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

Breast Cancer Detection using Machine Learning

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - In this project, I am attempting to use machine learning to identify cases of breast cancer efficiently and accurately from images of mammogram scans.
- What industry/realm/domain does this apply to?
    - This project relates to life sciences, specifically medical imaging.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - Besides the obvious life-saving benefits of cancer detection, this project could be a tool used by medical professionals to identify cancer quickly, support the doctor in making an accurate diagnosis, and improving both patient and doctor experience.

Data Understanding

- What data will you collect?
    - I will use a dataset of mammograms from the Radiological Society of North America.
- Is there a plan for how to get the data (API request, direct download, etc.)?

    - The data will be downloaded from Kaggle.
- What are the features you'll be using in your model?
    - I will use the mammogram images themselves, along with some metadata such as the age of the patient, whether they have had breast implants, density of the breast, and more.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - I will convert the images from DICOM format to PNG and likely perform some augmentation of the images, such as cropping, reflecting/rotating randomly, and resizing.
- What are some of the cleaning/pre-processing challenges for this data?
    - I am new to preprocessing images, so that will pose a challenge for me. Converting from DICOM to PNG will be computationally intensive for such a large dataset and I have already struggled with this process and saving the preprocessed images in the correct file structure.

Modeling

- What modeling techniques are most appropriate for your problem?
    - The most appropriate techniques would be to use deep learning for classification of image data. I will use transfer learning on a pre-existing model to save time and produce better results.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - The target variable is whether the patient is positive or negative for cancer.
- Is this a regression or classification problem?
    - This is a binary classification problem.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - I will use accuracy and F1 score to determine success

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - I will use a convolutional neural network built with Keras. I have not yet decided what pre-existing model I will build off of with transfer learning.