빅데이터 분석 프로젝트 실무

목 차 Contents

방법론
ML 프로젝트는 어떻게 진행되나?

방법론 – CRISP-DM

1. CRISP-DM이란

- Cross- Industry Standard Process for Data Mining은 분석 프로젝트를 위한 모형 중에 가장 잘 알려진 방법론
- 데이터 마이닝을 위해 만들어 진 방법론이나 예측분석,머신러닝 등 여러 분석적 프로젝트에 이용 될 수 있을 만큼 유연하고 빈틈 없다.[1)]
- Solution-independent, More focus on the business context
- 크게 6가지 단계로 이루어 지며 각 단계별로 하위 3~4과제로 구성
- 화살표는 단계간의 주요 의존관계를 표시, 외부의 원은 데이터 분석의 본질적으로 가지고 있는 순환적 특성을 의미
- 자세한 내용은 https://the-modeling-agency.com/crisp-dm.pdf 참고

1) R로 마스터하는 머신러닝 2/e, 코리 레스마이스터 저, 에이콘

방법론 – CRISP-DM

Data Mining 방법론

1. 비즈니스 이해

- 업무 목적 결정
- Data Mining 목적 결정
- 프로젝트 계획 수립

2. 데이터 이해

- 초기 데이터 수집
- 데이터 기술/탐색
- 데이터 품질 검증

3. 데이터 준비

- 데이터 선택/정제
- 데이터 생성
- 데이터 통합

4. 모델링

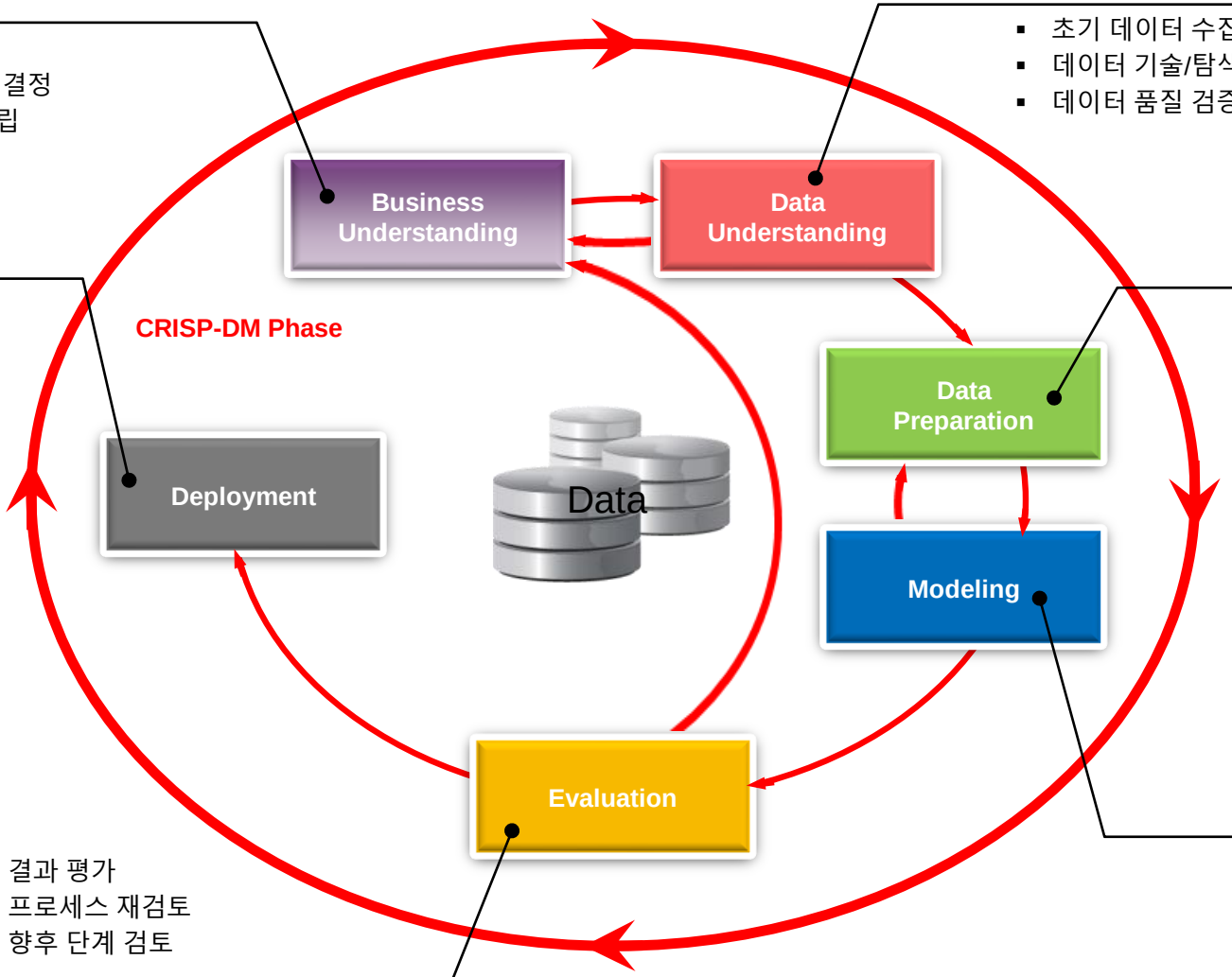
- 모델링 기법 선택
- 모델 생성
- 모델 평가

5. 평가

- 결과 평가
- 프로세스 재검토
- 향후 단계 검토

6. 전개

- 전개 계획 수립
- 최종보고 작성
- 프로젝트 재검토



방법론 – CRISP-DM

1. 비즈니스 이해
   - 분석 프로젝트의 목적을 비즈니스(업무)의 시각에서 부터 시작하고 그 문제를 머신러닝의 문제로 전환하여 구체화 하는 것이 핵심이 되며, 합리적인 성공의 기준 정의 및 그 목표를 달성 하도록 계획을 세우는 것이 핵심

2. 데이터 이해
   - 분석을 위한 데이터의 파악 및 수집 (ETL)과 데이터가 가진 의미 파악하고 데이터의 품질을 확인, 이후 기초탐색을 통해 의미있는 데이터 발견과 가설검증

3. 데이터 준비
   - 수집된 각각의 데이터를 머신러닝에 적합한 형태의 데이터로 만들며 최종 데이터 셋을 만드는 과정으로 필요한 데이터를 선택하고 여러 데이터를 조합하여 의미있는 데이터로 정제, 가공

4. 모델링

- 다양한 머신러닝 기법의 선택 및 평가를 위한 방안(데이터 및 평가지표)설정. 정해진 평가방법을 통해 최적의 알고리즘 선택 및 파라미터 최적화를 통해 최종 모델을 도출

5. 평가

- 선택된 모델이 비즈니스의 목표에 맞는지 확인, 중요한 비즈니스적 문제가 있었음에도 이를 반영하지 않은 부분이 있는 지 평가와 최종적으로 모델링 결과를 사용할 지 여부 결정.
- 이후 남은 일정과 자원을 고려하여 모델을 전개(적용)할지 반복을 통해 모델을 더 향상 시킬지, 후속 프로젝트를 할지 결정

