CHEATING DATASET DESIGN

Camila Barbagallo, Ryan Daher,
Paula Garcia and Rocio Gonzalez

01 PSI and Chi-Square

02 Data Collection

03 Data Transformation

04 Metrics in the Dataset

05 The Dataset

06 Metric Interpretation

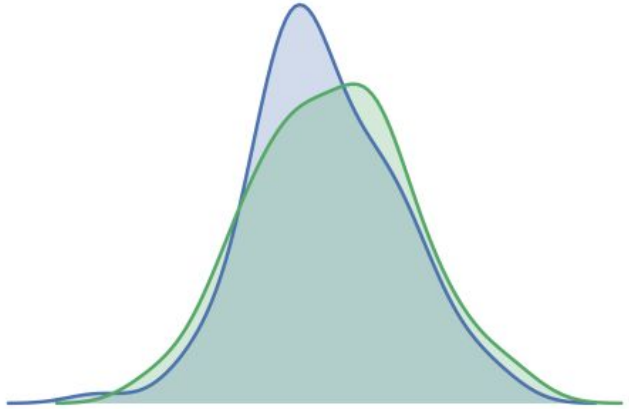
01
PSI & Chi-Square

PSI: Population Stability Index

How different two sample distributions are.

PSI = (%Actual - %Expected) x ln (%Actual / %Expected)

- Can be applied through binning numerics into categories.



Used to monitor population change and for diagnosing possible problems in model performance.

➔ PSI <= 0.1 indicates little change [no action required]

➔ 0.1 - 0.25 is little change but too small to determine [still no action required]

➔ PSI > 0.25 is a significant shift [action required - merits further investigation]

Chi-Square

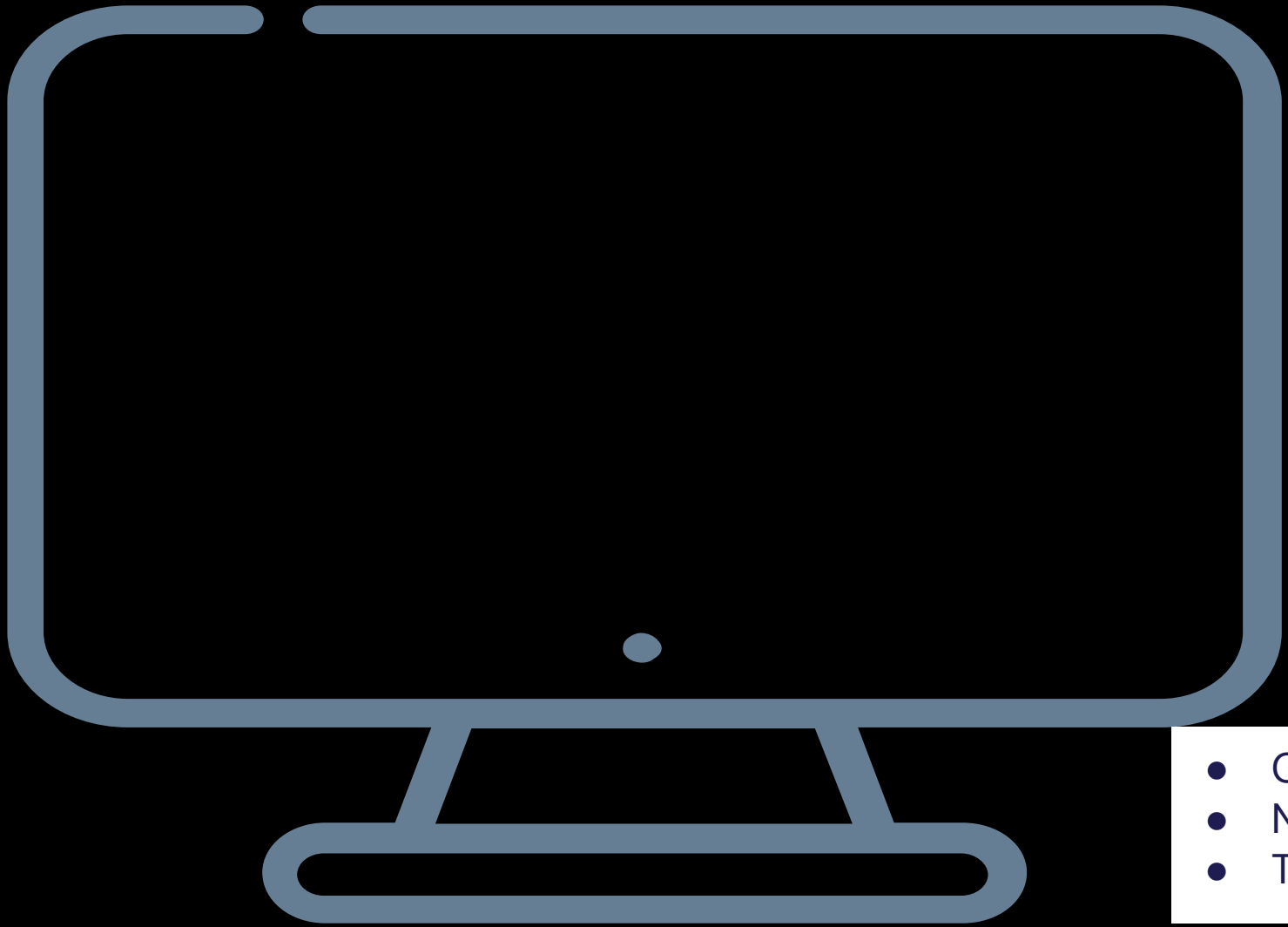
Non-parametric statistical test that describes the magnitude of discrepancy between the observed data and the data expected to be obtained with a specific hypothesis.

