CISC 5352 Financial Data Analytics Project (3) [1]

[1] Each group must choose at least one project.

A: OptionDataWebGleaner2 (100 points)

OptionDataWebGleaner from Yahoo/Google Finance or Nasdaq.com

- Write a python program called OptionDataWebGleaner or optionCrawl to download/crawl all option data from Yahoo, Google Finance, Nasdaq to accomplish AT LEAST the following functionalities

  1. Your program input should be all tickers in Yahoo,/Google Finance or Nasdaq. The file yahoo_ticker_list.csv provides all ticker lists from Yahoo finance in July 2015. You are welcome to update it to current ticker list.
  2. It should provide an interface to allow users to pick any tikers to download.
  3. It can randomly download select 2000 tickers and write the results in an excel/csv file. It further has a function to remove missing data (incomplete option data) and outliers (you can view option data with very large implied volatility (e.g. >2) as outliers. The final "clean" data are written in a csv file.
  4. It provides visualization and data summary for the clean data. Please make your data summary as much as you can
  5. Separate option data you crawled into European options and non-European options and write them into different csv files. You need to allow users to choose cutoff, which is the distance between the calculated implied volatility and true implied volatility, to collect European options, say 0.01 or 0.001
  6. Provides graphics interface so that user can select a certain number of Auto, IT, Bank, Retail and fashion stock option he/she wants.
  7. Other functionalities you want to add

B) Endowment Analytics: (150 points)[2]

Basic data information

- The dataset in the file: *endowment_2018.csv* consists of a total of 232 institutions
  - It includes the top-100 richest universities and 132 institutions with at least $50 million in endowments.
- There are nine variables
  - high school GPA, average SAT scores, and graduation and acceptance ratios (*academic characteristic variables)*
  - public/private, enrollments (undergraduate enrollment): *(Institutional characteristic variables)*
  - endowment, tuition, and loan rate (*Financial characteristic variables)*
- Money used is the label for these universities: 0: poorly used, 1: fairly used, 2 well-used.

Visualizations

- **Do visualizations so that to at least dig knowledge or disclose the following relationships** (you can do more visualizations to do more explorations)

  If private school students have lower loan rates than public school students?

  If private schools will get more endowments than public schools

  If schools with more enrollments, would it get more endowments?

  If school with low acceptance ratios, it will get more endowments?

[2] More credits but more work

PCA analysis

- **Do PCA analysis to generate biplot as follows and interpret its meaning and outliers** (Note: you don't need to use the label info when doing such a PCA)
  - You need to create another dataset called endowments_2018a.csv where all universities names are replaced by their initials such as using NYU to represent New York University
  - Your university initial must use (you can crawl this file to get all universities' initials)
    - https://en.wikipedia.org/wiki/List_of_colloquial_names_for_universities_and_colleges_in_
  - You need to use initial name instead of the full name for each institution in your PCA biplot.
  - Generate tri-plot for the dataset.

