Analysis of loan customers' characteristics
(대출 고객 특성 분석)

Abstract

이 연구의 목표는 신생 은행의 마케팅 부서에 더 많은 대출 고객을 확보하기 위한 새로운 캠페인을 시작할 수 있는 정보를 제공하는 것이다. 관심 있는 구체적인 질문은 매개 변수의 조합으로 인해 고객이 개인 대출을 받을 가능성이 높아지고 교차 판매 기회를 지원하는 온라인 서비스, 보안 계정, 신용 카드와 같은 특별 제안 중 어떤 연관성이 있는가 하는 것입니다.

사용된 데이터 마이닝 기술은 설명 데이터 분석, 엔트로피 분류 트리, 신경망, 4가지 평균 클러스터링 및 주요 구성요소 분석이다. 이 분석은 대출을 받을 가능성이 있는 개인들의 세 가지 다른 특성을 산출한다.

'신세대 자수성가'라는 이름의 한 그룹은 높은 소득과 학부 수준, 높은 신용카드 지출을 가진 젊은이들을 포함하고 있다. [new generation of self-made man]

두 번째 그룹은 높은 교육 수준과 높은 수입을 가지고 있고 은행의 다른 시설에 관심이 있는 '개방적인 사람들'이다. [open minded people]

세 번째 그룹은 "보수주의자"를 만들고, 그들은 고학력자이며, 한 가정에 살고 있고, 추가 은행 서비스에 관심이 없다. [conservative]

Table of contents

I 표 및 표

1. 소개

이 연구의 목표는 신생 은행 마케팅 부서에 더 많은 대출 고객을 확보하기 위한 새로운 캠페인을 시작할 수 있는 정보를 제공하는 것이다. 이 연구는 전반적인 고객 확보 측면에서 빠르게 성장하고 있는 한 젊은 은행에 대한 것입니다. 그 은행의 고객은 크게 두 그룹으로 나뉜다.

첫 번째는 가장 큰 그룹을 형성하는 책임 고객입니다. 책임 고객은 은행의 계좌로 돈을 입금하고, 은행은 고객의 요청이 있을 때 갚아야 한다. 보통 은행에서는 예치된 돈에 대해 소량의 이자를 준다.

두 번째 그룹은 개인 대출 고객입니다. 그들은 은행에서 돈을 빌리는 고객들이다. 계약이 체결되면 고객은 추가 이자와 함께 돈을 돌려받을 의무가 있다. 이 금리는 예치금(예치금의 금리) 보다 높다.

따라서, 대출은 은행의 수입원이며 그들은 대출 고객 수를 늘리는 데 관심이 있다. 게다가, 그 은행은 부채 고객을 대출 고객으로 전환하는 것을 목표로 한다. 그 은행이 작년에 책임 고객을 대상으로 실시한 캠페인은 전환율이 9%가 넘는 성공률을 보였다. 전반적인 관심은 이전 캠페인의 데이터를 기반으로 대출 고객의 개선과 변수 사이의 연관성을 찾는 것이다.

2. 재료 및 방법

이 장에서는 데이터와 변수의 특징을 범주 및 단위로 설명합니다. 데이터를 더 잘 알고 나면 목표가 더 정확하게 정의됩니다. 또한 사용할 데이터 마이닝 기법이 제시되고 사용되지 않는 기법에 대한 확장이 짧게 제공된다.

2.1. 데이터 설명

데이터 집합에는 14개의 변수를 4개의 다른 측정 범주로 나눈 5000개의 관측치가 포함됩니다.

이진수 범주는 대상 변동형 개인대출, 증권계좌, CD계좌, 온라인뱅킹, 신용카드 등 5가지 변수가 있다.
구간 범주에는 연령, 경력, 소득, CC 평균 및 주택담보대출의 다섯 가지 변수가 포함됩니다.
순서형 범주에는 변수 패밀리와 교육이 포함됩니다.
마지막 범주는 ID 및 Zip 코드와 함께 공칭입니다. 가변 ID는 개인(ID로 표시)과 대출 사이의 개별 연결과 같은 흥미로운 정보를 추가하지 않는다. 따라서, 그것은 TEST에서 무시될 것이다.

