STAT 422/722, Predictive Analytics Syllabus

Class Info:	Both sections meet on Monday/Wednesday, 01/15/20 - 03/04/20: Section 401: 12:00 - 1:30 pm in JMHH 255 Section 403: 3:00 - 4:30 pm in JMHH 255
Prerequisite:	Stat 102 for undergraduates and Stat 613 for MBAs
People:	Instructor: Kam Hamidieh, hkam@wharton.upenn.edu , JMHH 447, 215-898-9477 TA: Elsa Yang, yachong@wharton.upenn.edu
Office Hours:	Kam: Mondays 4:30-6:00 and Tuesdays 2:00-4:00 in JMHH 423 or by appointment Elsa: Tuesdays 4:00-6:00 in JMHH 452
Text:	You are not required to purchase any textbooks. I will provide the course slides before each class. The main reference book is the highly readable *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. You can download a free pdf version at: http://faculty.marshall.usc.edu/gareth-james/ISL/. The first author uses the book for an MBA data course at USC.
Software:	Our main tool will be **R**, https://www.r-project.org, along with RStudio Integrated Development Environment (IDE), https://rstudio.com. You are welcomed to use another IDE. I do not assume you have any experience with **R**. We will post sample codes, tutorials, and hold plenty of office hours to provide the necessary help with programming. You will be required to bring your laptop to each class; we will have hands on work.
Description:	You will: • Learn the fundamental principles behind predictive analytics. • Be introduced to the most widely used prediction algorithms and techniques. • Examine how predictive analytics can be used in decision making. • Learn about the cutting edge technologies and their potential impacts in business. The main topics include: Review of multiple regression, cross validation, variable selection procedures, shrinkage methods such as lasso, logistic regression, ROC curves and confusion matrix, trees, resampling techniques, random forests, boosting, neural networks & deep learning. Time permitting, we may cover principal component analysis, support vector machines, and clustering.
Grading:	• 30%: Participation which includes attendance, classroom engagement, and occasionally submitting short assigned work. • 70%: There will be at least 4 prediction competitions held in www.Kaggle.com . Your score for each challenge will partly depend on beating set benchmarks.
Important Stuff:	• You'll get an F for the entire course if you cheat on anything. • There will be no make-ups, extensions, or extra credit opportunities under any circumstances. • Collaboration is encouraged. However, you must create and submit your own prediction results and turn in all the assigned work individually.