자동차 보험 가입 예측

Health Insurance Cross Sell Prediction

3기 이형주

프로젝트 개요

- 건강 보험 가입 고객이 차량 보험에 가입할 가능성 예측

데이터셋 설명

- 데이터 셋: Kaggle Health Insurance Cross Sell Prediction

- 주요 수치형 변수

변수명	설명
Age	나이
Annual_Premium	연간 보험료 금액
Vintage	고객의 건강 보험 지속기간

데이터셋 설명

- 주요 범주형 변수

변수명	설명
Gender	성별(Male, Female)
Driving_License	운전 면허 보유 여부(No, Yes)
Vehicle_Age	차량 연식(<1년, 1-2년, >2년)
Vehicle_Damage	차량 사고 경험 여부
Response	차량 보험 가입 의향

데이터 전처리

- 연간 보험료: 이상치 제거

	Annual_Premium
	381109.000000
	30564.389581
	17213.155057
Min	2630.000000
	24405.000000
Median	31669.000000
	39400.000000
Max	540165.000000

```
cost = df["Annual_Premium"]
q1 = cost.quantile(0.25)
q3 = cost.quantile(0.75)

IQR = q3 - q1

lower = q1 - 1.5 * IQR
upper = q3 + 1.5 * IQR

outliers = df[(cost < lower) | (cost > upper)]
print(f'lower bound : {lower}, upper bound: {upper}

# 이상치 제거 : 상한 및 하한 설정
filtered = df[(cost >= lower) & (cost <= upper)]
```

```
lower bound : 1912.5, upper bound: 61892.5
 이상치 개수: 10320, 이상치 비율 = 0.03
```

