DETECTION OF FAKE ACCOUNTS IN INSTAGRAM USING MACHINE LEARNING

Ananya Dey[1], Hamsashree Reddy[2], Manjistha Dey[3] and Niharika Sinha[4]

[1]National Institute of Technology, Tiruchirappalli, India
[2]PES University, Bangalore, India
[3]RV College of Engineering, Bangalore, India
[4]Manipal Institute of Technology, Karnataka, India

ABSTRACT

With the advent of the Internet and social media, while hundreds of people have benefitted from the vast sources of information available, there has been an enormous increase in the rise of cyber-crimes, particularly targeted towards women. According to a 2019 report in the [4] Economics Times, India has witnessed a 457% rise in cybercrime in the five year span between 2011 and 2016. Most speculate that this is due to impact of social media such as Facebook, Instagram and Twitter on our daily lives. While these definitely help in creating a sound social network, creation of user accounts in these sites usually needs just an email-id. A real life person can create multiple fake IDs and hence impostors can easily be made. Unlike the real world scenario where multiple rules and regulations are imposed to identify oneself in a unique manner (for example while issuing one's passport or driver's license), in the virtual world of social media, admission does not require any such checks. In this paper, we study the different accounts of Instagram, in particular and try to assess an account as fake or real using Machine Learning techniques namely Logistic Regression and Random Forest Algorithm.

KEYWORDS

Logistic Regression, Random Forest Algorithm, median imputation, Maximum likelihood estimation, k cross validation, overfitting, out of bag data, recall, identity theft, Angler phishing.

1. INTRODUCTION

Instagram is an online photo and video sharing social networking platform that has been available on both Android and iOS since 2012. As of May 2019, there are over a billion users registered on Instagram.

In the recent years, Instagram has been found to be using third party apps, called bots. While these can definitely impersonate a user and tarnish their reputation leading to 'identity theft', there has also been greater instances of malicious ways of promoting the brand image of a company known as "influencer marketing". These days a number of businesses are using social media to heed to their customers' needs which has led to yet another malpractice called Angler phishing. All these malpractices have made it vital to implement strong fraud detection techniques and hence we propose our solution.

2. LITERATURE SURVEY

Previously a lot of work has been done on other platforms like Facebook and Twitter, but not much work has been done for Instagram. Each of these Social Medias are different in terms of features that have to be

considered, strategies used, etc. Few of the past work include [1] Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram: This particular paper concentrates on analysing likes and identifying the genuine ones to reduce the effect of fake likes on Instagram influencer market. They used a simple feed-forward neural network Multi-Layer Perceptron (MLP) which obtained a precision about 83%. [2] Identifying Fake Profiles in LinkedIn: A number of features were considered to train the dataset using neural networks, SVMs, and principal component analysis. The precision rate achieved was 84%. [3] Detection of Fake Profiles in Social Media: This is a literature review to detecting fake social media accounts classified into the approaches aimed on analysing individual accounts. So our 2 proposed method is a novel approach in terms of the platform chosen and the algorithms used like Random Forest for classification.

