IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FAKE ACCOUNT DETECTION USING MACHINE LEARNING

T. Om Prathyusha, N. Sai Kumar, E. Vishnu Priya
Department of Computer Science and Engineering
St. Peter's Engineering College
Hyderabad, India

T. Vijaykanth Reddy (Assistant Professor)
Department of Computer Science and Engineering
St. Peter's Engineering College
Hyderabad, India

***Abstract*— In today's world, Online Social Media is king in a number of forms the number of people who use the service is growing every day. The use of social media is skyrocketing. The primary benefit is that we can easily communicate with people via online social media and communicate with them in a more effective manner. This opened up a new avenue of a possible attack, such as a forged identity, false information and so on. According to a recent study, the number of accounts in the number of people who use social media is much higher than the number of people who use it. These fake accounts are difficult to detect for online social media providers. Since social media is flooded with false information, ads, and other types of content, it is essential to recognise these fake accounts.**

From an online social media dataset, we offer a method for detecting fraudulent accounts. We employed boosting methods to improve the accuracy of the standard technique, rather than employing typical machine learning classifiers.

By boosting weak learners, this method has resulted in a large improvement in accuracy. In this paper we will use accuracy comparison of Xgboost Classifier, and Gradient boosting Classifier. Xgboost performed brilliantly when compared with the previous work.

Keywords*— *machine learning, fake account detection, gradient boosting, extreme gradient boosting, Accuracy.

I. INTRODUCTION

Nowadays, Online Social Media is dominating the world in several ways. Day by day the number of users using social media is increasing drastically. The main advantage of online social media is that we can connect to people easily and communicate with them in a better way. This provided a new way of a potential attack, such as fake identity, false information, etc.A recent survey suggest that the number of accounts present in the social media is much greater than the users using it.This suggest that fake accounts have been increased in the recent years. Online social media providers face difficulty in identifying these fake accounts. The need for identifying these fake accounts is that social media is flooded with false information, advertisements, etc.

EXISTING SYSTEM:

The existing systems use very fewer factors to decide whether an account is Fake or not. The factors largely affect the way decision making occurs. When the number of factors is low, the accuracy of the decision making is reduced significantly. There is an exceptional improvement in fake account creation, which is unmatched by the software or application used to detect the fake account. Due to the advancement in creation of fake account, existing methods have turned obsolete. The most common algorithm used by fake account detection Applications is the Random forest algorithm. The algorithm has few downsides such as inefficiency to handle the categorical variables which has different number of levels. Also, when there is an increase in the number of trees, the algorithm's time efficiency takes a hit.

PROPOSED SYSTEM:

The gradient boosting algorithm is similar to the random forest algorithm in that it relies heavily on decision trees. We have modified the way we find fake accounts, using new approaches to locate them. Spam commenting, interaction rate, and artificial behaviour are some of the techniques used. The gradient boosting algorithm uses these inputs to build decision trees, which are then used in the gradient boosting algorithm. Even if some inputs are missing, this algorithm produces a result. This is the primary reason for using this algorithm. We were able to obtain extremely accurate results by using this algorithm. XGBoost and GBM performed brilliantly when compared with the previous work. It outperforms the accuracy of fake account identification by a large margin even with the default values of the hyperparameters provided in. Finally, we achieved a better result than earlier attempts.

OBJECTIVES:

The main objective of this paper is to detect fake accounts. Gradient boosting algorithm is used in this project to detect fake accounts accurately.

Identifying these fake accounts is that social media is flooded with false information, advertisements.

II. IMPLEMENTATION

As proposed earlier, this detection method uses gradient boost and extreme gradient boosting algorithms to detect fake accounts. The technologies we have used are Python and Python's Standard libraries like Numpy, Pandas, Matplotlib, Scipy and Sklearn.

A. Python

Python is our major programming language. It is used for our dataset to detect fake accpounts. It provides various tools and libraries that help in detecting fake accounts upto high accuracy levels.

B. Module description:

TensorFlow

TensorFlow may be a free and ASCII text file software package library for dataflow and differentiable programming across a spread of tasks. It's a symbolic scientific discipline library, and is additionally used for machine learning applications like neural networks. It's used for each analysis and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use. It had been discharged belowthe Apache two.0 ASCII text file license on November nine, 2015.

Pandas

Pandas is a associate degree ASCII text file Python Library providing superior knowledge manipulation andanalysis tool victimization its powerful knowledge structures. Python was majorly used for knowledge munging and preparation. It had little contribution towards knowledge analysis. Pandas resolved this drawback. victimization Pandas, we will accomplish 5 typical steps within the process and analysis of information, in spite of the origin of information load, prepare, manipulate, model, and analyze. Python with Pandas is employed in a very wide selection of fields as well as educational and industrial domains as well as finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a collection of functions that make **matplotlib** work like MATLAB . Each **pyplot** function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels.

