Name of Learners:- Shailaja Ramteke
Batch No:- 310
IT Vedant Institute of Bangalore
Internship Report Task 4

==

Objective: Enhance the dataset with new features and improve predictive power.
Tasks: ○ Perform feature selection using correlation and feature importance. ○ Create new features based on domain knowledge (e.g., length of URL, presence of special characters).
Deliverables: A refined dataset with engineered features and documentation

Week 4 Deliverables

Feature Selection Report: Correlation Analysis:
A heatmap or table summarizing feature correlations, highlighting highly correlated features and their implications for redundancy or multicollinearity.
Feature Importance Analysis:- A ranked list of features based on importance scores derived from techniques such as tree-based models, mutual information, or statistical tests.
Identification of irrelevant or less impactful features for potential removal.
Feature Engineering Summary:- New Features Created:
A list and description of newly created features, explaining their relevance to the business problem.
Examples include the length of the URL, number of special characters, or domain age for phishing detection.
Impact of New Features:
Preliminary analysis (e.g., correlation with the target variable or basic EDA) showing the potential contribution of the new features.
Refined Dataset:-A refined dataset with selected features and engineered features added.
Documentation of the feature selection process and rationale for keeping or discarding specific features.
Insights and Recommendations:- Key findings from the feature selection process, such as features strongly correlated with the target variable or new features with high predictive potential.

Report

Report Juyperte Notebook link :-https://github.com/Shaila92/INTERNSHIP-TASK4

Objective: Enhance the dataset with new features and improve predictive power.

Feature Selection:- Calculate correlation between features and the target variable to identify highly correlated features.

Use feature importance from a model (e.g., Random Forest or XGBoost) to rank features and eliminate low-importance ones.

Feature Engineering:- Create new features using domain knowledge about phishing detection:

url_length: Length of the URL.

- has_ip: Binary flag — whether the URL contains an IP address.
- count_dots: Number of dots in the URL.

  - count_hyphens: Number of hyphens in the URL.
  - has_at_symbol: Binary flag — whether the URL contains '@'.
  - count_special_chars: Count of special characters like ?, =, &, %, etc.

Deliverables:- A refined dataset with:

- Only selected features (after feature selection).
- Additional engineered features.

Feature Selection Report :-

The dataset has **11,430 rows** and **89 columns**. The target variable is **status** (values: phishing / legitimate).

1)Correlation Analysis:-

Performed correlation analysis to identify features strongly associated with the target variable (status: phishing = 1, legitimate = 0).

Top correlated features

Feature	Correlation
google_index	+0.73
page_rank	-0.51
nb_www	-0.44
ratio_digits_url	+0.36
domain_in_title	+0.34
nb_hyperlinks	-0.34
phish_hints	+0.34
domain_age	-0.33
ip	+0.32
nb_qm	+0.29

Features like google_index, page_rank, and nb_www show strong relationships with phishing status.

Correlation heatmap:- Visualized correlations among the top features and target using a heatmap to check for redundancy and multicollinearity:

- Features with high inter-correlation might introduce multicollinearity.
- This can impact certain models (e.g., logistic regression).

Random Forest Feature Importance

Trained a Random Forest to compute feature importance:

Feature	**Importance**
google_index	0.171
page_rank	0.114
nb_hyperlinks	0.088
web_traffic	0.073
domain_age	0.034
nb_www	0.032
longest_word_path	0.029
phish_hints	0.028
safe_anchor	0.027
ratio_extHyperlinks	0.027

The results align with correlation analysis, highlighting the same key features.

Implications

- Features with high correlation (e.g., google_index) are critical for prediction.
- Strongly inter-correlated features should be monitored for multicollinearity in linear models.
- Random Forest importance confirms these features contribute most to model decisions.

2)Feature Importance Analysis:

The goal of this task was to rank features by their predictive power using different techniques and identify irrelevant or less impactful features for potential removal.

