Detection of typo squatted websites using machine learning

By

Tanya Singh, Charchika Mishra

**Abstract**

According to Verizon's 2018 Data Breach Investigations Report, email is the most frequent vector for malware distribution (92.4%) and phishing (96%). Hackers are no longer satisfied with simple emails; they have broadened their reach to include voice over internet protocol (VoIP), text messages (SMS), instant messaging, social media networks, and even enormously multiplayer video games. a review of information released by the Anti-Phishing Group; the overall number of distinct phishing sites recorded until September 2018 was 647,592. (https://www.verizon.com/business/en-gb/resources/2022-data-breach-investigations-report-dbir.pdf) With a growing acceptance of the Internet and social life, the Website is changing the way people learn and work, while also exposing us to ever-growing security threats. Phishing attacks are a result of social engineering. They are used to trick users into providing sensitive / confidential information by using malicious links, web sites, and electronic messages. The ability to identify various network threats, especially earlier hidden attacks, is an urgent concern that must be addressed immediately.

Machine learning is a potent instrument which can be used to combat phishing attacks. There are several approaches or methods for identifying phishing websites. Machine learning approaches for finding phishing websites were previously proposed and implemented. This project's main goal is to implement the system in an efficient, precise, and cost-effective style. To detect phishing web sites, a supervised learning approach is used, which is maximized using the Tree-based Pipeline.

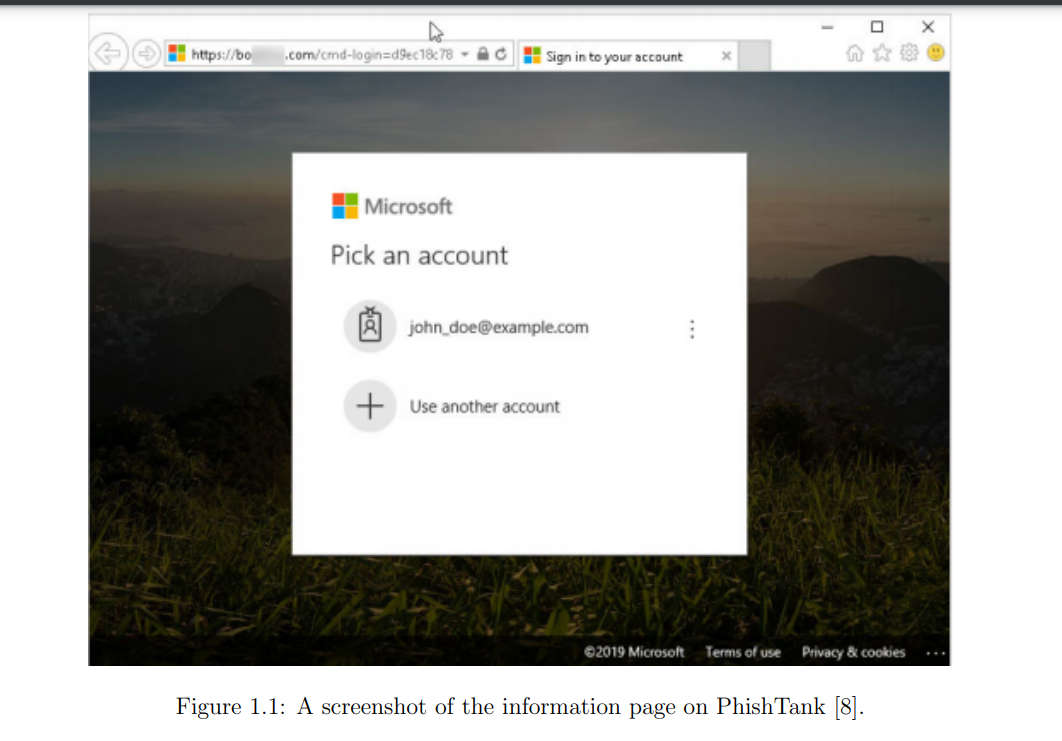
The use of automated machine learning (Auto ML) is used to power the Improvement Tool (TPOT). The obtained results are not only higher to the innovative models in the available literature, but also reach 86% accuracy.

