•Start with a DQR

•Clean: treat the frivolous fields (SSN, Address, Phone, DOB)

•Create all variables

•Scale all variables (typically z scale)

•Separate data into modeling/Out of Time (OOT) -- Notebook

•Use the modeling data (trn,tst) for feature selection

•Select 15 – 25 best variables -- Thu HW

•Build a baseline linear model (logistic regression)

•Build several nonlinear models, tune parameters somewhat

•I suggest neural net, boosted trees, random forest, SVM

•Select your favorite model, fine tune the parameters

•Build the final performance tables with your final, best model -- Sun meeting

•Write report

Sun lock down algorithms

1 person: 3 tables

•**Cover Page**: Title, project team number, names, date.

•**Table of Contents**. Make sure all pages are numbered.

•**Executive Summary** (few paragraphs). High level summary of project and results.

•**Description of Data** (few pages). Include most important (not all) distributions/histograms, put the full DQR in appendix.

•**Data Cleaning** (few paragraphs or pages). How did you handle exclusions, outliers, missing fields, frivolous field values.

•**Candidate Variables** Before calculating variables, Z-scale. (few pages). Describe in complete detail the formulas and logic for creating all the candidate model variables. Include a list of all variables; if it’s more than about 5 pages it could be in an appendix. (entity, different types of variables)

•**Feature Selection Process** (few pages). Describe what you did to select features. List of final variables.

•**Model Algorithms** (a few pages). Short explanation of all algorithms tried. Table of high level results for each algorithm (FDR for training, testing, oot).

•**Results** (few pages). For your final algorithm and parameter selection, show the three tables for training, testing, oot populations.

•**Conclusions** (few paragraphs). Summarize everything you did and what else you might do with more time.

•**Appendix.** DQR

Chao + proofread

Minglu

Cheng + format

Selene

Ocean + table

Sun meeting 5p-8p

Finish Model running, Thu night

Fri night report draft ready

Sat night finish