Techniques Used Random Forest Feature Importance and Mutual Information (MI) with target

Random Forest Importance (Top 10 Features)

Feature	Importance
google_index	0.171
page_rank	0.114

Feature	Importance
nb_hyperlinks	0.088
web_traffic	0.073
domain_age	0.034
nb_www	0.032
longest_word_path	0.029
phish_hints	0.028
safe_anchor	0.027
ratio_extHyperlinks	0.027

Mutual Information (Top 10 Features)

Feature	MI Score
google_index	0.21
page_rank	0.14
nb_hyperlinks	0.09
web_traffic	0.08
domain_age	0.05
nb_www	0.04
phish_hints	0.03
safe_anchor	0.03
ratio_extHyperlinks	0.03
longest_word_path	0.03

Low-Impact / Irrelevant Features Identified:- Features with very low importance (< 0.005 in RF or MI):

- Examples: nb_eq, nb_semicolons, prefix_suffix, path_extension, etc.
- These features contribute little to predictive power and can be considered for removal to simplify the model.

Implications

- Removing low-importance features can reduce dimensionality, improve model efficiency, and reduce overfitting risks.
- Important features (like google_index, page_rank, nb_hyperlinks) should be retained in the refined dataset.

3)Feature Engineering: -

New Features Created

Feature	Description	Relevance to Phishing Detection
url_length	Total length of the URL string	Phishing URLs often have long and complex URLs to confuse users
num_dots	Number of dot characters (.) in the URL	Excessive dots may indicate subdomain misuse common in phishing URLs
num_hyphens	Number of hyphens (-) in the URL	Many phishing domains use hyphens to mimic legitimate domains
num_special_chars	Count of special characters (e.g., !@#$%^&*()) in the URL	Suspicious URLs often contain special characters to evade filters
has_https	Indicator if URL uses HTTPS (1 if HTTPS present, else 0)	Legitimate sites are more likely to use HTTPS for security

The engineered features were created based on domain knowledge of phishing tactics:

- Phishing sites often use obfuscation techniques like long URLs, hyphens, and special characters.
- Lack of HTTPS is a known red flag in phishing detection.
- These features provide additional signals that improve model prediction of phishing status.

4)Impact of New Features:

Preliminary analysis (e.g., correlation with the target variable or basic EDA) showing the potential contribution of the new features.

Preliminary Correlation Analysis

New Feature	Correlation with status_binary
url_length	+0.42
num_dots	+0.31
num_hyphens	+0.27

New Feature	Correlation with status_binary
num_special_chars	+0.25
has_https	-0.33

Observations:- url_length, num_dots, and num_hyphens show positive correlation with phishing status — phishing URLs tend to be longer and contain more dots/hyphens.

And has_https is negatively correlated — phishing URLs are less likely to use HTTPS.

Boxplot EDA Findings:-

- URL Length: Phishing URLs generally have significantly greater length than legitimate ones.
- Number of Dots / Hyphens: Higher counts are more typical of phishing URLs.
- Special Characters: Phishing URLs tend to contain more special characters.
- HTTPS Presence: A smaller proportion of phishing URLs use HTTPS compared to legitimate URLs.

The preliminary analysis suggests that these new features contribute meaningful signals for detecting phishing URLs. These features should be retained for modeling.

5)Refined Dataset :-

Feature Selection Process

We combined multiple techniques to select the final features:

- Correlation Analysis: Removed highly correlated redundant features.
- Random Forest & Mutual Information Importance: Retained features with significant predictive contribution (importance > 0.005).
- Domain Knowledge: Added engineered features known to enhance phishing detection (e.g., URL structure attributes).

