APPLIED DATA SCIENCE UG Minor

DSCI 353-453: Data Science Modeling, Prediction and Inference: for Energy & Manufacturing

Spring 2016 Tuesday, Thursday 11:30 am to 12:45 pm Olin 303

Prof. Roger H. French, rxf131@case.edu, 3 Credits

This counts as a 5th level class in the **Applied Data Science UG Minor**

For more information see http://datascience.case.edu/minor

Prerequisites:

Textbook: **OpenIntro Statistics**, D. M. Diez, C. D. Barr, M. Cetinkaya-Rundel

An Introduction to Statistical Learning with Applications in R, G. James, D. Witten, T. Hastie, R. Tibshirani

The Art of Data Science, R. D. Peng, E. Matsui

Report Writing for Data Science in R, R. D. Peng

Course description: Data science methods for inference, modeling and prediction.

In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference, modeling and prediction of the behavior of real energy and manufacturing systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assembly and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will also introduce machine-learning approaches for classification and tree-based methods. Results will be interpreted, visualized and discussed.

We will introduce the basic elements of data science and analytics using R Project for Statistical Computing. R Analytics will be applied to the case of energy systems (such as PV power plant degradation, and building energy efficiency) over time. And it will be applied to manufacturing systems to understand the principles of statistical process control and identify critical factors of variability and uniformity.

DSCI353/453: **Data Science Modeling, Prediction and Inference: for Energy & Manufacturing**

Roger French, Spring 2016
3 credit course, Undergraduate/Graduate
Instructor permission required
Format: Seminar

Course Description: Data Sci. Models & Prediction

In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference, modeling and prediction of the behavior of real energy and manufacturing systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assembly and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will also introduce machine-learning approaches for classification and tree-based methods. Results will be interpreted, visualized and discussed.

We will introduce the basic elements of data science and analytics using R Project for Statistical Computing. R Analytics will be applied to the case of energy systems (such as PV power plant degradation, and building energy efficiency) over time. And it will be applied to manufacturing systems to understand the principles of statistical process control and identify critical factors of variability and uniformity.

We will introduce the basic elements of data science and analytics using R Project for Statistical Computing. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and assembly functions, and a rich selection of statistical packages, used for data analytics, model development, prediction, inference and clustering. This will include an introduction to R data types, reading and writing data, looping, plotting and regular expressions, so that one can start performing variable transformations for linear fitting and developing structural equation models, while exploring for statistically significant relationships.

R Analytics will be applied to the case of energy systems (such as PV power plant degradation, and building energy efficiency) over time, by analyzing system responses, combined with results of experiments to identify fundamental principles that are statistically significant in the observed system performance. And it will be applied to manufacturing systems to understand the principles of statistical process control and identify critical factors of variability and uniformity.

Learning Outcomes:

Familiarity with an open-data tool chain including
R Statistics, scripting, functions, packages, automated data analysis,
git versioning and Rmarkdown reproducible data science.

Familiarity with exploratory data analysis to guide data analysis

Familiarity with inference and significance of sample results to populations

Familiarity with regression and linear and non-linear statistical model building
Including training, testing and validating dataset strategies

Applications of domain knowledge and statistical analytics
To identify important predictors and develop initial predictive models

Familiarity with clustering, self-similarity methods
For categorization by different distance metrics

Introduction to machine-learning approaches such as tree-based methods

Data types include:
Time-series, spectral, image and higher order datatypes,
And their assembly to produce augmented and derivative datasets.

Data set characteristics will include:
Variety: Of types of information, including both structured and unstructured data,
Volume: Data from human sources (vendors, suppliers, distributors, customers, etc.) and sensor networks of the energy system of factory, both small and large data volumes.
Velocity: Energy system and manufacturing supply chains changes will be included.

Topical outline:

Week	Topic
1	Data Science and Open Access
2	Data science tool chain and
3	Markdown documents for collaborative, reproducible research
4	Statistical Analysis for Data Science, study design
5	Inference for numerical data
6	Inference for categorical data
7	Linear regression and model selection
8	Multiple regression
9	Fixed and mixed effects models
10	Logisitic regression and non-linear models
11	Model selection and validation: Testing and Training datasets
12	Classification and self-similarity, applied to diverse data types
13	Resampling and Cross-validation methods
14	Machine learning and tree-based methods
15	Wrap up