개인 대출	이 고객은 지난 캠페인에서 제공한 개인 대출을 수락했습니까?
증권계정	고객이 은행에 증권 계좌를 가지고 있습니까?
CD 계정	고객이 은행에 예금증서(CD) 계좌를 가지고 있습니까?
온라인 뱅크	고객이 인터넷 뱅킹 시설을 이용합니까?
신용카드	고객은 유니버설 은행에서 발급한 신용카드를 사용합니까?
나이	완료된 연도의 고객 연령
경력	다년간의 직업 경험
소득	고객의 연간 수입($000)
CC 평균	월별 신용 카드 지출($000)
저당권	주택담보대출의 가치(있는 경우) ($000)
가족	고객의 제품군 크기
교육	교육 수준. 1: 학부, 2: 졸업, 3: 고급/전문
우편번호	집 주소 ZIP 코드
아이디	고객ID

데이터 변수를 도입한 후 연구 목표를 보다 구체적으로 정의할 수 있습니다.
1) 고객이 개인 대출을 더 잘 받을 수 있도록 하는 매개 변수는 무엇입니까?
2) 온라인 서비스, 보안 계정, 신용카드 등과 같은 특별 행사 중 교차 판매 기회를 찾기 위한 연관성이 있는가?

2.2. 데이터마이닝 방법

이 파트는 사용된 다섯 가지 데이터 마이닝 기법 (즉, EDA, 분류 트리, 신경망, 클러스터 분석, 주요 구성요소 분석 세부 사항.) 이러한 방법의 프로세스뿐만 아니라 아이디어도 설명되고 사용에 대한 정당성도 제시됩니다.

탐색적 데이터 분석[EDA]은 보다 발전된 분석을 조사하기 전에 데이터를 파악하는 데 매우 유용한 방법입니다. 예를 들어, 데이터가 이 방법보다 더 쉽게 사용할 수 없다는 가정을 충족하지 못하는 경우, 그 결과를 신뢰할 수 없기 때문에 분석에서 실수를 피하기 위해 데이터를 아는 것이 중요하다. 변수의 분포를 조사합니다. 이 과정에서는 평균, 분산, 정규성 및 대칭이 중요하며 정규성을 최대화하기 위한 변환도 가능합니다. 또한 변수뿐만 아니라 변수와 대상 변수 간의 연관성도 발견할 수 있습니다. 따라서 그래픽 도구를 적용하는 변수 간의 상관 관계를 찾으려고 한다.

다음으로, 우리는 분류 트리를 사용하여 독립 변수로 대상 가변 개인 대출을 예측하고 연관성을 발견한다. 의사결정 트리는 변수를 사용하여 변수부터 시작하여 목표 변수를 가장 잘 구분하는 두 그룹 대출과 비대출자를 구분합니다. 각 나눗셈은 의사 결정 노드로 끝납니다. 변수가 남아 있지 않을 때까지 생성되는 노드가 점점 더 많아져 반응 변수를 유의하게 분리할 수 있습니다. 분리를 계산하는 데는 지니, 엔트로피 및 CHAID의 세 가지 다른 측정값이 있습니다. 우리는 분석해야 하는데, 그 중 어느 것이 가장 작은 오류율을 산출하는지 분석해야 한다. 이 방법과 다음 방법의 경우, 데이터를 교육, 검증 및 테스트 세트로 분할하는 것이 가장 좋습니다. 분류 트리는 몇 가지 장점이 있습니다. 해석하기 쉽고, 데이터의 오류가 결과에 영향을 줄 가능성이 낮으며, 불필요한 변수를 자동으로 제거합니다. 또한 이 방법은 변수 간의 교호작용을 잘 처리할 수 있습니다.

신경망 방법의 아이디어는 뇌의 신경 네트워크의 기능에 기초한다. 신경세포는 서로 연결되어 있고 모델이 미래 예측을 위해 사용할 수 있는 경험(인식된 패턴)에서 학습한다. 이 보고서에서는 여러 입력, 하나의 숨겨진 계층 및 단일 출력을 가진 네트워크를 사용한다. 은닉 계층은 조합 함수, 즉 입력 변수의 가중 합계와 변환 함수로 구성된 활성화 함수를 사용한다. 변환 함수의 경우 대상이 이항 변수이므로 로지스틱 함수를 선택합니다. 이 방법의 장점은 상호작용뿐만 아니라 선형 관계를 처리할 수 있다는 것이다. 따라서 예측 정확도는 높고 결과는 강력하다. 단점은 모델의 복잡성입니다. 결과가 어떻게 평가되고 어떤 변수가 반응 변수를 구별하는 데 중요한지 이해하는 것이 항상 쉬운 것은 아닙니다. 그럼에도 불구하고 몇 개의 숨겨진 계층과 노드를 통해 그 결과를 이해할 수 있다.

