**Machine Learning 을 통한**

**예금 가입 고객 예측 모형**

**: 포르투갈 은행 텔레마케팅 캠페인을 중심으로**

**Data Mining**

**김희경 교수님**

**2019 년 05월**

**동국대학교 이과대학**

**통계학과**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 김 경 록 |  | 박 권 수 |  |

**< 목 차 >**

**제 1 장 서 론**

**제 1 절 연구목적 및 배경**

**제 2 절 연구 방법**

**제 2 장 본 론**

**제 1 절 변수 파악**

**제 2 절 데이터 정제 및 처리**

1. **변수 제거 및 이상 관측치 제거**
2. **범주형 변수 재범주화**
3. **새로운 변수 생성 및 결측치 처리**
4. **데이터셋 분리**
5. **변수 변환 및 최종 변수 선택**

**제 3 절 모형 구축 및 평가**

1. **로지스틱 회귀 모형 구축**
2. **로지스틱 회귀 모형 평가**
3. **최종 모형 선택**

**제 3 장 결 론**

**제 1 절 최종 모형 의미 해석**

**제 2 절 향후 과제**

**부 록**

**< 표 목 차 >**

**<표 1-1> 데이터 변수 구성**

**<표 1-2> 연령 기술 통계량**

**<표 1-3> 채무 불이행 여부와 종속 변수 교차 분할표**

**<표 1-4> 평균 연간 잔고 기술 통계량**

**<표 1-5> 마지막 연락 연도별 월별 빈도표**

**<표 1-6> 지속 시간 기술 통계량**

**<표 1-7> 캠페인 진행 중 수행된 연락 수의 퍼센타일 표**

**<표 1-8> 마지막으로 고객에게 연락 후 경과일수의 퍼센타일 표**

**<표 1-9> 이전 캠페인 진행 중 수행된 연락 수의 퍼센타일 표**

**<표 1-10> 직업 변수 재범주화**

**<표 1-11> 학력 변수 의사결정나무 분류 결과표**

**<표 1-12> 잠재고객 범주형 변수 검정 결과**

**<표 1-13> 잠재고객 평균 연간 잔고 변수 변환 로지스틱 모형**

**<표 1-14> 잠재고객 최종 선택된 변수**

**<표 1-15> 기참여고객 범주형 변수 검정 결과**

**<표 1-16> 기참여고객 평균 연간 잔고 변수 변환 로지스틱 모형**

**<표 1-17> 기참여고객 최종 선택된 변수**

**<표 1-18> 잠재고객 Type3 분석 결과 선택된 변수**

**<표 1-19> 잠재고객 단계적 변수선택법 최종선택 변수**

**<표 1-20> 기참여고객 Type3 분석 결과 선택된 변수**

**<표 1-21> 기참여고객 단계적 변수선택법 최종선택 변수**

**<표 1-22> 잠재고객 Training Data Set 십분위 분석표**

**<표 1-23> 잠재고객 Test Data Set 십분위 분석표**

**<표 1-24> 잠재고객 Training, Test Set별 0분위수 그룹과 9분위수 그룹의 비율**

**<표 1-25> 기참여고객 Training Data Set 십분위 분석표**

**<표 1-26> 기참여고객 Test Data Set 십분위 분석표**

**<표 1-27> 기참여고객 Training, Test Set별 0분위수 그룹과 9분위수 그룹의 비율**

**<표 1-28> 잠재고객 로지스틱 회귀 모형 분류 결과표**

**<표 1-29> 기참여고객 로지스틱 회귀 모형 분류 결과표**

**<표 1-30> 잠재고객 이익도표**

**<표 1-31> 기참여고객 이익도표**

**<표 1-32> 잠재고객 로지스틱 모형 분산팽창지수**

**<표 1-33> 기참여고객 로지스틱 모형 분산팽창지수**

**< 그 림 목 차 >**

**<그림 1-1> 은행 이용 고객 연령 히스토그램**

**<그림 1-2> 은행 이용 고객 직업 막대 그래프**

**<그림 1-3> 혼인 상태별 예금 구매 여부 비율**

**<그림 1-4> 은행 이용 고객 학력 막대 그래프**

**<그림 1-5> 학력별 예금 구매 여부 비율**

**<그림 1-6> 채무 불이행 여부 막대 그래프**

**<그림 1-7> 평균 연간 잔고 히스토그램**

**<그림 1-8> 주택담보대출 여부별 예금 구매 여부 비율**

**<그림 1-9> 개인 대출 여부별 예금 구매 여부 비율**

**<그림 1-10> 연락방식 막대 그래프**

**<그림 1-11> 예금 구매 여부별 지속시간 박스플랏**

**<그림 1-12> 이전 캠페인 결과별 예금 구매 여부 비율**

**<그림 1-13> 타깃 변수 막대 그래프**

**<그림 1-14> 변환 후 연령 변수 분포**

**<그림 1-15> 잠재고객 마지막 통화 시간 변수 변환 전후 Q-Q plot**

**<그림 1-16> 잠재고객 평균 연간 잔고 변수 변환 전후 Q-Q plot**

**<그림 1-17> 기참여고객 마지막 통화 시간 변수 변환 전후 Q-Q plot**

**<그림 1-18> 기참여고객 평균 연간 잔고 변수 변환 전후 Q-Q plot**

**<그림 1-19> 잠재고객 모델 이익도표 그래프**

**<그림 1-20> 기참여고객 모델 이익도표 그래프**

**Ⅰ. 서 론**

**제 1 절 연구목적 및 배경**

시간이 지남에 따라 기술 발전에 부합하는 여러 마케팅 방법들이 나타나고 있지만 아직까지도 비즈니스를 성장시킬 수 있는 가장 효과적인 마케팅 전략 중 하나는 텔레마케팅이다. 텔레마케팅은 다른 마케팅들에 비해 어렵지 않은 방법으로 잠재적 고객들에게 쉽게 접근할 수 있는 강점을 가지고 있다. 또한 개인의 소비 개념이 중요해짐에 따라 텔레마케팅을 이용한 개인 맞춤형 상품과 서비스를 개발하는 것이 중요시되고 있다. 시장 조사 또는 다른 마케팅으로의 연결을 위한 방법으로 텔레마케팅이 종종 사용되곤 했다. 특히나 국내보다도 해외에서의 텔레마케팅은 더 전문적이고 꾸준한 성장을 보이고 있을 정도로 중요한 마케팅의 하나로 자리매김하고 있다.

이에 저자는 해외에서의 텔레마케팅의 중요성을 인지하고 사전에 진행된 텔레마케팅 캠페인이 추후 고객들의 예금 가입 여부에 미치는 영향에 대해 관심을 가지게 되었으나 국내 텔레마케팅 관련 논문 중 머신러닝 모델을 적용하여 상품 판매를 예측한 사례가 많지 않아 저자가 직접 모델을 구축하여 고객들에게 예금 상품을 판매하기 위해 미치는 텔레마케팅의 영향과 원인 변수를 알아보고자 한다.

**제 2 절 연구방법**

포르투갈 금융기관(Portuguese Banking Institution)으로부터 마케팅 캠페인과 관련된 데이터를 받아 연구를 하였다. 이 캠페인 데이터는 고객들에게 유선전화(tele-phone)와 휴대폰(cellular phone)으로 접촉한 텔레마케팅 고객들에 관련된 데이터이며, 고객들의 개인정보(연령, 직업, 학력 등)와 금전 관련 정보(평균 연간 잔고, 채무 불이행 여부 등), 최근 캠페인 정보(캠페인을 위해 접촉한 수, 캠페인 결과 등)를 추가적으로 받아 독립변수로 사용하고, 최종적으로 캠페인 결과 정기 예금 상품에 가입하였는 지에 대한 여부를 종속변수로 하여 분석을 실시한다. 종속변수의 경우 단변량에서의 이상치와 다변량에서의 이상치를 탐색하여 처리하고 변수 상의 결측치는 많은 경우 해당 변수를 제거, 적은 경우는 다른 변수들을 활용하여 채우도록 한다. 연속형 변수 사이에 단위 차이가 많이 나는 경우 표준화 또는 변수 변환을 실시해주고, 필요하다면 가지고 있는 변수를 활용한 새로운 변수를 생성한다. 모델 구축 전 변수들 간의 상관관계를 파악해 모델 구축 간에 다중공선성을 미리 파악하고, 전처리가 끝난 독립변수들을 가지고 머신러닝 기법인 로지스틱 회귀(Logistic Regression), 랜덤 포레스트(Random Forest)를 활용한 정기 예금 구매 고객 예측 모델을 만들고 모델 평가를 통한 최적의 모형을 선택하고, 최종적으로 타깃변수 예측을 통해 창출할 수 있는 경제적 수익과 예금 구매 성공에 미치는 주요한 특성을 확인하고 좀 더 효율적인 마케팅 자원 관리에 기여하고자 한다. 전체적인 분석은 R 3.6.0 을 통해 분석할 예정이며 필요에 따라 패키지를 사용할 것이다.또한 모든 검정에서 유의수준은 0.05로 설정할 것이다.

**Ⅱ. 본 론**

**제 1 절 변수 파악**

포르투갈 금융 기관으로부터 전해 받은 데이터인 bank-full.csv파일로 세미콜론(;)을 구분자로 가지고 있으며 데이터는 2008년부터 2011년 사이의 자료라고 한다. 우선 전체 데이터는 45,211개의 관측치와 17개의 변수로 이루어져 있다. 우선 데이터 제공자에 의해 전체 변수에 결측치가 없음을 파악했다. 제공받은 데이터의 변수는 아래의 <표 1-1>와 같다. 변수유형별로 나누어 보면 크게 독립변수 16개와 종속변수 1개로 구성되어 있다. 독립변수에서도 숫자형 변수가 7개가 있는데 그 중 이산형 변수로 연령(age)과 마지막 연락 때 통화 시간(duration), 현재 캠페인 중 수행된 연락 수(campaign), 마지막 연락 후 경과일(pdays), 과거 캠페인에서 수행된 연락 수(previous)가 있고, 연속형 변수로 평균 연간 잔고(balance), 숫자로 구성된 날짜형 변수로 마지막 연락한 달의 날짜(day)가 있다. 다음으로 범주형 변수는 9개가 있는데 그 중 범주가 3개 이상인 변수로 직업(job), 혼인 상태(marital), 연락 방식(contact), 이전 캠페인 결과(poutcome)가 있고 이진 변수(binary variable)로는 주택담보대출 여부(housing), 개인 대출 여부(loan)가 있고, 범주형 변수로 구성된 날짜형 변수로 마지막 연락한 달(month)이 있다. 종속 변수(y)는 이진 변수이며, 고객이 최종적으로 예금 상품에 가입 여부에 관한 yes/no로 이루어진 변수이다. 데이터 제공 시 종속변수에 불균형이 있음을 전달받았고, 작은 범주에 맞춰 달라는 요구에 따라 큰 범주에서 샘플링(sampling)하여 작은 범주에 맞추어야 한다.

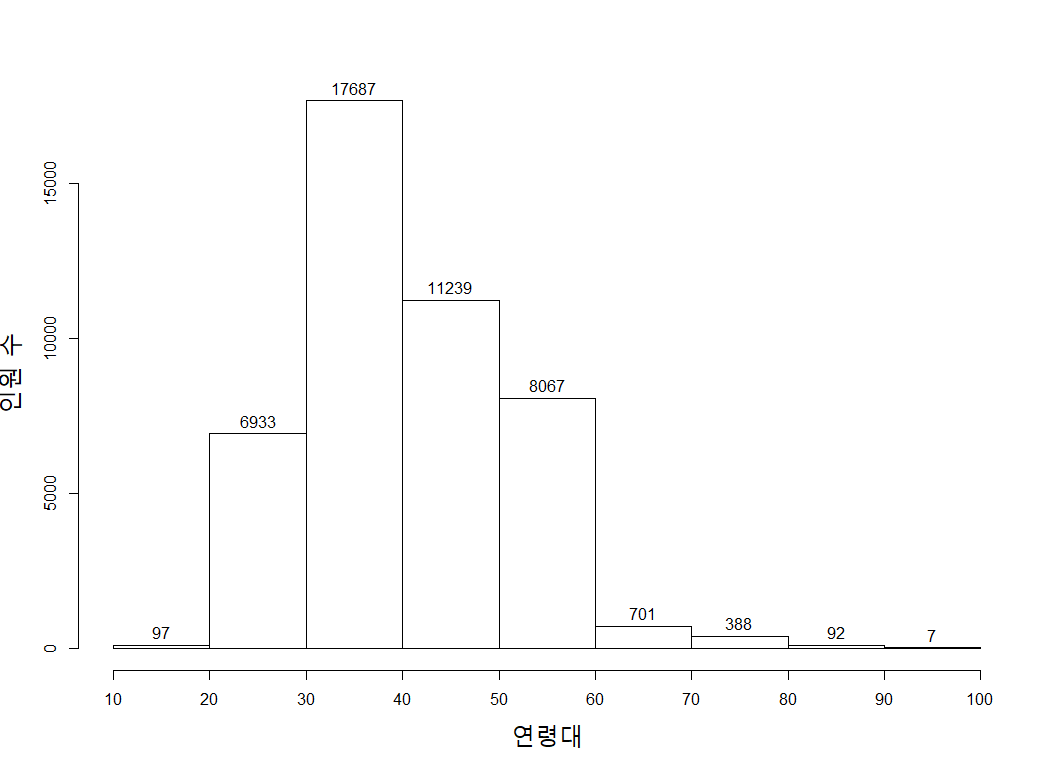
<표 1‑1> 데이터 변수 구성

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **변수명** | **변수유형** | **변수설명** | **변수명** | **변수유형** | **변수설명** |
| age | numeric | 연령 | day | numeric | 마지막 연락한 날짜 |
| job | factor | 직업 | month | factor | 마지막 연락한 달 |
| marital | factor | 혼인 상태 | duration | numeric | 마지막 연락 시간 |
| education | factor | 학력 | campaign | numeric | 수행된 연락 수(현재) |
| default | factor | 채무 불이행 여부 | pdays | numeric | 마지막 연락 후 경과일 |
| balance | numeric | 평균 연간 잔고 | previous | numeric | 수행된 연락 수(과거) |
| housing | binary | 주택담보대출 여부 | poutcome | factor | 이전 캠페인 결과 |
| loan | binary | 개인 대출 여부 | y | binary | 종속 변수 |
| contact | factor | 연락 방식 |  |  |  |

|  |  |
| --- | --- |
| **Minimum** | 18 |
| **1st Quan.** | 33 |
| **Median** | 39 |
| **Mean** | 40.94 |
| **3rd Quan.** | 48 |
| **Maximum** | 95 |

변수별로 더 자세히 파악하기 위해 위의 구성표에 따라 왼쪽부터 오른쪽 순으로 변수를 확인해 보면, 은행 이용 고객 연령(age)은 다음 <그림 1-1>과 같고, 연령의 기술 통계량은 <표 1-2>와 같다.

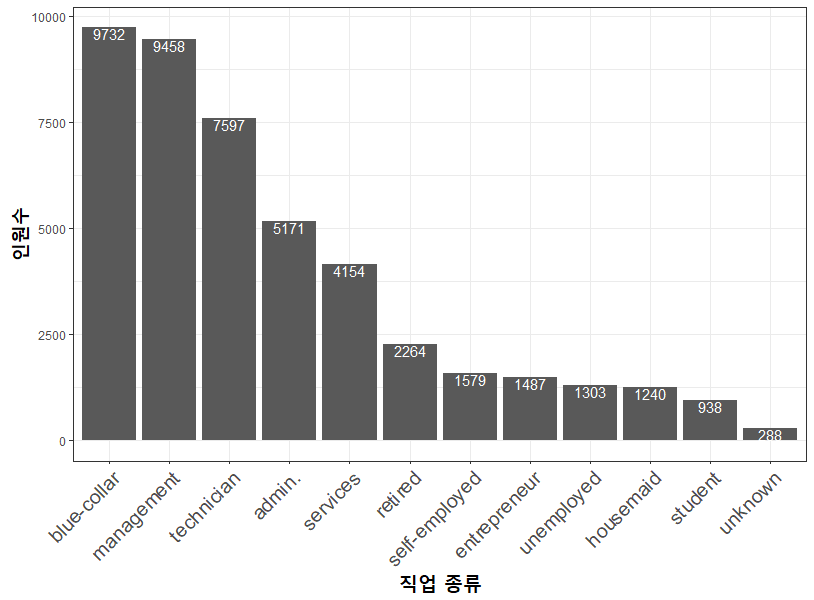
<그림 1‑1> 은행 이용 고객 연령 히스토그램



**<표 1-2> 연령 기술 통계량**

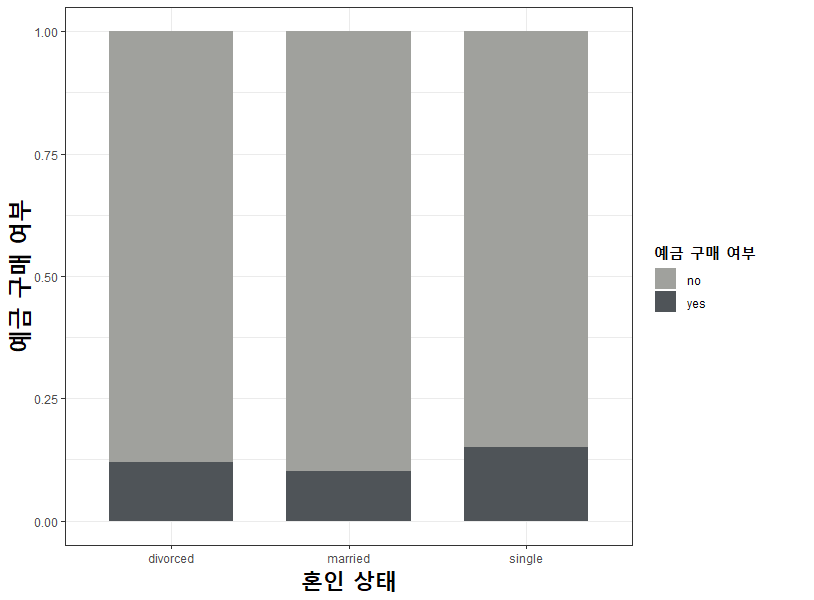
연령의 분포 먼저 확인해 보면 20대에서 60대가 전체 데이터의 97.16%를 차지하는 것을 확인할 수 있다. 가장 많은 30대가 17,687명이 있는 것으로 나타났다. 기술 통계량을 확인해 보면 18세부터 95세까지 분포해 있는 것을 알 수 있었지만 연령별 분포가 고르지 않아 범주화를 시켜주는 것이 좋다고 판단했다.

다음으로 직업 변수(job)를 가지고 그린 막대 그래프는 <그림 1-2>와 같다. 총 12개의 범주로 이루어져 있으며, 노동자(blue-collar)가 9,732명으로 가장 많았으며, unknown을 제외한 범주 중에서 학생(student)이 938명으로 가장 적었다. 여기서 unknown은 데이터 상 결측치는 아니지만 최종 모형을 구축하고 중요 변수들을 해석하는 데에 있어 다소 문제가 있을 것으로 예상되어 결측치로 간주한다. 또한 직업의 범주 개수가 많아 비슷한 직군끼리 대분류로 재범주화가 필요해 보인다.



**<그림 1-2> 은행 이용 고객 직업 막대 그래프**

다음은 혼인 상태(marital)에 대한 변수를 보려고 한다. 혼인 상태 변수 요인별 종속변수 비율을 그린 막대 그래프는 <그림 1-3>과 같다. 총 3개의 범주로 이혼 또는 사별(divorced)이 5,207명, 기혼(married)이 27,214명, 미혼(single)이 12,790명으로 나타났다. 예금 상품 가입 고객 비율은 이혼 또는 사별이 11.95%, 기혼이 10.12%, 미혼이 14.95%로 나타나 이혼 또는 사별과 기혼의 비율이 비슷하게 나타났고, 미혼이 더 높은 비율을 차지하는 것으로 나타났다. 해당 변수를 2개의 범주로 재범주화가 필요해 보인다.



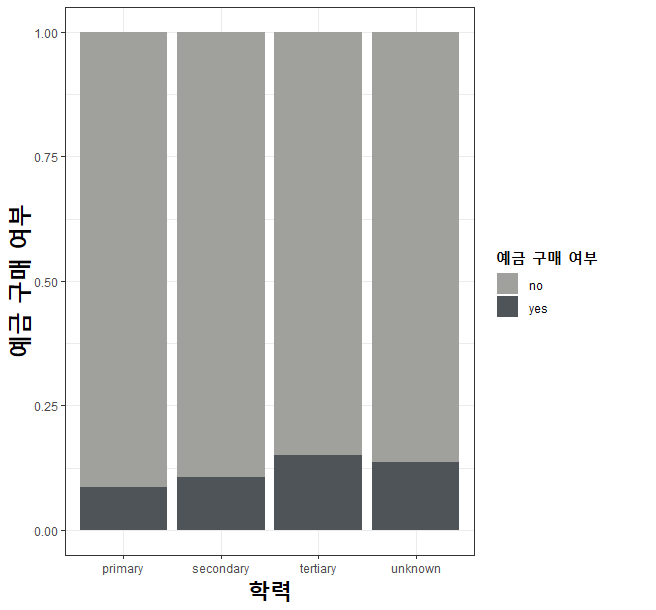
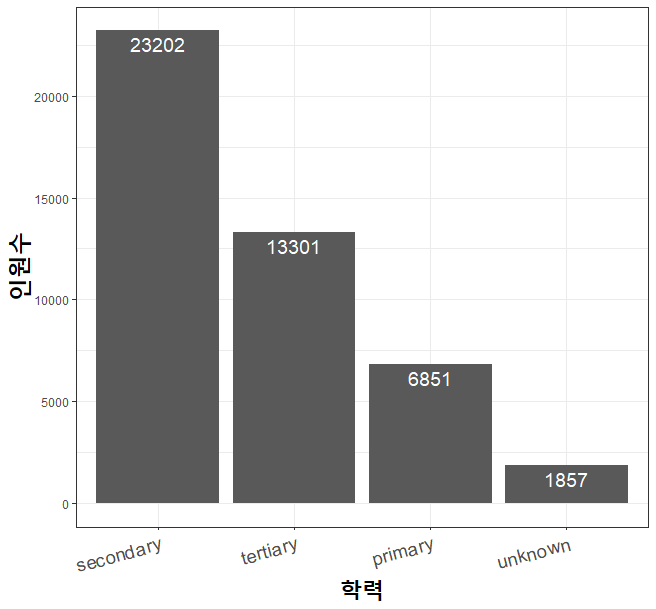
**11.95% 10.12% 14.95%**

**5,207 27,214 12,790**

**<그림 1-3>** **혼인 상태별 예금 구매 여부 비율**

다음은 학력 변수(education)에 대해 보려고 한다. 학력 변수로 그린 막대 그래프는 <그림 1-4>과 같다. 중등교육(secondary)이 23,202명으로 가장 많이 나타났고, unknown을 제외하면 초등교육(primary)이 6,851명으로 가장 적게 나타났다. Unknown을 무응답으로 보고 단순히 제거하기엔 데이터 손실이 크기 때문에 학력별 예금 구매 여부 비율을 보았다. <그림 1-5>에서 확인할 수 있다. 보다시피 학력별 예금 구매 비율이 각각 다르게 나타나는 것을 보아 학력 변수를 제거하는 것보다 unknown을 다른 변수들을 통해 채우는 것이 좋을 것 같다고 판단했다.

