**Syllabus for Fall 2021**

**MACHINE LEARNING FOR THE SOCIAL SCIENCES (GR5073)**

Time: Fri. 2:10pm-4:00pm

Location: 312 Math Building

Michael D. Parrott

[mp3675@columbia.edu](mailto:mp3675@columbia.edu)

Office Hours: IAB 509c, Friday, 4-5:30 (by appt. please email for time and Zoom link)

**Teaching Assistants:**

Garda Ramadhito, mr3536@columbia.edu,  
Jie Chen, jc5230@columbia.edu,  
Prajwal Seth, ps3190@columbia.edu,  
Han Lin, hl3199@columbia.edu

Office Hours: TBD (by appt. please email for time and Zoom link)

TA review sessions once a week (Time TBD)

**Course Goals**

Social scientists need to fully engage with machine learning approaches that are found in computer science, engineering, AI, tech and in industry. This course will provide a comprehensive overview of machine learning as it is applied in a number of domains. Every effort will be made to draw comparisons and contrasts between this machine learning approach and more traditional regression-based approaches in the social sciences. Emphasis will also be on opportunities to synthesize these two approaches. The basis of this course comes from the W4995 - Applied Machine Learning course taught by Andreas Mueller. The course will start with an introduction to Python, Jupyter Notebooks, and the scikit-learn package. After that, there will be some discussion of data exploration, visualization in matplotlib, preprocessing, feature engineering, variable imputation and feature selection. Supervised learning methods will be considered, including OLS models, linear models for classification, support vector machines, decision trees and random forests, and gradient boosting. Calibration, model evaluation and strategies for dealing with imbalanced datasets will be considered next. This will be followed by unsupervised techniques: PCA, clustering and cluster evaluation, and manifold learning. Lastly, we will consider neural networks and convolutional neural networks for image classification.

Prerequisites are basic probability and statistics, basic linear algebra and calculus. The course will use Python, and so if students have programed in at least one software language, that will make it easier to keep up with the assignments.

You will learn by doing. In class work will require the use of a laptop with the latest version of Python 3 using the Anaconda distribution or Google Colab. We will install this software, so there is no need to attempt this on your own prior to the first week of class. ***Please bring a laptop to each class.***

**Course Expectations**

*Format of the Class*:  The plan for the class is for it to operate in a “hyflex” classroom. This classroom will be set up with cameras and microphones, along with access to live Zoom. This will allow students to be in the classroom during class, and for remote students to have access to the class live and to be able to participate by asking questions and interacting with their peers. The number of students who will be present during each class will depend upon the actual specific room that we are provided with and the official rules and regulations governing social distancing within that room. As a first approximation, the hope would be that a live in person option will be available for students to come every week or every other week, depending upon the ultimate size of the class and classroom.

*Expectation of Regular Participation and Utilization of Courseworks*:  We will be monitoring student participation and completion of assignments throughout the semester.  We want to make sure that students are consistently engaged, and if that becomes difficult, that students alert us to their situations.  There will also be opportunities for asynchronous participation too, where students will respond to a prompt/question either via video or text chat/memos.

Exams. We will have two take home exams that will ask you to apply what you have learned in lectures and homework assignments.

Homework Assignments. Students will have four homework assignments due throughout the semester. They will be based on writing up the results of performing the commands learned during the lectures. Specific instructions, format and deadlines will be given as the semester progresses.

Plagiarism and Academic Dishonesty: Students must do all their work within the boundaries of acceptable academic norms. See the Academic Honesty page of the CU website regarding college policy on plagiarism and other forms of academic dishonesty - <http://www.columbia.edu/cu/history/ugrad/main/handbook/academic_honesty.html>. Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary action, which may include reduction of grade, a failure in the course, suspension or expulsion.

Late Assignments. Students will lose points for handing in late assignments, at the discretion of the instructor and teaching assistant.