6. 전개(배포)

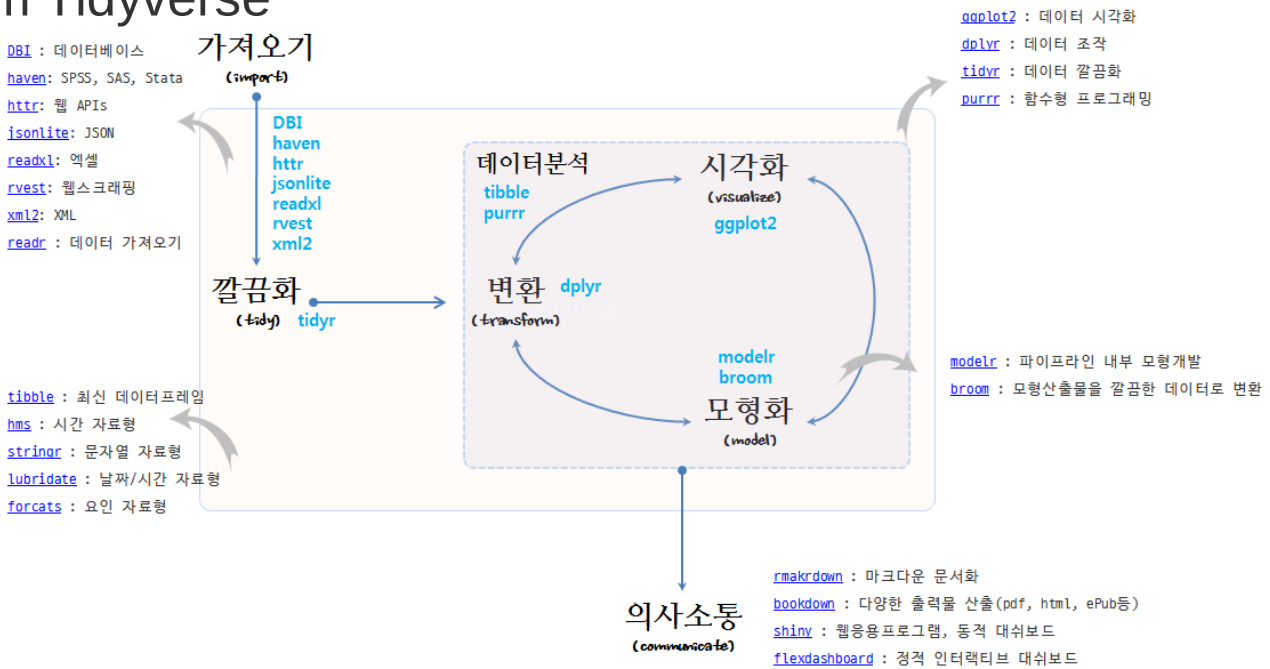
- 최종 서비스를 위한 준비, 시스템화에 필요한 정비와 모델의 모니터링 주기/ 평가지표 정의하여 유지보수 시점 및 방안 도출. 이후 보고 및 인수 인계를 위한 문서 작성

[기타] 데이터 분석 방법론 비교

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

[출처]https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview

Hadley Wickham Tidyverse



https://statkclee.github.io/data-science/ds-tidyverse.html

Case Study I
보험사 보험설계사 추천모형

1. 비즈니스 이해

어떻게 하면 상품을 잘 팔고 유지 할 수 있을까?

상품이 좋아야지!!

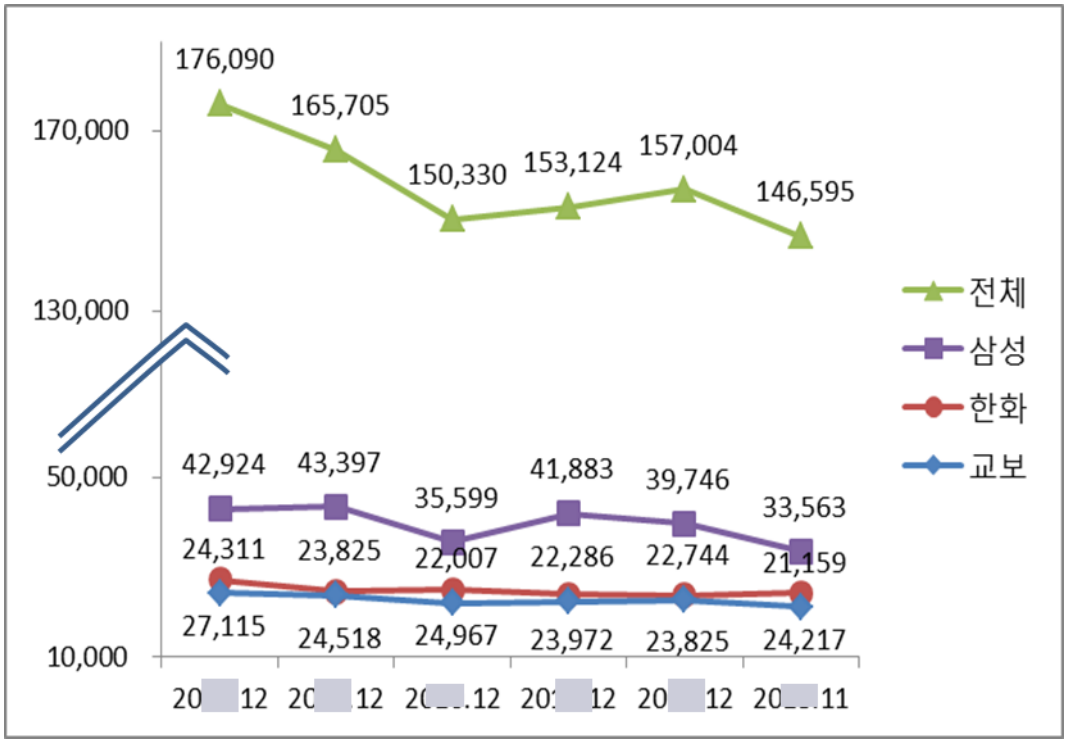
설계사가 잘해야지!!

그렇다면, 보험 설계사를 어떻게 하면 잘 고용할 수 있을까?

