# Problem Statement

Phishing is a type of social engineering attack often used to steal user data. Phishing attacks are becoming more and more sophisticated, and our algorithms are suffering to keep up with this level of sophistication. They have low detection rate and high false alarm especially when novel phishing approaches are used. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier. Moreover, a comprehensive blacklist can ensure a perfect up-to-date database. Various other techniques such as page content inspection algorithms have been used to combat the false negatives but as each algorithm uses a different approach, their accuracy varies. Therefore, a combination of the two can increase the accuracy while implementing different error detection methods.

Problem statements:

1. How to process dataset for phishing detection?
2. How to reduce false negative rate in phishing websites algorithm?
3. What are the best combinations of classifiers that can efficiently detect phishing attacks?
4. What data preprocessing techniques are going to be employed such that missing/NULL values and ambiguities in data are eliminated?
5. What train-cross validate-test split is going to be used?
6. What feature selection, standardization and one hot encoding techniques are going to be used to make the features/dimensions model ready?
7. Choosing an appropriate error metric such that it works well for an imbalanced dataset.
8. Model building- What appropriate model would be well suited for this classification? (GLM - linear models or non-linear models like Decision Trees, RandomForest and XGBoost)
9. Finding ideal hyperparameter values and figuring out the best ROC-AUC score and F1- score.
10. Deployment of model in the application and evaluate its performance to real world data