핀테크 기업의 대출 고객 기준 마련

서선우, 서수아, 이주노

목차

소 개

- 팀원 소개
- 일정표
- 주제 소개 및 선정 배경

팀원 소개

dropna

서선우

조장

주제선정
발표
데이터 수집
데이터 전처리
피쳐 선정 / 모델링
평가 및 검증
인사이트 도출

서수아

조원

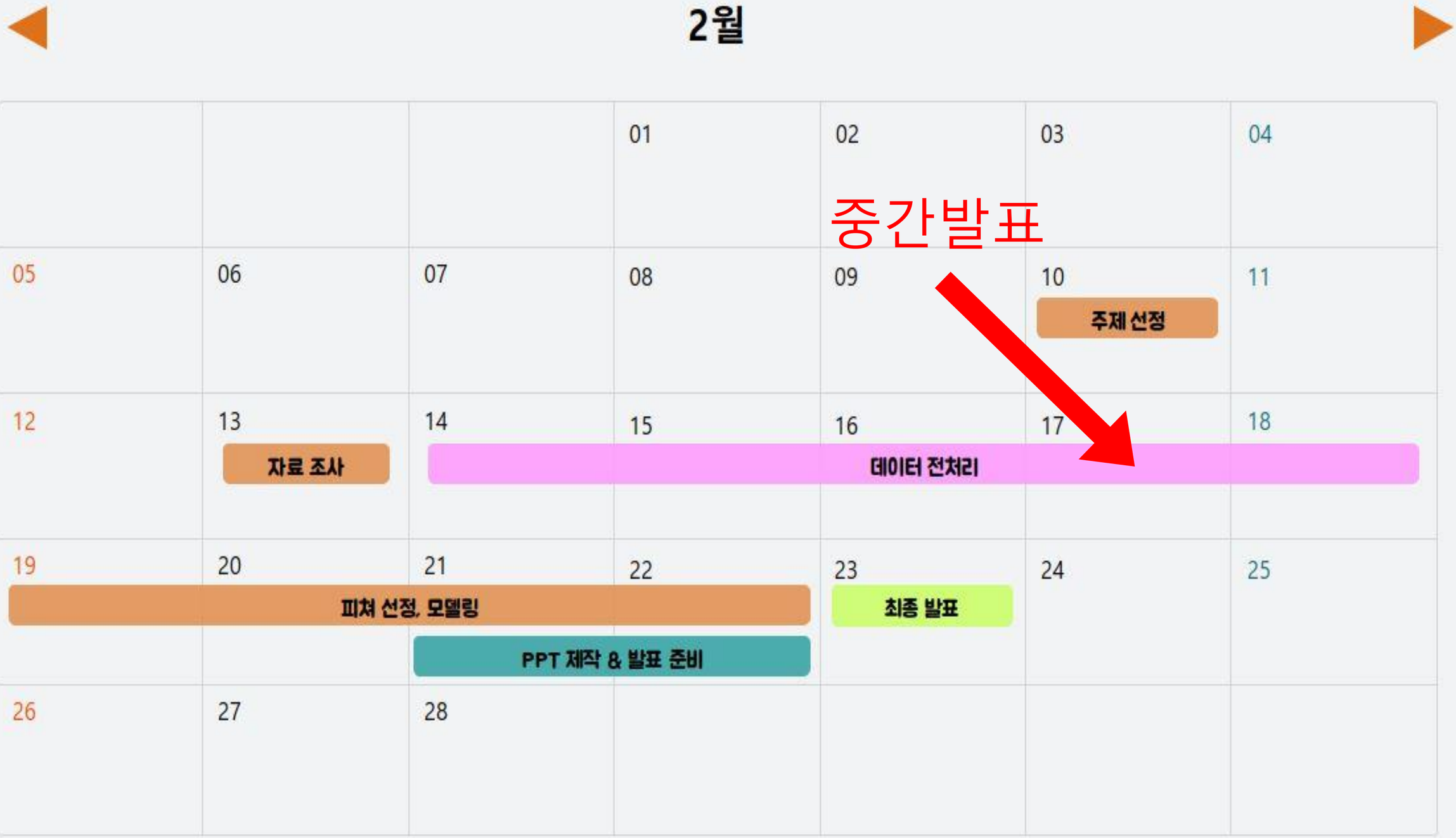
주제선정
PPT 제작
데이터 수집
데이터 전처리
피쳐 선정 / 모델링
평가 및 검증
인사이트 도출

이주노

조원

주제선정
논문탐색/보고서 작성
데이터 수집
데이터 전처리
피쳐 선정 / 모델링
평가 및 검증
인사이트 도출

일정표


2월
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
중간발표
주제 선정
자료 조사
데이터 전처리
피쳐 선정, 모델링
PPT 제작 & 발표 준비
최종 발표

주제 선정

- 주제 선정 배경
- 목표

주제 선정 배경

코로나로 인한 **경기 침체**

↓

미국의 무제한적 **양적완화**

↓

과도한 **물가 상승**

↓

연준 **금리 인상**

↓

한국 시중은행들의 **과점**

연합뉴스 PiCK

금감원, 5대은행 과점 체제 깬다…'완전 경쟁' 유도 검토(종합)

입력 2023.02.15. 오전 10:27 · 수정 2023.02.15. 오전 10:29 기사원문

심재훈 기자 · 임수정 기자

81 121

가가

금감원장, '돈잔치' 논란에 '은행 과점 완화' 검토 지시
인가단위 세분화·인터넷뱅크 확대·핀테크 금융업 진출 등 유력
5대 은행, 예금·대출 시장 점유율 60~70%대…'그들만의 리그'

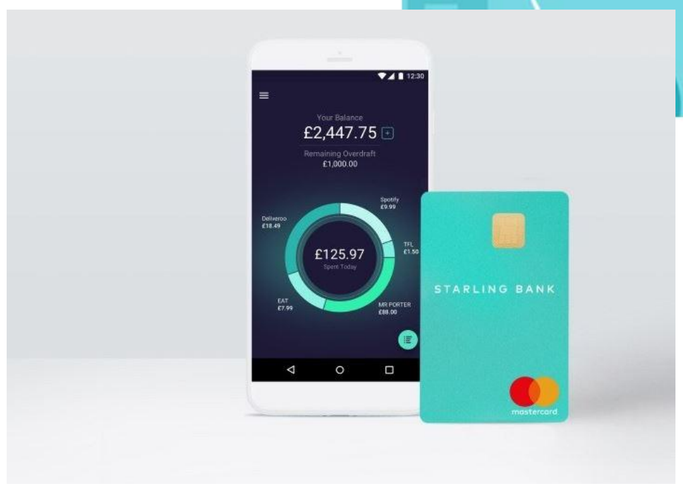
HOME > 경제 > 금융

금융감독원, 은행 과점 체제를 깬다...영국식 '챌린저 은행' 도입할까

나희재 기자 | 승인 2023.02.15 16:47 | 댓글 0

목표

신생 핀테크 기업의 대출 고객 기준 마련



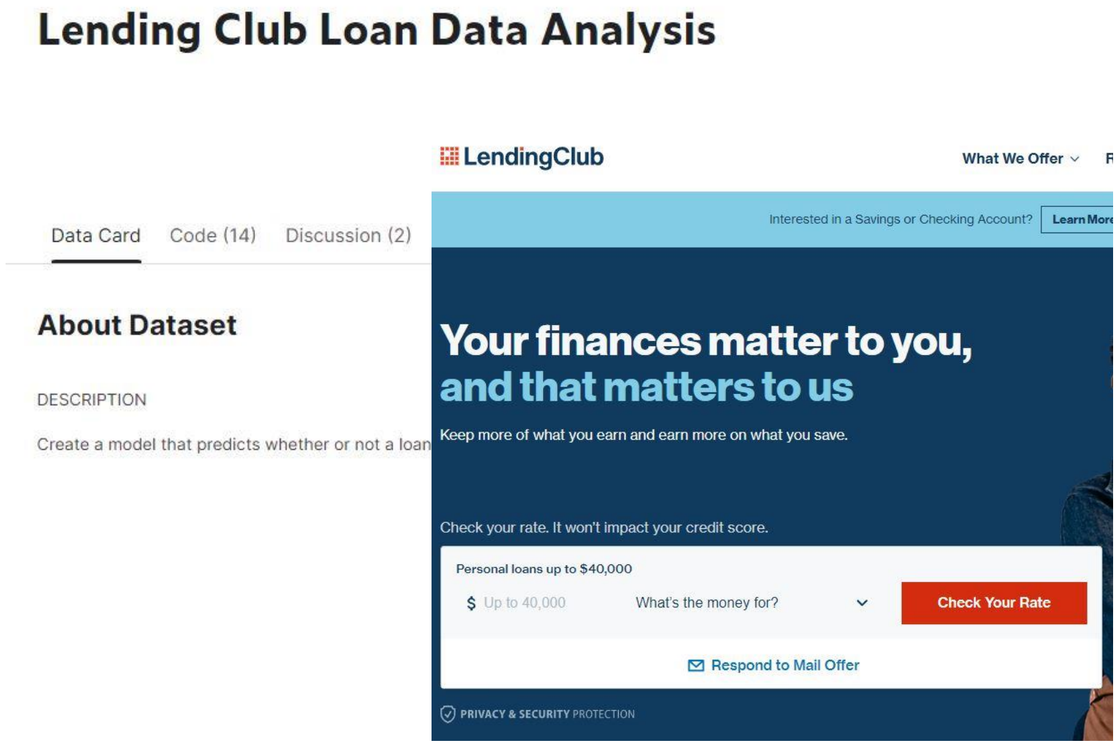
Revolut Atom bank N26
monzo STARLING BANK
TANDEM monese

