GRADUATE COURSE SYLLABUS
IST-718 Big Data Analytics

Professor: Willard Williamson

Professor Email: wewillia@syr.edu

Office Hours:

- Tuesday, 8 PM - ?, Hall of Languages, Room 205
- Thursday, 8 PM - ?, Lyman Hall, Room 229
- Saturday morning, by appointment, Hinds Hall, Room 239

Faculty Assistant: Yash Suresh Pasar

Faculty Assistant Email: yspasar@syr.edu

Faculty Assistant Office Hours:

- Tuesday, 10 – 12, Hinds Hall, Room 239
- By appointment

Tuesday Class: Hall of Languages, Room 205, 5 – 7:45 PM

Thursday Class: Lyman Hall, Room 229, 5 – 7:45 PM

Class Discussion Board Sign Up Link: www.piazza.com/syr/spring2020/ist718

Course Description:

A broad introduction to analytical processing tools and techniques for information professionals. Students will develop a portfolio of resources, demonstrations, recipes, and examples of various analytical techniques.

Additional Course Description:

This course will prepare you to participate as a Data Scientist on big data and data analytics projects

Prerequisite / Co-requisite:

Familiarity with command-line interfaces, quantitative skills including statistics, a basic knowledge of linear algebra, basic probability, basic statistics, basic calculus, strong algebra skills, and strong programming skills in Python or some other language. This is not an introductory course; most students who take this course have already taken IST-687 Introduction to Data Science. We will review of all the prerequisites outlined above at the start of the course.

If you have never programmed before, you should probably take an introductory programming course first (preferably in python). If it's been a long time since you took calculus, linear algebra, probability, and statistics, it's probably not going to be a problem. You need to

understand the high-level concept of a derivative and partial derivative, what a matrix is, how to multiply matrices, and basic probability and statistics. We will use the computer to crunch the numbers but we need to understand the high-level math concepts in order to understand the data science topics presented in class. It is assumed that all students are excellent with algebra and this topic will not be reviewed. If you don't have any of the math skills, you might consider taking some prerequisite math courses before you take this class. You should have at least a subset of the recommended math skills.

One test for math skills is to look through some of the assigned reading in Introduction to Statistical Learning in R (ISLR) and Deep Learning (DL) textbooks. If you are totally lost and have never seen any of the math before, this course is probably not for you. ISLR is an especially good test because it is the most used textbook in the class and is very representative of the typical mathematical skill level needed to be successful. The Deep Learning book is used for the math review and only used sparingly beyond that. If you can read and understand the ISLR book, you probably have the mathematical maturity needed to be successful in the class.

Audience:

- Any student in the School of Information Studies
  - Pursuing an MS in the Applied Data Science program
  - Pursuing a Certificate of Advanced Study in Data Science in the School
- Any student with an interest in big data analytics.

Credits: 3

Learning Objectives:

After taking this course, the students will be able to:

- Translate a business challenge into an analytics challenge
- Use linear and logistic regression, decision trees, and neural networks to make predictions
- Use data science to gain actionable insights
- Use Python and Apache Spark to build big data analytics pipelines
- Learn classic and state of the art machine learning techniques
- Explain how advanced analytics can be leveraged to create a competitive advantage

Texts / Supplies – Required:

- **Python Data Science Handbook (PDSH)** by Jake VanderPlas (Free), https://jakevdp.github.io/PythonDataScienceHandbook/
- **An introduction to Statistical Learning with Applications in R (ISLR)** by Gareth James, Daniela, Witten, Trevor Hastie, and Robert Tibshirani (Free)
  http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
  http://fs2.american.edu/alberto/www/analytics/ISLRLectures.html

- **Deep Learning (DL)** by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (Free) http://www.deeplearningbook.org/
- **Apache Spark: The Definitive Guide (ASDG)** by Chambers and Zaharia https://www.amazon.com/Spark-Definitive-Guide-Processing-Simple/dp/1491912219/ref=sr_1_1?crid=SY2OHTSAXNQB&keywords=apache+spark+definitive+guide&qid=1564328939&s=gateway&sprefix=apache+spark+definitiv%2Caps%2C150&sr=8-1. This is the only book in the course that is not available for free.
- **Open Intro Statistics (OIS)** by David Diez et al., 4th edition (https://www.openintro.org/stat/textbook.php, Texts / Supplies– Additional:
- Will be very useful to have a computer as powerful as a late model Macbook Pro or Windows / Linux equivalent.
- SPARK can be installed on your local computer and assignments can be done on your local computer.
- If you choose not to install SPARK on your own computer, the Datbricks community edition (https://databricks.com/product/faq/community-edition, https://databricks.com/try-databricks) is the official compute environment for the class. If using Databricks exclusively then all you need is a computer capable of running a web browser.

