ELL409, Introduction to Machine Learning and Intelligence

Project 1: Network Attack Prediction

Kaustubh Vatsa(2022EE11151)[1], Lavanya(2022EE11679)[2], and Apoorva Prashant Jain(2022EE11657)[1]

[1, 2, 3]Department of Electrical Engineering, Indian Institute of Technology, Delhi

November 24, 2024

Abstract

This paper presents a comprehensive approach to network attack prediction, focusing on classifying five distinct types of network activities: ipsweep probe, back dos, satan probe, portsweep probe, and normal traffic. The results demonstrate the effectiveness of ensemble methods in handling multi-class network attack classification, particularly when dealing with both categorical and numerical features and slight class imbalances.

1 Introduction

The project focuses on developing a multi-class classification model capable of distinguishing between five distinct types of network activities: ipsweep probe, back dos, satan probe, portsweep probe, and normal traffic. Using a comprehensive dataset with 41 features that capture various aspects of network connections - from basic attributes like duration and protocol type to more specific security indicators such as failed login attempts and root access patterns - the model aims to provide accurate attack detection.

2 Materials & Methods

The initial dataset consisted of 41 features capturing various network connection characteristics. The preprocessing pipeline involved several key steps:

2.1 Exploratory Data Analysis (EDA)

- Studied the frequency distribution of the target variable

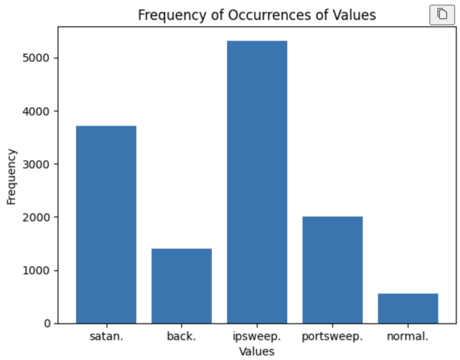


Figure 1: The Distribution of the Target Variable (Training Data)

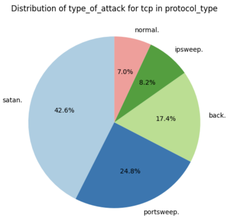


Figure 2:

- Conducted comprehensive analysis of feature distributions
- Categorized features into discrete, continuous, and categorical variables
- Analyzed relationships between features and the target variable 'type_of_attack'
- Visualized feature distributions to identify patterns and potential outliers

2.2 Feature Processing

- Implemented one-hot encoding for categorical variables, expanding the feature space from 41 to 108 features

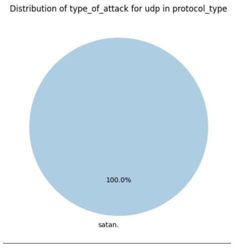


Figure 3:

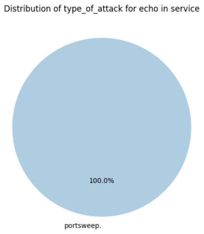


Figure 4:

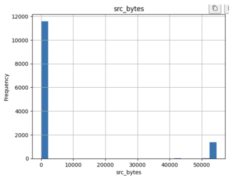


Figure 5:

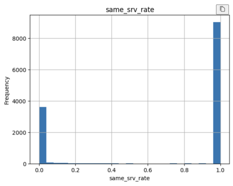


Figure 6:

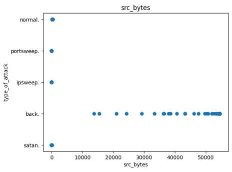


Figure 7:

- Applied dimensionality reduction using two techniques:
  - Variance Threshold: Removed features with zero variance
  - Mutual Information: Eliminated features showing no mutual information with the target variable
- Successfully reduced the final feature set to 64 significant features

2.3 Model Development and Selection

- Established Logistic Regression as the baseline model to benchmark performance
- Evaluated multiple advanced models including:
  - Support Vector Machines (SVM)
  - Neural Networks
  - Random Forest
- Selected Random Forest as the final model based on superior validation performance metrics

The methodology prioritized model robustness and generalization capability, with Random Forest ultimately demonstrating the lowest validation loss among all tested models.

