STA551-Foundations of Data Science

Instructor: Cheng Peng

Summer 2023

1 Contact

- Email: cpeng@wcupa.edu
- Office: 111 UNA 125
- Phone: 610-436-2369
- Office Hours:

Day	Time	Location
Monday/Thursday	4:00 PM-6:30 PM	UNA111 and ZOOM
By appointment		ZOOM only

- ZOOM Link: Available on the course web page and D2L as well.
- Course Web Page: https://pengdsci.github.io/STA551/

2 Course Description

This is a data science survey course. The first part of this course will be dedicated to data science foundations. Topics include statistical models, machine learning algorithms, model performance metrics, and major resampling algorithms. The second part will focus on data science processes. Topics include data science project life cycle, model selection, validation, performance evaluation, and data science ethics. The last part of the course will discuss data science infrastructure and pipelines.

3 Learning Objectives and Outcomes

3.1 Course Objectives

Upon successful completion of this course, students will be equipped with skills and strategies

- to understand the basic notion that a data science project is, in general, a process that builds a system of multiple models and/or algorithms.
- to convert practical problems accurately to data science problems and prepare data sets from possibly multiple data sources for analytics.
- to understand basic statistical methods and machine algorithms and their corresponding scope of applications.
- to identify the best solutions to a data science process from candidate models and algorithms.
- to effectively present the results and execute the project to extract actionable insights for decision-making.
- to identify and resolve potential ethical and privacy issues in data science practice.

3.2 Learning Outcomes

After finishing this course, students will be able to

- Describe and apply good practices for storing, manipulating, summarizing
- perform data visualization and exploratory data analysis.
- Use standard software packages and formal programming languages data management and visualization.
- Apply basic techniques from descriptive and inferential statistics and machine learning.

- Interpret and describe the output from such analyses.
- Critically evaluate data-driven methods and claims from case studies, in order to identify and discuss
  - potential ethical issues, and
  - the extent to which stated conclusions are warranted given the evidence provided.
- Complete a data science project and write a report describing the question, methods, and results.

4 Logistics and Required Materials

- **Textbooks:** No required textbook for this class.
- **Class Notes:** Class notes will be provided.
- **Computational Tools:**
  - Programming languages and Software: R, SAS, and Tableau Public
  - Platforms: RStudio, SAS Studio (OnDemand), Anaconda, Github
  - Typesetting: LaTex and Markdown
- **Coverage:** See the list of tentative topics

5 Assessments

5.1 Assessment Components

The course grade consists of the following components:

- Weekly projects (small data analysis, coding, etc) (80%)
- Class attendance and participation (10%)
- Project presentation (10%)

5.2 Grade Scales

The final course grade will be calculated based on the above components. A letter grade will be assigned according to the following scale.

Grade	Quality Points	Percentage Equivalents	Interpretation
A	4.00	[93%, 100%]	Superior graduate attainment
A-	3.67	[90%, 93%)	
B+	3.33	[86%, 90%)	Satisfactory graduate attainment
B	3.00	[83%, 86%)	
B-	2.67	[80%, 83%)	
C+	2.33	[76%, 80%)	Attainment below graduate expectations
C	2.00	[73%, 76%)	
C-	1.67	[70%, 73%)	
F	0	$< 70\%$	Failure

D grades are not used. Refer to the Graduate Catalog for the description of NG (No Grade), W, & other grades.

6 Class Policies

6.1 Attendance and Participation

Attendance in the class is mandatory. Actively participating in class discussion is required in this class and is one of the components of the final course grade.

6.2 Late Homework and Assignments

Late assignments will be accepted. However, all late assignments will be subject to a small penalty deduction.

7 Tentative Topics

Topic 1: Introduction and Tools - A Big Picture [Mon, 7/3]

- Introduction
  - Course logistics and coverage
  - Course policies
- Software and Platforms
  - Computing: R/Rstudio and SAS Studio (via SAS OnDemand)
  - Tech Writing and Reporting: Markdown and LaTex.
  - DBMS and Data Acquisition Tools
  - Viz: Tableau Public and R graphic libraries
  - Repository and Collaboration: Github
- Data Science foundations - Key Pillars
  - Math and stats foundation - models and algorithms
  - Machine learning algorithms
  - Databases technology
  - Effective communication
- Data Science Roles and Skills
  - Different roles with different skills
  - Programming languages: R/Python/Julia, SAS, SQL
  - Data science technologies for effective communication
  - Domain knowledge
  - Strategies and technologies for effective communication
- Data Science Process
  - Identification of practical questions
  - Data acquisition: data sources, security, and ethics
  - Data management (wrangling): missing values, inconsistency, etc.
  - Feature engineering: feature creation, extraction, transformation, etc.
  - iterative Model/algorithm building process: EDA, statistics and machine learning
  - Visualization and communication: make the business implications intuitive
  - Implementation and strategies
  - Model maintenance and updating