클러스터 분석(군집분석)은 데이터 개체 집합을 클러스터로 그룹화하는 데 사용되며, 동일한 클러스터에 있는 개체의 유사성과 다른 클러스터의 개체와의 차이를 최대화합니다. 목적은 주로 대출자를 포함하는 클러스터를 찾는 것이므로 이 그룹의 특성에 대한 통찰력을 얻을 수 있습니다. 결과는 거리 측정 방법에 따라 달라집니다. 데이터에는 구간, 순서 및 이항 변수가 포함되어 있는 다양한 변수에 대한 다양한 측정값이 있습니다. 간격과 순서형에서 SAS는 이진 변수에 대해 동일한 측정값을 적용하지만 다른 측정값을 적용합니다. 군집화 전에 변수의 각 차이가 전체 거리 값에 동일하게 기여하도록 데이터를 표준화해야 합니다. K-평균 군집을 사용하는 클러스터를 구축하기 위해, 그 이유는 분석 부분에 제시되어 있다. 먼저 다수의 클러스터를 자체적으로 결정해야 하며, 그런 다음 초기 중심이 선택되고 관측치가 가장 가까운 센터에 할당되며, 클러스터가 더 이상 변경되지 않을 때까지 반복됩니다. 범용 뱅크 데이터 세트는 5000개의 관측치와 공칭 데이터가 없으므로 빅 데이터 세트를 잘 처리할 수 있지만 공칭 데이터가 없으므로 이 방법을 사용하는 것이 좋습니다. 단점은 클러스터 수를 분석가가 미리 정의해야 한다는 것입니다.

주성분 분석은 모형이 해석하기 쉽도록 변수 수를 더 낮은 차원으로 줄이려고 합니다. 이러한 치수는 가중 원래 변수의 선형 함수인 주성분이라고 하는 새로운 변수입니다. 방법을 수행하기 전에 데이터를 표준화하는 것이 중요합니다. 그렇지 않으면 계산된 가중치가 거짓이 됩니다. 이러한 상관 관계

가 없는 각 구성 요소는 데이터의 변동성을 설명합니다. 우리의 목표는 차원을 변동성의 약 80%~90%를 설명하는 몇 가지 구성 요소로 줄이는 것입니다. 그런 다음 이러한 성분을 해석하고 대상 변수와의 연관성을 찾을 수 있습니다.

2.3. 추가 데이터 마이닝 방법

이 절에서는 본 연구에서 사용되지 않는 추가 데이터 마이닝 방법에 대해 설명합니다. 이것들은 다섯 가지 기법 연관 분석, 로지스틱 회귀 분석, 번들링 기법, 메모리 기반 추론 및 텍스트 마이닝이다. 그것들은 간략하게 제시되고 우리는 왜 그러한 방법들이 이 연구에 선택되지 않았는지 설명한다.

일반 용어 시장 바구니 분석은 두 가지 방법 연관성과 시퀀스 분석을 다룬다. 두 변수 모두 변수 사이에서 빈번한 패턴을 찾는 데 유용합니다. 연결 방법은 함께 발생하는 변수를 식별하고 그에 따라 규칙을 만드는 데 유용합니다. 규칙은 변수가 데이터에 단독으로 조합되어 나타나는 빈도를 계산하여 개발됩니다. 순서 지정은 변수와 변수의 확률의 연결 외에도 관계가 발생하는 순서도 고려합니다. 따라서 분석에 타이밍 요소가 포함됩니다. 전반적으로 시장 바구니 분석은 변수가 함께 나타날 확률을 알아내는 데 유용합니다. 불행히도, 이 분석은 우리에게 두 가지 이유 중 어떤 결과도 주지 않는다. 첫째, 5% 신뢰 수준에서 알 수 없는 이유로 유의한 연관성을 만들 수 없었다. 둘째, 시퀀스 검색을 수행하는 데 필요한 시간 요소가 없습니다.

