Predicting Phishing Attempts by Disassembling and Analyzing URLs

1st Logan Warren
University of Pittsburgh
Information Science Undergraduate
Pittsburgh, United States
LAW234@pitt.edu

2nd Xavier Bear
University of Pittsburgh
Information Science Undergraduate
Pittsburgh, United States
XMB1@pitt.edu

3rd Charles Tran
University of Pittsburgh
Information Science Undergraduate
Pittsburgh, United States
CHT126@pitt.edu

***Abstract*—Our project's goal is to create and compare machine learning models to detect phishing scams. The question we are trying to solve is: Can we predict the likelihood of a website being a phishing site, based on details within the URL's structure with high accuracy given the correct model? We felt that this question was interesting because we are involved in the cybersecurity club, and wanted to demonstrate how data and cybersecurity can be used to help solve cyber-related crime. Our project is useful because as technology advances, so does the amount of sophisticated phishing scams, especially with the use of AI. With our project, we aim to use a model to help detect these new phishing scams easier. Companies that stand to benefit from our model include cybersecurity firms, web browsers, and end-users. We also believe that universities such as Pitt, could benefit from a phishing model to help bring down the recent surge of phishing emails. Other schools or institutions could benefit from this as well, to help fight against students getting scammed.**
Our plan is to preprocess the dataset, check for missing values, and clean the data, checking and potentially dropping any extraneous data. We'll extract features from URLs, such as domain details and lexical patterns, which are indicative of phishing attempts. After a brief exploratory analysis to understand the data better, we'll train various machine learning models, comparing their effectiveness. The best model will then be fine-tuned and evaluated to ensure its reliability in detecting phishing threats

***Index Terms*—Phishing, Cybersecurity, URL, Data, Random Forest, Logistic Regression**

I. Introduction

The threat of cyber attacks is continuously escalating as the world around us becomes more interconnected with digital landscapes. One of the largest running cyber threats has been phishing scams. We are aiming to address these scams by being able to predict the likelihood of a website being a phishing site, using URL analysis. Our approach utilizes the "Phishing Websites Dataset", featuring 88,647 instances of both legitimate and phishing sites. Using this data we develop multiple models that can discern phishing attempts based on URL structures.

Despite the advancements in cybersecurity, phishing remains a challenging problem due to its continually evolving nature. Traditional methods of phishing detection are often unable to keep pace with sophisticated techniques employed by attackers. In this study, we explore the efficacy of URL structure analysis combined with machine learning techniques as a way of approaching phishing detection. The potential impact of our models is significant. By developing a reliable model to detect phishing, we can contribute to the broader cybersecurity efforts in mitigating the risks associated with online scams.

II. Methodology

We began with obtaining our data, "Phishing Website Dataset" from Mendeley Data and uploading it into our Jupyter Notebook. The dataset holds a comprehensive collection of both legitimate and phishing website instances. We began initial data exploration using Pandas, to review the dataset's structure and check for missing values. Once confirming no missing values in the dataset, we began to identify and handle outliers.

We calculated Q1, Q3, and IQR for each column and identified the outliers, putting them into a data frame for analysis. This found that there were 75,209 outliers in our data. To look closer at the outliers, we selected five columns based on their variance and relevance to phishing. Plotting these columns, we were able to see each feature showed varying levels of dispersion and outlier presence. Given the nature of cybersecurity data, particularly in phishing detection, outliers could represent real scenarios. Therefore instead of removing the outliers, we chose to keep them in the data. Since there was a large amount of outliers we recognized their potential statistical significance, and how removing them could create skewed results or potential biases.

To gain a more granular understanding of these outliers, we selected columns based on their high variance and relevance to phishing, 'qty_dot_url', 'qty_hyphen_url', 'qty_slash_url', 'qty_questionmark_url', and 'qty_equal_url'. We generated Boxplots for these selected features, which showed their distribution and extent of outliers.

In our analysis, we used Q-Q plots to assess the distribution of key features against a normal distribution. These plots are insightful for understanding the distribution characteristics of our data and identifying potential outliers.

A. Q-Q Plot of qty_dot_url

This plot compares the distribution of the number of dots in URLs against a theoretical normal distribution. Deviations from the line in the plot indicate that this feature does not follow a normal distribution. This could be indicative of unusual URL structures in the data which are often employed in phishing attempts.

