**Model**

**Step.1**

Remove Addresses -> "srcIP", "dstIP","srcMac", "dstMac"

**Step.2**

Remove Time -> "timeFirst", "timeLast"

As the files are collected at different days, the model will only use these variables to determine if they are included.

**Step.3**

Remove columns that are hard to sample ->"srcIPCC", "dstIPCC", "srcIPOrg", "dstIPOrg", "srcMac\_dstMac\_numP", "dstPortClass", "srcMacLbl\_dstMacLbl"

These columns have many unique values which only appear once or twice, making the sampling process to be hard.

If they are included, it will be common to find some layers in the testing set but not in the training test and vice versa.

**Step.4**

Remove all columns that only have NA value or only have 1 unique value

R will think these columns make no sense while training the model and reject them.

After these 4 steps, we got 80 columns left.

**Train**

70% of the data (220 flows) are used as training set, and 30%(529 flows) are used as the testing set.

For the training set, R will automatically use 70% of it for training and 30% of it for 10-fold cross validation.

**Tree -> only 3 columns are used in the tree**

tcpWS <= 0

| tcpAnomaly <= 4096: 1 (106.0)

| tcpAnomaly > 4096

| | numPktsSnt <= 4: 0 (5.0)

| | numPktsSnt > 4: 1 (4.0)

tcpWS > 0: 0 (105.0)