CAPSTONE:

1. Mortgage Default
   1. Problem Statement
      1. Build a classification model to determine whether a loan will default
   2. Metrics
      1. Accuracy versus industry baseline for loan class
   3. Data
      1. Kaggle - <https://www.kaggle.com/skihikingkevin/online-p2p-lending>
      2. Prosper API - <https://developers.prosper.com/>
      3. Lending Club API - <https://www.lendingclub.com/developers>
      4. FannieMae -
   4. Audience
      1. Consumer Credit Companies
   5. Challenges
      1. Time
      2. Interpretability

<https://www.kaggle.com/c/home-credit-default-risk/data>

<https://quant.stackexchange.com/questions/30231/interest-rate-implied-probability-of-default>

<https://www.lendingclub.com/info/download-data.action>

<https://www.kaggle.com/c/predicting-loan-default/data>

Fannie Mae:

1. <https://www.dataquest.io/blog/data-science-portfolio-machine-learning/>
2. <https://riskspan.com/news-insight-blog/hands-on-machine-learning-predicting-loan-delinquency/>
3. <https://economics.ucr.edu/job_candidates/Bagherpour-Paper.pdf>
4. <https://degravek.github.io/project-pages/project1/2016/11/12/New-Notebook/>
5. <https://cs230.stanford.edu/projects_spring_2018/reports/8289346.pdf>

Questions:

1. Predict default based on pre origination data, post origination data, or combined?
2. Current Loan Delinquency Status?
3. How to deal with huge datasets?
4. Condense Performance dataset?

Considerations:

1. Undersample rather than oversample
2. Inner join
3. Confusion matrix – minimize false negatives
4. AUC-ROC
5. Corr between target categories and nulls\*\*\*
6. Visualizations
   1. Distribution of target
   2. Correlation Heatmaps
   3. Outliers