UNIVERSIDAD DEL VALLE DE GUATEMALA

Informe Laboratorio 1

Andre Marroquin Tarot - 22266

Security data science

1. ¿Qué ventajas tiene el análisis de una URL contra el análisis de otros datos, como el tiempo de vida del dominio o las características de la página web?

- El análisis de URLs permite extraer información directamente del enlace, sin acceder al sitio web. Esto lo hace rápido, apto para detección en tiempo real y efectivo ante ataques zero-hour. Además, evita riesgos al no cargar contenido malicioso, a diferencia del análisis de HTML o dominio, que requiere descargas y consultas externas.

2. ¿Qué características de una URL son más prometedoras para la detección de phishing?

- Longitud de la URL y del hostname: las URLs excesivamente largas se usan para ocultar el dominio real y distraer al usuario.

- Uso de caracteres especiales: gran número de puntos, guiones, arrobas, signos de interrogación, ampersands, iguales, guiones bajos o slashes suele emplearse para confundir visualmente el dominio y manipular su lectura.

- Uso de direcciones IP en lugar de dominios: cuando el hostname es una IP y no un nombre de dominio, suele ser un indicador de phishing.

- Palabras sensibles en la URL: términos como login, signin, secure, account, confirm o banking se incluyen para generar confianza falsa y engañar al usuario.

- Redirecciones sospechosas: la presencia de varios // en el path o URLs embebidas dentro de otras URLs indica intentos de redirección encubierta.

- Uso engañoso de https: incluir la palabra https dentro del nombre del dominio, en lugar de usarla como esquema real, busca dar una falsa sensación de seguridad.

- Proporción alta de dígitos y mezcla de caracteres: dominios o rutas con muchos números y combinaciones aleatorias suelen indicar ofuscación.

- Entropía de caracteres no alfanuméricos: distribuciones desordenadas de símbolos especiales son típicas de URLs maliciosas, medir su entropía resulta más efectivo que usar solo conteos simples.

3. ¿Qué columnas o características fueron seleccionadas y por qué?

En el notebook se seleccionaron 15 características derivadas directamente de la URL:

- urlLength

- hostnameLength
- pathLength
- dotCount
- dashCount
- atCount
- questionCount
- digitCount
- digitRatio
- hasIpAddress
- hasSensitiveWords
- hasDoubleSlashRedirect
- hasHttpsTokenInDomain
- nanEntropy
- nanRelativeEntropy

Estas columnas no son constantes ni redundantes y muestran variación clara entre URLs de verdad y de phishing. Además, están respaldadas por la literatura como features efectivas basadas en URL.

Las métricas de entropía y entropía relativa capturan el desorden de símbolos especiales, mejorando el desempeño del modelo y reduciendo el costo computacional. En el análisis exploratorio, estas variables presentan mayor correlación con la etiqueta y reflejan estrategias reales de ofuscación usadas por atacantes.

4. ¿Cuál es el impacto de clasificar un sitio legítimo como phishing? (falso positivo)

- Cuando un sitio legítimo es clasificado como phishing, se bloquea o se marca como peligroso sin serlo realmente. Esto genera molestia en el usuario al impedir el acceso normal, reduce la confianza en el sistema de detección y puede incrementar los costos de soporte por reportes o reclamos. Si se ve en entornos corporativos, los falsos positivos también pueden afectar la productividad al bloquear aplicaciones o herramientas de trabajo legítimas.

5. ¿Cuál es el impacto de clasificar un sitio de phishing como legítimo? (falso negativo)

- Cuando un sitio de phishing es clasificado como legítimo, el usuario accede a un entorno malicioso creyendo que es seguro. Esto puede provocar robo de credenciales, compromiso de cuentas corporativas y facilitar ataques posteriores como movimiento lateral o ransomware. Además, puede generar pérdidas económicas y daño reputacional.

En términos de seguridad, los falsos negativos representan un riesgo mayor que los falsos positivos.

6. En base a las respuestas anteriores, ¿qué métrica elegirías para comparar modelos similares de clasificación de phishing?

