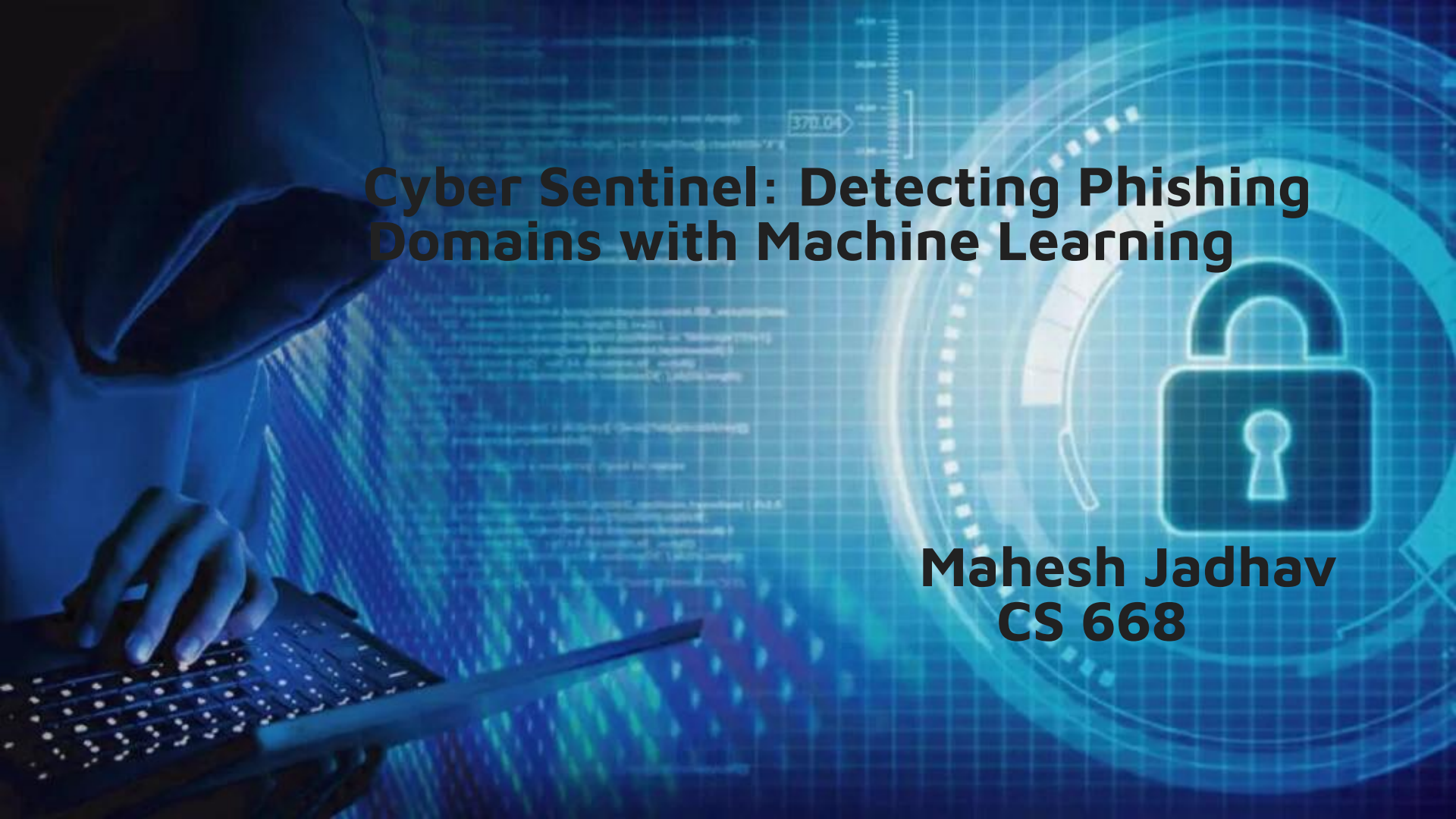


Cyber Sentinel: Detecting Phishing Domains with Machine Learning

Mahesh Jadhav
CS 668

Project proposal

- Targeted Problem
  - Phishing is a form of fraud in which an attacker pretends to be a legitimate business or individual in order to get sensitive data via email or other communication channels, such as login credentials or account information. Attackers frequently use phishing because it is simpler to trick someone into clicking a malicious link that looks legitimate than it is to get past a computer's security mechanisms.
  - Predicting whether the domains are malicious or real is the major objective.
- Research Question
  - what factors contribute most significantly to the accurate detection of phishing websites?

- DataSet
  - https://data.mendeley.com/datasets/72ptz43s9v/1
  - These data consist of a collection of legitimate as well as phishing website records. Each website is represented by the set of features which denote, whether website is legitimate or not.
  - Total number of entries: 88,647
  - Total number of features: 111
- Motivation
  - Project Motivation: To enhance online security and protect against phishing attacks.
  - Technical Motivation: Leveraging machine learning pattern recognition for accurate detection.
  - Personal Motivation: Committed to contributing to a safer digital environment.

Literature Review

Detection of Phishing Website using Machine Learning

- Goal:
  - Detect and classify phishing websites using machine learning.
- Dataset:
  - Utilized the Phishing Websites Data Set with 11,055 observations and 32 variables from the UCI Machine Learning Repository.
  - Link: https://ieeexplore.ieee.org/document/9617482
- Methodology:
  - Employed three machine learning algorithms: Decision Tree, KNN, and Random Forest.
  - Data was divided into an 80% training set and a 20% testing set.
  - Feature selection using the 'Boruta' package.
- Results:
  - First Experiment (Decision Tree):* Achieved an accuracy of 91.16%.
  - Second Experiment (KNN):* Attained the highest accuracy of 97.6% but concerns of overfitting.
  - Third Experiment (Random Forest):* Achieved 94.44% accuracy and had the lowest false-negative rate.
- Limitations:
  - While KNN achieved high accuracy, it raised concerns about overfitting.
  - The research does not include a comprehensive evaluation of overfitting and model complexity.
  - Other potential issues, like class imbalance, have not been explored.

