데이터 증강 및 통계적 가설 검정을 통한 불균형 보험 데이터 예측

일등 (최) 해원 – 김보경 김현우 이준희 임예림 최해원

1. 주제 선정 이유

머신러닝 기법을 활용한 예측의 과정에서 사용되는 데이터의 품질은 최종 예측 성능에 굉장히 많은 영향을 끼치는 것으로 알려져 있다. 하지만 데이터의 수집 과정과 수집 방법에 따른 데이터의 품질이 예측의 성능을 좌우하는 경향성을 불균형 데이터에서는 찾아보기 어렵다. 일례로 예측의 평가 지표로 가장 많이 사용되는 정확도(Accuracy)는 모델의 학습 양상과 무관하게 불균형 데이터에서 대부분 높은 수치를 기록한다. 두 범주의 비율이 굉장히 편향되어있기에 복잡한 데이터의 형태를 학습하여 예측하지 않고 단순히 다수의 범주로 통일하여 예측할 경우에도 높은 정확도를 기록할 수 있기 때문이다. 이러한 상황을 Paradox of Accuracy라 일컫기도 한다.

정확도 지표를 통해 예측 성능을 측정하는 대다수의 경우에서는 해당 문제를 간과할 수 있다. 하지만 실생활에서는 다수의 경우를 높은 확률로 예측에 성공하는 것 만큼 소수의 경우를 판별하여 정확하고 예리하게 예측하는 것이 중요한 경우들이 존재한다. 신용카드 사기와 같이 개인에게 심각한 악영향을 끼치는 예시를 생각해볼 수 있다. 하루에 신용카드를 통한 결제가 이루어지는 사례는 셀 수 없이 많고 이 중 사기 거래가 이루어지는 경우는 극히 드물 것이라는 것을 알 수 있다. 따라서 어떤 모델이 특정 기간 동안의 신용카드 거래 정보를 수집하여 학습하고 정상/사기 거래를 예측하는 과정에서 모든 거래가 정상이라 예측하여도 정확도는 99퍼센트 이상을 상회할 것이다. 하지만 높은 정확도의 이면에는 사기 거래의 피해자를 수혜할 수 있는 방안이 사라지는 단점이 존재하게 되는 것이다.

이를 해결하기 위해서 정확도 뿐만 아니라 정밀도(Precision), 재현율(Recall) 그리고 이 두 지표를 적절히 혼합한 F1 지표를 사용하여 다각화된 분석을 진행하고 있다. 다양한 평가지표를

통해 분석의 완결성을 더하는 것은 분명 불균형 데이터를 통한 예측 문제의 한계점을 해결하는 데에 도움을 줄 수 있으나 다분히 사후적이라는 판단에서 벗어날 수 없다. 따라서 본 분석에서는 [보험 데이터]를 선택하여 여러 데이터 증강 기법을 적용해 예측 문제를 해결하며 결과에 대한 검정을 통해 실질적인 해결책을 제시하고자 한다.

2. 데이터 설명

표 1과 표2를 통해 데이터의 개수, 결측치 존재 유무, 각 열의 Data Type을 확인할 수 있다. 본 분석에 사용된 데이터는 382,154개의 행과 12개의 열로 이루어져 있다. Gender, Driving License, Region Code, Previously Age, Vehicle Damage, Vehicle Age, Annaul Premium 변수들은 범주형 변수들이며 그 외 변수들은 연속형 변수이다. 또한 표 1을 통해 해당 데이터에는 결측치가 존재하지 않는 것을 확인할 수 있다.

분석에 사용된 데이터는 11개의 설명변수와 Response라는 종속변수로 구성되어있다. 설명변수 중 Id, Gender, Age, Region Code 변수들을 통해 고객들의 기본 정보를 알 수 있으며, Driving License, Vehicle Age, Vehicle Damage 변수들을 통해 고객들의 자동차 관련 정보들을 알 수 있다. 또한 고객들이 자사 보험과 1회 이상 접촉한 후에 가입 의향을 조사한 데이터이므로 Policy Sales Channel 변수와 Vintage변수를 통해 자사 보험 접촉 경로 및 접촉 시기 등을 알 수 있다. 추가적으로 접촉했을 당시 책정된 보험료를 Annual Premium 변수에서 확인할 수 있으며 Response는 종속변수로 고객이 자사 보험에 가입하고 싶은지에 대한 여부에 대한 정보를 담고 있다.

표 3 을 통해 각 변수들의 평균, 표준편차, 최대/최소 값 등을 확인할 수 있다. 통계량을 시각화 하기 위해 범주형 변수는 Countplot 및 Histogram을 바탕으로, 연속형 변수는 Boxplot, Histogram을 기반으로 개별 변수의 특징을 확인했다.

그림 4의 Boxplot을 통해 Age, Vintage 변수 모두 IQR기준의 이상치는 없는 것으로 확인

되며 Age 변수의 중위수가 다소 왼쪽으로 치우쳐져 있는 것을 제외하고 Vintage 변수의 Boxplot은 비교적 이상적인 형태를 띄고 있다. 이에 비해 Annual Premium 변수는 Boxplot의 형태가 편향되어 있음에 따라 다른 값들이 모두 IQR기준의 이상치로 나타나고 있다. Age 변수는 20, 30대와 40, 50대에 데이터가 밀집되어 있다. 따라서 데이터 전처리 과정에서 연령대를 grouping 하는 변수를 추가했다. 또한 Annual Premium의 경우 Histogram을 통해 특정 값의 빈도가 굉장히 높은 것을 파악할 수 있는데 해당 값은 표 6에서 확인할 수 있듯이 Annual Premium 내의 최소값인 2,630이다. 또한 이는 추가적인 옵션 없이 기본 보험료를 내는 고객의 특성으로 해석할 수 있으며 그림 5와 같은 분포를 통해 시각화할 수 있다. 따라서 기본 보험료를 지불하는 고객과 그렇지 않은 고객들을 다른 범주로 구분하는 변수를 추가했다. 마지막으로 Vintage 변수의 histogram은 고른 분포를 보이고 있음을 알 수 있다.

그림 7에 표현된 범주형 변수인 Gender 변수와 Previously Insured, Vehicle Damage 변수는 각 범주의 빈도수가 비슷하다. Driving License 변수는 1에 해당하는 값, 즉 면허를 소지한 고객들이 소지하지 않은 고객들 보다 많이 존재함을 알 수 있다. 표 8에서 확인할 수 있듯이 면허를 소지하지 않은 고객은 723명으로 전체 데이터의 0.2%에 불과하다. 해당 케이스를 전처리 과정에서 의미론적 이상치라 판단한 후 값을 제거하였다.

Policy Sales Channel 변수는 고객이 이용한 채널에 대한 정보를 담은 변수이다. 이 변수를 통해 특정 소수의 채널(152, 26, 124 채널)을 이용하는 고객들이 타 채널을 이용하는 고객들보다 상대적으로 많이 분포하는 것을 알 수 있다. 따라서 해당 채널들을 주요 채널들로 설정하여 새로운 변수로 추가하였다. 다음으로 Vehicle Age 변수는 본인 소유 차량 연식이 1년 미만의 경우와 1년에서 2년 사이인 경우의 고객 수는 비슷하지만 2년 초과인 경우의 고객 수는 다른 경우에 비해 적다. 마지막으로 Region Code 변수는 Policy Sales Channel과 마찬가지로 특정 지역의 데이터 개수가 많은 것을 확인할 수 있다. 표10에서 알 수 있듯이 전체 데이터 중 지역 코드가 28인 경우가 107199 개로 전체 데이터의 28%를 차지한다. 따라서 인구 밀집 지역에 사는 고객들을 따로 분리하도록 추후에 변수를 새로 추가하였다.

최종적으로 종속변수인 Response의 분포를 살펴보면 Response가 0 인 경우, 즉 고객이 자사 보험 가입 의향이 없는 경우가 의향이 있는 경우보다 많다. 의향이 없는 경우는 전체 데이터의 83.6%, 의향이 있는 경우는 전체 데이터의 16.4%로 종속 변수의 분포가 imbalance 한 것을 확인할 수 있다.

