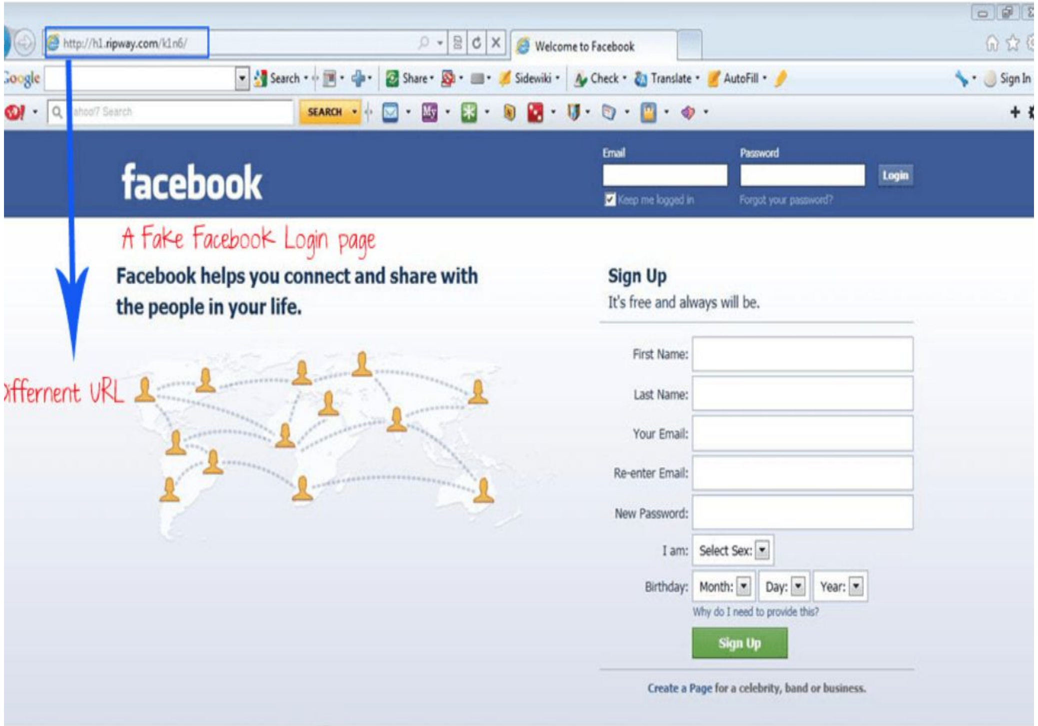
RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

# Phishing Website Detection: A Machine Learning Approach



Facebook helps you connect and share with the people in your life.

Sign Up

It's free and always will be.

Create a Page for a celebrity, band or business.

Presented by,
Siddhi Patil
Madhura Daptardar
Aishwarya Srikanth
Mehanaz Mohammed Iqbal

Date:
04/17/2018

# Phishing

- Phishing is the attempt to obtain sensitive information such as usernames, passwords, and credit card details (and money), often for malicious reasons, by disguising as a trustworthy entity in an electronic communication.
- In short, phishing steals identities and wrecks lives. It affects everyone, from a senior bank manager to a minor who has never heard of Internet scams.
- **Website Phishing** tricks you into believing you are on a legitimate website.

