Ⅰ. 목적 설정하기

[성공적인 목표 모형 단계]

1. 목적 설정
2. 데이터 선택
3. 모형 구축을 위한 데이터 준비
4. 변수 선택 및 변환
5. 모형 구축 과정
6. 모형 평가
7. 모형의 실행 및 유지

[프로젝트 목표 정하기]

<목표모형>

- 예측모형 : 미래의 행위 예측 계산
- 기술모형 : 자료의 모양을 기술통계량으로 표현

<고려사항>

- 새로운 고객 획득 : 특정 고객들에게만 반응 모형 적용
- 새로운 고객들이 기업에 이익이 되기를 원함 : 평생고객가치모형 적용
- 신뢰할 수 없는 고객 피하기 : 리스크 또는 승인모형
- 현재의 고객 특성 이해 : 고객유지모형 또는 고객이탈방지모형 이용_이탈할 고객을 현재 시점에서 식별.
- 판매량 늘이기 : 새로운 고객 획득 모형 또는 교차판매
- 비용 줄이기 : 새로운 고객 획득 모형 또는 고객 관계 관리 모형 이용 -> 목표 고객 분류

<고객 프로파일 분석>

- 관심있는 고객 모집단의 공통 특징
- (인구통계변수) 평균연령, 성별, 결혼 여부, 평균거주기간
- 고객 거래기간, 평균 리스크, 기간내의 평균판매액, 평균판매수, 평균순이익

<고객 분류>

- 일반적으로 수익성 또는 잠재적인 시장성에 근거하여 분류함.
- ex) 소매시장에서 소비자들의 구매 행위에 근거하여 분류
- ex) 신용카드나 대출회사에서 리스크와 예금잔고 기준에 근거하여 고객 분류

<반응 모형>

- 어떤 제품이나 서비스에 응답할 사람이 누구인지 예측하기.
  (과거에 비슷한 모집단 or 응답행위에 기초하여)

<리스크 모형>

- 새로운 고객이 대출한 돈을 갚지 않거나 파산 할 확률 예측
- 은행 or 보험회사

<활성화 모형>

- 어떤 성향의 예비 고객이 완전한 고객이 되는지의 여부 예측 모형
- 활성화 모형 구축 방법
- 1) 응답 예측 모형, 2) 응답후 사용할 것인가 예측 모형 ⇨ 응답부터 사용할 확률 : 두 모형값의 곱

<교차판매>
- 기존의 고객이 같은 회사의 다른 상품이나 서비스를 이용할 확률 예측

<상향판매>
- 같은 상품이나 서비스를 더 많이 이용할 확률 예측

<순현재가NPV 모형>
- 미리 정한 앞으로의 기간동안 어떤 상품의 전체적인 수익성 예측
- 순현재가 : 향후 몇 년간의 가격을 현재의 원화로 계산한 것.

<평생가치 모형>
- 미리 정한 앞으로의 기간동안 고객의 전체적인 수익성 예측
- 순현재가와 마찬가지로 향후 몇 년간의 가격을 원화로 계산.

[모형 구축 방법 선택]
- 전체 프로젝트의 성공이 모형 구축 단계에서 어떤 모형을 선택하느냐의 문제보다 더 중요함.

<선형 회귀분석>
- 연속적인 두 변수 사이의 관계를 알아보는 통계적인 방법

<로지스틱 회귀분석>
- 종속변수가 연속이지 않고 범주형일 때 (보통 두 개의 범주)
- 로지스틱 회귀식에서 우리가 예측하고자 하는 확률 p는 일반적으로 반응변수가 1일 확률
- $\log(\frac{p}{1-p})$는 독립변수들과 선형적인 관계가 있음.