반응 변수를 예측하는 또 다른 유용한 방법은 회귀 모형입니다. 목표 변수 대출이 이항이므로 이벤트 대출이 발생하지 않을 확률에 대해 발생할 확률을 예측하는 로지스틱 회귀 분석을 사용해야 합니다. 이 방법은 정규 분포 변수를 가정한 것입니다. 따라서 모형을 구축하기 전에 변수를 변환해야 할 수도 있습니다. 또한 변수를 자동으로 삭제하지 않습니다. 따라서 역방향 제거와 같은 변수 선택 노드 또는 변수 선택 방법을 사용해야 합니다. 회귀 모형의 장점은 변수 간의 선형 관계를 잘 처리할 수 있다는 것입니다. 일반적으로 각 개인이 개별 출력을 받기 때문에 분류 트리보다 더 정확한 방법이다. 이 방법을 사용하지 않는 이유는 분류 트리의 오류율이 낮기 때문입니다. 또한, 이 모델은 트리 모델에 아직 포함되지 않은 추가 정보를 제공하지 않습니다. 기술 번들의 일반적인 절차는 여러 모델의 결과를 평균 출력에 결합하는 것이다. 이 평균은 더 정확합니다. 즉, 개별 오류가 취소될수록 총 오차가 감소하고 측정 집합마다 차이가 작으면 결과가 더 안정적입니다. 이 기법의 장점은 새로운 데이터를 채점할 때 예측이 향상된다는 것이다. 단점은 결과를 해석하기 어렵다는 것이다. 대출자가 될 가능성이 높은 사람을 결정하기 위해 중요한 변수를 식별하는 우리의 목표에 대해, 우리는 분석의 결과를 해석할 수 있어야 하고 새로운 데이터의 채점에 관심이 덜해야 한다. 따라서, 새로운 데이터의 점수를 잘 매기는 이 방법의 이점은 우리의 목적에 유용하지 않으며 우리는 분석 도구로 번들을 사용하지 않는다.

메모리 기반 추론은 K-가장 가까운 이웃 방법을 사용하여 새로운 데이터를 예측한다. 이진 목표 변수의 경우, 이 방법은 사전 정의된 이웃 K 수의 로컬 영역을 검색하고 가장 가까운 이웃에 새 개체를 할당합니다. 우리의 목표 측면에서, 단점은 이것이 예측 방법이며 대출자의 특성을 설명하는 중요한 변수를 찾는 데 도움이 되지 않는다는 것이다. 분명히, 텍스트 마이닝은 기사나 다른 서면 문서의 패턴을 탐지하는 데 사용되므로 범용 은행 데이터 세트에는 아무런 쓸모가 없다.

3. 결과 및 해석

이 파트는 5가지 기법에 대한 SAS 분석의 결과를 보여줍니다. 더욱이, 그 결과는 주로 대출자의 특성 측면에서 해석된다.

3.1.EDA

이 섹션에서는 데이터를 자세히 살펴볼 수 있으므로 데이터를 보다 잘 파악할 수 있습니다. 우리는 데이터에 우리가 알아야 할 것이 있고 변수들 사이에 어떤 관계가 있는지, 변수가 어떻게 분포되어 있는지, 평균값이 얼마인지, 순서형 변수에 대한 비율 등을 알고 싶다.

부록 A의 히스토그램은 순서형 변수 패밀리와 교육뿐 아니라 연령, 경험, 소득, 공동 주택담보대출 등 간격 변수에 대한 정규성을 극대화하는 분포와 변환된 분포를 보여준다. 이러한 각 변수는 높은 왜도, 첨도 또는 둘 모두를 나타내므로 변환이 변수를 개선합니다. 가변 연령은 23세에서 67세 사이이며, 연령에 따라 거의 동일한 비율의 사람들이 있다. 제곱근 변환은 첨도를 개선하지만 왜도는 더 악화됩니다. 수년간의 경험에 대한 히스토그램에는 -1, -2, -3의 음수 값이 있는 것으로 나타납니다. 일반적으로 마이너스 연도를 측정할 수 없기 때문에 데이터 입력 오류가 발생할 수 있습니다. 그러나 데이터에서 해당 값의 비율은 1% 미만이며 음수 값의 이유를 찾을 수 없으므로 삭제해야 합니다. 변동 수입과 관련하여 최소값은 $8000이고 최대값은 $224000입니다. 대다수의 개인들은 2만 달러에서 9만 달러 사이의 수입을 가지고 있다. 일반적으로 재무 데이터의 경우 오른쪽으로 치우쳐 있으므로 변환을 통해 왜도 수준을 0.84에서 -0.08로 개선할 수 있습니다. 평균 신용카드 지출은 월 0에서 1만 달러에 이르는 광범위한 범위를 가지고 있으며, 대부분의 지출은 2천5백 달러 미만이다. 로그 변환은 왜도 수준을 향상시킵니다. 주택담보대출과 관련하여 개인들의 70%가 4만 달러 이하의 주택담보대출을 가지고 있지만, 최대 주택담보대출은 635,000 달러까지 높다. 제곱근 변환은 왜도를 2.1에서 1.2로 개선하고 첨도는 4.76에서 0.03으로 개선할 수 있습니다. 변수 패밀리와 교육은 순서형 변수이지만 EDA의 경우 EDA에서 사용할 수 있도록 간격 변수로 처리합니다. 이 연구에 참여한 인구의 30%는 1인 가구, 27%는 2인 가구, 20%는 3인 가구, 23%는 4인 가구입니다. 따라서, 가족의 분포는 균등하게 분포되어 있습니다. 교육에서 40%는 학부, 대학원, 전문직에서 각각 약 30%를 차지하고 있다.

