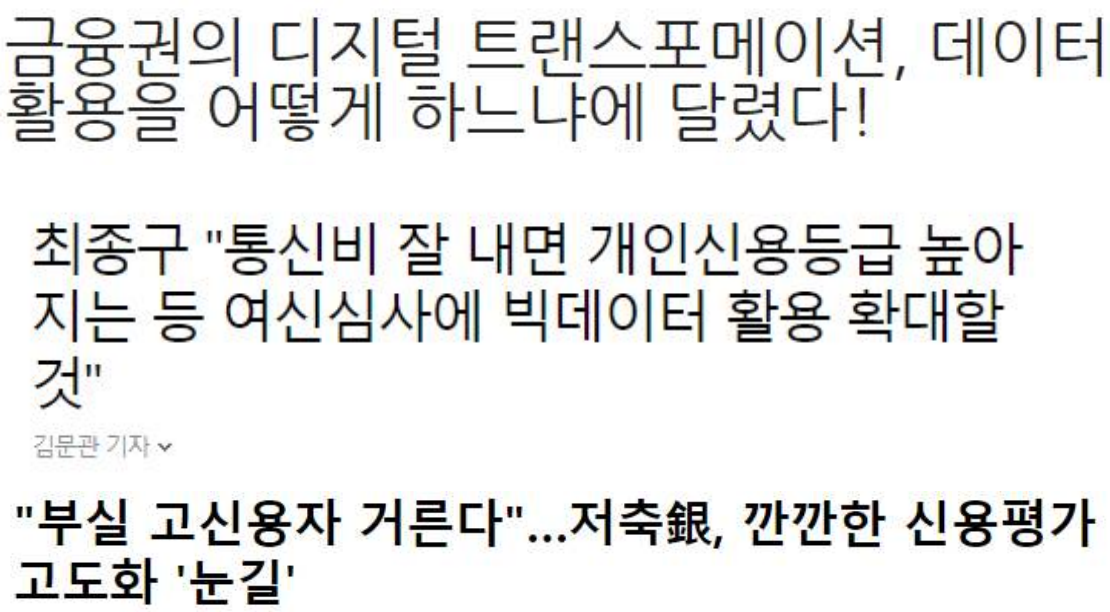
정확한 신용 평가를 위한 머신러닝 모델링

Ⅰ. 서론

1. 분석 배경 및 목적

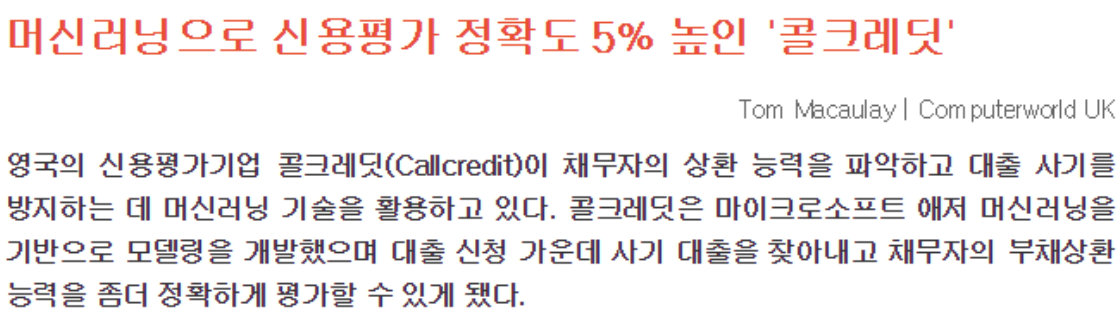

금융권의 디지털 트랜스포메이션, 데이터 활용을 어떻게 하느냐에 달렸다!

최종구 "통신비 잘 내면 개인신용등급 높아지는 등 여신심사에 빅데이터 활용 확대할 것"

김문관 기자

"부실 고신용자 거른다"...저축銀, 깐깐한 신용평가 고도화 '눈길'

4차 산업 혁명의 흐름과 함께 현재 우리 나라 금융권에서도 '디지털 트랜스포메이션'을 위해 빅데이터 등의 4차 산업 혁명 핵심 기술을 도입하는데 열중하고 있다. 이러한 과정에서 개인의 신용을 평가하는 방법에 빅데이터를 활용하는 기업들이 점차적으로 증가하고 있는 추세에 있다.


