HUDK5053: Feature Engineering Studio
Spring 2015
Professor Ryan Shaun Baker

SYLLABUS

Instructor Info
Phone: 212-678-8329
Email: baker2@exchange.tc.columbia.edu
Office: Grace Dodge Hall 464
Office hours: 3pm-6pm Wednesday, or by appointment

Number of points: 3

Required Texts:

- Kelley, T. (2001) *The Art of Innovation: Lessons in Creativity from IDEO, America's Leading Design Firm.*

Information on how to obtain course readings will be provided in class.

Course Goals: This course is a design studio-style course teaching how to distill and engineer features for data mining. We will cover the process of feature engineering and distillation, including brainstorming features, deciding what features to create, and criteria for selecting features. Students will learn how to create features in Excel, Java, Google Refine, the EDM Workbench, and other relevant tools. Students will learn skills for brainstorming and for brainstorming preparation.

Course Pre-requisites: HUDK4050 Core Methods in Educational Data Mining, or instructor permission

Assignments:

All assignments will be due three hours before class. The student may miss three assignments without penalty, except for the Final Project Presentation, which cannot be missed (and which counts extra). Please note that students **cannot** do additional assignments, and then take the top grades from the assignments they completed. Only the first nine assignments handed in will be graded. In special circumstances, for instance accommodation of special needs, alternate assignments may be provided by mutual agreement of the instructor and student; however, provision of alternate assignments cannot occur after assignment hand-in, and is solely at the instructor's discretion.

No extensions will be granted, except in case of instructor error or extreme circumstances (assignments in other classes, research studies, and so on do not count as extreme circumstances; serious injury, illness, or death in the family do count as extreme circumstances). Outside of these circumstances, late hand-ins will not be accepted (e.g. zero credit will be given). Students must be prepared to present every week's assignment in class. Assignments can involve either a data set of the instructor's choice, or a data set of the student's choice (with approval from the instructor).

Beyond presenting their own work, students are required to regularly participate in critique and discussion of other students' work, in general class discussions, and other classroom activities, as part of their grade. Attendance is not sufficient for a high grade for class participation.

Grading

Regular Assignments (9)	6% each = 54% total, plus 1% = 55%
Final Project Presentation	30%
Class Participation	15%

Course Schedule

Feature Engineering Studio
Professor Ryan S. Baker

Introduction
Wednesday, January 21

Readings

- None

Lab Session: Finding a Data Set
Monday, January 26

Readings

- None

Problem Proposal
Monday, February 2

Readings

- None

Assignment: 1. Problem Proposal

Lab Session: Using RapidMiner
Wednesday, February 4

Readings

- None

No Class
Monday, February 9

No Class
Wednesday, February 11

Data Cleaning
Monday, February 16

Readings

- Romero, C., Romero, J.R., Ventura, S. (2013) A Survey on Pre-Processing Educational Data. In Ayala, A.P. (2013) *Educational Data Mining: Applications and Trends.* Ch.2, pp. 29-64.

Assignment: 2. Data Cleaning

Feature Distillation in Excel
Monday, February 23

Readings

- Online Excel Pivot Table Tutorials
- Online Excel Vlookup Table Tutorials

Assignment: 3. Data Familiarization

Advanced Feature Distillation in Excel
Monday, March 2

Readings

- Online Excel Equation Solver Tutorials

Assignment: 4. Feature Engineering 1

Lab Session: More Advanced Feature Distillation in Excel
Wednesday, March 4

Readings

- None

Google Refine
Wednesday, March 11

Readings

- Google Refine User Guide

Assignment: 5. Feature Engineering 2

Lab Session: RapidMiner Practice Session
Monday, March 23

Readings

- None

Feature Reuse
Monday, March 30

Readings

- Rodrigo, M.M.T., Baker, R.S.J.d., McLaren, B., Jayme, A., Dy, T. (2012) Development of a Workbench to Address the Educational Data Mining Bottleneck. *Proceedings of the 5th International Conference on Educational Data Mining*, 152-155.