Course Requirements and Expectations:

Course assignments will consist of 6 equally weighted homework assignments and one group project. The preliminary grading table below describes the points breakdown in more detail. All assignments must run on databricks. If the assignment runs on your own computer but fails to run on databricks, that could lead to point reductions in your assignments. Please verify your assignments run on databricks before submission.

Grading:

Assignment	Description	Points
6 Homework Assignments	100 points per homework assignment	600

Assignment	Description	Points
Project Abstract	A one-page preliminary summary of the project – see assignment for details. Note that by the time you write your abstract, you should have actually worked with the data set you plan to use for your project. If you make major changes to your project like, your abstract grade will be changed to a 0. If you have to make a major chant to your project, it shows that you didn't really work with the data and think about your project before you wrote your abstract; but rather, were just trying to satisfy a requirement and check a box. For example, you originally decided to do image recognition but after you submit your abstract, you find that the images in the data set you chose are too large for your compute environment. Minor changes, additions, deletions, changes in direction are allowed.	50
Project Report	A 10 plus page comprehensive project report – see assignment for details.	100
Project Code	Well designed / commented code which runs without error.	50
In Class Project Presentation	A 20 minute in class project presentation where all students in the project group participate in the presentation. Individual students will not receive credit for the project presentation if the student doesn't participate in the presentation, leaves class early, or arrives significantly late.	75
Class / Piazza Participation	Did the student register on the piazza class discussion board, participate in piazza discussions, and participate in class discussions.	25
Attendance	Attendance is not directly tied to student grades. However, student attendance is considered when awarding final course grades where a student's grade is less than 1 point away from the next higher grade. For example, if a students final grade is 94.1 but has a good attendance record, the professor will likely round the grade up to a 95 and award the student an A.	0

Assignment	Description	Points
Final Exam	Blackboard online final exam. The final exam will not be math based; but rather, test the students understanding of high level conceptus. Throughout the semester, we will emphasize important high level concepts and the goal of the final exam is to test the students understanding of the high level concepts.	100
Total Points	**Note: Final Average % = Total Earned Points / Total Points**	**1000**

Grading Tables

Note that the A grade range is a little narrower than typical university grading standards which makes it slightly harder to get an A in the course. The typical A grade range is 93 – 100 but in this class an A grade is 95 – 100. However, the A-, B+, B, and below grade ranges are extended. For example, the B grade range extends down to 80%. Please consider the extended grade ranges below before making a decision to drop the class.

Grades	Grade points /credit	Percentage Range
A	**4.000**	95 – 100
A-	**3.667**	90 – 94
B+	**3.333**	85 - 89
B	**3.000**	80 - 84
B-	**2.667**	75 - 79
C+	**2.333**	70 - 74
C	**2.000**	65 - 69
C-	**1.667**	60 - 64
F	**0**	0 - 59

University Attendance Policy

Attendance in classes is expected in all courses at Syracuse University. Students are expected to arrive on campus in time to attend the first meeting of all classes for which they are registered. Students who do not attend classes starting with the first scheduled meeting may be academically withdrawn as not making progress toward degree by failure to attend. Instructors set course-specific policies for absences from scheduled class meetings in their syllabi.

It is a federal requirement that students who do not attend or cease to attend a class to be reported at the time of determination by the faculty. Faculty should use “ESPR” and “MSPR” in Orange Success to alert the Office of the Registrar and the Office of Financial Aid. A grade of NA

is posted to any student for whom the Never Attended flag is raised in Orange SUccess. More information regarding Orange SUccess can be found here, at:
http://orangesuccess.syr.edu/getting-started-2/
Students should also review the University's religious observance policy and make the required arrangements at the beginning of each semester

Course Specific Policies on attendance, late work, make up work, examinations

The professor realizes that unexpected things come up during the semester. To compensate for unexpected life events, all students will get 5 free late days for homework assignments excluding the project. Late days are allotted in whole late days. For example, assuming that homework 1 is due at midnight (12:00 AM) on Tuesday, the student currently has 5 late days remaining, and the assignment is turned in at 12:00 PM on Tuesday (12 hours after the deadline), one late day will be used and the student will have 4 late days left. Late days are always applied on whole day increments.

