TCP Network: Application Name Prediction

Yagnesh Patel
Ahmedabad University
Ahmedabad, India
yagnesh.p@ahduni.edu.in

Nityam Dixit
Ahmedabad University
Ahmedabad, India
nityam.d@ahduni.edu.in

Kavan Gondalia
Ahmedabad University
Ahmedabad, India
kavan.g@ahduni.edu.in

Tej Thakar
Ahmedabad University
Ahmedabad, India
tej.t@ahduni.edu.in

***Abstract*—Many research have lately been conducted on network traffic analysis and forecasting, which has a wide range of applications. Many experiments are conducted, reported on, and used to address a variety of concerns with contemporary computer network applications. In this project, the TCP network dataset is the main emphasis, and we use a variety of techniques to try to forecast the application. Application prediction is a critical task in network security and monitoring since it aids managers in determining the nature of traffic flowing over the network and can be further implemented for selecting an appropriate congestion control protocol. The goal of this project is to assess the effectiveness of various algorithms for predicting the application using TCP network data, including Random Forest, Logistic Regression, and Decision Tree. Accuracy of the models will be examined. Also, the measure of different features will be measured. The project's objective is to gain a deeper understanding of the algorithms' effectiveness in predicting the application from TCP network data and to identify the features that significantly affect the accuracy of the prediction model.**

***Index Terms*—TCP Network, Computer Networks, Classification, Congestion Control**

I. Introduction

The objective of the project is to propose a machine learning approach for application prediction on TCP Network Dataset. The features we observed from the flows included time, transport-layer data, packet and byte counts, and so on. Random Forest and Logistic Regression are two machine learning models that have been built in this study to forecast the application. Accuracy and F1 scores were two of the several metrics used to gauge how effectively these models performed.

II. Literature Review

This research suggests a machine learning-based method for identifying Applications. The paper proposes the use of several linear and non-linear machine learning methods, such as logistic regression, decision trees and random forest, on a dataset of network traffic recorded from a local network for classification. Analysis and Prediction of Real Network Traffic suggest the use of a different linear model that can be used for predicting the flow of the network. According to the results, the suggested method outperforms standard signature-based network classification.

III. Implementation

A. Data preprocessing

Data cleaning and preprocessing are done to make it usable for the machine learning algorithms. In this removal of irrelevant or missing data, also conversion of non-numeric data to numeric form. We also removed the features that were directly connected to the classes.

B. Under Sampling

Under Sampling is done to make data unbiased. For Example, in our dataset Google(9,59,110) have the highest number of packets while NFS(1) has the least number of packets. Under sampling helps in removing the bias and classes which have less number of packets.

C. Feature Selection

We selected some features to improve our machine learning algorithms in order to get more accuracy. We removed some features such as Flow.ID, Source.IP, Label, Timestamp, Destination.IP and various flags from our dataset to enhance the accuracy. Further the selection was done on the basis of correlation matrix.

D. Label Encoding

Label Encoding is a technique used for handling categorical variables. It is used in converting string to numeric form. Here, we used in order to convert the Protocol Name(string) to encoded numeric form.

E. Algorithms

We have implemented following machine learning algorithms on our dataset for our classification,

1) Logistic Regression
2) Decision Tree
3) Random Forest

(A) Logistic Regression
   a) With Feature Selection
   b) With PCA
   c) With K-Cross Fold Validation

(B) Decision Tree
   a) With Feature Selection
   b) With PCA
   c) With K-Cross Fold Validation

(C) Random Forest
   a) With Feature Selection
   b) With PCA
   c) With K-Cross Fold Validation

	Random Forest	Logistic Regression	Decision Tree
Feature Selection	0.981412	0.793118	0.464735
PCA	0.686647	0.299088	0.201471
K-Cross Folding	0.945059	0.183459	0.464312
Original	0.955647	0.180941	0.466971

Fig. 1. Accuracy of each model

	Random Forest	Logistic Regression	Decision Tree
Feature Selection	0.062971	0.059706	0.058088
PCA	0.057265	0.059706	0.058088
K-Cross Folding	0.945059	0.183459	0.464312
Original	0.955647	0.059706	0.058088

Fig. 2. F1 Score of each model

F. Model Selection

Random Forest is often used for high-dimensional datasets. It is known for handling complex relationships among features. Random Forest generates large number of decision trees and produces the final output by combining them. It is not prone to overfitting as each tree is generated using random subset of features. For our dataset, Random Forest exhibits high accuracy, precision, recall and F1 score compared to other implemented algorithms.

G. PCA

PCA is a method used in machine learning and data analysis to reduce the dimensionality of data. In order to capture the most variance, it converts high-dimensional data into a lower-dimensional representation.

H. HyperParameter Tuning

Hyperparameter tuning is the process of selecting the best hyperparameter values for a machine learning model. Some hyperparameters in random forest are:

- Number of trees
- Maximum depth
- Minimum number of samples required to split a node
- Maximum number of features for best split

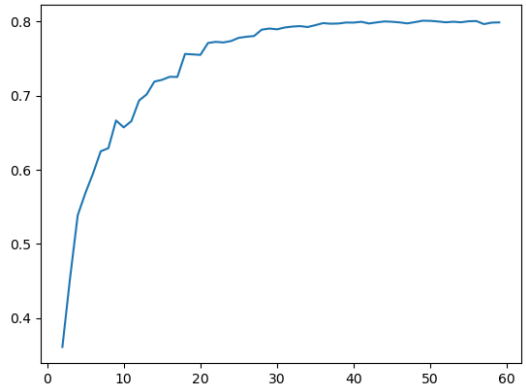

Fig. 3. Accuracy vs No. of Features

IV. Results

As we can infer from the graph, the accuracy of random forest model increases as the feature count increases. For random forest, PCA gives less accuracy than that with feature selection. Hence, our feature selection technique i.e. using correlation matrix gives better model performance. After Feature Selection, we were left with 28 features. Also, from above graph we can infer that after 28 features the accuracy almost remains constant.

V. Conclusion

In summary, the study evaluated the performance of three machine learning models - Logistic Regression, Decision Tree, and Random Forest - based on accuracy, F1-score, precision, and recall rate. Random Forest had the highest accuracy, F1-score, and precision, indicating its better predictive ability. The accuracy was highest for Random Forest, followed by Decision Tree and Logistic Regression. Further analysis revealed that a train-test split ratio of 20:80 prevented overfitting in the Random Forest model. Also, Logistic Regression is prone to overfitting for the given train-test split ratio. These findings can help in selecting the best machine learning model for a particular problem and dataset. The results may vary based on the dataset.

References

[1] Rojas, J.S. (2018) IP network traffic flows labeled with 75 apps, Kaggle. Available at: https://www.kaggle.com/datasets/jsrojas/ip-network-traffic-flows-labeled-with-87-apps (Accessed: March 11, 2023).
[2] Joshi, M. & Hadi, T. H. (2015). A review of network traffic analysis and prediction techniques. arXiv preprint arXiv:1507.05722. https://arxiv.org/abs/1507.05722
[3] Zhani, Mohamed Faten Elbiaze, Halima & Kamoun, Farouk. (2009). Analysis and Prediction of Real Network Traffic. Journal of Networks. 4. 10.4304/jnw.4.9.855-865.