Can you...

Spot A Bot?

galvanize

Haven Gumucio

@controlinfmanip A lot of life curves wise, a little 'selfhelp for women

@controlinfmanip A lot of life curves wise, a little 'selfhelp for women

did I miss the memo about a fireworks show south of the pier or is that just typical “rich bored guy with a boat…'

did I miss the memo about a fireworks show south of the pier or is that just typical “rich bored guy with a boat…'

Process

Explore the Data

- 4.5 million profiles
- 38 features
- Requested from Crowdflower

Spark

A programming software for parallel processing of large data

Classification models

- Random Forest Model for profile-based research
- Logistic Regression Model for text analysis of tweets

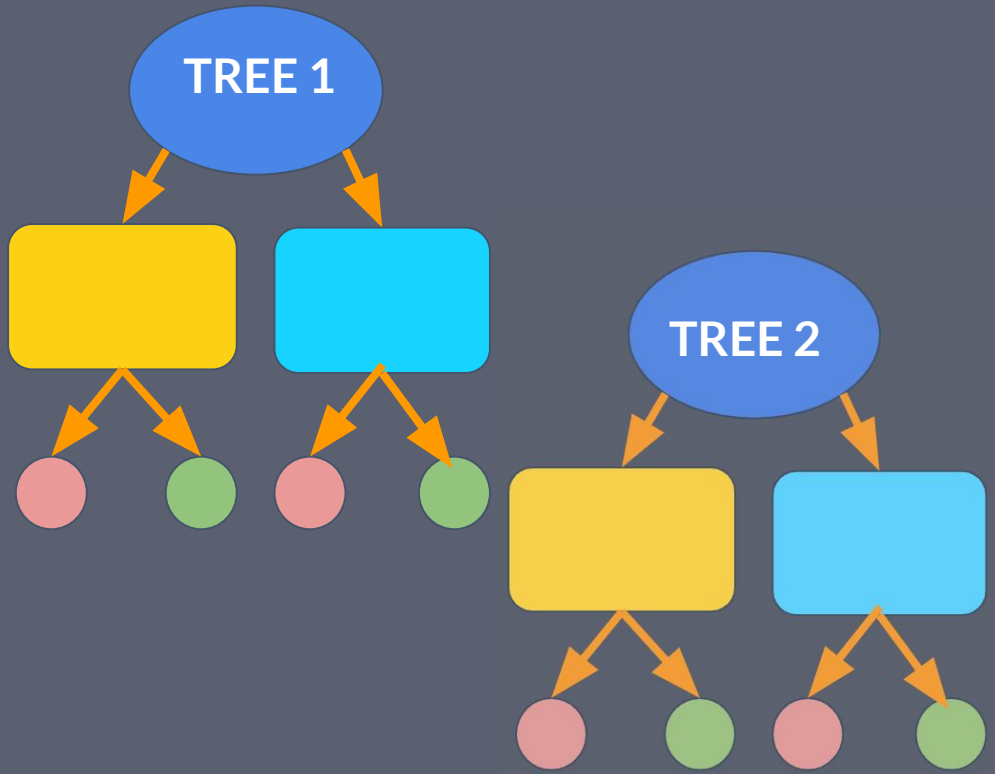
#FeaturesofFakes

- More friends
- Fewer followers

- Twitter Bots also tag things as ‘favorite’ more often
- 3x’s More Retweets (rt)

- Spambots predictively generate on a schedule

Random Forest Determined Features



Status, Friends and Followers

Language and Created_At times

Favorites and Listed Counts

Profiles with Background Images

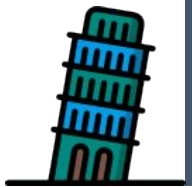
… to get a 92% accuracy score

Text Analysis

Logistic Regression is a statistical model used to categorize data in terms of its likelihood of belonging to a particular class. In our case, it was built to determine whether a tweet was ‘tweeted’ by a bot… or not.

Bias

This is the difference between the predicted and actual values.



REAL ITALIANS

Voglio	I want
Vero	true
Piacere	pleasure
Dev vero	totally
Li pensavo	I thought
Bisogno	need
Vorrei	would like

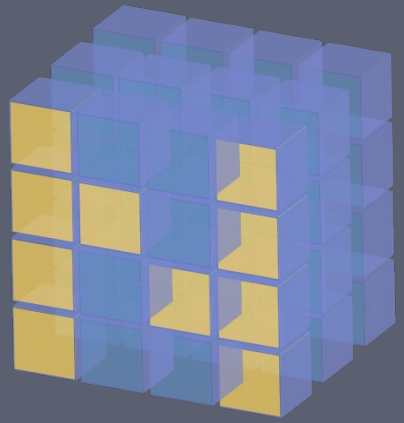
IMPOSTER ITALIANS

RETWITTA capitato Retweet
Sedemmo torto We sat wrong
Grandi dubbi great doubts
Torto posti wrong place
Posti occupati occasional place

With Great Power Comes Great Responsibility...