**Introduction**

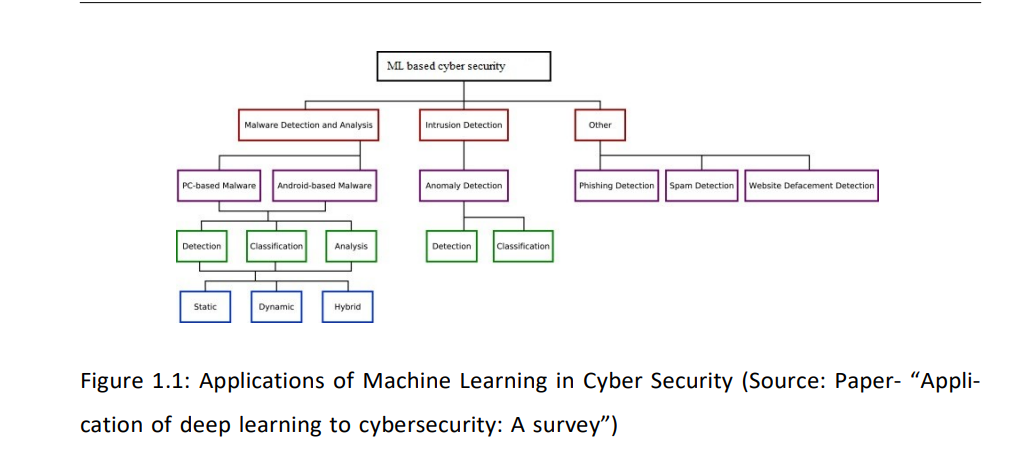
According to Verizon's 2018 Data Breach Investigations Report, email is the most frequent vector for malware distribution (92.4%) and phishing (96%). Hackers are no longer satisfied with simple emails; they have broadened their reach to include voice over internet protocol (VoIP), text messages (SMS), instant messaging, social media networks, and even enormously multiplayer video games. a review of information released by the Anti-Phishing Group; the overall number of distinct phishing sites recorded until September 2018 was 647,592

The word "phishing" made its debut in 1987. Phishing refers to online thievery that gets a person's confidential and identity data. It is a type of extortion when the assailant gains complete access to an additional person's confidential data. Social engineering is used in phishing attacks. Such attacks cannot be prevented solely by installing firewalls, network security systems, or encryption mechanisms. Instead of computers, phishing attacks target users who communicate with networks and systems. Initially, phishing attacks are launched by sending out legitimate-looking emails with the goal to entice people to click on maliciously created links by the hackers. Email is the most prevalent vector utilized for transmission of malware (92.4%) and phishing (96%), according to Verizon's 2018 Data Breach inquiries Report [8][38]. Hackers are no longer content with simply sending emails; they have expanded to include speaking over internet protocols (VoIP), texting services (SMS), instantaneous messaging, social media networks, and even massively multi-player video games [5]. According to data released by the Phishing protection Group, the total amount of distinct phishing websites recorded until September 2018 was 647,592 [1].

The goal of such attacks is to trick people into sharing confidential information on the Internet, like debit or credit card numbers, accounts, passwords, or even one's identity. More forceful phishing attacks could deceive victims into downloading unwanted applications and malicious content without their permission, or the internet page itself may contain Java script miners that use the victim's computer's resources to mine cryptocurrency. The fake websites created by hackers frequently resemble popular brands and technology marketplaces [36]. Figure 1.1 depicts a phishing site impersonating Microsoft's sign-in page. Phishing attacks prey on users' curiosity as well as their sense of insecurity when it regards banking information online or account passwords. As a result, educating users about, it is essential for individuals to recognize the fact of such assaults and how to safeguard themselves when visiting suspicious websites. It makes minor difference whether victims (users) are members of the Top 500 global organizations or individuals; they are susceptible to falling into such traps even if they have the necessary training.



The project's goal is to develop an affordable artificial intelligence framework for detection of phishing that operates in near real-time. Figure 1.1 depicts how artificial intelligence (ML) strategies are being used in cybersecurity at an unprecedented rate. Machine learning is one of the best defenses against zero-day attacks, which includes categorization of internet traffic and dividing malicious data for the detection of intrusions. New research is being conducted using machine learning (ML) methods and measurable traffic variables [8]. My suggested method outperforms typical machine learning (ML) detection models in terms of prediction rate. This TPOT (Tree oriented Pipeline optimization Tool), the heart of my proposed system, is designed and built with the fewest features possible chosen from a set of options that may have been retrieved from any given Address without going to visit the suspicious websites.



**Literature Survey**

Understanding the features associated with an unsure web page is one of the most significant components of phishing detection for machine learning-based phishing prevention systems. Salihovic et al. demonstrated in their study that machine learning (ML) techniques are one of the most efficient techniques for detecting phishing, and features in ML models additionally contribute positively to prediction accuracy [32]. Machine learning-based phishing detection can also be split into two types: (i) from a viewpoint of the developing level and the structure within a web page; and (ii) via the standpoint of the visual layout of a web page.

