이학석사 학위논문

다양한 앙상블 모델을 이용한 개인신용평가 모델 개발

2020년 2월

서울대학교 대학원

통계학과

정 민 규

다양한 앙상블 모델을 이용한 개인신용평가 모델 개발

지도교수 Myunghee Cho Paik

이 논문을 이학석사 학위논문으로 제출함
2019년 12월

서울대학교 대학원
통계학과
정 민 규

정민규의 이학석사 학위논문을 인준함
2019년 12월

위 원 장 박 병 욱 (인)

부위원장 Myunghee Cho Paik (인)

위 원 임 채 영 (인)

국문초록

Home credit 은행의 데이터를 이용해 개인 신용평가 모델을 만들어 'X일' 이상 연체할 사람을 예측해 보았다. 전처리 과정은 계층적 구조로 이루어진 데이터를 가공하여 상위 데이터를 기준으로 하위 데이터의 통계량을 만드는 것에서 시작하였다. 각 대출자마다 주어진 다수의 대출, 그리고 각 대출마다 주어진 월별 정보를 활용하였고, 다중 대체법을 이용해 결측치를 대체했다. 그리고 불균형 데이터 문제를 해결하고자 5-fold 교차 검증(cross validation)을 할 때마다 학습 데이터에 오버샘플링을 했다. 모델은 logistic, boosting, DNN을 사용했고, 각 모델을 최적화한 뒤에 두 가지의 간단한 앙상블 모델을 만들었다. 이러한 과정을 통해 신용등급을 만들어 등급별 차이를 비교해 보았고, 대출 간 기간 차이가 클수록 신용등급이 높다는 점을 발견할 수 있었다

주요어 : 신용평가, 다중 대체법, 앙상블 모델
학 번 : 2017-21680

목 차

표 목 차

그 림 목 차

제 1 장 도입

최근 은행에 가지 않아도 카카오, 토스 등의 어플을 이용해 간단하게 자신의 신용등급을 확인할 수 있게 되었다. 이렇게 간편해진 신용등급 평가 기능 도입의 배경에는 대출 절차의 간소화와 대출의 접근성 증가에 따른 비대면 대출의 증가에 있다.

해외의 기업들은 다양한 채널을 이용한 소액 중심의 간편한 대출을 지원하고 있다. 특히 credit karma라는 회사는 신용평가 서비스를 통해 다양한 금융상품과 카드를 추천해주는 대출 상품 비교 서비스를 제공하며 많은 수익을 내고 있다. 한국의 경우에도 개인신용정보 이동권 도입을 통한 본인신용정보관리업, 마이 데이터 산업의 도입이 예정되어 있어 한국판 credit karma의 등장이 기대된다. 실제로 핀다, 핀셋, 비바리퍼플리카, 마이뱅크, 핀테크. 뱅크샐러드 등 다양한 관련 핀테크 업체들이 등장하고 있다. 대출 상품 추천 이외에도 기존의 신용조회업으로 정의 혹은 금지되었던 개인정보 자기결정권의 대리행사(프로파일링 대응권), 빅데이터 분석, 투자자문과 같은 신용평가가 필요한 서비스들이 마이데이터 산업으로 지정이 될 예정이다. 아직 관련 법안이 통과되지는 않았지만 통과된다면 이처럼 시장에 많은 변화가 있을 예정이다. 또한 신용평가에 금융 이력 뿐만 아니라 SNS 이용정보, 통신, 전기, 수도 등의 요금납부 정보, 주거정보, 쇼핑 이용실적 및 대금지불기록 등을 활용하게 되면서 금융 이력이 없는 금리단층의 신용평가 또한 가능하게 되었다. 미국의 경우, 실제로 다양한 비금융정보를 활용하는 FICO Score를 도입해 금융상품/차주 특성별 평가가 가능해져 신용평가의 고도화를 이룰 수 있었다. 국내의 신용평가사에서도 약 1000만명 정도로 추정되는 금리단층에 대한 신용 평가를 위해 비금융정보를 이용한 신용평가 모형 개발을 시도하고 있다.