$$\chi^2 = (\text{Observed}_i - \text{Expected}_i)^2 / \text{Expected}_i$$

The computed value of χ^2 is compared with the table value of χ^2 for a given degree of freedom and at a given significance level.

- ➔ χ^2 > the table value, it is a significant shift [action required - merits further investigation]
- ➔ χ^2 < the table value, it indicates little change [no action required]

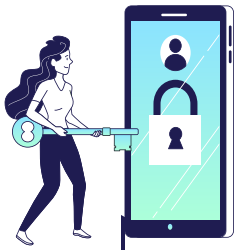
02
Data Collection



- Gps
- Netstat
- Tasklist

Our Sample

73%
Cheating



2 Computers

49
Observations

Netstat
Tasklist

Raw Data

Netstat

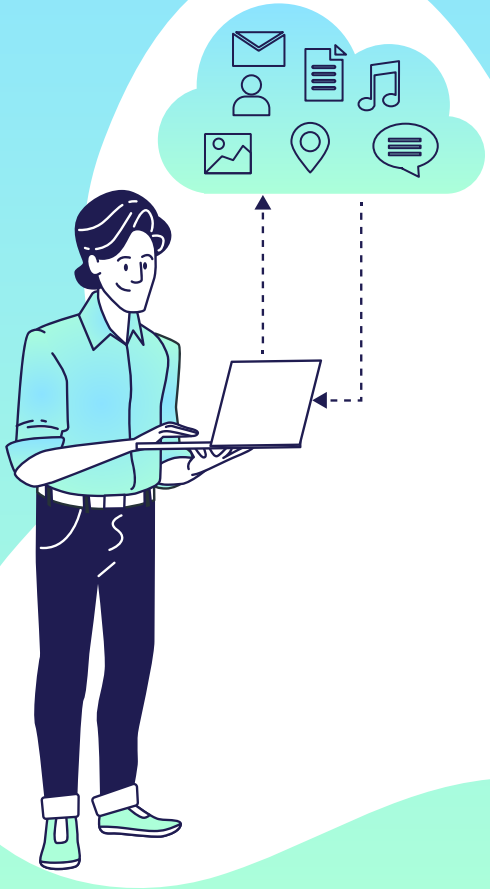
Proto	Dirección local	Dirección remota	Estado	PID
TCP	0.0.0.0:135	0.0.0.0:0	LISTENING	836
TCP	0.0.0.0:445	0.0.0.0:0	LISTENING	4
TCP	0.0.0.0:5040	0.0.0.0:0	LISTENING	6188
TCP	0.0.0.0:49664	0.0.0.0:0	LISTENING	816
TCP	0.0.0.0:49665	0.0.0.0:0	LISTENING	736
TCP	0.0.0.0:49666	0.0.0.0:0	LISTENING	1580
TCP	0.0.0.0:49667	0.0.0.0:0	LISTENING	2160
TCP	0.0.0.0:49668	0.0.0.0:0	LISTENING	3564
TCP	0.0.0.0:49669	0.0.0.0:0	LISTENING	804
TCP	0.0.0.0:54950	0.0.0.0:0	LISTENING	4732
TCP	127.0.0.1:5939	0.0.0.0:0	LISTENING	4652
TCP	192.168.1.41:139	0.0.0.0:0	LISTENING	4

Tasklist

Nombre de imagen	PID	Nombre de sesión	Núm. de ses	Uso de memor
System Idle Process	0	Services	0	8 KB
System	4	Services	0	1.384 KB
Registry	96	Services	0	70.020 KB
smss.exe	416	Services	0	1.112 KB
csrss.exe	648	Services	0	3.532 KB
wininit.exe	736	Services	0	4.040 KB
services.exe	804	Services	0	9.208 KB
lsass.exe	816	Services	0	17.364 KB
fontdrvhost.exe	1012	Services	0	1.688 KB
svchost.exe	1020	Services	0	2.080 KB
svchost.exe	324	Services	0	35.420 KB
svchost.exe	836	Services	0	17.388 KB
svchost.exe	1064	Services	0	7.496 KB
svchost.exe	1192	Services	0	6.236 KB
svchost.exe	1268	Services	0	8.432 KB

