랜덤포레스트를 이용한
미국 중소기업 대출 상환 여부 예측 모형

TEAM02

통계학과 1610768 김호정

통계학과 1611680 이지영

경제학부 1614681 조유민

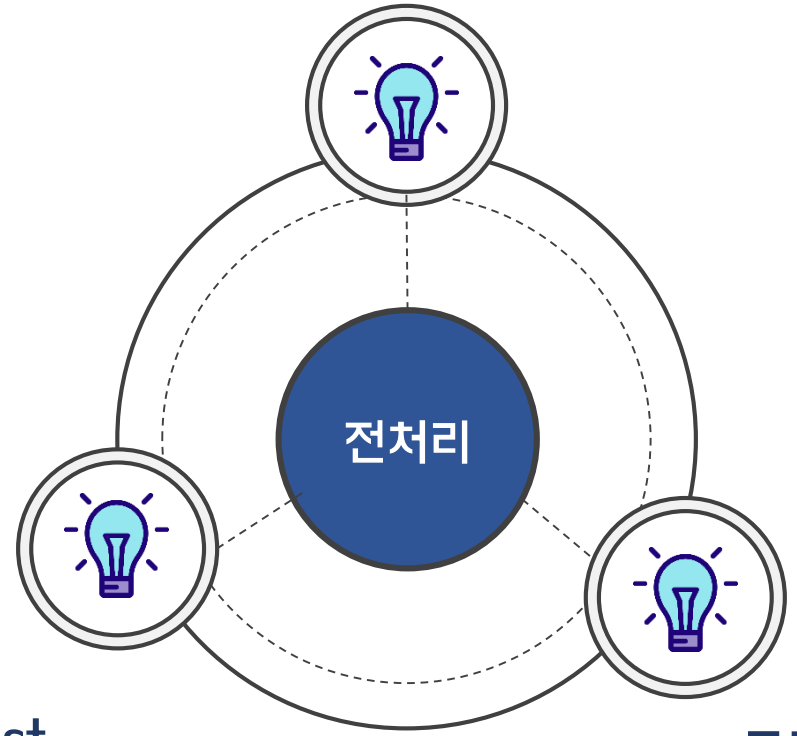
목 차

데이터 전처리

NA 처리

outcome, NewExist, DisbursementDate

NA인 행 제거



이상 범주 처리

RevLineCr, LowDoc, NewExist

원래의 범주 이외의 값들은 모두 0으로 통합

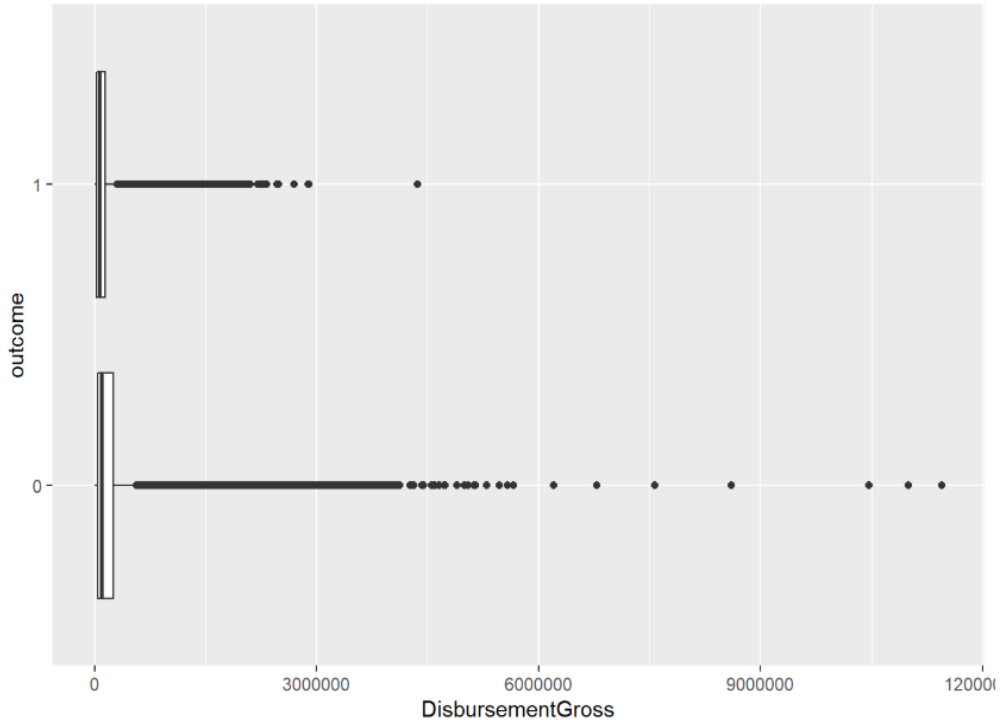
