Homework 5

We have two different datasets for question 1 & 2 in **hw5_student.RData** file. Both datasets are partitioned into training and test dataset. The data frame whose name starts from student is for question 1 regression problem and one starts from credit is for question 2 classification problem. For the test dataset, you do not have outcome variable, instead you can measure the performance of your result via a website:

https://hbchoi.shinyapps.io/datascience_hw5/

Put all the prediction results of question 1 & 2 into your RData file and turn it in using the website and Edmodo.

The Rdata file should be named **"st+<your student ID>.RData"** for example "st21100123.RData", otherwise the website cannot evaluate your work.

You also need to turn in the reporting document (PDF, WORD, TXT, PPT, or any format) to Edmodo as well as RData file. The document includes all the answer of questions. This time you do not have submit r script file, but you have to put sufficient explanation of your work in the in the reporting document instead.

Students with excellent record in the ranking charts on the website will get some extra benefit.

This homework could be a good practice for the final test, since the tasks are quite similar.

It took me 2 whole days to make up this homework, so I hope you enjoy.

(Korean Translation)

여러분이 숙제에서 다루어야할 데이터는 **hw5_student.RData**파일에 있습니다. student관련 data.frame은 1번이고 credit 관련 data.frame은 2번 문제입니다. 각각 학습/테스트 데이터로 나누어져 있으며, 테스트 데이터의 경우는 목적변수(outcome)가 포함되어 있지 않습니다. 여러분의 예측 모델의 결과를 평가하기 위해서 아래 웹사이트를 이용할 수 있습니다.

https://hbchoi.shinyapps.io/datascience_hw5/

여러분의 예측 결과를 모두 RData 파일에 넣어서 웹사이트에 제출하면, 성능을 평가해 줄 것입니다. RData 파일은 st<학번>.RData 의 형식을 지켜줘야 합니다. 예를 들어 "st21100123.RData" 그렇지 않으면 웹사이트가 여러분의 제출물을 평가할 수 없습니다.

에드모도에 제출할 것은 웹사이트에 제출한 것과 같은 RData file과 아래 질문들에 대한 답을 작성한 리포트를 제출하면 됩니다. 이번 숙제는 r script파일은 제출하지 않아도 되며, 대신 리포트에서 여러분이 어떻게 했는지에 대해서 조금 더 상세히 작성해주면 됩니다.

여러분에 웹사이트에 제출한 결과는 기록으로 남으며, 상위에 기록된 학생들은 추가로 혜택을 받을 수 있습니다.

이번 숙제는 기말고사와 매우 유사한 형식이므로 여러분이 기말고사 공부를 하는데 좋은 연습이 될 것입니다.

이번 숙제 만드느라 꼬박 이틀을 썼는데, 즐겁고 유익한 경험이 되면 좋겠네요.

In the RData file that you submit to Edmodo, you only need to include the prediction result of testdata of your best models in each question.

RData 파일에는 여러분의 가장 성능이 좋은 최종 모델의 예측결과만 포함해서 에드모도에 제출하면 됩니다.

Question 1

For the first question, you will be working with dataset containing student information who took a class of Math and Portuguese. The dataset contains following variables:

(Korean Translation) 첫번째 문제에서는 수학과 포루투갈어 수업을 수강한 학생들의 정보를 담고 있는 데이터를 사용할 것입니다. 데이터에 포함된 변수는 아래와 같습니다.

Variables

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2 sex - student's sex (binary: "F" - female or "M" - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: "U" - urban or "R" - rural)

5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

12 guardian - student's guardian (nominal: "mother", "father" or "other")

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

31 class – course subject: Math or Portuguese

these grades are related with the course subject, Math or Portuguese:

32 G3 - final grade (numeric: from 0 to 20, outcome variable)

1. Build a linear regression model with lm function to predict the final grade of student (**G3**) using all variables in the dataset. Describe the process, including data preparation if necessary, to obtain your model.

학생의 최종 성적을 예측하는 선형 회귀 모델을 만들어라. (주어진 모든 변수를 사용). 모델을 만드는 과정을 설명하고, 필요하다면 전처리도 수행하고 전처리 과정도 설명하여라.

2. What is the RMSE and R^2 of your model for both training and test dataset?

3. Interpret the linear model you got in Q1. Explain what are the variables that affect the Final Grade G3, positively or negatively. You do not have to explain every variable's influence, but only variables that you think is significant for the model.

앞서 문제에서 얻은 선형 회귀모델을 해석해보시오. 최종성적에 긍정적인 영향을 주는 변수와 부정적인 변수를 주는 변수는 무엇인가요? 모든 변수의 영향력을 다 설명할 필요는 없고, 모델에서 성적에 상당한 영향을 끼친다고 생각되는 변수만 설명하면 됩니다.

4. Try to add new features (input variables) to the linear model in order to improve the model's performance. How does it affect your model's performance? Find RMSE and R^2 of your improved model. Is your model improved or degraded? Try to infer the reason of improvement or degradation.

