PHISHING OR NOT?

by Sakura Lin

A MODEL THAT IMPROVES INFORMATION SECURITY

Why phishing identification matters ?

Ineffective Firewall

Important Data in
Employees' Computers

Phishing Scams Cost American Businesses

500 MILLION /One year

SOURCE:WWW.FORBES.COM

CURRENT SITUATION

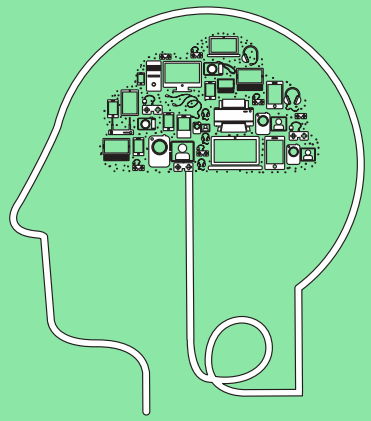
Phishing :
Block the domain

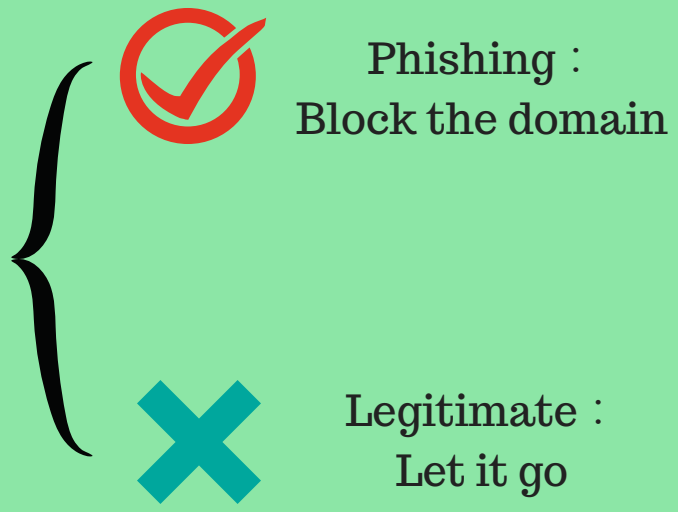
Legitimate :
Let it go

Actual phishing:
90%

20-70 alarms/day

MY SOLUTION





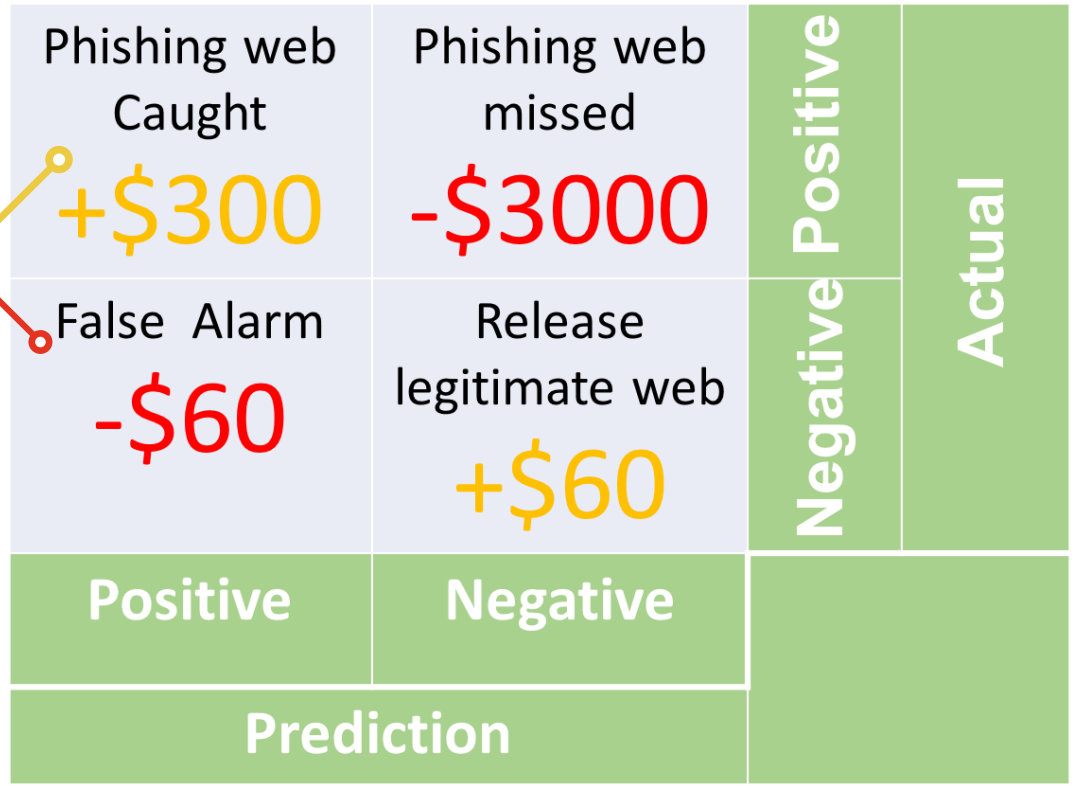
machine learning model

No more alarms!

We don't want to miss one phishing website, even when it means we may get 10 more false alarms.

MY GOAL: MAXIMIZE THE BENEFIT

Cost-Benefit Equation:
Benefit from one observation=
300 * TPR +
-60 * FPR +
-3000 * FNR +