- 미국의 P2P 대출 업체인 Lending Club의 기준을 참고
- 고객들의 재무적 데이터를 통해 타겟 고객들의 안정성 증대 목적
- 안전한 고객들은 어떤 경향을 보이는지 확인

데이터 전처리

- 데이터 소개
- 데이터 확인

데이터 소개(1)



credit.poli	purpose	int.rate	installmen	log.annual	dti	fico	days.with.c	revol.bal	revol.util	inq.last.6m	delinq.2yrs	pub.rec	not.fully.pa
1	debt_cons	0.1189	829.1	11.350407	19.48	737	5639.9583	28854	52.1	0	0	0	0
1	credit_carc	0.1071	228.22	11.082143	14.29	707	2760	33623	76.7	0	0	0	0
1	debt_cons	0.1357	366.86	10.373491	11.63	682	4710	3511	25.6	1	0	0	0
1	debt_cons	0.1008	162.34	11.350407	8.1	712	2699.9583	33667	73.2	1	0	0	0
1	credit_carc	0.1426	102.92	11.299732	14.97	667	4066	4740	39.5	0	1	0	0
1	credit_carc	0.0788	125.13	11.904968	16.98	727	6120.0417	50807	51	0	0	0	0
1	debt_cons	0.1496	194.02	10.714418	4	667	3180.0417	3839	76.8	0	0	1	1
1	all_other	0.1114	131.22	11.0021	11.08	722	5116	24220	68.6	0	0	0	1
1	home_imp	0.1134	87.19	11.407565	17.25	682	3989	69909	51.1	1	0	0	0
1	debt_cons	0.1221	84.12	10.203592	10	707	2730.0417	5630	23	1	0	0	0
1	debt_cons	0.1347	360.43	10.434116	22.09	677	6713.0417	13846	71	2	0	1	0
1	debt_cons	0.1324	253.58	11.835009	9.16	662	4298	5122	18.2	2	1	0	0
1	debt_cons	0.0859	316.11	10.933107	15.49	767	6519.9583	6068	16.7	0	0	0	0
1	small_busi	0.0714	92.82	11.512925	6.5	747	4384	3021	4.8	0	1	0	0
1	debt_cons	0.0863	209.54	9.4879721	9.73	727	1559.9583	6282	44.6	0	0	0	0
1	major_pur	0.1103	327.53	10.738915	13.04	702	8159.9583	5394	53.4	1	0	0	0
1	all_other	0.1317	77.69	10.522773	2.26	672	3895.9583	2211	88.4	0	0	0	0
1	credit_carc	0.0894	476.58	11.608236	7.07	797	6510.9583	7586	52.7	1	0	0	0
1	debt_cons	0.1039	584.12	10.491274	3.8	712	2760	8311	59.8	0	0	0	0
1	major_pur	0.1513	173.65	11.0021	2.74	667	1126.9583	591	84.4	3	0	0	0
1	all_other	0.08	188.02	11.225243	16.08	772	4888.9583	29797	23.2	1	0	0	0
1	all_other	0.0863	474.42	10.819778	2.59	797	11951	5656	27.6	0	0	0	0
1	credit_carc	0.1355	339.6	11.512925	7.94	662	1939.9583	21162	57.7	0	0	0	0

○ **[Kaggle] Lending Club Loan Data Analysis**

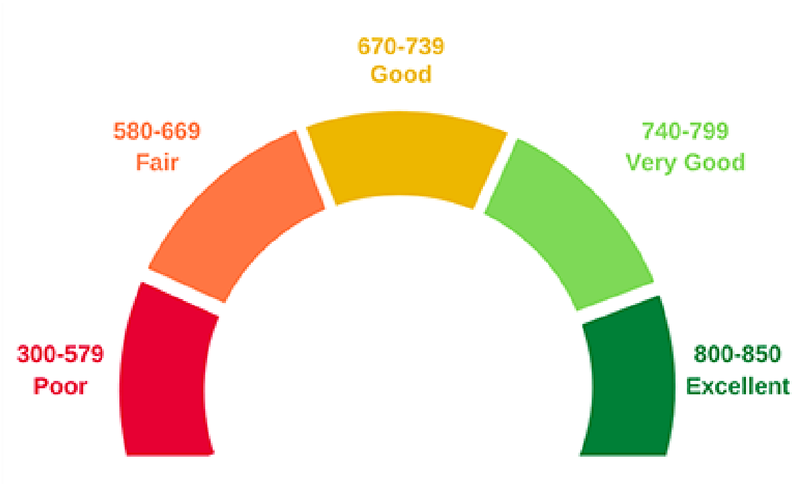
○ **Data 형태 : 9578 Rows, 14 Columns**

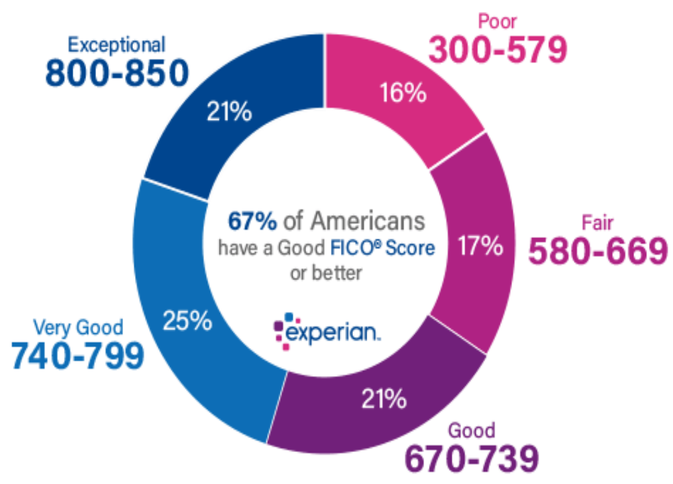
피쳐 갯수 14개

- 신용 정책
- 대출 목적
- 대출 금리
- 월분할 불입금
- 연간수익의 자연로그
- DTI(총소득대비부채)
- FICO 스코어
- 신용한도일수
- 리볼빙 잔액
- 리볼빙 이용률
- 신용조회횟수
- 연체횟수
- 부적절한 공공기록
- 전액지급여부

데이터 소개(2)

Fico Score와 KCB 신용점수 비교





신용점수		2021년 말
950~1000점	22%	10,639,706명
900~949점	18%	8,943,447명
850~899점	9%	4,301,649명
800~849점	7%	3,287,942명
750~799점	12%	5,587,137명
700~749점	16%	7,595,314명
600~699점	11%	5,275,306명
300~599점	2%	1,085,546명
300점 미만	4%	1,798,557명