새로운 변수를 추가해서 모델의 성능을 향상해보시오? 변수를 추가하는 것이 모델의 성능에 어떠한 영향을 주는지 설명하시오. 향상된 모델의 RMSE와 R^2는 무엇인가요? 성능이 향상되나요? 아니면 저하되나요? 그 원인을 추론해봅시다.

5. Try to remove some variables that you might think unnecessary or irrelevant to the final grade from the model. You can choose to modify the model either from Q1 or Q4. Find RMSE and R^2 of your new model. Is your model improved or degraded? Try to infer the reason of improvement or degradation.

필요 없거나 성적과 관련 없다고 생각되는 변수들을 모델로부터 제외해봅시다. 1번문제 혹은 4번문제의 모델 중에서 원하는 것을 수정하면 됩니다. 새 모델의 RMSE와 R^2는 얼마인가요? 모델의 성능이 향상되었나요? 저하되었나요? 그 원인은 무엇인가요?

6. Try to build the linear regression model with best RMSE and R^2. and make prediction for the students in the test dataset. Make a numeric vector of your prediction result named "**pred_grade_test**" and include the vector object in the RData file. "**pred_grade_test**" is a numeric vector that contains 183 predicted grade for the student in the test dataset. Report your final RMSE and R^2 through the website. The students who are in high rank will get some benefit.

https://hbchoi.shinyapps.io/datascience_hw5/

(You may add new features or remove irrelevant features as you did in Q4 and Q5 to get the best R^2 and RMSE. Or you may try other approach for improvement.)

가장 좋은 RMSE와 R^2를 갖는 선형 회귀 모델을 만들어 테스트 데이터 학생들의 최종성적을 예측

하시오. 최종 모델의 예측결과를 **pred_grade_test**라는 이름의 벡터로 만들어 RData에 저장한 후 웹사이트에 업로드하여 자신의 최종 모델 성능을 등록하시오. 상위에 등록된 학생들은 상응하는 혜택에 있을 예정입니다.

https://hbchoi.shinyapps.io/datascience_hw5/

(성능 향상을 위해 변수를 추가하거나 제거하는 방법을 사용할 수도 있고, 또 다른 방법을 사용할 수도 있습니다.)

Question 2

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.

variable name	description
default.payment.next.month*	채무불이행 여부, 1=채무불이행 0=채무불이행 아님 default payment, 1 = default, 0 = 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	교육수준 1=대학원 2=대학 3=고등학교 4=그외 Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
MARRIAGE	결혼여부 1=기혼 2=미혼 3=기타 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: payment of last month Sept. PAY_2: payment of the month before the last month August. … -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. 1달전부터 6달전까지 상환내역 -1=제때 상환, 1=한달연체 2=두달연체 3=세달연체, …… , 9=9달연체 혹은 그 이상
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

* outcome variable

** currency is NT dollar. (단위는 타이완 달러)

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.

2. What is the AUC of your model for both test and training dataset?

3. Set the threshold (i.e. cutoff) to 0.5 what are the accuracy, precision and recall? Explain how the difference in cost of false positive and false negative. According to the cost comparison, how do you adjust the threshold to obtain "better" precision and recall? what are the new accuracy, precision and recall? (If you think 0.5 is good enough, explain why)

* this question is for the training dataset only.

threshold를 0.5로 정하고 accuracy, precision, recall을 계산해보아라. false positive와 false negative의 비용이 어떻게 다른지 설명하라. false positive와 false negative의 비용을 생각했을 때 threshold를 어떻게 조정하면 좋을지 설명하고, 새로운 accuracy, precision과 recall을 계산하라.

* 이 문제는 학습데이터에 대해서만 수행하면 됩니다.

4. Find the logistic regression model with the best AUC for the test dataset. You may add some new features and remove irrelevant features to find the best one. Make a numeric vector containing 5,000 predicted probability of default payment named **prob_default_test** and include the vector in

the RData file. Report your prediction result to the website and check your rank. Highly ranked students will get some benefits.

Describe the process to get the best logistic model in your report as well.

가장 높은 AUC 값을 가지는 최선의 logistic regression 모델을 찾으라. 변수를 추가하거나 제거하면서 시도해볼 수 있을 것이다. 모델을 이용한 예측 확률 값을 길이 5000의 vector로 저장해서 RData에 포함시켜라. vector의 이름은 **prob_default_test**으로 하라. 웹사이트에 RData를 업로드해서 AUC 기록을 등록하라. 상위 랭크된 학생들은 혜택이 있을 것이다.

5. Suggest the proper threshold of the best logistic model from Q4. Make binary classification (Yes/No) for the test dataset based on the threshold. Make a logical vector of length 5,000 of prediction named **pred_default_test** included in RData. The website will find the accuracy, precision, and recall for you. What are they?

4번에서 찾은 모델에 적당한 threshold를 적용하여 테스트 데이터에 대해 True/False 예측을 수행하여라. 예측 결과를 길이 5000의 TRUE/FALSE logical 벡터(**pred_default_test**)로 만들어 RData에 저장하라. RData를 웹사이트에 업로드하면 accuracy, precision, recall을 계산하여 줄 것이다. 어떤 값을 얻었는가?