DAT 554 Final Project: Classification of Fake Instagram Accounts

Carissa Hicks
University of Illinois at Springfield, Data Analytics

Abstract—**social media can be one of the most dangerous platforms on the internet. Illegitimate accounts can be used for a variety of purposes including increasing popularity of a main account to join an affiliate program, emotional scams, or phishing. The purpose of this study is to explore a dataset consisting of 994 real accounts and 200 fake accounts for features of publicly available information to identify trends indicative of fake accounts. Machine learning algorithms including logistic regression, support vector machine, random forest, naïve bayes, and neural network will be applied to build a model to detect these accounts.**

Keywords - machine learning, Instagram, social media, fake accounts, binary classification

I. Introduction

Social media is a very important part of many people's daily lives. It is a communication medium for business to market their products, for influencers to interact with their fans, and where family and friends share and like posts amongst each other. Some of the most notable social media applications include Facebook, Twitter, Instagram, and YouTube [10]. However, as the importance of keeping up with relevant people, businesses, and happenings increase, so do methods of taking advantage of these systems.

Instagram is a free photo and video sharing app where a user can follow people, gain followers, share, comment and like posts. The popularity of an Instagram account can be measured by account engagement. Engagement includes total followers, total posts, total comments on each post, and total likes on each post. [11] Fake accounts to increase these metrics can be created using automated generators. These accounts can be programmed to like posts, follow accounts, and even comment and share posts automatically to increase the account's engagement. The popularity and even reliability of an account depends on these metrics. These accounts can engage in Sybil attacks (operating many fake identities simultaneously), scams, or sending malicious links and spam to other users. The detection of these kinds of accounts is vital to avoid these events.

It is important to note that there is a distinction between the usage of creating fake accounts for identity theft, and accounts used solely for increasing metrics and spam. There would be more effort going into an account that wants to impersonate an individual such as a celebrity or target's family member. These accounts would include a profile picture and style their username to match the impersonated individual's real username as closely as possible. However, these accounts can still be detected by their follower count and number of posts. This study will focus on detecting accounts made for increasing metrics and spam.

Instagram has already introduced measures to combat against bots and scammers. "We've built machine learning tools to help identify accounts that use these services and remove the inauthentic activity." [19] Although, services that offer Instagram automation are publicly available and easy to use. Colleen Christison, a writer at Hootsuite, demonstrated various apps available that use Instagram automation and their ease of use [8]. In recent years, several research articles have been released introducing fake account detection. Akyon et al [2] provided the dataset that was utilized in this analysis. The authors make a distinction between fake and automated accounts and were able to classify them with 86% accuracy using a neural network model.

Sheikhi et al [21] introduced a machine learning model based on features such as username length, profile picture, the number of followers vs following, if it is a public or private account, post count, and others. They found that most accounts that do not have a profile picture and/or are following more people than they have followers, are usually fake accounts. They were able to obtain a 98% classification accuracy using bagged decision trees.

This project will answer the following questions: What are the characteristics of a fake account? Can fake accounts be detected with publicly available data?

II. Dataset

The dataset was obtained from GitHub uploaded by Faith Cagatay Akyon. The author of the dataset along with Esat Kalgaoglu also provided their own analysis of the data. This data was hand labeled by the authors using publicly available information. Private accounts do not show the personal information of the profile such as their posts or comments. The goal of this analysis is to analyze data that can be quickly mined.