은 행	구 분	CB사 신용점수별 금리(%)											평균신용점수
		1000~951점	950~901점	900~851점	850~801점	800~751점	750~701점	700~651점	650~601점	600점 이하	평균금리	서민금융제외 평균금리	
NH농협은행 재 소	대출금리	6.46	6.78	7.31	7.74	8.43	9.08	8.26	7.97	8.39	7.14	7.13	895
신한은행 재 소	대출금리	6.35	6.57	7.01	7.32	7.87	8.36	8.68	9.67	10.89	6.84	6.60	906
우리은행 재 소	대출금리	6.30	6.48	6.87	7.40	7.75	7.72	8.89	9.62	10.21	6.63	6.46	922
하나은행 재 소	대출금리	6.36	6.66	7.22	7.83	8.30	8.93	9.16	9.61	10.41	7.10	6.32	895
KB국민은행 재 소	대출금리	6.11	6.60	7.22	7.53	8.13	8.37	9.19	9.94	10.42	6.88	6.57	901
카카오뱅크 재 소	대출금리	5.73	6.19	6.68	7.14	7.77	8.55	9.40	10.82	11.56	8.04	9.06	770
토스뱅크 재 소	대출금리	6.77	7.27	7.94	8.68	9.67	10.65	11.58	12.35	12.69	8.47	8.71	857

금융회사	대출종류	금리구분	900점초과	801~900점	701~800점	601~700점	501~600점
JT친애저축은행	일반신용대출	대출금리	-	12.03%	12.06%	12.11%	-
한화저축은행	일반신용대출	대출금리	12.53%	12.76%	13.34%	13.86%	14.19%
JT저축은행	일반신용대출	대출금리	12.76%	13.44%	13.71%	14.10%	-

FICO score 기준을 참고해 Very Good 이상 / Good 이상 / Good 미만 3단계로 범주화

0 : 양호 / 1 : 보통 / 2 : 미흡

데이터 확인(1)

기초 통계량 확인 및 인코딩

	credit.policy	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yrs	pub.rec	not.fully.paid
count	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000	9.578000e+03	9578.000000	9578.000000	9578.000000	9578.000000	9578.000000
mean	0.804970	0.122640	319.089413	10.932117	12.606679	710.846314	4560.767197	1.691396e+04	46.799236	1.577469	0.163708	0.062122	0.160054
std	0.396245	0.026847	207.071301	0.614813	6.883970	37.970537	2496.930377	3.375619e+04	29.014417	2.200245	0.546215	0.262126	0.366676
min	0.000000	0.060000	15.670000	7.547502	0.000000	612.000000	178.958333	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.103900	163.770000	10.558414	7.212500	682.000000	2820.000000	3.187000e+03	22.600000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.122100	268.950000	10.928884	12.665000	707.000000	4139.958333	8.596000e+03	46.300000	1.000000	0.000000	0.000000	0.000000
75%	1.000000	0.140700	432.762500	11.291293	17.950000	737.000000	5730.000000	1.824950e+04	70.900000	2.000000	0.000000	0.000000	0.000000
max	1.000000	0.216400	940.140000	14.528354	29.960000	827.000000	17639.958330	1.207359e+06	119.000000	33.000000	13.000000	5.000000	1.000000

