STAT 8730 Machine Learning Competition

Predicting the supercomputer usage for NSF awarded projects

Some NSF awarded projects need intensive computation in their research. Per request, the supercomputer resources over the country can be allocated to those projects. However, not every project made fully usage of their allocation. The overarching goal of this machine learning contest is to predict the supercomputer usage rate for each NSF projects with allocation. If we can make an accurate prediction on the usage rate, then better allocation and services can be provided to those projects.

Description of variables

There are 113 variables in this data set. They are introduced below in groups.

- Grant_Number: ID for the NSF awarded projects.
- StartDate_award, EndDate_award, StartYear_award, EndYear_award, TotalDays_award: Date/year and duration of the project.
- StartDate_usage, EndDate_usage, StartYear_usage, EndYear_usage, TotalDays_usage: Date/year and duration of the supercomputer usage.
- PI_id: ID for the principle investigators (PI).
- PI_New: Whether the PI used the supercomputer resources for the first time.
- PI_PrevUsage: Total amount of usage for any previous projects by the same PI.
- PI_PrevUsageRate: Median usage rate of previous projects by the same PI.
- Organization: Name of the PI's organization. The organization can be a university, college, institution, laboratory, governmental department, NGO, company, etc. inside and outside the U.S.
- Org_PrevFoS: Number of unique fields of science in this organization prior to this project.
- Org_New: Whether the Organization used the supercomputer resources for the first time.
- Org_PrevUsage: Total amount of usage for any previous projects by the same organization.
- Org_PrevUsageRate: Median usage rate of previous projects by the same organization.
- Carnegie: Carnegie classification (https://carnegieclassifications.iu.edu) if the organization is a college or university in the U.S. The detailed classification levels can be found in the `Labels` sheet of the Excel document (https://carnegieclassifications.iu.edu/downloads/CCIHE2018-PublicData.xlsx). The "2018 Carnegie Basic Classification" is used for this data. `-9` is created for organizations that are not included by the Carnegie classification.
- Carnegie15: Whether the Carnegie classification is Level 15 (Doctoral Universities: Very High Research Activity).
- Carnegie151617: Whether the Carnegie classification is in Levels 15 (Doctoral Universities: Very High Research Activity), 16 (Doctoral Universities: High Research Activity), or 17 (Doctoral/Professional Universities).
- ChampionsBefore: Whether the same organization had any campus champions (https://www.xsede.org/community-engagement/campus-champions) at least two years before the start year of the awarded project.
- ChampionsNow: Whether the same organization had any campus champions in the previous year or the current year of the project start year.
- isChampion: Whether the awarded project is a campus champion.

- FoS_AdvSciComp, . . . , FoS_HumanArts: Whether the awarded project belonged to the corresponding field of science. Note that one project can belong to multiple fields of science.
- Type_Startup, . . . , Type_DAC: Whether the awarded project had the corresponding type of work. Note that one project can have multiple types of work.
- Resource_anl, . . . , Resource_roam_T: Total amount of initial allocation that the awarded project received from the corresponding supercomputer resource. Note that one project can use multiple resources.
- Transaction_new_award, Transaction_renewal_award: Whether the awarded project is a new project or renewal project. Note that one project can be both new and renewal due to different status of the project.
- Transaction_new_usage, . . . , Transaction_supplement: Total amount of initial allocation for a project from a certain transaction type of the usage. Note that one project can have multiple usage transaction types.
- count_Project, . . . , count_ProjectTitle: Count of unique values from the corresponding columns when grouping the data by the grant number.
- Initial_Allocation: Total initial allocation for a project.
- Final_Allocation: Total final allocation for a project.
- Used_Amount: Total used amount for a project.
- UsageRate: Created by `Used_Amount/Final_Allocation`. If `Used_Amount` $= 0$, then `UsageRate` $= 0$. If `Used_Amount` > 0 but `Final_Allocation` $= 0$, then `UsageRate` $= 1$. This column is truncated by 6, i.e., if `UsageRate` > 6 then we assign `UsageRate` by 6.