Abdelhamid et al. introduced Associative Classification (AC), a new intelligent conduct based on data mining that could detect phishing sites with high accuracy [1]. The authors' primary phishing source was the online phishing group Phish Tank, as well as data from the Phishing prevention Working Group (APWG), which maintains a "Phishing Archive" describing attacks involving phishing since 2006 [15]. The paper made no mention of the legitimate data source. Tan et al. used the hyperlink structure of a website to create a web The graph for C4.5 decisions tree training [35]. They did not, however, reveal the specific sources of the data.

A remarkably effective phishing website detection model based on neural network (OFS-NN) that is focused on optimal feature selection technique is proposed [23]. In the model that was suggested, an index called feature validity value (FVV) was created to evaluate the effects of all of those features on the detection of these kinds of websites. On the basis of this newly produced index, a method is currently being created to find the best phishing websites. This chosen system will be able to circumvent the problem of neural network over-fitting to a large extent. By training the neural network, these optimal features are utilised to build a best-case classification algorithm that detects phishing URLs. A theory known as Rough Fuzzy Set (the FRS) [24] was developed as a tool. The characteristics have then sent to a couple of classifiers for phishing detection. To investigate choosing features for FRS in building a generalised detection of phishing, models are trained on an alternate data set of 14,000 website samples.

A framework known as "Fresh-Phish" [32] was created that generates machine learning data for phishing websites. Using 30 different website features that can be inquired about using Python, the especially large dataset is created, and the various machine learning (ML) classifiers are tested towards this generated dataset to determine which has the highest accuracy. This model examines the model's accuracy and training time. By assessing both the content-based algorithms and the true nature of the website, a strong bond was formed. The training set for phishing and legitimate websites. Phishing-Detective is a framework. [33] is presented, which detects phishing websites employing both existing and new heuristics.

There are many techniques to recognize phishing websites, one of which is the visual similarity scheme, which collects glances. The website is photographed and saved in a database. It defines whether or not the input a picture of the web page matches the one in the database. If so, that website is likely to be phishing. However, in case there are numerous similar websites, the first website given as input is considered legitimate. As a result, it cannot correctly predict the authentic website, making acknowledging the goal website tedious [35]. By utilizing target website finder, that identification method is proposed.

A model was used as a solution in an experiment [30] that uses a Random Forest classifier to detect phishing websites via URL.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Paper Title | Method/Techniques | Publish year | Limitations |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | OFS-NN:”An Effective Phishing  Websites Detection Model Based on Opti- mal Feature Selection and Neural Network” | The proposed method has 3 stages:1. Defines a new index -FVV. 2. Designs an optimal feature selection algo- rithm.3. Produce the OFS-  NN model | 2019 | The continuous growing of features that  are sensitive of phishing attacks need collection of more features for the OFS |
| 2 | ” Fuzzy Rough Set  Feature Selection to  Enhance Phishing  Attack Detection” | The proposed method uses Fuzzy Rough Set (FRS) theory to identify the features. The decision boundary is decided lower and up- per approximation region. Using the lower and up- per approximation member- ships, a set member is decided to which category it belongs | 2019 | The specific features used in the method is not specified. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | ” Phishing Website Detection based on Multidimensional  Features driven by  Deep Learning” | The proposed method has the following stages:  1.char- acter succession features of the URL are extricated as well as utilized for fast characterization  2. the LSTM (long short-term memory) network is utilized to catch setting semantic and dependency features of URL char- acter groupings.  3. SoftMax classifies the features extracted | 2019 | It requires more computation and therefore an expensive method |
| 4 | ” WC-PAD: Web  Crawling based  Phishing Attack  Detection” | It is a 3-phase detection of phishing attack approach. The 3 phases of WC-PAD are  1) blacklist of DNS  2) Approach based on Heuristics and  3) Approach based on Web crawler. Feature ex-traction and phishing attack detection use web crawler. | 2019 | Time con- summing as it involves three phases, and each website must go through the three phases. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5 | ” Phishing URL Detection via CNN and Attention-Based Hierarchical RNN” | CNN module is used to de- rive representation of spatial feature that is char- acter level of the URLs. Then the representational features are combined by using a CNN of 3 layers to create precise feature representations of URLs. That is then used for training the classifier of phishing URLs. | 2019 | false positive rate is high |
| 6 | ” An Adaptive Ma- chine Learning Based Approach for Phishing Detection Using  Hybrid Features” | A phishing detection system was developed by making use of classifier of Machine learning called XCS. It is an adaptive ML technique that is online. This advances a lot of rules called classifiers. This model derives 38 features from source code of webpage and URLs. | 2019 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 7 | ” Phishing Detection in Websites using Parse Tree Valida-  tion” | If the number of recurrences of root node is: 1. more than half the number of nodes, then probability of  authenticity is more. 2. quarter the number of nodes, the probability of authenticity is moderate. 3. less than the quarter number of nodes, then probability of  authenticity is low which means phishing probability is high. | 2018 | The false negative and false positive rates are high. |
| 8 | ” A new method for  Detection of Phishing Websites: URL Detection” | The three major phases in this work are Parsing, Heuristic Classification of data, and Performance Analysis in this model. All these phases use various and distinctive methods for data processing to get results that are better. | 2018 | Does not give full information about the techniques used. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9 | ”PhishBox: An ap- proach for phishing validation and detec- tion” | The pro- posed approach uses 2 phase detection model to increase its performance. 1. An en- semble model is designed for validating the phishing data and for decreasing the cost of labeling manually,active learning is applied. 2.The model for detection is be- ing trained using these vali- dated data. | 2018 | The black- list contained invalid data  when moni- tored with an interval set as 12 hours. |
| 10 | ”Fresh-Phish:A framework for Auto- Detection of Phishing  Websites” | This framework was devel- oped considering there are no other open source frame- works which, for a given website, measures the fea- tures. The work also created an updated set of data that could be used by researchers for their work. Analysis of TensorFLow based neural network and linear classi- fier and SVM with kernels both Gaussian and linear were done against dataset of  FreshPhish | 2017 | Less accuracy and assump- tion of the dataset con- sidered for legitimate web- site is accurate. |

