Scamdog Millionaire: Detecting E-commerce Scams in the Wild

December 6th, 2023

Platon Kotzias, Kevin Roundy, Michalis Pachilakis, Iskander Sanchez-Rola, Leyla Bilge

Norton Research Group

E-commerce scams definition

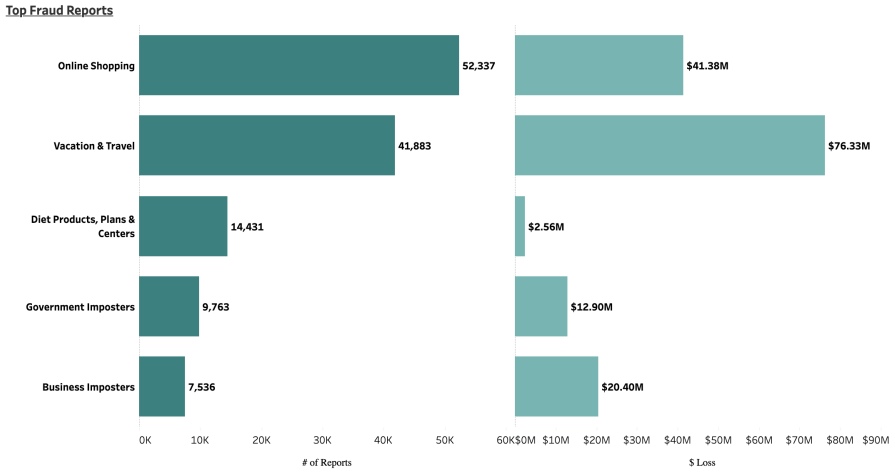
Online e-shops with a huge gap between what consumers are promised and what they receive

When consumers interact with these e-shops:

- Do not receive anything
- Receive a counterfeit/cheap replicas despite claims of originality
- Receive a completely different product from the one ordered

Impact on Consumers

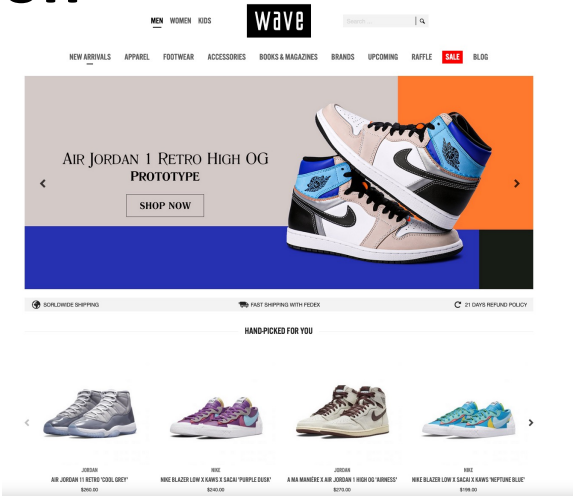
- FTC reported losses of over $41M in 2020 and $392M in 2021 due to online shopping scams
- BBB ranks online purchase scams as the top consumer risk for second consecutive year (2022)
- ACCC reports losses of $8.0M from 20.6K consumers

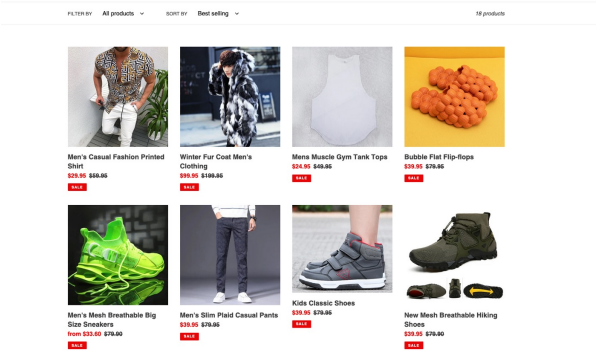


Top Frauds reported by FTC for 2021

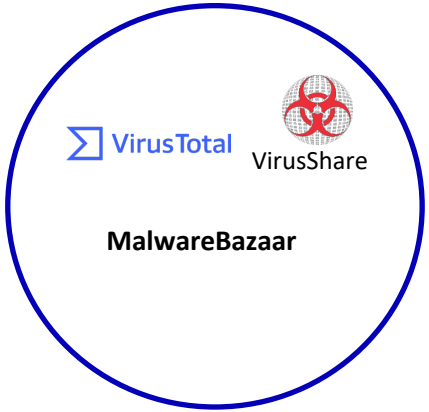
Challenge #1 – Increased sophistication

- Look and feel very similar to benign e-shops
- Ease of deployment and maintenance allows scammers to scale their operations
- Target various business sectors