After all late days are used, 1 letter grade will be deducted for each late day. For example, assume that a student used all 5 late days (0 late days left), turned in an assignment 12 hours late, and got a 95% on the late assignment. The assignment will be graded as 95% – 10% = 85%; one full letter grade would be deducted from the assignment.

Late days do not apply to the project code, report, or in class presentation.

My advice is to use your late days wisely and don't waste them on trivial matters. Late days are intended to be used for unexpected life events, religious holidays, sports team events, etc.

How to Succeed in This Course

- Start all assignments early. It's better to have free time at the end of the assignment (so you can work on your project 😊) because you finished it early than be trying to work as fast as possible to get the assignment done.
- Starting early allows you to get questions answered which could be blocking progress. You don't want to burn late days waiting to get questions answered.
- Show up regularly for lecture.
- Participate in the piazza discussion forum.
- The course textbooks are excellent, do the assigned reading.
- Don't think of the class as something where you are just completing assignments, try to achieve a high level of understanding.

Piazza Discussion Forum

Participation on the piazza discussion forum is highly encouraged and tied to your final grade. Everyone is allowed to ask and answer questions on piazza. If you know the answer to a question, go ahead and answer the question. However, do not post complete solutions to homework questions on piazza.

Syracuse University Policies:

Syracuse University has a variety of other policies designed to guarantee that students live and study in a community respectful of their needs and those of fellow students. Some of the most important of these concerns:

Diversity and Disability (ensuring that students are aware of their rights and responsibilities in a diverse, inclusive, accessible, bias-free campus community) can be found here, at https://www.syracuse.edu/life/accessibilitydiversity/

Religious Observances Notification and Policy (steps to follow to request accommodations for the observance of religious holidays) can be found here, at: http://supolicies.syr.edu/studs/religious_observance.htm

Orange SUccess (tools to access a variety of SU resources, including ways to communicate with advisors and faculty members) can be found here, at: http://orangesuccess.syr.edu/getting-started-2/

Disability-Related Accommodations:

Syracuse University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. There may be aspects of the instruction or design of this course that result in barriers to your inclusion and full participation in this course. I invite any student to meet with me to discuss strategies and/or accommodations (academic adjustments) that may be essential to your success and to collaborate with the Office of Disability Services (ODS) in this process.

If you would like to discuss disability-accommodations or register with ODS, please visit their website at http://disabilityservices.syr.edu./ Please call (315) 443-4498 or email disabilityservices@syr.edu for more detailed information.

ODS is responsible for coordinating disability-related academic accommodations and will work with the student to develop an access plan. Since academic accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible to begin this process.

University Academic Integrity Policy:

Syracuse University's Academic Integrity Policy reflects the high value that we, as a university community, place on honesty in academic work. The policy defines our expectations for academic honesty and holds students accountable for the integrity of all work they submit. Students should understand that it is their responsibility to learn about course-specific expectations, as well as about university-wide academic integrity expectations. The policy governs appropriate citation and use of sources, the integrity of work submitted in exams and

assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The policy also prohibits students from submitting the same work in more than one class without receiving written authorization in advance from both instructors. Under the policy, students found in violation are subject to grade sanctions determined by the course instructor and non-grade sanctions determined by the School or College where the course is offered as described in the Violation and Sanction Classification Rubric. SU students are required to read an online summary of the University's academic integrity expectations and provide an electronic signature agreeing to abide by them twice a year during pre-term check-in on MySlice.

Course Specific Academic Integrity Policy

Students found violating Syracuse University's academic integrity policy including but not limited to cheating or plagiarism will receive a grade of 0 on the assignment for which the policy is violated. The professor also reserves the right to report academic integrity violations to the university which could result in sanctions including being removed from the course.

The basic guiding principal of academic integrity in this course is that you must submit work written entirely by you. You are allowed to discuss problems with your class mates. However, everything you submit must be written by you except for short code snippets you may have obtained from web sites like stackoverflow.com. Code presented in class is always acceptable to copy and submit in homework assignments or in the project. Essentially, you can get ideas for how to solve problems from class mates, instructors, and the internet. **However, the work you submit in your homework assignments must be written by you.**

I have been a professional software engineer for over 25 years. That means I have been reading and reviewing code in my professional life for over 25 years. After working with a group of software engineers for a short period of time, I can oftentimes tell who wrote a piece of code just by the way it is written. People have unique ways of expressing themselves in code, much like people express themselves with the written word. I am very skilled in identifying unique patterns and signatures in code. If you copy your class mate's code, I will see that when I grade your homework. Unfortunately, I have identified students copying code in IST-718 in the past. In one case, I was even able to tell who provided the code and who received the code based on the patterns and how the software behaved at run time.