1. 비즈니스 이해

금융환경 변화, 개인정보보호 강화 및 설계사 리크루팅 시장의 경쟁 심화로 후보발굴에 어려움을 겪고 있으며, 이로 인해 생명보험시장 전체적으로 설계사 수가 20XX년 이후 감소하고 있는 추세임

생명보험회사 보험 설계사 수 변동 추이



- 여성의 직업 선택 폭 넓어짐
- 업계내 리크루팅 경쟁 치열
- 리크루팅 관련 규제 강화

자료: KLIA, 생명보험협회에 등록된 인원수 기준, 교차설계사 포함

1. 비즈니스 이해

리크루팅 관점에서 고객 접근

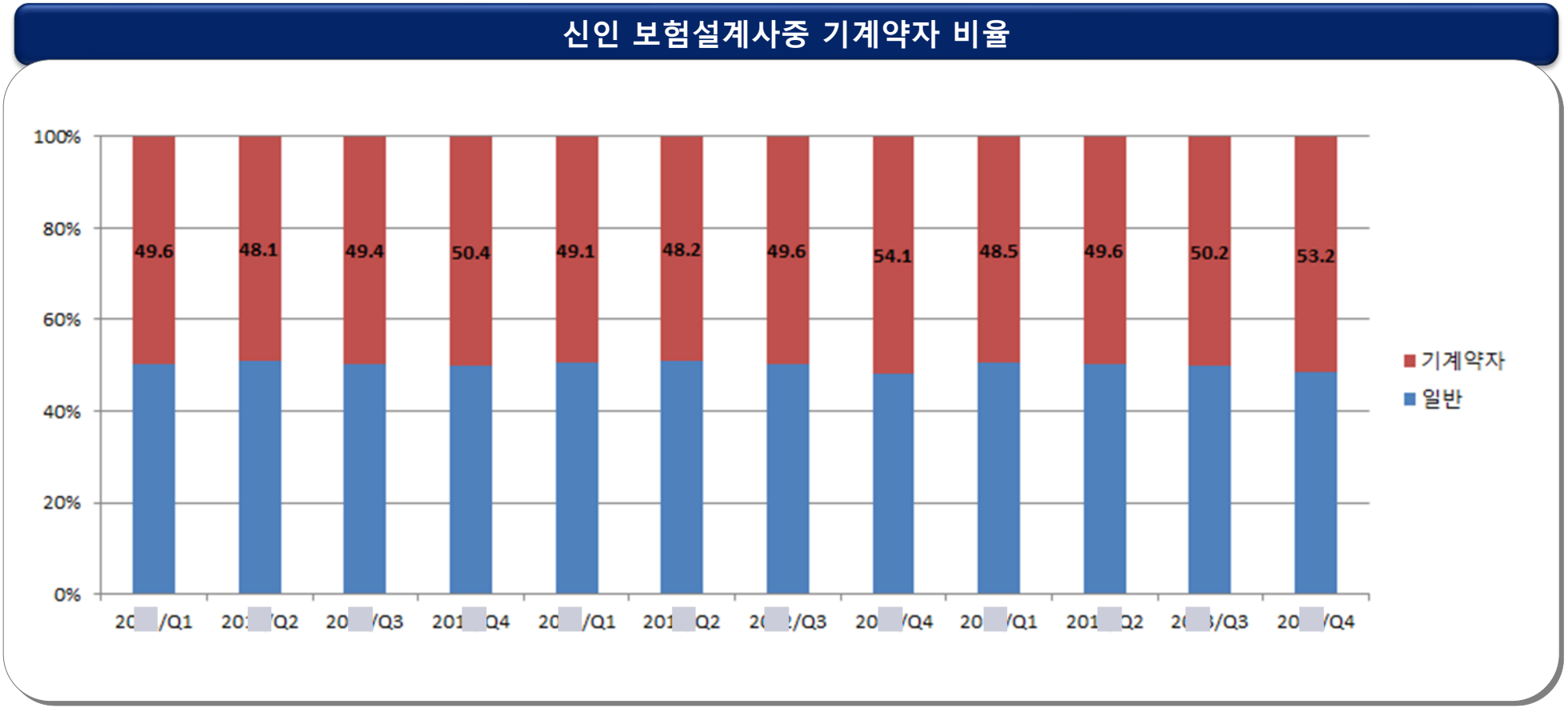
보험설계사가
지속적으로 줄고 있네
어떻게 해결해야 하나?

기존의 리크루팅 접근
방법 외에 다르게
접근할 방법은 없나?

고객들 중에서
보험 설계사를
찾아보는 것은
어떨까?

1. 비즈니스 이해

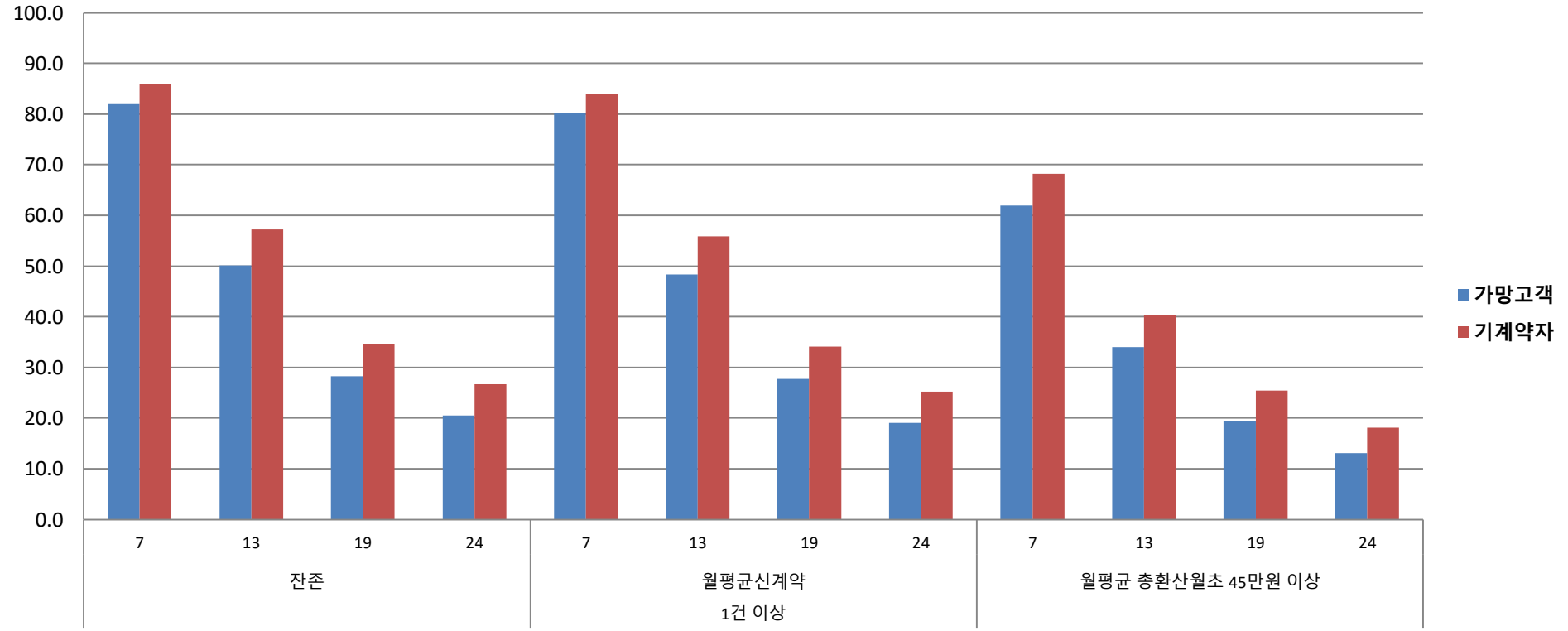
최근 3년 ㅁㅁ생명에서 전환된 신인 보험설계사의 약 47.5%가 기계약자에서 전환 되었음