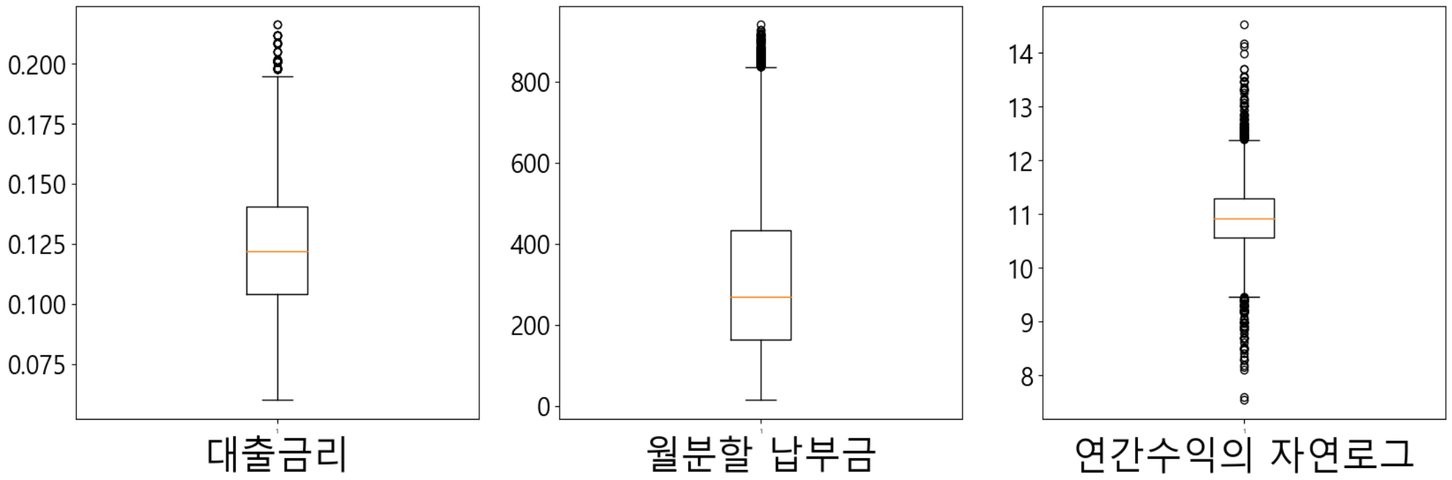
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9578 entries, 0 to 9577
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   credit.policy      9578 non-null   int64
 1   purpose            9578 non-null   object
 2   int.rate           9578 non-null   float64
 3   installment        9578 non-null   float64
 4   log.annual.inc     9578 non-null   float64
 5   dti                9578 non-null   float64
 6   fico               9578 non-null   int64
 7   days.with.cr.line  9578 non-null   float64
 8   revol.bal          9578 non-null   int64
 9   revol.util         9578 non-null   float64
 10  inq.last.6mths     9578 non-null   int64
 11  delinq.2yrs        9578 non-null   int64
 12  pub.rec            9578 non-null   int64
 13  not.fully.paid     9578 non-null   int64
dtypes: float64(6), int64(7), object(1)
memory usage: 1.0+ MB
```

	purpose
0	debt_consolidation
1	credit_card
2	debt_consolidation
3	debt_consolidation
4	credit_card
...	...
9573	all_other
9574	all_other
9575	debt_consolidation
9576	home_improvement
9577	debt_consolidation

	purpose
0	2
1	1
2	2