**<그림 1-4>** **은행 이용 고객 학력 막대 그래프**

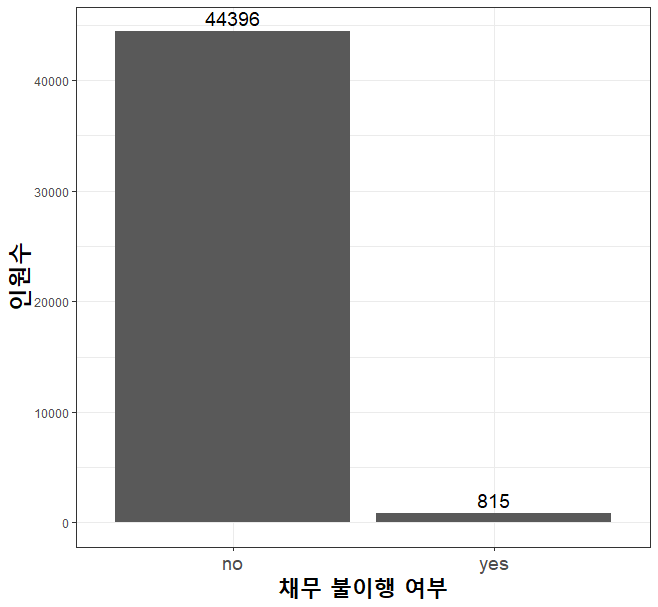


**<그림 1-5>** **학력별 예금 구매 여부 비율**

**8.6% 10.6% 15.0% 13.6%**

다음은 채무 불이행 여부(default)변수를 보면 채무 불이행이 있다(yes)가 815명, 채무 불이행이 없다(no)가 44,396명으로 나타났다. 두 범주의 빈도수가 차이가 많이 나는 것을 볼 수 있다. 종속 변수와의 교차분할표를 그려본 결과, 두 범주의 종속변수 yes 비율이 각각 11.8%, 6.4%로 차이가 나지만 빈도수를 고려해 보았을 때 유의미하다고 볼 수 만은 없어 채무 불이행 변수 제거도 고려해 보아야 한다.

**<그림 1-6>** **채무 불이행 여부 막대 그래프**

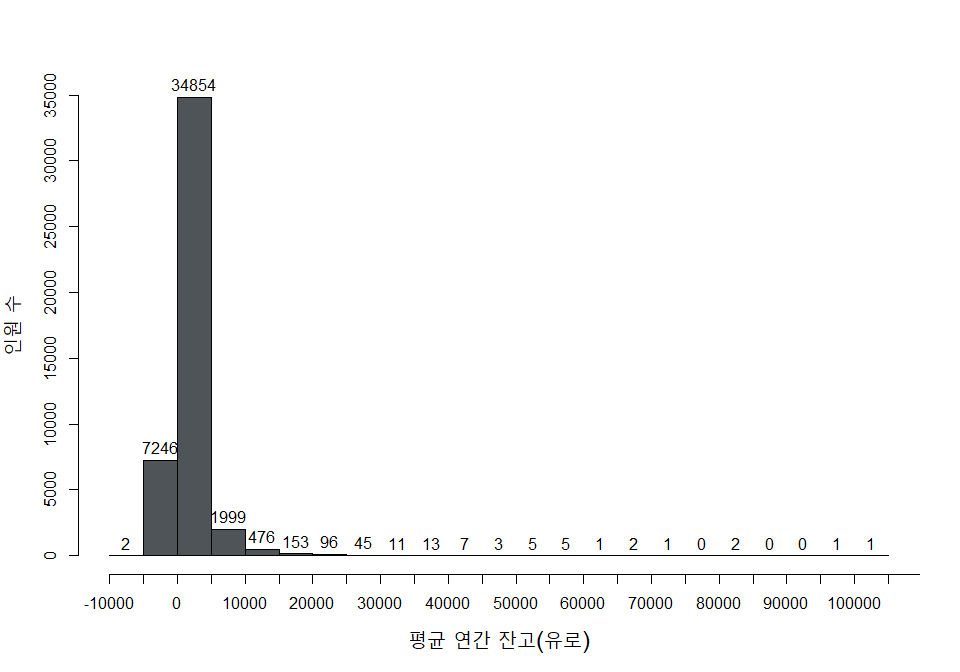


|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **종속 변수** | |
|  |  | **No** | **Yes** |
| **채무 변수** | **No** | 39159  (88.2%) | 5237  (11.8%) |
| **Yes** | 763  (93.6%) | 52  (6.4%) |

**<표 1-3>** **채무 불이행 여부와**

**종속 변수 교차 분할표**

다음으로 평균 연간 잔고(balance)변수는 연속형 변수여서 히스토그램과 기술 통계량을 기술해 보았다. 각각 <그림 1-7>과 <표 1-4>를 참고하면 된다. 히스토그램을 먼저 확인해 보면 0을 중심으로 42,100명이 분포해 전체 고객의 93%가 몰려 있는 것을 알 수 있다. 또한 오른쪽으로 꼬리가 길게 분포해 있는 것을 볼 수 있다. 그리고 기술 통계량을 보면 최저값이 -8,019유로, 최댓값이 102,127유로로 각각 한화로 환산(2008년 기준으로 1EUR=1,617KRW)[[1]](#footnote-1)하면 -12,966,723원과 165,139,359원이다. 2008년도 포르투갈의 1인당 GDP가 24,791.847(USD)[[2]](#footnote-2)이고 대한민국 1인당 GDP가 20,430.64(USD)[[3]](#footnote-3)으로 비슷한 점을 고려하면 평균 연간 예금 잔고로 불가능한 금액들은 아니다. 또한 예금 금액이 마이너스에 상관 없이 숫자가 높을수록 은행 입장에서는



**<표 1-4> 평균 연간 잔고**

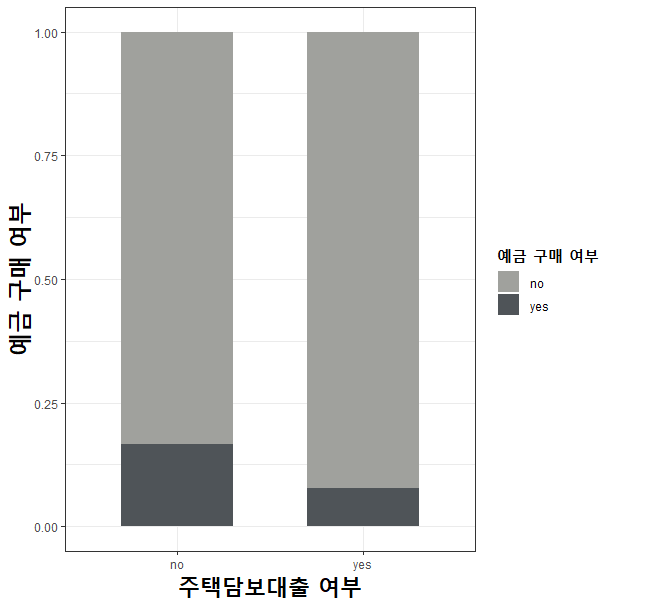
**기술 통계량**

**<그림 1-7> 평균 연간 잔고 히스토그램**

|  |  |
| --- | --- |
| **Minimum** | -8,019 |
| **1st Quan.** | 72 |
| **Median** | 447 |
| **Mean** | 1,360 |
| **3rd Quan.** | 1,421 |
| **Maximum** | 102,127 |

이익구조가 발생하기 좋은 고객들이므로 평균 연간 잔고가 아주 높거나 낮은 고객을 이상치로 제거하는 것보다는 0 주변을 잘 핸들링하고 변환을 해주는 것이 좋겠다고 판단했다.

다음은 주택담보대출 여부(housing)에 관한 변수에 대해서 보려고 한다. 해당 변수는 yes/no로 나누어진 이진 변수라 주택담보대출 여부별로 예금 구매 여부의 비율을 보았다. 그 결과는 <그림 1-8>에서 확인해 볼 수 있다. 주택담보대출은 없다(no)가 20,081명, 있다(yes)가 25,130명으로 다른 이진 변수들에 비해서 차이가 많이 나는 편은 아니다. 각각의 예금 구매 여부 비율은 주택담보대출이 있는 경우 예금 구매 비율이 16.7%, 주택담보대출이 없는 경우 7.7%로 나타나 주택담보대출이 없는 고객들이 예금 구매에 더 적극적인 것을 확인할 수 있다. 해당 변수에는 별다른 결측치가 보이지 않아 전처리가 따로 필요해 보이지 않는다.



**<그림 1-8> 주택담보대출 여부별**

**예금 구매 여부 비율**

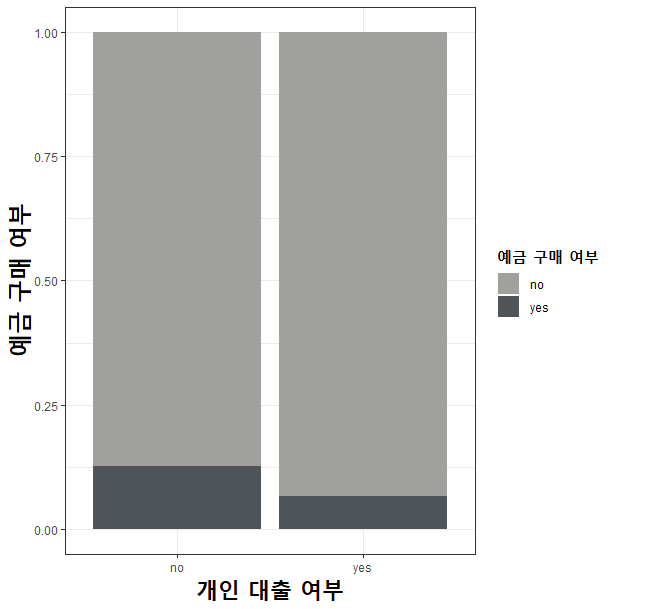
**20,081 25,130**

**16.7% 7.7%**

다음은 개인 대출 여부(loan)도 주택담보대출 변수와 똑같이 파악하려고 한다. 개인 대출 여부별 예금 구매 여부 비율은 <그림 1-9>에서 확인할 수 있다. 개인 대출이 없다(no)가 37,967명, 있다(yes)가 7,244명으로 개인 대출이 없는 인원이 5배 정도 더 많은 것으로 나타났고, 개인 대출 여부별 예금 구매 여부 비율을 확인한 결과 개인 대출이 없는 경우 예금을 구매한 고객의 비율이 12.7%, 개인 대출이 있는 경우 6.7%의 비율을 가지는 것을 확인할 수 있었다. 개인 대출이 없는 고객들이 예금 신청에 더 적극적인 것을 알 수 있다.

**<그림 1-9> 개인 대출 여부별**

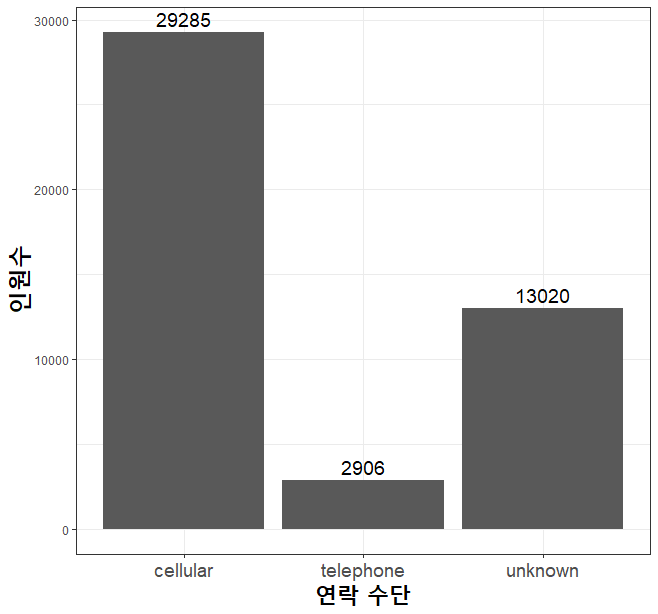
**예금 구매 여부 비율**



**37,967 7,244**

**12.7% 6.7%**

현재 진행된 캠페인의 마지막 연락방식(contact)의 변수를 가지고 그린 막대그래프는 <그림 1-10>과 같다. 휴대전화(cellular)로 연락을 취한 경우는 29,285명이고, 유선전화(telephone) 경우 2,906명이고 나머지 13,020명은 unknown으로 되어 있다. 여기서 unknown을 결측치로 생각하면 해당 데이터가 텔레마케팅 데이터임을 고려하면 휴대전화와 유선전화로 나누어야 하는데 다른 변수를 가지고 채우기 어렵다고 판단되었고, 또한 캠페인 동안의 연락수단이 아닌 마지막 연락수단인 점을 고려하면 해당 변수를 제거하는 것이 좋다고 생각했다.



**<그림 1-10> 연락방식 막대 그래프**

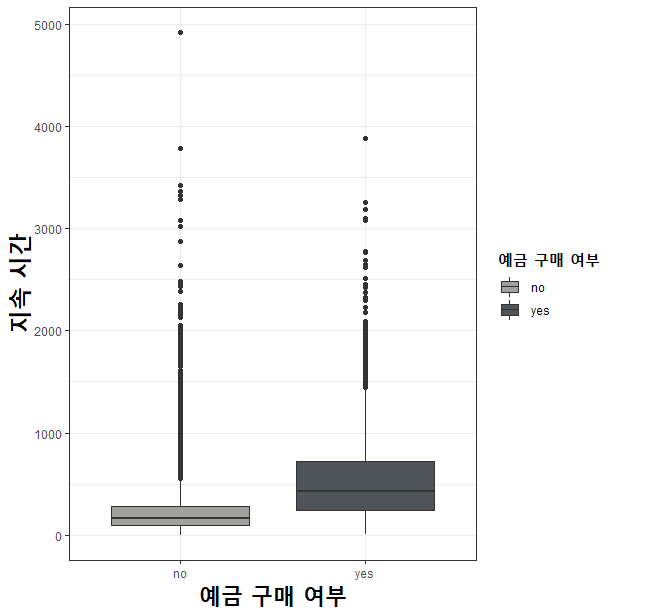
현재 진행된 캠페인에서 마지막으로 연락한 요일(day)과 달(month)변수는 해당 데이터에서 유일한 날짜 데이터이므로 같이 확인해 보기로 한다. 요일은 숫자형 변수로 달은 범주형 변수로 저장되어 있지만 데이터가 월과 요일 순으로 정렬되어 있고 5월부터 시작하여 1년 주기를 지나 11월까지 있다는 점을 발견했고, 데이터를 제공받았을 때 데이터가 2008년 5월부터 2010년 11월까지라는 것과 연관지어 생각하면 데이터가 과거부터 현재 순서대로 작성되어 있다고 판단했다. 따라서 새롭게 년도(year)라는 변수를 만들어 연도별 월별 빈도수를 보았다. 빈도표는 <표 1-5>에 있다.

**<표 1-5> 마지막 연락 연도별 월별 빈도표**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **2008** | **2009** | **2010** | **합계** |
| **Jan** |  | 1,176 | 227 | 1,403 |
| **Feb** |  | 2,296 | 353 | 2,649 |
| **Mar** |  | 258 | 219 | 477 |
| **Apr** |  | 2,718 | 214 | 2,932 |
| **May** | 7,957 | 5,575 | 234 | 13,766 |
| **Jun** | 4,486 | 642 | 213 | 5,341 |
| **Jul** | 6,380 | 207 | 308 | 6,895 |
| **Aug** | 5,215 | 772 | 260 | 6,247 |
| **Sep** | 0 | 282 | 297 | 579 |
| **Oct** | 80 | 438 | 220 | 738 |
| **Nov** | 3,598 | 297 | 75 | 3,970 |
| **Dec** | 13 | 201 |  | 214 |
| **합계** | 27,729 | 14,862 | 2,620 | 45,211 |

연도별 월별로 고르게 분포되어 있는 모습은 아니다. 2008년 5월이 7,957명으로 가장 많이 연락을 했으며, 2008년 합계가 27,729명으로 년도 중에 가장 많이 연락을 했다. 현재 마케팅 연락이 시작한 시점인 2008년 5월에 많은 고객들과 마지막으로 연락을 시도하고 이후에 계속 연락을 시도한 점이 점점 해가 지나면서 마지막 연락 횟수가 줄어든 원인으로 판단된다.

다음으로 마지막 연락 시 지속시간(duration)변수를 확인해 본다. 지속시간 변수에 대해서 기술 통계량과 예금 구매 여부별 박스플랏을 그렸다. 박스플랏과 기술 통계량 표는 각각 <그림 1-11>과 <표 1-6>에 나와 있다.



**<표 1-6> 지속 시간 기술 통계량**

**<그림 1-11> 예금 구매 여부별**

**지속시간 박스플랏**

|  |  |
| --- | --- |
| **Minimum** | 0 |
| **1st Quan.** | 103 |
| **Median** | 180 |
| **Mean** | 258.2 |
| **3rd Quan.** | 319 |
| **Maximum** | 4,918 |

먼저 지속 시간의 기술 통계량을 확인해 보면 단위가 초(sec)인 점을 감안하면 마지막 연락 시에 가장 오래 통화한 시간이 4,918초이다. 약 1시간 21분정도 통화를 한 셈이다. 또한 가장 짧은 통화가 0초로 기록되어 있는데 3명의 관측치가 존재했다. 통화 시간이 0초였던 관측치 3명에 대해서는 캠페인 성공 여부를 확인해 본 결과 모두 알 수 없음으로 나타나 전화를 받고 바로 끊은 것으로 간주하여 이상치로 보지 않고, 지속 시간 변수에 대해서는 데이터 손실율을 최소화하면서 모델 구축에 영향이 가지 않도록 극값 즉, 상위 1%를 제거하는 방향을 고려 보는 것이 좋을 듯 하다.

다음으로는 현재 캠페인을 진행하기 위해 수행된 연락 수(campaign)변수에 대해서 확인해 본다. 이산형 변수여서 예금 구매 여부별 수행된 연락 수를 박스플랏을 그려 확인하려고 했지만 이상치로 간주될 만큼의 최댓값을 가지고 있으며 수행된 연락 수가 1 또는 2에 편중되어 있어 퍼센타일 표를 그렸다. 표는 <표 1-7>와 같다.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **퍼센타일** | **1%** | **2.5%** | **5%** | **25%** | **50%** | **75%** | **95%** | **97.5%** | **99%** | **100%** |
| **연락 수** | 1 | 1 | 1 | 1 | 2 | 3 | 8 | 11 | 16 | 63 |

**<표 1-7> 캠페인 진행 중 수행된 연락 수의 퍼센타일 표**

표를 확인해 보면 현재 캠페인을 진행하기 위해 수행된 연락 수 대부분이 10번 이내인 것을 볼 수 있다. 실제로 수행된 연락 수가 1번인 경우가 전체의 38%를 차지했고, 2번까지는 66%였다. 4번 이상이 25%를 차지했으며, 가장 많이 수행된 연락 수는 63번이다. 캠페인 기간이 2008년 5월부터 2010년 11월까지 31개월인 점을 감안하면 평균적으로 월 2회의 연락을 수행한 셈이다. 현실 불가능한 이상치는 아니지만 위의 지속시간(duration)변수와 마찬가지로 모델 구축 시 많은 영향을 미칠 것으로 판단되어 이 변수 또한 상위 1%의 변수를 제거하는 방법을 고려해 볼 필요가 있어 보인다.

