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 proposal submission is 3/12**

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 charles-rice) 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

Predicting Median House Value

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - I am trying to predict the median house value in California, given the US Census variables.
- What industry/realm/domain does this apply to?
    - Real-estate and finance would apply the most for development and arbitrage reasons.
    - Mortgage lenders also would benefit as they could easily see an estimated value of a house before writing a mortgage on it.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - With COVID 19 affecting interest rates, we saw huge demand for houses that drove prices to record highs, which interested me. Why would people pay that much for a house and what really was the value?
    - Real estate has also always been interesting to me as it is commonly seen as an extremely safe investment with seemingly guaranteed appreciation, so finding out programmatically what makes a house valuable is the natural extension for me

Data Understanding

- What data will you collect?
    - I plan on using the California Housing Dataset available on Kaggle
    - This data is census data, so it is not collected on a per household basis but rather on a larger scale
    - The data is also only for a district in California so the applicability of the results will be subject to that caveat
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - I will use Kaggle to get my dataset via a direct download
- What are the features you'll be using in your model?
    - I think the biggest to consider are income, proximity to the shore, and bedrooms.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - I will possibly have to do some one hot encoding to deal with some features, and I will almost certainly have to drop or clean up the data to make it more sensible for the model.
- What are some of the cleaning/pre-processing challenges for this data?
    - While some features may not have a correlation, it is difficult to determine if they are truly important and deciding which features to use will be a large challenge.
        - i.e. Maybe ultra-high value homes have too many bathrooms so the model will not be able to understand the impact of single bathroom increments

Modeling

- What modeling techniques are most appropriate for your problem?
    - Regression will be the most relevant for 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)
    - Median house price.
- Is this a regression or classification problem?
    - Regression

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - I will use RMSE and R^2

Tools/Methodologies

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

Since it is a regression problem, I plan on using linear regression primarily and random forests to expand upon it, time permitting.