B. Q-Q Plot of qty_hyphen_url

The Plot of qty_hyphen_url similarly visualized the distribution of the number of hyphens in URLs. Deviations found in this plot could suggest unusual URL patterns, potentially signaling phishing URLs.

C. Q-Q Plot of qty_slash_url

With qty_slash_url, this plot helps with understanding the distribution of slashes in the URLs. A non-normal distribution in this feature could signify specific URL patterns that are characteristic of either legitimate or phishing sites.

D. Q-Q Plot of qty_questionmark_url

From the Plot of qty_questionmark_url, we analyze the distribution of question marks in URLs. Significant deviations from the normal distribution line could point to specific query structures in URLs. This can be a common sign that a link is being used for phishing.

E. Q-Q Plot of qty_equal_url

Lastly, the Q-Q Plot of qty_equal_url focuses on the occurrence of equals signs in URLs. Non-normal distributions can be indicative of complex query parameters, which are sometimes a feature of phishing URLs.

In all these plots, points that deviate significantly from the line suggest a departure from the normal distribution, meaning there is a presence of outliers. These outliers are of interest in our study as they might represent unconventional and potentially malicious URL structures employed in phishing attacks. Insights into the outliers are invaluable in enhancing the robustness of our phishing detection models.

Our approach to feature engineering involved identifying '-1' values in numeric columns, which were treated as missing indicators. We transformed these indicators by creating new boolean columns to flag missing data. The '-1' values were then replaced with zeros. Following this process helped to preserve the data's integrity. We also removed any boolean columns that only contained single values since they offered no variability and therefore no predictive value to our models.

In our analysis, we utilized various visualizations to look further at the insights of the dataset. The first data visualization used was a Scatter Plot of the Quantity of Dots vs Hyphens in URLs. This was used to examine the relationship between the quantity of dots and hyphens used in phishing and legitimate sites.

