Kaggle Report-Team ABBABA OH NO ABBBAB

Nick Halliwell, Aina Lopez, Yaroslav Marchuk

March 14, 2016

Introduction

Our team consists of Nick Halliwell, Aina Lopez, and Yaroslav Marchuk. In this competition, we were given both a training and test set consisting of features of various web links from mashable.com. We were asked to predict whether the website link would fall under one of five potential categories: Obscure (1), Mediocre (2), Popular (3), Super Popular (4) and Viral (5).

Method

1. Create new variables The url of the articles include useful information that can potentially imporve our prerformance in the competition. An example of the url is the following:

http://mashable.com/2014/01/12/game-of-thrones-season-4-trailer

From the url took the year, day and month it was published as well as some keywords of the content of the article.

Using text mining techniques we collected all the keywords of the article titles. With this we created new variables indicating whether a specific word appears in the title of the article. Given that we have many keywords, we created variables only for those keywords that have appeared in at least 150 observations.

2. Feature Selection
3. Create the model
4. Evaluate the model
5. Upload to Kaggle

Evaluation

In order to evaluate our models, we decided to use 5-fold Cross Validation

explain 5-fold CV

Variable Selection: RFF

Models

- Random Forest
- K-nearest neighbors
- SVM
- Boosting

Results

(plot ?) Kaggle score vs CV score