domain knowledge

gaurav and suriyan

Why predict the category of content hosted by a domain?

Why predict the category of content hosted by a domain?

- malware and phishing

Why predict the category of content hosted by a domain?

- malware and phishing
- adult content

Why predict the category of content hosted by a domain?

- malware and phishing
- adult content
- personalization

Why predict the category of content hosted by a domain?

- malware and phishing
- adult content
- personalization
  - you can have your *cookie* and eat it too!

But if we were to learn from domain names, how would we?

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.
- meta tags—self reports

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.
- meta tags—self reports
- metadata: how the pages are laid out, how many images, etc.

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.
- meta tags—self reports
- metadata: how the pages are laid out, how many images, etc.
- content understanding: text, images, etc.

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.
- meta tags—self reports
- metadata: how the pages are laid out, how many images, etc.
- content understanding: text, images, etc.
- signal in the url

But if we were to learn from domain names, how would we?

- manual labeling—`PhishTank`, `DMOZ`, etc.
- meta tags—self reports
- metadata: how the pages are laid out, how many images, etc.
- content understanding: text, images, etc.
- signal in the url

Considerations

Considerations

- computational burden, speed

Considerations

- computational burden, speed
  $>$ 2B domains
- privacy

Considerations

- computational burden, speed
  > 2B domains
- privacy
- accuracy

Intuition

Intuition

- Patterns in names.

Intuition

- Patterns in names.

  words like *xxx, porn*, *adult*, *sex* etc. common in domains carrying pornographic content

Intuition

- Patterns in names.
  words like *xxx*, *porn*, *adult*, *sex* etc. common in domains carrying pornographic content

- $>$ 200 `PhishTank` URLs have the word *paypal* in them.

Intuition

- Patterns in names.
  words like *xxx*, *porn*, *adult*, *sex* etc. common in domains carrying pornographic content

- $>$ 200 `PhishTank` URLs have the word *paypal* in them.
  compare with Alexa 1M to build classifiers that can tell those apart

How to Classify Text
- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus
  - Learn the context well

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus
  - Learn the context well
  - Preserve tens of hundreds of vectors and pass them to a model

How to Classify Text

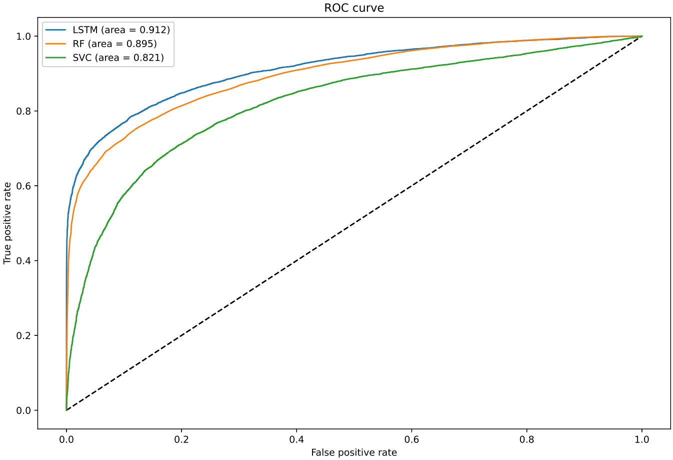
- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus
  - Learn the context well
  - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus
  - Learn the context well
  - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:
  - Embeddings of common bi-chars

How to Classify Text

- Text $\rightsquigarrow$ Embeddings $\rightsquigarrow$ Classifier
  - Embeddings leverage the adage:
    You are the company you keep.
  - Use a large corpus
  - Learn the context well
  - Preserve tens of hundreds of vectors and pass them to a model
- In Our Case:
  - Embeddings of common bi-chars
  - LSTM


ROC curve
LSTM (area = 0.912)
RF (area = 0.895)
SVC (area = 0.821)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

PhishTank

Application

Application

- Are disadvantaged people most at risk online?

Application

- Are disadvantaged people most at risk online?

Application

- Are disadvantaged people most at risk online?

Data

Application

- Are disadvantaged people most at risk online?

