Ethereum Fraud Detection

Data Mining

Group 48

נדב ישר

דניאל רביב

הקדמה

בשנים האחרונות שוק המטבעות הקריפטוגרפיים עלה בצורה דרסטית, מה שגרם לכל אדם להכיר את המושג קריפטו, לחשוב על הקעות פוטנציאליות במטבעות הקיימים, ובעיקר לשמוע על קשר כזה או אחר לפעילות עבריינית.

מטרת הפרויקט

בפרויקט שלנו נבצע קלסיפיקציה (סיווג) עבור טרנזקציות של מטבע איתריום שהתבצעו, ונחליט האם לסווג טרנזקציה כהונאה (1) או כטרנזקציה לגיטימית (0)

FLAG = Our Label

Some of the features

FLAG	Avg min between sent tnx	Avg min between received tnx	Time Diff between first and last (Mins)	Sent tnx	Received Tnx	Number of Created Contracts	Unique Received From Addresses	Unique Sent To Addresses	min value received	max value received	avg val received	min val sent	max val sent	avg val sent
0	844.26	1093.71	704785.63	721	89	0	40	118	0.000000	45.806785	6.589513	0.00	31.220000	1.200681
0	12709.07	2958.44	1218216.73	94	8	0	5	14	0.000000	2.613269	0.385685	0.00	1.800000	0.032844
0	246194.54	2434.02	516729.30	2	10	0	10	2	0.113119	1.165453	0.358906	0.05	3.538616	1.794308
0	10219.60	15785.09	397555.90	25	9	0	7	13	0.000000	500.000000	99.488840	0.00	450.000000	70.001834
0	36.61	10707.77	382472.42	4598	20	1	7	19	0.000000	12.802411	2.671095	0.00	9.000000	0.022688

היכרות עם בסיס הנתונים

Features, and features, and more features.....

- Index: the index number of a row
- Address: the address of the ethereum account
- FLAG: whether the transaction is fraud or not
- Avg min between sent tnx: Average time between sent transactions for account in minutes
- Avg *min* between *received* tnx: Average time between received transactions for account in minutes
- Time *Diff* between *first* and_last(Mins): Time difference between the first and last transaction
- Sent_tnx: Total number of sent normal transactions
- Received_tnx: Total number of received normal transactions
- Number *of* Created_Contracts: Total Number of created contract transactions
- Unique *Received* From_Addresses: Total Unique addresses from which account received transactions
- Unique *Sent* To_Addresses20: Total Unique addresses from which account sent transactions
- Min *Value* Received: Minimum value in Ether ever received
- Max *Value* Received: Maximum value in Ether ever received
- Avg *Value* Received5Average value in Ether ever received
- Min *Val* Sent: Minimum value of Ether ever sent
- Max *Val* Sent: Maximum value of Ether ever sent
- Avg *Val* Sent: Average value of Ether ever sent
- Min *Value* Sent *To* Contract: Minimum value of Ether sent to a contract
- Max *Value* Sent *To* Contract: Maximum value of Ether sent to a contract
- Avg *Value* Sent *To* Contract: Average value of Ether sent to contracts
- Total *Transactions (Including* Tnx *to* Create_Contract): Total number of transactions
- Total *Ether* Sent:Total Ether sent for account address
- Total *Ether* Received: Total Ether received for account address
- Total *Ether* Sent_Contracts: Total Ether sent to Contract addresses
- Total *Ether* Balance: Total Ether Balance following enacted transactions

- Total *ERC20* Tnxs: Total number of ERC20 token transfer transactions
- ERC20 *Total* Ether_Received: Total ERC20 token received transactions in Ether
- ERC20 *Total* Ether_Sent: Total ERC20token sent transactions in Ether
- ERC20 *Total* Ether *Sent* Contract: Total ERC20 token transfer to other contracts in Ether
- ERC20 *Uniq* Sent_Addr: Number of ERC20 token transactions sent to Unique account addresses
- ERC20 *Uniq* Rec_Addr: Number of ERC20 token transactions received from Unique addresses
- ERC20 *Uniq* Rec *Contract* Addr: Number of ERC20token transactions received from Unique contract addresses
- ERC20 *Avg* Time *Between* Sent_Tnx: Average time between ERC20 token sent transactions in minutes
- ERC20 *Avg* Time *Between* Rec_Tnx: Average time between ERC20 token received transactions in minutes
- ERC20 *Avg* Time *Between* Contract_Tnx: Average time ERC20 token between sent token transactions
- ERC20 *Min* Val_Rec: Minimum value in Ether received from ERC20 token transactions for account
- ERC20 *Max* Val_Rec: Maximum value in Ether received from ERC20 token transactions for account
- ERC20 *Avg* Val_Rec: Average value in Ether received from ERC20 token transactions for account
- ERC20 *Min* Val_Sent: Minimum value in Ether sent from ERC20 token transactions for account
- ERC20 *Max* Val_Sent: Maximum value in Ether sent from ERC20 token transactions for account
- ERC20 *Avg* Val_Sent: Average value in Ether sent from ERC20 token transactions for account
- ERC20 *Uniq* Sent *Token* Name: Number of Unique ERC20 tokens transferred
- ERC20 *Uniq* Rec *Token* Name: Number of Unique ERC20 tokens received
- ERC20 *Most* Sent *Token* Type: Most sent token for account via ERC20 transaction
- ERC20 *Most* Rec *Token* Type: Most received token for account via ERC20 transactions

ניתוח ראשוני

ערכים עבור כל עמודה
(מה זה סטיית תקן 0?)

