

데이터 전처리 기본

1. 데이터 전처리 개념

1. 데이터 전처리 개념

데이터 품질 향상

데이터 전처리는 데이터의 품질을 높이는 데 필수적이다. 잘못된 값이나 결측치, 이상치 등을 수정하거나 제거하면 데이터의 정확성이 올라가고, 정확한 데이터를 사용해야만 믿을 수 있는 분석 결과를 얻을 수 있다.

분석 결과의 신뢰성 확보

전처리가 제대로 되어 있지 않으면 분석 결과가 왜곡될 수 있다. 예를 들어, 이상치가 포함된 데이터를 그대로 분석하면 평균이나 분산 같은 통계값이 실제보다 크게 나타날 수 있다. 따라서 전처리를 통해 데이터를 정제하면 분석 결과의 신뢰성이 높아질 수 있다.

1. 데이터 전처리 개념

효율적인 데이터 분석

데이터 전처리를 통해 분석에 필요한 형태로 데이터를 변환하면 분석 과정이 더 효율적으로 진행될 수 있다. 예를 들어, 데이터를 표준화하거나 정규화하면 머신러닝 알고리즘이 더 빠르고 정확하게 학습할 수 있다.

데이터 통합 및 일관성 유지

여러 출처에서 수집된 데이터를 통합하는 과정에서 데이터 전처리가 필요하다. 이를 통해 일관된 형식과 단위로 데이터를 맞추면 분석이 더 쉽게 이루어질 수 있다. 일관성 있는 데이터를 사용하면 분석 결과를 비교하고 해석하기도 쉬울 수 있다.

1. 데이터 전처리 개념

데이터 시각화 및 탐색 용이

전처리된 데이터는 시각화나 탐색 과정에서 더 유용하다. 깨끗한 데이터를 사용하면 그래프나 차트가 더 명확하게 나타나고, 패턴이나 이상점을 쉽게 발견할 수 있다. 이는 데이터의 특성을 이해하고, 인사이트를 도출하는 데 큰 도움이 될 수 있다.

예측 모델의 성능 향상

전처리된 데이터는 예측 모델의 성능을 향상시키는 데 기여할 수 있다. 결측치나 이상치를 제거하고 데이터를 표준화하거나 정규화하면 모델이 더 정확하고 안정적으로 학습할 수 있다. 이는 예측 결과의 정확도를 높이는 데 중요한 역할을 할 수 있다.

1. 데이터 전처리 개념

데이터 전처리는 분석의 기본이자 필수 단계라고 할 수 있다. 이 과정을 통해 데이터의 품질을 높이고 분석 결과의 신뢰성을 확보할 수 있다. 따라서 전처리를 소홀히 하지 않고 철저하게 수행하는 것이 중요하다.

2. 데이터 전처리 기법

[1] 데이터 클렌징(Data Cleansing)

데이터 클렌징은 데이터를 깨끗하게 만드는 과정이다. 잘못된 값, 중복된 데이터, 불일치하는 데이터를 찾아내고 수정하거나 제거한다. 예를 들어, 잘못된 날짜 형식이나 잘못된 범주의 값을 찾아서 올바르게 수정해주는 작업을 포함한다. 이 과정을 통해 데이터의 정확성을 높이고 분석의 신뢰성을 확보할 수 있다.

2. 데이터 전처리 기법

1. 결측치 처리(Missing Data Handling)

결측치는 데이터에서 빠진 값을 말한다. 이를 처리하지 않으면 분석 결과가 왜곡될 수 있다. 결측치를 처리하는 방법은 크게 세 가지가 있다:

- 1. 삭제:** 결측치가 포함된 행이나 열을 삭제한다. 데이터의 손실이 크지 않을 때 유용하다.
- 2. 대체:** 결측치를 다른 값으로 채워 넣는다. 평균값, 중앙값, 최빈값 등으로 대체하거나, 더 복잡하게는 예측 모델을 사용해 결측치를 예측하여 채울 수 있다.
- 3. 예측:** 머신러닝 모델을 사용해 결측값을 예측하는 방법도 있다.

