# Phishing Website Detection: A Machine Learning Approach


**Submitted By,**
Group 5

Siddhi Patil (sp1508)
Madhura Daptardar (md1250)
Aishwarya Srikanth
Mehanaz Mohammed Iqbal(mm2456)

**Master of Science**
**Graduate Program in Electrical and Computer Engineering**
**2018**

**Graduate School - New Brunswick**
**Rutgers, The State University of New Jersey**

## Problem Statement

Phishing is an unlawful activity of creating fake websites to deceive users. The aim of such websites is to acquire confidential information such as username, passwords, account numbers, etc. from people. It is a vital issue especially in e-banking and e-commerce industry taking the number of online transactions involving payments. These websites look very similar to the original websites; users can land on one just because of mistyping the URL as well. Thus detection of phishing websites is a very important safety measure for websites that contain confidential information. Most of the websites today aim to have a robust phishing detection system.

The aim is to use the dataset to classify whether a website is a phishing website or not by applying the best Machine Learning algorithm on the dataset. The result can be further viewed dynamically on a predictive service which will be constructed using the best Machine Learning algorithm for our dataset.

## What are phishing websites?

A phishing website, tries to steal user's account password or other confidential information by tricking the users into believing that they are on a legitimate website. One can even land on a phishing site by mistyping a URL (web address).

***What to look for in a phishing website?***

1. **Poor resolution:** Phishing websites are often poor in quality, since they are created with urgency and have a short lifespan. If the resolution on a logo or in text strikes you as poor, be suspicious.
2. **Forged URL:** Even if a link has a name you recognize somewhere in it, it doesn't mean it links to the real organization. Read URLs from right to left — the real domain is at the end of the URL. Also, websites where it is safe to enter personal information begin with "https" — the "s" stands for secure. If you don't see "https" do not proceed. Look out for URLs that begin with an IP address, such as: http://12.34.56.78/firstgenericbank/account-update/ — these are likely phishes.

## Deliverables

1. Classify whether a website is a phishing website or not
2. Compare 6 Machine Learning techniques to find out which one provides better results.
3. Construct a predictive service using the best algorithm to see the results dynamically.

## Data set

The dataset has 30 features. Types of features are:

1. Address bar based features (12)
2. Abnormal based features (6)
3. HTML and JavaScript based features (5)
4. Domain based features (7)

Some of the features which have been considered to detect phishing websites are:

1. **IP Address**: If IP address is used in place of the domain name in the URL, then the users can be sure that someone is trying to steal their personal information.
2. **URL Length**: Phishing websites tend to have a long URL to hide doubt. If the URL length is more than 53 characters, then it can be classified as phishing website.
3. **URLs having "@" symbol**: Everything before the @ symbol in the URL can be ignored and the real address follows the @ symbol. Therefore URLs with "@" symbol have a high probability of being phishing websites.
4. **URLs with "//" symbol**: "//" is used to redirect the website.This often can be used to redirect the user to a phishing website.
5. **Adding prefix or suffix separated by '-' in the URL.**
6. **HTTPS** (HyperText Transfer Protocol with Secure Sockets Layer).

7. **Submitting information to email** : A web form allows a user to submit details to server for processing but a phishing website will redirect personal information to his/her personal email.
8. **Right Click Disabled** : Phishing websites use JavaScript to disable right click so that the user cannot save the web page source code.
9. **Use of pop up windows** : If pop up window contains text fields, the website is a fraudulent one else it has more chances of being legitimate.

## Approach

We built a machine learning model in Azure Machine Learning Studio and studied 6 algorithms to determine which one is more reliable in predicting if a website is phishing or not.

## Azure Machine Learning Model

We have implemented our project using Azure Machine Learning and R programming. Microsoft Azure Machine Learning Studio is a collaborative, drag-and-drop tool you can use to build, test, and deploy predictive analytics solutions on your data. Machine Learning Studio publishes models as web services that can easily be consumed by custom apps or BI tools such as Excel.

Machine Learning Studio is where data science, predictive analytics, cloud resources, and your data meet.