03

Data Transformation



Raw Data: Tasklist

```
Nombre de imagen              PID Nombre de sesión Núm. de ses Uso de memor
========================= ======== ================ ===========
System Idle Process             0 Services                   0         8 KB
System                          4 Services                   0     1.384 KB
Registry                       96 Services                   0    70.020 KB
smss.exe                      416 Services                   0     1.112 KB
csrss.exe                     648 Services                   0     3.532 KB
wininit.exe                   736 Services                   0     4.040 KB
services.exe                  804 Services                   0     9.208 KB
lsass.exe                     816 Services                   0    17.364 KB
fontdrvhost.exe              1012 Services                   0     1.688 KB
svchost.exe                  1020 Services                   0     2.080 KB
svchost.exe                   324 Services                   0    35.420 KB
svchost.exe                   836 Services                   0    17.388 KB
svchost.exe                  1064 Services                   0     7.496 KB
svchost.exe                  1192 Services                   0     6.236 KB
svchost.exe                  1268 Services                   0     8.432 KB
```

Raw Data: Tasklist

```
Nombre de imagen              PID Nombre de sesión Núm. de ses Uso de memor
========================= ======== ================ ===========
System Idle Process             0 Services                   0         8 KB
System                          4 Services                   0     1.384 KB
Registry                       96 Services                   0    70.020 KB
smss.exe                      416 Services                   0     1.112 KB
csrss.exe                     648 Services                   0     3.532 KB
wininit.exe                   736 Services                   0     4.040 KB
services.exe                  804 Services                   0     9.208 KB
lsass.exe                     816 Services                   0    17.364 KB
fontdrvhost.exe              1012 Services                   0     1.688 KB
svchost.exe                  1020 Services                   0     2.080 KB
svchost.exe                   324 Services                   0    35.420 KB
svchost.exe                   836 Services                   0    17.388 KB
svchost.exe                  1064 Services                   0     7.496 KB
svchost.exe                  1192 Services                   0     6.236 KB
svchost.exe                  1268 Services                   0     8.432 KB
```

Raw Data: Tasklist

```
Nombre de imagen              PID Nombre de sesión Núm. de ses Uso de memor
========================= ======== ================ ===========
System Idle Process             0 Services                   0         8 KB
System                          4 Services                   0     1.384 KB
Registry                       96 Services                   0    70.020 KB
smss.exe                      416 Services                   0     1.112 KB
csrss.exe                     648 Services                   0     3.532 KB
wininit.exe                   736 Services                   0     4.040 KB
services.exe                  804 Services                   0     9.208 KB
lsass.exe                     816 Services                   0    17.364 KB
fontdrvhost.exe              1012 Services                   0     1.688 KB
svchost.exe                  1020 Services                   0     2.080 KB
svchost.exe                   324 Services                   0    35.420 KB
svchost.exe                   836 Services                   0    17.388 KB
svchost.exe                  1064 Services                   0     7.496 KB
svchost.exe                  1192 Services                   0     6.236 KB
svchost.exe                  1268 Services                   0     8.432 KB
```

CSV: Tasklist

```
def tasklist_csv(path):
    csvconv = pd.read_fwf(path,encoding='UTF-16')
    csvconv = csvconv.drop([0,1],axis=0)
    csvconv.columns = ['Image Name','PID','Session Name','Session Number','Memory Usage']
    return(csvconv)
```

Raw Data: Netstat

Proto	Dirección local	Dirección remota	Estado	PID
TCP	0.0.0.0:135	0.0.0.0:0	LISTENING	836
TCP	0.0.0.0:445	0.0.0.0:0	LISTENING	4
TCP	0.0.0.0:5040	0.0.0.0:0	LISTENING	6188
TCP	0.0.0.0:49664	0.0.0.0:0	LISTENING	816
TCP	0.0.0.0:49665	0.0.0.0:0	LISTENING	736
TCP	0.0.0.0:49666	0.0.0.0:0	LISTENING	1580
TCP	0.0.0.0:49667	0.0.0.0:0	LISTENING	2160
TCP	0.0.0.0:49668	0.0.0.0:0	LISTENING	3564
TCP	0.0.0.0:49669	0.0.0.0:0	LISTENING	804
TCP	0.0.0.0:54950	0.0.0.0:0	LISTENING	4732
TCP	127.0.0.1:5939	0.0.0.0:0	LISTENING	4652
TCP	192.168.1.41:139	0.0.0.0:0	LISTENING	4

CSV: Netstat

```
def netstat_csv(path):
    netstatcsv =pd.read_csv(path,header=None,skiprows=4,encoding='UTF-16',delim_whitespace=True)
    netstatcsv[4][netstatcsv[0]=='UDP'] = netstatcsv[3]
    netstatcsv[3][netstatcsv[0]=='UDP'] = 'NA'
    netstatcsv.columns = ['Protocol','Local Direction','Remote Direction','State','PID']
    netstatcsv = netstatcsv.drop(['State','Protocol'],axis=1)
    return(netstatcsv)
```

CSVs

Tasklist

	Image Name	PID	Session Name	Session Number	Memory Usage
2	System Idle Process	0	Services	0	8 KB
3	System	4	Services	0	1.536 KB
4	Registry	120	Services	0	54.664 KB
5	smss.exe	468	Services	0	1.124 KB
6	csrss.exe	712	Services	0	5.104 KB
...	...	...	...	...	...
227	SkypeApp.exe	10228	Console	13	214.340 KB
228	RuntimeBroker.exe	9656	Console	13	21.412 KB
229	Microsoft.Notes.exe	15000	Console	13	110.988 KB
230	RuntimeBroker.exe	18976	Console	13	20.804 KB
231	tasklist.exe	8632	Console	13	8.624 KB

