Spammer detection in Twitter using Social Network Analysis techniques

Tsoy Roman ID:20202040

Department of Industrial Engineering [romatsoy26@unist.ac.kr](mailto:romatsoy26@unist.ac.kr)

**Abstract**

For social networks, such as Twitter, Facebook, it is crucially important to effectively determine spammers, and remove them from social network. As spammers cause harm to users both directly and indirectly. In this project, Twitter dataset was used, with the main 8 features describing user activity, his/her content features etc. By using social network concepts, new features were found which show greater difference between spammer and legitimate users. Using new features, different Machine Learning models, such as k-NN, Random Forest, Logistic Regression, SVM, were used and evaluated using F-1 score, to see how well our ML model classifies spammers and legitimate users. F1-score was used due to imbalance ratio between 2 types of users in the dataset. In addition, some advanced models, such as XGBoost Classifier, were used, and compared with Traditional ML models based on F1- score, and ROC-AUC curve. As a result, the findings of new features with the help of social network analysis techniques, combined with best model, led to better and more effective way of differentiating between spammer and legitimate user.

1. **INTRODUCTION**

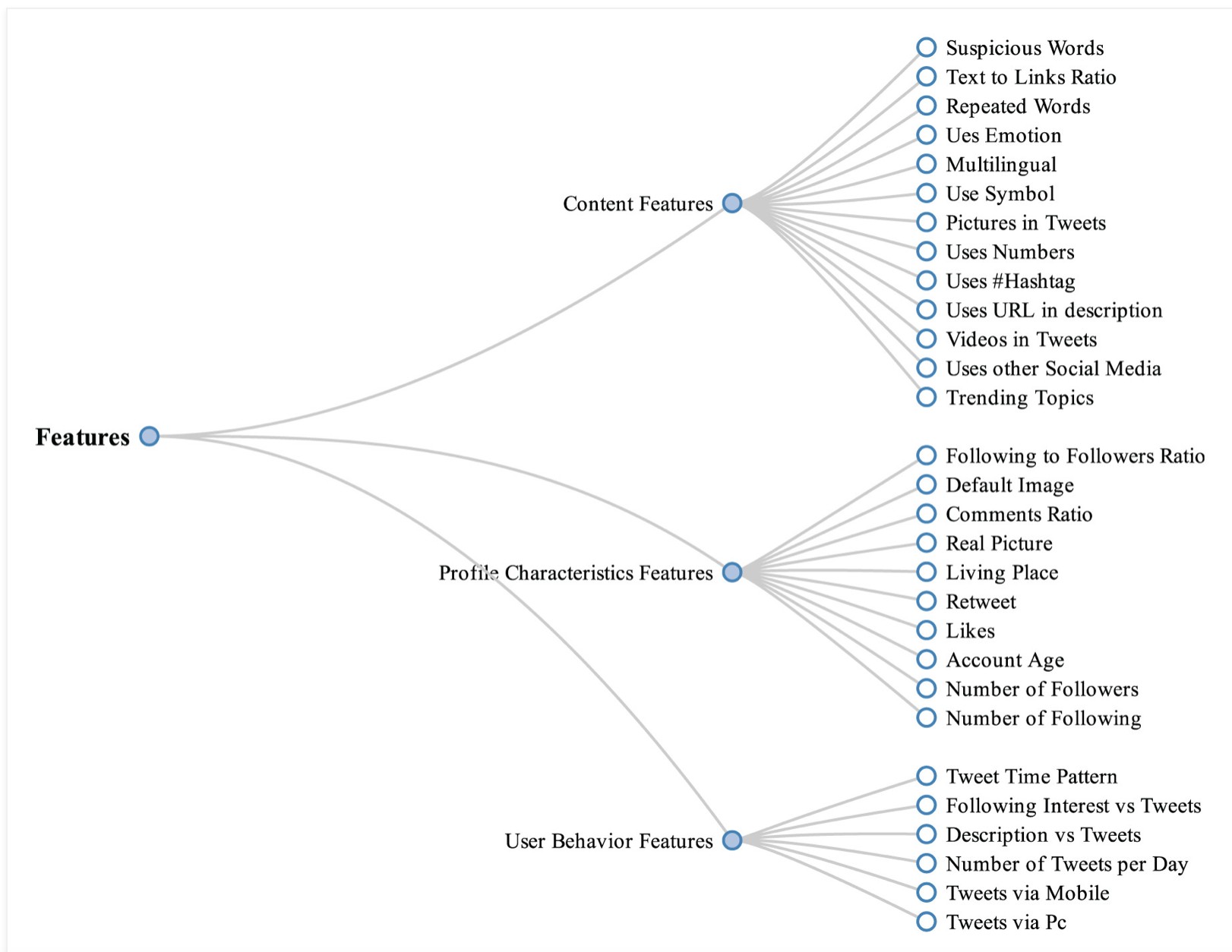
Nowadays, the number of new users in social networks increases very rapidly. With increased number of users, number of spammers increases too. Spam first appeared in the 1990s on the Internet in the form of emails through the emailing systems and has rapidly grown afterwards. Spammer is the person or organization who sends to many users irrelevant or undesirable contents using social media in purpose of advertising, promotion, criminality and etc. Spammers cause harm to both user and social network itself, as they do not use it for primary purposes. Moreover, with the advance in IT, today spammers can even steal your personal info, and damage network.

