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 1/16

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 dodgy719) 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

Lung Cancer Screening

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer? A healthcare company wants to determine someone's likelihood of Lung Cancer, given the factors of gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, and chest pain.
- What industry/realm/domain does this apply to? Healthcare
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation) To create an early screening for lung cancer

Data Understanding

- What data will you collect? Lung Cancer | Kaggle
- Is there a plan for how to get the data (API request, direct download, etc.)? Kaggle, direct download
- What are the features you'll be using in your model? gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty, and chest pain.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)? check for null values and encode any null values
- What are some of the cleaning/pre-processing challenges for this data? Clean up null values

Modeling

- What modeling techniques are most appropriate for your problem? confusion matrix
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target) whether the person has diabetes or not
- Is this a regression or classification problem? classification

Evaluation

- What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)? Accuracy and confusion matrix

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)? Decision trees and random forests