Prevention of Unwanted Calls Over Telephony Network

Gaurav Malviya, Vatsal Singh

Abstract. *The main motive is to increase the malicious call prevention while not relying on any underlying architecture of the phone . Challenges faced here are -how to gather useful information on how to decrease benign calls from being blocked . The first part of the work is* ***to gather data concerning*** *nuisance callers* ***so as to create an efficient interference*** *mechanism, with the help of Machine Learning Algorithms. Gathering of the dataset is based on client statistics. In that, there are three varieties of lessons like telemarketing, Unwanted (Malicious), robocalls.If a call is Malicious or not is judged based on few functions .So compare the effectiveness of these capabilities with the aid of numerous cutting-edge models. For this, usage of both Area under curve (AUC) score, and layout a brand new metric, average first prediction (AFP) is done. Average first prediction is designed to assess the averaged amount of nuisance calls that desires to be located before a user can expect it as a nuisance caller, without affecting benign traffic. Evaluation shows that segregating the planned options, a random forest version can do Associate in Nursing AUC score of at 0.98 ; additionally , it reduces the averaged important determined nuisance calls via up to 88% from a black-listing method, whilst making sure that over 98% of the benign calls will now no longer be banned from connecting . The system would be fast, effective, efficient, UI would be lightweight ,and this would work without accessing the telephony network architecture.*

Keywords: Robocalls, telephony network, unwanted calls, network calls, voice-over-IP

I.INTRODUCTION

The study overcomes all the issues mentioned in the Abstract furthermore, reveals new insight into the elemental plan, which is the center issue of this work. Results uncover that the common pernicious calls affects a country's GDP more than benign calls; Nuisance calls are significantly more liable to show in a workday and at working hours than gainful calls; what's more, the degree of approaching and activity from the two is demonstrative enough to separate malignant calls from benevolent calls.Malevolent calls can be done by voice over IP(VOIP) which can be untraceable because of the available IP addresses and Proxy .

Inspired by the activity study, the five features selected for the nuisance call prevention problem should include dynamic information about the current call, and also using information by examining previous records related to the decision of associate incoming call by referencing multiple records present in the database .

Manuscript received on April 02, 2020.
Revised Manuscript received on April 20, 2020.
Manuscript published on May 30, 2020.

* Correspondence Author
Gaurav Malviya*, UG Student Computer Science and Engineering SRM Institute of Science and Technology, Chennai, India
Vatsal Singh, UG Student Computer Science and Engineering SRM Institute of Science and Technology, Chennai, India

Now we have to decide which model to use out of the numerous dynamic models available to us .For this , use of both the primary AUC rating, and design average first prediction . Average First Prediction is designed to calculate the average number of nuisance calls that needs to be discovered earlier than an technique can predict it as a nuisance caller, without eating bandwidth of benevolent call traffics. The analysis indicates the mistreatment of other projected alternatives, a random forest version reduces the mean vital determined malicious calls by up to 88% from a blocking method . In distinct phrases, the random forest variant utilizing our 5 proposed highlights can lessen 80% of the unblocked pernicious calls. Also, the analysis suggests that a neural network version can achieve the same accuracy overall performance due to the fact the exceptional random forest, however it takes one additional millisecond process time than random forest .This shows the models in our assessment can be effectively actualized over the present foundation to accomplish both high precision and high productivity.

II.LITERATURE SURVEY

Algorithmic detection of malevolent code supported the normalized supervisor call instruction[1]. The paper comes up with a replacement linguistics detection of malevolent code methodology supporting the renormalized supervisor instruction .This achieves excellent malicious code system sequence and connected parameters through the control of virtual setting. So as to successfully affirm the vindictive code, it tends to set up a very affordable unique conduct vector's data of malevolent code. By an outsized scope of malignant codes test confirmation, the strategy is contrasted and existing methodologies which will be a great deal of right portrayal of the malevolent code assaults upheld framework choice, and successfully in particular obscure vindictive code..

"Detection of malicious package on supported multiple equations of API-calls sequences" [2].Improvement and dispersal of malicious bundle needs the production of new ways for their identification.This can be done by observing the symptoms often occuring in malwares . Because of the very reality that examined the program being checked is most likely unsafe, its execution must be in an isolated domain. This paper talks about proactive ways upheld Programming interface choice examination and propose a substitution procedure utilizing a numerous grouping arrangement to spot basic characteristics in malware. The main objective is about the subject to watch malevolent bundle, in light of Programming interface calls, every one of which is executed in programming. Additionally, another malware discovery topic upheld various succession Programming interface calls arrangement.

