확하게 예측하여 사기 거래 피해자를 보호하는 것이 중요

02 주제 선정 이유

주제 선정 이유

기존 해결방안

Accuracy 뿐만 아니라 Recall, Precision, F1 score을 사용하여 다각화된 분석 진행

다양한 평가지표를 통해 분석의 완결성을 더하는 것은 좋은 해결 방안이지만, 다분히 사후적

본 분석의 방향성

- 선정 데이터 : 보험 데이터 (불균형)
- 불균형 데이터의 문제를 근본적으로 해결할 수 있는 데이터 증강 기법 다수 적용
- 단순히 평가지표 수치 해석에만 그치지 않고 통계적 가설 검정을 통해 실질적으로 기존 평가지표 활용의 문제점에 대한 해결책 제시

데이터 설명

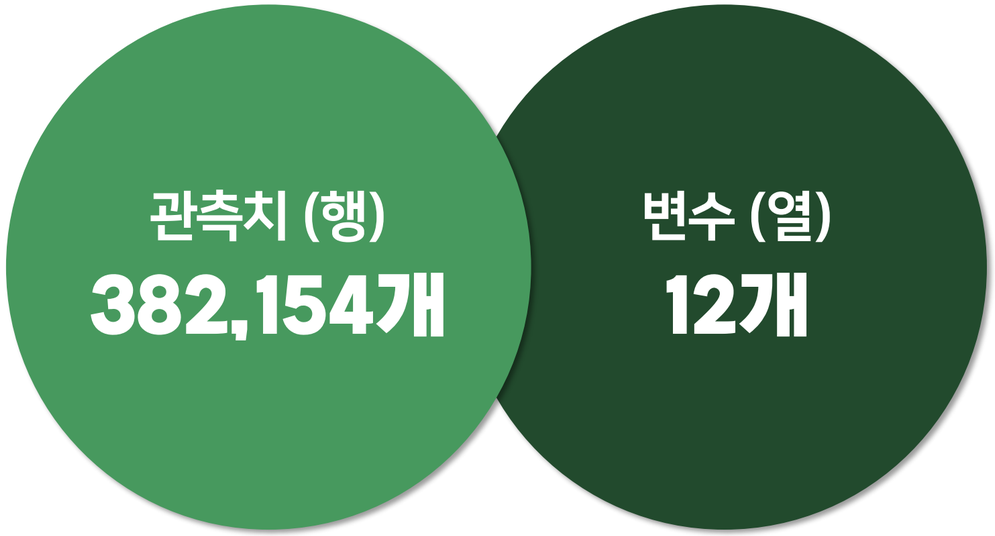
03

데이터 요약 / 변수 설명 / 기초 통계량 시각화 / 변수 관계 해석

03 데이터 설명

데이터 요약

Learning from Imbalanced Insurance Dataset



종속 변수 : Response

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 382154 entries, 0 to 382153
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   id                    382154 non-null  int64
 1   Gender                382154 non-null  object
 2   Age                   382154 non-null  int64
 3   Driving_License       382154 non-null  int64
 4   Region_Code           382154 non-null  float64
 5   Previously_Insured    382154 non-null  int64
 6   Vehicle_Age           382154 non-null  object
 7   Vehicle_Damage        382154 non-null  object
 8   Annual_Premium        382154 non-null  float64
 9   Policy_Sales_Channel  382154 non-null  float64
 10  Vintage               382154 non-null  int64
 11  Response              382154 non-null  int64
dtypes: float64(3), int64(6), object(3)
memory usage: 35.0+ MB
```

결측값 존재하지 않음

03 데이터 설명

변수 설명

변수명	설명	구분
ID	조사 응답자 식별 번호	정수형
Gender	성별	범주형
Age	나이	정수형
Driving_License	운전면허 소지 여부 0 : 운전면허 취소 혹은 정지 / 1 : 운전면허 소지	범주형
Region_Code	거주 지역 번호	범주형
Previously_Insured	타사 보험 가입 여부 0 : 타사 보험 미가입 / 1 : 타사 보험 가입	범주형

03 데이터 설명