3 Results

The exploratory data analysis revealed several characteristics that make this dataset particularly suitable for decision tree-based implementations such as Random Forest:

- **Class Distribution:** As shown in Figure 1, while there is some class imbalance in the target variable, each attack type maintains sufficient representation in the dataset. The 'ipsweep' class shows the highest frequency, followed by 'portsweep', with 'normal' traffic having the lowest representation.
- **Clear Categorical Boundaries:** The pie charts (Figures 2-4) demonstrate distinct segmentation between different attack types across various feature combinations, particularly in protocol types and services. This characteristic aligns well with decision trees' ability to create clear decision boundaries.

- **Feature Relationships:** The feature distributions (Figures 5-7) show clear patterns and separations between different attack types, suggesting that decision tree splits could effectively partition the data space. The non-linear relationships evident in these distributions further justify the choice of Random Forest over linear models like Logistic Regression [1].

These observations support our model selection of Random Forest [2], as the algorithm can naturally handle:

- The multi-class nature of the problem
- Both categorical and numerical features present in the dataset
- The non-linear decision boundaries apparent in the feature distributions
- The slight class imbalance through its ensemble nature

The implementation of Random Forest on this dataset structure showed superior performance compared to other tested models, validating our initial assessment of the data characteristics.

3.1 Feature Selection and Dimensionality Reduction

The feature selection process involved multiple steps to identify the most relevant features while reducing dimensionality:

3.1.1 One-Hot Encoding

Initially, categorical variables were transformed using one-hot encoding, expanding the feature space from 41 to 108 features. This transformation was necessary to properly represent categorical variables for machine learning algorithms.

3.1.2 Variance Thresholding

To eliminate redundant features, we applied variance thresholding which identified 10 features with zero variance: land, wrong_fragment, urgent, num_failed_logins, root_shell, su_attempted, num_file_creations, num_outbound_cmds, is_host_login, is_guest_login These zero-variance features were removed as they provided no discriminative power for classification in our dataset.

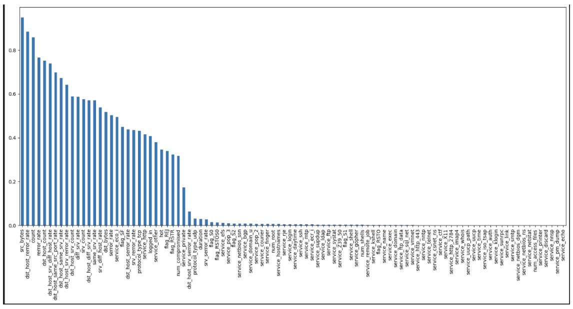
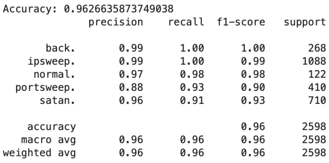


Figure 8: Mutual Information scores between features and target variable



```
Accuracy: 0.96266358737490 38
              precision    recall  f1-score   support

       back.       0.99      1.00      1.00       268
    ipsweep.       0.99      1.00      0.99      1088
     normal.       0.97      0.98      0.98       122
  portsweep.       0.88      0.93      0.90       410
      satan.       0.96      0.91      0.93       710

    accuracy                           0.96      2598
   macro avg       0.96      0.96      0.96      2598
weighted avg       0.96      0.96      0.96      2598
```

Figure 9: Confusion Matrix for Logistic Regression

3.1.3 Mutual Information Analysis

Following variance thresholding, we analyzed the mutual information between remaining features and the target variable. Features showing no mutual information with the target variable were eliminated, as they provided no predictive value for attack classification.