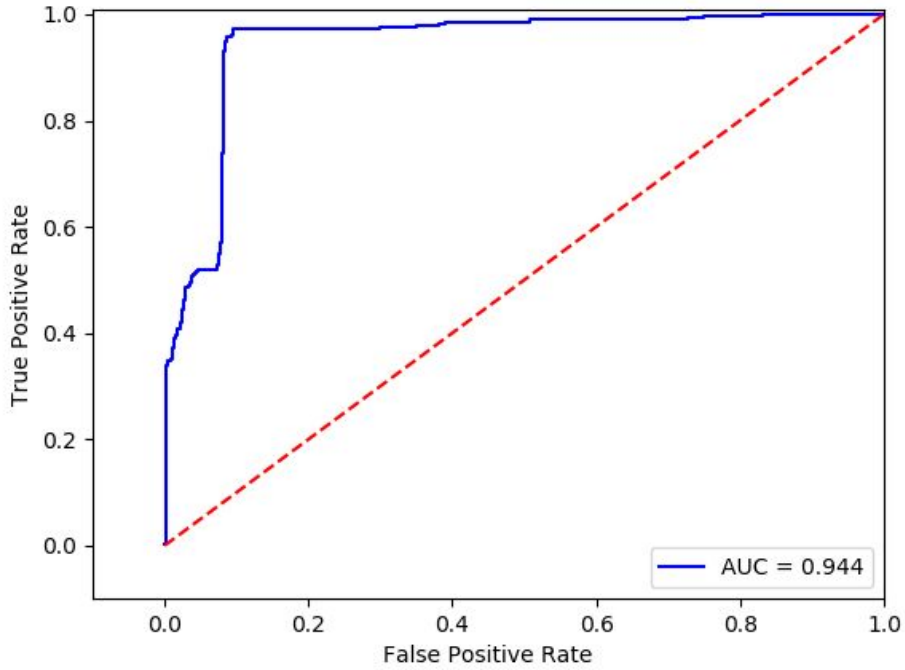
As highly influential platforms, online social networks should be held accountable for monitoring the content of postings and endorsements.

Questions?



NumPy

haven.gumucio@gmail.com
linkedin.com/in/haven-gumucio
GitHub.com/HM618



Final Logistic Regression

94 %

While my model continued to predict real profiles more often than bots, thorough cleaning, balancing and vectorizing suggests further text analysis for classifying spambots is worth pursuit.

'RT @GasperiniLuca: Secondo la matita di #Giannelli #Renzi cerca di "issare" #bandiera80 in prossimità

The World is Changing...

No one can argue the incredible impact social media platforms have made on our day to day lives. Yet as the landscape of technology has continued to expand many of these outlets have grown into main sources for news, information and a general barometer of cultural outlook.

Let's Keep Up

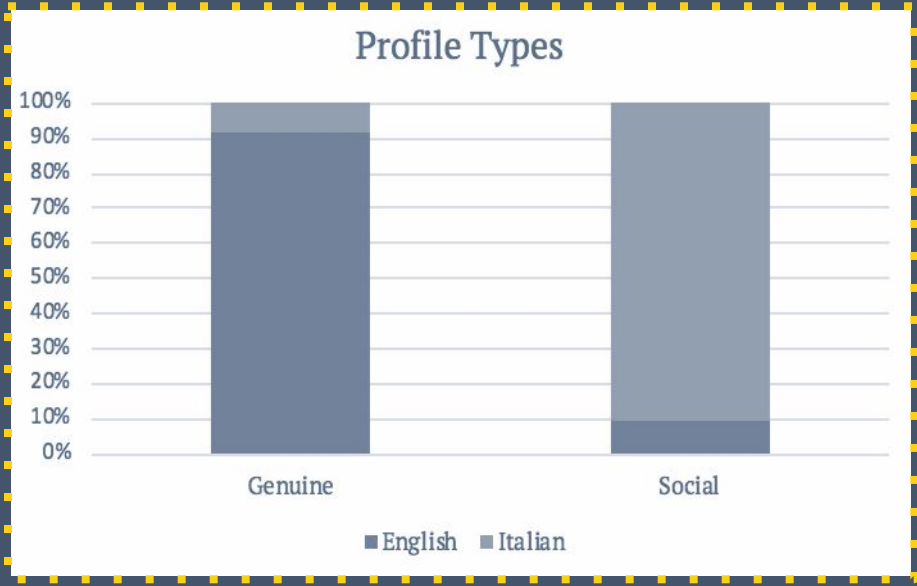
A Brief History:

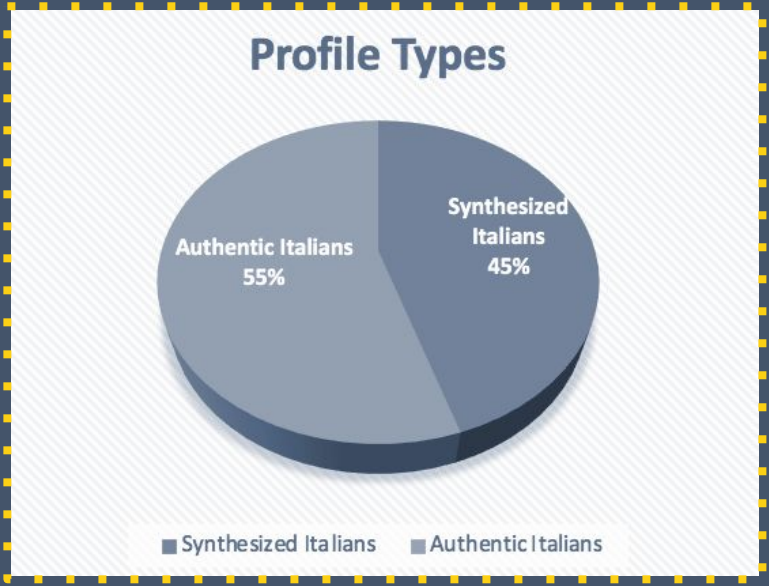
Social Media is the natural evolution of communication technologies.

- 1960's & 70's SuperComputers arise
- 1980's Internet Relay Chats
- 1997 Six Degrees, the first self-professed social media site is born
- 1999 the term 'WebBlog' is coined
- Early 2000's usher in MySpace, LinkedIn, Flickr and Photobucket
- 2005 YouTube is born, followed shortly by the birth of Facebook and Twitter in 2006

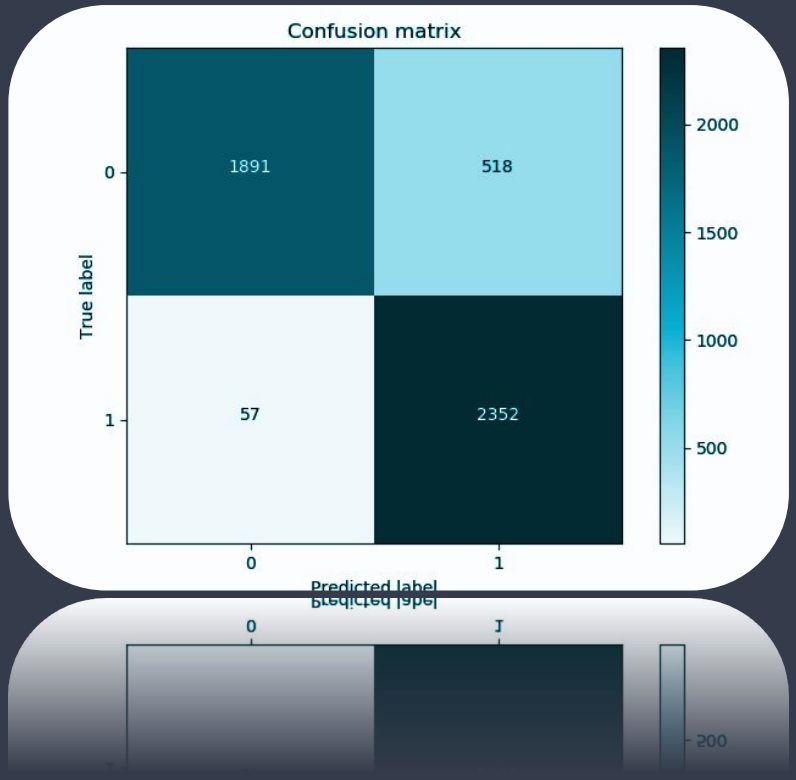
Balanced and Only Italians

While I was able to feed my model data that was only in Italian, the ratio of bots to nots was still highly imbalanced. Spark programming language does not have a package to automatically balance the data and so I manually adjusted the data samples to train and fit my model.





Random Forest Classifier using:



Status, Friends and Followers

Time Zone and Language

Favorites and Listed Counts

Profiles with Background Images

… to get a 92% accuracy score