מציאת ערכים חסרים

```
#Important information about each feature
df.describe()
```

	FLAG	Avg min between sent tnx	Avg min between received tnx	Time Diff between first and last (Mins)	Sent tnx	Received Tnx	Number of Created Contracts	min value sent to contract	max val sent to contract	avg value sent to contract	max va recei
count	9841.000000	9841.000000	9841.000000	9.841000e+03	9841.000000	9841.000000	9841.000000	9841.000000	9841.000000	9841.000000	9841.000
mean	0.221421	5086.878721	8004.851184	2.183333e+05	115.931714	163.700945	3.729702	0.000003	0.000008	0.000005	523.152
std	0.415224	21486.549974	23081.714801	3.229379e+05	757.226361	940.836550	141.445583	0.000225	0.000516	0.000323	13008.821
min	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	0.000000	0.000000	0.000000	3.169300e+02	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000
50%	0.000000	17.340000	509.770000	4.663703e+04	3.000000	4.000000	0.000000	0.000000	0.000000	0.000000	6.000
75%	0.000000	565.470000	5480.390000	3.040710e+05	11.000000	27.000000	0.000000	0.000000	0.000000	0.000000	67.067
max	1.000000	430287.670000	482175.490000	1.954861e+06	10000.000000	10000.000000	9995.000000	0.020000	0.046029	0.023014	800000.000

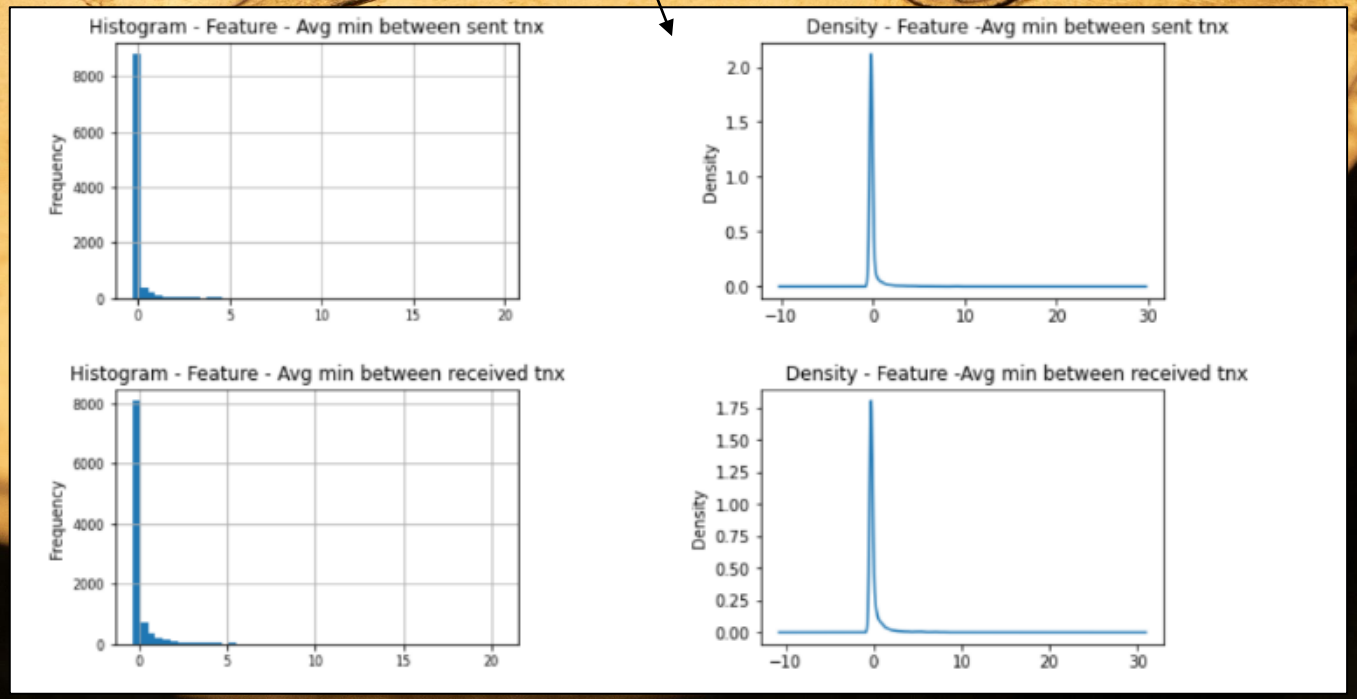
```
df.isnull().sum()[df.isnull().sum()>0]
```