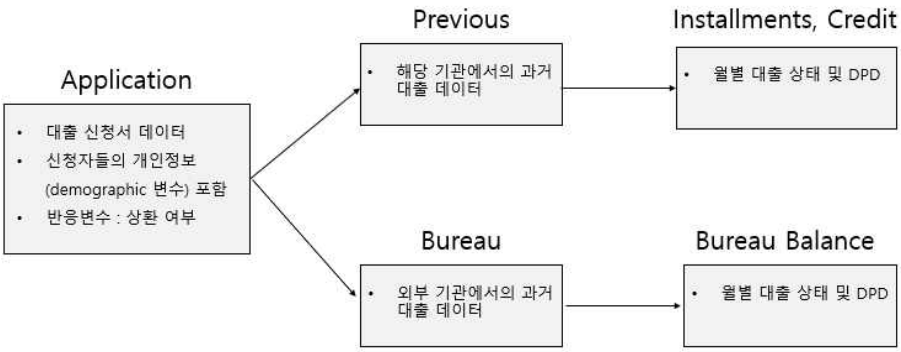
이러한 다양한 변화는 보다 많은 사람들이 일상생활 속에서 간단하게

대출을 접할 수 있게 할 것이고, 자연스레 신용평가의 필요성이 증대될 것이다. 이에 통계적 분석기법들을 이용해 직접 신용등급평가 모형을 만들어 보는 것을 목표로 하였다. 모형의 기본적인 아이디어는 데이터로부터 분류모형을 만들고 추정된 확률의 구간을 나누어 신용등급을 만드는 것이다. 특정 인물의 신용등급을 알고 싶다면 해당 인물의 추정된 확률이 들어가는 구간을 신용등급으로 책정한다. 또한 이를 위한 데이터와 그에 대한 전처리, 모델링에 대해서도 기재하고 최종 모형으로부터 의미있는 결과를 해석하고자 한다.

제 2 장 데이터

제 2.1 절 데이터 소개

사용한 데이터는 Home credit 은행에 대출을 신청한 사람들의 데이터로 대상들의 개인정보와 이전 대출정보를 포함한다. 반응변수는 상환여부로 대출을 신청한 사람들이 X일 이상(X : 명시되지 않음) 늦게 상환할 경우 1, 기간 내 상환할 경우 0인 값을 가지는 binary 변수로 분석의 목적은 classification이다. 데이터는 총 7개 파일(Application(train), Application(test), Previous, Bureau, Installments, Credit, Bureau balance)로 이루어져 있으며 2단계의 계층적(hierarchical) 구조를 가진다. Application은 대출 신청자들의 지원서 데이터로서 신청자가 Key인 메인 데이터이다. Application의 하위 데이터인 Previous, Bureau는 각각 Home credit, 그 외의 금융기관에서의 대출 신청 건에 대한 정보(신용융자 연체 일수, 남은 연체액, 정보 제공 여부, 신용융자 년수, 대출종류 등)를 포함하고, 대출 신청건이 Key인 데이터이다. 마지막으로 Installments, Credit, Bureau balance는 Previous와 Bureau의 하위 데이터로서 대출 신청건의 월별 대출 상태를 포함하고 있다. 즉 이 데이터들의 Key는 대출 신청건과 월(month) 이다. 또한 Installments는 현금대출일 경우, Credit은 카드관련 대출일 경우의 월별자료이다.



[그림 1]

위에서 언급한 각 데이터의 크기는 다음과 같다. Application의 학습 데이터와 시험 데이터의 개수는 각각 307,511과 48,744이고, 122개의 변수를 가지고 있다. Previous, Bureau는 각각 약 170만개의 과거 대출 정보를 포함하고, 변수 는 각각 37, 17개를 가진다. 마지막으로 월별 데이터인 Installments의 데이터 개수는 13,605,401개이고 Bureau Balance의 경우 27,299,925개이고, Credit은 3,840,312개이다.

데이터 종류	N	변수 개수
Application(train/test)	307,511 / 48,744	122
Previous/Bureau	1,670,214 / 1,716,428	37 / 17
월별데이터	13,605,401 / 27,299,925	8 / 3

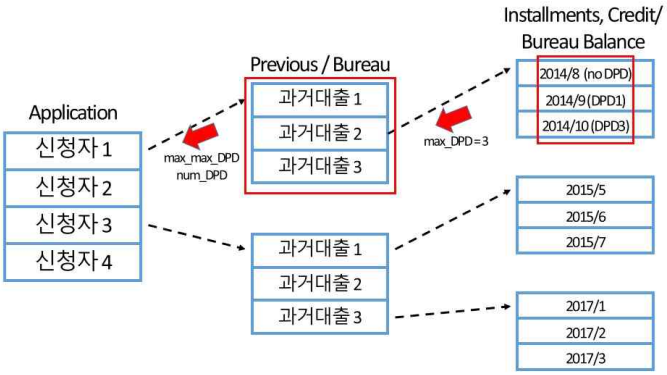
[표 1]

제 2.2 절 데이터 전처리

데이터가 계층적인 구조를 가지므로 하나의 데이터로 통합해야 할 필요가 있다. 이를 위해 상위 데이터의 key를 기준으로 하위 데이터들을 단계적으로 통합시켰다.