<신경망 분석>
- 패턴인식의 한 분야로 과거의 경험이나 지식을 습득하여 오류를 최소화하는 과정 포함.
- 가장 많이 사용하는 신경망 분석은 각 노드를 최적화하기 위한 S자형 함수(sigmodel fuction)
- 내재되어있는 로지스틱 모형을 연속적으로 사용한 것
- 반응변수가 두 개의 범주인 경우 신경망 분석이 우수하다고 알려짐
- 자료의 비선형적인 관계를 찾아낼 수 있음.
- 과대적합하는 경향이 있어서 새로운 자료가 주어졌을 때 적당하지 않을 수 있음.
- 모형 결과 해석 어려움

<분류의사 결정나무>
- 어미마디와 자식마디 사이의 종속변수 차이를 가장 크게 하면서 데이터를 연속적으로 분리하는 것.
- 종속변수의 값에 가장 많은 차이를 만들어내는 그룹이나 소집단으로 분류해나가는 것.
- 고객 행위를 유발시키는 요인을 찾아내는데 유용함.
- 비선형적인 관계도 알아낼 수 있음.
- 독립변수들 사이의 상호작용도 찾아낼 수 있음.
- 결과를 설명하기 쉬움

Ⅱ. 데이터 출처 선택

[데이터의 종류]

<인구통계 데이터>
- 개인이나 가구별 특성
- 성별, 나이, 결혼여부, 소득, 주택소유여부, 거주종류, 교육수준, 종교, 자녀 등
- 안정적인 데이터값 -> 예측 모형 구축 유용

<고객행위 데이터>
- 고객의 행위 측정
- 판매액, 상품종류, 판매일, 지불액, 지불날짜, 고객 서비스행위, 보험료 청구, 수납 불이행 등
- 고객의 미래 행동을 예측하는데 정확하고 예측력이 뛰어남

<고객 심리나 태도 데이터>
- 개인의견, 생활습관, 개인적인 가치
- 결혼, 자녀출산, 대학입학, 퇴직등과 같이 삶에 있어서 발생하는 일련의 사건들과 관련된 상품이나 서비스 개발을 위해서 중요한 데이터
- 고객의 실제 행위와 직간접적으로 연관될 수 있는 고객의 잠재적인 행동.

[표본추출방법]
- 기존모형에 대한 대체 모형 개발할 때 :
  기존 모형에 의해 선택된 그룹 + 기존의 모형에 선택되지 않은 그룹에서 무작위 표본추출
- 계층표본추출은 어떤 독립변수 중 한 특성이 현저하게 많을 때 사용하기 좋음
- ex) 성별 변수가 대부분이 남성인 데이터. 모형결과를 이용하여 점수화시킬 때 남녀의 비율이 비슷하게 되기를 원함. 고객 목록을 추출할 때 남성으로부터 1/1000 표본 추출. 여성으로부터 1/100 표본추출.
  남성, 여성의 표본 데이터세트를 각각 만들고 합칠 때, 원래 모집단의 비율로 다시 맞추기 위하여
  가중치를 다르게 준다. (남성일 때 weight = 1000, 여성일 때 weight = 100)

Ⅲ. 모형구축을 위한 데이터 준비

[모형구축을 위한 데이터준비]
- 데이터의 특성을 이해하는 것이 좋은 모형을 위한 첫 번째 단계
- 데이터 정제과정 : 데이터 오류, 이상값과 결측값들을 조사하여 처리함.
- 변수요약, 변수들끼리의 비율을 통해 기존의 변수를 통합, 제거 또는 새로운 변수를 생성함.

[데이터세트 생성_표본추출]
- 목표모형에서 관심그룹은 일반적으로 전체에 비해서 작은부분을 차지함. (10% 이내)
- > 관련 그룹을 모두 추출하고 비관심 그룹으로부터 임의추출을 실시함. (50,000 – 75,000개의 관측값)
- ex) 13,868(1062 + 12806)명의 응답자와 715,360명의 비응답자. -> 응답자 모두 추출하고 비응답자 중 1/10을 임의추출함.