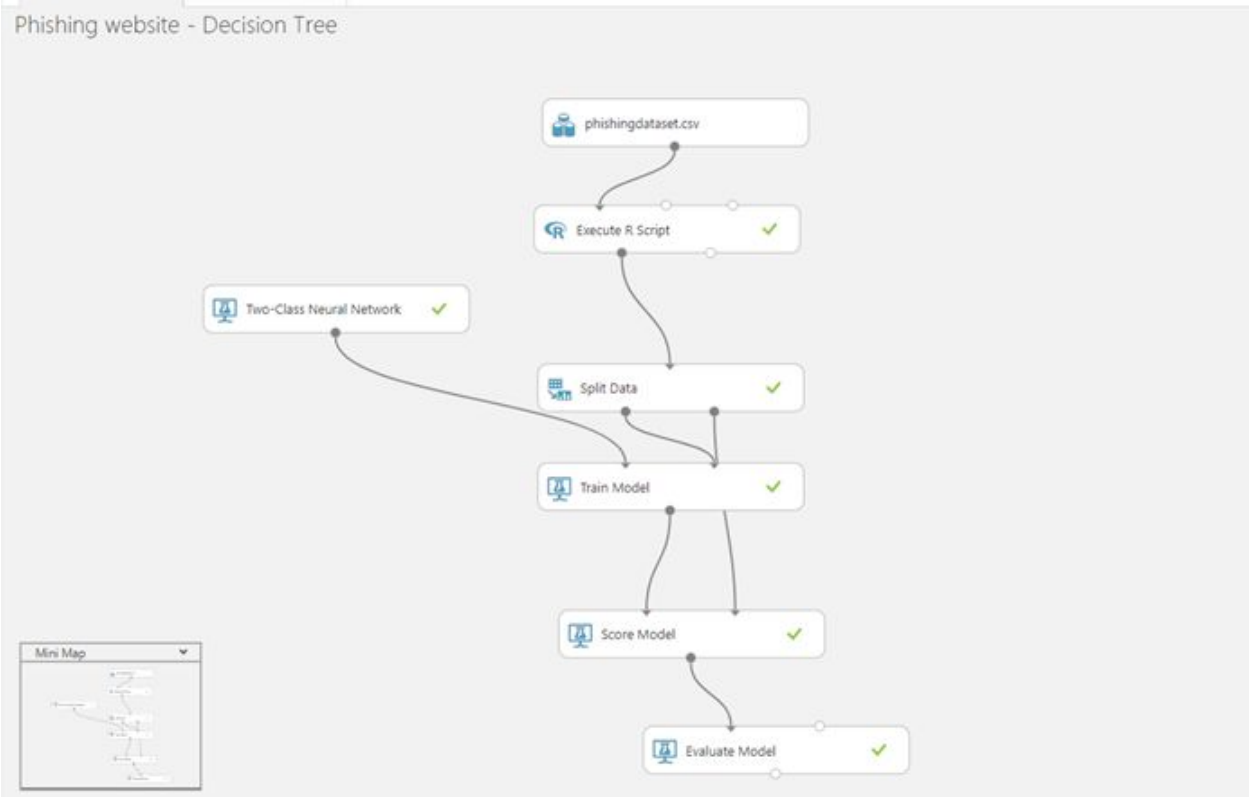


Figure 1: Two-Class Neural Network in Azure Machine Learning Studio

1. As the first step, the dataset is uploaded as a csv file
2. After uploading, the dataset is dragged and dropped in the work area. It will be under Saved Datasets option.
3. The dataset can be visualized by right clicking on it and then clicking on the visualize option.

