Name: Zhiqing Xu
Student No.: 1000226249

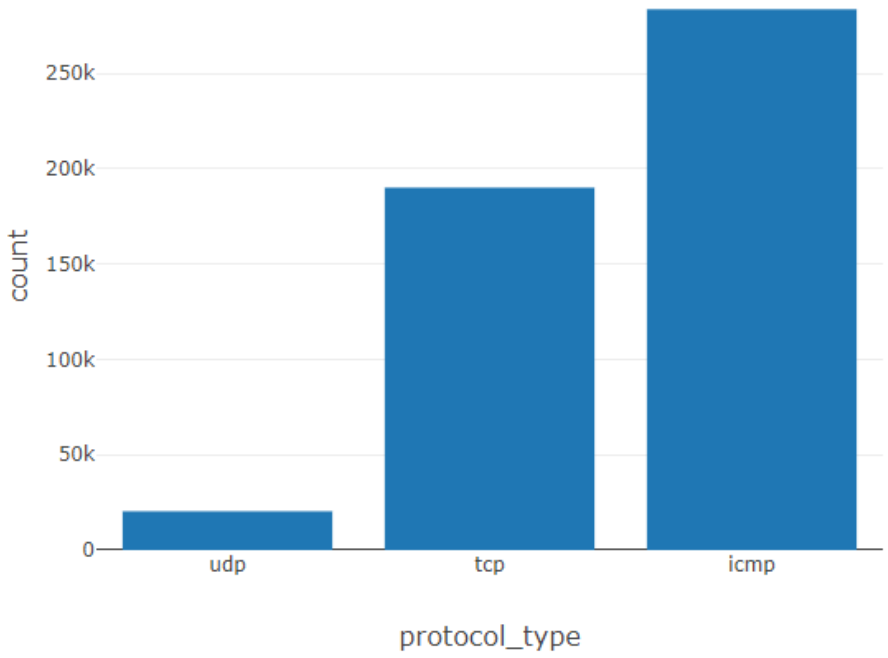
Part A – Question 1:
According to the paper, Intrusion detection system (IDS) is a system that monitors and analyzes data to detect any intrusion in the system or network. It is possible to implement such a system on this dataset. The workflow can be summarized as follows, (1) load and export dataset to Resilient Distributed Datasets (RDD) and and DataFrame in Apache Spark. (2) Process the data. (3) Feature Selection through ChiSqSelector. (4) Train using Support Vector Machine model on the processed dataset. (5) Test the performance and examine the predictive capability.

Question 2-6: See the notebook file for details.

Question 6: See code from the notebook, figures shown here,

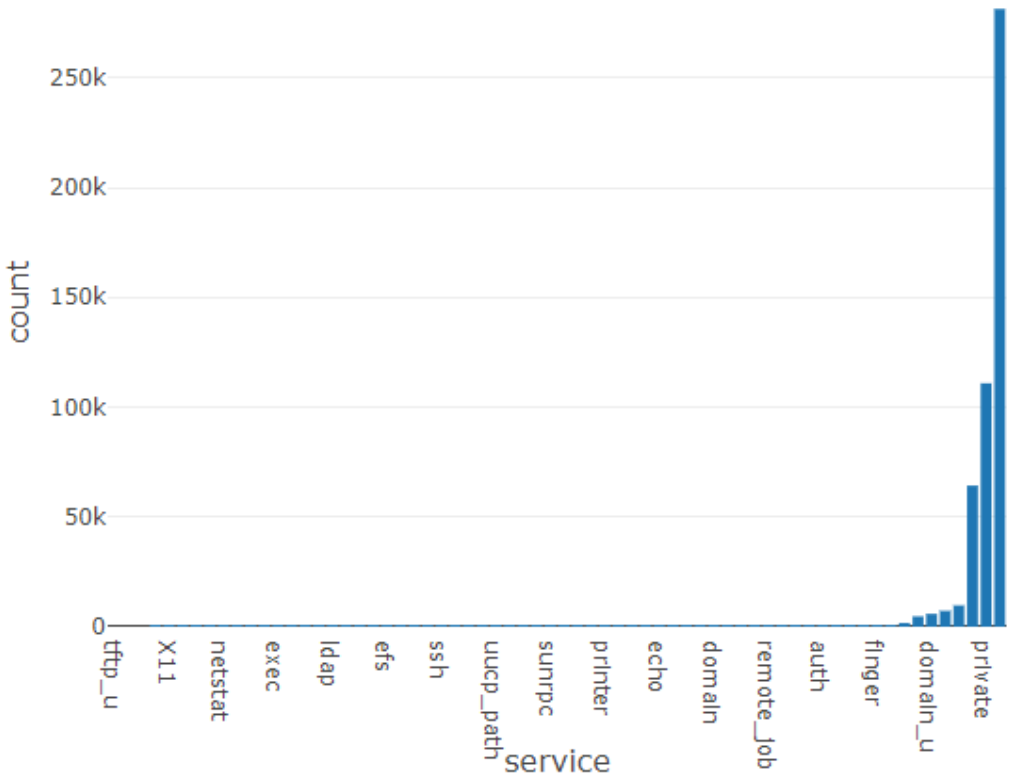
```
Part A - Question 6:


Get the total number of connections based on the protocol_type:
+-------------+------+
|protocol_type|count |
+-------------+------+
|udp          |20354 |
|tcp          |190065|
|icmp         |283602|
+-------------+------+
```