Assignment: 6. Feature Reuse

Lab Session: Building Prediction Models
Wednesday, April 1

Readings

- None

Brainstorming
Monday, April 6

Readings

- Kelley, T. (2001) *The Art of Innovation: Lessons in Creativity from IDEO, America's Leading Design Firm.*

Assignment: 7. Brainstorming

Construct Validity in Feature Selection
Monday, April 13

Readings

- Sao Pedro, M., Baker, R.S.J.d., Gobert, J. (2012) Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012), 249-260.

Assignment: 8. Construct Validity

Iterative Feature Refinement
Monday, April 20

Readings

- None

Assignment: 9. Iterative Feature Refinement

Collaboration in Feature Engineering
Wednesday, April 22

Readings

- Fischer, G. (2004) Social Creativity: Turning Barriers into Opportunities for Collaborative Design. Proceedings of the Participatory Design Conference (PDC'04), 152-161.
- Veeramachaneni, K., O'Reilly, U., Taylor, C. (2014) Towards feature engineering at scale for data from massive open online courses. arXiv preprint 1407.5238.

Assignment: 10. Problem Shift

Feature Adaptation
Wednesday, April 29

Readings

- Selected by each student

Assignment: 11. Feature Adaptation

Sustained Iteration
Wednesday, May 6

Readings

- None

Assignment: 12. Sustained Iteration

Final Project Presentations
Monday, May 11

Readings

- None

Assignment: 13. Final Project Presentation

UNIVERSITY POLICIES

1. All examinations, papers, and other graded work and assignments are to be completed in conformance with TCs Academic Integrity Policy (http://www.tc.columbia.edu/administration/diversity/index.asp?Id=Civility+Resources+and+Policies&Info=Civility+Resources+and+Policies&Area=Student+Miscon duct+Policy). Students who intentionally submit work either not their own or without clear attribution to the original source, fabricate data or other information, engage in cheating, or misrepresentation of academic records may be subject to charges. Sanctions may include dismissal from the college for violation of the TC principles of academic and professional integrity fundamental to the purpose of the College.

2. The College will make reasonable accommodations for persons with documented disabilities. Students are encouraged to contact the Office of Access and Services for Individuals with Disabilities for information about registration (166 Thorndike Hall). Services are available only to students who are registered and submit appropriate documentation. As your instructor, I am happy to discuss specific needs with you as well.

3. The grade of Incomplete will be assigned only when the course attendance requirement has been met but, for reasons satisfactory to the instructor, the granting of a final grade has been postponed because certain course assignments are outstanding. If the outstanding assignments are completed within one calendar year from the date of the close of term in which the grade of Incomplete was received and a final grade submitted, the final grade will be recorded on the permanent transcript, replacing the grade of Incomplete, with a transcript notation indicating the date that the grade of Incomplete was replaced by a final grade. If the outstanding work is not completed within one calendar year from the date of the close of term in which the grade of Incomplete was received, the grade will remain as a permanent Incomplete on the transcript. In such instances, if the course is a required course or part of an approved program of study, students will be required to re-enroll in the course including repayment of all tuition and fee charges for the new registration and satisfactorily complete all course requirements. If the required course is not offered in subsequent terms, the student should speak with the faculty advisor or Program Coordinator about their options for fulfilling the degree requirement. Doctoral students with six or more credits with grades of Incomplete included on their program of study will not be allowed to sit for the certification exam.

4. Teachers College students have the responsibility for activating the Columbia University Network ID (UNI) and a free TC Gmail account. As official communications from the College – e.g., information on graduation, announcements of closing due to severe storm, flu epidemic, transportation disruption, etc. -- will be sent to the student's TC Gmail account, students are responsible for either reading email there, or, for utilizing the mail forwarding option to forward mail from their account to an email address which they will monitor.

5. It is the policy of Teachers College to respect its members' observance of their major religious holidays. Students should notify instructors at the beginning of the semester about their wishes to observe holidays on days when class sessions are scheduled. Where academic scheduling conflicts prove unavoidable, no student will be penalized for absence due to religious reasons, and alternative means will be sought for satisfying the academic

requirements involved. If a suitable arrangement cannot be worked out between the student and the instructor, students and instructors should consult the appropriate department chair or director. If an additional appeal is needed, it may be taken to the Provost.