1. 비즈니스 이해

고객에서 전환된 설계사들의 성과와 잔존기간이 조금 더 좋았다

신인 보험설계사중 기계약자 비율



2. 계획 (time table) 및 R&R

전체 소요 일정 : 12 주

구분		M1				M2				M3			
단계	업무	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
	Milestone		▲착수보고					▲중간보고				▲종료보고	
착수 및 분석	분석환경 세팅 /업무 요건 및 요구사항 정의/ 데이터 준비												
탐색 및 설계	분석 마트 설계 및 데이터 분석												
개발	모형 개발 및 평가/ 프로세스 설계												
구현	모형 배치 개발 및 운영												
테스트	단위 테스트 및 통합 테스트												
안정화	사용자 및 운영자 교육, 인수 인계												

주요 인력

- 데이터 분석 leader 1 명, Junior 1명 * 3M
- CRM 담당자 1 명 * 1.5 M
- IT 담당(ETL 및 화면설계) * 3 M

3. 데이터 분석 프로세스

설계사 리크루팅 예측모형의 개발(CRISP-DM 방법론에 기반한 데이터 생성, 모형개발, 적용)

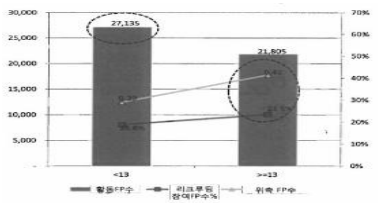
분석 데이터 및 대상정의

- 계약자, 상품, 계약유지, 지급/약대, 접촉 데이터
- Target: 기계약자 -> 설계사전환(O)
  Control: 기계약자 -> 설계사환(X)

20XX ~ 20XX
고객 데이터

데이터분석

- 변수탐색
- 새로운 변수 생성



예측 모형 개발

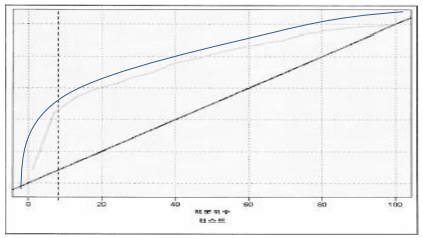
모형1-1
설계사전환가능성

모형2
우수설계사가능성

모형1-2
평점표

예측 모형 평가

- 모형 평가



모형 적용(Pilot)

- 결과 적용

등급	FP 후보	Target	Target의 누적건수	Target 포함비율	누적 Target비율	lift	누적 lift
10%	1,144	290	290	③ 25.35%	② 36.43%	③ 3.64	3.64
20%	1,144	124	414	10.84%	52.01%	1.56	2.60
30%	1,144	78	492	6.82%	61.81%	0.98	2.06
40%	1,144	60	552	5.24%	69.35%	0.75	1.73
50%							
60%							
70%							
80%	1,144	37	713	3.23%	89.57%	0.46	1.12
90%	1,144	45	758	3.93%	95.23%	0.57	1.06
100%	1,145	38	796	3.32%	100.00%	0.48	1.00
전체	11,441	796		6.96%		1.00	

Sample

Case Study I (보험사 보험설계사 추천모형)

주요 Issue

1. 모형을 단일 모형으로 만들것인가 or 세분화 하여 모델을 각기 만들것인가 ?

2. Target은 어떻게 정할 것인가 ?

3. 몇 년의 데이터를 사용할 것인가 ?

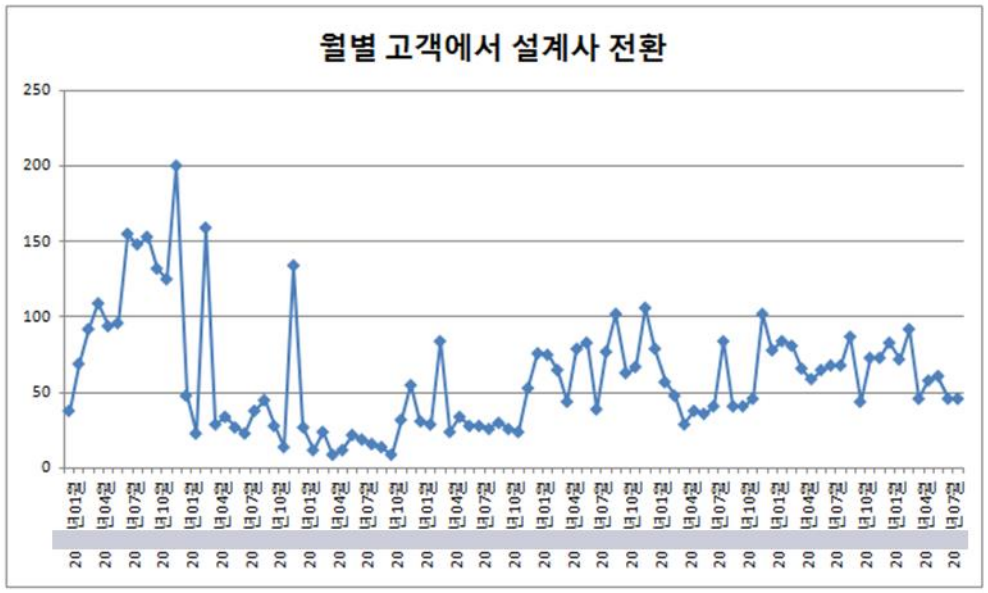
4. 어떤 알고리즘을 사용할 것인가?

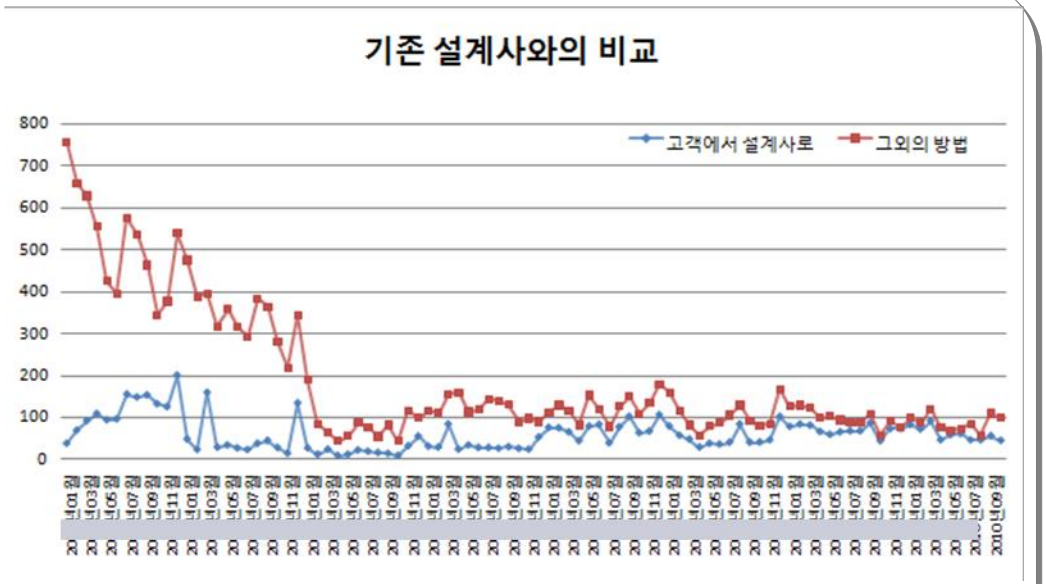
5. 얼마나 자주 모델을 update 할 것인가 ?

