5조

머신러닝을 활용한 대출자 연체 여부 예측

2020380501 이현지
2020380722 이승주

통계학과머신러닝 PBL

목차

01

데이터 탐색 및 전처리

02

모델링

03

모델 평가 및 결론

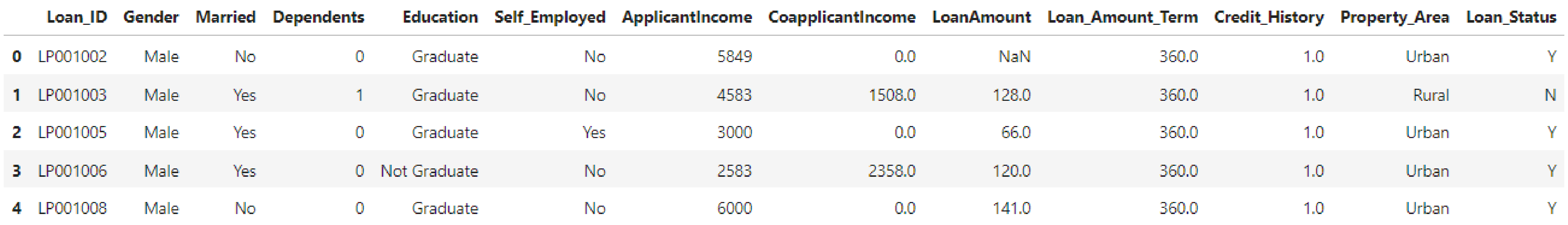
1

데이터 설명

DATA EXPLANATION

대출 연체 데이터 (Loan Data Set)

(614, 13)



	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

변수 ●

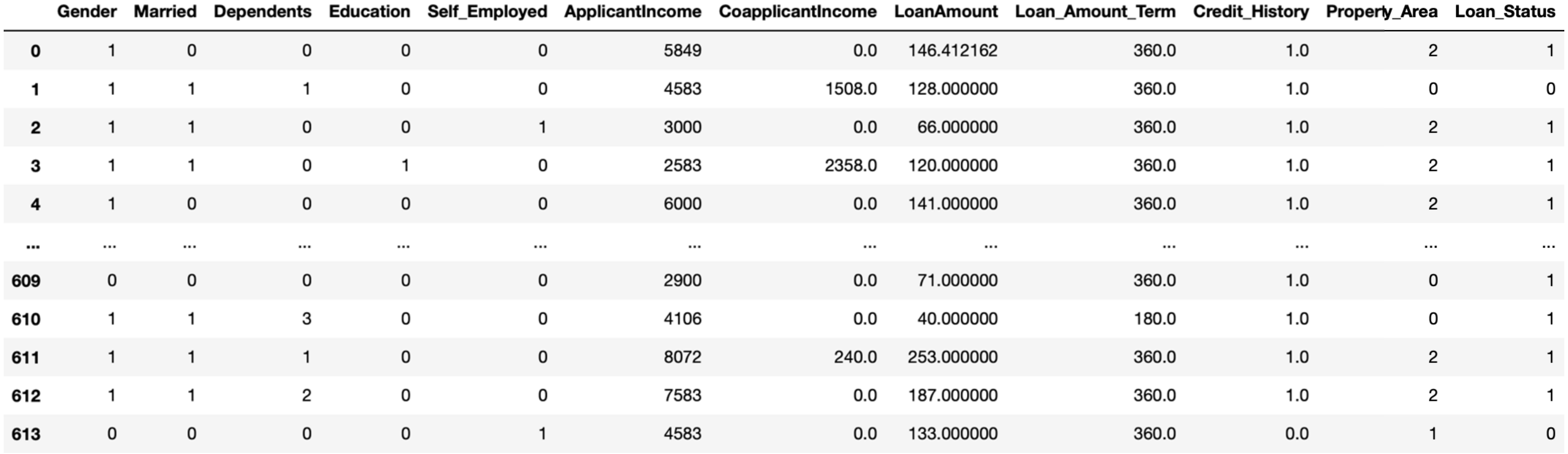
Loan_ID	아이디
Gender	성별
Married	결혼 여부
Dependents	부양가족 수
Education	교육 수준
Self_Employed	자영업 여부
ApplicantIncome	대출신청자 소득
CoapplicantIncome	공동대출신청자 소득
LoanAmount	대출금 총액
Loan_Amount_Term	대출 기간
Credit_History	과거 신용기록
Property_Area	부동산 지역
Loan_Status	연체 상태