2. 데이터 전처리 기법

결측치를 처리해줘. (결측치예제데이터.csv)



데이터셋의 결측치가 다음과 같이 분포되어 있습니다:

- **학생_ID:** 결측치 없음
- **수학:** 3개
- **영어:** 3개
- **과학:** 3개
- **사회:** 3개

결측치를 처리하는 방법에는 여러 가지가 있습니다. 일반적으로 사용하는 방법은 다음과 같습니다:

1. **삭제:** 결측치가 있는 행이나 열을 삭제합니다.
2. **평균값으로 대체:** 결측치가 있는 열의 평균값으로 대체합니다.
3. **중앙값으로 대체:** 결측치가 있는 열의 중앙값으로 대체합니다.
4. **최빈값으로 대체:** 결측치가 있는 열의 최빈값으로 대체합니다.
5. **특정 값으로 대체:** 결측치를 특정 값으로 대체합니다.

어떤 방법을 사용하여 결측치를 처리하고 싶으신가요? 예를 들어, 평균값으로 대체하는 방법을 원하시면 말씀해 주세요. [~]

2. 데이터 전처리 기법

결측치를 0으로 채워줘.



결측치가 0으로 채워진 데이터

	학생_ID	수학	영어	과학
1	1	85.0	0.0	88.0
2	2	0.0	75.0	92.0
3	3	78.0	82.0	0.0
4	4	92.0	90.0	85.0
5	5	0.0	85.0	79.0
6	6	88.0	0.0	94.0

결측치를 0으로 채운 데이터셋을 처리하였습니다. 데이터셋을 확인하실 수 있습니다. 추가로 필요한 작업이 있으면 말씀해 주세요. [\[?\]](#)

🔊 📄 ↺ 🔍 ⭐

2. 데이터 전처리 기법

결측치를 평균값으로 채워줘



결측치가 평균값으로 채워진 데이터



	학생_ID	수학	영어	과학
1	1	85.0	84.0	88.0
2	2	86.14285714285714	75.0	92.0
3	3	78.0	82.0	88.0
4	4	92.0	90.0	85.0
5	5	86.14285714285714	85.0	79.0
6	6	88.0	84.0	94.0

결측치를 평균값으로 채운 데이터셋을 처리하였습니다. 데이터셋을 확인하실 수 있습니다. 추가적으로 필요한 작업이 있으면 말씀해 주세요. [-]



2. 데이터 전처리 기법

결측치를 머신러닝으로 예측하여 채워줘.



결측치를 머신러닝 기법을 사용하여 예측하고 채우는 과정을 진행하겠습니다. 이를 위해서 각 열에 대해 회귀 모델을 학습시켜 결측치를 예측하겠습니다. 간단한 회귀 모델인 랜덤 포레스트 회귀 모델을 사용하겠습니다.

먼저, 결측치 예측을 위한 라이브러리를 불러오고, 필요한 전처리 및 모델 학습 과정을 진행하겠습니다.

이를 위해 각 열에 대해 다른 열의 데이터를 사용하여 결측치를 예측하는 모델을 생성하겠습니다.

Step 1: 데이터 전처리

1. 데이터셋에서 각 열의 결측치를 찾아 채울 데이터를 준비합니다.
2. 결측치가 없는 행들을 학습 데이터로 사용하고, 결측치가 있는 행들을 예측 대상으로 설정합니다.

Step 2: 머신러닝 모델 생성 및 예측

1. 각 열에 대해 다른 열들을 사용하여 결측치를 예측하는 랜덤 포레스트 회귀 모델을 학습합니다.
2. 학습된 모델을 사용하여 결측치를 예측하고, 이를 원래 데이터셋에 채워 넣습니다.

이를 통해 결측치를 머신러닝으로 예측하여 채워겠습니다.

