Basic Details of the Team and Problem Statement

PSID: KVH-004

Problem Statement Title: Phishing Detection Solution

Team Name: Narwhal Sentinels

Team Leader Name: Swayam Tejas Padhy

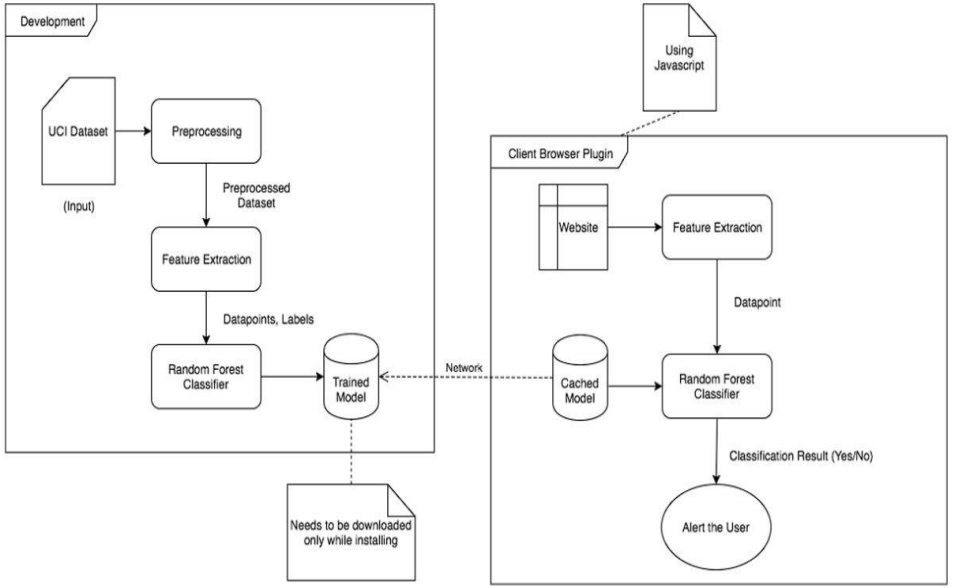
Institute Code (AISHE): C-7252

Institute Name: Manipal Institute of Technology, Manipal

Idea/Approach Details

Project Details

1. Our Project's main ambition is to bring AI powered phishing detection solution to every layman such that they can use this without worrying about their device specifications.
2. This solution is divided into two parts - backend and the frontend that co-exist with each other.
3. The backend will remain in a web server. Here Arff datasets will be preprocessed and features and attributes of the dataset are extracted into a numpy array.This array is then fed into the random forest classifier which in turn creates the trained model.This trained model in the form of a JSON fle is hosted on the server for users to download into their devices.In future iterations we will create a reporting systems in which users can report a phishing site (after which that site's data will be fed into our model).
4. The frontend is a web plugin installed on the user's device . On the first run it will download the cached model from our web server. Then everytime a site opens, it will extract the site's features and then use the model to figure out if it is a phishing site or not. Appropriately it will show a popup on the user's browser.
5. Our model currently has an accuracy score of 94.7401%.We are using a dataset created by UCI which has 30 different attributes and over 2456 instances of data.



Technology stack

- BACKEND-

1. Python with usage of different libraries such as Pandas,numpy and Sklearn
2. Json files to store the prepared model and configuration data

- FRONTEND-

1. Javascript for feature extraction
2. Manifest.json file for chrome plugin configuration

Idea/Approach Details

Use Cases

1. In our humble opinion , we believe that our product has extremely huge business potential.Phishing is an extremely common and extremely damaging cyber attack vector in the common world.We aim on ending that status quo.

2. Machine learning against phishing is a relatively new concept and is extremely resource intensive. Thus models cant be created on simple home devices.As our model is created on our servers, users can just download them and run them irrespective of their devices.Thus getting the power of ML with little to no effort.

3. Now our system can only detect phishing websites but in the future we are planning to expand into text messages,social media messages,emails,and voicecalls.

4. We also plan to have a subscription based model in which there would be three tiers- Free, Small and Enterprise.Free would be for normal users , small for small businessess and enterprise would be for companies with more than 200 employees.Differences between these tiers would be the efficacy of the models and support provided.

Dependencies

1. **Pandas** - Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.
2. **JSON** - The Json library is ued to create the json file of the model and it's configuration.
3. **Numpy** - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
4. **Sklearn** - scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
5. **Dump** - The Dump Document Library Object (DMPDLO) command is used primarily for problem analysis. It copies the contents and attributes of folders, documents, or internal document library system objects to a spooled printer file named QPSRVDMP.
6. **Random tree classification** - Random Forest Trees (RFT) is a machine learning algorithm based on decision trees. Random Trees (RT) belong to a class of machine learning algorithms which does ensemble classification. The term ensemble implies a method which makes predictions by averaging over the predictions of several independent base models.

Team Member Details

Sr. No.	Name of Team Member	Branch (Btech/Mtech/PhD etc):	Stream (ECE, CSE etc):	Year	Position in team (Team Leader, Front end Developer, Back end Developer, Full Stack, Data base management etc.)
1	Swayam Tejas Padhy	Btech	CSE	2	Team Leader
2	Anushtha Shalin Choudhary	Btech	CSE	2	Front End Developer
3	Harshit Verma	Btech	CSE	2	Back-end Developer
4	Arin Gupta	Btech	CSE	2	Technical Lead
5	Smayan Bohidar	Btech	CSE	2	Product Tester
6	Ketan Goel	Btech	CSE	2	UI/UX Designer