데이터 전처리

DATA PREPROCESSING

1. 결측치 처리

2. Label Encoding

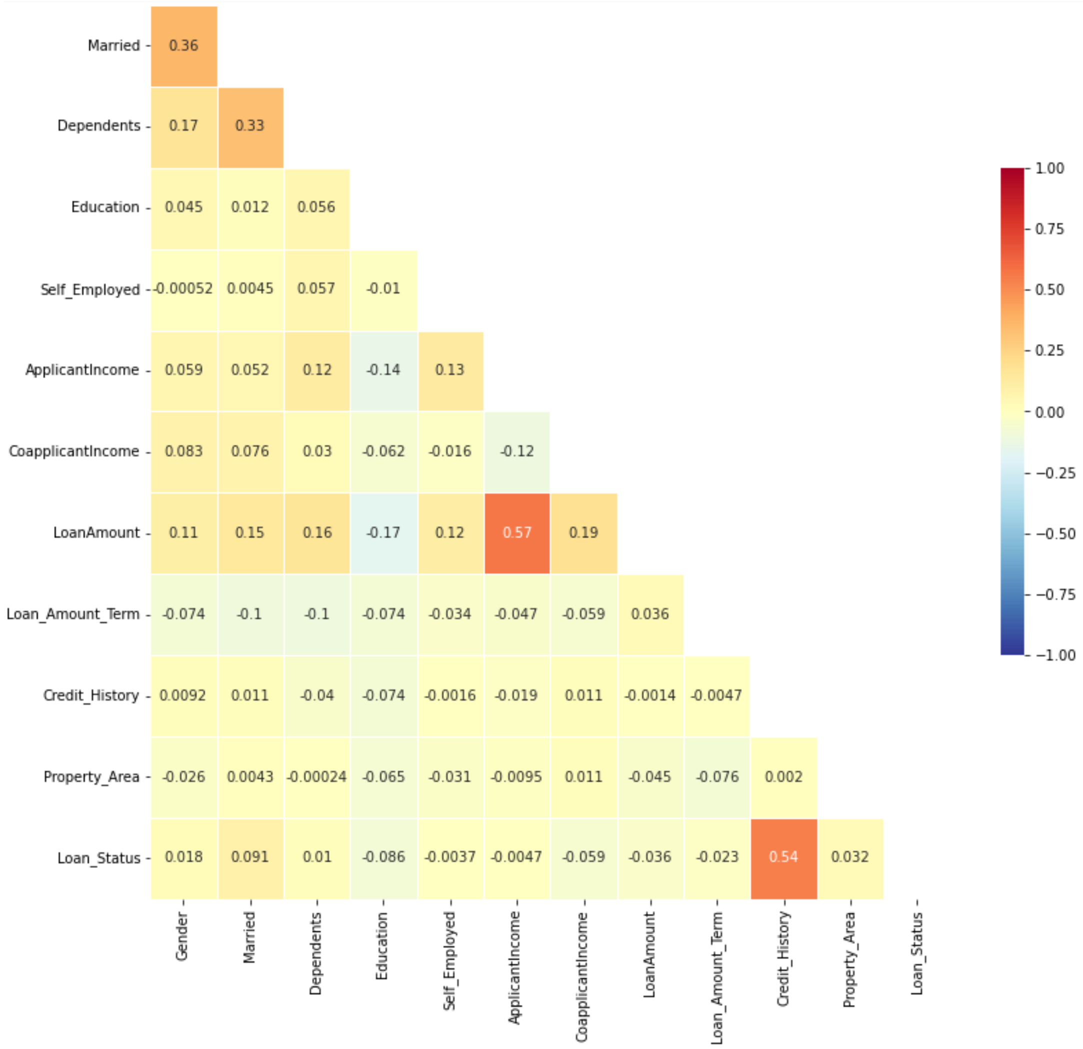


	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	1	0	0	0	0	5849	0.0	146.412162	360.0	1.0	2	1
1	1	1	1	0	0	4583	1508.0	128.000000	360.0	1.0	0	0
2	1	1	0	0	1	3000	0.0	66.000000	360.0	1.0	2	1
3	1	1	0	1	0	2583	2358.0	120.000000	360.0	1.0	2	1
4	1	0	0	0	0	6000	0.0	141.000000	360.0	1.0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...
609	0	0	0	0	0	2900	0.0	71.000000	360.0	1.0	0	1
610	1	1	3	0	0	4106	0.0	40.000000	180.0	1.0	0	1
611	1	1	1	0	0	8072	240.0	253.000000	360.0	1.0	2	1
612	1	1	2	0	0	7583	0.0	187.000000	360.0	1.0	2	1
613	0	0	0	0	1	4583	0.0	133.000000	360.0	0.0	1	0