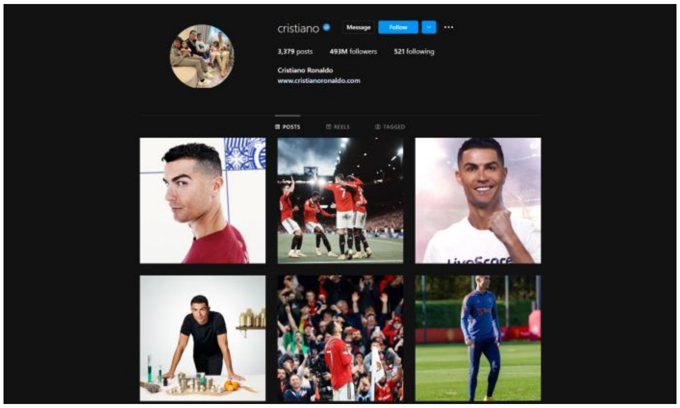


Figure 1 - Example of a real public Instagram account

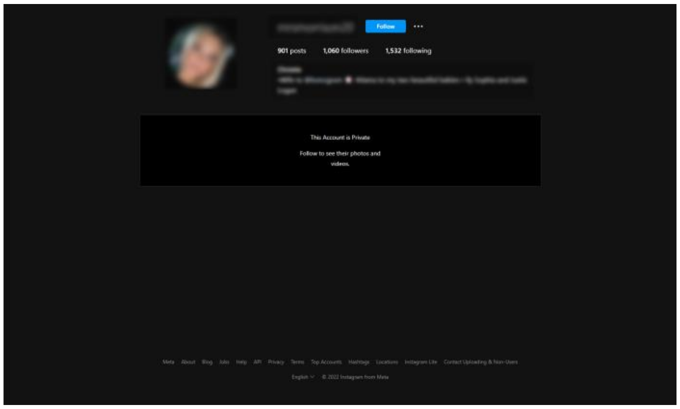


Figure 2- Example of a real private Instagram account

The original package contained csv files of fake accounts and real accounts separated. The fake account file contained 200 observations of 9 variables while the real account file contained 994 observations of 9 variables. The 9 variables are as follows:

- isFake: is the account fake or real (binary)
- userBiographyLength: length of characters in the biography
- userFollowerCount: number of users the account follows
- userFollowingCount: number of users the account is following
- userHasProfilPic: does the account have a profile picture (binary)
- userIsPrivate: is the account private (binary)
- userMediaCount: number of media the account has uploaded
- usernameDigitCount: number of digits in the username
- usernameLength: length of the username

During preprocessing of the dataset, the two separate files were combined into one dataset of 1194 observations. The data did not contain any missing variables or null observations. For initial exploration of the data, isFake, userHasProfilPic, and userIsPrivate were converted to factors since they are binary categorical variables. Then, real/fake was renamed as factor levels to or yes/no respectively for ease of identification.

