|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | 결측치\_tr | 결측치\_te | 제거이유 |  |
| 1 | data | 데이터 구분자(train, test)-사용X |  |  |  |  |
| 2 | CustomerId | 고객ID |  |  |  |  |
| 3 | Gender | 성별 | X |  |  |  |
| 4 | Age | 연령(년) | X |  |  |  |
| 5 | Education | 교육수준 | X |  |  |  |
|  |  |  |  |  |  |  |
| 6 | MaritalStatus | 결혼상태(결혼,이혼,독신 등) | X |  |  |  |
| 7 | FamilyCount | 가족수 | X |  |  |  |
| 8 | ChildCount | 자녀수 | X |  | 결혼상태와 가족수로 도출 가능 |  |
|  |  |  |  |  |  |  |
| 9 | IncomeType | 수입 형태(유형)(서비스,사업) | X |  |  |  |
| 10 | IncomeClass | 수입등급(년) | X |  |  |  |
| 11 | Occupation | 고객 직업군(영업,IT개발) | X |  | 너무 많은데 어케처리할까 |  |
| 12 | OrgType | 직장유형(일반,자영업,서비스) | X |  | 퇴직자 반영함 |  |
| 13 | EmployedYears | 근속년수(년) | 19318 |  | 얼마나 버는지 보려는거 -> x |  |
|  |  |  |  |  |  |  |
| 14 | HouseOwnedYN | 주택 소유 여부(Y:소유, N:없음) | X |  | 7.3/8 -> 상관계수 그리면 나갈 듯 |  |
| 15 | DwellingType | 주거형태(전세,월세) | X |  | 하은이 그래프 함 봐야됨 |  |
| 16 | HousingType | 주택형태(아파트,빌라) | X |  | 하은이 그래프 함 봐야됨 |  |
| 17 | ResidenceClass | 주거형태(양호,평균)????? | X |  | 이건 영향 있음. 가져가야 함 | 범주형 |
| 18 | HouseAge | 주택 나이(년)-결측 많음 | 61092 |  | 버릴거임~ |  |
| 19 | CarOwnedYN | 대출신청 차량 이외의 자동차 소유 여부(Y:소유, N:없음) | X |  | 영향 안보임. 상관계수 찾아서 날리기. | 범주형 |
|  |  |  |  |  |  |  |
| 1 | data | 데이터 구분자(train, test)-사용X |  |  |  |  |
| 2 | CustomerId | 고객ID |  |  | X |  |
| 3 | LoanId | 고객대출ID | X |  | X |  |
| 4 | Default | 대출 상환여부(1:대출 미상환, 0:정상 상환) | X | 결측 | 목표변수 |  |
| 5 | ActiveLoanYN | 대출신청 당시 활성 대출 유무(Y:있음, N:없음) | X |  | 유의할 듯. 근거좀 |  |
| 6 | LoanTyp | 대출유형(할부금융(장/단기)-소비자와 판매자, 금융회사 3자 간 계약. 오토론-소비자가 금융회사에서 대출받은 자금으로 자동차를 구입하고 대출 원리금을 금융회사에 상환하는 방식. 소비자와 금융회사 양자 계약 | X |  | 필요 없음. 근거 제시좀 |  |
|  |  |  |  |  |  |  |
| 7 | ApplWeek | 대출신청요일(월~금) | X |  | 제거. 근거: 해결책 제시 불가 |  |
| 8 | ApplHour | 대출신청시간대 | X |  | 제거. 근거: 해결책 제시 불가 |  |
| 9 | Accompany | 대출 신청시 동행자 | X |  | 제거. 근거: 해결책 제시 불가(이상치 제거용)  데이터 기록 과정에서 왜곡이 일어났음을 추측할 수 있다. 즉, 왜곡된 데이터 행임. |  |
|  |  |  |  |  |  |  |
| 10 | CarPrice | 차량 가격(원,1500만원~6100만원) | X |  | 16번과 높은 상관관계 – 제거  16번이 캐피탈에 필요한 정보에 더 적합함 |  |
| 11 | Deposit | 현금비율(%,0,10,20,30) | X |  | 제거. 밀도 그래프에서 의미 없음 확인. |  |
| 12 | LoanTerm | 대출기간(월,12,24,36,48) | X |  | 제거. 도메인 정보에서 금리와 상관관계가 있음을 알 수 있음. |  |
| 13 | LoanRemainTerm | 대출잔여기간(월) | X |  | 13-18까지 보류(상관계수 + 도메인) |  |
| 14 | InterestType | 금리유형(고정,변동,혼합) | X |  |  |  |
| 15 | InterestRate | 대출금리(%) | X | 결측 |  |  |
| 16 | LoanAmount | 대출금(원) | X |  |  |  |
| 17 | InstallAmount | 월납입액(원) | X | 결측 |  |  |
| 18 | LoanRemainAmount | 대출잔액(원) | X |  |  |  |
|  |  |  |  |  |  |  |
| 19 | HomeAddMatchedYN | 집주소 일치 여부(Y:일치, N:불일치) | X |  | 제거. 막대그래프 그려보면 의미없음 |  |
| 20 | WorkAddMatchedYN | 직장주소 일치  여부(Y:일치, N:불일치) | X |  | 제거. 막대그래프 그려보면 의미없음 |  |
| 21 | InquiryCount | 대출 전년도 문의건수(건) | X |  | 제거. 상자그림에서 평균치가 같다. 의미없다. |  |
|  |  |  |  |  |  |  |
| 22 | IdChangedYears | 신분증 변경 후 경과기간  (년,대출 신청 전 변경시) | 889 | 414 |  |  |
| 23 | InfoChangedYears | 고객정보 변경 후 경과기간  (년,대출 신청 전 변경시) | X |  |  |  |
| 24 | PhoneChangeYears | 휴대전화 변경 후 경과기간  (년,대출 신청 전 변경시) | 963 |  |  |  |
|  |  |  |  |  |  |  |
| 25 | ScoreA | 신용점수(A사).  평가사별 해당 대출건 평가 점수  (3사 또는 2사 중복) | 46157 | 결측 |  |  |
| 26 | ScoreB | 신용점수(B사).  평가사별 해당 대출건 평가 점수  (3사 또는 2사 중복) | 1845 | 결측 |  |  |
| 27 | ScoreC | 신용점수(C사).  평가사별 해당 대출건 평가 점수  (3사 또는 2사 중복) | 9626 | 결측 |  |  |
|  |  |  |  |  |  |  |
|  |  |  | 37 |  | Default\_Days\_60 | 대출 미상환 일수(최근 60일) |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