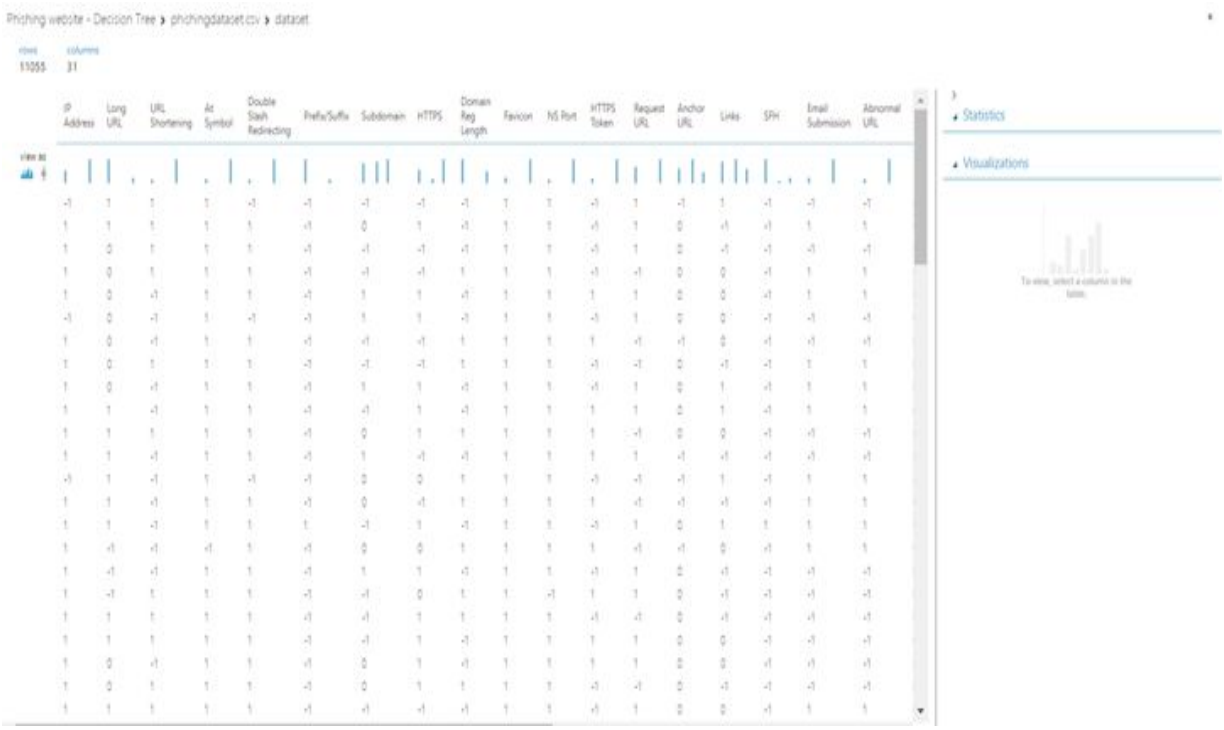


Figure 2: Visual Representation of our dataset

4. All the numeric columns are then converted to factor columns using an **Execute R Script** Module.

***R Script for converting from numeric columns to categorical:***

```
# Map 1-based optional input ports to variables
    mydata<- maml.mapInputPort(1) # class: data.frame
```

```
cols<-c("IP Address","Long URL","URL Shortening","At
Symbol","Double Slash Redirecting","Prefix/Suffix",
"Subdomain","HTTPS","Domain Reg Length","Favicon","NS
Port","HTTPS Token","Request URL","Anchor URL","Links","SFH",
"Email Submission","Abnormal URL","Redirect","onMouseOver",
"RightClick","PopUp","Iframe","Age of domain","DNS Record",
"Website Traffic","Page Rank","Google Index","Links Pointing Page","Stat
Reports","Result")
mydata[cols] <- lapply(mydata[cols], factor)
maml.mapOutputPort("mydata");
```

5. The dataset is then split into training and testing set. 75% of the data goes into training set and 25% goes into testing set.
6. We then train the model using the "Result" column as label and the algorithm. In the above diagram, we have used **Two Class Neural Network** Algorithm. If we want to train using any other algorithm, the process is simple. We just have to delete the Two Class Neural Networks module and replace it with any other module.
7. The model is then scored using **Score Model** module.
8. The model is finally evaluated using **Evaluate Model** module.