그림 3.1의 상자 그림은 대상 변수 개인 대출과 설명 변수 패밀리와 구내 사이의 관계를 보여줍니다. 왼쪽의 그림은 중위수가 3인 가정이 대출을 받을 가능성이 더 높다는 것을 나타냅니다(대출자 수 = 1, 비대출자 수 =0). 1인 가구나 2인 가구가 대출을 받을 가능성이 낮다. 이것은 미래의 캠페인을 고려할 때 유용한 연관성이 될 수 있다. 예를 들어, 캠페인은 자녀가 있는 가족을 목표로 할 수 있다. 우리는 평균이 두 분포의 왜도 때문에 비교에 유용하지 않다는 것을 언급해야 한다. 평균 신용카드 지출의 영향과 관련해서는 분배가 크게 중복되지 않아 대출과 무대출 분포가 뚜렷하게 구별된다. 일반적으로 평균 신용카드 지출이 3800달러로 높을수록 개인대출의 확률이 높다는 것을 의미한다. 신용 카드 지출이 중간값 1400달러로 낮아지면 대출받을 가능성이 낮아진다. 이것은 예를 들어 대출 광고를 보낼 사람을

선택할 때 유용한 정보가 될 수 있습니다. 나머지 변수의 경우 상자 그림은 대출 대상자와 비대출 대상자를 구별하는 데 도움이 되는 분포를 제공하지 않습니다. 상자 그림은 각 변수 내에서 너무 많이 겹칩니다.