Phishing websites detection

- Goal:
  - To predict whether domains are malicious or legitimate in order to enhance online security and protect against phishing attacks.

- Dataset:
  - Dataset provided by Vrbančič, Fister Jr, and Podgorelec (2019).
  - wo variations: 58,645 and 88,647 websites labeled as legitimate or phishing.
  - 111 features, including URL properties, URL resolving metrics, and external services.
  - Link: https://www.sciencedirect.com/science/article/pii/S2352340920313202#section-cited-by

- Methodology:
  - Machine learning and data mining techniques for classification.
  - Application in building phishing detection systems.

- Results:
  - Valuable resource for developing classification models and systems.
  - Benchmark for evaluating state-of-the-art machine learning methods in phishing website classification.

- Limitations:
  - Did not explore Deep Learning Models

PhishAri: Automatic real time phishing detection on twitter

- Goal:
  - Investigate and understand the current state of research on phishing detection in Twitter.
- Dataset:
  - The literature review is based on a wide range of academic and industry sources that discuss phishing attacks, detection methods, and Twitter security.
  - Link: https://ieeexplore.ieee.org/document/6489521
- Methodology:
  - Review and summarize key findings from existing research, organizing them into thematic sections.
  - Analyze the strengths and limitations of prior studies in the field.
- Results:
  - Identified trends and gaps in the existing literature.
  - Found that phishing detection methods have evolved, with a growing focus on machine learning.
  - Recognized a need for more research in the context of social media, particularly Twitter.
- Limitations:
  - Reliance on the availability and relevance of existing literature.
  - Possible bias in the selection of sources.
  - Limited to the scope and quality of previous research.

Phishing URL Detection Using URL Ranking

- **Goal:**
  - Enhance online security by accurately detecting and classifying phishing websites.
  - Develop a predictive model to determine the legitimacy of domains, contributing to user protection.
- **Dataset:**
  - Utilize a dataset comprising 88,647 website records, including both legitimate and phishing websites.
  - Facilitate machine learning for pattern recognition and accurate detection.
  - Link: https://ieeexplore.ieee.org/document/7207281
- **Methodology:**
  - Leverage lexical and host-based features inspired by "URL Classification and Categorization for Phishing Detection."
  - Incorporate clustering, classification, and URL ranking mechanisms.
  - Extend the existing system with URL clustering, categorization, and ranking.

- **Results:**
  - Achieve an impressive 98.46% accuracy rate in distinguishing between legitimate and malicious domains.
  - Enhance model performance with novel features (e.g., bigrams and cluster labels).
  - Provide meaningful user feedback to enhance understanding of URL risk.
- **Limitations:**
  - Model performance may be influenced by the quality of training data due to the short-lived nature of phishing campaigns.
  - Performance may vary over time as phishing techniques evolve, necessitating regular model updates.

Detecting Phishing Websites Using Machine Learning

- **Goal:**
  - Enhance detection of phishing attacks
  - Develop accurate phishing website identification
- **Dataset:**
  - Sourced from Kaggle
  - 11,430 URLs
  - 1:1 ratio of legitimate and phishing websites
  - Rich feature set (89 attributes)
  - Link: https://ieeexplore.ieee.org/document/9799917

- **Methodology:**
  - Machine learning techniques and algorithms
  - SVM, XGBoost, NB, KNN, AdaBoost, Gradient Boosting, DT
  - Comprehensive evaluation of algorithm effectiveness

- **Results:**
  - XGBoost outperforms
  - Test accuracy: 96.4%
  - Recall: 96.3%
  - Precision: 96.5%
  - Minimizes false positives and false negatives

- **Limitations:**
  - Doesn't explore deep learning
- **Future work**:
  - Larger datasets and deep learning exploration

Project proposal

Targeted Problem
- Phishing is a form of fraud in which an attacker pretends to be a legitimate business or individual in order to get sensitive data via email or other communication channels, such as login credentials or account information. Attackers frequently use phishing because it is simpler to trick someone into clicking a malicious link that looks legitimate than it is to get past a computer's security mechanisms.
- Predicting whether the domains are malicious or real is the major objective.

Research Question
- what factors contribute most significantly to the accurate detection of phishing websites?

DataSet
- https://data.mendeley.com/datasets/72ptz43s9v/1
- These data consist of a collection of legitimate as well as phishing website records. Each website is represented by the set of features which denote, whether website is legitimate or not.
- Total number of entries: 88,647
- Total number of features: 111

Motivation
- Project Motivation: To enhance online security and protect against phishing attacks.
- Technical Motivation: Leveraging machine learning's pattern recognition for accurate detection.
- Personal Motivation: Committed to contributing to a safer digital environment.

Machine learning based phishing detection from URLs

- **Goal**:
  - The paper aims to develop a real-time anti-phishing system using machine learning algorithms and natural language processing (NLP) based features to detect phishing URLs.