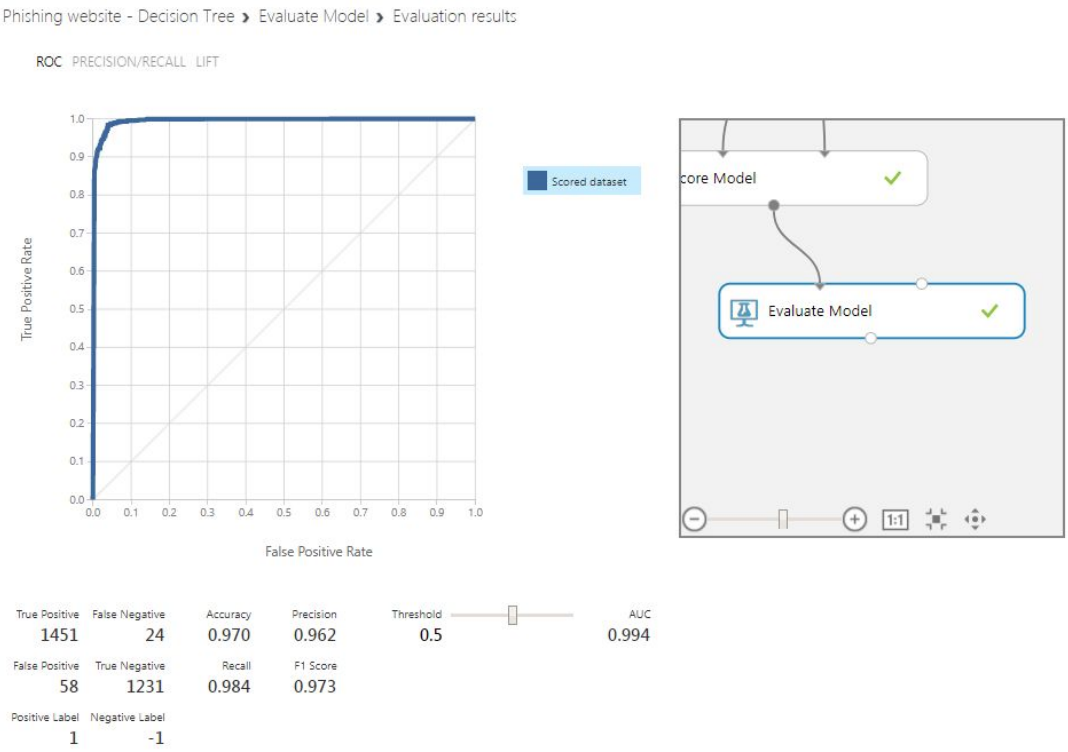


Figure 3: Evaluation result

# Algorithms

1. **Two-Class Logistic Regression:**

   Logistic Regression is a predictive analysis method in statistics. The algorithm predicts the probability of the occurrence of the event by fitting the data to a logistic function. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. [8]

In this model of Logistic Regression, the classification algorithm is optimized for dichotomous (nominal variables which have only two categories or levels) or binary variables.

To train this model a data set containing a class column was provided. Since our data set had only 2 values for the class column (i.e. 1 if the website is a phishing website, or 0 if it is not), we choose Two-Class models.[7]

2. **Two-Class Decision Forest:**

***What is a Decision Tree?***

A decision tree is a map of possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs or probabilities. A decision tree typically starts with a single node and branches out into the possible outcomes. [9]

Decision forest is an ensemble learning method for classification problems, that operates by constructing a multitude of decision trees at the training time and outputting the mode of the classes or the mean prediction of the individual trees. With decision forests you can get better results and a more generalized model by creating multiple related decision trees and combining them. [7]

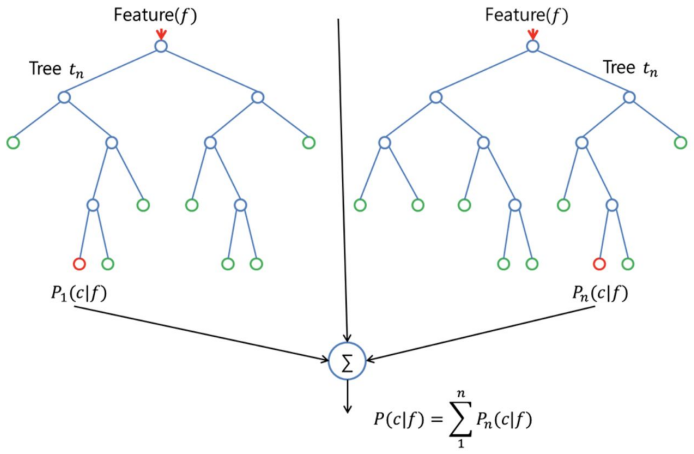


Figure 4: Decision Forest [10]

Instead of searching for the best feature while splitting a node, the decision forest searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model. [10]

## 3. Two-Class Boosted Decision Tree:

### *What is a Decision Tree?*

A decision tree is a map of possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs or probabilities. A decision tree typically starts with a single node and branches out into the possible outcomes. [9]

Just like Decision Forest, Boosted Decision Tree is also a ensemble learning method for classification, which constructs a multitude of decision trees. But unlike in Decision Forest, this model uses boosting to improve the results.

In boosted decision tree, the second tree corrects the errors of the first, the third tree corrects the errors of the second and the first, and so on. Predictions are based on the entire ensemble of trees together. [7]

4. **Two-Class Bayes Point Machine:**

Bayes Point Machine is a type of Bayesian Classifier. Bayesian Classifier is a probabilistic model that can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayes Point Machine uses a Bayesian approach to linear classification.

This algorithm efficiently approximates the theoretically optimal Bayesian average of linear classifiers by choosing one average classifier "the Bayes Point". Since the Bayes Point Machine is a Bayesian classifier, it is not prone to overfitting of the training data. [7] Another added advantage is that, Two-Class Bayes Point Machine does not require the data to be normalized.