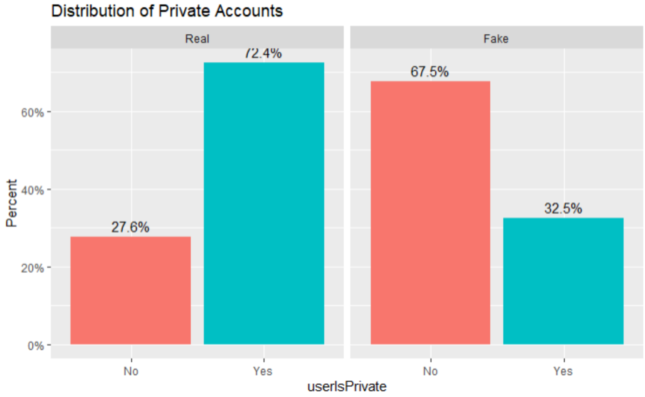


Figure 3 - Data distribution of private vs public accounts

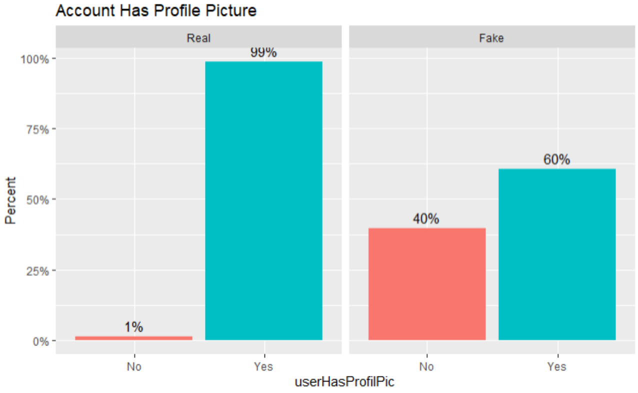


Figure 4 - Data distribution of "profile picture" feature

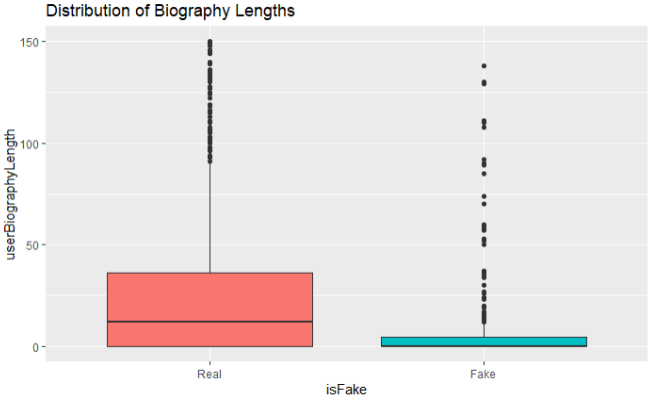


Figure 5 - Data distribution of biography lengths

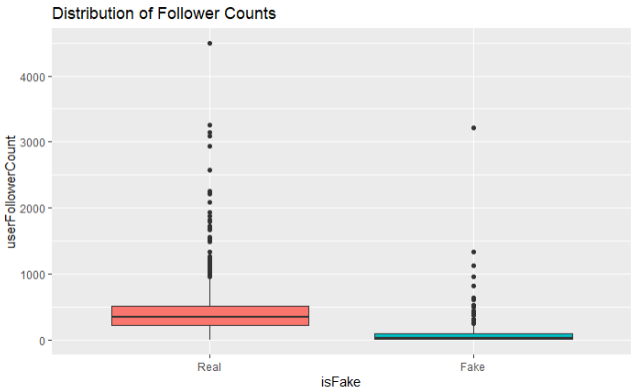


Figure 6 - Data distribution of follower counts

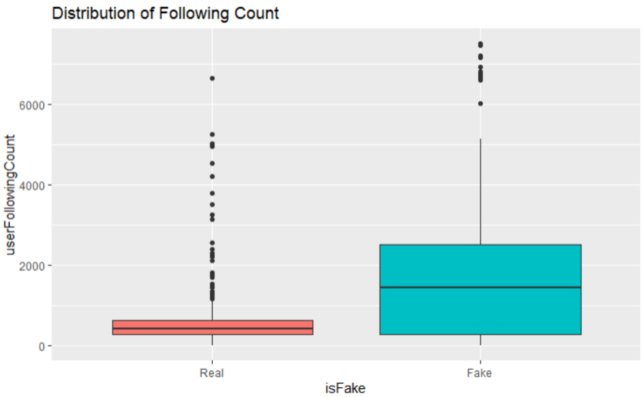


Figure 7 - Data distribution of following counts

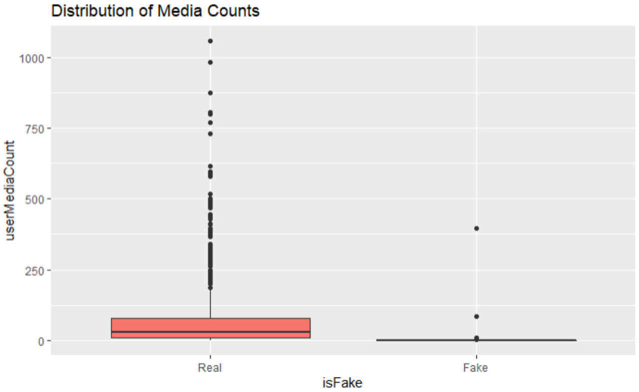


Figure 8 - Data distribution of media counts

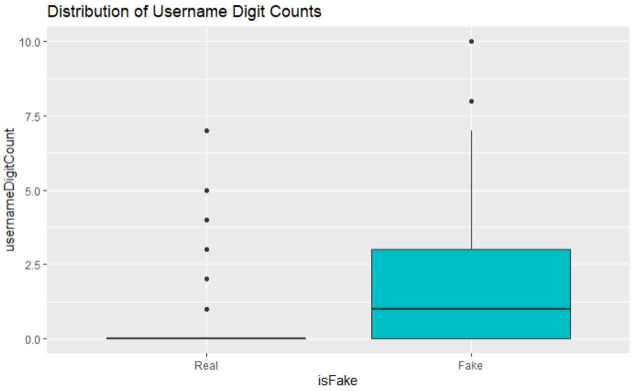


Figure 9 - Data distribution of username digit counts

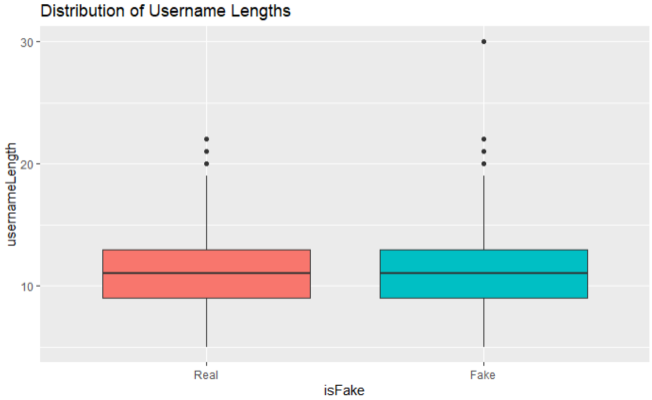


Figure 10 - Distribution of username lengths

Figures 3 – 9 demonstrate the distribution of the data between real and fake accounts. About 72% of Real accounts were private while only 32% of Fake accounts were private. In the next figure it is shown that 99% of real accounts had a profile picture while there was a 40/60 split on the fake accounts. Having a profile picture is a characteristic of a real account. It appears that the length of the biography of real accounts are longer than fake accounts. Real accounts have more followers. Fake accounts follow more people. Real accounts have more posts and media. Fake accounts had more numerical digits in their usernames but also have about the same length of usernames as real accounts.

The results of these visualizations were expected. As discussed previously, automated fake accounts are expected to have less followers than a real account and follow more than a real account as they are usually meant to distribute spam or increase the popularity of a base account. The distribution of media counts is expected as creating posts is not one of their goals. Spreading spam through posts would not be an effective way of distributing their