2. 데이터 전처리 기법

랜덤 포레스트 회귀 모델은 NaN 값을 처리하지 못하므로, 이를 해결하기 위해 `sklearn.impute` 모듈의 `SimpleImputer`를 사용하여 데이터 전처리를 하겠습니다. 이 과정에서 각 열의 결측치를 임시로 평균값으로 채운 후 모델을 학습하고 예측하겠습니다.

이제 결측치를 머신러닝으로 예측하여 채우는 과정을 다시 진행하겠습니다. [\[↗\]](#)

결측치를 머신러닝으로 예측하여 채운 데이터				
	수학	영어	과학	사회
1	85.0	87.39	88.0	87.24
2	84.23	75.0	92.0	83.0
3	78.0	82.0	91.45	78.0
4	92.0	90.0	85.0	88.0
5	90.16	85.0	79.0	90.0
6	88.0	80.79	94.0	83.44

결측치를 머신러닝으로 예측하여 채운 데이터셋을 처리하였습니다. 데이터셋을 확인하실 수 있습니다. 추가로 필요한 작업이 있으면 말씀해 주세요. [\[↗\]](#)

2. 이상치 처리(Outlier Handling)

이상치는 데이터 분포에서 벗어난 비정상적으로 큰 값이나 작은 값을 말한다. 이상치는 분석 결과를 왜곡할 수 있기 때문에 처리해야 한다.

1. 삭제: 이상치를 포함한 데이터를 삭제한다. 데이터 손실이 크지 않다면 유용하다.

2. 변환: 이상치를 다른 값으로 변환한다. 예를 들어, 상한값이나 하한값을 설정해 이상치를 그 값으로 변환할 수 있다.

2. 데이터 전처리 기법

이상치 처리 예제 데이터 셋 : [credit_card.xlsx](#)

