Loan approval prediction /대출 승인 예측/

24510112 자르갈사이칸 나몽

A Table of Contents.

Preview

- 문제 정의:

대출 승인 여부를 결정할때 은행이 고객의 대출 상환 능력을 정확히 평가하는게 중요하다.
대출 신청자의 정보를 바탕으로 분석하여 대출 불이행 위험을 최소화할 수 있다.
따라서 다양한 요인을 고려하여 데이터 분석을 통해서 대출 승인 여부를 예측할 수 있다.

- Hypothesis:

대출 승인 결정의 중요한 변수로 신청자의 소득 수준, 고용 상태, 자산 가치 등을 사용하여 분석하면 대출 승인 여부를 예측할때 도움이 될 것이다.

사용된 데이터셋

Features	Dataset 1 Size: 4,269 records 2,656 approved loans out of all	Dataset 2 Size: 615 records 422 approved loans out of all	Dataset 3 Size: 5,000 records 480 approved loans out of all
1	loan_id	Loan_ID	ID
2	no_of_dependents	Gender	Age
3	education	Married	Experience
4	self_employed	Dependents	Income
5	income_annum	Education	ZIP Code
6	loan_amount	Self_Employed	Family
7	loan_term	ApplicantIncome	CCAvg
8	cibil_score	CoapplicantIncome	Education
9	residential_assets_value	LoanAmount	Mortgage
10	commercial_assets_value	Loan_Amount_Term	Personal Loan
11	luxury_assets_value	Credit_History	Securities Account
12	bank_asset_value	Property_Area	CD Account
13	loan_status	Loan_Status	Online
14			CreditCard

Data preprocessing

데이터의 공통필드들을 머지:

Features	Dataset 1 Size: 4,269 records 2,656 approved loans out of all	Dataset 2 Size: 615 records 422 approved loans out of all	Dataset 3 Size: 5,000 records 480 approved loans out of all
1	loan_id	Loan_ID	ID
2	no_of_dependents	Gender	Age
3	education	Married	Experience
4	self_employed	Dependents	Income
5	income_annum	Education	ZIP Code
6	loan_amount	Self_Employed	Family
7	loan_term	ApplicantIncome	CCAvg
8	cibil_score	CoapplicantIncome	Education
9	residential_assets_value	LoanAmount	Mortgage
10	commercial_assets_value	Loan_Amount_Term	Personal Loan
11	luxury_assets_value	Credit_History	Securities Account
12	bank_asset_value	Property_Area	CD Account
13	loan_status	Loan_Status	Online
14			CreditCard

Data preprocessing

1	2	3	4
공통 필드들의 • feature name를 동일하게 변환 • 내용을 동일한 format로 변환 /Ex: loan status의 내용이 Approved/Rejected, Y/N라고 dataset마다 다르게 표현되어 있다/	• 3개의 데이터셋을 하나로 결합 • 중요하지 않는 feature들을 제거 /loan_id, age, online etc./	• Duplicated data를 제거 • 범주형 데이터를 수치형으로 변환	최종 데이터셋: • 13개의 feature • 총 9459개의 대출 신청자 정보