Overview of data

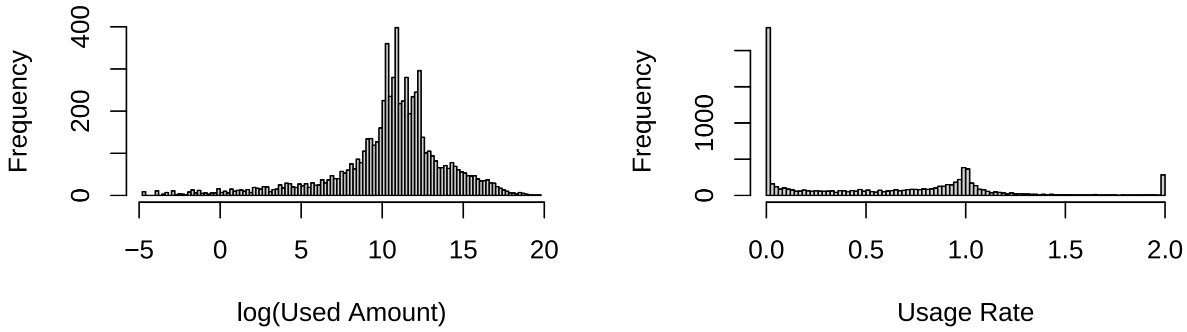
Here is a quick look at the training set. The training data contains 8466 funded projects between 2003 and 2015.

```r
train = read.csv("data/train.csv",stringsAsFactors=FALSE)
dim(train)
```

```
## [1] 8466  113
```

```r
train$StartDate_award = as.Date(train$StartDate_award, format="%Y-%m-%d")
summary(train$StartDate_award)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max. 
## "2003-06-30" "2007-11-06" "2010-10-01" "2010-08-20" "2013-06-28" "2015-12-28"
```



Now let us look at the test set. The test data contains 2770 projects between 2016 - 2019. Comparing to the training data, the test set does not include the two response variables `Used_Amount` or `UsageRate`.

```
test = read.csv("data/test.csv",stringsAsFactors=FALSE)
dim(test)
```

```
## [1] 2770  111
```

```
test$StartDate_award = as.Date(test$StartDate_award, format="%Y-%m-%d")
summary(test$StartDate_award)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2016-01-01" "2016-09-02" "2017-05-18" "2017-06-04" "2018-03-12" "2019-11-22"
```

Task (Individual Work)

1. Create the most accurate regression model that you can for the data, as measured by the **root mean squared error (RMSE)** on the test data.
2. Write the slides (up to 15 pages) to summarize your approach of
   (a) exploring data
   (b) formulating the model design matrix via variable creation & selection,
   (c) building models, tuning parameters, validating models,
   (d) results from all attempts,
   (e) your findings from the data.

Kaggle website

The data can be downloaded from the Kaggle site and you will have to participate the competition through the link will launch soon as it is not open to public.

Notes:

- The maximum daily submission number is 15. So you will need to wait until the next UTC day after submitting 15 results.
- Please use your real name as the "team" name when making submissions. You will work individually so you are the only member of your team.

Format of submission

Your submission file should be in the csv format with two columns: `Grant_Number` and `UsageRate`. Note that the `Grant_Number` column must be exactly same as in the `sample_submission.csv` file. The `UsageRate` column should be replaced by your prediction within $[0, 2]$. Example of the submission:

```
Grant_Number,UsageRate
ASC160034,0.143811
ASC160042,0.223481
MCB150133,0.883532
...
CIE160015,0.951223
```

Deadlines

- December 5 (11:59 pm): Final prediction submission on Kaggle.
- December 6 (4 pm): Presentations. Each presentation should be around 5 minutes.
- December 6 (11:59 pm): Slides and code submission.

Grading

- Total points: 20 (+1)
  - Accuracy of the model: 10
    - Your score will be based on your rank in class.
  - Progress made from multiple submissions: 2
    - Improvement of accuracy: 1
    - Number of submissions with progress: 1
  - Slides: 7 (+1)
    - Data exploration: 1
    - Variable creation/selection: 1
    - Validation approach: 1
    - Method selection: 1
    - Parameter tuning: 1
    - Result table for all models & parameters: 1
    - Findings & summary: 1
    - Special analysis via methods from class (+1 extra credit)
  - Met the deadlines: 1 (Late submissions on Kaggle are not accepted.)