우선 대출 신청건마다 여러 개의 월별 대출 상태 데이터가 있으므로 월별 데이터 의 통계량을 만들어 대출 신청건 데이터에 통합했다. 예를 들면 과거 대출의 월 별 데이터마다 연체일(DPD)과 연체액이 있으므로 이들의 최댓값(max_DPD, max_연체액) 등의 통계량을 대출 데이터에 통합했다. 물론 통합할 때의 Key는 대출 신청 번호를 사용했다.

마찬가지로 대출 신청자 한 명마다 여러 개의 대출 신청 건이 있으므로 위와 같은 과정을 반복했다. 예를 들어 각 대출 신청 건은 앞선 과정에서 만들어진 max_DPD 칼럼을 가지고 있고, 각 대출 신청자마다 여러 개의 대출 신청 건이 있으므로 이들의 최댓값(max_max_DPD)을 계산해 대출 신청자 데이터에 통 합했다. 통합할 때의 Key는 대출 신청자 번호를 사용했다.



[그림 2]

이처럼 연체 일수의 최댓값의 최댓값(max_max_DPD) 뿐만 아니라 대출 횟수, 연체된 대출 횟수, 현재 시점에서 각 대출의 종료/진행 의 비율 변수, 모든 대출이 종료됐는지 여부, 대출 사이 일 수의 최솟값과 합, 연체 일수의 최솟값, 연체된 최대 금액, 각 대출마다 연체된 횟수의 최댓값, 납기일의 연장횟수의 최댓값, 최솟값, 개수, 신용대출의 여러 종류의 비율 등의 변수를 만들어 대출 신청자 데이터에 포함시켰다. 하지만 Previous(Home credit에서의 대출 신청 건)과 Bureau(외부 기관에서의 대출 신청 건)의 칼럼이 달라 하나의 통합된 대출 신청자 데이터를 만드는 것이 어려운 문제가 있었다. 두 데이터의 칼럼이 다르면 대출 신청자 데이터에서의 통계량, 즉 통합(join)되는 칼럼 또한 달라지기 때문이다. 이에 하나의 통합된 데이터를 만들기 위해 각 칼럼의 의미를 파악해 공통되지 않은 칼럼은 제외했다. 이처럼 상위 key를 기준으로 하위 데이터의 통계량을 만들어 상위 데이터에 통 합했다. 그 결과 총 196개의 새로운 변수를 Application, 대출 신청자 데이터에 생성할 수 있었다.

제 2..3 절 결측치

월별 대출 상태 데이터의 경우 결측치가 존재하지 않았지만 대출 신청자와 대출 신청건 데이터에는 많은 결측치가 존재했다. 우선 대출 신청건의 통계량을 만들어 대출 신청자 데이터에 통합하기 이전에 대출 신청

건 데이터의 결측치를 처리할 필요가 있었다.

칼럼 이름	결측치 개수
DAYS_CREDIT_ENDDATE	105553
DAYS_ENDDATE_FACT	633653
ANT_CREDIT_MAX_OVERDUE	1124488
ANT_CREDIT_SUM	257669
ANT_CREDIT_SUM_LIMIT	591780
ANT_ANNUITY	1226791

[표 2]

Bureau의 경우 ant_credit_max_overdue(최대 연체액), annuity(대출 년차) 각각의 결측치의 개수는 각각 1124488개, 1226791개였다. Bureau의 row 개수가 1716428개임을 고려할 때, 결측치가 굉장히 많으므로 해당 칼럼을 제거 했다. ant_credit_max_overdue(최대 연체액)의 경우, Bureau_balance(월별 대출 상태)로 부터 생성되는 최대 연체 금액 통계량으로 대체 가능했다.