# Goal

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

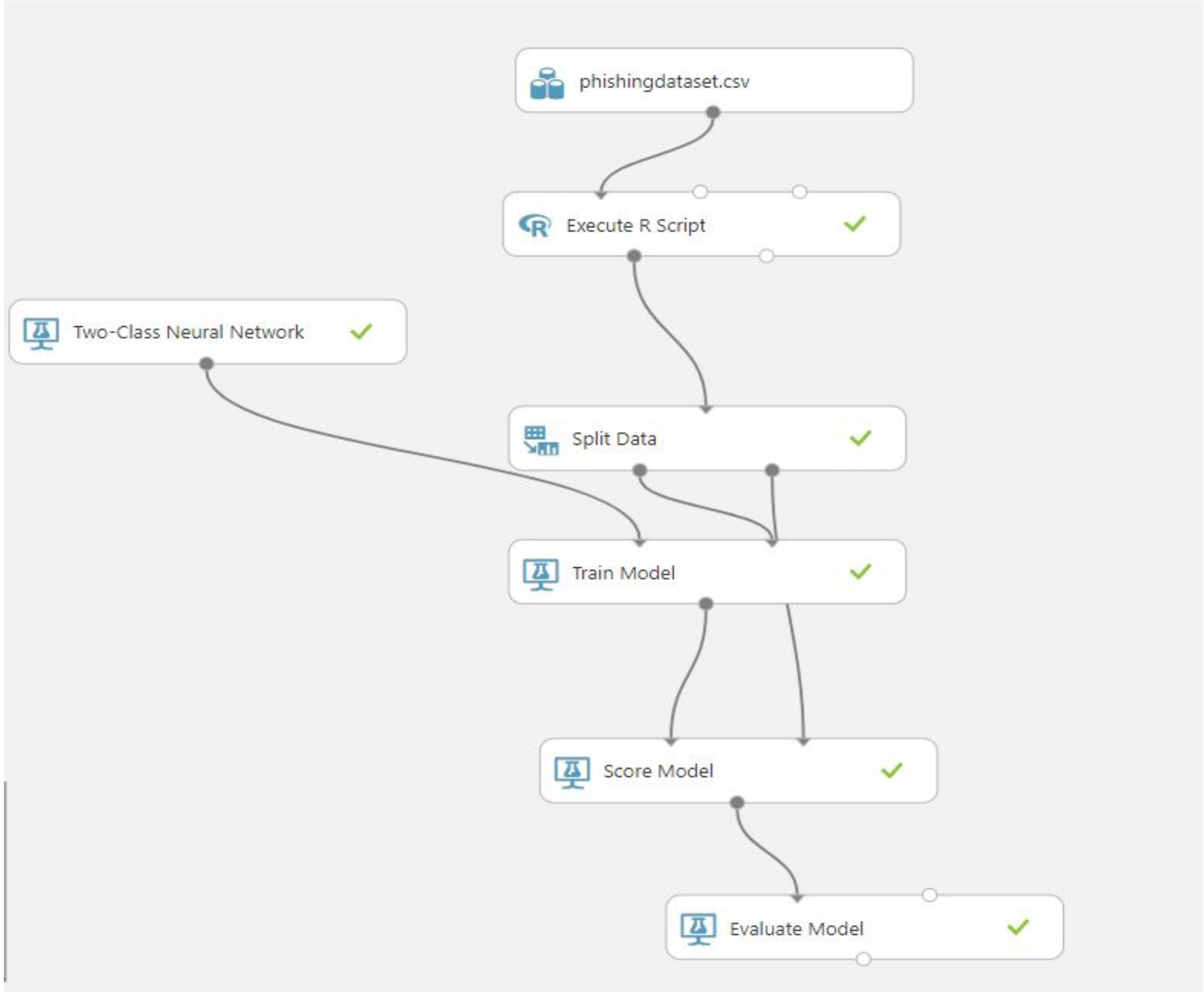
# Dataset

- https://archive.ics.uci.edu/ml/datasets/phishing+websites#
- It has 30 features
- Types of features: Address bar based features (12), Abnormal based features(6), HTML and JavaScript based features (5), Domain based features (7)

# Our Approach

- We built a machine learning model in Azure Machine Learning and used 6 algorithms to predict which one is more reliable in predicting if a website is phishing or not

# Azure ML Model



# Algorithms

1. Two-Class Logistic Regression
2. Two-Class Decision Forest
3. Two-Class Boosted Decision Tree
4. Two-Class Bayes Point Machine
5. Two-Class Support Vector Machine
6. Two-Class Neural Network

# Terminologies and Concepts (1)

- **Confusion matrix**: TP, FP, TN and FN.
- **Accuracy**: The number of correct predictions made by an algorithm. (TP+FP/Total Number of Samples).
- **Precision**: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

