# Machine Learning Final Project

## Project Topic

* Clear explanation of what this project is about.
* Clearly state which type of problem.
  + Clearly state type of learning.
  + Clearly state type of task.
* The motivation or the goal of project clearly stated.

## Data

* Properly cite and describe data source (including links and brief explanations).
  + Where data is from.
  + How data was gathered.
  + Cites dataset using APA format.
* Explain the data description properly.
  + Number of samples/rows.
  + Number of features, columns.
  + Data Type for each feature.
  + Description of features.
  + Whether data is multi-table form or gathered from multiple sources.

## Data Cleaning

* Clear explanations on how and why cleaning is performed.
  + For example, “author decided to drop a feature because it had too many NaN values and the data cannot be imputed.”
  + For example, “author decided to impute certain values in a feature because the number of missing values were small and he/she was able to find similar samples OR, he/she used an average value or interpolated value, etc.”
  + For example, “author removed some features because there are too many of them and they are not relevant to the problem, or he/she knows only a few certain features are important based on their domain knowledge judgement.”
  + For example, “author removed a certain sample (row) or a value because it is an outlier.”
* Has conclusions or discussions?
  + Data cleaning summary
  + Findings
  + Discussing foreseen difficulties
  + Analysis strategy
* Has proper visualizations?
  + Meeting the benchmark for moderate data cleaning could include:
    - Data type munging. (Converting and mapping unprocessed data into different format to improve suitability and value for downstream uses.)
    - Drop NA.
    - Impute missing values.
    - Check for imbalance.
    - Utilize visualizations to look for any data-specific potential problems and address issues found.

## Exploratory Data Analysis

* Clear explanations on how and why cleaning steps were performed
  + Has proper visualizations (simple plots)
  + Has proper analysis and conclusions.
  + Correlation matrix with analysis.
    - Also, histograms, feature importance
  + Has conclusions or discussions
    - Data Cleaning Summary
    - Findings
    - Discussing foreseen difficulties
    - Analysis strategy

## Models

* Proper single model
* Addresses multilinear regression/collinearity
* Feature engineering
* Multiple ML models
* Hyperparameter tuning
* Regularization or other training techniques such as
  + cross validation
  + oversampling/under-sampling/SMOTE or similar for managing data imbalance
* Uses models not covered in class

## Results and Analysis

* Summary of basic results and analysis
* Good amount of visualizations
  + For example, tables, graphs/plots, heat maps, statistics summary with interpretations.
* Use different kinds of evaluation metrics properly and why metric is chosen.
* Iterate the training/evaluating process to improve performance.  Address selecting features through the iteration process.
* Compare results from multiple models and make appropriate comparisons.

## Discussion and Conclusion

* Discussion of learning and takeaways.
* Discussion of why something didn’t work.
* Suggests ways to improve.