C744 – Course 8 "Data Analytics: Data Mining and Analytics 2" (3cr)

- **43hr/cr – 131hr (4.5wk) – 55pg paper/code (100hr) – 3hr Video Lectures – 27hr Textbook & Labs**
- ***Definition:*** *Examines application of descriptive and predictive data mining techniques to reveal info; factor analysis, cluster analysis, classification methods, and neural networks to limit human subjectivity in decision making process. Implement descriptive data mining methods. Assess data mining model performance and application. Methods and software for data mining projects. Implement classification and prediction data mining methods..*
- **Course Setup / Plan of Action / Installation:** 2.75hr (2d) – 6/10 : 6/11
  - DONE – 0.50 hr – Update this doc all over, close prior course + planning next
  - DONE – 2.25 hr – Prepare my Course Study Plan by reviewing all materials and listing out below
- **Course Material, Labs & Quizzes, Take Notes**, 12hr (4d) – 11.85hr Textbooks – 06/12 : 06/15

  Section 1 & 2, Descriptive Methods – 4hr, Reading=4hr
  - DONE – 0.10 hr – LINK – Intro Text
  - DONE – 0.50 hr – LINK – Section 2 PPT Overview
  - DONE – 0.50 hr – LINK – Topic 1, Chapter 7, Factor Analysis; Monkeyed around with browser = Edge not scrolling for 15min; then switched to Google Chrome with no problems
  - DONE – 0.75 hr – LINK – Topic 1, Lab Exercises (take notes on the code for R, but dry lab the rest due to issues with environment)
  - DONE – 0.50 hr – LINK – Topic 2, Overview + Chapter 9, Cluster Analysis (turning on read aloud allowed me to use fast forward to read and scroll…work around broken UI)
  - DONE – 0.75 hr – LINK – Topic 2, Lab Exercises (read everything / take notes, but dry lab it due to env issues; will "do" when building project and re-walk the lab steps then)
  - DONE – 1.00 hr – LINK – Topic 3, Overview + Chapter 10, Association Analysis

  Section 3, Classification and Prediction Methods – 6hr, Reading=6hr
  - DONE – 1.75 hr – LINK – Topic 1, Overview & Chapter 11.4/5/6
  - DONE – 2.25 hr – LINK – Topic 2, Overview Linear Regression & Chapter 11.7 + Lab Exercises
  - DONE – 1.00 hr – LINK – Topic 2, Overview Logisitc Regression & Chapter 11.8 + Lab Exercises
  - DONE – 1.00 hr – LINK – Topic 2, Chapter 11.9, 11.10, 11.11, 11.12, and 8.1 (yes out of order)

  Section 4, Model Performance and Application – 1.75hr, Reading=1.75hr
  - DONE – 1.25 hr – LINK – Section 4 Overview & Chapter 11.14, 11.15
  - DONE – 0.50 hr – LINK – Topic 2, Chapter 11.16.xx
- **Supplemental Training to Course Materials,** 15.25hr (4d) – 2.75 hr Video – 12.5hr Textbooks – 06/15 : 06/18

  Mentor & Instructor Early Emails Project Suggestions – Reading=1.65hr
  - DONE – 0.25 hr – LINK – **Sample for Project**: Advanced Modeling in R to Predict Customer Churn – Logistic Regression, Decision Tree, & Random Forest; Consider Random Forest in PA project
  - DONE – 0.15 hr – LINK – Code for all the big Data Science to run all the models – outstanding! (Reference, use it in the project below and probably forever going forward)
  - DONE – 0.25 hr – LINK – Wikipedia definition of Linear Algebra (level set, skim read…should have read back in earlier courses; esp. statistics)
  - DONE – 1.00 hr – LINK – sthda.com - Another PCA Guide with R code (xref in project)