3. 데이터의 이해

분석 기간을 얼마로 정할 것인가 ?

설계사 리크루팅 추이





월별 계약자 또는 비계약자에서 어드바이저로 위촉 건수

Summary

- 20XX년 이후 계약자에서 어드바이저로 위촉한 사람은 점차 늘고 있는 추세를 보임.
- 비계약자에서 어드바이저로 리크루팅 되는 추이는 2005년에 급감하여 계약자에서 위촉되는 추이와 유사한 형태를 보임.

3 데이터의 이해

Target 을 어떻게 정의할 것인가

Target 정의

Target	•계약자(기존보유고객)에서 설계사로 전환한 고객 •계약자 주민번호와 Advisor의 주민번호와 일치 **제외조건 1: 위촉일자 < 성립일자** **제외조건 2: 재위촉 설계사 제외(재입사 Advisor)**
Control	•계약자에서 설계사로 전환하지 않은 고객
데이터 정보 **(Target vs. Control**	•계약 기간 : **성립일자 20XX년 1월 1일 ~ 20XX년 8월 31** •계약 상태 : **정상유지, 당월실효, 기실효 (Target 위촉시점 계약상태 사용)** •설계사 위촉기간 : **20XX년 1월 1일 ~ 20XX년 8월 31일**

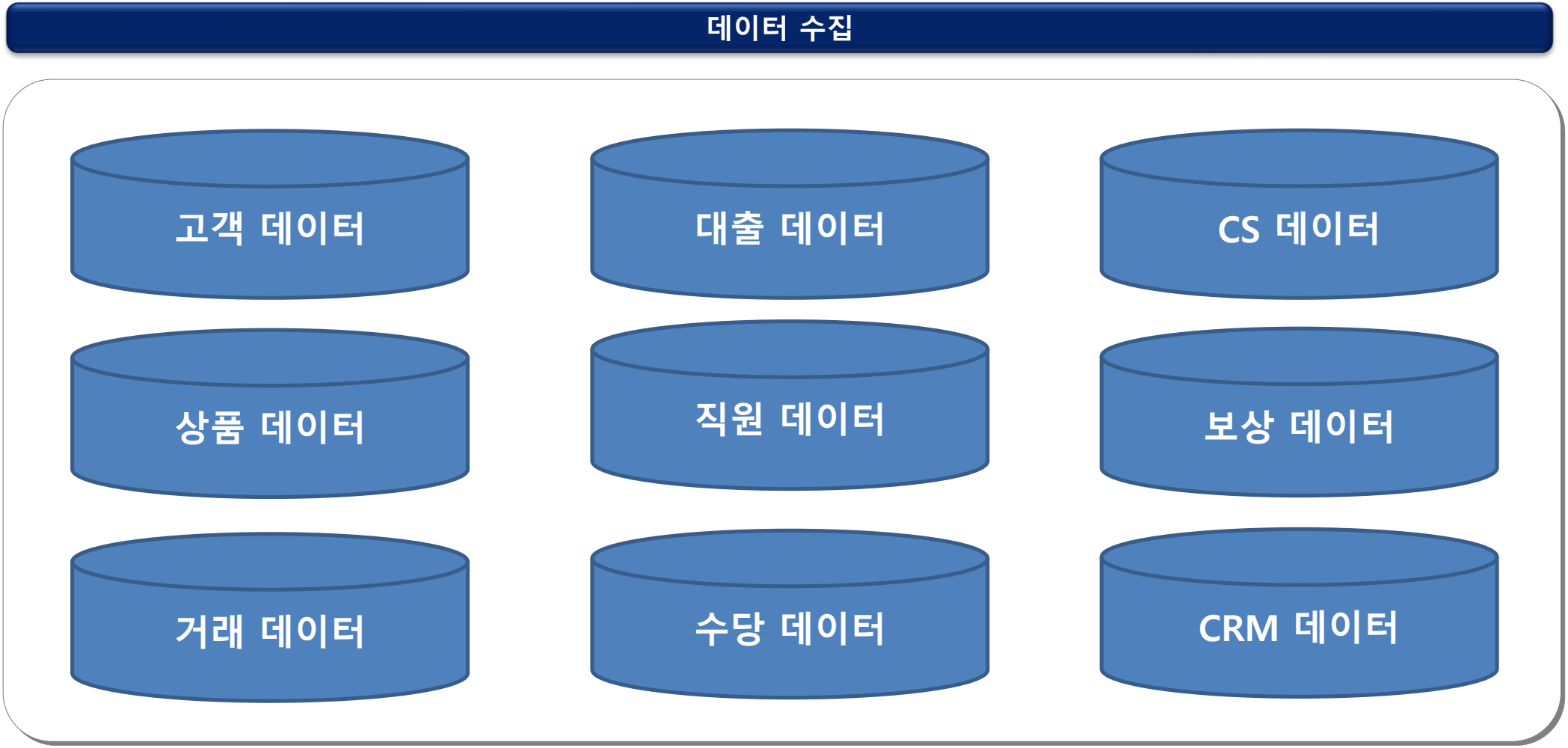
	Target	Control	보유고객
Count	470,000	530,000	1,000,000
비율	**47%**	53%	-

Summary

- ❖ Target과 Control의 분포는 Target이 약47%를 차지하고 있으며 Control은 53%를 차지함.
- ❖ 20XX년 8월 마감 보유 고객 수는 1,000,000명임.

3 데이터의 이해

어떤 데이터를 사용할 수 있나 ?