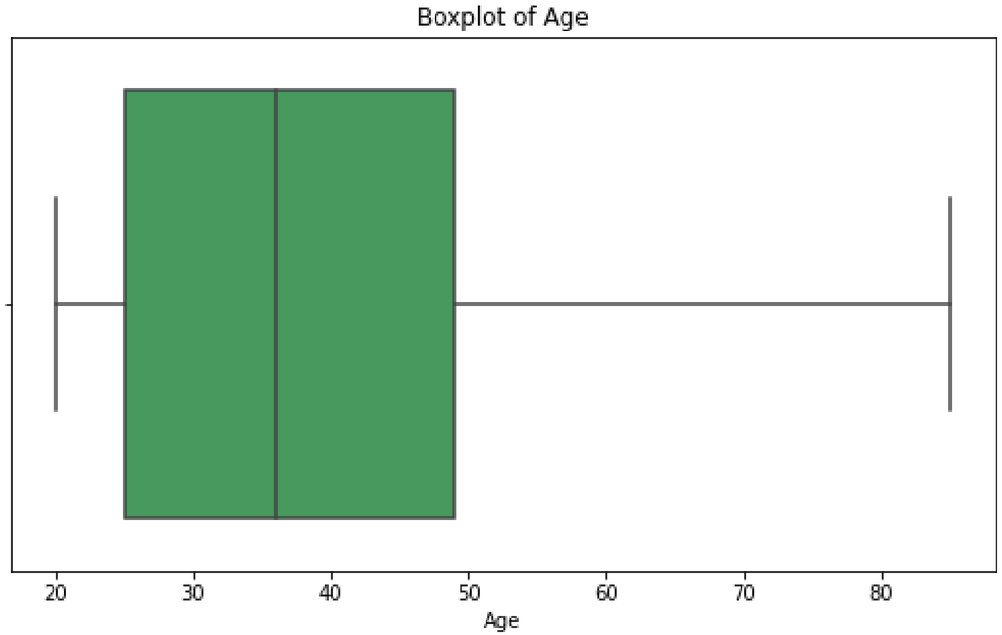
변수 설명

변수명	설명	구분
Vehicle_Age	본인 소유 자동차 연식 '1< Year' : 1년 미만 / '1-2 Year' : 1년 이상, 2년 이하 / '2 > Years' : 2년 초과	범주형
Vehicle_Damage	자동차 파손 이력 여부 0 : 자동차 파손 이력 없음 / 1 : 자동차 파손 이력 있음	범주형
Annual_Premium	연간 책정된 보험료	실수형
Policy_Sales_Channel	이용한 보험 판매 채널	실수형
Vintage	자사 보험 인지 후 경과 일	정수형
Response	자사 보험 가입 의향 (target) 0 : 부정적 / 1 : 긍정적	범주형

03 데이터 설명

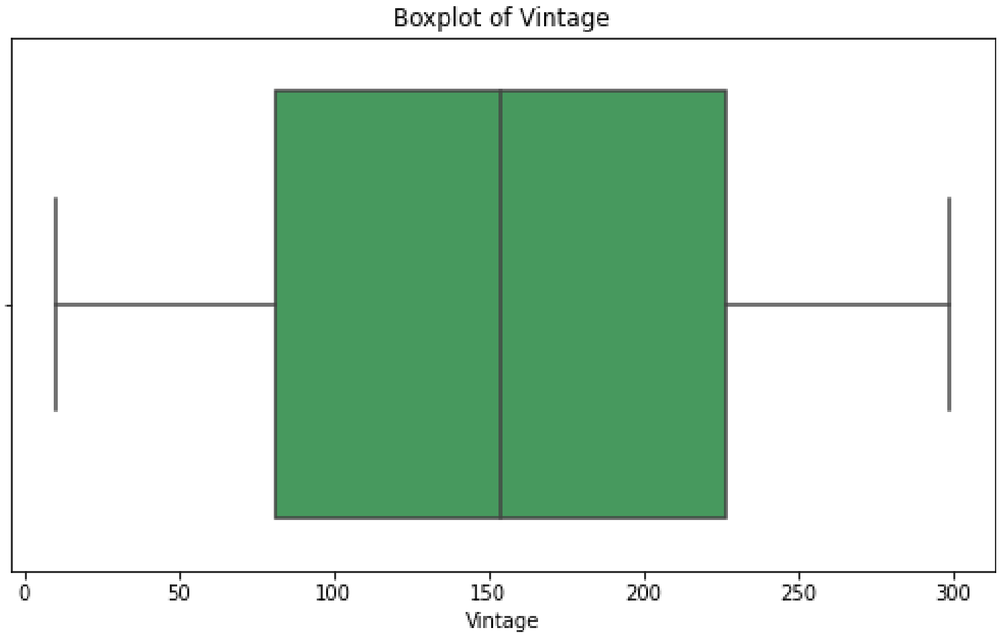
기초통계량 시각화

Age



- 이상치 존재하지 않음
- 중위수가 다소 왼쪽으로 치우친 형태

Vintage

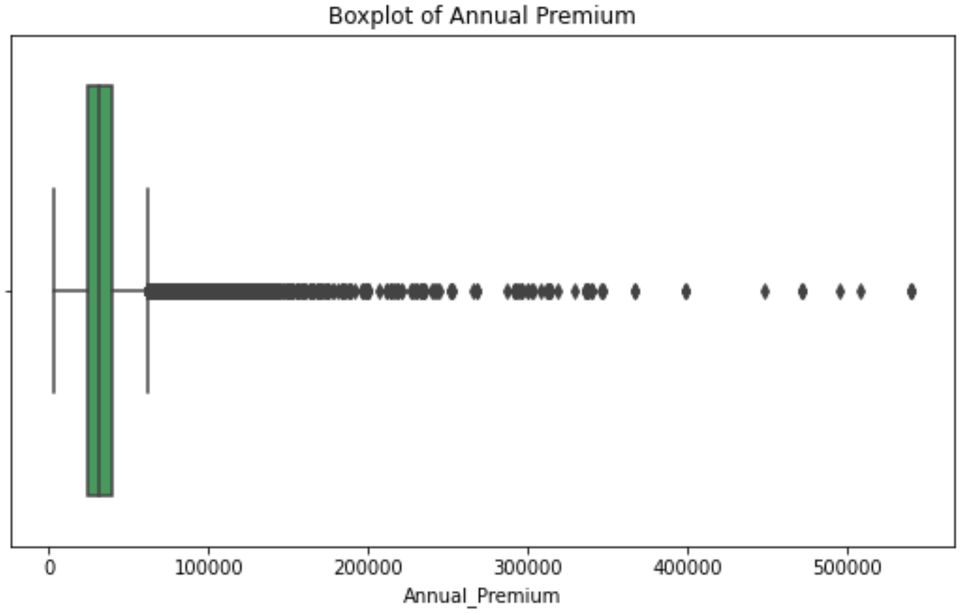


- 이상치 존재하지 않음
- 이상적 형태의 Boxplot

03 데이터 설명

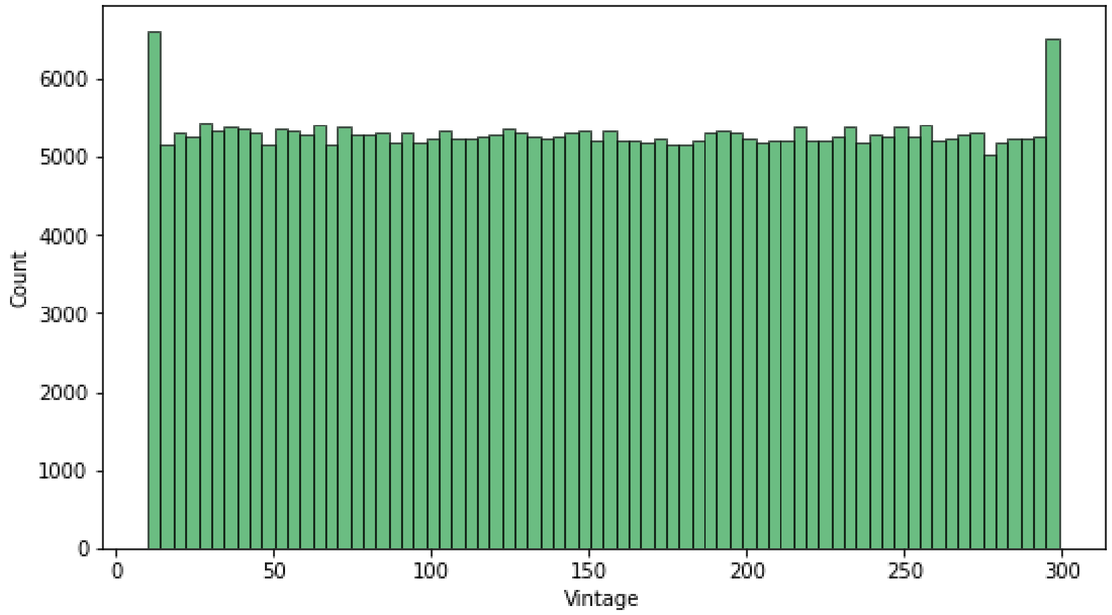
기초통계량 시각화

Annual_Premium



- Boxplot이 한 쪽으로 치우친 형태
- 다수의 이상치 발견 (IQR 기준)

Vintage

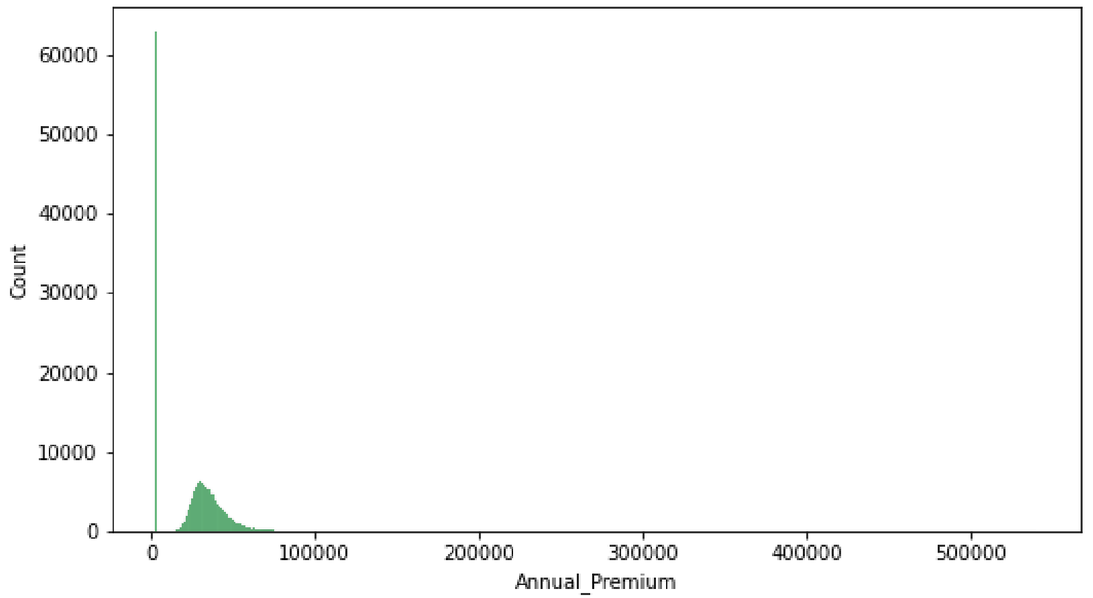


- 고른 분포를 보임

03 데이터 설명

기초통계량 시각화

Annual_Premium



- 특정 값 하나의 빈도가 높은 것 확인 가능
- 해당 값 : 2,630 (최솟값)

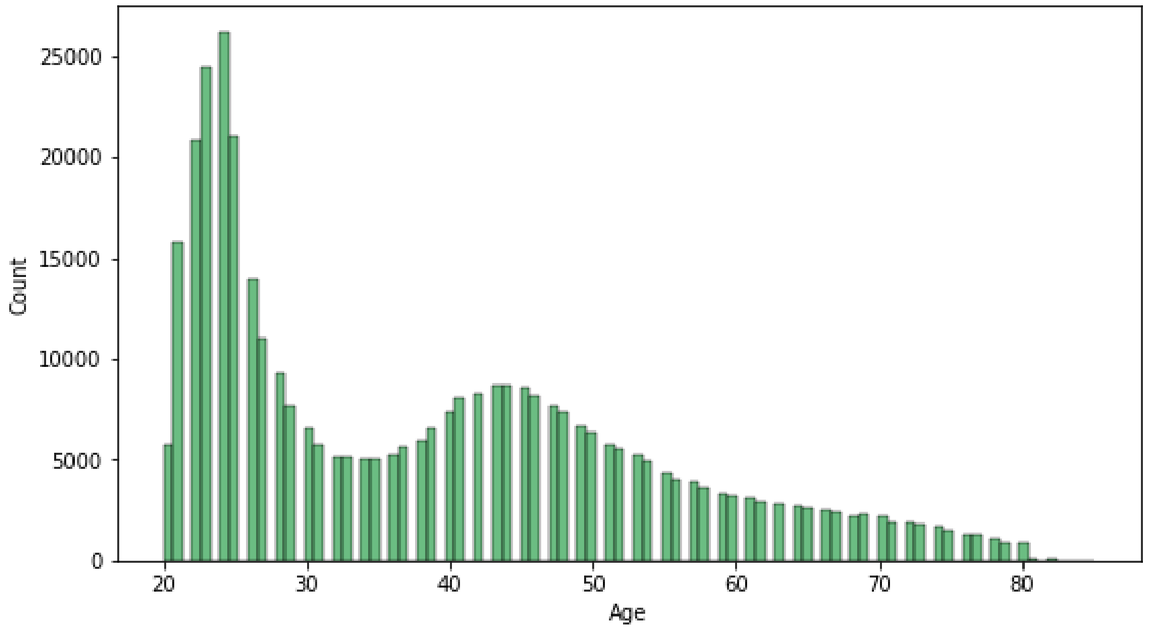
추가 옵션을 선택하지 않고,
책정된 기본 보험료만 납부하는
보험 이용 고객 특성으로 인해
왼쪽과 같은 분포가 나타남

기본 보험료가 책정된 고객과 아닌
고객을 분리하는 전처리 과정 추가

03 데이터 설명

기초통계량 시각화

Age



- 특정 연령대에 데이터 분포가 밀집된 형태
- 특정 연령대 : 20-30대 / 40-50대

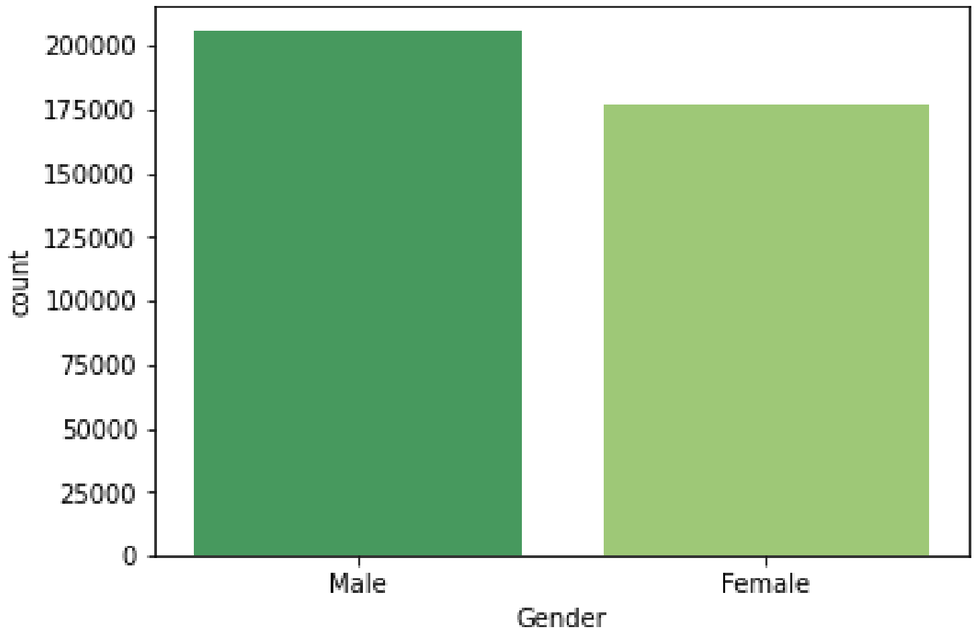
개별 나이를 나타내는 Age
변수를 grouping을 통해
연령대를 나타내는 변수로 변경

연령대와 타 변수들의 관계를
확인하여 유의미할 경우,
이를 활용해 새로운 변수 추가

03 데이터 설명

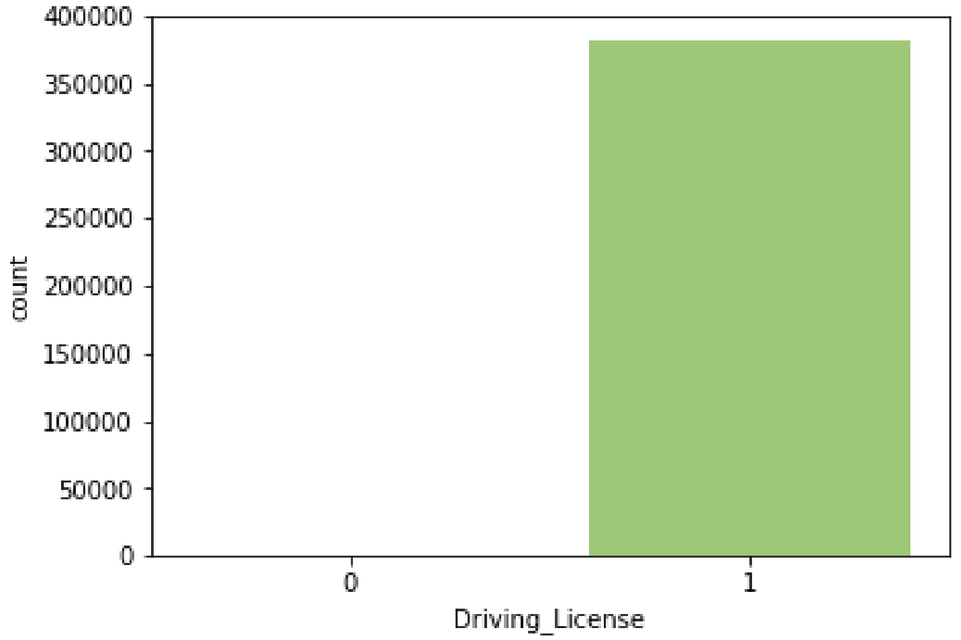
기초통계량 시각화

Gender



- 범주형 변수
- 각 범주의 빈도수가 비슷함

Driving_License

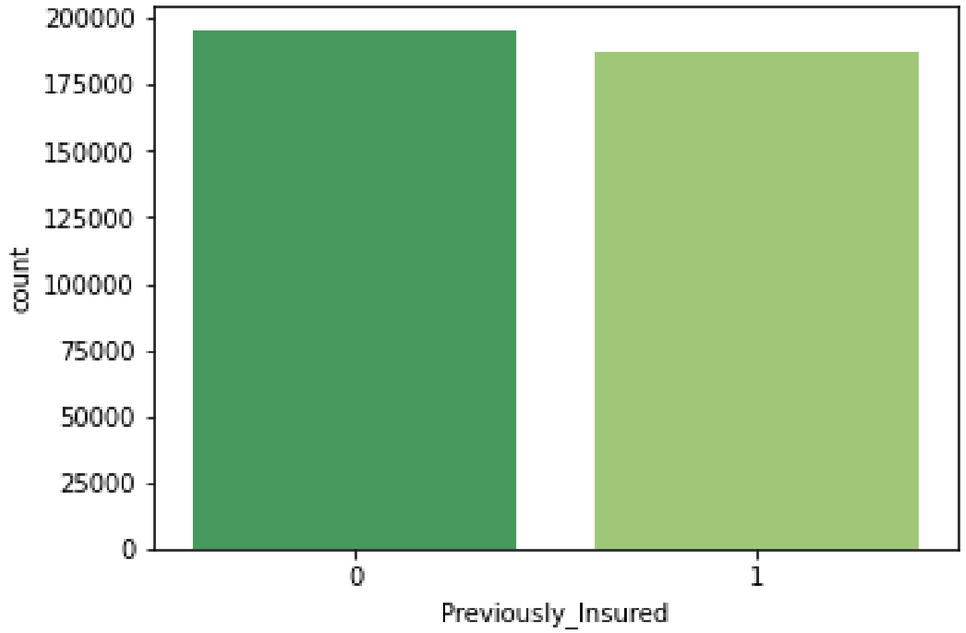


- 면허를 소지한 고객(1)이 면허를 소지하지 않은 고객(0)보다 매우 많음

03 데이터 설명

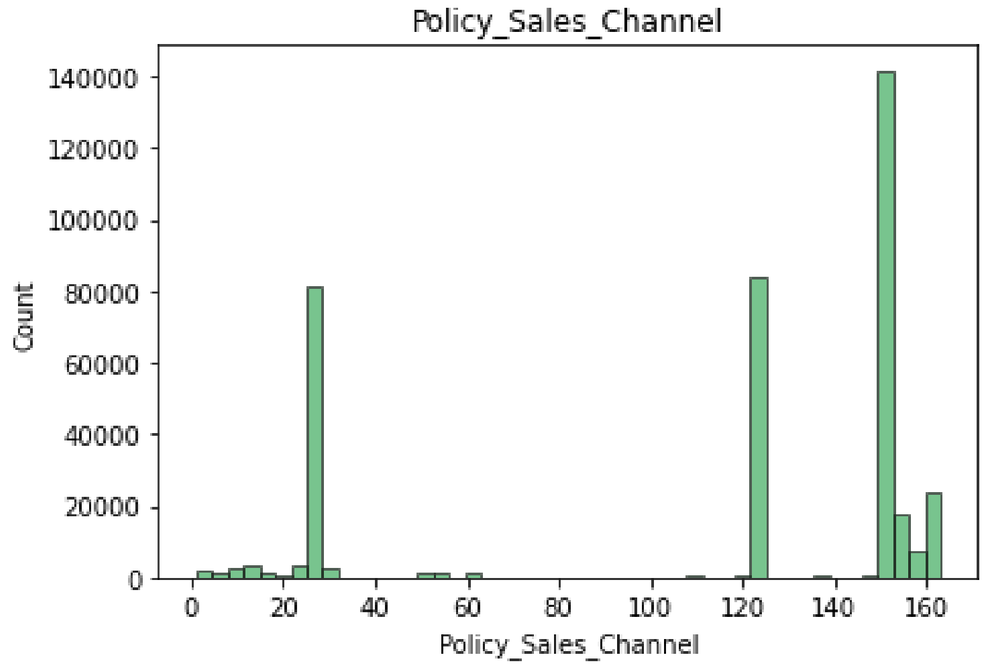
기초통계량 시각화

Previously_Insured



- 범주형 변수
- 각 범주의 빈도수가 비슷함

Policy_Sales_Channel