- Cuando el objetivo es detectar phishing, el mayor riesgo es permitir que un sitio malicioso pase como legítimo, es decir, cometer falsos negativos. Es por eso que, la métrica principal para comparar modelos debe ser el recall de la clase phishing, ya que mide qué tan bien el modelo identifica los sitios realmente maliciosos.

  Por otro lado, optimizar solo el recall puede aumentar los falsos positivos. Por eso, también es importante considerar la precisión, que indica qué proporción de las alertas generadas corresponde realmente a phishing, y el AUC-ROC, que permite evaluar el balance general entre detecciones correctas y errores.

7. ¿Qué modelo funcionó mejor para la clasificación de phishing? ¿Por qué?

- La regresión logística obtiene en el conjunto de prueba algo cercano a:
  - precision ≈ 0.82
  - recall ≈ 0.68
  - AUC ≈ 0.83
- El random forest obtiene en prueba aproximadamente:
  - precision ≈ 0.83
  - recall ≈ 0.83
  - AUC ≈ 0.92
- Por lo tanto, el Random Forest presenta un mejor desempeño en la detección de phishing.

  Funciona mejor porque alcanza un mayor recall para la clase phishing, logrando identificar una proporción más alta de URLs maliciosas. Además, mantiene una precisión un poco mejor, lo que evita un incremento significativo de falsas alarmas. Y por último, su AUC-ROC es mucho más alto, lo que indica un mejor balance global entre verdaderos positivos y falsos positivos.

8. Una empresa desea utilizar su mejor modelo, debido a que sus empleados sufren constantes ataques de phishing mediante e-mail. La empresa estima que, de un total de 50,000 correos electrónicos, un 15% son phishing. ¿Qué cantidad de alarmas generaría el modelo? ¿Cuántas serían positivas y cuántas negativas? ¿Funciona el modelo para el base rate propuesto? En caso negativo, ¿qué se propone para reducir la cantidad de falsas alarmas?

Base rate

- Total de correos: 50,000
- Phishing (15%): 7,500
- Legítimos: 42,500

Se utiliza el mejor modelo Random Forest con métricas aproximadas según la corrida del código:

- Recall phishing ≈ 0.83
- Precision phishing ≈ 0.83

Resultados estimados

- **Verdaderos positivos (TP)**:
  ≈ 0.83 × 7,500 ≈ **6,257**

- **Falsos positivos (FP)**:
  ≈ 6,257 / 0.83 − 6,257 ≈ **1,256**

- **Alarmas totales generadas predicción phishing**:
  TP + FP ≈ **7,513**

    - Positivas verdaderas: ≈ 6,257
    - Positivas falsas: ≈ 1,256

- **Falsos negativos phishing no detectado**:
  ≈ 7,500 − 6,257 ≈ **1,243**

- **Verdaderos negativos**:
  ≈ 42,500 − 1,256 ≈ **41,244**

¿Funciona el modelo para el base rate propuesto?

Sí. El modelo detecta aproximadamente el 83% de los correos de phishing, manteniendo una precision alta, por lo que la mayoría de las alarmas corresponden a amenazas reales. Sin embargo, 1,256 falsos positivos pueden generar una carga operativa considerable para el equipo de seguridad.

Propuestas para reducir falsas alarmas

- Ajustar el umbral de decisión para etiquetar phishing, reduciendo falsos positivos a costa de una leve disminución del recall.

- Implementar un segundo nivel de verificación como el análisis de contenido HTML o reglas adicionales solo para correos de riesgo medio.

- Incorporar listas blancas de dominios confiables tales como bancos, proveedores y servicios internos.

Referencias

- Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029
- Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B., & Joga, S. R. K. (2023). Phishing detection system through hybrid machine learning based on URL. IEEE Access, 11, 1–16.
- Calzarossa, M. C., Giudici, P., & Zieni, R. (2023). Explainable machine learning for phishing feature detection. Quality and Reliability Engineering International. https://doi.org/10.1002/qre.3411
- Hannousse, A., & Yahiouche, S. (2020). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. arXiv preprint arXiv:2010.12847. https://arxiv.org/abs/2010.12847
- Aung, E. S., & Yamana, H. (2019). URL-based phishing detection using the entropy of non-alphanumeric characters. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019) (pp. 1–8). https://doi.org/10.1145/3366030.3366064