신용카드 고객이탈분석

멘티 이은서

목차

01
개요
- 과제 정의

02
분석 방법
- 분석 절차
- 예측 모델

03
모델 개발
- 모델 생성 – 로지스틱 회귀모형, 의사결정나무, 랜덤포레스트

04
분석 결과 및 결론
- 분석 결과
- 결론

1. 과제 정의

문제 정의

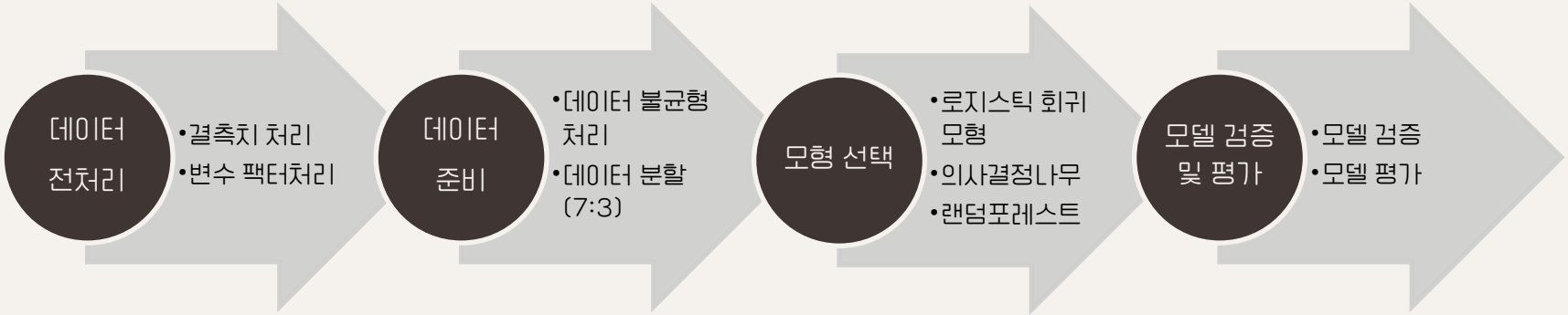
- 현재 데이터를 이용한 모델로는 이탈하지 않은 고객이 이탈고객으로 분류될 가능성이 있기에 더 정확한 분류를 하는 모델의 구축의 필요가 있음
- 고객이 사용할 카드를 선택하는 데에 있어 영향을 미칠 고객 특성에 대한 분석이 필요함

분석 목적 및 기대 효과

- 기존에 분류하지 못했던 숨은 이탈고객을 분류하여 분류 원인 발견 가능
- 이탈 가능성이 있는 고객의 상황과 이탈이유를 분석하여 향상된 고객서비스 제공
- 다양한 카드 서비스 제공을 위한 고객 특성 분석 및 희망 서비스 예측 가능

1. 분석 절차

- 신용카드 고객 데이터셋을 활용함 (https://www.kaggle.com/sakshigoyal7/credit-card-customers)
- 데이터 전처리 후 모델을 설계하고, 가설검증을 통해 모델의 적합성 판단