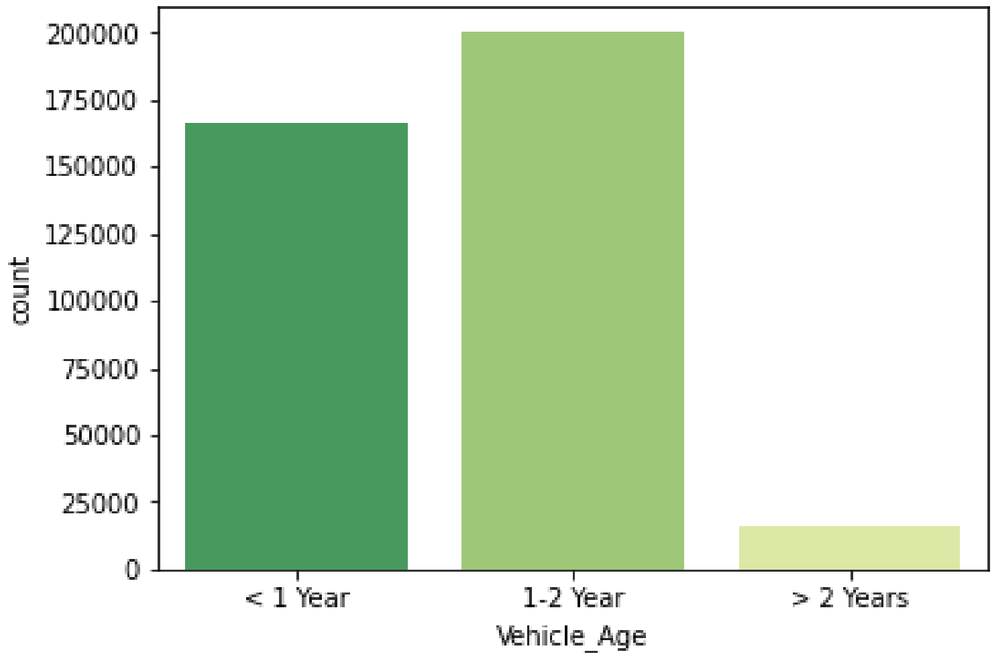


- 다수 고객이 이용하는 채널을 주요 채널로 고려하여 새로운 변수 추가 시 활용함

03 데이터 설명

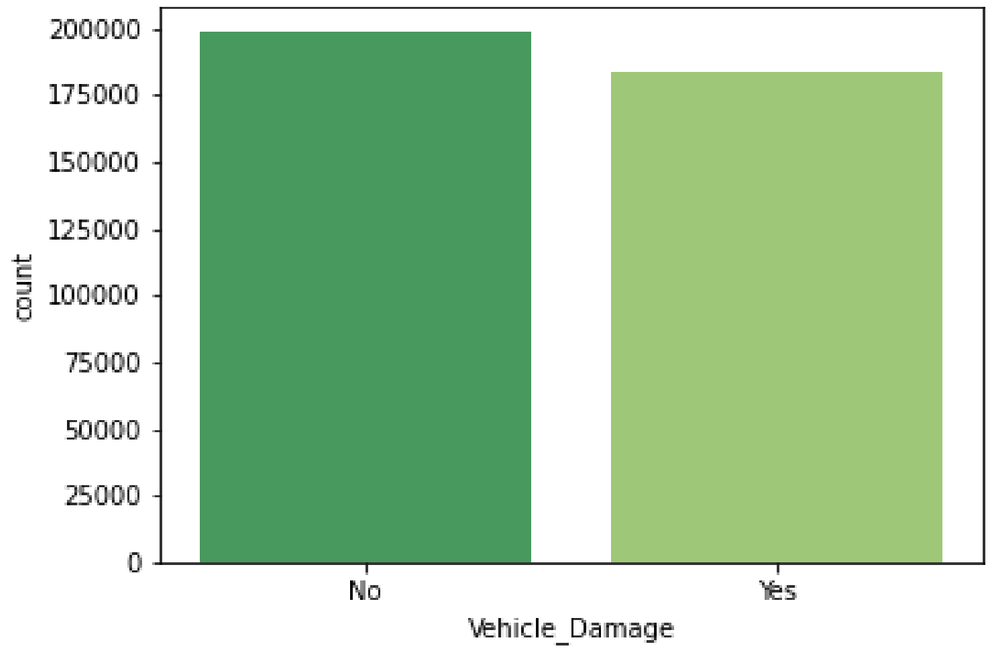
기초통계량 시각화

Vehicle_Age



- 2년 이상인 경우 타 경우보다 수가 적음
- Case를 나누어 타 변수와의 관계 파악

Vehicle_Damage

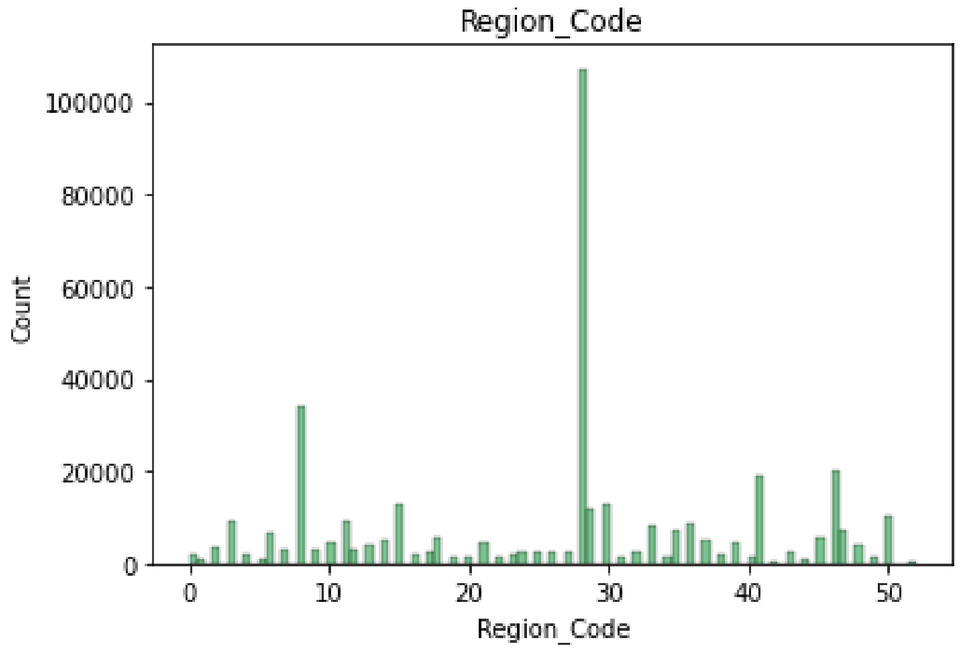


- 범주형 변수
- 각 범주의 빈도수가 비슷함

03 데이터 설명

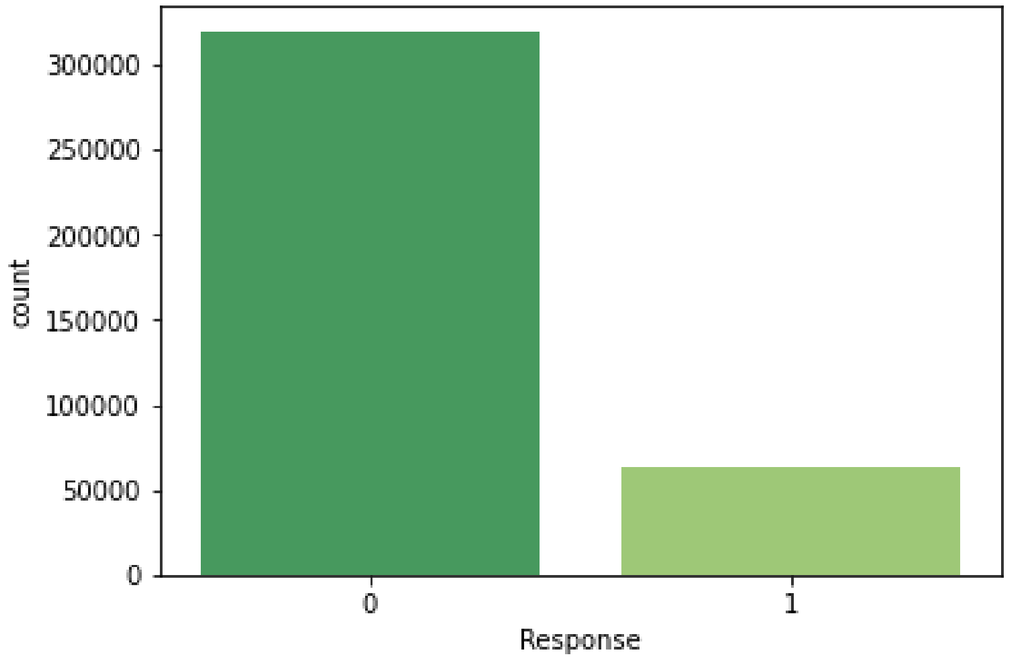
기초통계량 시각화

Region_Code



- 특정 지역의 데이터 개수가 많음
- 인구 밀집 지역 고객 따로 분리하여 고려

Response

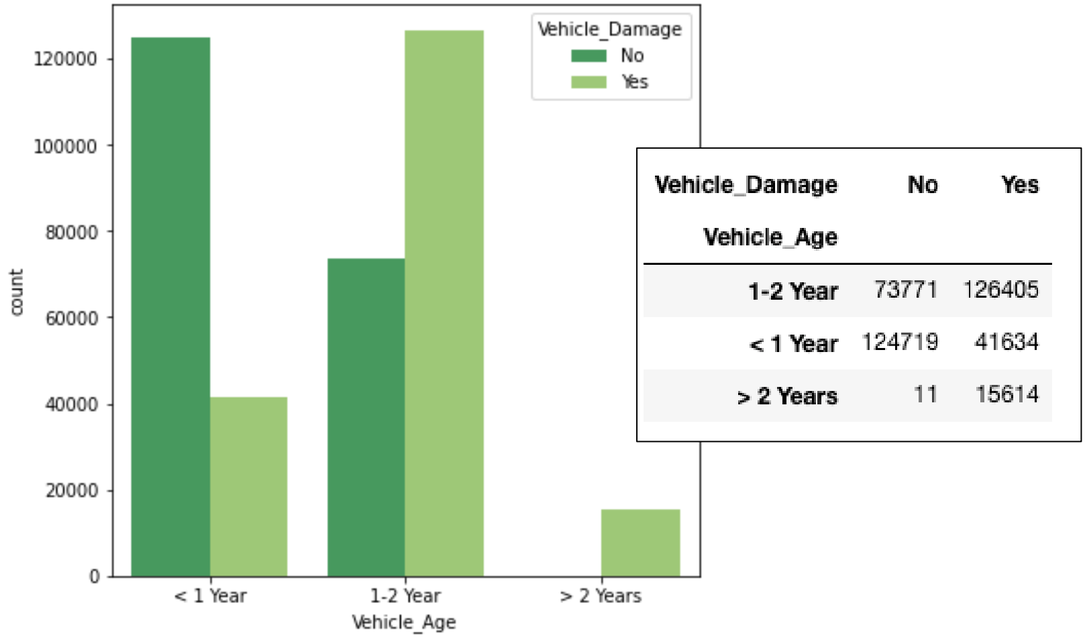


- 데이터 분포 비율이 83.6 : 16.4
- 분포가 Imbalanced

03 데이터 설명

독립변수 간 관계 해석

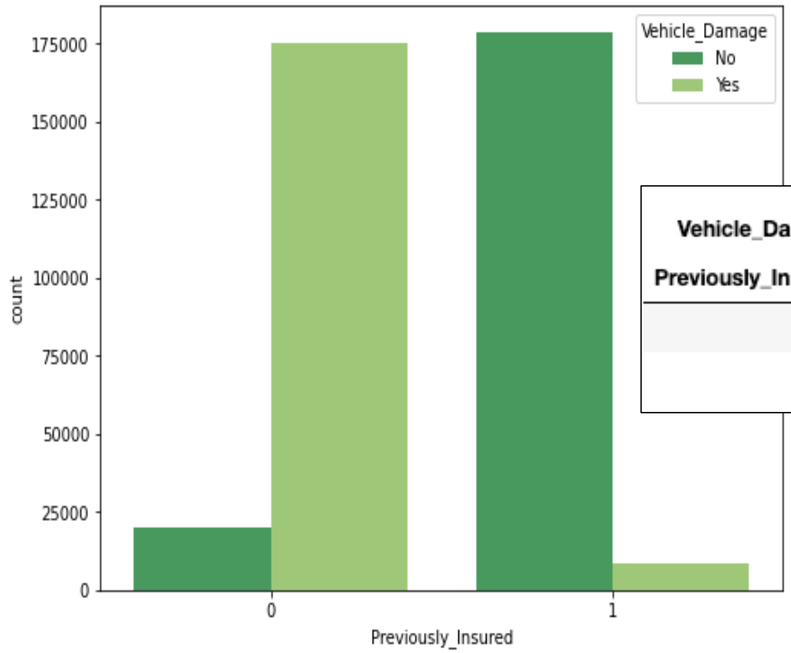
Vehicle_Age & Vehicle_Damage



Vehicle_Damage	No	Yes
Vehicle_Age		
1-2 Year	73771	126405
< 1 Year	124719	41634
> 2 Years	11	15614

- 자차 연식에 따라 사고 경험 유무 비율이 달라짐
- 자차 연식이 높을수록 사고 경험자 비율이 높아짐

Previously_Insured & Vehicle_Damage



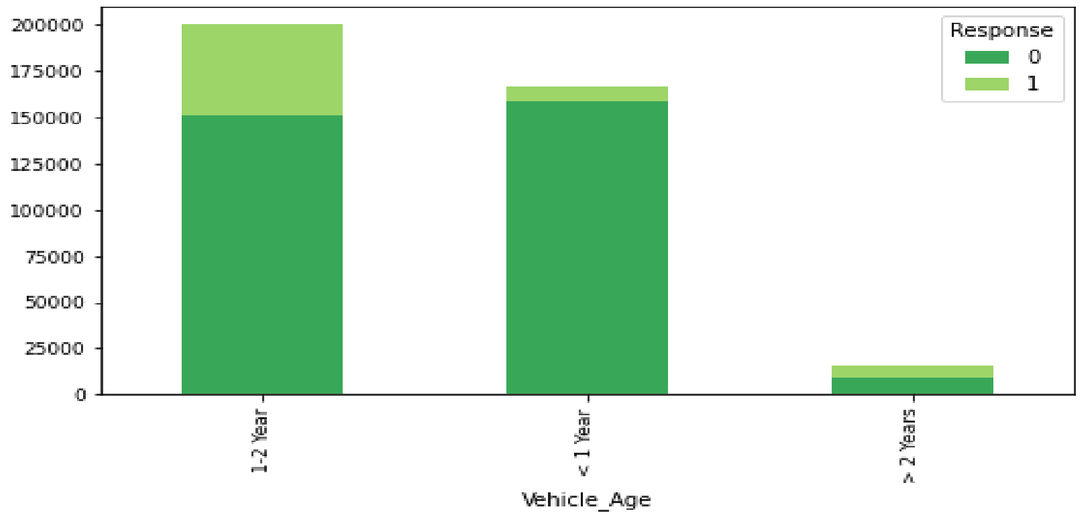
Vehicle_Damage	No	Yes
Previously_Insured		
0	19929	175282
1	178572	8371

- 자동차 파손 경험 유무에 따라 타사 보험 가입 이력 여부에 큰 차이가 있음을 확인 가능

03 데이터 설명

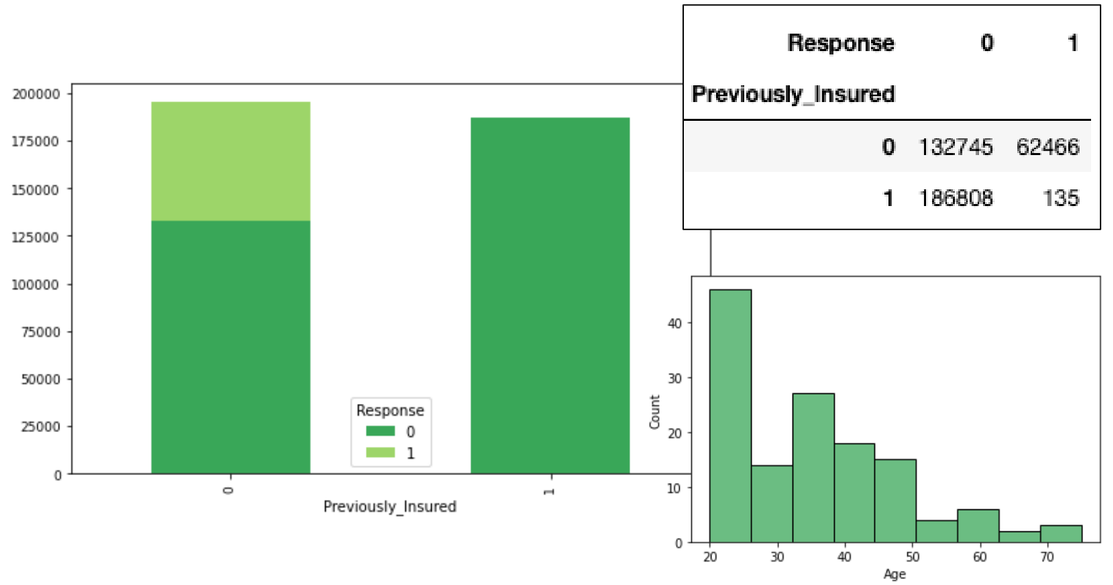
독립변수와 종속변수 간 관계 해석

Vehicle_Age & Response



- 자동차 연식에 따라 종속 변수 비율이 달라짐
- 이후 서브 모델링 진행 시 활용

Previously_Insured & Response



Response	0	1
Previously_Insured		
0	132745	62466
1	186808	135

- 타사 보험 가입 경험이 있는 고객 중 자사 보험 가입 의향이 있는 고객 연령대가 전체 데이터 대비 낮음

전처리 과정

04

이상치 처리 / 인코딩 / 변수 추가

04 전처리 과정

이상치 처리

Boxplot

- Annual_Preminum 변수 외에는 Boxplot 내에서 이상치 확인 불가
- 변수 특성상 기본 보험료 값인 2,630의 수가 압도적으로 많아 이상치 기준 모호

이상치 제거 기준에서 Boxplot 제외

의미론적 이상치

- 운전면허를 소지하지 않았음에도 자동차가 파손된 이력이 있는 경우 논리적으로 성립하지 않는다고 판단
- 즉, 의미론적 이상치로 판단

해당 관측값을 Train 데이터에서 삭제