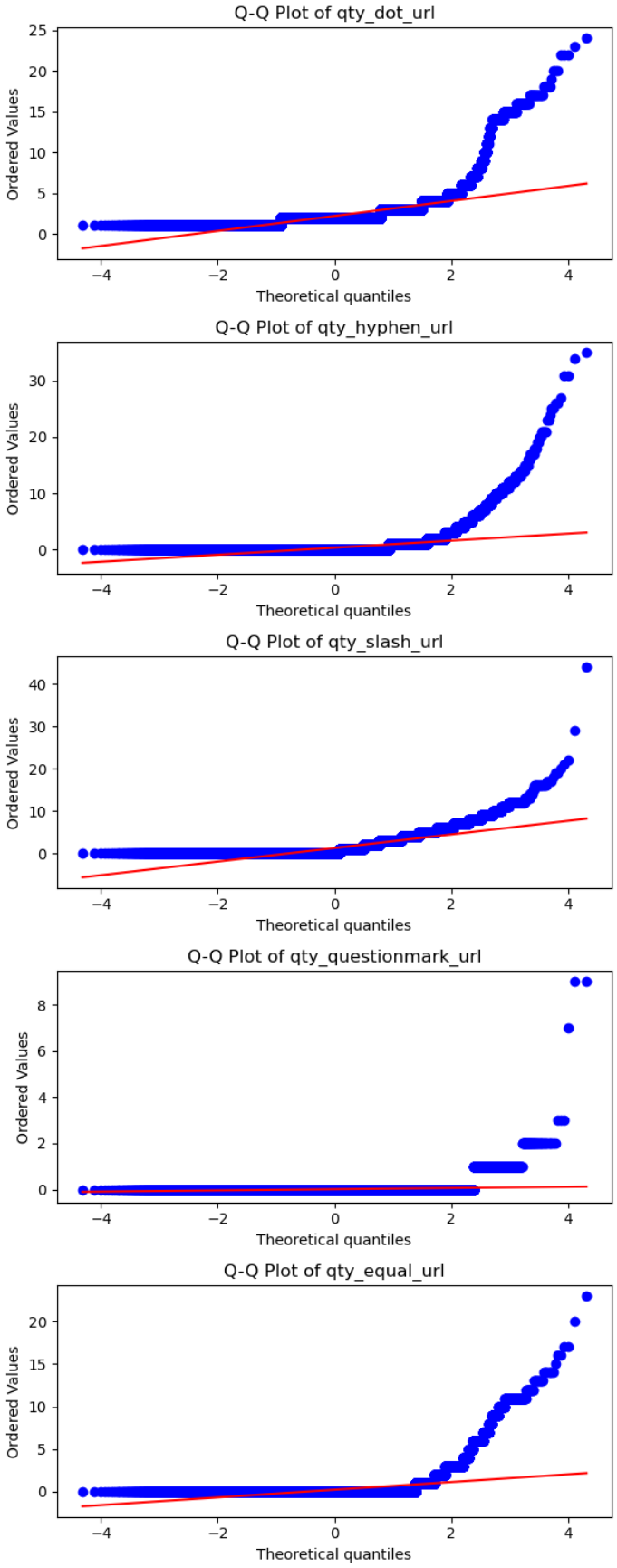

Fig. 1. Quantile-Quantile plots for selected features. Each plot compares the distribution of a specific URL feature (dots, hyphens, slashes, question marks, equals signs) with a theoretical normal distribution, highlighting deviations and potential outliers of phishing URLs.

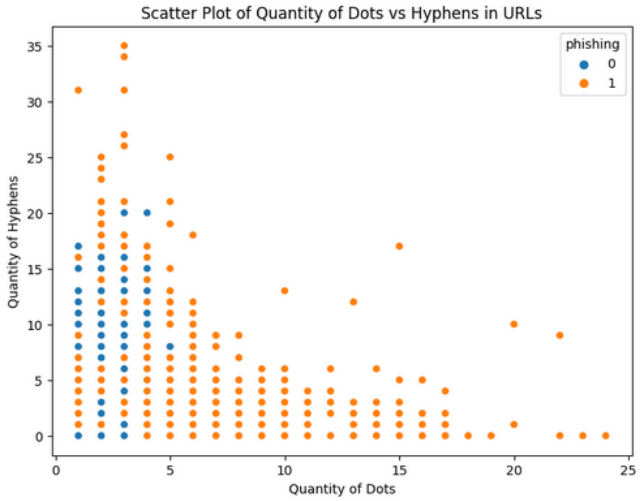

Fig. 2. Scatter plot illustrating the relationship between the quantity of dots and hyphens in URLs. This visualization aids in identifying patterns and anomalies in URL structures.

The next one was a Box Plot of the Quantity of Equals in URLs by Class. This was another analysis of the relationship between equal signs used in phishing and legitimate sites. From these two graphs, we were able to gain a better understanding of the characteristics seen in phishing sites vs legitimate.

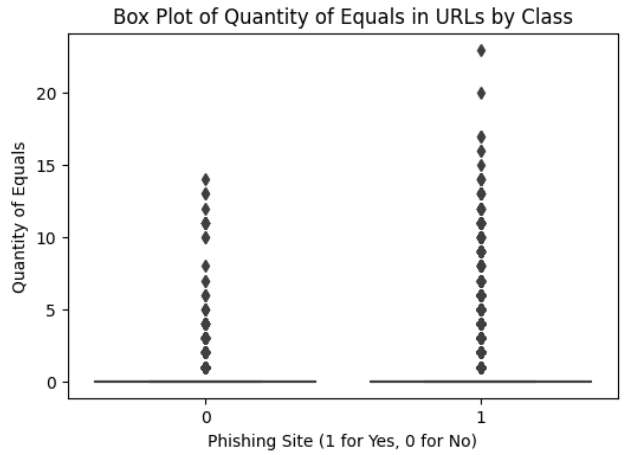

Fig. 3. Box plot showing the distribution of the quantity of equals signs in URLs. This plot provides the variation and outliers of this feature, which is useful for understanding phishing URL characteristics.

When choosing our models, we decided that the Logistic Regression Model and the Random Forest Model (RF) were the best ones to use for our topic. After doing a quick test with other models (Naive Bayes and SVM) we decided that the Logistic Regression and Random Forest were the best. The Logistic Regression Model is a great option for our topic due to its ability to handle high-dimensional data. Additionally, the Logistic Regression model has the power to make linear relationships which could prove useful in finding a simple relationship between certain URL features and the likelihood of phishing. We decided to use the Random Forest Model as well. The RF model's ability to capture non-linear relationships and find patterns within the data is incredible when looking into phishing URLs. The model's feature importance score helps give us insight into what features are most important when it comes to phishing attempts.