그림 12는 연속형 설명변수들 간의 Correlation plot이다. 그림 12를 확인했을 때 Age 변수와 Policy Sales Channel의 상관관계가 0.58인 것을 제외하고 다른 연속형 변수들 간의 상관관계 값이 절대적으로 작아 유의한 결과를 찾아볼 수 없다.

그림 13과 표 14는 Vehicle Damage변수와 Vehicle Age변수의 관계에 대한 정보를 담고 있다. 본인 소유 자동차 연식이 1년 미만임에도 불구하고 자동차 파손 이력 여부가 있는 고객들이 41,634명이 있는 것을 확인할 수 있으며 해당 고객들을 고위험군으로 분류할 수 있다. 이에 반해 본인 소유 자동차 연식이 2년 초과일 때 자동차 파손 이력 여부가 없는 고객들이 11명이 있는 것을 볼 수 있고 해당 고객들을 저위험군으로 분류할 수 있다.

그림 15와 표 16은 Vehicle Damage 변수와 Previously Insured변수의 관계에 대한 정보를 담고 있다. 타사 보험을 가입한 적이 존재하지 않지만(Previously_Insured : 0) 자동차 파손을 한 경험이 존재하는 (Vehicle_Damage : Yes) 고객들이 총 175,282명이 있음을 확인할 수 있다. 해당 고객들은 보험을 가입하지 않고 자동차 파손을 경험한 적이 있기에 위험군으로 분류할 수 있다.

그림 17과 표 18은 Vehicle Age 변수와 종속변수인 Response변수의 관계에 대한 정보를 담고 있다. 본인 소유 자동차 연식이 1년 미만일 경우 자사 보험 가입 의향이 없는 비율 (Response = 0)이 95%, 본인 소유 자동차 연식이 1년 이상 ~ 2년 이하일 경우 자사 보험 가입 의향이 없는 비율이 76%, 본인 소유 자동차 연식이 2년 초과일 경우 자사 보험 가입 의향이 없는 비율이 60%이다. 이 결과 자동차 연식에 따라서 종속변수의 비율이 변한다고 판단할 수 있으며, 해당 분석 결과를 이용하여 Sub Modeling을 진행했다.

마지막으로 그림 19와 표 20은 Previously Insured 변수와 종속변수인 Response 변수의 관계에 대한 정보를 담고 있다. 타사 보험 가입을 하고, 자사 보험 가입 의향이 있는 135명의 고객이 존재하며 해당 고객의 연령대 분포는 그림 21과 같이 나타난다. 해당 고객들의 연령대가 전체 데이터의 연령대에 비해 낮은 것을 알 수 있다. 따라서 타사 보험 가입을 하고 자사 보험까지 가입한 고객들은 젊고 자산이 많을 가능성이 있음을 파악했다.

3. 전처리 과정

- 이상치 처리

본 분석에서는 총 3가지 데이터 전처리 기법을 적용했다. 처음 과정은 이상치 처리 과정이다. 그림 4에서 확인할 수 있듯이 Boxplot을 확인했을 때는 Annual_Premium 변수 외에는 IQR 기준의 이상치를 확인할 수 없었다. 또한 Annual Premium 변수는 최소값인 2,630이 전체 데이터의 16%를 차지하고 있을 뿐만 아니라 다수의 값들이 상대적으로 작은 값을 가진다. 따라서 값들의 분포가 몰려 있으며 이에 따라 이상치의 개수가 굉장히 많은 것을 확인할 수 있다. 하지만 해당 이상치들을 다 제거하게 될 경우 변수의 고유한 분포를 왜곡시키는 상이한 분포를 형성할 수 있기 때문에 IQR 기준의 이상치 값들을 제거하지 않았다. IQR 기준 이상치를 삭제하지 않은 반면 개별 데이터 자체의 논리가 성립하지 않는 경우를 탐색하여 데이터 수집 과정에서 생긴 오류라고 판단한 후 해당 케이스를 제거했다. 운전면허를 소지하지 않았는데 본인 소유 자동차가 파손된 고객들의 경우가 그 예시이며 이는 논리적으로 성립할 수 없다고 판단하여 해당 데이터를 의미론적 이상치로 판단한 후 제거했다.

- 인코딩

인코딩 과정에서는 Label Encoding과 One-hot Encoding 총 2가지 방법론을 채택하였다. Label Encoding은 각 범주들을 알파벳 순서대로 숫자를 할당하여 변환하는 방법론이다. 이에 달

리 One-hot Encoding은 해당 변수의 각 범주들에 따라 Dummy Variable을 추가하는 방법이다.

- 변수추가

1) Population 변수

본 분석에서는 개별 변수의 분포적 특징 및 변수들 간의 분포적 특징을 바탕으로 총 10개의 새로운 변수를 제작하여 추가하였다. 그림 22는 Region Code변수의 Countplot이다. Region Code 가 26 일 경우 전체의 28%를 차지하므로 주요 지역인 26에 사는 고객들을 하나의 범주로 취급하여 Region Code가 26이면 "Main" 의 값을, 26이 아닐 경우 "Not Main" 의 값을 가지는 Population 변수를 추가했다.

2) Annual Basic 변수

그림 23은 Annual Premium 변수의 Countplot이다. 앞서 언급한 바와 같이 연간 책정된 보험료의 최소값인 2,630을 지불하는 고객이 62,876명으로 대다수를 차지하고 있음을 알 수 있다. 해당 정보를 통해 최저 보험료, 즉 기준 보험료를 지불하는 사람과 옵션을 추가하여 기준 보험료보다 높은 금액을 지불하는 고객을 분류하고자 했다. 따라서 2,630의 보험료를 지불한 고객을 "Basic"의 값, 최소값인 2,630의 보험료보다 많은 금액을 지불하는 고객을 "Option"의 값을 가지는 Annual Basic 변수를 추가했다.

3) Beneficiary 변수

그림 15에서는 사전에 타사 보험을 가입한 적이 있으며 자동차 파손을 당한 이력이 있는 고객을 확인할 수 있다. 본 분석 과정에서는 해당 고객을 타사 보험으로부터 보험료를 수령했던 고객으로 판단하여 "Benefit"의 값을 그렇지 않은 데이터는 "Not Benefit" 의 값을 가지는 Beneficiary 변수를 추가했다.

4) Danger 변수

그림 13에서는 자차 연식이 1년 미만임에도 불구하고 자동차 파손 경험이 있는 고객들을 확인할 수 있으며 해당 고객들의 운전 습관이 위험하다고 판단했다. 따라서 보험의 필요성이 높은 고객으로 분류하여 "High"의 값을 가지도록 설정했다. 이에 비해 본인 소유 자동차 연식이 2년 초과이고 자동차 파손 경험이 없는 고객들을 운전 습관이 안전한 고객으로 분류하여 "Low"의 값을 가지도록 설정했다. 이 외 해당 범주에 속하지 않은 고객들은 "Mid"의 값을 부여하여 총 3개의 범주를 가지는 Danger 변수를 추가했다.

5) N Daner 변수

그림 15를 통해 타사 보험을 가입한 적이 없는 고객 중(Previously_Insured : 0) 자동차 파손을 한 적이 있는 (Vehicle_Damage : Yes) 고객들이 총 175,282명이 있음을 확인할 수 있다. 해당 고객들은 보험을 가입하지 않고 자동차 파손을 경험한 적이 있기에 또 다른 위험군으로 분류하여 "High"의 값을 부여했고 케이스에 부합하지 않은 고객들은 "Low" 의 값을 부여하여 N_Danger의 변수를 추가했다.

6) Age Damaged 변수

그림 24를 통해 연령대에 따라 사고 유경험자의 비율이 상이해지는 경향성을 확인할 수 있다. 25세 이하 연령대의 사고 유경험자 비율에 비해 37세~49세, 50세 이상 연령대의 사고 유경험자 비율이 높은 것을 볼 수 있다. 해당 분석 결과를 이용하여 각 연령대의 사고 유경험자 비율을 바탕으로 Age Damaged 변수를 추가하였다. 25세 이하의 연령대는 0.27, 26세 이상 36세 이하의 연령대는 0.38, 37세 이상 49세 이하의 연령대는 0.69, 50세 이상의 연령대는 0.39의 값을 지니고 있음을 그림 23을 통해 확인할 수 있다.