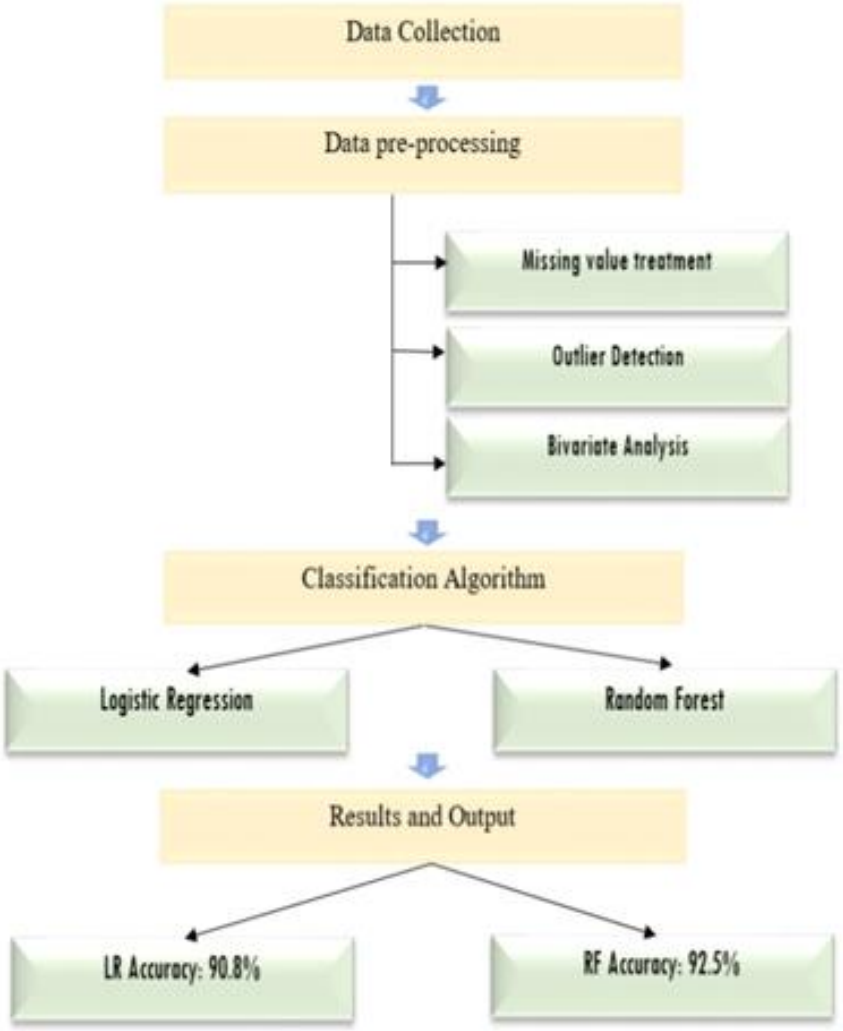

Figure1: Proposed System

3. MATERIALS AND METHODS

In this section we present the materials and methods used for the research work.

Data set information:

The dataset has been taken from https://www.kaggle.com/free4ever1/instagram-fake-spammer-genuine-accounts.It consists of two CSV files- train.csv (19 KB) and test.csv (4 KB) .The dependent variable, which is whether it is a fake or not fake account is categorical and it takes two values 0 (not fake) and 1 (fake) profile. The distribution of the training dataset is such that 50% is fake and the rest 50% is legitimate. Below is a table to denote the parameters that have been considered (denoted in column Profile feature), their range of values, each of their mean values and what each of the features denote.1. Collecting the IP addresses which will be used as the training dataset

Table 1: Dataset Features

	A	B	C	D	E	F	G	H	I	J	K	L
1	profile pic	nums/len	fullname	nums/len	name==us	descriptio	external U	private	#posts	#followers	#follows	fake
2	1	0.27	0	0	0	53	0	0	32	1000	955	0
3	1	0	2	0	0	44	0	0	286	2740	533	0
4	1	0.1	2	0	0	0	0	1	13	159	98	0
5	1	0	1	0	0	82	0	0	679	414	651	0
6	1	0	2	0	0	0	0	1	6	151	126	0
7	1	0	4	0	0	81	1	0	344	669987	150	0
8	1	0	2	0	0	50	0	0	16	122	177	0
9	1	0	2	0	0	0	0	0	33	1078	76	0
10	1	0	0	0	0	71	0	0	72	1824	2713	0
11	1	0	2	0	0	40	1	0	213	12945	813	0
12	1	0	2	0	0	54	0	0	648	9884	1173	0
13	1	0	2	0	0	54	1	0	76	1188	365	0
14	1	0	2	0	0	0	1	0	298	945	583	0
15	1	0	2	0	0	103	1	0	117	12033	248	0
16	1	0	2	0	0	98	1	0	487	1962	2701	0

Figure 2: Snapshot of Training Dataset

Exploratory Data Analysis: This is a critical process of initial data investigation done so as to discover patterns in the dataset and spot anomalies with the help of summary statistics and graphical representation. Below are the various sub processes that were done.

