International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Enron Corpus Fraud Detection

Lucky Mohanty, Kirtika Thakur, G. Manju

Abstract: ***The main motive behind this work is to identify the person of interest based on the email data from the Enron corpus which is made public for research. Fraud detection is done using artificial neural network (ANN) with Adam optimizer and ReLU activation functions which is a machine learning approach. With advancements in the field of Artificial Intelligence the fraud detection can done effectively in python environment. This work achieves greater accuracy in terms of recall, precision and F1 score. The work can prove useful to various firms that maintain accounting data of the financial transactions that take place in the given organization. The goal is to devise a method that can be implemented on accounting data of an organization, company or firm to identify the individuals susceptible of committing fraudulent activities by manipulating the financial statements to mislead the investors and shareholders. This ultimately aims to reduce the losses suffered by the investors and shareholders by detection of various fraudulent entities in the given organization.***

Index Terms: Enron corpus, Artificial Neural network, Adam optimizer, ReLU activation function, Fraud Detection

I. INTRODUCTION

Enron was one of the largest companies in the United States and in 2000 it had collapsed because of the corporate fraud in the business units [1]. During Federal investigation IN 2002, many confidential information was made public record that includes emails and detailed financial data. It contains all personal and official emails. Some of the emails have been deleted and the remaining were put up in the data-set for researchers (*http://www-2.cs.cmu.edu/~enron/*). This new dataset contains 517,431 emails with 3500 folders from 151 users. The folder contained information of all 151 employees. The message in the folders contains the receiver and the senders email address, date and time, body, subject, text and email specific technical details. Also, it contains various information about the salary and also the stock records. The Enron data set is used as valuable training and testing ground for machine learning researchers who try to develop models which can identify the persons of interests (POIs) from the features present within the Enron email data set [1].

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Lucky Mohanty*, Student, B.Tech in Computer Science and Engineering, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology. (Tamil Nadu) India.
Kirtika Thakur, Student, B.Tech in Computer Science and Engineering, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology. (Tamil Nadu) India.
Dr. G. Manju, Associate Professor, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technologyd. (Tamil Nadu) India.

The persons of interest are the one who were suspected to be fraud in Enron investigation which even includes higher level executives. Thus, the objective of this work is to create a machine learning model that could identify out the POIs.

II. NEURAL NETWORKS OPTIMIZATION

In the realm of AI, artificial neural systems models are utilized which is enlivened by natural neural systems. They are used to approximate functions that can depend on many inputs which are generally unknown.
ANN is a system of interconnected neurons which is capable of exchanging messages among themselves. The connections have weights which is tuned making neural nets adaptive to inputs and capable of learning. An ANN [2] is modeled by three kinds of parameters: They are Architecture (The interconnection design between the layers of neurons), The learning procedure for updating the weights of the interconnections and the activation function which converts a neuron's weighted input to its output activation. Nowadays Optimization is the most essential part of deep machine learning algorithms. It starts with defining a kind of loss function/cost function and ends with minimizing it.

A. Artificial Neural Network with ADAM Optimizer

Adam [3] is one of the popular gradient descent algorithms in the field of deep learning as it accomplishes great outcomes quick.. It keeps up a consistent learning rate for all weight refreshes amid preparing. The learning rate is kept up for every parameter and is independently refreshed as learning unfurls. Adam analyzer computes individual adaptive learning rates for various parameters from assessments of first and second moments of the gradients. In summary, each update of Adam involves the following steps:

1. Computing the gradient and its element-wise square using the current parameters.
2. Updating the exponential moving average of the 1st-order moment and the 2nd-order moment.
3. Computing an unbiased average of the 1st-order moment and 2nd-order moment.
4. Computing weight update: 1st-order moment unbiased average divided by the square root of 2nd-order moment unbiased average (and scale by learning rate).
5. Apply the update to the weights.

Adam utilizes the benefits of both Adaptive Gradient Algorithm and Root Mean Square Propagation.

A. Artificial Neural Network with ADAM Optimizer and ReLU Activation functions.

The main purpose of the activation functions in neural network is to convert an input signal of a node to an output signal.

That output signal now is used as an input in the next layer in the stack. Rectified Linear units [3] has become very popular in the past couple of years as it showed 6 times more improvement in convergence from the hyperbolic tangent function. Hence it avoids and rectifies vanishing gradient problem.

III. SYSTEM MODEL FOR ENRON FRAUD DETECTION

A. Importing Libraries

To perform this machine learning under the python environment certain libraries must be loaded. The required libraries are Scikit-learn, pandas and matplotlib.

B. Uploading the base data set

The Enron corpus data set [1] contains the features of three categories. They are Salary Features, Stock Features and Email Features.