5. **Two-Class Support Vector Machine**

The support vector machine is a supervised learning model that requires labeled dataset. The algorithms analyses the training data and recognizes patterns in a multi-dimensional feature space called hyperplane.[7] In simple words, the support vector machine take the input data and generates an optimal hyperplane ( which can also be called a decision boundary),

which best separates the different categories. The best hyperplane is one that maximizes the margins from both categories. In other words, a hyperplane whose' distance to the nearest data point in each categories is largest is the optimal hyperplane.

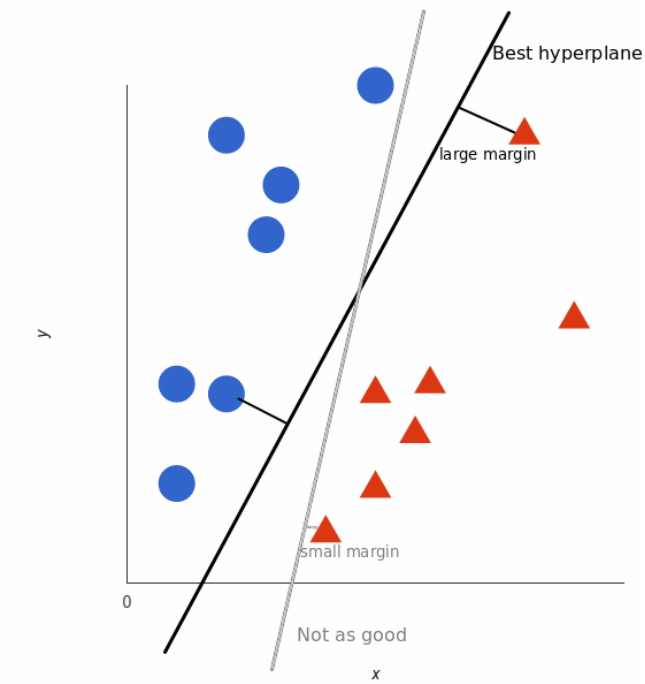


Figure 5: Best hyperplane is the one is black color [11]

The hyperplane depends on the the dimension of the feature space, and the type of kernel used. For example, in a two dimensional feature space the hyperplane is a line.

## 6. Two-Class Neural Network

A artificial neural network is set of interconnected layers of nodes called neurons. A neural network model consists of a input layer which is the first layer and an output layer which is the final most layer, connected by an

acyclic graph. Between the input and the output layer there can be multiple hidden layers. The input data has weights.

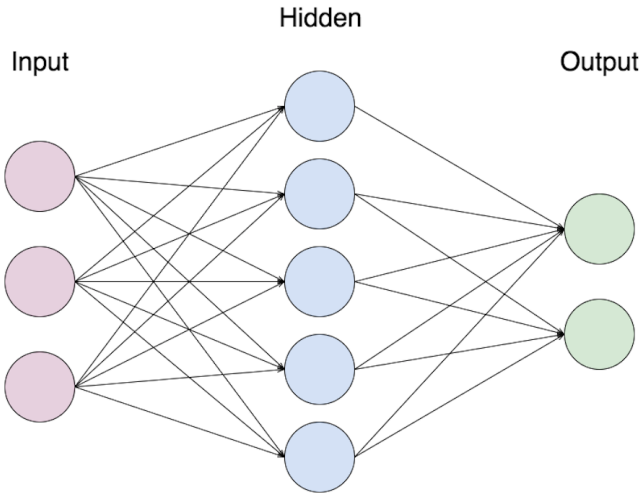


Figure 6: Artificial Neural Network [12]

The relationship between the inputs and outputs is learned from training the neural network on the input data. To compute the output, a value is calculated at each node in the hidden layer and the output layer. The value is set by calculating the weighted sum of the values of the nodes from the previous layer, followed by application of an activation function. [7]

## Terminologies and Concepts

### 1. Confusion Matrix

In the field of machine learning, confusion matrix also known as error matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

| | Predicted (Yes) | Predicted (No) |
|---|---|---|
| Actual (Yes) | TP | FN |
| Actual (No) | FP | TN |

Table 1: Confusion Matrix

The basic terms are defined below.

- True Positive (TP): The Observation is positive, and is predicted to be positive.
- False Negative (FN): The Observation is positive, but is predicted negative.
- True Negative (TN): The Observation is negative, and is predicted to be negative.