다음으로 볼 변수는 이전 캠페인에서 마지막으로 고객에게 연락 후 경과한 일 수(pdays)이다. 해당 변수도 위의 수행된 연락 수(campaign)와 마찬가지로 퍼센타일 표를 그려보았다. 표는 <표 1-8>를 보면 된다.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **퍼센타일** | **1%** | **2.5%**  **<표 1-8> 마지막으로 고객에게 연락 후 경과일수의 퍼센타일 표** | **5%** | **25%** | **50%** | **75%** | **95%** | **97.5%** | **99%** | **100%** |
| **연락 수** | -1 | -1 | -1 | -1 | -1 | -1 | 317 | 354.75 | 370 | 871 |

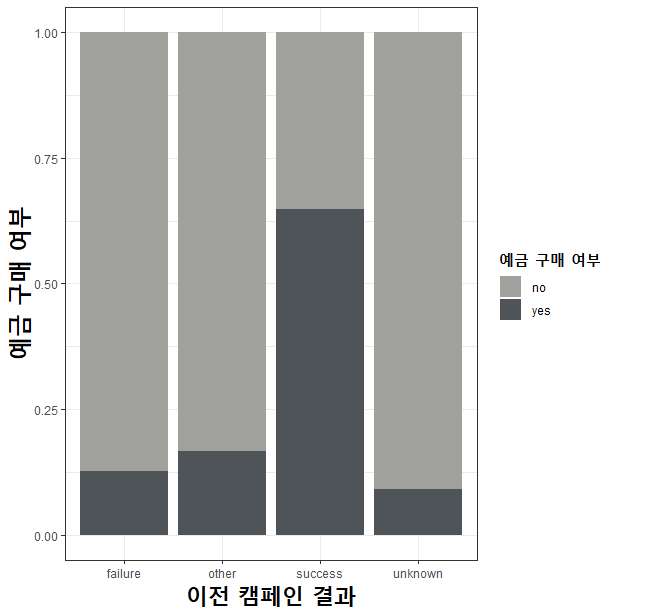
표를 보면 -1이 상당 수를 차지하고 있는데 이 고객들은 이전 캠페인에서 대상이 아니었던 사람들이다. 여기서도 31개월의 캠페인 기간을 고려하면 871일이 경과한 점은 현실성이 있는 데이터라고 볼 수 있다. -1인 고객 수가 36,954명으로 전체 고객의 82%가량을 차지하고 있어 이 부분에 있어서는 잠재고객과 기참여고객으로 데이터셋을 나누어 따로 모델을 구축할 필요가 있어 보인다.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **퍼센타일** | **1%** | **2.5%**  **<표 1-9> 이전 캠페인 진행 중 수행된 연락 수의 퍼센타일 표** | **5%** | **25%** | **50%** | **75%** | **95%** | **97.5%** | **99%** | **100%** |
| **연락 수** | 1 | 1 | 1 | 1 | 2 | 4 | 9 | 12 | 23 | 275 |

다음은 과거 캠페인 진행하기 위해 수행된 연락 수(previous)변수로 앞서 현재 캠페인에서 수행된 연락 수(campaign)변수와 비슷한 느낌의 변수이다. 그러나 위에서 보다시피 전체 82%의 고객이 잠재고객이므로 이전 캠페인에서 수행된 연락 수 대부분이 0을 차지하고 있는 모습을 볼 수 있다. 따라서 이전 캠페인 대상이었던 고객들을 대상으로 변수를 퍼센타일 표로 확인해 보기로 한다. 표는 <표 1-9>와 같다.

표를 보면 <표 1-7>에서 나타났던 현재 캠페인 진행 중 수행된 연락 수와 비슷한 분포를 보이고 있으나 최댓값이 275로 나타났다. 이전 캠페인을 진행한 진행기간을 알 수가 없어 현실 가능한 값인지는 알 수 없지만 두 번째로 가장 많이 수행된 연락 수가 58번, 세 번째가 55번인 것을 보면 입력 오류로 인한 이상치라고 생각된다. 해당 관측치 제거가 필요해 보인다.

다음으로 마지막 독립변수인 이전 캠페인 결과(poutcome)변수를 보려고 한다. 이전 캠페인 결과별로 예금 구매 여부 비율을 그린 <그림 1-12>를 보면 이전 캠페인 결과는 각각 실패(failure)가 4,901명, 성공(success)이 1,511명, 그 외(other)가 1,840명으로 나왔고, unknown이 36,959명으로 나타났다. 여기서 unknown의 수가 현재 캠페인에서 마지막 연락 후 경과일수(pdays)변수에서의 -1의 수와 비슷해 비교해 본 결과 5명의 차이가 있었지만 36,954명의 잠재고객이 모두 unknown변수에 속해 있는 것을 확인했다. 따라서 저자는 차이가 있는 5명의 고객을 이상치로 간주하여 제거하고 이전 캠페인의 결과 unknown인 고객은 잠재고객층으로 분류할 것이다. 또한 이전 캠페인에 참여하여 예금 상품을 신청하지 않은 즉, 실패(failure)고객 중 12.6%가 현재 캠페인에서 예금 상품을 신청한 것으로 나타났다. 그리고 그 외(other)고객 중 16.7%가 예금 신청을, 성공(success)고객 중 64.7%가 현 캠페인에서 예금 신청을 한 것을 볼 수 있다. 이전 캠페인에서 성공한 고객층에서 현재 캠페인 예금 신청 비율이 높은 것을 보면 기존에 캠페인에 호의적이었던 고객들이 지속적으로 캠페인에 호의적인 반응을 보인다는 것을 알 수 있다. 이는 앞으로도 캠페인에 성공했던 고객들을 중심으로 캠페인을 적극적으로 활용하면 예금 신청 고객을 늘릴 수 있다는 것으로 해석될 수 있다.



**<그림 1-12> 이전 캠페인 결과별**

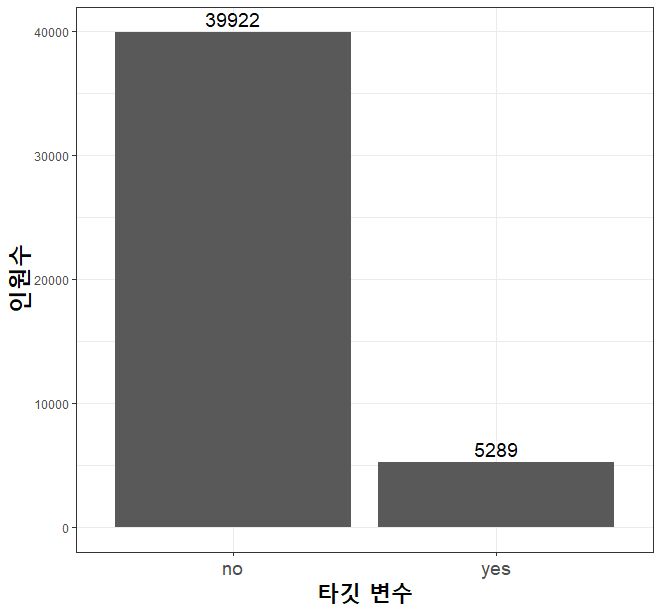
**예금 구매 여부 비율**

**12.6% 16.7% 64.7% 9.2%**

**4,901 1,840 1,511 36,959**

마지막으로 종속변수인 예금 상품 구매 여부(y)변수를 보면 이미 데이터를 제공받을 때 종속변수의 불균형이 있음을 전달받았기 때문에 불균형 정도를 파악하기 위해 막대 그래프를 그렸다. 막대 그래프는 <그림 1-13>을 확인하면 된다. 그림을 보면 39,922명의 고객이 예금 상품을 신청하지 않은 것으로 나타났다. 전체 고객의 88%이므로 이 상태로 모델을 구축할 경우 불균형으로 인해 실제로 예금에 가입할 고객임에도 불구하고 가입하지 않을 것이라고 예측할 가능성이 높아지게 된다. 따라서 예금을 가입하지 않은 고객(no) 중 샘플링을 통해 예금을 가입한 고객(yes) 수에 맞추는 과정이 필요하다.

**<그림 1-13> 타깃 변수 막대 그래프**



여기까지 포르투갈 금융 기관의 텔레마케팅 데이터의 변수 파악이 끝이 났다. 앞서 변수를 탐색한 토대로 다음 절에서 변수들의 결측치(일부 unknown)를 채우고, 이상치를 처리할 예정이다.

**제 2 절 데이터 정제 및 처리**

앞서 1절 변수 파악에서 진행하면서 전처리가 필요하다고 느낀 변수 또는 관측값에 대해서 본 절에서 처리할 것이다. 결측치라고 판단한 몇몇 변수들의 unknown을 채우기 전 먼저 임의로 제거할 변수와 이상 관측치들을 제거한 후 결측치를 채우고 잠재고객과 이전에 캠페인에 참여한 경험이 있는 고객층을 나누어 모델을 구축하기 전 데이터 준비를 마칠 것이다.

1. **변수 제거 및 이상 관측치 제거**

먼저 unknown의 수가 많을 뿐만 아니라 채우는 것도 어렵다고 판단된 변수를 제거한다. 포르투갈 텔레마케팅 데이터에서는 현재 캠페인에서 고객에게 마지막으로 연락한 수단(contact)변수가 unknown이 13,020개로 전체 데이터의 28%가량을 차지했으며 다른 변수를 통한 채우는 것도 어렵고 또한 마지막 연락 수단에 대한 정보는 고객들과 꾸준히 연락했던 수단이 아닌 마지막에만 초점이 맞춰져 있는 변수이므로 해석이 다소 어렵다고 느껴져 과감히 변수를 제거하기로 하였다.

다음으로 위에서 이전 캠페인 결과(poutcome)에서 unknown이었던 36,959명을 잠재고객으로 파악했고, 현재 캠페인에서 마지막 연락 후 경과일수(pdays)변수에서 -1인 36,954명이 이전 캠페인에 참여하지 않은 똑 같은 잠재고객이라고 했는데 이 둘의 차이가 있는 5명을 잘못 기록된 이상치로 판단하고 제거하기로 하였다.

위에서 변수 파악할 때는 이상치로 생각하지 못했던 관측치들이었는데 여러 변수를 복합적으로 보니 이상치로 판단된 관측치들이 있었다. 나이(age)와 직업(job), 학력(education)에서 학력이 초등교육(primary)이고 직업이 학생(student)인데 나이가 18세부터 36세인 고객이 44명이 있었다. 초등교육이라함은 우리나라에서 초등학교 졸업이다. 이 상태에서 직업이 학생이라고 하면 중학교 재학중이거나 중등 검정고시생이어야 한다는 얘기다. 이 부분은 입력 오류라고 판단하고 44명에 대해서 모두 제거하기로 하였다. 이런 비슷한 경우로 나이가 40세 미만인 경우에서 직업을 은퇴(retired)라고 기재한 고객이 2명이 나타났다. 직업에 고용되지 않음(unemployed)이 있음을 고려하면 이 부분도 잘못 기재된 것으로 판단하여 제거하였다.

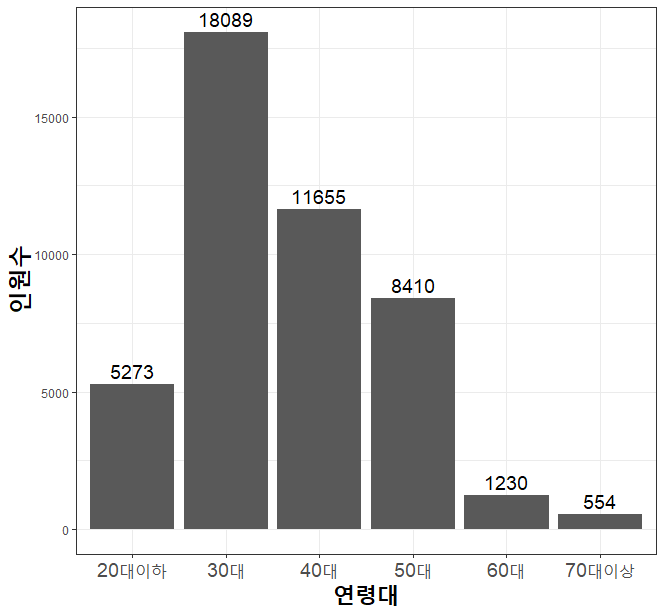
다음은 변수 해석상 이상치는 아니었지만 값들의 차이가 커 극값을 제거하기로 고려한 두 개의 변수(campaign, previous)에 대해서 극값을 처리한다. 먼저 campaign변수는 현재 캠페인 진행 중 연락 횟수이므로 데이터가 0이상인 값들을 갖게 되는데 <표 1-7>를 참고하면 최댓값이 63회로 99%값인 16과 차이가 너무 커 좋은 모델을 생성하는데 영향을 미칠 것으로 보아 99%보다 큰 1%의 자료에 대해서 제거하기로 했다. 또한 previous변수는 과거 캠페인 진행 중 연락 횟수인데 <표 1-9>를 참고하면 잠재고객층을 제외한 기존 참여고객 중에서 최댓값이 275회인 것에 비해 두 번째와 세 번째 값은 각각 58회, 55회로 차이가 심해 입력 오류로 판단하고 1개의 관측치에 대해서 제거해주었다. previous변수도 campaign과 마찬가지로 극값 1%를 제거하고 싶었으나 극값 1% 즉, 이전 캠페인에서 연락한 횟수가 많았던 상위 1% 고객들 중에서 예금 가입 비율이 높게 나와 상대적으로 비가입 고객보다 적은 수를 가지고 있는 가입 고객 데이터 손실이 클 것으로 예상돼 한 개의 관측치만 제거하기로 했다.

마지막으로 직업(job)변수에 존재하고 있던 unknown의 수는 288명으로 전체 데이터의 1% 이내의 관측치이기 때문에 데이터 손실이 적어 제거하기로 했다.

변수와 이상 관측치를 제거하고 남은 데이터는 변수가 16개, 관측치는 44,401개로 구성되어 있다.

1. **범주형 변수 재범주화**

예금 가입에 관련된 고객층을 다루다 보니 주로 수입원이 많거나 대출 신청 가능성이 높아보이는 고객들인 20대에서 50대가 데이터의 주를 이루고 있었다. 이에 연령(age)변수를 범주화하여 보는 것이 좋다고 판단했다. 따라서 같은 나이대를 묶어주되 10대와 20대를 같이 묶어 20대 이하로, 70대와 80대, 90대를 70대 이상으로 묶어 총 6개의 범주로 나타내 주었다. 변환 후의 연령 변수를 가지고 그린 막대 그래프는 <그림 1-14>과 같다. 연령대별로 비율이 비슷하지는 않지만 비교적으로 잘 묶인 것을 확인할 수 있었다.



**<그림 1-14> 변환 후 연령 변수 분포**

직업(job)변수는 범주가 많고 비슷한 범주로 묶일 수 있는 직업들이 존재하고 있어 범주의 개수를 줄이는 재범주화 과정을 거치려고 한다. 기존의 직업 변수에 있던 범주는 11개였으나 미국의 표준직업분류표(SOC)[[4]](#footnote-4)와 대한민국의 표준직업분류(KSCO)[[5]](#footnote-5)를 참고하여 나누어 준 범주는 7개이다. 재범주화는 <표 1-10>를 참고하면 된다. 재범주화 이후 기업가(entrepreneur)와 자영업자(self-employed)가 하나로 묶여 기업가 범주가 되었고, 서비스업(services)과 가정부(housmaid)가 핑크칼라(pink-collar)라는 범주로 묶였고, 나머지 학생(student), 은퇴(retired), 미취업(unemployed)가 미취업으로 묶였다. 학생은 직종이 달랐지만 수입원이 없다는 점에서 위와 같이 묶었다. 7개의 범주로 묶인 이후 각 범주의 백분율이 비슷하지는 않지만 상대적으로 빈도수가 작은 범주의 직업끼리 묶여 묶이기 전보다는 차이가 심하게 나지는 않는다.

**<표 1-10> 직업변수 재범주화**

|  |  |  |  |
| --- | --- | --- | --- |
| **재범주화 전** | **재범주화 후** | **빈도** | **백분율** |
| admin | admin | 5,127 | 11.5 |
| management | management | 9,344 | 21.0 |
| entrepreneur | entrepreneur | 3,042 | 6.9 |
| self-employed |
| blue-collar | blue-collar | 9,615 | 21.7 |
| technician | technician | 7,509 | 16.9 |
| services | pink-collar | 5,350 | 12.0 |
| housemaid |
| student | unemployed | 4,414 | 9.9 |
| retired |
| unemployed |

1. **새로운 변수 생성 및 결측치 처리**

앞서 우리는 데이터셋에 두 개의 날짜 변수(day, month)가 있는 점과 데이터가 날짜 순으로 기록되어 2008년도부터 2010년도 순으로 저장되어 있다는 점을 알았다. 이에 따라 month데이터가 12월(Dec)에서 1월(Jan)로 바뀌는 지점을 기준으로 새롭게 year라는 연도 변수를 생성해 주었다. 그리고 여기서 day와 month는 마지막으로 연락한 날짜를 기록한 변수이므로 자체적으로는 예금 신청에 큰 영향을 미치지 않을 것으로 보여 마지막으로 연락한 날짜가 주중인지 주말인지에 대한 변수를 새롭게 만들었습니다. 그리고 주중 중에서도 국가 공휴일[[6]](#footnote-6)을 반영하여 휴일 또한 주말과 같은 범주로 묶어 새로운 day2라는 변수를 생성해 주었습니다.

이미 데이터를 받았을 당시에 결측치가 따로 없음을 전달 받았지만 변수 파악 중 unknown으로 기록되어 있는 몇몇 데이터를 확인했고, 그 중 학력(education)변수에서의 unknown을 결측치로 판단하고 처리하려고 한다. 결측치를 처리하는 방법에는 여러 가지 방법이 있지만 저자는 그 중 다른 변수를 이용하여 결측치를 예측하는 모델을 만들어 채우려고 한다. 모델은 의사결정나무(Decision Tree)를 사용하여 만들려고 한다. 설명변수로 사용될 변수들은 인구통계학적 변수들 중 예측력을 높게 하는 변수들을 선택한다. 모델을 구축하기 전 범주별로 불균형이 있어 샘플링을 통해 가장 작은 범주였던 초등교육(primary)에 맞춘 뒤 예측률 검정을 위해 Training Data Set과 Test Data Set을 7:3으로 나누어 Training Data Set으로 모델을 구축한 결과 예측률이 가장 좋은 모델로 재범주화된 연령(age)과 혼인 상태(marital), 재범주화된 직업(job)변수가 선택되었다. 예측률은 69.03%가 나왔고 분류 결과표는 <표 1-11>를 보면 된다. 구축된 의사결정나무 모델을 토대로 결측치라고 판단했던 unknown을 가진 관측치들에 대입하여 각각 학력을 채워주었다.

**<표 1-11> 학력 변수 의사결정나무 분류 결과표**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **69.03%** | **실제값** | | | |
| **예측값** |  | **primary** | **secondary** | **tertiary** |
| **primary** | 1,623 | 703 | 114 |
| **secondary** | 212 | 1,092 | 401 |
| **tertiary** | 202 | 241 | 1,459 |

1. **데이터셋 분리**

저자는 이전 캠페인 관련 변수들을 가지고 잠재 고객과 과거 캠페인 참여 고객으로 나뉜다는 것을 발견했다. 두 고객층을 하나로 묶어서 모델을 구축해도 되지만 두 고객층을 분리해서 각각 모델을 구축하면 이전 캠페인 결과(poutcome)변수에서의 unknown을 따로 처리하지 않아도 되고 두 그룹이 다른 모델이 생성되어 해석에 더 용이하다고 판단해 나누기로 했다. 이전 캠페인 결과(poutcome)변수를 가지고 unknown이었던 고객들을 잠재고객, 그 외 고객들을 기참여고객으로 나누었다. 그 결과 잠재고객은 36,204명, 기참여고객은 8,197명으로 각각 나뉘었다. 이후 잠재고객 데이터에서는 과거 캠페인 관련 변수들(pdays, previous, poutcome)이 필요가 없어지게 되어 제거하여 종속변수 포함 13개의 변수가 남았다. 이후부터 모델을 구축하기 위한 변수 변환과 변수 선택 과정과 모델 구축 과정을 잠재고객 데이터와 기참여고객 데이터에서 각각 실행하여 각각의 모델을 구축할 예정이다.

1. **변수 변환 및 최종 변수 선택**

모델을 구축하기 이전 독립변수들에 대한 모든 전처리가 끝나고 분석자의 주관에 따라 데이터셋을 두 개의 그룹으로 나누었다. 이제 모델을 구축하기에 앞서 두 그룹으로 나누어진 데이터셋에서 변수 선택을 위해 독립변수가 종속변수에 미치는 영향을 검정하고 연속형 변수의 경우 변환과 검정을 통해 최종적으로 가장 영향이 큰 변환 변수를 선택할 것이다.

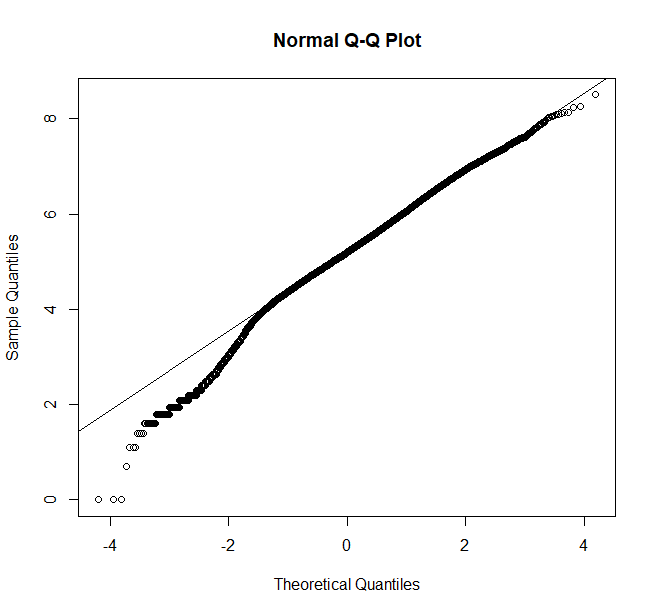
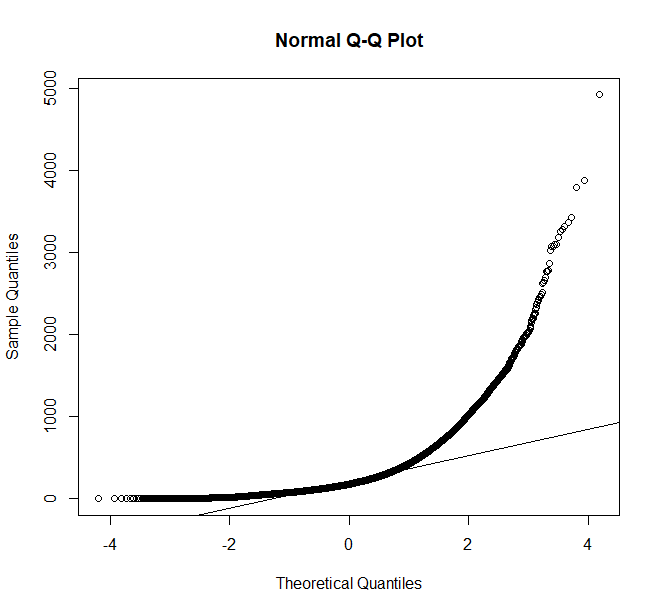
위 과정을 잠재고객층 먼저 확인해 본다. 검정에 있어서 분석자는 유의수준을 0.05로 설정한다. 또한 혼인 상태(marital)와 학력(education)은 변수가 3개 있기 때문에 본페로니 교정(Bonferroni Correction)을 통하여 각각 3번의 모비율 검정을 할 예정이고 나머지 범주형 자료들은 범주가 2개씩이므로 그냥 모비율 검정을 진행하면 된다. 모든 검정의 결과는 <표 1-12>를 참고하면 된다.

