



Kaggle ML 프로젝트

대출금 상환 확률 예측

머신러닝 기반의 연구방법론



목차 소개

Table of Contents

01

데이터셋 개요

02

데이터 전처리

03

모델 학습

04

결과



01. 데이터셋 개요

Predicting Loan Payback

» 데이터셋 건수

- Train_set: 593,994(rows)
- Test_set: 254,569(rows)

» 기본 기술통계

Statistic	annual_income	debt_to_income_ratio	credit_score	loan_amount	interest_rate	loan_paid_back
count	593,994	593,994	593,994	593,994	593,994	593,994
mean	48,212.20	0.1207	680.92	15,020.30	12.36	0.7988
std	26,711.94	0.0686	55.42	6,926.53	2.01	0.4009
min	6,002.43	0.0110	395.00	500.09	3.20	0.00
25%	27,934.40	0.0720	646.00	10,279.62	10.99	1.00
50% (median)	46,557.68	0.0960	682.00	15,000.22	12.37	1.00
75%	60,981.32	0.1560	719.00	18,858.58	13.68	1.00
max	393,381.74	0.6270	849.00	48,959.95	20.99	1.00

» 데이터셋 정보

No	Column Name	Dtype	Non-Null Count	설명
1	annual_income	float64	593,994	연소득
2	debt_to_income_ratio	float64	593,994	부채 비율
3	credit_score	int64	593,994	신용 점수
4	loan_amount	float64	593,994	대출 금액
5	interest_rate	float64	593,994	대출 이자율
6	gender	object	593,994	성별
7	marital_status	object	593,994	결혼 상태
8	education_level	object	593,994	학력 수준
9	employment_status	object	593,994	고용 상태
10	loan_purpose	object	593,994	대출 목적
11	grade_subgrade	object	593,994	등급/서브등급
12	loan_paid_back	float64	593,994	상환 여부(타겟)



01. 데이터셋 개요

Predicting Loan Payback

» 데이터셋 예시

	annual_income	debt_to_income_ratio	credit_score	loan_amount	interest_rate	gender	marital_status	education_level	employment_status	loan_purpose	grade_subgrade	loan_paid_back
id												
0	29367.99	0.084	736	2528.42	13.67	Female	Single	High School	Self-employed	Other	C3	1.0
1	22108.02	0.166	636	4593.10	12.92	Male	Married	Master's	Employed	Debt consolidation	D3	0.0
2	49566.20	0.097	694	17005.15	9.76	Male	Single	High School	Employed	Debt consolidation	C5	1.0
3	46858.25	0.065	533	4682.48	16.10	Female	Single	High School	Employed	Debt consolidation	F1	1.0
4	25496.70	0.053	665	12184.43	10.21	Male	Married	High School	Employed	Other	D1	1.0

✓ Target

» Object 컬럼 고유타입

- gender : ['Female' 'Male' 'Other']
- marital_status : ['Single' 'Married' 'Divorced' 'Widowed']
- education_level : ['High School' 'Master's' 'Bachelor's' 'PhD' 'Other']
- employment_status : ['Self-employed' 'Employed' 'Unemployed' 'Retired' 'Student']
- loan_purpose : ['Other' 'Debt consolidation' 'Home' 'Education' 'Vacation' 'Car' 'Medical' 'Business']
- grade_subgrade : ['C3' 'D3' 'C5' 'F1' 'D1' 'D5' 'C2' 'C1' 'F5' 'D4' 'C4' 'D2' 'E5' 'B1' 'B2' 'F4' 'A4' 'E1' 'F2' 'B4' 'E4' 'B3' 'E3' 'B5' 'E2' 'F3' 'A5' 'A3' 'A1' 'A2']



01. 데이터셋 개요

Predicting Loan Payback

✓ Winsorizing 기반
이상치 처리 필요

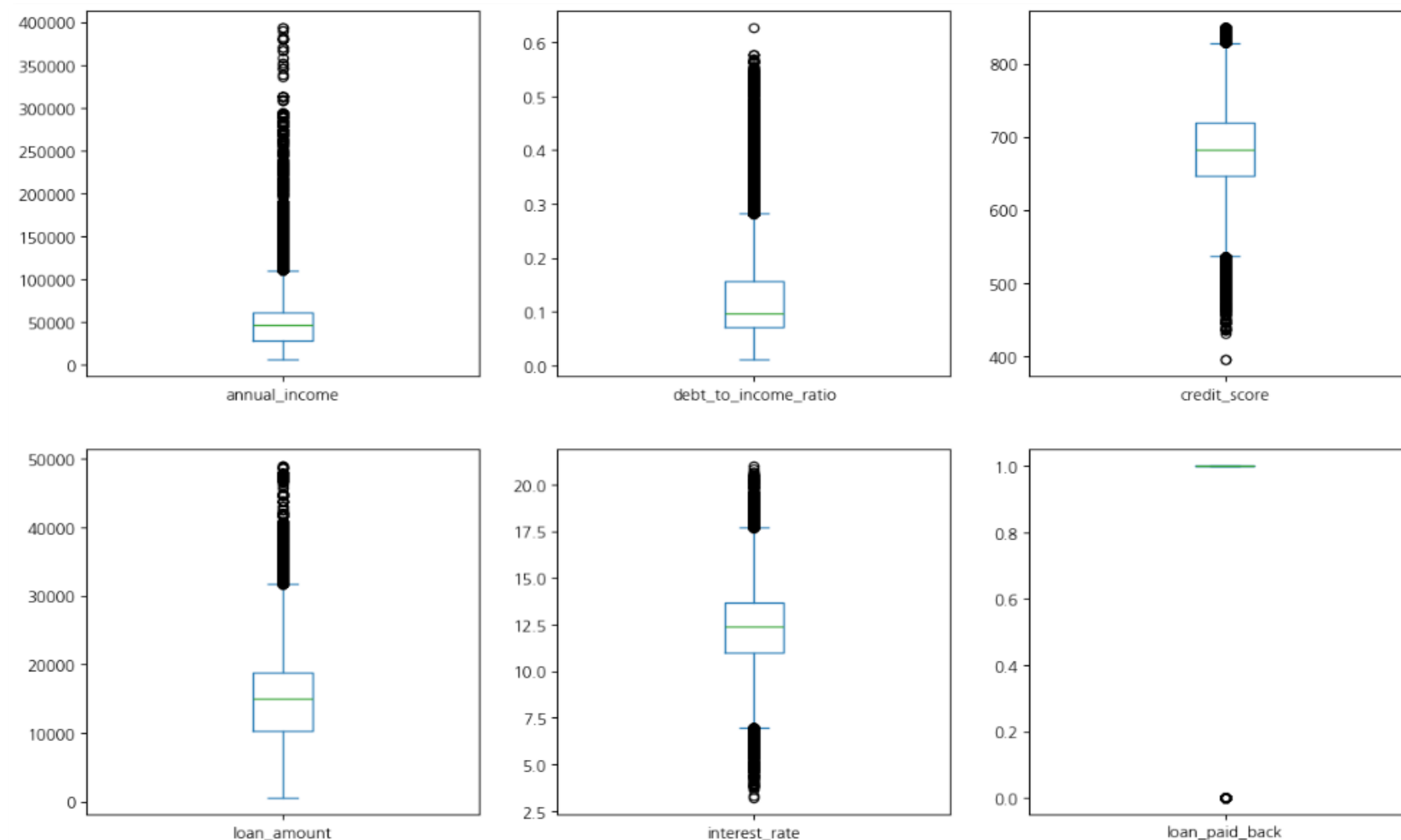
» 결측치 확인

```
=====결측치 확인=====
annual_income      0
debt_to_income_ratio 0
credit_score        0
loan_amount         0
interest_rate       0
gender              0
marital_status      0
education_level     0
employment_status   0
loan_purpose          0
grade_subgrade      0
loan_paid_back      0
dtype: int64
```