1

상관분석

CORRELATION ANALYSIS



	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Married	0.36											
Dependents	0.17	0.33										
Education	0.045	0.012	0.056									
Self_Employed	-0.00052	0.0045	0.057	-0.01								
ApplicantIncome	0.059	0.052	0.12	-0.14	0.13							
CoapplicantIncome	0.083	0.076	0.03	-0.062	-0.016	-0.12						
LoanAmount	0.11	0.15	0.16	-0.17	0.12	0.57	0.19					
Loan_Amount_Term	-0.074	-0.1	-0.1	-0.074	-0.034	-0.047	-0.059	0.036				
Credit_History	0.0092	0.011	-0.04	-0.074	-0.0016	-0.019	0.011	-0.0014	-0.0047			
Property_Area	-0.026	0.0043	-0.00024	-0.065	-0.031	-0.0095	0.011	-0.045	-0.076	0.002		
Loan_Status	0.018	0.091	0.01	-0.086	-0.0037	-0.0047	-0.059	-0.036	-0.023	0.54	0.032	

LoanAmount	**ApplicantIncome**	**0.57**
Loan_Status	**Credit_History**	**0.54**
Married	**Gender** **Dependents**	**0.36** **0.33**
Dependents	**Gender**	**0.17**

데이터 모델링

DATA MODELING

Logistic Regression

KNN

Decision Tree

Bayes Classifier

Support Vector Machine

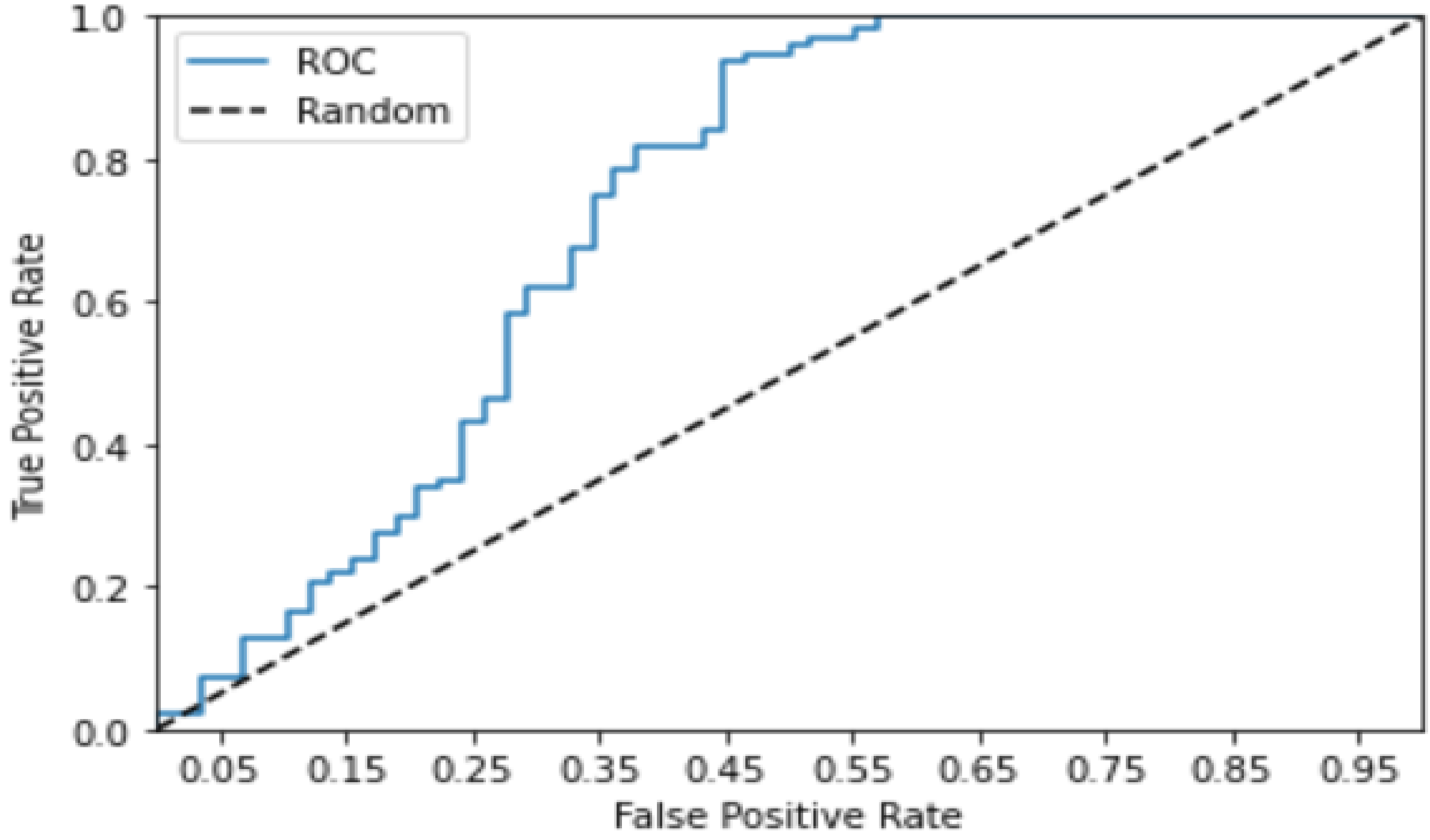
Ensemble

데이터 모델링

DATA MODELING

AUC = 0.715517

cut-off = 0.676



데이터 모델링

DATA MODELING

	Train	Test
Logistic Regression	0.807	0.822
KNN (k=5)	0.821	0.773
Decision Tree (max_depth=5)	0.844	0.789
Bayes Classifier (Linear)	0.808	0.816
Bayes Classifier (Quadratic)	0.808	0.816
Bayes Classifier (Multinomial)	0.808	0.816
Bayes Classifier (Gaussian)	0.809	0.800
Support Vector Machine	0.807	0.811