By sending the same content, links to enormous number of users, spammer makes lots of users to see undesirable content, wastes their time, and can even violate their privacy issues. So, it is crucially important to determine spammer as soon as possible, and decrease harm it causes to social network’s community. The huge number of spammers will lead to decreasing user’s trust in using the social network, it also decreases the level of performance of social network.

In this project, it was proposed to consider new features, that were obtained using social network concepts. By setting and checking different hypotheses based on new features, ML models were trained and tested to result in better investigation of spammer in social network such as Twitter.

1. **LITERATURE REVIEW**

The main reference that was used for this project was “Spam profiles detection on social networks using computational intelligence methods: The effect of the lingual context” paper written by Ala’ M Al-Zoubi. This article addresses the nature and the characteristics of spam profiles in a social network like Twitter to improve spam detection. This paper was chosen as main reference, due to the way of how they do feature engineering to improve space detection system.



In this paper, they obtain 29 features, and categorized them in 3 groups: Content Features, Profile Characteristics Features and User Behavior features. Using 3 types of features, they improved their spam detection system, as it captured the important features between legitimate user and spammer.

1. **EDA AND METHOD**

In Twitter data, total 8 features were presented. In terms of social network concepts, the user is defined as node. Each user has in degree due to number of people following the user. It has out degree, the number of people he is following - friends. In dataset, AvgRetweet and AvgMention are 2 variables describing how many nodes were involved and connected by the user during his activity. Avg Hashtag, AvgURLCount were used as - two content based features. While User Description Length - number of words written in description of his/ her profile. The last feature that was presented - TweetCount, representing number of total posts done by the user.

First to mention, the dataset is imbalanced, as number of spammers tends to be much less in social networks like Twitter.

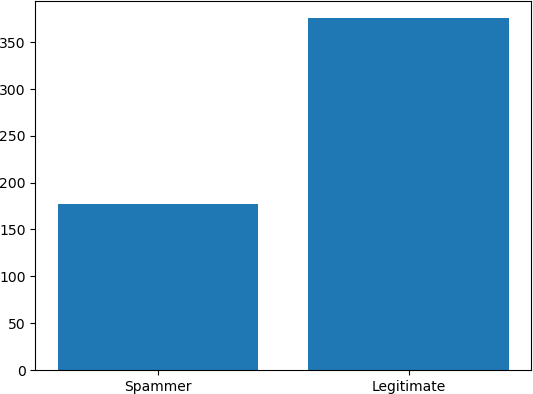
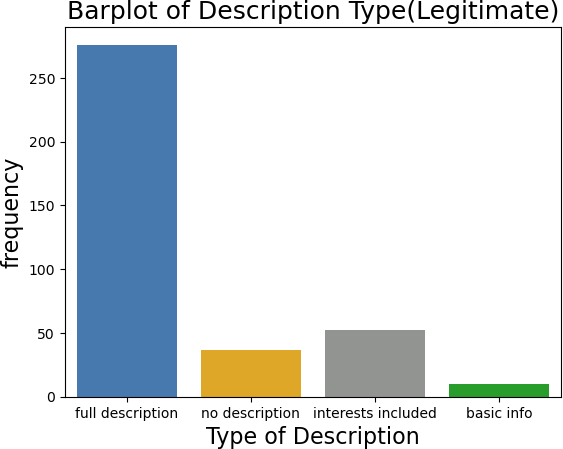


Figure 2. Ratio between Spammer & Legitimate user

It was set the hypothesis that AvgRetweet should be high for spammers, as in equal time, spammer involves, connects more nodes in his activity rather than legitimate user. And the hypothesis was confirmed:

It makes to consider AvgRetweet value as significant one. While Hashtag and URL count variables were found to be insignificant features. It could be explained with the fact that usually spammers tend to share 1-2 links in the messages they sent, the same can be applicable to legitimate users.