Topic 2: Understanding Data Generation Process, Storage and Preparation[Thu, 7/6]

- Data Generation Process and Storage
  - How data were generated/collected
  - what are data types
  - Where data are stored
- Querying Databases

  - Basics of rational databases and DBMS
  - Query databases: Basics of SQL
- Best practices of data wrangling
  - Potential issues of data:
  - Exploratory data visualization and analysis for EDA
  - Mistakenly recorded data and inconsistency
  - Missing value handling
- Feature Engineering
  - Handling categorical variables - aggregation
  - Handling extremely skewed numerical variable - discretization
  - Feature extraction - PCA
- Data Preparation Process
  - Iterative process
  - Roles of EVA and EDA

Topic 3: Statistical Methods for Data Science [Mon, 7/10]

- Mathematical and statistical foundations of data science
  - Calculus and matrix algebra:vectorization and optimization
  - Statistical models and algorithms
  - Methods of estimation: Frequentist and Bayesian approaches
  - Model/algorithm assumptions
- Traditional regression models for prediction
  - Linear regression models
  - Generalized linear regression models
  - Time series forecasting
- Re-sampling algorithms
  - Overview of re-sampling algorithm
  - Bootstrapping methods
  - Extensions of Bootstrapping algorithms

Topic 4: Cross Validation and Performance Measures [Thu, 7/13]

- Cross-validation
  - The logic of cross-validation
  - The use cases of cross-validation
- Performance Measure
  - Error-based measures
  - likelihood-based
- Global Performance measures and visualizations
  - Profit curve
  - ROC and AUC
  - Lift curves (optional)

Topic 5: Case-Study 1 and Project 1 [Mon, 7/17]

- Case-study: Predicting bank loan default
  - Data set link

Topic 6 Survey of Supervised Algorithms [Thu, 7/20]

- Overview machine learning and its relation to statistics
  - What is machine learning?
  - Difference between machine learning and statistics
  - Three major types of machine learning
- Decision Tree-based Classification
  - Decision tree classification
  - Bootstrap Aggregation (BAGGING)

  - Random forests
- Regression-based classification
  - Logistic regression
  - Artificial neural networks (ANN)
  - Extensions of ANN
- Mathematical Algorithms
  - Naive Bayes
  - KNN
  - Support vector machines

Topic 7: Survey Unsupervised learning algorithm [Mon, 7/24]

- Unsupervised learning algorithms
  - Clustering
  - K-means
- Dimensional Reduction
  - Principle component analysis
  - Regularization approaches
- Project #1 due: Monday,

Topic 8: Case Study: [Thu, 7/27]

**

Topic 9: Model/Algorithm Deployment [Mon, 7/31]

- Pre-deployment: Evaluating models in staging environment
  - dependency between data sources and model
  - pre-deployment testing and updates
  - The two-language problems
- Model management in production: champion and challengers
  - Static and real-time system
  - Auditing and version control
  - Tracking accuracy and gains
- Model Maintenance/updating
  - Monitoring
  - Evaluating
  - Rebuilding/Retraining
- Challenges
  - interdisciplinary
  - (near) real-time model updating

Week 10: Data Science Ethics [Mon, 8/3]

- Potential Ethical Issues in Data
  - Privacy issues
  - Biases in study design and data collection
- Potential Ethical Issues in Reporting
  - Failure to reporting negative results
  - Cherry-picking impressive partial results
  - Failure to investigate the model performance
- Potential Ethical Issues in Algorithms and Models
  - Biases in algorithm
  - Algorithm and model interpretability
  - Model and algorithm reproducibility
- Project #2 Due Sunday, July 2.

8 University Policies and Resources

8.1 ACADEMIC & PERSONAL INTEGRITY

It is the responsibility of each student to adhere to the university's standards for academic integrity. Violations of academic integrity include any act that violates the rights of another student in academic work, that involves misrepresentation of your own work, or that disrupts the instruction of the course. Other violations include (but are not limited to): cheating on assignments or examinations; plagiarizing, which means copying any part of another's work and/or using ideas of another and presenting them as one's own without giving proper credit to the source; selling, purchasing, or exchanging of term papers; falsifying of information; and using your own work from one class to fulfill the assignment for another class without significant modification. Proof of academic misconduct can result in automatic failure and removal from this course. For questions regarding Academic Integrity, the No-Grade Policy, Sexual Harassment, or the Student Code of Conduct, students are encouraged to refer to the Department Graduate Handbook, the Graduate Catalog, the Ram's Eye View, and the University website at www.wcupa.edu.