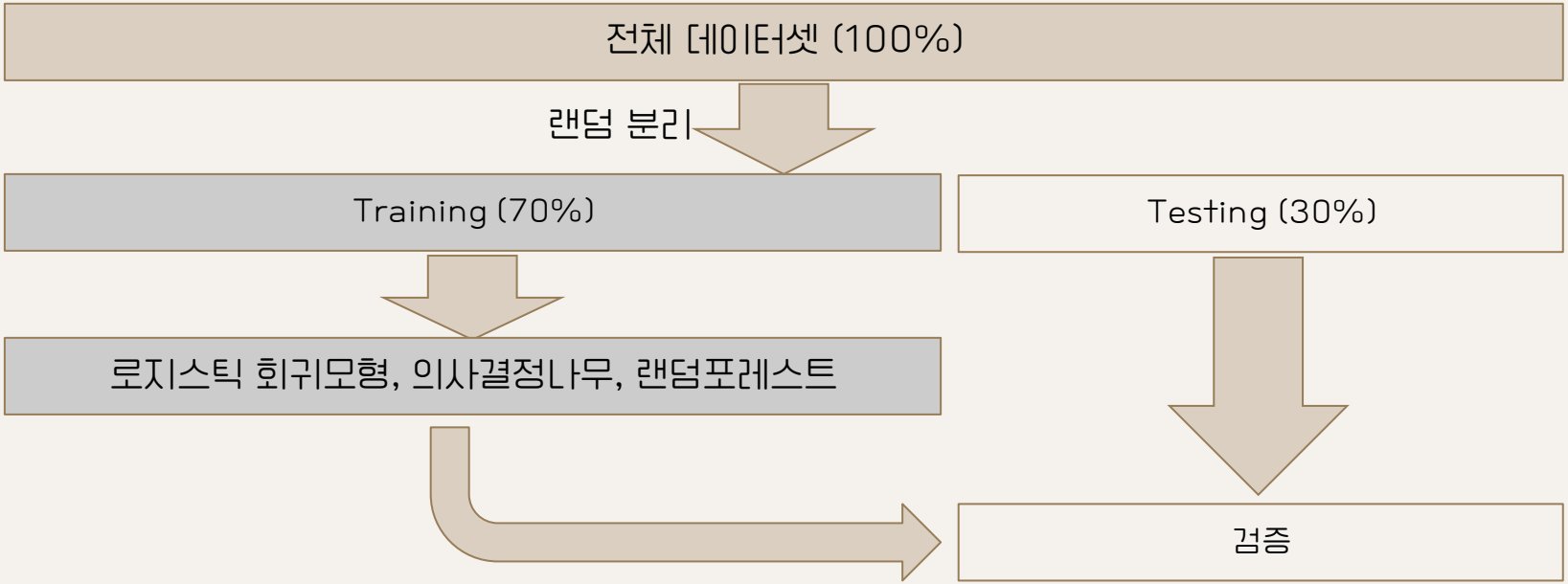
2. 분석 데이터셋

- 수집 대상 : 신용카드 고객 10,127명
- 고객번호를 제외한 범주형 변수 6개(이탈 여부, 성별, 부양 가족수, 학력, 결혼 여부, 수입)와 수치형 변수 14개로 구성
- 분석에 이용할 독립변수는 label(0: 비이탈, 1: 이탈)로 선정
- 변수 목록

번호	변수명	번호	변수명
1	"Customer_Age"	2	"Gender"
3	"Dependent_count"	4	"Education_Level"
5	"Marital_Status"	6	"Income_Category"
7	"Card_Category"	8	"Months_on_book"
9	"Total_Relationship_Count"	10	"Months_Inactive_12_mon"
11	"Contacts_Count_12_mon"	12	"Credit_Limit"
13	"Total_Revolving_Bal"	14	"Avg_Open_To_Buy"
15	"Total_Amt_Chng_Q4_Q1"	16	"Total_Trans_Amt"
17	"Total_Trans_Ct"	18	"Total_Ct_Chng_Q4_Q1"
19	"Avg_Utilization_Ratio"	20	"label"

3. 예측 모델

데이터셋을 7:3으로 나누어 훈련용과 시험용으로 분리한 후 로지스틱 회귀분석, 의사결정나무, 랜덤포레스트를 활용하여 분석 예정



1. 모델 생성 – 로지스틱 회귀분석

회귀모형 생성 후 후진제거법(Backward Elimination)을 이용하여 최적모형 생성

```
step(model1, direction = "backward")
```

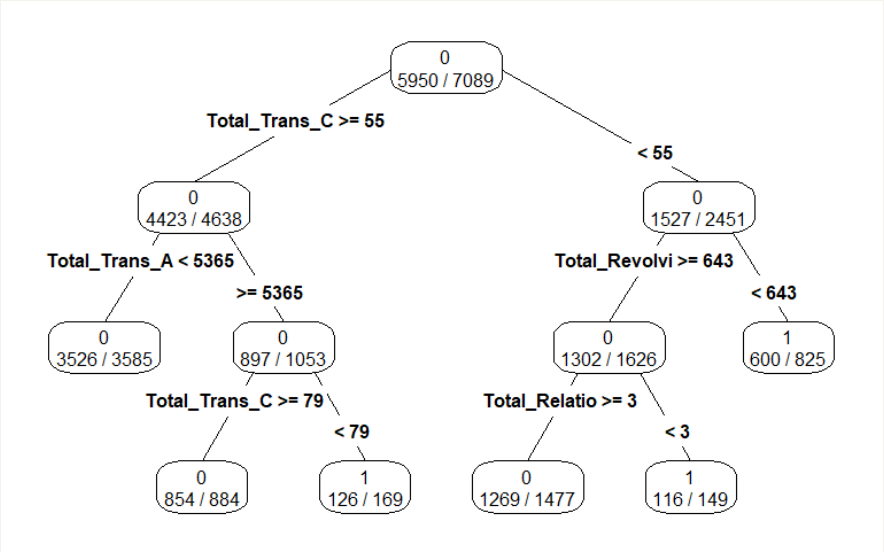
```
## Call:  glm(formula = label ~ Customer_Age + Gender + Dependent_count +
##     Marital_Status + Income_Category + Card_Category + Total_Relationship_Count +
##     Months_Inactive_12_mon + Contacts_Count_12_mon + Credit_Limit +
##     Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt +
##     Total_Trans_Ct + Total_Ct_Chng_Q4_Q1, family = "binomial",
##     data = train)
```

