Artificial Intelligence

Mini-Project 1
Modern Low Footprint Cyber Attack Detection

Sai Sri Meghana Dharmapuri (220250135)

Modern Low Footprint Cyber Attack Detection

Sai Sri Meghana Dharmapuri

Computer Science department, California State University Sacramento
6000 J street, Sacramento, CA - 95825

***Abstract*— This paper aims in building a binary classification model that can identify good normal connections and bad connections like intrusions or attacks. Data has been played with models like SVM, KNN, Logistic Regression and Fully connected Neural Networks.**

***Keywords*— Intrusion Detection, Data Pre-processing, Data Analysis, Logistic Regression, SVM, Neural Network, Nearest Neighbor(KNN)**

I. INTRODUCTION

Information is considered as an organisation's biggest asset. Information is generally handled on network-based systems, due to which network intrusion detection holds an integral part in today's world. An intrusion detection system can

A network intrusion is any unauthorized activity on a computer network. Network intrusions or attacks can be defined as a set of events, transmitted through network packets that cannot be detected or analysed by firewalls present today. Due to this sensitive information, integrity and availability of information are compromised

Intrusion detection systems (IDS's) are the watchdogs of information systems. IDS's are software and hardware systems designed and programmed to automate the process of monitoring events happening in a computer system or network and analysing them for potential security issues. The goal of an IDS is to detect cyberattacks by analysing the signature of data packets as they traverse the network.

II. METHODOLOGY

The dataset (UNSW-NB 15) was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. The training and the test data have been given separately.

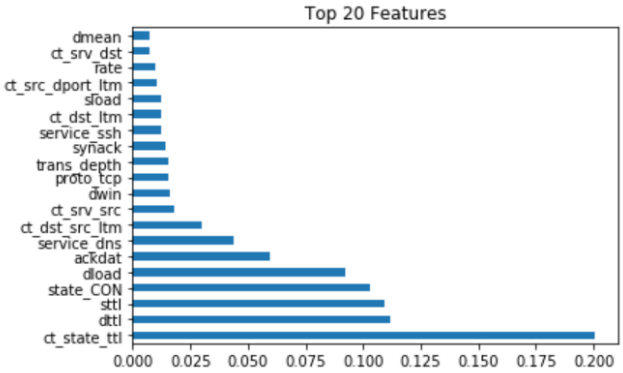
A. Data Preprocessing

Data Preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Data processing services require skilled professionals to apply different techniques for analysing and processing data. For every business organization, data has become the most important tool to make critical decisions.

The first step we performed was to import our datasets, which are available in the form of csv files and we stored them into dataframes. A Dataframe is the most common Structured API and simply represents a table of data with rows and columns. We cross checked if the file was imported properly and then performed a search for null values and removed the rows containing null values from both the training data and the test data. In order to train the model with no bias, it is important that there exists no duplicate data. We then perform one hot encoding for all the categorical variables, they are converted into a form that could be provided to ML algorithms to do a better job in prediction (categorical data are variables that contain label values rather than numeric values). Normalisation is then performed on numeric attributes to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. While there are other normalisation techniques, we performed Zscore normalisation for our data.

B. Feature Selection

Feature selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. It is performed so that your model is not complicated with unnecessary features increasing the training time. We tried different feature selection techniques and we implemented ExtraTreesClassifier, this aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. At first, we built and trained the model, then we computed the importance of each feature and selected the top 20 features and plotted the model with barh plot. We used the first 15 columns based on largest values in descending order. The importance of features might have different values because of the random nature of feature samples.



III. EXPERIMENTS, RESULTS AND ANALYSIS

We used four machine learning models i.e Logistic Regression, K Nearest Neighbor, Support Vector Machine, fully connected neural networks in order to detect bad connections. To proceed, before starting training the model, we must test the model on some test dataset. To do so, we used sklearn library to split the data to two different datasets, one for the independent features x, and one for the dependent variable y. Then We split the dataset x into two separate sets xTrain and xTest. Similarly, we split the dataset y into two sets as well yTrain and yTest.

A. Logistic Regression

Is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis and is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

	precision	recall	f1-score	support
0.0	0.98	0.74	0.84	9625
1.0	0.91	1.00	0.95	25554
accuracy			0.92	35179
macro avg	0.95	0.87	0.90	35179
weighted avg	0.93	0.92	0.92	35179

B. K nearest neighbor

Is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

Classification Report is:

	precision	recall	f1-score	support
0.0	0.90	0.82	0.86	9625
1.0	0.93	0.97	0.95	25554
accuracy			0.93	35179
macro avg	0.92	0.89	0.90	35179
weighted avg	0.92	0.93	0.92	35179

C. Support Vector Machine

Is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Classification Report is:

	precision	recall	f1-score	support
0.0	1.00	0.71	0.83	9625
1.0	0.90	1.00	0.95	25554
accuracy			0.92	35179
macro avg	0.95	0.85	0.89	35179
weighted avg	0.93	0.92	0.92	35179

Comparison of the metrics for all the three models above:

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	92%	93%	92%	92%
Nearest Neighbor	93%	94%	93%	93%
Support Vector Machine	92%	93%	92%	92%

D. Fully connected Neural Network

Is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria.

On trying different activation layers and changing the optimiser between ‘adam’ and ‘sgd’, we found the best results when the activation layer was ‘tanh’ and the last layer was ‘sigmoid’ with the optimiser ‘sgd’. The accuracy was found to be 96%, which is the highest amongst other observations that were made.

	precision	recall	f1-score	support
0.0	0.98	0.88	0.93	9625
1.0	0.96	0.99	0.97	25554
accuracy			0.96	35179
macro avg	0.97	0.94	0.95	35179
weighted avg	0.96	0.96	0.96	35179

Activation layer	**Optimizer**	**F1 Score**
[model 1] Relu, sigmoid	Adam	93%
[model 2] Relu, sigmoid	sgd	93%
[model 3] Tanh, sigmoid	Adam	92%
[model 4] Tanh, sigmoid	sgd	96% [BEST RESULT}

IV. TASK DIVISION AND PROJECT REFLECTION

A. Task Division

- Data Preprocessing – *Combined effort*
- Training and testing models (KNN & Logistic Regression) – *Abrar Ali*
- Training and testing models (SVM and Neural network) – *S.S.Meghana Dharmapuri*
- Documentation – *Combined effort*

B. Project Reflections

- We learnt how important data pre processing as we worked through the project.
- We understood why data normalization and one hot encoding is performed in depth.
- We noticed that the performance for SVM was slow in comparison to the other models.
- We learnt that the accuracy decreased when we played around with the parameter min_delta. When it

was le-7, we got an accuracy of 94%. When it is le-3, we got an accuracy of 93%

- We observed that on reducing one layer and manipulating the no of neurons, the accuracy for our best model [model 4] has changed from 0.959 to 0.955

V. Challenges

- Choosing a certain method to perform feature selection, we tried correlation matrix, Laso regression and chose to proceed with feature importance property.
- We faced difficulty initially in trying to understand the flow of the project but we got better with it as we encountered errors. We understood how and when to drop a few columns.

VI. Conclusions

We reviewed several influential algorithms for intrusion detection based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to changing malicious behaviors.