3	2
4	1
...	...
9573	0
9574	0
9575	2
9576	4
9577	2

데이터 확인(2)

수치형 데이터 이상치 확인



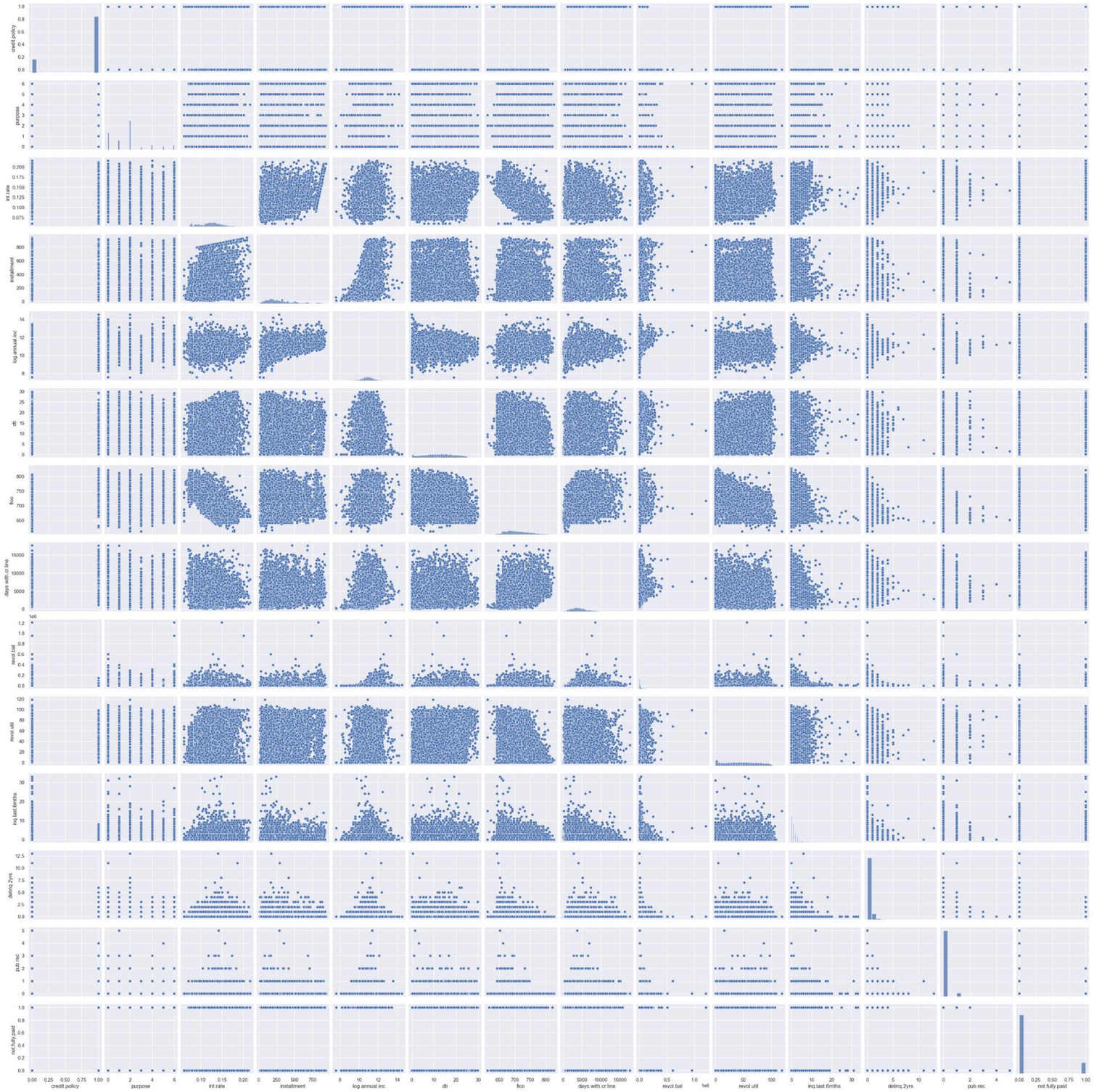
리볼빙 잔액의 경우 편차가 심해

로그화 시켜 편차를 줄임

	리볼빙 잔액
count	9578.00
mean	16913.96
std	33756.19
min	0.00
25%	3187.00
50%	8596.00
75%	18249.50
max	1207359.00

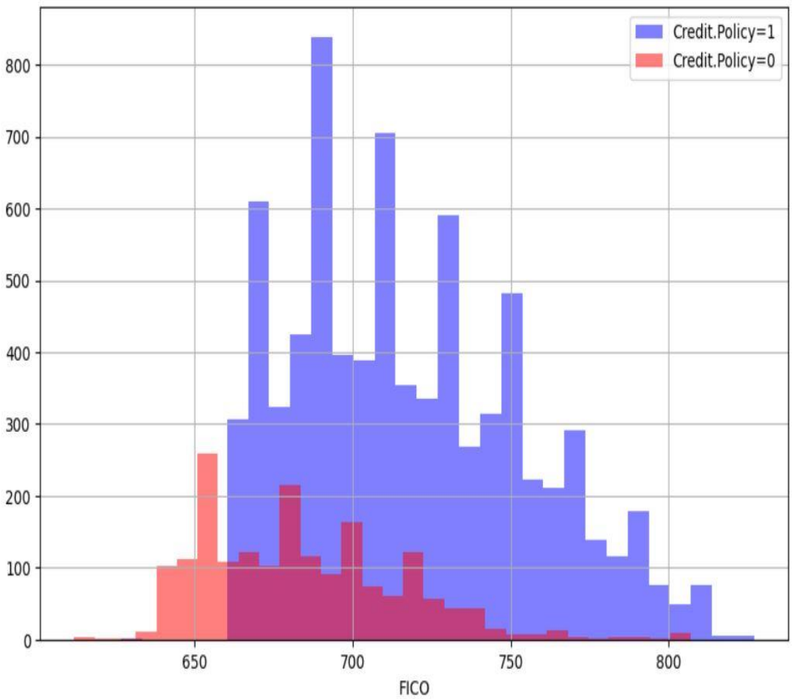
데이터 확인(3)

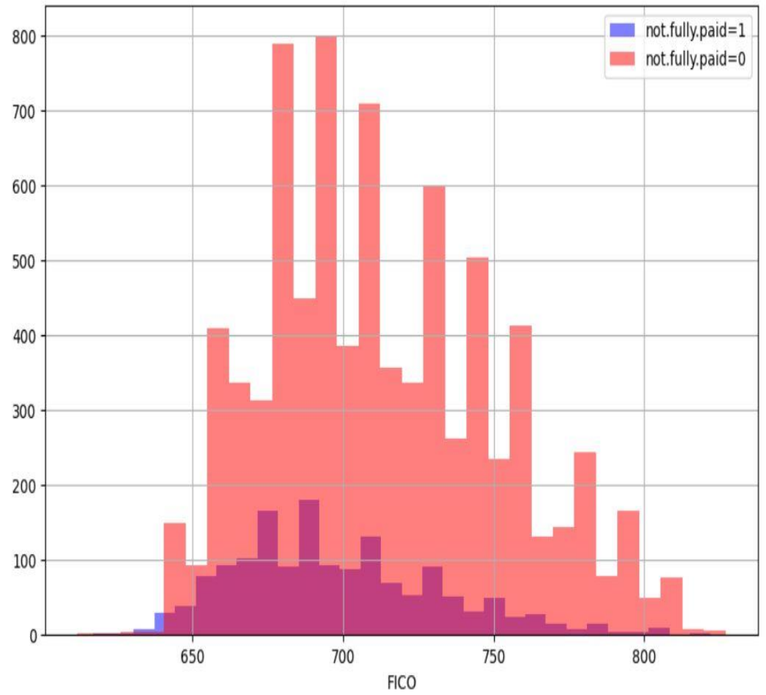
Pairplot 활용 데이터간 관계 확인

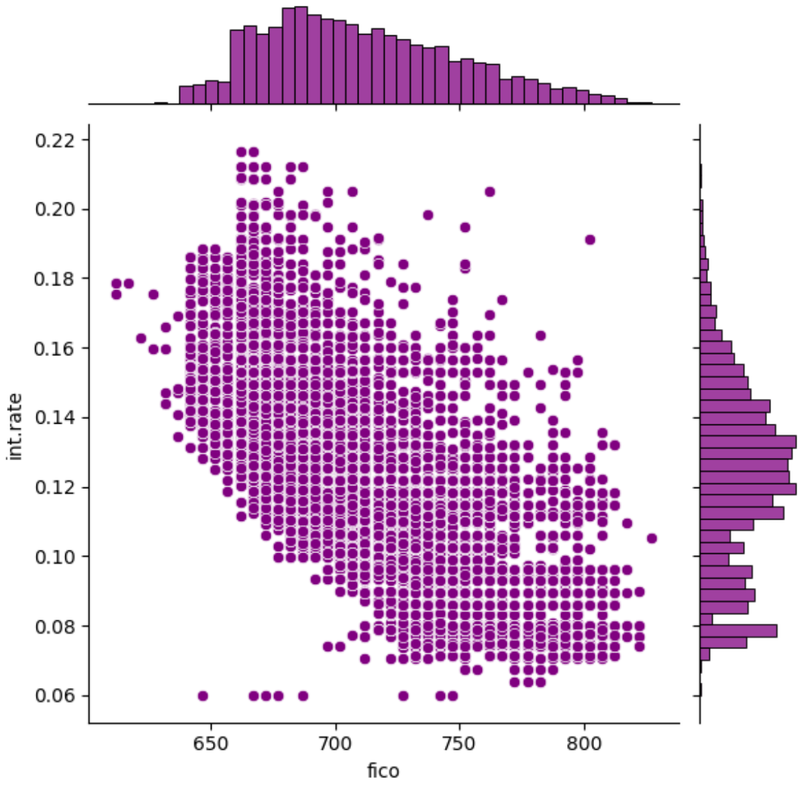


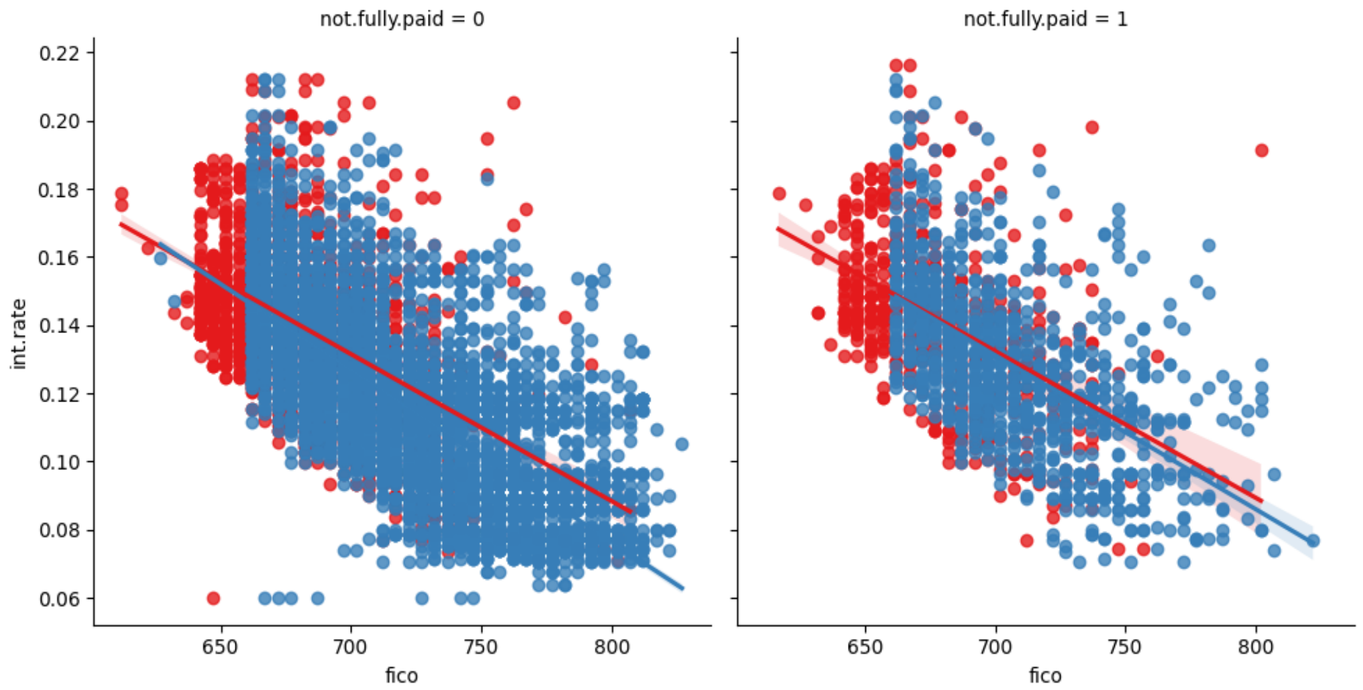
데이터 확인(4)

Jointplot, lmplot 활용 데이터간 관계 확인









EDA 결론

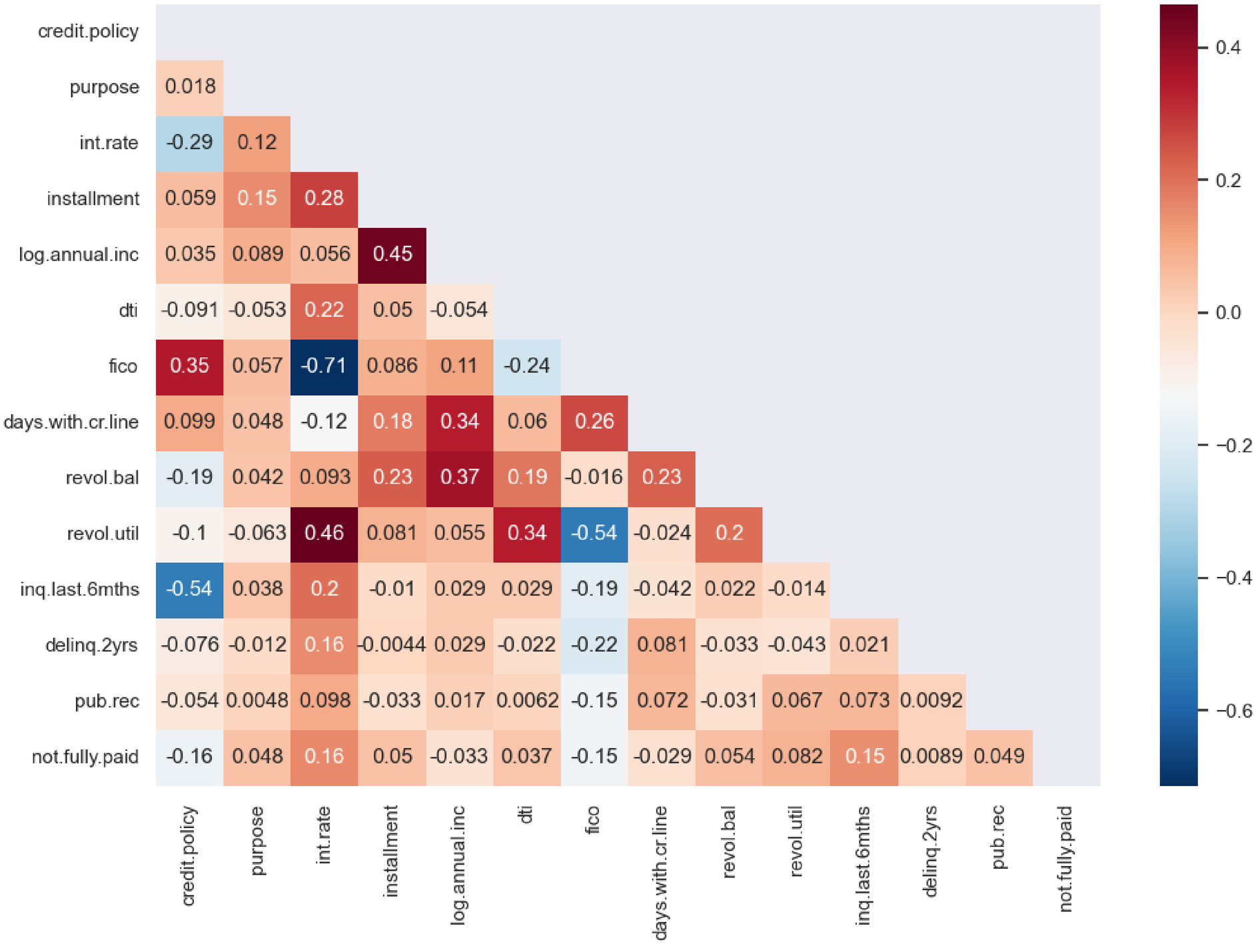
미국의 Lending Club 대출 기준의 경우

FICO 점수의 영향도 있겠지만 더욱 영향을 주는 다른 피쳐가 있을 것으로 판단.

피쳐 선정

피쳐 선정(1)

히트맵 활용 피쳐간 상관관계 확인



Fico와 대출금리(-0.71)

Fico와 리볼빙 이용률(-0.54)

신용정책과 신용조회기록(-0.54)

리볼빙이용률과 대출금리(0.46)

연간소득과 월분할 납부금(0.45)

피쳐 선정(2)

VIF 활용 피쳐간 다중공선성 확인

	VIF Factor	features
4	378.022311	log.annual.inc
6	288.442353	fico
2	36.812683	int.rate
0	8.493502	credit.policy
9	5.753433	revol.util
7	5.272656	days.with.cr.line
5	5.106430	dti
3	4.302080	installment
1	2.476473	purpose
10	2.222577	inq.last.6mths
8	1.663446	revol.bal
13	1.252136	not.fully.paid
