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 dodg719) 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

Qar' (Arabic for "Reader")

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - The problem that I am trying to solve is how to create an initial transliteration of a handwritten Arabic document without the need of a trained human translator. To this end, I am trying to create an initial model that can reliably recognize individual Arabic characters.
- What industry/realm/domain does this apply to?
    - This project could easily apply to various realms, though I foresee that this could have an application in the law enforcement, homeland security, or defense space.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - I minored in Arabic in my undergraduate studies, so I can read basic Arabic texts; I also studied homeland security topics as part of my International Politics major, so I understand that translators can receive a ton of documents for review. Therefore, if a machine could transliterate or translate these documents or triage these documents before arriving at a translator, that could certainly help reduce their workload.

Data Understanding

- What data will you collect?
    - I plan to use a dataset from Kaggle which has a 16800 sample of handwritten Arabic.[1]
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - The dataset was retrieved via direct download.
- What are the features you'll be using in your model?
    - I plan to use the images and labels in the dataset to initially train the dataset. From there, I can use the images in the testing dataset to validate whether the model is performing appropriately.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - Encoding could certainly be an issue since I intend to transliterate the data; therefore, I would likely have to create something in the dataset or program which returns an English/Latin character(s) for a given Arabic character.
- What are some of the cleaning/pre-processing challenges for this data?
    - Ensuring that the classifications are correct is a potential challenge for this data, but I do not expect to encounter this issue.

Modeling

- What modeling techniques are most appropriate for your problem?
    - Given that the dataset provides labels for training the model, I would assume that supervised learning techniques would be most applicable to this problem.
- 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 model appropriately classifies an Arabic character.
- Is this a regression or classification problem?
    - This is a classification problem.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?
    - This is a classification problem, and I foresee that I am going to be concerned with both precision and recall. Therefore, I am probably going to be concern with the F1-score which combines both precision and recall. however, it is more likely that Precision are more appropriate as characters in a language often tend to

[1] Loey, Mohamed. 2020. "Arabic Handwritten Characters Dataset." Accessed via Kaggle. https://www.kaggle.com/datasets/mloey1/ahcd1 Accessed 14 January 2022.

have unequal frequency distributions in a dataset; for example, in English, the letter e tends has the highest frequency in writing samples.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - I'm leaning towards decision trees or random forests as this is a classification problem. Such a model could then be extensible if I found a dataset of handwritten Arabic words. K-nearest neighbors is also a possibility, though I think that doesn't provide enough growth potential for this product.