04 전처리 과정

인코딩

원본 데이터

	Gender	Vehicle_Age	Vehicle_Damage
0	Male	< 1 Year	No
1	Male	1-2 Year	Yes
2	Male	1-2 Year	Yes
3	Female	< 1 Year	No
4	Female	< 1 Year	No

One-Hot Encoding

	encode_gender	encode_v_age	encode_v_dam
0	1	1	0
1	1	0	1
2	0	0	1
3	0	1	0
4	1	1	0

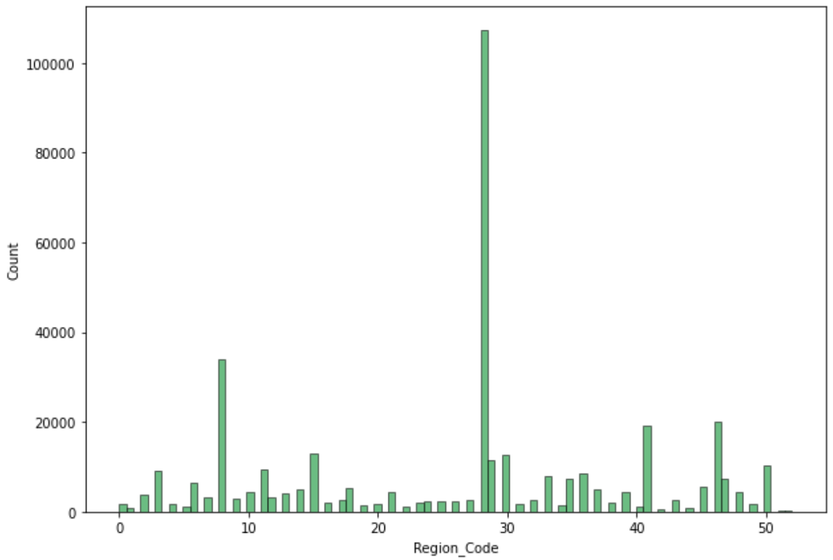
Label Encoding

	encode_Female	encode_Male	encode_1-2 Year	encode_< 1 Year	encode_> 2 Years	encode_No	encode_Yes
0	0	1	0	1	0	1	0
1	0	1	1	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	1	0	1	0
4	0	1	0	1	0	1	0

04 전처리 과정

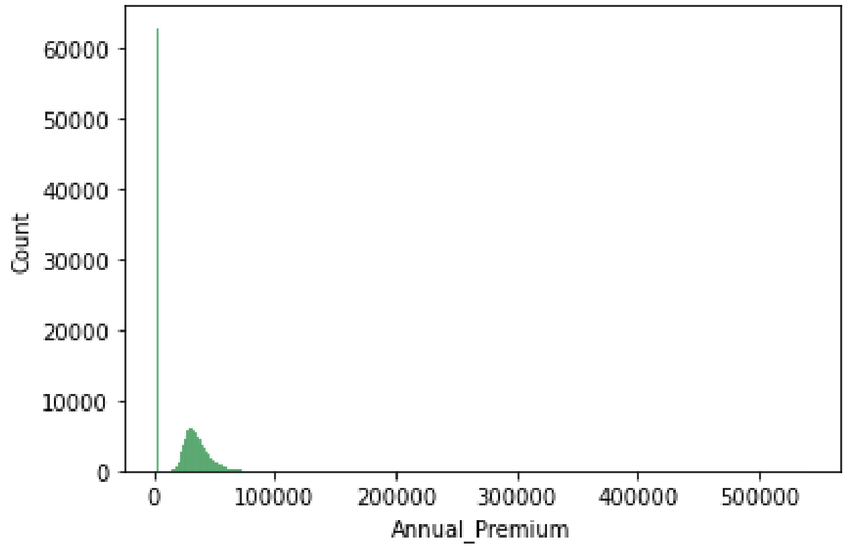
변수 추가

Population



- Region_Code의 Histogram 상에서 '26'이 전체의 약 28%를 차지하는 것 확인
- 해당 범주를 인구 밀집 지역으로 간주하여 변수 추가
  → 인구 밀집 지역(26)에 해당 시 main, 아니면 notmain

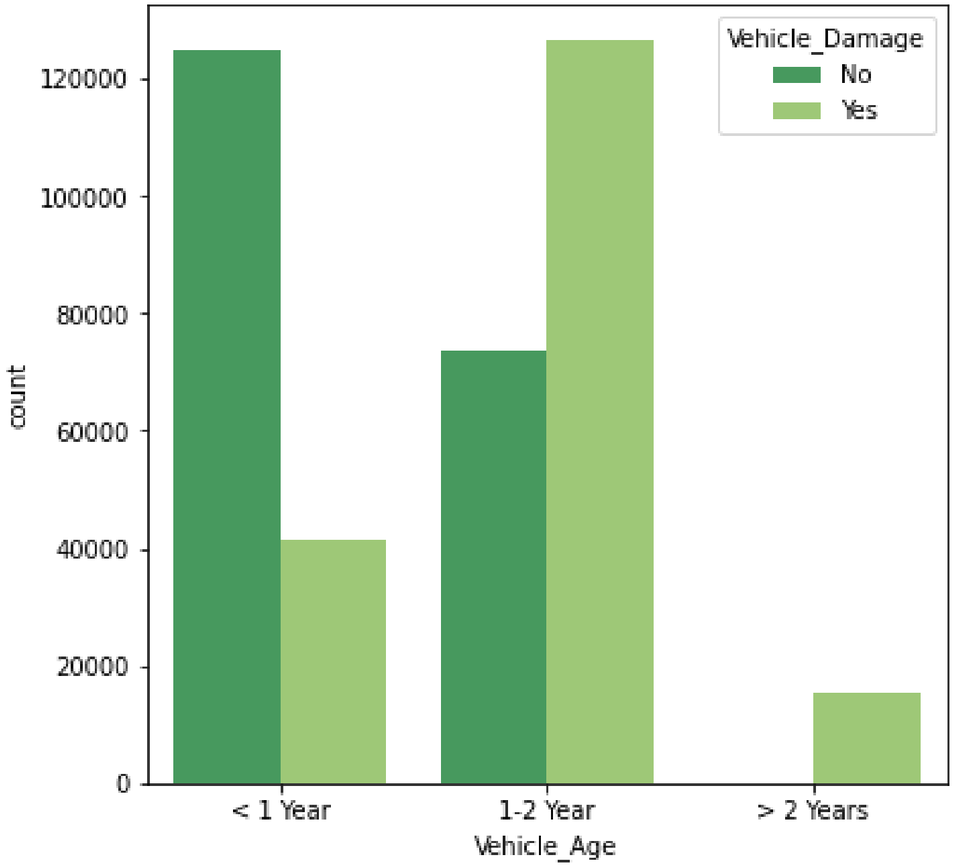
Basic_Annual



- Annual_Premium의 Histogram 상에서 최솟값 '2630'이 전체 데이터의 16%를 차지하는 것 확인
- 해당 범주를 기본 보험만 적용한 것으로 취급하여 변수 추가
  → 기본 보험료만 지불 : basic / 옵션 추가 : option

04 전처리 과정

변수 추가



Danger

- 자차 연식이 1년 미만임에도 자동차 파손 경험이 있는 경우 운전 습관이 위험, 즉 보험의 필요성이 높은 고객으로 분류
- 반대로 자가 연식이 2년 초과이고 파손 경험이 없는 경우 운전 습관이 안전한 고객으로 분류

→ 고위험군 : high / 저위험군 : low / 그 외 : mid

N_Danger

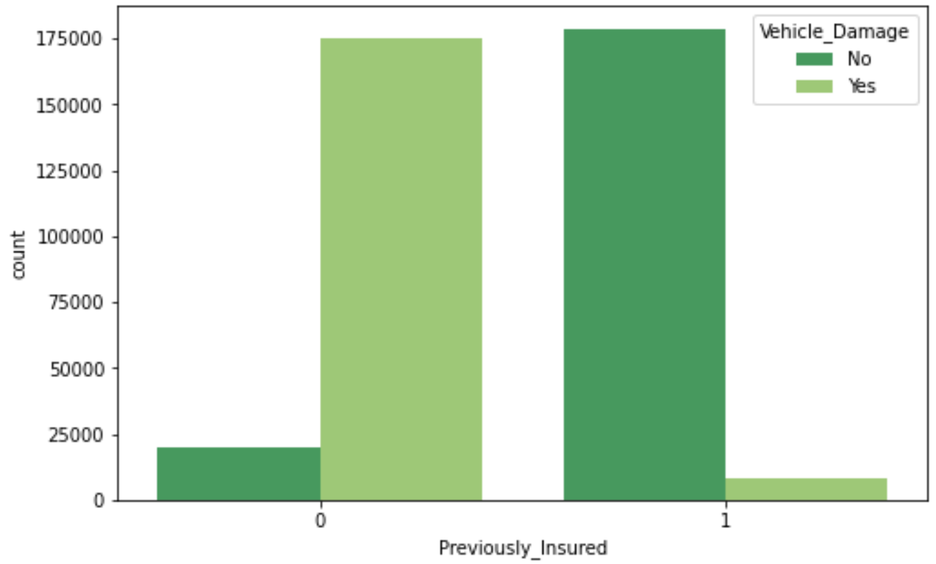
- 타사 보험을 가입한 이력이 없는 고객 중 자동차 파손 경험이 있는 고객은 또 다른 위험군으로 판단

→ 위험군에 속하는 경우 high, 아닌 경우 low

04 전처리 과정

변수 추가

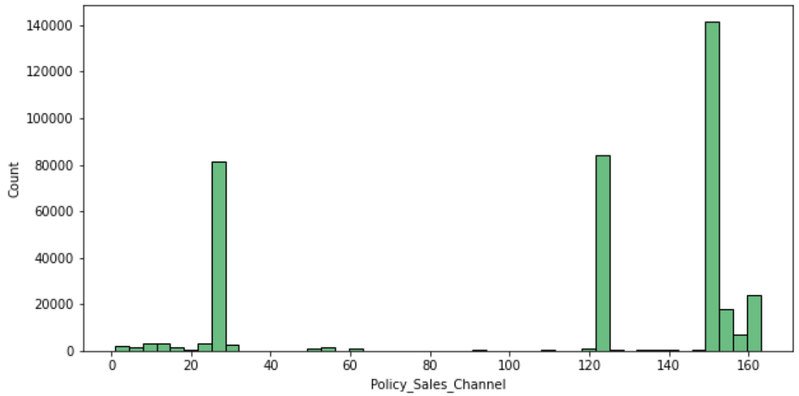
Beneficiary



- 위의 Histogram을 통해 타사 보험 가입 이력이 있고, 자동차 파손 경험도 있는 고객을 확인 가능
- 해당 고객을 보험 수혜 경험이 있는 것으로 고려
  → 보험 수혜자이면 benefit, 아니면 not_benefit

Main_Channel

152.0	137422
26.0	81566
124.0	73315
160.0	21045
156.0	10106
122.0	9745
157.0	6739
154.0	5883