이상치 처리

수치형 변수의 이상치 확인

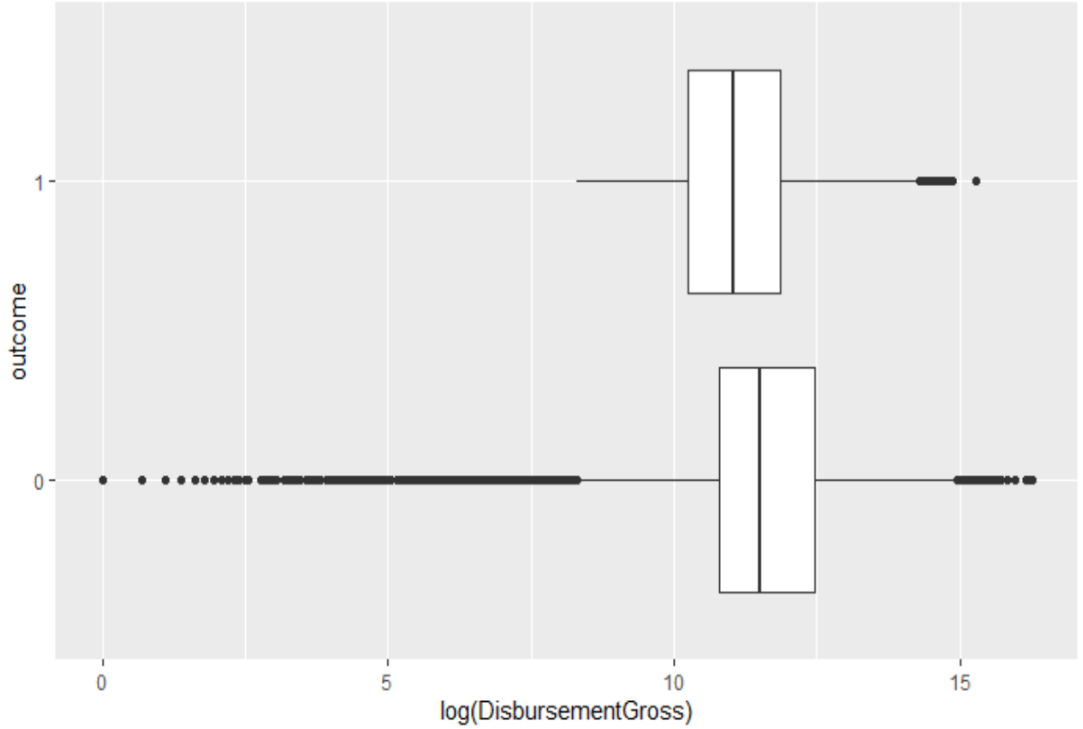
로그 변환을 했으나 큰 효과는 없었음

원래의 변수를 그대로 사용

데이터 전처리



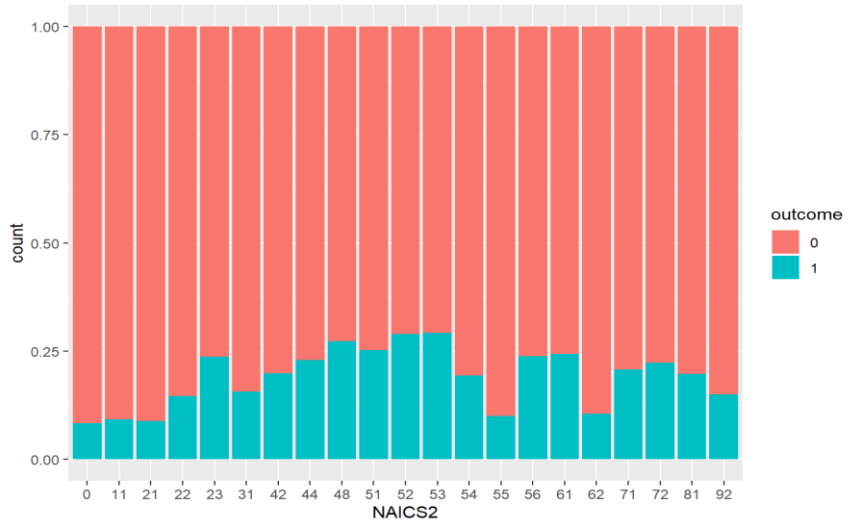
DisbursementGross 변수의 boxplot



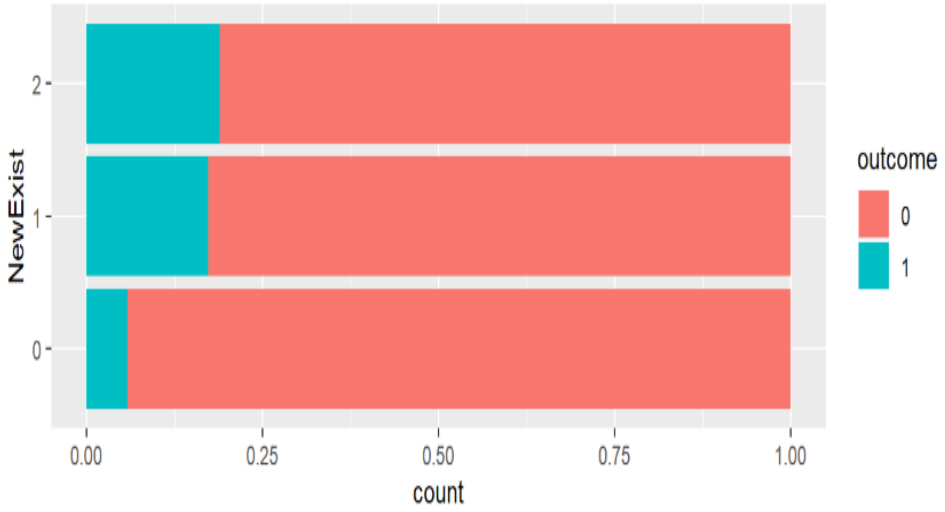
로그변환 후, DisbursementGross 변수의 boxplot

파생변수 생성