Challenge #2 – Lack of labelled datasets



Malware repositories

Phishing blocklists

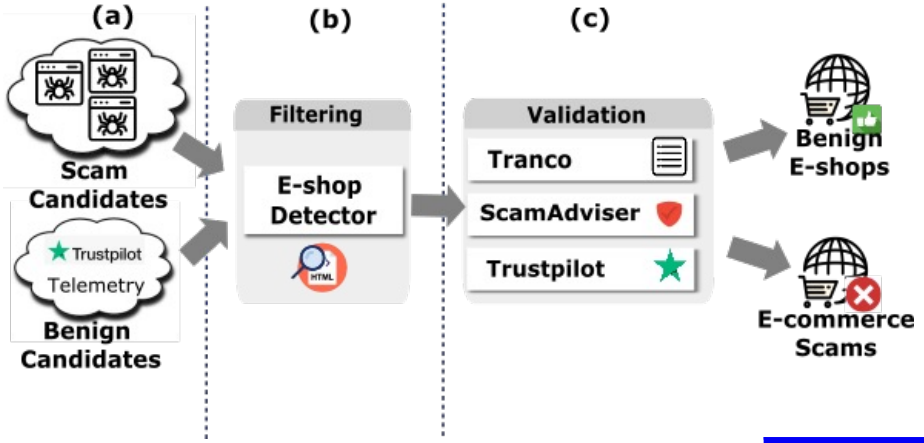
Online Scams

Only 4.2% of ecommerce scams in our dataset is detected by at least one URL scanners on VirusTotal

Contributions

- Automated and replicable pipeline for creating labeled dataset of e-commerce scams and benign e-shops
- Measure e-commerce scam traffic volumes on real consumers
- Build an ML classifier for detecting e-commerce scams using features attuned to this type of threat
- Evaluate classifier and show that it achieves high performance both on groundtruth data and in a large-scale real-world evaluation

Ground-truth Dataset pipeline



- Scam candidates from 4 scam repositories:
  - 2 with reports from volunteer scam analysts
  - 2 with victims' complaints
- Benign candidates:
  - Scrape popular 3rd party review platforms
  - Very popular shopping domains in our telemetry

Dataset	E-shops
Benign	5,179
Scams	3,765
Total	**8,944**

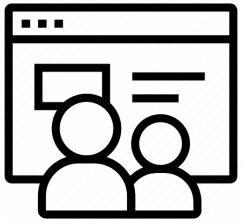
Dataset is 9 to 19 times larger than that of prior work

E-commerce scams traffic

Measure traffic of 3.7K scam domains on Norton's telemetry for 7 months

(January 2021 – July 2021)