- False Positive (FP): The Observation is negative, but is predicted positive.

**Accuracy**

The accuracy is the ratio of the correctly predicted observations to the total observations. Accuracy is a great measure only when we have symmetric datasets where values of False Positive is equal to that of the False Negative.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

**Precision**

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate.

Precision = TP/(TP+FP)

**Recall (Sensitivity)**

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/(TP+FN)

**F1 score**

F1 Score is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account. F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best

if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, then it's better to look at precision and F1 score.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

## Analysis

In the previous section, different machine learning terminologies were introduced. In this section, we will describe which parameters we had chosen to be important for our model and the reason for choosing those parameters.

Precision is more important than recall when you would like to have less False Positives in trade off to have more False Negatives. Meaning, getting a False Positive is very costly, and a False Negative is not as much. In our case, recall is more important than precision because getting a False Negative is costly whereas getting a False Positive isn't as important.

For example, if a website is not a phishing website but it has been predicted to be a phishing website, it is just a false alarm. There will not be much damage due to the prediction. It will just make us exercise some additional caution.

On the other hand, if a website is actually a phishing website but it has been predicted as a legitimate website, it would lead to enormous damage. When people enter their personal details like their name, phone number or even worse, their credit card details, it would affect the user's life big time. That is why we feel that for this dataset Recall is more important than precision.

We have also considered F-1 Score to be an important parameter. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an

uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, the cost of false negative is more compared to a false positive. Therefore, we have decided to consider F-1 Score over accuracy.

## Result

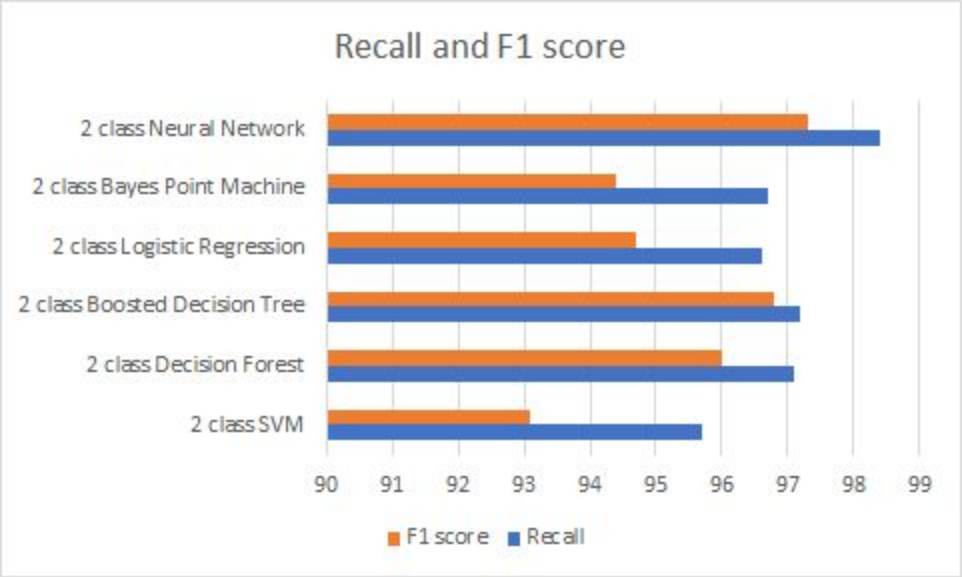


Figure 7: Recall and F1 Score

- As seen in the Figure 7, Two-Class Neural Network performs the best in terms of both the Recall as well as the F1 Score.
- There is not much significant difference in the Recall of Two-class Decision Forest and Two-Class Boosted Decision Tree, but the F1 score of Two-Class Boosted Decision Tree is significantly greater than Two-Class Decision Forest.
- Two-Class SVM, surprising performs the worst in terms of Recall as well as F1 score.