독신 00 10 20

f, c=0

결혼 20 31 42 53

f=c+2

미혼 10 21 32 43

사별 10 21 32 43

f=c+1

재직자

기존 대출 없을 때: 1443/22324 = 6.46%

기존 대출 있을 때: 5215/71917 = 7.26%

* 평균(8.2%)보다 낮으므로 중요인자 아님.

할부금융: 6027/69143 = 8.72%

오토론: 631/9979 = 6.32%

* 할부금융의 경우 평균(8.2%)보다 높으므로 중요인자일 가능성 존재. 할부기간 및 적용금리 종류 검토 필요.

할부기간 12개월: 1233/15705 = 7.85%

할부기간 24개월: 2501/31376 = 7.97%

할부기간 36개월: 1864/23643 = 7.88%

할부기간 48개월: 1849/23517 = 7.86%

* 평균(8.2%)보다 낮으므로 중요인자 아님.
* 적용금리 종류에 따른 영향력 확인 필요.

금리종류에 대한 전체평균: 6658/79122 = 8.41%

고정금리가 적용됐을 때: 696/12053 = 5.77%

변동금리가 적용됐을 때: 5639/62948 = 8.96%

* 변동금리는 연체에 영향을 미치는 인자로 확인.

퇴직자

기존 대출 없을 때: 1443/22324 = 6.46%

기존 대출 있을 때: 5215/71917 = 7.26%

* 평균(8.2%)보다 낮으므로 중요인자 아님.

할부금융: 726/13824 = 5.25%

오토론: 63/1295 = 4.86%

* 할부금융의 경우 평균(5.2%)보다 높으므로 중요인자일 가능성 존재. 할부기간 및 적용금리 종류 검토 필요.

할부기간 12개월: 144/2542 = 5.66%

할부기간 24개월: 258/4932 = 5.23%

할부기간 36개월: 194/3787 = 5.12%

할부기간 48개월: 193/3858 = 5.00%

* 12개월 및 24개월에서 평균(5.2%)보다 높기 때문에 검토 필요. 검정 결과 할부기간보다는 금리 종류가 더 큰 영향을 미치는 것으로 확인.