Netstat

	Local Direction	Remote Direction	PID
0	0.0.0.0:135	0.0.0.0:0	1144
1	0.0.0.0:445	0.0.0.0:0	4
2	0.0.0.0:3306	0.0.0.0:0	6020
3	0.0.0.0:5040	0.0.0.0:0	8056
5	0.0.0.0:7680	0.0.0.0:0	12444
6	0.0.0.0:8733	0.0.0.0:0	13104
8	0.0.0.0:49664	0.0.0.0:0	948
9	0.0.0.0:49665	0.0.0.0:0	836
10	0.0.0.0:49666	0.0.0.0:0	1800

Preprocessing

```
def standardizer(nocheat_netIN,new_netIN,nocheat_taskIN,new_taskIN):
    nocheat_netdf = pd.DataFrame(nocheat_netIN['Remote Direction'].value_counts())
    new_netdf = pd.DataFrame(new_netIN['Remote Direction'].value_counts())
    nocheat_imagedf = pd.DataFrame(nocheat_taskIN['Image Name'].value_counts())
    new_imagedf = pd.DataFrame(new_taskIN['Image Name'].value_counts())
    nocheat_memdf = pd.DataFrame(nocheat_taskIN['Memory Usage'].value_counts())
    new_memdf = pd.DataFrame(new_taskIN['Memory Usage'].value_counts())
    nocheatlist = [nocheat_netdf,nocheat_imagedf,nocheat_memdf]
    newlist = [new_netdf,new_imagedf,new_memdf]
```

Value Counts

```
*:*                    14
0.0.0.0:0              12
40.67.254.36:443        2
52.114.74.88:443        1
2.17.133.7:443          1
127.0.0.1:25340         1
204.79.197.200:443      1
52.114.75.19:443        1
172.217.19.138:443      1
52.157.234.37:443       1
Name: Remote Direction, dtype: int64
0.0.0.0:0              14
*:*                    13
52.114.128.73:443       3
93.184.220.29:80        1
40.67.254.36:443        1
[::]:0                  1
173.194.76.188:443      1
95.100.101.33:80        1
Name: Remote Direction, dtype: int64
```

Preprocessing

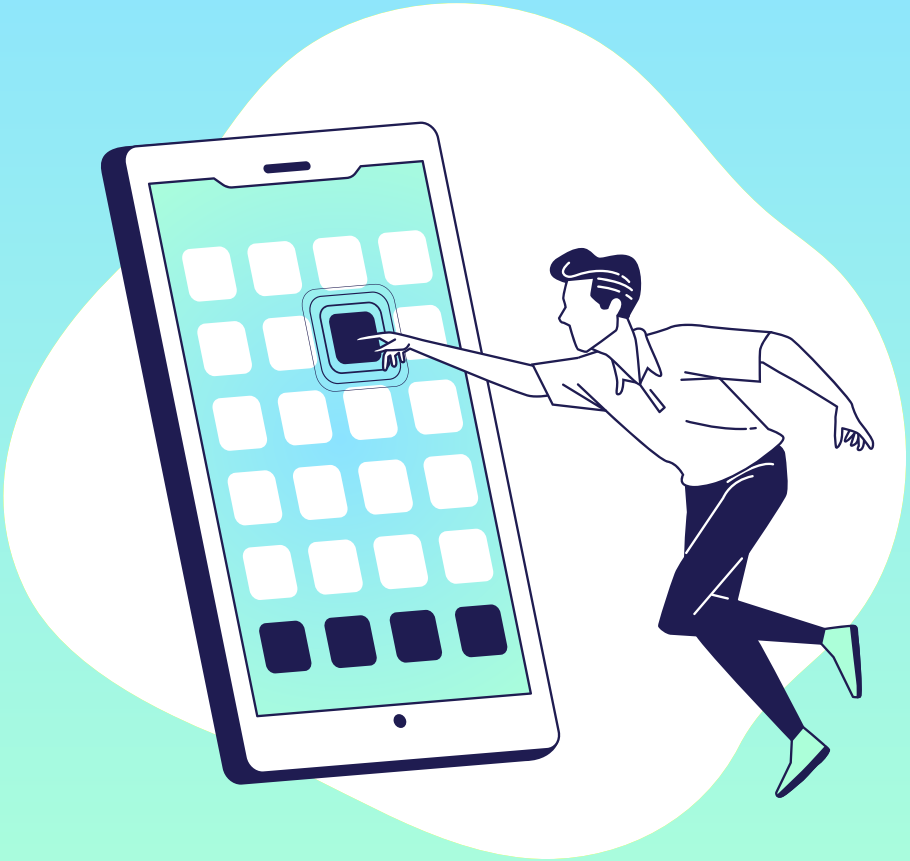
```
for i in range(0,len(nocheatlist)):
    nocheataddon = newlist[i].index.difference(nocheatlist[i].index)
    newaddon = nocheatlist[i].index.difference(newlist[i].index)
    for j in newaddon:
      newlist[i].loc[j] = [0]
    for j in nocheataddon:
      nocheatlist[i].loc[j] = [0]
    newlist[i] = newlist[i].reindex(nocheatlist[i].index)


 return [nocheatlist,newlist]
```