  Precision = TP/TP+FP

# Terminologies and Concepts (2)

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

Recall = TP/TP+FN

- **F1 score**: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

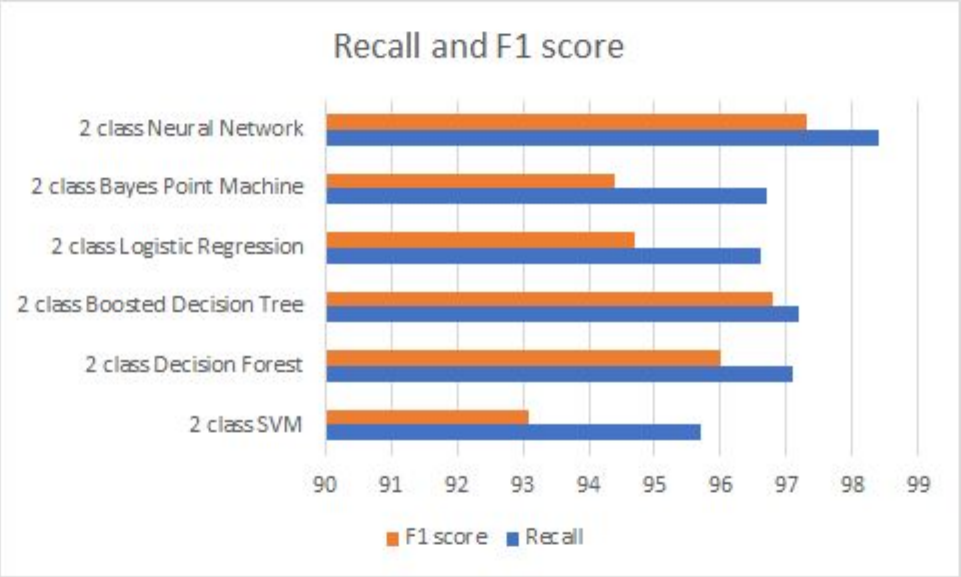
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

# What is more important?

- For our problem and dataset, recall is more important than precision
- F1 score is more important than accuracy
- Why F1 score? False Negative has a higher cost than False Positive

# Result

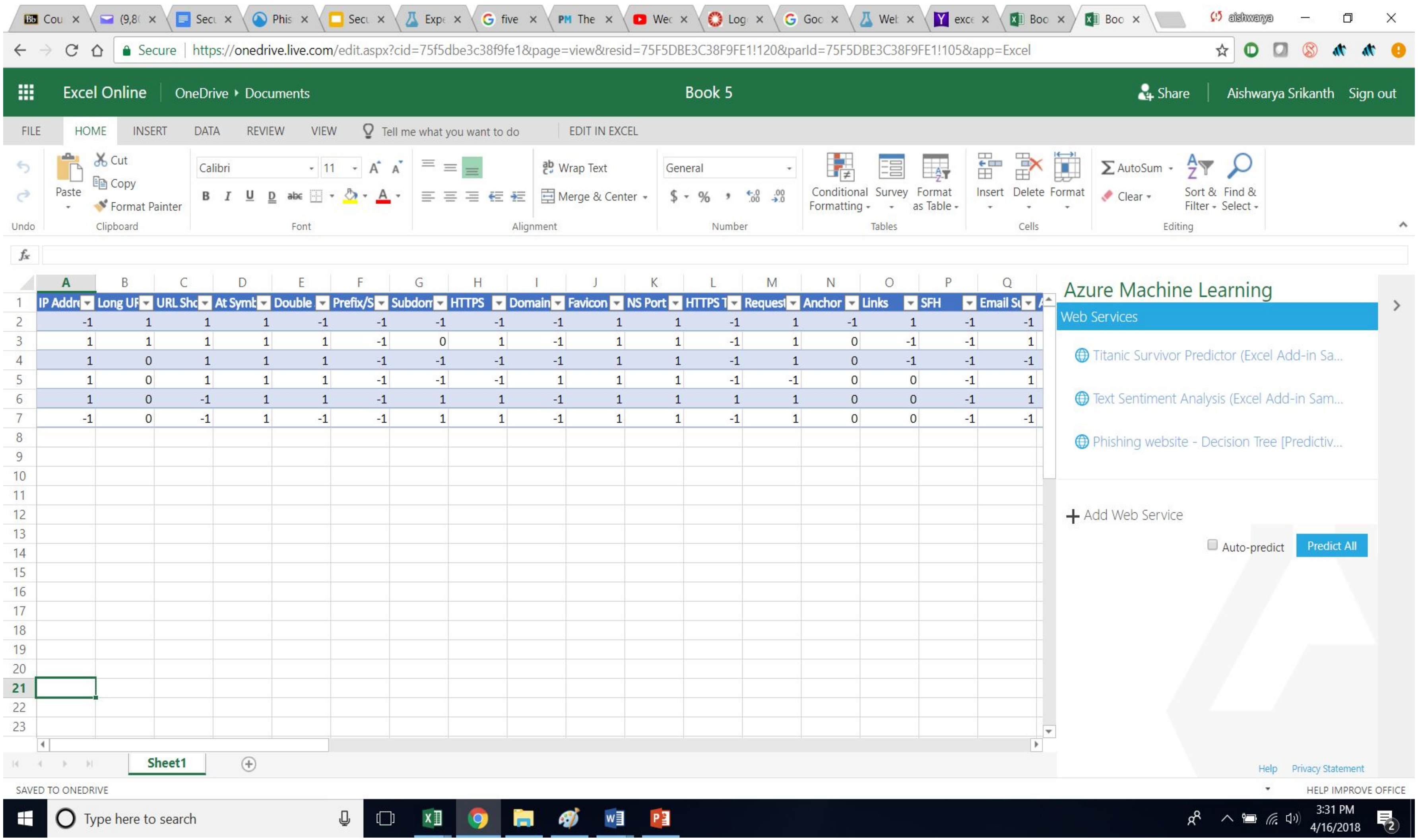
- 2 class Neural Network performs the best in terms of both F1 score and Recall



