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

Found in a Random Forest

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  - *I am trying to solve the problem of fraud detection and fraud prevention.*
- What industry/realm/domain does this apply to?
  - *Fraud detection and prevention applies to all industries that involve financial transactions.*
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
  - *I conduct white-collar criminal investigations, and this directly applies to my line of work.*

Data Understanding

- What data will you collect?
  - *Financial / transactional data.*
- Is there a plan for how to get the data (API request, direct download, etc.)?
  - *I will be gathering the data from a Kaggle library – 'Credit Card Fraud Detection'.*
- What are the features you'll be using in your model?

    - *Due to confidentiality issues, the features are labeled V1 – V28 along with 'Time' and 'Amount'. V1 – V28 contains numerical variables from a prior PCA transformation.*

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - *Correlation matrices to see which features have a high correlation to fraud.*
    - *Information gain / gini index*
- What are some of the cleaning/pre-processing challenges for this data?
    - *Will need to scale 'Time' and 'Amount' to V1-V28 as V1-V28 have undergone a PCA transformation.*
    - Will need to remove non-fraud data to create a more balanced data set (fraud cases only account for 0.172% of the data).
    - *Will need to split the data so we can test the data on the original data set.*

Modeling

- What modeling techniques are most appropriate for your problem?
    - *Decision Trees*
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - *Fraud detection*
- Is this a regression or classification problem?
    - *Classification*

Evaluation

- What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?
    - AUC and Confusion Matrix.
    - F1-score and precision/recall could also be helpful.

Tools/Methodologies

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