인공지능을 활용한 예금 가입 여부 예측 모델 및 성능 평가

써니C 1기, 신은영

INDEX

1.개발 개요

a. 정의: 본 프로젝트는 마케팅의 효과를 높이기 위해 예금 가입 대상자를 예측하는 목적으로 수행되었다.

b. 개발 목적: 예금 가입 여부 예측

c. 기대 효과

1) 예금 가입 확률이 높은 고객을 대상으로 마케팅을 진행하여 기존보다 시간과 비용이 절약될 수 있다.
2) 효율적으로 마케팅이 진행되어 성공률을 높일 수 있다.

d. 개발 계획

1) 탐색적 데이터 분석을 통해 해당 프로젝트에 대한 Data Set 파악

2) Feature Engineering을 통한 데이터 개선

3) 인공지능 모델링

4) 데이터 개선과 하이퍼파라미터 조정을 통해 모델 고도화

5) 프로젝트 모델 평가 레포트 제작

2. 데이터 설명 및 탐색적 분석

a. 파일 형식: csv 파일

b. 데이터 크기: 4.63 MB

c. Data: bank-additional-full (rows total 41,188개, comlumns 20개)

고객과의 전화통화 기반 마케팅 캠페인을 통해 수집.

d. 독립변수 설명**(Features)**: 20개 (수치형: 10개, 범주형: 10개)

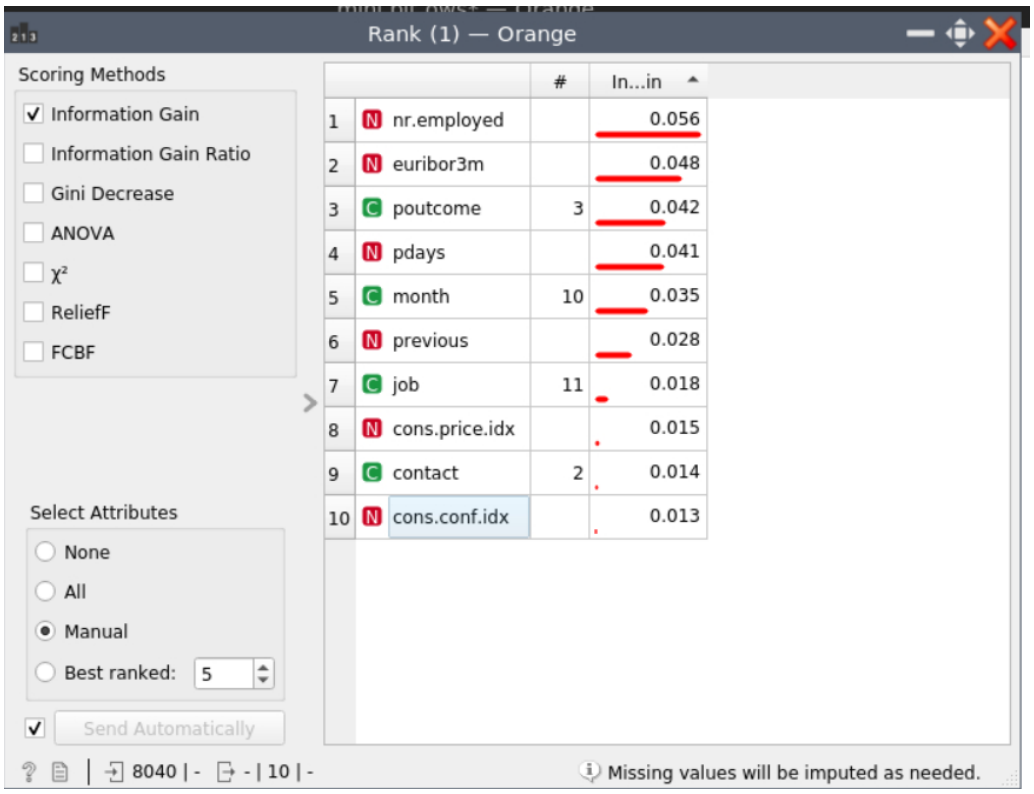
1) age: 나이
2) job: 직업 종류 12종
3) marital: 결혼 여부 4종
4) education: 학력 8종
5) default: 신용불량 여부 3종
6) housing: 주택 대출 여부 3종
7) loan: 개인 대출 여부 3종
8) contact: 통신 유형 2종
9) month: 전화 연결된 월
10) day_of_week: 전화 연결된 요일
11) duration: 통화 연결 시간
12) campaign: 이 캠페인 동안 해동 고객에게 접촉한 횟수
13) pdays: 이전 캠페인에서 고객에게 마지막으로 연락한 이후 얼마나 지났는지
14) previous: 이 캠페인 이전에 고객에게 접촉한 횟수
15) poutcome : 이전 마케팅 캠페인의 결과
16) emp.var.rate: 고용 변동률
17) cons.price.idx: 소비자 물가지수
18) cons.conf.idx: 소비자 신뢰 지수
19) euribor3m: 유로보 3개월 금리
20) nr.employed: 직원들의 수