최종 로지스틱 회귀모형

```
glm(label ~ Gender + Dependent_count + Marital_Status + Income_Category +
    Card_Category + Months_on_book + Total_Relationship_Count +
    Months_Inactive_12_mon + Contacts_Count_12_mon + Credit_Limit +
    Total_Revolving_Bal + Total_Trans_Amt + Total_Trans_Ct +
    Total_Ct_Chng_Q4_Q1, family = "binomial", data = train)
```

2. 모델 생성 – 의사결정나무

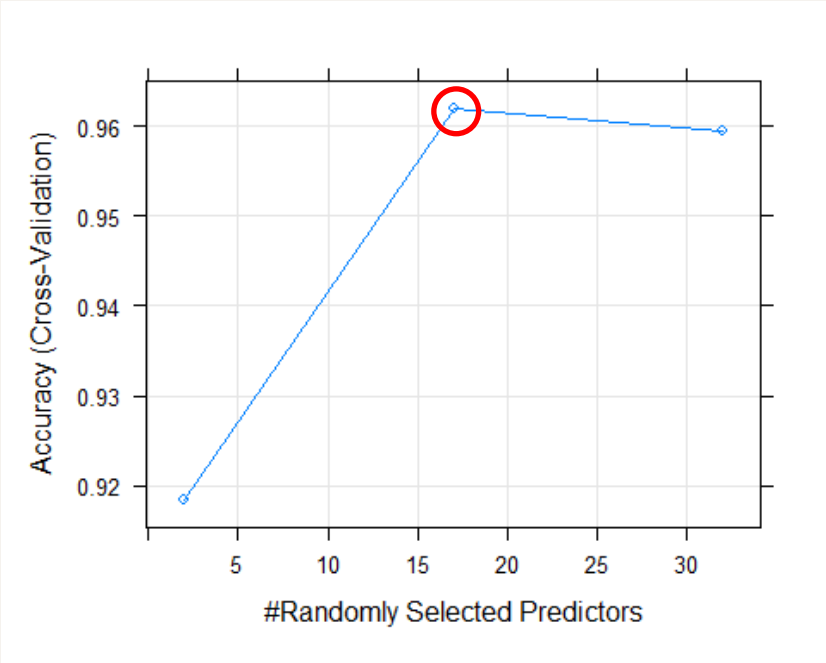
과적합을 방지하기 위해 가지치기(pruning)을 진행한 후 8개의 변수 (Total_Trans_Ct, Total_Revolving_Bal, Total_Relationship_Count, Total_Ct_Chng_Q4_Q1, Credit_Limit, Contacts_Count_12_mon, Months_Inactive_12_mon, Customer_Age)을 이용하여 모델을 생성함



3. 모델 생성 – 랜덤포레스트

의사결정나무 보다 예측 정확성을 높이기 위해 랜덤포레스트 모델을 생성, 변수가 17일 때 정확성이 96%로 가장 높게 나오는 것을 확인 가능

```
## Random Forest
##
## 7089 samples
##   19 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 6380, 6380, 6380, 6380, 6380, 6380, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9184630  0.6322065
##   17    0.9619126  0.8544032
##   32    0.9595155  0.8464446
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 17.
```

