Course number: Feature Engineering Studio
Semester 201X
Professor Ryan Shaun Joazeiro de Baker

SYLLABUS

Instructor Info
Phone: 212-678-8329
Email: baker2@exchange.tc.columbia.edu
Office: Grace Dodge Hall 290
Office hours: Wednesdays 1pm-4pm
Course time: Monday, 11am-12:40pm
Some special optional classes TBD Wednesday, 11am-12:40pm

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: Core Methods in Educational Data Mining (or instructor approval)

Assignments:

An assignment will be due every class session after the first class session, 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). 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 participate in critique and discussion of other students' work, in general class discussions, and other classroom activities, as part of their grade.

Grading

- Regular Assignments (12) 4% each = 48% total, plus 2% = 50%
- Final Project Presentation 30%
- Class Participation 20%

SCHOOL 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.

Course Schedule

Feature Engineering Studio
Professor Ryan S.J.d. Baker

Class 1: Introduction
Monday, September 9

Readings

- None

Special Session 1A: Finding a Data Set
Wednesday, September 11

Class 2: Problem Proposal
Monday, September 16

Readings

- None

Assignment: 1. Problem Proposal

Class 3: Feature Distillation in Excel
Monday, September 23

Readings

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

Assignment: 2. Data Familiarization ("Mucking Around")

Special Session 3A: Using RapidMiner
Wednesday, September 25

Class 4: Advanced Feature Distillation in Excel
Monday, September 30

Readings

- Online Excel Equation Solver Tutorials

Assignment: 3. Feature Engineering 1 ("Bring Me a Rock")

Special Session 4A: More Advanced Feature Distillation in Excel
Wednesday, October 2

Class 5: Advanced Feature Distillation with Excel Equation Solver, Google Refine
Monday, October 7

Readings

- Google Refine User Guide

Assignment: 4. Feature Engineering 2 ("Bring Me Another Rock")

Special Session 5A: RapidMiner Practice Session
Wednesday, October 9

Class 6: Iterative Feature Refinement
Monday, October 14

Readings

- None

Assignment: 5. Iterative Feature Refinement ("Keep Running!")

Class 7: Feature Adaptation
Monday, October 21

Readings

- Selected by each student

Assignment: 6. Feature Adaptation ("This One's For Nikolai Ivonavich Lobachevsky**")**

Special Session 7A: Building Prediction Models
Wednesday, October 23

Class 8: Feature Reuse
Monday, October 28

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: None

Class 9: External Critique

Monday, November 4

Readings

- None

Assignment: 7. Posters ("The Charrette")

Class 10: Construct Validity in Feature Selection

Monday, November 11

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 ("One Who Visions Must Be Steeped in Data")

Class 11: Brainstorming

Monday, November 18

Readings

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

Assignment: 9. Brainstorming ("Ideation")

Class 12: Collaboration in Feature Engineering

Monday, November 25

Readings

- Fischer, G. (2004) Social Creativity: Turning Barriers into Opportunities for Collaborative Design. Proceedings of the Participatory Design Conference (PDC'04), 152-161.

Assignment: 10. Problem Shift ("The Fresh Mind")

Class 13: Sustained Iteration Part 1

Monday, December 2

Readings

- None

Assignment: 11. Sustained Iteration 1 ("The Slog")

Class 14: Sustained Iteration Part 2

Monday, December 9

Readings

- None

Assignment: 12. Sustained Iteration 2 (“Son of Slog”)

Class 15: Final Project Presentations

Monday, December 16

Readings

- None

Assignment: Final Project Presentation (“Finally”)