데이터 전처리

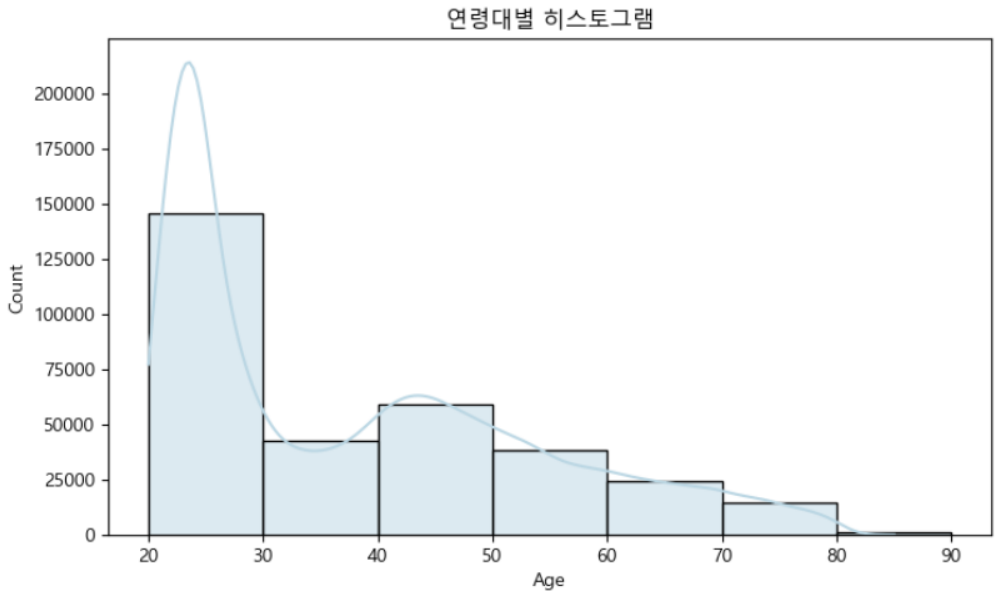
- 라벨 인코딩

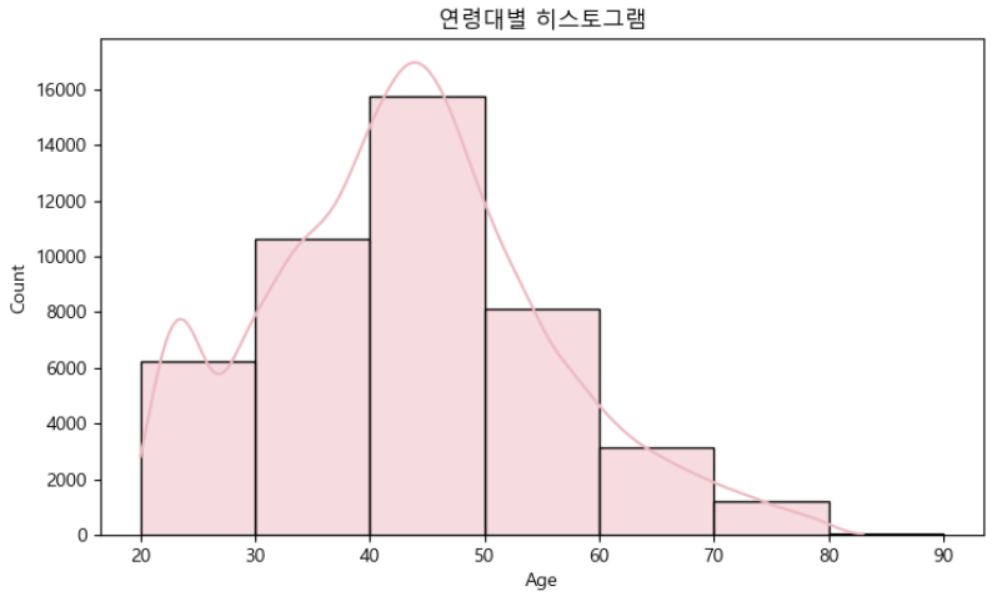
```
Gender: ['Female' 'Male']
Vehicle_Age: ['1-2 Year' '< 1 Year' '> 2 Years']
Vehicle_Damage: ['No' 'Yes']
```

```python
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

# 범주형 데이터 라벨인코딩
df = train.copy()
df['Gender'] = encoder.fit_transform(df['Gender'])
df['Vehicle_Age'] = encoder.fit_transform(df['Vehicle_Age'])
df['Vehicle_Damage'] = encoder.fit_transform(df['Vehicle_Damage'])
```

Gender	Age		Vehicle_Age	Vehicle_Damage
1	44		2	1
1	76		0	0
1	47	...	2	1
1	21		1	0
0	29		1	0

EDA 요약 (1)