데이터 모델링

DATA MODELING

	Train	Test
Logistic Regression	**0.807**	**0.822**
KNN (k=5)	0.821	0.773
Decision Tree (max_depth=5)	0.844	0.789
Bayes Classifier (Linear)	**0.809**	**0.816**
Bayes Classifier (Quadratic)	0.809	0.816
Bayes Classifier (Multinomial)	0.809	0.816
Bayes Classifier (Gaussian)	0.809	0.800
Support Vector Machine	**0.807**	**0.811**

Voting Classifier

	Train	Test
Logistic Regression	0.807	0.822
Bayes Classifier (Linear)	0.809	0.816
Support Vector Machine	0.807	0.811

	Train	Test
Voting Classifier (hard)	0.809	0.816
Voting Classifier (soft)	0.809	0.816

앙상블 모형

ENSEMBLE MODEL

Random Forest

Number of Trees = **181**
Maximum Depth of Individual Tree = **6**

AdaBoost

Number of Trees = **3**
Maximum Depth of Individual Tree = **2**

Gradient Boosting

Number of Trees = **5**
Maximum Depth of Individual Tree = **4**

XGBoost

Number of Trees = **2**
Maximum Depth of Individual Tree = **3**

LightGBM

Number of Trees = **30**
Maximum Depth of Individual Tree = **4**

CatBoost

앙상블 모형

ENSEMBLE MODEL

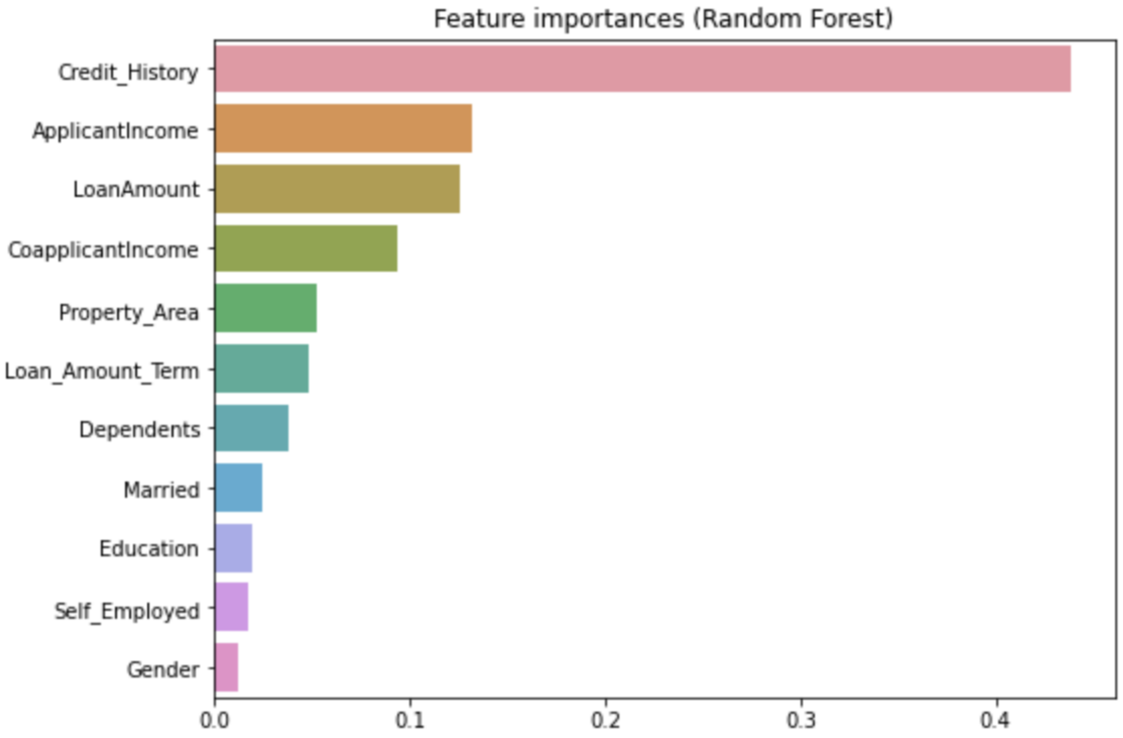
	Train	Test
Random Forest	0.848	0.821
AdaBoost	0.825	0.800
Gradient Boosting	0.822	0.805
LightGBM	0.828	0.816
XGBoost	0.823	0.806
CatBoost	0.828	0.822