e. 종속변수 설명**(Label)**:

Y: 고객과의 전화 통화 기반 마케팅 캠페인을 통해 실제로 예금 가입을 했는지 여부

f. Rank: 종속변수('Y')와 다른 독립변수(features)와 비교

1) 범주형 변수를 많이 포함하고 있어 rank를 통해 연관도 확인
2) preprocess위젯으로 전처리 한 이후 rank를 통해 시각화 진행
3) 분석내용: rank 실행결과 가장 높은 연관도를 가진 feature도 0.056 밖에 되지 않는다. 큰 의미가 없다고 생각하여 rank 실행결과는 반영하지 않았다.



g. 데이터의 특징

1) 범주형 데이터와 수치형 데이터가 섞여 있다.
2) label과 feature들의 연관도가 전체적으로 낮다.

3. 데이터 전처리 및 **Feature Engineering**

a. select column

feature 중 duration은

1) 시간이 지나치게 짧으면 가입하지 않음을 알 수 있어 결과가 지나치게 좋게 나온다.
2) 통화 이후에 얻을 수 있는 데이터이기 때문에, 이용 시점인 통화 전에는 얻을 수 없는 데이터이다.

위와 같은 문제를 가지고 있어 select column을 이용하여 제거해준다.

b. feature statistics & impute

1) feature statistics를 이용해 어떤 feature에 결측치가 얼마나 있는지 확인한다.
2) 다음과 같은 이유와 방법으로 결측치를 처리한다.
    1. default (8597개, 20%) -> Model-based imputer : 결측치의 양이 많아 삭제하지 않았다. 또한

2. education (1731개, 4%) -> random : 학력과 예금 가입은 큰 관련이 없다고 생각하였고, 기존의 데이터의 분포가 고르게 형성되어있어 random배정하였다.
3. housing, loan (990개, 2%) -> remove : 2%는 적은 양에 해당하고, 대부분이 housing과 loan이 같이 없는 경우라 삭제해주었다.
4. marital (80개, 0%) -> remove : 삭제해도 결과의 큰 영향을 주지 않을 정도의 적은양이라 삭제해주었다.

c. select column

과적합을 방지하기 위해 큰 의미가 없는 변수들은 제거하고자 하였으나, 거의 모든 변수들이 모델 평가에 유의미한 영향을 주는것으로 나타났다. 따라서 duration 외에는 변수를 제거하지 않았다.

d. Feature Constructor

데이터를 개선하여 모델의 정확도를 높이기 위해 비슷한 속성을 가지고 있는 변수끼리 합쳐 새로운 변수를 만들고자 하였으나 유의미한 결과를 얻지 못하였다.

