Capstone Project Proposal Template

Notes:

- This should take no more than one hour to complete – the clearer you are about the business problem you're working to solve with your ML-driven solution, the easier your proposal will be to complete
- This will be uploaded to your repo, which will be a part of your final submission
- Due date for submission is 12/9

Instructions:

1. Download this document as a Word Doc
2. Answer each question using a few sentences, at most
3. Save your completed proposal as a PDF
4. Create a project GitHub repo (if you have yet to do so)
5. Add your instructor as a collaborator (username nickmccarty) to your project repo
6. Add your mentor as a collaborator
7. Push your proposal PDF (created in Step 3) up to your repo
8. Copy the URL corresponding to the location of the PDF in your repo
9. Submit the copied URL using this link

Trevor Flanagan – Predicting MLB Player Performance

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - I'd like to predict which MLB players are going to have good seasons and which one's are going to have poor seasons. This labeling description sounds categorical, but I intend to approximate statistics in the season, which will be a regression problem.
- What industry/realm/domain does this apply to?
    - Statistic's role in modern sports has quickly become very impactful and found across many different layers of the industry, from box office decisions, to training conditions, to lineup ordering ect.
    - In predicting performance of MLB hitters, I hope to **model results that could lead front office decision making** in the offseason as well as before the trade deadline. These results could also be useful in the **growing sports betting industry**
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)

  - I’ve grown up playing and watching baseball. The statcast era and modern effect of statistics on the industry in general is very interesting. I imagine more sports will continue to adopt and embrace it’s impact.

Data Understanding

- What data will you collect?
  - Baseball-reference.com is a reputable source of player data, both active and retired. For this project I’m concerned with metrics of hitting performance.
    - Plate Appearances (PA)
    - Batting Average (BA)
    - Strikeouts(SO)
    - Homeruns(HR)
    - Etc..
  - Additional stats that might be a factor in batting performance such as age, player salary, are also available to be looked at in baseball reference databases
- Is there a plan for how to get the data (API request, direct download, ect.)?
  - I’ve done a project in the past that used **web-scraping** to obtain publicly available data. I feel that will be appropriate here, as you can run the script over the course of an MLB season to feed newer, more current data to the model.
- Are the features that will be used described clearly?’
  - Yes

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, ect.)
  - There are a multitude of data preparation steps I foresee needing to work through. Primarily I will be **pulling player data from many different pages** of the baseball-reference.com website. So I will **need a list of players names that I want to pull**. After defining a list of names and webscraping I’ll have to **join all the different player tables together.** And finally, I’ll have to check and clean up any null values and
- What are some of the cleaning/pre-processing challenges for this data?
  - I need to solve cases where **players share the same first and last names**. For example will smith is an active pitcher for the Houston Astros, but there is also an active Catcher for the Dodgers named will smith, with another 4 retired Will Smith’s in the baseball reference database

Modeling

- What modeling techniques are most appropriate for your problem?
  - I want to **build a deep neural network** that predicts performance statistics, using concepts like **Principal Component Analysis** and topics covered in modules 2 and 3

- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - There are two statistics I'd be interested as assigning as target variables when building a model.
        - WAR (Wins above replacement)
        - OPS (On base + Slugging)
- Is this a regression or classification problem?
    - This is a regression problem

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - Regression Analysis – R^2
    - I'll prefer RMSE over MAE because I don't expect an abnormal number of outliers in the data I pull.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - For webscraping, which is a concept we didn't cover in this course I'll be using Beautiful Soup in python
    - I'll be using the keras module in python to build my neural network