151.0	3760
163.0	2972



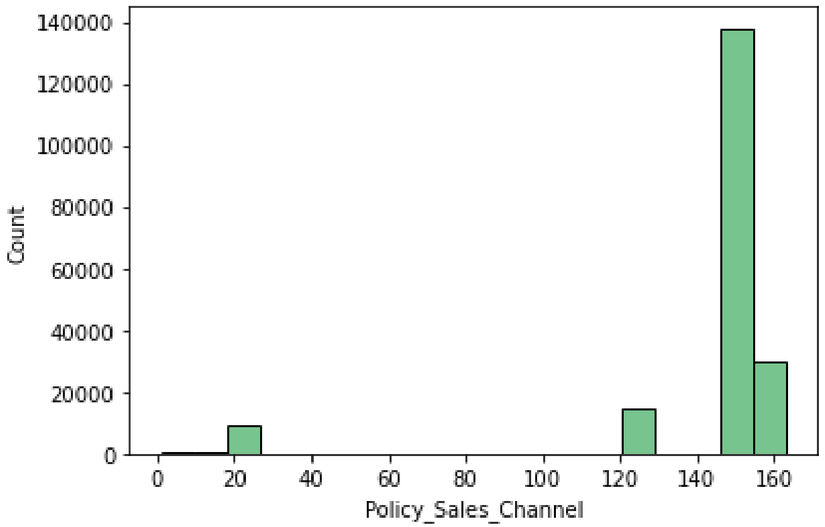
- '152', '26', '124' 등 특정 채널에 해당하는 고객의 수가 압도적으로 많은 것을 확인 가능
- 이를 주요 채널로 간주하여 변수로 추가
  → 주요 채널 이용 : main_ch / 아닌 경우 : notmain_ch

04 전처리 과정

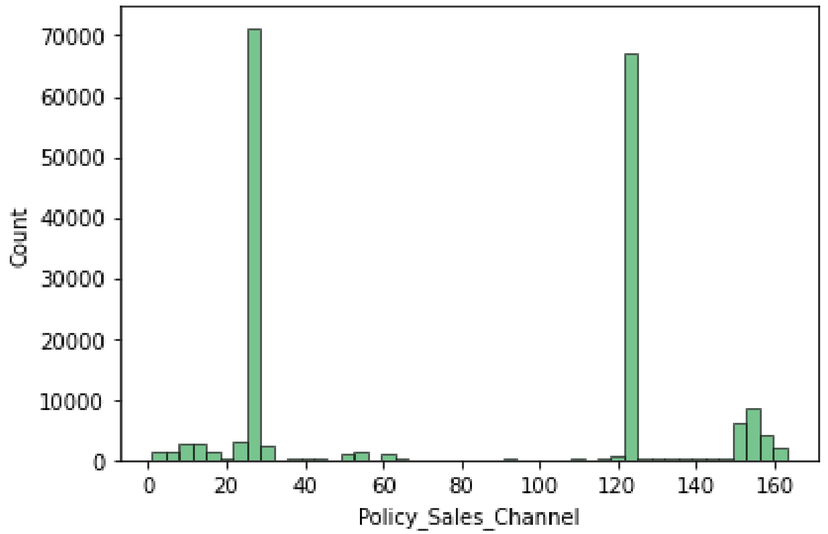
변수 추가

Age_Channel

* 36세 이하



* 36세 초과



- 36세를 기준으로 나눈 두 연령대 그룹에 따라 주요 보험 판매 채널이 달라진다는 것 확인
- 각 그룹별 주요 채널 이용 여부를 나누어 변수로 추가

  → 채널 '26', '124' : main_over / 채널 '152', '160' : main_under / 그 외 : channel

04 전처리 과정

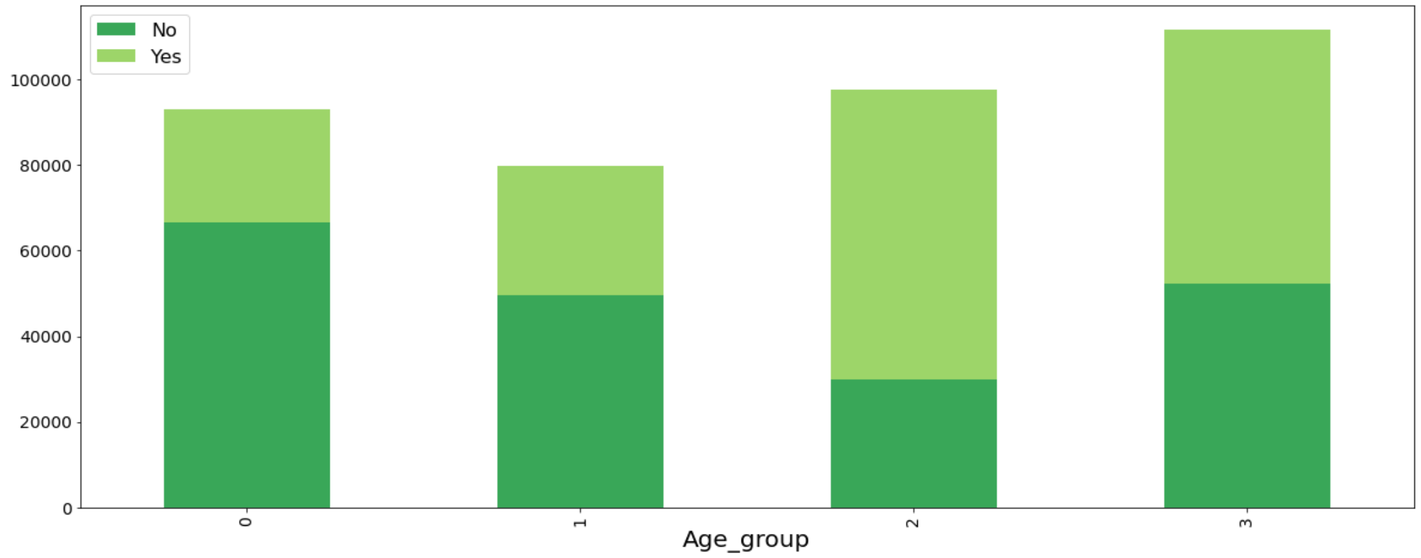
변수 추가

Age_Group

- 연령대와 타 변수 사이에 분포적 특징이 다수 나타남
- 연령대 기준으로 아래와 같이 4개의 그룹 생성
  1. 25세 이하
  2. 26세 이상 36세 이하
  3. 37세 이상 49세 이하
  4. 50세 이상

각 그룹마다 ~25, 26~36,37~49, 50~의 값을 부여해 새로운 변수 추가

Age_Damaged

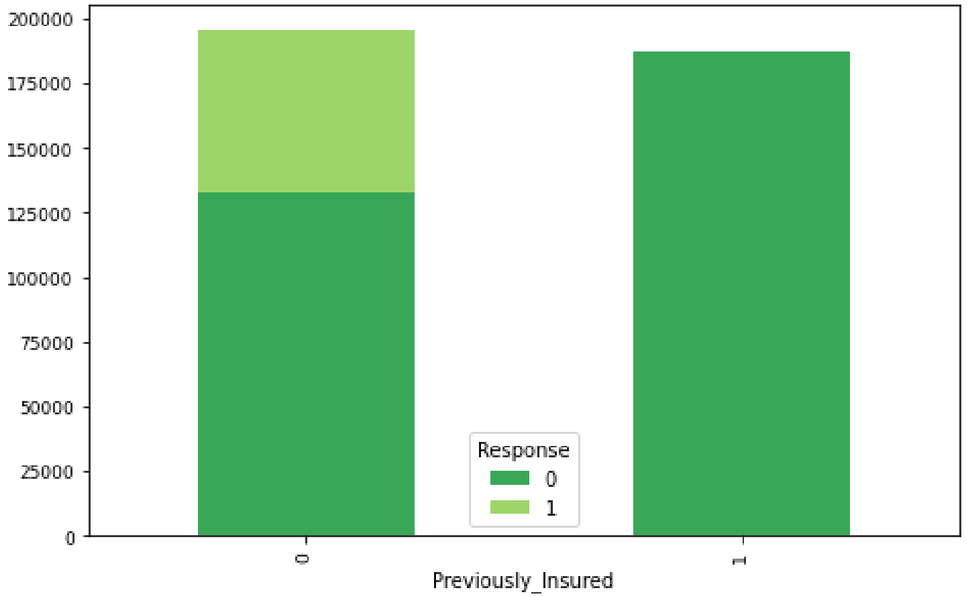


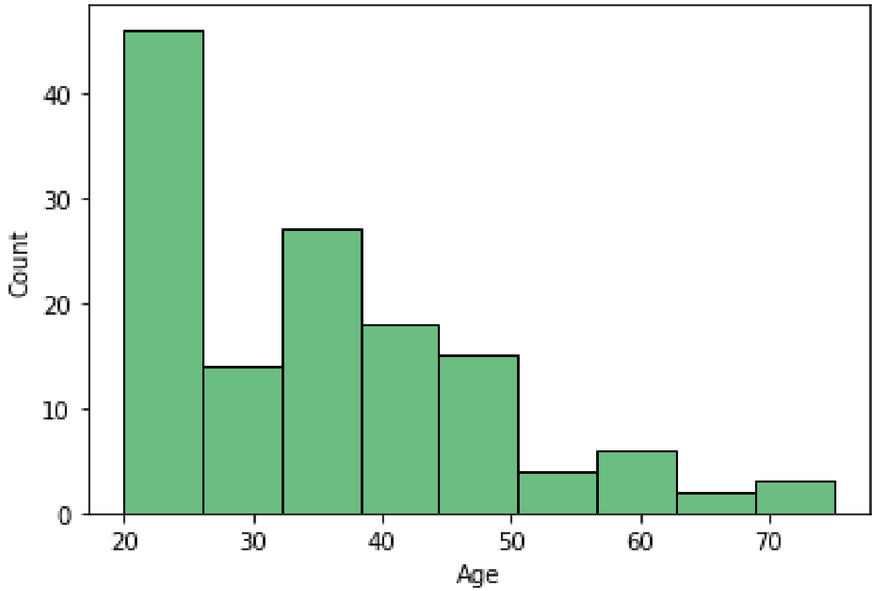
- 연령대에 따라 사고 유경험자의 비율이 달라지는 것 확인
  → 각 연령대에 따른 사고 유경험자의 비율 변수로 추가

04 전처리 과정

변수 추가

Young_Rich





- 타사 보험 가입 이력이 있는 고객 중 자사 보험 가입 의향이 있는 고객의 수 및 연령대가 전체 대비 낮은 것을 확인
- 해당 고객을 보험 추가 가입에 거리낌이 없는 재력을 소유한 것으로 판단하여 변수로 추가

  → 타사 보험 가입 이력이 있고 자사 보험 가입에도 긍정적이면 YR, 그렇지 않은 경우 Not_YR

04 전처리 과정

최종 전처리 결과

사용할 데이터 종류 : 총 6가지

Notation	서브 모델링 적용		서브 모델링 미적용
	데이터 균형	데이터 불균형	
변수 추가 O	ob	oi	of
변수 추가 X	vb	vi	vf

적용할 전처리 종류 : 총 4가지

① Label Encoding + 이상치 유지

② Label Encoding + 이상치 삭제

③ One-Hot Encoding + 이상치 유지

④ One-Hot Encoding + 이상치 삭제

>> 6 X 4 = 24가지의 데이터 생성

모델링 과정

05

Model & Variable Selection / Sub-Modeling / Data Augmentation

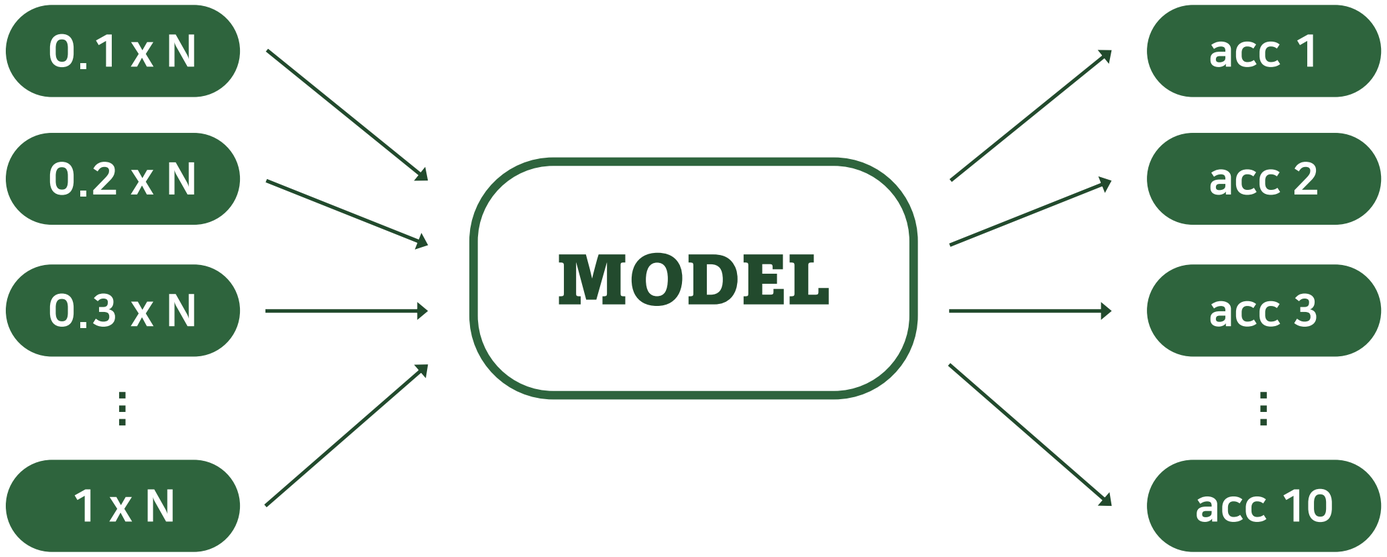
05 모델링 과정
Model Selection

크기가 N X P인 데이터가 있을 때…

① 데이터 개수가 전체의 10%부터 100%까지 점차 늘어가는 부분집합들을 생성