Get the total number of connections based on the service:

service	count
red_i	1
pm_dump	1
tftp_u	1
tim_i	7
X11	11
urh_i	14
IRC	43
Z39_50	92
netstat	95
ctf	97
kshell	98
name	98
netbios_dgm	99
http_443	99
exec	99
ldap	101



Question 7:

Here is the total number of connections based on the label. There are considerable amount of normal and abnormal (attack) connections which can be observed from the figure.

```
Part A - Question 7:


Get the total number of connections based on the label:
+---------------+------+
|label          |count |
+---------------+------+
|smurf          |280790|
|neptune        |107201|
|normal         |97278 |
|back           |2203  |
|satan          |1589  |
|ipsweep        |1247  |
|portsweep      |1040  |
|warezclient    |1020  |
|teardrop       |979   |
|pod            |264   |
|nmap           |231   |
|guess_passwd   |53    |
|buffer_overflow|30    |
|land           |21    |
```

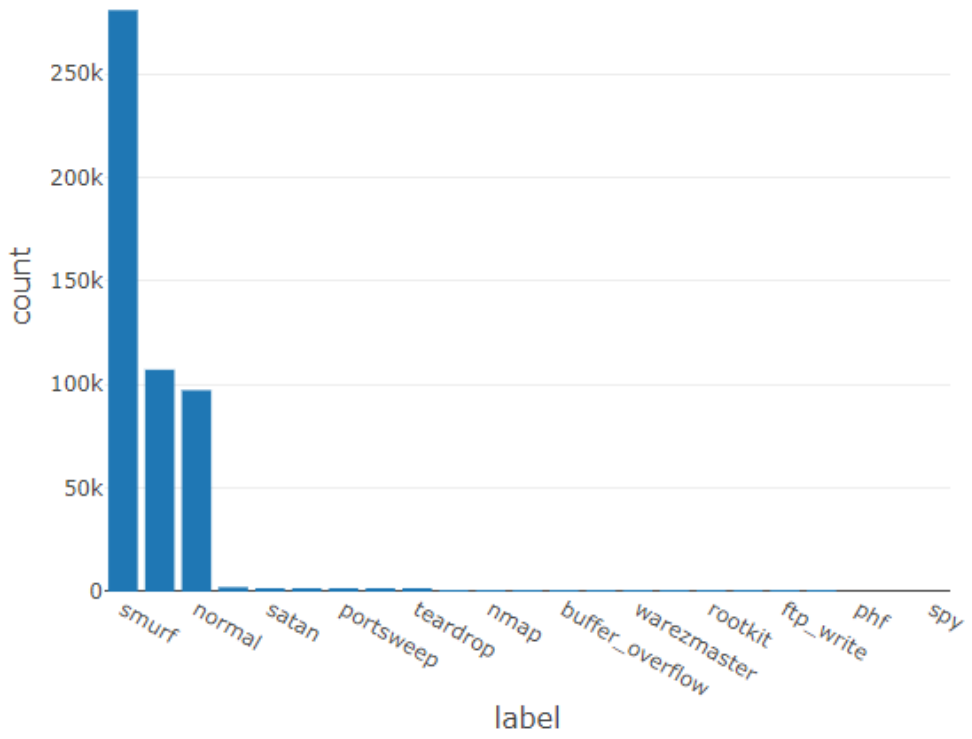
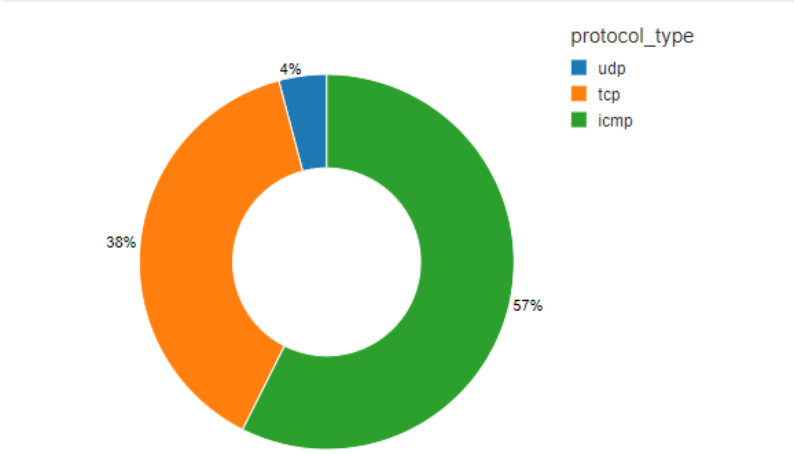