After training our models, we created a Feature Importance Plot for RF. This visualization helped to identify which features were most influential in predicting phishing sites. The plot shows the top 10 features that Random Forest found significant.

For model development, we divided the data into features, X, and the target variable phishing, Y. We then proceeded to split the data into training and test sets. Next, the features were scaled using the StandardScaler to normalize the data. Once the data was ready, two models were trained: Logistic Regression and Random Forest Classifier. Both models were evaluated based on accuracy and Area Under the Curve scores. This determined their effectiveness in predicting phishing attempts. Lastly, we created a comparative evaluation to allow us to assess the strengths and weaknesses of each model in handling our dataset.

With the goal of finding out the best way to deal with a large amount of outliers present in our data set (75209 outliers total), we tested the accuracy of the Logistic Regression and Random Forest Classifier Models when applying three separate techniques for dealing with outliers. The chosen techniques are: (1) Keep Outliers, (2) Drop Outliers, and (3) Apply median and modal imputation on outliers for numeric and categorical variables respectfully.

III. Results

A. Predictive Power of Models

A crucial step in executing the Random Forest Model is to identify the most important features. These features guide the model and have a heavy role in influencing the model's predictions. To determine which features had the most weight on the predictions, we decided to create a bar graph of the 10 most important features. A bar graph is easy to interpret and gives a clear visualization of the importance of different features. On the next page are the results:

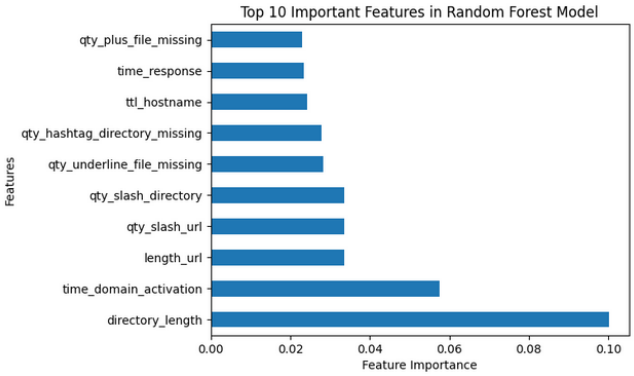

Fig. 4. Bar graph depicting the top 10 most important features as identified in the Random Forest model. This chart highlights the importance of each feature in predicting phishing URLs.

We see that out of the ten most important features, the directory length carries the most importance out of all the features with an importance of 0.10. It is noticeably the most influential feature and significantly surpasses the other features. Notably, "time domain activation" came out to be the second most important feature with a score of 0.06. From the graph, we can conclude that the directory length has the biggest impact on forming the model's predictions

The directory length being the most important feature makes the most sense for a model's prediction. Phishers often make their URLs longer to make it harder for automated systems to discern between fraudulent and credible websites. Additionally, the longer URL allowed phishers to employ different phishing strategies that can not be done with a shorter URL. These strategies include: using IP addresses as URLs for domain concealment, redirecting traffic with @ symbols, or using hexadecimal codes to mask phishing URLs. Knowing that phishing websites employ large URLs, there is a heavy emphasis on URL lengths being less than or equal to 35 characters. Anything over that threshold is most likely to be a phishing website [3]. The Random Forest model made the directory length the most important feature because it can easily look for noticeable patterns within it. It can capture the characteristics of phishing URLs and recognize different patterns associated with phishing emails.

After looking at the most important feature of the Random forest model, we evaluated our different models. We created a summary table of the performances of the Random forest model and our Logistic Regression model. After that, we compared the model's accuracy scores and AUC scores:

TABLE I
MODEL EVALUATION (OUTLIERS DROPPED (ZSCORE > 3))

Model	**Accuracy**	**AUC Score**
Logistic Regression	0.910234	0.902895
Random Forest	0.953425	0.952145

TABLE II
MODEL EVALUATION (IMPUTATION APPLIED TO OUTLIERS (ZSCORE > 3)

Model	**Accuracy**	**AUC Score**
Logistic Regression	0.914707	0.903981
Random Forest	0.967748	0.965736

TABLE III
MODEL EVALUATION (OUTLIERS KEPT)

Model	**Accuracy**	**AUC Score**
Logistic Regression	0.933805	0.928593
Random Forest	0.970986	0.968656

In Table 1, outliers were removed from the dataset. The Random Forest model shows better performance compared to the Logistic Regression model, both in terms of accuracy and AUC score. The accuracy of over 95 percent for Random Forest indicates that it correctly predicts phishing attempts 95.34 percent of the time. The AUC score, which measures the model's ability to distinguish between phishing and non-phishing URLs, is also higher for Random Forest, indicating better overall performance.