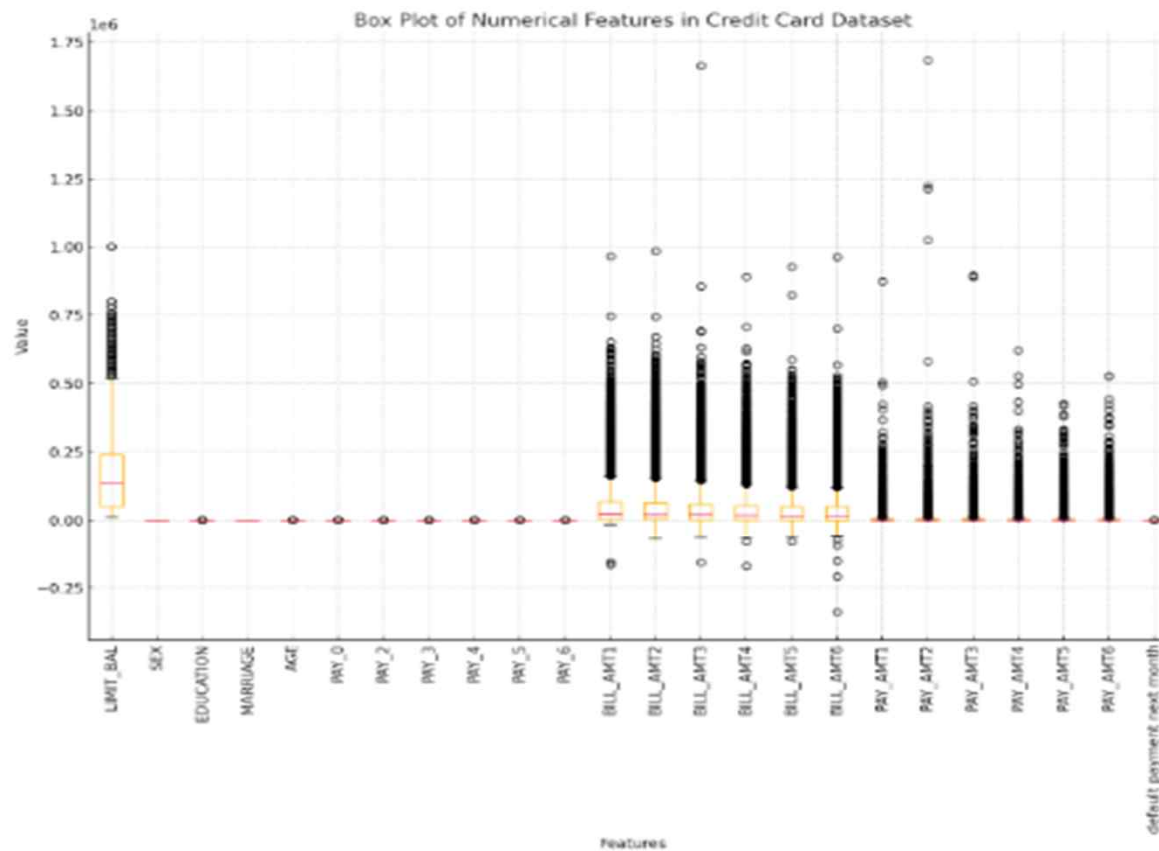
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	LIMIT_B	SEX	EDUCAT	MARRIA	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	default	payment	next mo	
2	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	0	1		
3	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1			
4	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	1000	5000	0		
5	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0			
6	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0			
7	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0			
8	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0			
9	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0			
10	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0			
11	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0			
12	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0			
13	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0			
14	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0			
15	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1			
16	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0			
17	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0			
18	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1			
19	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0			
20	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
21	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
22	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0			
23	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	0	632	316	0	316	0	1		
24	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1			
25	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1			
26	90000	1	1	2	23	0	0	0	0	-1	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0			
27	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0			
28	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	425	259	57	127	-189	0	1000	0	500	0	1000	1			
29	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0			
30	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0			
31	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	0	0			
32	230000	2	1	2	27	-1	-1	-1	-1	-1	-1	16646	17265	13266	15339	14307	36923	17270	13281	15339	14307	37292	0	0			
33	50000	1	2	2	33	2	0	0	0	0	0	30518	29618	22102	22734	23217	23680	1718	1500	1000	1000	1000	716	1			
34	100000	1	1	2	32	0	0	0	0	0	0	93036	84071	82880	80958	78703	75589	3023	3511	3302	3204	3200	2504	0			
35	500000	2	2	1	54	-2	-2	-2	-2	-2	-2	10929	4152	22722	7521	71439	8981	4152	22827	7521	71439	981	51582	0			
36	500000	1	1	1	58	-2	-2	-2	-2	-2	-2	13709	5006	31130	3180	0	5293	5006	31178	3180	0	5293	768	0			
37	160000	1	1	2	30	-1	-1	-2	-2	-2	-1	30265	-131	-527	-923	-1488	-1884	131	396	396	565	792	0	0			
38	280000	1	2	1	40	0	0	0	0	0	0	186503	181328	180422	170410	173901	177413	8026	8060	6300	6400	6400	6737	0			
39	60000	2	2	2	22	0	0	0	0	0	-1	15054	9806	11068	6026	-28335	18660	1500	1518	2043	0	47671	617	0			
40	50000	1	1	2	25	1	-1	-1	-2	-2	-2	0	780	0	0	0	0	780	0	0	0	0	0	1			
41	500000	1	1	2	31	1	1	0	1	0	1	408	9875	4641	8876	17876	8437	8875	0	8876	8880	8525	781	0			

2. 데이터 전처리 기법

박스 플롯 그려줘. (credit_card.xlsx)



Box Plot Of Numerical Features In Credit Card Dataset



2. 데이터 전처리 기법

이상치가 많은 순서대로 5개 컬럼의 이름과 각각의 이상치의 개수를 알려줘.