Random Forest

AUC = 0.896875

Train	precision	recall	f1-score
0 (N)	0.53	0.97	0.69
1 (Y)	0.99	0.82	0.90

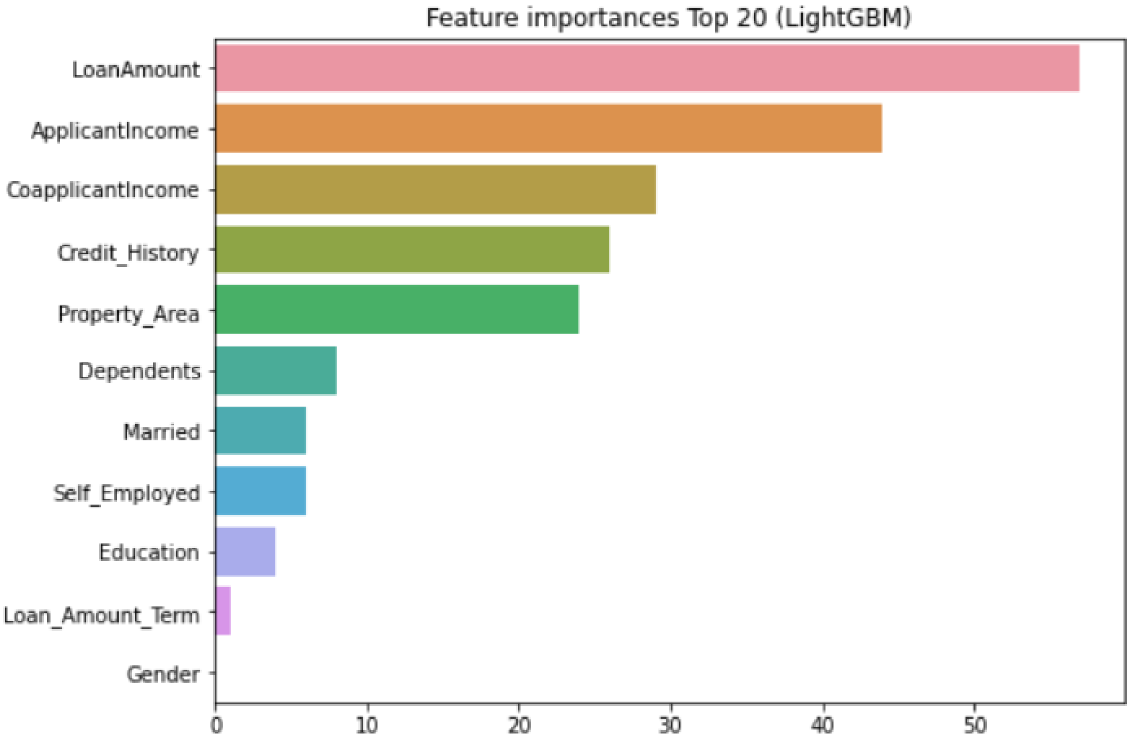
Test	precision	recall	f1-score
0 (N)	0.43	1.00	0.60
1 (Y)	1.00	0.79	0.89