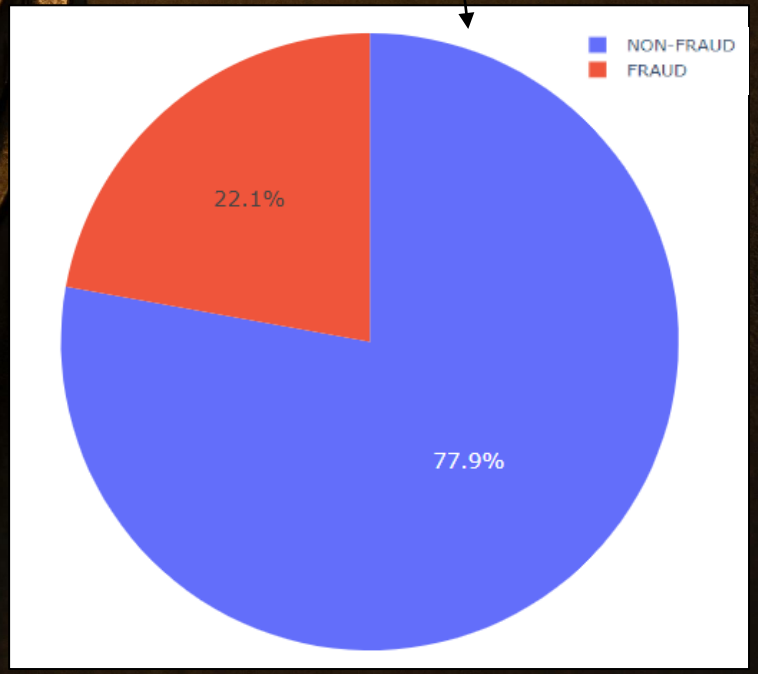
```
Total ERC20 tnxs                    829
ERC20 total Ether received          829
ERC20 total ether sent              829
ERC20 total Ether sent contract     829
ERC20 uniq sent addr                829
ERC20 uniq rec addr                 829
ERC20 uniq sent addr.1              829
ERC20 uniq rec contract addr        829
ERC20 avg time between sent tnx     829
ERC20 avg time between rec tnx      829
ERC20 avg time between rec 2 tnx    829
ERC20 avg time between contract tnx 829
ERC20 min val rec                   829
ERC20 max val rec                   829
ERC20 avg val rec                   829
ERC20 min val sent                  829
ERC20 max val sent                  829
ERC20 avg val sent                  829
ERC20 min val sent contract         829
ERC20 max val sent contract         829
ERC20 avg val sent contract         829
ERC20 uniq sent token name          829
ERC20 uniq rec token name           829
ERC20 most sent token type          841
ERC20_most_rec_token_type           851
dtype: int64
```

ניתוח ראשוני

נראה שאין התפלגות נורמלית,
רוב הערכים סביב הערך 0..

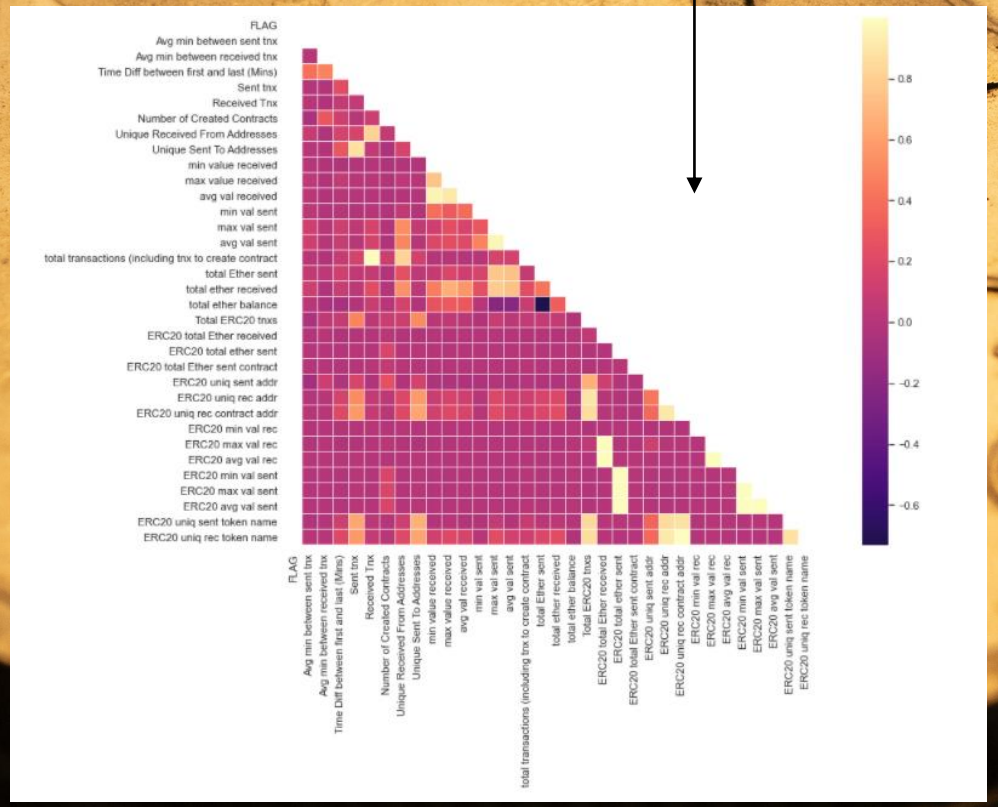
בסיס הנתונים אינו מאוזן!
יותר טרנזקציות לגיטימיות
מהונאות



