

Project 2: Predictive Modeling (Classification)

Assigned: 9/22/2016
Due: 10/20/2016 (via Canvas)
Points: 100

Please submit your report in **PDF format**.



We will work again with the data set from Project 1. Questions that we would like to answer are:

- Can we predict which university will improve over time in the ranking and which university will fall back?
- What are the important variables?

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Submit your R code (if necessary also a description of how you used other tools) in a separate file.

Follow the CRISP-DM framework

Steps 1 and 2 have already been performed in Project 1.

3. Data Preparation [35]

- Define and prepare your class variables used for the different questions. You may decide to discretize or aggregate (i.e., combine values) for your class and/or other features. [10]
- Select features that might be useful for modeling. Create features if necessary (e.g., transformation to rates, time differences, etc.). Include additional data from other rankings, external data, etc. [20]
- Describe the final dataset that is used for classification (include the scale/range of new features) [5]

4. Modeling [50 points]

- Create at least 5 different classification models (e.g., use different techniques, different parameters, different class variables, etc.). [20]
- Discuss the advantages of each model for this classification task. [5]
- What are the most important features found by each model. Are they the same. Discuss what this means. [5]
- Assess how well each model performs (use training/test data, cross validation, etc. as appropriate). [20]

5. Evaluation and Deployment [5 points]

- How useful is your model for the stake holders (e.g., students, university administrators)? How could stake holders act on the model. [5]

Exceptional Work [10 points]