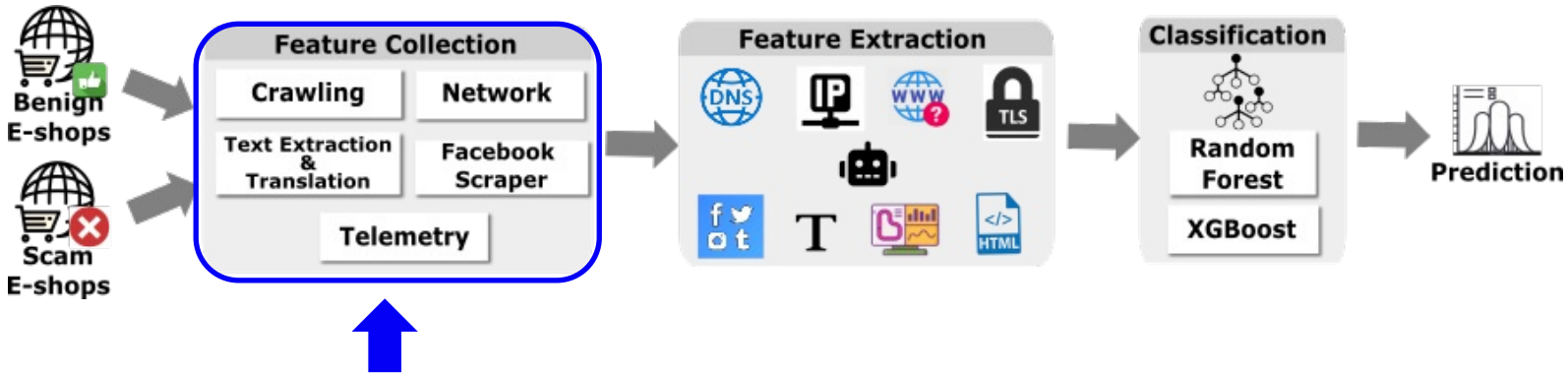
- 75% of scams visited by real users



- **6.3M** visits
- **30K** average visits per day

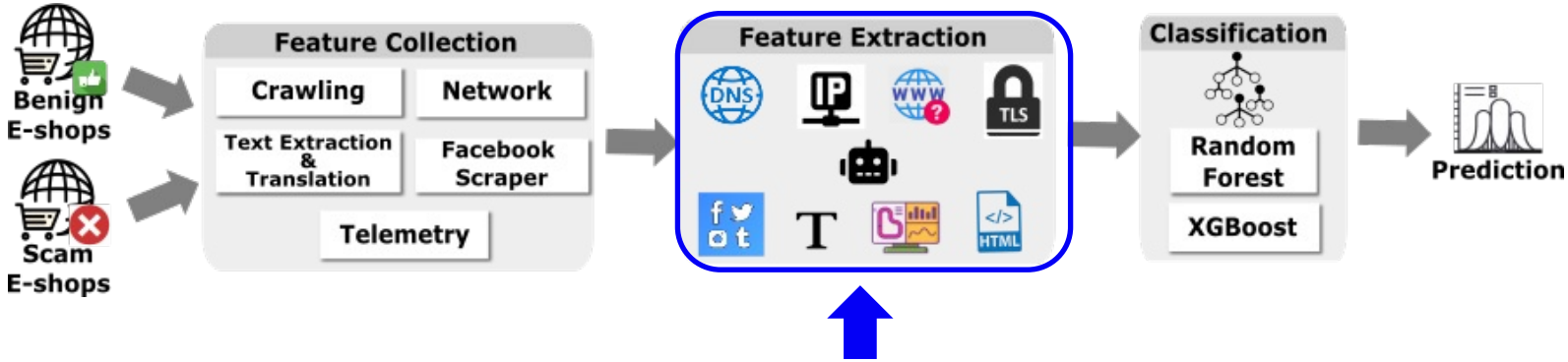
- 56.7% of all scam visits
- 7 out 10 sell clothes
- 4 out 7 target women

E-commerce scam detection pipeline



- Crawling of main, secondary (e.g., About-Us, Contact, etc.), and robots.txt pages
- Network module collects DNS, RDAP, IP, and TLS related information
- Text traslation uses LibreTranslate and supports 30 languages
- Facebook scraper module detects scammers' FB profiles and crawls them
- Telemetry statistics from millions of real users

E-commerce scam detection pipeline



- **111 features** groupped into five categories:
  - Network (DNS, RDAP, IP, TLS)
  - HTTP
  - Robots.txt
  - Content (HTML, Social media, Text)
  - Domain (domain characteristics, telemetry)

See our **294** reviews on Trustpilot

"Excellent"

See our **294** reviews on Trustpilot

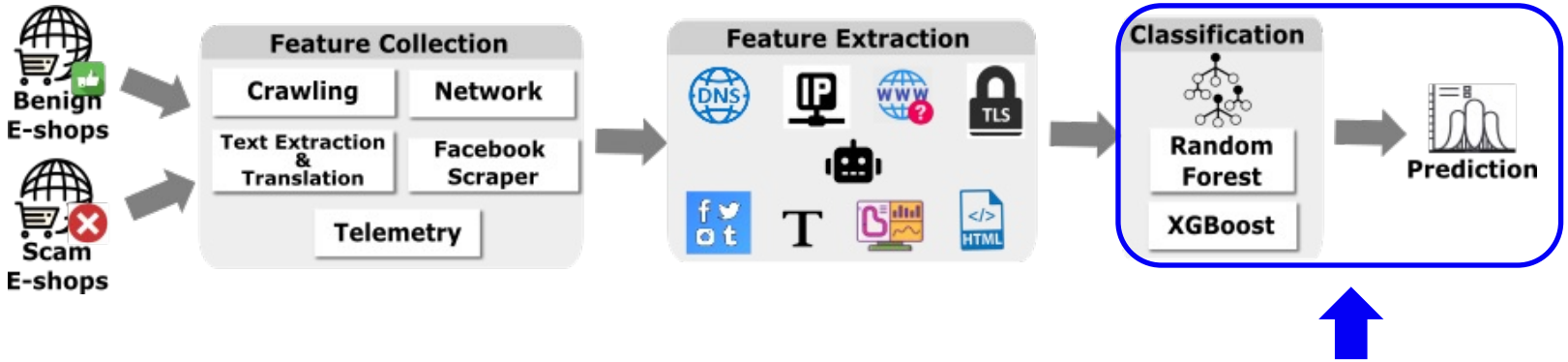
Integration with external review platforms found on 80% of benign e-shops compared to 1% of e-commerce scams

Social media integration far more common on benign e-shops

- **44 are novel** and targetting specific e-commerce scam characteristics

Benign e-shops provide more contact information

E-commerce scam detection pipeline



- Trained a Random Forest classifier
- Pipeline outputs the probability [0, 1] that an e-shop is scam