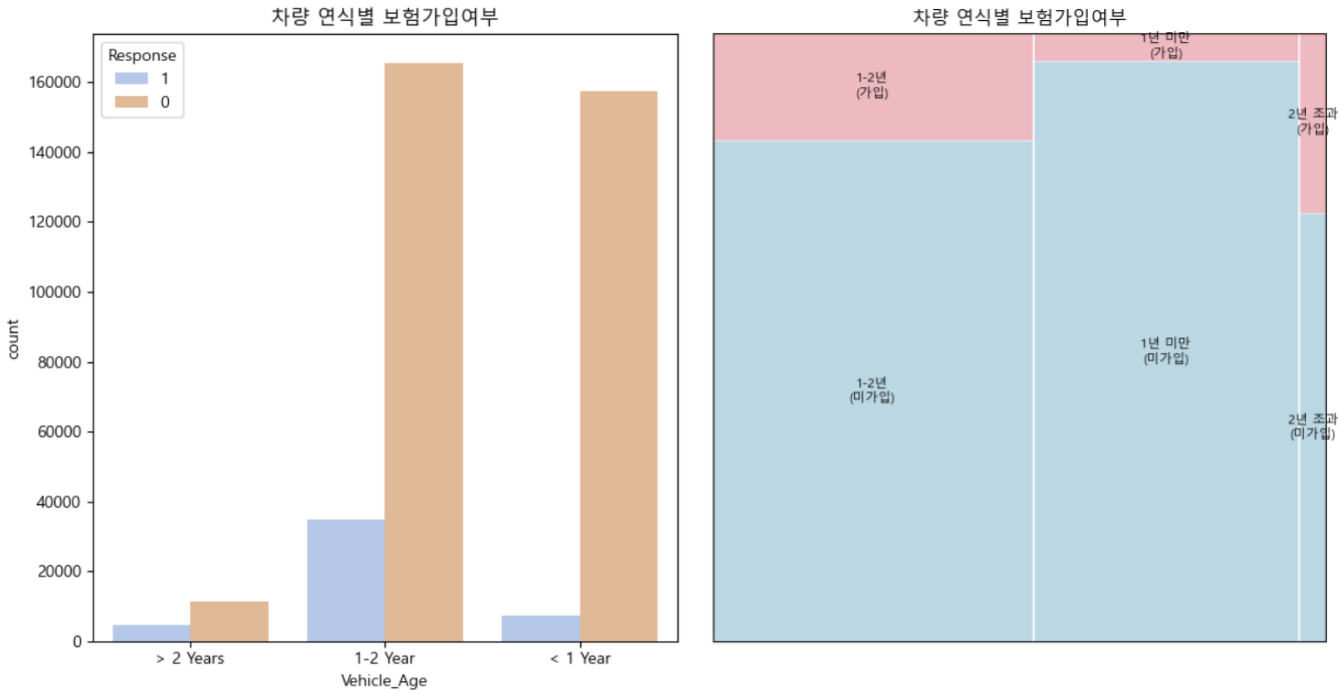




- 가입 O: 40대 > 30대 > 50대
- 가입 X: 20대 > 40대 > 30대

→ 연령대 → 보험 가입에 영향

EDA 요약 (2)



- 차량이 오래될수록 보험에 가입하는 경향이 높음

Feature Engineering

- **표준화** : StandardScaler
- **파생변수** 생성
  - Age_Premium: 나이에 따른 보험료 변화 경향

    → 나이가 많을수록 보험료가 높다는 보험 자체의 특징에 기반
  - Vintage_Year: 지속 일수를 100일 단위로 3개의 범주로 구분

Feature Engineering

- 파생변수 생성

```
filtered['Age_Premium'] = filtered['Annual_Premium']/filtered['Age']
```

```
bins = [0, 100, 200, 300]
labels = [0,1,2]
filtered['Vintage_Long'] = pd.cut(df['Vintage'], bins = bins, labels = labels, right = False)
```

```
filtered.head()
```

riving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response	Age_Premium	Vintage_Long
1	28.0	0	2	1	40454.0	26.0	217	1	919.409091	2
1	3.0	0	0	0	33536.0	26.0	183	0	441.263158	1
1	28.0	0	2	1	38294.0	26.0	27	1	814.765957	0
1	11.0	1	1	0	28619.0	152.0	203	0	1362.809524	2
1	41.0	1	1	0	27496.0	152.0	39	0	948.137931	0

Feature Engineering

- 최종 데이터

```
data.head()
```

	Gender	Age	Vehicle_Age	Vehicle_Damage	Annual_Premium	Vintage	Age_Premium	Vintage_Long	Response
0	1	0.345182	2	1	0.758959	0.748826	0.083977	2	1
1	1	2.417701	0	0	0.289720	0.342470	-0.799130	1	0
2	1	0.539480	2	1	0.612449	-1.521990	-0.109293	0	1
3	1	-1.144442	1	0	-0.043793	0.581503	0.902912	2	0
4	0	-0.626312	1	0	-0.119965	-1.378570	0.137038	0	0

결과 요약 및 인사이트

- 분석 결과 요약 및 인사이트
  - 중년층에서의 가입률이 가장 높고, 청년층의 가입률이 낮음
  - 차량이 오래될수록 보험 가입률이 높음

    → 고객 군 별로 보험 필요성을 강조한 마케팅 활용
- 한계점
  - 일차원적으로 X와 Y 두 변수 간의 관계만 분석
  - 크게 유의미한 관계를 발견하지 못함

향후 계획 및 개선 방향

- 추가 분석 아이디어
  - Y와의 관계와는 상관 없이 X변수 간에 조합하여 분석해보기
- 데이터 보완 계획
  - 가입 경로에 대한 정보 확보해보기

마무리

- 새로운 데이터를 혼자 살펴본 경험은 처음이라 어떤 방향으로 분석을 진행해야 할 지 막막..
- 다양한 그래프 그리는 방식에 대해 좀 더 알아보고 싶음