  Tip Sheet #2 – Videos=0.85hr; Reading=0.40hr
  - DONE – 0.60 hr – LINK – PCA Course by French instructor (www.r-bloggers.com, 2x speed, 3 videos + Handling missing values, skim b/c goes beyond what need)
  - DONE – 0.25 hr – LINK – More PCA Training videos (Video#1 same as above, Video #2=FactoMiner, Video #2=FactoShiny, 2x speed, skim b/c goes beyond what need)
  - DONE – 0.10 hr – LINK – Article (Reading) sthda.com PCA videos (same 5 videos in two links above + better written R code) – Use this for coding project
  - DONE – 0.30 hr – LINK – Slides (Reading) Sebastien Ledien Unofficial FactoMineR; PCA, CA, MCA, several other links; LINK to R Code Using FactomineR for project

  Student Advice Guide – Reading=2.00hr
  - DONE – 0.50 hr – LINK – Logistic Regression Video with R (sample code at 43min & 45min) – Use in project
  - DONE – 0.25 hr – LINK1, LINK2, LINK3, LINK4 – Google "Rmisc multiplot" … or use grid.arrange(p1, p2, ncol=2)
  - DONE – 0.25 hr – LINK1, LINK2 – Google "variable/model selection with stepwise (try step function in base library)" (link 1 = return here when writing code for project)
  - DONE – 0.25 hr – LINK1, LINK2, LINK3 – Google "r goodness of fit confusion matrix"; link1 = using caret to fit, and train to test different models; link2 = calc confusion matrix
  - DONE – 0.25 hr – LINK1, LINK2 – Google "data science" vs. "big data analytics"
  - DONE – 0.25 hr – LINK1, LINK2 – Google "data engineer vs. business intelligence"
  - DONE – 0.25 hr – LINK1, LINK2, LINK3, LINK4, LINK5 – Google "how to choose statistical tests"

Course Chatter – Videos=1.75hr, Reading=2.10hr

Reading: (2.10hr)

- DONE – 0.25 hr – LINK – Which is better discussion: R or Python for Data Science project
- DONE – 0.10 hr – LINK – Stack Overflow note regarding resolving Caret library installation issue
- DONE – 0.50 hr – LINK – Lynda Foster's Reddit Comments (excellent), pull ideas out for my project checklist/self-rubric (plus added 8+ links below from here)
- DONE – 0.35 hr – LINK – Huge, detailed data mining guide – The CRISP-DM Process Model (xref video below) (use pg 12, 13, 17, 20, 23, 26, & 28 diagrams to help guide my project work; pg 48 for technique selection)
- DONE – 0.15 hr – LINK – Kaggle competition that closely matches this assignment (skim read thru, maybe return later for deeper dive)
- DONE – 0.25 hr – LINK – Statistical Identity of Data
- DONE – 0.15 hr – LINK – Comparing data mining methods for binary response variables (skim read, great idea to use stats like boxplot to compare results of several models)
- DONE – 0.20 hr – LINK – Working with categorical variables (good article, must "treat" cat. Vars: convert to numeric, mean if range bucket…or lower + upper bound new vars, reduce levels eg: rollup zipcodes to cities, dummy codes)
- DONE – 0.15 hr – LINK – Guide to Efficient Model Building (copied flow chart + code to model my project off)

Videos: (1.75hr)

- DONE – 0.50 hr – LINK – YouTube CRISP-DM Framework playlist (68 videos, at 2x speed)
- DONE – 0.50 hr – LINK – Strategies for building predictive models (58min, 2x speed, took screenshot notes for project)
- DONE – 0.60 hr – LINK1, LINK2 – Building predictive analytics capabilities (1hr 52min, 2x speed); Link2 = Using R (just super quick skip thru slides)
- DONE – 0.15 hr – LINK – YouTube Video on Churn (20min at 2x) Good end to end model shown (4th example = prediction table + logistic regr. + decision tree)

Suggestion = Research Churn Predictions Papers or with R (Google) – Reading = 6.35 hr

Important: *skim read now, heavier re-read while doing project and picking out ideas in section below*