» 중복치 확인

```
=====중복값 확인=====
중복된 행의 개수: 0
중복된 행이 없습니다.
```

» 이상치 확인

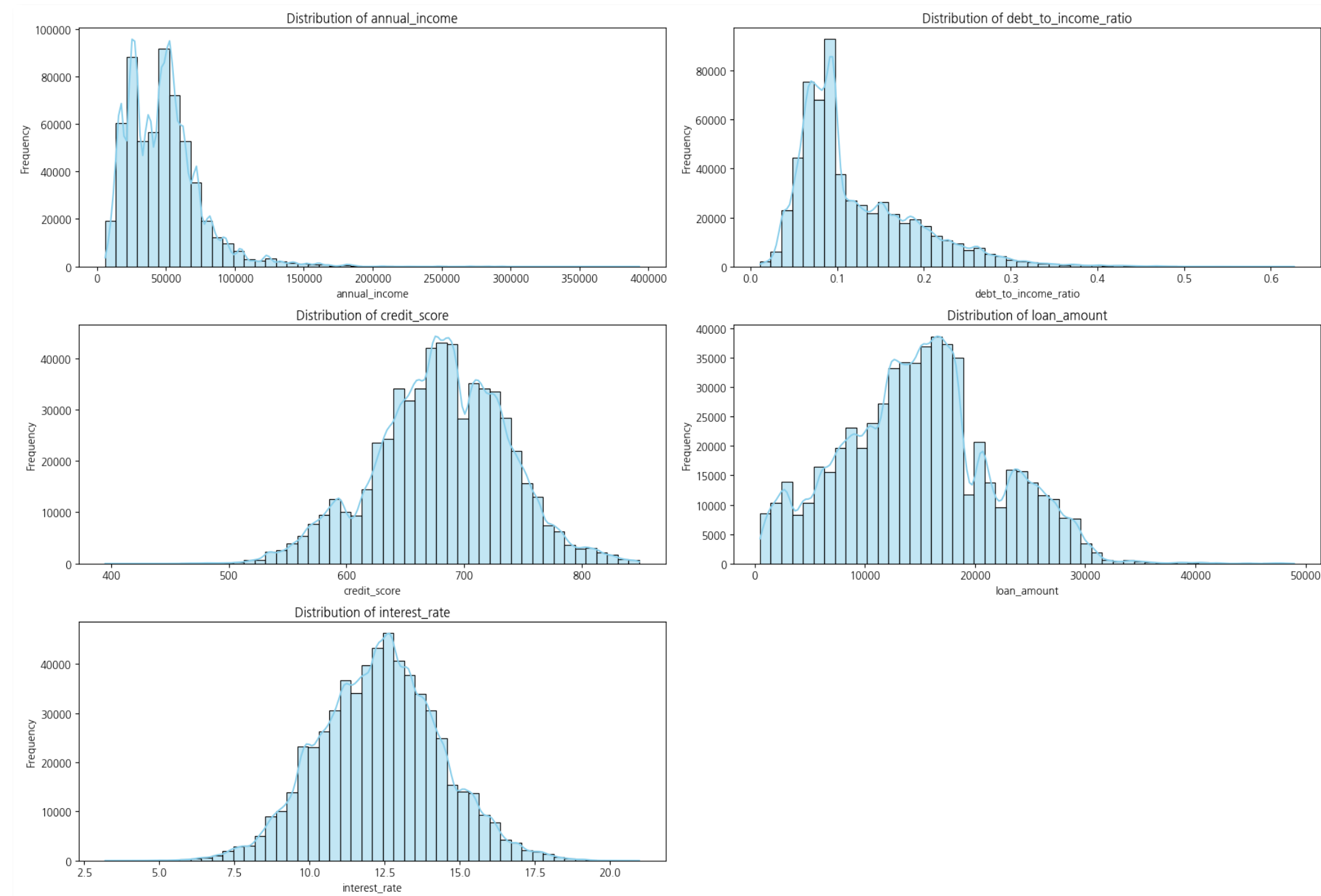




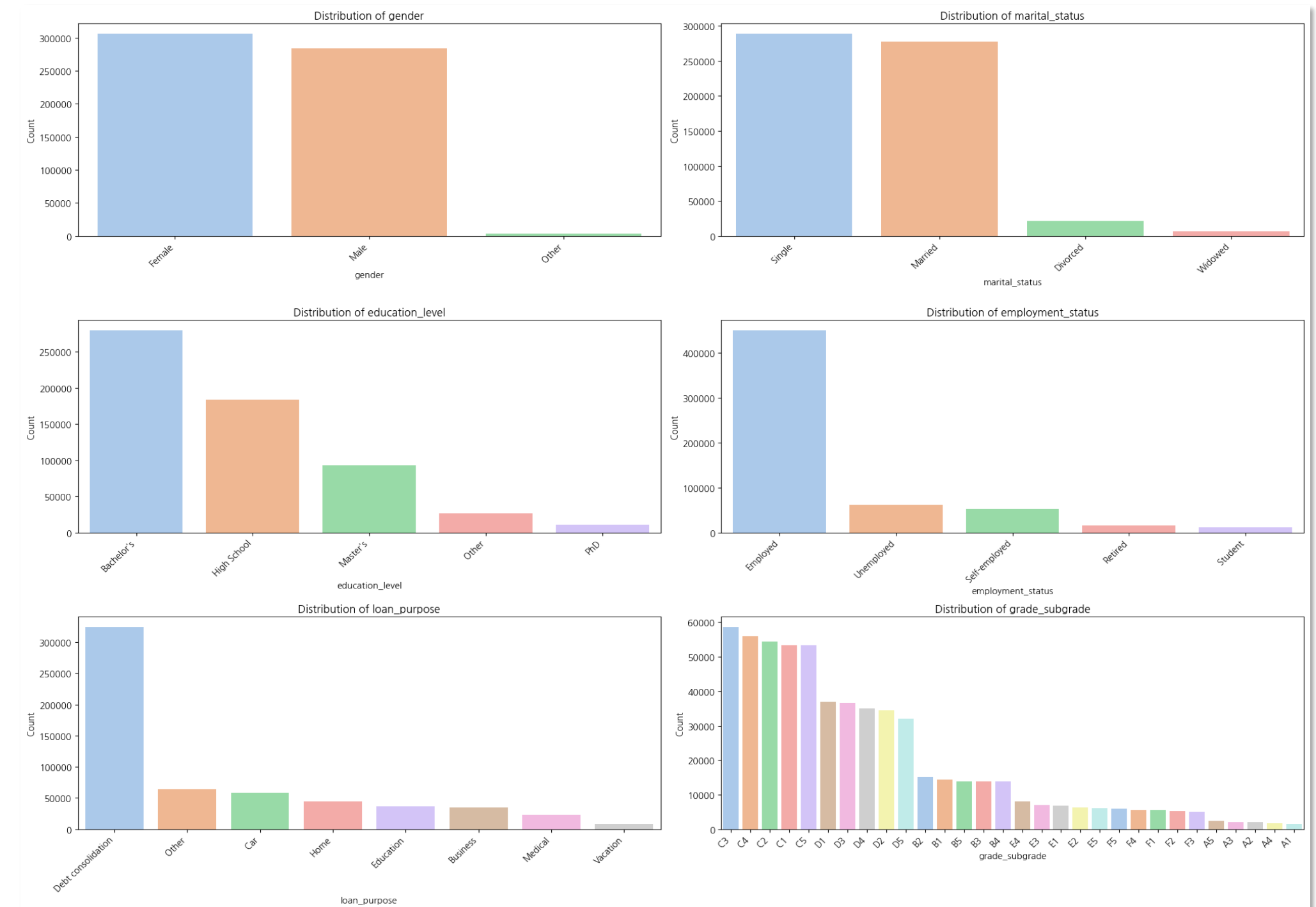