4. Orange를 활용한 AI 모델링

a. 분석 플랫폼: Orange3

b. 아래 알고리즘들을 이용해 분석 및 예측 수행

1) Logistic Regression
   구체적인 수치를 예측하는 모델이 아닌 예금 가입 여부를 예측하는 모델이기 때문에 분류 모델인 Logistic Regression을 사용하였다.
2) Random Tree
   Logistic Regression과 같은 이유로 분류모델 중 하나인 Random Tree를 사용하였다. Tree의 수는 평가지표가 가장 좋게 나오는 150으로 설정하였다.
3) Test and Score
   설계한 모델을 평가하기 위해 Test and Score 위젯을 사용하였다. 10개의 그룹을 랜덤으로 설정하여 테스트 하도록 설정하였다.
4) Confusion Matrix
5) ROC Analysis

c. 평가지표

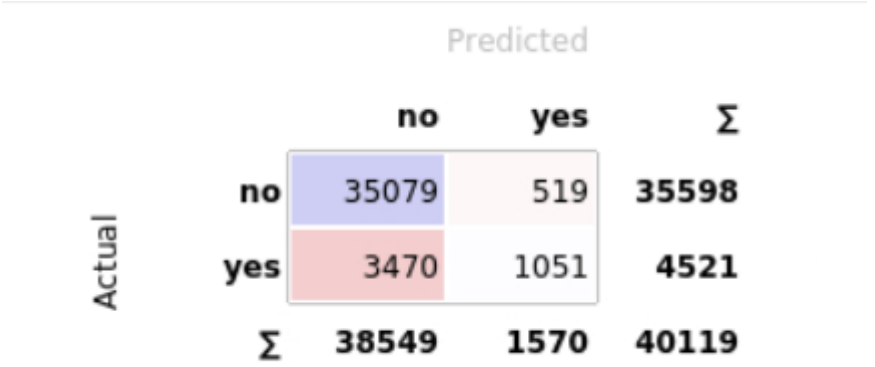
1) AUC
2) CA(Classification Accuracy)
3) Precision
4) Recall

d. 결과

1) Logistic Regression
   1. Test and Score

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.786	0.901	0.878	0.883	0.901