In Table 2, imputation was applied to outliers, both models show improved performance. Random Forest's accuracy and AUC score are notably higher than in the previous table, indicating that this model benefits more from the imputation technique. The Random Forest model's accuracy of approximately 96.77 percent and AUC score of 96.57 percent suggest a high level of precision in predicting phishing URLs.

In Table 3, where the outliers were kept in the dataset, both of the models perform at their highest levels compared to the previous approaches. The Random Forest model, in particular, shows a remarkable accuracy of 97.10 percent and an AUC score of 96.87 percent. This suggests that retaining the outliers, rather than removing or imputing them, provides a more realistic and challenging dataset, which the Random Forest model is exceptionally well-equipped to handle.

B. Analysis of Confusion Matrices

To further analyze our results we created the confusion matrices on the next page:

TABLE IV
CONFUSION MATRIX OF LOGISTIC REGRESSION MODEL

	Actual Positive	**Actual Negative**
Predicted Positive	61.8611%	3.5670%
Predicted Negative	3.0523%	31.5193%

TABLE V
CONFUSION MATRIX OF RANDOM FOREST MODEL

	Actual Positive	**Actual Negative**
Predicted Positive	63.8713%	1.5567%
Predicted Negative	1.3447%	33.2273%

The confusion matrices of the Logistic Regression and Random Forest models provide insights into their predictive abilities, particularly regarding false positives and false negatives.

For the Logistic Regression model, the confusion matrix shows that it correctly identifies 61.8611% of the phishing sites (True Positives), it incorrectly labels 3.5670% of legitimate sites as phishing (False Positives). On the other hand, it fails to identify 3.0523% of phishing sites (False Negatives) but correctly labels 31.5193% of the legitimate sites (True Negatives). This suggests that while the model is relatively strong in identifying non-phishing sites, there is room for improvement in reducing false positives.

In contrast, the Random Forest model had a higher True Positive rate of 63.8713% and a lower False Positive rate of 1.5567%. Its False Negative rate stands at 1.3447%, and it correctly identifies 33.2273% of legitimate sites as non-phishing (True Negatives). These figures indicate a more balanced performance, with particular strength in minimizing false alarms (False Positives) while maintaining a high detection rate of phishing sites.

The comparison between the two models shows that Random Forest is more effective in distinguishing phishing sites from legitimate ones. This is seen from its lower rates of False Positives and False Negatives. Logistic Regression, while still competent, shows a slightly higher tendency to misclassify legitimate sites as phishing (False Negatives). This could be critical in practical applications, as no alarm can lead to a potential leak of personal or confidential information.

Our findings show the importance of choosing the right model for phishing detection. The lower false positive rate of the Random Forest model makes it a more suitable choice for environments where falsely labeling legitimate sites as phishing could have significant consequences.

IV. DISCUSSION

A. Challenges and Limitations:

One of the primary challenges was reformatting the data for proper and efficient model training. Our data was skewed and had a significant number of outliers, which is characteristic of many phishing URLs. The high volume of these outliers made approaches such as dropping or imputing impractical without adding a bias.

Further problems occurred from the mixed columns in our data, which consisted of a combination of ratio and categorical data points. We had to separate these combinations without misrepresenting the original data set. This step was important for preserving the accuracy of our analysis. This ensured that our models were trained on data that reflected the real-world scenarios captured in the data set. In addition to the challenges already outlined, we faced the complexity of ensuring model scalability and adaptability. Given the rapid evolution of phishing techniques, our models need constant updates to stay relevant. This evolving nature of cyber threats requires a dynamic approach to model training and maintenance. Our research highlights the need for continuous data collection and model retraining, emphasizing the ongoing commitment required to maintain efficacy in the field of cybersecurity.

B. Implications and applications:

The practical implications of our study are significant. By identifying phishing URLs, organizations and regular people alike can proactively defend against one of the most common cyber threats, potentially saving resources and protecting sensitive data. Our model can be integrated into preexisting cybersecurity systems for businesses where it would offer real-time analysis and alerts for potential attempts at phishing. Additionally, our model can be utilized by educational institutions, which are frequently targeted by phishing attacks. By implementing our model, these institutions can better protect both their networks and their students from phishing scams.