Scikit – learn

Scikit – learn Scikit-learn provides a spread of supervised and unattended learning algorithms via an identical interface in Python. It's accredited below a permissive simplified BSD license and is distributed below several UNIX system distributions, encouraging educational and industrial use. The library is madeupon the SciPy (Scientific Python) that has to be put in before you'll use scikit-learn. This stack that includes:

- **NumPy**: Base n-dimensional array package
- **SciPy**: Fundamental library for scientific computing
- **Matplotlib**: Comprehensive 2D/3D plotting
- **I Python**: Enhanced interactive console
- **Sympy**: Symbolic mathematics
- **Pandas**: Data structures and analysis
- Extensions or modules for SciPy care conventionally named SciKits. As such, the module

C. Algorithm's:

GRADIENT BOOSTING MACHINE (GBM):

To produce final predictions, a Gradient Boosting Machine (GBM) combines predictions from multiple decision trees. Keep in mind that in a gradient boosting machine, all of the poor learners are decision trees.But how is using a hundred decision trees better than using a single decision tree if we're using the same algorithm? What are the different ways that different decision trees capture different signals/information from data?

The trick is that each node in the decision tree uses a different subset of features to choose the best split. This means that the individual trees aren't all the same, and they can absorb different signals from the data as a result.

EXTREME GRADIENT BOOSTING MACHINE (XGBM):

Another widely used boosting algorithm is XGBoost (Extreme Gradient Boosting). In reality, XGBoost is just a tweaked GBM algorithm! XGBoost follows the same steps as GBM in terms of operation. XGBoost builds trees in a sequential fashion, attempting to fix previous trees errors.

However, there are a few features that render XGBoost slightly superior to GBM:One of the most significant differences between XGBM and GBM is that XGBM uses parallel pre-processing (at the node level), making it faster.

Regularization strategies in XGBoost help to reduce overfitting and improve overall efficiency. Setting the XGBoost algorithm's hyperparameters allows you to choose the regularisation technique.

ALGORITHM:

GBM, XGBoost and AdaBoost classifiers were trained and validated with training and validation sets after feature selection and then accuracy was tested on the training set.

INPUT:

TrainData = The labeled training set (70%);

ValidationData = The validation dataset (10%)

TestData = Unlabeled dataset (20%)

OUTPUT:

Predictions = prediction from classifiers used.;

//ValidationData is used to validate the classifier predictions

1. Load TrainData

2. for all instances in TrainData

3. for each feature matrix fed to the CLASSIFIER [LR, RANDOM FOREST XGB, ADB, GBM]

4. train classifier

5. Accuracy,precision = PREDICTION.metrics

6. RESULT COMPARISON

III. METHODOLOGY

A.Uploading the data:

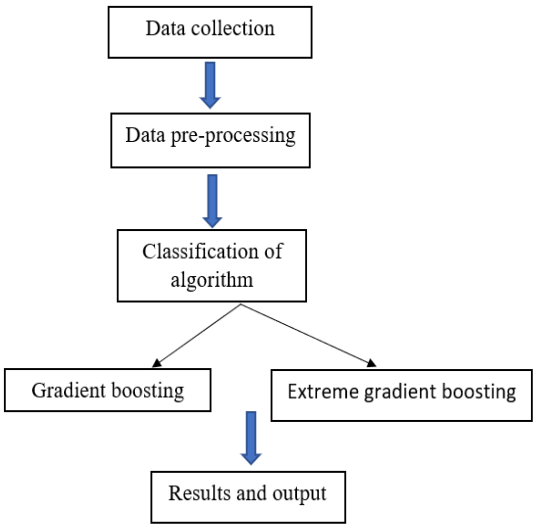
A collection of instances is a dataset and when working with machine learning methods we typically need a few datasets for different purposes.

- Training Dataset: A dataset that we feed into our machine learning algorithm to train our model.
- Testing Dataset: A dataset that we use to validate the accuracy of our model but is not used to train the model. It may be called the validation dataset.

B. Dataset pre-processing:

It is an important step to detect fake account. In this step data is processed in an appropriate form which can be inputted for detection process. the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model.

3. Detecting fraudulent accounts and comparing the results.

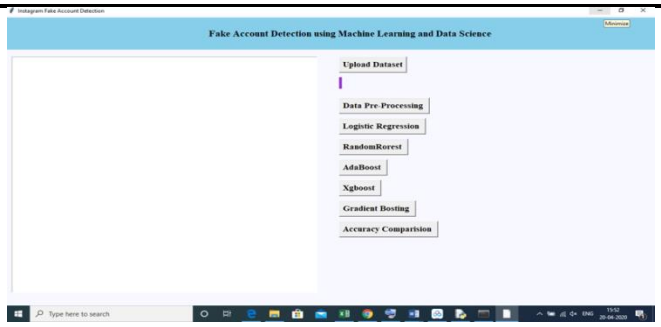


C. EXPERIMENT AND RESULT:

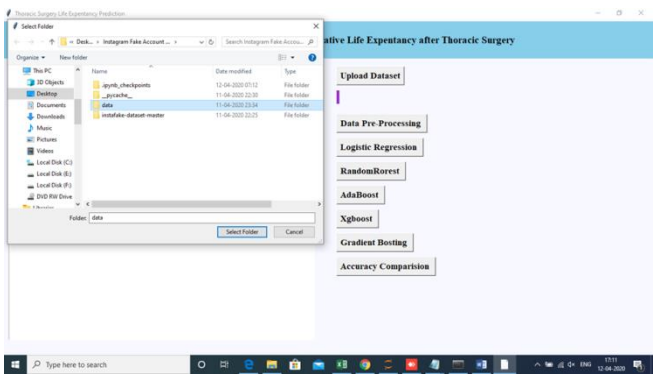
Table 1: Boosting Classifier Performance

