PREDICTING THE ‘VIRALITY’ OF TWEETS

DAVID FASTOVICH

INTRODUCTION

- Seemingly random topics become viral and flood the internet for days on end (e.g. Momo challenge)
- The ability for a tweet to go viral is enabling nefarious work, such as election interference by foreign nations
- Practically, advertising agencies would want to better predict when a user or tweet is likely to become viral to better advertise on Twitter for a lower cost – optimizing the amount of money spent per retweet
- Research questions: What makes a viral tweet? Is the popularity (e.g. the number of followers) of the individual tweeting? Is it the content of the tweet? Is it an association with a social movement?

TWITTER DATA USED

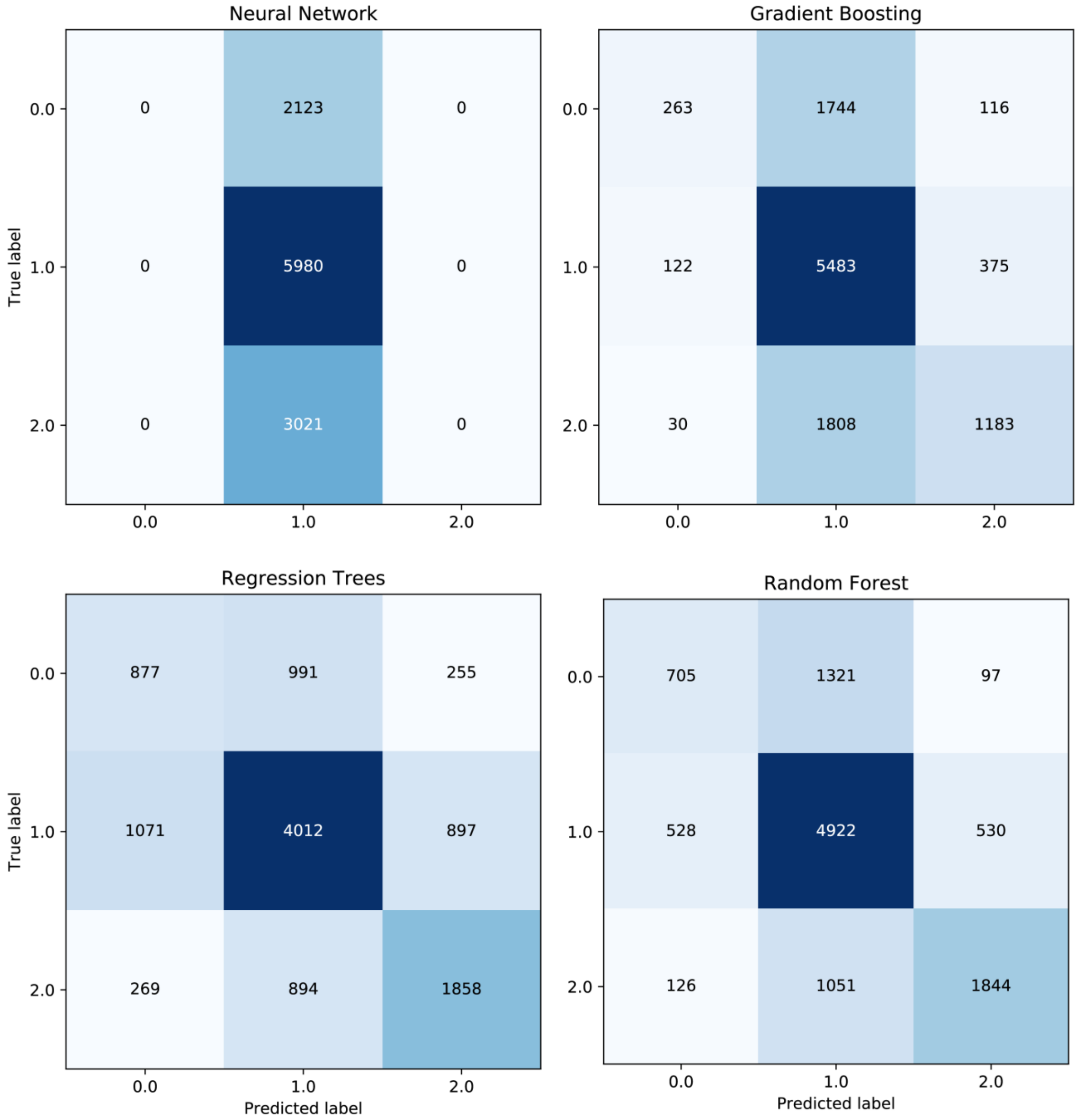
- Collected 100,000 tweets on a single topic: coronavirus
- No defined date or geographic boundary
- Classified tweets in three categories for modelling: not viral (less than 10 retweets), becoming viral (10 to 1000 retweets), and viral (more than 1000 retweets)
- Predictors used for modelling: number of followers, number of favorited tweets, number of people user follows, sentiment of the tweet, sentiment of the user description