Please be advised that blatant cases of academic dishonesty will be awarded a grade of 0 in all cases with absolutely no exceptions. The professor also reserves the right to report academic integrity violations to the University which could result in sanctions including being removed from the course.

Allowed Activities

- Discuss approaches on how to solve homework problems.
- Discuss if problems seem like they are correct or not.
- Help a class mate understand a key concept that will help solve a homework or project problem.
- Use web sites like stackoverflow.com to help solve homework and project problems

- Use short code snippets from stackoverflow.com to help solve homework and project problems.
- Help each other with python and spark runtime problems. It's fine to show a class mate your code to get help with a run time problem if you trust that your class mate will not copy your code and submit it as their own.
- Think of it like this: The goal of the class is for YOU to develop skills to solve data science problems and write code that implements solutions to those problems. You can get some ideas from others on how to solve the problems but YOU are the one that needs to turn the ideas into code that implements the solution.

Prohibited Activities

- **Submitting code or written answers which were written by someone other than you.**
- Copying a solution from a class mate, changing it a little, and submitting it as your own.
- Coding solutions to homework problems in pairs or groups then slightly changing the answers across the assignments in the group and submitting.
- Copy a complete homework or project solution from the internet.
- Submitting project work that you did in another class. If you previously turned in a project for another class, you are not allowed to resubmit it for this class. You are also not allowed to slightly modify a project you submitted for another class and submit it for this class. It is possible that you could perform a major extension to a previous project from another class. Please consult with the professor to avoid the possibility of academic integrity violations.
- Solving problems by submitting homework questions in their entirety to stackoverflow.com. You are allowed to use stack overflow to get ideas and help for really specific coding issues but not to answer entire homework questions. For example, if the homework problem says implement a logistic regression algorithm to predict survival in the titanic data set, you could use stack overflow to help solve some specific logistic regression coding problem. However, submitting the entire homework question is prohibited.

Note that students are required to complete an ungraded academic integrity quiz on blackboard and obtain 100% on the quiz. Students who do not complete the quiz by the due date risk being removed from the course.

Educational Use of Student Work

Student work prepared for University courses in any media may be used for educational purposes, if the course syllabus makes clear that such use may occur. You grant permission to have your work used in this manner by registering for, and by continuing to be enrolled in, courses where such use of student work is announced in the course syllabus.

The professor may use academic work that you complete this semester for educational purposes in this course during this semester. Your registration and continued enrollment constitute your permission.

The professor may also use academic work that you complete this semester in subsequent semesters for educational purposes. Before using your work for that purpose, I will either get your written permission or render the work anonymous by removing all your personal identification.

As a generally accepted practice, honors theses, graduate theses, graduate research projects, dissertations, or other capstone projects submitted in partial fulfillment of degree requirements are placed in the library, University Archives, or department for public reference.

Discrimination or Harassment

The University does not discriminate and prohibits harassment or discrimination related to any protected category including creed, ethnicity, citizenship, sexual orientation, national origin, sex, gender, pregnancy, disability, marital status, age, race, color, veteran status, military status, religion, sexual orientation, domestic violence status, genetic information, gender identity, gender expression or perceived gender.

Any complaint of discrimination or harassment related to any of these protected bases should be reported to Sheila Johnson-Willis, the University's Chief Equal Opportunity & Title IX Officer. She is responsible for coordinating compliance efforts under various laws including Titles VI, VII, IX and Section 504 of the Rehabilitation Act. She can be contacted at Equal Opportunity, Inclusion, and Resolution Services, 005 Steele Hall, Syracuse University, Syracuse, NY 13244-1120; by email: titleix@syr.edu; or by telephone: 315-443-0211.

Federal and state law, and University policy prohibit discrimination and harassment based on sex or gender (including sexual harassment, sexual assault, domestic/dating violence, stalking, sexual exploitation, and retaliation).

If a student has been harassed or assaulted, they can obtain confidential counseling support, 24-hours a day, 7 days a week, from the Sexual and Relationship Violence Response Team at the Counseling Center (315-443-4715, 200 Walnut Place, Syracuse, New York 13244-5040). Incidents of sexual violence or harassment can be reported non-confidentially to the University's Title IX Officer (Sheila Johnson Willis, 315-443-0211, titleix@syr.edu, 005 Steele Hall). Reports to law enforcement can be made to the University's Department of Public Safety (315-443-2224, 005 Sims Hall), the Syracuse Police Department (511 South State Street, Syracuse, New York, 911 in case of emergency or 315-435-3016 to speak with the Abused Persons Unit), or the State Police (844-845-7269).

