Load the data

Action is our Y variable, what trying to predict

4, less than 1/10th of 1% on reset, highly imbalanced data set, stored as float 64, allocating too much memory used converted to Int 32

5, reduce from 6,291,200 bites to 4456304

8, create dummies for following features source port…..

9, check for duplication, get rid of duplicates

11,18,19, 57694 to 54958 columns, extremely wide data set reason for getting ride of 2731 (check math)

21, saved data set as wb file instead of CSV, that way opens exactly as have it, example won’t just default to float 64? Check that

Linear SVC only predicted allowed/denied, classified everything into 2 classes, did not do very well

SVM, scaled the non Boolean in data set, Boolean are dummies we created, after split data 80/20 ran SVM classifier

60/40 split Allow 100% drop 100%, deny 98%,

Tuned dataset using 60/40 split to find best parameters

SDG predicting 2 classes

Classification reports for each method, which were not right, walk through problems and mention time took to run