[데이터 정제]
- 결측치
- 이상치

Ⅳ. 변수 선택 및 변환

[새로운 변수 만들기]
- 변수들을 합치거나 분할하여 예측력있는 새로운 변수 찾기

[변수 선택하기]
- 모든 사용가능한 변수들을 생성하였다면 가장 영향력이 있는 변수들을 선택하는 과정이 필요함.
- 연속형 변수 : 로지스틱 모형 적합, 카이제곱 검정
- 범주형 변수 : 분할표 검정, 카이제곱 검정 (p102-103)
    - 카이제곱 통계량 유의적(24.817 / 0.001) -> active와 pop_den은 서로 관련있음
      -> 최종모형에 pop_den이 후보변수로 채택될 수 있음.
    - 카이제곱 통계량 = 0.352 / 0.553 이므로 두 변수는 서로 통계적으로 독립
      -> 최종모형을 위한 변수로 선택되지 않음. 두 변수가 독립이라는 것은 분할표로 알 수 있음.
      -> active에 영향을 주지 못하는 변수. 두 변수는 관계가 없음.

[선형적인 독립변수 개발하기_범주형변수]
- 로지스틱 회귀분석에서는 모든 독립변수를 연속형으로 간주하기 때문에 범주형 변수의 경우에는 연속형에 맞게 변수를 다시 조정해야함.
- 가장 많이 사용하는 방법은 지시변수를 생성하는 것.
- 지시변수를 생성할 때는 지시변수의 수가 항상 원래 범주형 변수의 계급 수보다 하나 적게 해야함.

[상호작용찾기]
- 변수들 사이에 상호작용 찾는 하나의 방법 : 두 변수의 가능한 모든 조합으로 이루어진 변수를 생성하여 유의성 검정 (변수의 개수가 많은 경우는 부적합)
- 의사결정나무를 수행하면 상호작용을 찾을 수 있는 빠른 방법을 제공함.

Ⅴ. 모형 구축 과정 및 평가

- 데이터 세트를 모형개발과 모형 평가를 위한 데이터세트로 나눔.
- 반응변수를 가장 잘 예측하는 독립변수들을 선택하는 여러 가지 기법들을 시도
- 구축된 모형을 평가하고 비교하기 위하여 십분위분석법 시행

[모형구축과정]
- 모형구축방법으로 로지스틱 회귀모형을 사용하는 이유
    1) 모형구축이 올바르다면 정확성이 우수함
    2) 구축과정이 용이하고 해석 쉬움
    3) 과대적합 가능성 적음
    4) 오차를 최소화하는 선형적인 관계를 찾는데 우수
- 변수선택이나 새로운 변수 생성과정을 통하여 최종 모형의 후보변수로 독립변수 선택

* 대량의 데이터를 이용하여 예측 모형을 구축하는 경우 다중공선성이 큰 문제가 아닐 수 있음. (p131-133)

[데이터 나누기]
- 모형구축 데이터세트와 평가용 데이터세트로 나눔.
- 모형개발에 사용되지 않는 데이터를 가지고 개발된 모형 평가

- 목표그룹이 비목표그룹에 비해 상대적으로 적은 경우
  목표그룹 : 모형구축과 평가용 데이터세트에 모두 사용
  비목표그룹 : 모형구축과 평가용 데이터세트로 분할함

<방법1 : 하나의 모형만 구축>
- 구축된 모형의 결과 평가방법 : 십분위분석법
- 데이터세트를 10개의 비슷한 크기의 그룹으로 나누기 (십분위)
- 먼저 데이터세트에서 가중치의 합계 계산
- 전체 데이터세트에서 구축된 모형의 결과인 활성화 확률을 정렬한 후 관찰치의 개수가 비슷한 그룹으로 10등분함

- 평가용 데이터세트를 이용하여 십분위 분석표를 작성. 이것은 구축된 모형의 로버스트 정도나 다른 데이터에 대해서도 모형이 잘 적용되는지 알아 볼 수 있는 좋은 방법이 됨.