머신러닝으로 신용평가 정확도 5% 높인 '콜크레딧'

Tom Macaulay | Computerworld UK

영국의 신용평가기업 콜크레딧(Callcredit)이 채무자의 상환 능력을 파악하고 대출 사기를 방지하는 데 머신러닝 기술을 활용하고 있다. 콜크레딧은 마이크로소프트 애저 머신러닝을 기반으로 모델링을 개발했으며 대출 신청 가운데 사기 대출을 찾아내고 채무자의 부채상환 능력을 좀더 정확하게 평가할 수 있게 됐다.

실제로 영국의 신용평가 기업 콜크레딧은 머신러닝 기술을 활용한 신용평가 방법의 도입으로 신용평가의 정확도는 5% 향상 시킬 수 있었다. 따라서 머신러닝을 활용한 신용평가 기법이 얼마나 정확하게 고객을 분류해 낼 수 있는지에 대해 실험해보고자 했고 어떠한 요소가 신용 상태를 정확하게 분류하는데 중요한 영향을 미치는지 알아보고자 했다. 또한, 이러한 분류 기법의 모델링은 신용평가 뿐만 아니라 질병 예측 등 분류 기법이 적용되는 다양한 분야로의 확장이 용이하기 때문에 신용 평가에 사용되는 데이터 셋을 확보하고 머신러닝 기법을 활용해 정확한 신용평가 모델링을 수행하게 되었다.

2. 데이터 출처 및 소개

1) 데이터 출처 : 분석에는 UCI에서 제공하는 '**German Credit Data**'를 이용하였다.
(https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29)

2) 데이터 소개 : **총 1,000 건**의 개인 정보가 담긴 신용 평가 데이터 셋으로 **반응 변수는 신용 등급 (양호 또는 불량)**이며 **20개의 독립 변수**로 구성되어 있다. 변수들의 세부적인 설명은 아래와 같다.

(1) Status of existing checking account (예금 계좌 상태) : 현재 이용 중인 예금 계좌 잔액 (범주형)

(2) Duration in month (기간) : 대출을 받은 이후 소요된 기간 (월 기준) (수치형)

(3) Credit history (과거 신용 정보) : 과거 신용 대출 이력 및 정해진 기간 내 상환 여부 (범주형)

(4) Purpose (신용 대출 목적) : 신용 대출을 받은 목적 (범주형)

(5) Credit amount (신용 대출 금액) : 총 대출을 받은 금액 (수치형)

(6) Savings account/bonds (저축 예금/채권) : 저축 예금 및 채권 계좌의 잔액 (범주형)

(7) Present employment since (재직 기간) : 현재 직장에서 근무한 기간 (범주형)

(8) Installment rate in percentage of disposable income (가처분 소득 대비 적금 비율) (수치형)

(9) Personal status and sex (결혼여부 및 성별) (범주형)

(10) Other debtors / guarantors (여타 채무 및 채권) : 다른 채무 및 채권의 존재 여부 (범주형)

(11) Present residence since (거주 기간) : 현재 거주지에서의 거주 기간 (수치형)

(12) Property (재산) : 부동산 등 재산의 보유 여부 (범주형)

(13) Age in years (나이) (수치형)

(14) Other installment plans (여타 적금) : 다른 적금 계좌의 존재 여부 (범주형)

(15) Housing (주거 형태) : 현재 거주지의 주거 형태 (범주형)

(16) Number of existing credits at this bank (해당 은행 신용 계좌 개수) (수치형)

(17) Job (직업) (범주형)

(18) Number of people being liable to provide maintenance for (부양 가족 수) (수치형)

(19) Telephone (휴대전화 소유 여부) (범주형)

(20) foreign worker (외국인 노동자 여부) (범주형)

(21) Credit (신용 등급) : 양호 또는 불량의 이진 변수로 이루어진 반응 변수 (범주형)

Ⅱ. 본론

1. 분석 목적

이번 모델링의 최종 목적은 생성한 모델을 바탕으로 고객의 신용 평가(양호 또는 불량)를 **정확하게** 맞추는 것이다. 따라서 **이진(Binary) 데이터를 정확하게 분류**하는 다양한 기법을 적용해 보기로 했다.

2. 분석방법 소개

분류 기법에 적용할 수 있는 다양한 머신러닝 기법을 활용 했고 사용한 분석 기법은 다음과 같다.

