# Data Science Analysis of Malicious Advertisements and Threat Detection Automation for Cybersecurity Progress

Sandra H. Nguyen, Dr. Doina Bein
Department of Computer Science, California State University, Fullerton

CALIFORNIA STATE UNIVERSITY FULLERTON

ASSURE-US

## Abstract

We are living in an era of unprecedented technology. Millions of users depend on information technology to carry out their daily lives and large-scale commercial and industrial operations are no exception. At the same time, the rapidly growing interconnectivity of IT systems and the surge in cybercrime since the pandemic have rendered industry-standard hardware and software components increasingly vulnerable to malicious attacks. Cyber defense is a coordinated act of resistance that intends to understand the capabilities and motives of attackers in order to secure our country's data and more importantly, the livelihoods of our citizens. This research aims to contribute to the progress of cybersecurity and defense technology as a whole by focusing on a dynamic aspect of malware: advertisements. It presents a novel approach to automating the analysis of malicious content on the internet by web scraping ads of the popular search engine Google to extract relevant data (URL, Company, Title, Product Desc.), building machine learning models (supervised & unsupervised) to classify and make predictions on that data, and creating a web application for end users to access. The results show that our tool can detect trends within the features with limited false positives, paving the way for us to make predictions on whether the advertisements are benign or malign. The research concludes that in this time and age, it is extremely important to protect against fraud, especially by adhering to cybersecurity's best practices and to think about threats in more global terms. Our hope with this research is to prompt action to ensure society continues to improve in IT resilience.

## Introduction

In recent times, online advertising has become the predominant strategy for business marketing, which greatly impacts the day-to-day activates of online users, such as entertainment, communication, and e-commerce. Unfortunately, due to the popularity of the Internet, online platforms are flooded with meaningless and malignant advertisements that devalue the user experience and deliver complex malicious activities—such as malware, scam, and phishing. Today, attackers are targeting unsuspecting users by purchasing Google Ads based on specific keywords that point to compromised websites. Studies have concluded that a significant proportion of users fail to recognize certain types of media as ads, which invokes a question: To what extent has cyberattacks evolved in regards to our ability to detect it? How can we, as a society, improve in cybersecurity vigilance and develop the necessary technologies to remedy this disparity?

In my research, I propose an exploratory detection system for potentially malicious advertisement URLs. It involves: (1) The automation of web-scraping Google Ads based on user-inputted keywords, (2) Feature extraction, data analysis, and supervised & unsupervised machine learning algorithms, such as Random Forest Classification and K-means Clustering for prediction and identifying patterns, and (3) A user-friendly web app for real-world application. The goal is to put the power back in online users' hands and allow them to understand the media they interact day-to-day.

## Technologies

We explored a range of developer tools, Python libraries, and Machine learning algorithms which are listed, to an extent, below:

- Technologies: Visual Studio Code, Jupyter Notebook, Streamlit, & Kaggle
- Libraries: Beautiful Soup, Pandas, Numpy, Scikit-learn, Matplotlib, & Seaborn
- Models: Logistic regression, Random forest, Decision tree, K-means, & Gaussian Mixtures

## Methodology

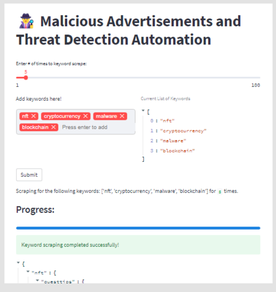
**Data Science Aspect:**
Scrape potentially malicious advertisements on popular search engine, Google, using BeautifulSoup, a parsing library. Extract relevant data (URL, Company, Title, and Product Description) to a csv file for data analysis, by accessing each web page's component. I structured the data with a dictionary so that for every key word, the top-performing companies would be recorded along with the number of times it appeared at the top (most impressions).

**Machine Learning-Focused Aspect:**
Employ an end-to-end Scikit-learn workflow: (1) Getting the data ready – cross-validation based on a 70-30% train/test split, (2) Handling NaN & categorical data – preprocessing, (3) Choosing the right ML algorithms, (4) Fitting the models & making predictions, (5) Evaluating the models – accuracy score & silhouette score, and (6) Displaying the results with plots.

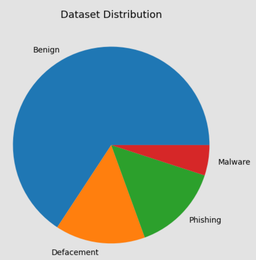
We examine two datasets, an extremely large collection of 651,191 URLS with the classification "benign", "defacement", "phishing", and "malware" published on Kaggle in 2021, and our own 615-count, scraped dataset with an assortment of sample key words, including "nft", "cloud computing services," and "artificial intelligence software".