- DONE – 0.10 hr – LINK – Matt Dancho Deep Learning model for predicting customer churn on same IBM educational telco churn dataset
- DONE – 0.75 hr – LINK – "Predict Customer Churn with R" (logistic regression, decision tree, and random forest) – AWESOME Example Code + concepts + explanations
- DONE – 0.25 hr – LINK – This seems like a copy of the R churn model article above…but adds a check for missing values (the precursor above did not) Review this
- DONE – 0.75 hr – LINK1, LINK2 – John Sullivan Example project in R (example of imputing missing values using mean, research Q's, try alter # variables, # trees, etc.); concept = optimize (tune) ML model for accuracy
- DONE – 0.10 hr – LINK – Create better data science projects with business impact: Churn Prediction with R
- DONE – 0.35 hr – LINK – Churn Analysis. Concepts: divide customers into 3 groups: active, inactive, about to churn; ignore involuntary churn; uses 5th model survival
- DONE – 0.15 hr – LINK – Job interview take home Churn Analysis quiz. Value adds = date munging (not seen in prior models), definition of churn (specific to problem solving, is flexible), seasonality (Nov/Dec = 2x higher); lead indicator vs. moving avg
- DONE – 0.10 hr – LINK – Article by Mandy Gu. Key Value Add = pair plot (churn vs. no churn) for analysis – now I see prior articles did same; is in python so convert to R
- DONE – 0.25 hr – LINK – Python Article, key take-away = spend time analyzing existing data / churn rate by variables to see whether any stand out graphically (esp. tenure in a scatter plot, plus some categorical vars in bar graphs)
- DONE – 0.10 hr – LINK – Churn Prediction Model Diagram – idea = make my own flow diagram after building my model, put it in the paper
- DONE – 0.10 hr – LINK – Concept = % Correlation between data points and attrition; leads to Q of whether to do PCA on the variables so still can whittle down to 3-4-5 variables, but since PCA then actually opinion of 6-8-10 weighing in
- DONE – 0.05 hr – LINK – Nice transformation list: drop irrelevant columns, handle missing values, convert objects to numerical values, convert categorical data into numerical values, split dataset (training/test)
- DONE – 0.50 hr – LINK – PowerPoint slides (Biz Logic) on customer churn (Outstanding details to use in project, slides 2,3,4,6,10,11,12,19)
- DONE – 0.50 hr – LINK – Churn Prediction (Python) – Learn how to train a decision tree model for churn prediction, Focus on the Data Preparation / Cleanup steps for ideas
- DONE – 1.00 hr – LINK – "Less Hacky" churn prediction (avoid the pitfalls) – Too advanced? Interesting tips (ie: predict non-churn as opposed to churn…invert the measure to actual events, like clicks or purchases or logins; predict TTE)
- DONE – 0.50 hr – LINK – "Using Deep Learning to Predict Customer Churn in a Mobile Telecom Network" – Too advanced (deep learning), but great idea: New >> Active (event) >> Inactive (no events, age) >> Churn (>q days w/o event)
- DONE – xref below – LINK – "Customer Churn – Logistic Regression with R" (hours included below during perf asmt)
- DONE – xref below – LINK – "Customer Churn – Exploratory Data Analysis" (hours included below during perf asmt)
- DONE – xref below – LINK – "Customer Churn – Prediciton with Logistic Regression and Random Forest" (hours included below during perf asmt)
- DONE – xref below – LINK – "Using MCA and variable clustering R for Insights in Customer Attrition" (hours included below during perf asmt)

Don't Bother: (Reading - 0.65 hr)

- DONE – 0.00 hr – LINK – Same Model Map as prior data mining course…skim read as needed
- DONE – 0.00 hr – LINK – FactoMineR Free Tutorials (Duplicate of all several above videos and links, a home page for it all…thus skip)
- DONE – 0.25 hr – LINK – Survey on Customer Churn Prediction using Machine Learning Techniques ("binary classification task", ie: #1 = logistic regression?)
- DONE – 0.10 hr – LINK – Kaggle Search on "Customer Churn Prediction" yields 2,436 results…skim thru some for ideas
- DONE – 0.25 hr – LINK – IJETR (Journal) PDF, nada

- **Performance Assessment (Project & Paper)**, 100.25hr (4.5wk) – 74.0 hr Researching – 26.25hr Writing 55pg Paper/Code – 06/18 : 07/10

