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](https://github.com/new) (if you have yet to do so)
5. [Add your instructor as a collaborator](https://docs.github.com/en/account-and-profile/setting-up-and-managing-your-personal-account-on-github/managing-access-to-your-personal-repositories/inviting-collaborators-to-a-personal-repository) (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](https://my.learn.co/courses/586/quizzes/6353?module_item_id=79223)

**Predicting Malignant Breast Cancer Tumors**

**Business Understanding**

* What problem are you trying to solve, or what question are you trying to answer?

Trying to classify breast cancer tumors are malignant or benign as accurately as possible.

* 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)

I enjoyed working on health-related questions when I was in school, so I picked this as my project because I thought it would be interesting.

**Data Understanding**

* What data will you collect?

Data on breast cancer tumors – size, features, etc.

* Is there a plan for how to get the data (API request, direct download, etc.)?

Download from Kaggle

* What are the features you’ll be using in your model?

(['diagnosis', 'texture\_mean', 'area\_mean', 'smoothness\_mean',

'compactness\_mean', 'symmetry\_mean', 'fractal\_dimension\_mean',

'texture\_se', 'area\_se', 'smoothness\_se', 'compactness\_se',

'symmetry\_se', 'fractal\_dimension\_se'],

dtype='object')

**Data Preparation**

* What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?

Cleaning the data of nulls, changing categorial variables, getting rid of collinearity.

* What are some of the cleaning/pre-processing challenges for this data?

Trying to find the collinearity.

**Modeling**

* What modeling techniques are most appropriate for your problem?

Linear regression, random forrest.

* What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)

Diagnosis

* Is this a regression or classification problem?

Classification

**Evaluation**

* What metrics will you use to determine success (MAE, RMSE, Accuracy, Precision etc.)?

Accuracy, precision, recall f1

**Tools/Methodologies**

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

Random forests