Summer Internship Presentation

Me, Myself, and My Models

Disha Dasgupta

About Me

- Rising senior at Stanford University
- Majoring in Data Science, Markets, and Management

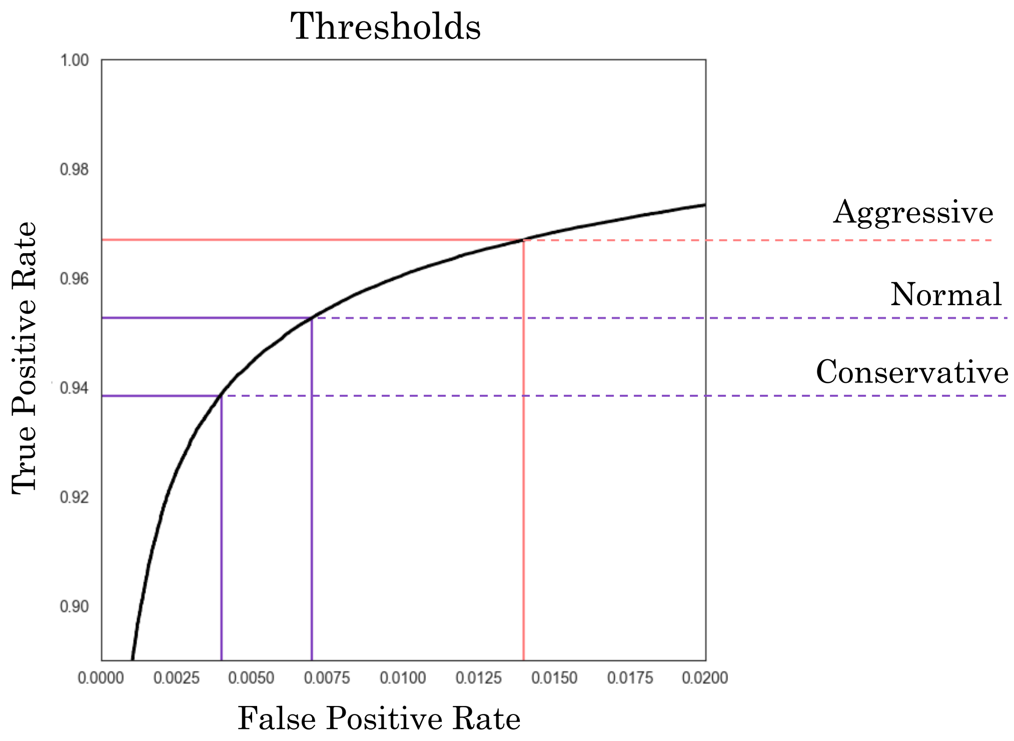
Projects

- MalwareScore Evaluation
- File Path based Malware Classification
- MalwareScore Data Reduction