information. The number of numerical digits in the fake account's usernames is also expected. As most of these accounts are generated automatically from software, the software would spend less computational time using random letters and numbers to make usernames rather than search for dictionary words or names. However, the length of the usernames of fake accounts was surprising. It appears both real and fake accounts are around 11 characters long.

Furthermore, an additional helpful feature was added that was derived from userFollowerCount and userFollowingCount called followRatio. Instagram does not allow users to follow more than 7,500 accounts [13]. Fake accounts will near this number to increase their chances of getting a "follow-back" which will increase their stats [11].

$$\text{followingRatio} = \text{userFollowerCount} \div \text{userFollowingCount}$$

There are instances in the data where userFollowerCount or userFollowingCount are zero. This would result in undefined values and create an error. To remedy this issue, all instances in which the counts equal zero will be changed to 1 for the followRatio calculation to take place. A higher ratio indicates that the account has more followers than they are following. The distribution of our follow ratio data indicated that fake accounts have a much lower follow ratio than real accounts as shown in the below figure with the mean indicated with the dotted black line.

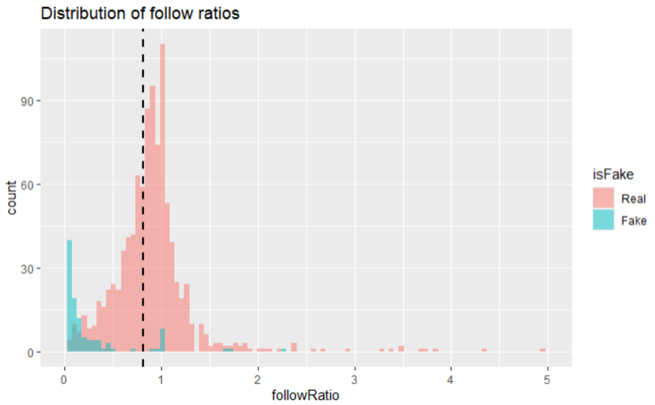


Figure 11 – Follow ratios

One issue with the data is that there is a class imbalance favoring real accounts which may throw off our prediction training. The majority class in this case is real accounts and the minority class are fake accounts. About **83.25%** of the data consists of real accounts whereas **16.75%** of the data consists of fake accounts. This analysis is interested in predicting the minority class – fake accounts.