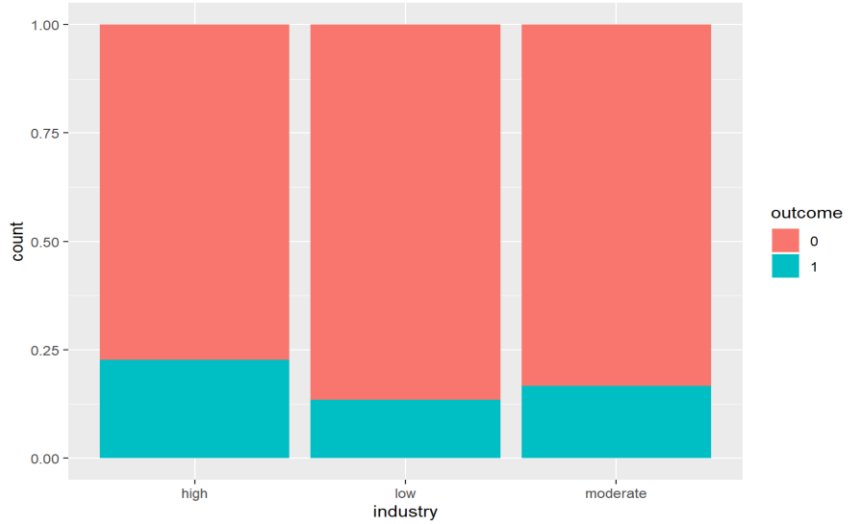
1. RealEstate : 부동산담보대출여부(1 - YES, 0 - NO)
2. NAICS2 : 기존 NAICS 코드의 앞 두 자리만 추출하여 상위의 범주로 나타낸 북미산업분류코드
3. Portion : SBA의 보증 비율(= SBA_Appv / GrAppv)
4. Industry : 위험 정도에 따른 산업분류(highrisk / lowrisk / moderate)
5. Financrisis : 금융위기(2007.12.01-2009.06.30) 여부(1 - YES, 0 - NO)