Improving Evaluation Method

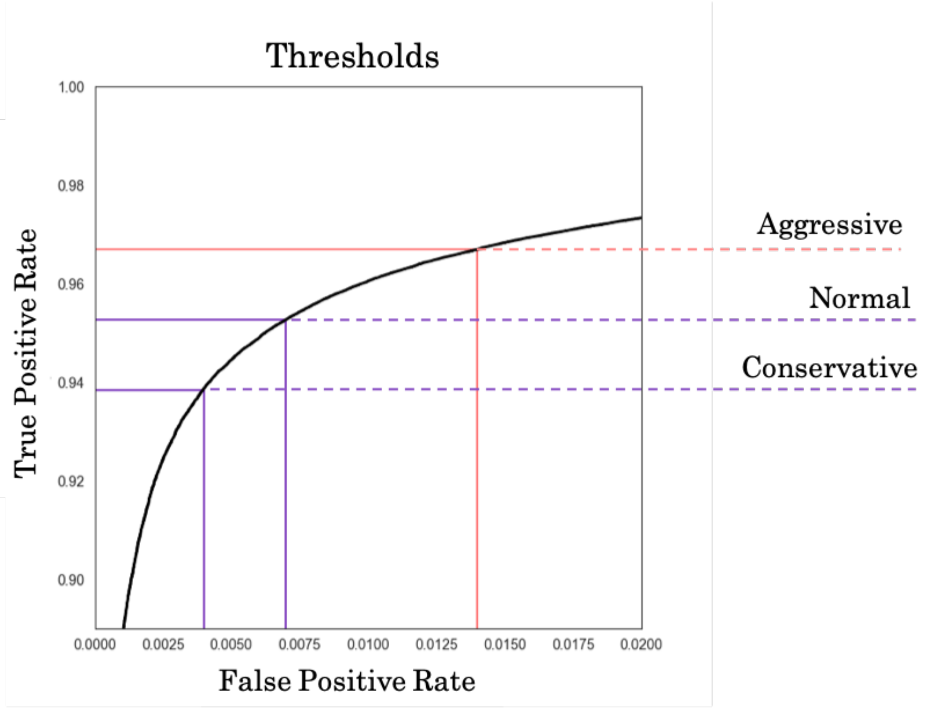
Motivation

- MalwareScore evaluation method was computationally and timewise expensive



Process

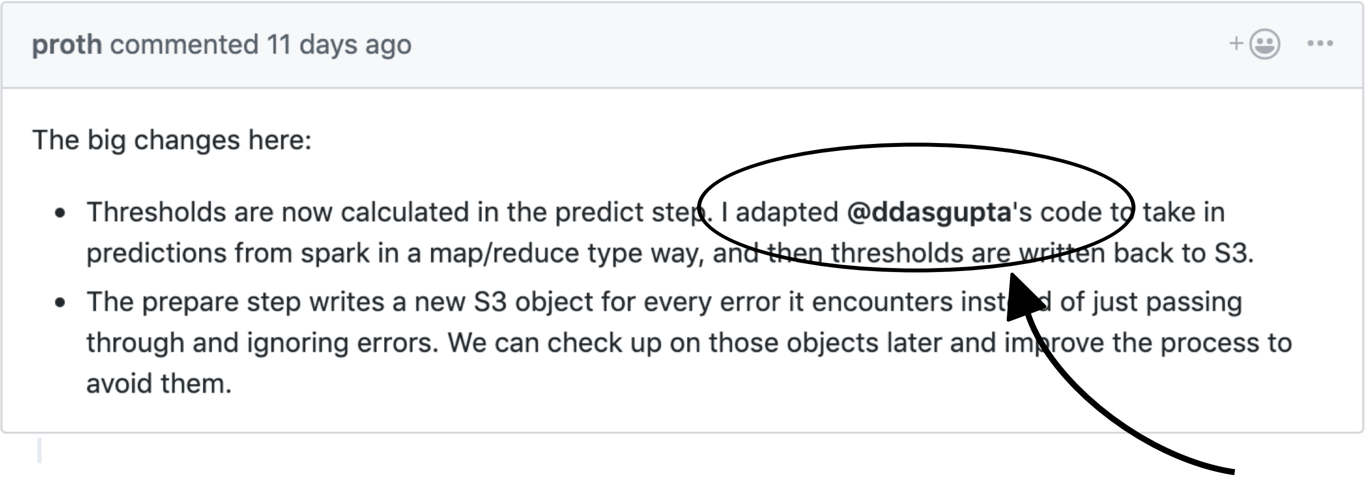
- Histogrammed benign and malicious files into cumulative bins for each threshold
- Used those values to calculate corresponding thresholds from desired false positive rates (or vice versa)



Improving Evaluation

Results and Moving Forward

- Evaluation method is faster and more computationally effective
- Step towards streamlining entire prediction and evaluation process into one step



proth commented 11 days ago

The big changes here:

- Thresholds are now calculated in the predict step. I adapted **@ddasgupta**'s code to take in predictions from spark in a map/reduce type way, and then thresholds are written back to S3.
- The prepare step writes a new S3 object for every error it encounters instead of just passing through and ignoring errors. We can check up on those objects later and improve the process to avoid them.