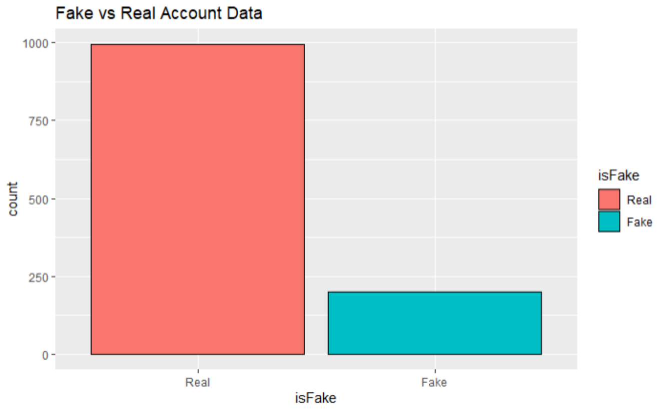


Figure 12 - Bar chart demonstrating the class imbalance

There are several methods to dealing with imbalanced data such as over-sampling the minority class, under-sampling the majority class, and generating synthetic data. Due to the low amount of data in general (under 2000 observations) I will choose to generate synthetic data using the ADASYN method [1]. ADASYN stands for Adaptive Synthetic algorithm which creates synthetic instances of the minority class to achieve a better classifier performance. ADASYN will be applied after splitting the data into train/validation and test sets and only applied on the training set.

Further data exploration was conducted comparing the relationships between the target variable “isFake” with the other variables in the dataset. Categorical variables were explored with chi-squared test and mosaic plot. Using a confidence level of 95%, an alpha of 5% was used for the duration of this analysis. The two categorical variables in the data are userHasProfilPic and userIsPrivate. Both resulted in p-values lower than our alpha and are therefore, relevant to our analysis. The rest of the features were numerical and were analyzed using t-tests. The only variable with a p-value higher than alpha was usernameLength. Referencing back to Fig. 10 it is shown that there is almost no distinguishable difference between the boxplots of real and fake account username lengths. As a result, usernameLength will be removed from our data before building the machine learning models.

III. Solution

First, the rows of the data frame should be shuffled randomly before splitting the data into test, train, and validation sets because the separate real and fake datasets were combined and may show a data separation bias. The data will be split into 80% training, 10% testing, and 10% validation sets. Models used include logistic regression, simple support vector machine, random forest, neural network, and naïve bayes. Then ADASYN, feature scaling using min-max normalization, and cross validation was applied on the training data for relevant models. Min-max normalization was used instead of R’s default normalization method which is z-score normalization.

Min-Max Normalization

$$x' = x - \min(x) \div \max(x) - \min(x)$$

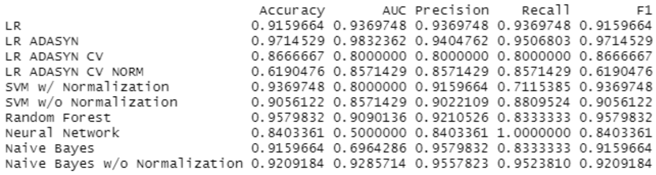
All models chosen are typically used for classification problems. Our baseline model, logistic regression was used to contextualize the results of the other trained models. The following is a list of the models used:

1. Logistic regression (baseline)
2. LR with adaptive synthetic minority oversampling technique on the fake accounts
3. LR, ADASYN, and cross validation
4. LR, ADASYN, CV, and min-max normalization
5. Simple support vector machine with normalization
6. Simple support vector machine without normalization
7. Random forest with ADASYN
8. Naïve Bayes with ADASYN and normalization
9. Naïve Bayes, ADASYN, and without normalization
10. Neural Network
    a. 32 units in layer 1 (ReLU activation)
    b. 32 units in layer 2 (ReLU activation)
    c. 1 unit in the output layer (sigmoid activation)
    d. Adam optimization
    e. Batch size = 32
    f. Epochs = 500

Of all these models, the logistic regression model with ADASYN, CV, and normalization performed the worst. This could be because of the feature scaling technique that was used. Referencing back to the boxplots created in figures 3-10 it is shown that the data contains a high number of outliers. This could be due to celebrity accounts that may have thousands of followers but zero following. These types of accounts are not typical of a normal everyday user's Instagram profile. Although the min-max normalization algorithm performs better than z-score normalization overall it does have the downfall of not handling outliers very well [12]. If this project were to be redone, it would be interesting to see if z-score normalization produced better results.

