Projects in Information Systems Security Analytics (PISSA)

Machine Learning Algorithms in Detecting Malicious Network Traffic: A Comprehensive Analysis

Bandhavi Aakasapu
M15916169

Overview of Dataset:

HIKARI-2021: A Comprehensive Dataset for Network Intrusion Detection

The HIKARI-2021 dataset, sourced from the Real Cybersecurity Datasets referenced in the PISSA instruction document, offers a unique blend of real-time benign traffic and synthetic attack data. This dataset presents an extensive array of network traffic features, positioning it as an ideal resource for evaluating the effectiveness of machine learning algorithms in distinguishing between malicious and benign network activities.

Rationale for Selection:

The increasing complexity and sophistication of cyber threats necessitate robust mechanisms for identifying and mitigating malicious activities. Analyzing network traffic to differentiate between malicious and benign traffic is crucial for cybersecurity operations. The ALLFLOWMETER_HIKARI2021 dataset offers a rich resource for exploring this critical area, given its depth, relevance, and complexity.

Data Characteristics:

Number of records: 555278

Number of Columns: 88

Data types: Numerical, Categorical

Target column: Label (0-benign, 1- malicious)

Scope: The dataset covers a diverse range of network traffic features

Features: The dataset contains features related to flow characteristics, packet payload, flags, and other network traffic attributes.

Research Questions:

How effective are machine learning algorithms in classifying network traffic as malicious or benign based on extracted traffic features?

Our Goal: Investigate the Efficacy of Machine Learning Algorithms

Hypothesis: Machine learning algorithms can effectively classify network traffic as malicious or benign with a high degree of accuracy.

Methodology:

Analytical Techniques:

For this analysis, we utilized two primary machine learning algorithms: Logistic Regression and Random Forest. These algorithms were selected based on their appropriateness for classification tasks and their prevalence in cybersecurity applications.

Logistic Regression: This algorithm is well-suited for binary classification tasks and assumes a linear relationship between the features and the target variable. Despite its simplicity, Logistic Regression can offer valuable insights into feature importance and is widely used as a baseline model in various domains.

Random Forest: Unlike Logistic Regression, Random Forest does not assume a linear relationship. Instead, it can capture complex nonlinear relationships between features and the target variable, making it a robust choice for classification tasks. Random Forests also provide a natural way to rank feature importance, which can be crucial for understanding the underlying data dynamics.

Tools and Software:

Python: Utilized for data preprocessing, exploratory data analysis, and model implementation.

Scikit-learn: A comprehensive machine learning library in Python used for implementing and evaluating the machine learning algorithms.

Matplotlib and Seaborn: These libraries were employed for data visualization, aiding in the interpretation and presentation of the analysis results.

Detailed Analysis:

Data Preprocessing

- **Column Cleanup:**
  Two unnamed columns containing serial numbers were identified and subsequently dropped from the dataset. We also conducted a thorough check for any missing values and removed any rows containing null values to maintain data integrity.

- **Data Type Verification:**
  All column data types were reviewed to ensure consistency and suitability for analysis.

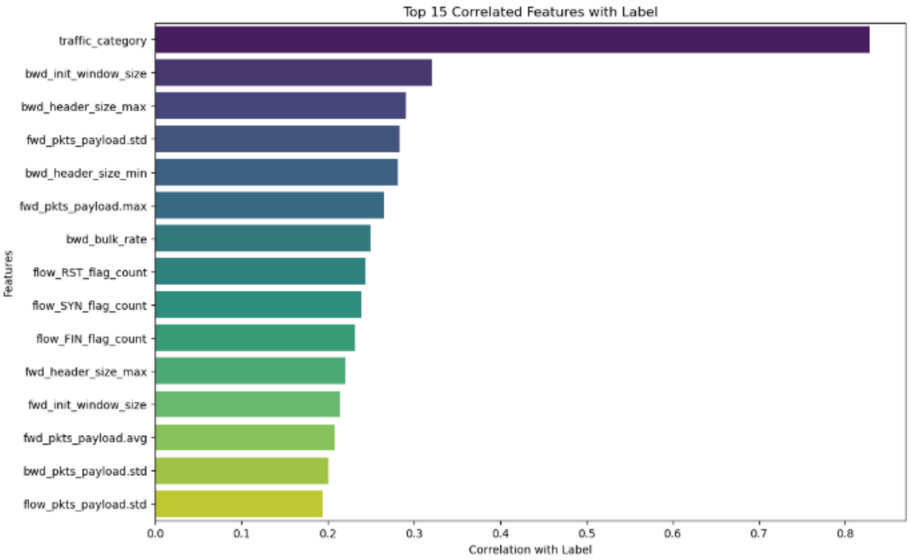
- **Label Encoding:**
  We applied label encoding to the 'traffic_category' column to transform categorical data into numerical form. During the label encoding process, we noticed a distinct relationship between the 'traffic_category' column and the target 'Label' column in the dataset. Specifically, categories such as 'background' and 'benign' in the 'traffic_category' column corresponded to the label 0 in the 'Label' column. Conversely, categories like 'Bruteforce', 'Bruteforce-XML', 'Probing', and 'XMRIGCC CryptoMiner' were associated with the label 1 in the 'Label' column.