01. 데이터셋 개요

Predicting Loan Payback

» 수치형 컬럼 시각화



» 범주형 컬럼 시각화

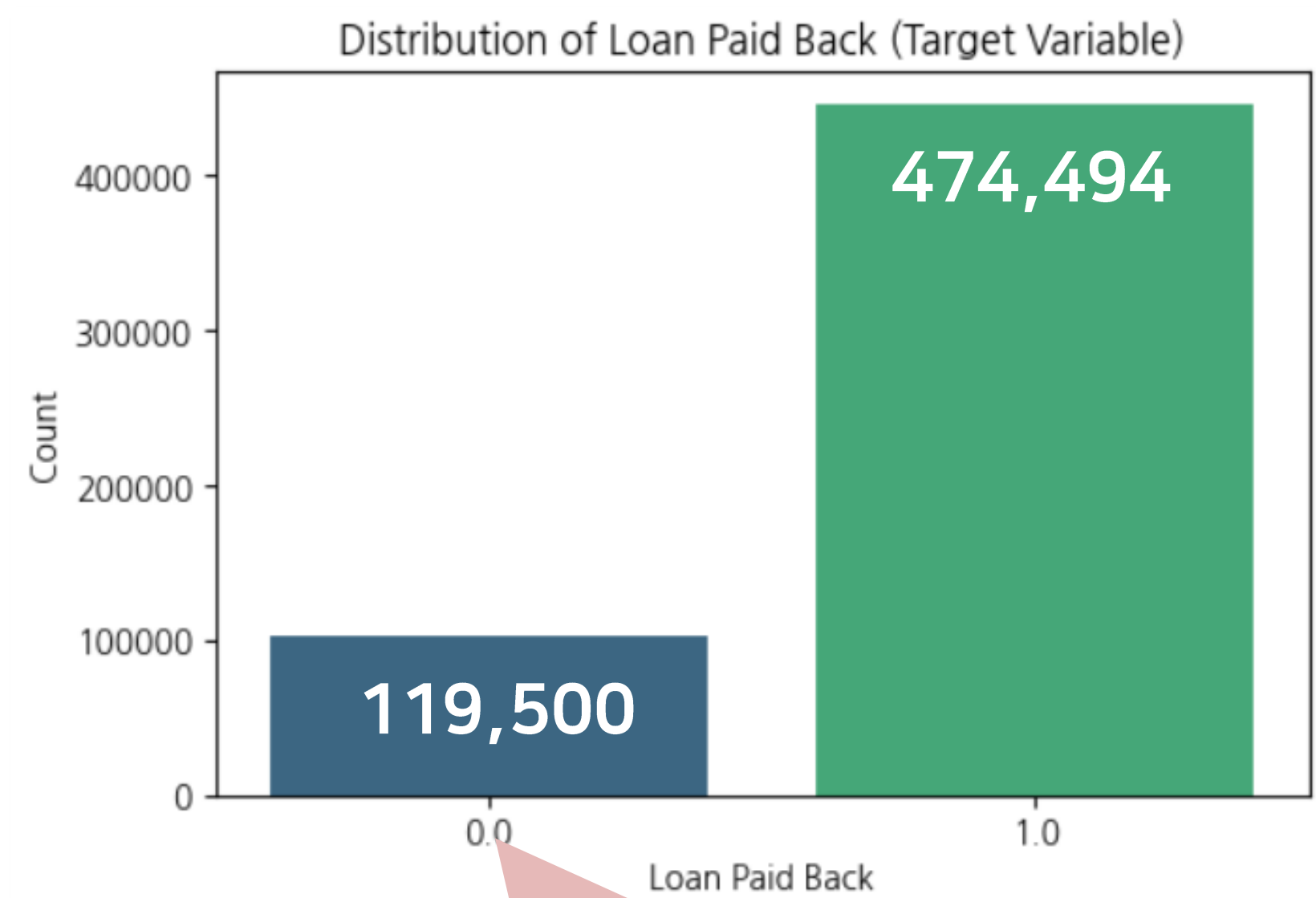




01. 데이터셋 개요

