

2019-2 DS LAB

Credit Scoring in IT company

목차

1. 신용 평가란?

- 1.1 신용 평가란
- 1.2 대안 신용 평가란

2. 신용 평가 모델링

- 2.1 모델 개발 프로세스
- 2.2 신용 평가 모델링 특징
- 2.3 IT에서의 신용 평가 모델링 / 데이터 분석

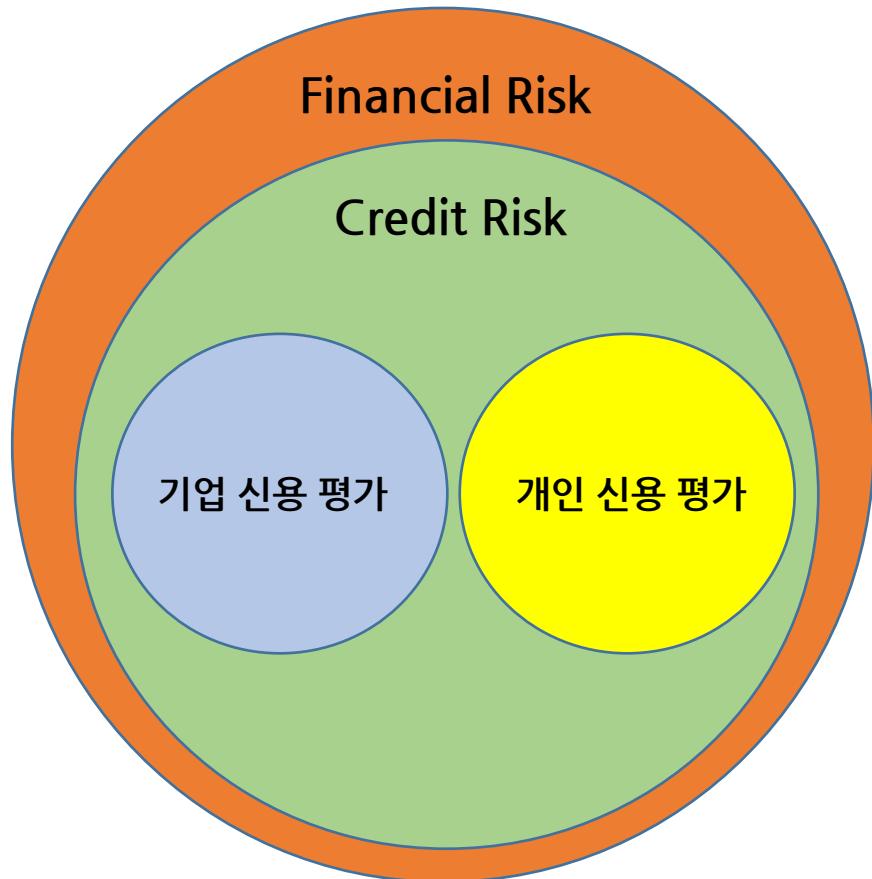
3. Get Ready

- 3.1 To be Data Scientist

1. 신용 평가란?

1.1 신용 평가란?

Financial Risk & Credit Risk



1.1 신용 평가란?

Credit Risk / Credit Score / Credit Scoring

Credit Risk 신용위험

차입자가 약정된 조건에 따라 채무를 이행
할 수 없게 될 가능성
금융회사의 가장 중요한 리스크 중 하나

Credit Scoring 신용평가

채무자가 빌려간 돈을 제때 잘 상환할
수 있는지 예측하여 판단하는 행위

Credit Score 신용 점수

신용평가를 통해 판단된 신용위험을
바탕으로 개인에게 부여되는 숫자로,
숫자가 높을수록 그 사람의 신용도가
높다는 것을 의미

1.1 신용 평가란?

Credit Score Model

정의

고객의 신용도(default 여부)를 예측하기 위해서 해당 고객의 과거 거래정보(설명변수)들을 활용하는 모델이다.

고객의 신용도를 예측하기 위한 평가 Tool으로, 일반적으로 금융기관 등에서 여신(대출)에서 활용된다.

필요성

금융기관 입장에서 대출 신청 고객에 대해 해당 고객에게 대출을 해도 되는지 의사결정이 필요하다.

이 때 의사결정의 판단기준은 ‘빌려준 돈을 잘 갚을 수 있는 고객인지’이다.

