Homework 11 Report

Vineela Bandi and Mark Chafin

11/3/2021

We started by pulling in and analyzing our data to see what we were working with. After the initial processing and normalization we decided the first model was going to be KNN. We thought it would be able to find similarities and group the emails together as either spam or not. The accuracy was not great and even after some tuning the accuracy only got to ~90%. We moved on and began trying other models such as the Naive Bayes, Decision Tree and Random Forest, each model becoming increasingly more accurate. With the Random Forest achieving an AUC of ~95%. Even with hyperparameter tuning the Random Forest model without tuning performed slightly better. We decided to do some research on email spam detection and we found that “Support Vector Machines (SVM) has proved over the years to be one of the most powerful and efficient state-of-the-art classification techniques for solving the email spam problem … SVM is a good classifier due to its sparse data format and satisfactory recall and precision value. SVM has high classification accuracy.” From this research we decided it might be a good idea to try SVM. Unfortunately the SVM model performed slightly worse than the Random Forest model but still a great model.

We decided to try more complex models such as Deep Learning and Ensemble models. After some tuning the highest we achieved with the Deep Learning model was ~94%. With everything we have learned we decided to implement an Ensemble Model since RFand SVM are performing the best and we wanted to improve our model even more. The best model we achieved with Ensemble is using the 3 models of Decision Tree, Random Forest and Linear SVC with an accuracy of 95.5%