7) Main Channel 변수

앞서 언급한 바와 같이 그림 25와 표26을 통해 3개의 채널의 고객 수가 타 채널보다 상대적으로 많은 것을 볼 수 있다. 고객의 수가 많은 채널은 152, 26, 124 채널이며, 본 분석에서는 해당 채널을 주요 채널로 선정하였다. 주요 채널은 "Main Ch"의 값을, 주요 채널이 아닌 값을

"Not Main Ch" 의 값으로 설정하여 Main Channel 변수를 추가하였다.

8) Age Channel

그림 27은 36세 이하의 고객을 Grouping한 후 Policy Sales Channels 값의 분포를 시각화한 자료이다. 해당 연령대의 고객들이 이용한 주요 보험 판매 채널을 확인했을 때 채널 152와 채널 160임을 알 수 있다. 그림 28은 36세 초과의 고객을 Grouping한 후 Policy Sales Channels 값의 분포를 시각화한 자료이다. 이를 기반으로 36세 초과의 고객들의 주요 보험 판매 채널이 채널 26, 채널 124임을 파악하여 주요 채널들을 연령대에 따라 그룹화했다. 따라서 채널 26, 124는 "Main Over"의 값, 채널 152, 160은 "Main Under" 의 값, 해당 그룹에 속하지 않는 타 채널들은 "Channel"의 값을 가지도록 Age Channel 변수를 추가했다.

9) Age Group

특정 변수 내에서 연령대에 따른 분포의 차이가 다수 나타나기 때문에 연령대를 기준으로 25세 이하의 고객, 26세~36세의 고객, 37~49세의 고객, 50세 이상의 고객으로 총 4개의 그룹을 생성하여 각 그룹마다 ~25 / 26~36 / 37~49 / 50~ 의 값을 부여했다. 해당 값을 이용하여 Age group 변수를 추가했다.

10) Young Rich

그림 19를 통해 타사 보험 가입을 하고, 자사 보험 가입 의향이 있는 135명의 고객을 확인할 수 있고 이들의 연령대 분포를 그림 21에서 확인할 수 있다. 이 135명 고객들의 연령대가 전체 데이터의 연령대에 비해 낮은 것을 알 수 있었고 이에 타사 보험 가입을 하고 자사 보험까지 가입한 고객들을 젊고 자산이 많을 것이라 판단하였다. 이를 통해 해당 그룹을 "YR" 의 값으로, 타 그룹을 "Not YR"의 값으로 설정하여 Young Rich 변수를 추가했다.

4. 모델링 과정 소개

본 분석의 모델링 과정에서 주안점을 둔 부분은 총 4가지이다. Model Selection, Variable

Selection, Sub-Modeling, Data Augmentation이 각 경우이다. 분석에서 사용할 모델을 선택하는 과정에서 임의성을 최대한 배제하기 위해 Model Selection의 과정을 거치게 되었다. 특정 모델을 분석에 사용하기 위해서는 모델과 데이터의 적합성을 평가해야 한다. 이를 위해 전체 데이터의 10%에 해당하는 부분집합부터 개수를 늘려가 전체 데이터의 100%에 해당하는 부분집합까지 총 10개의 부분집합을 생성한 후 데이터의 개수가 증가함에 따라 모델의 성능이 상승하는 관계성을 검토한다. 데이터의 개수와 모델의 성능이 양의 상관관계를 지니고 있음이 파악되면 해당 데이터는 이 모델로 평가하기에 적합하다는 결론을 내릴 수 있게 되는 것이다. 본 분석에서는 그림 29의 결과를 바탕으로 CatBoostClassifier, LGBMClassifier, RandomForestClassifier, LogisticRegression, SGDClassifier의 5개 모델을 선정하였다.

두번째는 Permutation Importance Method를 통한 변수 선택이다. Permutation Method는 Filter Method의 일례로서 개별 변수들의 중요도를 내림차순으로 정렬해 특정 임계점을 넘는 변수만을 선택하여 최종 모형에 사용하는 방식으로 동작한다. 개별 변수의 중요도는 해당 변수가 포함되었을 때의 모델 성능에서, 배제되었을 때의 모델 성능을 빼는 과정을 통해 도출할 수 있다. 해당 변수를 제거하여 성능을 도출하기 위해서는 적합(fit)하는 과정을 변수 별로 두번씩 진행해야 하기 때문에 Permutation Importance는 해당 변수를 Shuffle하여 제거와 동일한 효과를 유도한다. 이 과정에서 중요도가 음수가 나오게 될 경우 해당 변수가 존재하지 않는 것이 모델의 성능을 향상한다고 판단할 수 있기 때문에 임계점은 '0' 으로 설정하는 것이 타당하다.

세번째는 Sub-Modeling이다. Sub-Modeling은 전처리가 완료된 데이터를 두 부분집합으로 분할하여 각각 다른 모델을 통해 병렬적으로 학습한 후 최종 결과를 종합하여 평가하는 방식을 일컫는 용어이다. 앞서 언급한 바와 같이 "Vehicle Age"를 기준으로 데이터를 분할했을 때 특정 부분집합에서 종속 변수가 균형적인 분포를 보이고 있음을 알 수 있었다. Sub-Modeling은 종속변수의 상이한 분포를 각기 다른 모델을 통해 학습하여 불균형한 데이터가 지니는 문제점을 해결할 수 있을 것이라는 기대효과가 존재한다.

네번째는 Data Augmentation 이다. 본 분석에서 주안점으로 두고 있는 부분이 Data Augmentation 이며 세가지 방법론과 세가지 비율을 계획하여 총 9 가지 증강 경우의 수를 생성했다. 각각의 방법론은 Random Over Sampling(이하 ROS), Synthetic Minority Over Sampling Technique(이하 SMOTE), Conditional Tabular Generative Adversarial Network(이하 CTGAN)이 해당되며 종속변수의 두 범주를 5:5, 6:4, 7:3 의 비율로 총 3 가지 경우의 수를 고려했다.

ROS 는 소수 범주에 속하는 데이터를 단순 복제하는 방식으로 불균형 문제를 해결한다. 가상의 데이터를 생성하지 않기 때문에 데이터의 품질은 보존될 수 있지만 동일 데이터로 학습한다는 한계점 때문에 테스트 데이터에 대해 성능이 떨어지는 과적합현상이 생길 수 있다. 이에 비해 SMOTE 는 가상의 데이터를 생성한다는 점에서 ROS 와 차이점이 있다. SMOTE 는 소수 범주에 속하는 데이터를 임의로 두개 선택한다. 이 두 데이터는 굉장히 근접한 위치에 존재해야 하며 두 데이터를 잇는 직선 상에 새로운 데이터를 생성하는 과정으로 샘플링이 진행된다. 직선상에 생성되는 데이터의 품질을 신뢰하기 어렵다는 단점이 있지만 대체적으로 ROS 보다 높은 성능을 보임이 알려져 있다.

생성 모델(이하 GAN)은 최근 인공지능 학계에서 활발하게 연구되고 있는 분야로서 Computer Vision 과 관련된 다양한 문제를 해결하는 데에 사용되고 있다. 생성모델에는 두 네트워크가 동시에 학습되며 학습 데이터의 분포를 학습하여 새로운 데이터를 해당 분포로부터 생성하는 네트워크와 생성된 데이터가 실재하는 데이터인지의 여부를 파악하는 네트워크가 그 종류이다. 보통 GAN 은 비정형 데이터를 기반으로 새로운 이미지를 생성하는 데에 사용된다. 하지만 Conditional Tabular Generative Adversarial Network 는 연속형 데이터와 이산형 데이터가 혼재되어 있는 정형 데이터에서 적용할 수 있게 설계된 생성모델로서 본 분석에서는 CTGAN 이 생성한 새로운 데이터를 Over Sampling 의 관점으로 해석하게 되었다. CTGAN 은 원본 데이터인 T 로부터 생성 데이터 Tsync 를 생성하는 것을 목표로 하며 T 의 변수들이 미지의 joint distribution 을 따른다는 가정을 한다. 원본 데이터는 T-train 과 T-test 로 구분되어 존재한다. 데이터를 생성하는 Generator 는 Variational Auto Encoder(이하 VAE)의 골자를 따르며 Likelihood

Fitness 와 Machine Learning Efficacy 라는 두 축을 이용해 학습을 하게 된다. Likelihood Fitness 는 Tsync 과 T-train 의 joint distribution 사이의 일치성에 대해 평가하게 되며 Machine Learning Efficacy 는 T-train 으로 학습한 모델의 성능과 Tsync 로 학습한 모델의 성능의 일치성에 대한 평가를 진행하게 된다.