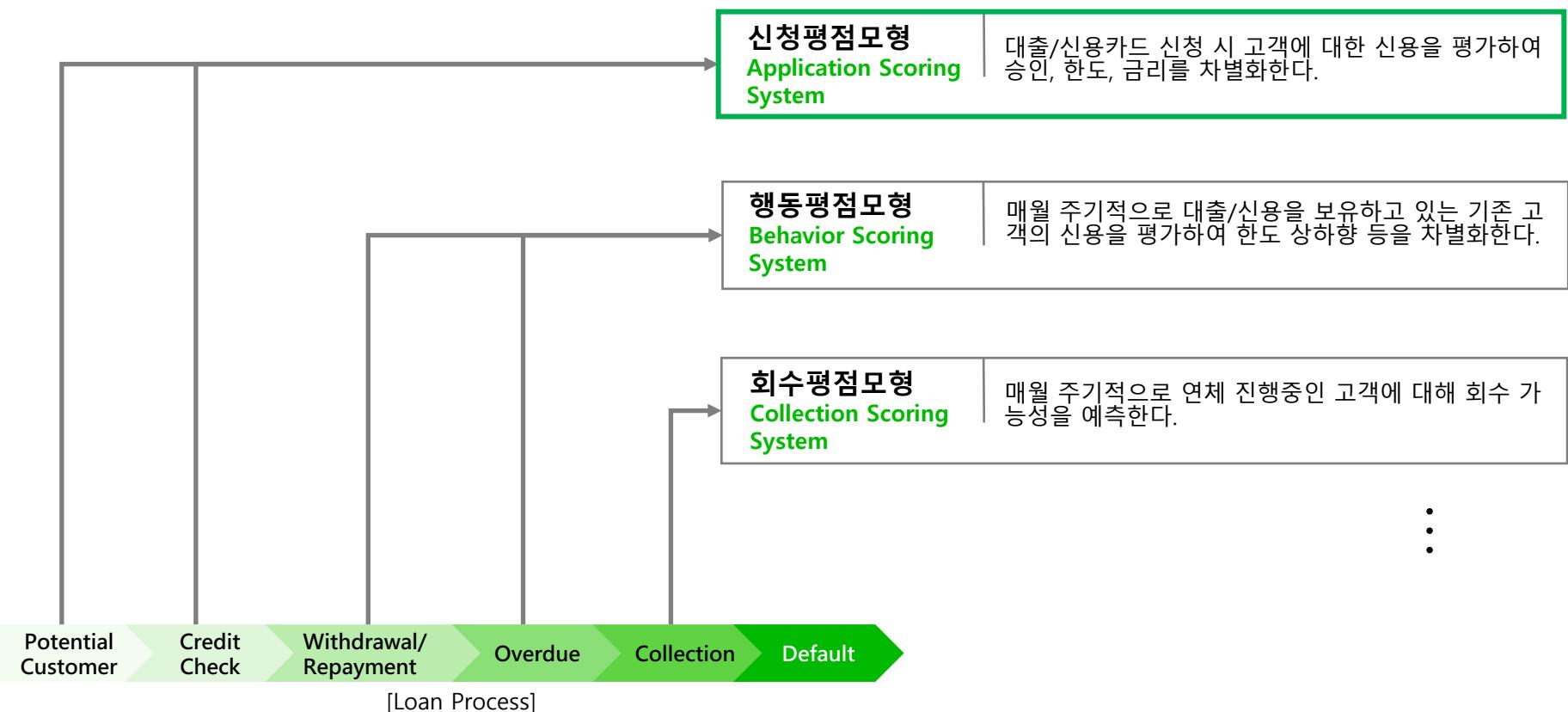
이러한 의사결정을 지원해주기 위해 고객의 신용도(default 여부)를 예측/산출하는 tool이 신용평가 모델이다.

예시

- 과거 1년내 연체 경험 30일 이상 연체 경험이 2번 이상 있는 고객은 신용도가 열위하다?
- 과거 6개월 동안 한도 대비 카드사용실적이 90% 이상인 고객은 신용도가 열위하다?
- 최근 3개월 이내 은행업권 이외에서 대출을 받은 경험이 있는 고객은 신용도가 열위하다?
- 과거 5년내 금융사에서의 상각, 파산면책, 개인회생, 신용회복 등의 경험이 있는 고객은 신용도가 열위하다?
- 최근 1년 내 현금서비스 이용 경험이 있는 고객은 신용도가 열위하다?