Data

- comScore panel

Application

- Are disadvantaged people most at risk online?

Data

- comScore panel
- domain level data for a machine in a household + household attributes

Concerns

Concerns

- Domain $!=$ URL

Concerns

- Domain ! = URL
- Net error is a function of FP - FN.

Concerns

- Domain $! =$ URL
- Net error is a function of FP - FN.
- The larger the number of domains, the larger the error in their score

Concerns

- Domain $!=$ URL
- Net error is a function of FP - FN.
- The larger the number of domains, the larger the error in their score
- Visits are right-skewed. Errors in the right tail are very costly.

Concerns

- Domain ! = URL
- Net error is a function of FP - FN.
- The larger the number of domains, the larger the error in their score
- Visits are right-skewed. Errors in the right tail are very costly.

Solutions

Concerns

- Domain $! =$ URL
- Net error is a function of FP - FN.
- The larger the number of domains, the larger the error in their score
- Visits are right-skewed. Errors in the right tail are very costly.

Solutions

- Calibration $\rightsquigarrow FP = FN$ locally

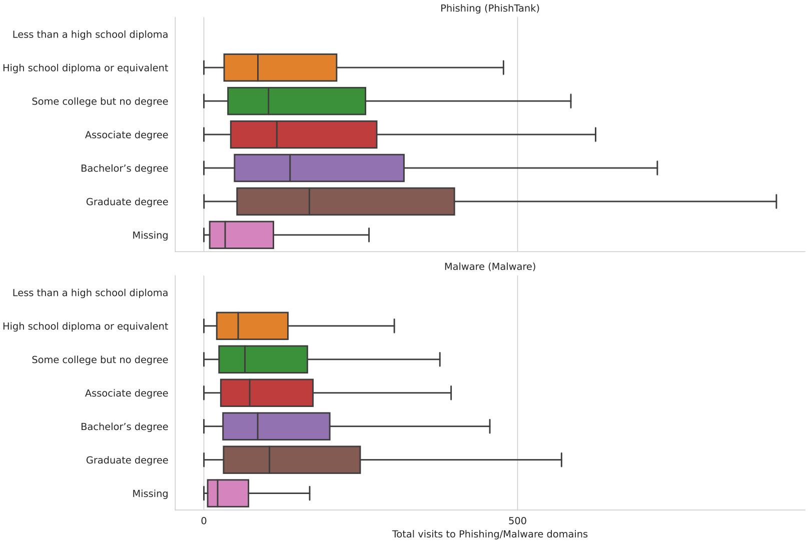
Concerns

- Domain != URL
- Net error is a function of FP - FN.
- The larger the number of domains, the larger the error in their score
- Visits are right-skewed. Errors in the right tail are very costly.

Solutions

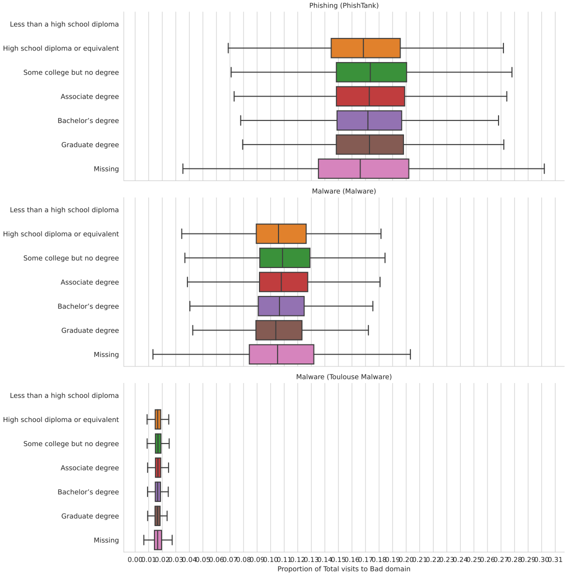
- Calibration $\rightsquigarrow$ $FP = FN$ locally
- Using curated lists (manual coding) for popular domains

The better educated visit phishing/malware domains most often



Visits to phishing/malware domains

But the less educated choose worse



Proportion of visits to phishing/malware domains