Figure below shows connections counts based on protocol type and normal connection counts based on protocol types. High percentage of icmp connection but low percentage of icmp normal connection shows that icmp protocol type come with lots of connections that are NOT normal. While comparison of the two figures show tcp protocol seems to come with lots of normal connections.

```
Ttotal number of connections based on the protocol type:
+-------------+------+
|protocol_type|count |
+-------------+------+
|udp          |20354 |
|tcp          |190065|
|icmp         |283602|
+-------------+------+


Ttotal number of normal connections based on the protocol type:
+-------------+-----+
|protocol_type|count|
+-------------+-----+
|icmp         |1288 |
|udp          |19177|
|tcp          |76813|
+-------------+-----+
```



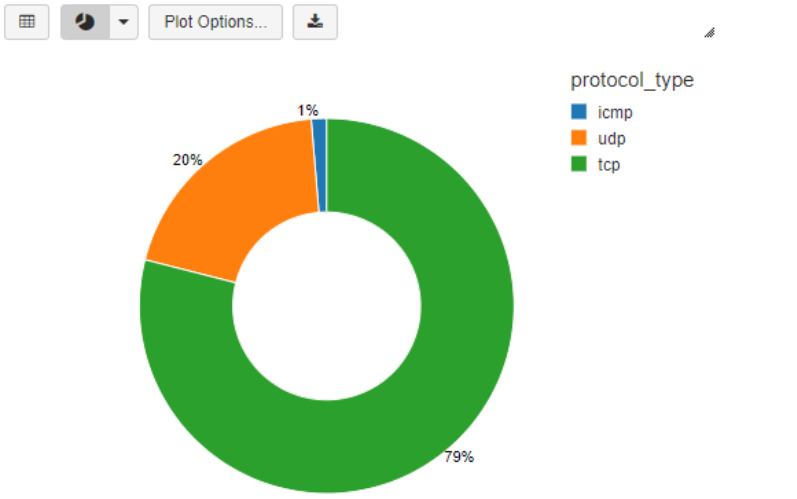
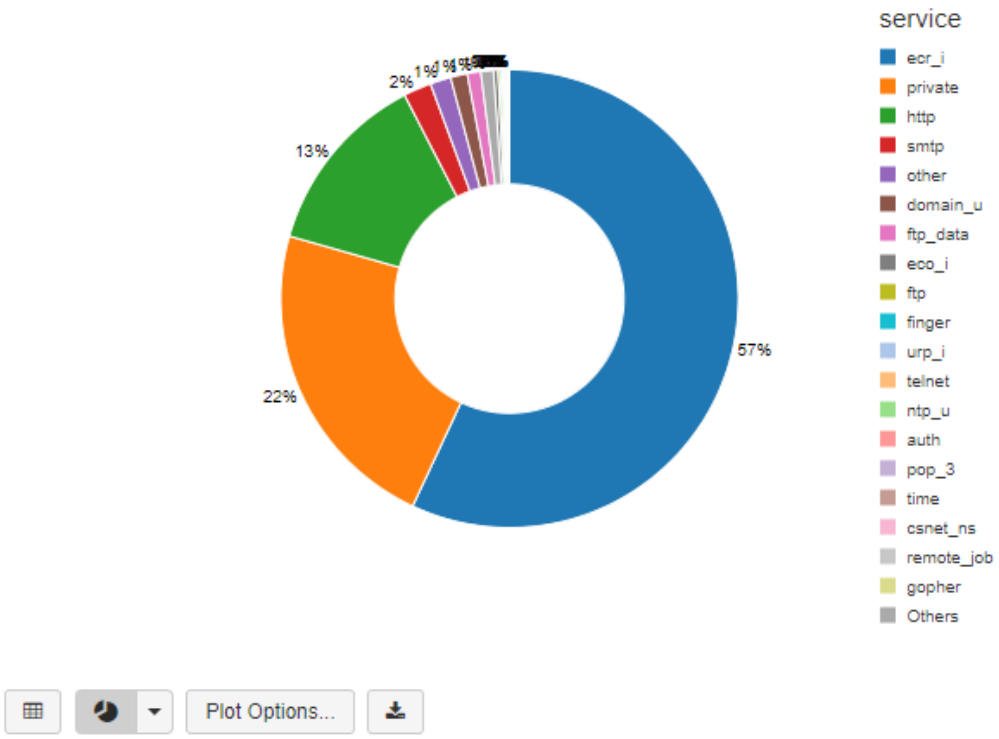
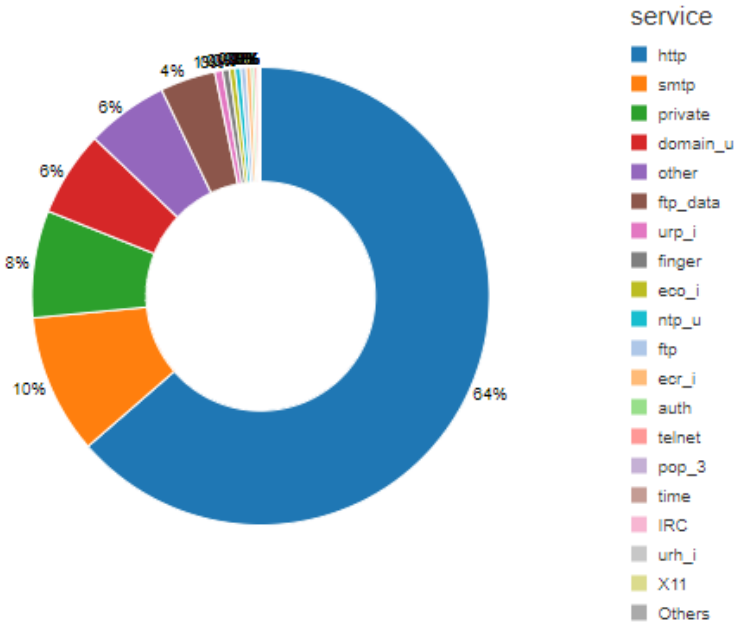


Figure below shows connections counts based on services and NORMAL connection counts based on services. High percentage of ecr_i connection but low percentage of ecr_i normal connection shows that ecr_i service come with lots of connections that are NOT normal. While comparison of the two figures show http service seems to come with lots of normal connections.





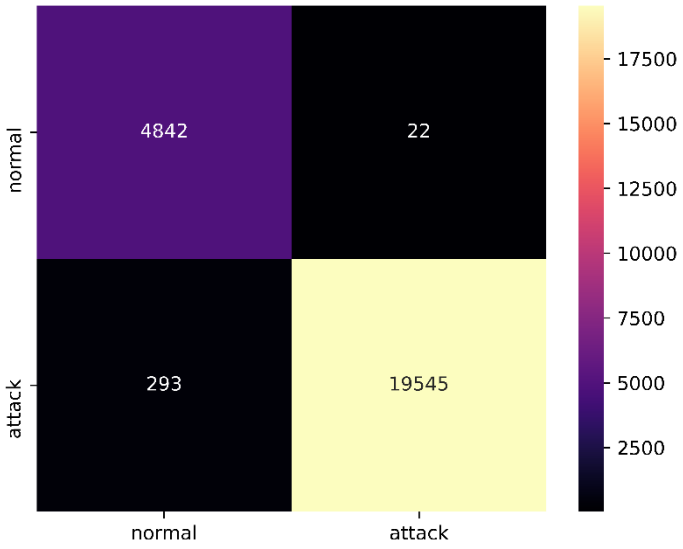
Question 8:

See python notebook file for code in databricks.