**<표 1-12> 잠재고객 범주형 변수 검정 결과**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **검정 변수** | |  | **자유도** | **p-value** |
| **marital** | **Divorced vs Married** | 33.27 | 1 | <0.001 |
| **Divorced vs Single** | 10.348 | 1 | 0.0013 |
| **Married vs Single** | 173.34 | 1 | <0.001 |
| **education** | **Primary vs Secondary** | 10.67 | 1 | 0.0011 |
| **Primary vs Tertiary** | 78.224 | 1 | <0.001 |
| **Secondary vs Tertiary** | 70.93 | 1 | <0.001 |
| **default** | **Yes vs No** | 7.1661 | 1 | 0.0074 |
| **housing** | **Yes vs No** | 393.01 | 1 | <0.001 |
| **loan** | **Yes vs No** | 91.299 | 1 | <0.001 |
| **day** | **주중 vs 주말** | 2.5684 | 1 | 0.109 |

검정 결과 혼인 상태(marital)와 학력(education)은 본페로니 교정에 의해 0.05/3 = 0.0167 기준으로 보았을 때 모든 검정에서 귀무가설을 기각하여 모비율에 차이가 있다고 볼 수 있다. 채무 불이행(default)과 주택담보대출 여부(housing), 개인대출 여부(loan)도 모두 유의수준 0.05보다 p-value가 작아 모비율에 차이가 있다고 볼 수 있다. 그러나 새롭게 만든 마지막 연락한 날의 휴일 여부(day)에서 p-value가 0.109로 귀무가설 채택 모비율이 비슷하다고 볼 수 있다. 참고로 잠재고객 36,204명 중에 39명에게만 휴일에 마지막으로 연락한 것으로 나타나 모델 구축 시에 day변수는 제외하기로 결정했다.

다음으로 마지막 통화 시간(duration)변수는 초단위로 기록된 이산형 변수이기 때문에 0이상의 값을 가지고 있다. 따라서 정규성을 만족하는지 확인하고 만족하지 않으면 만족하는 방향으로 변환을 실시할 예정이다. 먼저 정규성을 만족하는지 확인하기 위해 duration에 대해서 Q-Q plot을 그려본 결과 정규성을 만족하지 않는 것으로 나타나 0값을 가지고 있어 1을 더해주고 로그변환을 실시한 이후 다시 그려본 Q-Q plot에서 정규성을 만족하는 것으로 나타났다. 변환 전후의 Q-Q plot은 <그림 1-15>을 보면 된다.



**<그림 1-15> 잠재고객 마지막 통화 시간 변수 변환 전후 Q-Q plot**

**로그변환 전**

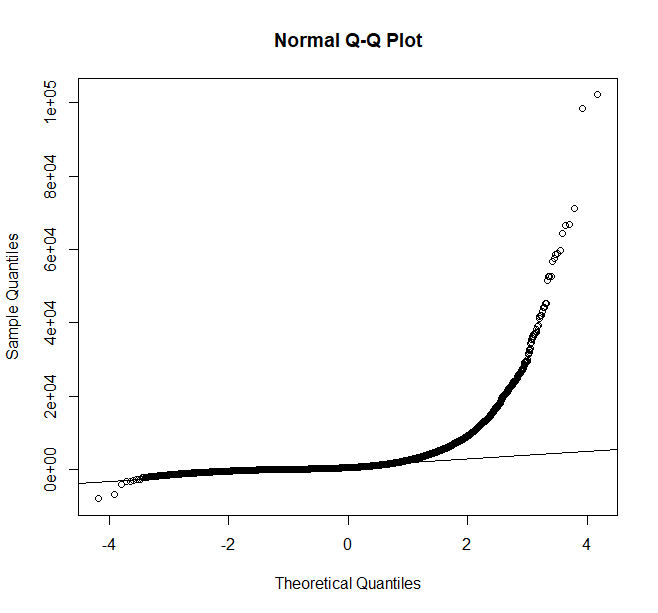
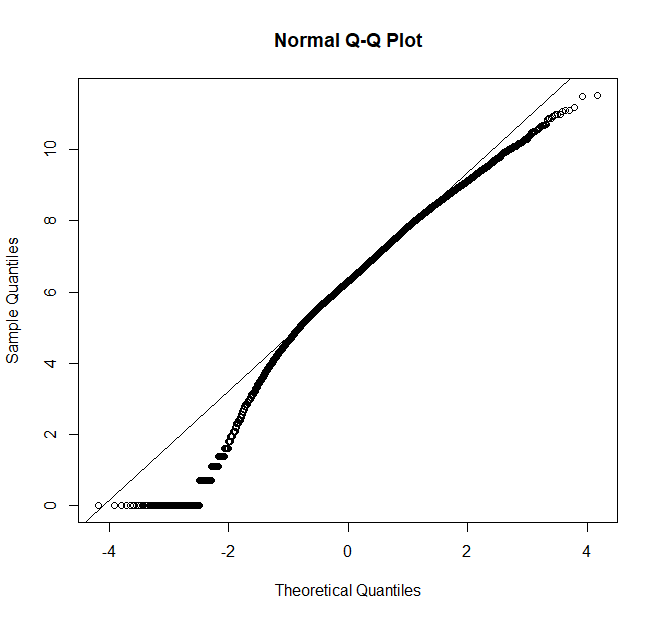
**로그변환 후**

다음으로 연간 평균 잔고(balance)에 대해 큰값을 갖는 관측치들이 몇몇 개가 있어 변수 변환을 해주어야 하는데 음의 값과 0, 양의 값을 가지고 있어 최소값보다 큰 값을 더해 모두 양수로 보내주어야 하지만 이 데이터가 은행 고객들의 대한 데이터이고 은행에서의 잔고는 수익에 직결되는 변수이다. 그리고 마이너스 잔고 또한 이자라는 수익이 발생하는 수익 구조를 가지고 있기 때문에 단순히 값들을 양수로 옮기는 방법보다는 0의 값을 갖는 고객들은 거래를 하지 않는 고객이라고 판단하여 제거 후 절댓값을 취해 음수값을 양수로 보내는 변환(balance)을 시켜준 뒤 로그 변환(balance0)과 제곱근 변환(balance1), 역변환(balance2)을 하여 로지스틱 회귀 모형을 돌려 가장 종속 변수에 영향을 많이 미치는 변수를 최종 변수로 선택한다. 로지스틱 모형을 구축한 결과는 <표 1-13>와 같다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **Z value** | **p-value** |
| **(Intercept)** | -3.427 | 0.1877 | -18.256 | <0.001 |
| **balance** | -0.000016 | 0.00002 | -0.784 | 0.4333 |
| **balance0** | 0.1883 | 0.04504 | 4.182 | <0.001 |
| **balance1** | -0.000359 | -0.004487 | -0.080 | 0.9361 |
| **balance2** | 1.007 | 0.337 | 2.988 | 0.0028 |

**<표 1-13> 잠재고객 평균 연간 잔고 변수 변환 로지스틱 모형**

로지스틱 모형 구축 결과 변환 변수 중에는 로그 변환 변수(balance0)와 역변환 변수(balance2)의 p-value가 유의수준 0.05보다 작은 것으로 나타났다. 그 중에서도 p-value가 더 작은 로그 변환 변수가 종속 변수에 더 큰 영향을 미칠 것으로 보여 최종 변환 변수로 선택했다. 선택된 로그 변환 변수를 가지고 변환 전과 후의 Q-Q plot을 그려 정규성도 검정했다. 변환 후의 Q-Q plot에서 정규성을 만족하는 것 같았다. 변환 전후의 Q-Q plot은 <그림 1-16>에서 확인하면 된다.



**<그림 1-16> 잠재고객 평균 연간 잔고 변수 변환 전후 Q-Q plot**

**로그변환 전**

**로그변환 후**

평균 연간 잔고 변수를 끝으로 잠재고객 데이터의 독립 변수들의 변환이 끝이 났다. 마지막으로 종속 변수의 불균형을 처리하기 위해 데이터 제공자의 요구대로 빈도수가 큰 범주에서 샘플링을 하여 작은 범주 빈도수에 맞출 것이다. 현재 종속 변수의 no가 30,072명, yes가 3,151명으로 큰 범주인 no에서 10%비율로 샘플링을 하였다. 또한 샘플링에 따르는 가중치 변수를 새로 만들어 no범주에는 9.5만큼의 가중치를 yes는 1로 두었다. 그 결과 잠재고객 데이터는 종속 변수의 비율이 5:5로 맞춰지고 총 관측치는 6,302개로 되었다.

잠재고객에서의 선택된 변수와 각 변환된 변수에 관련된 정보는 아래 <표 1-14>에서 확인하면 된다.

**<표 1-14> 잠재고객 최종 선택된 변수**

|  |  |
| --- | --- |
| **변수명** | **변수 변환 설명** |
| age | 나이대로 범주화 |
| job | 표준직업분류 참고해서 재범주화 |
| marital | 그대로 사용 |
| education | 그대로 사용 |
| default | 그대로 사용 |
| balance | 절댓값 이후 로그 변환 변수 사용 |
| housing | 그대로 사용 |
| loan | 그대로 사용 |
| month | 그대로 사용 |
| duration | 로그변환 후 사용 |
| campaign | 그대로 사용 |

잠재고객과 마찬가지로 이전에 캠페인에 참여했던 고객들에 대해서도 똑 같은 변수 변환 및 선택 과정을 거친다. 먼저 범주형 변수들에 대해서 모비율 검정을 실시했다. 기참여고객층의 모비율 검정에 대한 검정결과는 <표 1-15>와 같다.

기참여고객에 대한 범주형 변수 검정 결과 혼인 상태(marital)에서 이혼(divorced)과 기혼(married)이 p-value가 0.1503으로 본페로니 교정을 통한 유의수준 0.0167에서 귀무가설을 채택 모비율의 차이가 없는 것으로 나타났다. 따라서 혼인 상태의 3개의 범주를 이혼과 기혼을 묶은 결혼 경험 있음과 미혼(single)으로 나누어 결혼 경험 유무에 관한 변수로 재범주화한다. 다음으로 학력(education)변수에서도 초등교육(primary)과 중등교육(secondary)이 p-value가 0.0119로 본페로니 교정을 통한 유의수준 0.0167에서 모비율의 차이가 없는 것으로 나타났다. 따라서 학력 또한 고등학생 과정까지의 초등교육과 중등교육을 묶고, 고등과정(tertiary)을 따로 두어 3개의 범주에서 2개의 범주로 재범주화한다. 채무 불이행(default)과 주택담보대출 여부(housing), 개인대출 여부(loan) 모두 유의수준 0.05에서 귀무가설을 기각하여 변수를 그대로 사용하기로 하였다.

**<표 1-15> 기참여고객 범주형 변수 검정 결과**

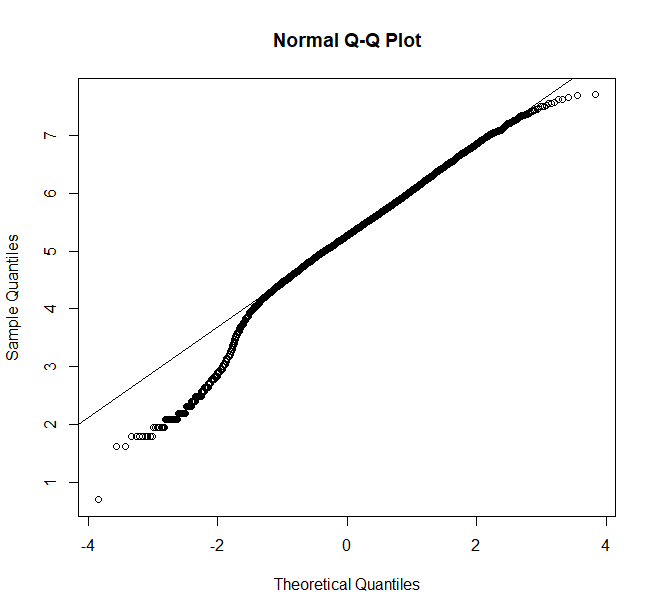
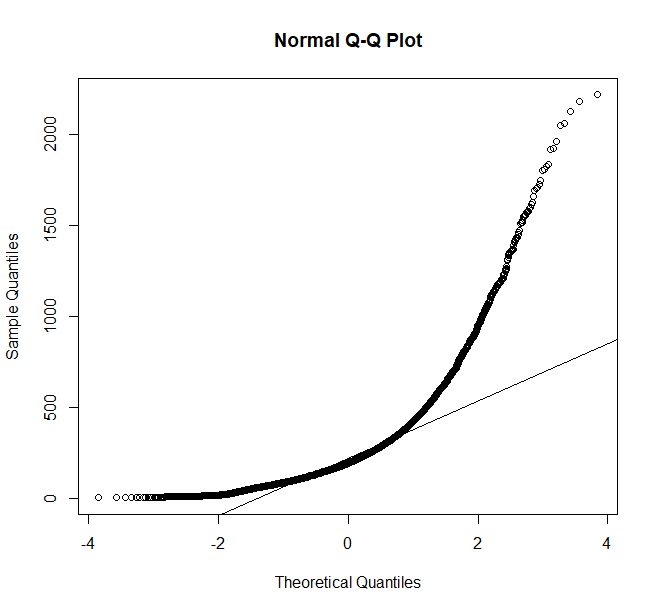
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **검정 변수** | |  | **자유도** | **p-value** |
| **marital** | **Divorced vs Married** | 2.0694 | 1 | 0.1503 |
| **Divorced vs Single** | 12.416 | 1 | <0.001 |
| **Married vs Single** | 12.214 | 1 | <0.001 |
| **education** | **Primary vs Secondary** | 6.3292 | 1 | 0.0119 |
| **Primary vs Tertiary** | 63.298 | 1 | <0.001 |
| **Secondary vs Tertiary** | 81.086 | 1 | <0.001 |
| **default** | **Yes vs No** | 5.7325 | 1 | 0.0167 |
| **housing** | **Yes vs No** | 821.33 | 1 | <0.001 |
| **loan** | **Yes vs No** | 110.13 | 1 | <0.001 |
| **day** | **주중 vs 주말** | 1.2613 | 1 | 0.2614 |
| **poutcome** | **Failure vs Other** | 18.795 | 1 | <0.001 |
| **Failure vs Success** | 1658.8 | 1 | <0.001 |
| **Other vs Success** | 797.8 | 1 | <0.001 |

주말 여부(day)변수에서는 p-value가 0.2614로 잠재고객과 마찬가지로 모비율에 차이가 없는 것으로 나타나 변수를 제외하고 모델을 구축하기로 한다. 추가로 기참여고객에는 이전 캠페인 결과(poutcome)변수가 있어 검정한 결과 세 범주 모두 본페로니 교정에 의한 유의수준 0.0167보다 p-value가 낮게 나와 모비율에 차이가 있는 것으로 나타났다.

다음으로 마지막 통화 시간(duration)변수는 정규성을 만족하는지 확인하기 위해 duration에 대해서 Q-Q plot을 그려본 결과 정규성을 만족하지 않는 것으로 나타나 0값을 가지고 있어 1을 더해주고 로그변환을 실시한 이후 다시 그려본 Q-Q plot에서 정규성을 만족하는 것으로 나타났다. 변환 전후의 Q-Q plot은 <그림 1-17>을 보면 된다.

**로그변환 후**

**로그변환 전**



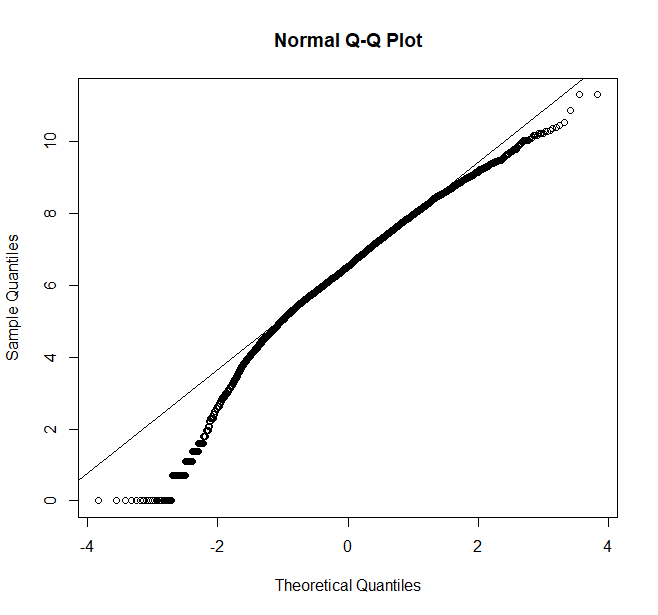
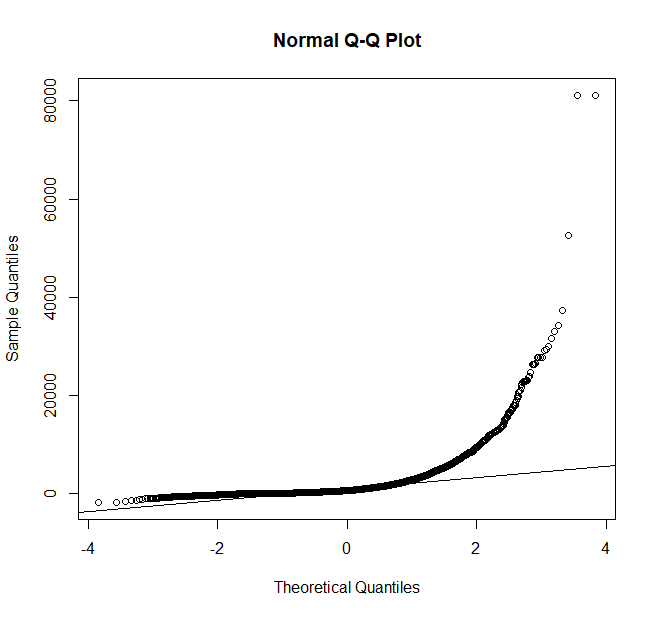
**<그림 1-17> 기참여고객 마지막 통화 시간 변수 변환 전후 Q-Q plot**

다음으로 연간 평균 잔고(balance)에 대해서는 잠재고객과 0의 값을 갖는 고객들은 제거 후 절댓값을 취해준 뒤 로그 변환(balance0)과 제곱근 변환(balance1), 역변환(balance2)을 하여 로지스틱 회귀 모형을 돌려 가장 종속 변수에 영향을 많이 미치는 변수를 최종 변수로 선택한다. 로지스틱 모형을 구축한 결과는 <표 1-16>와 같다.

**<표 1-16> 기참여고객 평균 연간 잔고 변수 변환 로지스틱 모형**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **Z value** | **p-value** |
| **(Intercept)** | -2.707 | 0.3013 | -8.984 | <0.001 |
| **balance** | 0.000048 | 0.00003 | 1.633 | 0.1026 |
| **balance0** | 0.2671 | 0.07026 | 3.802 | <0.001 |
| **balance1** | -0.01029 | -0.006667 | -1.544 | 0.1226 |
| **balance2** | 1.909 | 0.5495 | 3.474 | <0.001 |

로지스틱 모형 구축 결과 변환 변수 중에는 로그 변환 변수(balance0)와 역변환 변수(balance2)의 p-value가 유의수준 0.05보다 작은 것으로 나타났다. 그 중에서도 로그 변환 변수 p-value가 0.000143, 역변환 변수 p-value가 0.000514로 더 작은 로그 변환 변수가 종속 변수에 더 큰 영향을 미칠 것으로 보여 최종 변환 변수로 선택했다. 선택된 로그 변환 변수를 가지고 변환 전과 후의 Q-Q plot을 그려 정규성도 검정했다. 변환 후의 Q-Q plot에서 정규성을 만족하는 것 같았다. 변환 전후의 Q-Q plot은 <그림 1-18>에서 확인하면 된다.



**<그림 1-18> 기참여고객 평균 연간 잔고 변수 변환 전후 Q-Q plot**

**로그변환 전**

**로그변환 후**

평균 연간 잔고 변수를 끝으로 기참여고객 데이터의 독립 변수들의 변환이 끝이 났다. 마지막으로 종속 변수의 불균형을 처리하기 위해 데이터 제공자의 요구대로 빈도수가 큰 범주에서 샘플링을 하여 작은 범주 빈도수에 맞출 것이다. 현재 종속 변수의 no가 5,965명, yes가 1,787명으로 큰 범주인 no에서 30%비율로 샘플링을 하였다. 또한 샘플링에 따르는 가중치 변수를 새로 만들어 no범주에는 3.33만큼의 가중치를 yes는 1로 두었다. 그 결과 기참여고객 데이터는 종속 변수의 비율이 5:5로 맞춰지고 총 관측치는 3,574개로 되었다.

기참여고객에서의 선택된 변수와 각 변환된 변수에 관련된 정보는 아래 <표 1-17>에서 확인하면 된다.