최종 모델링 과정에서는 전처리가 완료된 데이터에 변수 선택과 데이터 증강을 순차적으로 적용하여 사전에 선택된 모델의 성능을 기록하게 된다. 또한 실험에 사용되는 데이터는 Model AB-trainN1-N2.csv 의 형태를 띄게 된다. Model 은 앞서 모델 선택에서 선정된 CatBoostClassifier, LGBMClassifier, RandomForestClassifier, Logistic Regression, SGDClassifier 의 다섯개 중 하나가 되며, A 는 변수 추가 여부에 따라 추가되었을 경우 V, 추가되지 않을 경우 O 가 된다. B 는 sub-modeling 여부에 따라 sub-modeling 을 진행했을 경우 bi, 진행하지 않았을 경우 f 가 된다. N1 은 전처리 경우의 수에 따라 1 부터 4 까지 경우의 수가 되며 N2 는 증강 기법에 따라 0 부터 9 의 값 중 하나를 할당받는다. 따라서 실험에 사용될 전체 데이터의 개수는 800 가지이다.

5. 결과 검정

Accuracy는 머신러닝의 분류 평가 지표 중 하나로 대부분의 분석에서 중요한 지표로서 사용된다. 하지만 불균형 데이터를 분류함에 있어 Accuracy는 더이상 좋은 측도가 될 수 없다. 그 이유는 분류기가 데이터 개수가 더 많은 Real Negative 집단에 편향된 학습을 진행하기에 Real Positive 집단은 거의 고려하지 못하기 때문이다. [1]

또다른 성능 지표로서 Precision 및 Recall과 이 둘의 조화평균인 F-measure가 불균형 데이터를 분류하는데 사용되었다. 하지만 이 세 지표 모두 오로지 True Positive만을 판단하는 기준이기에 True Negative는 전혀 고려하지 않는다는 단점이 있다. [2] 특히 F-measure은 Precision과 Recall의 가중평균으로 표현되고 이 가중치가 분류기에 따라 달라지기에 성능 지표로서 약점을 지니고 있다. [3]

결국, 기존 평가 지표로서 Accuracy와 F-measure은 불균형 데이터를 다루는데 있어 신뢰할 수 없는 결과를 낳는다. [6] 하지만 MCC는 Confusion Matrix의 모든 값을 고려하였기에 다른 성능 지표보다 더 많은 정보를 담고 있다. [4] 또한, 불균형 데이터를 분류할 때 가장 신뢰할 만한 평가 지표로 알려져 있다. [5]

$$Bookmaker\ Informedness = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$$

$$Markedness = \frac{TP}{FN+TP} + \frac{TN}{TN+FN} - 1$$

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

MCC는 Bookmaker Informedness와 Markedness의 기하 평균으로 나타낼 수 있다. Confusion Matrix을 이용하여 피어슨 카이제곱 검정을 통해 파악된 Bookmaker Informedness 및 Markedness의 추정치를 바탕으로 MCC 추정치 또한 계산이 가능하다. 따라서 P-value를 통하여 해당 추정치가 예상된 범위 내의 값인지 판단할 수 있다. [7]

본 분석에서는 MCC 추정치와 $\chi 2(0.99,1) \approx 6.63$을 기준으로 적합도 검정을 실시하였고 이를 바탕으로 예상과 실제로 발생한 빈도의 차이가 적은 결과들을 선별하였다. 그리고 그 결과들 중 MCC가 높은 상위 100개를 추출한 후 Accuracy가 높은 모형을 최종 결과물로 선정하였다.

6. 결과 분석

본 실험은 총 800개의 데이터에 대해서 진행되었으며 각각은 사용한 모델의 개수 5가지, 전처리 경우의 수 16가지, 데이터 증강 기법 10가지 경우의 수의 모든 조합으로 구성되어 있다. 이 중 앞서 언급한 바와 같이 chi검정을 통과한 결과 중 MCC 상위 100개의 결과를 분석하였다.

분석의 대상인 상위 100 개의 결과들은 Accuracy 의 한계점을 극복했다고 판단했으며 그림 32 와 표 33 에 나타나 있는 바와 같이 해당 결과 중 LGBMClassifier 을 사용하였을 때 Accuracy 91.04% 로 가장 좋은 성능을 거두었음을 알 수 있었다.

CatBoostClassifier, LGBMClassifier, RandomForest, LogisticRegression SGDClassifier 의 5 개 모델 중 CatBoostClassifier 은 상위 100 개의 결과물 중 한 개도 포함되지 못했으며 다른 모델은 각각 19 개, 24 개, 43 개, 14 개가 포함되었다. 상위 100 개에 포함될 기대값인 20 개를 기준으로 판단했을 때 LogisticRegression 이 가장 유의한 결과를 기록했음을 알 수 있다. 그림 34 는 모델 별 정확도를 Boxplot 으로 표현하고 있다. 이를 통해 LGBMClassifier 와 RandomForestClasifier 가 여타 모델에 비해 좋은 성능을 거두고 있음을 확인할 수 있다. 또한 모델 별 정확도 지표의 평균 등수를 계산했을 때에는, RandomForest 가 31 등을 기록해 평균적으로 가장 좋은 성능을 기록함을 알 수 있었고, LGBMClassifier 가 43 등, LogisticRegression 이 58 등, SGDClassifier 가 62 등을 기록했다. Boxplot 통해 모델 별 정확도의 분포를 파악할 수 있다.

이후 데이터를 처리한 방법론에 따른 세가지 결과 분석을 진행했다. 첫번째는 변수 추가 여부에 따른 결과 차이 분석이고 두번째는 인코딩/이상치 처리 여부에 따른 결과 차이 분석이며 마지막은 Sub-modeling에 따른 결과 차이 분석이다. 변수 추가 여부에 따른 결과는 새로운 변수를 추가하지 않은 데이터 44개와 새로운 변수를 추가한 데이터 56개가 최종 상위 100개의 결과에 포함되었다. 또한 그림 35의 Accuracy Boxplot을 통해 새로운 변수를 추가하는 것은 성능 향상에 도움이 되었다는 사실을 파악할 수 있었다. 인코딩/이상치는 총 4가지 세부 경우의 수로 분할되는데 1번 경우의 수는 Label Encoding과 이상치 처리 적용이며 2번 경우의 수는 Label Encoding과 이상치 처리 미적용, 3번 경우의 수는 One-Hot Encoding과 이상치 처리 적용, 4번 경우의 수는 One-Hot Encoding과 이상치 처리 미적용이 그 각각의 경우이다. 이 경우 1번과 2번 데이터가 각각 53개, 30개 포함되어 기대값인 25개를 넘어서는 경향성을 보였다. 하지만 그림 36의 Boxplot을 통해 알 수 있듯이 3번 전처리 결과 (One-Hot Encoding / 이상치 처리 적용)가 평균적으로 성능이 가장 좋은 것을 알 수 있었다. 마지막으로 Sub-modeling여부에 따른 분석에서

는 Sub-Modeling을 한 경우가 81개로 압도적인 차이를 보였다. 그림 37의 Boxplot에서 볼 수 있듯이 Sub Modeling을 진행했을 경우의 성능이 압도적으로 높게 측정되었음을 확인할 수 있다.

데이터 증강의 경우에는 사용 방법론 3 가지와 종속변수의 비율 3 가지, 그리고 증강을 적용하지 않은 경우의 수를 포함한 10 가지 데이터가 파생되었다. 데이터를 증강하지 않은 경우는 오직 한 개가 상위 100 개 결과에 포함되었기 때문에 증강을 통해 Imbalance Data 를 효과적으로 처리할 수 있음을 알 수 있었다. 또한 SMOTE 방법론을 적용했을 때의 데이터가 각각 17, 16, 13 개 포함되어 기대값인 10 개를 넘어서는 양상을 보였다. 그림 38 은 증강 기법에 따른 성능을 Boxplot 으로 표현한 자료인데, SMOTE 방법론을 적용한 5, 6, 7 번의 평균 성능이 여타 기법보다 우수함을 알 수 있다.

