3.데이터전처리\_김미래

EDA를 마친 뒤, 분석을 위해 다음과 같은 3가지 작업을 통해 데이터 전처리를 진행하였다.

1. 불필요한 변수 제거
   1. EDA를 위해 만들었던 파생변수 중 타겟 변수(***‘default’***)에 유의미한 차이를 보여주는 변수는 기존 변수를 대체하여 사용하고 유의미한 차이를 보이지 않은 변수는 기존 변수를 그대로 사용하기로 하였다. <Table 1>에서 하이라이터로 표시된 변수가 분석에서 사용되었고 그 외의 변수들은 모두 데이터에서 컬럼 삭제 처리를 해주었다.
   2. <Table 1>에서 ***‘BILL\_AMTX\_bin(월별 청구서 구간별 금액)’***은 타겟 변수(***‘default’***)에 대해서 유의미한 차이를 보이지 않았기 때문에 컬럼 삭제 처리를 해주었다. 이때, ***‘BILL\_AMTX’***에서 음수값은 Overpaid(초과납부)를 의미하는데, 수치상 의미를 잘 나타내기 위해서 ***‘BILL\_AMTX\_new’***를 사용하고 ***‘BILL\_AMTX(월별 청구서 금액)’***는 컬럼 삭제 처리를 해주었다.

|  |  |
| --- | --- |
| **기존 변수** | **파생 변수** |
| *Age* | *AgeBin* |
| *LIMIT\_BAL* | *LimitBin* |
| *PAY\_X* | *PAY\_X\_NEW* |
| *BILL\_AMTX* | *BILL\_AMTX\_bin* |
| *BILL\_AMTX\_new* |
| *PAY\_AMTX* | *PAY\_AMTX\_bin* |

Table 1

1. 연속형 변수 표준화(scaling)
   1. 분석에서 사용되는 변수 중 연속형 변수 ‘BILL\_AMTX\_new’는 <Figure 1>에서 모두 오른쪽으로 치우친 분포(right-skewness)를 띄고 있어 표준화 작업으로 평균 0, 표준편차 1로 조정을 해주었다.
   2. ‘Age’ 또한 오른쪽으로 치우친 분포(right-skewness)를 따르고 있지만, 표준화를 했을때 z-score[[1]](#footnote-1)가 절댓값 2를 넘어가는 수치는 이상치(outlier)로 인식하고 삭제하기 때문에 다양한 연령대의 데이터 정보 손실을 막기 위해 표준화 작업을 해주지 않았다.

(1)

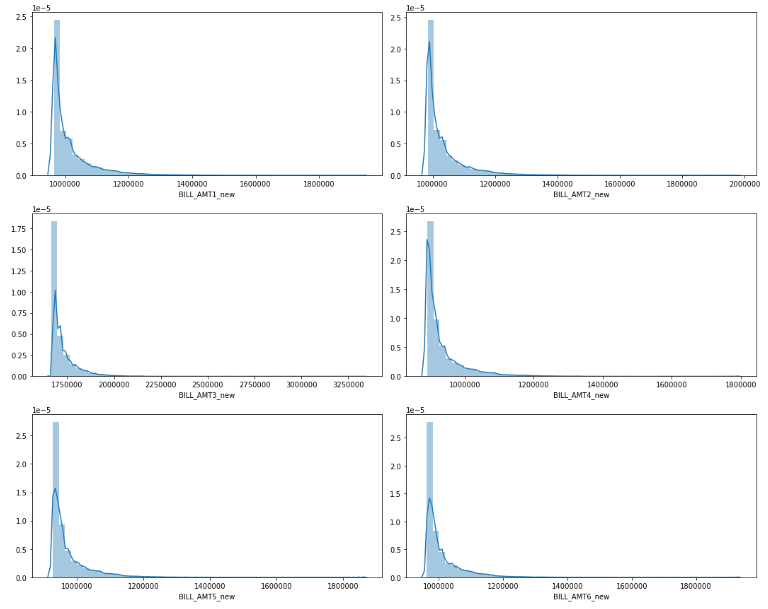


Figure 1

1. 가변수 작업
   1. 분석에서 사용되는 변수 중 범주형 변수(***‘LimitBin’, ‘PAY\_AMTX\_bin’***)는 범주가 3개 이상인 다항 변수(multiclass variable)이기 때문에 가변수(dummy variable) 변환 작업을 통해 기저 범주(base line)를 제외한 변수들에 대해서 이항 변수(binary variable)로 변환하였다.
   2. 범주형 변수 ***‘PAY\_X\_new’***는 가변수로 변환해주기 전에 변수 타입을 정수형(int)에서 문자형(str)로 바꿔주었다. 이후에 동일한 방식으로 가변수 변환 작업을 처리해주었다.
   3. 가변수 변환 작업을 마친 변수들의 기존 변수들(***‘LimitBin’, ‘PAY\_AMTX\_bin’, ‘PAY\_X\_new’***)은 모두 컬럼 삭제 처리를 해주었다.

1. z-score는 표준화(standardization)을 위해서 사용되는 지표이다. 식(1)에서 은 평균, 는 표준 편차를 나타낸다. [↑](#footnote-ref-1)