In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
from sklearn.linear_model import LogisticRegression
```

In [3]:
```
df=pd.read_csv("C3_bot_detection_data.csv")
df
```

Out[3]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinsto
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sandersto
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfu
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezber
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachovill
...	...	...	...	...	...	...	...	...	.
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lak Kimberlyburg
49996	739297	jessicamunoz	Provide whole maybe agree church	18	5	9900	False	1	Greenbur

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
			respond most ...						
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfo
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephensid
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakber

50000 rows × 11 columns

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   User ID         50000 non-null  int64 
 1   Username        50000 non-null  object
 2   Tweet           50000 non-null  object
 3   Retweet Count   50000 non-null  int64 
 4   Mention Count   50000 non-null  int64 
 5   Follower Count  50000 non-null  int64 
 6   Verified        50000 non-null  bool  
 7   Bot Label       50000 non-null  int64 
 8   Location        50000 non-null  object
 9   Created At      50000 non-null  object
 10  Hashtags        41659 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.9+ MB
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count',
       'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At',
       'Hashtags'],
      dtype='object')
```

In [6]:
```
feature_matrix=df[['User ID',  'Retweet Count', 'Mention Count',
       'Follower Count','Bot Label']]
target_vector=df[ 'Verified']
```

In [7]:
```
feature_matrix.shape
```

Out[7]: (50000, 5)

In [8]:
```
target_vector.shape
```

Out[8]: (50000,)

In [9]:
```
from sklearn.preprocessing import StandardScaler
```

In [10]:
```
fs=StandardScaler().fit_transform(feature_matrix)
```

In [11]:
```
logr=LogisticRegression()
logr.fit(fs,target_vector)
```

Out[11]: LogisticRegression()

In [12]:
```
observation=[[1,2,3,4,5]]
```

In [13]:
```
prediction=logr.predict(observation)
print(prediction)
```

```
[ True]
```

In [14]:
```
logr.classes_
```

Out[14]: array([False, True])

In [15]:
```
logr.predict_proba(observation)[0][0]
```

Out[15]: 0.4875957520146553

In [16]:
```
logr.predict_proba(observation)
```

Out[16]: array([[0.48759575, 0.51240425]])