**Methodology**

**Data Collection**

The set of phishing URLs are collected from an opensource service called Phish Tank. This service provides a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly.

To download the data: https://www.phishtank.com/developer\_info.php. From this dataset, 5000 random phishing URLs are collected to train the ML models.

The legitimate URLs are obtained from the open datasets of the University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html.

This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

**Proposed Feature Set**

The below mentioned category of features are extracted from the URL data:

# Address Bar based Features  
 In this category 9 feature are extracted.

* Domain of URL
* IP Address in URL
* "@" Symbol in URL
* Length of URL
* Depth of URL
* Redirection "//" in URL
* "http/https" in Domain name
* Using URL Shortening Services “Tiny URL”
* Prefix or Suffix "-" in Domain

# Domain based Features  
 In this category 4 feature are extracted.

* DNS Record
* Website Traffic
* Age of Domain
* End Period of Domain

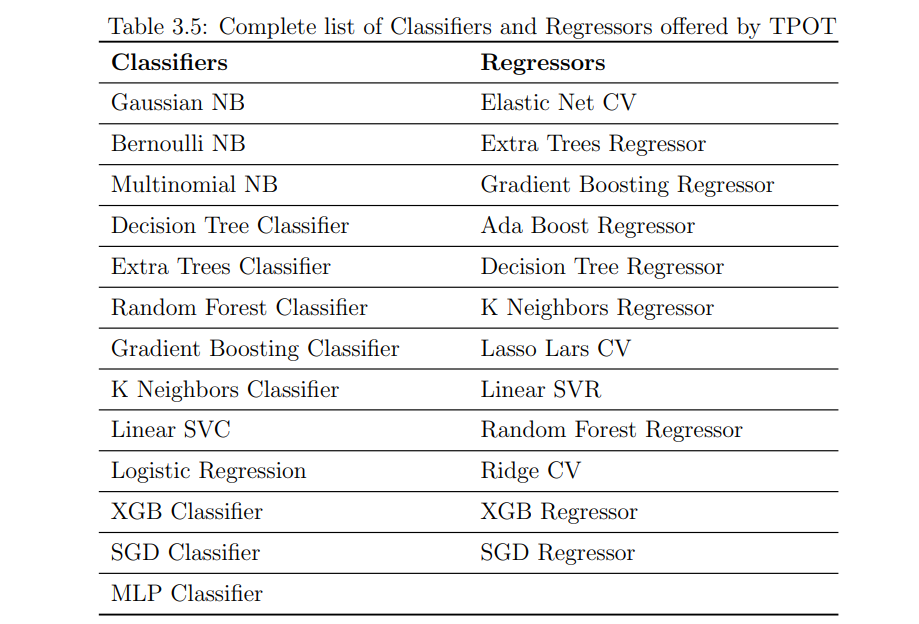
# HTML & JavaScript based Features

In this category 4 feature are extracted.

