I210269: Muhammad Faizan Karamat

I210330: Nibtahil Nafees

AI-K

Critical Analysis of Machine Learning Models for Malicious URL Classification

1. Introduction

With the rapid proliferation of malicious URLs, detecting harmful web links is crucial for cybersecurity. This study evaluates three machine learning models—Random Forest (RF), XGBoost, and LSTM—for classifying URLs into five categories: benign, defacement, phishing, malware, and spam. Initially, SVM was considered but was excluded due to difficulties in finding optimal support vectors. BERT was not used in this study due to computational constraints and the focus on traditional ML and deep learning methods.

2. Data Preprocessing and Feature Engineering

Data Cleaning and Imbalance Handling

- The dataset contained 653,046 URLs with five labels.
- Class imbalance was handled using SMOTE to ensure a balanced dataset.
- Duplicates were removed, and missing values were checked.

Feature Extraction

- Structural Features: URL length, presence of special characters, number of digits, and subdomains.
- NLP Features: Tokenization and analysis of top-level domains (TLDs).
- Path Analysis: Extracted path lengths and query parameters.

3. Model Implementation and Performance Analysis

3.1 Random Forest

- **Pros:** Handles feature importance well, resistant to overfitting.
- **Cons:** Struggles with capturing URL sequences and patterns.
- **Accuracy:** 92.3%
- **Observations:** RF performed well but lacked sequential learning capabilities.

3.2 XGBoost

- **Pros:** Robust against overfitting, strong feature selection.
- **Cons**: Computationally expensive, especially on large datasets.
- **Accuracy**: 94.1%
- **Observations:** Achieved the highest accuracy among traditional ML models.

3.3 LSTM

- **Pros:** Effective at sequence learning, capturing patterns in URL structures.

- **Cons**: Requires more computational resources, sensitive to hyperparameters.

- **Accuracy:** 91.8%

- **Observations:** Despite being slightly less accurate than XGBoost, LSTM excelled in identifying malicious behavior within URLs.

4. Results Comparison and Insights

Model	Accuracy	Key Strength	Key Weakness
Random Forest	92.3%	Strong feature importance handling	Lacks sequential pattern recognition
XGBoost	94.1%	High accuracy, robust feature selection	Computationally expensive
LSTM	91.8%	Captures sequential URL patterns	Sensitive to hyperparameters

- XGBoost outperformed RF and LSTM, making it the best-performing model.

- LSTM demonstrated strong sequential learning, beneficial for URL classification but required significant tuning.

- SVM was initially considered but was excluded due to difficulties in finding support vectors for high-dimensional data.

5. Challenges and Improvements

Challenges Faced:

- Computational Complexity: LSTM and XGBoost required significant processing power.

- SVM Performance Issues: It struggled with URL-based features, making it unsuitable.

- Feature Engineering: Ensuring extracted features were relevant to malicious behavior.

Potential Improvements:

- Transformer-Based Models: Future work could include BERT-based approaches for improved performance.

- Hybrid Models: Combining XGBoost and LSTM could yield better results.

- Real-Time Testing: Deploying models in real-world conditions for further validation.

6. Conclusion

Among the models tested, XGBoost achieved the best accuracy (94.1%), followed by Random Forest (92.3%) and LSTM (91.8%). While LSTM performed well for sequential learning, XGBoost balanced accuracy and efficiency effectively. Future enhancements could explore deep learning transformers for even better malicious URL detection.