>> >> >>

Data preprocessing

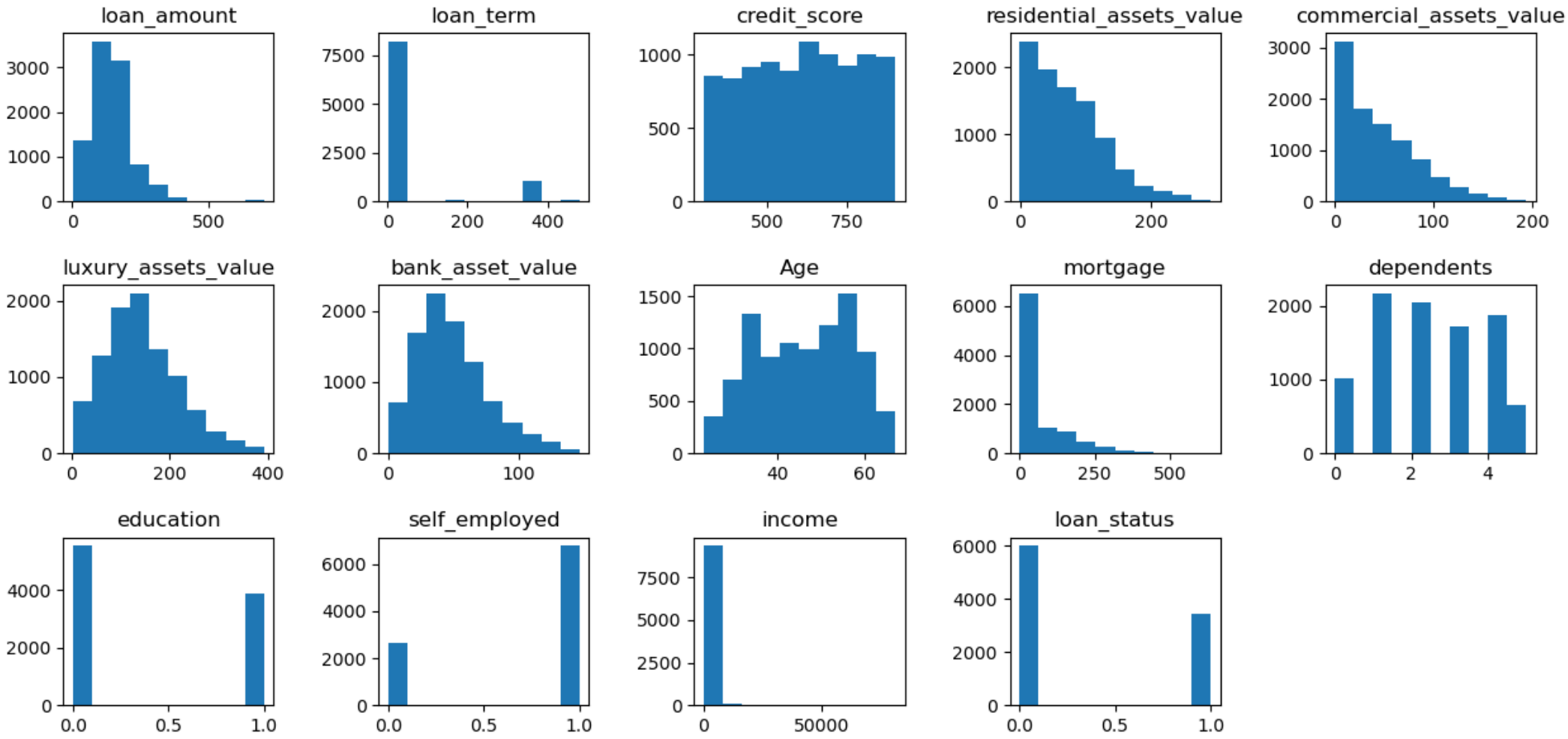
Checking missing value

features	count	datatype
dependents	15	float64
education	0	int64
self_employed	32	float64
income	0	int64
loan_amount	4777	float64
loan_term	4769	float64
credit_score	5364	float64
residential_assets_value	5364	float64
commercial_assets_value	5364	float64
luxury_assets_value	5364	float64
bank_asset_value	5364	float64
loan_status	0	int64
Age	4704	float64
mortgage	4704	float64

- Missing value 수가 적은 dependents랑 self_employed의 missing value를 평균값으로 대체
- Missing value 수가 많은 다른 feature를 다중모델 MICE(Multiple Imputation by Chained Equations) 방법으로 대체