- **Dataset**:
  - A new dataset constructed with phishing and legitimate URLs
  - Link:https://www.sciencedirect.com/science/article/abs/pii/S0957417418306067

- **Methodology**:
  - Seven classification algorithms are employed
  - Use of NLP-based features for analyzing URL semantics

- **Results**:
  - Random Forest algorithm using NLP-based features achieves 97.98% accuracy
  - Demonstrates the system's effectiveness in real-time detection

- **Limitations**:
  - Evolving nature of phishing attacks poses a challenge
  - Emphasis on the need for hybrid detection models

Phishing URL Detection Using Machine Learning: A Survey

- **Goal:**
  - Enhance online security and protect against phishing attacks.
- **Dataset:**
  - Link: https://ieeexplore.ieee.org/document/10074337

- **Methodology:**
  - Phishing Detection Techniques
    - Whitelisting and Blacklisting
    - Heuristic Approaches (URL-based and Content-based)
  - Mitigation of Phishing Attack
    - Offensive Defense
    - Correction
    - Prevention
  - Machine Learning for Phishing Detection
    - Deep Learning
    - Support Vector Machine
    - Random Forest
- **Results:**
  - Varied accuracy levels achieved by different approaches:
  - Whitelisting and Blacklisting
  - Heuristic Approaches (URL-based and Content-based)
  - Offensive defense tools like BogusBiter and Humboldt
  - Correction strategies and their impact
  - Machine learning techniques' effectiveness
- **Limitations:**
  - The challenge of user awareness in combating phishing
  - Variability in accuracy levels of different detection techniques
  - Practical challenges with offensive defense

Mid-Semester Presentation

Cyber Sentinel: Detecting Phishing Domains with Machine Learning

Mahesh Jadhav

GitHub:https://github.com/mahesh15698/My_Capstone_Project

Mail_id : mj35806n@pace.edu

Research question

- **Question**
  - what factors contribute most significantly to the accurate detection of phishing websites?

- **Targeted Problem**
  - Phishing is a form of fraud in which an attacker pretends to be a legitimate business or individual in order to get sensitive data via email or other communication channels, such as login credentials or account information. Attackers frequently use phishing because it is simpler to trick someone into clicking a malicious link that looks legitimate than it is to get past a computer's security mechanisms. Predicting whether the domains are malicious or real is the major objective.

Literature Review

- Key Findings:
  - Phishing Detection Methods: Various methods exist, with a growing emphasis on machine learning.

- Positioning Our Work
  - Leveraging Research: Project will build on the methodologies and findings from these studies:
    - Phishing Website Detection
    - Detecting Phishing Websites Using Machine Learning
  - Addressing Gaps: Focusing on addressing limitations, including overfitting, model complexity.

- Identified Needs:
  - Deeper Evaluation: Robust evaluation of overfitting and model complexity.
  - Exploration Ensemble: Exploring Ensemble Algorithms for enhanced performance.

- Significance:
  - Enhancing Cybersecurity: My work contributes to the fight against phishing, a critical aspect of online security.

Dataset

- Dataset Link: https://data.mendeley.com/datasets/72ptz43s9v/1
- Description :
  - These data consist of a collection of legitimate as well as phishing website records. Each website is represented by the set of features which denote, whether website is legitimate or not.
  - Total number of entries: 88,647, Total number of features: 111
- Parameters:

```
data.head()
```

✓ 0.0s Python

e_url	...	qty_ip_resolved	qty_nameservers	qty_mx_servers	ttl_hostname	tls_ssl_certificate	qty_redirects	url_google_index	domain_google_index	url_shortened	phishing
0	...	1	2	0	892	0	0	0	0	0	1
0	...	1	2	1	9540	1	0	0	0	0	1
0	...	1	2	3	589	1	0	0	0	0	0
0	...	1	2	0	292	1	0	0	0	0	1
0	...	1	2	1	3597	0	1	0	0	0	0

EDA & Methodology

Preliminary EDA results:

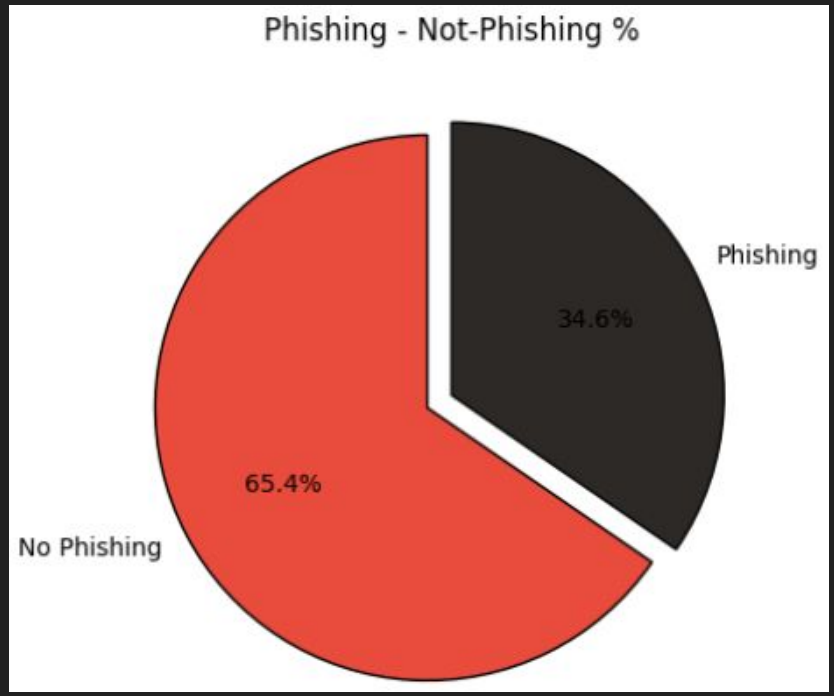
1.No Null Values(Dataset is Cleaned):

```
data.isnull().sum()
```

