Analysis and Clean

Comments Dataset

Contents index

1 Comments Dataset

Number of features	36
Number of examples	1,278,205
Number of bots	260,469
Number of trolls	6,671
Targets	2. (is_troll and is_bot)
Number of author	15924

2 Cleaning Dataset

2.1 Changes by dataset columns

Columns	**Change for the clean dataset**
id	Object (string) encoded to int64
author_link_karma	without changes
author_comment_karma	without changes
author_created_at	without changes
author_verified	Object (boolean) encoded to int64 (1 and 0)
author_has_verified_email	Object (boolean) encoded to int64 (1 and 0)
subreddit_id	Object (string) encoded to int64
approved_at_utc	Eliminated due to missing data
edited	without changes
mod_reason_by	Eliminated due to missing data
banned_by	Eliminated due to missing data
author_flair_type	Object (string) encoded to int64
removal_reason	Eliminated due to missing data
link_id	Object (string) encoded to int64
author_flair_template_id	Object (string) encoded to int64
likes	Eliminated due to missing data

banned_at_utc	Eliminated due to missing data
mod_reason_title	Eliminated due to missing data
gilded	without changes
archived	Object (boolean) encoded to int64 (1 and 0)
no_follow	Object (boolean) encoded to int64 (1 and 0)
author	Object (string) encoded to int64
num_comments	without changes
score	without changes
over_18	Object (boolean) encoded to int64 (1 and 0)
controversiality	without changes
body	without changes because this cause for calculate the entropy
link_title	without changes because this cause for calculate the entropy
downs	Eliminated due to data homogeneity
is_submitter	Object (boolean) encoded to int64 (1 and 0)
subreddit	Object (string) encoded to int64
num_reports	Eliminated due to missing data
created_utc	without changes
quarantine	Eliminated due to data homogeneity
subreddit_type	Object (string) encoded to int64
ups	without changes
is_bot	Object (boolean) encoded to int64 (1 and 0). Delete and add in targets.
is_troll	Object (boolean) encoded to int64 (2 and 0). Delete and add in targets.

2.2 Clean Dataset description

Number of features	26
Number of examples	1,278,205
Number of bots	260,469
Number of trolls	6,671
Targets	3. (0 is a normal user, 1 is a bot and 2 is a troll)
Number of author	15924

3 Create new dataset Group by “author”

3.1 Changes by clean dataset columns

Columns	**Change for the clean dataset**
id	Eliminated by irrelevance
author_link_karma	without changes
author_comment_karma	without changes
author_created_at	Eliminated because it is the same value for each author
author_verified	without changes
author_has_verified_email	without changes
subreddit_id	Eliminated because I think it irrelevant
edited	without changes
author_flair_type	without changes
link_id	Eliminated because I think it irrelevant
author_flair_template_id	Eliminated because I think it irrelevant
gilded	without changes
archived	Deleted because it is deprecated
no_follow	without changes
author	Eliminated because the new dataset will be created from this feature

num_comments	without changes
score	without changes
over_18	without changes
controversiality	without changes
body	Eliminator because it is not possible to operate on this feature (“ It could be considered for entropy ”)
link_title	It is postponed because it is considered for entropy
is_submitter	without changes
subreddit	Eliminated because I think it irrelevant
created_utc	It is postponed because it is considered for entropy
subreddit_type	Eliminated because I think it irrelevant
ups	without changes
target	without changes

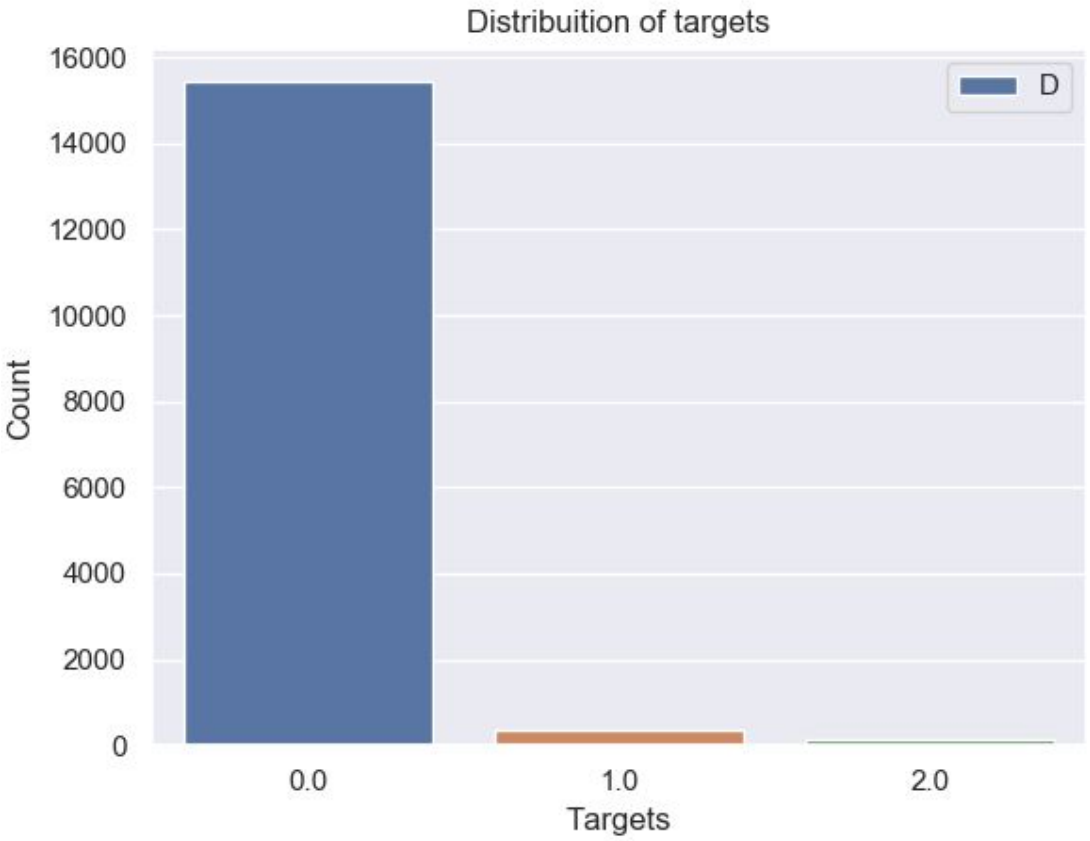
***Note**: For the creation of new examples. The data are grouped by author and the mean of the selected features was obtained.

