Data Ventures: Practical 2

Toxic Comments Classification

Spring 2018

1 Description

A big issue in AI lately is building tools that improve online conversation. In this competition, you'll build a classifier to detect different types of hateful comments. Your goal is to predict a probability of a comment being in the categories 'toxic','severe toxic','obscene','threat', 'insult', and 'identity_hate'.

2 Rules

Outside packages are allowed, but outside data isn't! No pretrained models. As you make submissions, the leaderboard will be updated. The leaderboard is only an estimate of your true score. Half of the test data will be set aside for the final score calculations (to avoid overfitting on the testing set, and incentivizing a large number of submissions).

Scoring will be done based on average "log loss" accuracy.

3 Baseline

In order to receive credit for the assignment you must beat the baseline, an off the shelf Random Forest model, with a log loss of 0.199. The code for this is on the github.

4 Submission

The upload format is a CSV with a 6 columns with the predicted probability of each category.

Click here: `https://goo.gl/forms/PDGqATnB3obWboxR2`