## Predictive Web Service

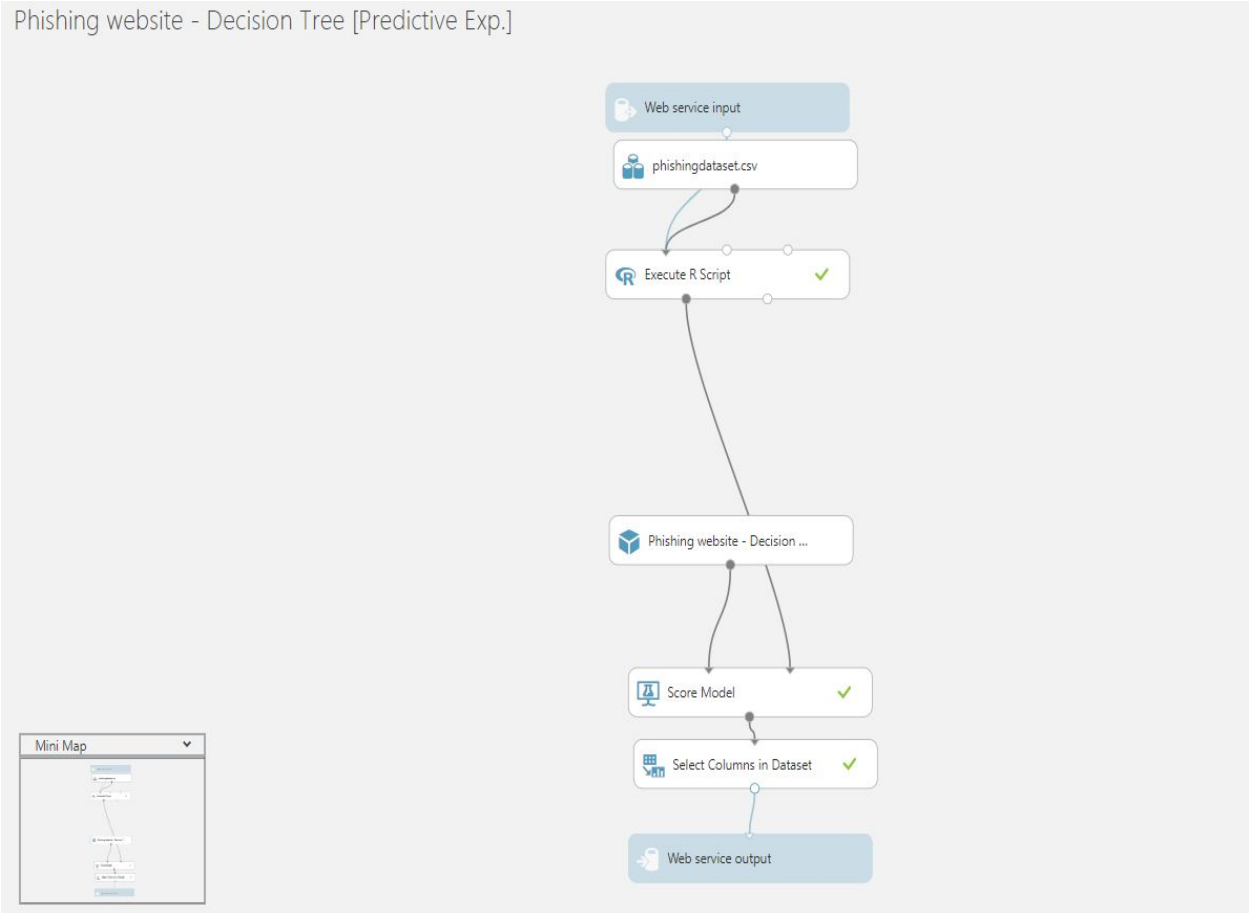


Figure 8: Predictive Web Service Model in Azure ML Studio

We have also built a predictive web service in Azure where the user can dynamically use our Machine Learning Model.

For using our web predictive service, the user has to

1. Open Excel Online
2. Go to Insert Menu -> Office Add-Ins and then add Azure Machine Learning.
3. Then click “Add Web Service”.

4. Enter API key and URL in the space provided.

   **API Key:**

   OS/AHuzZ4FsArwG6Fmo4daGmvx2KoiPGklh2j57I9XzysEltpJUXLmCJD EuFz0htSsuQHiYj5v0OVDt+I1masg==

   **URL Key:**

   https://studio.azureml.net/apihelp/workspaces/02ef4b27ba794b5eac02b495b6da3275/webservices/8d06fa0acd5e4ea4b3b16e18584b8c9a/endpoints/11680744ed5a4ca78fddb85484c69e58/score

5. Then click “Add”.
6. Then click “Use Sample Data”.
7. The service can be tested by copying a few rows from the dataset(except “Result column”.
8. We should then select the input range and a single cell for output and then click “Predict”.