7. 한계점 및 의의

분석을 진행하는 과정 속에서 3가지 한계점과 3가지 의의를 찾을 수 있었다. 우선 모든 모델링 과정에서 진행한 하이퍼 파라미터 튜닝 과정이 오로지 국한된 범위 내에서만 진행된 것이 첫번째 한계점이다. 하이퍼 파라미터를 올바르게 튜닝하기 위해서는 모든 범위 내에서 시각화를 병행하며 가장 좋은 성능을 보이는 지점을 찾아야 했으나 범용적으로 사용되는 범위 내에서만 이 과정을 진행하였다. 두번째 한계점은 데이터 증강과 통계적 가설 검정이 보험 데이터가 지니는 특수성 때문에 다른 데이터에도 일반적으로 적용하기 어렵다는 것이다. 마지막 한계점은 변수를 선택하는 과정에서 오직 Permutation Importance 방법론만 적용했다는 것이다. Drop Column Importance, IBS 등 다양한 방법을 적용해서 비교하지 못했기 때문에 다소 편향적인 결과가 도출되었을 가능성을 배제하기 어렵다.

반면 데이터를 증강하는 과정에서 종속변수의 두 범주의 비율을 5:5로 통일하지 않고, 6:4의 비율과 7:3의 비율로 경우의 수를 생성한 것이 본 분석의 장점 중 하나라고 생각한다. 또한 통계적 가설 검정을 추가적으로 진행하면서 불균형한 데이터를 평가하는 과정에서 Accuracy를 사

용하는 과정에 당위를 부여할 수 있었던 것이 본 분석의 장점이라고 판단했다. 마지막으로는 Sub Modeling을 진행하면서 상이한 분포의 데이터의 학습을 가능케 하여 성능을 향상했다는 점에서 의의를 찾을 수 있었다.

[참고자료]

[1] Classification_of_imbalanced_data_a_review _ However, for classification with the classimbalance problem, accuracy is no longer a proper measure since the rare class has very little impact on the accuracy as compared to that of the prevalent class.

[2] What the F-measure doesn't measure _ both (or multiple) classes tend to be significant for us, but neither Recall nor Precision take the TN cell of the contingency table into account.

[3] A note on using the F-measure for evaluating data linkagealgorithms _ In short, when looked at from the weighted arithmetic mean perspective, we see that theF-measure is equivalent to using different performance criteria for different linkagemethods.

[4] Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics _ Lastly, the Matthews correlation coefficient (MCC) is a strong metric that considers both accuracies and error rates on both classes, since all the four values in the confusion matrix are in-volved in this formula

[5] Evaluation Measures for Models Assessment over Imbalanced Data Sets _ MCC is considered by some authors as the best singular assessment metric , especially suitable to the case of imbalanced data learning

[6] The advantages of the Matthews correlation coefficient (MCC) over F1 score and

accuracy in binary classification evaluation _ F1and accuracy, instead, generate reliable results only when applied to balanced datasets, and produce misleading results when applied to imbalanced cases.

[7] EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION _ The Matthews/Pearson correlation is expressed in reduced form as the Geometric Mean of Bookmaker Informedness and Markedness // The Geometric Mean of these two overall estimates for the full contingency table is χ2KBM // This is simply the total Sum of Squares Deviance (SSD) accounted for by the correlation coefficient BMG

[시각화 자료]

```
<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  target                382154 non-null  int64
dtypes: float64(3), int64(6), object(3)
```

표 1

Id	조사 응답자 식별번호 (정수형 데이터)
Gender	성별 (범주형 데이터)
Age	나이 (정수형 데이터)
Driving_License	운전면허 소지 여부 (범주형 데이터)
Region_Code	거주 지역 번호 (범주형 데이터)
Previously_Insured	타사 보험 가입 여부 (범주형 데이터)
Vehicle_Age	본인 소유 자동차 연식 (범주형 데이터)
Vehicle_Damage	자동차 파손 이력 여부 (범주형 데이터)
Annual_Premium	연간 책정된 보험료 (실수형 데이터)
Policy_Sales_Channel	이용한 보험 판매 채널 (실수형 데이터)
Vintage	자사 보험 인지 후 경과 일 (정수형 데이터)
Response	자사 보험 가입 의향 (범주형 데이터)

표 2

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
count	382154.000000	382154.000000	382154.000000	382154.000000	382154.000000	382154.000000	382154.000000	382154.000000	382154.000000
mean	234392.953477	38.545691	0.998108	26.406867	0.489182	30711.271362	111.939812	154.189429	0.163811
std	139527.487326	15.226897	0.043455	13.181241	0.499884	17061.595532	54.286511	83.735107	0.370104
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	1.000000	10.000000	0.000000
25%	115006.250000	25.000000	1.000000	15.000000	0.000000	24546.000000	26.000000	81.000000	0.000000
50%	230461.500000	36.000000	1.000000	28.000000	0.000000	31692.000000	145.000000	154.000000	0.000000
75%	345434.750000	49.000000	1.000000	35.000000	1.000000	39447.750000	152.000000	227.000000	0.000000
max	508145.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	163.000000	299.000000	1.000000