Final Feature List

Type	Feature	Rationale
Original	google_index	Strongest indicator of phishing, high importance
Original	page_rank	Important signal from site reputation
Original	nb_hyperlinks	Significant positive correlation with phishing
Original	web_traffic	Low traffic associated with phishing
Original	domain_age	Younger domains often phishing

Type	Feature	Rationale
Original	nb_www	Phishing sites often manipulate subdomains
Original	longest_word_path	Helps detect obfuscated URLs
Original	phish_hints	Specific indicators of phishing
Original	safe_anchor	Indicates safe linking practices
Original	ratio_extHyperlinks	Excessive external links common in phishing
Engineered	url_length	Longer URLs often phishing
Engineered	num_dots	Subdomain abuse common in phishing
Engineered	num_hyphens	Mimics legitimate domains
Engineered	num_special_chars	Obfuscation via special characters
Engineered	has_https	Phishing sites less likely to use HTTPS

Features Discarded:- Features with very low importance or redundancy:
nb_eq, nb_semicolons, prefix_suffix, path_extension, nb_dots_directory, https_token, nb_slash_directory, etc.

Rationale: These features contributed little predictive value or introduced multicollinearity without improving model performance.

A refined dataset was produced with 15 features + target variable. This balanced set includes high-importance original features and engineered features that enhance predictive power.

6)Insights and Recommendations :-

Key Findings from Feature Selection:- Features strongly correlated with phishing status (status_binary)

- google_index ➞ Highly predictive of phishing; phishing URLs often not indexed by Google.
- page_rank ➞ Low page rank is a strong signal of phishing.
- web_traffic ➞ Low traffic sites are more likely to be phishing sites.
- url_length ➞ Longer URLs are commonly used to obfuscate phishing URLs.
- has_https ➞ Absence of HTTPS is associated with phishing.

Engineered Features with High Predictive Potential:-

- url_length — Showed a positive correlation (+0.42) with phishing status.

- num_dots, num_hyphens — These features revealed significant differences in phishing vs legitimate URLs, with phishing URLs tending to contain more of these.
- num_special_chars — Helped identify suspicious URLs employing obfuscation techniques.

Recommendations:- Keep high-importance original features

- Retain features like google_index, page_rank, web_traffic, nb_hyperlinks, phish_hints, and safe_anchor as they contribute substantial predictive power.

Leverage engineered features in the model

- Incorporate url_length, num_dots, num_hyphens, num_special_chars, has_https as they enhance detection of phishing patterns.

Remove low-impact features

- Discard features with negligible predictive contribution (importance < 0.005) and those contributing to multicollinearity.

Future enhancements

- Consider adding features derived from WHOIS data (e.g., domain registration country, registrar reputation).
- Explore temporal features like URL creation date to catch fast-moving phishing campaigns.

Conclusion:- The combined feature selection and engineering process has strengthened the dataset's predictive capability. The refined features should improve model accuracy and generalizability when detecting phishing websites.

<u>Overall Conclusion for Feature Engineering & Selection Process</u>

Objective :- The goal of this project phase was to enhance the phishing dataset through thoughtful feature selection and engineering, aiming to improve model predictive performance for phishing detection.

Key Outcomes:- Feature Selection :-

- We identified critical features such as google_index, page_rank, web_traffic, and nb_hyperlinks through correlation analysis and feature importance ranking (Random Forest, Mutual Information).
- Features with high multicollinearity or low predictive value (e.g., nb_eq, nb_semicolons, prefix_suffix) were flagged for removal to reduce redundancy and overfitting risk.

Feature Engineering :-

- New features were created based on domain knowledge:
    - url_length, num_dots, num_hyphens, num_special_chars, has_https
- Preliminary analysis showed these features provided meaningful signals:

    - url_length (correlation +0.42) and has_https (correlation -0.33) were particularly valuable in distinguishing phishing from legitimate URLs.

Refined Dataset

- A final refined dataset was created with:
    - 10 key original features (high importance, low redundancy)
    - 5 engineered features (proven value in EDA/correlation)
- Low-impact or redundant features were removed, enhancing model efficiency and interpretability.

Insights & Recommendations:-

- Combining feature selection and engineering strengthened the dataset's predictive capability.
- Engineered features like url_length and num_dots are strong phishing indicators.
- Recommend retaining high-importance features and continuing to explore new features (e.g., WHOIS, temporal attributes) in future work.

The refined dataset is well-prepared for model training and evaluation.
The process balanced predictive power, efficiency, and explainability, laying a strong foundation for building a robust phishing detection model.

==