| IP Addn | Long Ur | URL Sho | At Symb | Double | Prefix/S | Subdom | HTTPS | Domain | Favicon | NS Port | HTTPS T | Request | Anchor | Links |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | -1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | 0 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 |
| -1 | 0 | -1 | 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 |
| 1 | 0 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 0 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 |
| 1 | 1 | -1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 1 |

| SFH | Email Sl | Abnorm | Redirec | onMou | RightCli | PopUp | Iframe | Age of d | DNS Re | Website | Page Ra | Google | Links Pc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 |
| -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 |
| -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 0 |
| -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 |
| -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 |
| -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 |
| -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 |
| -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 0 |
| -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 0 |
| -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 |

| Stat Rep | Result | Scored |
|---|---|---|
| -1 | -1 | -1 |
| 1 | -1 | -1 |
| -1 | -1 | -1 |
| 1 | -1 | -1 |
| 1 | 1 | 1 |
| -1 | 1 | 1 |
| -1 | -1 | -1 |
| 1 | -1 | -1 |
| 1 | 1 | 1 |
| 1 | -1 | -1 |

Figure 9: Result of predictive web service in Excel Online

## Way to avoid phishing scams

Phishing scams exists ever since the inception of the internet and nobody wants to become its victim. There are several ways by which we can avoid becoming a victim to the phishing scams.

1. **Keep Informed About Phishing Techniques**

   It is very much essential to know about the latest phishing techniques as early as possible. This would prevent us being a prey to the cybercriminals. For IT administrators, ongoing security awareness training and simulated phishing for all users is highly recommended in keeping security top of mind throughout the organization.

2. **Think Before You Click!**

   It is not a good idea to click on the links that appear in random emails and messages. The phishing emails may look like as it is from a legitimate company and it will ask to fill out the information but it will not be addressed to your name. It may start with “Dear Customer” or something like that in general. When in doubt, go directly to the source rather than clicking a potentially dangerous link.

3. **Install an Anti-Phishing Toolbar**

   Most popular Internet browsers can be customized with anti-phishing toolbars. Such toolbars run quick checks on the sites that you are visiting and compare them to lists of known phishing sites. If you stumble upon a malicious site, the toolbar will alert you about it.

4. **Verify a Site's Security**

   It is always a good idea to check the website's URL before submitting any personal sensitive information. It should begin with "https" and there should be a closed lock icon near the address bar. Always check the security certificate of the website. Never download files from the suspicious emails or websites. If you get a message stating a certain website may contain malicious files, do not open the website.

5. **Check Your Online Accounts Regularly**

   If you don't visit an online account for a while, someone could be having a field day with it. Even if you don't technically need to, check in with each of your online accounts on a regular basis. Change your passwords regularly. To prevent bank phishing and credit card phishing scams, you should personally check your statements regularly. Get monthly statements for your financial accounts and check each and every entry carefully to ensure no fraudulent transactions have been made without your knowledge.

6. **Keep Your Browser Up to Date**

   Security patches are released for popular browsers all the time. They are released in response to the security loopholes that phishers and other hackers inevitably discover and exploit. So, whenever an update is available it is necessary to download and install it as soon as possible.

7. **Use Firewalls**

   High-quality firewalls act as buffers between you, your computer and outside intruders. You should use two different kinds: a desktop firewall and a network firewall. The first option is a type of software, and the second

option is a type of hardware. When used together, they drastically reduce the odds of hackers and phishers infiltrating your computer or your network.

8. **Be Wary of Pop-Ups**

   Pop ups are often phishing scams. If we click on any pop ups, chances are very high that it may lead to the phishing websites. Hence as a precaution, we can block pop-ups using the browser capabilities.

9. **Never Give Out Personal Information**

   Never share personal or sensitive financial information over the internet. It is also not advisable to share the sensitive information through emails.

10. **Use Antivirus Software**

    Special signatures that are included with antivirus software guard against known technology workarounds and loopholes. Just be sure to keep your software up to date. Anti-spyware and firewall settings should be used to prevent phishing attacks and users should update the programs regularly. Firewall protection prevents access to malicious files by blocking the attacks. Antivirus software scans every file which comes through the Internet to your computer. It helps to prevent damage to your system.

## REFERENCES


1. http://www.phishing.org/10-ways-to-avoid-phishing-scams
2. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
3. http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
4. https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio
5. https://www.quora.com/When-is-precision-more-important-over-recall?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa
6. https://www.phishtank.com/what_is_phishing.php?view=website
7. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/
8. https://www.statisticssolutions.com/what-is-logistic-regression/
9. https://www.lucidchart.com/pages/decision-tree
10. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
11. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
12. https://blog.webkid.io/neural-networks-in-javascript/