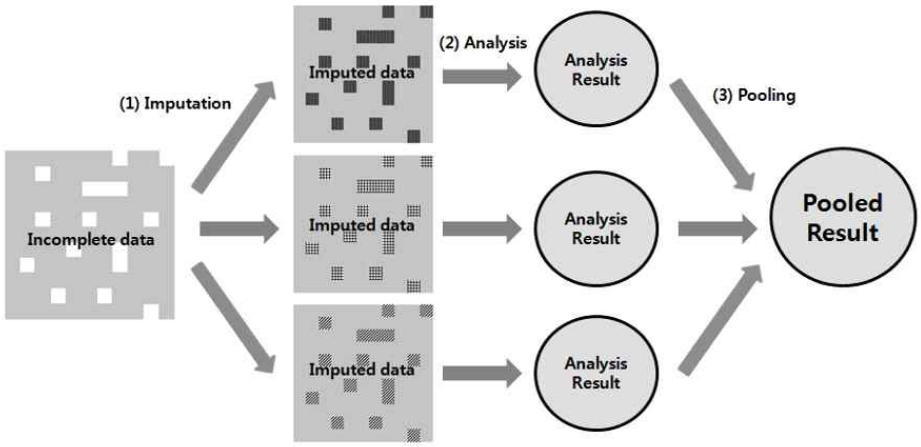
칼럼 이름	결측치 개수
AMT_ANNUITY	372235
AMT_CREDIT	1
AMT_DOWN_PAYMENT	895844
RATE_DOWN_PAYMENT	895844
RATE_INTEREST_PRIMARY	1664263
RATE_INTEREST_PRIVILEGED	1664263
NAME_TYPE_SUITE	820405

[표 3]

Previous에서 amt_down_payment(계약금), rate_down_payment(계약금 비율), rate_interest_primary(정규화된 이자율), rate_interest_privileged(정 규화된 이자율2), name_type_suite(대출 신청할 때, 동행한 사람)등의 경우, Previous의 row 개수가 1670214개임을 고려할 때, 결측치가 상당히 많으므로 해당 칼럼을 제거했다.

여기서 언급되지 않은 나머지 칼럼의 전처리는 칼럼들의 관계를 이용하거나 다중대체법(Multiple imputation)(Schafer, J. L. 1999.)을 이용해

대체했다. 이렇게 대출 신청건 데이터에 대한 결측치를 처리한 후에 대출 신청자 데이터에 통계량들을 결합했다.



[그림 3]

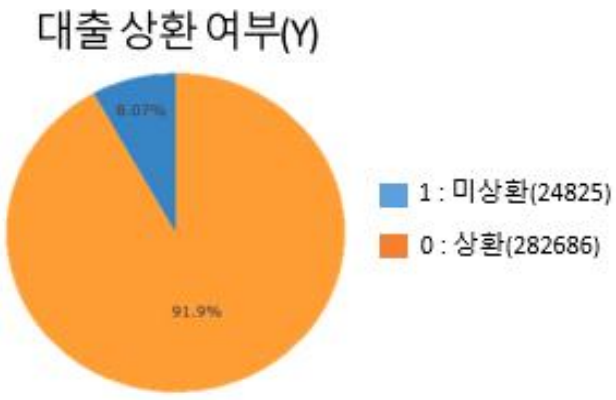
다중대체법은 시뮬레이션을 반복하여 누락된 데이터를 채워 넣는 방법이다. 구체적으로 시뮬레이션을 통하여 누락된 자료를 채운 데이터 셋을 3-10개 정도 만든다. 이들의 평균, 즉 몬테카를로 방법을 사용해서 누락 자료를 채운다. 이 후, 대체 결과를 바탕으로 신뢰구간을 계산할 수 있다. Application, 즉 대출 신청자 데이터의 경우 많은 칼럼이 결측치를 가지고 있었다. 우선 시험 데이터의 경우에 결측치 비율이 40% 이상인 변수가 49개가 있었는데 이를 활용할 경우 예측에 큰 차질을 줄 것이라고 판단해 해당 변수를 제거했다. 제거한 후에 학습 데이터의 경우에도 결측치 비율이 50% 이상인 칼럼이 몇 개 있어 추가적으로 제거하였다. 나머지 칼럼의 결측치는 변수들 간의 관계나 다중대체법(multiple imputation)을 이용해 대체했다.

결측률	Application_train	Application_test
0.5 이상	41개	29개
0.4 이상	49개	49개

[표 4]

제 2.4 절 불균형 데이터

반응변수는 상환여부로 대출을 신청한 사람들이 X일 이상(X : 명시되지 않음) 늦게 상환할 경우 1, 기간 내 상환할 경우 0인 값을 가지고, 0과 1의 비율은 9:1 정도로 불균형 데이터 문제를 가진다고 할 수 있다. 아무런 처리 없이 모형을 바로 적용하면 거의 모든 예측값이 0이 나오는 경우가 생길 수 있기 때문에 불균형 자료에 대한 처리가 필요하였다. 실제로 처리 없이 모형을 적합한 결과 정확도(Accuracy)는 비교적 높게 나왔지만 제2종 오류가 고려되는 F1 score나 Recall 등의 손실 함수를 고려했을 때, 성능이 좋지 않았다.