# Predictive Web Service (1)

- **URL:**
  https://studio.azureml.net/apihelp/workspaces/02ef4b27ba794b5ea
  c02b495b6da3275/webservices/8d06fa0acd5e4ea4b3b16e18584b
  8c9a/endpoints/11680744ed5a4ca78fddb85484c69e58/score
- **API:**
  OS/AHuzZ4FsArwG6Fmo4daGmvx2KoiPGklh2j57I9XzysEltpJUXL
  mCJDEuFz0htSsuQHiYj5v0OVDt+I1masg==

# Predictive Web Service (2)



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IP Addr | Long UF | URL Shc | At Symb | Double | Prefix/S | Subdom | HTTPS | Domain | Favicon | NS Port | HTTPS T | Request | Anchor | Links | SFH | Email Su |
| 2 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | -1 | 1 | -1 | -1 |
| 3 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | 1 |
| 4 | 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 |
| 5 | 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | 0 | -1 | 1 |
| 6 | 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | 1 |
| 7 | -1 | 0 | -1 | 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | -1 |

Azure Machine Learning

Web Services

- Titanic Survivor Predictor (Excel Add-in Sa...
- Text Sentiment Analysis (Excel Add-in Sam...
- Phishing website - Decision Tree [Predictiv...

+ Add Web Service

□ Auto-predict **Predict All**

Help Privacy Statement

# How to avoid phishing scams? (1)

- Be informed about phishing techniques
- Think before you click!!!!
- Install an Anti-Phishing Toolbar
- Verify a site’s security
- Check your online accounts regularly

# How to avoid phishing scams? (2)

- Keep your browser up to date
- Use firewalls
- Be cautious of pop-ups
- Never give out personal information
- Use antivirus software

# References

1. https://en.wikipedia.org/wiki/Phishing
2. http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
3. http://resources.infosecinstitute.com/category/enterprise/phishing/phishing-as-a-risk-damages-from-phishing/#gref
4. https://www.quora.com/When-is-precision-more-important-over-recall?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa
5. https://stats.stackexchange.com/questions/49226/how-to-interpret-f-measure-values?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa
6. http://www.phishing.org/10-ways-to-avoid-phishing-scams