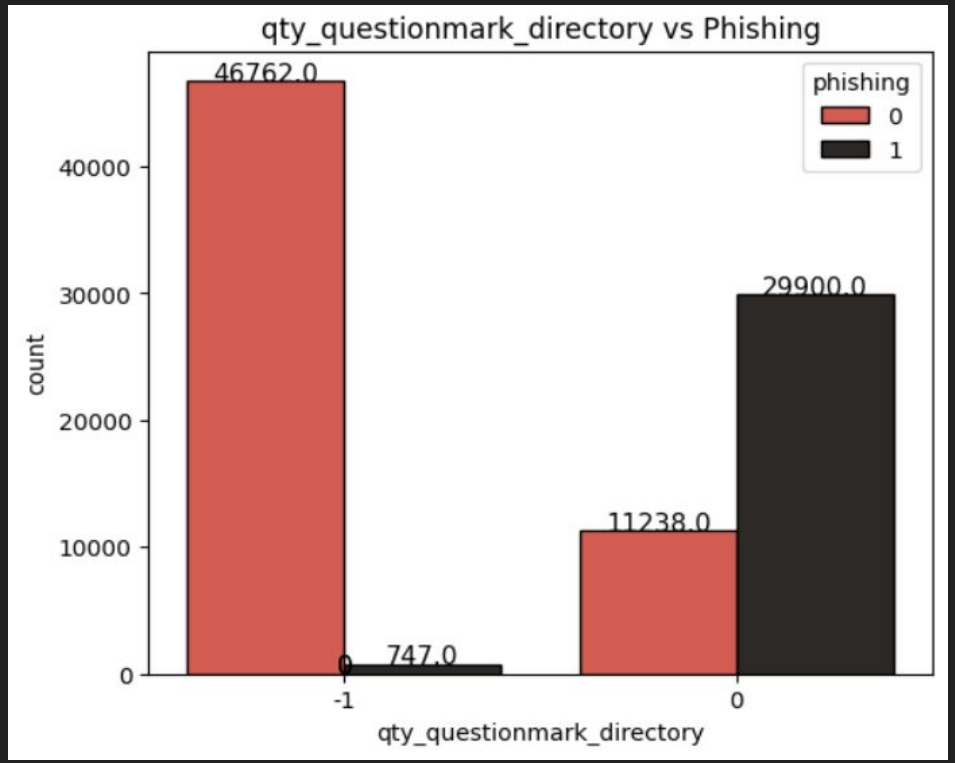
✓ 0.0s

```
qty_dot_url              0
qty_hyphen_url           0
qty_underline_url        0
qty_slash_url            0
qty_questionmark_url     0
                        ..
qty_redirects            0
url_google_index         0
domain_google_index      0
url_shortened            0
phishing                 0
Length: 112, dtype: int64
```

2. Unbalanced Dataset:



Categorical Feature VS Target Variable



- **Total 46 Categorical Variables and 65 Numerical Variables**
- **In Fig,If Question mark Dictionary is 0 then it has high chances of detecting Phishing.**
- **There are more Categorical and Numerical Features those are significant for Detection of Phishing.**

Next Step

1.Check Outliers of Numerical Features Using Box Plot

2. Data Scaling(Standardization and Normalization if Needed)

3. Feature Engineering

4. Feature Selection

5.Train Test Split

6.Training and Evaluation Machine Learning Ensemble Techniques

7.Hyperparameter Tuning to get more accurate prediction

References

- Website:https://www.sciencedirect.com/science/article/pii/S2352340920313202#bib001
  - Paper: Phishing websites detection

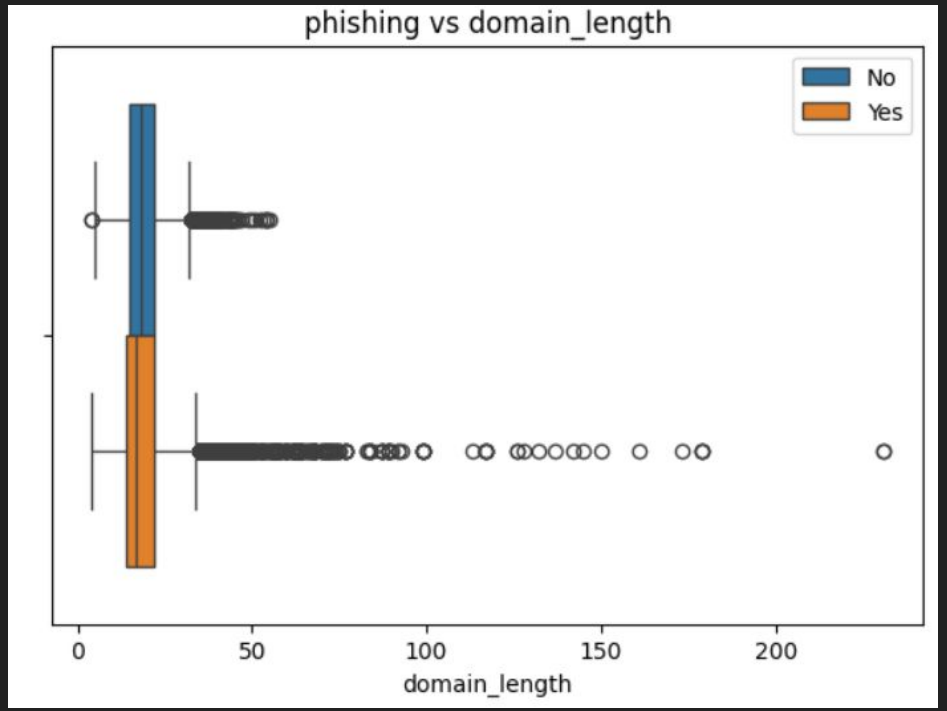
- Website: https://ieeexplore.ieee.org/document/9617482
  - Paper: Detection of Phishing Website using Machine Learning

- Website: https://ieeexplore.ieee.org/document/9799917
  - Paper: Detecting Phishing Websites Using Machine Learning

Methodology And Experimentation

Check Outlier:

- Since the Dataset has huge number of Features. Not able to depict each every column's outlier over here.
- However, Huge number outlier found in quantitative features.