The plan is depicted in detail and actualized in programming. A look at bundle and the authenticity of the irresistible operator nature. Testing has demonstrated it can identify malicious package with high precision “Sequencing malicious system calls”[3]. Jointly recognizes the use of productive methods to detect malwares.

One of the key challenges that is faced using this method is the large overhead of system-call . The main contributions are: (i) Proposal of a novel way to deal with malware framework call arrangement portrayal . This leads to a more efficient detection of individual malware.. This results in a significant decrease in the number of malware instances and the program failing to “dummy insertion attacks”. (ii) Continuing on (i), propose a novel managed learning based structure for discovery of vindictive framework call successions in already concealed programming programs. This structure can be used to identify belovelent system programs

Bray-Curtis Weighted Automaton which is used for Distinguishing Vindictive Code Through Framework Call Examination[4]. Vindictive code discovery is one of the most noteworthy subjects of enthusiasm for interruption identification frameworks in the present PC security examination regions. The work proposes another heuristic technique for distinguishing pernicious code through framework call coordinating, that furthermore takes in thought the hour of the choice.To accomplish this pattern matching is done .It also discuss how this procedure can be applied to enhance the arrangement of existing principles from the information base for improving the identification rate. “Android Malicious Application Detection Based on Ontology Technology Integrated with Permissions and System Calls” [5]. The work is focused on sharing the security information on advanced cell applications and distinguishing the pernicious applications . The new strategy used here is an advanced dependent on cosmology innovation which considered consents and framework calls data with Jess engine(a JAVA platform). So as to instigate last component data rundown and define SWRL(Semantic web rule language) ,it was necessary to extract system call information. The made application remarked application domain data together with permissions and system calls etc so unequivocal and implicit information could be shared. By choosing characterized SWRL rules and running JESS engine ,this could classify efficiently malicious and benign. Experimental results showed that the accuracy reached 95% . In addition, through a similar examination, it could be seen that the application security identification dependent on ontology strategy beat two existing Android malware discovery schemes .

III.SYSTEM OVERVIEW

A. Data Collection/Pre-processing

The dataset collects to the customer call, robot calls and unknown. The dataset gives to the machine, the processing of randomly selected data to be trained and testing . The randomly splitted into those values of size in the customer dataset. The investigation prescribes for producers to settle on proactive choices in distinguishing which components are the most critical to increment of calls. There are 999 rows in the dataset, with 13 columns, 1 being the class. In the class column, 2 variables are used, 1 indicating the presence of malicious data and 0 being the absence of it. 12 other features that have been used are: Ticket ID, Time of Issue, Form, Method, Issue, Caller ID Number, Type of Call/Message, Advertiser Business Number, State, Zip and Location.

B. Pre-processing

Pre-process is sort of a usual C pre-processor, but it can be used for multiple languages. Languages for which it is efficient consist of: C++, Python, Tcl, XML, JavaScript, CSS, IDL, PHP, Java, Shell scripts etc.. Here pre-processing is used to find null data if it exist and check for any duplicate entries .The duplicate entries can be found by making a set of data entries .If this set has N elements and N=number of entries then it has no duplicate entries because sets prohibits duplicate entries .

C. Feature Extraction

Highlights are chosen to apply characterization calculations. Properties are chosen as highlights in the event that they are not subject to different characteristics and they increment productivity of the grouping.Here 5 features are selected out of 12 which was adequate to identify malicious calls while not having a big overhead . After determination of qualities, the dataset that are now delegated positive or negative are required for the training purpose. So the extraction of significant highlights in the dataset.

D. Classification

Characterization is the way toward classifying an information object into classifications called classes dependent on highlights/characteristics related with that information object.Arrangement utilizes a classifier, a calculation that forms the qualities of every datum (singular form of date) article and yields a class dependent on this data. In this project, we use Logistic regression, naive bayes , Random Forest and Decision Tree as a classifier. Random Forest is a rich and efficient strategy for working on a huge informational index.

IV.ALGORITHMS USED

A. **Logistic regression**

This binary provision model is employed to estimate the likelihood of a binary response supported one or additional predictor (or independent) variables (features).

It permits one to mention that the presence of a risk issue will increase the likelihood of a given outcome by a particular share.

Like all regression analyses, the provision regression could be a prognostication analysis.

Logistic regression is hired to provide an explanation for data and to make clear the link among one established binary variable and one or extra nominal, ordinal, c programming language or ratio-degree freelance variables.Fig 2 shows the comparison between Logistic regression to Random forest

B. **Naive Bayes**

A Naive approach assumes that the presence of a specific function in a very class is independent from the existence of the other function. Bayes’ Theorem is stated as:

P(A|B) = (P(A|B) * P(A)) / P(B)

Where

P(A|B) - the probability of speculation A and B given that B already occurs. This is known as the posterior chance.

