데이터 사이언스 입문 – 기말고사

2016-06-16

11:30 ~ 12:50

기말시험을 위해 히즈넷 과제물 페이지에 train dataset ("trainData.csv")과 test dataset("testData.csv")이 제공됩니다. 주어진 데이터는 대만의 한 신용정보 회사에서 2005년 고객별로 채무불이행(default) 정보를 나타내는 데이터입니다. train dataset에서는 각 고객들의 채무 불이행 여부가 포함되어 있고, test dataset에서는 채무 불이행 정보가 나타나있지 않습니다. 여러분들은 train dataset으로부터 채무불이행 여부를 예측하는 모델을 만들어서 test dataset의 고객들에 대해 채무불이행 여부를 예측하여야 합니다.

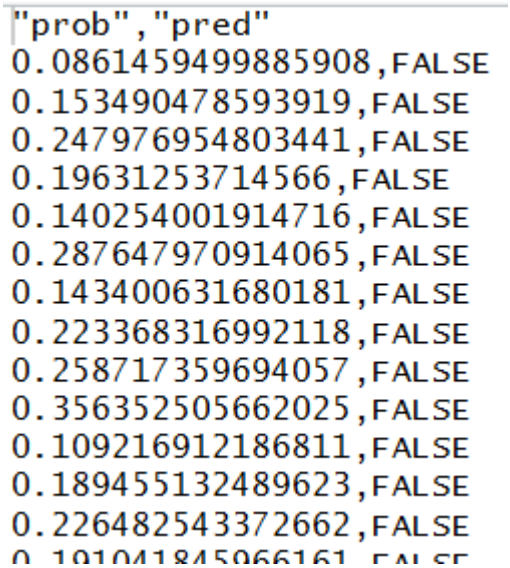
다음은 dataset에 포함된 변수들의 구체적인 설명입니다.

변수명	설명
default.payment.next.month*	채무불이행 여부, 1=채무불이행 0=채무불이행 아님
LIMIT_BAL**	신용한도
SEX	성별 1=남자 2=여자
EDUCATION	교육수준 1=대학원 2=대학 3=고등학교 4=그외
MARRIAGE	결혼여부 1=기혼 2=미혼 3=기타
AGE	나이
PAY_1 ~ PAY_6	1달전부터 6달전까지 상환내역 -1=제때 상환, 1=한달연체 2=두달연체 3=세달연체, …… , 9=9달연체 혹은 그 이상
BILL_AMT1 ~ BILL_AMT6**	1달전부터 6달전까지 청구금액
PAY_AMT1 ~ PAY_AMT6**	1달전부터 6달전까지 상환금액

* 예측하고자 하는 변수 (test dataset에는 빠져있음)

** 금액과 관련된 단위는 대만달러입니다 (NT dollar)

과제를 통해 여러분이 제출해야 하는 것은 2가지 입니다. 하나의 csv파일과 하나의 보고서(r 스크립트를 포함), csv파일은 두 개의 column을 구성됩니다('**prob'** 와 '**pred'**). prob column에는 고객이 default가 날 확률 값을 가지게 하고 pred column에는 default이냐 아니냐 (TRUE/FALSE)를 나타내야 합니다. csv파일의 예시는 다음과 같습니다.



```
"prob","pred"
0.0861459499885908,FALSE
0.153490478593919,FALSE
0.247976954803441,FALSE
0.19631253714566,FALSE
0.140254001914716,FALSE
0.287647970914065,FALSE
0.143400631680181,FALSE
0.223368316992118,FALSE
0.258717359694057,FALSE
0.356352505662025,FALSE
0.109216912186811,FALSE
0.189455132489623,FALSE
0.226482543372662,FALSE
0.191041845966161,FALSE
```

	prob	pred
1	0.086145950	FALSE
2	0.153490479	FALSE
3	0.247976955	FALSE
4	0.196312537	FALSE
5	0.140254002	FALSE
6	0.287647971	FALSE
7	0.143400632	FALSE
8	0.223368317	FALSE
9	0.258717360	FALSE
10	0.356352506	FALSE
11	0.109216912	FALSE

Figure 1 제출 csv파일 예시

파일의 형식이나 column의 이름을 정확하게 지켜주기 바랍니다. 채점이 힘들어지기 때문에 minor한 감점이 있을 수 있습니다. **csv파일에 포함 되어야 하는 확률과 예측 값은 주어진 testData.csv파일에 대한 채무불이행 예측입니다 (5000건).**

(중요!!) 다음으로 제출해야 하는 파일은 보고서 파일로서 **예측에 사용한 모델에 대한 자세한 설명**, 그 **모델을 사용하게 된 논리적인 근거**, **최적의 모델을 선택하게 된 과정에 대한 충분한 설명**이 있어야하고 필요한 경우에 사용한 r code도 포함할 수 있습니다. 또한 **train data에서 측정한 성능**에 대해서 기록하고 **어떤 과정으로 성능 개선을 하였는지에 대한 설명**도 필요합니다.

제출하는 csv파일과 보고서 파일은 히즈넷에 별도로 업로드할 수 있도록 과제를 생성하였습니다. **압축파일로 올리지 말고 각각 따로 올려주세요**.

채점에 사용할 performance measure는 확률에 대해서는 AUC, TRUE/FALSE prediction에 대해서는 F1 score입니다. F1 score는 다음과 같이 계산합니다.

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

F1 score는 precision과 recall이 비슷한 수준으로 높아야 높게 나오며 둘 중에 하나가 값이 현저히 낮아지면 F1 score는 값이 낮아집니다. F1 score는 값이 높을수록 (1에 가까울수록) 좋은 성능

을 의미합니다.

여러분이 목표로 해야 하는(=뛰어 넘어야 하는) 성능은 AUC의 경우 0.7216454, F1 score의 경우 0.4982243입니다. 이는 특별한 데이터 가공 없이 logistic regression을 이용해서 test dataset에 대해 얻은 성능 값입니다.

좋은 성적을 위해서는 다양한 방법으로 예측 모델의 성능을 향상해서 정확한 예측 결과를 얻어야 합니다. 그러나 성능 개선을 위해 사용한 방법들의 과정과 근거가 충분히 합리적이고 논리적이라면 성능 개선에 크게 성공하지 못하더라도 만족할만한 성적을 받을 수 있을 겁니다. 반대로 좋은 예측 결과를 얻었다고 하더라도 과정에 대한 설명이나 근거가 불충분하면 좋은 점수를 받을 수 없습니다.