Capstone check-in #1

The Capstone Project

Your capstone project is the culmination of your time at GA. You will:

- Formulate an interesting question
- Collect the data required to address the question
- Develop the best possible model or analysis
- Communicate your findings

You'll present your findings on the last day of the course, but there will be due dates and check-ins leading up to the end of the course to help you structure your approach.

The check-ins:

In summary:

- Check-in 1: Topic proposals
- Check-in 2: Problem statement & preliminary EDA
- Check-in 3: Progress report & preliminary findings
- Check-in 4: Report write-up & technical analysis (last week of course)
- Check-in 5: Capstone Presentation (last day of the class)

Capstone Check-In - Part 1

What?

- Presenting three potential topics and problems
- Describe your:
  - Goals & criteria for success
  - Potential audience(s)
- Ideally identify 1-2 potential datasets or data sources for each problem

When?

- TBD

How?

- Lightning Talks! A 5-7 minute presentation that covers **three potential topics**, including potential sources of data, goals, metrics and audience.

Things to Include in the Talk

1. What is your problem statement? What will you actually be doing?
2. Who is your audience? Why will they care?
3. What is your success metric? How will you know if you are actually solving the problem in a useful way?
4. What is your data source? What format is your data in? How much cleaning and munging will be required?
5. What are potential challenges or obstacles and how will you mitigate them?
6. Is this a reasonable project given the time constraints that you have?

Do your best to come up with three ideas that you'd genuinely be happy to work on, even if you have one idea you're very set on!

Slides template

- The next slides include a template for capstone check-in #1, as well as some examples based on the template.
- You’re not required to use this template, but make sure you’re presenting all the same information as in the template!

{ Project Idea }

- Data will be collected from:
  - Source 1
  - Source 2
- My MVP is: a model and analysis.
- My stretch goals include:
  - Goal 1
  - Goal 2
- My observations will be _____ and my target will be ____.

I will use ___{what}___ data

to build a ___{type}___ model

that predicts _{target}_ values

in order to _{value prop}_.

___

Additional Notes

- Some potential roadblocks are...
- Something I want to research more is...
- I’m not sure if I can even accomplish...
- If anyone has recommendations on how to find _____, please let me know!

EXAMPLE: Hit Streak Predictor

- Data will be collected from ESPN API and Some Stats Website
- My MVP is an daily scraper and the main classification model.
- My stretch goal is an automated pipeline that emails me every morning with the top 5 predictions for the day.
- My observations will be batters, representing a single matchup. My target will be binary, whether or not they got a hit that day.

I will use batter performance data to build a binary classification model that predicts whether or not a batter will get a hit in order to try and win the MLB “Beat the Streak” competition.

Additional Notes

- Data collection/wrangling will be an issue due to the abundance of data. Each observation will need to be a single day for the batter so I will need to reformat a lot of the information I will have.
- I need to research more expert analysis to see what the important features might be.
- My stretch goal will be difficult, I will need help on automating the process and running it each day at a specific time.
- If anyone has recommendations on how to send emails with python, please let me know!

EXAMPLE: Produce Image Classifier

- Data will be collected from the flickr API
- My MVP is a NN that beats baseline accuracy for broad produce categories.
- Stretch goal: deploying the model to a webapp.
- My observations will be single images and the target will be the fruit label.

I will use produce images to build a multi-class classification model that predicts produce type in order to improve a frustrating part of the checkout process..

Additional Notes

- Image data is “fun” to work with
- We haven’t learned neural networks yet, so I’ll have to learn a lot on my own and implement it pretty quickly.
- Image data is computationally intensive to handle. I’ll probably have to use AWS.
- I don’t know much about hosting models on Flask and will have to learn that for my stretch goal.

EXAMPLE: Star Trek Classifier

- Data will be collected from fan transcripts of the show (Star Trek: Deep Space Nine)
- My MVP is a model to predict who said what lines, and an analysis of the dialog
- My stretch goal is analyzing the ratings for each episode as well.
- My observations will be text from the show, and I'll be predicting the speaker.

I will use script data to build a multi-class classification model that predicts which Star Trek character said a line of dialog in order to learn more about the best TV show that ever aired.

Additional Notes

- This doesn't have any business application, so I'll make sure my write-up is really impressive. I want to make sure readers understand my methods and approach.
- NLP modeling can be slow and computationally intensive. I might have to run models overnight or use AWS.

Capstone Check-in #1

You should be able to convince people to care about your capstone, but it doesn't have to be a full business plan. **Remember, you're pitching *yourself* to an employer -- not your project**!