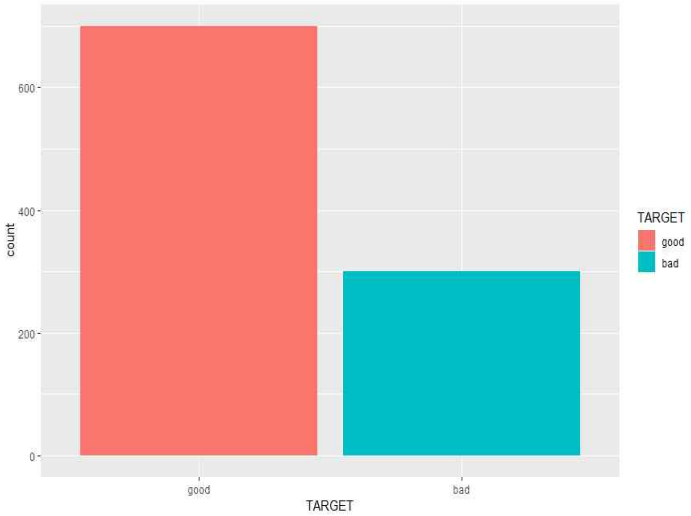
로지스틱 회귀모형(Logistic Regression), SVM(Support Vector Machine), 랜덤포레스트(Random Forest), GBM(Gradient Boosting Model), XGB(eXtreme Gradient Boosting)

위의 5가지 분류 기법을 적용한 뒤 **가장 정확도가 높은 모형**을 최종 모형으로 선정했다.

3. 탐색적 데이터 분석(EDA)

1) 반응 변수 분포 확인

: 분류 하고자 하는 변수의 CLASS가 균형 잡힌 분포인가?



총 1,000개의 관측치 중 신용 평가 변수의 값 (양호 or 불량)을 기준으로 분포를 살펴보았다.

양호(good)에 속하는 관측치는 **700개**, 불량(bad)에 속하는 관측치는 **300개**로 이루어져 있는 것을 확인할 수 있었다.

신용등급이 불량한 사람 대비 양호한 사람의 빈도수가 2배 이상이므로 분석 과정에서 **불균형 분포 대한 고려**가 필요해 보인다.

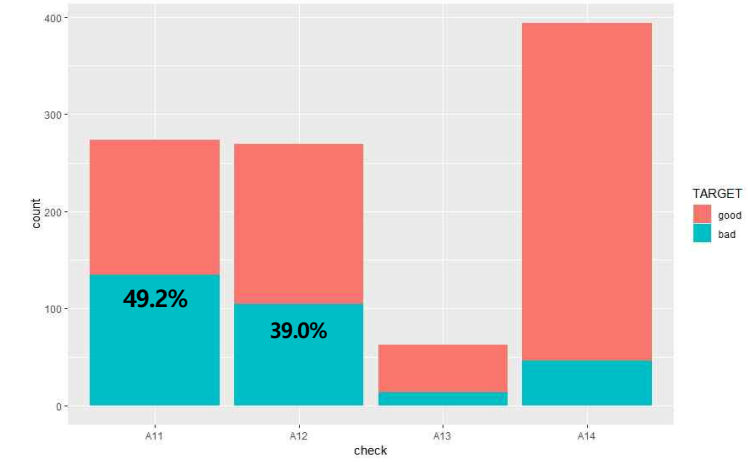
2) 독립 변수에 따른 반응 변수의 비율 확인

: 독립 변수마다 반응 변수에 미치는 영향이 다르지 않을까?

: 다르다면 어떠한 분포를 가지고 있고, 어떤 변수가 영향력이 크다고 할 수 있을까?

따라서 모든 독립 변수에 대해서 반응 변수의 비율을 알 수 있는 탐색적 자료 분석을 시행해 보았고, 그 중 반응 변수의 비율이 명확하게 구별이 되는 몇 가지 변수를 찾아낼 수 있었다.

(1) Status of existing checking account (예금 계좌 상태)



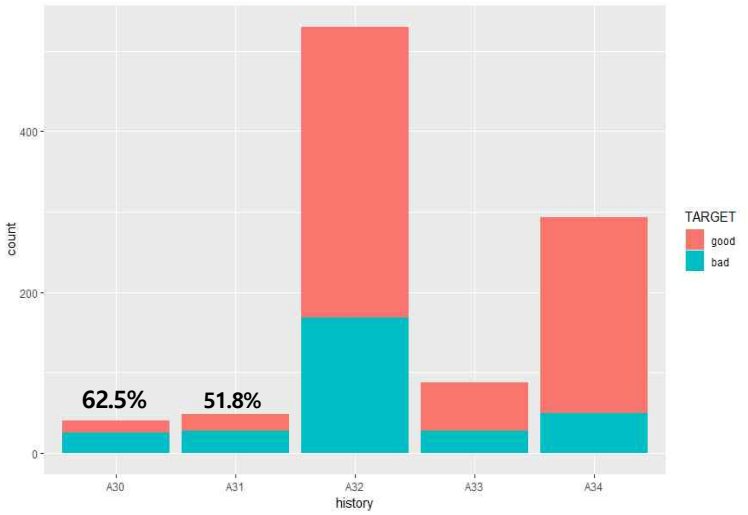
- 변수 설명

최소 1년 이내의 급여 할당량
A11 : 0 DM 미만
A12 : 0 DM 이상 200 DM 미만
A13 : 200 DM 이상
A14 : 예금 계좌 없음