11	1.202985	delinq.2yrs
12	1.100423	pub.rec

	VIF
fico	34.933226
int.rate	30.691195
credit.policy	8.483429
revol.util	5.496663
days.with.cr.line	5.202053
dti	5.056806
installment	4.177329
purpose	2.473059
inq.last.6mths	2.200376
revol.bal	1.561919
not.fully.paid	1.250041
delinq.2yrs	1.179850
pub.rec	1.094062

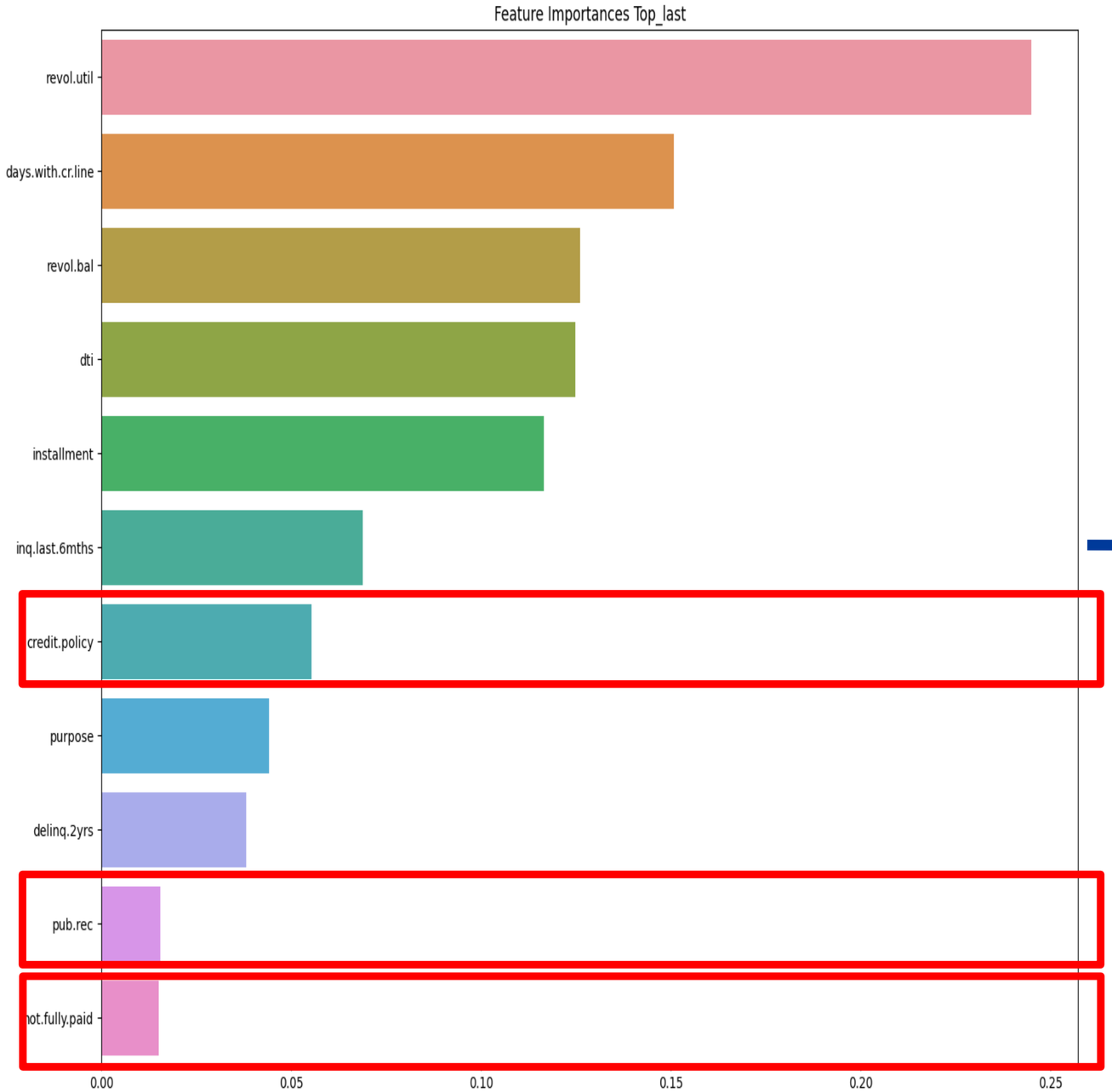
	VIF
int.rate	17.311525
credit.policy	5.575682
revol.util	5.259745
dti	4.979539
days.with.cr.line	4.563993
installment	4.149744
purpose	2.466461
inq.last.6mths	2.059992
revol.bal	1.553499
not.fully.paid	1.250018
delinq.2yrs	1.161191
pub.rec	1.087245

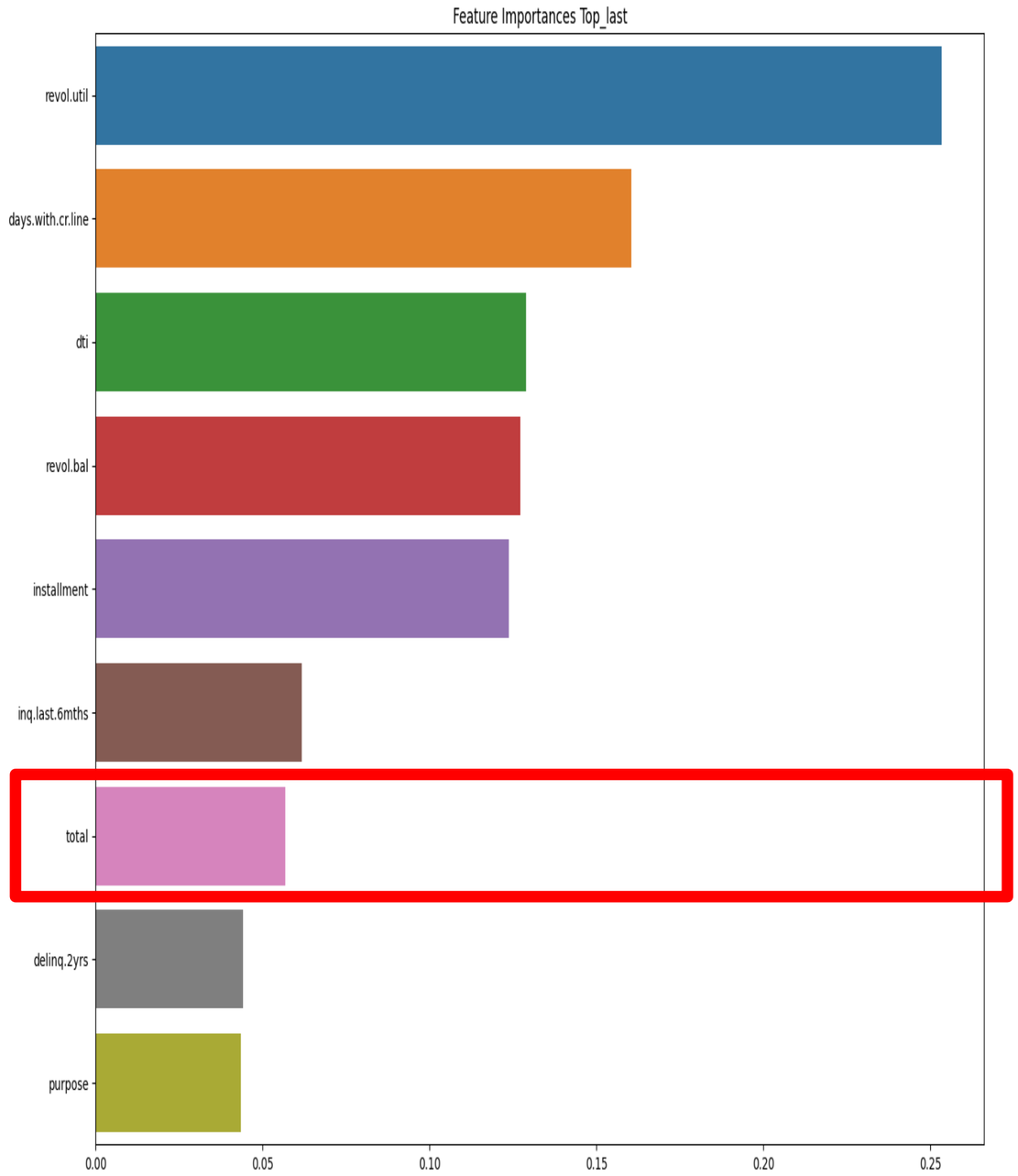
	VIF
credit.policy	4.663260
dti	4.552334
days.with.cr.line	4.538208
revol.util	3.874206
installment	3.688452
purpose	2.307742
inq.last.6mths	1.689021
revol.bal	1.548284
not.fully.paid	1.234858
delinq.2yrs	1.099721
pub.rec	1.082742

VIF가 높은 연간소득의 자연로그 → Fico → 대출금리 순서대로 삭제 후 모든 피쳐의 VIF가 5미만으로 감소했다

피쳐 선정(3)

랜덤포레스트 피쳐 임포턴스 활용 변수 중요도 파악





1차 : VIF가 높은 연간소득의 자연로그, Fico, 대출금리 삭제 후 피쳐 임포턴스 확인

2차 : 0과 1로만 이루어진 피쳐들을 더해서 재확인 (신용정책, 전액지급여부, 부적절공공기록)

최종 피쳐 선정(4)

부채

- DTI

연체

- 연체횟수
- 리볼빙 잔액