P(A) - the probability of hypothesis A being genuine . And is called the earlier chance of h.

P(E) - the probability of the event E occurring.

C. Decision Tree

When schooling a dataset to categorise a variable, the motive of using the Decision Tree is to divide the statistics into folds primarily determined by a positive function value till the target variables are all less than a fixed class . While the human brain decides to pick the "splitting feature" based on the experience

a pc splits the dataset supported the most data gain.

D. Random Forest Classifier

Random forest, consists of a large number of individual decision trees that operate as a group or collection. Every individual tree in the Random Forest lets out a class expectation and the class with the most votes turns into our model's forecast Random backwoods calculation can utilize both for arrangement and relapse sort of issues. A lot of trees within the forest the a lot of sturdy forest appears like. In the same manner within the random forest classifier, the upper the amount of trees within the forest offers the higher precision results.

E. Stacking

Stacking is an group/collection learning technique . In the Stacked Generalization technique, two or three models would be used as First Level Classifiers.The joined results from this level would be used in the 2nd Level Classifier models. The outcomes leaving this subsequent level would be utilized as the outcome. The primary level student utilized in the paper was Random Forest and Extreme Gradient Boosting. The 2nd level student utilized was Extreme Gradient Boosting as it were. Following the architecture shown in Fig 2 , we serve the output we get to stacking implementation for better accuracy and precision . This is a compelling method to expand the precision and has been a fascinating subject of enthusiasm with regards to the examination field. Bit by bit process-

Algorithm for Stacking

1. Shuffle the dataset randomly and divide the data set into N groups
eg- if data set is {2,3,4,5,6,1} and n =3
fold 1= {2,3}
fold 2={4,5}
fold 3={6,1}
(Folds can be calculated by starting with an index 0 , elements from index 0 till N/n are in 1st fold . Second fold comprises of elements from N/n to 2*N/n index and so on)

2. For each unique group as a test data , take rest of the group as training data .Fit a model on this set and evaluate it on test set .Store the evaluated value now with N models we will have N-1 train data and 1 test data model 1 : Trained on 1st fold +fold 2 , tested on 3 model 2 : Trained on 2nd fold +fold 3 , tested on 1 and so on
3. A base model is fitted on the n-1 parts and predictions are made for the nth part , do this for each part of the training data to calculate the performance eg- one can use Random forest as first classification and XGB as second classifier models=[RandomForestClassifier(),XGBClassifier()]
4. Repeat the last 3 steps for other models like Logistic regression
5. Results from train set serve as a guide for second level model eg-Accuracy : 0.61 (+/-0.01)[KNN] Accuracy : 0.62 (+/-0.02)[Random forest] Accuracy : 0.65 (+/-0.03)[Stacking classifier]
6. Second level model is used to make a prediction on the test set . The second level classifier is called the meta classifier which gives us the best result

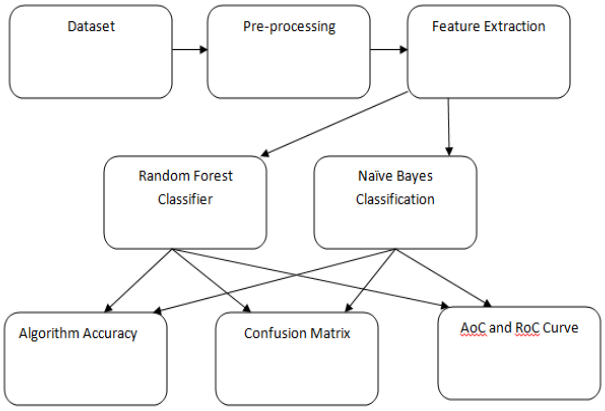


Fig. 1. Shows Architectural Diagram

V.RESULTS

Stacking has performed comparatively better than the other models namely Naïve Bayes, Logistic Regression and Decision Tree. On the other hand, the performance of this model is slightly better than that of the Random Forest Model, which is an ensemble learning technique that was considered to be most accurate till now. The maximum prediction accuracy obtained by stacking was 63.15% as seen in Fig 3.