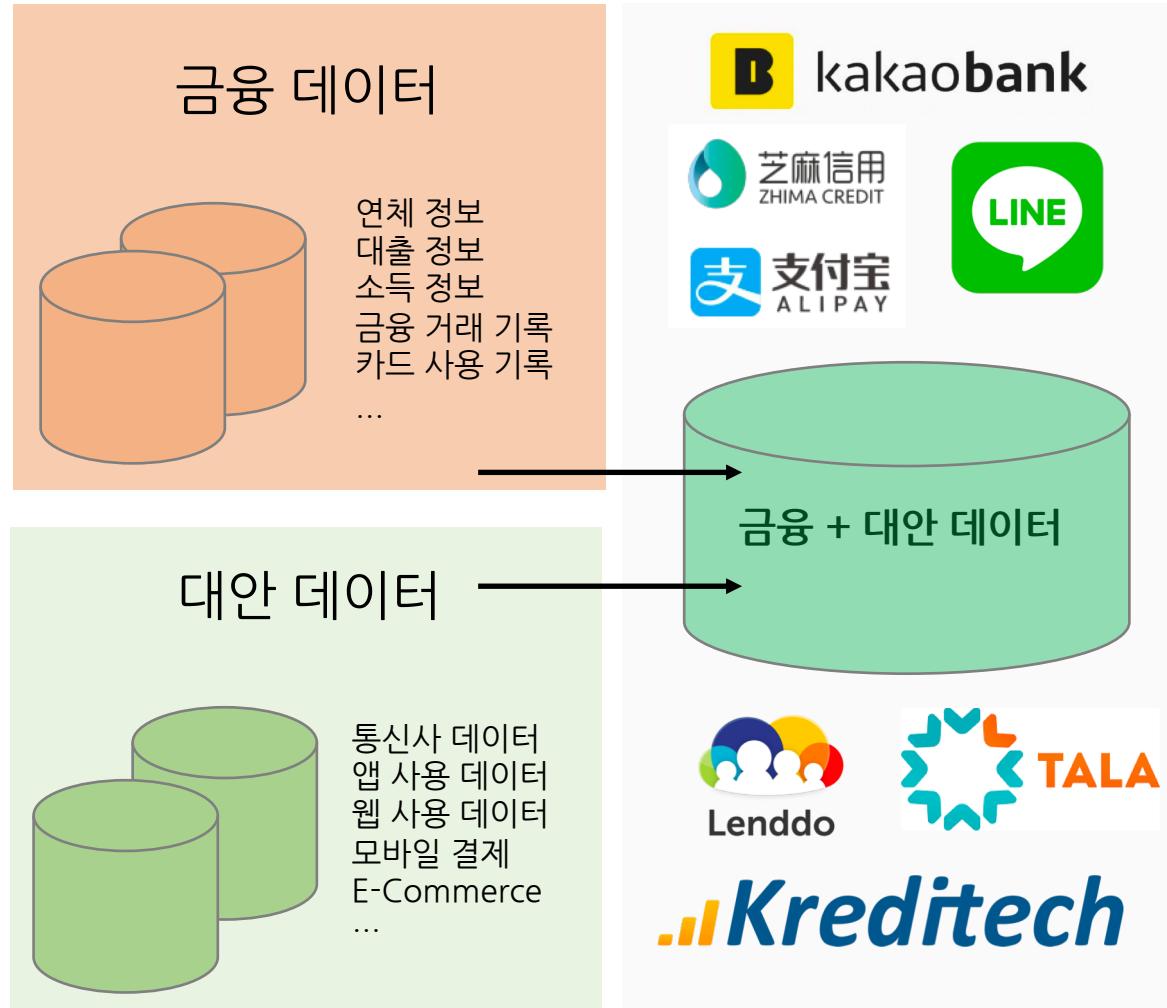
1.1 신용 평가란?

Credit Scoring 종류



1.2 대안 신용 평가란?

