Part 1

Perform data analysis, what is the data telling you

What were your top 3-5 observations from initial data exploration?

Describe and provide supporting screen shots

What features look most promising for modeling?

Describe how you selected these features, screen shots to support your discussion, if appropriate

Do you need to deal with missing values?

Scale the attributes?

Make initial attempt to model and predict, set your own baseline

Part 2

Construct initial model, post on Kaggle to set your team baseline

Screenshot of position on leaderboard

Compare with multiple classifiers, model settings

Opportunity of feature selection, feature transformation such as binning or clustering, etc

Describe the data cleaning, transformation steps you selected and describe the why and how. Run at least two to three new model evaluations, submit at least one for new submission to Kaggle contest site

Part 3

Repeat part 2 as necessary during part 3 of the assignment

Compare with multiple classifiers, different model settings

Write up description of modeling strategy, screen shots of both results in WEKA or tool of choice to illustrate exploration phase, screen shots of all submissions to Kaggle contest site

Can you identify an opportunity for model ensembles, if so describe and evaluate

Story

Show best Kaggle score, describe model that generated it

How to go from big, crappy text data file to our best Kaggle score?

Overall, team working strategy

Divide, conquer, and ensemble!

Work separately, but share results as soon as you have them

Each create our unique models

Ensemble predicted probabilities with simple average for final team submission

Well, what is this big, crappy data file?

X GB = Y number of rows by Z number of variables

Variables m, n, p, etc have been anonymized

All categorical-really though?

Problem: it’s really not scary big but too big to work on most consumer focused computers

16 GB isn’t big enough to load it and have space for computations

48 GB on the other hand is

Going to need to sample to be able to be efficient

Need to know what we’re sampling

Summary of big data

Counts

Conditional probabilities

A chip off the old block

Now time to sample the dataset to an easily handled size

Sampling strategies

Feature selection

Group plot matrix

Correlation

Wrapper methods

Other transformations needed

Models

Summary of individual models

Ensemble plans

Log loss