This syllabus is subject to change. Students will be notified of changes by email.

Instructor: B. D. McCullough, bdmccullough@drexel.edu
Office: GHall 742, office hours: Tu, 5:30-6:30.

If you send me an email, put “STAT 642” in the subject line. Also: if you send me an email from a non-Drexel account and my reply to that email bounces, I will not try to resend it.

**REQUIRED SOFTWARE:**

1. R/Rstudio
2. Rattle
3. Tableau

**Tableau Information:**

- Desktop Key: TDKQ-8166-E1E0-D9CD-AA4B
- Instructions: Click on the link above and select Get Started. On the form, enter your university email address for Business email; and under "Organization", please input the name of your school.
- Tutorials: [http://www.tableau.com/learn/training](http://www.tableau.com/learn/training) and use “On-Demand Training” which is free.

**REQUIRED BLOG:** www.r-bloggers.com

**REQUIRED TEXTS:**


**RECOMMENDED TEXTS:**


**WEEK 1: Introduction**

R/Rstudio, missing data; ggplot2; dplyr; reshape; Tableau; train-test-validate

ASSIGNMENT: watch tableau and dplyr videos; install Rstudio, dplyr, ggpplot2, Tableau. Explore the home equity dataset.

WEEK 2: data exploration, Trees, more missing data, the mi package
Ledolter ch. 13, Williams ch. 11, Linoff ch. 7

ASSIGNMENT: do something about the missing values in the home equity dataset
READEINGS: Gschwind (2007), Nie, Rowe, Zhang, Tian and Shi (2011)

WEEK 3: Logistic Regression and Model Assessment
Ledolter ch. 7; Ledolter ch. 8, Linoff pp. 180-190, Williams ch. 15


WEEK 4: Neural Networks and the Lasso
Linoff ch. 8, “neuralnet: Training of Neural Networks” by Frauke Gunther and Stefan Fritsch,” The R Journal 2(1), 2010, pp. 30-38; Ledolter ch. 6


WEEK 5: Boosting, Bagging and Random Forests
Williams ch. 13 (boost) and 12 (rf).

READEINGS: Ranney (2010), Koh and Tan (2005), Liaw and Weiner (2002), Berk (2008), Altendorf et al. (manuscript) watch Schapire’s 1 hour video on boosting: [http://videolectures.net/ml](http://videolectures.net/ml)

WEEK 6: Nearest Neighbors and Clustering, Network Analysis
Ledolter ch. 9 (nn), 15 (cl), 20 (net), Williams ch. 9 (cl), Linoff ch. 9 (nn), 13 (cl), 16 (net)

READEINGS: Shih, Chiang Hu and Chen (2011)

WEEK 7: Discriminant Analysis and Dimension Reduction
Ledolter ch. 12 (da) and 17-18 (dr), Linoff ch. 20 (dr)

READEINGS: Elder (2009)

WEEK 8: Support Vector Machines; using AWS in the cloud

READEINGS: Ben-Hur and Weston (2010)

WEEK 9: Marketbasket Analysis
Ledolter ch. 16, Williams ch. 10, Linoff ch. 15
READINGS: Kaufman, Rosset and Perlich (2012)

Week 10, week of 25 Nov: Naive Bayes, Reject Inference
Ledolter ch. 10, Linoff pp. 210-213; readings for Reject Inference to be specified

Final Exam Week: Each team will make a presentation of its project, as well as turn in its written report.

Grading: Grades will be based on a final project that is due at the time of the final exam. The class will be split into teams, and will work on a common dataset. The team will prepare a report on the project, including an executive summary that is suitable for reading by a non-technical manager. The report is worth 1/3, the summary is worth 1/3, and the predictive accuracy is worth 1/3.

The project must include:

- At least one non-trivial Tableau graphic
- the “data.table” function must be used
- The primary means of producing graphics must be ggplot2
- dplyr must be used

Reading List

Altendorf, Eric, Peter Brende, Josh Daniel and Laurent Lessard (manuscript), “Fraud Detection for Online Retail using Random Forests”


Elder, Jon (2009), “Top Ten Data Mining Mistakes”, Chapter 20 in Handbook of Statistical Analysis and data Mining Applications by Nisbet, Elder and Miner


Nie, Guangli, Wei Rowe, Lingling Zhang, Yingjie Tian and Yong Shi (2011), “Credit card churn forecasting by logistic regression and decision tree,” Expert Systems with Applications 38, 15273-15285


