Data Science

Homework 11

Practice 11: Logistic Regression

Loading data into R

```
dim(train_df)
## [1] 25000  24
dim(test_df)
## [1] 5000  24
```

prediction for customer payments

For this question, you will be working with dataset about customer default payments from a Taiwan credit card company. We aim to predict the probability of default payment for the customer with high accuracy with logistic regression model. The variables in the dataset are described as follows:

이번 문제는 타이완 신용회사의 고객정보 데이터를 가지고 분석하게 됩니다. 고객의 채무 불이행 확률을 예측하기 위한 logistic regression model을 학습하게 됩니다. 변수의 설명은 아래와 같습니다.

변수(variable)	의미(Description)
default.payment.next.month	(목적변수 target var.) 채무불이행 여부, 1=채무불이행(default) 0=채무불이행 아님(not default)
LIMIT_BAL**	신용한도 Amount of the given credit(NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
SEX	성별 1=남자 2=여자 Gender(1=male; 2=female)
EDUCATION	Education (1=graduate school; 2=university; 3=high school; 4=others)
MARRIAGE	Marital status (1 = married; 2=single; 3=others).
AGE	나이 Age(year).
PAY_1 ~ PAY_6	History of past payment. We tracked the past monthly payment records (from April to September, 2005)

	PAY_1 = the repayment status in Sept. 2005; PAY_2 = in Aug. 2005; …; PAY_6 = the repayment status in April, 2005. -1 = 제때 상환, 1=한달연체 2=두달연체 3=세달연체…, 9=9달연체 혹은 그 이상 The measurement scale for the repayment status is: -1=pay duly; 1=payment delay for one month; 2=payment delay for two months; . . .; 8=payment delay for eight months; 9=payment delay for nine months and above.
BILL_AMT1 ~ BILL_AMT6**	1달전부터 6달전까지 청구금액 Amount of bill statement from 6 months ago to the last month BILL_AMT1: amount of bill statement in September, 2005 BILL_AMT2: amount of bill statement in August, 2005… BILL_AMT6: amount of bill statement in April, 2005.
PAY_AMT1 ~ PAY_AMT6	1달전부터 6달전까지 상환금액 Amount of previous payment PAY_AMT1: amount paid in September, 2005 PAY_AMT2: amount paid in August, 2005… PAY_AMT6: amount paid in April, 2005

*currency is NT dollar. (단위는 타이완 달러)

Question 1

Build a logistic regression model with glm function to predict the probability of default payment using all given variables in the dataset. Describe the process, including data preparation if necessary, to obtain your model.
고객의 채무 불이행 확률을 예측하는 로지스틱 회귀 모델을 만들어라. (주어진 모든 변수를 사용). 모델을 만드는 과정을 설명하고, 필요하다면 전처리도 수행하고 전처리 과정도 설명하여라.

Question 2

What is the AUC of your model for both test and training dataset for the prediction model of QI?
For the testdataset, you use the website and Attach a captured image Of the evaluation of AUC.
1번 문제에서 학습한 모델의 AUC를 측정하시오.

Question 3

Looking at the summary of logistic model of QI, what are the variables increase the risk of default mostly?
What are the variables reduce the risk of default?
Explain how much and in what direction they change the risk of customers' default.
문제 1에서 학습한 모델의 해석을 보면, 어떤 변수가 채무불이행 확률을 높이는데 많은 영향을 주는가? 채무불이행 확률을 낮추는 변수는 어떤 것이 있는가?
변수들의 값의 변화에 따라 채무불이행의 확률이 어떻게 변하는지 해석해보자.

Question 4

For logistic model of QI, calculate accuracy, precision, and recall with threshold of 0.5.
Considering the cost of two different types of error, false positive and false negative, adjust threshold and find new accuracy, precision, and recall.

- for this question, use only training dataset.

1번 문제의 모델에서 train data에서 threshold를 0.5로 했을 때, accuracy, precision, recall을 계산하여 보라.
false positive와 false negative의 비용을 고려하여 threshold를 조정한 후 새로운 accuracy, precision과 recall을 계산하라.

- 4번 문제에서 threshold의 조정은 train data에 대해서만 수행한다.

Question 5

Apply the threshold you chose in Q4, and find accuracy, precision, and recall. Are they similar in both test and train datasets?
4번 문제에서 조정한 threshold에 대해서 test 데이터의 accuracy, precision과 recall을 계산한 후 train data에서의 결과와 비슷한지 비교하여보라.

Question 6

Use any possible way to improve your logistic regression model such as adding new variables and removing irrelvant ones.
Find the best logistic regression model that have maximal AUC for testdataset.
And adjust threshold of your best model if necessary to find your final AUC, accuracy, precision and recall for test dataset.

- You may take different way to improve your linear regression model other than adding or removing variables.

변수를 추가하거나 제거하는 등의 다양한 방법을 이용해서 모델의 성능(AUC)이 가장 높은 best model을 찾으시오.
AUC가 가장 높은 모델에 대해서 threshold 조정을 통해 최종 test data의 AUC, accuracy, precision과 recall을 계산하시오.

- best model을 얻기위해 변수를 추가하거나 제거하는 방법외에 다른 방법을 사용할 수도 있다.

Question 7

For your best model's prediction result of training dataset, investigate two groups of samples of false positive and false negative to understand the reason why they were mis-classified.

Describe your inference why your model's prediction did not work properly on those customers. For example, are they have unique distribution on certain variables?

best model에서 train data에 대한 false positive와 false negative로 나타나는 sample들을 살펴보시오.
어떠한 특징이 있는지 (변수들의 분포 등), 어떤 이유로 예측 모델이 예측에 실패하는지를 설명해보시오.