## STA 551 Foundations of Data Science
## Spring Semester

**Professor**: _ **Phone**: 610-436-_
**Email**: _@wcupa.edu **Office**: Building Room
**Office Hours**: Day 1
Day 2
Day 3

**Prerequisites:** STA 503 and 506

**Required Materials**: None

**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, and performance evaluation, and data science ethics. The last part of the course will discuss data science infrastructure and pipelines.

**Applicable Programmatic Student Learning Outcomes:**

1. Demonstrate an understanding of probability and statistical inference, including the fundamental laws of classical probability, discrete and continuous random variables, expectation theory, maximum likelihood methods, hypothesis testing, power, and bivariate and multivariate distribution theory.
2. Demonstrated the ability to apply the elementary methods of statistical analysis, namely those based on classical linear models, categorical methods, and non-parametric ideas to perform data analysis for the purposes of statistical inference.
3. Demonstrate proficiency in the effective use of computers for research data management and for analysis of data with standard statistical software packages, particularly SAS.
4. Learn to develop and critically assess design of experimental studies and the collection of data.
5. Apply one or more methods of statistical inference to a particular area of interest, particularly the program in the elective concentration.
6. Gain practical experience in statistical consulting and communicating with non- statisticians, culminating with interaction with research workers at a local company as part of the internship practicum.

**Course Student Learning Outcomes:**

Students will be able to:

1. know the basic notion that a data science project is, in general, a process that builds a system of multiple models and/or algorithms. (PSLO 5)
2. convert practical problems accurately to data science problems and prepare data sets from possibly multiple data sources for analytics. (PSLO 5)
3. understand basic statistical methods and machine algorithms and their corresponding scope of application. (PSLO1, PSLO2)
4. identify the best solutions to a data science process from candidate models and algorithms. (PSLO2)

5. effectively present the results and execute the project to extract actionable insights for decision making. (PSLO 6)
6. work effectively in teams on data science projects to identify and resolve potential ethical and privacy issues in data science practice. (PSLO 6)

**Meeting & Assessing Student Learning Outcomes:**

The course learning outcomes will be evaluated in the following components:

(1). Three take-home assignments: data management, statistical modeling, and machine learning algorithms - assessing the subset of skills for the term team project. (20% each)
(2). Class participation and Contribution (10%)
(3). Final team project: reporting and presentation (30%)

**Attendance Policy:** Attendance is class is expected. Attendance will be recorded each class session. One unexcused absence is allowed with no penalty; every unexcused absence after the first will result in a deduction from the participation component of the course grade.

**Tentative Course Outline:**

## Part I. Overview and Foundations of Data Science (Weeks 1-4)

1. Introduction to Data Science and Tools
   - 1.1. Types of data science problems
   - 1.2. Core tools for data scientists
   - 1.3. Programming languages for modeling and data preparation
   - 1.4. Machine learning and data mining
   - 1.5. Data visualization and communication
   - 1.6. Basic understanding of software and database tools

2. *Quick Review* of Statistical Methods for Data Science
   - 2.1. Regression models
     - 2.1.1. Basic testing hypothesis
     - 2.1.2. Linear regression models
     - 2.1.3. Generalized linear regression models
     - 2.1.4. Regularized regression models
     - 2.1.5. Regression models with incomplete data
   - 2.2. Time series modeling: ARIMA methods and Non-parametric smoothing
   - 2.3. Basic nonparametric methods – smoothing techniques and A/B testing
   - 2.4. Imputation methods
   - 2.5. Basic Bayesian models and predictive analytics
   - 2.6. Methods for correlated data

3. Introduction to Machine Learning/Data Mining Algorithms for Data Science
   - 3.1. Decision-tree based regression and classification algorithms
     - 3.1.1. Averaging
     - 3.1.2. Boosting

- 3.2. Neural Networks
- 3.3. Probabilistic methods
- 3.5. Clustering methods
- 3.4. Other heuristic classification algorithms

4. Model KPIs
   - 4.1. Validation KPIs
   - 4.2. Testing and Production KPIs
   - 4.3. Structure of KPIs
     - 4.3.1. Probability-based KPIs
     - 4.3.2. Bias based KPIs
     - 4.3.3. Discriminatory KPIs

**Part II. Data Science Process and Case Studies (Weeks 5-10)**

1. Converting Business Questions to Data Science Problems.
   - 1.1. What is the business question?
   - 1.2. Whether the question is a stand-alone or part of serial questions?
   - 1.3. What is the business process/data generation process?
   - 1.4. What is the corresponding data science problem? Association? Prediction?
   - 1.5. Can the problems be solved using traditional statistical approaches?
   - 1.6. Whether the machine learning algorithm brings additional values to the problems?
   - 1.7. Whether the solution to the problem is a single model/algorithm or a set of inter-related models/algorithm (system)?
   - 1.8. Is this a real-time solution?
   - 1.9. Case Study

2. Data Collection/Sampling Design
   - 2.1. Where are data sources: internal and external data sources?
   - 2.2. Types of data: structured, semi-structured and unstructured data.
   - 2.3. Data governance, security, and storage (Data Lake, etc.).
   - 2.4. Data integration: ETL.
   - 2.5. Data cleaning and data pre-processing
   - 2.6. Case Studies

3. Exploratory Data Analysis
   - 3.1. Whether there are response variables? What are the response variables?
   - 3.2. Missing value issues - imputation?
   - 3.3. Distributions of attributes and modifying attributes for the needs of model building
   - 3.4. Simple association between variables.
   - 3.5. Data dependence and correlations.
   - 3.6. EDA techniques
   - 3.7. Identifying insights to support model selection.
   - 3.8. Insights for new feature creation.

