NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp

Machine Learning Kaggle Project

Introduction

So, you've proven to your client that you can analyze data from a descriptive standpoint. Whether it be through static or interactive visualization, provided or collected data, you have shown that you can lead an audience through various insights and themes informed by your data. But exploring the data on a surface level is only part of the story. The client for whom you've been working is so impressed by your skills that they have launched you into an interview process for the lead data scientist role at the company -- but you are among a slew of viable candidates. Now you're in the hot seat: you must show that your skills extend beyond interpretation and into the predictive realm. Can you implement machine learning to make accurate predictions? Can you research and implement new machine learning skills? The client wants to know. You still need to tell a story, but the chapters must extend beyond the introduction. Once again, what story will you tell?

What We're Looking For

You're on an accelerated path to becoming an accomplished statistician: machine learning topics abound and the anticipation of building your own models excites, but there are so many algorithms, tuning parameters, and assumptions of which to keep track. How are you going to survive?

The bottom line is: **You will find a way.**

For this project, your primary task is to employ machine learning techniques to accurately make predictions given a dataset. The framework will be through the lens of the Higgs Boson Machine Learning Challenge from Kaggle. We implore you to consult the

website for more detailed information. For your convenience, an excerpt from this particular competition is provided below:

The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the experiment. No knowledge of particle physics is required. Using simulated data with features characterizing events detected by ATLAS, your task is to classify events into "tau tau decay of a Higgs boson" versus "background."

While the primary goal of Kaggle competitions is generally focused on predictive accuracy, you will be expected to lead your audience through descriptive insights as well. For the purposes of your project you will aim to not only create a model that predicts well, but also allow yourself to describe data insights drawn from exploration. Although you have the data at hand, it still seems like you're doing double-duty. And time is short. How will this all be possible?

The bottom line is: **You will still find a way.**

As always, preparation will be key. Successful projects will encompass a plethora of skills including, but not limited to, the following:

- Submission in respect to the deadline.
- Background knowledge of dataset(s).
- Communication of motivation: why do we care?
- Research questions of interest: what do you want to find out?
- Answers to research questions: what have you uncovered?
- Presentation skills.
- Time management (not going over the allotted time).
- Ability to answer audience questions effectively and efficiently.
- Balance of complexity and simplicity.
- Explanation of future work: what would you do if given more time, data, etc.?
- Demonstration of EDA skills:

    - Numeric methodology.
    - Graphic methodology.
- Demonstration of machine learning skills:
    - Supervised methodology.
    - Unsupervised methodology.
- Ability to research and implement new machine learning skills including, but not limited to, the following:
    - Sensitivity
    - Specificity
    - Receiver Operating Characteristic Curve
    - Area Under the Curve Metric
- Ability to assess model weaknesses and identify improvements.
- Ability to manage a team workflow.

The Details

Your project proposal declaration is due uniformly by the beginning of the Pulse Check on **Friday, August 19**. **No exceptions.** You must declare your team on the project proposal document.

This is a **team project** in respect to the final deliverable. **No exceptions.** Every student must work with at least one other student for the project and presentation; we encourage collaboration and knowledge generation, so it is possible that teams may want to merge and/or specialize after the team proposal deadline.

All code, data, etc. used to generate your graphics and/or Shiny app and any slides, markdown files, etc. intended for your presentation are due to the project GitHub repository uniformly by **Sunday, August 28 at 11:59pm**. **No exceptions.** Only one teammate needs to submit on behalf of each group; please identify by team name upon submission.

You will be required to deliver a **10 - 20 minute presentation** dependent upon group size (up to 10 minutes for groups of size ≤ 2, up to 15 minutes for groups of size > 2 and

size ≤ 4, up to 20 minutes for groups of size ≥ 5) and respond to any audience questions. Time slots will be randomly assigned on this calendar, so all projects must be submitted on time. **No exceptions.**

An associated blog post will be due by **Monday, September 5 at 11:59pm**. **No exceptions.** Remember, this is a living and breathing document. You may continue to develop and edit your project far beyond the deadline, as no project will ever truly be complete. You may choose to co-author a single blog post describing your whole project, or submit individual blog posts highlighting your own personal project workflow (e.g., if your team specifically delineated responsibilities, etc.).

For inspiration, take a look at our previous students' blog posts (here's a link specifically to the Machine Learning Kaggle category).

For any lingering questions, please do not hesitate to reach out; we are always here to help!

Good luck!