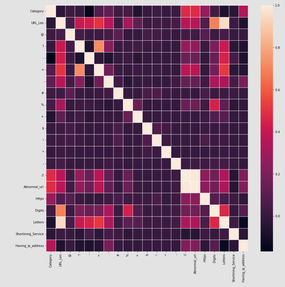
Feature engineering includes finding the URL length number of digits & letter characters, domain name, symbols (@, ?, #, %, and more), the presence of an HTTP, shortening service, and IP address. All models are built and evaluated according to their accuracy or silhouette score.

**Web Development Aspect:**
Develop a web application to allow users to input certain keywords and the number of times to scrape into the AI model, with the goal of receiving a more unbiased sample. Users will also be able to display and view their results through the data structure.

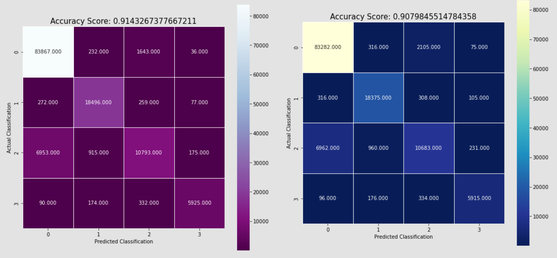



## Results

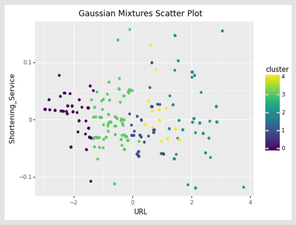


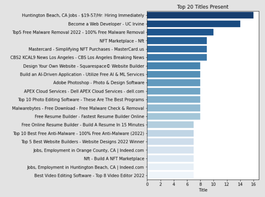



For the Kaggle dataset, the target distribution is 428,103 benign, 96457 defacement, 94,111 phishing, & 32,520 malware URLS. Our relevant features and their patterns are depicted with a seaborn heatmap. The closer a value is to 1.0 (light color), the stronger their positive correlation is.




Our Random forest classification model (left plot) gives the best accuracy score of 91.4% compared to Decision tree (right plot) and Logistic regression (not pictured). A confusion matrix for each model is generated to visually represent the Actual vs. Predicted values.




For our scraped URLs dataset, the top 20 Companies, i.e. Adobe, Amazon, & Ziprecruiter are identified. Furthermore, the top 20 Titles bar graph depicts an interesting trend, whereas our keywords spit back ads that are location-specific, "Huntington Beach", "UC Irvine", "KCAL9 News", etc.

We build unsupervised learning models Kmeans Clustering and Gaussian Mixtures based on the features "URL" and "Shortening Service" which are often by spammers to evade detection and blacklisting. Our Gaussian Mixtures model gives finds four clusters and gives the best silhouette score of 0.421.

## Conclusion

Since the mid-1990s, the Internet has developed rapidly and impacted almost if not all aspects of society, for the better or for the worse. In investigating malware and the relationships between, for example, the surge in cryptocurrency and the rise in cyberattacks, as well as cybersecurity best practices, it led me to realize that we have reached a point of no return in technology. Our digital footprint, the things we shop for, our ideas and opinions, our locations—everything is tracked and stored in some company's database ready to become the new baseline. As users, our data has become the new gold. And hackers want to steal it.

The components of this research kindly lends the everyday user a hand in helping them understand the online opponents they face everyday: advertisements, friendly or not-so-friendly. The results achieved from the machine learning aspect of this project were anticipated to a degree, as the sampling of data is plenty and non-biased. The web application, furthermore, will only become more functional and intuitive as I continue my research in this field.

With this study, we can now conclude that the key to preventing cyberattacks is to remain vigilant of everyday processes – even something seemingly safe such as Google Ads; and keeping your staff trained and systems up to date, in order to protect the common user placing their trust in your services.

## References

1. *Gottsegen, Gordon. "Machine Learning Cybersecurity: How It Works and Companies to Know." Built In, 30 June 2019.*
2. *Staff, the Premerger Notification Office, and DPIP and CTO Staff. "Blurred Lines: An Exploration of Consumers' Advertising Recognition in the Contexts of Search Engines and Native Advertising: A Federal Trade Commission Staff Report." Federal Trade Commission, 5 Aug. 2021.*
3. *Zhang, Xichen, et al. "Classifying and Clustering Malicious Advertisement Uniform Resource Locators Using Deep Learning." ResearchGate, Nov. 2020.*
4. *EasyDmarc. "URL Shorteners: Pros and Cons." EasyDMARC, 6 Sept. 2022.*
5. *Web application adapted from Andrew-FungKinHo on Github*

## Acknowledgements

The research presented in this poster is part of the ASSURE-US Summer Research Experiences and is entirely supported by the National Science Foundation for the project Building Capacity: Advancing Student Success in Undergraduate Engineering and Computer Science Award # 1832536.