표 3

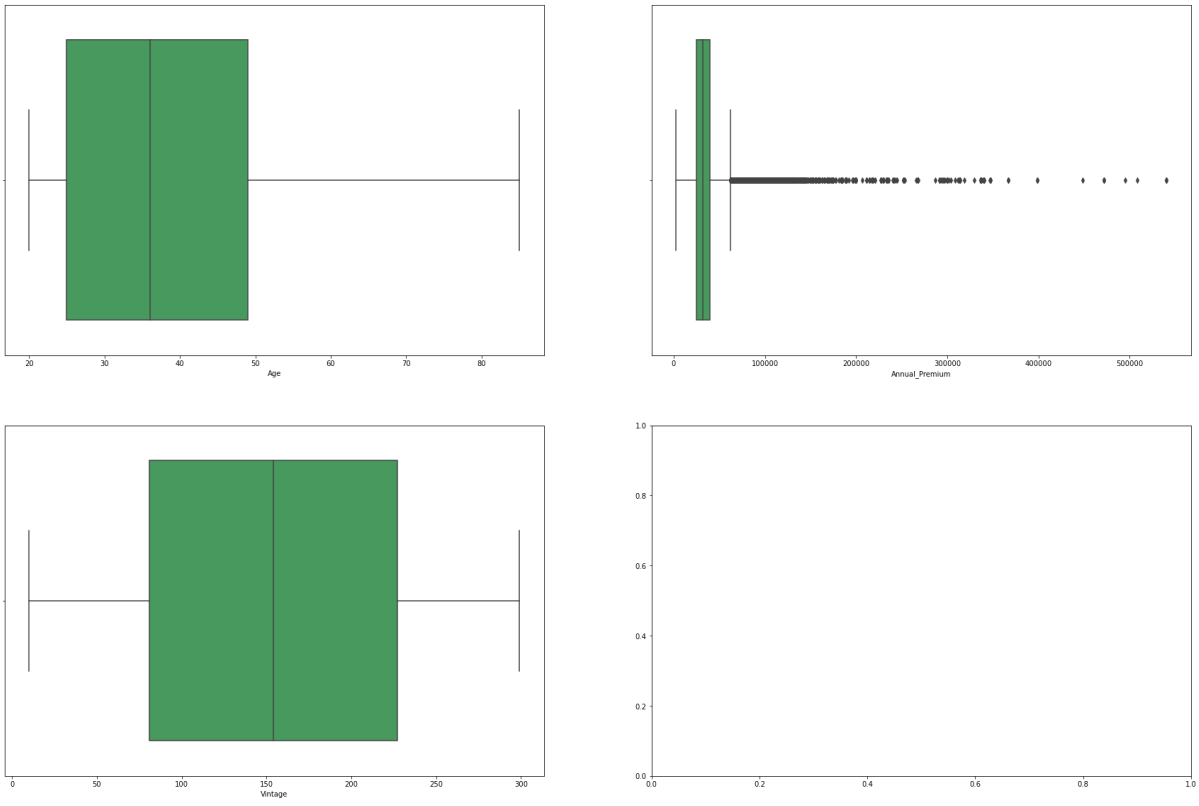

그림 4

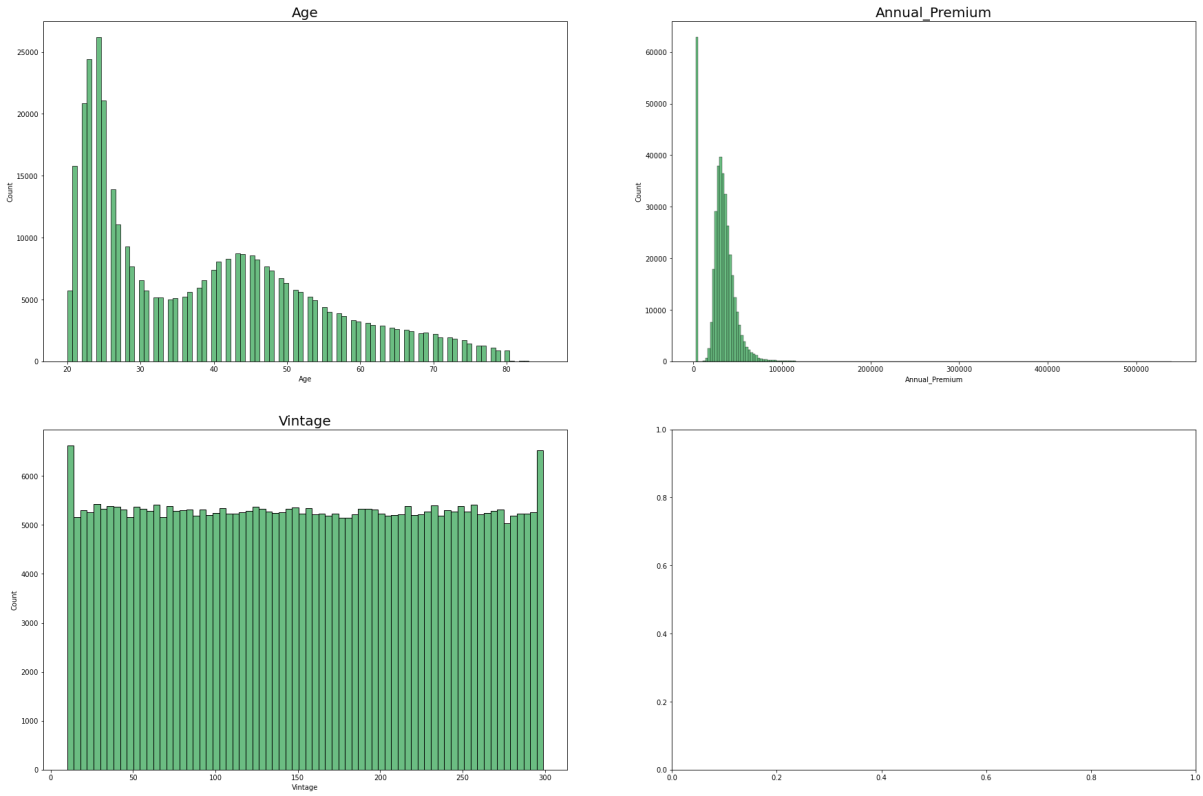

그림 5

```
df['Annual_Premium'].value_counts()
```

2630.0	62876
69856.0	133
38452.0	47
45179.0	45
36086.0	40

표 6

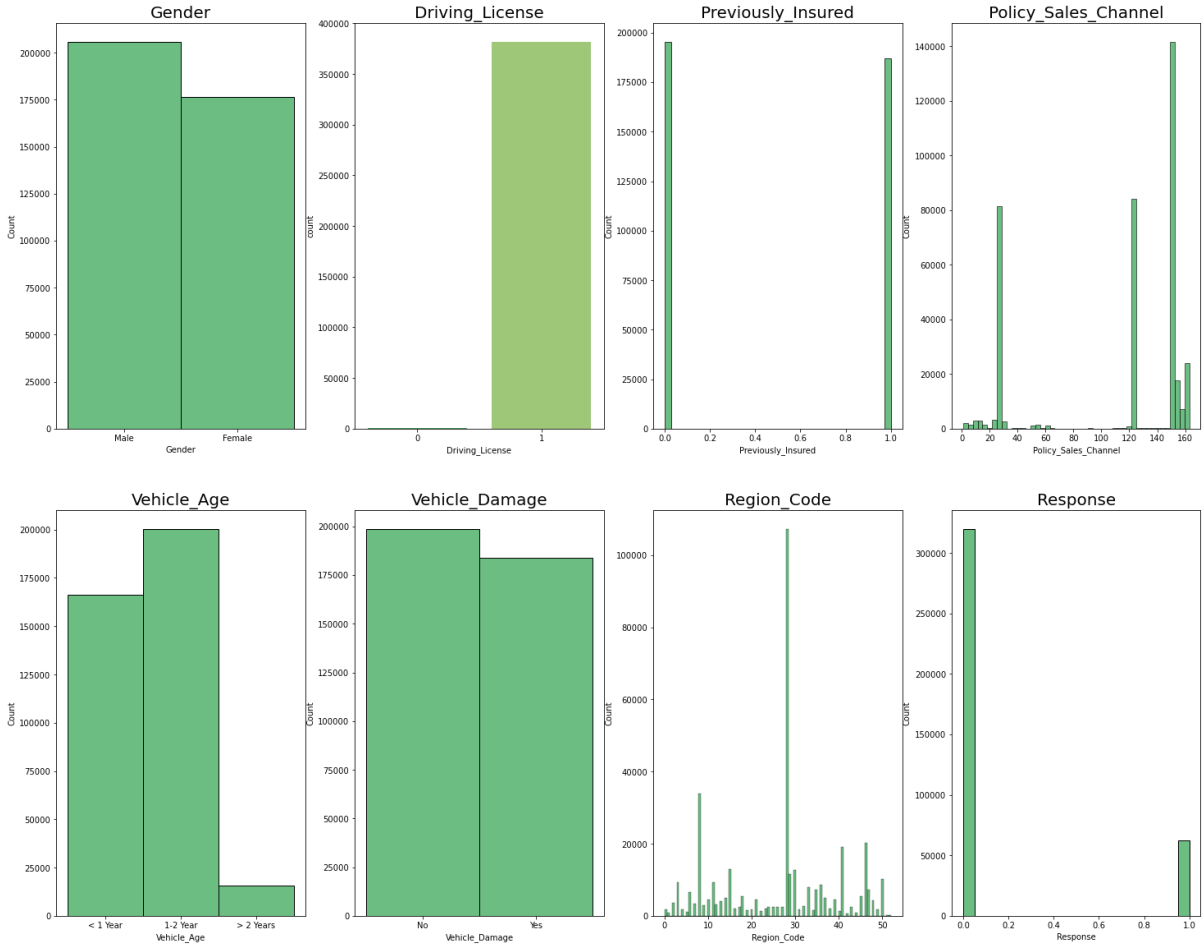

그림 7

```
df['Driving_License'].value_counts()
```

```
1    381431
0       723
Name: Driving_License, dtype: int64
```

표 8

```
df['Policy_Sales_Channel'].value_counts()
```

```
152.0    137422
26.0      81566
124.0     73315
160.0     21045
156.0     10106
```

표 9

```
df['Region_Code'].value_counts()
```

```
28.0    107199
8.0      33941
46.0     20203
41.0     19090
15.0     13071
```

