**Detecting Phishing URL’s**

Business Problem Scenario

Our company, at the request of its employees, has just allowed access to personal emails. Originally, we have prevented access to personal emails to prevent outside cyber-attacks and malicious content. However, now that access has been granted, we want to put a team together to help determine if URL links are phishing or legitimate URLs. Phishing attacks are a growing threat, and our companies’ emails spam filter system has been doing a good job of preventing spam emails with phishing URL’s from making it through to our employees. But with access to personal emails, the potential for phishing URL’s goes up. Our job is to create a model that will determine if a URL is phishing or a legitimate URL. Our key stakeholders will not only be our employees for their safety, but also our cyber security team, networking team, and upper management.

The dataset that I chose to solve this problem is from UC Irvine Machine Learning Repository. It contains a total of 235,795 URLs. With those URL’s, 134,850 are legitimate and 100,945 are phishing. This dataset is universally applicable to all companies and beneficial to anyone who uses a computer. Phishing attempts are on the rise. From emails to text messages, they are more common and are an easy way for malicious data to be downloaded if not caught properly.

Since this is about preventing phishing attacks success will be measured by the recall of the model. The more phishing emails that slip through, the more trouble the company will be in. So a better recall value, even if it means some legitimate URLs don’t make it through is what is most important for the safety of the company. We will also use other evaluation metrics, but right now the main priority is preventing as many phishing URLs as possible.

Problem Solving Process

1. Data Acquisition and Understanding

* The data was acquired from a known machine learning dataset website UC Irvine. I searched through many datasets before coming across this one. This hits home for me as I once worked in IT and phishing emails and URLs were a very big subject at the time. I found it interesting.
* Upfront the data seems pretty good, but after stumbling on dataset after dataset, this seemed to be one of decent enough size that has a good amount of variables to use for the project.
* Preliminary I will look through the variables and graph them all out if numeric via a histogram to see what they appear to look like throughout the datafile.

1. Data Preparation and Feature Engineering

* First, I will go through the data, make sure everything is clean and fill in missing values or remove missing values where necessary. Remove any duplicate values from the dataset Then I will see how the different variables are related and what is most impactful towards a phishing URL
* Based on the impact each feature has on the phishing label variable, features will be selected and used for the model. Any categorical data will be encoded to be used for the model as well
* I will use sklearn pipeline to streamline the data before modeling. Scale any data that needs to be scaled, possibly need to use PCA with 55 variables, encoding.

1. Modeling Strategy

* For the model I will be trying different models for this assessment, such as the logistic regression model, random forest model, and XGBoost model. I will use recall as the deciding factor on model performance since malicious URLs are not wanted.
* For cross validation I will be using a k-fold of 5 to maintain class balance.
* I will use GridSearch for hyperparameter tuning to find the parameters that create the best model. If the GridSearch is too expensive and time consuming then RandomSearch will be used.
* The primary evaluation metric will be recall as stated above. Secondary will be accuracy and the f1 score to find a good balance if needed.
* After the models are created I will use feature importance and SHAP to view the importance of each feature, where as before I will likely be using a correlation matrix, this will show the true importance of each feature on the model.

1. Results Interpretation and Communication

* I will take the results of the model and present it to the key stakeholders, in this case cyber security team and upper management to explain to them the model and exactly how well it performs and the benefits it will provide to the company.
* Upon implementation a new presentation would need to be put together for the employees on the benefits and the use of the model to determine malicious URLs.
* To explain the results visuals will be simple and easy to interpret on percentage of what could have been if no model was implemented on what the new results are after implementing the model to explain how successful the implementation would be to users.

1. Conceptual Framework

[Data Acquisition]

↓

[Exploratory Data Analysis]

↓

[Data Cleaning]

↓

[Feature Preparation]

↓

[Model Training]

↓

[Evaluation & Validation]

↓

[Model Tuning]

↓

[Re-Evaluation & Validation]

↓

[Documentation]

↓

[Business Insights & Visualization]

↓

[Deployment Strategy]

Timeline and Scope

* **Dataset finalization and problem formulation – 3 Days**
  + Dataset acquisition and initial exploration
  + Business problem definition refinement
  + Project repository setup
* **Exploratory Data Analysis – 2 Day**
  + Comprehensive data profiling
  + Statistical analysis of relationships
  + Creation of informative visualizations
  + Documentation of insights
* **Data Preprocessing – 1 Day**
  + - Challenges can arise with data preprocessing. Until you’re comfortable with the data and understand what you’re working with it could be difficult to manage
  + Data cleaning implementation
  + Feature engineering
  + Pipeline development
  + Data splitting (train/validation/test)
* **Model Development – 1 Day**
  + Implementation of baseline models
  + Algorithm comparison
  + Hyperparameter tuning
  + Cross-validation
* **Model Evaluation and Refinement – 2 Days**
  + - Fine tuning model can seem problematic and troublesome. To find the best model may take some time and research to see what modifications would be most useful
  + Final model selection
  + Performance evaluation on test data
  + Business metric calculation
  + Interpretation of results
* **Documentation and Reporting – 1 Day**
  + Code commenting and cleanup
  + Technical report writing
  + Executive presentation development
* **Final Review and Submission – 1 Day**
  + - Finding visualizations for non-technical people and just presenting good visualizations that are easily readable can be a challenge to create.
  + Quality assurance
  + Video recording
  + Final submission preparation