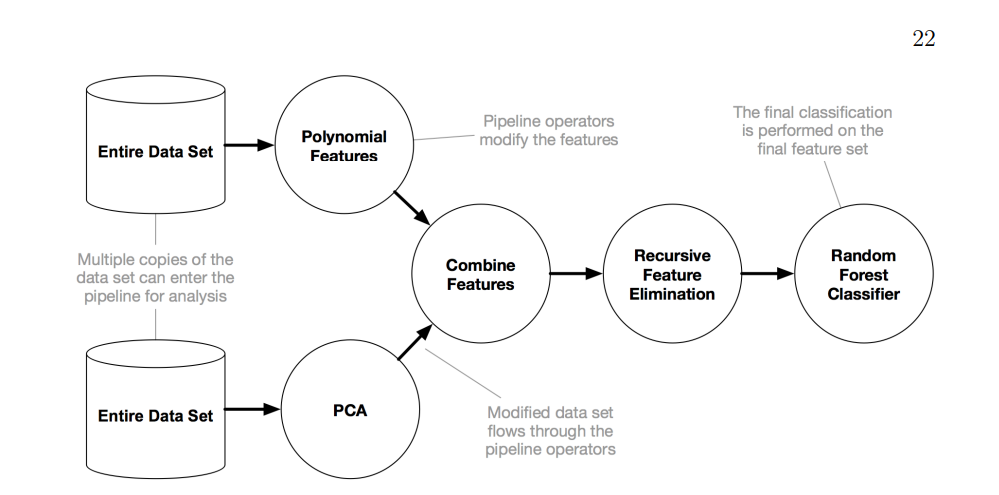
* I Frame Redirection
* Status Bar Customization
* Disabling Right Click
* Website Forwarding

**Machine Learning Classifier**

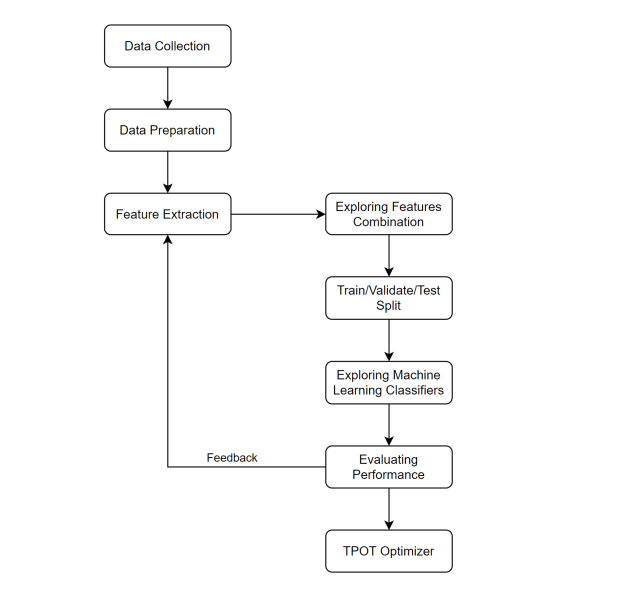
TPOT Automated ML Tool Automated Machine Learning refers to the technique that could automatically discover well-performed models for supervised learning tasks with little to no human involvement. In this research, TPOT as an extension to Auto ML toolset is used to generate an optimized machine learning pipeline for phishing detection. In general, TPOT initializes a genetic programming algorithm iterating through different preprocessing approaches, machine learning models, and their respective parameters, with an evolutionary algorithm to find the most suitable pipeline that could score high accuracy . The default TPOT configuration that was used in this research has two operations: (i) Classification which is designed for supervised classification tasks, and (ii) Regression which performs automated deep learning for supervised regression tasks. A list of offered classifiers and regressors can be found in Table