- 그래프 설명

예금 계좌 상태에 따른 신용 평가 상태의 비율을 살펴보았을 때, 최소 1년 이내의 급여 할당량이 0 DM 미만인 첫 번째 그룹에서 신용 등급이 불량한 사람의 비율이 많은 것을 알 수 있었다.

(2) Credit history (과거 신용 정보)



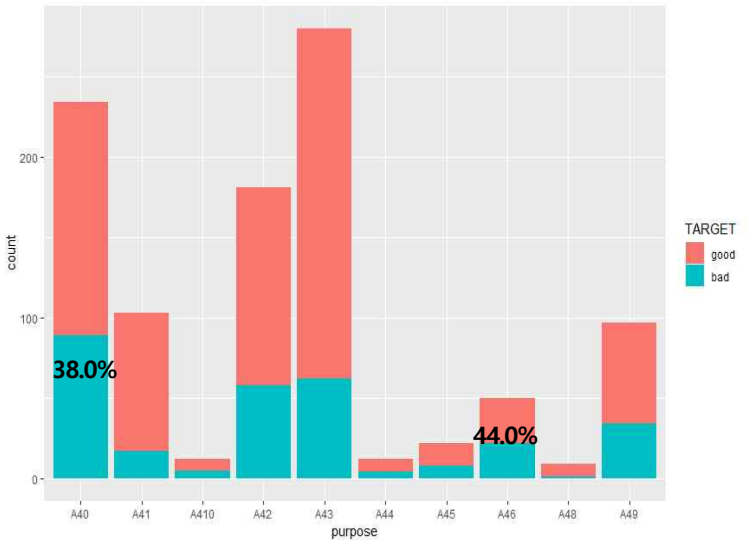
- 변수 설명

A30 : 이력 없음
A31 : 당행의 신용 대출 모두 상환
A32 : 현재까지 상환분 상환 완료
A33 : 연체 사실 존재
A34 : 중요 계정 / 타행 신용대출 존재

- 그래프 설명

과거 신용 정보를 살펴보았을 때, 과거 신용 정보에 대한 이력이 없거나 당행의 신용 대출을 모두 상환한 경우에 신용 등급이 불량한 사람의 비율이 상대적으로 많았다. 하지만 두 경우 모두 관측치 자체가 적은 경우에 속해서 이 부분에 대한 고려를 해야 한다.

(3) Purpose (신용 대출 목적)



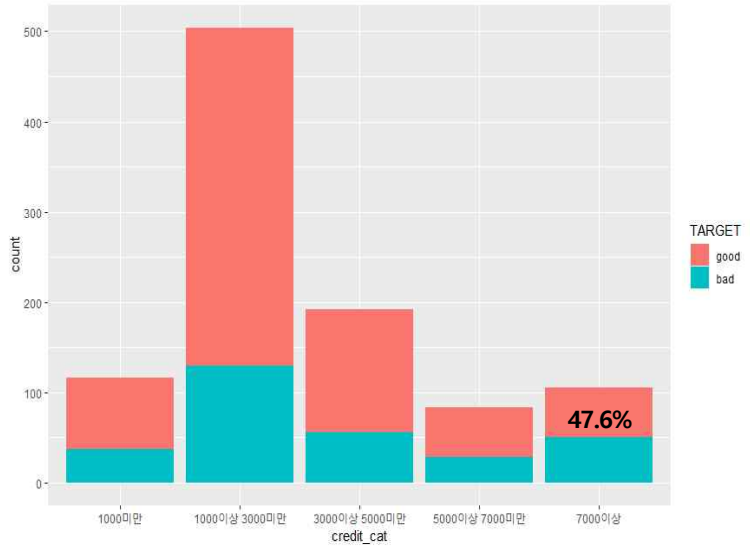
- 변수 설명

A40 : 신차 구매
A41 : 중고차 구매
A42 : 가구류 구매
A43 : 라디오/TV 구매
A44 : 가전기기 구매
A45 : 수리비
A46 : 교육비
A47 : 휴가비
A48 : 재훈련비
A49 : 비즈니스
A410 : 그 외

- 그래프 설명

신용 대출 목적을 기준으로 반응 변수를 살펴 보면, 신차 구매와 교육비 항목에서 신용 평가가 불량인 사람이 상대적으로 많이 분포하고 있음을 알 수 있다. 또한 휴가비 항목(A47)에 해당되는 관측치가 없는 것으로 보아 변수를 제거해 줄 필요성을 알 수 있었다.