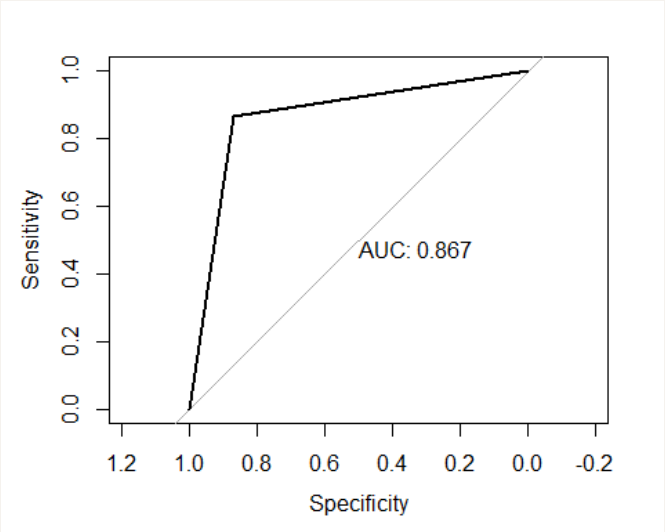


1. 분석결과 – 로지스틱 회귀분석

결과 – 컨퓨전 매트릭스와 ROC 커브

예측 \ 실제	비이탈	이탈
비이탈	1906	62
이탈	284	392

정확도 (Accuracy)	민감도 (Sensitivity)	특이도 (Specificity)
0.8691	0.8634	0.8703



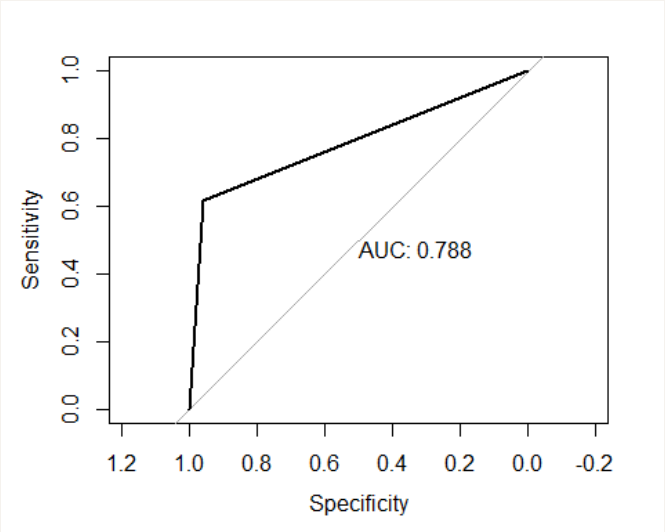
- 정확도 87%, 민감도 86%로 실제 이탈 고객을 제대로 식별하고 있음
- AUC = 0.867로 데이터 분석모델로서 적합함

2. 분석결과 – 의사결정나무

결과 – 컨퓨전 매트릭스와 ROC 커브

예측 \ 실제	비이탈	이탈
비이탈	2451	188
이탈	99	300

정확도 (Accuracy)	민감도 (Sensitivity)	특이도 (Specificity)
0.9055	0.61475	0.96118



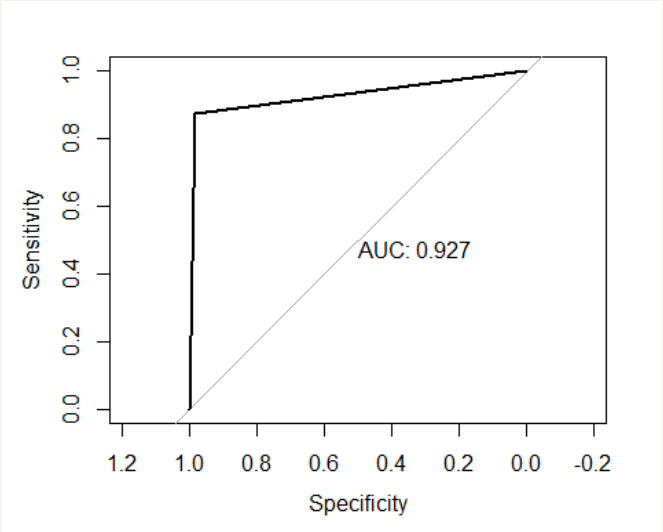
- 정확도 91%, 민감도 61%로 로지스틱 회귀분석보다는 더 높은 예측 정확도를 보여줌
- AUC = 0.788로 로지스틱 회귀분석보다 데이터 분석모델로서 덜 적합함

3. 분석결과 – 랜덤포레스트

결과 – 컨퓨전 매트릭스와 ROC 커브

예측 \ 실제	비이탈	이탈
비이탈	4259	62
이탈	284	392

정확도 (Accuracy)	민감도 (Sensitivity)	특이도 (Specificity)
0.9658	0.9839	0.8703



- 정확도 97%, 민감도 98%로 세 모델 중 제일 정확성이 높고 실제 이탈고객 분류 예측이 잘 이루어짐
- AUC = 0.927로 데이터 분석모델로서 적합함

4. 결론

- 총 20개의 변수를 사용
- 후진제거법을 이용한 로지스틱 회귀분석 모델을 생성 (정확도 87%, 민감도 86)
- 가지치기 후 8개의 변수를 이용한 의사결정나무 모델 생성 (정확도 91%, 민감도 61%)
- 17개의 변수를 이용한 랜덤포레스트 모델 생성 (정확도 97%, 민감도 98%)
- 분석 결과, 생성한 3개의 모델 중 랜덤포레스트 분석모델이 가장 정확성이 높고 민감도가 높아 데이터의 특성에 부합하는 결과 도출