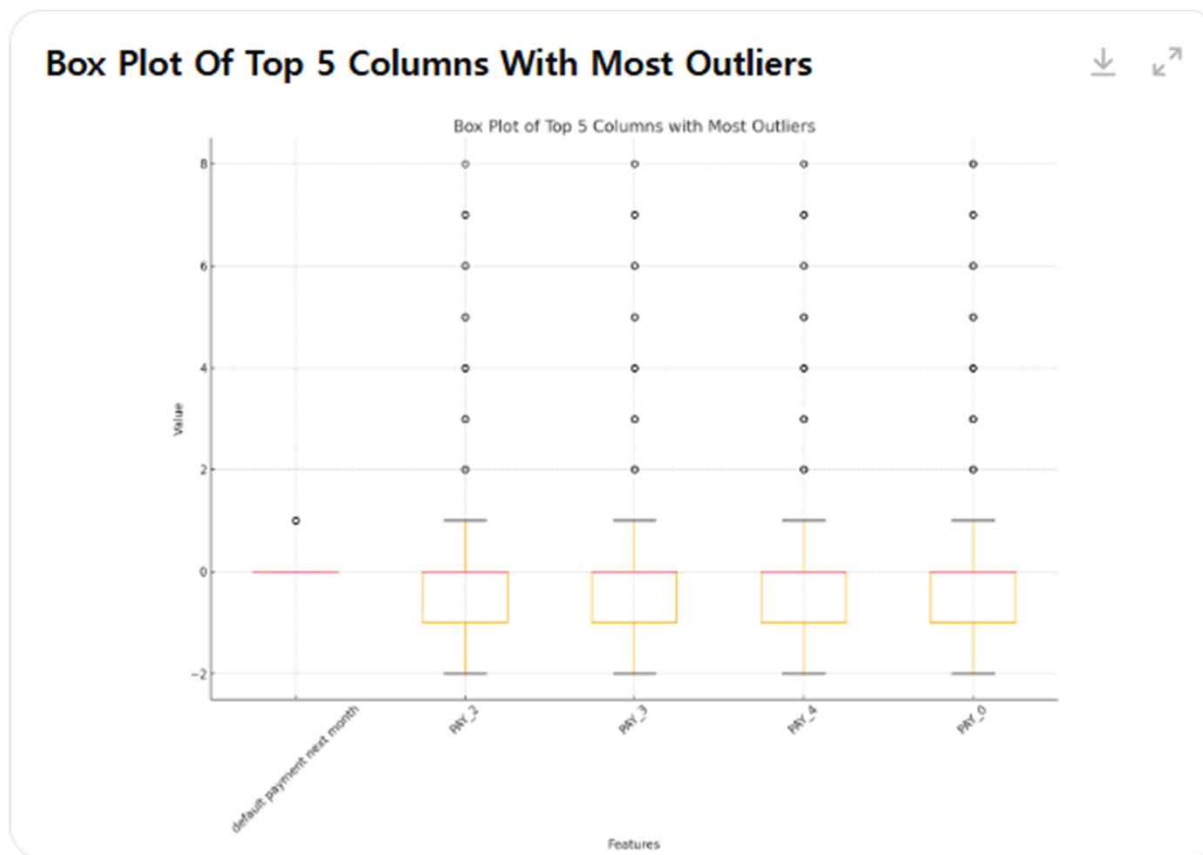
Top 5 Columns With Most Outliers



	Column	Number of Outliers	
1	default payment next month	6636	
2	PAY_2	4410	
3	PAY_3	4209	
4	PAY_4	3508	
5	PAY_0	3130	

2. 데이터 전처리 기법

이 5개의 박스플롯 그려줘.



2. 데이터 전처리 기법

PAY2,PAY_3,PAY_4,PAY_0 컬럼의 이상치를 포함한 데이터를 제거해줘



Cleaned Credit Card Data				
		LIMIT_BAL	SEX	EDUC
22586	29992	10000	1	3
22587	29993	100000	1	1
22588	29995	220000	1	3
22589	29996	150000	1	3
22590	29998	80000	1	3
22591	29999	50000	1	2

2. 데이터 전처리 기법

제거된 데이터의 총 개수와 제거된 행 번호를 모두 알려줘(한글로)



제거된 데이터의 총 갯수는 7409개입니다. 제거된 행 번호는 다음과 같습니다:

[0, 1, 8, 10, 13, 15, 16, 22, 31, 39, ...]