a. Missing Value Treatment

The given dataset had no missing value. Missing values could occur in a dataset due to mostly real world problems and can be treated either through deletion or imputation. The presence of missing values reduces the data available to be analysed, compromising on the statistical power of the study, and eventually the reliability of its results.

b. Outlier Detection

Outliers are extreme values that deviate from the usual data values in the dataset. If outliers are present in the dataset, then the accuracy is reduced significantly as the training dataset learns from this noise in the data and could give an over-fit model. After careful analysis using graphs, we conclude that the following features had outliers present in them- nums/length username, full name words, description length, #posts, #followers and #follows. To deal with these outliers, we used median imputation by calculating the median of these set of values. Note that we do not include the outliers in the median calculation. Once we get the median, we replace all the outliers with the calculated median value.

c. Bivariate Analysis

This is done to understand the relationship between two variables and the strength of association between them. We calculated the correlation matrix and concluded absence of high multicollinearity between the variables. This is one of the assumptions before building a logistic regression model.

Once data pre-processing is done, we can safely move into the algorithms. We have been provided with a labelled training dataset and can therefore proceed with applying supervised learning algorithms that map the input to the output .For the scope of this paper we have considered two commonly used classification algorithms, i.e, Logistic Regression and the Random Forest algorithm. Each of them has been explained in depth below.

1. Logistic Regression

The assumptions of this model include absence of outliers in the dataset and absence of high correlations between the predictors, which have been taken care of in the preceding steps. In logistic regression, the probabilities predicted using the logit function. The values greater than or equal to the decision boundary belong to one class while the values lower than it belong to the other.

We first run the GLM function in R to perform regression and find out the beta coefficients and p values for each of the features. The Beta coefficient, which is calculated based on maximum likelihood estimation, is an indicator of how strongly the predictor variable indicates the dependent variable. Based on the p values, we remove those variables whose 5 values are greater than 0.05 and re run the model. Finally we performed K cross validation to check overfitting

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6918  -0.1939   0.0000   0.0864   3.1230

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        5.243e+00  1.249e+00   4.199 2.69e-05 ***
profile.pic1      -5.203e+00  1.176e+00  -4.425 9.66e-06 ***
nums.length.username  8.557e+00  1.316e+00   6.502 7.91e-11 ***
fullname.words    -2.581e-01  1.974e-01  -1.308  0.19097
nums.length.fullname  4.392e+00  4.686e+00   0.937  0.34858
name..username1    1.737e+01  1.304e+03   0.013  0.98937
description.length  -5.399e-03  5.435e-03  -0.993  0.32054
external.URL1     -3.120e+01  1.918e+03  -0.016  0.98702
private1          -7.076e-01  3.674e-01  -1.926  0.05410 .
X.posts           -1.225e-02  3.953e-03  -3.098  0.00195 **
X.followers       -2.823e-03  5.693e-04  -4.958 7.12e-07 ***
X.follows          8.720e-04  2.685e-04   3.248  0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 798.51  on 575  degrees of freedom
Residual deviance: 209.86  on 564  degrees of freedom
AIC: 233.86