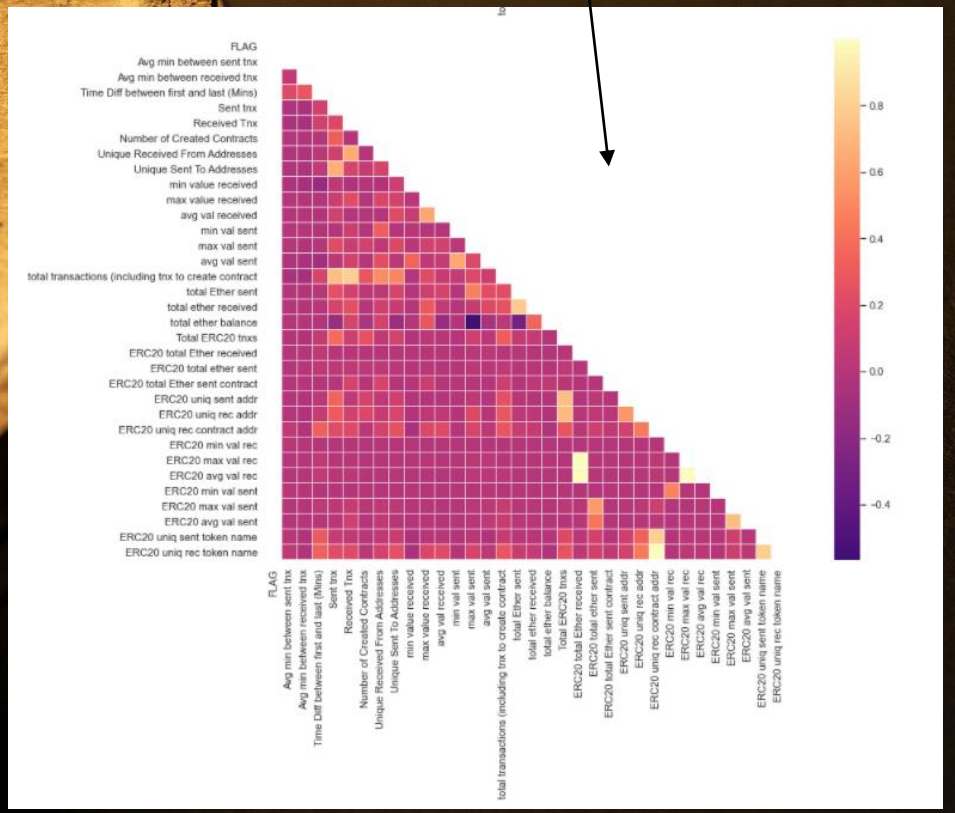


ניתוח ראשוני

קורלציה בין העמודות עבור כל הטרנזקציות שהן הונאה (1= FLAG)

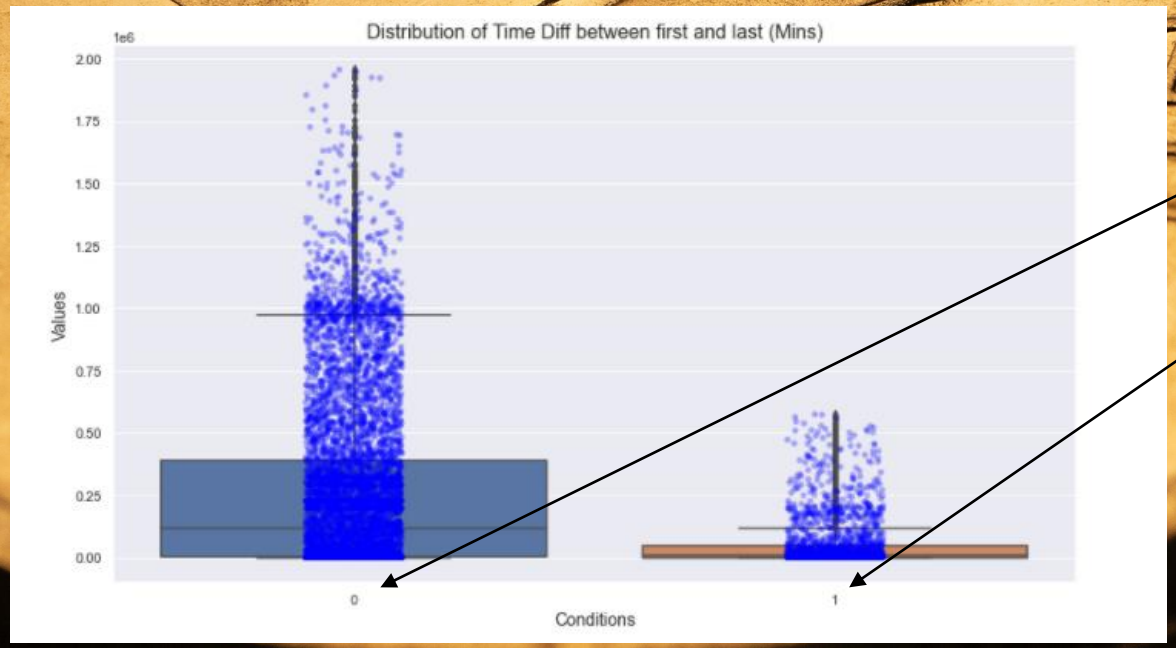


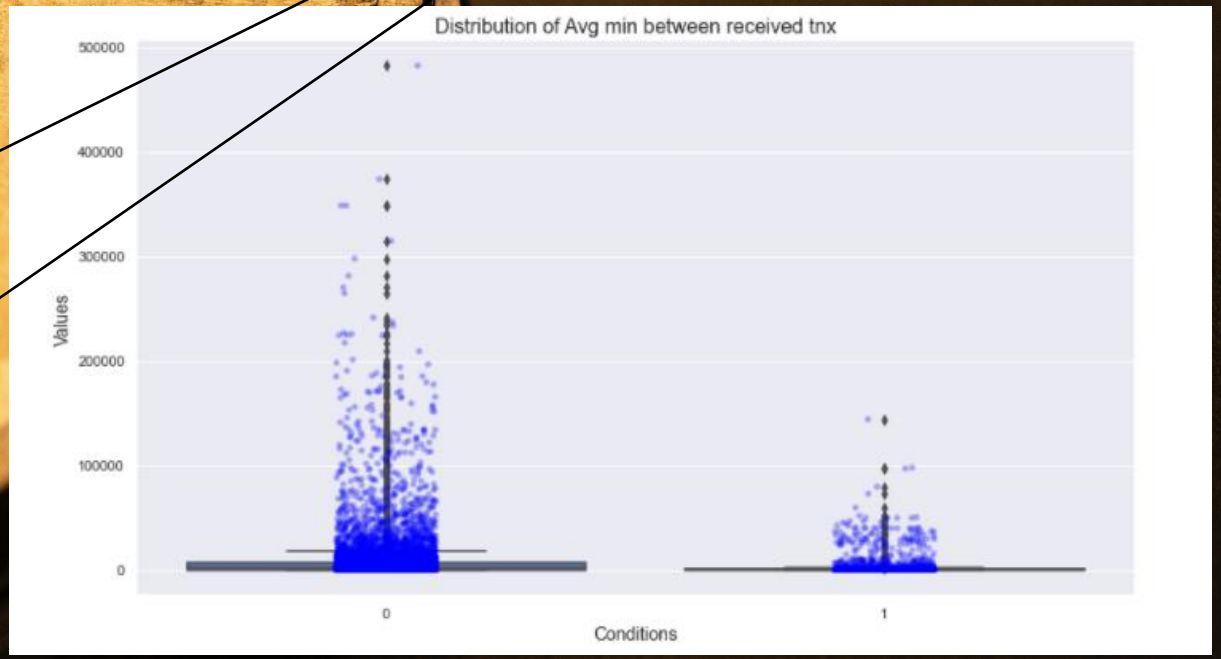
קורלציה בין העמודות עבור כל הטרנזקציות הלגיטימיות (0= FLAG)