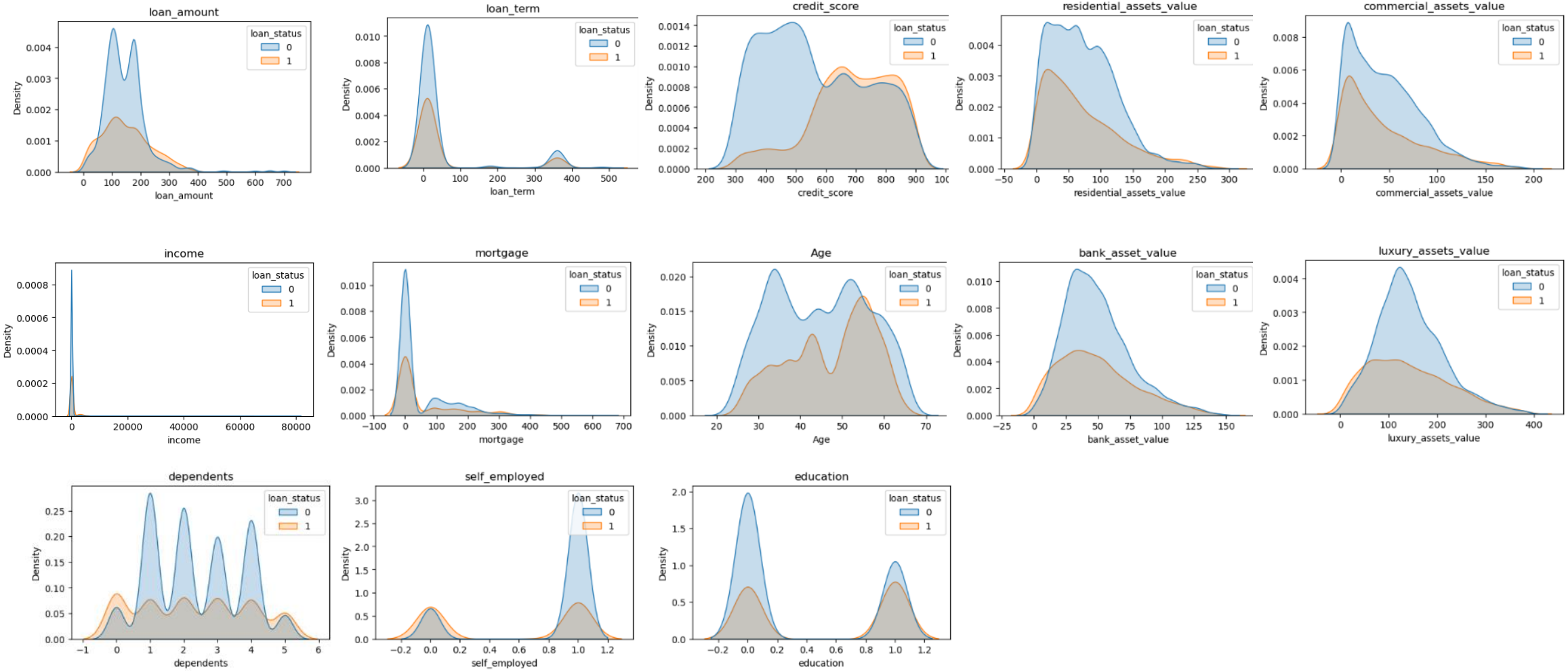
EDA

Histogram



EDA

Relative frequency graph



EDA

description

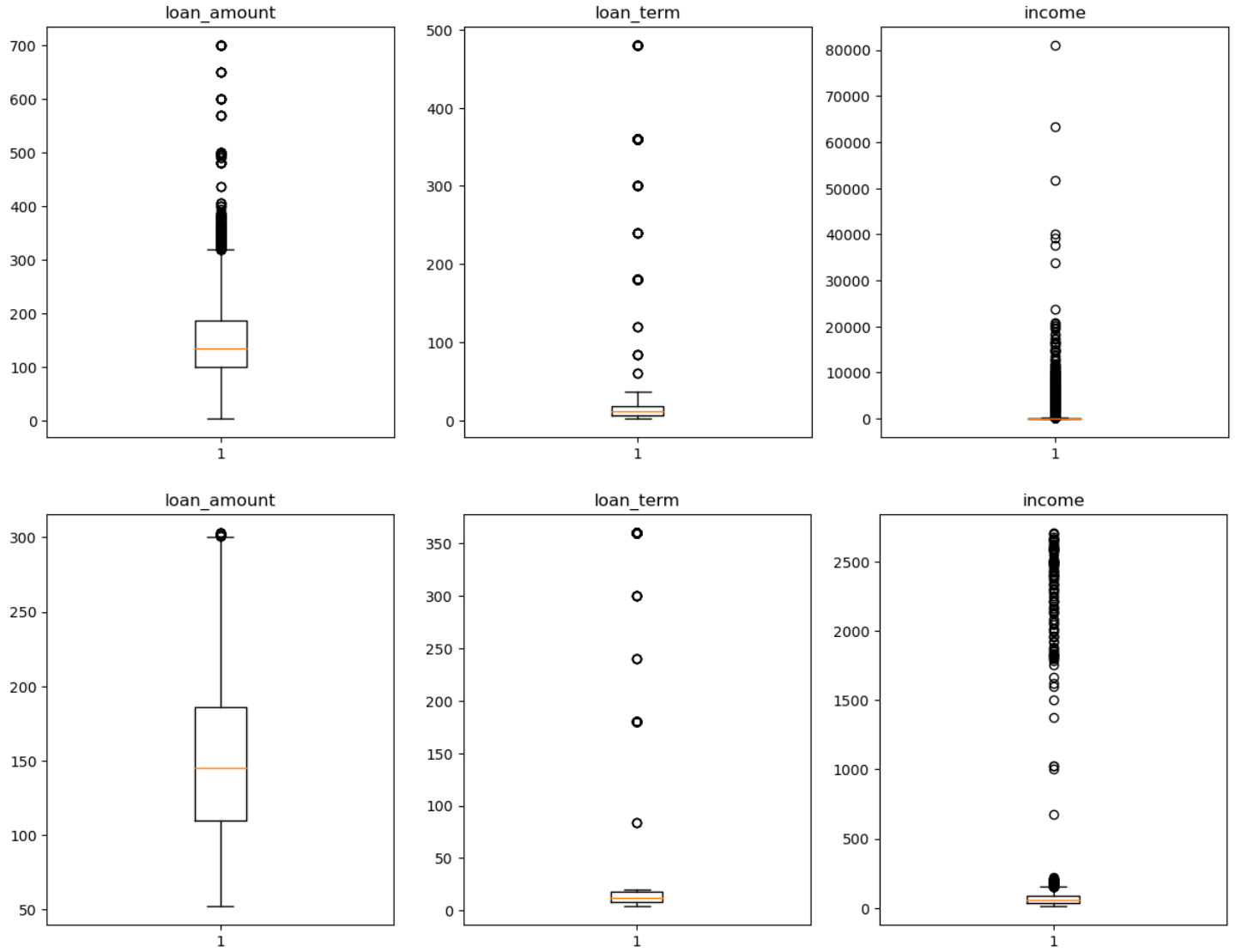
	loan_amount	loan_term	credit_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value
count	9459	9459	9459	9459	9459	9459	9459
mean	146.380907	53.90125806	607.7472249	71.78042076	44.82936886	143.0569828	49.09927054
std	79.5189415	113.6532549	171.0411486	54.36153667	38.49495495	74.9552473	27.55567191
min	3	2	300	-1	0	3	0
25%	96	6	464	28	12	89	29
50%	136	12	613	63	37	133	45
75%	184	18	752	105	68	189	65
max	700	480	900	291	194	392	147