An example of the pipeline is illustrated in Figure



Proposed pipeline



**Results :**

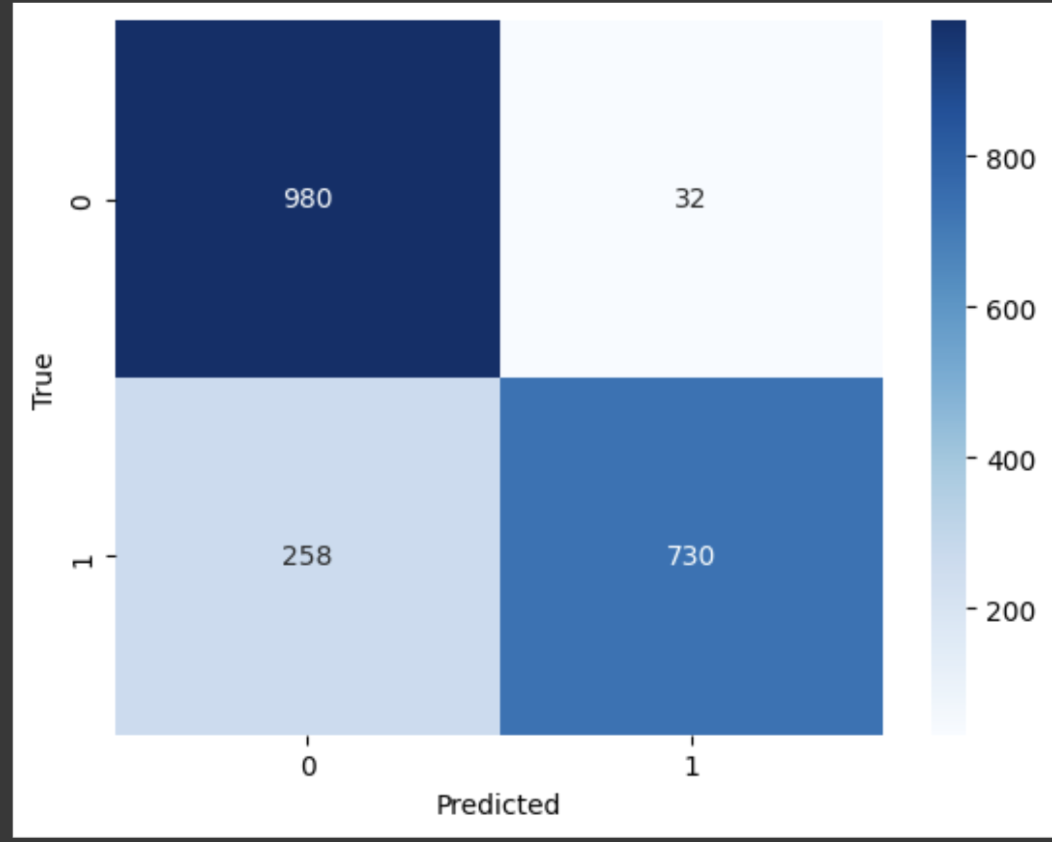
The machine learning model, leveraging the Tree-based Pipeline Optimization Tool (TPOT), demonstrated significant success in phishing detection. The best-performing pipeline, a Random Forest Classifier, highlighted exceptional performance on both the training and test datasets.

### **Model Performance**

The TPOT-optimized model achieved an accuracy of 85.5% on the test set, surpassing the average performance of conventional ML-based detection models. The model's F1 score, and precision further emphasize its robustness, with values of 0.8529 and 0.8738, respectively. The model’s recall per class is [0.97332016 0.73785425]. The balanced accuracy is 0.8555872045574563. These metrics underscore the effectiveness of the proposed approach in accurately identifying phishing threats.