(4) Credit Amount (신용 대출 금액)

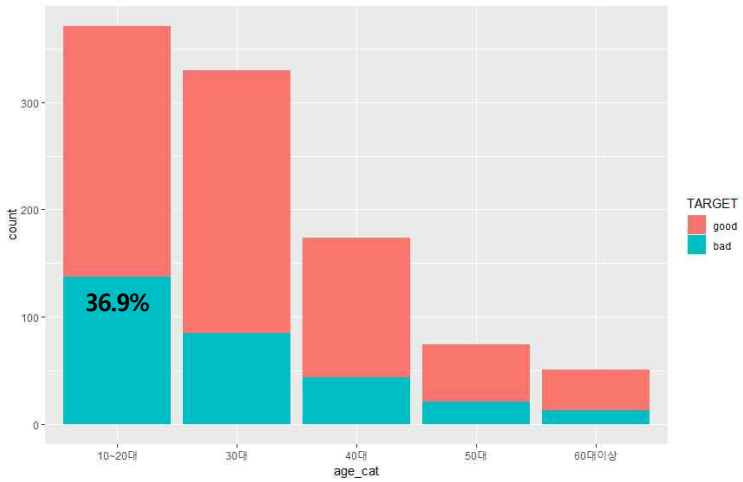


- 그래프 설명

연속형 변수이기 때문에 비율 분포를 제대로 확인할 수 없었다. 따라서 이를 범주화 하는 작업을 했고 그 결과 왼쪽과 같은 그래프를 얻을 수 있었다.

1000미만의 소액 대출을 제외하고 신용 대출 금액이 높을수록 '신용 불량'의 비율이 높아졌고, 7000이상의 범주에서는 47.6%라는 높은 비율을 가지는 것을 알 수 있었다. 따라서 대출 금액에 따라 신용 평가에 영향을 미칠 수 있을 것이라는 추측을 해볼 수 있었다.

(5) Age (나이)



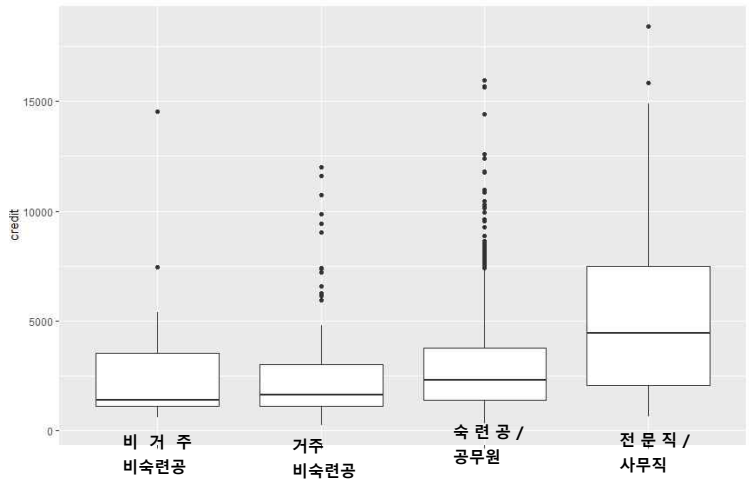
- 그래프 설명

마찬가지로 연속형 변수이기 때문에 이를 범주화 하는 작업을 했고 그 결과 왼쪽과 같은 그래프를 얻었다.

10-20대 그룹에 속하는 범주에서 가장 높은 '신용 불량' 비율이 있다는 것을 알 수 있었고 이러한 수치는 나이대가 증가할수록 상대적으로 감소하는 것을 알 수 있었다. 따라서 나이 변수 역시 신용평가에 있어서 중요한 변수로 작용할 확률이 높을 것이라고 예상할 수 있었다.

3) 독립 변수 간 분포 교차 확인

- Job (직업) & Credit amount (신용 대출 총 금액)



- 설명

범주형 변수와 수치형 변수를 두 가지를 조합하여 Boxplot으로 분포를 살펴보았다. 특정 범주 별로 서로 다른 분포를 갖는 수치형 변수가 있다면, 표준화 또는 범주화 작업이 필요하다고 판단했기 때문이다.