	Age	mortgage	dependents	education	self_employed	income	loan_status
count	9459	9459	9459	9459	9459	9459	9459
mean	45.6461571	55.47372872	2.336509953	0.411037108	0.72027156	407.9096099	0.364414843
std	10.942499	97.8393543	1.461082199	0.492047965	0.44812951	2035.584635	0.481291132
min	23	0	0	0	0	2	0
25%	36	0	1	0	0	34	0
50%	46	0	2	0	1	61	0
75%	55	92	4	1	1	90	1
max	67	635	5	1	1	81000	1

EDA

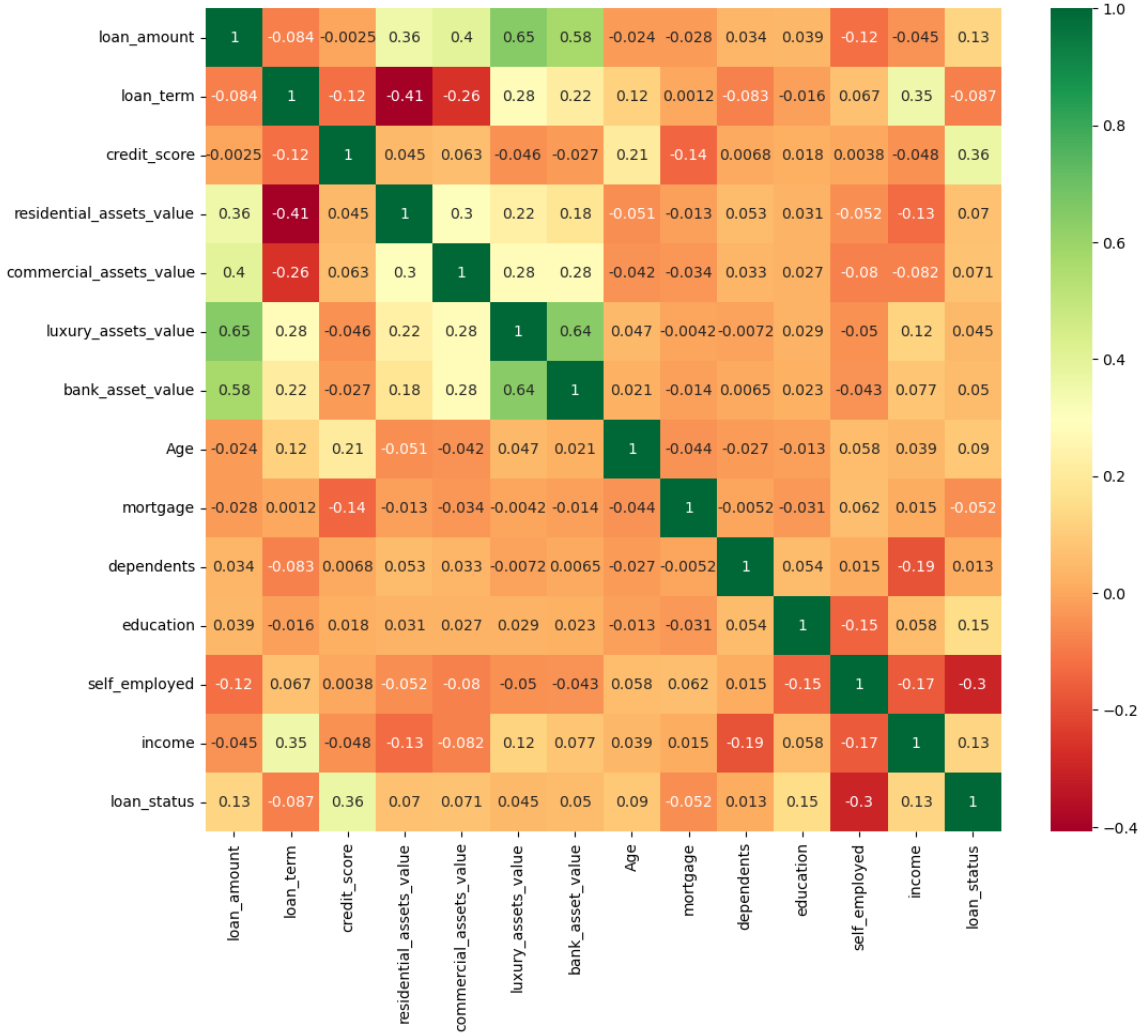
Handling outliers

- IQR 방법을 사용해여 이상치 처리
- 이상치 처리한 데이터셋: 총 6917개의 대출 신청자 정보



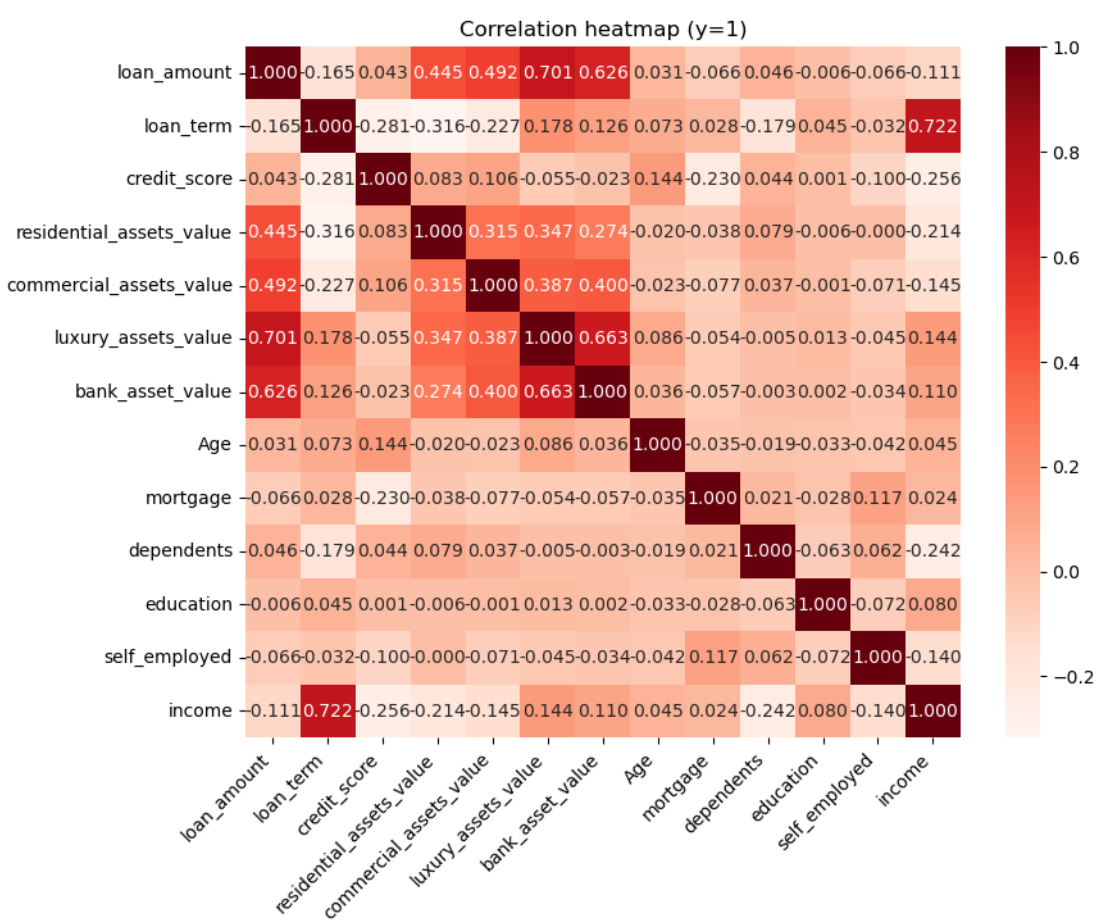
EDA

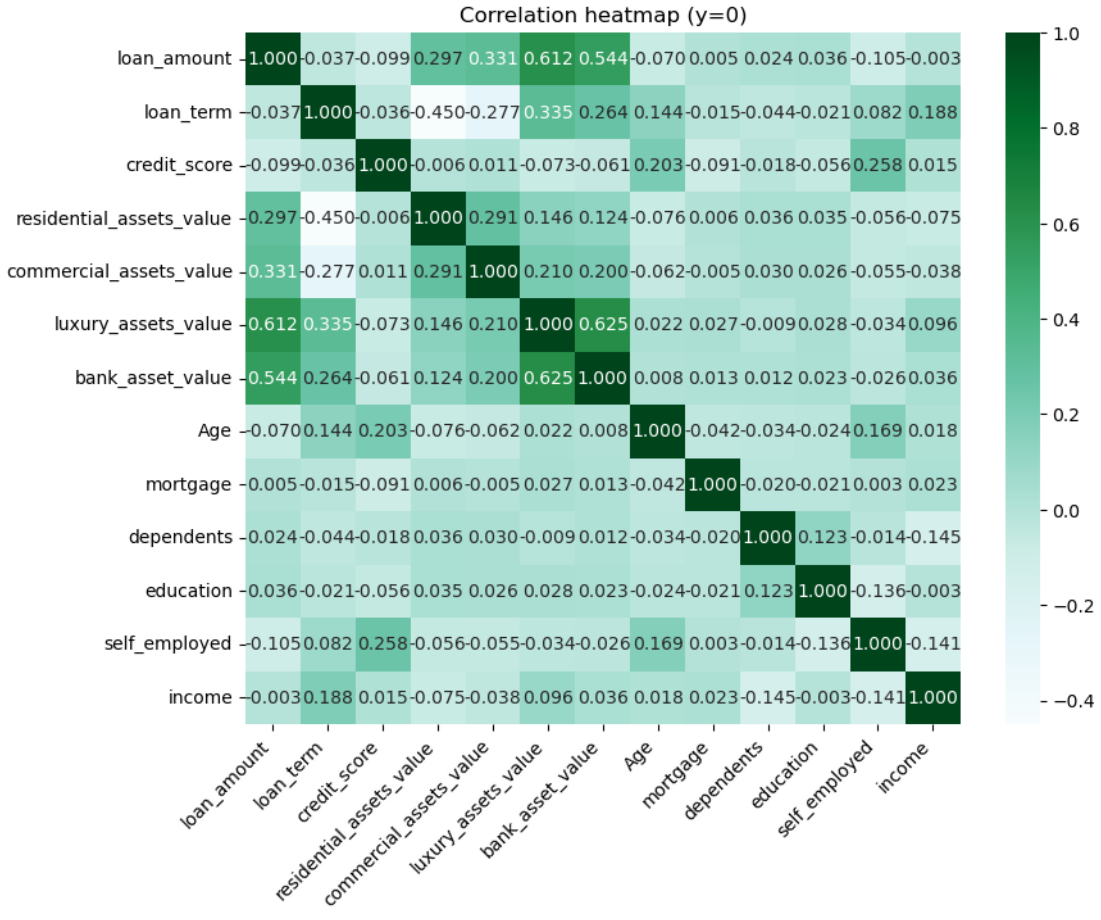
Correlation heatmap (전체 dataset)



EDA

Correlation heatmap (y=1, y=0일 때)





Performance

예측 성능 비교

	accuracy	weighted F1	F1 score	auc score
Decision tree	0.9075	0.9047	0.8388	0.8683
Random forest	0.8880	0.8830	0.7974	0.8368
Logistic regression	0.6835	0.6409	0.3070	0.5592
K-Nearest Neighbors	0.8179	0.8040	0.6675	0.7545

- 성능이 가장 높은 모델: Decision tree

Conclusion

- 3개의 데이터셋을 결합하여 세로운 데이터셋 만들었다.
- 4개의 모델을 활용하여 대출 승인 예측했을때 성능이 가장 높은 모델은 Decision tree.
- 성능이 가장 낮은 모델은 Logistic regression.