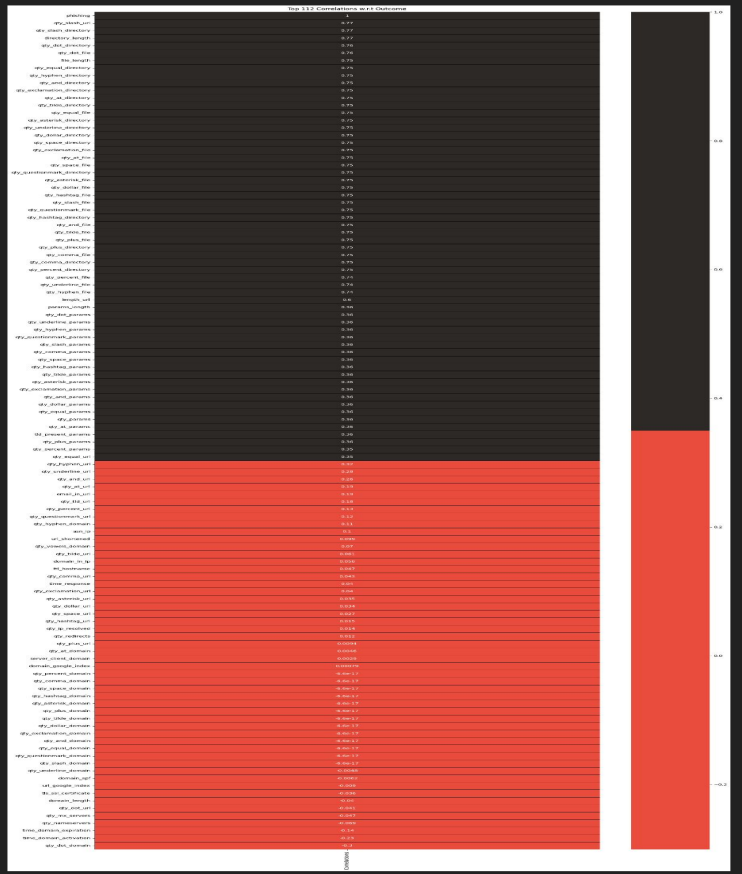
Data Scaling:

ttl_hostname	tls_ssl_certificate	qty_redirects	url_google_index	domain_google_index	url_shortened	phishing
-0.152391	-5.199338	-0.316330	-0.002509	-0.001255	-5.199338	1
0.602124	5.199338	-0.316330	-0.002509	-0.001255	-5.199338	1
-0.322932	5.199338	-0.316330	-0.002509	-0.001255	-5.199338	0
-0.664287	5.199338	-0.316330	-0.002509	-0.001255	-5.199338	1
0.240671	-5.199338	0.855287	-0.002509	-0.001255	-5.199338	0

- Data Scaling is essential for this data set as some features contains very high values as compare to the Other features.
- As per EDA insights, Dataset contains huge outliers and also some features not normally distributed over the scale. It is necessary to normalize.
- To resolve this problem,used Quantile Transformer, which can normalize the features distribution and handle the outliers in the dataset without changing their actual meaning.

Correlation matrix of Whole Dataset:

- This correlation matrix showed valuable insights related target variable.
- Parameters that showed in Black color are highly correlated to Phishing column.
- Rest variables are less correlated to phishing parameters.
- [Check Github to see more clear diagram]



Feature Engineering

- Performed Feature selection step, I am going drop the column those are contributing less than 20% to predict the Target Variable.

- Around 38 features are dropped which are not Contributing to the target variable(Phishing)

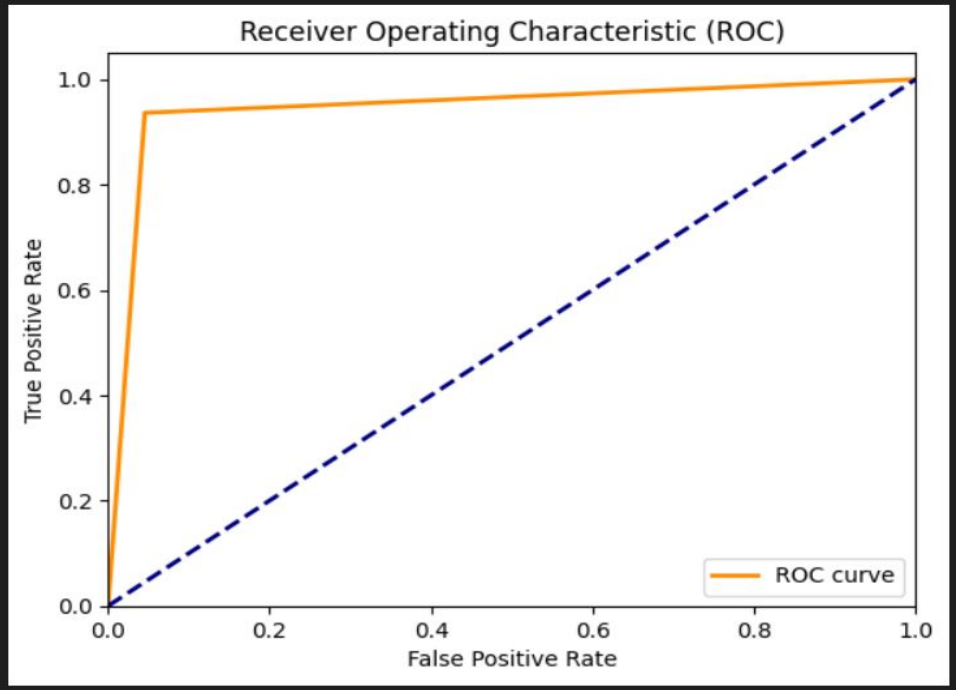
- This Dimensionality reduction step will increase Model Performance on test Data.
- In the modeling step,we are trying to detect phishing domain using these important features.