Alternative Credit Scoring



기존 신용 평가 고도화

Unbanked

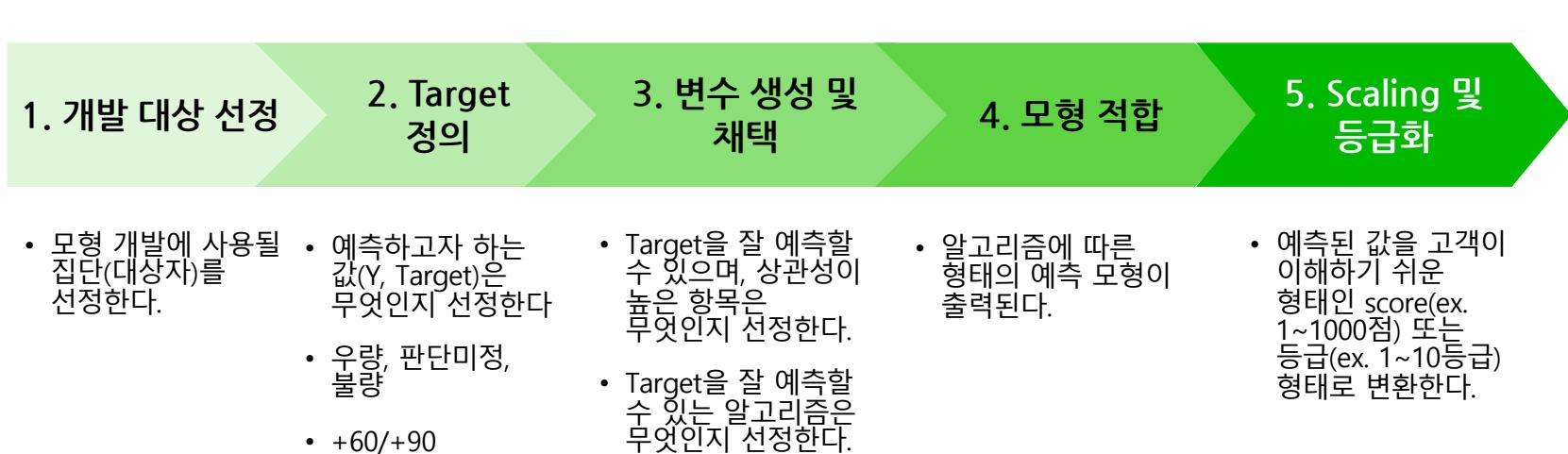
No Financial history

2. 신용 평가 모델링

2.1 모델 개발 Process

모델 개발 Process - Overview

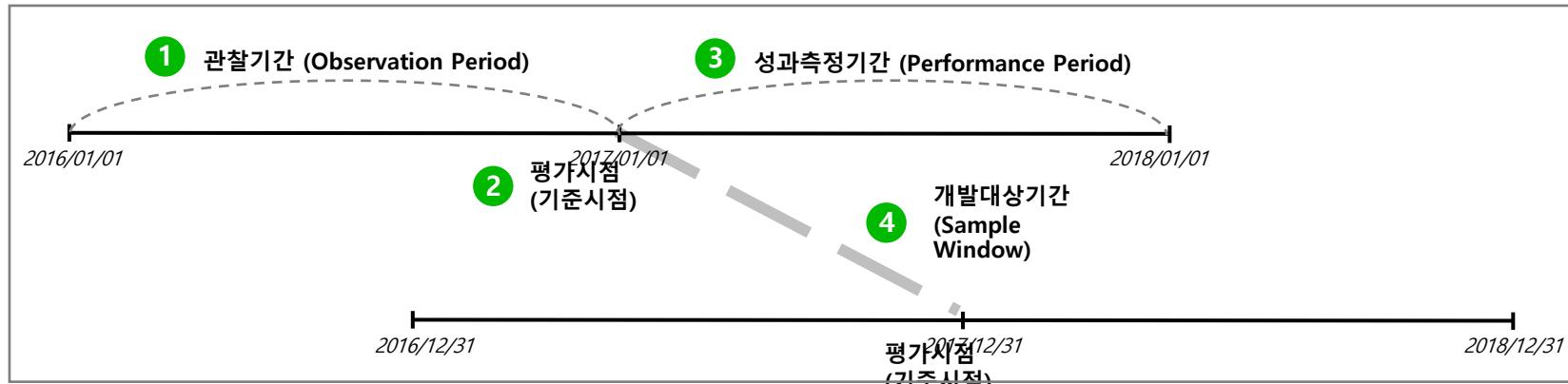
Credit Scoring Model 개발 프로세스



2.1 모델 개발 Process

개발 Process - Framework

Credit Scoring Model 개발 요건을 정의하기 위한 기본 Framework



1. 관찰 기간 (Observation Period):

- 평가시점에서부터 과거로의 기간으로, 모델의 변수로 사용되는 데이터의 기간이다.

2. 평가 시점:

- 고객신용을 평가하기 위한 기준시점으로, 대출신청평점모형에서는 신청일을 말한다.

3. 성과측정 기간 (Performance Period):

- 평가시점에서부터 미래로의 기간으로, 모델의 Target을 예측하기 위한 기간이다
- 예: 2018/01/01일에 대출을 신청한 대상자는 2018/12/31까지의 Target 발생 유무를 확인

4. 개발 대상 기간 (Sample Window): Credit Scoring Model을 개발하기 위해 정의한 복수개의 평가 시점을 말한다.

- 예: 2017년 1월 1일부터 2017년 12월 31일 까지 1년간 대출을 신청한 고객의 데이터를 바탕으로 모델을 개발

2.1 모델 개발 Process

개발 Process – 변수 생성

가설 1. 친구는 나를 비추는 거울, Strong-relationship의 사람들과 성향이 비슷할 것이다.

1. 가설 제시 이유/배경

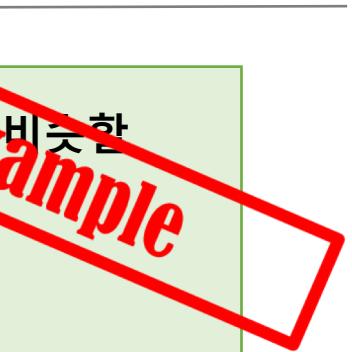
- 나와 가장 가까운/교류가 많은 사람들의 성향이 나의 성향과 비슷할 것이라고 추론
- 상환능력 혹은 상환 의지와 관련된 여러가지 지수(오락/결제/금융서비스/쿠폰)를 반영하여 다양한 관점에서 성향을 파악
- 새로운 변수로 뿐 아니라 결측치에 대한 대체 변수로 사용 가능

