신용카드사 사례

이 번 사례는 약간 기술적인 내용을 다루어 보겠습니다. 김대리는 머신러닝 석사를 취득하고, 신용카드 사에 입사한 인재입니다. 김대리는 고객데이터 분석을 통해 신용카드 신청자의 연체가능성을 추정하고, 이에 따라 신용카드의 신용한도를 결정하는 신용관리 부서에서 일하고 있습니다. 보통 신용한도는 연체 가능성에 따라 결정하게 됩니다. 연체가능성이 낮으면 높은 신용한도를, 연체가능성이 높으면 낮은 신용한도를 받게 되는 것입니다. 연체 가능성을 파악하기 위해서 신용카드 신청자의 다양한 정보를 분석합니다. 크게 두 가지 데이터 소스가 있습니다. 카드 발급 신청서에 기입한 개인 정보와 크레딧 뷰로우(신용 금융 정보를 집중 관리 하는 기관) 데이터입니다. 신용카드 신청서에는 연령, 성별, 주소, 직업 등의 정보를 기입하도록 되어 있습니다. 크레딧 뷰로우에서는 신청자의 타 은행 신용 정보가 공유되고 있습니다. 대출을 받아보신 분은 경험하셨을 것이라고 생각합니다. 대출 신청을 하면, 담당 금융사는 타 금융사의 대출정보도 조회할 수 있습니다. 이 정보를 크레딧뷰로우가 제공하게 됩니다. 신용 대출액 혹은 카드 현금서비스 사용여부 등이 대표적인 예입니다. 신용카드사에서는 이런 모든 정보를 종합하여 정교한 연체 확률 모델을 개발, 관리하고 있으며, 신규 가입 요청이 들어오면, 신청자의 데이터가 연체 확률 예측 모델의 입력변수로 들어가게 되고, 예측 모델은 연체 확률을 추정하게 됩니다. 연체 예측 모델을 만들기 위해 신용관리 부서에서는 다양한 가설 검증과 데이터 분석으로 예측모델을 개발하고 운영하고 있습니다.

어느날, 재미교포 카스트로씨가 카드사를 방문을 했습니다. 미국에 어렸을 적에 부모님과 같이 이민을 갔다가 다시 한국에 역이민을 왔는데 신용카드가 필요하다는 것입니다. 김대리는 고민에 빠졌습니다. 카스트로씨 같은 경우는 아직 국내 신용거래가 없어, 크레딧 뷰로에 데이터가 존재하지 않습니다. 카드 신청서 개인 정보도 제한적입니다. 예를 들어, 자가 보유 여부, 주택담보대출 여부 등의 정보가 없습니다. 즉, 운용하고 있는 연체 확률 모델을 활용할 수 가 없는 것입니다.

김대리는 이 문제를 해결하기 위해 사내에 있는 통계자료를 수집했습니다. 수집된 통계자료는 연령대별, 성별, 거주지별, 직장별로 평균 연체확률(아래 그림) 이 있습니다. 카스트로씨가 제공할 수 있는 개인정보는 (1) 연령, (2) 성별, (3) 거주지, (4) 직장정보가 전부였습니다. 김대리가 수집한 통계자료는 아래와 같습니다. 그리고 카스트로씨에 해당하는 부분을 회색으로 표시했습니다. 아래 정보를 이용하여 카스트로씨의 연체 확률을 추정할 수 있을까요?

연령대	연체확률
20대	3.1%
30대	2.5%
40대	2.0%
50대	2.0%
60대	1.0%
전체	2.6%

성별	연체확률
남자	2.8%
여자	2.5%
전체	2.6%

거주지	연체확률
서울	2.8%
경기	2.5%
광역시	2.6%
지방	1.5%
전체	2.6%

직장	연체확률
대기업	2.1%
중소기업	2.5%
스타트업	3.5%
무직	5.1%
전체	2.6%

좋은 아이디어가 떠오르지 않았습니다. 김 대리는 이전 직장상사 오 부장님께 전화를 걸었습니다. 오 부장님은 신용분석으로 경력이 20년이상 되신 분이라, 비슷한 경험이 있으실 것이라고 생각했는데, 정말 좋은 해결책을 알려주셨습니다. 오즈(odds) 를 이용한 방법이였습니다. 오즈(odds)란 '이벤트가 일어나지 않을 확률' 대비 '이벤트가 일어날 확률'을 의미합니다. 김대리의 문제에서 오즈(odds) 는 (연체할 확률 / 연체하지 않을 확률) 로 계산이 될 수 있습니다. 아래는 오즈(odds) 계산 결과입니다.

