# 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

## [Credit Card Fraud Detection]

**Business Understanding**

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

Purchases made through websites or in situ stores represent an imminent danger to credit/debit cardholders throughout the world, given their exposure to cybernetic dangers that may generate unrecognized fraud transactions.

Thus, it is important to generate a trustworthy AI model that detects potential threats based on transaction datasets collected on a daily-basis routine.

- *What industry/realm/domain does this apply to?*

This applies particularly to the finance industry, with an emphasis on the fraud monitoring techniques that are implemented on a bigger scale at banks worldwide.

- *What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)*

I'm interested in how a bank can detect frauds and quickly notifies its users based on your daily financial activities, and this basic approach could insert me in the world of basic fraud detection.

**Data Understanding**

- What data will you collect?

The dataset I collected consists of a CVS type file with transactions made by credit cards in September 2013 by European cardholders.

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

I obtained the data through a direct download form the Kaggle website, result of an extensive search of machine learning datasets. The web link to this dataset is: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

- What are the features you'll be using in your model?

This dataset consists of numerical input variables only. Most features (V1 to V28) are the result of PCA transformation of the original features due to confidentiality issues; the rest ('Time' and 'Amount') haven't been modified at all.

**Data Preparation**

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

As all features and label are numerical variables, there is no need to use One Hot Encoding. However, it will be prudent to explore for N/A or Categorical-type values and fill them with the average of each feature.

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

The authors warn that the dataset is highly unbalanced: there are 492 frauds about of 284,807 transactions (only 0.172% of the total 'Class' labels). Doing a quick search, one possible solution could be resampling the minority / majority class or applying SMOTE (Synthetic Minority Oversampling Technique using k nearest neighbour).

**Modeling**

- What modeling techniques are most appropriate for your problem?

Logistic Regression model, knowing it's used to predict categorical target variables through linear combination of the selected features.

Random Forest Classifier model, also known to solve regression or classification problems (I need to differentiate between fraud and no fraud).

Support Vector Machine model, effective in high dimensional spaces (dataset consists of 30 features).

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

My target variable is the column called 'Class', where 1 represents fraud and 0 the opposite.

- Is this a regression or classification problem?

Both. I'm trying to predict if a transaction is fraud or not, so the output is a categorical variable.

**Evaluation**

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

The authors recommend measuring accuracy through the Area Under the Precision-Recall Curve, given the unbalanced nature of the dataset. The rest of the traditional metrics will be incorporated (precision, F1-score, recall).

**Tools/Methodologies**

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

Random forests, logistic regression, support vector machine. I need the 3 of them to either build a pipeline or just to obtain metrics and compare them all to choose the best-fitted model.