2. 가설 핵심 포인트

- ① Strong relationship 안에 있는 몇 명의 사람을 반영 대상으로 삼을 것인가?
 - 기본적으로 Strong relationship 안에 있는 대상 중 최근 인터랙션 비중이 높은 멤버
 - Strong relationship 안의 사람이 너무 적을 경우: Strong relationship의 기준을 0.4로 높이기
 - Strong relationship 안의 사람이 너무 많을 경우: Strong relationship의 기준을 0.2로 낮추기
- ② Strong relationship 안에 있는 사람들이 서로 너무 다를 경우는 어떻게 대비할 것인가?
 - 아웃 라이어 제거
 - 박스 플롯을 그려 상위 25-75% 사람들만 반영
 - 평균값 혹은 중위값 이용

3. 예상 시나리오

- ① Strong relationship의 사람들이 오락/결제를 많이 할수록 해당 유저의 부도 확률 증가 (+)
- ② Strong relationship의 사람들이 금융서비스/쿠폰을 많이 이용할수록 해당 유저의 부도 확률 감소 (-)



2.1 모델 개발 Process

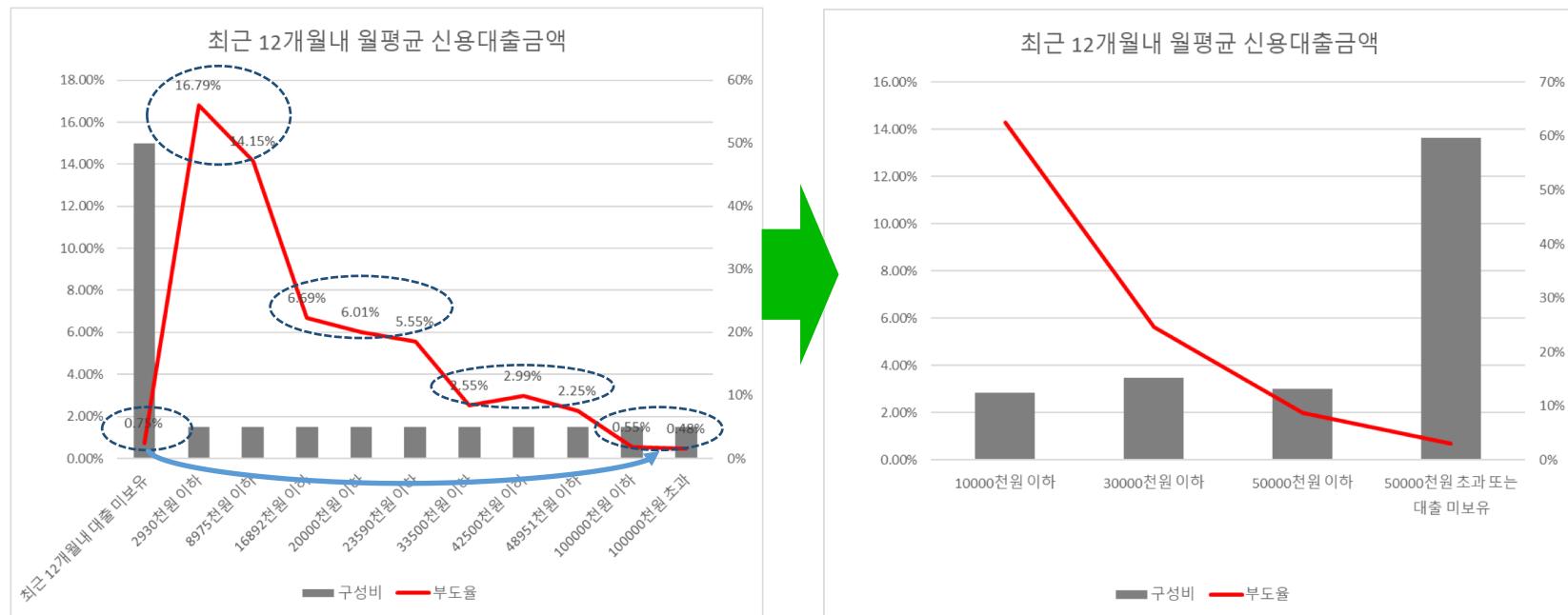
개발 Process – 변수 생성

Feature	Desc.	Remarks
strong_relation_entertain_avr_time	최근 xxx일 Strong relationship 그룹의 엔터테인먼트 일 평균 사용 시간	엔터테인먼트 게임 + 만화 + 라이브
strong_relation_entertain_cul_time	최근 xxx일 Strong relationship 그룹의 엔터테인먼트 누적 사용 시간	
strong_relation_entertain_avr_count	최근 xxx일 Strong relationship 그룹의 엔터테인먼트 일 평균 접속 횟수	
strong_relation_entertain_cul_count	최근 xxx일 Strong relationship 그룹의 엔터테인먼트 누적 사용 횟수	
strong_relation_billing_sum	최근 xxx일 Strong relationship 그룹의 누적 결제 금액	결제: 포춘 + 만화 + 게임 + 라이브 + 뮤직 + 스티커 결제
strong_relation_billing_avr	최근 xxx일 Strong relationship 그룹의 주 평균 결제 금액	
strong_relation_billing_sum_count	최근 xxx일 Strong relationship 그룹의 누적 결제 횟수	
strong_relation_billing_avr_count	최근 xxx일 Strong relationship 그룹의 주 평균 결제 횟수	