Textbooks. The following books will help you further your understanding of the material:

* Müller, Guido: Introduction to machine learning with python (IMLP) (available **for free** for Columbia Students via [Columbia](http://proquest.safaribooksonline.com.ezproxy.cul.columbia.edu/book/programming/machine-learning/9781449369880) library website)
* Kuhn, Johnson: Applied predictive modeling (APM) (available **for free** [via Columbia library website)](http://link.springer.com.ezproxy.cul.columbia.edu/book/10.1007%2F978-1-4614-6849-3)
* Tibshibani, Hastie, Friedman: Elements of Statistical Learning (ESL)
* Goodfellow, Bengio, Courville - Deep Learning (DL) (Available **for free** online [here](http://www.deeplearningbook.org/))

The course will closely follow IMLP, which also comes with Python code and uses scikit-learn (as we will). APM goes into more detail than IMLP but only contains R code. We will not use any R code in this course. ESL contains a rigorous mathematical treatment of the machine learning methods. DL will be our go to book for detailed explanations of neural network and convolutional neural network models.

Additional Materials. Other articles and materials will be distributed via Courseworks that cover additional topics in more depth.

Grade Distribution. The distribution of the parts for your grade is as follows:

Two Exams = 30%

Homework Assignments = 60%

Attendance and Participation = 10%

Changes: There may be adjustments in the scheduling of assignments, exams, and classrooms. Changes will be posted on Courseworks along with other announcements.

**Calendar of Class Sessions and Assignments**

Class 1 (September, 10th). Introduction; How can Machine Learning help social scientists?

|  |
| --- |
| **Reading Assignments** |
| IMLP Ch 1, APM Ch 1-2 |

Class 2 (September, 17th). Software Infrastructure: *Python and Jupyter Notebooks. Pandas. Matplotlib and visualization.*

|  |
| --- |
| **Reading Assignments** |
| IMLP Ch 1 |

Class 3 (September, 24th). Introduction to supervised learning, basic model selection. Linear models for Regression. Note: **HW 1 is due.**

|  |
| --- |
| **Reading Assignments** |
| IMLP p25-44, APM Ch 4-4.3, IMLP p251-262, APM Ch 4.4-4.8, |
| IMLP p45-55, APM Ch 6 |

Class 4 (October, 1st). Linear models for Classification. Preprocessing and feature engineering.

|  |
| --- |
| **Reading Assignments** |
| IMLP p55-68, APM Ch 12.1-12.2, 12.5, IMLP p132-140, IMLP p211-220, APM Ch 3 |

Class 5 (October, 8th). Imputation and Feature Selection. Support Vector Machines

|  |
| --- |
| **Reading Assignments** |
| IMLP p236-241, APM Ch 19, IMLP p92-103, APM Ch 13.4 |

Class 6 (October, 15th). Decision Trees and Random Forests. Gradient Boosting and Calibration. Note: **HW 2 is due.**

|  |
| --- |
| **Reading Assignments** |
| IMLP p70-88, APM Ch 14.1-14.4, IMLP p89-92, |

Class 7 (October, 22nd). **Mid-Term Exam**

Class 8 (October, 29th). Ensemble Models. Model evaluation and imbalanced datasets.

|  |
| --- |
| **Reading Assignments** |
| APM Ch 14.5, IMLP p275-302 |

Class 9 (November 5th). Dimensionality reduction using PCA, Clustering, Manifold Learning.

|  |
| --- |
| **Reading Assignments** |
| IMLP p140-155, p163-187, APM p35-40 |

Class 10 (November 12th). Resampling strategies for Imbalanced Data.

|  |
| --- |
| **Reading Assignments** |
| APM Ch16, SMOTE |

Class 11 (November 19th). Working with Text as Data. Note: **HW 3 is due**

|  |
| --- |
| **Reading Assignments** |
| IMLP p323-336 |

Class 12 (December, 3rd). Neural Networks; Convolutional neural networks for image classification

|  |
| --- |
| **Reading Assignments** |
| IMLP p104-119, DL Ch 6, DL Ch 7.12, Ch 9, keras docs |

Class 13 (December, 10th). Even more on Neural Networks. Note: **HW 4 is due.**

|  |
| --- |
| **Reading Assignments** |
| DL Ch 9 |

**Take Home Final Exam Due by December 17th**