Firenze: Model Evaluation using Weak Signals

Bhavna Soman

Security Researcher, AWS

Michael Morais

AWS

Ali Torkamani

AWS

Jeffery Bickford

AWS

Baris Coskun

AWS

Agenda

- Challenges of Current Evaluation methods
- Firenze, Introduction and Key Constructs
- Practical applications
- Results and Limitations

Challenges of Current Evaluation Methods

Often, strong performance seen in PR curves/ROC curves does not translate to the real world

Why does this happen?

- Real world data distributions are different, complex and not always represented in the training data
  - Complexity in the universe
  - Concept Drift
- Labels are noisy, sparse or absent
- Feedback is infrequent and imperfect

How we handle this today

- Lengthy manual evaluation by domain experts
  - Shadow mode
  - Replay mode
- **Limitation:** Time taken
- **Limitation:** Uses scarce security talent
- **Limitation:** Impact on ability to innovate

Firenze: Key Constructs

What is a Marker?

"A marker is a weak signal that is cheaply obtained and is associated with the maliciousness or benignity of a sample, instance, or event."

- Based on Domain Expertise
- Cheap to obtain
- Weak signal/Imperfect accuracy
- Combine information from many markers over populations to better evaluate a model

A Toy Example: Domain Classification

```
if domainAge < 1 day then domain is likely malicious
```

Marker	Type	Description	Sample "Marker Function"
Domain Age	Malicious Signal	Malicious domains likely have lower age	1 if domain age < 1 day, else 0
Popularity	Benign Signal	Benign domains likely appear on popular lists like Alexa top X	-1 if domain appears in Alexa top 10k, else 0
Known good registrar	Benign Signal	Benign domains likely registered via reputable providers.	-1 if domain registered with one of list of known good registrars, else 0

- Marker verdicts $m_j(s)$ indicate the verdict of the j^{th} marker for sample s and $m_j(s) \in \{-1, 0, 1\}$
- Markers can abstain

Combining Scores for a single sample

- Combining multiple markers to provide a stronger verdict

Domain	ML mode Score	IsDomAge Marker	IsDomPopular	IsKnownReg	Marker Score z_i
amazon.com	TBD	0	-1	-1	-1
ibcojed.ga	TBD	1	0	0	1

Intuition: For two samples s_i and s_j, if $z_i > z_j$, then s_i is ***more malicious*** *than* s_j.

- Emulates how human experts build confidence
- Using Majority Voting, naïve but suitable for low signal density
- Other methods can be explored for future work

Comparing Sets of Samples

- Compute the Average Marker Score Z(S) for the group of samples

$$Z(S) = \frac{1}{N} \sum_{i=1}^{N} z_i$$

Intuition: If $Z(Set_1) > Z(Set_2)$ then Set_1 contains more malicious samples.

Toy Example: Defining the Reference Set and the Test Set

Common universe of domains

Domain
ibcojed.ga
kwoe.us
mj5f.ddns.net
m.likarooxsmile.com
0-007.ws
a6kn1judi41rob3.ws
brajrasik.org
jxbnpoveb.org
dpstream.biz
xn--gamebi-mta.com
328-bfz-688.mktoresp.com
⋮
droscarundurraga.com

Top K malicious domains scored by Reference

Domain	Score by Old Model
ibcojed.ga	0.993
mj5f.ddns.net	0.993
dpstream.biz	0.987
328-bfz-688.mktoresp.com	0.965
kwoe.us	0.921

Reference Set

Top K malicious domains scored by Test

Domain	Score by New Model
kwoe.us	0.891
mj5f.ddns.net	0.852
0-007.ws	0.85
jxbnpoveb.org	0.85
dpstream.biz	0.80

Test Set

Toy Example: Comparing two sets of samples

Common universe of domains

Domain	Marker Score z_i
ibcojed.ga	1
kwoe.us	1
mj5f.ddns.net	1
m.likarooxsmile.com	-1
0-007.ws	1
a6kn1judi41rob3.ws	1
brajrasik.org	0
jxbnpoveb.org	1
dpstream.biz	1
xn--gamebi-mta.com	-1
328-bfz-688.mktoresp.com	0
⋮	
droscarundurraga.com	1

Top K malicious domains scored by Reference

Domain	Score by Old Model	Marker Score z_i
ibcojed.ga	0.993	1
mj5f.ddns.net	0.993	1
dpstream.biz	0.987	1
328-bfz-688.mktoresp.com	0.965	0
kwoe.us	0.921	1

Reference Set

Top K malicious domains scored by Test

Domain	Score by New Model	Marker Score z_i
kwoe.us	0.891	1
mj5f.ddns.net	0.852	1
0-007.ws	0.85	1
jxbnpoveb.org	0.85	1
dpstream.biz	0.80	1

Test Set

$$m_j \in \{\mathrm{IsDomAgeMarker}, \mathrm{IsDomPopular}, \mathrm{IsKnownReg}\}$$