2.1 모델 개발 Process

개발 Process – 변수 Binning

‘최근 12개월내 월평균 신용대출금액’ 을 5 Percentile 씩 구분하여 Default와의 관계를 탐색, 구성비와 부도율을 고려하여 구간화(Binning)을 수행하고 모형 개발(Fitting)에 사용함

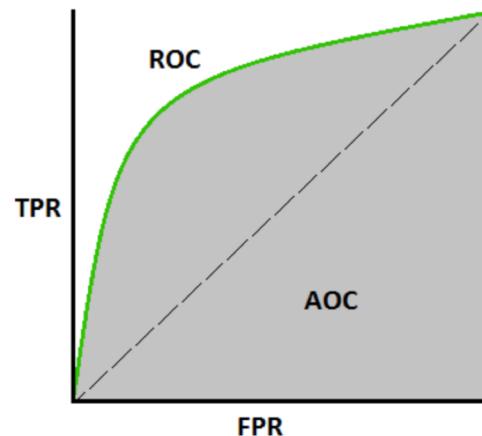


2.1 모델 개발 Process

개발 Process – 모델 성능 평가

해당 평가 지표들을 통해 모델의 변별력과 모델의 안정성을 평가한다.

ROC : 변별력 지표



PSI : 모델 안정성 지표

$$PSI = \sum_{\text{등급별}} (\%O - \%E) \cdot \ln \frac{\%O}{\%E} \quad (\%E : \text{기준시점 구성비}, \%O : \text{현재 구성비})$$

신용 등급	기준시점 고객수	현재 고객수	기준시점 구성비(%E)	현재 구성비(%O)	%O-%E	$\ln(\%O/\%E)$	PSI
1	600	700	20.0	21.9	1.9	0.0896	0.0017
2	1,000	900	33.3	28.1	-5.2	-0.1699	0.0088
3	1,000	1,100	33.3	34.4	1.0	0.0308	0.0003
4	400	500	13.3	15.6	2.3	0.1586	0.0036
합계	3,000	3,200	100%	100%	-	-	0.0144

$\rightarrow PSI=0.0144$

$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{\text{FP}}{\text{TN} + \text{FP}}$$

2.1 모델 개발 Process

개발 Process – 모델 output 확인

완성된 모델을 output으로 표현한다. Output은 모델의 알고리즘별로 상이하다.

예)

Score Card 형태

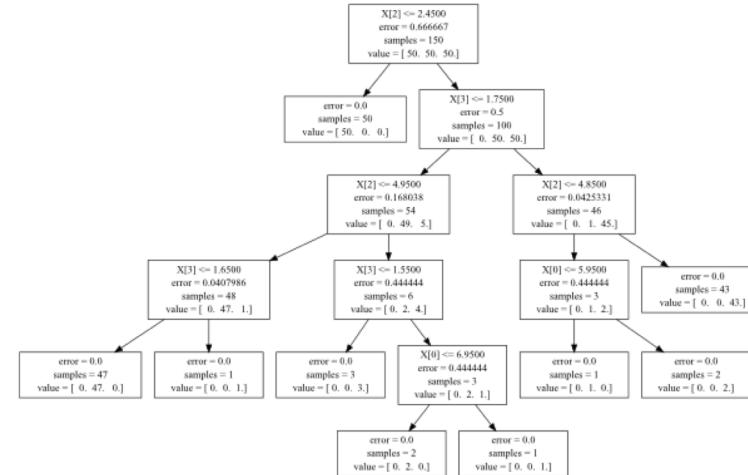
Logistic Regression을 사용한 경우

항목	상세	배점
결혼여부	미혼	0
	기혼	5
근속년수	2년이하	0
	3~5년	68
6개월내 카드 개설	6년이상	158
	없음	0
타금융기관 총대출건수	있음	51
	3건이상	0
타금융기관 총대출건수	2건	17
	1건	45
	0건	60

예)

Tree 형태

Decision Tree를 사용한 경우



2.1 모델 개발 Process

개발 Process – Score 및 등급화

예측된 값을 고객이 이해하기 쉬운 형태인 score(ex. 1~1000점) 또는 등급(ex. 1~10등급) 형태로 변환한다.