② 각 부분집합을 모델에 적용하여 Accuracy 산출

③ 데이터 개수가 증가함에 따라 모델의 성능, 즉 Accuracy 또한 증가하면 해당 모델이 평가에 적합하다 판단



- acc 1 < acc 2 < … < acc 10 →모델 채택
- 채택된 모델 : 총 5가지
  1. CatBoost
  2. LightGBM
  3. RandomForest
  4. LinearRegression
  5. SGDClassifier

→ 24 x 5 = 120가지의 데이터 생성

05 모델링 과정

Variable Selection

변수 선택 방식 : Permutation Importance Method

→ 개별 변수의 중요도를 내림차순으로 정렬하여 특정 임계점을 넘는 변수만 최종 모형에 사용

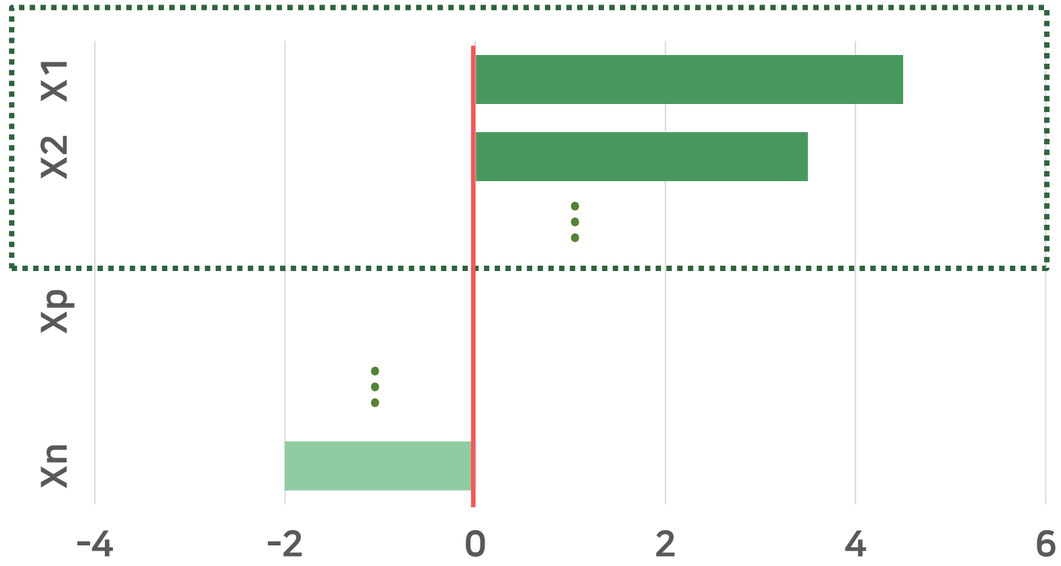
Dataset

X				Y
Gender	···	Age	Vintage	Target
0		22	16	1
1	···	42	135	0
1		28	253	1
⋮		⋮	⋮	⋮

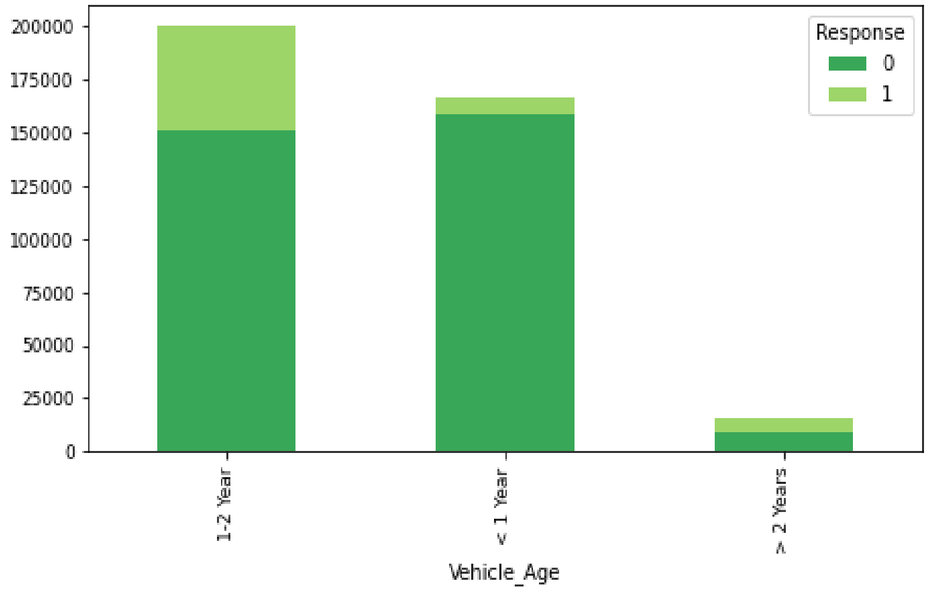
- Feature Importance가 음수인 경우 해당 변수가 존재하지 않는 것이 모델의 성능을 향상시킴
  → Feature Importance가 0 이상인 경우만 포함

Modeling

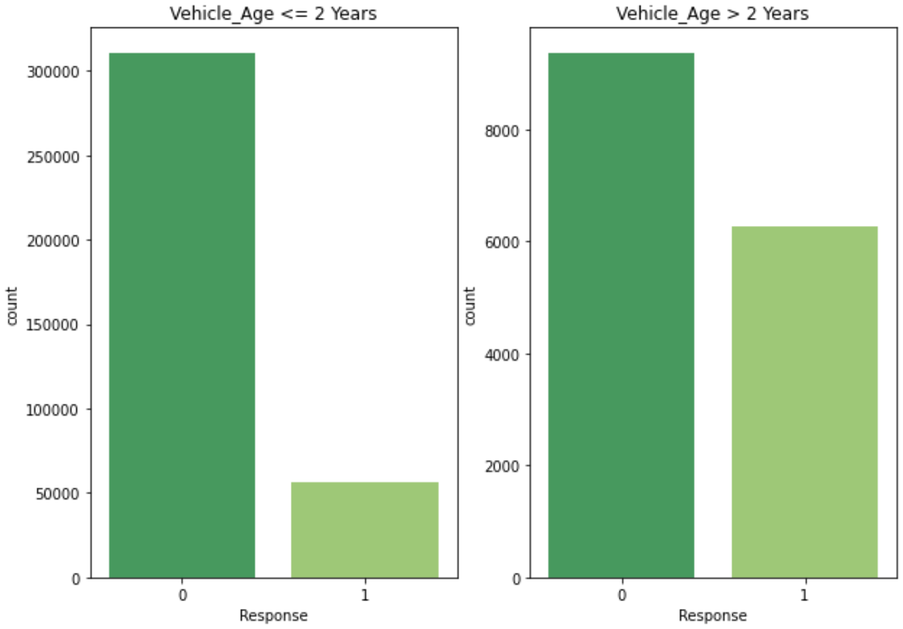
[Feature Importance 평균]



05 모델링 과정
Sub-Modeling



‘1-2 Year’과
‘< 1 Year’ 병합



- ‘1-2 Year’과 ‘< 1 Year’는 Target 값이 불균등한 분포를 나타내는 반면, ‘> 2 Years’는 Target 값이 균등한 분포를 보임
- 변수들 중 불균형과 균형의 차이를 가장 분명하게 나타냄
  → 해당 변수로 Sub-Grouping 진행

05 모델링 과정
Sub-Modeling

Sub-Group Data

Imbalanced Data
Vehicle_Age != '> 2 Years'

Balanced Data
Vehicle_Age == '> 2 Years'

Model

Confusion Matrix

		Real
Predict	TP	FP
	FN	TN

+

		Real
Predict	TP	FP
	FN	TN

Final Confusion Matrix

		Real
Predict	TP	FP
	FN	TN

Sub-Group의 Confusion Matrix는 삭제하고,
위의 Matrix로 검정 진행

05 모델링 과정

Data Augmentation

아래와 같이 총 9가지 증강 경우의 수 생성

Random Over Sampling (ROS)	Synthetic Minority Over Sampling Technique (SMOTE)	Conditional Tabular Generative Adversarial Network (CTGAN)
종속변수 두 범주의 비율	종속변수 두 범주의 비율	종속변수 두 범주의 비율
5:5	5:5	5:5
6:4	6:4	6:4
7:3	7:3	7:3
1	4	7
2	5	8
3	6	9

05 모델링 과정