3 데이터의 이해

번호	고객 데이터
1	계약자 현재 나이
2	계약자 성별
3	계약자 직업
4	계약자 거주지
5	소득
6	최초 가입일

번호	상품 데이터
1	상품 명
2	상품 카테고리
3	계약 금액
4	특약 종류
5	상품 판매 시작 일시
6	상품 판매 종료 일시

번호	CRM 데이터
1	고객 명
2	담당자
3	담당 지점
4	접속 유형
5	제공 내역

번호	거래 데이터
1	계약자 명
2	피보험자 명
3	관계
4	보험금액
5	회차
6	현재 상태
7	담당자
8	담당 지점
9	계약 일시
10	종료 일시
11	일시납 여부

번호	대출 데이터
1	대출자 명
2	담보 상품
3	대출 금액
4	이자율
5	대출 시작 일자
6	현재상태

번호	직원 데이터
1	직원 ID
2	취업 일시
3	취업 경로
4	계약 건수
5	계약 금액 합계

번호	CS 데이터
1	CS 일시
2	고객명
3	CS 카테고리
4	문의 내역
5	처리 내역

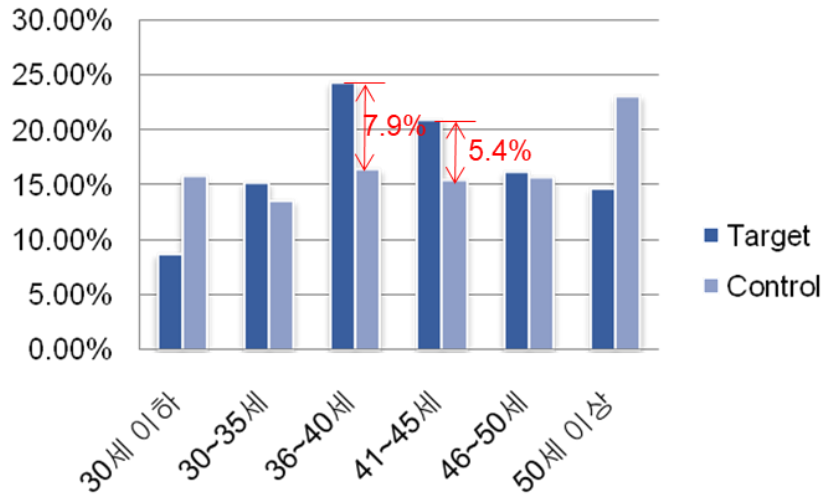
번호	보상 데이터
1	보험 계약 ID
2	보상 접수 일시
3	접수 채널
4	보상 유형
5	보상 금액
6	담당자

3 데이터의 이해

번호	변수명	정의
1	계약자 현재 연령	계약자의 현재 연령
2	age	CRM 구분나이,
3	계약자 성별	계약자의 성별 구분
4	계약자 직업	계약자의 직업종류 중분류 구분
5	추정소득	계약자의 월수입 추정소득(CRM 고객세분화 소득추정사용)
6	계약자 거주지	계약자의 거주지역 도시구분(서울,광역시,...)
7	모집지점과거주지 동일여부	계약자의 최근 계약의 계약자 우편번호와 모집지점 우편번호 동일여부
8	수금지점과거주지 동일여부	계약자의 최근 계약의 계약자 우편번호와 수금지점 우편번호 동일여부
9	모집수금동일여부	계약자가 가입한 보유 계약중 계약의 모집/수금 모든 계약에 대한 설계사 동일여부
10	상품 카테고리	계약자가 최근에 가입한 상품이 속하는 카테고리
11	월환산합계보험료	계약자가 가입한 계약 중에 계약상태가 정상유지, 당월실효, 기실효에 대한 합계보험료의 합계(SUM).
12	총계약수	계약자가 보유한 모든 계약의 수
13	정상보유계약수	계약자가 보유한 계약상태가 정상유지, 당월실효, 기실효인 계약의 수
14	해지계약수	계약자가 보유한 계약중 계약상태가 정상유지, 당월실효, 기실효가 아닌 계약의 수
15	약관대출이용여부	계약자가 보유한 보험을 통해 약관대출을 이용했는지 여부
16	약관대출이용횟수	계약자가 보유한 보험을 통해 약관대출을 이용한 횟수
17	약관대출합계급액	계약자가 보유한 보험을 통해 약관대출을 받은 금액의 합계
18	보험금 지급여부	계약자가 보유한 보험의 보험금을 지급받았는지 여부
19	보험금 지급횟수	계약자가 보유한 보험의 보험금을 지급받은 횟수
20	보험금 지급금액	계약자가 보유한 보험을 통해 지급받은 보험금의 합계
21	자녀수	계약자의 자녀의 수(캠페인 세대정보 활용)
22	활동여부	계약자의 담당설계사가 활동 발행 여부(가입설계, 입시고객 제외)
23	활동건수	계약자의 담당설계사가 활동 발행 건수(가입설계, 입시고객 제외)
24	가입설계여부	계약자의 담당설계사가 계약 가입설계 발행 여부
25	가입설계건수	계약자의 담당설계사가 계약 가입설계 발행 건수
26	활동(금년운세)여부	계약자의 담당설계사가 금년운세 발행 여부
27	활동(금년운세)건수	계약자의 담당설계사가 금년운세 발행 건수
28	계피동일여부변수	계약자가 최근에 가입한 보유 계약중 계약자/피보험자 모든 계약에 대한 동일여부
29	일시납여부	계약자의 계약중 일시납 계약 보유여부

3 데이터의 이해

데이터 탐색 (계약자 연령대)



Summary

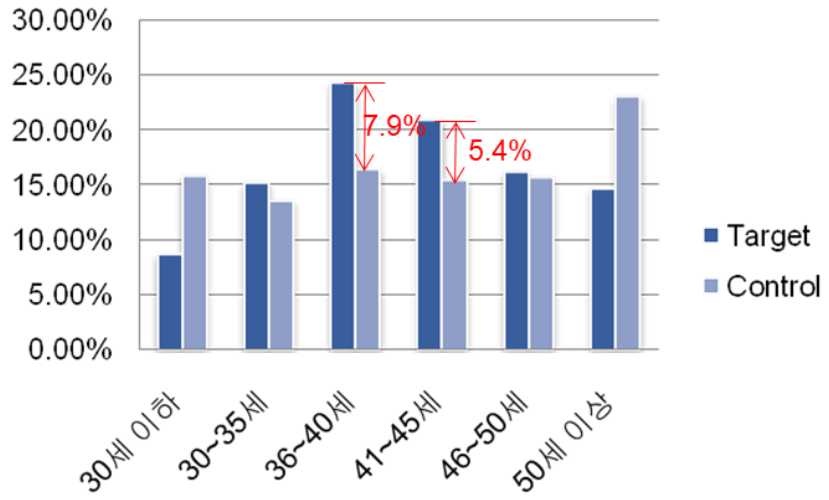
- ❖ 기존 고객에서 설계사로 전환한 고객의 연령대는 36~45세가 45.2%로 가장 높은 비율을 보임.
- ❖ 36~40세의 Target의 비율은 24.3%로 Control 보다 7.9% 높으며, 이는 Control 대비 1.48배 높음.
- ❖ 41~45세의 Target의 비율은 20.9%로 Control에 비해 5.4%. 높으며, 이는 Control 대비 1.35배 높음.