3.9. Case Study

4. Feature Extraction/Selection

4.1. Re-thinking about the data generation process of the underlying problem;

4.2. Longitudinal and cross0sectional feature creation

4.3. Extracting/creating features

4.3.1. Statistical model-based feature creation (cross-sectional data)

4.3.2. Feature extraction based on sequence data

4.3.3. Addition indicator features via clustering analysis – heterogeneity reduction

4.3.4. Feature extraction for time series in machine learning

4.3.5. Introduction to automated feature engineering

4.4. Feature Selection methods

4.4.1. Domain knowledge based on feature selection

4.4.2. Information-based based feature selection

4.4.3. Regularized approaches such as LASSO related approach.

4.5. Case Study

5. Model building

5.1. Identify model type according to the business questions and information in the data.

5.2. Can traditional statistical model address the modeling question?

5.2.1. Association problem – regression models (cross-sectional and

5.2.2. Classification model – multi-variate regression models

5.2.3. Forecasting models – time series related models

5.2.4. Bayesian predictive models

5.2.5. Model diagnostics

5.3. Whether the question is better addressed by a machine learning algorithm?

5.3.1. Machine learning algorithms?

5.3.2. Various ensemble models.

5.4. Whether a combination of traditional statistical models and ML algorithms?

5.5. Rare event modeling and learning

5.5.1. Rarity Adjustments – sampling approaches

5.5.2. Likelihood/information approaches

5.6. Modeling building loop – general steps

5.6.1. Creation of training, validation and testing sets

5.6.2. Cross-validation for model selection and hyper-parameter tuning

5.6.3. Overfitting detection and prevention

5.7. Case Study

6. Communication and Visualization

6.1. Data science presentation styles

6.2. Communicate results with appropriate audience

6.3. Effective presentation

7. Model Deployment Best Practice

7.1. Specify the target model performance requirements

- 7.2. Separate prediction algorithm from model parameters
- 7.3. Develop automated tests to maintain model performance
- 7.4. Develop back-testing and now-testing infrastructure – maintain model performance
- 7.5. Challenge then trial model updates

8. Model Monitoring and improvement/revising
   - 8.1. Track and report progress against goals
   - 8.2. Learn from the model and find new insights
   - 8.3. Refine and update the objectives and the model

**Part III. Data Science Ethics (Weeks 11-12)**

1. Potential Ethical Issues in Data
   - 1.1. Privacy issues
   - 1.2. Biases in study design and data collection
2. Potential Ethical Issues in Reporting
   - 2.1. Failure to reporting negative results
   - 2.2. Cherry-picking impressive partial results
   - 2.3. Failure to investigate the model performance
3. Potential Ethical Issues in Algorithms and Models
   - 3.1. Biases in algorithm
   - 3.2. Algorithm and model interpretability
   - 3.3. Model and algorithm reproducibility

**Part IV. Overview of Data Science Infrastructure and Pipelines (Weeks 13-14)**

1. Data Science Infrastructure
   - 1.1. Data Storage
   - 1.2. Processing
   - 1.3. Platforms and (Analytic) Software
   - 1.4. Networking tools for data science components

2. Data Science Pipelines
   - 2.1. Data science platform
   - 2.2. Data science pipelines
   - 2.3. Real time analytics architecture

**Evaluation & Grading:**

A letter grade will be assigned based on performance in the course, according to the following scale:

| Grade | Quality Points | Percentage Equivalents | Interpretation |
|---|---|---|---|
| A | 4.00 | | Superior graduate attainment |
| A- | 3.67 | | |
| B+ | 3.33 | | Satisfactory graduate attainment |
| B | 3.00 | | |
| B- | 2.67 | | |
| C+ | 2.33 | | Attainment below graduate expectations |
| C | 2.00 | | |
| C- | 1.67 | | |
| F | 0 | < 70% | Failure |

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

## Statements Common to All WCU Graduate Syllabi:

### 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 the 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.

### 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.

### 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 meet this commitment and to comply with Title IX of the Education Amendments of 1972 and guidance from the Office for Civil Rights, the University requires faculty members to report incidents of sexual violence shared by students to the University's Title IX Coordinator, Ms. Lynn Klingensmith. 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 the webpage for the Office of Social Equity at http://www.wcupa.edu/_admin/social.equity/.

### 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.

### 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.

### 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.