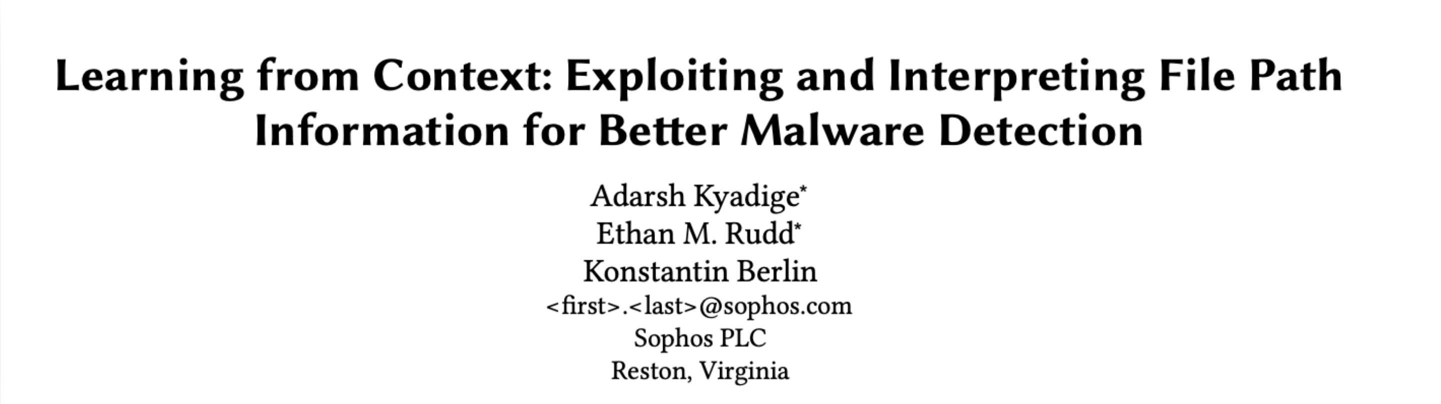
my peak!

File Path Model

File Path Model

Motivation

- Inspired by similar work by Kyadige, Rudd, and Berlin
- Could we use contextual information to improve static malware classification (i.e. MalwareScore)?

