Data Science

Homework 3

Perform following tasks and submit your work in html, word, or pdf generated by RMarkdown. (Chance to practice RMarkdown). The datasets you need for this homework are in **"PRSA_data.csv"** and "**bankcruptcy.RData**".

여러분은 아래 주어진 task들을 수행하여 그 과정과 결과 및 설명을 담은 **보고서**를 RMarkdwon으로 작성하여 html, word, 또는 pdf 형식으로 제출하세요. 과제를 수행하기 위해 필요한 데이터는 각각 **"PRSA_data.csv"**와 **bankcruptcy.RData** 파일에 포함되어있습니다.

Single Variable Model for Regression

PRSA is an hourly data set contains the PM2.5 data of US Embassy in Beijing. Meanwhile, meteorological data from Beijing Capital International Airport are also included.

PRSA 데이터는 2010.1.1 ~ 2014.12.31 기간 동안 중국 베이징의 미세먼지 농도 및 날씨 관련 정보를 기록한 데이터이다. 변수의 설명은 아래와 같다.

No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: hour of data in this row

pm2.5: PM2.5 concentration (ug/m^3) ---- *목적 변수(outcome variable)

DEWP: Dew Point

TEMP: Temperature

PRES: Pressure (hPa)

cbwd: Combined wind direction

Iws: Cumulated wind speed (m/s)

Is: Cumulated hours of snow

Ir: Cumulated hours of rain

- Let us use the dataset from 2010 to 2013 to train prediction model. To test the model, use subset of data in year of 2014. (remove observation with missing values in PM2.5 Variable)

- 2010년부터 2013년까지의 데이터를 학습 데이터(train data), 2014년데이터를 테스트 데이터로 사용하자. (PM 2.5에 NA가 존재하는 경우 해당 행은 삭제하고 사용하자)

1. Perform data exploration on training dataset (2010 – 2013). Describe **at least 10 findings** that you have discovered from data exploration.

   (findings here do not mean that simple the number of rows or columns, but it means likes correlation of two variables or distribution of certain variables. for example PM2.5 seems to be more related to certain Month or season. How temperature is related to the PM2.5.)

   For data exploration purpose, try **at least 3** visualization of dataset.

   학습데이터에 대해서 data exploration을 수행하시오. exploration을 통해 알 수 있는 사실을 10가지 이상 기술하시오.

   (이 때, 발견한 사실은 행의 개수, 열의 개수와 같은 단순한 정보가 아닌, 변수들 사이의 상관관계와 같은 것입니다. 예를 들어 월별 미세먼지 농도와 가장 미세먼지가 높은 달은 언제인지, 기온이 높을수록 미세먼지가 높아지는지 등.)

   data exploration을 위해서 시각화는 최소 3가지 이상 수행하시오.

2. Find the best single-variable prediction model in terms of RMSE that predicts PM2.5. As outcome variable (PM2.5) is continuous it would be regression model and the model should pick one variable as predictor. Include the process how you find the best model in the report. (You possibly use a new variable as predictor derived from other variables to improve the performance.)

   학습데이터를 사용하여 미세먼지 수치를 가장 잘 예측하는 (단일 변수 모델) single variable model(Best Model)을 찾으시오(**RMSE 기준**). 목적 변수가 연속형이므로 regression model이고 단일 변수 모델이므로 하나의 입력 변수만을 모델에 사용해야 합니다. 가장 성능이 좋은 단일 변수 모델을 찾아가는 과정을 보고서에 함께 기술하시오. (최선의 모델을 만들기 위해서 기존의 변수 외에 새로운 변수를 추가하는 것도 가능합니다.)

3. What are the train data RMSE and R square for the model you found in question 2.

   2번에서 찾은 모델의 Train 데이터에서의 RMSE와 R square 값을 계산하시오.

4. Verify your best model with test dataset. What are the RMSE and R square for the test dataset?

   2번에서 찾은 모델을 테스트 데이터를 이용해 검증하시오. 테스트 RMSE와 R square의 값은 어떻게 계산되나요?

5. Do you think that your best model is overfitting or not? Explain the reason why you think so.

   찾은 모델은 overfitting(과적합)이라고 할 수 있나요? 할 수 있거나 혹은 할 수 없거나 왜 그런지 이유를 설명하시오.

6. Compare your result of test data in question 5 with at least one of your classmate. If his number is better than yours, find the reason why it is better and improve yours. (include this process in the report as well) If you do not have one to compare the result who are taking this class, ask TA to introduce someone.

   여러분이 찾은 모델의 성능을 Test 데이터 기준으로 최소 한명 이상의 수업 듣는 친구와 비교하시오. 여러분의 모델이 동료의 모델의 성능보다 낮다면 왜 그런지 생각해보고, 여러분의 모델을 개선하시오. 비교 과정/결과와 개선 방법/개선 결과를 보고서에 기술하시오. 혹시 성능을 비교할 친구가 없다면 TA에게 요청해서 소개를 받도록 합시다.

7. Draw a scatter plot that represent outcome variable (Actual value of PM2.5) on y-axis and predicted value (predicted PM2.5) on x-axis. Describe what you can infer from that chart.

   학습 데이터와 테스트 데이터에서, 각각 y축에 목적변수, x축에 목적변수에 대한 모델의 예측 값을 그린 산점도(scatter plot)을 그리시오. 그래프로부터 알 수 있는 것이 있다면 설명해 보시오.