C. Future Direction:

Our future research could explore integrating Natural Language Processing techniques to analyze the content within the URLs for a more comprehensive phishing detection approach where phishing indicators are identified from the text. In addition to NLP we plan to try experimenting with other machine learning techniques and algorithms such as artificial neural networks or the Naive Bayes algorithm to find the most accurate model for our purpose.

The integration of diverse datasets from various geographical could be used to improve the accuracy of our model. This expansion would allow for a stronger understanding of global phishing trends, helping to models that are accurate but also applicable to the rest of the world. Additionally, exploring collaborations with cybersecurity experts could provide valuable knowledge and resources, to further the accuracy of the model.

V. CONCLUSION

Our research aimed at developing an effective model to predict phishing attempts through the analysis of URL structures has yielded promising results. The methodology we employed, involving the preprocessing and analysis of the ”Phishing

Websites Dataset," has demonstrated the potential of machine learning techniques in cybersecurity applications.

The results from our Random Forest and Logistic Regression models, especially the results from the Random Forest model with an accuracy score of 0.97 and an Area Under the Curve score of 0.96, represent the value of sophisticated data analysis tools in identifying phishing URLs. The fact that Random Forest captured the patterns hidden in the data shows the use for the application of machine learning for cybersecurity.

Ultimately, our study has demonstrated the applicability of machine learning in identifying phishing threats through URL analysis. As failing to detect phishing can have a detrimental effect on the lives of the individuals who fall for it, it is important to create a detection model that with almost perfect accuracy. While our results are promising, it's important to acknowledge the continuous nature of research and development in this field.

VI. Contribution

A. Logan Warren

My contributions were primarily in the coding portions and write-ups. I took part in creating the visualizations and training the two models in our project using linear regression and random forest. I trained and scaled the data using data Xavier transformed. I tested a few different models, such as SVM and Naive Bayes but found them to be nowhere near as accurate as the original two. I made multiple visualizations throughout our project to showcase each set visually and wrote out explanations for each.

I also contributed heavily to the final report. I worked on writing out our Methodology, Introduction, and Abstract. While also incorporating the visualizations into our final writeup.

B. Xavier Bear

My contributions to this research project were particularly in the data handling and model development phases. I was primarily responsible for the coding tasks that involved sorting and organizing the large dataset. This process was crucial for ensuring the accuracy and efficiency of the subsequent analysis.

In terms of the model development, I worked on adjusting the models created by Logan to test what the best outlier technique was. This not only simple adjustments to the original model generation but also a testing phase where different techniques for handling outliers were applied. My efforts were aimed at optimizing the models' performance to ensure the most accurate predictions possible while using the proper techniques.

Beyond the technical aspects, I contributed significantly to conceptualizing the project. This included the initial stages of formulating the research question and objectives, as well as drafting the introduction, methodology, and discussion sections of the paper alongside my group members.

In terms of the paper, I worked to create many of the tables and visualizations as well as write the majority of the discussion session.

C. Charles Tran

For the final project, I helped the team pick the topic and dataset that would be the main focus of our project. Additionally, I researched phishing websites to get a better understanding of the terminology and how phishing websites are detected. This helped guide the path of our project and gave us better insight into how we should handle the data. I helped draft the final paper and talked about our findings and conclusions from our data. I also helped in creating some of the visualizations of the data. I helped to create some different graphs that gave a clear visual of the different patterns in our data.

References

[1] G. Vrbančič, "Phishing Websites Dataset", Mendeley Data, V1, 2020. [Online]. Available: https://data.mendeley.com/datasets/72ptz43s9v/1. doi: 10.17632/72ptz43s9v.1

[2] G. Vrbančič, I. Fister, and V. Podgorelec, "Datasets for phishing websites detection," Data in Brief, vol. 33, pp. 106438, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920313202. doi: https://doi.org/10.1016/j.dib.2020.106438.

[3] Alkhalil Z, Hewage C, Nawaf L and Khan I "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy". Front. Comput. Sci. 3:563060. doi: 10.3389/fcomp.2021.563060 https://www.frontiersin.org/articles/10.3389/fcomp.2021.563060/full