- **Email Features**: "to messages", "email address", "from poi to this person", "from messages", "from this person to poi", "shared receipt with poi".
- **Stock Features**: "Exercised Stock Options", "Restricted Stock", "Restricted Stock Deferred", "Total Stock Value".
- **Finance Features**: "Salary", "Loan Advances", "Expenses", "Deferral Payments", "Total Payments", "Bonus", "other", "Long Term Incentive", "Director", "Fees", "Deferred income".

A. Converting the python dictionary into data frame

The data set is loaded in the form of a Python dictionary with key and the information as values, so it must be converted it to a pandas data frame for further processing the data.

B. Segregating features by its categories

Since the data set contain features based on e-mail, stock and finace. So, the features are categorized based on the domains.

C. Data Pre-Processing

By the observing the data set we could infer there were NaN values and also certain features which are not so significant are removed in the preprocessing stage. The main data preprocessing done are as follows.

1. Replacing NaN values in the entire data set into nan
2. Replacing nan in the financial feature by 0
3. Replacing nan in the email feature by its median
4. Removing Total from the Enron corpus dataset
5. Adding new features to the dataset
   The additional features added are
   1.ratio_messages_to_poi = from_this_person_to_poi/from_messages
   2.ratio_messages_from_poi = from_poi_to_this_person / to_messages

D. Visualization

The distribution certain features like salary deferred payments and total payments are visualized using the scatter plot which in turn helps in feature selection as shown in the Fig. 1

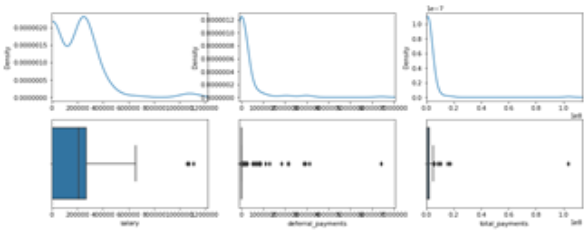

Fig. 1. Visualization of the features in the data set

E. Feature Selection

The significant feature that contains more vital information are selected to be given as input for machine learning process. The selected features are salary', 'total_payments', 'bonus', 'deferred_income', 'total_stock_value', 'exercised_stock_options', 'long_term_incentive', 'shared_receipt_with_poi', 'restricted_stock' and 'ratio_messages_to_poi'.

F. Normalization and One-hot Encoding

The selected features were normalized and the Enron corpus data set for training and testing the neural network was spitted in the ratio of 80-20. Also, the initialization of Parameter process taken place. One Hot Encoding is the popular encoding methodology which is going to be applied on the available data set to make it ready to fed into the neural network learning and evaluation process.

G. Machine Learning

The core model of the machine learning approach was done using ANN multilayer perceptron with adam optimizer and ReLU activation function. The model and its parameters of the ANN can be explained using the Fig. 2 and Fig. 4.

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 128)	2816
activation_1 (Activation)	(None, 128)	0
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 128)	16512
activation_2 (Activation)	(None, 128)	0
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 2)	258
activation_3 (Activation)	(None, 2)	0

Total params: 19,586
Trainable params: 19,586
Non-trainable params: 0

Fig 2. ANN Architecture

From the above figure, it is inferred that there were totally 19586 parameters that can be trainable.

H. Training and Validation

From the Fig.3, the ANN used a 30 epoch with 128 hidden layers for training with validation of 0.1 and dropout of 0.2.

```
epoch = 30
batch = 128
classes = 2
optimizer = RMSprop()
hidden = 128
validation = 0.1
dropout = 0.2
```

Fig. 3 ANN Specifications

During training and validation, the loss and accuracy at every epoch were plotted as shown in the Fig.4.
From the Fig.4, it is observed that training accuracy and validation accuracy is around 0.85 and 0.8 which is constant throughout the epochs. But it is observed that training accuracy is spiking at some epochs and constant at few from the range of 1.8 to 2.20 whereas validation loss is around 2.7 which is constant throughout the epochs.

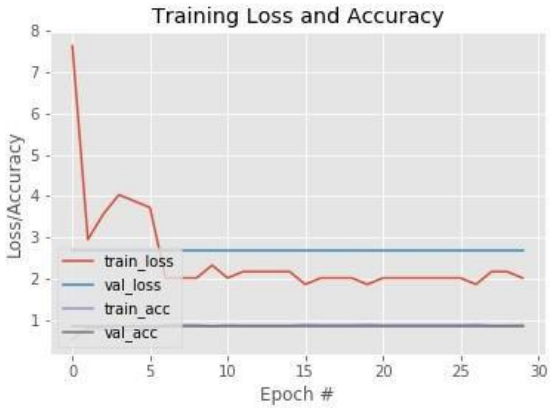

Fig. 4. Loss/Accuracy during training and validation

IV. RESULTS AND DISCUSSIONS

The trained ANN is validated and can be tested for classifying the fraud in the Enron corpus data. The efficiency of the modeled ANN can be evaluated by observing the confusion matrix which is used to observe the performance of the classifier using the testing data in all possibilities of classification. Fig.5 shows the confusion matrix of the models ANN.