최종 데이터 형태

Model AB-trainN1-N2.csv

Model	A	N1
CatBoost	변수 추가 : v	전처리 경우의 수에 따라 1~4 사이 값 할당
LightGBM	변수 추가 X : o	
RandomForest	**B**	**N2**
LogisticRegression	서브 모델링 : bi	증강 기법에 따라 0~9 사이 값 할당
SGDClassifier	서브 모델링 X : f	

>>

실험에 사용할 전체 데이터 수

800개

결과 검정 및 해석

06

06 결과 검정 및 해석

기존 평가지표의 한계점

기존 평가지표 산출식

	Real	
Predict	TP	FP
	FN	TN

- Accuracy = (TP + TN) / (TP + FP + FN + TN)
- Precision = TP / (TP +FP)
- Recall = TP / (TP +FN)
- F-measure = 2 * (Precision * Recall) / (Precision + Recall)

한계점

- 불균형 데이터 분류 시 데이터 개수가 많은 Real Negative 집단에 편향된 학습을 진행하여 Real Positive 집단은 거의 고려하지 못하므로 Accuracy는 더 이상 좋은 측도가 될 수 없음
- Precision, Recall 그리고 이의 조화평균인 F-measure은 불균형 데이터 분류 결과를 평가하는데 주로 사용되지만 이 세 지표는 True Positive만 고려하며 True Negative는 전혀 고려하지 않는다는 단점이 존재

06 결과 검정 및 해석

Matthews Correlation Coefficient (MCC)

MCC

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Confusion Matrix의 모든 값을 고려하므로 기존 성능지표보다 더 많은 정보를 담고 있으며, 불균형 데이터 분류 시 가장 신뢰할 만한 평가지표로 알려져 있음

Bookmaker Informedness (B)와 Markedness(M)의 기하 평균은 MCC이며, B와 M은 피어슨 카이제곱 검정을 통해 추정치를 구할 수 있음

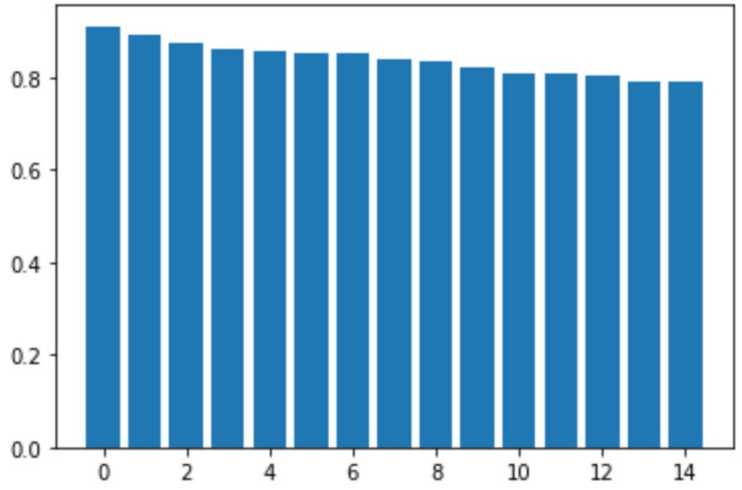
- MCC의 추정치를 계산할 수 있으므로, 통계적 가설 검정 가능
- 연구 진행 과정
  1. MCC 추정치와 $\chi^2(0.99, 1) \approx 6.63$을 기준으로 적합도 검정 실시하여 예상과 실제 빈도 차이가 작은 결과 선별
  2. MCC 상위 100개 추출
  3. Accuracy가 가장 높은 모델 최종적으로 선택

06 결과 검정 및 해석

실험 결과 분석

- 본 실험은 모델 수 5가지, 전처리 경우의 수 16가지, 데이터 증강 기법 10가지의 조합으로 구성된 800개의 데이터에 대해 진행
- 카이제곱 검정을 통과한 결과 즉, 예상과 실제 빈도 차이가 적다고 볼 수 있는 결과 중 MCC 상위 100개의 결과 분석

Result



	DATA_PATH	MCC	ACC
0	LGBM vbi_train3-6-res.csv	0.367636	0.910482
1	LGBM vbi_train4-6-res.csv	0.367912	0.891251