ניתוח ראשוני

Box Plot – עבור כל עמודה,
כאשר מסתכלים בנפרד על
הונאות וטרנזקציות לגיטימיות





הכנה וניקוי הנתונים - חריגים

איך מוצאים ערכים חריגים?
IQR?

```
# finding more outliers and trying to improve the model
for i in df_copy.drop('FLAG',axis =1).columns:
    IQR = np.percentile(df_copy[i],75) -  np.percentile(df_copy[i],25)
    lower_limit = np.percentile(df_copy[i],25) - 1.5*IQR
    upper_limit = np.percentile(df_copy[i],75) + 1.5*IQR
    outliers_a = df_copy[i][df_copy[i] > upper_limit].shape \
    + df_copy[i][df_copy[i] < lower_limit].shape
    if outliers_a[0]/df_copy['Unique Sent To Addresses'].shape[0] <= 0.07:
        print(i)



outliers_a = df_copy['avg val sent'][df_copy['avg val sent'] > upper_limit].shape \
    + df_copy['avg val sent'][df_copy['avg val sent'] < lower_limit].shape
outliers_a




outliers_a = df_copy['avg val sent'][df_copy['avg val sent'] > upper_limit].index
df_droped_corr.drop(outliers_a,axis = 0)
```

```
avg val sent
```

לא מספיק טוב... ☹

הכנה וניקוי הנתונים - חריגים

קורלציה בין כל עמודה לבין
עמודת הסיווג – ערכים נמוכים

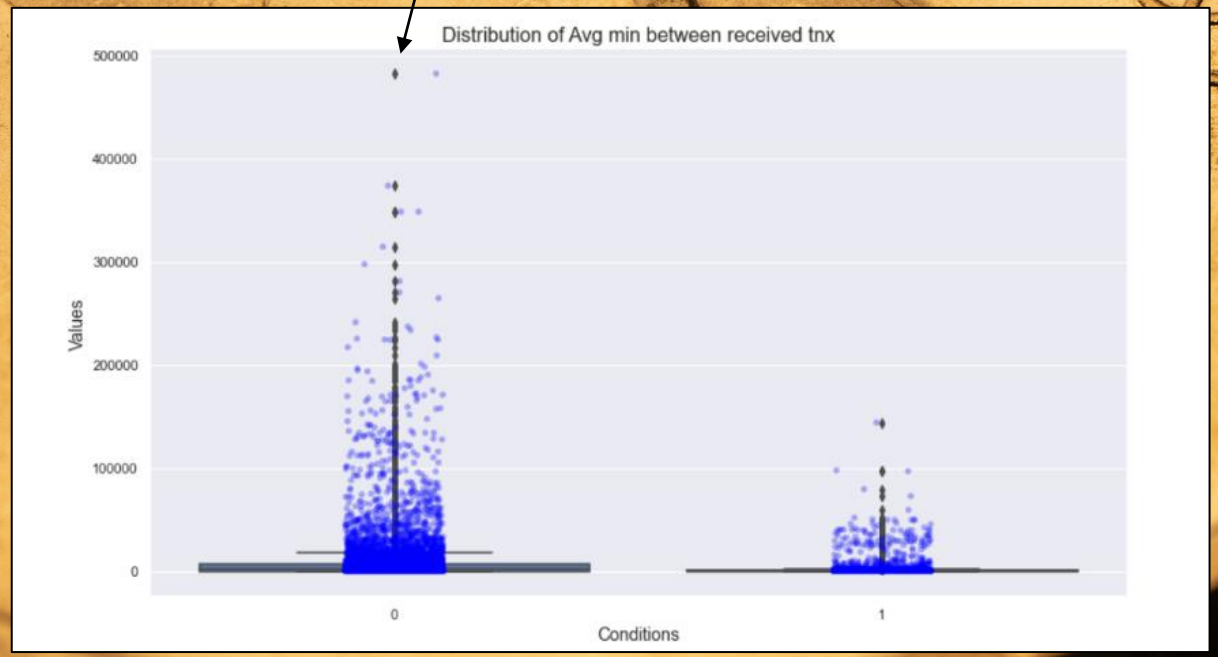
```
df_copy.corr()['FLAG'].sort_values(ascending= False)
```