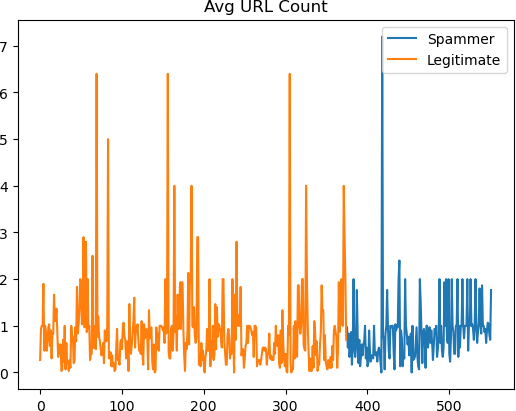
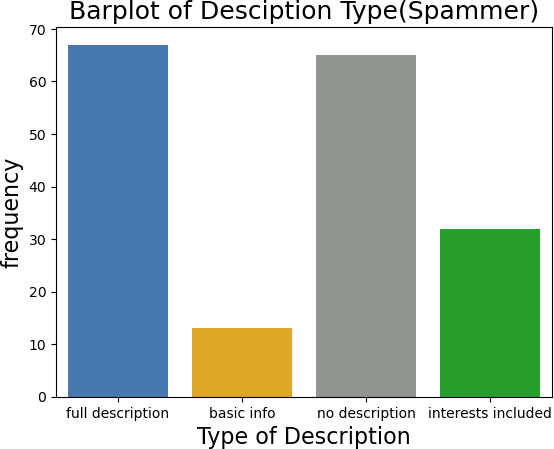
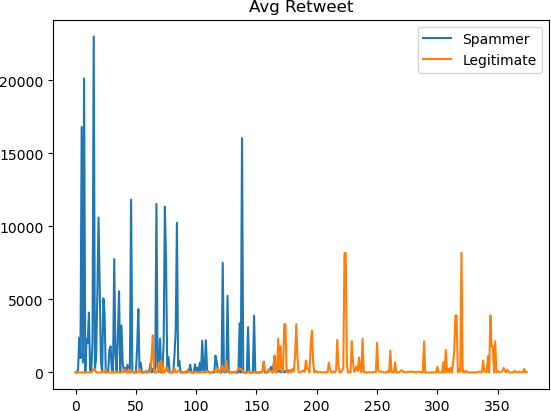


Figure 4. Avg URL count Considering User Description Length variable, the

number of words in description were split in 4 groups: 0

- no description, 1- 10 - Name, 10-50 - Name, hobby, 50-160 - Full description. Although, it was found that spammers also fake their user description, the second highest number of spammers tend to have no. description. While for legitimate users, top 2 was 10-50 words and full description.





Now, Feature engineering starts, and in the following part, methods of obtaining new significant features will be described. The first 2 features were developed using degree ranking methods. First one is Copeland score, which equals to Outdegree (Number of friends) - Indegree (Number of followers). Second was degree ratio score which was calculated as:

Degree ratio score = Outdegree / Indegree

Both scores were thought to be high for spammers, and low for legitimate users. Due to the fact, that out degree tends to be very high compared to in degree for spammer, because not many people follow spammer account. It was proved with the following charts:

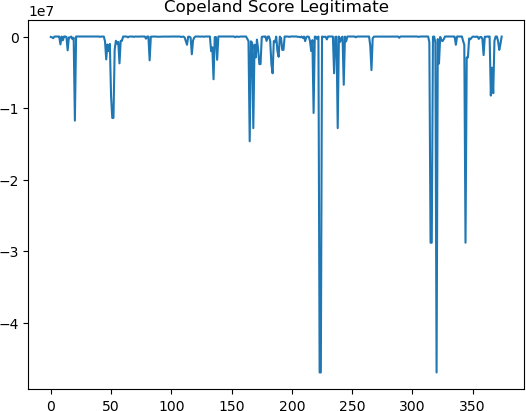
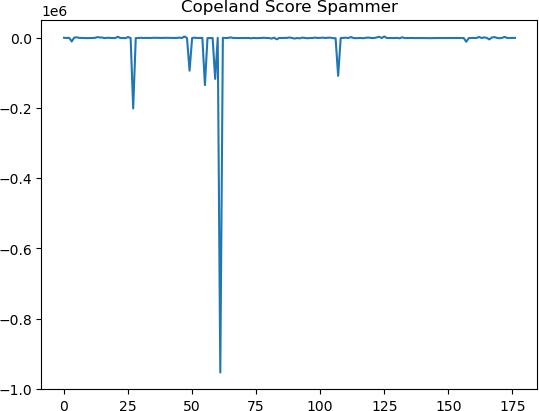


Figure 6. Copeland Score Copeland score for both are negative scores, but

Copeland score values for spammer are to the power of 10^6, while for Legitimate is 10^7, which proves our hypothesis that Copeland degree for spammers is much higher.

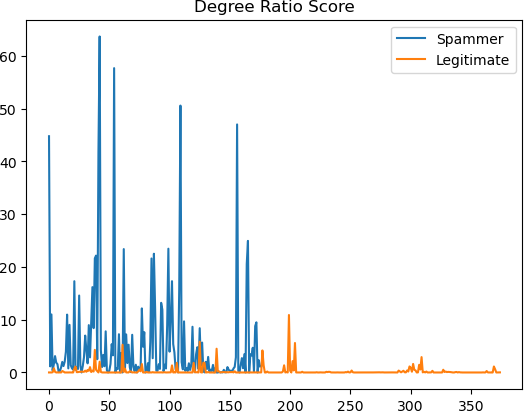


Figure 7. Degree ratio score

The same was true for Degree ratio score, where values for legitimate user is much less than values for spammer.

As next feature, the account age in days was calculated, by subtracting from current time the time when account was created. It was hypothesized that account age for legitimate users tends to be much higher, as sooner or later, spammer nodes are found in social network, and removed.

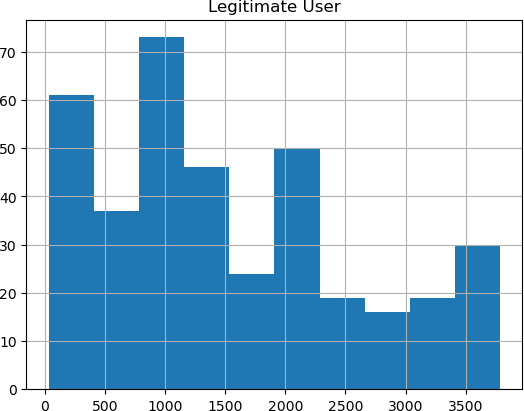
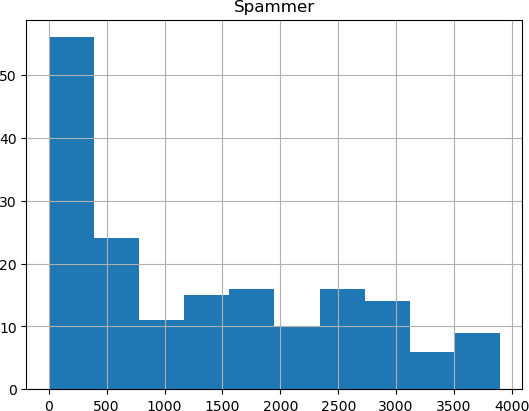


Figure 8. Account Age

As, it is clearly seen, most of the spammers have quite small account age, while for legitimate users results vary a lot. The next feature that was created is Tweet per day: Tweet Count / Age of Account. It was hypothesized to be high for spammers, as usually spammers have a very high activity for short period of time.