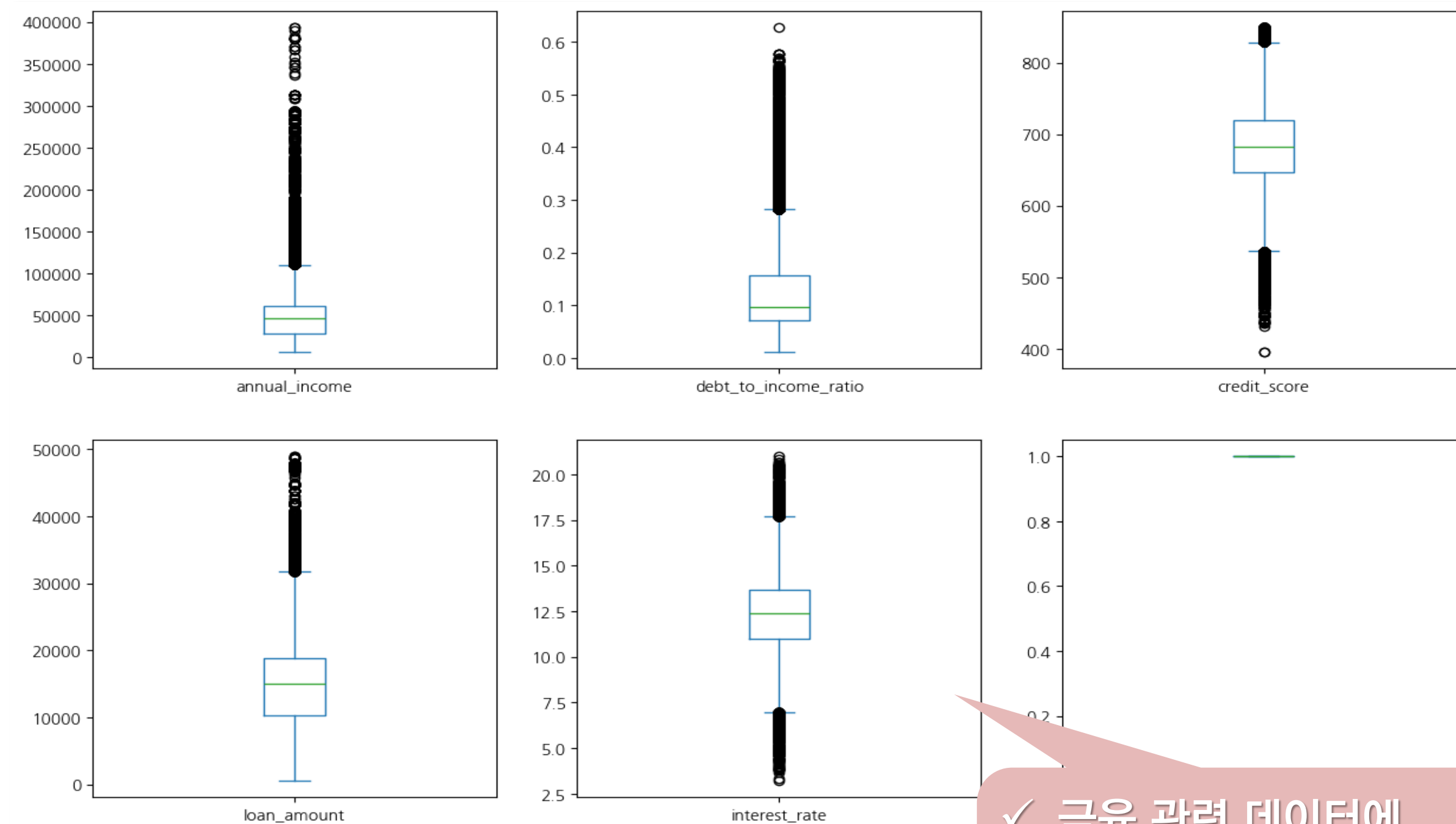
Predicting Loan Payback

» Target 값 시각화



- ✓ 클래스 불균형 처리 필요
- ✓ 모델 하이퍼파라미터에서 설정

» 이상치 확인(Boxplot)