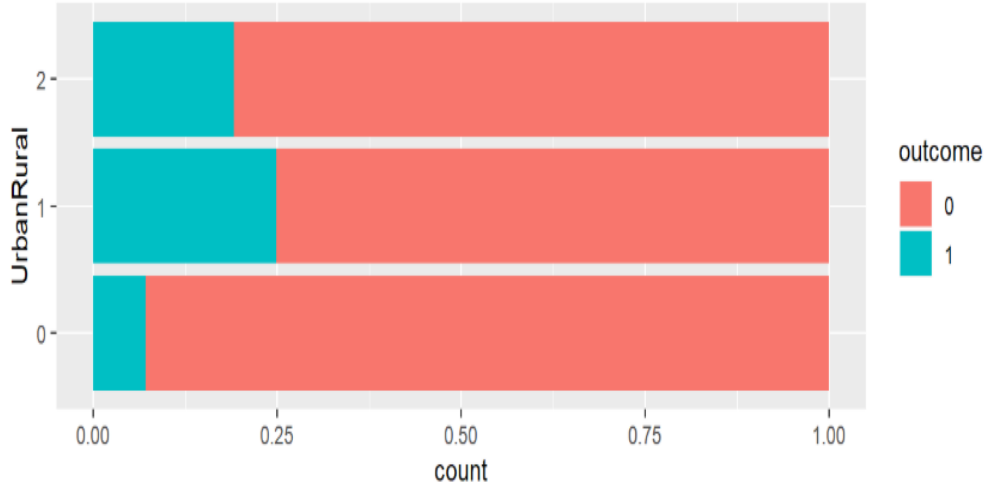
산업분류(NAICS2)에 따른 대출상환 여부 비중



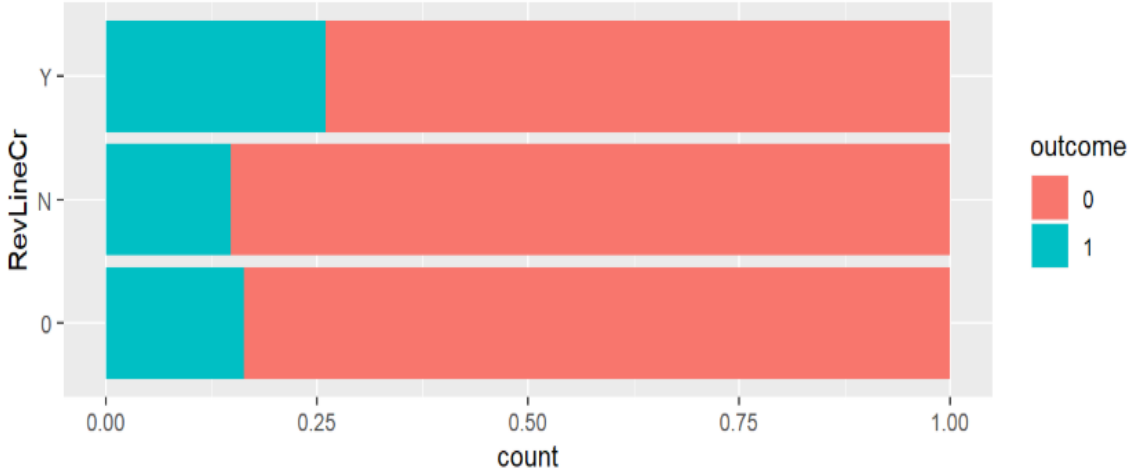
신생사업여부(NewExist)에 따른 대출상환 여부 비중



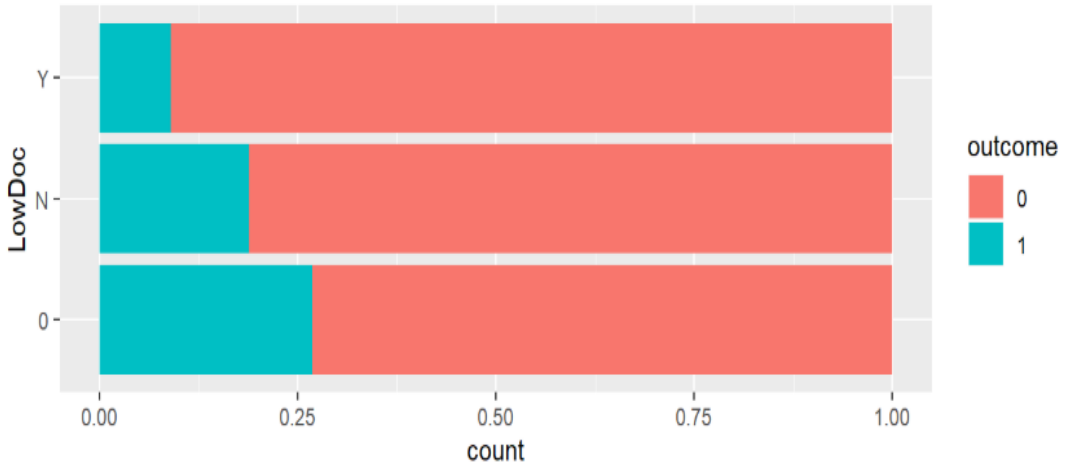
산업위험도(industry)에 따른 대출상환 여부 비중



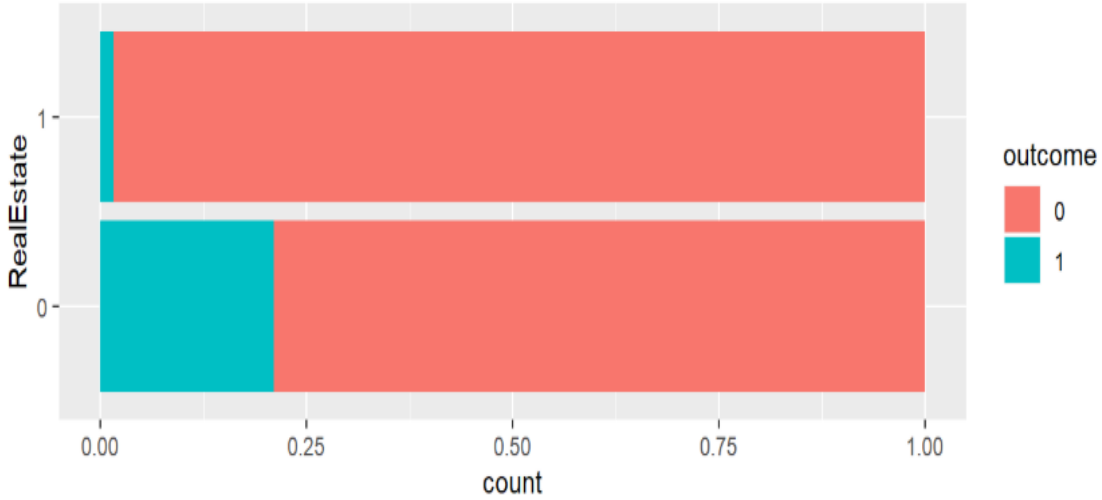
지역(UrbanRural)에 따른 대출상환 여부 비중



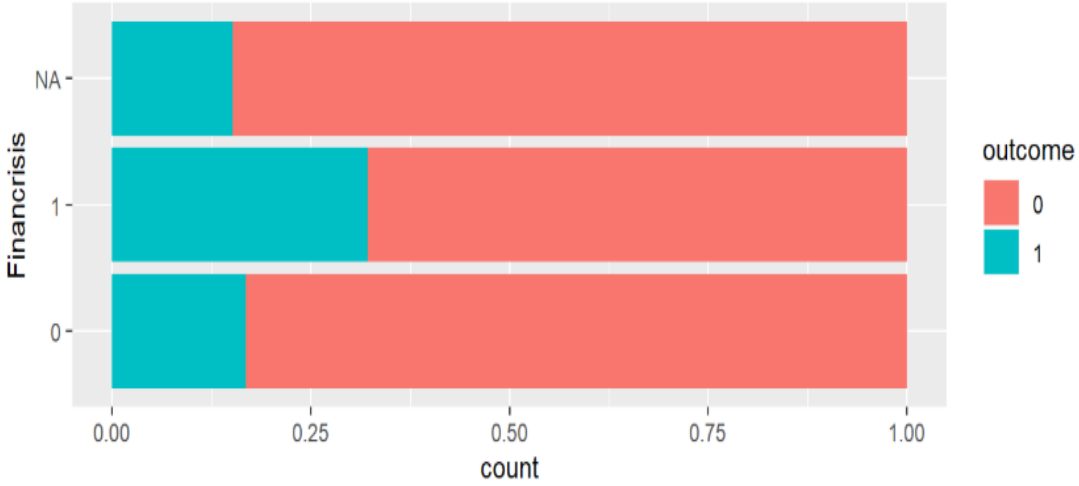
회전한도대출여부(RevLineCr)에 따른 대출상환 여부 비중



LowDoc 여부에 따른 대출상환 여부 비중



부동산담보대출여부(RealEstate)에 따른 대출상환 여부 비중



금융위기여부(Financrisis)에 따른 대출상환 여부 비중

클래스 불균형 : sampling