예) Score로 변환시

	PD (Probability of Default)	Score
1	0.000%	1,000 ~ 876
2	0.125% ~ 0.249%	875 ~ 851
3	0.250% ~ 0.374%	850 ~ 826
4	0.375% ~ 0.499%	825 ~ 801
5	0.500% ~ 0.624%	800 ~ 776
6	0.625% ~ 0.749%	775 ~ 751
7	0.750% ~ 0.874%	750 ~ 726
8	0.875% ~ 0.999%	725 ~ 701
9	1.000% ~ 1.249%	700 ~ 676
10	1.250% ~ 1.499%	675 ~ 651
11	1.500% ~ 1.749%	650 ~ 626
12	1.750% ~ 1.999%	625 ~ 601
13	2.000% ~ 2.499%	600 ~ 576
14	2.500% ~ 2.999%	575 ~ 551
15	3.000% ~ 3.499%	550 ~ 526
16	3.500% ~ 3.999%	525 ~ 501
17	4.000% ~ 4.499%	500 ~ 481
18	4.500% ~ 4.999%	480 ~ 461
19	5.000% ~ 5.499%	460 ~ 441
20	5.500% ~ 5.999%	440 ~ 421
21	6.000% ~ 6.499%	420 ~ 401
22	6.500% ~ 6.999%	400 ~ 381
23	7.000% ~ 7.499%	380 ~ 361
24	7.500% ~ 7.999%	360 ~ 341
25	8.000% ~ 8.499%	340 ~ 321
26	8.500% ~ 8.999%	320 ~ 301
27	9.000% ~ 9.499%	300 ~ 281
28	9.500% ~ 9.999%	280 ~ 261
29	10.000% ~ 10.499%	260 ~ 241
30	10.500% ~ 10.999%	240 ~ 221