I will seek to keep information you share with me private to the greatest extent possible, but as a professor I have mandatory reporting responsibilities to share information regarding sexual misconduct, harassment, and crimes I learn about to help make our campus a safer place for all.

Course evaluations:

There will be an end of course evaluation for you to complete this term. This evaluation will be conducted online and is entirely anonymous. You will receive an official notification in your

email account with the evaluation website link and your passcode. Please take the time and fill out this evaluation as your feedback and support of this assessment effort is very much appreciated. The school carefully reviews ratings and comments that you submit, and these factor into decisions about course, program and instructor development.

Use of Blackboard:

This course makes minimal use of Syracuse University's Blackboard system as an online tool. To access Blackboard, [http://blackboard.syr.edu] use your Syracuse University NetID & Password. This specific course will appear in your course list.

To search for answers to your Blackboard questions, visit the Answers self-help knowledge [https://answers.syr.edu/display/blackboard01/Blackboard]. If you have problems logging in or need assistance with Blackboard, contact the ITS Service Center at: help@syr.edu or 315.443.2677. The Syracuse University Blackboard support team will assist you.

Course Schedule

The following is a preliminary course schedule.

Week	Class #	Topics	Suggested Reading / Notes	Events
Jan 13	1	**Course overview and python programming:** Anaconda installation, jupyter notebooks, databricks, git, Numpy, Pandas, Matplotlib	PDSH: Ch 1 PDSH: Ch 2.2 – 2.7 PDSH: Ch 3.2 – 3.4 PDSH: Ch 4.1 – 4.9	
Jan 20	2			Homework 1 Released
Jan 27	3	**Review:** Probability, statistics, linear algebra.	DL: Ch 2.1 – 2.6 OIS: Ch 3 OIS: Section 4.1, 4.3	Form Project Teams – Students pick their own teams of 3 to 4
Feb 3	4	**Introduction to Spark, Hadoop, MapReduce, and YARN**	ASDG: Pages 13 – 18 ASDG: Ch 4, 5, 12	Project team members due Homework 1 Due Homework 2 Released

Week	Class #	Topics	Suggested Reading / Notes	Events
Feb 10	5	**Introduction to Spark DataFrames and Spark ML** Calculus Review	ASDG: Chapter 6	
Feb 17	6	**A Statistical Perspective on Machine Learning:** Introduction to linear regression, mean squared error, reducible and irreducible error, regression vs. classification, parametric vs. non parametric models.	ISLR: Ch 1 ISLR: Ch 2.1 ASDG: Chapter 24	Homework 2 Due Homework 3 Released
Feb 24	7	**Assessing Model Accuracy:** Bias variance tradeoff, confusion matrix, receiver operating characteristic, model selection.	ISLR: Ch 2.2 ISLR chapter 3.1, 3.2 ISLR 5.1	Homework 3 Due Homework 4 Released
Mar 2	8	**Sentiment Analysis Case Study:** Supervised learning, model selection, logistic regression, regularized logistic regression, elastic net regularization, model interpretation	ISLR Ch 4.1, 4.2, 4.3 ISLR Ch 6	
Mar 9	9			Project Abstract Due
Mar 16	N/A	**Spring Break – No Class**		
Mar 23	10	**Course Recommendation Case Study:** Unsupervised Learning, Principal component analysis, K-Means clustering.	ISLR Ch 10.1 – 10.3.1 PDSH: Chapter 5 – In Depth – Principal Component Analysis	Homework 4 Due Homework 5 Released
Mar 30	11		PDSH: Chapter 5 – In Depth – K-Means Clustering	
Apr 6	12	**Predicting Credit Score Case Study:** Decision trees, bagging, boosting, random forests. ISLR chapter 5.2	ISLR chapter 8 ASDG p 465 – 469 (trees)	Homework 5 Due Homework 6 Released
Apr 13	13			
Apr 20	14	**Deep Learning**	DL Ch 6 ASDG Ch 31	Homework 6 Due

Week	Class #	Topics	Suggested Reading / Notes	Events
Apr 23	**15**	**Thursday Class Project Presentations**	**Last Thursday Class of the Semester**	**Project Report and Code Due**
Apr 28	**15**	**Tuesday Class Project Presentations**	**Last Tuesday Class of the Semester**	**Project Report and Code Due**
May 4	**N/A**	**Online Blackboard Final Exam**		