60 * TNR



MODELING PROCESS

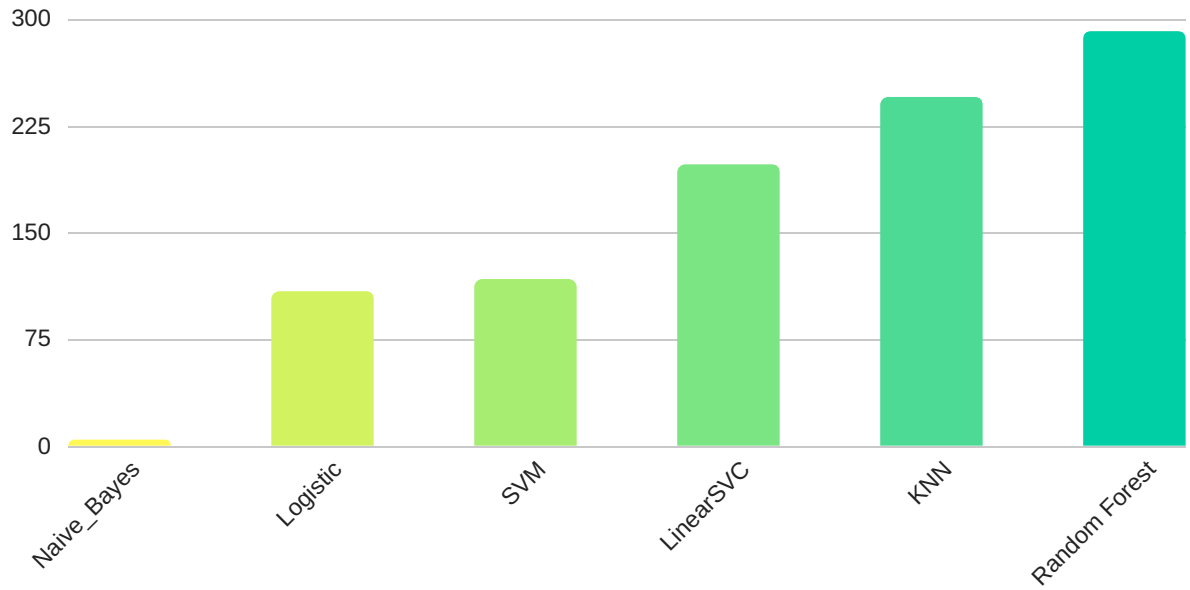
DATA OVERVIEW

- Source: Kaggle
- 11055 Entries
- 30 Features

FINAL MODEL PERFORMANCE

- Benefit:291 dollars
- Number of alarms: 20000
- Total Benefit:
  5,800,000 dollars/one year

MODEL SELECTION



RANDOM FOREST WINS !

INTRODUCE THE MOST COST-EFFECTIVE MODEL

PROJECT DEPLOYMENT

"Wait...this model will need 2-3 months to be deployed, or even longer..."

" Well, let's figure out a plan ! "

INTRODUCE THE MOST COST-EFFECTIVE MODEL

USEFUL & FUN FACTS

1. Sub-domain

"http://www.hud.ac.uk/students/"

vs

"http://www.hud.ac.uk.st.acc/students/"

2. Suspicious symbols

"Amazon/customer"

vs

"Amazon—customer"

1. —

2. @

3. HTTP//....HTTP//....

THANK YOU FOR LISTENING! ^_^

Questions?