표 10

```
df['Response'].value_counts()
```

```
0    319553
1     62601
```

표 11

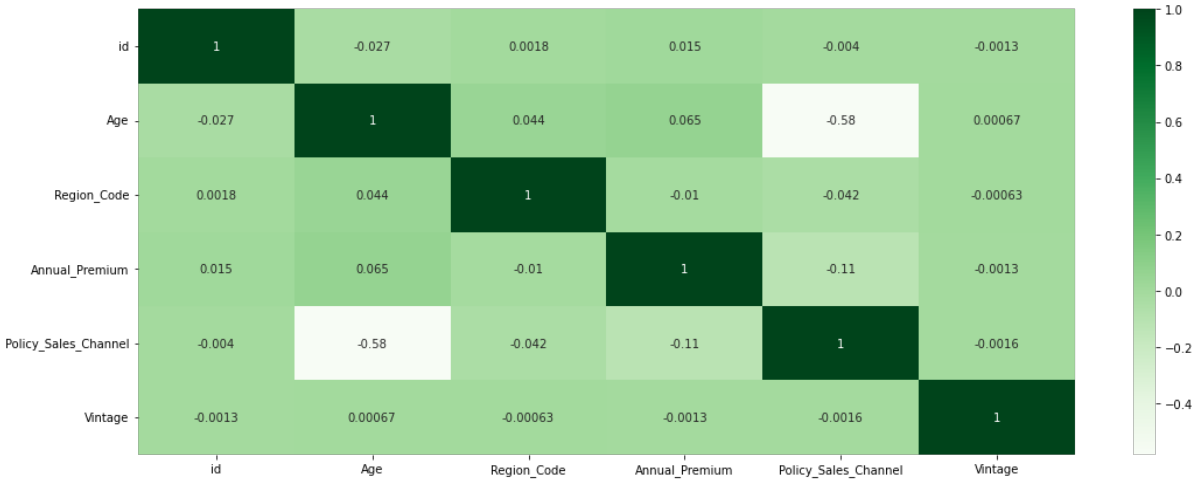

그림 12

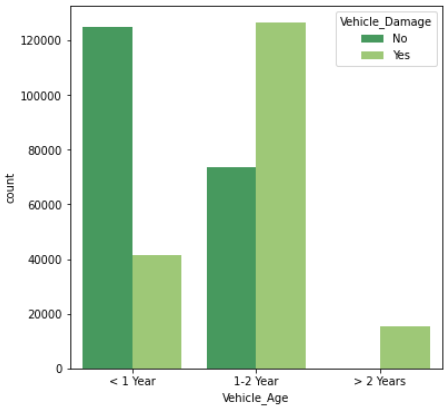

그림 13

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

표 14

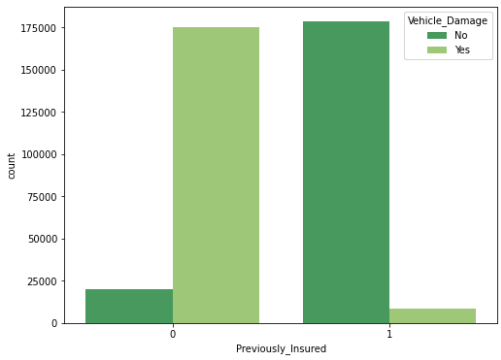

그림 15

Vehicle_Damage	No	Yes
Previously_Insured		
0	19929	175282
1	178572	8371

표 16

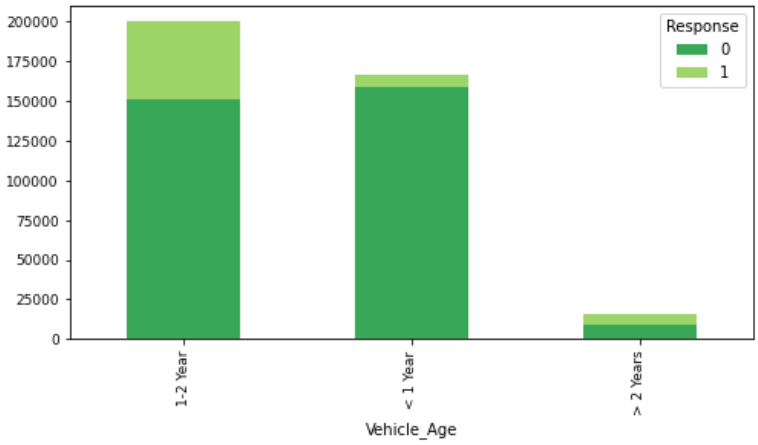

그림 17

Response	0	1
Vehicle_Age		
1-2 Year	151384	48792
< 1 Year	158809	7544
> 2 Years	9360	6265