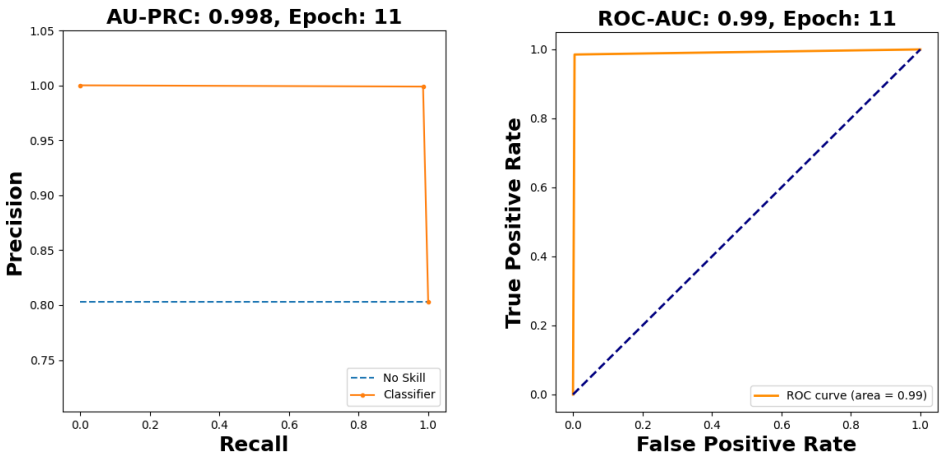
For better prediction, I have also written a 3-layer feedforward neural network to train on the dataset and shows a good predictive performance. The code is also attached in the CLF.py and output.txt shows all information about the hyperparameters and details of the neural network.

I choose SVM from MLlib because it is reported to be the best model in the paper. But due to databricks community version limitation, I cannot tune the model's hyperparameters. Therefore, I trained a 3-layer feedforward neural network here and results (after 10 epoches) are shown below.