Number of Fisher Scoring iterations: 19
```

Figure 3: Model Summery

2. Random forest Algorithm

This is a regression model performing well on classification model. Since there are a very few assumptions attached to it, so data preparation is less challenging. We used the random Forest function in

R and set the n-tree parameter (denoting the number of trees) as 500 and the number of variables tried at each split as 3. Random Record Selection is the first task in which each tree is trained on ⅔ of the total training data and some variables are selected at random(say m) out of all the variables and these m variables are used to split the node. For each tree, using the leftover (36.8%) data, the misclassification rate is calculated. This gives us the Out Of Bag (OOB) error rate. The forest chooses the classification having the most votes over all the trees in the forest. This is the RF score and the percent YES votes received is the predicted probability.

4. RESULTS AND ANALYSIS

In this section we determine the results of the two classification models.

After creating the models using the training datasets, we apply the models on unseen data, i.e., and the test dataset. We create the confusion matrix based on these and calculate various performance parameters as discussed below.

1. Logistic Regression

	1 (Predicted= Yes)	**0 (Predicted= No)**
1 (Actual = Yes)	57 (TP)	3 (FN)
0 (Actual = No)	8 (FP)	52 (TN)

Figure 4: Confusion matrix for Logistic Regression (T = True, F=False, P=Positive, N=Negative)

We calculate the different model metrics as follows-

- Precision- We divide the total number of correctly classified positive examples by the total number of predicted positive examples.
  Precision= TP / (TP+FP) = 57/ (57+8) = 87.6 %
- Recall- The ratio of the total number of correctly classified positive examples divide to the total number of positive examples. Recall= TP/((TP+FN) = 57/ (57+3) = 95 %
- F1 score- It is the harmonic mean of recall and precision. F1 score= $(2*Precision*Recall)/(Precision+Recall)$ = 91.15
- Accuracy- It is calculated using the following formula
  (TP+TN) /(TP+TN+FP+FN) = (57+52) / (57+52+8+3) = 90.8 %

Based on this curve, we infer that using k=3 will give optimal results.

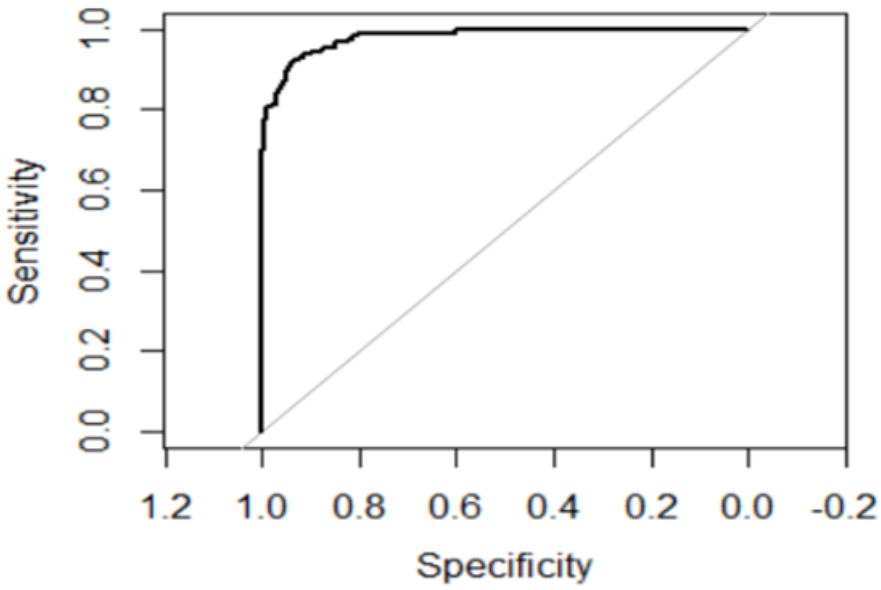

Figure 5: ROC Curve for Logistic Regression

2. Random Forest Algorithm

	1 (Predicted= Yes)	**0 (Predicted= No)**
1 (Actual = Yes)	55 (TP)	5 (FN)
0 (Actual = No)	4 (FP)	56 (TN)

Figure 6: Confusion matrix for Logistic Regression (T = True, F=False, P=Positive, N=Negative)

On a similar note, we calculate the various model metrics for Random forest algorithm.