Same number of rows in no-cheating and cheating

Differencing

	Remote Direction
0.0.0.0:0	15
:	14
52.114.128.73:443	0
93.184.220.29:80	0
173.194.76.188:443	1
[::]:0	1
40.67.254.36:443	1
95.100.101.33:80	0
185.199.111.154:443	1
40.101.92.178:443	1
40.90.137.120:443	1
52.114.133.61:443	1
92.123.129.198:443	1

	Remote Direction
0.0.0.0:0	14
:	13
52.114.128.73:443	3
93.184.220.29:80	1
173.194.76.188:443	1
[::]:0	1
40.67.254.36:443	1
95.100.101.33:80	1
185.199.111.154:443	0
40.101.92.178:443	0
40.90.137.120:443	0
52.114.133.61:443	0
92.123.129.198:443	0



04
Metrics in the Dataset

Pre-Processing

calc_df(nocheat_netIN_1,nocheat_netIN_2,new_netIN,nocheat_taskIN_1,nocheat_taskIN_2, new_taskIN,appendata,TARGET):

tasklist_csv(path)

netstat_csv(path)

standardizer(nocheat_netIN,new_netIN,nocheat_taskIN,new_taskIN)

PSI

```
def sub_psi(og_perc, new_perc):
    if new_perc == 0:
      new_perc = 0.0001
    if og_perc == 0:
      og_perc = 0.0001
    subpsi = (og_perc - new_perc) * np.log(og_perc / new_perc)
    return (subpsi)
```

CHI

```
def sub_chi(og_perc, new_perc):
    if new_perc == 0:
      new_perc = 0.0001
    if og_perc == 0:
      og_perc = 0.0001
    subchi = ((og_perc - new_perc)**2)/og_perc
    return (subchi)
```

PSI & Chi Computations

```
psi_value = 0
    for i in range(0, len(og_perc)):
      psi_value += sub_psi(og_perc[i], new_perc[i])
    psilist.append(psi_value)


chi_value = 0
    for i in range(0, len(og_perc)):
      chi_value += sub_chi(og_perc[i], new_perc[i])
    chilist.append(chi_value)
```

Creating the Dataset

```
finallist =
calc_chi_psi(nocheat_netIN_1,nocheat_netIN_2,new_netIN,nocheat_taskIN_1,nocheat_taskIN_2,new_taskI
N) + [TARGET]


appendata = appendata.append(pd.Series(finallist,index=appendata.columns),ignore_index=True)


return(appendata)
```

05

The Dataset

Our Dataset

Remote Direction_PSI_1	Image Name_PSI_1	Memory Usage_PSI_1	Remote Direction_PSI_2	Image Name_PSI_2	Memory Usage_PSI_2
2.255758	0.165668	7.089758	4.371513	2.306637	7.277827
1.962120	0.120315	7.053344	4.174531	2.328434	7.531000
2.611957	0.125612	7.196870	4.251042	2.339715	7.543256
2.770529	0.139884	7.017070	4.235008	2.349817	7.450522
1.985350	0.090635	7.029805	4.431992	2.421653	7.246365

Remote Direction_CHI_1	Image Name_CHI_1	Memory Usage_CHI_1	Remote Direction_CHI_2	Image Name_CHI_2	Memory Usage_CHI_2	TARGET
36.337110	2.861258	41.390571	42.268964	59.507115	41.570181	1.0
30.664180	1.466135	40.803390	38.184019	61.584761	42.471574	1.0
54.778019	1.449352	42.669335	46.659293	61.094803	43.582866	1.0
54.466522	1.487045	42.131149	46.593919	62.124444	43.625442	1.0
31.848733	1.098694	41.195872	37.671781	80.526935	41.562405	1.0

