Data set: This data set includes 3673 of samples, with 86 features which are all numerical except that x80 is categorical variable. Target variable denoted by "Financial Distress" falls in a continuous domain containing both positive and negative number..

Problem specification: This is intended to find one optimal model among three approaches to predict whether a company would be considered as financially distressed with given features. The question is, does these pre-optimized three methods behave significantly differently, and if yes, which one would achieve best performance given this dataset.

Approaches: There are 3 approaches to be tested: linear

regression with l2 regularization, stochastic gradient descent and batch gradient descent. These are three commonly used algorithms for regression problems. The interesting thing is that, this problem could be transform to a classification problem after some modification of the target value, because, from the description of the data file, one company will be regarded as healthy (0) “if its target value will be greater than -0.50. otherwise, it would be regarded as financially distressed (1).”

Design of experiments: First of all, the data set will be split by 5-fold CV given the relatively small sample size. Besides, for each algorithm being tested, internal CV is applied in each test sets, in order for tuning of

parameters such as lambda in l2 regularization, step-size in gradient descent.

Finally, statistical significant test is essential to check whether these learning algorithms behave differently. If so,

Adjusted R-squared will be used as metric for performance measure.