<방법2 : 두 개의 모형 구축하기>

Ⅵ. 모형평가

- 이익도표
- 구축된 모형의 결과를 다른 데이터세트에 적용시켜 검정
- 재표본 방법을 사용하여 구축된 모형 계수들의 신뢰구간을 추정

[이익도표]
- 모형구축을 하지 않았을 때 또는 평균적인 모형의 적합성보다 구축된 모형이 얼마나 잘 적용되는지를 각 십분위에서 알아보는 것.
- 십분위 0에서는 평균보다 실제 활성화된 관측값을 더 포함하고 있음.
- 십분위 2까지는 구축된 모형이 평균모형보다 더 좋음.
- 누적리프트에 의하면 십분위 3까지의 모형의 적합도는 모형구축을 하지 않았을 때 또는 평균 모형보다 더 적합함.

[데이터세트 점수화하기]
- 목표모형을 구축하는 일반적인 목적은 모형을 구축한 데이터세트에 있는 관측값들에 대해서가 아니라 다른 데이터세트에 있는 관측값들에 대해서 예측하기 위한 것.
- 모형의 안정성이나 로버스트 정도를 알아보는 가장 좋은 방법은 구축된 모형의 목적을 가장 근접하게 반영할 수 있는 다른 캠페인에 대해 구축된 모형을 적용시켜 보는 것.

[재표본 resampling]
- 구축된 모형들의 안정성이나 로버스트 정도를 검정하는 기법
- 분포에 근거를 둔 모수적인 추정법이 아니라 경험이나 관찰에 의한 경험적인 추정법
- 장점1 : 반복적인 표본추출에 사용함으로써 모형 계수나 모수 추정에서 과대적합을 피할 수 있음.
- 장점2 : 구축된 모형의 결과를 평가하기 위하여 반복적인 표본추출을 사용함으로써 과대적합을 발견할 수 있음.
- 모형평가의 한 방법으로써 재표본법을 사용하여 주어진 추정값의 신뢰구간을 계산하는 방법에 대해서 알아보고자 함.
- 잭나이프 방법과 부스트랩 방법

<잭나이프 방법>
- 하나의 관측값만 남겨두는 방법에 근거를 둔 재표본 방법
- N개의 관측값을 가진 데이터세트에서 잭나이프 방법은 N-1개의 관측값을 가진 각기 다른 N-1개의 표본에서 추정값들을 계산
- 관측값이 많을땐 잭나이프 방법의 변형된 형태를 적용함. 변형된 형태는 하나의 관측값만 제외하는 대신 다수의 관측값을 제외하는 방법

<부스트랩 방법>
- 잭나이프 방법과 중요한 차이점은 원하는 표본을 원래의 전체 표본 데이터세트로부터 복원추출하는점.
- ex) 표본크기가 N인 원래 표본으로부터 표본크기가 N인 표본이 복원추출하여 구해진다.
- 대용량의 데이터일땐 같은 수의 데이터 추출 불가능하므로 변형된 형태의 부스트랩 방법 사용함.
- 부스트랩 재표본과정 : 전체 데이터에서 1%의 관측값을 추출하는데 이러한 과정을 복원추출로 100번 반복함. (25번 이상의 반복은 결과에 큰 차이가 없음) -> 하나의 부스트랩 표본 구성함.

[중요변수에 대한 십분위분석]
- 모형에서 정말 중요한 몇몇 요소가 무엇인지 알고자 할 때
- 현업에서는 모형들은 계수를 해석하는 것보다 마케팅에 도움을 줄 수 있는 요소들을 찾고 예측하는데 더 중점을 두고 있따.
- 중요한 몇몇 요소들을 탐색하는 다른 기법이 필요함.
- 중요변수들의 이익도표를 만들어서 가장 좋은 십분위에 있는 예비고객들의 행위/특성을 알 수 있음.