LightGBM

Train	precision	recall	f1-score
0 (N)	0.50	0.91	0.64
1 (Y)	0.98	0.81	0.89

Test	precision	recall	f1-score
0 (N)	0.45	0.93	0.60
1 (Y)	0.98	0.80	0.88

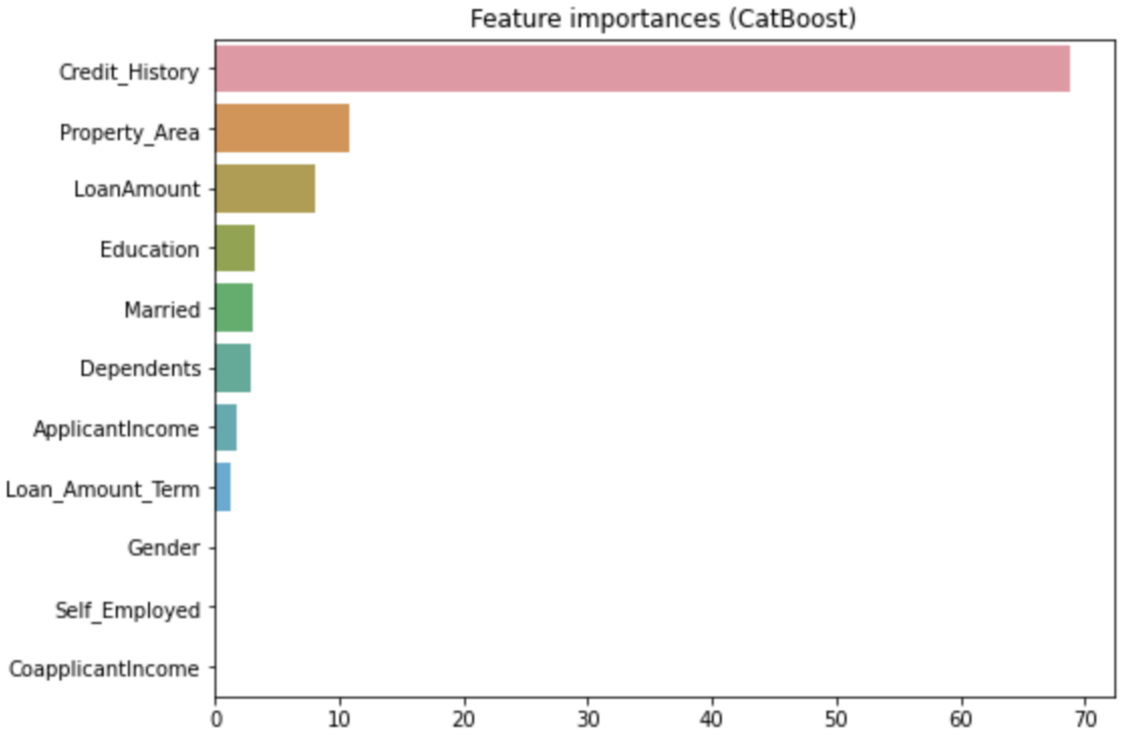


CatBoost

AUC = 0.896875

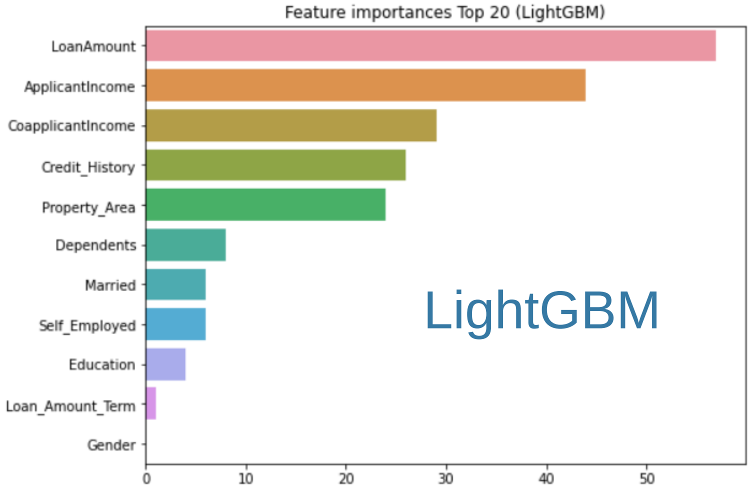
Train	precision	recall	f1-score
0 (N)	0.53	0.97	0.69
1 (Y)	0.99	0.82	0.90