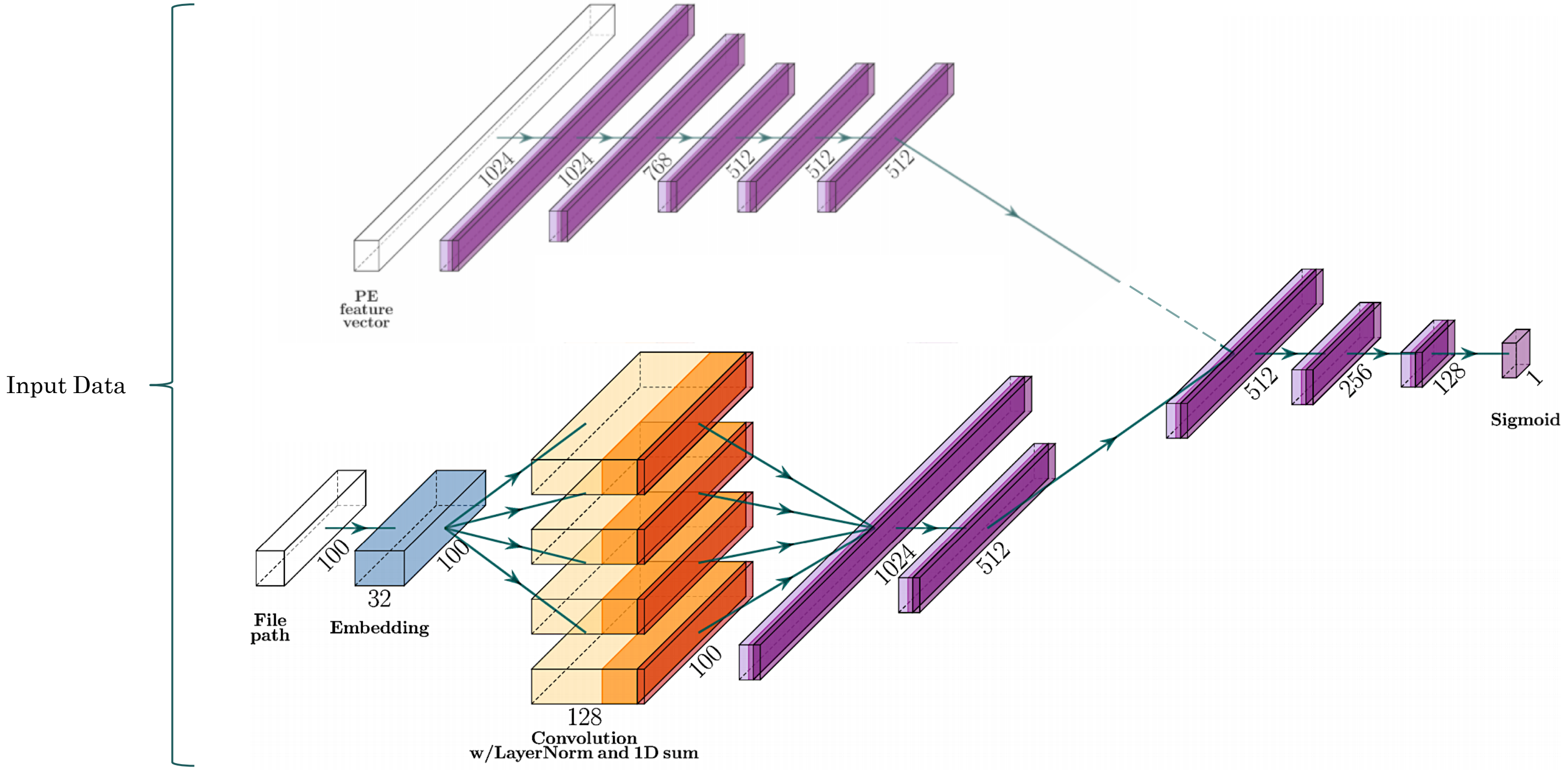


Learning from Context: Exploiting and Interpreting File Path Information for Better Malware Detection

Adarsh Kyadige*
Ethan M. Rudd*
Konstantin Berlin
<first>.<last>@sophos.com
Sophos PLC
Reston, Virginia