그중에서 직업 변수와 신용 대출 총 금액 변수 간의 관계에서 이러한 사항을 찾아 낼 수 있었고, 신용 대출 총 금액 변수를 범주화 해주는 작업을 했다.

4. 모형 적합 과정

(1) 로지스틱 회귀 모형 (Logistic Regression)

Stepwise 변수 선택법을 이용해서 모형을 생성하고 로지스틱 회귀 모형을 적합한 결과 총 14개의 변수가 분류에 사용되었다.

로지스틱 회귀 모형의 성능을 알아보기 위해 Cost Matrix를 그린 후 분류기의 성능을 나타내는 여러 측도 값을 계산하여 비교해 보았다.

예측값 / 실제값	양호	불량
양호	660	40
불량	174	126

- 선택된 변수

1. cheking account
2. duration in month
3. credit history
4. purpose
5. credit amount
6. savings account
7. installment rate
8. personal status
9. other debtors
10. present residence
11. other installment
12. housing
13. telephone
14. foreign

- 분류기의 성능 파악

1. Accuracy : 0.786
2. Miss-classification Rate : 0.214
3. Specificity : 0.42
4. Sensitivity(=Recall) : 0.94
5. Precision : 0.79
6. F1 score : 0.86

전반적으로 높은 예측률을 보이는 것을 알 수 있었다. 특히 Precision과 Recall의 조화평균 값인 F1 score가 0.86으로 안정적이고 높은 예측률을 보인다고 할 수 있다. 하지만 이보다 더 높은 성능을 가지는 모형에 대해 알아보기 위해 다른 알고리즘을 적용해 보기로 했다.

(2) 서포트 벡터 머신 (SVM)

서포트 벡터 머신의 Cost Parameter를 수정해가며 최적 값을 찾고 이를 통해 얻어낸 최적 모형을 데이터 분류에 적용했다.

최종적으로 Cost = 0.1 의 최적 값을 얻어 낼 수 있었고 이를 바탕으로 모형을 구축해서 데이터를 분류한 결과를 나타내는 Cost Matrix는 아래와 같다.

예측값 / 실제값	양호	불량
양호	642	149
불량	58	151

- Cost parameter에 따른 에러율

- 0.001 0.300
- 0.01 0.301
- **0.1 (최적)** 0.244
- 1 0.251
- 10 0.250
- 50 0.250

- 분류기의 성능 파악

1. Accuracy : 0.793
2. Miss-classification Rate : 0.207
3. Specificity : 0.722
4. Sensitivity(=Recall) : 0.812
5. Precision : 0.917
6. F1 score : 0.861

로지스틱 회귀 모형보다 더 높은 예측률을 얻었음을 알 수 있다. 정확도(Accuracy)는 물론 나머지 다섯 개 값이 모두 더 높은 수치를 기록했다. 비록 로지스틱 회귀 모형처럼 변수 선택에 대한 정보를 얻을 수 없긴 하지만 더 정확한 예측률을 목표로 한 분석이었기 때문에 유의미한 결과를 얻었다고 할 수 있다.

(3) 랜덤 포레스트 (Randomforest)

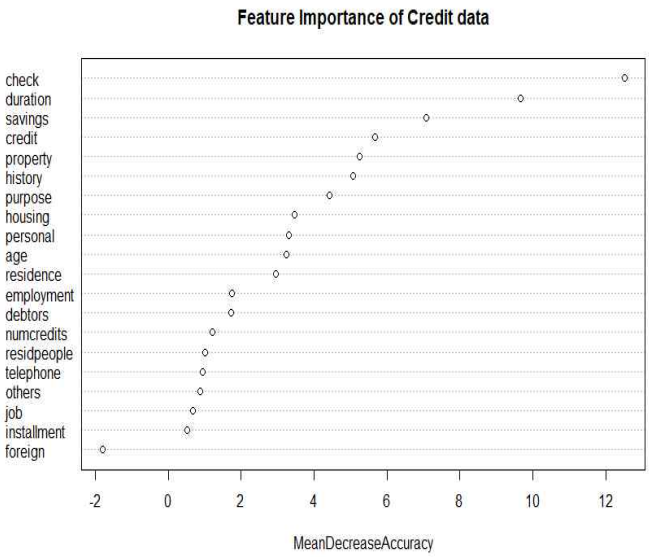
분류 문제를 정확하게 수행하는 데 있어서 변수 간의 상관관계에 영향을 적게 받고, 보다 안정적인 모형을 생성해 낼 수 있는 트리 앙상블 모형을 적용해 보았다.