연령대	Odds
20대	0.032
30대	0.026
40대	0.020
50대	0.020
60대	0.010
전체	0.027

성별	Odds
남자	0.029
여자	0.026
전체	0.027

거주지	Odds
서울	0.029
경기	0.026
광역시	0.027
지방	0.015
전체	0.027

직장	Odds
대기업	0.021
중소기업	0.026
스타트업	0.036
무직	0.054
전체	0.027

또한, 오즈비(odds ratio) 에 대한 이해가 필요합니다. 카스트로씨 연령에 대한 오즈(odds) 는 0.026 이고, 전체 오즈(odds) 는 0.027 이므로, 전체 대비 연령의 오즈비(odds ratio) 는 0.026 를 0.027 로 나눈 0.961 입니다 . 이 값을 의미는 카스트로씨의 연체에 대한 odds 는 전체 odds 대비 96.1% 낮다고 해석할 수 있습니다. 따라서 아무런 정보가 없는 카스트로씨의 오즈는 0.027 이였지만, 카스트로씨가 30대라는 사실을 알면 우리는 오즈를 조금 줄일 수 있습니다. 카스트로씨가 30대라는 정보를 입수하면, 카스트로씨의 오즈(odds) 는 0.027 * (0.026/0.027) 로 변경됩니다.

같은 방법으로 많은 정보가 많을 수록 카스트로씨의 odds 가 구체화 됩니다. 연령, 성별, 거주지, 지역 정보를 반영하면 카스트로 씨의 odds ratio 는 아래 공식으로 계산을 할 수 있습니다. 카스트로씨의 odds = (전체 odds) * (연령 odds ratio) * (성별 odds ratio) * (거주지 odds ratio) * (직장 odds ratio). 이 공식의 계산 결과는 0.029 가 됩니다. 오즈(odds) 는 P / (1-P) 즉, '연체할 확률' / '연체하지 않을 확률'이므로 P 로 풀어쓰면, 연체 확률 P 는 odds / (1 + odds) 가 됩니다. P 를 계산하면 카스트로씨가 연체할 확률은 2.79% 가 됩니다. 따라서 김대리는 연체확률 2.79% 에 해당하는 신용한도를 부여하면 합리적인 결정이라고 할 수 있습니다. 특히 신용관리 부서는 "왜 그런 결론을 내렸는지?" 에 대하여 고객에게 설명을 할 수 있어야 합니다. 예를 들면 "내 옆집은 신용한도가 천만원인데 나는 오백만원이가요? 등의 민원이 있을 수 있습니다. 머신러닝 기반 모델 들은 명확한 해석이 불가능해서 이런 종류의 민원을 근본적으로 해결하지 못합니다. 통계 기반의 모델은 결과에 대한 설명이 가능하므로 이런 종류의 민원에 대응이 가능합니다. 따라서 많은 금융기관이 통계적인 모델링 방식을 아직 선호하고 있습니다.

이 사례에서 데이터 분석을 공부하셨던 분들은 로지스틱 회귀분석과 비슷하다고 느끼 셨을 것입니다. 맞습니다. 위 해결책은 로지스틱 회귀분석 모델링과 동일합니다. 참고로 로지스틱 회귀모델은

$\ln(odds)$ 를 $_X$ 의 선형조합 $(b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4)$ 의 형태로 설명하는 모델입니다.

$(\text{카스트로씨 } odds)$ 는 $(\text{전체 } odds) * (\text{연령 } odds\ ratio) * (\text{성별 } odds\ ratio) * (\text{거주지 } odds\ ratio) * (\text{직장 } odds\ ratio)$ 로 추정할 수 있다고 사례에서 설명드렸습니다.

양변에 log 를 씌우면,

$$\ln(\text{카스트로씨 } odds) = \ln(\text{전체 } odds) + \ln(\text{연령 } odds\ ratio) + \ln(\text{성별 } odds\ ratio) + \ln(\text{거주지 } odds\ ratio) + \ln(\text{직장 } odds\ ratio) + \ln(\text{직장 } odds\ ratio)$$

따라서,

$\ln(\text{전체 } odds)$ 는 b_0, $\ln(\text{연령 } odds\ ratio)$ 는 $(b_1 \cdot x_1)$ 에 해당한다는 것을 알 수 있습니다.
x_1 이 (0,1) 의 바이너리 값이라면 b_1 은 해당 연령의 $odds\ ratio$ 에 log 를 한 값임을 알 수 있습니다.

* 주석: 각 정보가 독립인 경우에 계산이 성립함