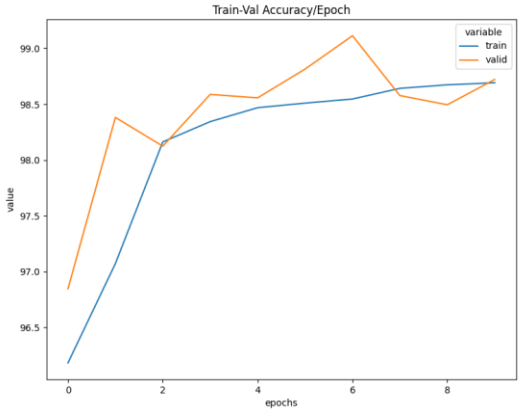
Here is the Confusion Matrix,

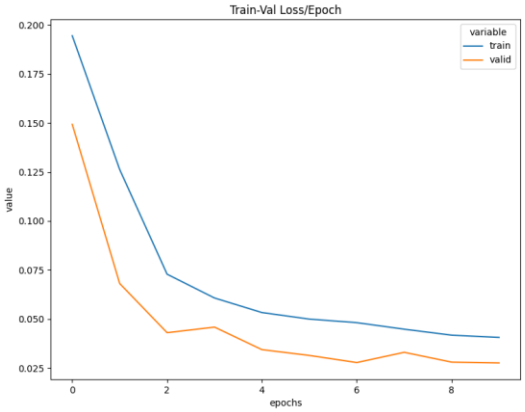


Here are the AUPRC and ROC-AUC plot



Here are the training details,





Epoch 001: | Train Loss: 0.19448 | Valid Loss: 0.14929 | Train Acc: 96.181| Valid Acc: 96.845
Epoch 002: | Train Loss: 0.12628 | Valid Loss: 0.06814 | Train Acc: 97.071| Valid Acc: 98.381
Epoch 003: | Train Loss: 0.07287 | Valid Loss: 0.04304 | Train Acc: 98.162| Valid Acc: 98.124
Epoch 004: | Train Loss: 0.06074 | Valid Loss: 0.04589 | Train Acc: 98.344| Valid Acc: 98.588
Epoch 005: | Train Loss: 0.05332 | Valid Loss: 0.03440 | Train Acc: 98.469| Valid Acc: 98.557
Epoch 006: | Train Loss: 0.04998 | Valid Loss: 0.03144 | Train Acc: 98.509| Valid Acc: 98.814
Epoch 007: | Train Loss: 0.04816 | Valid Loss: 0.02782 | Train Acc: 98.546| Valid Acc: 99.113
Epoch 008: | Train Loss: 0.04483 | Valid Loss: 0.03304 | Train Acc: 98.642| Valid Acc: 98.577
Epoch 009: | Train Loss: 0.04179 | Valid Loss: 0.02803 | Train Acc: 98.674| Valid Acc: 98.495
Epoch 010: | Train Loss: 0.04063 | Valid Loss: 0.02763 | Train Acc: 98.692| Valid Acc: 98.722

Here is the classification report, with precision, recall and f1-scores.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	4864
1	1.00	0.99	0.99	19838
accuracy			0.99	24702
macro avg	0.97	0.99	0.98	24702
weighted avg	0.99	0.99	0.99	24702

All metrics examined here (AUPRC, ROC-AUC, precision, recall, f1-score, etc.) shows very strong predictive ability of the model and no overfitting problems in the classification.

Part B – Question 1:

Statement 1 is yes.

According to the link provided in the Assignment file, https://azure.microsoft.com/en-us/overview/what-is-paas/, PaaS solution in Azure provides a framework that developers can build upon to develop or customize cloud-based applications. PaaS development tools in Azure can cut the time it takes to code new apps with pre-coded application components built into the platform. At the same time, PaaS solutions in Azure offers the capabilities of supporting the complete web application lifecycle: building, testing, deploying, managing, and updating within the same integrated environment. This allows continuously adding features to custom application.

Statement 2 is yes.

According to the Azure documentation, https://docs.microsoft.com/en-us/azure/azure-sql/azure-sql-iaas-vs-paas-what-is-overview, PaaS database offered by Azure allows built-in features and functionality that requires extensive configuration and therefore, features built-in high availability, intelligence, and management.

Question 2: D

A is wrong because dynamic schema is associated with non-relational/distributed database.

B is wrong because relational database stores data in Tabular Format with Predefined schema. It organizes data into one or more tables (or relations) of columns and rows, with a unique key identifying each row. NOT key/value pair.

C is wrong because images and videos are unstructured data, while relational database is used to store structured data (organised data format with a fixed schema).

D is correct answer because relational database is used to store structured data (organised data format with a fixed schema) where strong consistency guarantees are required.

Question 3: D

SaaS is product that is run and managed by the service provider (for example, google doc). The user will NOT need to worry about maintaining, provisioning, installing hardware or solution. Therefore, A, B, C is not correct. The only thing user will be responsible for is configuring the SaaS solution itself.

Question 4:

Statement 1: No

What a company could do is to combine a public cloud with an on-premises infrastructure in order to implement a hybrid cloud rather than migrating from a private cloud model.

Statement 2: Yes

Extending the capacity of internal network through using the public cloud is very common. When more capacity is needed, a company would only need to configure a cloud environment.

Statement 3: No

Anyone with an account in a public cloud model (Azure for example) can access to the cloud

resources.

Question 5:

a. A cloud service that remains available after a failure occurs: Fault Tolerance

b. A cloud service that can be recovered after a failure occurs: Disaster recovery

c. A cloud service that performs quickly when demand increases: Dynamic Scalability

d. A cloud service that can be accessed quickly from the internet: Low Latency

Explanations: cloud service w/ Disaster recovery for example, RDDs track the transformations used to build them (their lineage) to recompute lost data. Dynamic scalability is an architecture based on a system with predefined scaling that allows the dynamic allocation of resources, and therefore will performs quickly when demand increases. And Low Latency means quick internet access.