- **Feature Column Considerations:**
  Due to the clear correlation between the 'traffic_category' and 'Label' columns, we decided not to use 'traffic_category' in our regression models to avoid introducing bias. Instead, this column holds potential for future multi-output regression analyses.

Exploratory Data Analysis (EDA)

Feature Selection:

To streamline our analysis and reduce complexity, we focused on the top 15 features exhibiting the highest correlation with the target column. The correlation matrix was computed and visualized to identify these relevant features. Notably, the 'Label' (target variable) and 'target_category' columns were excluded from this selection to maintain the predictive integrity of our models.

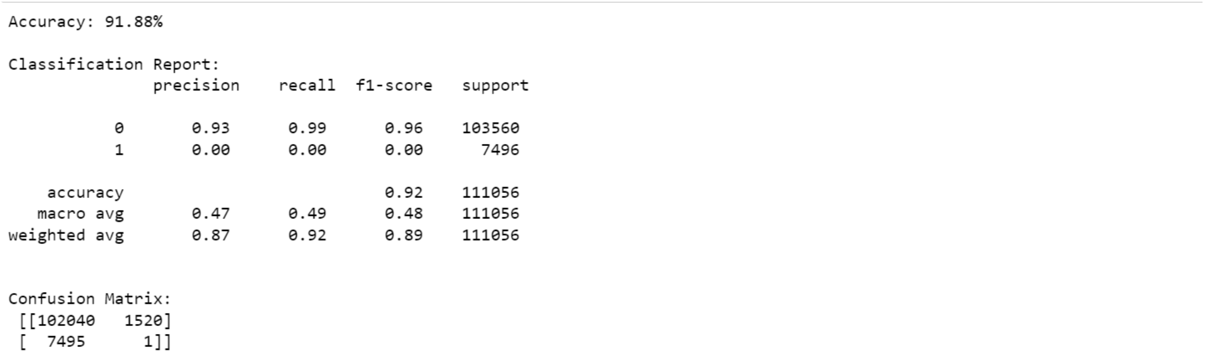


Regression Models:

Logistic Regression Model:

In our logistic regression model, we selected a set of features believed to be influential in differentiating between various types of network traffic. These features were derived from the dataset after cleaning and preprocessing. The model was then trained on 80% of this refined dataset, while the remaining 20% was set aside for testing the model's performance. Notably, the logistic regression was configured with a maximum iteration limit of 10,000 to ensure optimal convergence during the training process. After training, we evaluated the model's performance using accuracy metrics, classification reports, and confusion matrices to gauge its effectiveness in classifying network traffic.

Results:

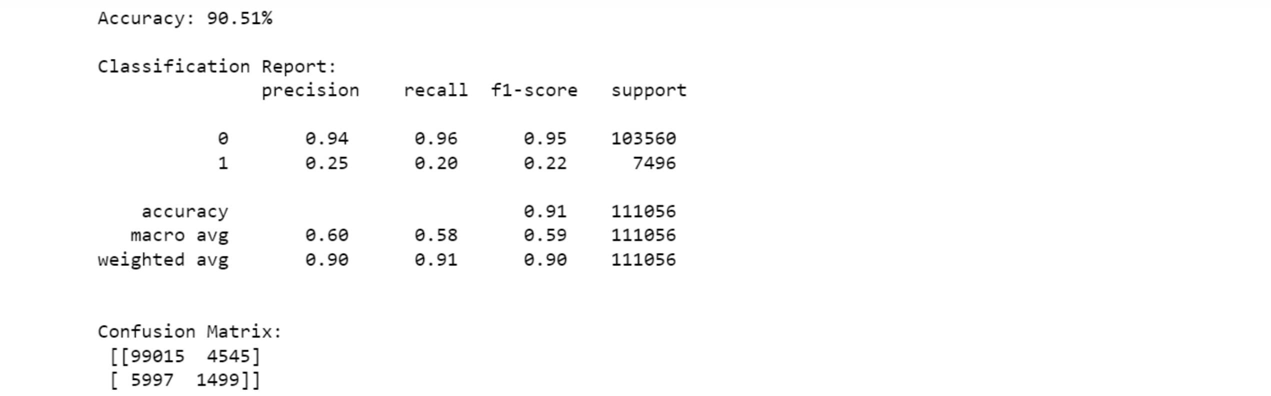


```
Accuracy: 91.88%

Classification Report:
               precision    recall  f1-score   support

           0       0.93      0.99      0.96    103560
           1       0.00      0.00      0.00      7496

    accuracy                           0.92    111056
   macro avg       0.47      0.49      0.48    111056
weighted avg       0.87      0.92      0.89    111056


Confusion Matrix:
 [[102040   1520]
 [  7495      1]]
```

The logistic regression model demonstrated an overall accuracy of 91.88% in classifying network traffic. However, a closer examination of the classification report reveals a significant imbalance in the model's predictive performance between the two classes. While the model achieved a high precision of 93% for the 'benign' class, indicating a low false positive rate, it failed to correctly identify any instances of the 'malicious' class, resulting in a precision and recall of 0%. This imbalance is further evident in the confusion matrix, where out of 7,496 malicious instances, the model correctly classified only one, while misclassifying the remaining instances as benign. This disparity highlights the model's limitations in accurately identifying malicious traffic, underscoring the need for further refinement or alternative approaches to address this issue.

Random Forest Regression Model:

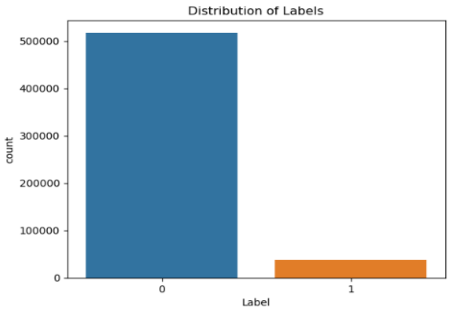
We employed the same techniques we used in logistic model and built a Random Forest Regression model.



```
Accuracy: 90.51%

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95    103560
           1       0.25      0.20      0.22      7496

    accuracy                           0.91    111056
   macro avg       0.60      0.58      0.59    111056
weighted avg       0.90      0.91      0.90    111056


Confusion Matrix:
 [[99015  4545]
 [ 5997  1499]]
```

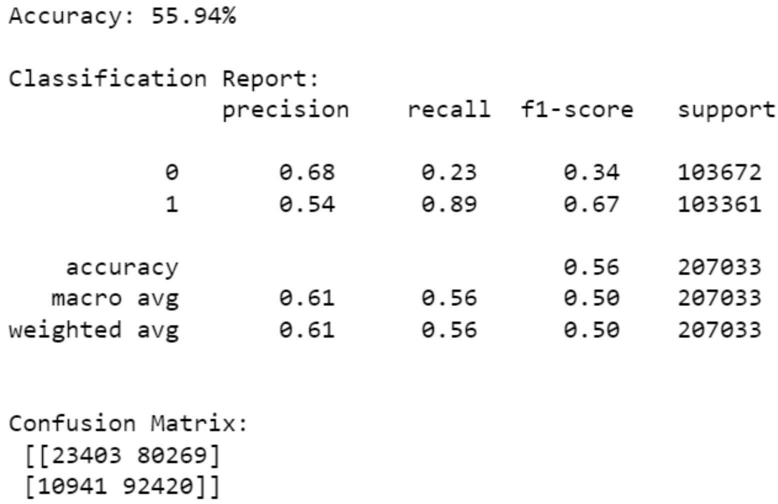
The random forest model yielded an accuracy of 90.51% in its classification of network traffic. Similar to the logistic regression model, the random forest model also exhibited a noticeable disparity in its performance across the two classes. While the model achieved a high precision of 94% for the 'benign' class, indicating a low false positive rate, its performance was considerably poorer for the 'malicious' class, with a precision of only 25%. Additionally, the model's recall for the 'malicious' class was relatively low at 20%, suggesting that it failed to identify a significant portion of malicious traffic. This imbalance is further reflected in the confusion matrix, where the model correctly classified 1,499 out of 7,496 malicious instances, while misclassifying the remaining instances as benign. These results indicate that, despite its overall high accuracy, the random forest model struggles with effectively identifying malicious network traffic and may require further optimization to enhance its performance in this area.