PREDICTION METHODS – SOME ML AND SOME TRADITIONAL STATISTICS

- Neural network with two sequential layers using *keras* and the *TensorFlow API*
  - Mean-squared-error loss function
  - 5 layers
  - 100 epochs
- Gradient boosting using *scikit-learn*
  - Tested 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1 as learning rate – 1 produce most accurate predictions
- Regression trees using *scikit-learn*
- Random forests using *scikit-learn*

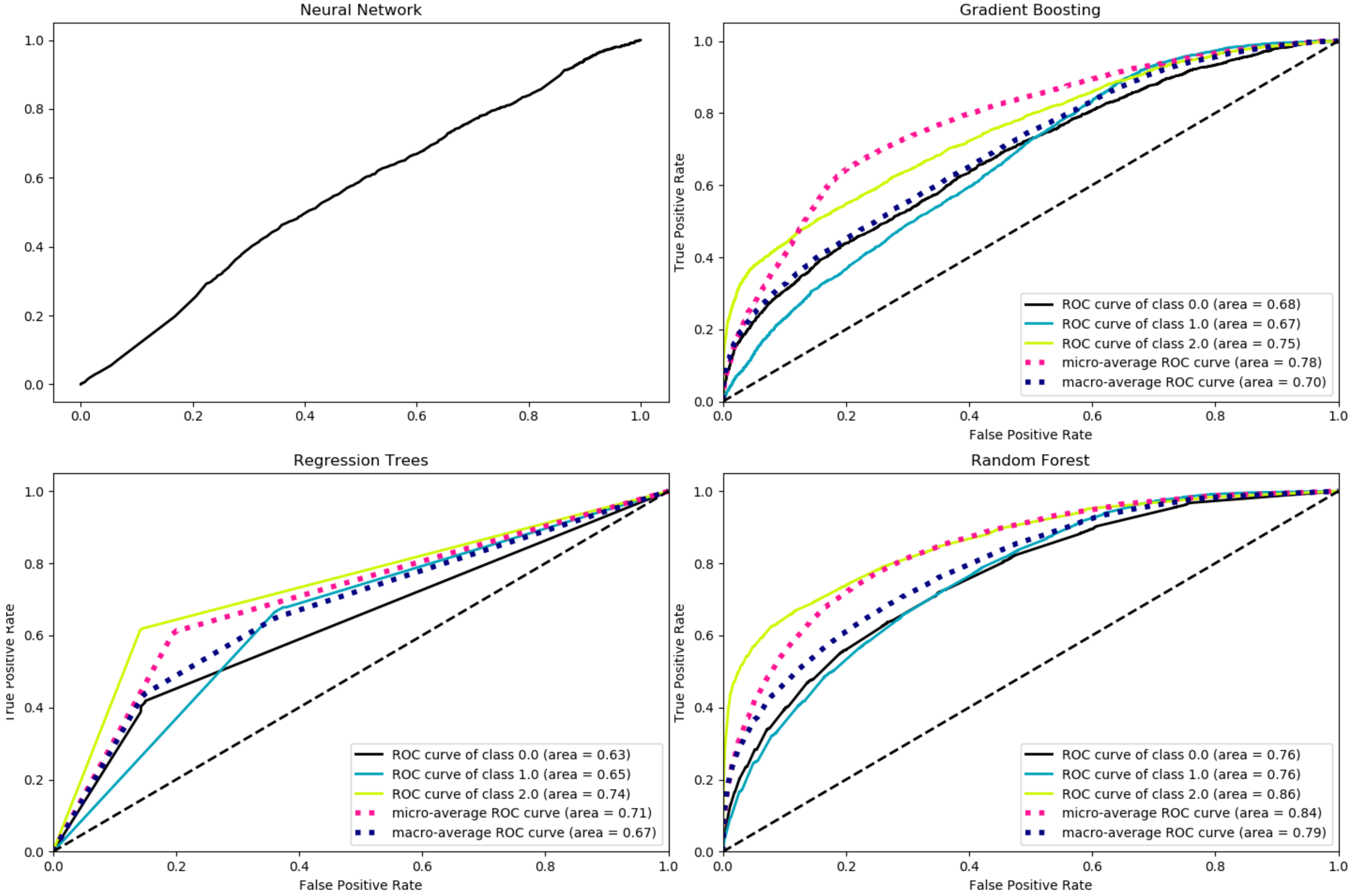
MODEL CONFUSION MATRICES

- Neural network performs the worst – likely operator error (me)
- Gradient boosting performs well but tends to underpredict viral tweets
- Regression trees predict virality the best
- Random forests performs best out of the models analyzed, but also frequently underpredicts virality



MODEL ROC AND AUC

- Random forests outperforms all other prediction methods
- Gradient boosting and regression trees perform similarly



WHY SUCH POOR PERFORMANCE?

- Poor predictors
  - Sentiment analysis is not enough to analyze and assess the topics of a tweet
  - Topic modelling using Latent Dirichlet Allocation to identify topics and use those topics as predictors
- Using a network analysis approach
  - Replace follower count with degrees of centrality and connectedness
- Throw everything but the kitchen sink – use every single piece of data associated with the tweet and weight the data according to how effective it is in prediction
- Maybe no way to predict a viral tweet