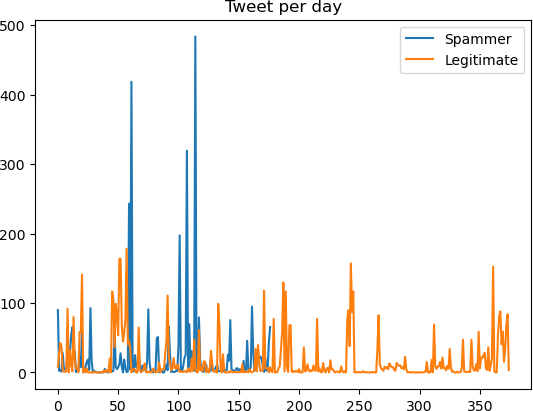


Figure 9. Tweet per day

It is clearly seen that range of values for legitimate user at maximum reach around 200 tweets, while for spammer it reaches 500 tweets day, which makes this feature also significant for our spammer detection model.

Finally, using social network concepts, the rate of creating new edges (connections) feature was made using UserFriendsCount / Age of account (Longetivity). And here, spammers tend to have higher rate of creating new edges, as for them it is much more beneficial to connect as many people as possible in shorter amount of time.

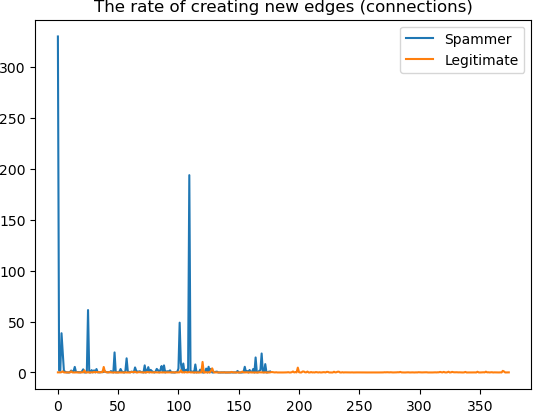


Figure 10. The rate of creating new edges

The rate of creating new edges is also significant for our spammer detection model.

1. **APPLICATION**

After setting and obtaining new significant features, Traditional Machine Learning models started to be applied. The total list of features that were used, is as following:

1.UserDescriptionLength 2. AvgRetweet 3.AvgMention 4.Copeland score 5.Degree ratio score

6.Account age: Current time - UserCreatedAt 7. Tweet per day: Tweet Count / Longetivity (!)

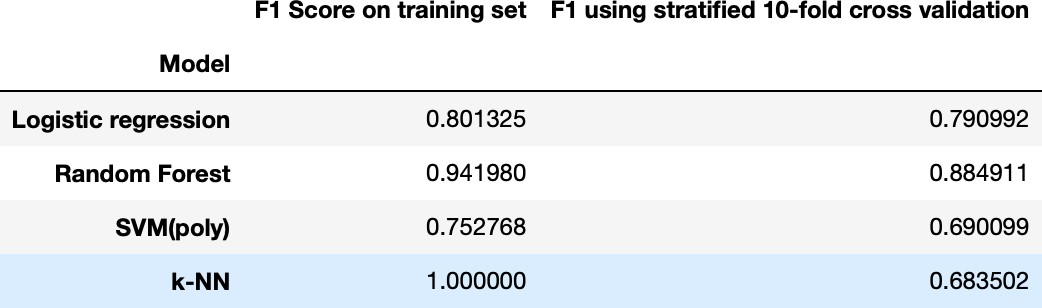
8. The rate of creating new edges

The following model were fitted with training data consisting of features above: k-NN,Logistic Regression, SVM with polynomial kernel, and Random Forest. All 4 models were tuned with hyperparameter class\_weight set to ‘balanced’ to take care of imbalance in the dataset.

Besides, ML models, also XGBoost Classifier was used to detect spammers, and it was tuned with the following hyper parameters: random\_state = 0, scale\_pos\_weight = 2, max\_depth = 1.

1. **RESULTS AND DISCUSSION**

Regarding the results for ML models, due to imbalance data, they were evaluated based on Stratified 10-fold CV on F1 score (metric dealing with imbalance data) The following results were obtained:



k-NN model is the only one that overfits, as F1 score on training set is 1, while F1 - score on 10-fold CV is 0.683502. Comparing other 3, Random Forest tend to perform better than others with F1 score on test set equal to 0.89189 with the following confusion matrix, where only 8 nodes(users) were not classified correctly:

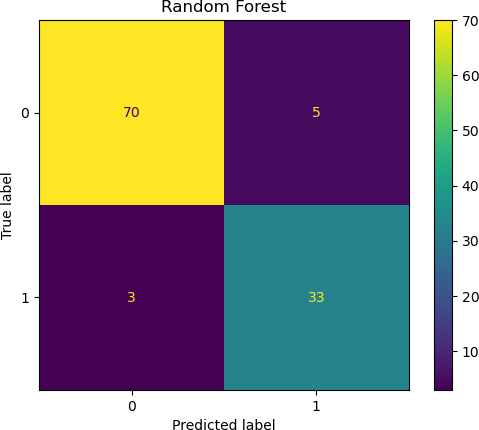
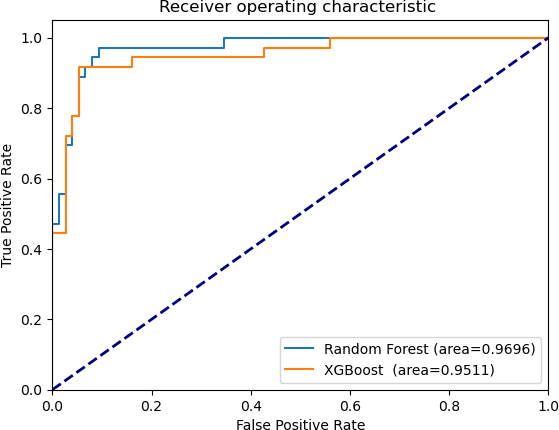
 

Figure 11. Random Forest confusion matrix

Regarding XGBoost Classifier model, it also performed very well with the same F1-score on test set - 0.89189.

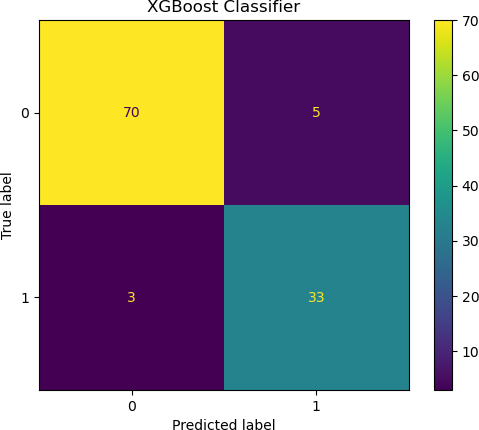
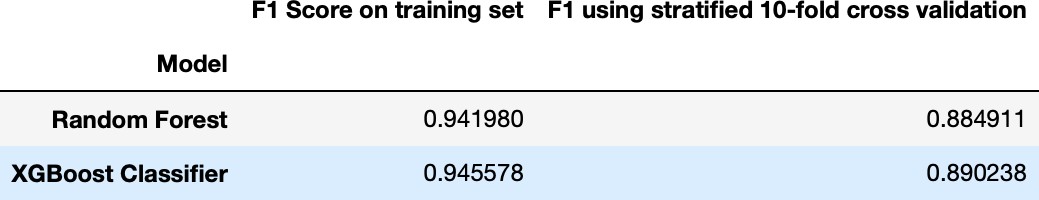


Figure 12. XGBoost confusion matrix Thus, it can be concluded that 2 best models from

chosen are Random Forest and XGBoost Classifier, as following:



2 models were also compared based on ROC-AUC curve, and they also perform mostly similar very well.

Figure 13. ROC-AUC

1. **CONCLUSION**

As conclusion, it can be said, that by deriving new important features using social network techniques, the F1 score for the same models (Random Forest and XGBoost Classifier) was improved from 0.84 to 0.89 constructively. It will allow to detect spammers in more effective and faster away, which will result in more friendly use of social network, decreasing number of privacy violations, crimes, irrelevant content, and damage for network. With such high increase of F1 score, the number of spammers eliminated will be much higher in the same period, and it will lead to better use of social networks such Twitter for legitimate users.

**REFERENCES**

1. Ala’ M Al-Zoubi. SPAM profiles detection on social networks using - sage journals. Retrieved December 12, 2022, from https://journals.sagepub.com/doi/full/ 10.1177/0165551519861599
2. DenBrink, R. van. (2020, July 3). The degree ratio ranking method for directed graphs. European Journal of Operational Research. Retrieved December 12, 2022, from https://[www.sciencedirect.com/science/](http://www.sciencedirect.com/science/) article/pii/S0377221720305567