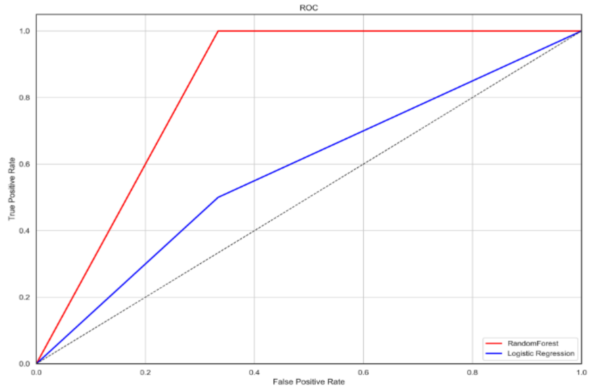
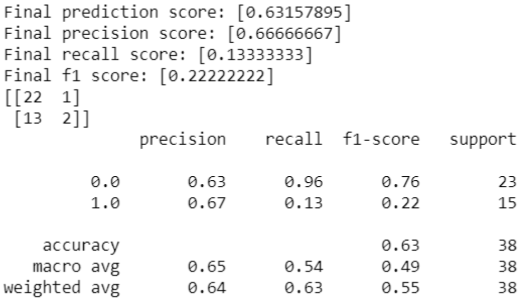


Fig. 2. Comparison of Random Forest and Logistic Regression using a ROC Curve.



```
Final prediction score: [0.63157895]
Final precision score: [0.66666667]
Final recall score: [0.13333333]
Final f1 score: [0.22222222]
[[22  1]
 [13  2]]
              precision    recall  f1-score   support

         0.0       0.63      0.96      0.76        23
         1.0       0.67      0.13      0.22        15

    accuracy                           0.63        38
   macro avg       0.65      0.54      0.49        38
weighted avg       0.64      0.63      0.55        38
```

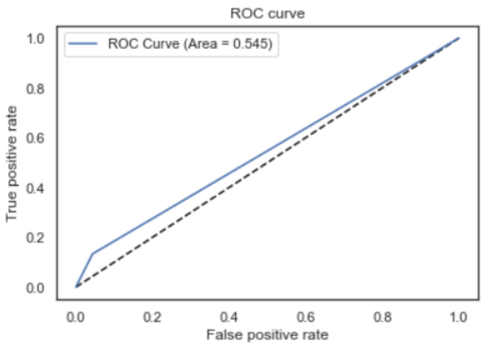


Fig. 3 Here shows details on the Stacking model with the ROC Curve.

VI.CONCLUSION

From the system that is studied ,it's concluded that when more than one model is made, using the averaging system of ensembles, or stacking, the final predictions that are done, are always better than the individual predictions of each of the models. There would be Random forest and the parameters of random forest that would be changed, such as learning rate, decay, momentum, weight etc. Random forests or random choice forests unit accomplice ensemble gaining knowledge of method for type, regression and other duties that operates by building a multitude of selection bushes at schooling time and outputting the magnificence this is the mode of the lessons (classification) or imply prediction (regression) of the person bushes.

REFERENCES

1. Liu, Z., Li, Y., Liu, D., & Shao, C. (2011, June). Semantic detection of malicious code based on the normalized system call. In *2011 International Conference on Computer Science and Service System (CSSS)* (pp. 1680-1683). IEEE.
2. Hachinyan, O. (2017, February). Detection of malicious software on based on multiple equations of API-calls sequences. In *2017 IEEE conference of Russian young researchers in electrical and electronic engineering (EIConRus)* (pp. 415-418). IEEE.
3. Madani, P., & Vlajic, N. (2016, October). Towards sequencing malicious system calls. In *2016 IEEE conference on communications and network security (CNS)* (pp. 376-377). IEEE.
4. Pungila, C. P. (2009, September). A bray-curtis weighted automaton for detecting malicious code through system-call analysis. In *2009 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (pp. 392-400). IEEE.
5. Chen, D., Zhang, H., Zhang, X., & Wang, D. (2016, October). Android Malicious Application Detection Based on Ontology Technology Integrated with Permissions and System Calls. In *2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)* (pp. 481-484). IEEE.
6. Phonion: Practical Protection of Metadata in Telephony Networks Stephan Heuser, Bradley Reaves, Praveen Kumar Pendyala, Henry Carter, Alexandra Dmitrienko, William Enck, Negar Kiyavash, Ahmad-Reza Sadeghi, and Patrick Traynor
7. A. Biryukov and I. Pustogarov. Proof-of-work as anonymous micropayment: Rewarding a Tor relay. In Financial Cryptography and Data Security 2015, 2015.

AUTHORS PROFILE

Vatsal Singh -Currently pursuing bachelors from SRM University Kattankulathur during the session 2016-2020. The author is a machine learning enthusiast .Author has completed certified courses from well reputed universities in Design and analysis of algorithms, Artificial intelligence, JAVA platforms, Android development.

Gaurav Malviya- Currently pursuing bachelors from SRM University Kattankulathur during the session 2016-2020. . The author is a coding enthusiast .Author has completed certified courses from well reputed universities in Python Programming, Artificial intelligence, Database Management, Graph Manipulations.