Overview of Modeling

- Goal: Detect Phishing using selected Features Data.
- Split the whole dataset into 80:20 train test split
  - Train Size = 70917
  - Test Size = 17730
- Experimented with Decision Tree and Ensemble Techniques to detect the Phishing and selected best model by performance metrics
- Applied Stratified K-Fold cross validation. Repeats Stratified K-Fold n times with different randomization in each repetition.
- Performance Metrics:
  - The ROC curve illustrates the tradeoff between true positive rate and false positive rate across varying classification thresholds.
  - The confusion matrix assesses the performance of a classification model by summarizing true positive, true negative, false positive, and false negative predictions

Experiment 1:Boosting Algorithms(XGBoost)

```
Cross Validation Score: 98.69%
ROC_AUC Score: 94.54%
```

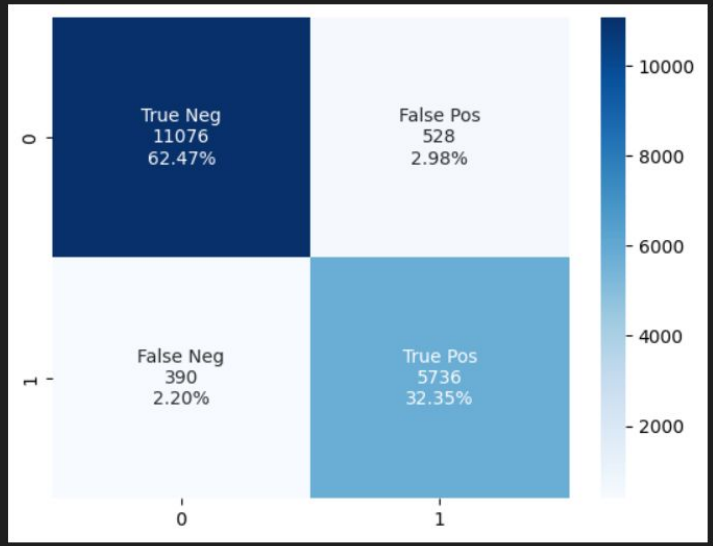


- Cross Validation Score : 98.69%
- ROC_AUC_Score: 94.54%

Model Evaluation:XGBoost

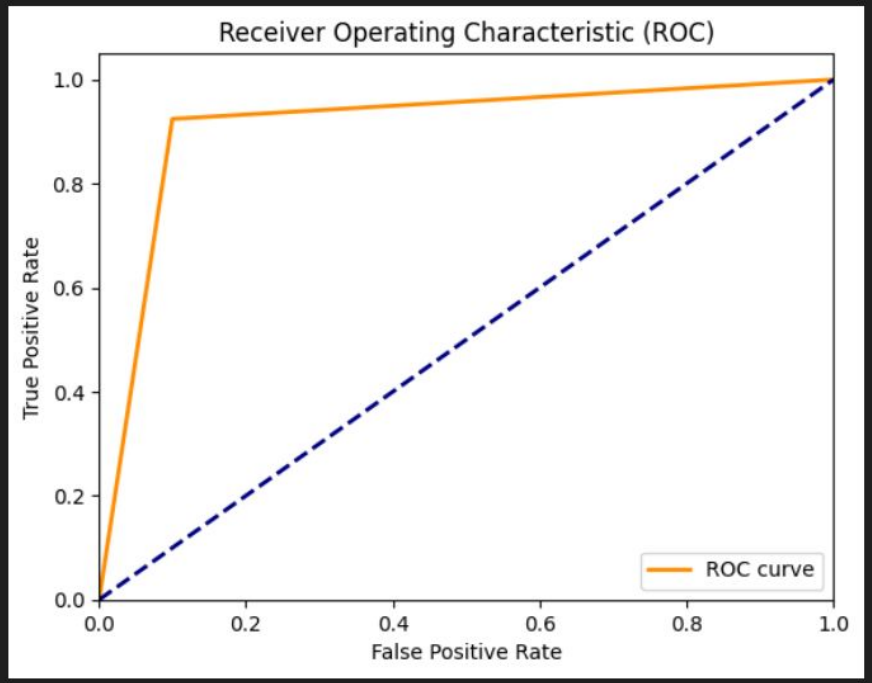
- XGBoost exhibited high accuracy, precision, and recall, indicating robustness in distinguishing phishing websites.

	precision	recall	f1-score	support
0	0.97	0.95	0.96	11604
1	0.92	0.94	0.93	6126
accuracy			0.95	17730
macro avg	0.94	0.95	0.94	17730
weighted avg	0.95	0.95	0.95	17730



Experiment 2: Bagging Algorithm(Random Forest)

```
Cross Validation Score: 97.35%
ROC_AUC Score: 91.23%
```