금리종류에 대한 전체평균: 789/15119 = 5.22%

고정금리가 적용됐을 때: 69/2428 = 2.84%

변동금리가 적용됐을 때: 681/11881 = 5.73%

* 변동금리는 연체발생에 영향을 주는 인자임을 확인. XGBoost 영향력 검토를 통해 이를 재확인하였음.

대인신용평가: 사회초년생, 주부 등 신 파일러(thin filer – 금융 거래 정보가 거의 없는 사람) 대상으로 신용평가를 하기 위해 비금융정보를 활용

통신비 납부 이력, 소셜미디어 활동 등 빅데이터를 통해 이들의 신용을 실제에 가깝게 평가해준다.

신한은행 – 배달부 데이터 분석, 음식점 특화모형 개발

KB국민은행 – 직장 변동 횟수, 부동산 데이터 자산 활용

우리은행 – 통신요금 납부 내역, 온라인 구매 정보 등 활용

네이버페이 – 네이버 스마트스토어 발품률 및 재구매율 활용

* 주어진 자료 중 근속년수, id, phone, info changed years, 거주형태 등이 해당되는 듯. 가중치를 부여하여 활용하려 했으나 영향력이 너무 적어 추가적인 자료 수집 필요함. (프로젝트 기간 내 불가능하다고 판단함.)

밀도그래프 분석 결과

재직자

AGE: 30대 후반부터 연체를 적게 한다. 나이가 많아질수록 연체 발생 감소

EmployedYears: 5년 이내에 연체가 많이 발생함. 해당 구간에 대한 특별관리 필요. 기간이 길어질수록 연체 발생 감소.

IdChangedYears: 11년 이후 연체 발생 감소. 10~13년 구간에서 shooting 발생으로 인해 일괄적용은 불가능하며, 기간 별 구분 적용 필요.

InfoChangedYears: 17년 이후 연체 발생 감소.

PhoneChangedYears: 3년 이후 연체 발생 감소

ScoreA: 0.45(450점) 이후 연체 발생 감소. 0.8(800점)부터 연체 발생빈도 낮음.

ScoreB: 0.45(450점) 이후 연체 발생 감소. 0.95(950점)부터 연체 발생빈도 매우 낮음.

ScoreC: 0.45(450점) 이후 연체 발생 감소. 0.95(950점)부터 연체 발생빈도 매우 낮음.

퇴직 재직자

AGE: 나이에 따른 연체발생과의 연관성이 확인되지 않음. 50~60대의 분포가 많다는 것을 확인.

IdChangedYears: 10~15년 사이에 분포가 집중되어있으며 12년 이후 구간에서 연체 발생 감소.

InfoChangedYears: 33년 이후 구간에서 연체 발생 감소.

PhoneChangedYears: 2년 이후 연체 발생 감소. 1년 전후 구간에 분포 집중.

ScoreA: 0.6(600점) 이후 연체 발생 감소. 0.9(900점)부터 연체 발생빈도 매우 낮음.

ScoreB: 0.5(500점) 이후 연체 발생 감소. 0.8(800점)부터 연체 발생빈도 낮음. 1.0구간에서 연체 거의 없음.

ScoreC: 0.5(500점) 이후 연체 발생 감소. 0.9(900점)부터 연체 발생빈도 매우 낮음.

결론: 재직자와 퇴직자 간에는 나이에 따른 연체 발생에 대한 차이가 존재한다. (영향 있음/없음) 또한, Score가 재직자/퇴직자 간 다른 기준이 적용됨을 알 수 있기 때문에 두 집단에 대해 별도의 영향력 분석이 필요하다.

결론2: 기존 신용점수(Score)에 대한 분석 결과, 비금융 데이터(대안신용평가)가 일부 반영된 것으로 판단된다. 겹치지 않는 독립적인 데이터 분석을 위해서는 파생변수 설정 혹은 추가적인 데이터 수집이 필요할 것으로 판단되나, 상관계수 분석 등을 통해 변수간 연관성을 찾지 못하였음. 분석자의 주관적인 판단에 따라 영향력을 구분짓고 적용시킴으로써 모델을 완성시켰으나, 완성도가 몹시 낮아 실용성이 낮음.