**<표 1-17> 기참여고객 최종 선택된 변수**

|  |  |
| --- | --- |
| **변수명** | **변수 변환 설명** |
| age | 나이대로 범주화 |
| job | 표준직업분류 참고해서 재범주화 |
| marry(marital | marital 재범주화 후 사용 |
| university(education) | education 재범주화 후 사용 |
| default | 그대로 사용 |
| balance | 절댓값 이후 로그 변환 변수 사용 |
| housing | 그대로 사용 |
| loan | 그대로 사용 |
| month | 그대로 사용 |
| duration | 로그변환 후 사용 |
| campaign | 그대로 사용 |
| pdays | 그대로 사용 |
| previous | 그대로 사용 |
| poutcome | 그대로 사용 |

앞서 잠재고객과 기참여고객 각각에서 최종적으로 선택되어진 변수들을 가지고 예금 구매 고객을 예측하는 로지스틱 회귀(Logistic Regression)를 구축하여 예금 구매에 미치는 주요한 변수가 무엇인지 파악해 보려고 한다.

**제 3 절 모형 구축 및 평가**

1. **로지스틱 회귀 모형 구축**

잠재고객 데이터와 기참여고객 데이터에 각각 로지스틱 회귀 모형을 구축하려고 한다. 모형 구축 이후에는 사후평가를 위해 전체 데이터셋을 Training Data Set과 Test Data Set으로 나누려고 한다. 비율은 70:30으로 나누어 Training Data Set을 가지고 모형을 구축하고 Test Data Set으로 예측하여 모형을 평가하려고 한다. 모형 구축 시에는 종속 변수(y)를 5:5로 맞추기 위해 샘플링했던 샘플 가중치를 반영하여 구축하고 Type3 분석을 통해 변수를 한 번 검정하고 그 중 검정결과 선택되어진 변수들을 최종적으로 로지스틱 모형에서 단계적 변수선택법을 사용하여 최종 모형을 도출할 것이다.

먼저 잠재고객에서 최종적으로 선택된 변수들을 모두 넣고 샘플 가중치를 넣어 만든 로지스틱 모형에서 먼저 Type3 분석을 통해 왈드(Wald)검정을 한 결과 선택된 변수 결과는 <표 1-18>와 같다.

**<표 1-18> 잠재고객 Type3 분석 결과 선택된 변수**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Df** | **Chisq** | **p-value** |
| **(Intercept)** | 1 | 1736.4185 | <0.001 |
| **age** | 5 | 73.5681 | <0.001 |
| **job** | 6 | 27.0275 | <0.001 |
| **marital** | 2 | 12.6239 | 0.0018 |
| **default** | 1 | 9.8384 | 0.0017 |
| **balance** | 1 | 24.8356 | <0.001 |
| **housing** | 1 | 116.9575 | <0.001 |
| **loan** | 1 | 39.9047 | <0.001 |
| **month** | 11 | 769.1324 | <0.001 |
| **duration** | 1 | 2312.3831 | <0.001 |
| **campaign** | 1 | 44.9207 | <0.001 |

Type3분석을 한 결과 학력(education)변수가 p-value가 0.0648로 유의수준 0.05에서 귀무가설을 채택하여 종속 변수에 미치는 영향이 거의 없는 것으로 나타나 제외하고 단계적 변수선택법을 통한 최종 모형은 <표 1-19>와 같다.

**<표 1-19> 잠재고객 단계적 변수선택법 최종선택 변수**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | | **Estimate** | **Std. Error** | **Z value** | **p-value** |
| **(Intercept)** |  | -14.2497 | 0.341962 | -41.6704 | <0.001 |
| **age** | **30대** | -0.40586 | 0.083018 | -4.88889 | <0.001 |
| **40대** | -0.56835 | 0.097832 | -5.80948 | <0.001 |
| **50대** | -0.48219 | 0.108084 | -4.46128 | <0.001 |
| **60대** | 0.410406 | 0.162079 | 2.532139 | 0.0113 |
| **70대이상** | 0.197076 | 0.193384 | 1.019095 | 0.3082 |
| **job** | **admin** | 0.178692 | 0.118737 | 1.504941 | 0.1323 |
| **blue-collar** | -0.26632 | 0.111595 | -2.38647 | 0.0170 |
| **entrepreneur** | 0.176547 | 0.136333 | 1.294967 | 0.1953 |
| **management** | 0.070546 | 0.105451 | 0.668997 | 0.5035 |
| **pink-collar** | -0.08469 | 0.120078 | -0.7053 | 0.4806 |
| **technician** | -0.08316 | 0.108846 | -0.76401 | 0.4449 |
| **marital** | **married** | -0.31132 | 0.08762 | -3.553 | <0.001 |
| **single** | -0.26898 | 0.100391 | -2.67928 | 0.0074 |
| **default** | **yes** | 0.692452 | 0.220764 | 3.136623 | 0.0017 |
| **balance** |  | 0.08295 | 0.016645 | 4.983533 | <0.001 |
| **housing** | **yes** | -0.70353 | 0.065053 | -10.8147 | <0.001 |
| **loan** | **yes** | -0.53964 | 0.085427 | -6.31702 | <0.001 |
| **month** | **feb** | 1.050377 | 0.184964 | 5.678813 | <0.001 |
| **mar** | 3.787963 | 0.234461 | 16.15607 | <0.001 |
| **apr** | 1.567805 | 0.180607 | 8.680749 | <0.001 |
| **may** | 0.194754 | 0.171919 | 1.132826 | 0.2573 |
| **jun** | 0.522617 | 0.175929 | 2.970621 | 0.0029 |
| **jul** | 0.400588 | 0.173442 | 2.309638 | 0.0209 |
| **aug** | 0.48247 | 0.174405 | 2.766375 | 0.0057 |
| **sep** | 1.872334 | 0.267199 | 7.007268 | <0.001 |
| **oct** | 3.165882 | 0.230386 | 13.74163 | <0.001 |
| **nov** | 0.256517 | 0.189817 | 1.351396 | 0.1766 |
| **dec** | 3.151451 | 0.348687 | 9.038051 | <0.001 |
| **duration** |  | 2.135194 | 0.044403 | 48.08724 | <0.001 |
| **campaign** |  | -0.10868 | 0.016216 | -6.70229 | <0.001 |

기참여고객에서 최종적으로 선택된 변수들을 모두 넣고 샘플 가중치를 넣어 만든 로지스틱 모형에서 먼저 Type3 분석을 통해 왈드(Wald)검정을 한 결과 선택된 변수 결과는 <표 1-20>와 같다.

**<표 1-20> 기참여고객 Type3 분석 결과 선택된 변수**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Df** | **Chisq** | **p-value** |
| **(Intercept)** | 1 | 451.3306 | <0.001 |
| **age** | 5 | 22.1191 | <0.001 |
| **job** | 6 | 29.3775 | <0.001 |
| **university** | 1 | 24.7646 | <0.001 |
| **balance** | 1 | 4.3089 | 0.0379 |
| **housing** | 1 | 44.5761 | <0.001 |
| **month** | 11 | 182.7999 | <0.001 |
| **duration** | 1 | 500.3138 | <0.001 |
| **pdays** | 1 | 5.2730 | 0.0217 |
| **poutcome** | 2 | 495.4838 | <0.001 |

Type3분석을 한 결과 결혼 경험 유무(marry)변수가 p-value가 0.5614로 유의수준 0.05에서 귀무가설을 채택하여 종속 변수에 미치는 영향이 거의 없는 것으로 나타났고 마찬가지로 채무 불이행 여부(default)의 p-value가 0.3573, 개인대출 여부(loan)의 p-value가 0.8069, 현재 캠페인 진행 중 연락수(campaign)의 p-value가 0.0628, 이전 캠페인 진행 중 연락수(previous)의 p-value가 0.4540로 이 모든 변수들을 제외하고 단계적 변수선택법을 통한 최종 모형은 <표 1-21>와 같다.

**<표 1-21> 기참여고객 단계적 변수선택법 최종선택 변수**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | | **Estimate** | **Std. Error** | **Z value** | **p-value** |
| **(Intercept)** |  | -11.4528 | 0.516969 | -22.1537 | <0.001 |
| **age** | **30대** | -0.22189 | 0.136847 | -1.62147 | 0.1049 |
| **40대** | 0.070843 | 0.153977 | 0.460087 | 0.6454 |
| **50대** | -0.35325 | 0.163543 | -2.15999 | 0.0308 |
| **60대** | 0.428503 | 0.212145 | 2.019858 | 0.0434 |
| **70대이상** | -0.3683 | 0.269003 | -1.36914 | 0.1709 |
| **job** | **admin** | 0.15669 | 0.175939 | 0.890593 | 0.3731 |
| **blue-collar** | -0.38638 | 0.181088 | -2.13364 | 0.0329 |
| **entrepreneur** | -0.76935 | 0.221681 | -3.47051 | <0.001 |
| **management** | -0.2048 | 0.176489 | -1.16042 | 0.2459 |
| **pink-collar** | -0.28525 | 0.188827 | -1.51065 | 0.1309 |
| **technician** | 0.086592 | 0.17224 | 0.502738 | 0.6151 |
| **university** | **univ** | 0.58204 | 0.119271 | 4.879987 | <0.001 |
| **balance** |  | 0.060996 | 0.028605 | 2.132317 | 0.0329 |
| **housing** | **yes** | -0.69215 | 0.103873 | -6.66344 | <0.001 |
| **month** | **feb** | 0.940003 | 0.248817 | 3.777888 | <0.001 |
| **mar** | 1.961955 | 0.318545 | 6.159109 | <0.001 |
| **apr** | 0.725559 | 0.243313 | 2.982001 | 0.0029 |
| **may** | 0.49014 | 0.23475 | 2.087926 | 0.0368 |
| **jun** | 1.820683 | 0.272253 | 6.687458 | <0.001 |
| **jul** | 1.517217 | 0.285533 | 5.313631 | <0.001 |
| **aug** | 1.905009 | 0.254978 | 7.47126 | <0.001 |
| **sep** | 2.261789 | 0.296499 | 7.628321 | <0.001 |
| **oct** | 1.778173 | 0.26846 | 6.623612 | <0.001 |
| **nov** | 0.671905 | 0.246681 | 2.723777 | 0.0065 |
| **dec** | 0.995401 | 0.333329 | 2.98624 | 0.0028 |
| **duration** |  | 1.53239 | 0.068255 | 22.45099 | <0.001 |
| **pdays** |  | 0.000944 | 0.000417 | 2.261258 | 0.0237 |
| **poutcome** | **other** | 0.037096 | 0.109924 | 0.33747 | 0.7358 |
| **success** | 2.27482 | 0.107382 | 21.18441 | <0.001 |

1. **로지스틱 회귀 모형 평가**

앞서 구축한 두 개의 로지스틱 모형을 평가하는 방법으로는 여러 방법이 있는데 그 중 저자는 사전평가 방법으로 십분위 분석(Decile Analysis)을 사용하고, 사후평가로는 정확도(Accuracy)와 이익도표(Gains Tables and Charts)를 이용해서 모델을 평가하려고 한다.

모형구축에서 최종적으로 선택된 변수로 구축한 모형을 평가하기 위해 십분위 분석법을 실시하기로 하였다. 구축된 모형이 실제 캠페인 성공 여부를 잘 분류하는가를 평가하기 위함이었다. 먼저 잠재고객에 대한 Training Data Set의 십분위 분석표는 <표 1-22>와 같다.

**<표 1-22> 잠재고객 Training Data Set 십분위 분석표**

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** | **prospects** | **Predicted Probability** | **Percent y** |
| **0** | 441 | 0.76697 | 0.96145 |
| **1** | 441 | 0.51532 | 0.90249 |
| **2** | 441 | 0.34738 | 0.83447 |
| **3** | 441 | 0.23141 | 0.77324 |
| **4** | 442 | 0.14512 | 0.63719 |
| **5** | 441 | 0.08321 | 0.45475 |
| **6** | 441 | 0.04058 | 0.27211 |
| **7** | 441 | 0.01924 | 0.11338 |
| **8** | 441 | 0.00777 | 0.05896 |
| **9** | 441 | 0.0017 | 0.00907 |
| **Total** | 4,411 | 0.21587 | 0.5017 |

0분위수 그룹과 9분위수 그룹의 캠페인 성공 여부가 1일 비율인 Percent y를 비교한 결과 0분위수 그룹은 캠페인 성공 여부가 1일 확률이 0.96145이다. 반면에 9분위수 그룹은 캠페인 성공 여부가 1일 확률이 0.00907로 0분위수 그룹이 9분위수 그룹보다 106배 더 높다고 이야기를 할 수 있다. 이렇게 0분위수 그룹과 9분위수 그룹의 확률 차이가 106배가 나온다는 것은 Training Data Set에서는 모형의 적합도가 꽤 높다고 이야기를 할 수 있어 보인다.

그 다음으로는 잠재고객에 대한 Training Data Set의 십분위 분석표는 <표 1-23>와 같다.

**<표 1-23> 잠재고객 Test Data Set 십분위 분석표**

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** | **Prospects** | **Predicted Probability** | **Percent y** |
| **0** | 189 | 0.77319 | 0.97354 |
| **1** | 189 | 0.51821 | 0.89947 |
| **2** | 189 | 0.35121 | 0.88889 |
| **3** | 189 | 0.23279 | 0.78836 |
| **4** | 190 | 0.14478 | 0.58947 |
| **5** | 189 | 0.08266 | 0.44974 |
| **6** | 189 | 0.04175 | 0.21164 |
| **7** | 189 | 0.02008 | 0.10582 |
| **8** | 189 | 0.00819 | 0.03704 |
| **9** | 189 | 0.00152 | 0.01587 |
| **Total** | 1,891 | 0.2174 | 0.49603 |

Training Data Set처럼 0분위수 그룹과 9분위수 그룹의 캠페인 성공 여부가 1일 비율인 Percent y를 비교한 결과, 0분위수 그룹의 경우 0.97354가 나왔으며, 9분위수 그룹의 경우는 0.01587가 나왔다. 0분위수 그룹이 9분위수 그룹보다 61.34배 더 높다고 이야기를 할 수 있다.

이후 모형이 잘 구축이 되어있는지 확인하기 위해서<표 1-22>와 <표1-23>의 결과를 통해 Training Data Set과 Test Data Set의 적합도를 비교한 결과는 <표 1-24>과 같다.

**<표 1-24> 잠재고객 Training, Test Set별 0분위수 그룹과 9분위수 그룹의 비율**

|  |  |
| --- | --- |
| **Training Data Set** | **Test Data Set** |
| 106 | 61.34 |

Training Data Set의 적합도가 Test Data Set의 적합도에 비해 높다는 것을 확인 할 수 있었다. 따라서 잠재고객에 대한 모형이 잘 구축 되었다고 할 수 있다.

기참여고객에 대한 Training Data Set의 십분위 분석표는 <표 1-25>와 같다.

**<표 1-25> 기참여고객 Training Data Set 십분위 분석표**

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** | **Prospects** | **Predicted Probability** | **Percent y** |
| **0** | 250 | 0.91068 | 0.952 |
| **1** | 250 | 0.76845 | 0.904 |
| **2** | 250 | 0.60316 | 0.836 |
| **3** | 250 | 0.41392 | 0.748 |
| **4** | 251 | 0.27754 | 0.59761 |
| **5** | 250 | 0.16742 | 0.456 |
| **6** | 250 | 0.09429 | 0.208 |
| **7** | 250 | 0.0538 | 0.132 |
| **8** | 250 | 0.02719 | 0.044 |
| **9** | 250 | 0.00826 | 0.012 |
| **Total** | 2,501 | 0.33245 | 0.489 |

0분위수 그룹과 9분위수 그룹의 캠페인 성공 여부가 1일 비율인 Percent y를 비교한 결과 0분위수 그룹은 캠페인 성공 여부가 1일 확률이 0.952이다. 반면에 9분위수 그룹은 캠페인 성공 여부가 1일 확률이 0.012로 0분위수 그룹이 9분위수 그룹보다 793.33배 더 높다고 이야기를 할 수 있다. 이렇게 0분위수 그룹과 9분위수 그룹의 확률 차이가 793배가 나온다는 것은 Training Data Set에서는 모형의 적합도가 꽤 높다고 이야기를 할 수 있어 보인다.

그 다음으로는 기참여고객에 대한 Test Data Set의 십분위 분석표는 <표 1-26>와 같다.

**<표 1-26> 기참여고객 Test Data Set 십분위 분석표**

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** |  | **Predicted Probability** | **Percent y** |
| **0** | 107 | 0.92721 | 0.99065 |
| **1** | 108 | 0.81185 | 0.87037 |
| **2** | 107 | 0.65128 | 0.79439 |
| **3** | 107 | 0.47976 | 0.75701 |
| **4** | 108 | 0.31795 | 0.76852 |
| **5** | 107 | 0.20542 | 0.53271 |
| **6** | 107 | 0.1214 | 0.28037 |
| **7** | 107 | 0.06244 | 0.18692 |
| **8** | 108 | 0.02834 | 0.06481 |
| **9** | 107 | 0.00885 | 0.00935 |
| **Total** | 1,073 | 0.36152 | 0.52563 |

Training Data Set처럼 0분위수 그룹과 9분위수 그룹의 캠페인 성공 여부가 1일 비율인 Percent y를 비교한 결과, 0분위수 그룹의 경우 0.99065가 나왔으며, 9분위수 그룹의 경우는 0.00935가 나왔다. 0분위수 그룹이 9분위수 그룹보다 105.95배 더 높다고 이야기를 할 수 있다.

이후 모형이 잘 구축이 되어있는지 확인하기 위해서<표 1-25>와 <표1-26>의 결과를 통해 Training Data Set과 Test Data Set의 적합도를 비교한 결과는 <표 1-27>과 같다.

**<표 1-27> 기참여고객 Training, Test Set별 0분위수 그룹과 9분위수 그룹의 비율**

|  |  |
| --- | --- |
| **Training Data Set** | **Test Data Set** |
| 793.33 | 105.95 |

Training Data Set의 적합도가 Test Data Set의 적합도에 비해 높다는 것을 확인 할 수 있었다. 따라서 기참여고객에 대한 모형도 잘 구축 되었다고 할 수 있다.

다음으로 모델들의 정확도를 보려고 한다. 앞서 잠재고객과 기참여고객에서 각각 Training Data Set으로 구축한 로지스틱 모형을 Test Data Set으로 예측하여 실제값과 비교하는 분류 결과표를 만들어 정확도를 계산한다. 먼저 잠재고객 로지스틱 모형의 분류 결과표는 <표 1-28>를 보면 된다.

**<표 1-28> 잠재고객 로지스틱 회귀 모형 분류 결과표**

|  |  |  |  |
| --- | --- | --- | --- |
| **64.41%** | **실제값** | | |
| **예측값** |  | **no** | **yes** |
| **no** | 939 | 659 |
| **yes** | 14 | 279 |

분류 결과표를 확인한 결과 정확도는 64.41%로 나타났다. 원래 데이터셋에서도 예금 구매하지 않는 고객(no)의 비율이 압도적으로 높아 모델에서도 샘플 가중치를 반영한 결과 실제값이 예금을 구매했음(yes)에도 불구하고 하지 않을 것으로 예측한 비율이 생각보다 높았다. 이 부분을 제외하고는 대체적으로 예측을 잘 하는 것으로 나타난다.

마찬가지로 기참여고객에 대해서 정확도를 계산하려고 한다. 기참여고객 로지스틱 모형의 분류결과표는 <표 1-29>를 보면 된다.

**<표 1-29> 기참여고객 로지스틱 회귀 모형 분류 결과표**

|  |  |  |  |
| --- | --- | --- | --- |
| **73.07%** | **실제값** | | |
| **예측값** |  | **no** | **yes** |
| **no** | 465 | 245 |
| **yes** | 44 | 319 |

분류 결과표를 확인한 결과 정확도는 73.07%로 나타났다. 기참여고객도 잠재고객과 똑같이 원래 데이터셋에서 예금 구매하지 않는 고객(no)의 비율이 압도적으로 높아 실제값이 예금을 구매했음(yes)에도 불구하고 하지 않을 것으로 예측한 비율이 높았다. 이 부분을 제외하고는 대체적으로 예측을 잘 하는 것으로 나타난다.

현재 잠재고객에 대한 모델과 기참여고객에 대한 모델로 총 2개의 모델이 구축 된 가운데 이 모델들을 마케팅 담장자나 경영자들의 이해를 돕기 위한 목적으로 이익도표 생성 및 모델 평가를 실시하였다. 잠재고객에 대한 이익도표는 <표 1-30>와 같다.

**<표 1-30> 잠재고객 이익도표**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Decile | count | cum % of count | Predicted Probabilty percent | Percent y | Cumulative percent y | count of y | percent of Total y | Cum count of y | Cum count of y percent | Lift % | Cum lift % |
| **0** | 441 | 10% | 76.697% | 96.145% | 96.145% | 423 | 19.16% | 423 | 19.16% | 490 | 564 |
| **1** | 441 | 20% | 51.532% | 90.249% | 93.197% | 397 | 17.98% | 820 | 37.14% | 217 | 353 |
| **2** | 441 | 30% | 34.738% | 83.447% | 89.947% | 368 | 16.67% | 1188 | 53.80% | 106 | 271 |
| **3** | 441 | 40% | 23.141% | 77.324% | 86.791% | 340 | 15.40% | 1528 | 69.20% | 67 | 220 |
| **4** | 442 | 50% | 14.512% | 63.719% | 82.177% | 281 | 12.73% | 1809 | 81.93% | 38 | 183 |
| **5** | 441 | 60% | 8.321% | 45.475% | 76.060% | 200 | 9.06% | 2009 | 90.99% | 36 | 159 |
| **6** | 441 | 70% | 4.058% | 27.211% | 69.081% | 120 | 5.43% | 2129 | 96.42% | 21 | 139 |
| **7** | 441 | 80% | 1.924% | 11.338% | 61.864% | 50 | 2.26% | 2179 | 98.69% | 7 | 122 |
| **8** | 441 | 90% | 0.777% | 5.896% | 55.645% | 26 | 1.18% | 2205 | 99.86% | 4 | 109 |
| **9** | 441 | 100% | 0.170% | 0.907% | 50.171% | 3 | 0.14% | 2208 | 100.00% | 4 | 100 |
| **Total** | 4,411 |  | 21.587% | 50.170% |  | 2208 | 100.00% |  |  |  |  |

주어진 잠재고객 이익도표에서 비전문가들의 이해를 돕기 위해 설명을 할 부분은 Cum Count of y percent와 Lift%, Cum lift %이다. 먼저 Cum Count of y percent는 일반 누적 확률 대비 누적 활성화 확률을 비교하는 방법으로 현재 50%의 누적인원으로 90.99%의 활성화 관찰치를 찾을 수 있기 때문에 모형을 구축을 하지 않을 때보다 40%의 인원을 더 찾을 수 있음을 의미 한다. 이를 시각화로 표현할 경우 <그림 1-19>와 같다.