전체 리스트는 너무 길기 때문에 앞의 10개만 표시했습니다. 더 자세한 리스트가 필요하면 말씀해주세요! [\[>\]](#)

2. 데이터 전처리 기법

7409개 모두 알려줘



Removed Row Indices			↓ ↗
	Removed Row Index		▲
1	0		○
2	1		
3	8		
4	10		
5	13		
6	15		

제거된 7409개의 행 번호를 모두 제공했습니다. 이 정보를 더 분석하거나 다른 요청이 있으시면 말씀해 주세요! [\[>\]](#)

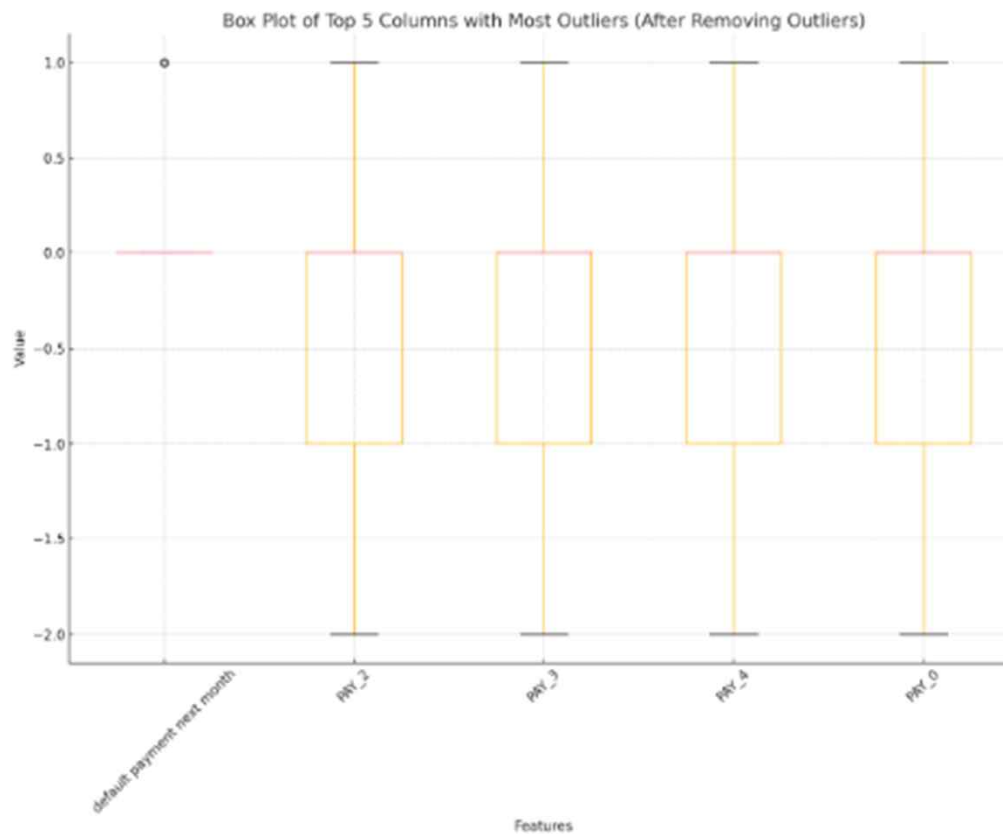
🔊 📄 ↺ 📌 ☆

2. 데이터 전처리 기법

이상치 제거후의 박스 플롯 그려줘.



Box Plot Of Top 5 Columns With Most Outliers (After Remo...



[2] 데이터 변환(Data Transformation)

데이터 변환은 데이터를 분석에 적합한 형태로 바꾸는 과정이다.
대표적인 방법으로는 표준화와 정규화가 있다

1. 표준화(Standardization): 데이터의 평균을 0, 표준편차를 1로 맞추는 과정이다. 데이터의 단위를 맞추고, 다양한 변수의 영향을 균등하게 하기 위해 사용한다.

