Data Mining in Foresight

Course Title	Data Mining and Analytics in Foresight
Course Time	5:30 - 8:30 PM, CST
Online	Canvas and Blackboard

Instructor

Anne Boysen
ahboysen@uh.edu
+512.568.4941
Skype: annehboysen

The course will prepare you to

- Understand key concepts in Data Mining
- Create basic Machine Learning models used in AI systems today
- Use algorithms that have helped the world's most profitable companies
- Extract and analyze text based data
- Master a skill in high demand by employers
- Take your own research to the next level

Overview

The world is drowning in data. How can we extract relevant data and use data mining techniques and machine learning to expand our knowledge base in Strategic Foresight?

This course is a pioneering approach to integrate Data Analytics with Foresight. It will allow you to not only speculate about the future of artificial intelligence, but teach you how to use AI-methods in your own foresight work. In essence, you'll learn what has been called the sexiest job skill of the 21st century - Data Science.

Prerequisites

While Data Science is deeply embedded in computer science, advanced math and statistics, this course does not require higher level mastery of these disciplines. However, to do well in this class a comprehension of basic concepts in algebra and statistics will be helpful. While you will learn to use basic tools in data analytics, this class will not make you a Data Scientist, which typically requires years of study and proficiency in calculus, linear algebra, probability statistics and mastery of several programming languages.

Approach

Weekly online lectures	• Discussion of last weeks assignment and topic • Demonstration on dataset in Rapidminer or R • Conceptual understanding of weekly topic
Weekly home assignments	• Replicate Data Mining Model. Submit screen capture to demonstrate model • Participate in weekly Canvas discussions
End of Semester Paper	• Demonstrate understanding of course material and the role of data mining in foresight

Grading

Weekly model assignments	50%
Weekly discussions & class participation	15%
End of Semester Paper	30%
End of semester survey	5%

Late submission will reduce 10% of the grade

Class 1: Introductions and objectives. Rapidminer & R, resolving lingering set-up issues

Objective: Get to know each other. Understand our backgrounds. What do you want to get out of this class? How do you hope to use it within your foresight domain?

Introduce myself and my background. Set realistic expectations on 1) what I can teach and what will be possible to learn in the amount of time; and 2) the opportunities and confines of what DM/ PA suggest for futurists.

General discussion on what Data Science can and cannot do for futurists now and in the future. We discuss why most disciplines, especially information intensive fields like foresight, are starting to include more statistical and automated approaches to handle data and overcome cognitive bias. Compare with how the business world and social scientists currently use these methods.

Discussion topic: Introduce yourselves on Canvas Discussion

Class 2: Data mining in exponential times. Types of data, types of models and processes

We will learn about types of data, models, cross validation methods for overfitting and underrating. Continue discussion on implications for foresight and other purposes.

Discussion topics: What types of data are useful to help environmental scanning? Structured? Unstructured? What machine learning techniques do you think are most useful - supervised or unsupervised? What are the benefits and drawbacks do you see with supervised and unsupervised learning? Why is cross validation important?

Class 3, Cluster Analysis and Association Rules

We dive into unsupervised learning and build our first models! Have you ever been recommended a product or service based on what you bought in the past? And did you know you can find natural segments with clustering techniques? We will learn these models.

Discussion topics: What patterns did you see between the community groups in the association model? How can association rules be used when looking for social trends or in business? How would you use clustering in your own research domain?

Class 4, Decision Trees

We begin our supervised learning session by building our first predictive model, a Decision Tree. We learn about pruning, performance testing and find with which confidence our model managed to predict future actions for each individual unit in our dataset.

Discussion topics: What can you learn about technology adoption from your decision tree? What characterizes an Early Adopter of technology customers? Which variables are associated with Late Majority? Look at the predictions for the target variable. Why is the use of the word prediction in this context, i.e. *Predictive Analytics*, different from most definitions of “prediction” used in foresight? Is the model still useful for futurists?

Class 5, Artificial Neural Networks

Did you know most AI systems today are based on the human brain? This week we explore the ML architecture behind self-driving cars, image recognition and speech recognition - Artificial Neural Networks (ANNs). We will build our own small neural networks and process datasets with thousands of rows! The class begins with a brief introduction of the history of neural nets, how they are built, and get an overview of the mathematical and statistical processes that make them work.

Discussion topics: Why are neural networks so useful for advanced data analytics and artificial intelligence. Can you find types of deep learning models used for image classification and voice recognition? What did you learn from your own neural networks? What did you learn about your churn model by looking at *recall* and *precision* in your Performance operator?

Class 6, Text Mining and NLP in R

Natural Language Processing is a major component of machine learning and AI today. People express themselves through free text, often in social media, and we can learn a lot about human behavior and attitudes by analyzing large quantities of these wells of unstructured data. A subfield of NLP and text mining is sentiment analysis and emotion analysis. This week we will extract real time data from Twitter around the topics you are most interested in. We can compare sentiments around your topic before and after a recent news event, in several languages of the Roman alphabet, and compare the various types of emotions people express.

Please be aware that doing the computing yourself is optional.
We move away from Rapidminer to the R coding environment this week, so taking part in the computing part in this class requires quite a bit of setup and preparation. You need developer account with Twitter, receiving an API key, setting up R and RStudio on your local computer, get familiar with the environment and manipulate the results in Excel or Google Sheets using some basic Excel functions. Since time is and this is not a CS class, I have decided to give students

the option of having the text mining operation completed on my end. You will still need to interpret your results, but you can use this analysis in your final paper. If you decide not to set up your local computing environment, you should still follow the instructions to get familiar with the text mining process.

Discussion topics: What does the ubiquity of text based data mean for the ratio between structured and unstructured data in the wild? Do you think the prevalence of user generated content contributes to the rapid growth of unstructured data? How can this resource help us glean weak signals insights in scanning?

Class 7, Term paper due & Foresight in Machine Learning

You should demonstrate that you know the different types of data and data mining methods. You should demonstrate that you understand the opportunities and limitations Data Mining can offer in Foresight. Do you see an opportunity for data mining in your own research domain/ industry/ line of business?

After a brief class discussion on how what you learned this semester, we turn the tables and explore if futurists could or should help ML professionals inform their algorithms to be more "futures directed" and fair.

Reading Materials

Required:

- "Data Science vs Machine Learning vs Data Analytics vs Business Analytics" on KDNuggets, by Iliya Valchanov, 365 Data Science at https://www.kdnuggets.com/2018/05/data-science-machine-learning-business-analytics.html
- *Data Mining for the Masses, Third Edition: With Implementations in RapidMiner and R*, by Matthew North

- *Extending the Knowledge Base of Foresight: The Contribution of Text Mining,* vorgelegt von, Victoria Kayser, M. Sc., geb. in Böblingen (will be emailed students), Potential of Text Mining in Foresight, p 3- 11 & 57 - 68

Recommended:

"Data Science Predicting The Future", at KDNuggets by Iliya Valchanov, 365 Data Science at https://www.kdnuggets.com/2018/06/data-science-predicting-future.html

- *Data Mining Techniques,* Third Edition by Gordon S Lindoff and Michael J.A. Berry

- *Data Mining for the Social Sciences: An Introduction,* First Edition, by Paul Attewell (Author), David Monaghan (Author)

- *Social Media Mining with R,* by Nathan Dannemann and Richard Heimann

- *Everybody Lies: Big Data, New Data and What the Internet Can Tell Us About who We Really Are*, by Seth Stephens-Davidowitz

- *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, by Eric Siegel