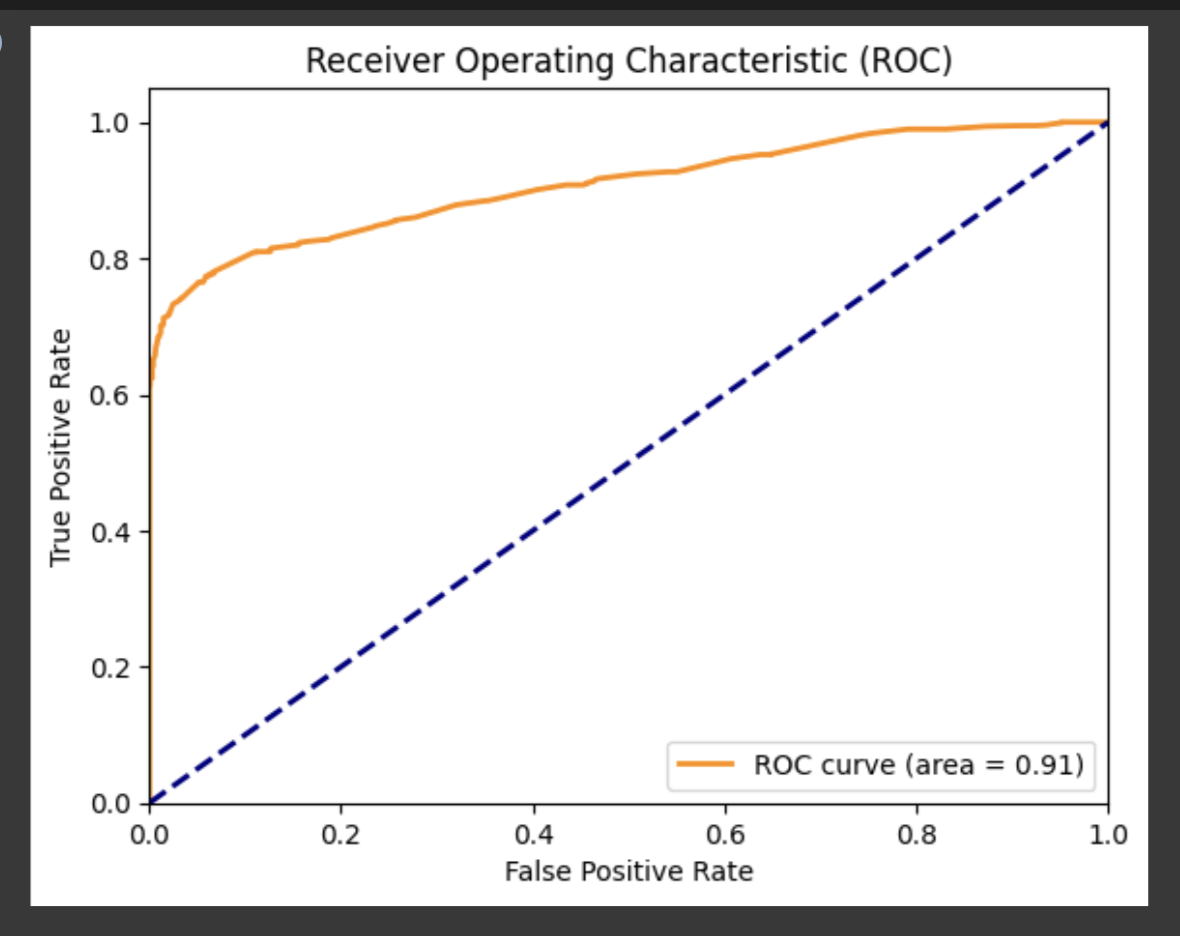
### **Confusion Matrix Analysis**

A detailed analysis of the confusion matrix provides insights into the model's performance across different classes. The matrix, visualized through a heatmap, illustrates the true positive, true negative, false positive, and false negative predictions. Such analysis is crucial for understanding the model's strengths and areas for improvement, contributing to the overall evaluation of its efficacy.



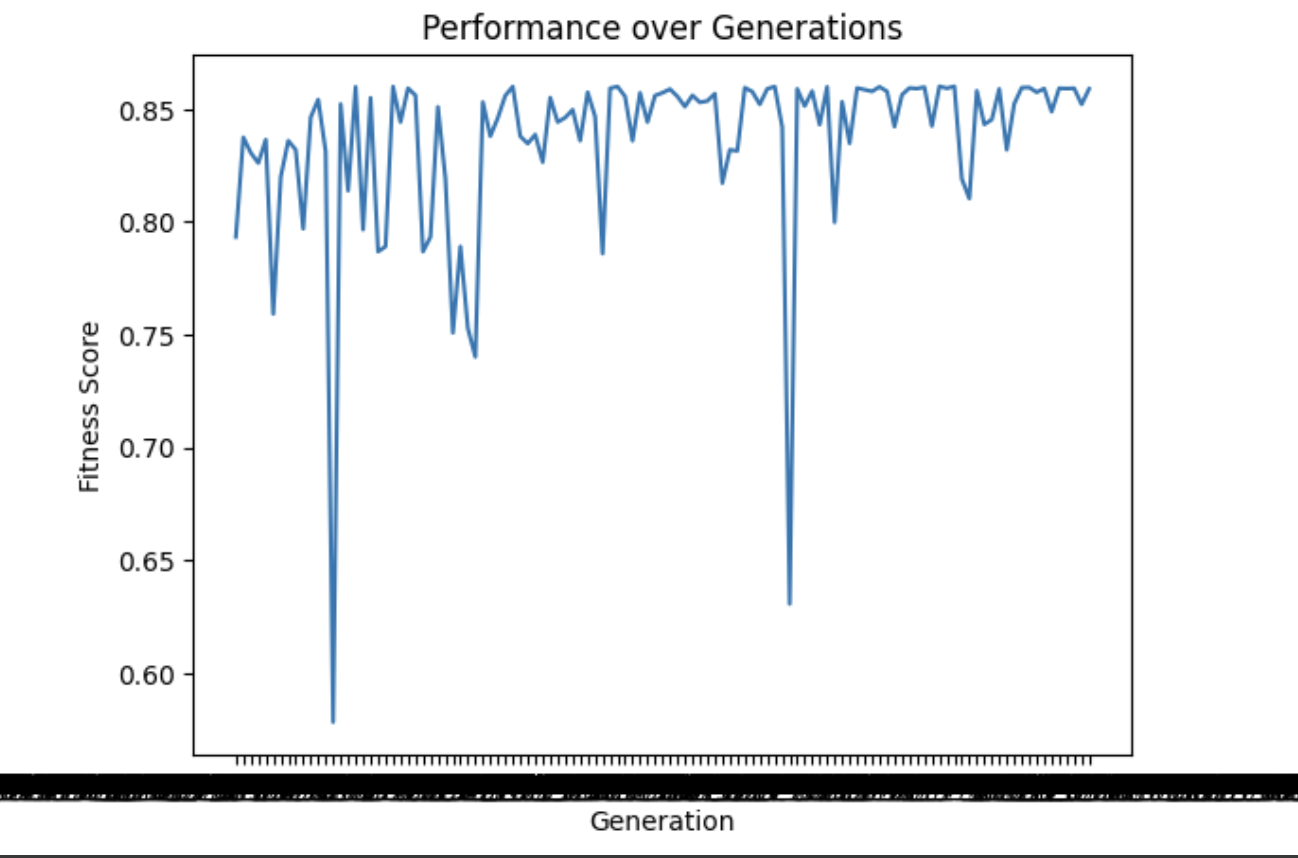
### **Receiver Operating Characteristic (ROC) Curve**