- ✓ 금융 관련 데이터에 따라 winsorized 기법 적용



02. 데이터 전처리

Predicting Loan Payback

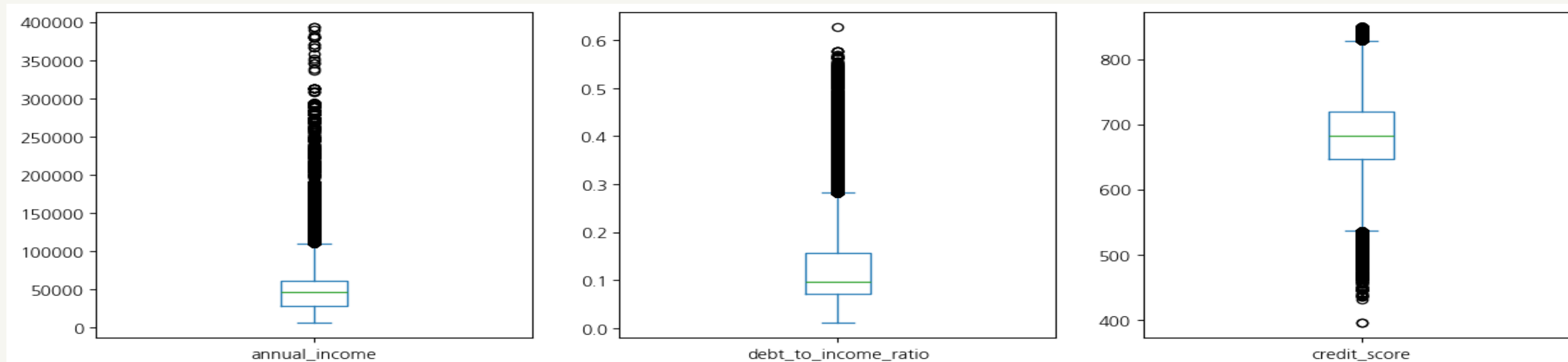
» 이상치 처리

Winsorizing(원저화)

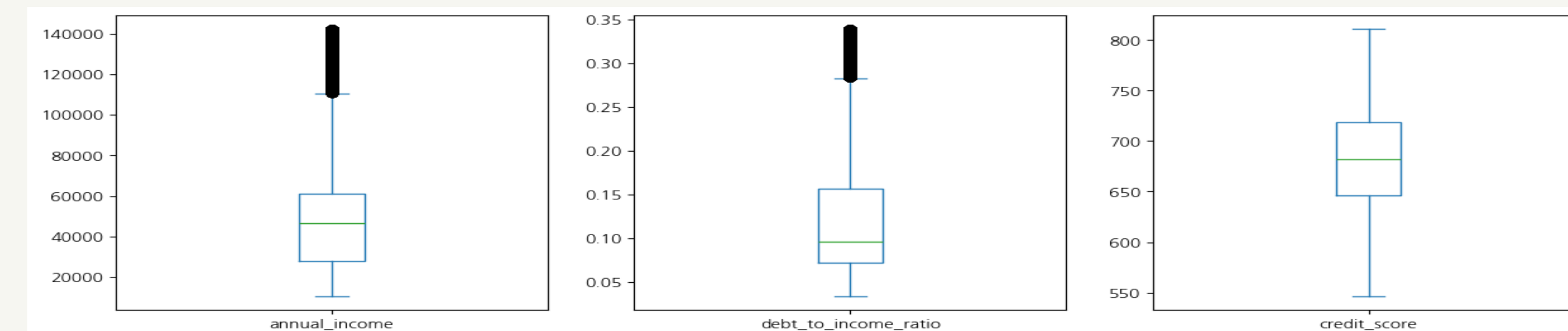
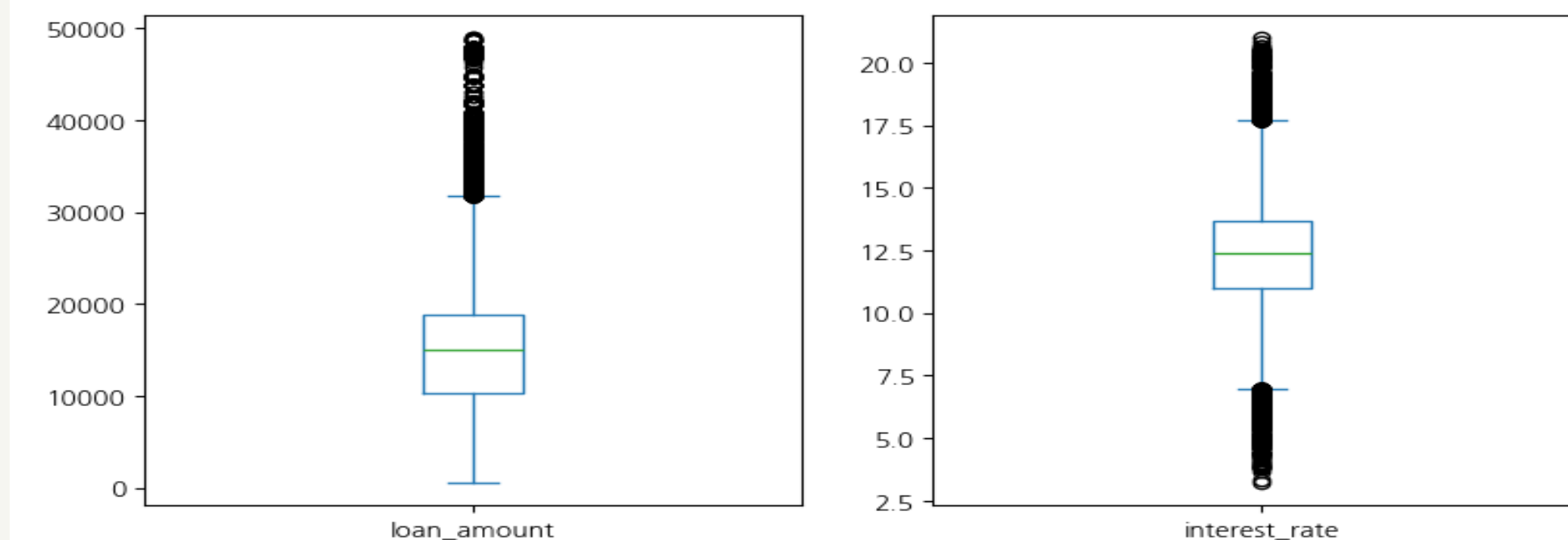
이상치(outlier)를 처리하는 기법으로
이상치를 특정 백분위수(percentile)에
값으로 치환하는 방식