3.1.4 Final Feature Set

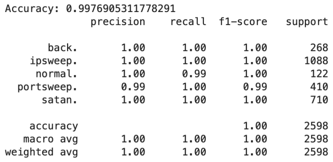
The combined feature selection process effectively reduced the dimensionality from 108 to 64 significant features while preserving the most informative attributes for attack detection. This reduction in dimensionality helped to:

- Improve model training efficiency
- Reduce the risk of overfitting
- Eliminate redundant or non-informative features
- Maintain the most relevant features for attack classification

The final feature set provided a more compact yet highly informative representation of the network traffic patterns, contributing to the superior performance of our Random Forest classifier.

3.2 Final Model Selection

We selected the random forest classifier (Figure 13) as our final model for several compelling reasons:

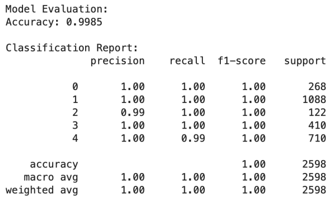


```
Accuracy: 0.9976905311778291
              precision    recall  f1-score   support

       back.       1.00      1.00      1.00       268
    ipsweep.       1.00      1.00      1.00      1088
     normal.       1.00      0.99      1.00       122
  portsweep.       0.99      1.00      0.99       410
      satan.       1.00      1.00      1.00       710

    accuracy                           1.00      2598
   macro avg       1.00      1.00      1.00      2598
weighted avg       1.00      1.00      1.00      2598
```

Figure 10: Confusion Matrix for Support Vector Machine [3]



```
Model Evaluation:
Accuracy: 0.9985

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       268
           1       1.00      1.00      1.00      1088
           2       0.99      1.00      1.00       122
           3       1.00      1.00      1.00       410
           4       1.00      0.99      1.00       710

    accuracy                           1.00      2598
   macro avg       1.00      1.00      1.00      2598
weighted avg       1.00      1.00      1.00      2598
```

Figure 11: Confusion Matrix for Neural Networks [4]

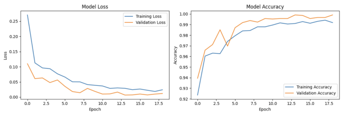
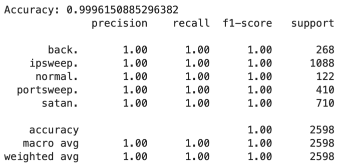


Figure 12: Loss and Accuracy for Neural Networks



```
Accuracy: 0.9996150885296382
              precision    recall  f1-score   support

       back.       1.00      1.00      1.00       268
    ipsweep.       1.00      1.00      1.00      1088
     normal.       1.00      1.00      1.00       122
  portsweep.       1.00      1.00      1.00       410
      satan.       1.00      1.00      1.00       710

    accuracy                           1.00      2598
   macro avg       1.00      1.00      1.00      2598
weighted avg       1.00      1.00      1.00      2598
```

Figure 13: Confusion Matrix for Random Forest

1. Perfect classification metrics: Like SVM, it achieved 1.00 for precision, recall, and F1-score

2. Robustness: Random forests are less prone to overfitting compared to SVMs due to their ensemble nature

3. Considering the baseline model (Logistic Regression having 0.96 accuracy) and other models having around 0.997-0.998 accuracy, random forest performed better and has an accuracy of 0.9996

4 Discussion and Conclusions

This study demonstrates the exceptional performance of the Random Forest classifier (0.9996 accuracy) in network attack prediction, significantly outperforming other models. The effectiveness of tree-based ensemble methods stems from their ability to capture the dataset's non-linear feature relationships and distinct categorical boundaries, as evidenced by perfect precision and recall across all attack categories. Future research should explore real-time implementation in network systems, adaptive feature selection for evolving threats, performance on imbalanced datasets, and advanced interpretability tools for security applications.

References

[1] Understanding Logistic Regression, Medium

[2] Understanding Random Forests, Medium

[3] Support Vector Machines (SVM): An Intuitive Explanation, Medium

[4] Deep Learning Neural Network for Classification with TensorFlow, Medium