**<그림 1-19> 잠재고객 모델 이익도표 그래프**

그 다음 Lift%는 현재 구둑된 모형이 얼마나 잘 적용이 되는지 각 십분위에서 알아보는 것이다. 현재 0분위수에서는 490%의 값을 가지고 있는데 이는 평균보다 4.9배 더 포함을 하고 있다는 것을 의미한다. 그리고 현재 2분위수 까지는 106%로 100%보다 높기 때문에 2분위수 까지는 구축된 모형이 평균 모형보다 좋음을 알 수 있다.

마지막으로 Cum lift %의 경우 모형을 구축 하지 않았을 때 보다 구축된 모형이 얼마나 잘 수행이 되는지를 누적으로 보는 것으로 현재 4분위수까지의 Cum lift%는 183%로 모형을 구축 하지 않았을 때보다 83%더 적합하는 것을 의미한다.

현재 이익도표를 통해서 살펴본 잠재고객 모델의 경우 모델을 구축하지 않을 때보다 Cum of y percent, Lift%, Cum lift%가 우수하기 때문에 정기구독자를 예측하기에 적절한 모델로 보인다. 그 다음 기참여고객에 대한 이익도표는 <표 1-31>과 같다.

**<표 1-31> 기참여고객 이익도표**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Decile | count | cum % of count | Predicted Probabilty percent | Percent y | Cumulative percent y | count of y | percent of Total y | Cum count of y | Cum count of y percent | Lift % | Cum lift % |
| **0** | 189 | 10% | 77.319% | 97.354% | 97.354% | 183 | 19.64% | 183 | 19.64% | 353 | 372 |
| **1** | 189 | 20% | 51.821% | 89.947% | 93.651% | 169 | 18.13% | 352 | 37.77% | 254 | 303 |
| **2** | 189 | 30% | 35.121% | 88.889% | 92.063% | 168 | 18.03% | 520 | 55.79% | 158 | 255 |
| **3** | 189 | 40% | 23.279% | 78.836% | 88.757% | 149 | 15.99% | 669 | 71.78% | 95 | 215 |
| **4** | 190 | 50% | 14.478% | 58.947% | 82.795% | 111 | 11.91% | 780 | 83.69% | 57 | 183 |
| **5** | 189 | 60% | 8.266% | 44.974% | 76.491% | 85 | 9.12% | 865 | 92.81% | 35 | 158 |
| **6** | 189 | 70% | 4.175% | 21.164% | 68.587% | 39 | 4.18% | 904 | 97.00% | 22 | 139 |
| **7** | 189 | 80% | 2.008% | 10.582% | 61.337% | 19 | 2.04% | 923 | 99.03% | 14 | 123 |
| **8** | 189 | 90% | 0.819% | 3.704% | 54.933% | 7 | 0.75% | 930 | 99.79% | 7 | 110 |
| **9** | 189 | 100% | 0.152% | 1.587% | 49.598% | 2 | 0.21% | 932 | 100.00% | 2 | 100 |
| **Total** | 1,891 |  | 21.740% | 49.603% |  | 932 | 100.00% |  |  |  |  |

먼저 Cum Count of y percent의 경우 현재 50%의 누적인원으로 92.81%의 활성화 관찰치를 찾을 수 있기 때문에 잠재고객 모델이 90.99%의 활성화 관측치를 찾는 것보다는 더 나은 것으로 보인다. 그리고 이를 시각화로 표현할 경우 <그림 1-20>와 같다.

**<그림 1-20> 기참여고객 모델 이익도표 그래프**

그 다음 Lift%는 현재 0분위수에서는 353%의 값을 가지고 있어 잠재고객 모델보다는 다소 떨어질 수 있지만, 현재 2분위수 까지는 158%로 100%보다 높은데다가 잠재고객의 경우는 2분위수까지의 값이 106%이기 때문에 기참여모델의 경우 2분위수까지는 평균 모형보다 좋으며 잠잠재고객모델 보다는 안정성이 더 있을 것으로 보인다.

마지막으로 Cum lift %의 경우 예시로 든 현재 4분위수까지의 Cum lift%가 183%로 모형을 구축 하지 않았을 때보다 83%더 적합하며 이 부분에서는 잠재고객 과 기참여고객 모델 모두 비슷한 것으로 보인다.

현재 이익도표를 통해서 살펴본 기참여고객 모델의 경우도 모델을 구축하지 않을 때보다 Cum of y percent, Lift%, Cum lift%가 우수하기 때문에 정기구독자를 예측하기에 적절한 모델로 보이며 잠재고객 모델과는 비교했을 때, 0분위수에서의 Lift% 값은 잠재고객보다 부족할 수 있지만 다른 분위수들을 합쳐서 누적으로 판단한다면 기참여고객 모델이 더 이익 측면에서는 더 나은 것으로 생각할 수 있어 보인다.

1. **최종 모형 선택**

잠재고객과 기참여고객에서 각각 모델을 구축하였다. 이 두 모델을 최종 모델로 확정하기 전에 변수들 간에 다중공선성을 확인하기 위해 독립변수들의 분산팽창지수(Variance Infalation Factor)를 확인하려고 한다. 일반적으로 분산팽창지수가 5 또는 10을 넘으면 변수사이에 다중공선성이 있다고 본다. 저자는 분산팽창지수가 5를 넘으면 다중공선성이 있다고 판단하겠다. 먼저 잠재고객 로지스틱 모형의 분산팽창지수표가 <표 1-32>와 같다.

**<표 1-32> 잠재고객 로지스틱 모형 분산팽창지수**

|  |  |
| --- | --- |
| **독립변수** | **VIF값** |
| age | 2.322950 |
| job | 1.778488 |
| marital | 1.486697 |
| default | 1.045199 |
| balance | 1.100934 |
| housing | 1.451143 |
| loan | 1.092661 |
| month | 2.392130 |
| duration | 1.255050 |
| campaign | 1.096350 |

분산팽창지수를 확인한 결과 모든 변수에서 5를 넘지 않아 독립변수들 사이의 다중공선성은 없는 것으로 나타났다. 잠재고객의 최종 모형은 다음과 같다.

다음은 기참여고객 로지스틱 모형의 독립변수들 분산팽창지수를 확인해 본다. 분산팽창지수는 <표 1-33>와 같다.

**<표 1-33> 기참여고객 로지스틱 모형 분산팽창지수**

|  |  |
| --- | --- |
| **독립변수** | **VIF값** |
| age | 1.682572 |
| job | 2.967001 |
| university | 1.867179 |
| balance | 1.079601 |
| housing | 1.489759 |
| month | 1.633015 |
| duration | 1.162394 |
| pdays | 1.337298 |
| poutcome | 1.192802 |

분산팽창지수를 확인한 결과 모든 변수에서 5를 넘지 않아 독립변수들 사이의 다중공선성은 없는 것으로 나타났다. 기참여고객의 최종 모형은 다음과 같다.

**Ⅲ. 결 론**

**제 1 절 최종 모형 의미 해석**

먼저 잠재고객에 대한 로지스틱 회귀분석에서의 계수들을 살펴보면 가장 눈에 띄는 점은 로그변환을 한 duration으로 로그변환한 duration의 값이 1이 증가한다면고객이 정기예금에 가입할 가능성이 기존보다 8.46배 증가한다는 것은 꽤나 주목할 만한 점이다. 그렇기 때문에 전화를 받은 고객이 아예 관심이 없어 빨리 끊는 경우가 아니라면 통화시간을 오래 유지해서 고객이 정기예금 가입에 관심을 가지도록 하는 것이 좋아보인다.

그 다음으로는 재범주화를 진행한 age 변수로 기준이 20대 이하인 상황에서 예금을 이미 어느 정도 해놨을 가능성이 있는 30,40,50대의 경우에는 20대보다 정기예금에 가입할 가능성이 거의 0.66,0.56,0.61배 수준으로 낮아지기 때문에 이 나이대의 사람들 보다는 사회초년생이 많은 20대에 혹은 정기예금에 가입할 확률이 1.5배,1.21배 높은 60대 이상 집단에 집중을 하는 것이 좋아 보인다.

직업 변수의 경우는 일을 하지 않는 사람들이 기준이기 때문에 이를 기준으로 비교해보자면 admin직업의 경우 1.18배, blue-collar의 경우 0.76배, entrepreneur의 경우 1.19배, management의 경우 1.07배 pink-collar, technician의 경우 0.92배의 정기예금 가능성을 가지고 있다. blue-collar를 제외하고는 이 가능성의 차이가 없기 때문에 전화를 받은 고객이 blue-collar인 경우에는 기존 직업들 보다는 약 0.76배 정도로 낮게 가입할 것으로 생각해야 될 것이다.

결혼 여부에 대한 변수의 경우 이혼,사별이 기준인 가운데 결혼한 사람과 싱글인 사람은 이보다 낮은 70%에서 76%정도의 정기예금 가능성을 가지고 있다. 따라서 결혼을 경험하고 현재 혼자인 대상들이 정기예금을 신청할 가능성이 높을 것으로 보인다.

채무 불이행 여부의 경우 채무 불이행을 한 사람의 경우가 약 2배가량 높은 정기예금 가능성을 가지고 있는 것을 볼 수 있는데 이는 채무 불이행을 한 사람의 경우 신뢰 등급이 낮아 가입이 쉽지 않기 때문에 만약에 마케팅 전화가 와서 예금 가입이 가능하다고 하면 이를 통해서 가입하려고 하는 것으로 생각된다. 하지만 이를 기준으로 채무불이행 대상들을 일부러 찾아서 하는 가입을 권유하기에는 리스크가 크기 때문에 주의가 필요하여 보인다.

그리고 로그변환을 한 balance변수의 경우 단위가 1이 증가한다면 정기예금에 가입할 가능성이 기존보다 1.1배정도 증가하기 때문에 balance가 낮은 값들 보다는 balance의 마이너스가 큰 고객이나 혹은 balance 자체가 고객들을 대상으로 적극적으로 가입권유를 하는 것이 좋아보인다.

집 담보 여부와 개인 담보 여부의 경우, 담보를 가지고 있는 고객의 경우 정기예금에 가입할 가능성이 각각 0.5배, 0.58배 정도로 떨어진다. 이는 담보 대출때문에 예금을 통해서 저축을 할 가능성이 별로 없기 때문이라고 생각이 되어 만약에 담보 대출을 가지고 있지 않은 고객에게 집중을 하는 것이 좋은 결과를 가지고 올 것이라고 본다.

월 변수의 경우 1월을 기준으로 2월은 2.85배, 3월은 43.81배, 4월은 4.76배, 5월은 1.2배, 6월은 1.68배 7월은 1.49배, 8월은 1.6배, 9월은 6.49배, 10월은 23.57배, 11월은 1.28배, 12월은 23.33배 더 높은 것이 확인이 되었다. 따라서 만약에 예금 가입에 집중할 달을 고른다면 3월, 10월, 12월, 9월, 4월 순으로 집중적으로 하는 것이 좋아 보인다.

마지막으로 campaign 변수의 경우 현재 캠페인을 진행하면서 횟수가 1번 증가하면 정기예금에 가입할 가능성이 0.9배로 낮아지는 것을 확인 할 수 있다. 따라서 전화연결을 무작정 많이 하는 것은 좋은 것이 아니어 보이며 예를 들어 현재 최소 기준인 1번보다 5번 더 많이 해서 6번의 전화연결을 한다면 그 고객이 정기예금에 가입할 가능성이 0.6배로 낮아지기 때문에 최소한의 연결로 많은 잠재고객들과 접촉하여 정기예금에 가입하도록 하는 것이 좋아 보인다.

잠재고객과 다르게 기참여고객은 이전에 참여한 기록인 pdays, previous, poutcome이 추가로 존재하였기 때문에 이 부분에 대해서 살펴보는게 좋아 보인다.

먼저 pdays의 경우 일자로 되어있어서 1일을 기준으로 하면 사실상 동일한 정기예금 가능성을 보이는 것처럼 보이지만 pdays가 다른 사람들보다 100일이나 더 지난 고객의 경우 정기예금에 가입할 가능성이 1.1배로 높아지는 것을 확인이 가능하다. 이전 캠페인에 참여한 것이 최근인 경우는 pdays의 값이 낮기 때문에 이러한 결과를 나온 것으로 파악이 가능해 보인다.

다음 poutcome의 경우는 이전 캠페인에서의 결과이기 때문에 실패를 기준으로 성공이 아닌 other의 경우에는 1.03배가 증가하는 것으로 보여 별 다른 차이가 없는 것으로 보인다. 하지만 이전 캠페인의 결과가 성공인 고객의 경우에는 실패한 고객보다 정기예금에 가입할 가능성이 무려 9.72배가 증가하는 결과가 나타난다. 이는 이전의 캠페인에 가입을 할 만큼 해당 은행에 신뢰성을 가지고 있기 때문에 이 신뢰성을 기반으로 현재 캠페인의 정기예금에도 가입하는 것으로 해석이 가능하여 보인다.

재범주화를 진행한 age 변수의 경우 20대 이하를 기준으로 30대는 0.8배, 40대는 1.07배, 50대의 경우 0.7배, 60대의 경우 1.53배, 70대 이상의 경우 0.7배의 정기예금 가입 가능성을 보이는데, 현재 30대, 40대, 70대 이상의 경우 20대 이하와는 차이가 없다고 z검정 결과가 나왔기 때문에 기참여고객의 경우에는 50대에게서는 기준보다는 낮은 정기예금 가입 가능성을, 60대에게서는 기준보다는 높은 정기예금 가입 가능성을 기대할 수 있어 보인다.

직업 변수의 경우는 일을 하지 않는 사람들이 기준이기 때문에 이를 기준으로 비교해보자면 admin직업의 경우 1.16배, blue-collar의 경우 0.67배, entrepreneur의 경우 0.46배, management의 경우 0.78배, pink-collar의 경우 0.75배, technician의 경우 1.08배의 정기예금 가능성을 가지고 있다. blue-collar, entrepreneur를 제외하고는 이 가능성의 차이가 없기 때문에 전화를 받은 고객이 blue-collar, entrepreneur인 경우에는 기존 직업들 보다는 각각 0.67배, 0.46배 정도로 낮게 가입할 것으로 생각해야 될 것이다.

로그변환을 한 balance변수의 경우 단위가 1이 증가한다면 정기예금에 가입할 가능성이 기존보다 1.06배정도 증가하기 때문에 잠재고객과 마찬가지로 balance가 낮은 값들 보다는 balance의 마이너스가 큰 고객이나 혹은 balance 자체가 고객들을 대상으로 적극적으로 가입 권유를 하는 것이 좋아 보인다.

월 변수의 경우 1월을 기준으로 2월은 2.55배, 3월은 7.11배, 4월은 2.07배, 5월은 1.63배, 6월은 6.17배 7월은 4.55배, 8월은 6.72배, 9월은 9.6배, 10월은 5.92배, 11월은 1.96배, 12월은 2.71배 더 높은 것이 확인이 되었다. 잠재고객보다는 차이가 많이 나지는 않지만, 만약에 예금 가입에 집중할 달을 고른다면 9월, 3월, 6월, 10월, 7월 순으로 집중적으로 하는 것이 좋아 보인다.

집 담보 여부의 경우, 담보를 가지고 있는 고객의 경우 정기예금에 가입할 가능성이 각각 0.5배 정도로 떨어진다. 이는 잠재고객의 경우와 동일하게 담보 대출때문에 예금을 통해서 저축을 할 가능성이 별로 없기 때문이라고 생각이 된다.

마지막으로 university 변수의 경우 대학을 나오지 않은 것을 기준으로 대학을 나온 고객이 정기예금에 가입할 가능성이 1.79배 더 높은 것으로 확인이 되었는데, 이는 학력과 직업의 연관성에서 비롯된 결과로 추측할 수 있어 보인다.

현재 잠재고객 모델과 기참여고객 모델 모두에게 들어 있는 변수는 age(연령대), 직업, 로그변환이 된 balance, housing, month 이다.

연령대의 경우 두 모델 모두 60대의 연령대의 고객들에게서 기준보다 높은 가입 가능성을 예측하였기 때문에 해당 연령대의 고객이 있는 경우에는 최대한 가입에 대한 노력을 시도해 보는 것이 좋아 보인다.

직업의 경우 blue-collar가 공통적으로 기준보다 낮은 가입 가능성을 예측하였기 때문에 blue-collar의 경우에는 다른 직업들 보다는 예금 가입을 위해서 덜 시도 하는 것이 좋아 보인다.

Balance의 경우는 공통적으로 마이너스 금액 혹은 플러스 금액이 높은 고객들에게 가입 시도를 높이는 것이 좋아 보이며, housing의 경우에는 집 담보 대출이 없는 고객들에게 시도를 높이는 것이 좋아 보인다.

월 변수의 경우는 두 모델에 공통적으로 나온 달은 3월, 9월, 10월로 이 기간에는 모든 고객을 대상으로 활발한 정기예금 가입 시도를 하는 것이 좋아 보이며, 잠재고객을 대상으로는 4월, 12월에, 기참여고객을 대상으로는 6월 7월에 추가적인 정기예금 가입 시도를 하는 것이 좋아 보인다.

잠재고객에만 있는 변수의 경우 결혼여부, default, loan, duration, campaign이 있는데 잠재 고객이라는 미지의 고객에게 해당 정기예금 상품을 권유 할 때는 전화연결을 많이 시도하기 보다는 적은 횟수를 많은 고객들에게 시도하여서 효율성을 높이는 것이 좋아 보이며, 1번의 통화에도 통화 지속시간을 길게 하여서 고객이 정기예금을 신청할 가능성을 높이는 것이 적절 하여 보인다.

기참여고객에게만 있는 변수의 경우 pdays, poutcome, university가 있는데 기참여고객에게만 있는 이전 캠페인에 대한 결과들은 가장 마지막으로 연락을 한 날짜가 길수록, 그리고 이전에 참여한 결과가 성공인 경우에는 현재 캠페인에서 정기예금을 신청할 가능성이 높기 때문에 이러한 특성을 이용하여서 고객들에게 정기예금을 권유한다면 더 많은 고객들이 해당 은행의 정기예금에 가입할 것으로 보인다.

**제 2 절 향후 과제**

|  |  |
| --- | --- |
| **모형 이름** | **PORBANK2019** |
| 모형 구축 기간  모형 구축가  전체 목적  구체적인 목표 고객  모형 구축 데이터명(날짜)  중요 변수 | 2019.4 – 2019.5  김경록 & 박권수  전체 예금 고객 증대  예금 가입 가능한 모든 고객  Bank-full(2011.10)  전화 통화 시간, 이전 캠페인에서 성공 |

포르투갈 은행 기관을 통한 텔레마케팅이 예금 가입에 미치는 영향을 로지스틱 회귀 분석을 통해 알아보는 프로젝트가 끝이 났다. 모형을 구축하는 과정보다는 변수를 하나하나 파악하고 전처리하는 과정이 더 많은 시간을 할애했다. 분석 기간이 한 달이 채 안되 더 깊이 있는 분석을 하기 어려웠지만 최대한 변수의 의미를 생각하며 전처리를 하려고 노력했다. 예를 들면 평균 연간 잔고 같은 경우에도 은행 입장에서의 수익구조를 파악하고 단순한 로그 변환이 아닌 분석자의 주관에 맞는 변수 변환을 실시하면서 구축된 모델의 해석이 좀 더 용이하고 현실 가능하도록 하려고 했다. 그럼에도 적당했던 인구통계학적 변수에 비해 캠페인 관련 변수들이 사용하기 어려운 형태로 기록되어 있었다. 캠페인에서의 마지막 연락의 중요성이 얼마나 되는지는 파악하지 못하였지만 마지막 연락에 관련된 변수보다는 캠페인 기간 중 모든 연락에 대한 변수가 있었다면 예금 가입 고객을 조금 더 캠페인에 초점을 맞춘 분석이 가능하게 하여 이 모델을 사용할 은행 기관이 확실한 행동을 취할 수 있게 할 수 있지 않았을까 하는 아쉬움이 있었다. 다음 캠페인에 모형을 구축할 계획이 있다면 좀 더 구체적인 변수 기록이 필요하다고 생각한다.