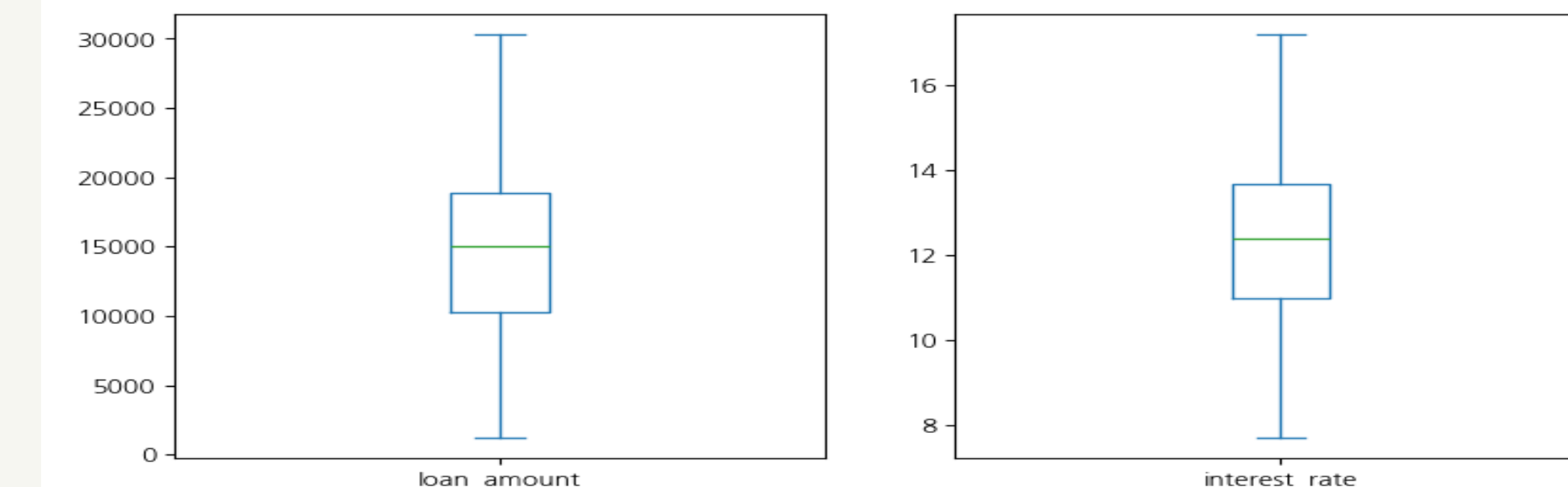
- 하위 1% 이하 값 → 하위 1% 분위수 값으로 치환
- 상위 1% 이상 값 → 상위 99% 분위수 값으로 치환



이상치 처리 전



이상치 처리 전





02. 데이터 전처리 > Feature 추가

Predicting Loan Payback

추가 전

$\text{loan_amount} / (\text{annual_income} + 1)$



추가 후

lti_ratio (Loan to Income Ratio): 연소득 대비 대출 비율

$\text{annual_income} * (1 - \text{debt_to_income_ratio})$



disposable_income: 추정 가처분 소득
(소득에서 세금 등 비소비지출을 제외하고 남은 금액을 추정한 것)

$\text{loan_amount} * (\text{interest_rate} / 100)$



interest_burden: 예상 이자 부담액

['gender_Female', 'gender_Male', 'gender_Other', 'employment_status_Employed', 'employment_status_Retired', 'employment_status_Self-employed', 'employment_status_Student', 'employment_status_Unemployed', 'marital_status_Divorced', 'marital_status_Married', 'marital_status_Single', 'marital_status_Widowed', 'grade_subgrade', 'education_level', 'loan_purpose', 'credit_score', 'debt_to_income_ratio', 'interest_rate', 'annual_income', 'loan_amount', 'lti_ratio', 'disposable_income', 'interest_burden'] → 23개 컬럼



02. 데이터 전처리 > 표준화 및 인코딩

Predicting Loan Payback

컬럼명	변수 유형	적용 전처리 기법
gender	범주형	One-Hot Encoding
employment_status	범주형	One-Hot Encoding
marital_status	범주형	One-Hot Encoding
grade_subgrade	범주형(서열)	Ordinal Encoding
education_level	범주형(서열)	Ordinal Encoding
loan_purpose	범주형(다중)	Target Encoding
credit_score	수치형	StandardScaler
debt_to_income_ratio	수치형	StandardScaler
interest_rate	수치형	StandardScaler
annual_income	수치형	log1p → StandardScaler
loan_amount	수치형	log1p → StandardScaler
lti_ratio (파생)	수치형	log1p → StandardScaler
disposable_income (파생)	수치형	log1p → StandardScaler
interest_burden (파생)	수치형	log1p → StandardScaler



