THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Department of Decision Sciences
DNSC 6290: Responsible Machine Learning

Instructor: Patrick Hall, Visiting Faculty

Email: jphall@gwu.edu

Class Time & Location: 4:30 PM ET on Thursdays, DUQUES 359

Office Hours: TBD, occasionally via Webex: https://gwu.webex.com/join/jphall.

Pre-requisite course: DNSC 6314 and 6315, MSBA program candidacy or instructor approval.

Class website: https://jphall663.github.io/GWU_rml/

Course Description
This is a technical, nuts-and-bolts course about increasing transparency, fairness, robustness, and security in machine learning.

Recommended Text

- *Machine Learning for High-risk Applications*, by Patrick Hall, James Curtis, and Parul Pandey (2023), O'Reilly
    - Current available (free) via Dataiku sponsorship
- *Elements of Statistical Learning*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2017), Springer

Supplemental Texts

- *An Introduction to Responsible Machine Learning*, by Patrick Hall, Navdeep Gill, and Benjamin Cox, O'Reilly Media

Tentative Course Calendar

Topic 1	**Interpretable models** - Elastic net linear models - Monotonic gradient boosting machines (GBM) - Generalized additive models (GAM), GA2M, and explainable boosting machines (EBM) - Explainable neural network (XNN)/Gami-net/Neural additive models
Topic 2	**Post-hoc explanation** - Shapley values - Surrogate decision trees - Local interpretable model-agnostic explanations (LIME) - Partial dependence and individual conditional expectation (ICE) - Best practices

Topic 3	**Discrimination testing and remediation**
	- Discrimination definitions
	- Discrimination testing
	- Discrimination remediation
	  - Feature and model specifications
	  - Data pre-processing, in-processing, data post-processing
	- Best practices

Topic 4	**Machine learning security** - Attacks: model and data stealing, data poisoning, backdoors, trojans, evasion - Counter-measures: debugging, throttling, authentication, robust machine learning - General concerns and counter-measures
Topic 5	**Model debugging** - Software testing - Sensitivity analysis - Residual analysis - Benchmark models
Topic 6	**Responsible machine learning best practices** - Technical best practices - Business best practices

Attendance and Participation
Attendance and participation is expected for those students in Washington DC. Class participation will be measured by a combination of in-person attendance, online class attendance, office hour attendance, peer evaluation and digital communications with the instructor.

Reading Assignments
Students are responsible for studying and understanding all assigned materials *before* class. If reading generates questions that are not discussed in class, the student has the responsibility of addressing the instructor privately or raising the issue in an appropriate digital medium.

Learning Outcomes
As a result of completing this course, students will:

- Extend their knowledge of machine learning to include algorithmic foundations of interpretable models, explainable AI, discrimination testing and remediation, machine learning security, and model debugging.
- Gain additional knowledge regarding model governance and compliance for large, regulated organizations.
- Familiarize themselves with important machine learning communities and technologies like GitHub.

Additional Assignments

- **Weekly Assignments:** As this will be a 6 week, workshop based course, students are expected to make weekly progress toward a completed project. Each week the instructor will provide an assignment building towards a completed model.
- **Project Performance:** Lecture materials and hands-on workshop materials will be geared toward application to the class project and contest. Students are expected to participate in this project as individuals or in groups and to do reasonably well. **Students are also expected to apply interpretable modeling, post-hoc explanation, discrimination testing and remediation, and model debugging in the class project.**
- **Public Github Contributions:** Students are expected to write code and generate other artifacts (i.e. notebooks, visualizations, markdown) and to store them in a publicly accessible GitHub repository (or other public location, i.e., personal website, Colab).

Grading Contributions

Weekly Assignments	60%
Public Github “Model Card”	30%
Class Participation	10%

Numeric Grade Scale

≥ 94.00	A
90-93.99...	A-
87-89.99...	B+
84-86.99...	B
80-83.99...	B-
77-79.99...	C+
74-76.99...	C
70-73.99...	C-
≤ 69.99...	F

Statement on Diversity and Belonging

To make the most of this course, the instructor and students must create a rigorous and lively forum of ideas that is welcoming to everyone. The opportunity to speak freely and know that you will be heard, even if not agreed with, is crucial. We must be careful to approach our discussions with empathy and mutual respect, irrespective of ideology, political views, or identity. Civility is a value in this course because it permits intellectual, personal, and professional exploration and growth, and we want to make sure those opportunities for exploration and growth include all members of our classroom community.

University Policy on Observance of Religious Holidays
In accordance with University policy, students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. For details and policy, see: http://provost.gwu.edu/policies-procedures-and-guidelines.

Academic Integrity
Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information. For details and complete code, see: http://studentconduct.gwu.edu/code-academic-integrity. Violations of the academic integrity code could result in serious disciplinary action.

Contact Hours Statement
This 1.5-credit course includes 2.5 hours of direct instruction (class meetings) and a minimum of 5 hours of independent learning (outside of class), totaling a minimum of 7.5 hours per week, totaling approximately 52 hours for the 7-week term, including exams period.

Disability Support Services (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact Disability Support Services in Rome Hall, 801 22nd Street, NW, Suite 102, to establish eligibility and to coordinate reasonable accommodations. For additional information see: http://disabilitysupport.gwu.edu. Phone: 202-994-8250.

Counseling and Psychological Services
GW's Student Health Center offers counseling and psychological services, supporting mental health and personal development by collaborating directly with students to overcome challenges and difficulties that may interfere with academic, emotional, and personal success. For additional information see http://healthcenter.gwu.edu/counseling-and-psychological-services. Phone: 202-994-5300.

Safety and Security
In an emergency: call GWPD 202-994-6111 or 911. For situation-specific actions: review the Emergency Response Handbook: http://safety.gwu.edu/emergency-response-handbook. In an active violence situation: Get Out, Hide Out, or Take Out: http://go.gwu.edu/shooterprep. Stay informed: http://safety.gwu.edu/stay-informed.

Class Policy Changes
The instructor reserves the right to revise any items on this syllabus, including, but not limited to any class policy, course outline or schedule, grading policy, test date, etc. Note that the requirements for deliverables may be clarified and expanded in class, via email, or on Blackboard. Students are expected to complete the deliverables incorporating such additions.