IV. Evaluation Metrics

Each model was tested on the test data and evaluated using accuracy, area under the ROC curve (AUC), precision, recall, and f1-score. Training accuracy of the models may be low because of the oversampling on the training data which made the training data balanced but the validation data imbalanced.



```
                               Accuracy       AUC Precision    Recall        F1
LR                            0.9159664 0.9369748 0.9369748 0.9369748 0.9159664
LR ADASYN                     0.9714529 0.9832362 0.9404762 0.9506803 0.9714529
LR ADASYN CV                  0.8666667 0.8000000 0.8000000 0.8000000 0.8666667
LR ADASYN CV NORM             0.6190476 0.8571429 0.8571429 0.8571429 0.6190476
SVM w/ Normalization          0.9369748 0.8000000 0.9159664 0.7115385 0.9369748
SVM w/o Normalization         0.9056122 0.8571429 0.9022109 0.8809524 0.9056122
Random Forest                 0.9579832 0.9090136 0.9210526 0.8333333 0.9579832
Neural Network                0.8403361 0.5000000 0.8403361 1.0000000 0.8403361
Naive Bayes                   0.9159664 0.6964286 0.9579832 0.8333333 0.9159664
Naive Bayes w/o Normalization 0.9209184 0.9285714 0.9557823 0.9523810 0.9209184
```

Figure 13 - Results chart

Since our data is imbalanced, accuracy is not a good metric for evaluating the models and more importance was placed on the f1-score because it combines precision and recall into a single metric by using their harmonic mean.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2\ x\ \frac{Recall\ x\ Precision}{Recall + Precision}$$

TP – True Positives
TN – True Negative
FP – False Positives
FN – False Negatives

The Receiver Operating Characteristic (ROC) curve is a curve that graphs the relationship between true positives and false positives, giving a single metric to cover both sensitivity and specificity. The area under the ROC curve (AUC) is used as a metric for classification and represents the model's predictive ability.

The logistic regression model with ADASYN scored the highest amongst all other models in every metric. The next best model would be the random forest model, then Naïve Bayes without normalization. It is interesting to note that the Neural Network model was overfitting by having a validation accuracy lower than the training accuracy and validation loss higher than the training loss. and produced a recall of 1 indicating that there were no false negatives. However, it also has an AUC of 0.5 meaning that it is making random guesses for predictions and has no class separation capacity. This could indicate that the model needs more training on a larger dataset. A future project may want to gather much more data and remove outliers for better training and relevancy.

V. Conclusion

This project discussed the detection of fake accounts used on Instagram and the dangers they pose to social media users. The characteristics of fake accounts included if the account was private or public, if the account has a profile picture, the length of the biography, the number of accounts they are following vs how many follow them, the number of media they have uploaded, and the number of numerical digits in their username.

Several machine learning algorithms, along with min-max feature scaling, and ADASYN algorithm were used to predict these fake accounts given publicly available features found on a user's profile. As a result, logistic regression with adaptive synthetic minority oversampling produced the best f1-score for detecting fake accounts at about 97%. In future work, more data could be collected to increase the complexity of the models. The small amount of data available in the dataset was lackluster for training. In addition, with the inclusion of more data, outliers could be erased to give better representation to the general public. Also, because of these outliers, min-max normalization could be swapped out with z-score normalization which has been seen to work better with outliers.

Acknowledgment

I would like to express my acknowledgement to the professors at the University of Illinois at Springfield for their guidance during this project. Specifically, Professor Liang Kong who led the capstone course for this project, and professor Elham Buxton who I went to for questions and guidance.