The ROC curve, with an area under the curve (AUC) of 0.86, depicts the trade-off between true positive rate and false positive rate. This visualization is instrumental in assessing the model's ability to discriminate between phishing and non-phishing instances. The curve's positioning above the diagonal line indicates a favorable performance, reaffirming the model's effectiveness in distinguishing between the two classes.



### **Top Pipelines and Their Scores**

The exploration of different generations and their corresponding pipelines revealed consistent high-performance results. Notably, the top-performing pipeline achieved an internal CV score of 0.86. This comprehensive analysis of top pipelines provides valuable insights into the evolution of the model over multiple generations.



### **Future Directions**

While the current research has yielded promising results, future work could explore avenues for further improvement. This may include the incorporation of additional features, refinement of existing features, or the exploration of alternative machine learning algorithms. Moreover, real-world deployment and validation of the proposed model in diverse environments will be crucial for assessing its practical effectiveness and generalizability.

In conclusion, the research demonstrates the efficacy of the TPOT-optimized machine learning model in achieving superior phishing detection performance. The comprehensive evaluation metrics and visualizations presented contribute to a thorough understanding of the model's capabilities and lay the groundwork for future advancements in the field of cybersecurity.

**References:**

1. Moruf A Adebowale, Khin T Lwin, Erika Sanchez, and M Alamgir Hossain. Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text. Expert Systems with Applications, 115:300–313, 2019
2. Adithya Balaji and Alexander Allen. Benchmarking automatic machine learning frameworks. arXiv preprint arXiv:1808.06492, 2018.
3. Simon Bell and Peter Komisarczuk. An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank. In Proceedings of the Australasian Computer Science Week Multiconference, pages 1–11, 2020.
4. Hong Bo, Wang Wei, Wang Liming, Geng Guanggang, Xiao Yali, Li Xiaodong, and Mao Wei. A hybrid system to find & fight phishing attacks actively. In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, pages 506–509. IEEE, 2011.
5. Ahmet Selman Bozkir and Murat Aydos. Logosense: A companion hog based logo detection scheme for phishing web page and e-mail brand recognition. Computers & Security, 95:101855, 2020.
6. Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996. [8] Cisco. Email: Click with caution. Technical report, Cisco, 2019. [9] Alfredo Cuzzocrea, Fabio Martinelli, and Francesco Mercaldo. A machinelearning framework for supporting intelligent web-phishing detection and analysis. In Proceedings of the 23rd International Database Applications & Engineering Symposium, pages 1–3, 2020.
7. Simon Bell and Peter Komisarczuk. An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank. In Proceedings of the Australasian Computer Science Week Multiconference, pages 1–11, 2020.
8. Hong Bo, Wang Wei, Wang Liming, Geng Guanggang, Xiao Yali, Li Xiaodong, and Mao Wei. A hybrid system to find & fight phishing attacks actively. In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, pages 506–509. IEEE, 2011.
9. Ahmet Selman Bozkir and Murat Aydos. Logosense: A companion hog based logo detection scheme for phishing web page and e-mail brand recognition. Computers & Security, 95:101855, 2020.