[그림 4]

이러한 불균형 데이터 문제를 해결하고자 크게 두 가지의 샘플링 방법을 고려하였다. 하나는 SMOTE(Synthetic Minority Over-sampling Technique)(Nitesh V., Chawla. et al., 2002)방법으로 오버샘플링을 적용해 Y=1인 데이터를 늘려 두 클래스의 비율을 맞춰주는 방식이고, 다른 하나는 SMOTE와 언더샘플링 기법 중 하나인 ENN(Edited Nearest Neighbor)(Donghai Guan et al., 2009)방법을 함께 이용해 비율을 조정하는 방법이다. 이후 설명할 교차검증(cross validation)(Richard R et al., 1983)을 통해 후자의 방법을 최종 모형에 적용했다. SMOTE(Synthetic Minority Over-sampling Technique) 방법은 오버샘플링 기법의 일종으로서

기존 소수 샘플들 사이의 내분점을 이용해 새로운 소수 데이터를 합성해 낸다. 구체적으로 소수 데이터 중 특정 데이터와 가장 가까운 이웃 사이의 차이를 계산하고, 이 차이에 0과 1 사이의 난수를 곱해 소수 클래스에 추가하는 방식이다. 이처럼 소수데이터들 사이를 보간해 오버샘플링을 하기 때문에 기존의 소수 데이터들 사이의 특성만을 반영하고 새로운 사례의 데이터 예측엔 취약할 수밖에 없다.

ENN(Edited Nearest Neighbours) 방법은 언더샘플링 기법의 일종으로 다수 클래스 데이터 중 가장 가까운 k개의 데이터가 모두 또는 다수 클래스가 아니면 삭제하는 방법이다. 즉 소수 클래스 주변의 다수 클래스 데이터는 사라진다.

제 3 장 학습

제 3.1 절 모델

분석에 사용한 기본모델은 심층신경망(DNN)(Donald F. Specht., 1991), Lasso logistic(Wang H, Xu Q, Zhou L, 2015), Gradient Boosting(Friedman, J. 2001.)이다. 교차검증을 이용해 각 모델을 조정(tuning)한 후, 두 개의 앙상블 (ensemble) 모델을 구축했다. 로지스틱 회귀분석(Logistic regression)은 독립변수와 종속변수 간의 관계를 계수(coefficient)에 학습한다. 이 모델을 학습할 때, 계수에 제약 조건을 추가해 과적합을 방지할 수 있고, 그 중 Lasso logistic 모형은 가중치의 절댓값의 합을 최소화하는 것을 제약 조건으로 한다.

심층신경망은 입력층(Input layer)과 출력층(output layer) 사이에 여러 개의 은닉층(hidden layer)들로 이뤄진 인공신경망(Artificial Neural network)이다. 심층 신경망은 일반적인 신경망과 마찬가지로 복잡한 비선형 관계식(non-linear relationship)들을 모델링할 수 있다.

주요 앙상블 알고리즘은 bagging(Breiman, L., 1996)과 boosting으로 나눌 수 있고, Gradient Boosting은 boosting 계열의 앙상블 알고리즘이다. f(x)에 대한 negative gradient에 적합(fitting)해서 다음 모델을 순차적으로 만들어 가는데, 이는 손실(loss)함수가 줄어들기 위한 f(x)의 방향이라고 볼 수 있다. 이 방향으로 새로운 모델을 순차 적으로 fitting한 후, 이전 모델과 결합하는 방법이 gradient boosting이다. 이를 통해 f(x)는 손실함수가 줄어드는 방향으로 업데이트 되었다고 할 수 있다.

Model1은 세 가지의 기본 모델을 서로 다른 샘플링된 데이터를 이용해 각각 학습시킨 후, 세 개의 추정 확률의 평균을 최종 추정 확률로 사용한다. 샘플링을 하는 이유는 종속 변수의 비율을 맞추기 위함이다. 구체적으로 대출 상환을 한 대출 신청자 중에 약 70000개를 샘플링 함과

동시에 오버샘플링 기법을 이용해 대출 상환을 하지 않은 신청자를 2000개에서 70000개로 늘렸다.