Toy Example: Comparing two sets of samples

Common universe of domains

Domain	Marker Score Z_i
ibcojed.ga	1
kwoe.us	1
mj5f.ddns.net	1
m.likarooxsmile.com	-1
0-007.ws	1
a6kn1judi41rob3.ws	1
brajrasik.org	0
jxbnpoveb.org	1
dpstream.biz	1
xn--gamebi-mta.com	-1
328-bfz-688.mktoresp.com	0
⋮	
droscarundurraga.com	1

Top K malicious domains scored by Reference

Domain	Score by Old Model	Marker Score z_i
ibcojed.ga	0.993	1
mj5f.ddns.net	0.993	1
dpstream.biz	0.987	1
328-bfz-688.mktoresp.com	0.965	0
kwoe.us	0.921	1

Reference Set

$$Z(R) = \frac{1}{N}\sum_{i=1}^{N} z_i = 0.8$$

Top K malicious domains scored by Test

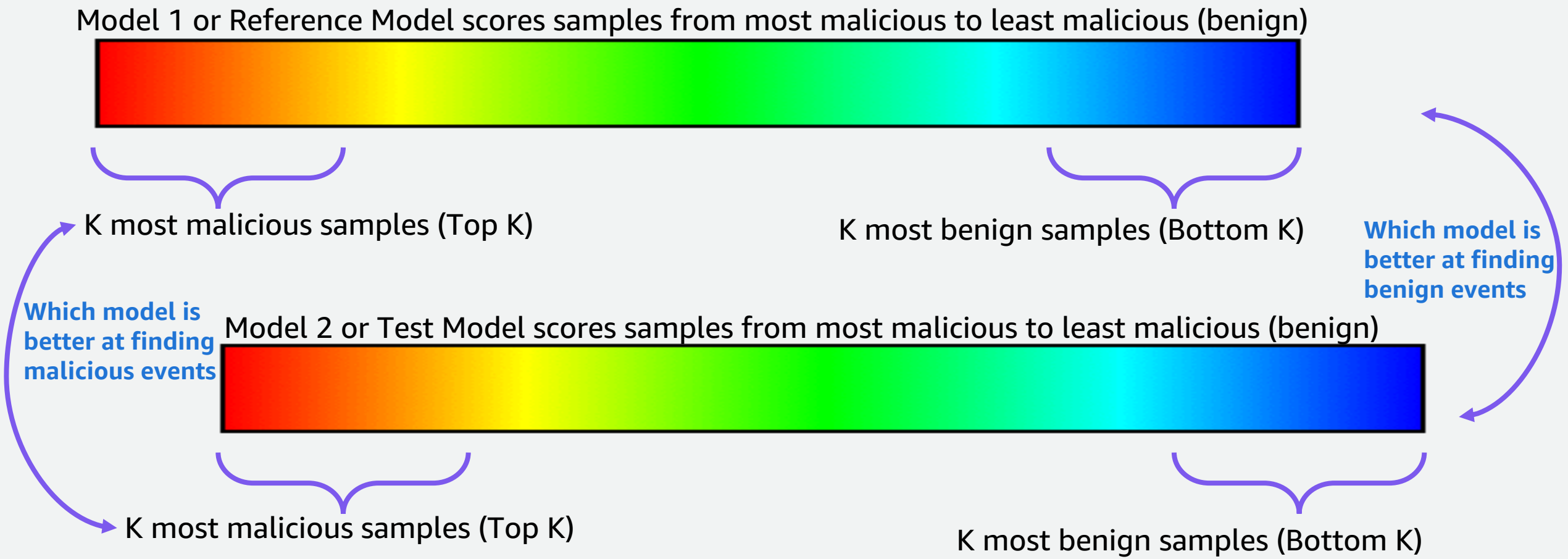
Domain	Score by New Model	Marker Score z_i
kwoe.us	0.891	1
mj5f.ddns.net	0.852	1
0-007.ws	0.85	1
jxbnpoveb.org	0.85	1
dpstream.biz	0.80	1

