Academic Calendar Entry DATA 101 (3) Basic Predictive Modelling An introduction to the techniques and software for handling real-world data. Topics include data cleaning, visualization, simulation, basic modeling, and prediction making. [3-1-0] Prerequisite: None. Course Overview, Content and Objectives The objective of the course is to introduce a combination of fundamental data science concepts to perform simple analyses of real-world and simulated data. The emphasis is on predictive validation rather than traditional statistical inference, which would involve p-values and confidence intervals. The course covers introductory aspects of supervised machine learning, such as regression, classification, and feature selection. Explorations in data collection, validity, and privacy are incorporated alongside other topics of the course to foster awareness of the growing importance of these issues. Students will be introduced to a variety of data sets, highlighting that the same concepts can be applied to several fields of study. Simulations, lab quizzes, and assignments are designed to emphasize the link between the application and the concept. Learning Outcomes After completing this course, students will be able to: a. apply analytic techniques to large and small data sets b. organize data using open source software c. use carefully constructed visualizations and appropriate numerical summaries to explore and/or present data. d. explain concepts and applications of simulations e. identify potential privacy and data concerns surrounding a data set f. identify and apply appropriate predictive techniques g. apply predictive validation techniques. Topics include 1 Introduction to Data Science and R, Objects, Computation and Manipulation, Software Basics I: installation, packages, data frames, arithmetic operators, reading external data, help files, Visualization Software Basics II: indexing, loops, apply functions, conditionals, logical operators, 4 Numerical Summaries Graphics: univariate, bivariate, multivariate/conditional plotting; basics of human visual perception and implications for graphic design, 5 Simulation: Regression Models Functions: declaring arguments, storing, results, using summary function and other, object-oriented methods; big data methods, Simulation: Nonlinear Models Simulations I: simple regression, based on additive simulated errors and noting effects on coefficient estimates and predictions, qqplots, dangers of extrapolation, Simulation: Tree models Simulations II: other random mechanisms that might lead to outliers and other forms of contamination; implications for data management and cleaning; nonlinear regression, including quadratic, exponential and some intrinsically nonlinear models. 8 Review and Midterm Simulations III: fitting classification and regression trees with CART using simulated and real data. 9 Measuring Errors in Prediction Scraping: rvest, sourcing public data, html and css selectors 10 Predictive: Multiple Regression and Diagnostics Predicting I: predict function, error calculations, such as mean-squared-error, mean-absolute-error, standard error, prediction sums of squares 11 Training and Testing Predicting II: multiple regression, diagnostics, visualization in two and three dimensions, including an introduction to RGL 12 Cross-Validation and Overfitting Predicting III: fitting and validation applied to both multiple regression and CART models 13 Review and Conclusion Model Selection: training/testing/validation, predicted error types; feature reduction