- 리볼빙 한도
- 월분할 불입금

내부

- 대출 금리
- 신용등급
- Credit Policy

개인

- 연간수익
- 신용조회 횟수
- 부적절한 공공기록
- 대출목적

모델링

데이터 분할 + 스케일링

Train Test Split

모델 선정(1)

모델 성능 비교

DecisionT	Precision	Recall	F1-Score	Support
0	0.62	0.62	0.62	268
1	0.82	0.82	0.82	1202
2	0.72	0.72	0.72	446
Micro Avg	0.77	0.77	0.77	1916
Macro Avg	0.72	0.72	0.72	1916
Weighted Avg	0.77	0.77	0.77	1916
Samples Avg	0.77	0.77	0.77	1916

RandomF	Precision	Recall	F1-Score	Support
0	0.84	0.51	0.64	268
1	0.83	0.92	0.87	1202
2	0.85	0.74	0.79	446
Micro Avg	0.83	0.82	0.83	1916
Macro Avg	0.84	0.72	0.77	1916
Weighted Avg	0.84	0.82	0.82	1916
Samples Avg	0.82	0.82	0.82	1916

XGB	Precision	Recall	F1-Score	Support
0	0.77	0.70	0.73	268
1	0.86	0.90	0.88	1202
2	0.83	0.81	0.82	446
Micro Avg	0.84	0.85	0.85	1916
Macro Avg	0.82	0.80	0.81	1916
Weighted Avg	0.84	0.85	0.84	1916
Samples Avg	0.83	0.85	0.83	1916

KNN	Precision	Recall	F1-Score	Support
0	0.20	0.04	0.07	268
1	0.67	0.76	0.71	1202
2	0.48	0.28	0.35	446
Micro Avg	0.62	0.55	0.58	1916
Macro Avg	0.45	0.36	0.38	1916
Weighted Avg	0.56	0.55	0.54	1916
Samples Avg	0.55	0.55	0.55	1916

결과가 가장 좋은

XGB 선택

모델 선정(2)

GridSearchCV