3.2 New Dataset description

Number of features	14
Number of examples	15924
Number of bots	342
Number of trolls	153
Targets	3. (0 is a normal user, 1 is a bot and 2 is a troll)

3.3 New Dataset Targets

Targets	Description	Count
0	Normal User	15429
1	Bots	342
2	Trolls	153



4 Apply Random Forest Algorithm

4.1 Initial Condition

Parameter	**Description**
Training data	70%
Test data	30%
Random State	16
N estimators	1000

4.2 Analysis of results

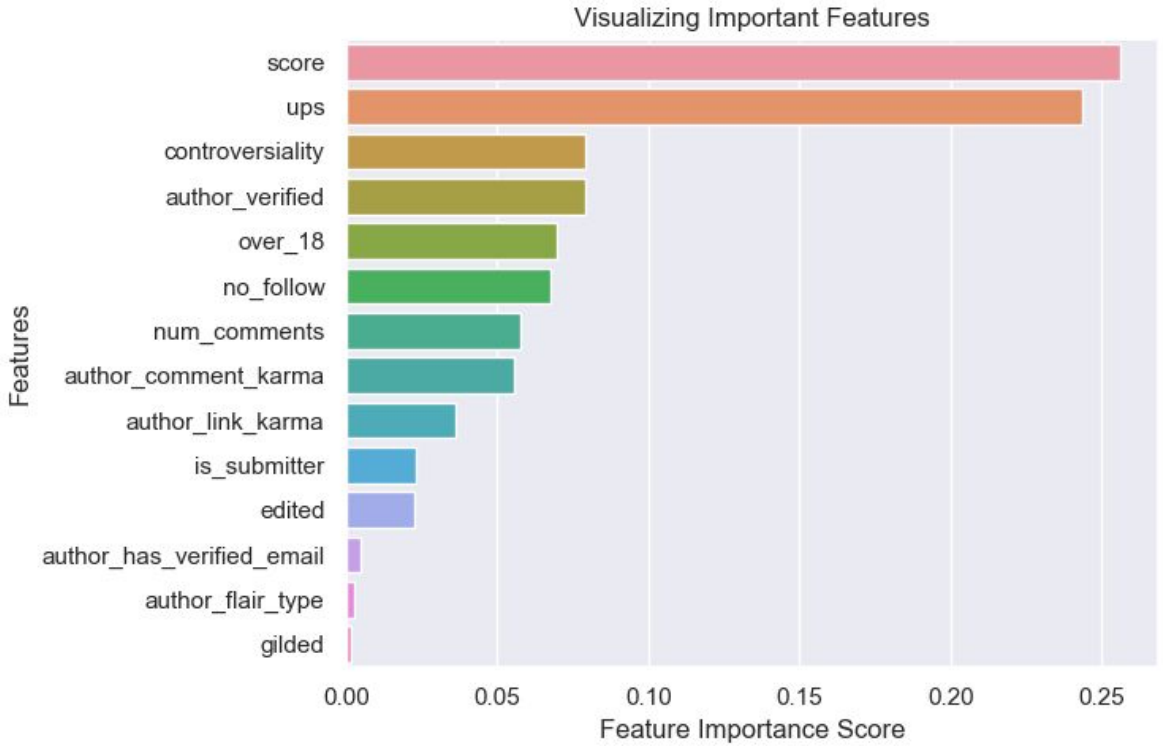
4.2.1 Confusion matrix

Users	**Predicted**	**Normal User**	**Bots**	**Trolls**	**All**
True					
Normal User		4632	1	1	4634
Bots		1	98	0	99
Trolls		6	0	39	45
All		4639	99	40	4778

4.2.1 Metrics

Metrics	Normal User	Boots	Troll
Accuracy	0.9981163666806195		
Matthews correlation coefficient	0.9676011774091072		
F1	0.99902944	0.98989899	0.91764706
Recall	0.99956841	0.98989899	0.86666667
Precision	0.99849105	0.98989899	0.975

4.3 Important features



4.4 Selected features

- ups
- score
- controversiality
- author_verified
- no_follow
- over_18
- author_comment_karma
- author_link_karma
- is_submitter

4.5 Analysis of results

4.5.1 Confusion matrix

Users	**Predicted**	**Normal User**	**Bots**	**Trolls**	**All**
True					
Normal User		4632	2	0	4634
Bots		1	98	0	99
Trolls		3	0	42	45
All		4636	100	42	4778

4.5.2 Metrics

Metrics	Normal User	Boots	Troll
Accuracy	0.9987442444537463		
Matthews correlation coefficient	0.9785426871459745		
F1	0.99935275	0.98492462	0.96551724
Recall	0.99956841	0.98989899	0.93333333
Precision	0.99913719	0.98	1