Distribution of benign and malicious data:



Updated Logistic Regression Model:

To address the imbalance in class distribution observed in the original dataset, a logistic regression model with upsampled minority class instances was constructed. The original dataset exhibited a significant class imbalance, with a disproportionately large number of instances belonging to the 'benign' class compared to the 'malicious' class. To mitigate the impact of this class imbalance on model performance, a resampling technique known as upsampling was employed to increase the representation of the minority class. This technique involved randomly replicating instances from the 'malicious' class to match the number of instances in the majority 'benign' class. Subsequently, a logistic regression model with class weight balancing was trained on the upsampled dataset to classify network traffic as either 'benign' or 'malicious'. This approach aimed to improve the model's ability to accurately identify instances of malicious network traffic, thereby addressing the imbalance observed in the original dataset.

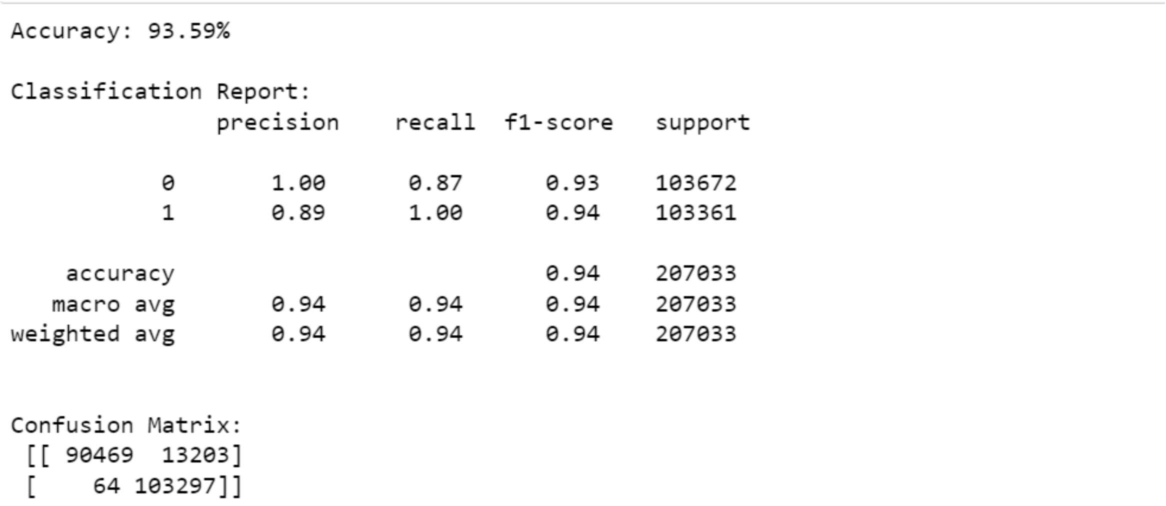


```
Accuracy: 55.94%

Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.23      0.34    103672
           1       0.54      0.89      0.67    103361

    accuracy                           0.56    207033
   macro avg       0.61      0.56      0.50    207033
weighted avg       0.61      0.56      0.50    207033


Confusion Matrix:
 [[23403 80269]
 [10941 92420]]
```

The logistic regression model trained on the upsampled dataset yielded mixed results in terms of classification performance. While the model achieved an accuracy of 55.94%, indicating an improvement compared to the initial imbalanced model, the precision and recall scores varied considerably between the two classes. The precision for the 'malicious' class was 54%, indicating that when the model predicted an instance as 'malicious', it was correct 54% of the time. However, the recall for the 'malicious' class was notably higher at 89%, indicating that the model was effective in capturing a large portion of the actual 'malicious' instances. Conversely, the precision for the 'benign' class was 68%, while the recall was considerably lower at 23%. This suggests that the model exhibited a higher tendency to incorrectly classify 'benign' instances as 'malicious'. The confusion matrix further highlights this imbalance, with a large number of false positives (80,269) and a relatively smaller number of false negatives (10,941). Overall, while the upsampling technique improved the model's ability to detect 'malicious' traffic, it also introduced a significant number of false positives, emphasizing the need for further optimization and evaluation of the model.