File Path Model

Process



Picture adapted from Kyadige, Rudd, and Berlin

Process

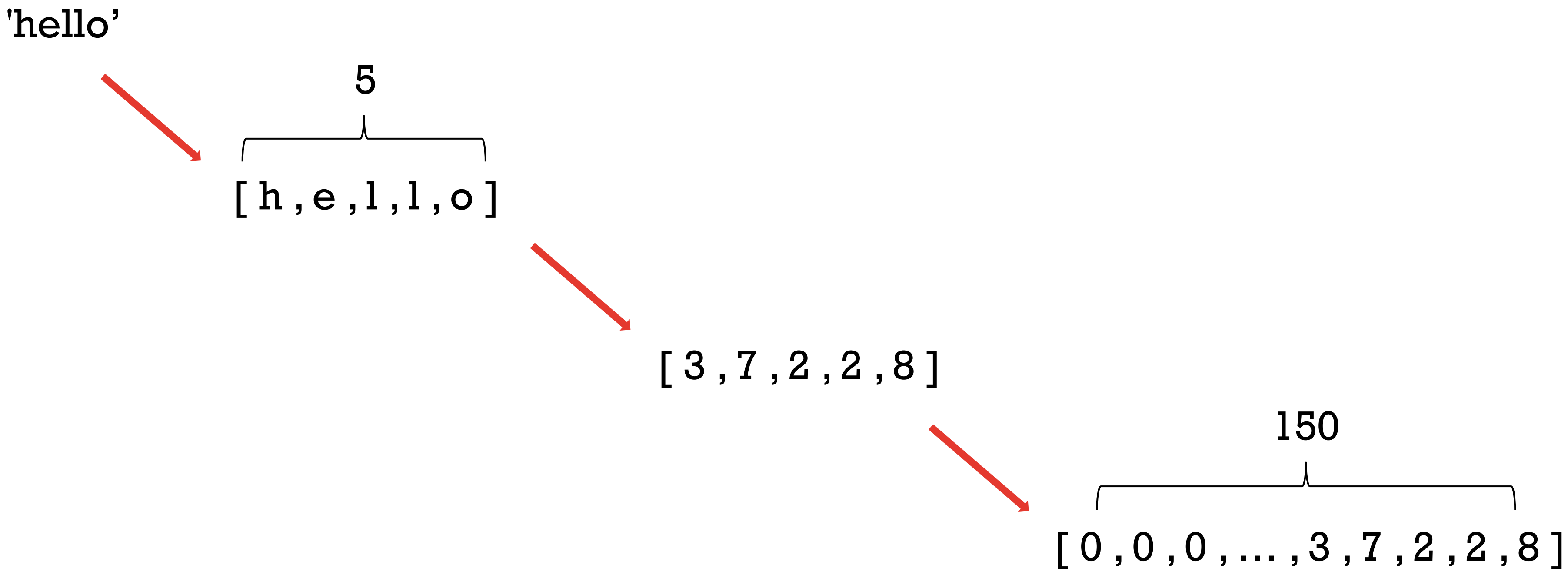
`C:\users\Bob\appdata\local\temp\rar\payment.scr.`

↓

`[drive]\users\[user]\appdata\local\temp\rar`

File Path
Model

Process



'hello'

$\overbrace{[h,e,l,l,o]}^{5}$

[3 , 7 , 2 , 2 , 8]

$\overbrace{[0,0,0,\ldots,3,7,2,2,8]}^{150}$

File Path Model

Process

- Training data (Thank you Rich and Response Team!!!)
  - Malicious data: Alerts from Metabase
  - Benign data: Process surveys on POC platforms
- Challenges:
  - Some malicious data comes from testing Endgame protections (i.e. Demo4)
  - Malware was launched in non realistic locations/paths

Let’s Demo!

- http://striker.endgames.local:3000/

File Path Model

Moving Forward

- Obtain more data
- Analyze effects of depth of file path
- Build baseline model for file path contextualization
- Integrate baseline model as part of MalwareScore
    - Elevate/suppress alerts on SMP

Data Reduction

Data Reduction

Motivation

- MalwareScore has had high predictive accuracy
- We suspect accuracy has decreased over time
- Can training data be aged off?

