**Project proposal**

**Network Intrusion Detection Using Logistic regression model**

**Dataset:**

I am using the UNSW-NB15 dataset, which is a public available dataset on Kaggle. It was developed by the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS), using the IXIA Perfect Storm tool for Cyber Security. The dataset consists of normal and attack records, and includes nine attack categories and malware such as analysis, backdoors, DoS, exploits, generic, reconnaissance, fuzzers for anomalous activity, shellcode, and worms. It has a total of 49 features, including basic features such as source and destination IP addresses and port numbers, as well as more advanced features such as packet and byte counts, duration of the connection, and the type of service being used. The dataset is divided into two parts: the training set and the testing set, with 175,341 and 82,332 records or observations, respectively. The 49 features are described in detail in Table 1.

Table 1: List of features, feature types, and its descriptions

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S/N | Name | Type | Description | **S/N** | Name | Type | Description |
| 1 | srcip | nominal | Source IP address | **25** | trans\_depth | integer | Represents the pipelined depth into the connection of http request/response transaction |
| 2 | sport | integer | Source port number | **26** | res\_bdy\_len | integer | Actual uncompressed content size of the data transferred from the server’s http service. |
| 3 | dstip | nominal | Destination IP address | **27** | Sjit | Float | Source jitter (mSec) |
| 4 | dsport | integer | Destination port number | **28** | Djit | Float | Destination jitter (mSec) |
| 5 | proto | nominal | Transaction protocol | **29** | Stime | Timestamp | record start time |
| 6 | state | nominal | Indicates to the state and its dependent protocol, e.g. ACC, CLO, CON, ECO, ECR, FIN, INT, MAS, PAR, REQ, RST, TST, TXD, URH, URN, and (-) (if not used state) | **30** | Ltime | Timestamp | record last time |
| 7 | dur | Float | Record total duration | **31** | Sintpkt | Float | Source interpacket arrival time (mSec) |
| 8 | sbytes | Integer | Source to destination transaction bytes | **32** | Dintpkt | Float | Destination interpacket arrival time (mSec) |
| 9 | dbytes | Integer | Destination to source transaction bytes | **33** | tcprtt | Float | TCP connection setup round-trip time, the sum of ’synack’ and ’ackdat’. |
| 10 | sttl | Integer | Source to destination time to live value | **34** | synack | Float | TCP connection setup time, the time between the SYN and the SYN\_ACK packets. |
| 11 | dttl | Integer | Destination to source time to live value | **35** | ackdat | Float | TCP connection setup time, the time between the SYN\_ACK and the ACK packets. |
| 12 | sloss | Integer | Source packets retransmitted or dropped | **36** | is\_sm\_ips\_ports | Binary | If source (1) and destination (3)IP addresses equal and port numbers (2)(4) equal then, this variable takes value 1 else 0 |
| 13 | dloss | Integer | Destination packets retransmitted or dropped | **37** | ct\_state\_ttl | Integer | No. for each state (6) according to specific range of values for source/destination time to live (10) (11). |
| 14 | service | nominal | http, ftp, smtp, ssh, dns, ftp-data, irc and (-) if not much used service | **38** | ct\_flw\_http\_mthd | Integer | No. of flows that has methods such as Get and Post in http service. |
| 15 | Sload | Float | Source bits per second | **39** | is\_ftp\_login | Binary | If the ftp session is accessed by user and password then 1 else 0. |
| 16 | Dload | Float | Destination bits per second | **40** | ct\_ftp\_cmd | integer | No of flows that has a command in ftp session. |
| 17 | Spkts | integer | Source to destination packet count | **41** | ct\_srv\_src | integer | No. of connections that contain the same service (14) and source address (1) in 100 connections according to the last time (26). |
| 18 | Dpkts | integer | Destination to source packet count | **42** | ct\_srv\_dst | integer | No. of connections that contain the same service (14) and destination address (3) in 100 connections according to the last time (26). |
| 19 | swin | integer | Source TCP window advertisement value | **43** | ct\_dst\_ltm | integer | No. of connections of the same destination address (3) in 100 connections according to the last time (26). |
| 20 | dwin | integer | Destination TCP window advertisement value | **44** | ct\_src\_ ltm | integer | No. of connections of the same source address (1) in 100 connections according to the last time (26). |
| 21 | stcpb | integer | Source TCP base sequence number | **45** | ct\_src\_dport\_ltm | integer | No of connections of the same source address (1) and the destination port (4) in 100 connections according to the last time (26). |
| 22 | dtcpb | integer | Destination TCP base sequence number | **46** | ct\_dst\_sport\_ltm | integer | No of connections of the same destination address (3) and the source port (2) in 100 connections according to the last time (26). |
| 23 | smeansz | integer | Mean of the row packet size transmitted by the src | **47** | ct\_dst\_src\_ltm | integer | No of connections of the same source (1) and the destination (3) address in in 100 connections according to the last time (26). |
| 24 | dmeansz | integer | Mean of the ?ow packet size transmitted by the dst | **48** | attack\_cat | nominal | The name of each attack category. In this data set , nine categories e.g. Fuzzers, Analysis, Backdoors, DoS Exploits, Generic, Reconnaissance, Shellcode and Worms |
|  |  |  |  | **49** | Label | binary | 0 for normal and 1 for attack records |

Table 2: Distribution of class labels in training and testing subsets

|  |  |  |
| --- | --- | --- |
| Class Label | Training | Testing |
| Analysis | 677 | 2000 |
| Backdoor | 583 | 1746 |
| DoS | 4089 | 12264 |
| Exploits | 11132 | 33393 |
| Fuzzers | 6062 | 18184 |
| Generic | 18871 | 40000 |
| Normal | 37000 | 56000 |
| Reconnaissance | 3496 | 10491 |
| Shellcode | 378 | 1133 |
| Worms | 44 | 130 |

**Research Questions:**

1. Is there a significant correlation or association between the input features (listed in Table 1) and network intrusion detection?
2. How accurately do all 49 input features predict intrusion?
3. How do you select the most significant features for intrusion detection?
4. How accurately do the most significant features predict intrusion?
5. How does the performance of the model using the most significant features compare to the model using all 49 features? Which model performs better?
6. What is the accuracy of the selected model for network intrusion detection using the most significant features?

**Model:**

As i previously mentioned, the class variable has ten categories (See in Table 2). However, for the purposes of our analysis, I will be treating the class variable as a binary variable, with '0' representing normal network activity and '1' representing an attack. Given this dichotomous nature of the class variable, I have decided to use a logistic regression model for the network intrusion detection. I will also carry out tuning of the regularization parameter (lambda) to optimize the performance of the model using k-folds cross-validation. The calculation procedure for the logistic regression model is as follows:

Step 1: Read in the training and testing CSV data files.

Step 2: Train the logistic regression model on the training dataset and estimate its coefficients.

Step 3: Use the trained logistic regression model to predict the class label (i.e., 1 for an attack or 0 for normal network activity) for the testing dataset.

Step 4: Compute the confusion matrix based on the ground truth label and the predicted class label.

Step 5: Calculate various performance metrics, including classification accuracy, sensitivity, specificity, and area under the curve (ROC curve), using the confusion matrix.

\*\*I would appreciate a feedback from you sir.