Model2는 Model1과 다르게 세 가지 기본 모델의 분석 결과가 서로 영향을 주도 록 구축했다. 먼저 샘플링 없이 심층신경망을 먼저 학습시킨 후, 실제로는 대출 을 상환했지만 심층신경망을 통해 추정된 종속 변수의 추정량의 분위수(quantile) 0.8 보다 큰 데이터를 제거했다. 즉 실제로는 대출을 상환했지만 상환하지 않는다고 추정되는 데이터를 제거한 것이다. 이 후, 이 데이터를 활용해 Lasso logistic 모델을 학습시킨다. 그리고 앞과 같이 데이터를 제거한 후, sampling을 이용해 종속변수의 비율을 맞춰준 다. 이를 이용해 Gradient Boosting으로 모 델을 학습시킨다.

[그림 5] Model2의 구조

제 3.2 모델

교차검증(validation)할 때, 각 모델의 모수(hyperparameter), 샘플링(sampling) 방법, 샘플링 비율을 사용했다. 또한 학습/껌증 데이터를 분리하기 이전에 샘플링을 하면 검증 과정에서 모수가 샘플링된 데이터에 맞춰져 과적합이 발생할 수 있다. 따라서 각 fold의 학습 데이터에만 샘플링을 시행했다. 다만, 학습 손실 함수를 계산할 때는 검증 데이터와 분포가 비슷한 기존의 학습 데이터를 이용했다. 왜냐하면 검증의 목적은 학습 데이터와 검증 데이터의 손실 함수의 크기를 비교해 과적합 여부를 판단하는 것이고, 따라서 두 데이터의 분포, 특히 종속변수의 비율에 크게 영향을 받기 때문이다. 사용한 손실함수는 AUC(Area under curve)(Jin H

uang et al. 2005)이다. AUC는 1-특이도(=specificity)와 민감도(= sensitivity)를 각각 x, y축으로 놓은 그래프인 ROC(Receiver Operating Characteristic curve) 의 밑면적 값이다. 고려한 샘플링 방법은 위에서 언급했듯이 SMOTE, SMOTE와 ENN을 동시에 사용하는 방법 2가지이고, 샘플링 비율은 SMOTE의 경우 1:1, 2:1, 4:1, 8:1, SMOTE와 ENN을 동시에 사용하는 방법의 경우 비율은 1:1로 고정했으나 데이터 개수를 다르게 설정했다. 교차검증을 통해 최종 선택한 샘플링 방법은 SMOTE와 ENN을 함께 사용하는 방법이며, 샘플링 비율은 두 클래스 모두 약 80000개 정도가 되도록 하는 방법이다.

제 4 장 결과 분석 및 논의

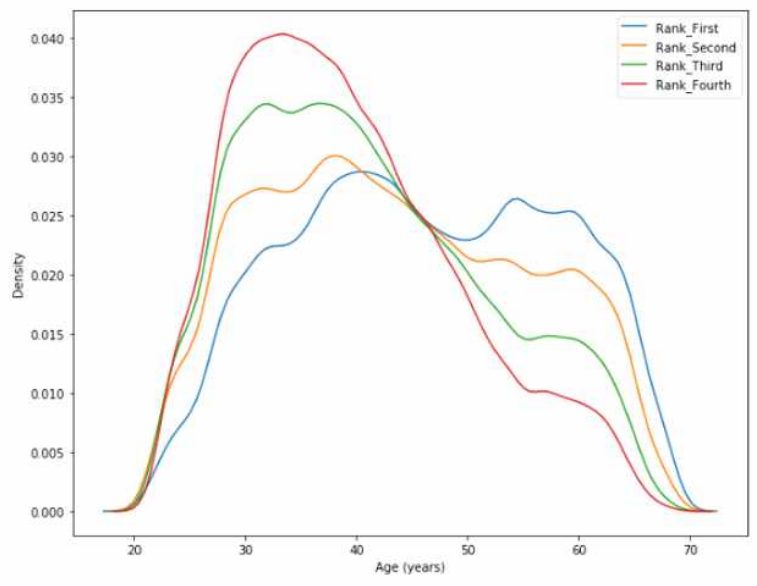
앞서 설명한 모델들의 학습 결과 및 테스트 결과는 다음과 같다. 전반적으로 학습 데이터에 과적합이 발생했으며 Gradient Boosting의 경우에 심한 과적합이 발생했다. 테스트 결과가 가장 좋은 모델은 Model1이며 이를 최종 모델로 사용 했다.