2. Confusion Matirx

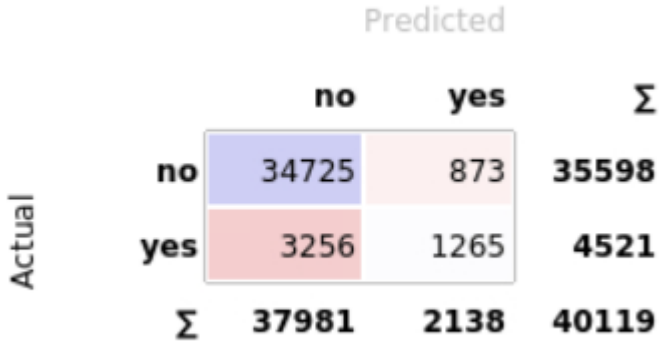


2) Random Forest

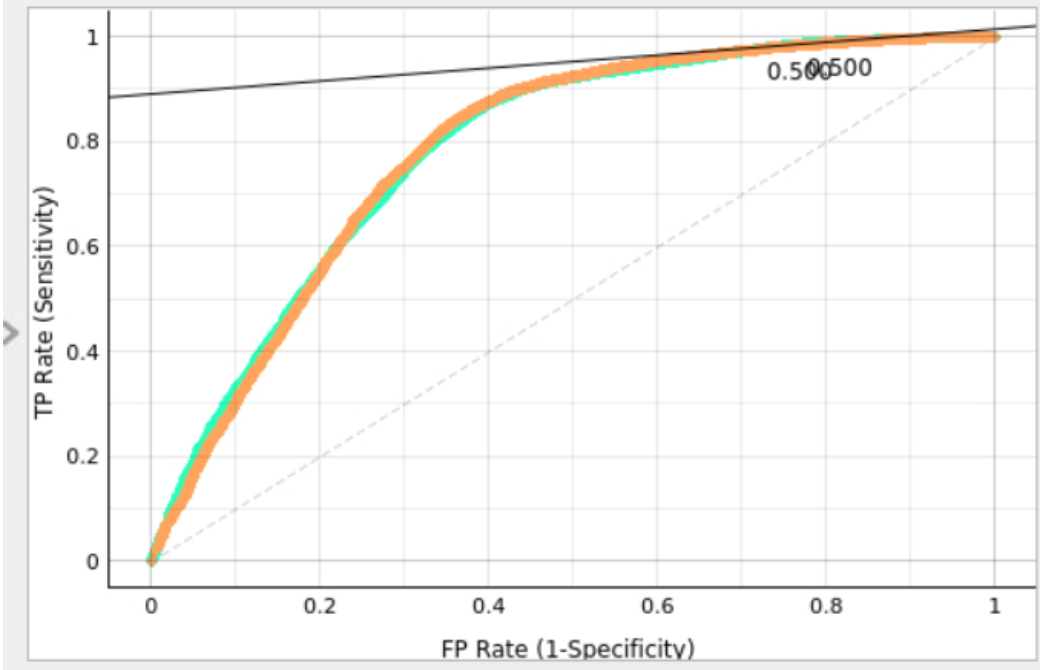
1. Test and Score

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.785	0.897	0.880	0.878	0.897

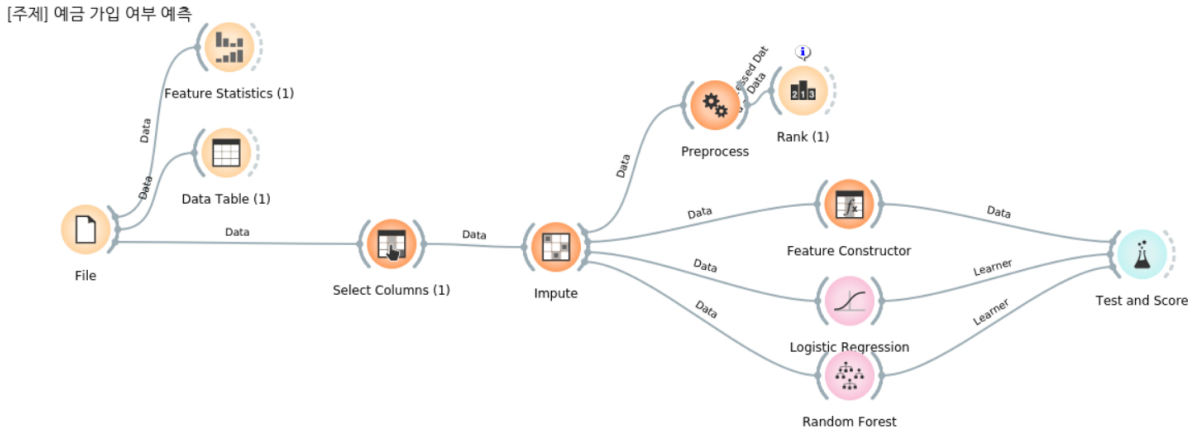
2. Confusion Matrix



3) ROC Analysis



e. Workflow



5. 향후 추가 개선 과제

a. 데이터 관점 고도화
feature가 19개로 상당히 많은 편이다. 여러 조합을 통한 파생 변수를 만든다.

b. 모델 관점 고도화
모델 관점의 고도화를 진행하였지만 AUC에 큰 변화가 없었다. AUC가 유의미하게 좋아지기 위한 추가적인 고도화가 필요하다.

6. 프로젝트를 통하여 얻은 교훈

a. 데이터 정제에 대한 인사이트
결과에 직접적인 영향을 주는 데이터는 제거해 주어야 정확한 모델을 만들 수 있다.

b. 인공지능 모델을 사용할 수 있다는 자신감
어려운 프로그램이 아닌 상대적으로 간단한 orange3를 이용해서도 충분히 모델을 만들 수 있으며 연습에 그치는 것이 아닌 실제로 사용할 수 있다는 자신감을 얻을 수 있었다.