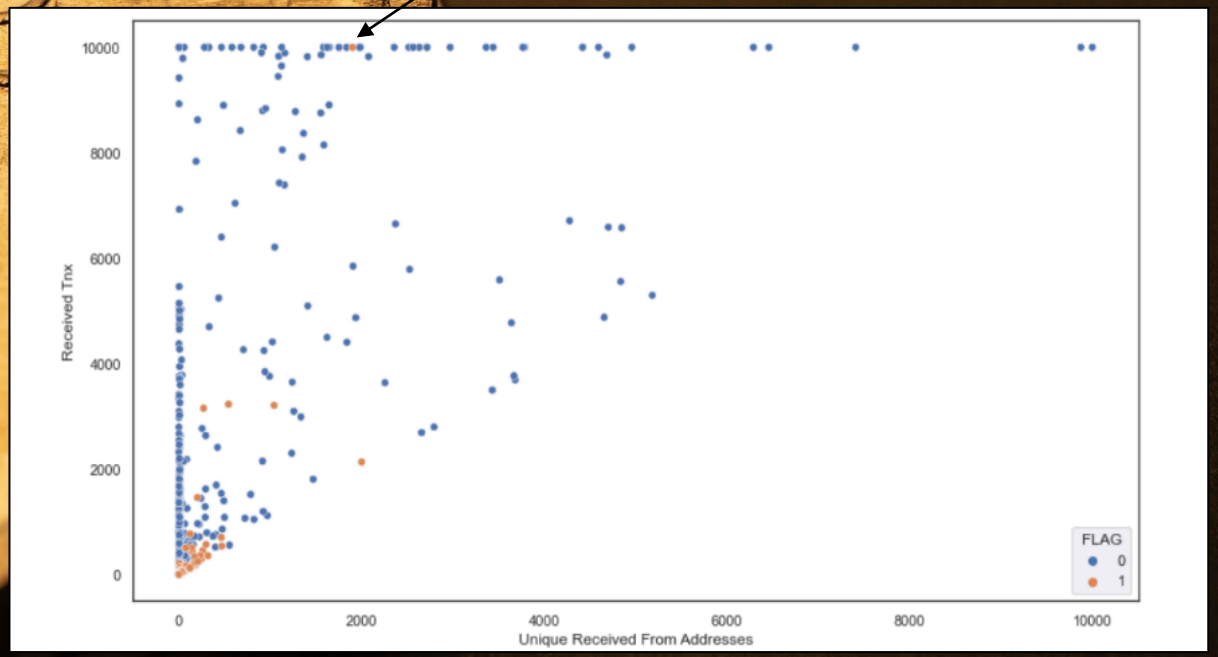
FLAG	1.000000
ERC20 min val sent	0.020860
ERC20 avg val sent	0.020597
ERC20 max val sent	0.020593
ERC20 total ether sent	0.020365
ERC20 total Ether sent contract	0.011106
ERC20 min val rec	0.009166
ERC20 uniq sent token name	0.007806
min val sent	0.006603
ERC20 avg val rec	0.006470
total ether balance	-0.003236
ERC20 max val rec	-0.003557
ERC20 total Ether received	-0.003690
avg val received	-0.011871
Number of Created Contracts	-0.013741
ERC20 uniq rec addr	-0.014456
ERC20 uniq rec token name	-0.014830
total Ether sent	-0.015022
ERC20 uniq rec contract addr	-0.015225
total ether received	-0.016934
ERC20 uniq sent addr	-0.016939
max value received	-0.019286
Total ERC20 tnxs	-0.021171
min value received	-0.021555
max val sent	-0.022436
Avg min between sent tnx	-0.029905
Unique Received From Addresses	-0.032031
Unique Sent To Addresses	-0.045654
avg val sent	-0.063358
Sent tnx	-0.078130
Received Tnx	-0.079493
total transactions (including tnx to create contract	-0.100485
Avg min between received tnx	-0.118750
Time Diff between first and last (Mins)	-0.269853

הכנה וניקוי הנתונים - חריגים

Outlier!

Outlier!





הכנה וניקוי הנתונים – ניתוח קורלציה

```python
#making a list of the features which is their variance is equal to 0
to_drop = list(df_copy.var()[df_copy.var() == 0].keys())
to_drop
```

```
['ERC20 avg time between sent tnx',
 'ERC20 avg time between rec tnx',
 'ERC20 avg time between rec 2 tnx',
 'ERC20 avg time between contract tnx',
 'ERC20 min val sent contract',
 'ERC20 max val sent contract',
 'ERC20 avg val sent contract']
```

מחיקת עמודות עם סטיית תקן 0 – אנחנו לא רוצים עמודות של קבועים !

```python
corr = df_copy.corr()
corr_df = corr[corr>0.6].dropna(axis = 1 ,thresh = 2)
corr_df = corr_df[corr_df != 1]
to_drop = corr_df[corr_df>0.9]
to_drop.values[to_drop.values>0]
to_drop.unstack().sort_values(kind="quicksort")[to_drop.unstack()>0]
```

```
ERC20 total ether sent         ERC20 min val sent              0.999311
ERC20 min val sent             ERC20 total ether sent          0.999311
ERC20 total ether sent         ERC20 avg val sent              0.999566
ERC20 avg val sent             ERC20 total ether sent          0.999566
ERC20 uniq rec contract addr   ERC20 uniq rec token name       0.999641
ERC20 uniq rec token name      ERC20 uniq rec contract addr    0.999641
ERC20 total ether sent         ERC20 max val sent              0.999649
ERC20 max val sent             ERC20 total ether sent          0.999649
ERC20 min val sent             ERC20 max val sent              0.999729
ERC20 max val sent             ERC20 min val sent              0.999729
ERC20 min val sent             ERC20 avg val sent              0.999785
ERC20 avg val sent             ERC20 min val sent              0.999785
ERC20 max val sent             ERC20 avg val sent              0.999952
ERC20 avg val sent             ERC20 max val sent              0.999952
ERC20 total Ether received     ERC20 max val rec               0.999967
ERC20 max val rec              ERC20 total Ether received      0.999967
dtype: float64
```

מחיקת עמודות עם קורלציה מעל 90 אחוז ! (מוחקים עמודה אחת מבין כל זוג).