**부 록 (R code & SAS code)**

**R Part**

**#0. 사용 패키지 정리**

**#install.packages("ggplot2")**

**library(ggplot2)**

**#install.packages("randomForest")**

**library(randomForest)**

**#install.packages("caret")**

**library(caret)**

**#install.packages("e1071")**

**library(e1071)**

**#install.packages("party")**

**library(party)**

**#install.packages("Epi")**

**library(Epi)**

**#install.packages("ROCR")**

**library(ROCR)**

**#install.packages("car")**

**library(car)**

**#1. 데이터 불러오기 및 전체적인 자료 파악**

**bank<read.csv("C:/Users/bluedice/Desktop/BankData/bank-full.csv",sep=";")**

**head(bank)**

**str(bank)**

**summary(bank)**

**#2. 전체적인 시각화 작업을 통한 이상치 파악 제거(기본적인 변수만 파악 작업)**

**bank$tk<-1**

**#ggplot에 갯수 추가하려고 잠깐 넣은 변수**

**#a. 연령**

**table(bank$age)**

**hist(bank$age,breaks=seq(10,100,by=10),axes=F,ann=F,labels=T)**

**axis(1,at=seq(10,100,10))**

**axis(2,ylim=c(0,20100))**

**title(xlab="연령대",cex.lab=2)**

**title(ylab="인원 수",cex.lab=2)**

**title(main="은행 이용 고객 연령 분포",cex.main=2)**

**#b. 직업**

**h<-ggplot(bank, aes(x=job)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(angle=45, hjust=1,size=15))**

**h2<-h+scale\_x\_discrete(limits=c("blue-collar","management","technician","admin.","services","retired","self-employed","entrepreneur","unemployed","housemaid","student","unknown"))**

**#h3<-h2+ ggtitle("은행 이용 고객의 직업 종류") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "black"))**

**h3 <- h2 + geom\_text(aes(label=..count..), stat="count", vjust=1.0, colour="white")**

**h3+ labs(x="직업 종류", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 18, color = "black"))**

**#c. 결혼 유무**

**table(bank$marital)**

**prop.table( table(bank$marital,bank$y),1)**

**bar<-ggplot(bank, aes(x=marital,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill", width=0.7)**

**bar3 <- bar2+theme\_bw()**

**#bar3<-bar2 + ggtitle("결혼 여부 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="혼인 상태", y="예금 구매 여부")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**bar5<-bar4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**bar6<-bar5+geom\_text(x=1, y=0.03, label="0.1194", alpha=.1)+geom\_text(x=2, y=0.03, label="0.1012", alpha=.1) +geom\_text(x=3, y=0.03, label="0.1494", alpha=.1)**

**bar5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**#이혼,사별-> 싱글**

**bank2<-bank**

**bank2[bank2$marital=="divorced",3]<-"single"**

**table(bank2$marital)**

**prop.table( table(bank2$marital,bank$y),1)**

**bar<-ggplot(bank2, aes(x=marital,y=bank2$tk, fill=bank2$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**bar3<-bar2 + ggtitle("결혼 여부 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="결혼 여부", y="현재 등록 비율")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**bar5<-bar4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**bar5 +geom\_text(x=1, y=0.08, label="0.1012", alpha=.1) +geom\_text(x=2, y=0.08, label="0.1408", alpha=.1)**

**#큰 차이 없음 #범주형 변수에서 또다른 그룹화 필요로 보임**

**#d 교육**

**e1<-ggplot(bank, aes(x=education)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(angle=15, hjust=1,size=14))**

**e2<-e1+scale\_x\_discrete(limits=c("secondary","tertiary","primary","unknown") )**

**#e3<-e2 + ggtitle("은행 이용 고객의 학력") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e4<-e2+labs(x="학력", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**e4+geom\_text(aes(label=..count..), stat="count", vjust=1.4, colour="white", size=5)**

**#e. default**

**e2<-ggplot(bank, aes(x=default)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(size=14))**

**#e3<-e2 + ggtitle("채무 이행 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e4<-e2+labs(x="채무 불이행 여부", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**e4+geom\_text(aes(label=..count..), stat="count", vjust=-0.4, colour="black", size=5)**

**bar<-ggplot(bank, aes(x=default,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**bar3<-bar2 + ggtitle("채무 이행 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="채무 이행 여부", y="현재 등록 비율")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**bar4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**#일단 갯수에서 차이가 심하게 나는데 이 변수를 기준으로 y의 yes,no를 확인해봐도 극단적인 변화차이는 없어보여서 #샘플링 과정에서도 차이가 없으면 제거하여도 될 것으로 보인다.**

**#f. balance**

**summary(bank$balance)**

**hist(bank$balance)**

**quantile(bank$balance, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.99))**

**#qunatile에서 이정도로 극단적인 값을 가지는 1%미만의 값들을 제거하고 수행하는게 오히려 분포 구성이나 표준화 작업에서 유리해보인다**

**#극이상치 제거작업**

**bank4<-bank**

**bank4<-bank4[(bank4$balance>=-627 & bank4$balance<=13164.9),]**

**dim(bank4[(bank4$balance<-627 | bank4$balance>13164.9),])**

**boxplot(bank4$balance)**

**boxplot(bank4$balance~bank4$y)#약간 차이**

**hist(bank$balance,axes=F,ann=F,labels=T,col="#4F5458")**

**axis(1,at=seq(-10000,110000,5000))**

**axis(2,ylim=c(0,36000))**

**title(xlab="평균 연간 잔고(유로)",cex.lab=1.5)**

**title(ylab="인원 수",cex.lab=1.5)**

**title(main="은행 이용 고객 잔고 분포",cex.main=2)**

**#g. housing,**

**table(bank$housing)**

**bar<-ggplot(bank, aes(x=housing,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill", width=0.6)**

**#bar3<-bar2 + ggtitle("집 채무 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar3<-bar2+theme\_bw()**

**bar4<-bar3+labs(x="주택담보대출 여부", y="예금 구매 여부")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**bar5<-bar4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**bar5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**e2<-ggplot(bank, aes(x=housing)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(angle=0, hjust=1,size=14))**

**e3<-e2 + ggtitle("집 채무 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e3+labs(x="집 채무 여부", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**#h. loan**

**table(bank$loan)**

**bar<-ggplot(bank, aes(x=loan,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**#bar3<-bar2 + ggtitle("개인 채무 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar3<-bar2+theme\_bw()**

**bar4<-bar3+labs(x="개인 대출 여부", y="예금 구매 여부")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**bar5<-bar4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**bar5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**e2<-ggplot(bank, aes(x=loan)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(angle=0, hjust=1,size=14))**

**e3<-e2 + ggtitle("개인 채무 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e3+labs(x="개인 채무 여부", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**#i. contact(마지막 연락 방식)**

**table(bank$contact)**

**bar<-ggplot(bank, aes(x=contact,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**#bar3<-bar2 + ggtitle("연락수단 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar3<-bar2+theme\_bw()**

**bar4<-bar3+labs(x="연락 수단", y="현재 등록 비율")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**bar4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**e2<-ggplot(bank, aes(x=contact)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(size=14))**

**#e3<-e2 + ggtitle("채무 이행 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e4<-e2+labs(x="연락 수단", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**e4+geom\_text(aes(label=..count..), stat="count", vjust=-0.4, colour="black", size=5)**

**#j. 일별, 월별 (day,month)**

**bank$year<-2008**

**bank[27730:42591,"year"]<-2009**

**bank[42592:45211,"year"]<-2010**

**bank$year2 <- paste(bank$year,bank$month)**

**table(bank$year2)**

**#k. duration(지속시간)**

**summary(bank$duration)**

**box2<-ggplot(bank,aes(x=y,y=duration,fill=y))+geom\_boxplot()**

**#box3<-box2 + ggtitle("현 캠페인 등록 여부 별 지속시간") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**box3<-box2+theme\_bw()**

**box4<-box3+labs(x="예금 구매 여부", y="지속 시간")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**box5<-box4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**box5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**quantile(bank$duration, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.99))**

**bank6<-bank**

**bank6<-bank6[bank6$duration<=1269,]**

**box2<-ggplot(bank6,aes(x=y,y=duration,fill=y))+geom\_boxplot()**

**box3<-box2 + ggtitle("현 캠페인 등록 여부 별 지속시간") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**box4<-box3+labs(x="현재 등록 여부", y="지속 시간")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**box4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**#l. campaign(현재 캠페인 연결 횟수)**

**summary(bank$campaign)**

**quantile(bank$campaign, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.99))**

**boxplot(bank$campaign)**

**boxplot(bank$campaign~bank$y) #99퍼에 시도 16번인데 그 이상의 이상치는 시각화를 하는데 방해가 되어보임**

**bank7<-bank**

**bank7<-bank7[bank7$campaign<=16,] #극 이상치 제거후**

**box2<-ggplot(bank,aes(x=y,y=campaign,fill=y))+geom\_boxplot()**

**box3<-box2 + ggtitle("현 캠페인 등록 여부 별 연결 횟수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**box4<-box3+labs(x="현재 등록 여부", y="연결 횟수")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**box4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**#중간값에서는 차이가 없지만 애초에 1,2,3의 값이 너무 많은 상태이기 때문에 차이가 없다고 하기에는 애매하며 #오히려 yes의 경우는 median과 3rd의 차이가 없는 것을 보아 yes의 경우가 평균적인 campaign 연락 횟수가 적을 것으로 보인다.**

**#m. pdays**

**summary(bank$pdays)**

**quantile(bank$pdays, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.99))**

**bank8<-bank**

**bank8$cu<-"잠재고객"**

**bank8[bank$pdays!=-1,19]<-"이전고객"**

**table(bank8[,19])**

**bar<-ggplot(bank, aes(x=cu,y=bank8$tk, fill=bank8$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**bar3<-bar2 + ggtitle("이전 참가 여부 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="이전 참가 여부", y="현재 등록 비율")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**bar4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**#n. previous**

**summary(bank$previous)**

**quantile(bank13$previous, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.995))**

**bank13<-bank[bank$previous>0,]**

**boxplot(bank13$previous~bank13$y) #이전 캠페인 횟수와 현재 y의 관계**

**table(bank13$previous,bank13$y)**

**prop.table( table(bank13$previous,bank13$y),1) #잠재고객 이전에 참여 고객 구분 확률**

**table(bank13$y)#전체 ; 5260/(5260+39766) #0.1168**

**#0의 경우는 잠재고객이고 1부터는 이전부터 존재하였던 고객**

**bar<-ggplot(bank13, aes(x=as.character(bank13$previous),y=bank13$tk, fill=bank13$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**bar3<-bar2 + ggtitle("이전 횟수 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="이전 횟수", y="현재 등록 비율")+theme(axis.title = element\_text(face = "bold", size = 18, color = "darkblue"))**

**bar5<-bar4+labs(fill="현재 등록 여부")+theme(legend.title=element\_text(face = "bold", size = 12, color = "black"))**

**bar5+scale\_x\_discrete(limits=c("0","1","2","3","4","5","6","7","8","9","10","11","12"))**

**box2<-ggplot(bank13,aes(y=previous))+geom\_boxplot(color="#4F5458")**

**#box3<-box2 + ggtitle("현 캠페인 등록 여부 별 연결 횟수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**box3<-box2+theme\_bw()**

**box3+labs(y="이전 캠페인 연락 수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**#o. poutcome**

**table(bank$poutcome,bank$y)**

**prop.table( table(bank$poutcome,bank$y) ,1)**

**bar<-ggplot(bank, aes(x=poutcome,y=bank$tk, fill=bank$y)) #바로 위에 prop.table을 시각화**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**#bar3<-bar2 + ggtitle("이전 캠페인 현황 별 현 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar3<-bar2+theme\_bw()**

**bar4<-bar3+labs(x="이전 캠페인 결과", y="예금 구매 여부")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**bar5<-bar4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**bar5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**#해석을 해보자면 이전 캠페인에 성공을 한 경우에 바로 다음에 진행되는 캠페인에 참가를 하려는 경향이 커보이며 #잠재고객이 대부분인 unknwon에 대해서는 다른 곳에 비해서 성공확률이 낮으며 #이전 고객의 결과를 중 나머지인 failure과 other의 경우에 대해서는 그래도 잠재고객들 보다는 성공확률이 높다.**

**#p. ### y**

**e2<-ggplot(bank, aes(x=y)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(size=14))**

**#e3<-e2 + ggtitle("채무 이행 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e4<-e2+labs(x="타깃 변수", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**e4+geom\_text(aes(label=..count..), stat="count", vjust=-0.4, colour="black", size=5)**

**#y에 대한 누적(필요한 경우에 사용)**

**prop.table( table(bank$education,bank$y),1)**

**bar<-ggplot(bank, aes(x=education,y=bank$tk, fill=bank$y))**

**bar2<-bar+geom\_bar(stat="identity", position="fill")**

**bar3<-bar2+theme\_bw()**

**#bar3<-bar2 + ggtitle("학력 별 캠페인 비율") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**bar4<-bar3+labs(x="학력", y="예금 구매 여부")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**bar5<-bar4+labs(fill="예금 구매 여부")+theme(legend.title=element\_text(face = "bold", size = 15, color = "black"))**

**bar5+scale\_fill\_manual(values=c("#A0A19D", "#4F5458"))**

**#3. 사용하지 않을 변수와 복합적인 이상치에 대한 삭제 작업**

**#a. 사용하지 않을 변수인 contact 삭제**

**bank<-bank[,-9]**

**#b. 이전 참여고객의 이전 결과가 unknown인 5명에 대해서 제거**

**bank<-bank[(bank$pdays==-1 & bank$poutcome=="unknown") | (bank$pdays!=-1 & bank$poutcome!="unknown") ,]**

**#c. 복합적 이상치 제거작업**

**#c1. 초졸에 나이가 좀 있는데 학생인 사람 #44명**

**bank<-bank[bank$education!="primary" | bank$job!="student",]**

**#c2. 나이가 40살 미만, 직업이 은퇴인 사람 #25명**

**bank<-bank[ bank$age>39 | bank$job!="retired",] #이상치로 이야기가 나온 부분으로 제거를 고려해 볼 필요가 있어보인다.**

**#c3. campaign(현 캠페인 연결 횟수) 극이상치 처리**

**quantile(bank$campaign, probs=c(0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95,0.975,0.99))**

**bank<-bank[bank$campaign<=16,] #극 이상치 제거**

**#c4. previous에서 제일 큰 275값 1개 제거**

**summary(bank)**

**bank<-bank[bank$previous!=275,] #극 이상치 제거**

**#4. age, job 재 범주화, day, month로 week변수 만든 후 day변수 제거, education 이상치 처리**

**#a. age 재범주화**

**bank$age2<-"-01"**

**bank[bank$age<=29,"age2"]<-"20대이하"**

**bank[bank$age>=30 & bank$age<=39,"age2"]<-"30대"**

**bank[bank$age>=40 & bank$age<=49,"age2"]<-"40대"**

**bank[bank$age>=50 & bank$age<=59,"age2"]<-"50대"**

**bank[bank$age>=60 & bank$age<=69,"age2"]<-"60대"**

**bank[bank$age>=70,"age2"]<-"70대이상"**

**bank$age<-as.factor(bank$age2)**

**bank<-bank[,-17]**

**### 재범주화한 age 막대 그래프**

**bank$age2 <- NA**

**bank[bank$age<=29,"age2"]<-"20대이하"**

**bank[bank$age>=30 & bank$age<=39,"age2"]<-"30대"**

**bank[bank$age>=40 & bank$age<=49,"age2"]<-"40대"**

**bank[bank$age>=50 & bank$age<=59,"age2"]<-"50대"**

**bank[bank$age>=60 & bank$age<=69,"age2"]<-"60대"**

**bank[bank$age>=70,"age2"]<-"70대이상"**

**bank$age2 <- as.factor(bank$age2)**

**e2<-ggplot(bank, aes(x=age2)) +geom\_bar() + theme\_bw() + theme(axis.text.x=element\_text(size=14))**

**#e3<-e2 + ggtitle("채무 이행 별 인원수") + theme(plot.title = element\_text(family = "serif", face = "bold", hjust = 0.5, size = 18, color = "darkblue"))**

**e4<-e2+labs(x="연령대", y="인원수")+theme(axis.title = element\_text(face = "bold", size = 20, color = "black"))**

**e4+geom\_text(aes(label=..count..), stat="count", vjust=-0.4, colour="black", size=5)**

**#b. job 재범주화 #unknown제거(285개)**

**bank <- bank[bank$job!="unknown",]**

**levels(bank$job)**

**for(i in 1:nrow(bank)){**

**if(bank$job[i]=="admin."){**

**bank$job2[i] <- "admin"**

**}else if(bank$job[i]=="management"){**

**bank$job2[i] <- "management"**

**}else if(bank$job[i]=="entrepreneur"){**

**bank$job2[i] <- "entrepreneur"**

**}else if(bank$job[i]=="self-employed"){**

**bank$job2[i] <- "entrepreneur"**

**}else if(bank$job[i]=="blue-collar"){**

**bank$job2[i] <- "blue-collar"**

**}else if(bank$job[i]=="technician"){**

**bank$job2[i] <- "technician"**

**}else if(bank$job[i]=="services"){**

**bank$job2[i] <- "pink-collar"**

**}else if(bank$job[i]=="housemaid"){**

**bank$job2[i] <- "pink-collar"**

**}else{**

**bank$job2[i] <- "1unemployed"**

**} }**

**bank$job <- as.factor(bank$job2)**

**bank<-bank[,-17] #job2 제거**

**str(bank)**

**#c. 순서 가정하에 연도변수 추가 후 day->평일,휴일로 변경**

**rownames(bank)<-seq(length=nrow(bank)) #index 초기화**

**bank$year<-2008**

**bank[27074:41820,"year"]<-2009**

**bank[41821:44401,"year"]<-2010**

**#date변환용 임시 투입**

**bank$month2<-"-01"**

**bank[as.character(bank$month)=="feb","month2"]<-"-02"**

**bank[as.character(bank$month)=="mar","month2"]<-"-03"**

**bank[as.character(bank$month)=="apr","month2"]<-"-04"**

**bank[as.character(bank$month)=="may","month2"]<-"-05"**

**bank[as.character(bank$month)=="jun","month2"]<-"-06"**

**bank[as.character(bank$month)=="jul","month2"]<-"-07"**

**bank[as.character(bank$month)=="aug","month2"]<-"-08"**

**bank[as.character(bank$month)=="sep","month2"]<-"-09"**

**bank[as.character(bank$month)=="oct","month2"]<-"-10"**

**bank[as.character(bank$month)=="nov","month2"]<-"-11"**

**bank[as.character(bank$month)=="dec","month2"]<-"-12"**

**bank$date<-paste0(as.character(bank$year),bank$month2,"-",as.character(bank$day))**

**bank$date<-as.Date(bank$date)**

**bank$day2<-weekdays(bank$date)**

**bank$day3<-"평일"**

**bank[as.character(bank$day2)=="토요일","day3"]<-"휴일"**

**bank[as.character(bank$day2)=="일요일","day3"]<-"휴일"**

**bank$day<-as.factor(bank$day3)**

**bank<-bank[,c(-17,-18,-19,-20,-21)] #평일,휴일 이후에 사용한 변수들 전부 제거**

**#기준점 잡기용 월 정리**

**bank$month2<-"1jan"**

**bank[as.character(bank$month)=="feb","month2"]<-"2feb"**

**bank[as.character(bank$month)=="mar","month2"]<-"3mar"**

**bank[as.character(bank$month)=="apr","month2"]<-"4apr"**

**bank[as.character(bank$month)=="may","month2"]<-"5may"**

**bank[as.character(bank$month)=="jun","month2"]<-"6jun"**

**bank[as.character(bank$month)=="jul","month2"]<-"7jul"**

**bank[as.character(bank$month)=="aug","month2"]<-"8aug"**

**bank[as.character(bank$month)=="sep","month2"]<-"9sep"**

**bank[as.character(bank$month)=="oct","month2"]<-"l0oct"**

**bank[as.character(bank$month)=="nov","month2"]<-"l1nov"**

**bank[as.character(bank$month)=="dec","month2"]<-"l2dec"**

**bank$month<-as.factor(bank$month2)**

**bank<-bank[,-17]**

**#d. education 이상치 처리**

**set.seed(1234)**

**bank\_edu <- bank[bank$education!="unknown",]**

**bank\_edu\_pri <- bank\_edu[bank\_edu$education=="primary",]**

**bank\_edu\_sec <- bank\_edu[bank\_edu$education=="secondary",]**

**bank\_edu\_ter <- bank\_edu[bank\_edu$education=="tertiary",]**

**#pri, sec, ter 비율 맞추기**

**idx <- sample(nrow(bank\_edu\_sec), nrow(bank\_edu\_sec)\*0.295)**

**bank\_edu\_sec <- bank\_edu\_sec[idx,]**

**idx <- sample(nrow(bank\_edu\_ter), nrow(bank\_edu\_ter)\*0.513)**

**bank\_edu\_ter <- bank\_edu\_ter[idx,]**

**bank\_edu <- rbind(bank\_edu\_pri, bank\_edu\_sec, bank\_edu\_ter)**

**#train,test 나누기**

**set.seed(1234)**

**idx <- sample(nrow(bank\_edu), nrow(bank\_edu)\*0.7)**

**bank\_edu\_train <- bank\_edu[idx,]**

**bank\_edu\_test <- bank\_edu[-idx,]**

**bank\_edu\_check <- bank\_edu[-idx, "education"]**

**#age + marital + job2 68.66215% 가장 좋은 조합**

**set.seed(1234)**

**bank\_edu\_dt <- ctree(education ~ age + marital + job, data=bank\_edu\_train)**

**bank\_edu\_pred <- predict(bank\_edu\_dt, newdata=bank\_edu\_test, type="response")**

**bank\_edu\_table <- table(bank\_edu\_pred, bank\_edu\_check)**

**sum(diag(bank\_edu\_table))/sum(bank\_edu\_table)**

**#가장 좋은 조합을 unknwon에만 적용**

**bank\_edu <- bank[bank$education!="unknown",]**

**bank\_edu\_test <- bank[bank$education=="unknown",]**

**set.seed(1234)**

**bank\_edu\_dt <- ctree(education ~ age + marital + job, data=bank\_edu)**

**bank\_edu\_pred <- predict(bank\_edu\_dt, newdata=bank\_edu\_test, type="response")**

**table(bank\_edu\_pred)**

**bank\_edu\_test$education<-bank\_edu\_pred**

**bank<-rbind(bank\_edu,bank\_edu\_test)**

**bank$education<-as.factor( as.character(bank$education) ) #factor 3개로 변경작업**

**str(bank)**

**#cf 필요없는 변수 contact 처리, 이상치 처리, 변수변환 된 자료 내보내기**

**#write.table(bank, "C:/Users/bluedice/Desktop/BankData/bank\_real2.txt", sep = "," )**

**#전처리 된 데이터 내보낸거 다시 불러오기**

**#bank<-read.table("C:/Users/bluedice/Desktop/BankData/bank\_real2.txt",sep=",")**

**str(bank)**

**#5. 모델 나누기**

**#5-1-1. 잠재 고객 모델**

**bank\_pre0<-bank[bank$poutcome=="unknown",]**

**summary(bank\_pre0)**

**bank\_pre0 <- bank\_pre0[,-c(13,14,15)] #remove pdays, previous, poutcome**

**#5-1-2 y변수에 대한 사전적인 변수처리**

**#a. 범주형 변수**

**(marital,education,default,housing,loan,day,month)**

**#job,month은 중요해보여 검정없이 들고감**

**#marital**

**tk<-table(bank\_pre0$marital,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[1,2],tk[3,2]),c(tk[1,1]+tk[1,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[2,2],tk[3,2]),c(tk[2,1]+tk[2,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**#education**