2	RF of_train1-6-res.csv	0.369186	0.874718
3	RF of_train1-5-res.csv	0.381814	0.861708
4	LGBM vbi_train2-5-res.csv	0.428462	0.857054
5	LGBM vbi_train3-5-res.csv	0.445409	0.853030
6	RF of_train1-4-res.csv	0.393283	0.851259
7	RF obi_train1-4-res.csv	0.358580	0.839720
8	LGBM vbi_train2-4-res.csv	0.433108	0.836331
9	RF of_train2-6-res.csv	0.379429	0.822859

- LightGBM을 사용했을 때 Accuracy 91.04%로 가장 좋은 성능을 거두었음
- 분석 대상인 MCC 상위 100개의 결과는 Paradox of Accuracy의 한계점을 극복했다고 판단할 수 있음

06 결과 검정 및 해석

실험 결과 분석

MCC 상위 100개의 결과 분석

- 데이터 증강 기법 (0~9)

	0	1	2	3	4	5	6	7	8	9
분할	80	80	80	80	80	80	80	80	80	80
기댓값	10	10	10	10	10	10	10	10	10	10
관측값	1	10	10	8	17	16	13	9	8	8

→ 데이터를 증강하지 않은 경우는 상위 100개 중 단 하나만 포함됨

→ SMOTE 방법론이 가장 우수한 결과를 도출

- 변수 추가(o) / 변수 미추가(v)

	변수 추가 (o)	변수 미추가 (v)
분할	400	400
기댓값	50	50
관측값	46	56

→ 변수 추가 여부에서는 비슷한 결과 도출

- Model

	Cat	LGBM	RF	LR	SGD
분할	160	160	160	160	160
기댓값	20	20	20	20	20
관측값	0	19	24	43	14
평균등수	Nan	43	31	58	62

→ 5개의 모델 중 CatBoost는 상위 100개 중 단 하나도 포함되지 않음

→ 기댓값을 기준으로 LogisticRegression이 가장 유의한 결과를 냄

- 전처리 방식(1~4)

	1	2	3	4
분할	200	200	200	200
기댓값	25	25	25	25
관측값	53	30	3	14

→ Label Encoding이 One-Hot Encoding보다 해당 분석에서 더 적절

- 서브모델링 O(bi) / X(f)

	bi	f
분할	400	400
기댓값	50	50
관측값	81	19

→ Sub-Modeling을 한 경우가 압도적으로 좋은 결과 도출

한계점 및 분석 의의

07

07 한계점 및 분석 의의

최종 결론

한계점

1. 하이퍼 파라미터 튜닝을 국한된 범위에서만 진행
2. 보험 데이터가 지니는 특수성 때문에 타 불균형 데이터에 해당 방법론을 적용하기 어려울 수 있음
3. 한 가지 변수 선택 방법론만 적용했기 때문에 편향된 결과가 나타날 수 있음

분석 의의

1. 데이터 증강의 비율을 다양화하여 우수한 성능을 거둠
2. 통계적 가설 검정을 통해 평가지표 분석을 다양화하였음
3. Sub-Modeling을 통해 상이한 분포의 데이터의 학습을 가능하게 함