- Precision= TP / (TP+FP) = 55 / (55+4) = 93.2 %
- Recall= TP / (TP+FN) = 55/(55+5) = 91.6 %
- F1 score= (2∗$Precision$∗$Recall$) / ($Precision$+$Recall$) = 92.42
- Accuracy= (TP+TN) / (TP+TN+FP+FN) = (55+56) / (55+56+4+5) = 92.5 %

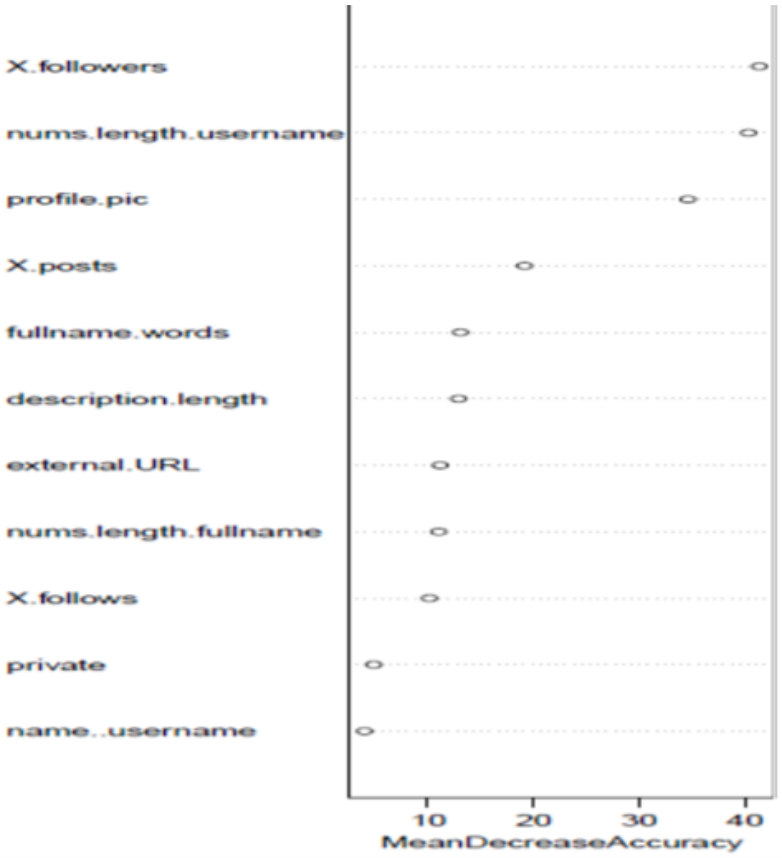

Figure 7: Variable Importance Graph

5. CONCLUSIONS

While going through previous similar research conducted on detection of fake profiles on social media platforms, we realized that not a lot has been done on Instagram as a social network platform in particular. Hence, we targeted our approach for the same. In this paper, we introduced a novel approach for detecting fake user profiles on Instagram based on certain features using concepts of machine learning. We used two models for this- Logistic Regression and Random Forest algorithms, achieving an accuracy of 90.8% and 92.5% respectively. Such high accuracies have not been attained in previous work conducted for other social media platforms (highest accuracy achieved before this was 86%).

REFERENCES

[1] Indira Sen,Anupama Aggarwal,Shiven Mian.2018."Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram". In ACM International Conference on Information and Knowledge Management.

[2] Shalinda Adikari, Kaushik Dutta. 2014. “Identifying Fake Profiles In LinkedIn”. In Pacific Asia Conference on Information Systems.

[3] Aleksei Romanov, Alexander Semenov, Oleksiy Mazhelis and Jari Veijalainen.2017. "Detection of Fake Profiles in Social Media”. In 13th International Conference on Web Information Systems and Technologies.

[4] https://telecom.economictimes.indiatimes.com/news/india-saw-457-rise-in-cybercrime-in-fiveyears-study/67455224

[5] Todor Mihaylov, Preslav Nakov.2016. "Hunting for Troll Comments in News Community Forums". In Association for Computational Linguistics.

[6] Ml-cheatsheet.readthedocs.io. (2019). Logistic Regression — ML Cheatsheet documentation. [Online] Available at: https://ml cheatsheet.readthedocs.io/en/latest/logistic_regression.html#binarylogistic-regression [Accessed 10 Jun. 2019].

[7] 3. Schoonjans, F. (2019). ROC curve analysis with MedCalc. [Online] MedCalc. Available at: https://www.medcalc.org/manual/roc-curves.php [Accessed 10 Jun. 2019].

[8] Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre,B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. Bus.Horiz., SPECIAL ISSUE: SOCIAL MEDIA 54, 241251. doi:10.1016/j.bushor.2011.01.005.

[9] Krombholz, K., Hobel, H., Huber, M., Weippl, E., 2015.Advanced Social Engineering Attacks. J Inf SecurAppl 22, 113–122. doi:10.1016/j.jisa.2014.09.005.