계약자 연령대	Target	Control	보유고객
30세 이하	481	97,018	123,434
30~35세	844	83,441	131,848
36~40세	1,349	100,768	179,516
41~45세	1,158	94,709	182,011
46~50세	901	96,400	202,795
50세 이상	816	141,862	417,752
Total	5,549	614,198	1,237,356

Target	Control	T-C	T/C	보유고객
8.7%	15.8%	-7.1%	0.55	10.0%
15.2%	13.6%	1.6%	1.12	10.7%
24.3%	16.4%	7.9%	1.48	14.5%
20.9%	15.4%	5.4%	1.35	14.7%
16.2%	15.7%	0.5%	1.03	16.4%
14.7%	23.1%	-8.4%	0.64	33.8%
100.0%	100.0%	0.0%	1.00	100.0%

3 데이터의 이해

데이터 탐색 (계약자 연령대)



Summary

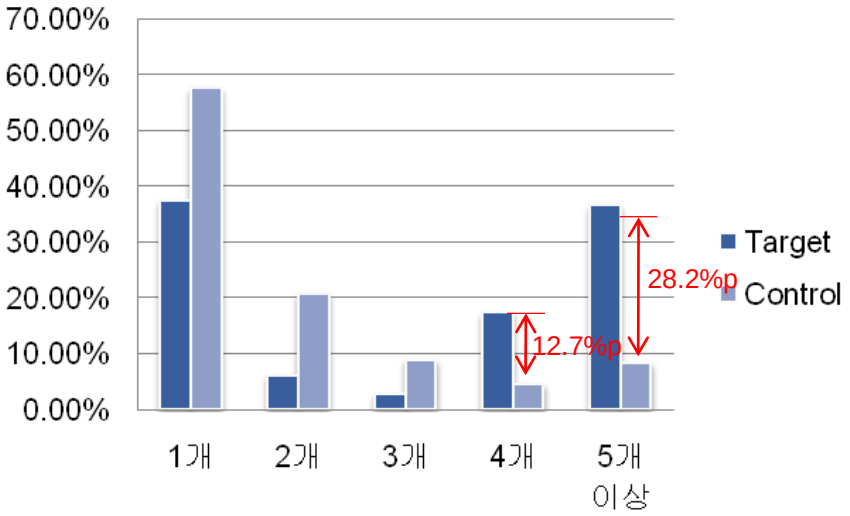
- ❖ 기존 고객에서 설계사로 전환한 고객의 연령대는 36~45세가 45.2%로 가장 높은 비율을 보임.
- ❖ 36~40세의 Target의 비율은 24.3%로 Control 보다 7.9% 높으며, 이는 Control 대비 1.48배 높음.
- ❖ 41~45세의 Target의 비율은 20.9%로 Control에 비해 5.4%. 높으며, 이는 Control 대비 1.35배 높음.

계약자 연령대	Target	Control	보유고객
30세 이하	481	97,018	123,434
30~35세	844	83,441	131,848
36~40세	1,349	100,768	179,516
41~45세	1,158	94,709	182,011
46~50세	901	96,400	202,795
50세 이상	816	141,862	417,752
Total	5,549	614,198	1,237,356

Target	Control	T-C	T/C	보유고객
8.7%	15.8%	-7.1%	0.55	10.0%
15.2%	13.6%	1.6%	1.12	10.7%
24.3%	16.4%	7.9%	1.48	14.5%
20.9%	15.4%	5.4%	1.35	14.7%
16.2%	15.7%	0.5%	1.03	16.4%
14.7%	23.1%	-8.4%	0.64	33.8%
100.0%	100.0%	0.0%	1.00	100.0%

3 데이터의 이해

데이터 탐색 (계약자 총 계약건수)



Summary

- ❖ Target의 총 계약건수는 1개가 37.4%로 가장 높음.
- ❖ 계약자가 보유한 총 계약건수가 5개 이상인 Target의 비율은 36.6%로 Control에 보다 28.2%p 높으며, 이는 Control 대비4.39배 높음.
- ❖ 총 계약건수가 4개 이상인 Target이 Control에 보다 높은 비율을 보임.

총 계약 수	Target	Control	보유고객
1개	2,073	353,458	595,984
2개	340	127,045	259,192
3개	149	54,389	127,051
4개	958	28,101	73,841
5개 이상	2,029	51,205	181,288
Total	5,549	614,198	1,237,356

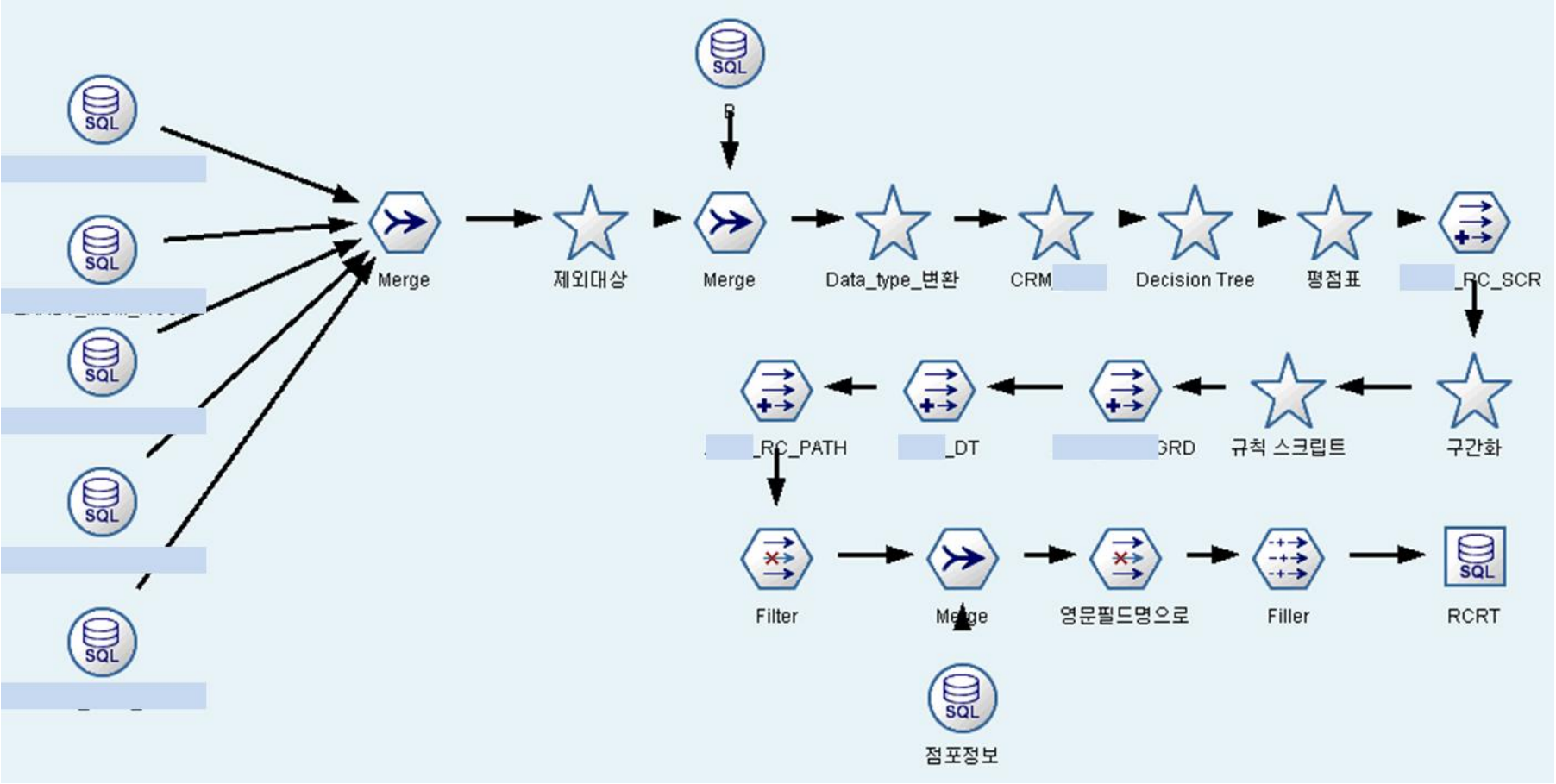
Target	Control	T-C	T/C	보유고객
37.4%	57.5%	-20.2%	0.65	48.2%
6.1%	20.7%	-14.6%	0.30	20.9%
2.7%	8.9%	-6.2%	0.30	10.3%
17.3%	4.6%	12.7%	3.77	6.0%
36.6%	8.3%	28.2%	4.39	14.7%
100.0%	100.0%	0.0%	1.00	100.0%