**tk<-table(bank\_pre0$education,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[1,2],tk[3,2]),c(tk[1,1]+tk[1,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[2,2],tk[3,2]),c(tk[2,1]+tk[2,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**#default**

**tk<-table(bank\_pre0$default,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#housing**

**tk<-table(bank\_pre0$housing,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#loan**

**tk<-table(bank\_pre0$loan,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#day**

**tk<-table(bank\_pre0$day,bank\_pre0$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#day의 경우 pvalue채택이고 휴일의 경우 자체 갯수가 적기 때문에 제외하고 모델을 구축하기로 결정함**

**#b. 연속형,이산형 변수로 취급되는 balance, duration**

**#b1. duration**

**qqnorm(bank\_pre0$duration)**

**qqline(bank\_pre0$duration)**

**bank\_pre0$logduration<-log(bank\_pre0$duration+1) #최소가 0이라서 1추가함**

**qqnorm(bank\_pre0$logduration)**

**qqline(bank\_pre0$logduration)**

**bank\_pre0$duration<-bank\_pre0$logduration**

**bank\_pre0<-bank\_pre0[,-14]**

**#b2. balance**

**#은행예금의 이익 시점을 위해서 0을 제거하고 절대값을 씌워서 변환을 실시함**

**qqnorm(bank\_pre0$balance)**

**qqline(bank\_pre0$balance)**

**bank\_pre0<-bank\_pre0[bank\_pre0$balance!=0,]**

**bank\_pre0$balance0<-log(abs(bank\_pre0$balance)) #로그**

**bank\_pre0$balance1<-sqrt(abs(bank\_pre0$balance)) #제곱근**

**bank\_pre0$balance2<-1/(abs(bank\_pre0$balance)) #역변환**

**model<- glm(y ~ balance+balance0+balance1+balance2 ,data=bank\_pre0, family=binomial(link="logit") )**

**summary(model) #pvalue가 가장 낮은 로그모델 선택**

**qqnorm(bank\_pre0$balance0)**

**qqline(bank\_pre0$balance0)**

**#정규성 어느 정도 만족**

**bank\_pre0$balance<-bank\_pre0$balance0**

**bank\_pre0<-bank\_pre0[,c(-14,-15,-16)]**

**#5-1-3 로지스틱**

**bank\_pre0 <- bank\_pre0[,-9] #remove day**

**#bank\_pre0(잠재고객) y 1:1 맞추기**

**bank\_pre0\_y0 <- bank\_pre0[bank\_pre0$y=="no",]**

**bank\_pre0\_y1 <- bank\_pre0[bank\_pre0$y=="yes",]**

**nrow(bank\_pre0\_y1)**

**set.seed(1234)**

**idx <- sample(nrow(bank\_pre0\_y0), 3151 )**

**bank\_pre0\_sample <- bank\_pre0\_y0[idx,]**

**summary(bank\_pre0\_sample)**

**bank\_pre0\_sample$wei<-9.5**

**bank\_pre0\_y1$wei<-1**

**bank\_pre0 <- rbind(bank\_pre0\_sample, bank\_pre0\_y1)**

**summary(bank\_pre0)**

**set.seed(12366777)**

**idx <- sample(nrow(bank\_pre0), nrow(bank\_pre0)\*0.7)**

**bank\_pre0\_train <- bank\_pre0[idx,]**

**bank\_pre0\_test <- bank\_pre0[-idx,]**

**bank\_pre0\_check <- bank\_pre0[-idx, "y"]**

**#나중에 10분위 분석용 자료 추출**

**#write.table(bank\_pre0\_train, "C:/Users/bluedice/Desktop/BankData/bank\_pre0\_train.txt", sep = "," ,quote=F,row.names=F)**

**#write.table(bank\_pre0\_test, "C:/Users/bluedice/Desktop/BankData/bank\_pre0\_test.txt", sep = "," ,quote=F,row.names=F)**

**bank\_pre0\_logit <- glm(y ~age+job+marital+education+default+balance+housing+loan+month+duration+campaign, data=bank\_pre0\_train, family=binomial(link="logit"),weights=wei)**

**Anova(bank\_pre0\_logit,type="III",test="Wald")**

**bank\_pre0\_logit <- glm(y ~age+job+marital+default+balance+housing+loan+month+duration+campaign, data=bank\_pre0\_train, family=binomial(link="logit"),weights=wei)**

**out <- step(bank\_pre0\_logit, direction = "both")**

**summary(out)**

**vif(bank\_pre0\_logit)**

**bank\_pre0\_logit\_pred <- predict(out, newdata=bank\_pre0\_test, type="response")**

**bank\_pre0\_logit\_pred[bank\_pre0\_logit\_pred > 0.5] <- 1**

**bank\_pre0\_logit\_pred[bank\_pre0\_logit\_pred <= 0.5] <- 0**

**tk<-table(bank\_pre0\_logit\_pred, bank\_pre0\_check)**

**sum(diag(tk))/sum(tk) #weight쓰면 64.41퍼**

**#5-2-1 이전 참여고객 로지스틱**

**bank\_pre1 <- bank[bank$poutcome=="failure"|bank$poutcome=="other"|bank$poutcome=="success",]**

**bank\_pre1$poutcome<-as.factor(as.character(bank\_pre1$poutcome))**

**summary(bank\_pre1)**

**#5-2-2 y변수에 대한 사전적인 변수처리**

**#a. 범주형 변수**

**(marital,education,default,housing,loan,day,month)**

**#job,month,poutcome은 중요해보이기 때문에 검정없이 들고감**

**#marital**

**tk<-table(bank\_pre1$marital,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=1-(0.05/3)) #divorce랑 marry랑 동일하게 나옴**

**prop.test( c(tk[1,2],tk[3,2]),c(tk[1,1]+tk[1,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[2,2],tk[3,2]),c(tk[2,1]+tk[2,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**bank\_pre1$marry<-"marry"**

**bank\_pre1[as.character(bank\_pre1$marital)=="single","marry"]<-"single"**

**tk<-table(bank\_pre1$marry,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#결혼여부로 차이 존재**

**#education**

**tk<-table(bank\_pre1$education,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=1-(0.05/3)) #primary와 secondary는 귀무가설 채택**

**prop.test( c(tk[1,2],tk[3,2]),c(tk[1,1]+tk[1,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[2,2],tk[3,2]),c(tk[2,1]+tk[2,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**bank\_pre1$university<-"no\_univ"**

**bank\_pre1[as.character(bank\_pre1$education)=="tertiary","university"]<-"univ"**

**tk<-table(bank\_pre1$university,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95) #primary와 secondary는 귀무가설 채택**

**#대학여부로는 차이 존재**

**#default**

**tk<-table(bank\_pre1$default,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#housing**

**tk<-table(bank\_pre1$housing,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#loan**

**tk<-table(bank\_pre1$loan,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#day**

**tk<-table(bank\_pre1$day,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=0.95)**

**#day의 경우 pvalue채택이고 휴일의 경우 자체 갯수가 적기 때문에 제외하고 모델을 구축하기로 결정함**

**#poutcome**

**tk<-table(bank\_pre1$poutcome,bank\_pre1$y)**

**prop.test( c(tk[1,2],tk[2,2]),c(tk[1,1]+tk[1,2],tk[2,1]+tk[2,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[1,2],tk[3,2]),c(tk[1,1]+tk[1,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**prop.test( c(tk[2,2],tk[3,2]),c(tk[2,1]+tk[2,2],tk[3,1]+tk[3,2]),alternative="two.sided",conf.level=1-(0.05/3))**

**#실제로도 각 그룹마다 차이 존재(즉 사용)**

**str(bank\_pre1)**

**bank\_pre1<-bank\_pre1[,c(-3,-4,-9)]**

**#marital, education, day제거(대체됨)**

**#5-2-b.**

**연속형,이산형 변수로 취급되는 balance, duration**

**#duration**

**qqnorm(bank\_pre1$duration)**

**qqline(bank\_pre1$duration)**

**bank\_pre1$logduration<-log(bank\_pre1$duration+1) #최소가 0이라서 1추가함**

**qqnorm(bank\_pre1$logduration)**

**qqline(bank\_pre1$logduration) #어느정도 정규성 만족**

**bank\_pre1$duration<-bank\_pre1$logduration**

**bank\_pre1<-bank\_pre1[,-16]**

**#balance**

**#은행예금의 이익 시점을 위해서 0을 제거하고 절대값을 씌워서 변환을 실시함**

**qqnorm(bank\_pre1$balance)**

**qqline(bank\_pre1$balance)**

**bank\_pre1<-bank\_pre1[bank\_pre1$balance!=0,]**

**bank\_pre1$balance0<-log(abs(bank\_pre1$balance)) #로그**

**bank\_pre1$balance1<-sqrt(abs(bank\_pre1$balance)) #제곱근**

**bank\_pre1$balance2<-1/(abs(bank\_pre1$balance)) #역변환**

**model<- glm(y ~ balance+balance0+balance1+balance2 ,data=bank\_pre1, family=binomial(link="logit") )**

**summary(model) #pvalue가 가장 낮은 로그 모델 선택**

**qqnorm(bank\_pre1$balance0)**

**qqline(bank\_pre1$balance0)**

**#정규성 어느 정도 만족**

**bank\_pre1$balance<-bank\_pre1$balance0**

**bank\_pre1<-bank\_pre1[,c(-16,-17,-18)]**

**#5-2-3 로지스틱**

**#bank\_pre1(이전 참여 고객) y 1:1 맞추기**

**summary(bank\_pre1)**

**bank\_pre1\_y0 <- bank\_pre1[bank\_pre1$y=="no",]**

**bank\_pre1\_y1 <- bank\_pre1[bank\_pre1$y=="yes",]**

**set.seed(1234)**

**idx <- sample(nrow(bank\_pre1\_y0), 1787)**

**bank\_pre1\_sample <- bank\_pre1\_y0[idx,]**

**summary(bank\_pre1\_sample)**

**bank\_pre1\_sample$wei<-3.33333**

**bank\_pre1\_y1$wei<-1**

**bank\_pre1 <- rbind(bank\_pre1\_sample, bank\_pre1\_y1)**

**summary(bank\_pre1)**

**#logistic(bank\_pre1)**

**set.seed(1234)**

**idx <- sample(nrow(bank\_pre1), nrow(bank\_pre1)\*0.7)**

**bank\_pre1\_train <- bank\_pre1[idx,]**

**bank\_pre1\_test <- bank\_pre1[-idx,]**

**bank\_pre1\_check <- bank\_pre1[-idx, "y"]**

**#나중에 10분위 분석용 자료 추출**

**#write.table(bank\_pre1\_train, "C:/Users/bluedice/Desktop/BankData/bank\_pre1\_train.txt", sep = "," ,quote=F,row.names=F)**

**#write.table(bank\_pre1\_test, "C:/Users/bluedice/Desktop/BankData/bank\_pre1\_test.txt", sep = "," ,quote=F,row.names=F)**

**bank\_pre1\_logit <- glm(y ~age+job+default+balance+housing+loan+month+duration+campaign+pdays+previous+poutcome+marry+university, data=bank\_pre1\_train, family=binomial(link="logit"),weights=wei)**

**Anova(bank\_pre1\_logit,type="III",test="Wald")**

**bank\_pre1\_logit <- glm(y ~age+job+balance+month+duration+housing+pdays+poutcome+university, data=bank\_pre1\_train, family=binomial(link="logit"),weights=wei)**

**out <- step(bank\_pre1\_logit, direction = "both")**

**summary(out)**

**vif(bank\_pre1\_logit)**

**bank\_pre1\_logit\_pred <- predict(out, newdata=bank\_pre1\_test, type="response")**

**bank\_pre1\_logit\_pred[bank\_pre1\_logit\_pred > 0.5] <- 1**

**bank\_pre1\_logit\_pred[bank\_pre1\_logit\_pred <= 0.5] <- 0**

**tk<-table(bank\_pre1\_logit\_pred, bank\_pre1\_check)**

**sum(diag(tk))/sum(tk) #weight쓰면 73.066**

**SAS Part**

**/\* 6-1 decile analysis 잠재고객 \*/**

**/\* train data 불러오기 \*/**

**data bank\_pre0\_train;**

**infile 'C:\Users\bluedice\Desktop\bank\_pre0\_train.txt' dlm=',' firstobs=2 missover;**

**input age$ job$ marital$ education$ default$ balance housing$ loan$ month$ duration campaign y$ wei;**

**if y='yes' then y2=1;**

**if y='no' then y2=0;**

**records=1;**

**adj=1;**

**wei2=wei\*adj;**

**run;**

**/\*test data 불러오기 \*/**

**data bank\_pre0\_test;**

**infile 'C:\Users\bluedice\Desktop\bank\_pre0\_test.txt' dlm=',' firstobs=2;**

**input age$ job$ marital$ education$ default$ balance housing$ loan$ month$ duration campaign y$ wei;**

**if y='yes' then y2=1;**

**if y='no' then y2=0;**

**records=1;**

**adj=.;**

**wei2=wei\*adj;**

**run;**

**data bank\_full;**

**set bank\_pre0\_train bank\_pre0\_test;**

**run;**

**/\* train decile analysis 과정 \*/**

**proc logistic data=bank\_full(keep=y2 age job marital default balance housing loan month duration campaign wei records adj wei2);**

**weight wei2;**

**class age job marital default housing loan month;**

**model y2(event='1')=age job marital default balance housing loan month duration campaign /link=logit aggregate scale=none selection=stepwise;**

**output out=k pred=pred;**

**run;**

**proc sort data=k;**

**by descending pred; run;**

**data training\_a ;**

**set k ;**

**if adj^=. ;**

**run ;**

**data training\_a;**

**set training\_a;**

**if \_n\_<=441 then q\_10=0;**

**else if \_n\_<=882 then q\_10=1;**

**else if \_n\_<=1323 then q\_10=2;**

**else if \_n\_<=1764 then q\_10=3;**

**else if \_n\_<=2205 then q\_10=4;**

**else if \_n\_<=2647 then q\_10=5;**

**else if \_n\_<=3088 then q\_10=6;**

**else if \_n\_<=3529 then q\_10=7;**

**else if \_n\_<=3970 then q\_10=8;**

**else q\_10=9; run ;**

**/\* decile analysis 표 생성하기 \*/**

**title1 'decile analysis - y - 이전 미참여 고객 train set ';**

**title2 ' training data - score selection';**

**proc tabulate data=training\_a;**

**class q\_10;**

**var records y2 pred;**

**table q\_10='Decile' all='Total',**

**records='Prospects'\*sum=' '\*f=comma10.**

**pred='Predicted Probability'\*(mean=' '\*f=11.5)**

**y2='Percent y2'\*(mean=' '\*f=11.5)**

**/rts=9 row=float; run;**

**/\* test decile analysis 과정 \*/**

**/\* test \*/;**

**data test\_a ; set k ; if adj=. ; run;**

**data test\_a;**

**set test\_a;**

**if \_n\_<=189 then q\_10=0;**

**else if \_n\_<=378 then q\_10=1;**

**else if \_n\_<=567 then q\_10=2;**

**else if \_n\_<=756 then q\_10=3;**

**else if \_n\_<=946 then q\_10=4;**

**else if \_n\_<=1135 then q\_10=5;**

**else if \_n\_<=1324 then q\_10=6;**

**else if \_n\_<=1513 then q\_10=7;**

**else if \_n\_<=1702 then q\_10=8;**

**else q\_10=9; run ;**

**/\* decile analysis 표 생성하기 \*/**

**title1 'decile analysis - y - 이전 미참여 고객 test set ';**

**title2 ' training data - score selection';**

**proc tabulate data=test\_a;**

**class q\_10;**

**var records y2 pred;**

**table q\_10='Decile' all='Total',**

**records='Prospects'\*sum=' '\*f=comma10.**

**pred='Predicted Probability'\*(mean=' '\*f=11.5)**

**y2='Percent y2'\*(mean=' '\*f=11.5)**

**/rts=9 row=float; run;**

**/\*confusion matrix \*/;**

**data k;**

**set k;**

**if pred>0.5 then pred\_y=1;**

**else pred\_y=0; run;**

**/\*training set\*/**

**proc freq data=k;**

**table y2 \* pred\_y;**

**where adj^=.; run;**

**/\*test set\*/**

**proc freq data=k;**

**table y2 \* pred\_y;**

**where adj=.; run;**

**/\* 6-2 decile analysis 참여 고객 \*/**

**/\* train data 불러오기 \*/**

**data bank\_pre1\_train;**

**infile 'C:\Users\bluedice\Desktop\bank\_pre1\_train.txt' dlm=',' firstobs=2 missover;**

**input age$ job$ default$ balance housing$ loan$ month$ duration campaign pdays previous poutcome$ y$ marry$ university$ wei;**

**if y='yes' then y2=1;**

**if y='no' then y2=0;**

**records=1;**

**adj=1;**

**wei2=wei\*adj; run;**

**/\* test data 불러오기 \*/**

**data bank\_pre1\_test;**

**infile 'C:\Users\bluedice\Desktop\bank\_pre1\_test.txt' dlm=',' firstobs=2;**

**input age$ job$ default$ balance housing$ loan$ month$ duration campaign pdays previous poutcome$ y$ marry$ university$ wei;**

**if y='yes' then y2=1;**

**if y='no' then y2=0;**

**records=1;**

**adj=.;**

**wei2=wei\*adj; run;**

**data bank\_full;**

**set bank\_pre1\_train bank\_pre1\_test;**

**run;**

**/\* train decile analysis 과정 \*/**

**proc logistic data=bank\_full(keep=y2 age job balance housing month duration pdays poutcome university wei records adj wei2);**

**weight wei2;**

**class age job housing month poutcome university;**

**model y2(event='1')= age job balance housing month duration pdays poutcome university /link=logit aggregate scale=none selection=stepwise;**

**output out=k pred=pred; run;**

**proc sort data=k;**

**by descending pred; run;**

**data training\_a ;**

**set k ;**

**if adj^=. ; run ;**

**data training\_a;**

**set training\_a;**

**if \_n\_<=250 then q\_10=0;**

**else if \_n\_<=500 then q\_10=1;**

**else if \_n\_<=750 then q\_10=2;**

**else if \_n\_<=1000 then q\_10=3;**

**else if \_n\_<=1251 then q\_10=4;**

**else if \_n\_<=1501 then q\_10=5;**

**else if \_n\_<=1751 then q\_10=6;**

**else if \_n\_<=2001 then q\_10=7;**

**else if \_n\_<=2251 then q\_10=8;**

**else q\_10=9; run ;**

**/\* decile analysis 표 생성하기 \*/**

**title1 'decile analysis - y - 기참여 고객 train set ';**

**title2 ' training data - score selection';**

**proc tabulate data=training\_a;**

**class q\_10;**

**var records y2 pred;**

**table q\_10='Decile' all='Total',**

**records='Prospects'\*sum=' '\*f=comma10.**

**pred='Predicted Probability'\*(mean=' '\*f=11.5)**

**y2='Percent y2'\*(mean=' '\*f=11.5)**

**/rts=9 row=float; run;**

**/\* test decile analysis 과정 \*/**

**/\* test \*/**

**data test\_a ;**

**set k ;**

**if adj=. ; run;**

**data test\_a;**

**set test\_a;**

**if \_n\_<=107 then q\_10=0;**

**else if \_n\_<=215 then q\_10=1;**

**else if \_n\_<=322 then q\_10=2;**

**else if \_n\_<=429 then q\_10=3;**

**else if \_n\_<=537 then q\_10=4;**

**else if \_n\_<=644 then q\_10=5;**

**else if \_n\_<=751 then q\_10=6;**

**else if \_n\_<=858 then q\_10=7;**

**else if \_n\_<=966 then q\_10=8;**

**else q\_10=9; run ;**

**/\* decile analysis 표 생성하기 \*/**

**title1 'decile analysis - y - 기참여고객 test set ';**

**title2 ' training data - score selection';**

**proc tabulate data=test\_a;**

**class q\_10;**

**var records y2 pred;**

**table q\_10='Decile' all='Total',**

**records='Prospects'\*sum=' '\*f=comma10.**

**pred='Predicted Probability'\*(mean=' '\*f=11.5)**

**y2='Percent y2'\*(mean=' '\*f=11.5)**

**/rts=9 row=float;**

**run;**

**/\*confusion matrix \*/**

**data k;**

**set k;**

**if pred>0.5 then pred\_y=1;**

**else pred\_y=0; run;**

**/\*training set\*/**

**proc freq data=k;**

**table y2 \* pred\_y;**

**where adj^=.; run;**

**/\*test set\*/**

**proc freq data=k;**

**table y2 \* pred\_y;**

**where adj=.;**

**run;**

1. 2008년도 유로 환율 (<https://spot.wooribank.com/pot/Dream?withyou=FXXRT0016>) [↑](#footnote-ref-1)
2. 포르투갈 2008년도 1인당 GDP (<https://www.ceicdata.com/ko/indicator/portugal/gdp-per-capita>) [↑](#footnote-ref-2)
3. 대한민국 2008년도 1인당 GDP (<http://datatopics.worldbank.org/world-development-indicators/>) [↑](#footnote-ref-3)
4. 미국의 표준직업분류표 (https://works.pe.kr/classam.htm) [↑](#footnote-ref-4)
5. 한국의 표준직업분류(http://kssc.kostat.go.kr/ksscNew\_web/kssc/main/main.do?gubun=1&pageChk=Y) [↑](#footnote-ref-5)
6. 포르투갈 국가공휴일 (<https://www.timeanddate.com/holidays/portugal/2008>) [↑](#footnote-ref-6)