모형	학습 결과	테스트 결과
Lasso Logistic	0.78	0.72
Gradient Boosting	0.86	0.75
심층신경망	0.83	0.69
Model1	0.87	0.76
Model2	0.76	0.71

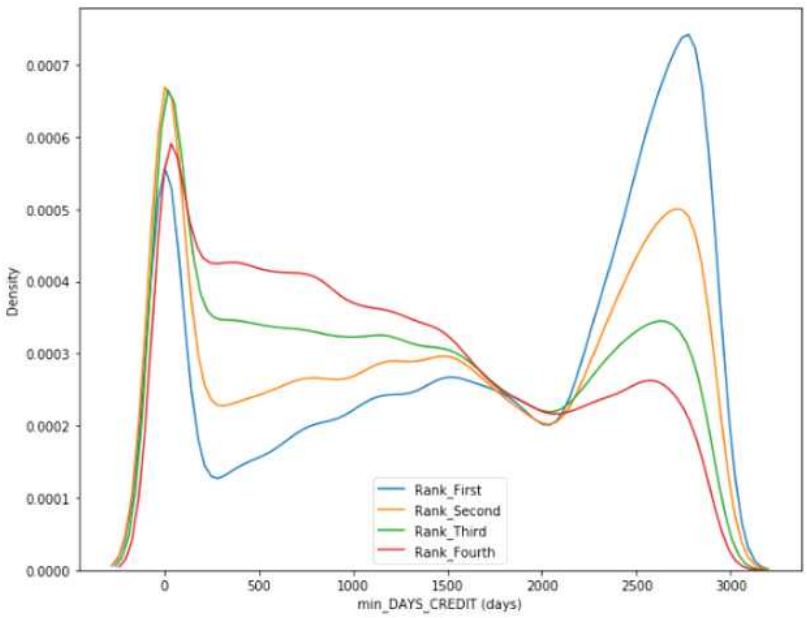
[표 5]

실제 여신 심사를 할 때, 신용평가 모형의 결과를 바탕으로 등급화 과정을 거치고, 이를 심사에 사용한다. Home credit 은행의 주 고객은 '저신용자'이기 때문에 신용등급을 1-10등급으로 나누는 것은 부적절하다고 판단해서 분위수 값을 기준으로 4개의 등급으로 나누어 제일 높은 등급부터 낮은 등급 구간을 만들었다. (rank_first, rank_second, rank_third, rank_fourth) 각 등급 간 차이를 알아보기 위해 특정 그룹의 등급 분포를 kernel density plot를 이용해 추정했다.

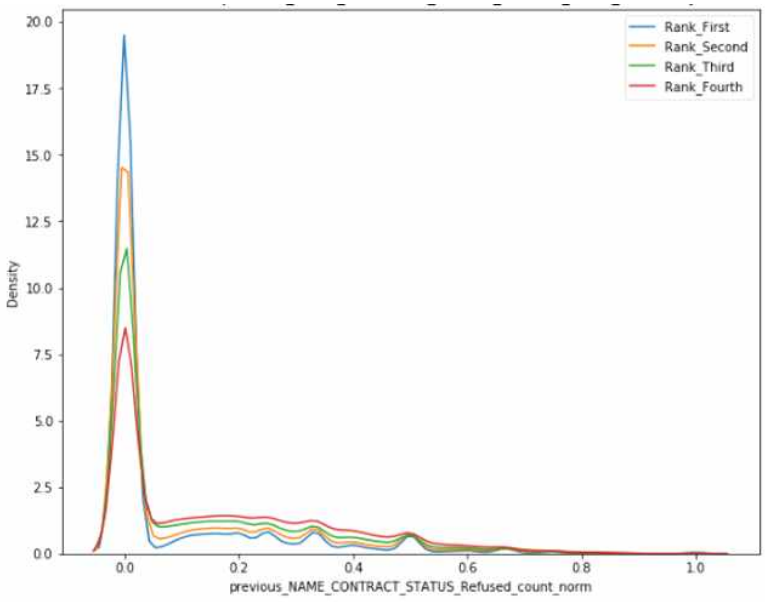
우선 나이가 많을수록 신용도가 높다는 것을 확인할 수 있었다. 또한 교육 수준이 높을수록 신용이 높다는 것을 확인할 수 있다. min_DAYS_CREDIT 변수는 각 대출마다 '이전 대출일과 대출 신청일과의 차이' 변수가 있고, 이들의 최솟값이다. 이 min_DAYS_CREDIT 변수가 클수록, 즉 이전 대출일과 대출 신청일간의 차이 일수가 클 수록 신용도가 높다는 것을 확인할 수 있다. 또한 이전 대출들의 통과, 취소, 거절된 것 중 거절의 비율이 낮을수록, 그리고 통과의 비율이 클 수록 신용이 높다는 것을 확인할 수 있다. 마지막으로 이전 대출 중에 서 이자율이 낮은 대출의 비율이 클 수록 신용이 높다는 것을 확인할 수 있다.