06
Metric Interpretation

PSI vs. Chi-Square

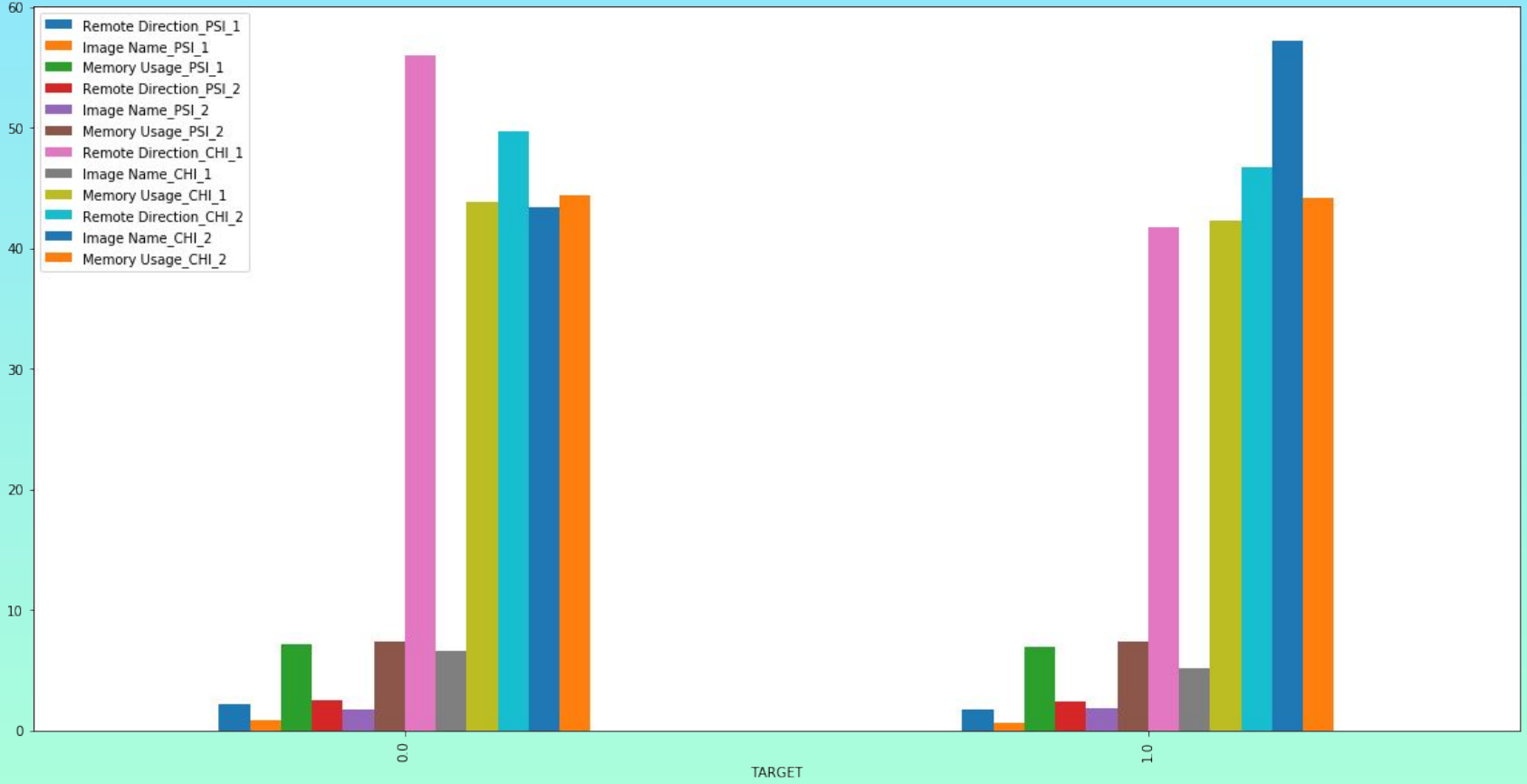
PSI

- <0.1 no change-> most probably no cheating
- 0.1-0.2 -> slight change -> maybe cheating
- ≥0.2 -> significant change -> most probably cheating

