

# SYST/OR 568. Applied Predictive Analytics

**Department of Systems Engineering and Operations Research  
George Mason University  
Fall 2016**

Introduces predictive analytics with applications in engineering, business, finance, health care, and social economic areas. Topics include time series and cross-sectional data processing, data visualization, correlation, linear and multiple regressions, classification and clustering, time series decomposition, factor models and causal models, predictive modeling performance analysis, and case study. Provides a foundation of basic theory and methodology with applied examples to analyze large engineering, social, and econometric data for predictive decision making. Hands-on experiments with R will be emphasized.

## Prerequisites

Graduate standing (Undergraduate engineering math: Calculus, probability theory, statistics, and some basic computer programming skills.)

## Course staff

### Course web page

**Lecture Notes:** Will be made available one-week in advance on [Bb](#)

**Instructor:** [Vadim Sokolov](#)

**Office:** [Engineering Building](#), Room 2242

[vsokolov\(at\)gmu.edu](mailto:vsokolov(at)gmu.edu)

**Tel:** 703-993-4533

**Office hours:** TBA.

## Lectures

**Location:** [Planetary Hall](#) 124

**Times:** 7:20-10pm on Thursday

## Grades

**Grade composition:** No in-class examination. Grade based entirely on participation in class, homework assignments, take-home midterm and final project.

## Textbooks

- Kuhn and Johnson, [Applied Predictive Modeling](#), Springer, 2013.
- Diez, Barr and Cetinkaya-Rundel [OpenIntro Statistics](#), OpenIntro, 2015
- James, Witten, Hastie and Tibshirani, [An Introduction to Statistical Learning with Applications in R](#), Springer, 2009.
- Hyndman and Athanasopoulos, [Forecasting: Principles and Practice](#), OTexts, 2013.

## Course Outline

- Data and Tools
- Predicting with probability
- Model diagnostics
- Linear regression
- Optimization
  - Regularization
  - Model Estimation
  - Robust Model Estimation
- Lasso and Model Selection
- Classification
- Time series forecasting

## Assignments

Students will have a take-home midterm exam and final project. There are approximately 5 homework assignments; students are encouraged to work in small groups. Each homework has 2-3 “theoretical questions” and 2-3 “hands-on” problems. Theoretical questions will be based on the material covered in class. Hands-on problems will require using verb|R| and routines provided by instructor to perform data analysis tasks. For the final project a student or a group of students can choose their own data set and a hypothesis to verify. Instructor will have 1-2 data sets/analysis problems, in case students have hard time identifying it on their own. Work on the final project can begin as soon as class starts. Each group will submit the final report (final week < 4 pages per student).

## Computing

You can choose which software you use. I recommend investing the time to learn R. Python is good choice as well. R is the dominant software package for real world Predictive Analytics and is used throughout other courses. This open-source software is available for free download at [www.r-project.org](http://www.r-project.org) and you can find documentation there.

A great way to start learning is to buy a book and start working through tutorials. A good guide is Adler’s [R in a Nutshell](#). They have many tutorials to help you get up to speed. You can browse other options by searching ‘R statistics’ on Amazon. If you are new to R (and even if not) you should complete a tutorial to familiarize yourself with the language. A great option is the [TryR](#) code school.

## Grading

Take home Midterm 30% Final project 40% + Homework 30%. Scores of each component are normalized to be out of 100. Grades will be posted on [Bb](#).

## Project

There are two options:

- Data analysis. You can deliver results in a form of a report, webpage (a blog post or a dedicated webpage), presentation, or video.
- Tutorial on a topic not covered in class. Same delivery options.

The data analysis project will be evaluated based on the following criteria:

- How adequate is the chosen analysis methodology for the problem at hand
- Level of sophistication of the analysis
- Correctness of the conclusions (or absence) made based on the data
- The presentation of the results

Results on the data analysis project should contain:

- Problem description and possible hypothesis
- Data description
- Methodologies used (larger groups should try different methodologies if applicable)
- Results and conclusion Final project attribution
- Each student documents his/her contribution
- Each student reports an evaluation for his/her group mates (blind)
- Individual grades will be based on full participation in the project