이 과정에서 train, test set 분리(70:30 비율)를 통해 과적합이 일어나는 것을 방지하는 과정을 거쳤고, 배깅(bagging) 기법 기반의 랜덤포레스트를 적용해 일반화 성능을 더욱 향상시킨 모형이라고 할 수 있다. 조정할 수 있는 parameter인 트리 수(ntree)를 조정하면서 최적의 값을 도출했고, 150이 F1-measure를 가장 최대로 만들어주는 최적 값임을 알 수 있었다.

최적의 parameter를 통해 모형을 적합하고 test set을 통해 확인해 보았을 때, 아래의 Cost Matrix와 같은 결과를 얻을 수 있었다.

또한 Feature Importance plot을 생성해본 결과는 그 아래의 그래프와 같다.

예측값 / 실제값	양호	불량
양호	183	60
불량	18	39



- Feature Importance 순위

1. Status of existing checking count
2. Duration in month
3. Savings account/bonds
4. Credit amount
5. Property

예금 계좌 잔액, 신용 대출 기간, 저축 예금/채권, 신용 대출 총액, 자산 순으로 모형에서 높은 설명력을 가지는 것을 알 수 있었다.

위의 5개의 변수가 대표적으로 분류의 정확도를 크게 향상 시켜준 중요한 변수 였고, foreign 변수를 제외한 나머지 변수들도 정확도 향상에 도움이 된 것으로 나타났다.

foreign 변수는 오히려 정확도를 낮추는 것을 확인할 수 있는데, 이에 대한 처리가 필요할 것으로 판단된다.

- 분류기의 성능 파악

1. Accuracy : 0.757
2. Miss-classification Rate : 0.243
3. Specificity : 0.724
4. Sensitivity(=Recall) : 0.764
5. Precision : 0.920
6. F1 score : 0.835

앞 선 두 개의 모형 대비 정확도(Accuracy)가 다소 하락했으나, 다른 성능 측도는 크게 감소하지 않은 것으로 나타났다. 랜덤포레스트가 가진 최대 장점인 일반화 성능 향상(Variance 감소)을 고려해 보았을 때, 이 정도의 분류 정확도 감소는 충분히 감수 할 만 하다고 판단된다.

특히 SVM과 비교해 보았을 땐, 랜덤포레스트에서는 변수의 중요도를 추가적으로 알 수 있어 어떤 변수가 모형에서 중요하게 쓰였는지를 판단할 수 있다. 따라서 모형을 설명할 수 있는 요소를 갖춘 랜덤포레스트 모형을 데이터에 적합 하는 것이 나쁘지 않다고 판단된다.

(4) 그레디언트 부스팅 (GBM)

랜덤포레스트는 배깅(bagging) 방법의 트리 앙상블 모형이라면 그레디언트 부스팅은 부스팅(boosting)기법이 적용된 트리 앙상블 모형이다.

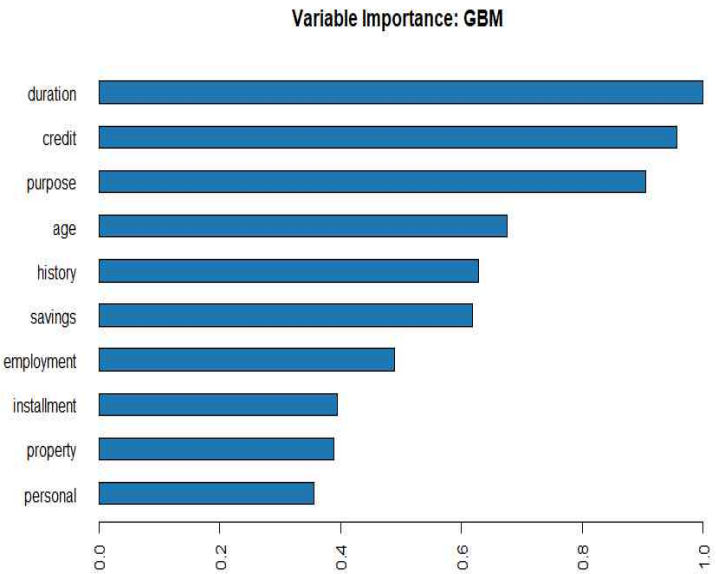
배깅이 분류기를 병렬적으로 학습해 개별 분류기에서 나온 결과를 voting 하는 형식으로 작동한다면, 부스팅은 이전 분류기의 오분류 된 관측치에 집중해서 다음 분류기가 오분류된 관측치에 가중치를 두는 방식으로 순차적으로 학습해 나가는 것이 특징이다.