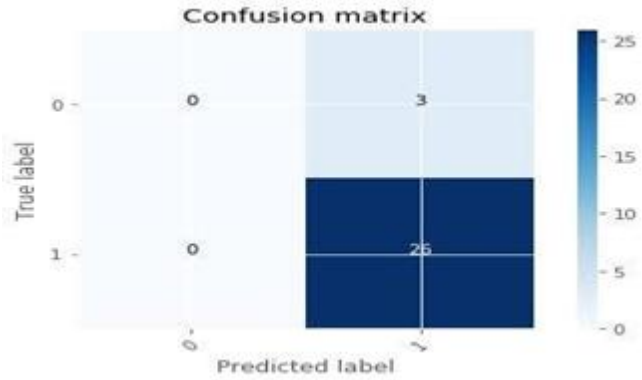

Fig. 5 Confusion Matrix

It is observed that the modeled ANN classifies the fraud correctly with the true label 26 times and with wrong label only 3 times. This shows that the modeled ANN work well in most of the scenarios of input during testing.

Also, precision, recall and f1 scores were calculated to examine the accuracy of the modeled ANN. Fig.6 shows the precision recall and f1 scores during testing.

	precision	recall	f1-score	support
0	1.00	0.90	0.95	29
1	0.00	0.00	0.00	0
avg / total	1.00	0.90	0.95	29

Fig. 6 Evaluating the classifier performance

Precision is the positive predictive value while recall is the sensitivity of the classifier. F1 score is the harmonic mean of precision and recall. From the Fig.6, it is observed that precision is 1.0 which is the maximum achievable one and recall is 0.90. Also, F1 score is 0.95.

V. CONCLUSION

It can be concluded that the modeled ANN can classify an employee as Person of interest -POI correctly with a precision accuracy of 100 % and recall of 90% shows that it can classify the POI Correctly with an accuracy of 90%. Also, it classifies the POI correctly with a F1 score accuracy of about 95%. Thus, the Enron Fraud detection can be done with greater accuracy using the modeled ANN.

REFERENCES

1. Bryan Klimt and Yiming Yang, “The Enron Corpus: A New Dataset for Email Classification Research,” Language Technologies Institute Carnegie Mellon University Pittsburgh, PA 15213-8213, USA.
2. Justin DeCunha, “Fraud Detection - A Machine Learning Approach” in April,2018
3. Diederik P. Kingma & Jimmy Lei Ba ArXiv, “The Berkeley VIEW,” in 2015.
4. Jason Browniee, “Gentle Introduction to the Adam Optimization Algorithm for Deep Learning,” 2017.
5. Allen Reyes, “Identifying fraud from enron emails and financial datas,” 2015.
6. Jitesh Shetty & Jafar Adibi,“The Enron Email Dataset Database Schema and Brief. Statistical Report, "University of Southern California.
7. Bryan Klimt & Yiming Yang, "The Enron Corpus: A New Dataset for Email Classification Research,” Machine Learning: ECML 2004 pp 217-226.
8. Kasthurirangan Gopalakrishnan, Siddhartha K.Khaitan Alok Choudhary & Ankit Agrawal, "Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection,” Construction and Building Materials Volume 157, 30 December 2017, Pages 322-330
9. Forest Agostinelli, Matthew Hoffman, Peter Sadowski, Pierre Baldi,“Learning Activation Functions to Improve Deep Neural Networks, "International Conference on Learning Representations (ICLR) 2015.
10. C. Zhang, Philip C. Woodland,“Parameterized Sigmoid and ReLU Hidden Activation Functions for DNN Acoustic Modelling,”16th Annual Conference of the International Speech Communication,2015.
11. Goutte C., Gaussier E, “A Probabilistic Interpretation of Precision, Recall and *F*-Score, with Implication for Evaluation “, In: Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval. ECIR 2005.

AUTHORS PROFILE

Lucky Mohanty Student, B.Tech in Computer Science and Engineering, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology.
E-mail:luckymohanty_su@srmuniv.edu.in

Kirtika Thakur Student, B.Tech in Computer Science and Engineering, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology.
E-mail: kittu070996@gmail.com

Dr. G. Manju Associate Professor, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology.
E-mail: manju.g@ktr.srmuniv.ac.in