여기서:

$$z = \frac{x - \mu}{\sigma}$$

- x 는 원본 데이터 값
- μ 는 데이터의 평균
- σ 는 데이터의 표준 편차
- z 는 표준화된 데이터 값

이 과정을 거치면 데이터의 분포는 원본 데이터의 분포를 유지하지만, 모든 특성들이 동일한 스케일을 가지게 되어 모델이 특정 특성에 대해 편향되지 않게 된다.

2. 데이터 전처리 기법

2. 정규화(Normalization): 데이터의 범위를 0과 1 사이로 맞추는 과정이다. 각 변수의 크기 차이를 줄이고, 데이터의 분포를 일정하게 만들기 위해 사용한다.

여기서:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- x 는 원본 데이터 값
- x_{\min} 는 데이터의 최소값
- x_{\max} 는 데이터의 최대값
- x' 는 정규화된 데이터 값

이 방법을 사용하면 데이터의 값이 항상 0과 1 사이에 위치하게 되어, 다양한 크기의 데이터를 동일한 범위로 변환하여 모델이 학습하는 동안 특정 특성이 너무 큰 영향을 미치지 않도록 한다.


2. 데이터 전처리 기법

표준화 예제 데이터 셋 : 표준화예제데이터.csv

	A	B	C
1	feature1	feature2	feature3
2	685	85	876
3	560	204	270
4	630	325	881
5	193	775	297
6	836	965	329
7	764	48	20
8	708	640	608
9	360	132	841
10	10	973	411
11	724	869	451
12	278	181	249
13	755	847	181
14	805	144	324
15	600	661	830
16	71	228	783
17	473	955	865
18	601	792	261
19	397	720	964
20	315	910	750

2. 데이터 전처리 기법

이 데이터를 표준화해서 출력해줘 (표준화예제데이터.csv)



	feature1	feature2	feature3	
1	0.5614196543593855	-1.2999771290071014	1.2531922204420973	
2	0.12455799602445526	-0.9089027847650074	-0.8300756941568793	
3	0.3692005246920162	-0.5112557624684246	1.2703808996054553	
4	-1.1580678328468998	0.9675967997915941	-0.7372568266747467	

표준화된 데이터를 표로 보여드렸습니다. 추가로 필요한 작업이나 분석이 있으시면 말씀해주세요.

[>-]

.1.

2. 데이터 전처리 기법

이 데이터를 정규화해서 출력해줘. (표준화예제데이터.csv)



Normalized Data				
	feature1	feature2	feature3	
1	0.6923076923076923	0.07588357588357589	0.8753799392097265	
2	0.5641025641025641	0.1995841995841996	0.26139817629179335	
3	0.635897435897436	0.32536382536382535	0.8804457953394125	
4	0.1876923076923077	0.7931392931392932	0.2887537993920973	
	0.84717948717948	0.99064449064449	0.32117527862208	

정규화된 데이터를 표로 보여드렸습니다. 추가로 필요한 작업이나 분석이 있으시면 말씀해 주세요.

[>-]



[3] 데이터 통합(Data Integration)

데이터 통합은 여러 출처의 데이터를 하나로 합치는 과정이다. 다양한 출처에서 수집된 데이터를 일관성 있게 결합해 분석에 사용할 수 있도록 해준다 이를 위해 데이터의 형식, 단위, 명명 규칙 등을 맞추는 작업이 필요하다. 예를 들어, 여러 파일에서 수집된 고객 데이터를 하나의 파일로 합치는 작업을 포함한다.

2. 데이터 전처리 기법

[요약]

이와 같은 전처리 기법을 통해 데이터의 품질을 높이고, 분석 결과의 신뢰성을 높일 수 있다. 전처리가 잘 되어 있으면 분석이 훨씬 더 효율적이고 정확하게 이루어질 수 있다.

감사합니다