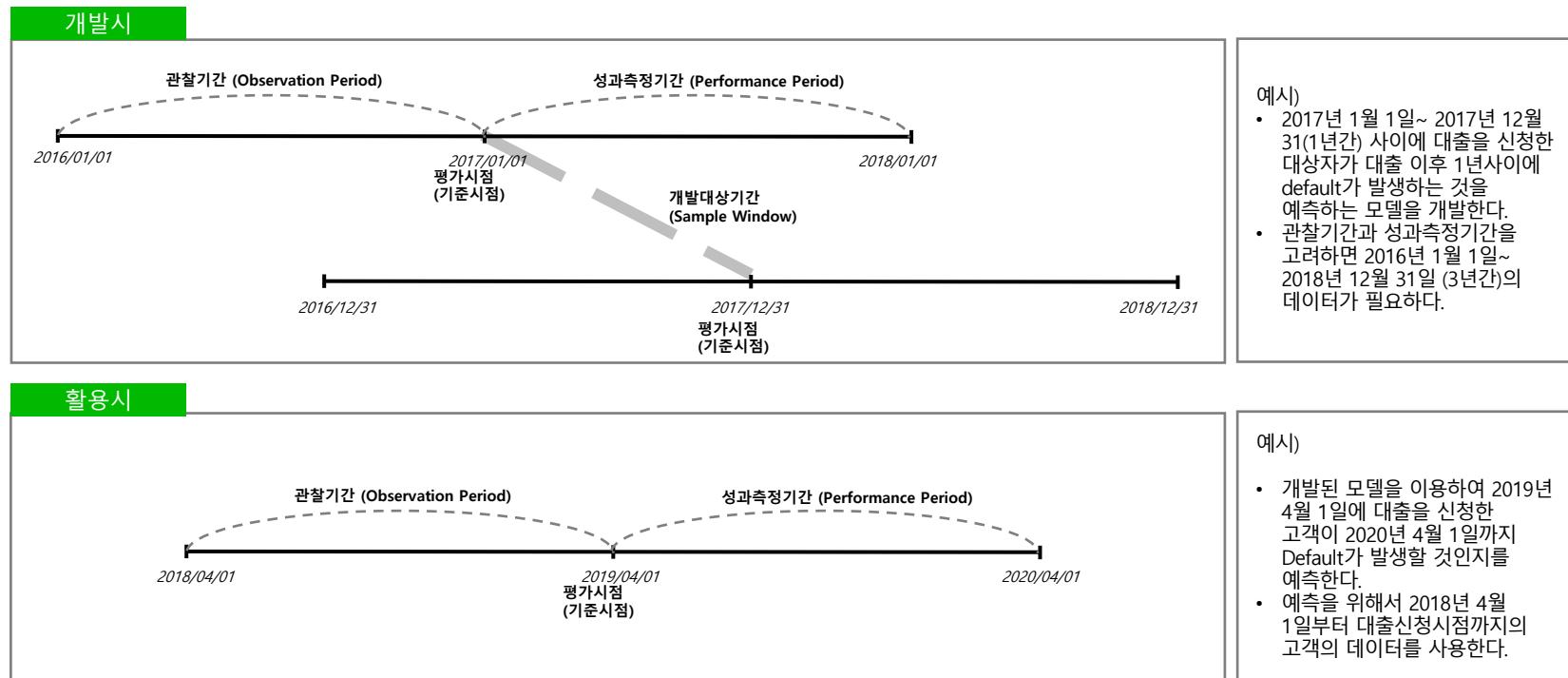
예) 등급으로 변환시

등급	평점구간
1등급	820점 이상
2등급	800점 이상 820점 미만
3등급	780점 이상 800점 미만
4등급	750점 이상 780점 미만
5등급	730점 이상 750점 미만
6등급	730점 이상 750점 미만
7등급	650점 이상 700점 미만
8등급	600점 이상 650점 미만
9등급	530점 이상 600점 미만
10등급	530점 미만

2.1 모델 개발 Process

개발 Process – 모델 활용

평가시점을 기준으로 과거의 데이터를 이용하여, 성과측정기간동안 타겟이 발생할 확률을 측정한다.



2.2 신용 평가 모델링 특징

1. 시간에 따른 변화 중요 & 샘플(row)마다 학습/예측 시기가 다름

- 장기간에 걸쳐 관찰되고 적재된 데이터 사용 ex) 과거, 개발 대상, 미래 : 3년
- 외부 변수를 잘 이해하는 것이 중요 ex) seasonal factor, 이벤트

2. 모델을 통해 어떤 대출 상품의 고객을 평가할 지가 중요

- 대출 상품에 따라 개발 모집단 선정 ex) 단기론, 일반금융권 대출, P2P 대출
- 단기/중기/장기 상품 or 대출 금액에 따라 다름

3. 다른 산업의 예측 모델에 비해, 설명 가능성이 중요

- 한국의 경우, 대출 승인 거절 사유에 대한 설명 의무가 있음 -> 알고리즘 제한
- 모델의 성능 이외에도 고려할 것들이 多

4. 도메인 지식 & 산업에 대한 이해가 필요

- 금융 데이터 법적 규제 多
- 대출 상품 및 신용 평가 구조에 대한 이해 필요

5. 대안 데이터 / 빅 데이터의 등장으로 도전해볼 수 있는 영역 多

- 다른 핀테크 산업들과 시너지

2.3 IT에서의 신용 평가 모델링 / 데이터 분석

1. 데이터 크기 Big & 개발 환경 Good

- 장기간에 걸쳐 관찰되고 적재된 데이터 사용
- 데이터 크기만큼 그에 걸맞는 개발 환경 지원 ex) 스파크, 제플린

2. 기획, 사업, UX/UI, 개발, 데이터 엔지니어들과의 협업

- 기획 - 어떤 대출 상품을 기획하는지, 해당 대출 상품에 대한 이해 필요 ex) 만기 60일 상품
- 사업 - 마케팅, 여신 전략 수립 협의 ex) 허용 부도율은 어디까지인가, 시뮬레이션
- 개발 - 모델이 java로 구현된 서비스 안에서 워킹 ex) PMML
- 데이터 엔지니어 - 데이터베이스, 데이터 구조, hadoop, spark, 컴퓨터 사이언스 지식 필요

3. 기본적으로 신용 평가에 대한 이해도가 높지 않음

- 금융권에 비해 신용 평가에 대한 이해가 상대적으로 부족
- 커뮤니케이션 비용 多

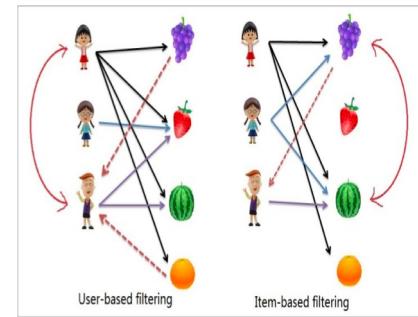
4. 다른 핀테크 서비스들과의 협업

3. Get Ready

3.1 To be Data Scientist



3.1 To be Data Scientist



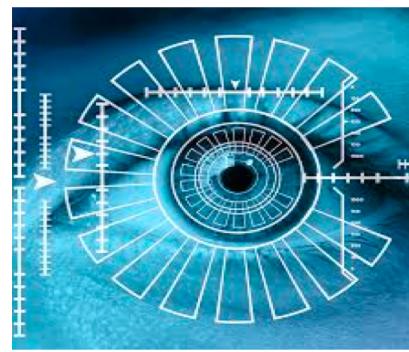
게임

광고 / 추천

다양한 분야



마케팅



머신 러닝 LAB

비전, 음성 인식, 텍스트

FDS

감사합니다.