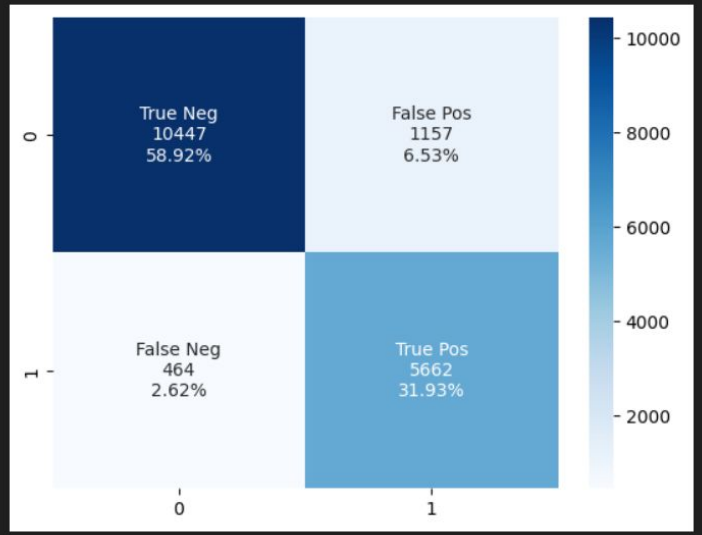


- Cross Validation Score : 97.35%
- ROC_AUC_Score: 91.23%

Model Evaluation:Random Forest

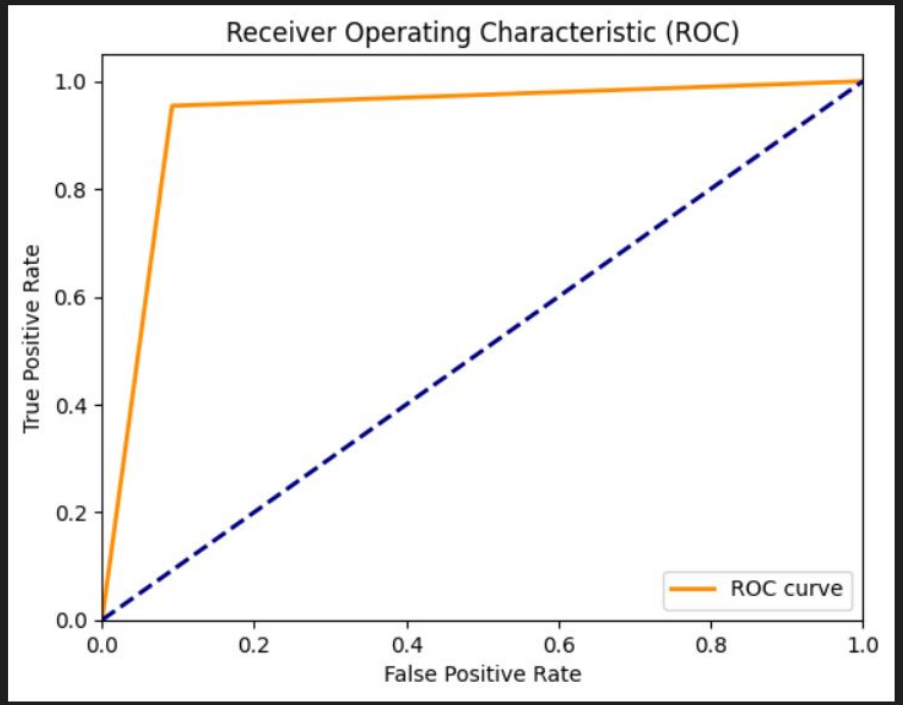
- Random Forest demonstrated good accuracy but slightly lower precision and recall compared to XGBoost.

	precision	recall	f1-score	support
0	0.96	0.90	0.93	11604
1	0.83	0.92	0.87	6126
accuracy			0.91	17730
macro avg	0.89	0.91	0.90	17730
weighted avg	0.91	0.91	0.91	17730



Experiment 3: Decision Tree

```
Cross Validation Score: 97.10%
ROC_AUC Score: 93.11%
```

