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 end-of-day 3/13 for Cohort 3b

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 jvntra) 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 for Cohort 3b

[Drug Review Classification]

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  I am trying to predict the patient's condition based on the Drug review.
- What industry/realm/domain does this apply to? **NLP**
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
  By doing this project, I will learn more about classification algorithms in Machine Learning and Natural Language Processing. It will help me in my day- to-day work at Deloitte.

Data Understanding

- What data will you collect?
  "UCI ML Drug Review" dataset will be used. The dataset is available on Kaggle.
- Is there a plan for how to get the data (API request, direct download, etc.)?
  I will get the data by downloading the data directly from Kaggle.
- What are the features you'll be using in your model?
  There are 7 features in the dataset. They are:

uniqueID, drugName, Condition, Review, Rating, Date, usefulCount

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
  Missing value, encoding of categorical value.
- What are some of the cleaning/pre-processing challenges for this data?
  Some of the review contains weird characters which need to be cleaned.

Modeling

- What modeling techniques are most appropriate for your problem?
  Classification algorithms
  What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
  "Condition" is the target variable.
- Is this a regression or classification problem?
  It is a classification problem.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
  Accuracy, Precision, Recall, F1-score, ROC-AUC score, etc.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
  Classification algorithms, such as Logistic regression, XGBoost, Random Forest, SVM, etc. If possible, some NLP classification algorithms.