표 18

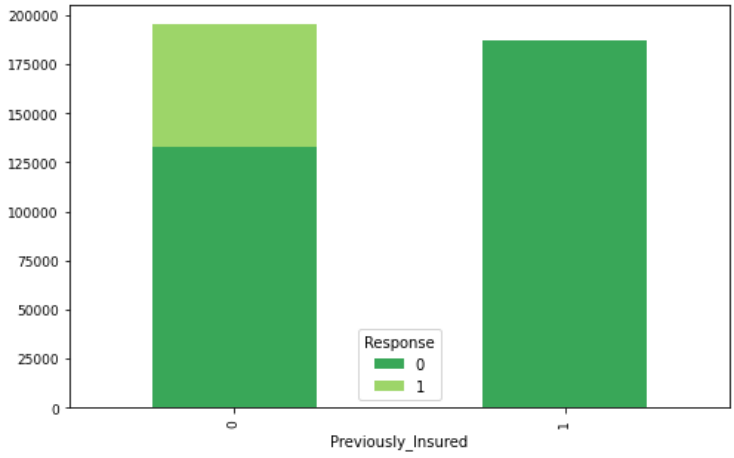

그림 19

Response	0	1
Previously_Insured		
0	132745	62466
1	186808	135

표 20

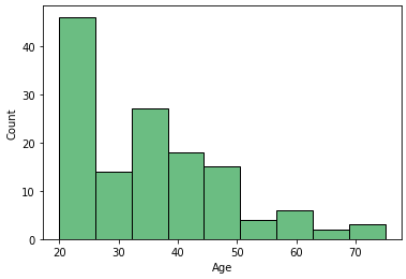

그림 21

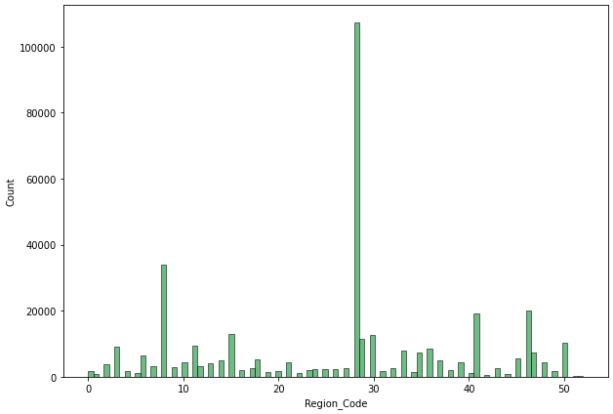

그림 22

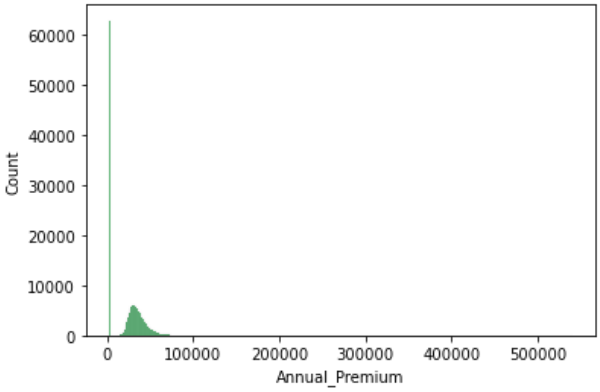

그림 23

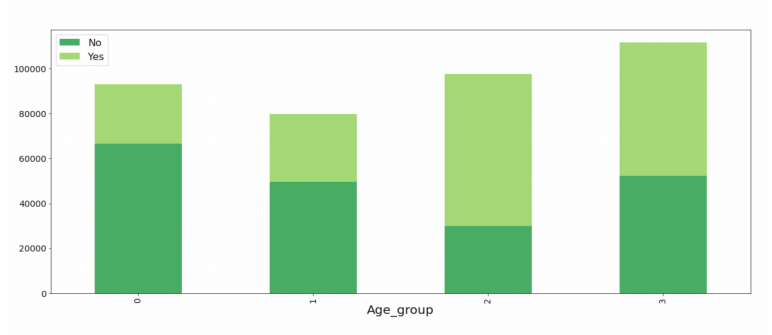

그림 24

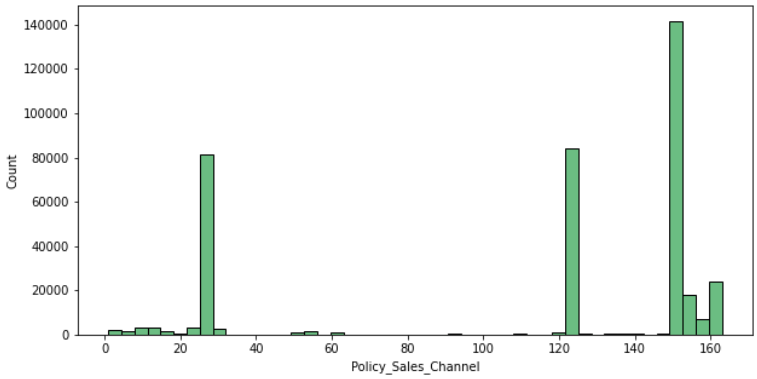

그림 25

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

Name: Policy_Sales_Channel, dtype: int64

표 26

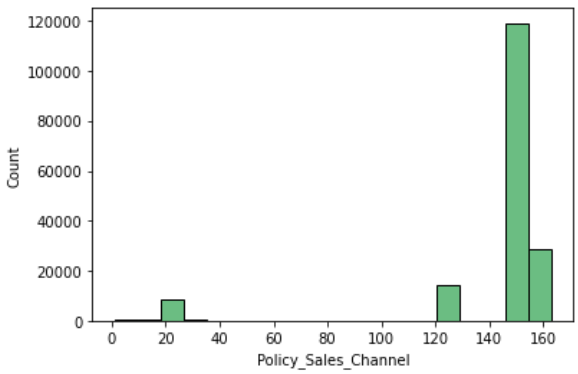

그림 27

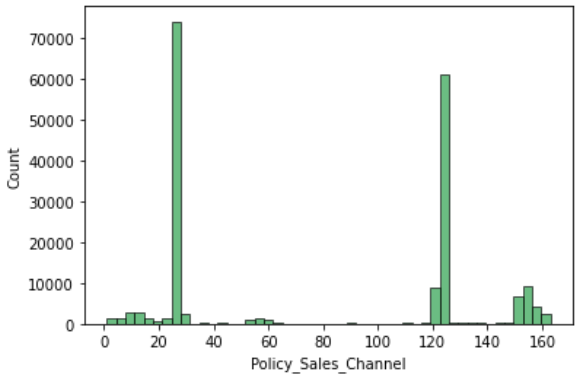

그림 28

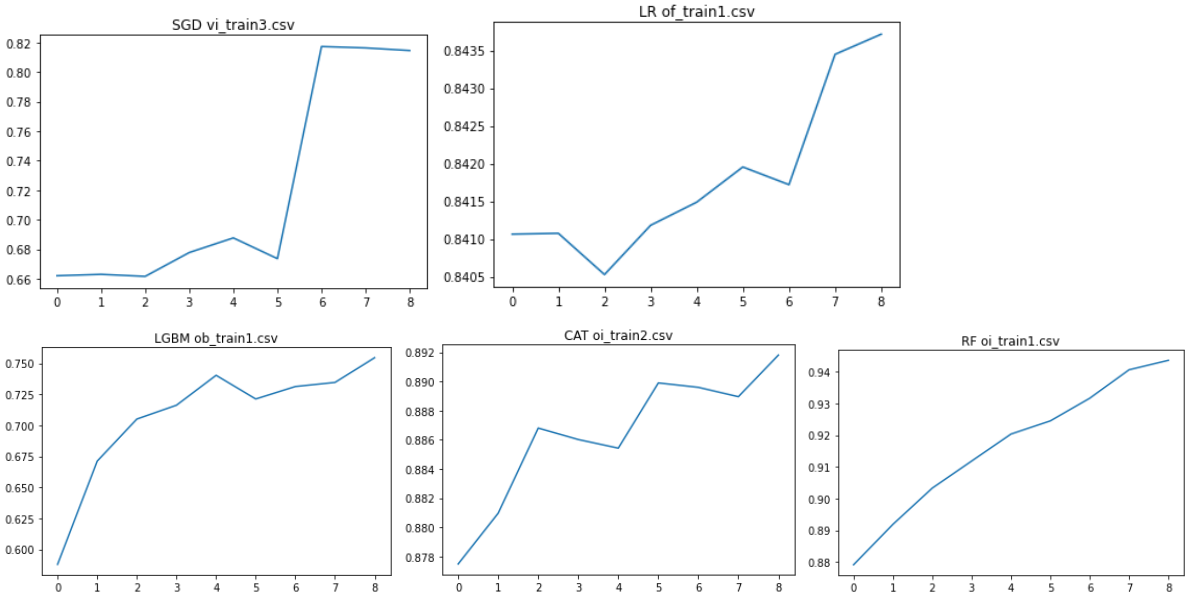

그림 29

Table 2. Confusion matrix.

	Predicted as Positive	Predicted as Negative
Actually Positive	True Positives (TP)	False Negatives (FN)
Actually Negative	False Positive (FP)	True Negatives (TN)

표 30

	+R	−R	
+P	tp	fp	pp
−P	fn	tn	pn
	rp	rn	1

	+R	−R	
+P	TP	FP	PP
−P	FN	TN	PN
	RP	RN	N

그림 31

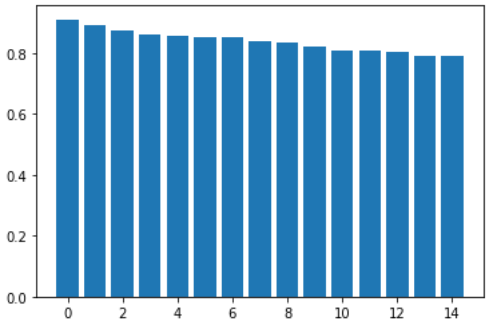

그림 32

	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

표 33

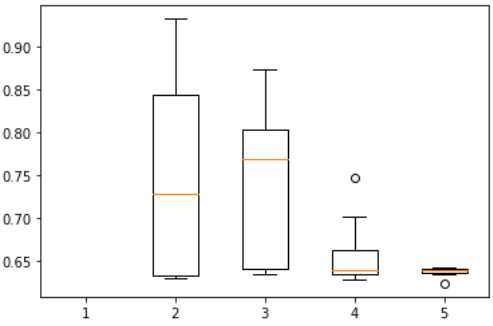

그림 34

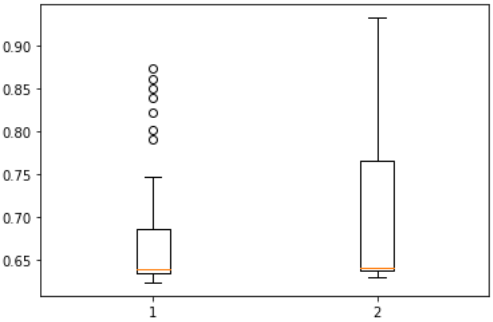

그림 35

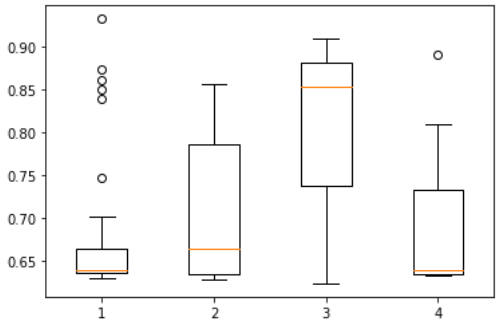

그림 36

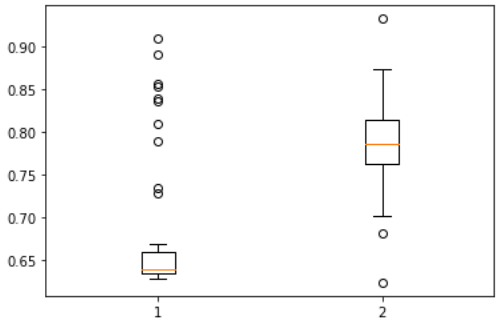

그림 37

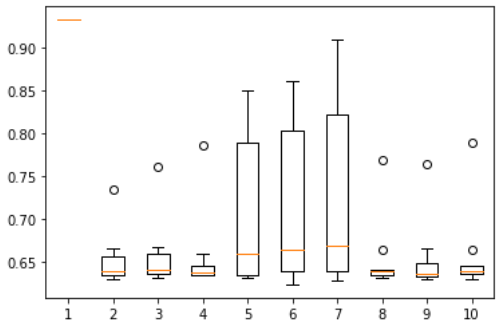

그림 38