# Intermediate Data Science Bootcamp

## Advanced Classification

8 hours of instruction (Python)

Build upon your classification skills to develop advanced classification models and improve your predictive accuracy. Learn how to implement powerful algorithms, such as logistic regression and ensemble learning methods to identify the likelihoods of events and trends. You'll learn how to develop advanced classification models with increased accuracy, avoid pitfalls of classification models, and fine-tune parameters.

### Syllabus & Topics Covered

1. Logistic regression
   - Introduction to logistic regression and its relation to neural networks
   - Assessing classifier performance
     *(ROC curve, AUC, cutoff value selection)*
   - Tuning & regularization of logistic regression
2. Ensemble methods - random forests and gradient boosting
   - Random forests
   - Boosting methods
   - Grid search and optimization of ensemble methods

### Prerequisites

- Attendees must be comfortable using Python to manipulate data and must know how to create basic visualizations with ggplot2.
- Attendees must have a foundation in classification models and model accuracy measures.

## Introduction to Text Mining

24 hours of instruction (Python)

Do you spend your days reading through reports and summaries? Do you want to be able to highlight key phrases and extract meaning automatically? This introduction to text mining provides the foundational skills you need in order to process, clean, and format text data for analysis. You'll learn how to integrate text mining into your work and extract key summary metrics and words from a corpus of documents.

### Syllabus & Topics Covered

1. Preparing data for text mining
   - What are text mining and Natural Language Processing (NLP)?
   - Applications and intuition of text mining
   - The process of cleaning and preparing text for analysis
   - Working with different data formats – PDF, CSV, TXT
2. Visualizing text and understanding text distributions
   - N-grams (bi-grams, tri-grams and quad-grams)
   - Word clouds
   - Histograms
   - Summary metrics of corpora

3. Topic modeling within text
   - Building a term document matrix
   - Implementing bag-of-words technique on text data
   - Building out TF-IDF
   - Summary of topic modeling and implementing LDA
   - Evaluating results and optimize number of topics
4. Text mining analysis
   - Description of cosine similarity in text
   - Building interactive network graphs to visualize similar documents
   - Applying hierarchical clustering to text data
   - Visualizing clustering as dendrogram and evaluate results

### Prerequisites

Students must be comfortable using Python to manipulate data and must know how to create basic visualizations.
Additionally, students must have a foundation in classification models and model accuracy measures.