4 모형 개발

데이터 품질검토 통계적 유의성 검정, 유사의미 변수 제거 등의 분석을 통해 선별했고, 최종적으로 27개의 변수가 사용됨

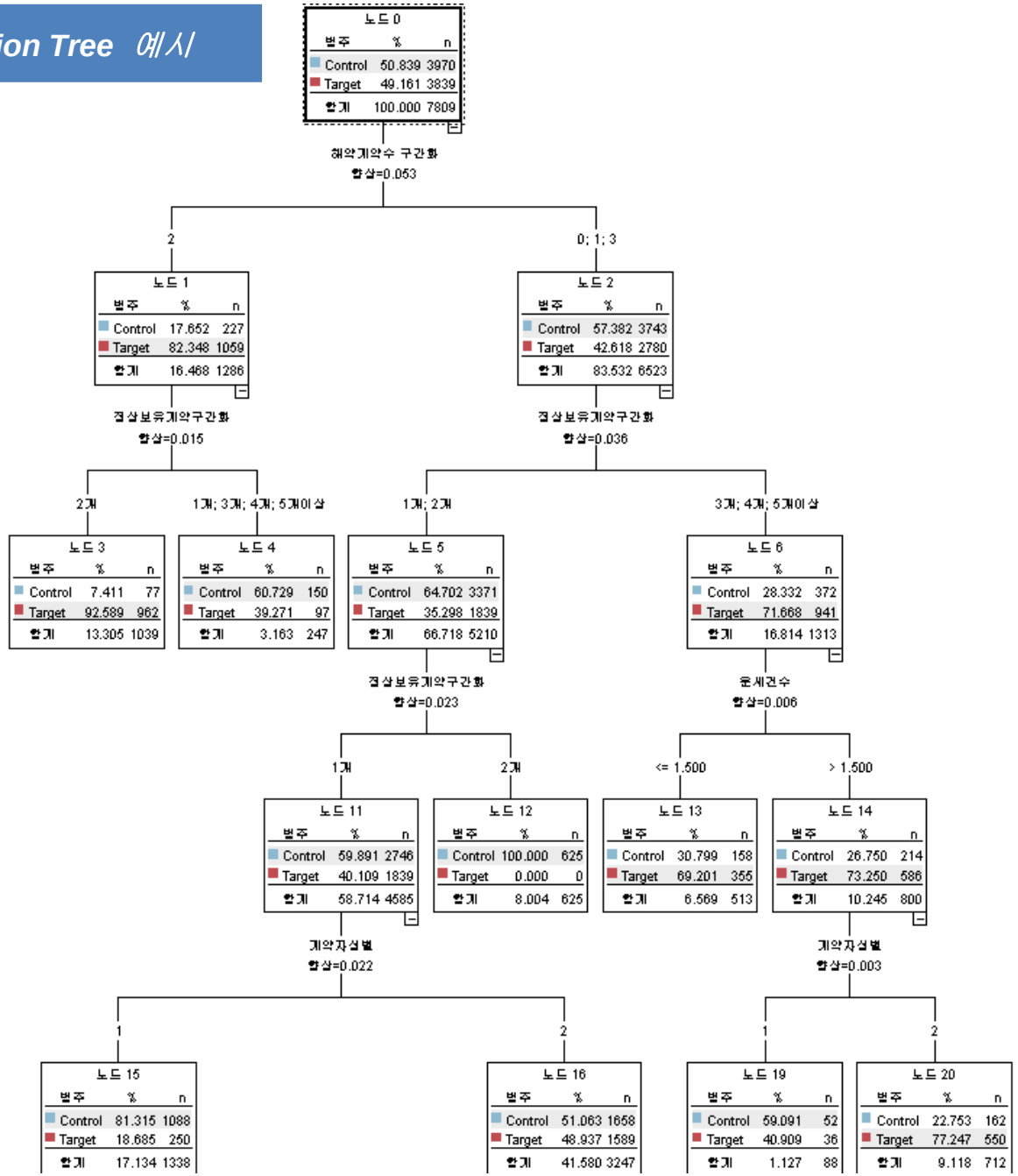
4 모형 개발

데이터 탐색 (계약자 연령대)



Case Study I (보험사 보험설계사 추천모형)

Decision Tree 예시



Summary

- ❖ if 해약 계약수 = 2
  and 정상 보유 계약수 =2
  then Target (설계사로 전환)
- ❖ If 해약 계약수 >= 3
  and 정상보유 계약수 >= 2
  and 계약자 성별 = 여성
  and CTGR= VOL, 건강보험,
  사고보험, 양로보험,
  어린이보험
  then Target (설계사로 전환)
- ❖ If 해약 계약수 >= 3
  and 정상유지 계약수 >= 3
  and 운세 건수 <= 1.5
  then Target (설계사로 전환)
- ❖ If 해약 계약수 >= 3
  and 정상유지 계약수 >= 3
  and 운세 건수 >= 1.5
  and 계약자 성별 = 여성
  then Target (설계사로 전환)

4 모형 개발 결과(1)

설계사 활동가능성 예측을 위한 데이터분석 모형은
의사결정나무와 로지스틱 회귀 모형을 채택함

모형 구분	정확도	민감도	상위 10% Target 비율	상위 10% 구간 Lift
CART	64.9%	43.1%	36.5%	3.64
CHAID	80.1%	50.0%	49.6%	4.95
C5.0	87.0%	53.5%	35.5%	5.18
로지스틱 회귀 모형	84.8%	49.0%	53.1%	5.28

로지스틱 회귀모형의 11개 예측변수 및 중요도

머신 러닝을 하고 싶다
but
데이터가 없다 ㅠㅠ

가장 유사한 데이터 – Kaggle !!