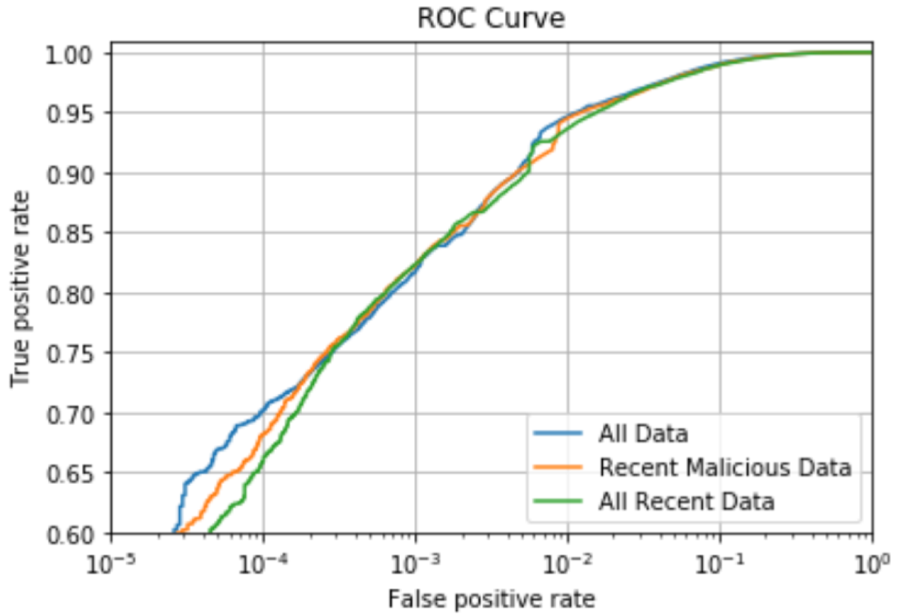
Data Reduction

Process

- Used 5% dataset
- Train/Test Data: Before/after February 1st, 2019
- Trained 3 models
  - All data
  - Most recent malicious data, all benign data
  - Most recent benign and malicious data
- Compared performance levels using Area Under the Curve (AUC) values

Data Reduction

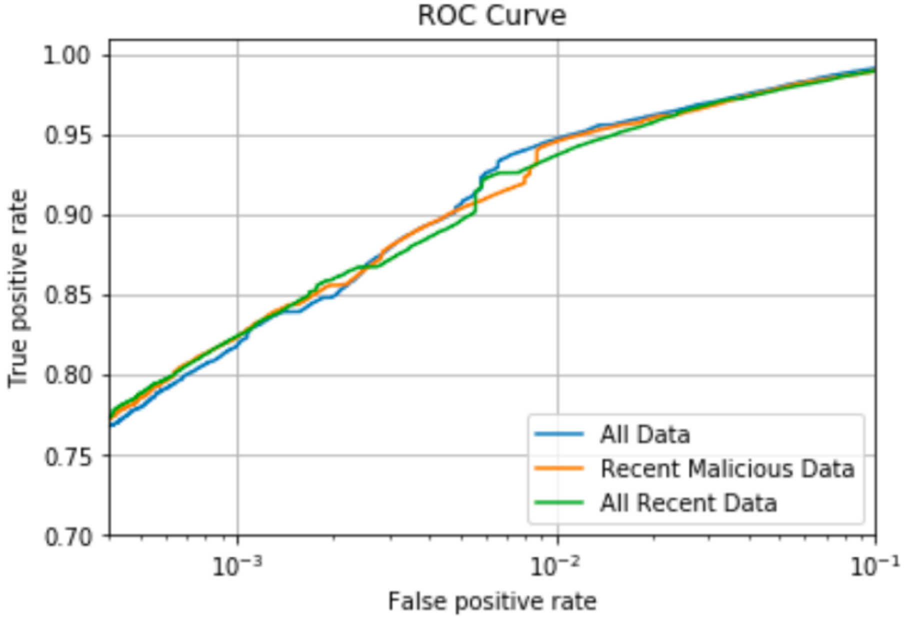
Results



Model	AUC Score
All Data	0.9959
Recent Malicious Data	0.9957
All Recent Data	0.9955

Data Reduction

Results



Model	AUC Score
All Data	0.9959
Recent Malicious Data	0.9957
All Recent Data	0.9955

Moving Forward

- Similar AUC values → similar performance
- Performance differs in the way we would expect
- Remove old data and retain similar performance accuracy

Special Thanks

- Bobby Filar, Phil Roth, & Rich Seymour
- Daniel Grant
- Research and Development

“Your slides
are boring”

“Add more
memes or gifs”

Thank You!

- Twitter: @DishaDasgupta
- Instagram: @disha_dasgupta
- Email: disha01@stanford.edu

~ Stay in touch ~