**ENGINEERED FEATURES – ANOMALY DETECTION**

1. **Connection-based features**: Rolling window for the 1000 previous packets when **SRCADDRESS** appears in the traffic.

* **How many times the SRCADDRESS has appeared in the last 1000 netflows**
  1. (size of the sub-dataframe)
  2. *(“CONN\_BASED\_SRCADDRESS\_OCCURENCES”)*
* For any of the flow records that SRCADDRESS has appeared within the last 1000 flow records , count the distinct destination ports
  1. *(“CONN\_BASED\_SRCADDRESS\_DISTINCT\_DSTPORTS”)*
* For any of the flow records that SRCADDRESS has appeared within the last 1000 flow records , count the distinct destination ip addresses
  1. For DSTADDRESS count the SRCADDRESS
  2. (“CONN\_BASED\_SRCADDRESS\_DISTINCT\_DSTADDRESS”)
* For any of the flow records that SRCADDRESS has appeared within the last 1000 flow records , count the distinct source ports
  1. *(“CONN\_BASED\_SRCADDRESS\_DISTINCT\_SRCPORTS”)*
* For any of the flow records that SRCADDRESS has appeared within the last 1000 flow records , average the packets
  1. *(“CONN\_BASED\_SRCADDRESS\_AVGPACKETIN”)*
* For any of the flow records that SRCADDRESS has appeared within the last 1000 flow records , average the bytes
  1. *(“CONN\_BASED\_SRCADDRESS\_AVGBYTEIN”)*
* How many times dstaddress and srcaddress appear in the last thousand flow records
  1. (“CONN\_BASED\_SRCADDRESS\_DSTADDRESS\_COUNT”)
* **Repeat Features for DSTADDRESS**

1. **Time-based features**: Rolling window for the 10 previous minutes when **SRCADDRESS** appears in the traffic.

* How many records exist in a given time frame?
  + Size of the dataframe within the 10 min frame
  + “*TIME\_BASED\_SRCADDRESS\_TOTAL\_OCCURENCES”*
* How many times has SRCADDRESS has appeared within the last 10 minutes
  + (size of the sub-dataframe)
  + “*TIME\_BASED\_SRCADDRESS\_OCCURENCES”*
* For any of the flow records that SRCADDRESS has appeared within the last 10 minutes , count the distinct destination ip addresses
  + For DSTADDRESS count the SRCADDRESS
  + (“*TIME\_BASED\_SRCADDRESS\_DISTINCT\_DSTADDRESS”)*
* For any of the flow records that SRCADDRESS has appeared within the last 10 minutes, count the distinct destination ports
  + *(“TIME\_BASED\_SRCADDRESS\_DISTINCT\_DSTPORTS”)*
* For any of the flow records that SRCADDRESS has appeared within the last 10 minutes, count the distinct source ports
  + “*TIME\_BASED\_SRCADDRESS\_DISTINCT\_SRCPORTS”*
* For any of the flow records that SRCADDRESS has appeared within the last 10 minutes, get the average packets
  + (“TIME\_BASED\_SRCADDRESS\_AVGPACKETIN”)
* For any of the flow records that SRCADDRESS has appeared within the last 10 minutes, get the average bytes
  + (“TIME\_BASED\_SRCADDRESS\_AVGPACKETIN”)
* How many times dstaddress and srcaddress appear in the last 10 minutes
  + (“TIME\_BASED\_SRCADDRESS\_DSTADDRESS\_COUNT”)
* **Repeat Features for DSTADDRESS**

**Each feature will result into a new column added.**