Test Set

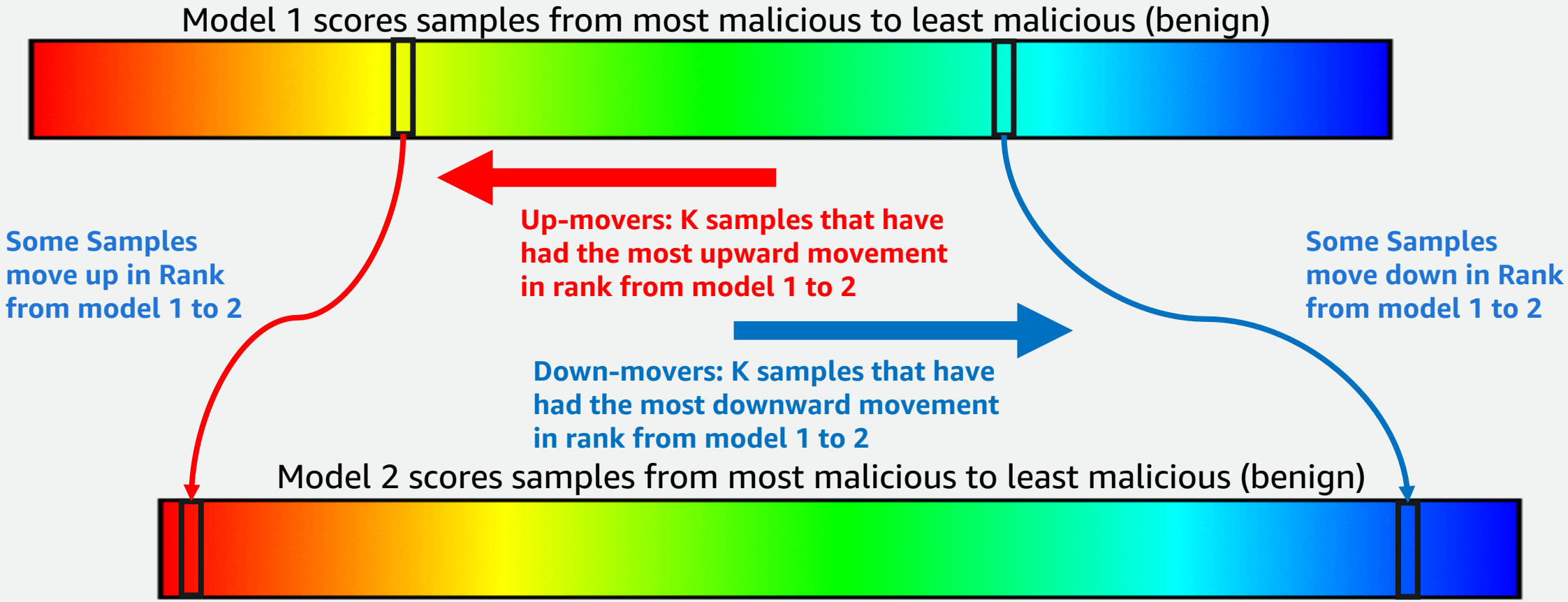
$$Z(T) = \frac{1}{N}\sum_{i=1}^{N} z_i = 1$$

Intuition: If Z(T) > Z(R) then the new model is better at finding malicious domains

How to define the sets: "Locally interesting Regions"



"Holistic comparison using global ranks"



Intuition: If average marker score of the Up-movers is greater than that of down-movers and passes significance then the new model is better.

What is the significance of the difference?

- **Question:** Do the Test Model's high-ranked samples have **significantly higher** marker-scores on average compared to the Reference model??
  - i.e. is the test model ***significantly*** better than the reference model?
- **Answer**: Hypothesis testing
- Averages of marker-scores Z(Set) will follow a normal or a t-distribution (our sample size is large)
  - Two-sample or paired statistical test (like Welch's t-test)

Firenze Tests

- Top-K Test → Z(RefTop) >= Z(TestTop)
- Bottom-K Test → Z(RefBottom) <= Z(TestBottom)
- Movers Test → Z(UpMovers) <= Z(DownMovers)
- **If assertion is false, and p-value < 0.05 then reject the null hypothesis**

Application: Malware Classification

Experimental Set up to Evaluate Malware Classifiers

- EMBER Malware Dataset
- Two models
  - NN (Reference) Model: Adversarially robust neural network (Erdemir et al, Neurips 2021)
  - Tree (Test) Model: Gradient-boosted decision tree (Anderson et al. in the EMBER paper, 2018)
- Trained on "past" samples (collected pre-Dec 2017, 600k files)
- Validated on "present" samples (collected in Dec 2017, 200k files)

```
Reference model on training data:
P=0.9829
Reference model on validation data:
TNR=0.9877, FPR=0.0123, FNR=0.0240, TPR=0.9760
Prec=0.9876, Rec=0.9760, F1=0.9817, AUC=0.9981
P=0.9819
```

```
Test model on training data:
P=0.9820
Test model on validation data:
TNR=0.9856, FPR=0.0144, FNR=0.0199, TPR=0.9801
Prec=0.9856, Rec=0.9801, F1=0.9828, AUC=0.9984
P=0.9829
```

Evaluating Malware classification with Firenze

- Using "future" data (unlabeled, collected in 2018, 200k files)
- K= 50k
- Designed 5 marker functions
- E.g. if the section name is random looking or contains UPX, then file is likely malicious
- E.g. if the file is signed then it is likely benign

Firenze Test Results

	NN Model	Tree Model		
	Average CMS Score	**Average CMS Score**	**p-value**	**Which is better**
TopK	0.11456	0.68445	$< 10^{-16}$	Test
BottomK	0.09788	-0.16862	$< 10^{-16}$	Test
	Up-Movers	**Down-Movers**		
	Average CMS Score	**Average CMS Score**	**p-value**	**Which is better**
Movers Test	0.42884	0.00868	$< 10^{-16}$	Test

Limitations

- This method does not preclude the need for good training data
- We rely on security experts to define markers
- Does not allow marker signals to overlap with those used in training to prevent bias
- Test sensitivity is varies with experiment parameters (e.g. K)
- Proves fitness for use by comparison

What we are working on now

- Testing new ways for marker aggregation
- Estimating single model performance with weak signals
- Formalizing explainability with markers
- .. And more.

Thank you!

Bhavna Soman