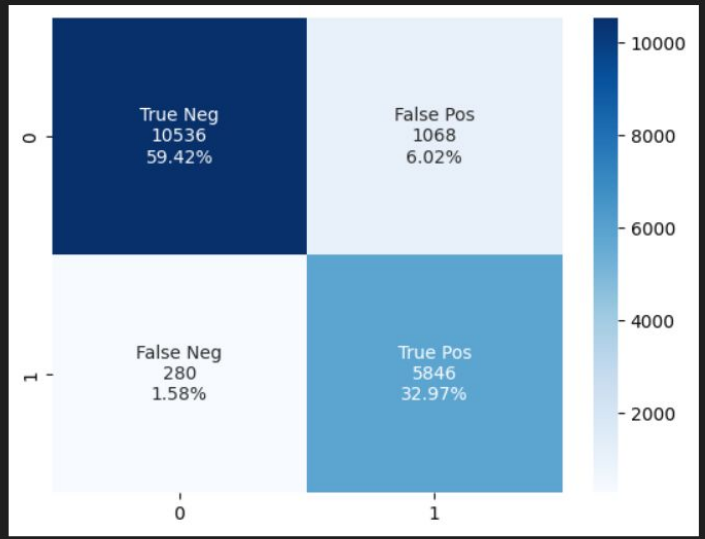


- Cross Validation Score : 97.10%
- ROC_AUC_Score: 93.11%

Model Evaluation:Random Forest

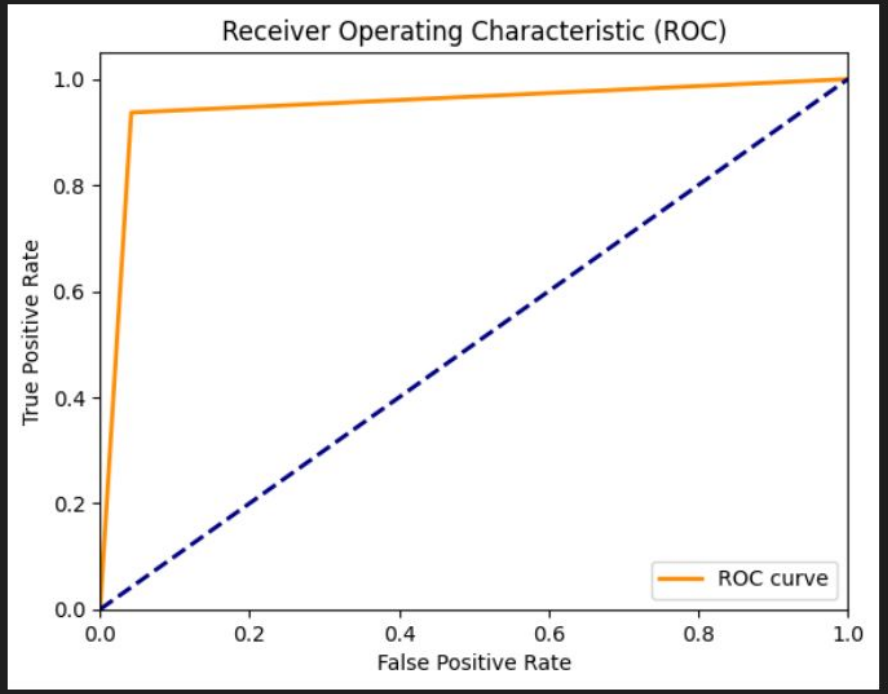
- Decision Tree showed strong recall for phishing websites but comparatively lower precision.

	precision	recall	f1-score	support
0	0.97	0.91	0.94	11604
1	0.85	0.95	0.90	6126
accuracy			0.92	17730
macro avg	0.91	0.93	0.92	17730
weighted avg	0.93	0.92	0.92	17730



Experiment 4: Stacking

```
Cross Validation Score: 98.77%
ROC_AUC Score: 94.76%
```

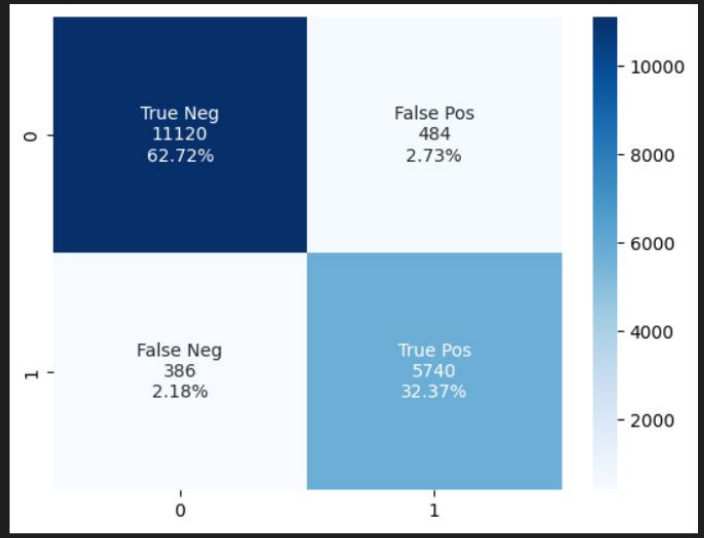


- Cross Validation Score : 98.77%
- ROC_AUC_Score: 94.76%

Model Evaluation:Stacking

- Stacking ensemble method produced high accuracy and balanced precision and recall, leveraging the strengths of individual models.

	precision	recall	f1-score	support
0	0.97	0.96	0.96	11604
1	0.92	0.94	0.93	6126
accuracy			0.95	17730
macro avg	0.94	0.95	0.95	17730
weighted avg	0.95	0.95	0.95	17730



Key Insights of Modeling:

- XGBoost demonstrated superior performance with high accuracy, precision, and recall in identifying phishing websites.

- Ensemble techniques, especially stacking, showcased enhanced performance by combining multiple models(Xgboost,Random Forest,Decision Forest).

Research Question Answer:

"What factors contribute most significantly to the accurate detection of phishing websites?"

- Key Findings:
  - Robust feature selection techniques identified crucial website characteristics vital in distinguishing between legitimate and phishing websites.
  - Ensemble methods, notably stacking, showcased improved performance by leveraging the strengths of individual models for enhanced accuracy.
  - The capability of models to identify and prioritize relevant features significantly impacted their effectiveness in discerning phishing websites.
- Insights:
  - Understanding essential website characteristics plays a pivotal role in accurate phishing detection.
  - Feature selection and ensemble techniques are critical in fortifying cybersecurity measures against phishing attacks.

Conclusion:

This research underscores the efficacy of machine learning techniques in fortifying cyber security against phishing attacks, contributing to a safer online environment. Through meticulous feature selection and leveraging advanced algorithms, critical factors instrumental in accurate phishing website detection were identified. The study showcased the superiority of XGBoost in distinguishing phishing websites, while ensemble techniques, particularly stacking, demonstrated enhanced performance. These findings lay the groundwork for developing more effective phishing detection systems. Continuous adaptation and updates are crucial to addressing evolving phishing tactics and maintaining cyber security effectiveness.

Appendix:

GitHub: Review Code and Graphs