```
print("estimator : ",gcv.best_estimator_)
print("params : ",gcv.best_params_)
```

[21] Pytho

```
estimator :  XGBClassifier(base_score=None, booster='gbtree', callbacks=None,
              colsample_bylevel=0.9, colsample_bynode=None,
              colsample_bytree=0.5, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=0, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=5, max_leaves=None,
              min_child_weight=1, missing=nan, monotone_constraints=None,
              n_estimators=50, n_jobs=None, nthread=4, num_parallel_tree=None,
              predictor=None, ...)
params :  {'booster': 'gbtree', 'colsample_bylevel': 0.9, 'colsample_bytree': 0.5, 'gamma': 0, 'max_depth': 5, 'min_child_weight': 1, 'n_estimators': 50, 'nthread': 4, 'objective':
'binary:logistic', 'random_state': 2, 'silent': True}
```

Best Params

booster : gbtree

colsample_bylevel : 0.9

colsample_bytree : 0.5

Gamma : 0

max_depth : 5

min_child_weight : 1

n_estimators : 50

nthread : 4

objective : binary:logistic

random_state : 2

Silent : True

모델 평가

XGB	Precision	Recall	F1-Score
0 (Poor)	0.77	0.70	0.73
1 (Standard)	0.86	0.90	0.88
2 (Good)	0.83	0.81	0.82
Micro Avg	0.84	0.85	0.85
Macro Avg	0.82	0.80	0.81
Weighted Avg	0.84	0.85	0.84
Samples Avg	0.83	0.85	0.83
Accuracy		0.80	

→ 낮은 신용등급의 사람들을 구분할때는

재무적인 요소만으로 판단할 수 없다

시행 착오

1. 데이터 전처리

- 레이블 인코딩 → get dummies() 가변수화

2. 피쳐 선택

- 분류모델의 경우 다중공선성 확인 불필요

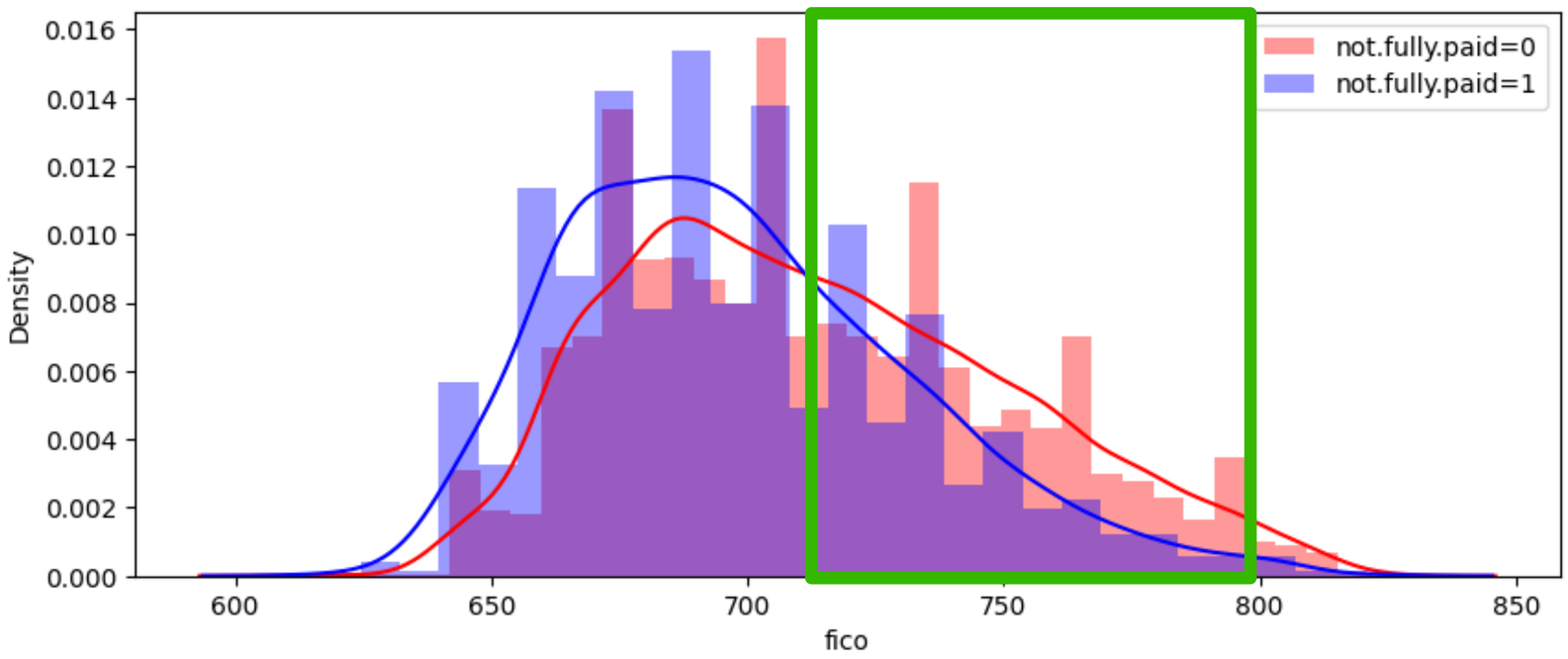
3. 모델 선정

- 차원축소의 경우 분류모델에서 무의미
- GridsearchCV를 돌려봤으나 디폴트 값을 넘는 하이퍼 파라미터 도출 실패

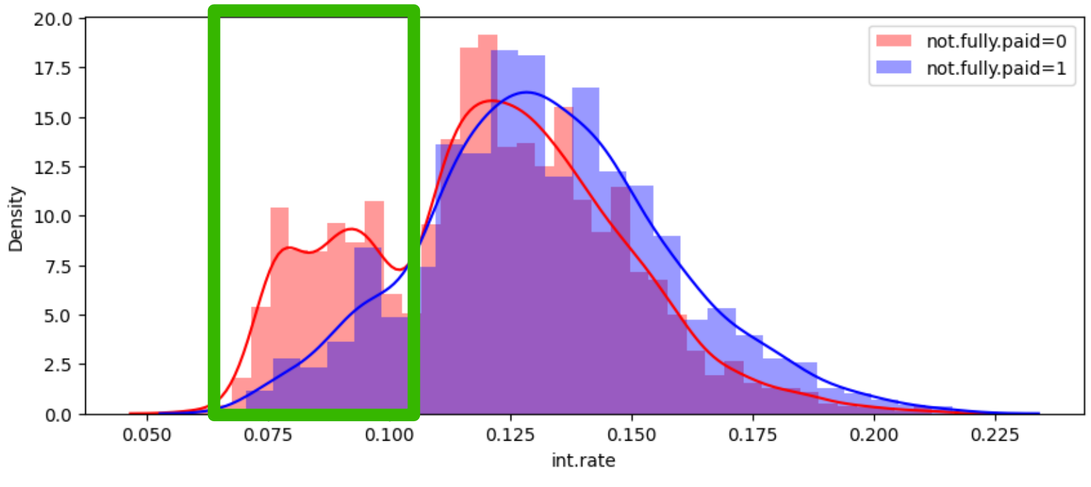
인사이트

경향분석(1)

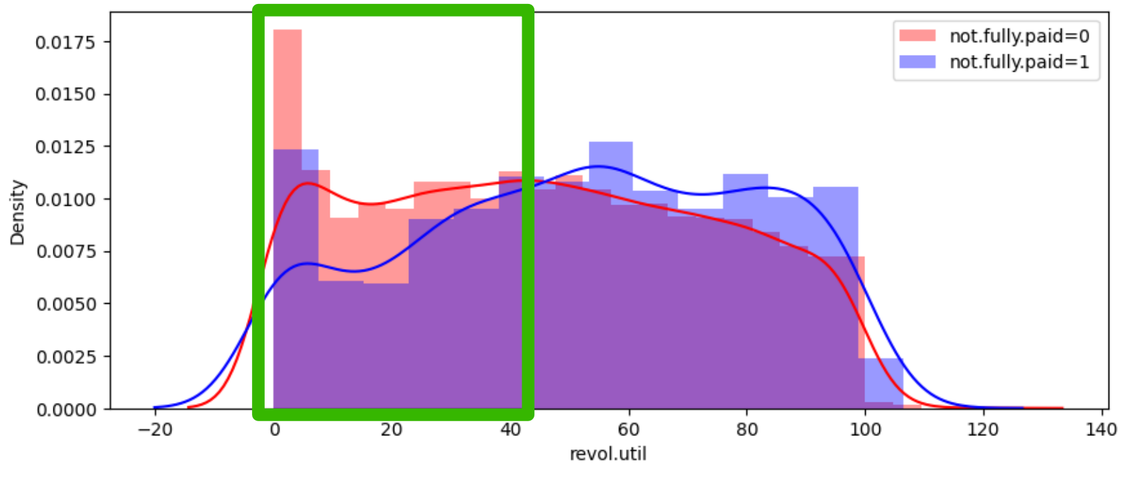
EDA를 통한 경향분석



FICO와 전액지급여부 상관관계



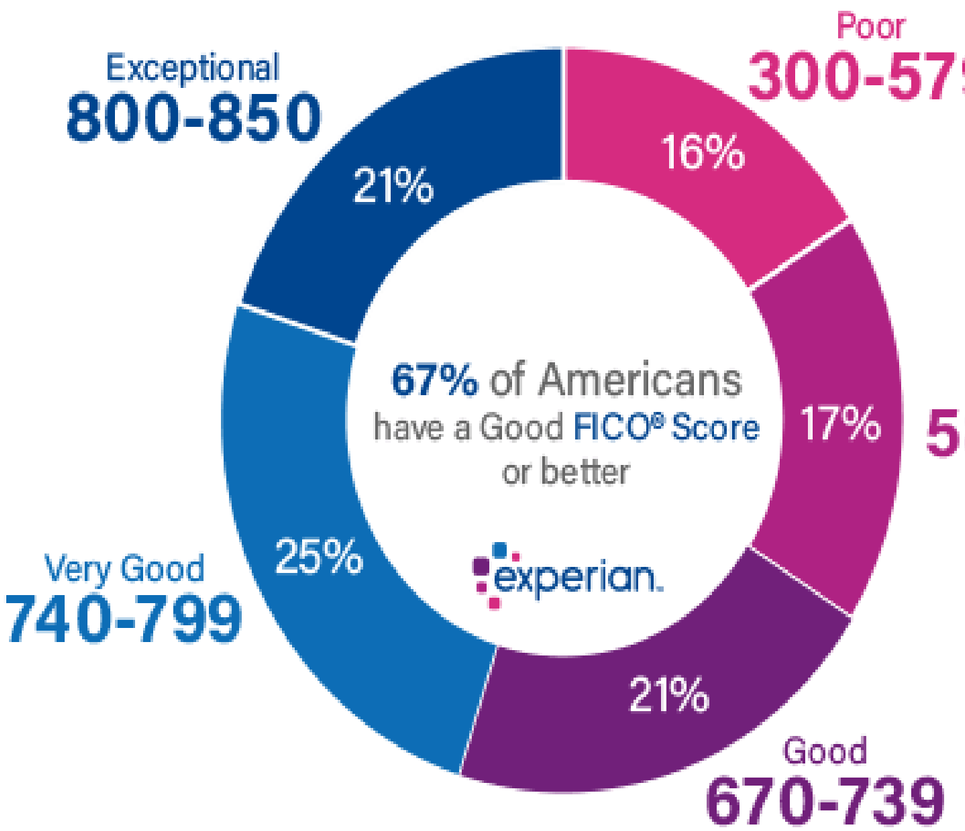
대출금리와 전액지급여부 상관관계



리볼빙 비율과 전액지급여부 상관관계

경향분석(2)

Fico Score와 KCB 신용점수 비교



신용점수		2021년 말
950~1000점	22%	10,639,706명
900~949점	18%	8,943,447명
850~899점	9%	4,301,649명
800~849점	7%	3,287,942명
750~799점	12%	5,587,137명
700~749점	16%	7,595,314명
600~699점	11%	5,275,306명
300~599점	2%	1,085,546명
300점 미만	4%	1,798,557명

FICO score 기준을 참고해 Very Good 이상 / Good 이상 / Good 미만 3단계로 범주화

0 : 양호 / 1 : 보통 / 2 : 미흡

경향분석(3)

Permutation Importances

Weight	Feature
0.1952 ± 0.0065	int.rate
0.0938 ± 0.0093	credit.policy
0.0591 ± 0.0031	revol.util
0.0542 ± 0.0072	inq.last.6mths
0.0403 ± 0.0074	installment
0.0152 ± 0.0053	days.with.cr.line
0.0120 ± 0.0040	delinq.2yrs
0.0104 ± 0.0060	dti
0.0076 ± 0.0021	pub.rec
0.0073 ± 0.0038	log.annual.inc
0.0061 ± 0.0047	revol.bal

비재무적인 요소가 중요한 피쳐임을 알 수 있다

인사이트

1. 핀테크 기업의 경우 재무적 평가요소 외에도 비재무적 요소의 비중이 크다

- Lending Club의 경우 Credit Policy, 신청서 평가 등 (Adam Nowak, Amanda Ross, Christoper Yencha, 2015)

2. 비재무적 평가 요소란?

- K-score (박소희, 최대선, 2019)
- 공과금 납부 기록, 통신비 납부 기록, 쇼핑, SNS 활동 내역, 요일별 통화 건수, 달력 관련 기록, 고객들이 사용하는 특정 단어의 빈도, 연락하는 사람 기록 등등 (이건희 ,이기환, 2022)

3. 향후 핀테크 업체의 전략

- 비재무 데이터 확보를 위해 빅테크 기업, 이동통신사, 전자상거래 업체, SNS업체 등과 협력 필요

4. 금융 데이터와 Lending club 금융 데이터의 위험 비교

- 채무불이행 위험 비율 차이 없음(0.03 ~ 0.41%) (박성우, 2017)

한계점

1. 데이터

- 데이터 양과 컬럼 부족(분석할 Feature의 갯수가 많지 않았다는 아쉬움)

2. 비재무 데이터

- 비재무 데이터를 확보하여 분석 및 평가가 필요했으나 데이터 확보가 어려움

3. 기타 등등

- 도메인 지식, 데이터 분석 기법 숙련도 부족

레퍼런스

김은미, 박지영, "Lending Club 데이터를 이용한 다분류 기반의 개인신용등급 예측", KMIS International Conference, pp.633-637, 2018

박성우, "개인 신용 평가 예측에 대한 다양한 머신러닝 기법 연구", 대한전기학회 정보 및 제어 논문집, pp.291-292, 2017

박소희, 최대선, "개인신용정보 표본DB 기반의 대출 현황 분석 및 채무불이행 예측성능 비교, Journal of KIISE. Vol. 46. No 7. pp.627-635, 2019

이건희, 이기환, "신용카드회사의 개인사업자 신용 평가 업무에 관한 연구: 머신러닝 모델의 도입", 신용카드리뷰 Vol 16-1, 2022

Adam Nowak, Amanda Ross, Christopher Yencha, "Small Business Borrowing and Peer to Peer Lending : Evidence from Lending Club", West Virginia University Working Paper No.15-28, 2015

Mohammad Rafiqul Islam, Tabitha Kemboi, "Project: Lending Club Data Analysis", Florida State University, https://www.researchgate.net/publication/340395124, 2019