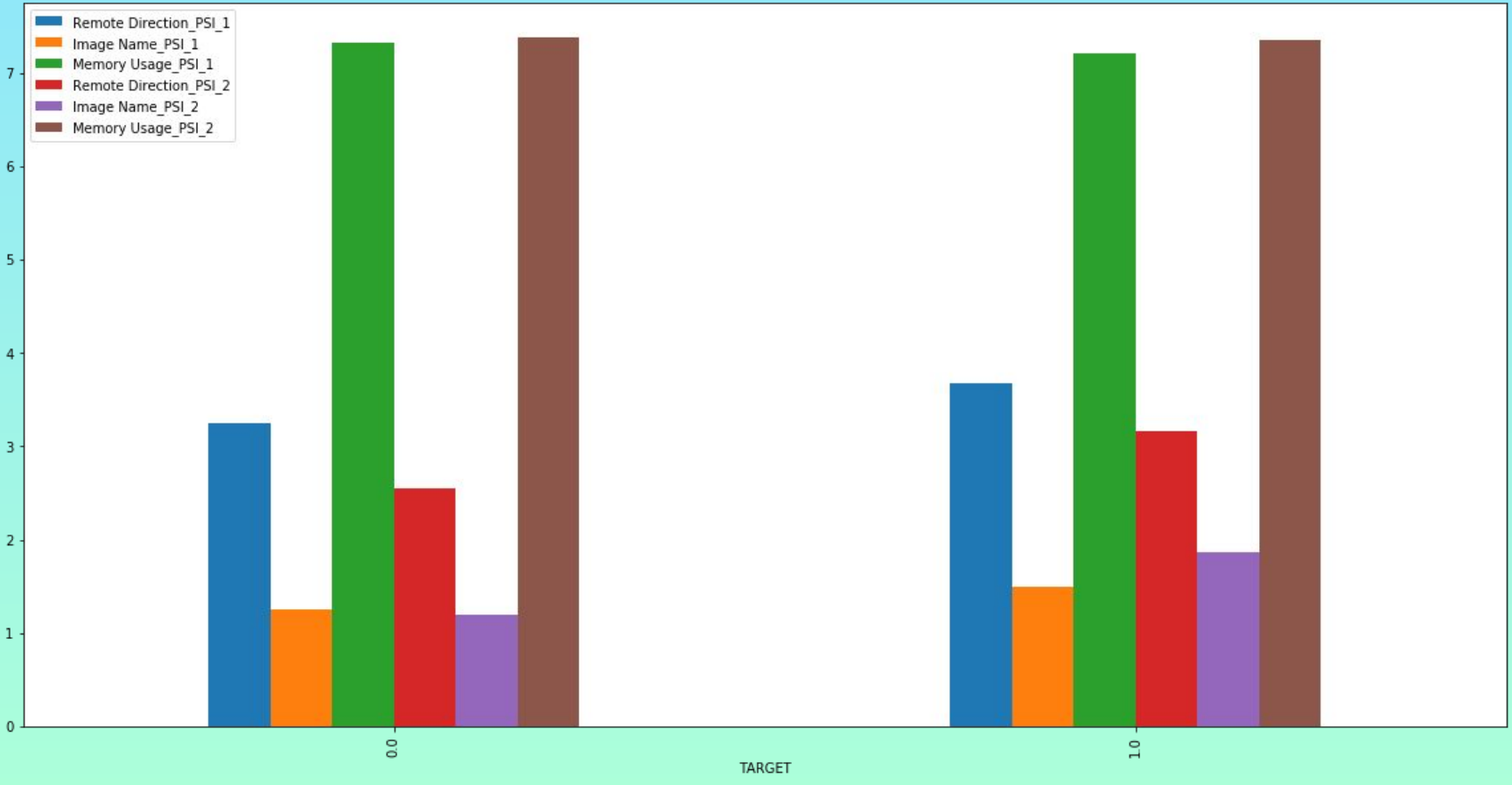
Chi-Square

- H0: expected fits observed -> no cheating
- Ha: expected doesn't fit observe -> cheating
- Calculate Critical Chi-square
  - Degrees of freedom
  - α