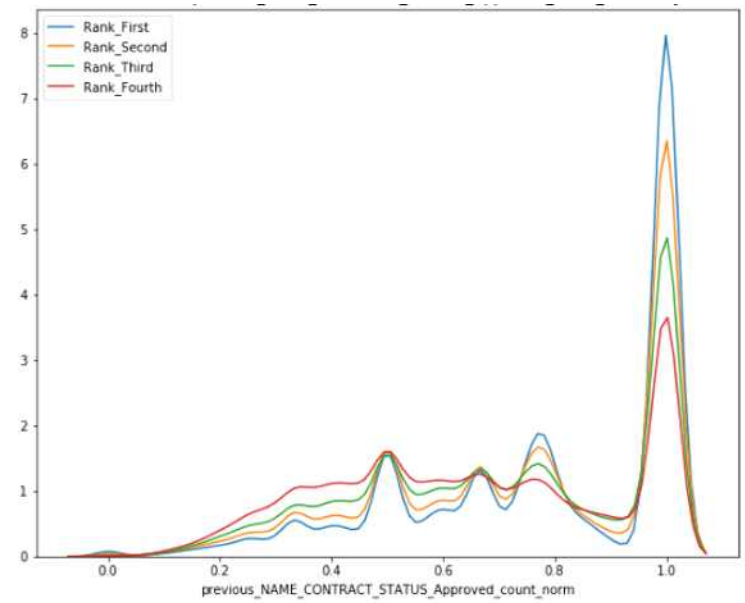
[그림 6] 나이 vs 신용등급



[그림 7] 대출간 기간차이 vs 신용등급



[그림 8] 이전 대출들의 거절 비율 vs 신용등급



[그림 9] 이전 대출들의 저금리 여부 vs 신용등급

참 고 문 헌

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research, 8(1), 3 - 15.

Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: Journal of artificial intelligence research 16 (2002), pp. 321 - 357.

Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM. 2016, pp. 785 - 794.

Marcus A Maloof. "Learning when data sets are imbalanced and when costs are unequal and unknown". In: ICML-2003 workshop on learning from imbalanced data sets II. Vol. 2. 2003, pp. 2 - 1.

Donghai Guan et al. "Nearest neighbor editing aided by unlabeled data". In: Information Sciences Volume 179, Issue 13, 13 June (2009), pp. 2273-2282, https://doi.org/10.1016/j.ins.2009.02.011

Richard R et al. "Cross-Validation of Regression Models". In: Journal of the American Statistical Association, Issue 387, 01 Mar (1983), pp. 575-583

Donald F. Specht. "A General Regression Neural Networ". In: IEEE TRANSACTIONS ON NEURAL NETWORKS. VOL. 2. NO. 6. NOVEMBER (1991) pp 568-576

Wang H, Xu Q, Zhou L (2015) Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. PLoS ONE 10(2): e0117844. https://doi.org/10.1371/journal.pone.0117844

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189-1232. Retrieved January 18, 2020, from www.jstor.org/stable/2699986

Breiman, L. Bagging predictors. Mach Learn 24, 123 - 140 (1996) doi:10.1007/BF00058655

Jin Huang et al. "Using AUC and accuracy in evaluating learning algorithms". In: IEEE Transactions on Knowledge and Data Engineering Volume: 17 , Issue: 3 , March 2005, pp 299-310

Abstract

Credit Evaluation Model Using Various Ensemble Models

Min-Kyu Jeong

Statistics

The Graduate School

Seoul National University

In this paper, we created a personal credit rating model to predi ct who would be overdue by more than X days. The process b egan by processing the hierarchical data to generate statistics o f the lower data based on the upper data. Multiple loans were given for each lender, and monthly information was given for e ach loan, and multiple imputation was used to replace missing values. And to solve the unbalanced data problem, we oversamp led the training data every 5-fold cross validation. We used logi stic, boosting, and DNN model, and after optimizing each model, we created two simple ensemble models. Through this process, we made credit ratings and compared the differences between g rades.

keywords : credit rating, Multiple imputation, Ensemble
Student Number : 2017-21680