Test	precision	recall	f1-score
0 (N)	0.43	1.00	0.60
1 (Y)	1.00	0.79	0.89

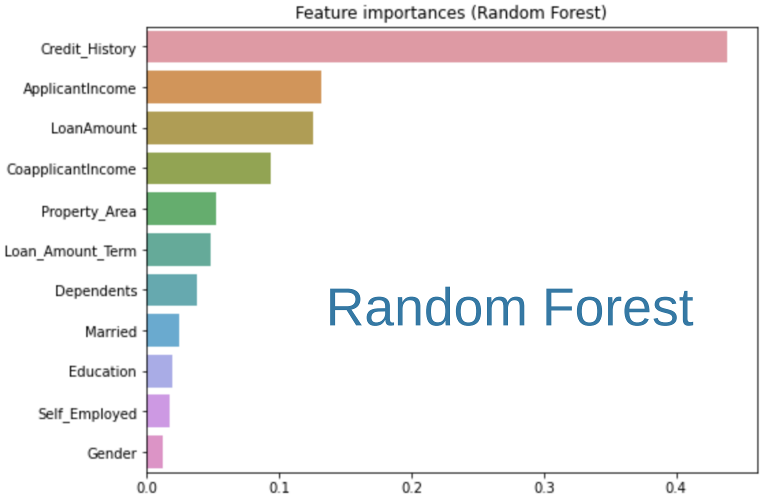


변수 중요도

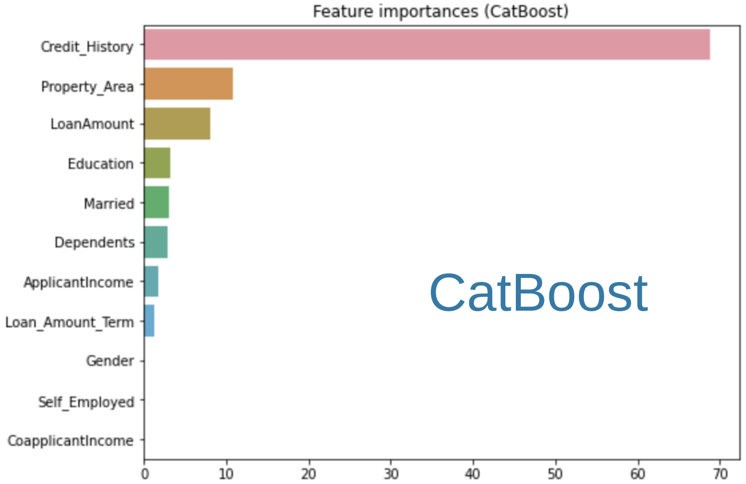
FEATURE IMPORTANCES



LightGBM

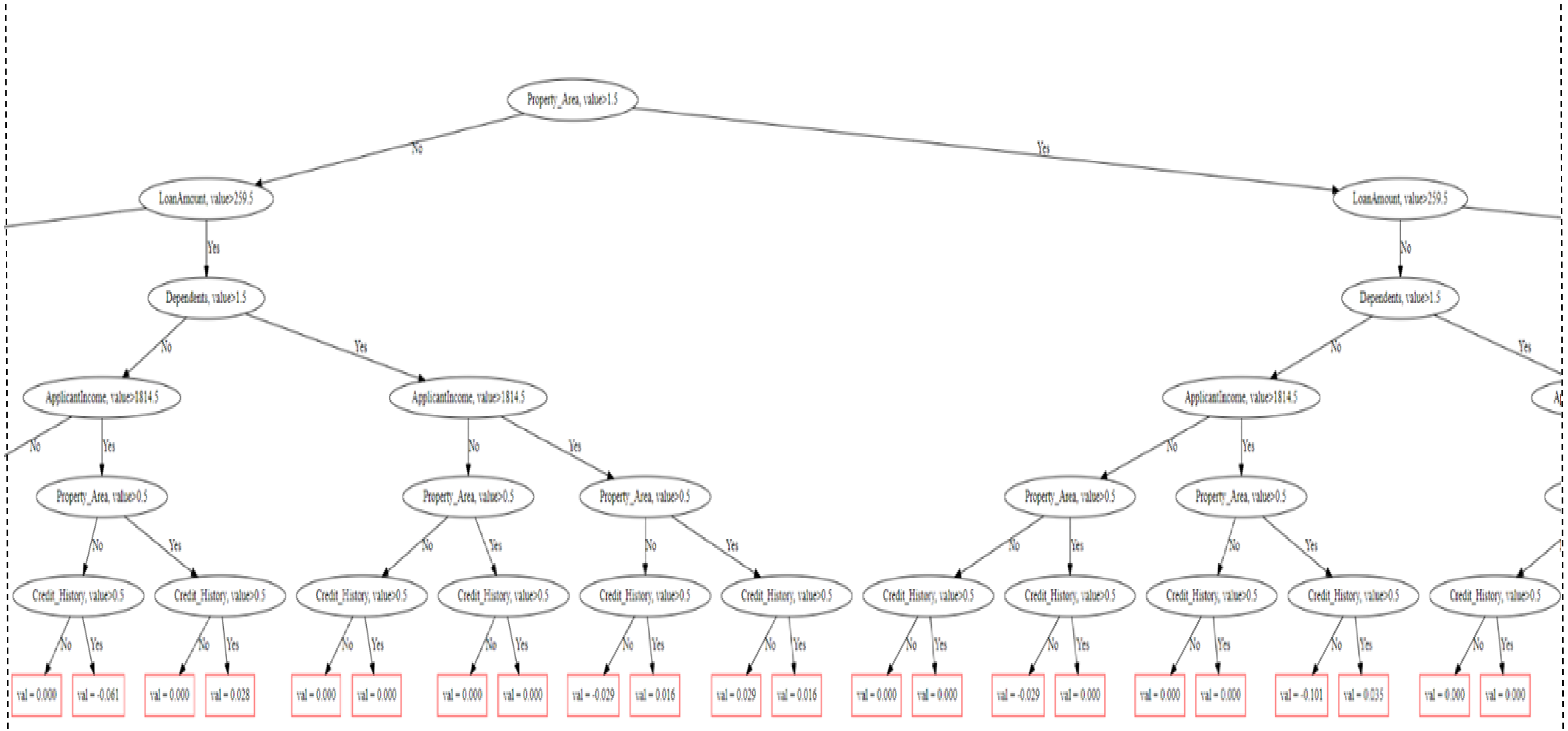


Random Forest



CatBoost

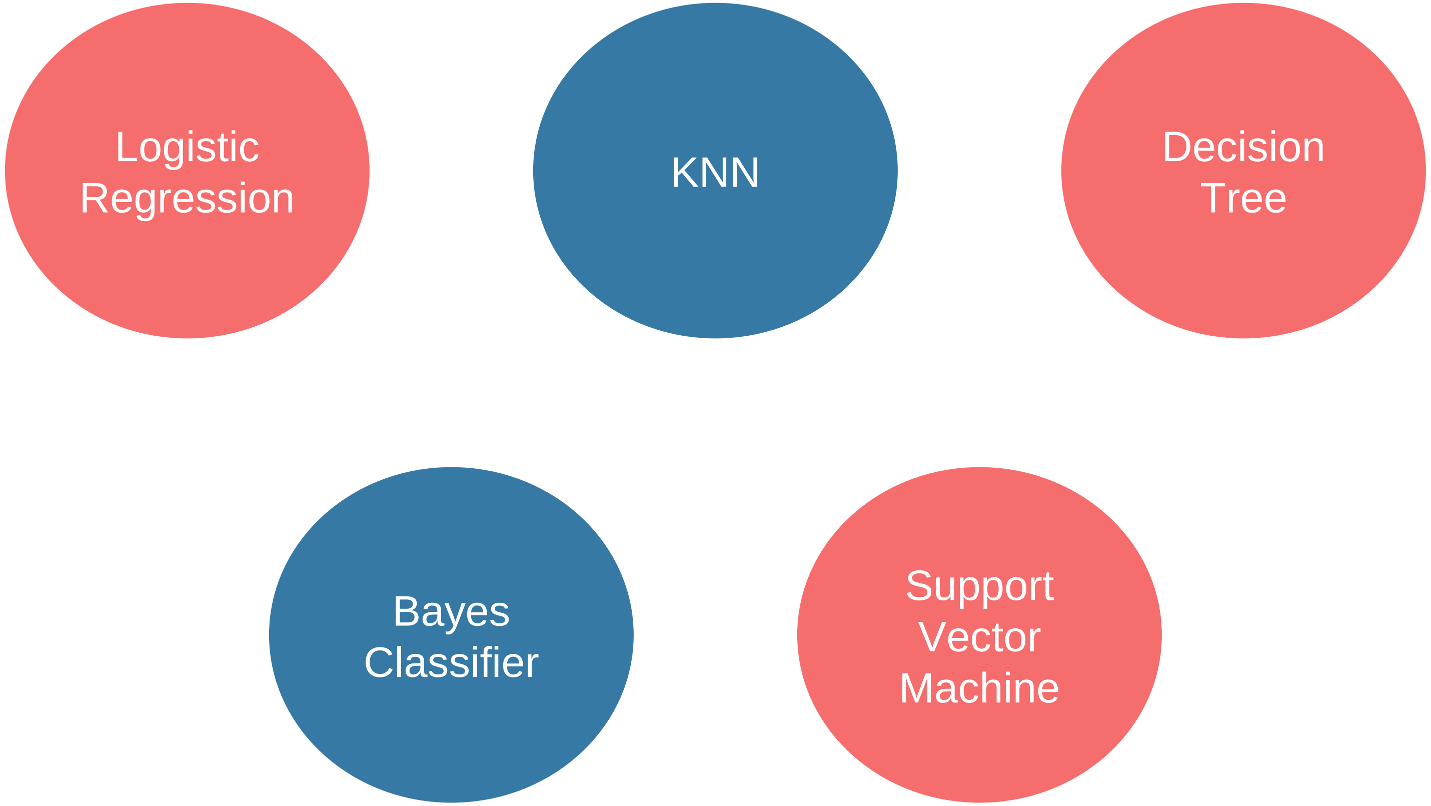
CatBoost


Property_Area, value>1.5
No
Yes
LoanAmount, value>259.5
LoanAmount, value>259.5
Dependents, value>1.5
Dependents, value>1.5
ApplicantIncome, value>1814.5
ApplicantIncome, value>1814.5
ApplicantIncome, value>1814.5
Property_Area, value>0.5
Property_Area, value>0.5
Property_Area, value>0.5
Property_Area, value>0.5
Property_Area, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
Credit_History, value>0.5
val = 0.000
val = -0.061
val = 0.000
val = 0.028
val = 0.000
val = 0.000
val = 0.000
val = 0.000
val = -0.029
val = 0.016
val = 0.029
val = 0.016
val = 0.000
val = 0.000
val = -0.029
val = 0.000
val = 0.000
val = 0.000
val = -0.101
val = 0.035
val = 0.000
val = 0.000

3

결론

CONCLUSION



Logistic Regression

KNN

Decision Tree

Bayes Classifier

Support Vector Machine

③

결론

CONCLUSION

Random Forest

Gradient Boosting

LightGBM

AdaBoost

XGBoost

CatBoost

5조

THANK YOU.