הכנה וניקוי הנתונים – ניתוח קורלציה

```
df_copy.corr()['FLAG'].sort_values(ascending= False)
```

FLAG	1.000000
ERC20 min val sent	0.020860
ERC20 avg val sent	0.020597
ERC20 max val sent	0.020593
ERC20 total ether sent	0.020365
ERC20 total Ether sent contract	0.011106
ERC20 min val rec	0.009166
ERC20 uniq sent token name	0.007806
min val sent	0.006603
ERC20 avg val rec	0.006470
total ether balance	-0.003236
ERC20 max val rec	-0.003557
ERC20 total Ether received	-0.003690
avg val received	-0.011871
Number of Created Contracts	-0.013741
ERC20 uniq rec addr	-0.014456
ERC20 uniq rec token name	-0.014830
total Ether sent	-0.015022
ERC20 uniq rec contract addr	-0.015225
total ether received	-0.016934
ERC20 uniq sent addr	-0.016939
max value received	-0.019286
Total ERC20 tnxs	-0.021171
min value received	-0.021555
max val sent	-0.022436
Avg min between sent tnx	-0.029905
Unique Received From Addresses	-0.032031
Unique Sent To Addresses	-0.045654
avg val sent	-0.063358
Sent tnx	-0.078130
Received Tnx	-0.079493
total transactions (including tnx to create contract	-0.100485
Avg min between received tnx	-0.118750
Time Diff between first and last (Mins)	-0.269853

מחיקת עמודות עם קורלציה
לעמודת הlabel מתחת ל -2 אחוז

הכנה וניקוי הנתונים

```python
for i in df_copy.columns:
    if df_copy[i].isnull().sum() == 829:
        df_copy[i].replace({np.NaN:df_copy[i].mean()},inplace=True)
    else:
        pass
```

מילוי ערכים ריקים בעזרת הממוצע

```python
from sklearn.decomposition import PCA
pca_model = PCA(n_components= 15)
X = pca_model.fit_transform(X)
```

הורדת מימד – לא עזר בשיפור המודל..

```python
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(X_train)
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

נורמליזציה בשיטת Min-Max

הכנה וניקוי הנתונים

Label Encoding

עמודות קטגוריות

ERC20 most sent token type	ERC20_most_rec_token_type
1	1
2	2
3	3
4	3
5	4
...	...
3	5
262	400
262	390
262	250
3	5