Evaluation on ground-truth data

Dataset	E-shops
Benign	5,179
Scams	3,765
Total	**8,944**

Features	Algo	Precision	Recall	F1-score
All	RF	0.988	0.959	0.973
No telemetry	RF	0.979	0.946	0.962
FaDe [94]	SVM	0.846	0.876	0.860
SOTA	RF	0.966	0.898	0.930

- All models evaluated using 10 x 10-fold cross validation
- High performance using all features achieving F1-score of 0.973
- Performance only 1% lower if telemetry features are not used
- Best-effort comparison with prior work [94] showing improvement of 11% on F1-score
- 44 novel features offer performance boost of 4.3% on F1-score (decrease FPR by 2.5 times)

[94] Thymen Wabeke, Giovane Moura, Nanneke Franken, and Cristian Hesselman. 2020. Counterfighting Counterfeit: detecting and taking down fraudulent web- shops at a ccTLD. In International Conference on Passive and Active Network Measurement. Springer, Cham, 158–174.

Real-world evaluation

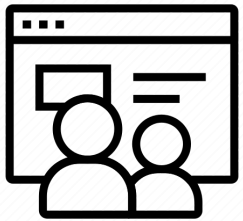
- Deploy our classifier in production for two months

760K

E-shops

10% (73K)

e-commerce scams



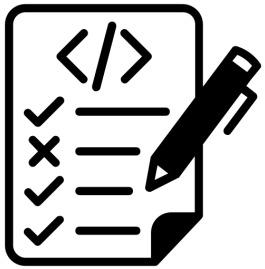
One week analysis:

17K e-commerce scams

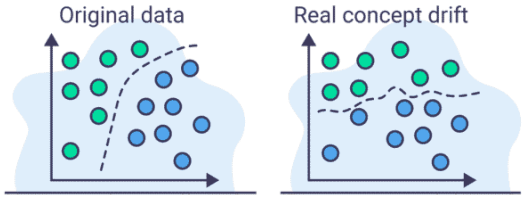
285K real visits

- Detection engine can protect hundreds of thousands of users
- Manual examination of top 100 most visited scams:
  - 2 FPs only (ranked 80th and 87th)
  - FPs received less than 0.2% of the visits

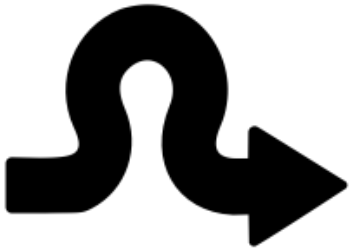
More evaluation



Feature Importance



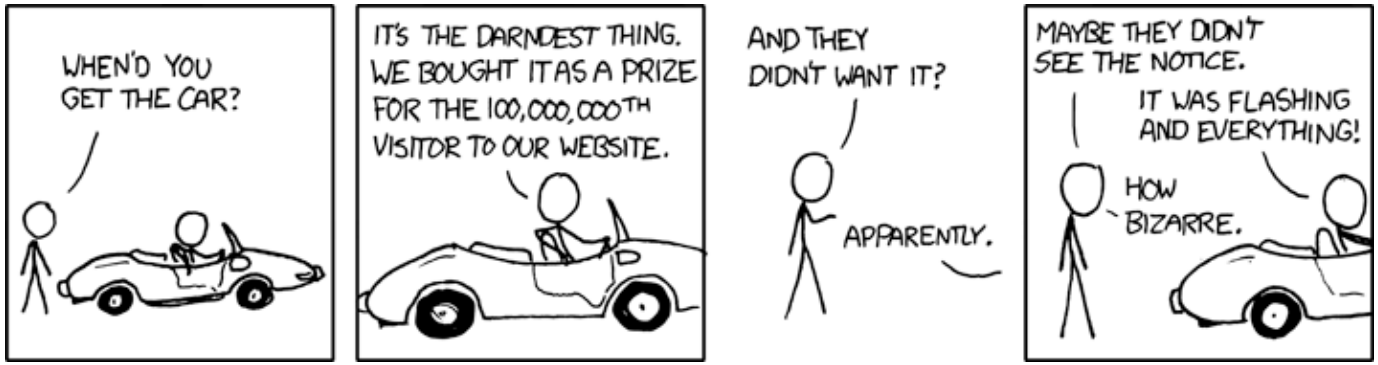
Concept Drift Impact



Model Evasion

Summary

- E-commerce scams is a real and important problem for Internet users
- We build an automated and replicable pipeline for creating labeled dataset of e-commerce scams and benign e-shops
- Build an ML classifier for detecting e-commerce scams that achieves high performance both on groundtruth data and in a real-world deployment



Thank you!

Contact: platon.kotzias@gendigital.com