References

[1] *ADASYN — Version 0.9.1*. (n.d.). Imbalanced Learn. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html

[2] Akyon, F. C., & Esat Kalfaoglu, M. (2019). InstaFake Dataset [Dataset; Python]. In *GitHub*. https://github.com/fcakyon/instafake-dataset

[3] Akyon, F. C., & Esat Kalfaoglu, M. (2019). Instagram Fake and Automated Account Detection. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. https://doi.org/10.1109/asyu48272.2019.8946437

[4] Bhargava, K. (2021, December 24). *Instagram Fake Account Detection using Machine Learning*. Medium. https://bhargavakapil24.medium.com/instagram-fake-account-detection-using-machine-learning-fd070f58b8f

[5] Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. Machine Learning Mastery. https://machinelearningmastery.com/what-is-imbalanced-classification/

[6] Chang, W. (2022, November 29). *R Graphics Cookbook, 2nd edition*. https://r-graphics.org/

[7] Chawla, N. V. (2002, June 1). *SMOTE: Synthetic Minority Over-sampling Technique | Journal of Artificial Intelligence Research*. https://www.jair.org/index.php/jair/article/view/10302

[8] Christison, C. (2022, September 16). *I Tried Instagram Automation (So You Don't Have To): An Experiment*. Social Media Marketing & Management Dashboard. https://blog.hootsuite.com/i-tried-instagram-automation-so-you-dont-have-to/

[9] Ersahin, B., Aktas, O., Kilinc, D., & Akyol, C. (2017). Twitter fake account detection. *2017 International Conference on Computer Science and Engineering (UBMK)*. https://doi.org/10.1109/ubmk.2017.8093420

[10] Fletcher, E. (2022, January 25). *Social media a gold mine for scammers in 2021*. Federal Trade Commission. https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021

[11] Hackett, S. (2022, October 21). *How To Identify Fake Instagram Followers & Put A Stop To Them*. Kicksta Blog | Tips & Tricks to Get More Real Followers on Instagram. https://blog.kicksta.co/how-to-identify-fake-instagram-followers-put-a-stop-to-them/

[12] Henderi, H. (2021, March 1). *Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer | Henderi | International Journal of Informatics and Information Systems*. https://ijiis.org/index.php/IJIIS/article/view/73

[13] *How many accounts can I follow on Instagram?* (n.d.). Instagram Help. https://help.instagram.com/408167069251249

[14] *Introducing new authenticity measures on Instagram*. (2020, August 13). Instagram Blog. https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram/

[15] Kondeti, P., Yerramreddy, L. P., Pradhan, A., & Swain, G. (2020). Fake Account Detection Using Machine Learning. *Evolutionary Computing and Mobile Sustainable Networks*, 791–802. https://doi.org/10.1007/978-981-15-5258-8_73

[16] Lantz, B. (2019). *Machine Learning with R* (3rd ed.). Packt Publishing.

[17] Luis Torgo ltorgo@dcc.fc.up.pt. (n.d.). *SMOTE function - RDocumentation*. https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE

[18] Qiang Cao, Michael Sirivianos, Xiaowei Yang, & Tiago Pregueiro. (2012). Aiding the detection of fake accounts in large scale social online services. *Networked Systems Design and Implementation*, 15.

[19] Ramalingam, D., & Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers &Amp; Electrical Engineering*, *65*, 165–177. https://doi.org/10.1016/j.compeleceng.2017.05.020

[20] *Reducing inauthentic activity on Instagram*. (2018, November 19). Instagram Blog. https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram

[21] Roy, P. K., & Chahar, S. (2020). Fake Profile Detection on Social Networking Websites: A Comprehensive Review. *IEEE Transactions on Artificial Intelligence*, *1*(3), 271–285. https://doi.org/10.1109/tai.2021.3064901

[22] Sheikhi, S. (2020). An Efficient Method for Detection of Fake Accounts on the Instagram Platform. *Revue D'Intelligence Artificielle*, *34*(4), 429–436. https://doi.org/10.18280/ria.340407

[23] Siriseriwan, W. (2019). *A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE* (1.3.1) [Software]. https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf

[24] Van Der Walt, E., & Eloff, J. (2018). Using Machine Learning to Detect Fake Identities: Bots vs Humans. *IEEE Access*, *6*, 6540–6549. https://doi.org/10.1109/access.2018.2796018