STAT 422/722: Predictive Analytics for Business
Syllabus, Fall 2017

Course Info

- **Course times**:
  Section 401: MW 9–10:30, 8/29 – 10/18, JMHH 265
  Section 403: MW 12–1:30, 8/29 - 10/18, JMHH 350
- **Instructor**: Yuancheng Zhu, `zhuyuanc@wharton.upenn.edu`
  Office hour: M 3–5 pm, JMHH 400
- **Teaching assistants**:
  - Justin Khim, `jkhim@wharton.upenn.edu`
    Office hour (R recitation): T 3–4 pm, JMHH F36
  - Hao Nguyen, `haong@wharton.upenn.edu`
    Office hour: TBD
- **Pre-requisite**: STAT 613/621, or knowledge of linear regression

Course Overview

This seven-week course introduces students to the statistical techniques that extend the ideas of regression analysis introduced in STAT 613. Digressing from traditional approaches that focus on carefully modeling how one or two chosen measurements relate to a response, we will take a "modern" approach applicable to managerial decision making in the presence of large data sets.

After a brief review of linear regression, we will round out our regression toolbox by learning how to build models for predicting categorical responses. Equipped with a solid foundation, we will switch our approach to the point of view of predictive modeling using automatic tools. The name of the game in predictive modeling is to be able to predict the behavior of new data. If, for example, we can show a bank how to predict who will default on a loan better than their existing system, the bank can increase profits. Similarly, if we help a company identify those in the market most interested in its products, then it can construct a much more focused product launch.

While our focus will center around prediction, managers still want to understand where their forecasts come from as there is still much hesitance in trusting pure "black-boxes" in many business scenarios. Hence, we will constantly explore and emphasize the trade-off between prediction power and model interpretability.

As the business world rapidly progresses towards a paradigm of data-driven decision making, the primary goal of this course is on understanding both the power and limitations of regression analysis. The course is designed to allow future managers–both data scientists and not–to communicate effectively with the data science team within an organization.

We are going to let software do the number crunching for us–our value-add comes from how to choose to tackle the problem and what insights we can draw from the model results.

Grading

- **Homework**: 60%

  There will be 5 weekly assignments.

  All assignments except the 5th one are due 11:59pm on Wednesday of that week. You need to submit your homework electronically via the Canvas site.

  The 5th homework contains two prediction problem and no written solution is required. The score will depend on the performance of the prediction submitted electronically. More details will be provided when the homework is posted.

  Late policy: Late assignments will be penalized at 10% of the maximum grade per day (24 hour) for up to two days (48 hours) and are ineligible to be handed in for credit after this time.

  Collaboration policy: Working together on homework is allowed and encouraged. However, students must write up their homework solutions by themselves. Names of collaborating students should be provided on the front page of each homework write-up.

- **Final exam**: 40%

  An in-class, open-book and open-computer final exam will be held in the last lecture.

  The exam will consist of small data analysis tasks. Students are supposed to bring their own laptop and perform data analysis using software during the exam.

Course Materials

- **Lecture slides** and **sample codes**

  The lecture slides will be the primary learning guide for the course and should be fairly complete. We will also make available tutorials on R related to the topics covered in class. The slides and codes will be posted by the morning of lecture (sometimes the night before) and you are encouraged to bring them to class as an aid.

- **Textbook**

  There is no required textbook for this course. Below are some recommended textbooks that cover the material presented in the context of much more extensive treatments of advanced modeling.

  – *An Introduction to Statistical Learning with Applications in R.*
    Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.
    Great textbook which covers basics of supervised learning (predictive analytics), with detailed examples and R codes.
    Free PDF version available at `http://www-bcf.usc.edu/~gareth/ISL`.

  - *Elements of Statistical Learning.*
    Trevor Hastie, Robert Tibshirani and Jerome Friedman
    More advanced textbook on machine learning.
    Free PDF version available `http://web.stanford.edu/~hastie/ElemStatLearn/`.
  - *Machine Learning: a Probabilistic Perspective.*
    Kevin Murphy.
    A more complete textbook with a background introduction in probability, linear algebra, calculus, and programming.

- **Computer software**

  The software to be used in this class is R. The focus of the course, however, will be on teaching predictive analytics rather than how to use R.

  Previous experience with R is *not* required. We will post sample codes, tutorials, and the TAs will hold recitations to provide necessary help with programming.

  The first R recitation 3-4 pm on Tuesday, Sept 5th, will be a tutorial on getting started with R.

Schedule (tentative)

Event	Date	Topics	Assignment
Lecture 1	Wed, Aug 30	Course overview	
Lecture 2	Wed, Sept 6	Linear regression	Assn 1 out
Lecture 3	Mon, Sept 11	Logistic regression	
Lecture 4	Wed, Sept 13	Classification	Assn 1 due, Assn 2 out
Lecture 5	Mon, Sept 18	Out-of-sample validation	
Lecture 6	Wed, Sept 20	k-fold cross validation	Assn 2 due, Assn 3 out
Lecture 7	Mon, Sept 25	Subset selection	
Lecture 8	Wed, Sept 27	Shrinkage method	Assn 3 due, Assn 4 out
Lecture 9	Mon, Oct 2	Model selection	
Lecture 10	Wed, Oct 4	Decision tree	Assn 4 due, Assn 5 out
Lecture 11	Mon, Oct 9	Random forest	
Lecture 12	Wed, Oct 11	Boosted trees	
Lecture 13	Mon, Oct 16	Review	
Lecture 14	Wed, Oct 18	Final exam	Assn 5 due