따라서 보다 높은 정확도를 가진 분류 모형을 생성해 낼 수 있어서 최근 데이터 분류 문제에서 자주 쓰이는 알고리즘이다.

비록 크지 않은 데이터 양(1,000개)이지만 조금 더 높은 정확도를 얻을 수 있지 않을까 하는 생각에 5-fold Cross validation을 함께 적용해 보았고 추가적으로 Extreme gradient boosting(XGB) 알고리즘도 적용해보았다.

결과적으로 GBM모형과 XGB모형 모두 비슷한 성능을 나타냈고, 이는 데이터 셋의 크기 때문이라고 판단되었다. (XGB모형은 GBM모형 보다 조금 더 빠른 처리를 가능하게 해주고 penalty를 통해 과적합을 방지해준다.)

따라서 조금 더 단순한 모형인 GBM모형이 주어진 데이터에 가장 적합한 알고리즘이라는 판단을 내렸고, GBM모형을 통해 얻은 Feature Importance는 아래와 같다.



- Feature Importance 순위

1. Duration in month
2. Credit amount
3. Purpose
4. Age
5. Credit history

신용 대출 기간, 신용 대출 총액, 신용 대출 목적, 나이, 과거 신용 대출 이력 순으로 모형에서 높은 설명력을 가지는 것을 알 수 있었다.

랜덤포레스트 모형에서 뽑혔던 중요 변수와 다소 차이가 있었으나, Duration in month와 Credit amount 변수가 중요 변수로 뽑힌 것은 동일 했다.

또한 앞선 EDA 과정에서 파악 했던 독립 변수들의 반응 변수 비율에 대한 설명 부분과도 일치하는 결과를 얻을 수 있었다.

- 분류기의 성능 파악

1. Accuracy : 0.734
2. Miss-classification Rate : 0.266
3. Specificity : 0.713
4. Sensitivity(=Recall) : 0.973
5. Precision : 0.734
6. F1 score : 0.837

부스팅 모형도 랜덤포레스트 모형에 비해 정확도(Accuracy)가 다소 하락했으나, 다른 성능 측도는 대부분 상승했다. 특히 Recall 값이 크게 증가해 Precision 값이 감소 했음에도 F1 score를 향상 시킬 수 있었다.

이러한 특징은 데이터 셋이 더 확장되었을 때, 다른 모형에 비해 새로운 데이터를 더 정확하게 분류해낼 수 있는 특징으로 연결된다. 따라서 확장성의 측면에서 다른 모형들에 비해 우수하다고 할 수 있고, 결과적으로 최종 모형은 GBM 모형으로 결정했다.

Ⅲ. 결론

간단한 로지스틱 회귀 모형부터 최종 모형인 부스팅 모형까지 적합해본 결과 분류 문제에 있어서 더욱 향상된 알고리즘이라고 평가 받는 분류 기법들의 높은 정확도를 확인할 수 있었다. 또한 부스팅, 랜덤포레스트 모형에서 얻어낸 Feature Importance에 대한 정보는 데이터를 신용 평가 기준으로 분류하는 데 있어서 어떤 요소들이 신용 평가의 정확한 분류에 중요한 역할을 하는지 파악할 수 있는 중요한 정보를 제공해주었고, 실제로 신용 평가 모델에 적용한다면 이러한 요소에 더 집중한 모델을 개발하는 방법을 통해서 보다 향상된 신용 평가 모형을 개발할 수 있을 것이다.

위의 분석 과정에서 얻어낸 결과를 바탕으로 더 복잡한 변수가 추가된 신용 평가 모형에 적용하거나 아니면 다른 산업 분야의 분류 문제에 다른 인사이트를 통해서 모델링을 하는 등 다양한 방법으로 위의 분류 모형을 확장할 수 있을 것이라고 생각된다. 다만, 주어진 데이터 셋이 분류 문제에 적합하게 사용될 수 있도록 처리 및 가공 되어있다는 점을 감안한다면 위의 분석 결과처럼 높은 성능을 가지는 분류 모형을 생성하기 위해서는 변수의 표준화, 파생 변수 생성 등 목적 적합한 처리와 가공이 매우 중요할 것이다. 또한, 모형 비교에서 알 수 있었듯이 각각의 알고리즘이 가지고 있는 특성이 조금씩 상이하다는 것을 알 수 있었다. 따라서 분류 문제에 접근하는 분석가는 알고리즘에 대한 장,단점 역시 명확히 파악해야하고 다루는 데이터의 특성과 분석의 목적에 맞는 분석 방법을 선택하는 것이 매우 중요하다.