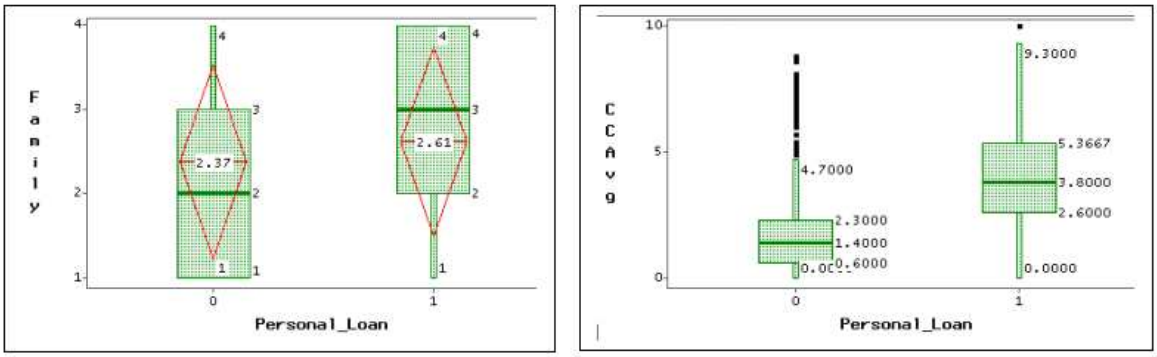

Figure 3.1: a) Box plot of family and personal loan b) CC avg and Personal loan

그림 3.2는 변수 경험과 연령 사이의 관계와 상관관계를 나타내는 산점도를 보여준다. 다년간의 업무 경험과 나이가 긍정적인 상관관계를 가지고 있다는 것을 나타내는데, 이는 합리적인 것 같다. 또한, 우리는 어떤 종류의 그룹화를 인식하며, 교육 수준 3(전문, 검정)은 교육 수준 2(대학원, 빨강) 및 1(학부, 녹색)와 구별된다. 레벨 3은 같은 양의 상관관계를 가지지만 전반적으로 경험이 적다. 아마도 이 그룹은 교육에 더 많은 시간을 소비하고 따라서 근무 경력이 더 짧을 것이다. 게다가, 40대 중반의 전문가들에게는 격차가 있는데, 아마도 그러한 사람들은 이 연구에 포함되지 않았기 때문에 누락되었을 것이다. 연구 대상자의 대다수는 학부 출신이며, 교육 수준은 가장 낮지만 경험이 가장 많은 사람들이다.

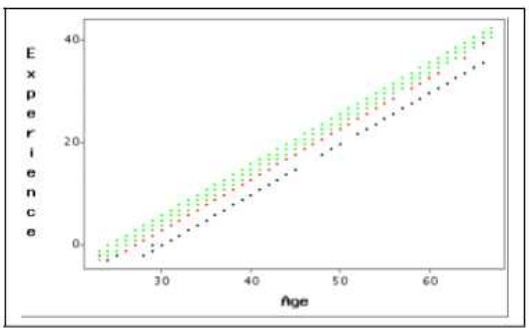

Figure 3.2: Scatter plot of age and experience

다음 그림 3.3, 소득, 공동 주택담보대출에 대한 산포도가 각각 표시된다. 신용카드 평균과 소득의 관계는 무관계에서 양관계로 달라진다. 긍정적인 상관관계에 대한 일반적인 진술은 더 높은 평균 신용카드 지출은 더 높은 소득을 나타내는 경향이 있다는 것이다. 소득이 적은 개인은 신용카드 지출이 제한적이다. 그러나 소득이 높다고 해서 반드시 신용카드 지출이 많은 것은 아니다. 불행히도, 어떠한 설명도 찾을 수 없었고, 어떤 요소가 무사용의 확산과 높은 신용카드 사용에 영향을 끼치는지 알 수 없었다.

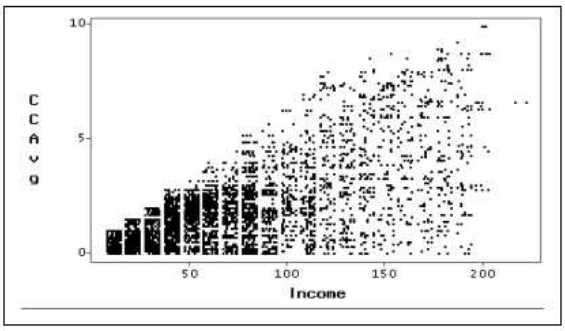

Figure 3.3: Scatter plot of income and CC avg

소득이 높은 사람이 신용카드와 고신용카드를 더 많이 쓰는 경향이 있다면 지출은 소득과 대출 사이에 간접적인 관계가 있을 수 있는 것보다 대출자가 될 확률이 더 높다는 것을 나타낸다. 대출에 대한 산포도를 그룹화하면(대출자가 빨간색으로 표시됨) 흥미로운 결과가 표시됩니다. 그림 3.4는 신용카드 지출이 4000달러 이상이고 소득이 10만 달러인 사람들이 대출을 받을 것을 제안한다. 소득 10만 달러 이상의 사람들은 그 신용카드로 얼마를 쓰는지와 무관하게 대출을 받는다.