8.2 STUDENTS WITH DISABILITIES

If you have a disability that requires accommodations under the Americans with Disabilities Act (ADA), please present your letter of accommodations and meet with me as soon as possible so that I can support your success in an informed manner. Accommodations cannot be granted retroactively. If you would like to know more about West Chester University's Services for Students with Disabilities (OSSD), please visit them at 223 Lawrence Center. The OSSD hours of Operation are Monday – Friday, 8:30 a.m. – 4:30 p.m. Their phone number is 610-436-2564, their fax number is 610-436-2600, their email address is ossd@wcupa.edu, and their website is at www.wcupa.edu/ussss/ossd.

8.3 EXCUSED ABSENCES POLICY

Students are advised to carefully read and comply with the excused absences policy, including absences for university-sanctioned events, contained in the WCU Graduate Catalog. In particular, please note that the "responsibility for meeting academic requirements rests with the student," that this policy does not excuse students from completing required academic work, and that professors can require a "fair alternative" to attendance on those days that students must be absent from class in order to participate in a University-Sanctioned Event.

8.4 REPORTING INCIDENTS OF SEXUAL VIOLENCE

West Chester University and its faculty are committed to assuring a safe and productive educational environment for all students. In order to comply with the requirements of Title IX of the Education Amendments of 1972 and the University's commitment to offering supportive measures in accordance with the new regulations issued under Title IX, the University requires faculty members to report incidents of sexual violence shared by students to the University's Title IX Coordinator. The only exceptions to the faculty member's reporting obligation are when incidents of sexual violence are communicated by a student during a classroom discussion, in a writing assignment for a class, or as part of a University-approved research project. Faculty members are obligated to report sexual violence or any other abuse of a student who was, or is, a child (a person under 18 years of age) when the abuse allegedly occurred to the person designated in the University Protection of Minors Policy. Information regarding the reporting of sexual violence and the resources that are available to victims of sexual violence is set forth at: https://www.wcupa.edu/_admin/diversityEquityInclusion/sexualMisconduct/default.aspx

8.5 INCLUSIVE LEARNING ENVIRONMENT AND ANTI-RACE STATEMENT

Diversity, equity, and inclusion are central to West Chester University's mission as reflected in our Mission Statement, Values Statement, Vision Statement and Strategic Plan: Pathways to Student Success. We

disavow racism and all actions that silence, threaten, or degrade historically marginalized groups in the U.S. We acknowledge that all members of this learning community may experience harm stemming from forms of oppression including but not limited to classism, ableism, heterosexism, sexism, Islamophobia, anti-Semitism, and xenophobia, and recognize that these forms of oppression are compounded by racism.

Our core commitment as an institution of higher education shapes our expectation for behavior within this learning community, which represents diverse individual beliefs, backgrounds, and experiences. Courteous and respectful behavior, interactions, and responses are expected from all members of the University. We must work together to make this a safe and productive learning environment for everyone. Part of this work is recognizing how race and other aspects of who we are shape our beliefs and our experiences as individuals. It is not enough to condemn acts of racism. For real, sustainable change, we must stand together as a diverse coalition against racism and oppression of any form, anywhere, at any time.

Resources for education and action are available through WCU's Office for Diversity, Equity, and Inclusion (ODEI), DEI committees within departments or colleges, the student ombudsperson, and centers on campus committed to doing this work (e.g., Dowdy Multicultural Center, Center for Women and Gender Equity, and the Center for Trans and Queer Advocacy).

Guidance on how to report incidents of discrimination and harassment is available at the University's Office of Diversity, Equity and Inclusion.

8.6 EMERGENCY PREPAREDNESS

All students are encouraged to sign up for the University's free WCU ALERT service, which delivers official WCU emergency text messages directly to your cell phone. For more information, visit www.wcupa.edu/wcualert. To report an emergency, call the Department of Public Safety at 610-436-3311.

8.7 ELECTRONIC MAIL POLICY

It is expected that faculty, staff, and students activate and maintain regular access to University-provided e-mail accounts. Official university communications, including those from your instructor, will be sent through your university e-mail account. You are responsible for accessing that mail to be sure to obtain official University communications. Failure to access will not exempt individuals from the responsibilities associated with this course.