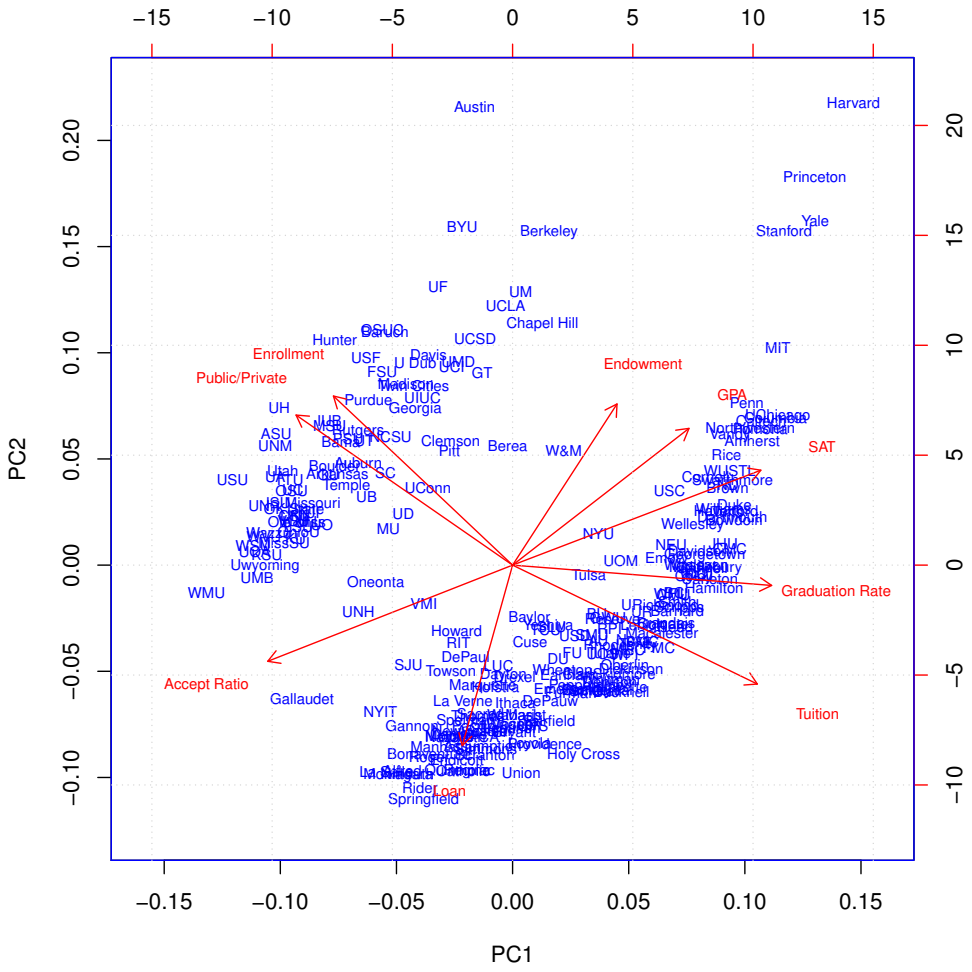

Fig 1. Biplot of 232 institutions

Twitter analytics

- Use python twitter API to collect the number of followers for the 232 rich universities in the file endowments_2018.csv and updated the file as endowments_ 2018_2.csv to include the the number of followers in twitter.

- Verify the following hypotheses via your analytics:

  The richer a university, the more followers in its twitter

Private universities have more followers than its public peers

Visualize the data via PCA by including info you get from twitter

Machine learning and selective learning

use data in train.csv and test.csv as training and test data for the following learning machines

1. Extremely randomized trees (ET)
2. Random forests
3. SVM
4. Gradient boosting
5. Deep neural networks (DNN): a neural network with at least two hidden layers (you #neurons in each hidden layer should be at least 200)

- Compute d-index, sensitivity, specificity, and accuracy, F-1 values for them.
- Use selective learning at least for Extremely randomized trees (ET), Random forests, and SVM (Note: this is a little bit tricky because it is a classification problem! HINT: Convert it as a regression one)

Learning validation

- Use Extremely randomized trees (ET) to classify the validation dataset and generate a csv file validation_ans.csv where the last column should have the predicted label for each institute
- Draw your conclusion according to knowledge you get: which kind of universities are more likely to misuse their endowment? Private or public?
- Use selective learning to this dataset and create validation_ans_sure.csv file to include all predicted. Draw your conclusion again.
- **Extra credits (50 points)**: you can crawl usnews to get more validation data to test and support your conclusion

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at least 40 pages)
  - your data
  - source files
  - corresponding related output.
  - A link of your group's presentation video
- 2. Please name your folder last_name1_last-name2_CISC5352_project_3. For example, Brown_Smith_CISC5352_project_3 if your group members with last names: Brown and Smith.
- 3. Send the zipped file (.zip instead of ,rar) of your folder to Blackboard before 11:59 pm Dec 12, 2018