Classifiers	Xgboost	GBM
Accuracy	0.958	0.952
Precision	0.951	0.939

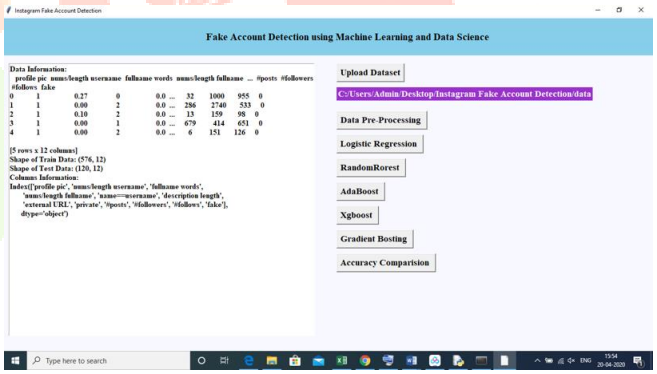
Boosting classifiers outperformed typical machine learning classifiers by a significant margin. The default parameter values for these boosting classifiers were used. XGBoost obtained the value of 95 percent, which is slightly higher than other algorithms.



After above screen will be opened and Need select the dataset directory by clicking on upload button.

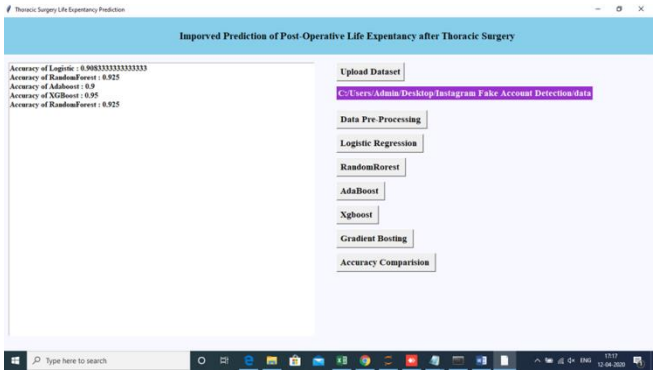


Data is uploaded.
Now click on pre-process button.

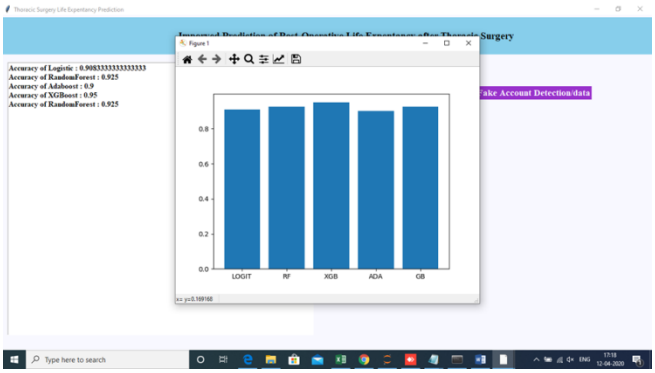


Data Pre-processing will be done and will show the data information.

Now click on Logistic Regression, Xgboost, Adaboost, Gradient boosting buttons.



GBM AND XGBOOST ACCURACY:



IV. CONCLUSION

We said before that there was a lack of a gold standard public dataset for analysis, thus we had to use active learning. The use of Extreme Gradient Boosting to detect fraudulent accounts is still relatively new and on the rise. There are numerous branches to investigate. As previously stated, we did not perform deep hyperparameter tuning in our suggested strategy or trials. Tuning hyperparameters is both expensive and time-consuming. Finding the optimal collection of parameters might be difficult. XGBoost, on the other hand, fared better with default values, achieving accuracy of up to 95%.

V.FUTURE SCOPE:

Finally, we intend to enrich the dataset further and look forward to observing the results of other elements of the boosting methods

V. REFERENCES

1.“Detection of Fake Twitter accounts with Machine Learning Algorithms” Ilhan Aydin, Mehmet sevi, Mehmet Umut salur January 2019.

2.“Detecting Fake accounts on Social Media” Sarah Khaled, Neamat el Tazi, Hoda M.O. Mokhtar January 2019.

3.“Detection of fake profile in online social networks using Machine Learning” Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Choudhury August 2018.

4.“Twitter fake account detection”, Buket Ersahin, Ozlem Aktas, Deniz kilinc, Ceyhun Akyol November 2017.