03. 모델 학습

Predicting Loan Payback

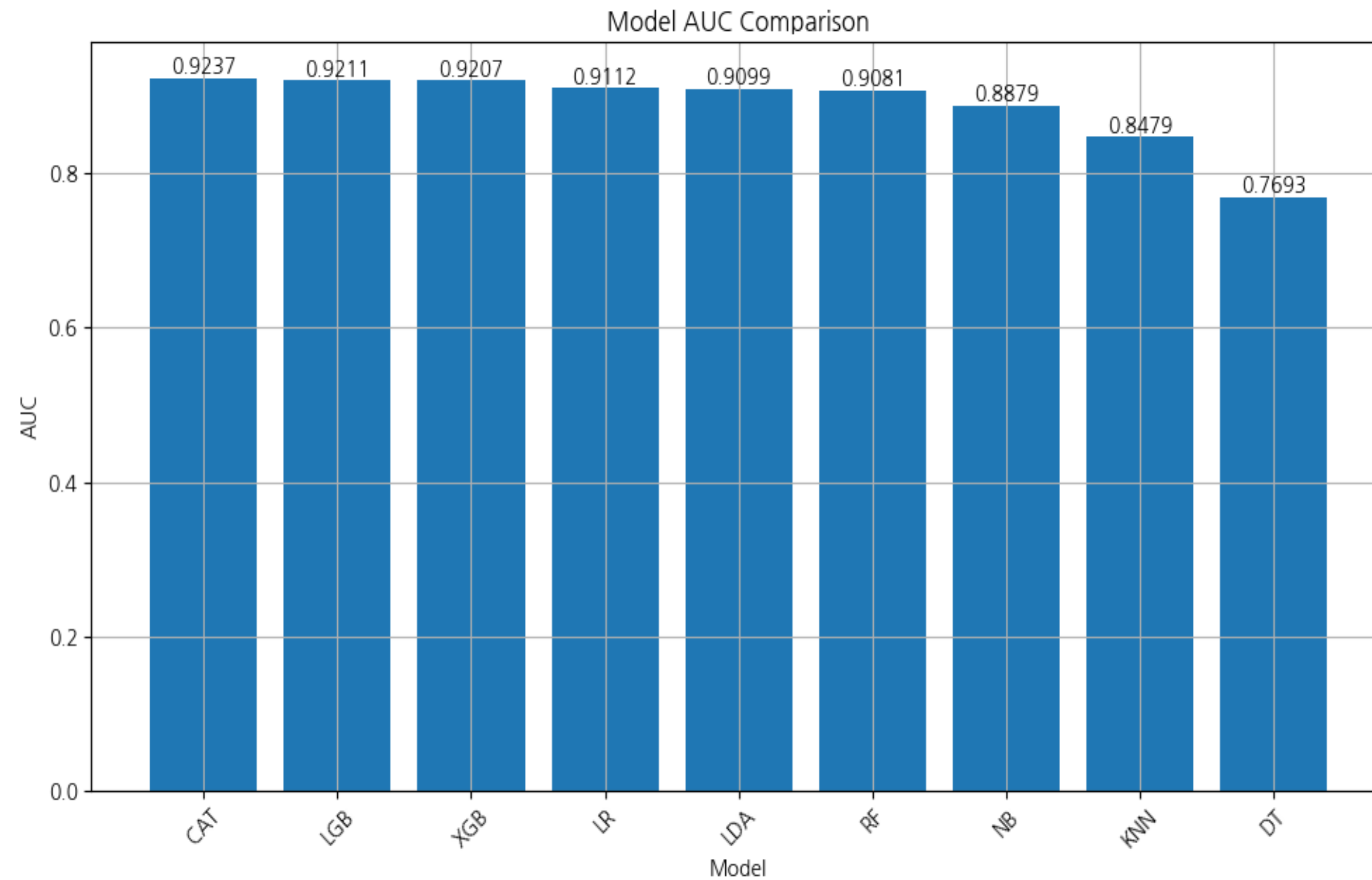
모델명	Accuracy	AUC	학습 소요 시간 (sec)
Logistic Regression	0.8602	0.9112	61.465
Naïve Bayes	0.8914	0.8879	0.316
Linear Discriminant Analysis (LDA)	0.8979	0.9099	1.238
Decision Tree	0.8520	0.7693	22.092
Random Forest	0.9020	0.9081	203.608
KNN	0.8880	0.8479	0.097
XGBoost	0.8667	0.9207	5.516
LGBM	0.8676	0.9211	8.990
CatBoost	0.8729	0.9237	132.374