train 데이터를 9:1의 비율로 train set과 validation set으로 분할

네 가지 방법의 샘플링을 시도한 뒤 각 방법의 F1 score를 비교

Raw Data	Under Sampling	Over Sampling	Both Sampling	ROSE Sampling
0.82	0.75	0.84	0.82	0.66

최종 모형 : random forest

랜덤 포레스트 결정 이유

1. 의사결정나무를 적합할 때 **모든 변수를 사용하지 않는다.**

- 설명변수 사이에 높은 상관관계가 존재

- 랜덤 포레스트의 다양한 설명변수의 조합으로 상관관계 문제를 해결할 수 있다고 판단

2. **변수 중요도**를 **계산해준다.**

- 현재 총 24개의 설명변수

- 랜덤 포레스트의 변수 중요도를 기반으로 변수 선택을 할 예정

최종 모형 : random forest 모수 설정

고려한 모수

- mtry : 의사결정나무 적합 시 고려할 설명변수의 개수
- num.trees : 총 의사결정나무의 개수
- min.node.size : 한 노드에 있을 최소한의 데이터 개수
- max.dept : 의사결정나무의 최대 깊이

최종 모형

데이터 oversampling

방법 random forest

모수 - mtry 6

- num.tree 1000
- min.node.size, max.depth NULL

formula outcome~. -Id -DisbursementDate -industry -LowDoc -NewExist -ApprovalDate

Private Score	Public Score
0.839	0.835

애로사항

- 데이터에 대한 심도 있는 이해가 부족했다.

- 다양한 변수 변환을 시도하지 못했다.

예) 부스팅 모형 적합 시, 레벨의 수가 너무 많은 범주형 변수들을 적절하게 조절하지 못함.

날짜형 변수들을 이용해서 분석에 도움이 될 만한 의미 있는 변수를 만들어내지 못함.

THANK YOU☺