Paper outline

Data and Data cleaning

* Actually 56 variables after 1 hot encoding
* Split 4 variables from integer valued to one hot encoded
* Unbalanced sets within data
* Outcome is unbalanced, apriori – 70% yes

Methodology

* Split the data into test and training subsets
* Check model performance on the training data
* Use cross validation if necessary to tune hyper parameters
* Test the model performance on test data

Models

* How the model works
* Significant variables
* Confusion matrix
* Coefficients of the models
* Logistic – regression with all parameters
* Stepwise parameter estimation- Forward and backwards logistic regression
* Ridge
* Lasso
* Random Forest