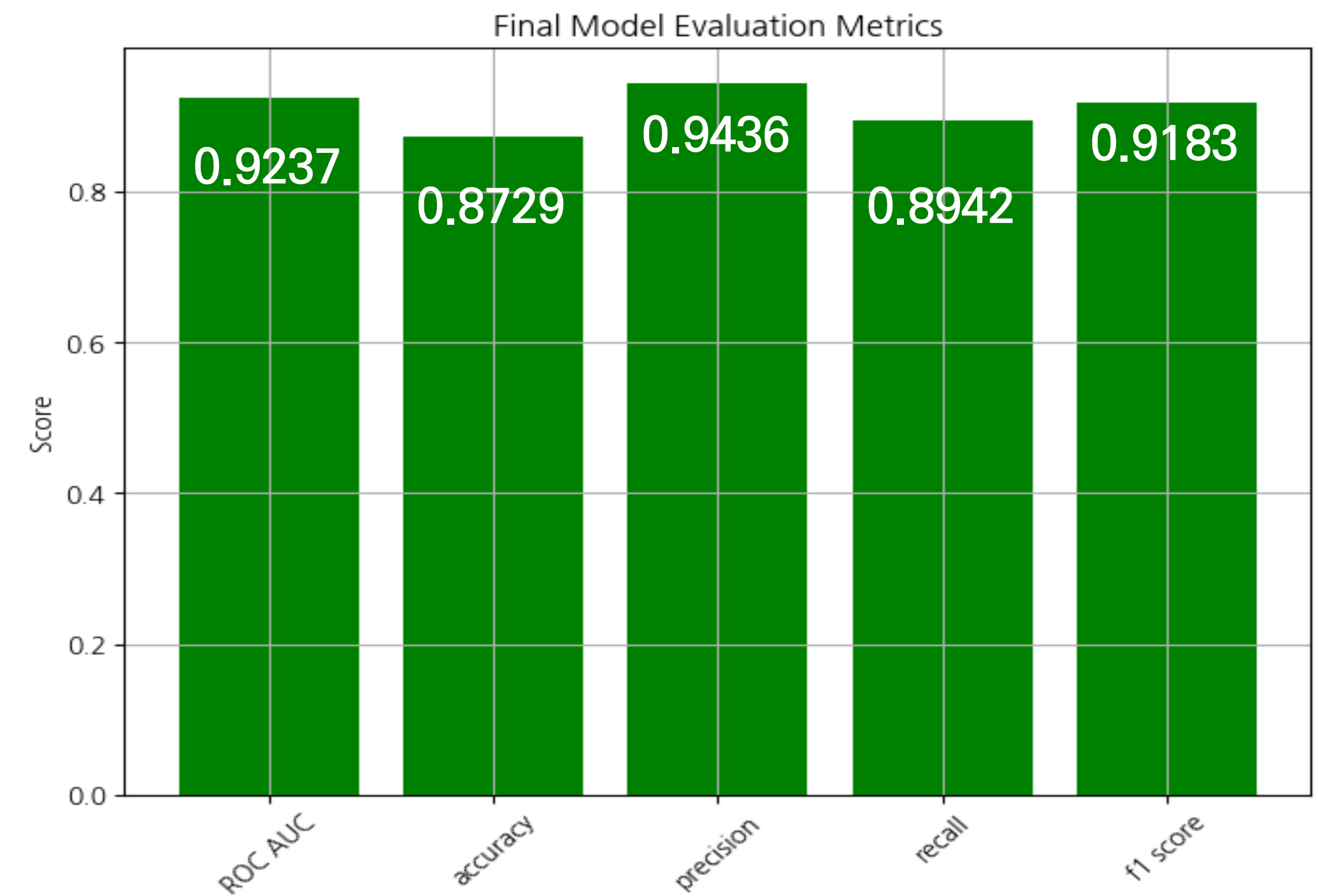
03. 모델 학습

Predicting Loan Payback

» 모델별 AUC



» CatBoost 예측에 대한 여러 지표 결과





03. 모델 학습

Predicting Loan Payback

» 앙상블(Stacking) 기법 적용

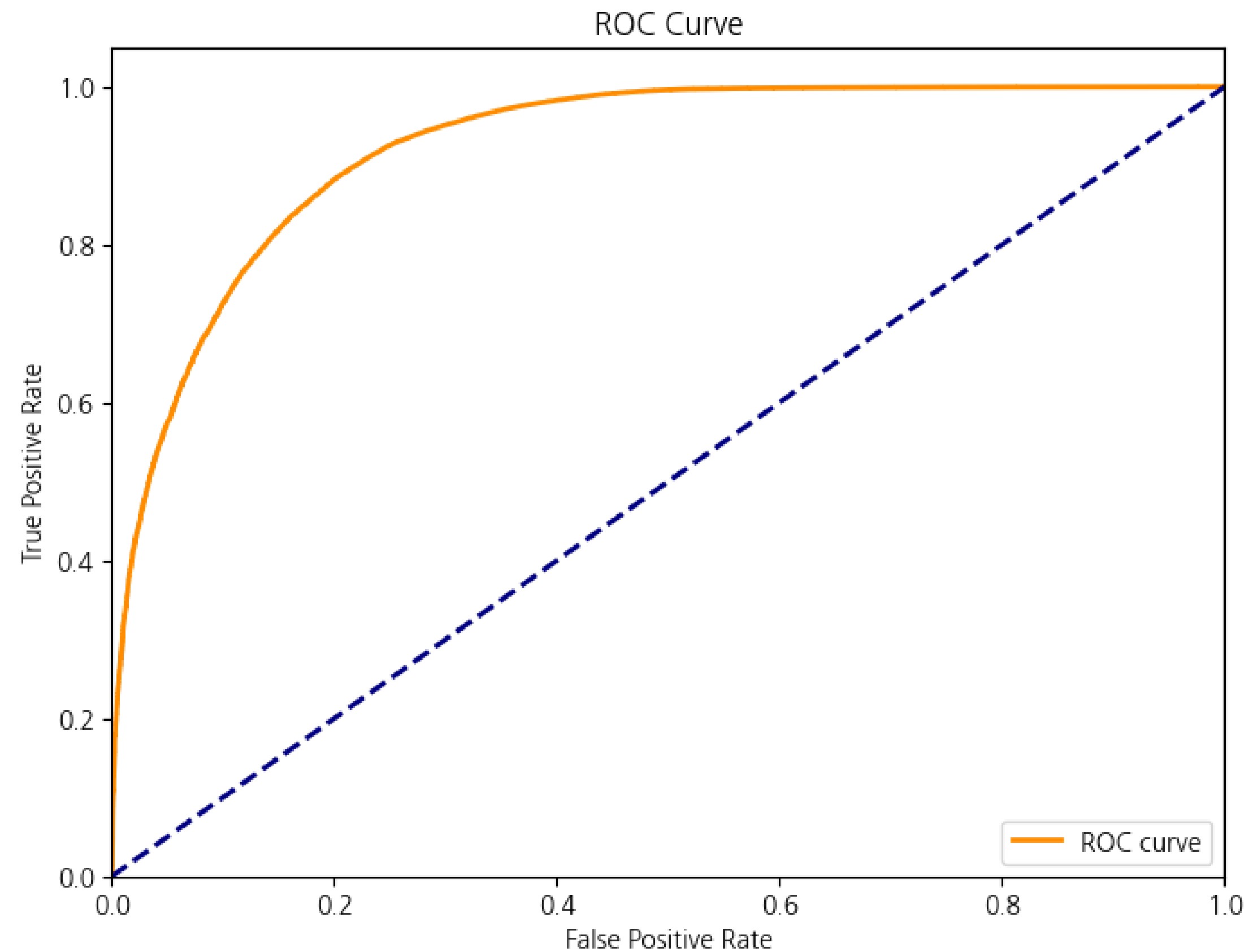
Meta Model: Logistic Regression

기본 모델

- ✓ Random Forest
- ✓ CatBoost
- ✓ LDA
- ✓ Logistic Regression
- ✓ XGBoost

Stacking 앙상블 정확도: 0.9050

Stacking 앙상블 AUC: 0.9229





04. 결과

Predicting Loan Payback



CatBoost

CatBoost, XGBoost, LGBM 등과 같이 부스팅 기반의 모델이 성능이 가장 좋았으며, Decision Tree의 성능이 가장 낮았다.

또한, 앙상블 기법인 Stacking을 적용하여 실험했을 때보다, CatBoost 단일 모델을 적용했을 때가 성능이 더 좋았다.

CatBoost는 대부분이 범주형변수로 이루어진 데이터셋에서 예측 성능이 우수하다. 기본 파라미터로도 강력한 성능을 보여주며, Ordered Boosting, Oblivious Tree, 내부 Regularization이 내장되어 있어 과적합을 방지하고 성능을 높인다.

머신러닝의 고질적인 문제인 Bias-Variance Trade-off가 있지만, CatBoost는 다른 모델보다 Trade-off를 완화시킬 수 있다.

추후, 정교한 전처리 및 하이퍼파라미터 튜닝을 적용하여, 최적의 모델을 구축하는 것이 필요