* 가장 좋은 모형평가 기법은 독립적인 다른 데이터세트를 사용하거나 재표본법이나 특정 변수들에 대한 십분위분석법을 통하여 모형이 실제 데이터에 대해 어떻게 수행되는가를 시험해 보는 것이다.

Ⅶ. 모형의 실행 및 유지
- 구축된 모형이 현업에 정확하게 실행되는 것이 중요함.
- 내주에서 직접 점수화하거나 외부의 전문회사로부터 점수화하는 단계와 함께 모형을 감시하는 기법
- 특정한 목적에 알맞은 다양한 종류의 모형을 선택하는 것
- 순현재가(NPV)를 계산하여 구축된 모형의 재정적인 효과 추정
- 회사의 입장에서 허용할 수 있는 재정적인 범위 내에서 모형의 효과를 평가할 수 있음.
- 십분위 분석을 통하여 마케팅 담당자나 매니저가 그들의 비즈니스 목적에 가장 알맞은 고객의 수를 선택할 수 있음.
- 마지막으로 구축된 모형의 수행정도를 지켜보고, 모형개발의 모든 과정을 문서화하는 것이 구축된 목표 모형의 효율성을 높이는 면에서 중요함.

※ 참고_인터넷

[데이터마이닝 절차]
sampling : 표본추출
exploration : 기초통계자료 획득, 다양한 데이터 시각화 도구를 이용해 데이터 이해
- 결측치 파악
- 기술통계량
- 의심되는 변수의 차트 그리기
- 상관관계 파악

modification : 변수의 변환을 통해 데이터 변환
- 범주형 변수 처리
- 상관관계 높은 변수 처리
- 범주형 변수 합치기
- PCA

modeling : 분석목적에 따라 적절한 기법을 사용하여 예측모형 만들기
assessment : 도표를 이용해 모형화 결과에 대한 신뢰성, 유용성등을 평가하는 단계(이익도표)

[범주형 예측변수 _ 범주형 반응변수]
1. 로지스틱 회귀분석
2. 신경망
3. 분류 나무
4. 나이브베이즈

[변수를 많이 사용하는게 더 나은 결과를 보장하지 않는다]
- 신뢰성 있는 모델을 만들기 위해 필요한 변수만 사용
- 많은 변수를 모델에 사용할 경우 변수간의 상관관계를 고려해야하는 번거로움이 생김.
- 두 변수간의 상관관계가 크면 하나의 변수만 포함시키는 것이 바람직함.
- 변수간의 관계를 그림으로 보여주는 산점도 행렬은 변수 선택시 유용함.
- 상관관계가 높은 변수들이나 결과변수와 관련없는 변수들을 포함시킬 경우에는 과적합현상 발생.

[데이터 마이닝 모형화 기법]
1. 지도화 모형 (내가 관리 감독하겠다. 과정과 결과에 대해 조절)
- 추정: 연속형 = 선형회귀, 신경망 (미래의 값 예측)
- 분류: 이산형 = 결정트리, 신경망 (그룹을 정해놓고 새로운 데이터가 왔을 때 어디 그룹에 속할지 예측)

2. 비지도화 모형 (자료를 던져주고 100% 맡기기)
- 군집화 = k-means, SOM
- 연관규칙 탐사 = apriori (장바구니분석)
  -> 무엇이 무엇과 잘 어울리는지 밝히기
  -> 교차판매에 적용 (상품 A에 상품 B를 얹어서 파는 것)

[차원축소방법]
- 차원 축소 전에 변수의 특성 파악부터 해야함 (기술통계)
1단계. 변수 삭제 - 상관관계에 근거하여 삭제함
2단계. 범주형 변수 내 범주 통합
3단계. PCA - 상관관계가 높은 변수들을 재조립함. 새로운 형태의 설명력을 보존한 변수집단 생성.