Single Variable Model for Classification

"bankruptcy_train", "bankruptcy_test" data.frame represent various attributes of 250 business companies (for train 200, for test 50) to predict bankruptcy of the business. Attributes are given as follows.

"bankruptcy_train", "bankruptcy_test" data.frame은 기업의 도산(bankruptcy or non- bankruptcy) 여부를 예측하기 위해서, 250개(train = 200, test = 50) 기업의 여러 지표들을 표현한 데이터이다. 표현된 기업의 지표들은 아래와 같다.

1 Industry risk (IR) :

Government policies and International agreements,

Cyclicality,

Degree of competition,

The price and stability of market supply,

The size and growth of market demand,

The sensitivity to changes in macroeconomic factors,

Domestic and international competitive power,

Product Life Cycle.

2 Management risk(MR):

Ability and competence of management,

Stability of management,

The relationship between management/ owner,

Human resources management,

Growth process/business performance,

Short and long term business planning,

achievement and feasibility.

3 Financial Flexibility(FF):

Direct financing,

Indirect financing,

Other financing

4 Credibility (CR):

Credit history,

reliability of information,

The relationship with financial institutes.

5 Competitiveness (CO):

Market position,

The level of core capacities,

Differentiated strategy,

6 Operating Risk (OP):

The stability and diversity of procurement,

The stability of transaction,

The efficiency of production,

The prospects for demand for product and service,

Sales diversification,

Sales price and settlement condition,

Collection of A/R,

Effectiveness of sale network.

7. Class: Non-Bankruptcy / Bankruptcy **(* 목적변수 outcome variable)**

1. Using the training dataset **bankruptcy_train,** find the best single variable model in terms of AUC. Since outcome variable is discrete, it will be classification model and model must use one variable as predictor. Include the description of process how to obtain the best model. (You possibly use a new variable as predictor derived from other variables to improve the performance.) What variable did you use as predictor? and what was AUC of your model?

   학습데이터를 사용하여 목적 변수를 가장 잘 예측하는 (단일 변수 모델) single variable

model을 찾으시오 (**AUC 기준**). 목적변수가 범주형이므로 classification model이며 단일 변수 모델이므로 하나의 입력 변수(지표)만을 모델에 사용해야 합니다. 가장 성능이 좋은 모델을 찾아가는 과정을 보고서에 함께 기술하시오. (최선의 모델을 만들기 위해서 기존의 변수 외에 새로운 변수를 추가하는 것도 가능합니다.) 어떤 변수를 사용하였나요? AUC는 얼마가 나오나요?

2. Find AUC of your best model found in question 1 for the test dataset.

   1번에서 찾은 모델을 **test** data frame에 대해서 test해서 AUC를 계산하시오.

3. Do you think that your best model is overfitting or not? Explain the reason why you think so.

   찾은 모델은 overfitting(과적합)이라고 할 수 있나요? 할 수 있거나 혹은 할 수 없거나 왜 그런지 이유를 설명하시오.

4. Draw a graph that shows how precision and recall changes over threshold in your best model. Based on you graph, pick up the best threshold and explain the reason why it is the best threshold (for both training and test dataset).

   Best model에서 Threshold를 변화시킬 때, precision과 recall 값이 어떻게 변화하는지 그래프로 표현하시오. 그래프를 바탕으로 어떤 threshold을 선택하는 것이 좋을지 여러분의 의견을 이유와 함께 제시하시오. (학습 데이터 테스트 데이터 각각에 대해)

5. Classification model typically trades off precision against recall. To compensate this issue, we sometimes use another measure call **F1 measure** that is combined with precision and recall as in the equation below. What is threshold that maximize F1 measure in your best model. (for both training and test dataset).

   Trade-off 관계에 있는 precision과 recall을 하나의 measure로 보기 위해서 F_1 Measure라는 것을 사용하기도 합니다. F_1 Measure를 계산하는 수식은 아래와 같습니다. F_1 값이 가장 크게 되는 threshold는 얼마인가요? (학습 데이터 테스트 데이터 각각에 대해)

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

6. Draw ROC curve of your best model for both for both training and test dataset. Describe what you can infer from the ROC curve.

   학습데이터와 테스트 데이터 각각에 대해서 여러분 모델의 ROC 커브를 그려보시오. 그래프를 통해서 알 수 있는 점이 있다면 설명해보시오.

7. What is threshold that maximize **accuracy** in your best model. (for both training and test dataset).

Accuracy가 가장 크게 되는 threshold는 얼마인가요? (학습 데이터 테스트 데이터 각각에 대해)

8. Compare the test result of your best model with at least one of your classmate. If his/her number is better than yours, find the reason why it is better and improve yours. (include this process in the report as well) If you do not have one to compare the result who are taking this class, ask TA to introduce someone.

   여러분이 찾은 모델의 성능을 Test 데이터 기준으로 최소 한 명 이상의 수업 같이 듣는 친구와 비교하시오. 여러분의 모델이 동료의 모델의 성능보다 낮다면 왜 그런지 생각해보고, 여러분의 모델을 개선하시오. 비교 과정/결과와 개선 방법/개선 결과를 보고서에 기술하시오.