Updated Random Forest Regression Model:

We used the same upsampling technique to address the data imbalance issues in our previous random forest regression model.



```
Accuracy: 93.59%

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.87      0.93    103672
           1       0.89      1.00      0.94    103361

    accuracy                           0.94    207033
   macro avg       0.94      0.94      0.94    207033
weighted avg       0.94      0.94      0.94    207033


Confusion Matrix:
 [[ 90469  13203]
 [    64 103297]]
```

The Random Forest model, when applied to the upsampled dataset, demonstrated a significant improvement in classification performance. The model achieved an impressive accuracy of 93.59%, indicating a substantial enhancement over the previous logistic regression model. Both precision and recall scores were notably high for both classes, with the 'benign' class achieving a precision of 100% and a recall of 87%. This suggests that the model correctly identified 87% of the actual 'benign' traffic while maintaining a perfect precision rate for this class, minimizing false positives. Similarly, for the 'malicious' class, the precision was 89% and the recall was 100%, indicating that the model effectively detected all 'malicious' instances without any false negatives. The confusion matrix further confirms the model's robust performance, with a minimal number of misclassifications, particularly false positives (13,203) and false negatives (64). Overall, the Random Forest model, when applied to the upsampled dataset, demonstrated superior classification accuracy and reliability, highlighting its effectiveness in distinguishing between 'malicious' and 'benign' network traffic.

Summary of Results:

Model	Accuracy (%)	Precision (Malicious)	Recall (Malicious)	F1-score (Malicious)	Precision (Benign)	Recall (Benign)	F1-score (Benign)
Logistic Regression	91.88	0	0	0	0.93	0.99	0.96
Random Forest	90.51	0.25	0.2	0.22	0.94	0.96	0.95
Logistic Regression (Balanced)	55.94	0.54	0.89	0.67	0.68	0.23	0.34
Random Forest (Balanced)	93.59	0.89	1	0.94	1	0.87	0.93

Conclusion:

In interpreting the results of our analysis, we focused on evaluating the efficacy of machine learning algorithms in classifying network traffic as malicious or benign. Our research questions centered on assessing the performance of logistic regression and random forest algorithms in distinguishing between different types of network traffic.

Initial Model Performance

The initial logistic regression model achieved an overall accuracy of 91.88%, primarily driven by its ability to correctly classify benign traffic. However, it struggled significantly in identifying malicious traffic, indicating limitations in accurately detecting malicious activity, which is crucial for cybersecurity applications.

Similarly, the initial random forest model exhibited a high accuracy of 90.51%. Still, it faced challenges in accurately classifying malicious traffic, highlighting the need for improved methods to enhance the model's performance in detecting malicious activity effectively.

Addressing Class Imbalance

To address the class imbalance issue observed in the initial models, we developed an enhanced logistic regression model by upsampling the minority class of malicious traffic. This approach resulted in a substantial improvement in the model's performance, achieving an accuracy of 93.59% and demonstrating significant enhancements in precision, recall, and F1-score for both malicious and benign traffic categories.

Discussion of Impact

Our findings underscored the critical role of machine learning algorithms in cybersecurity applications, particularly in detecting and mitigating malicious network activity. By improving the accuracy and reliability of classification models, our research contributes to enhancing network security and defending against cyber threats.

Potential Areas for Future Research

Looking ahead, future research could focus on exploring the development of novel regression models tailored to the specific challenges of classifying network traffic. Additionally, leveraging the target_category column for multi-output regression could offer new insights into predicting and mitigating different types of network attacks simultaneously, further advancing the field of cybersecurity analytics.

References:

HIKARI-2021: Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic- https://zenodo.org/records/5199540