Step 1 - Initial Research to Populate Project Checklist: 10.25 hrs

Flush Out Project Checklist details before even writing 1 line of code

- DONE – 1.75 hr – Read Course Rubric, download data file "telco CSV" from project website; decision use XL for checklist (high level waypoints)
- DONE – 2.25 hr – Add to Project Checklist, Rubric elements in as base framework, read it, understand it, re-organize it into execution plan
- DONE – 0.75 hr – Add to Project Checklist, review prior course checklists, pull in any good general ideas, add file APA ref
- DONE – 1.00 hr – Add to Project Checklist, TIPS sheets 1 and 2 from mentor– ESPECIALLY the Guide #2 "Quick Start R Code"
- DONE – 0.50 hr – Add to Project Checklist, Student Advice Guide notes
- DONE – 1.25 hr – Add to Project Checklist, Pull out anything relevant from Course Chatter (giant list, I know)
- DONE – 0.25 hr – Add to Project Checklist, Videos & Articles from Mentor emails + Tip Sheets
- DONE – 0.75 hr – Add to Project Checklist, Videos & Articles from Advice Guides & Course Chatter
- DONE – 1.25 hr – Add to Project Checklist, Videos & Articles from Research Papers on Churn on the Web
- DONE – 0.50 hr – Add to Project Checklist, My Notes from Reading Course Material and watching videos, add anything pertinent to design checklist

Step 2 – Execute & Answer Rubric Tasks (R Code + extra details + screenshots to Project Checklist)**: 63.50 hrs**

This worked out really well using XL as a giant checklist to gather all the research info before reporting on it.
Copy-pasted code snippets, screenshots, text from research, etc. into the checklist first then again later into the final report.
3 columns: (a) Status (TODO, IP, DONE), (b) the information, (c) APA Reference

- DONE – 0.50 hr – Task "0" – **Coding** to Import Data
- DONE – 2.00 hr – Task "A" – **Research** for Paper: Download file, Import to data frame in "R", List Benefits; get process established for working thru tasks
- DONE – 3.00 hr – More research around churn prediction models
- DONE – 1.50 hr – Task "B" – **Research** for Paper: Goal / Objectives
- DONE – 2.75 hr – Task "C" – **Research** for Paper: Prescribed (Planned) Analysis
- DONE – 0.75 hr – Task "D" – **Research** for Paper: Target Variable
- DONE – 1.50 hr – Task "E" – **Research** for Paper: Predictor Variables
- DONE – 0.50 hr – Task "F" – **Research** for Paper: Data Exploration, Goal
- DONE – 1.00 hr – Task "G" – **Research** for Paper: Data Exploration, Statistical Identity
- DONE – 8.75 hr – Task "H" – **Coding + Research** for Paper: Data Exploration, Cleaning
- DONE – 2.25 hr – Task "I" – **Coding + Research** for Paper: Univariate Statistics Summary
- DONE – 14.50 hr – Task "J" – **Coding + Research** for Paper: Bivariate Statistics Summary + Redo Prior Work above as learn from mistakes downstream
- DONE – 17.50 hr – Task "K" – **Research** for Paper: Analytic and Evaluative Models Fit + MANY iterations of re-doing the work above as find errors and omissions
- DONE – 6.25 hr – Task "L" – **Research** for Paper: "Analysis Justification" for Model Selected (Confusion Matrix for Model Quality)
- DONE – 0.25 hr – Task "M" – **Research** for Paper: "Visual Justification" for presenting data
- DONE – 0.25 hr – Task "N" – **Research** for Paper: "Model Phenomenon Detection" – Discriminating or Not?
- DONE – 0.25 hr – Task "O" – **Research** for Paper: "Best Predictor Variables"
- DONE – 0.00 hr – Task "P" – APA References – Have been building this all along as I go…time spread across prior tasks
- DONE – 0.00 hr – Wrote Code as did each step above, so time built in there

Step 3 – Write Paper from Checklist & Code Above: 25.75 hrs

- DONE – 25.75 hr – Paper + Finalize + Submit

Step 4 – Cleanup Notes & Update Course Chatter: 0.75 hrs

- DONE – 0.75 hr – Tabulate final results, update my docs; add to "Study Plan" and "Project Checklist Template" (empty) to Course Chatter