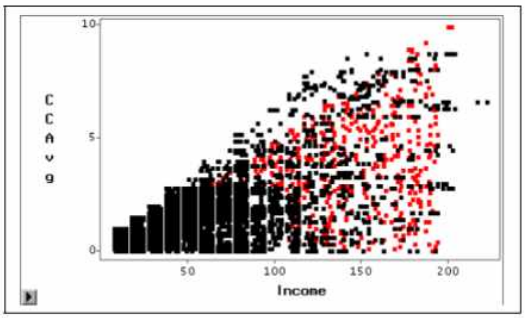

Figure 3.4: Scatter plot for income and CC avg grouped for personal loan

아마도 우리가 가족, 대출, 소득 사이에서 비슷한 결과를 얻는지 확인하는 것이 합리적일 것이다. 그림 3.5의 상자 그림은 흥미로운 결과를 다시 보여 준다. 가족의 규모가 실제로 대출의 가능성에 영향을 미치는 것 같지는 않지만 그것은 수입의 양보다 영향이 더 크다. 소득이 10만 달러 이하인 가정은 가족 규모와 상관없이 소득이 높은 가정보다 대출받을 가능성이 낮다. 우리는 신용카드 평균에 대해 비슷한 결과를 얻었습니다. 그럼에도 불구하고, 우리는 이 두 결과 중 어떤 결과(가족 크기가 영향을 미치는지 또는 영향을 미치지 않음)가 더 적합한지 알 수 없지만, 추가 분석에서 이 연관성에 대해 더 자세히 알 수 있을 것이다.

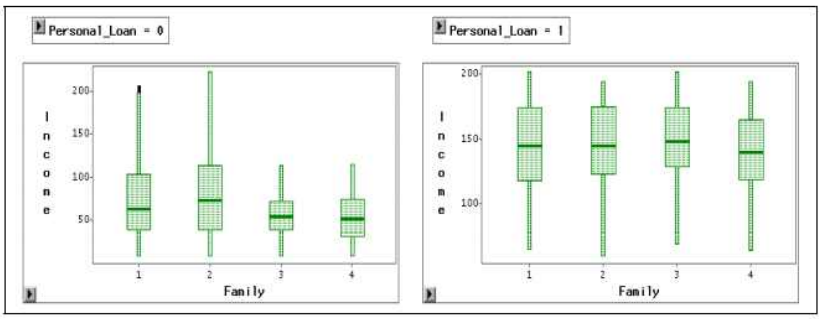

Figure 3.5: Box plot of family and income grouped for personal loan

담보대출과 소득의 상관관계에 대해, 우리는 신용카드 평균과 동일한 패턴을 관찰한다(그림 3.6). 그 관계는 소득과 주택담보대출 사이의 매우 긍정적인 상관관계에 대한 무관심 사이를 부채질한다. 긍정적인 관계는 높은 저당권이 높은 소득을 의미한다는 것을 암시한다. 우리는 팬아웃에 대한 설명을 찾

을 수 없었다. 게다가, 0달러와 75,000달러 사이의 격차가 눈에 띄게 보인다. 이는 주택 담보대출의 최소 가치가 약 75,000달러이기 때문이다. 빨간 점은 다시 대출받는 사람을 보여주고, 대출과 대출 사이에는 아무런 관계가 보이지 않지만, 우리는 이전과 같은 소득과 대출의 연관성을 볼 수 있다.

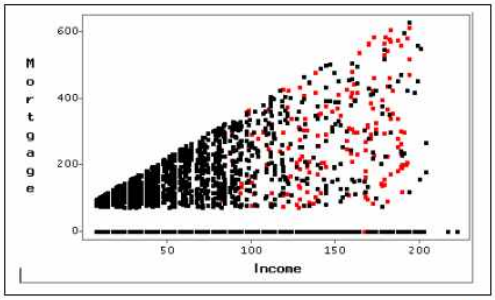

Figure 3.6: Scatter plot of income and mortgage grouped for personal loan

요약하면, EDA는 다음과 같은 결과를 제공한다. 전체적으로 정규성 변환은 왜도와 첨도를 개선합니다. 이는 정규 분포 데이터가 가정된 경우 추가 조사에 유용합니다. 상자 그림 그래프에서 대가족(구성원 3명 이상)이 대출을 받을 가능성이 더 높다는 결론을 내린다. 반대로, 다른 상자 그림은 대출과 가족 크기 사이에는 관계가 없으며 소득과 대출 사이에는 양의 연관성이 있음을 나타냅니다. 게다가, 우리는 소득이 높을 때 신용카드 지출이 중요하지 않은 반면, 낮은 소득과 결합된 높은 신용카드 지출은 대출을 받을 가능성이 높다는 것을 알고 있다. 이러한 결과 외에도, 이 데이터 세트의 대부분의 사람들은 학부 수준의 교육을 받고 있으며 대부분의 사람들은 2만 달러에서 9만 달러 사이의 수입을 올리고 있습니다.

개인적인 생각) EDA란 데이터를 시각화하여 변수간의 상관관계를 따져보며 영향이 있는지 있다면 어떤 영향이 있는지 알아내는 분석방법인 것 같다.