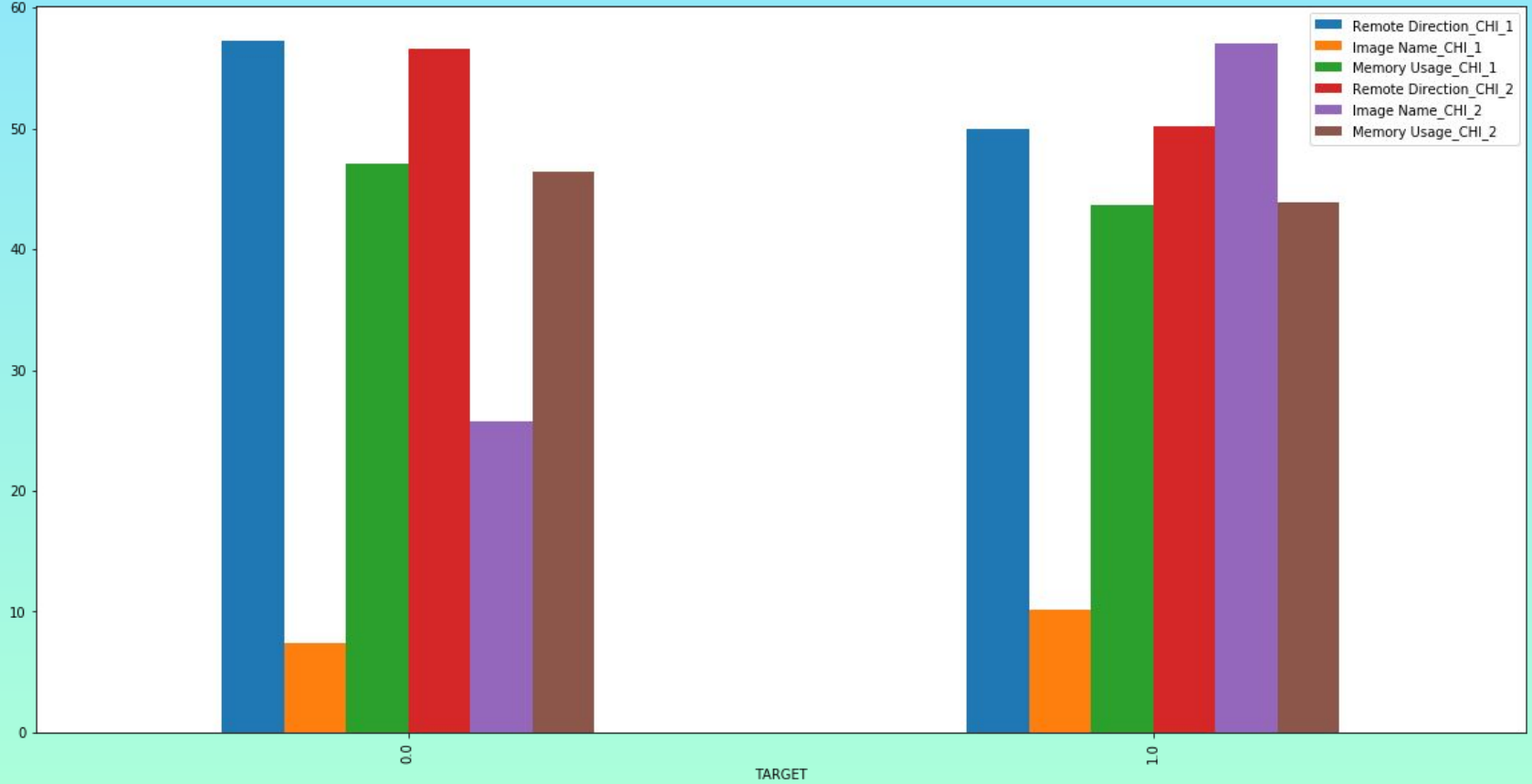
Full Comparison



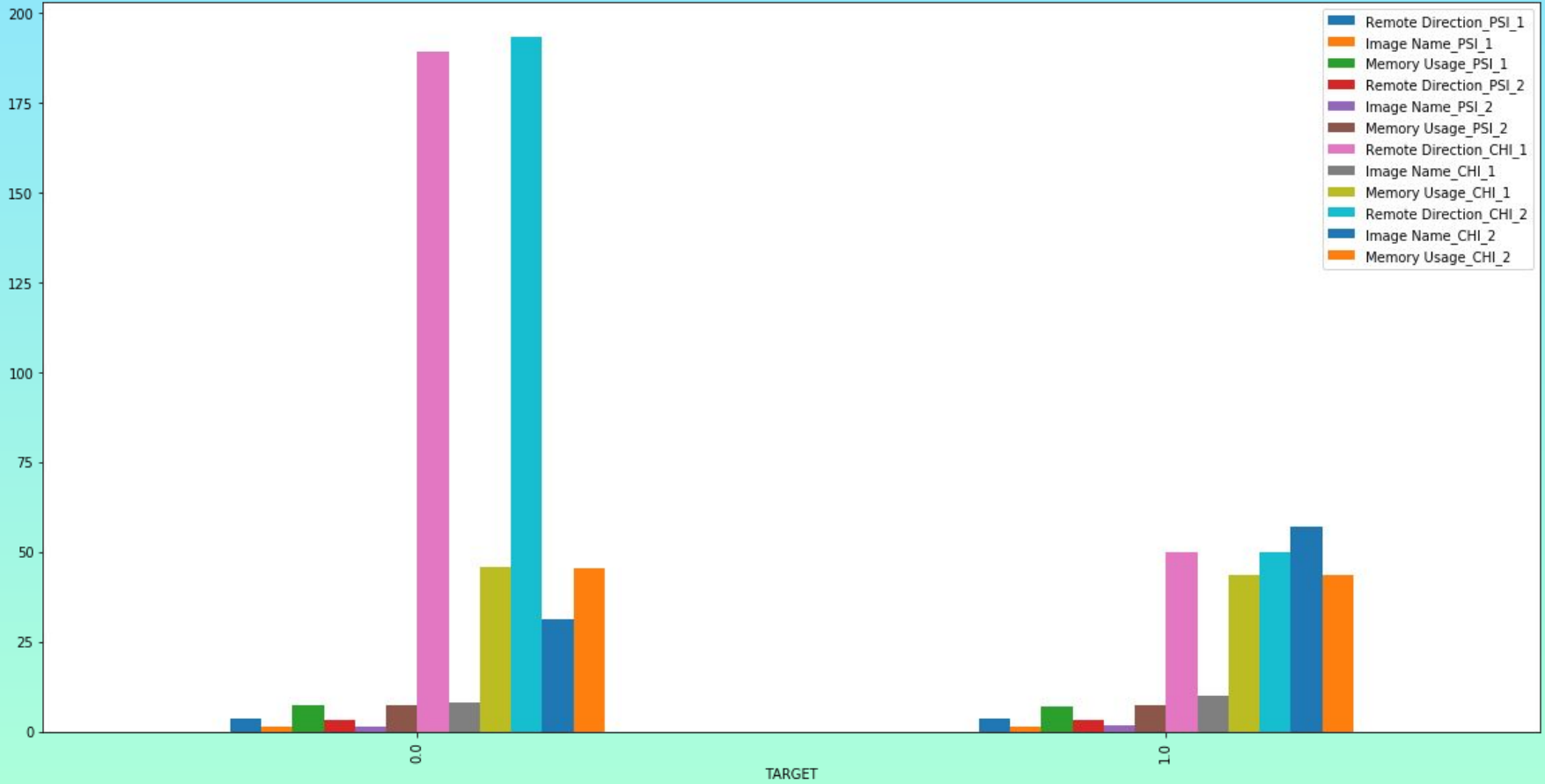
PSI


Remote Direction_PSI_1
Image Name_PSI_1
Memory Usage_PSI_1
Remote Direction_PSI_2
Image Name_PSI_2
Memory Usage_PSI_2
0
1
2
3
4
5
6
7
0.0
1.0
TARGET

Chi


Remote Direction_CHI_1
Image Name_CHI_1
Memory Usage_CHI_1
Remote Direction_CHI_2
Image Name_CHI_2
Memory Usage_CHI_2
60
50
40
30
20
10
0
0.0
1.0
TARGET

No-Cheat Disparities



THANKS!

Do you have any questions?