**Introduction to Data Science Course Outline**

Andrew Wheeler, PhD, [andrew.wheeler@hms.com](mailto:andrew.wheeler@hms.com)

# Lesson 01: Data Science 101

* Intro to Data Science Team
* What is Data Science
  + Predictive modeling
  + Cost-benefit analysis
  + Experimental Evaluation (e.g. A/B testing)
  + Augmenting Human Processes
* Overview of Types of Prediction Tasks
  + Supervised learning (know the outcome)
  + Unsupervised learning (unlabeled data)
  + Reinforcement learning
* Typical Data Science Process Work Flow
* Simple Data Analysis Example in Python
  + Reading in data
  + Exploratory Data Analysis (EDA)
  + Fitting a regression model
  + Making Predictions

# Lesson 02: Machine Learning 101

* Prediction vs Inference
  + Bias/Variance trade-off
  + Cost function
  + Stats vs Machine Learning
  + IIA
* Supervised Learning Models
  + Regression
    - Linear for continuous outcomes
    - Logistic for categorical outcomes
  + Tree based models
    - Random Forest & Ensembles
    - Boosting
  + K-nearest Neighbors / SVM
  + Neural Networks/Deep Learning
* Other models
  + Unsupervised Learning
    - Latent Variables
  + Reinforcement Learning
    - Multi-armed bandits
  + Recommender Systems
* Interpretability vs Black Box
* Example Regression vs Random Forests in Python

# Lesson 03: Evaluating Predictions

* Evaluating Predictive Models
  + Test/train (in sample is optimistic)
  + Weighing false positives & false negatives
  + ROC and AUC
  + Positive Predictive Value is dependent on prevalence (“mix %”)
  + Simple models as baseline
  + Continuous Loss functions
* Example binary prediction in Python
  + Creating test and training samples
  + Fit logistic model and random forest model
  + Compare in-sample vs out-of-sample
  + Cost-benefit analysis of false-positives vs false-negatives

# Lesson 04: Intro Data Transformation in Python

* Data Types
  + Numeric, Categorical
* Data Wrangling
  + Duplication
  + Aggregation
  + Reshaping data (Pivot)
  + Stacking & Merge
* Data Normalization
  + Outliers
  + Transformations (Log, Square Root, Box-Cox)
  + Standardizing [0-1 vs 0-100]
  + Z-Scoring
* Intro to Data Pipeline / ETL
* Example in Python

# Lesson 05: Data Visualization 101

* Visual Processing
* Hierarchy of Data Visualization
* Color Advice
  + Color blindness
  + Printing / Presentation
* Making nice tables
  + Comparisons across rows vs columns
  + Aligning numbers
  + Limiting Digits
* Examples in Python

# Lesson 06: Feature Engineering

* Motivation
  + Understanding causal mechanisms
  + What impacts outcome, as well as functional form
  + Importance of Business Domain Knowledge
* Creating new data
  + Polynomial & Spline terms
  + Dummy variables
  + Interactions
  + When it is necessary (regression) vs not (tree-based)
* EDA
  + Smoothed plots
  + Binning for interactions
  + Small multiple plots
* Understanding feature importance in Machine Learning Models
  + Feature importance for prediction
  + Marginal Effects
* Example in Python
  + Feature Tools (??Python??)

# Lesson 07: Missing Data

* Understanding why missing data occurs
  + Missing = 0, or missing is unknown, or missing is N/A
* Ways of Encoding Missing Data
  + Dummy variable and interaction trick
* Imputation Strategies
  + Caution with using mean/mode imputation
  + Dropping cases/columns
  + Predicting missing cases using Machine Learning
  + Multiple imputation is for inference, not for prediction
* Example in Python

# Lesson 08: Big Data and Parallel Computing Intro

* Subsampling (working with data in chunks)
  + Stratified sampling for rare outcomes
  + Adjusting predictions based on sampling
    - Case/control
    - Raking to population
    - Weighting ML models
* SQL vs inside Python
  + Working with already aggregated data
  + Turning models into SQL code
* HDF5 & MapReduce
* Hive/Spark/Clusters
* Sparse matrices
* NoSQL solutions
* Example in Python

# Lesson 09: Dimension Reduction and Unsupervised Learning

* Too many independent variables
  + Feature selection (regularized models)
  + Dimension reduction via Principle Components Analysis (PCA)
* Unsupervised Learning
  + Latent Categories (Clustering)
  + Latent Continuous values (IQ)
* Example PCA in Python

# Lesson 10: High Cardinality (Many Categories)

* Types of Many Category Data
  + General concept of handling high cardinality
  + Diagnoses Codes
  + Geographic Data
* Many Categories for Outcomes
  + Multinomial Logistic Regression
  + Reformulating as a Logit model
  + Posterior probabilities and Assigning a category
* Many Categories for Independent Variables
  + Theory of why traditional encoding does not work
  + Reduced encoding of data subsets
  + Hierarchical Models for predicting new categories
  + Association Rules
* Examples in Python

# Lesson 11: Intro to Forecasting

* Goals of Forecasting
  + Resource Allocation
  + Outlier Identification
* Simple models for forecasting
  + Last value forward
  + Exponential smoothing
  + Simple count statistics
* ARIMA modelling framework
* Prediction Intervals
* Time Series Forecasting and Feature Engineering
* Example in Python

# Lesson 12: Conducting Experiments

* Purpose of doing experiments
  + Knowing whether a change in strategy works
* A/B testing framework
  + Hypothesis Testing
  + Power analysis upfront
  + Testing continuous outcomes
  + Testing binary outcomes
* Continuous Monitoring of Outcomes
  + CuSum charts
  + Stopping Early based on results
* Alternatives to random experiments when not possible
  + Historical analysis is difficult
  + Stratified experiments
  + Can’t cherry pick
* Example in Python

# Lesson 13: Recommender Systems

* ?????????????

# Lesson 14: Genetic Algorithms

* ?????????