```
print(df[' ERC20_most_rec_token_type'].nunique())
print(df[' ERC20 most sent token type'].nunique())
```

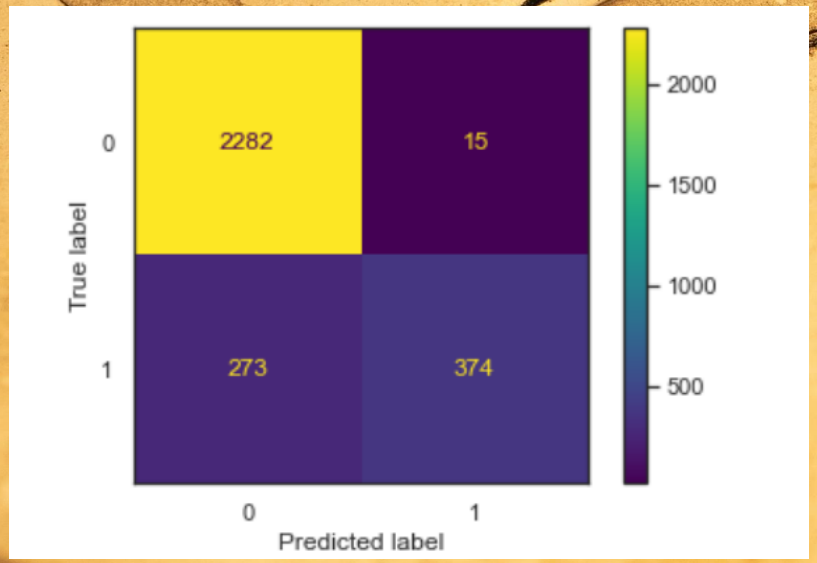
```
467
305
```

One-Hot?

ERC20 most sent token type	ERC20_most_rec_token_type
Cofoundit	Numeraire
Livepeer Token	Livepeer Token
None	XENON
Raiden	XENON
StatusNetwork	EOS
...	...
	GSENetwork
	Blockwell say NOTSAFU
	Free BOB Tokens - BobsRepair.com
None	OmiseGO
	INS Promo1

Models – Logistic Regression

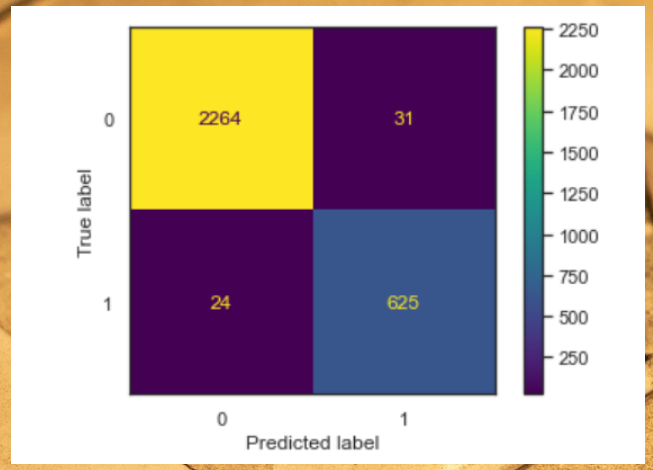
מאחר ובסיס הנתונים לא מאוזן – הריקול הוא הערך שנחפש למקסם



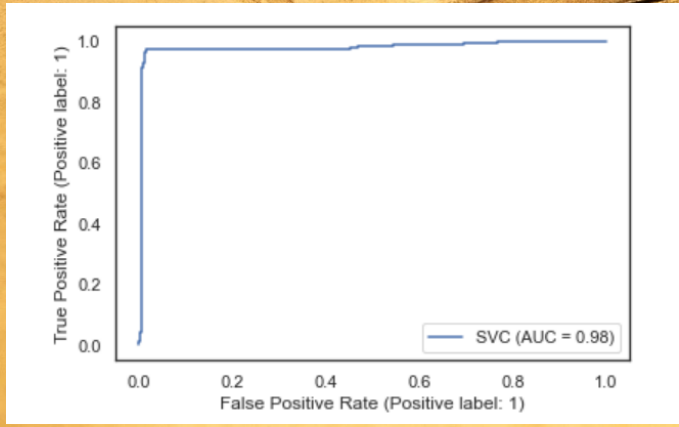
	precision	recall	f1-score	support
0	0.89	0.99	0.94	2297
1	0.96	0.58	0.72	647
accuracy			0.90	2944
macro avg	0.93	0.79	0.83	2944
weighted avg	0.91	0.90	0.89	2944

Models – SVM

מאחר ובסיס הנתונים לא מאוזן – הריקול הוא הערך שנחפש למקסם



	precision	recall	f1-score	support
0	0.99	0.99	0.99	2295
1	0.95	0.96	0.96	649
accuracy			0.98	2944
macro avg	0.97	0.97	0.97	2944
weighted avg	0.98	0.98	0.98	2944



Models – SVM

מציאת – Grid Search
הפרמטרים שממקסמים
את הריקול שלנו

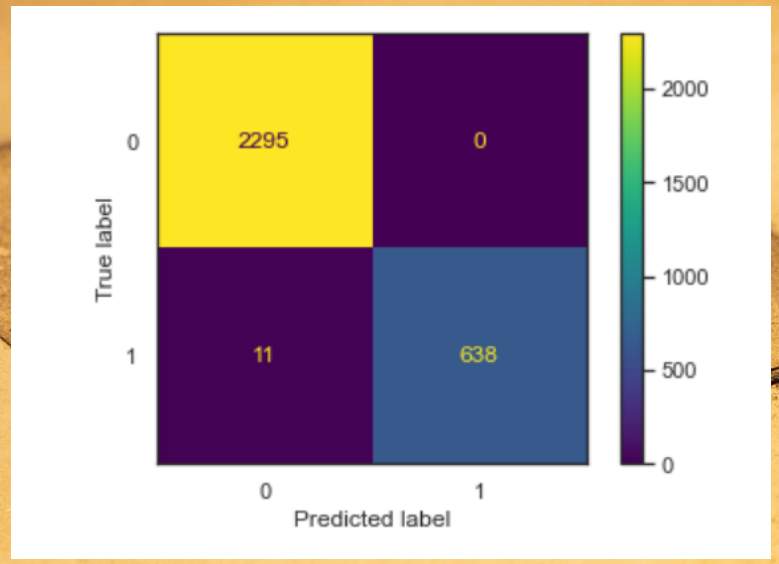
```
# Grid search for SVM
param_grid = {'C':[i for i in range(1,10,1)],'kernel':['linear','rbf','poly']}
grid = GridSearchCV(model,param_grid,scoring = 'recall')
grid.fit(X_train,y_train)
```

```
grid.best_params_
```

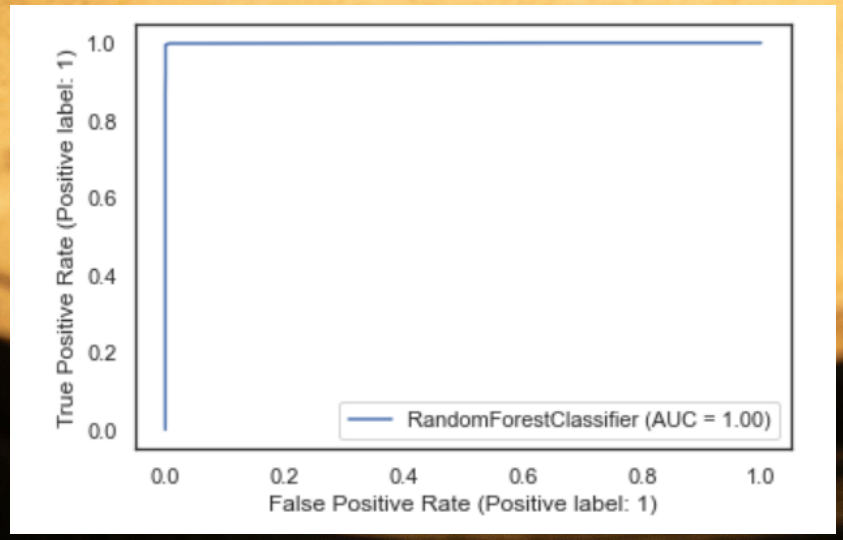
```
{'C': 9, 'kernel': 'rbf'}
```

Models – Random Forest

מאחר ובסיס הנתונים לא מאוזן – הריקול הוא הערך שנחפש למקסם



	precision	recall	f1-score	support
0	1.00	1.00	1.00	2295
1	1.00	0.98	0.99	649
accuracy			1.00	2944
macro avg	1.00	0.99	0.99	2944
weighted avg	1.00	1.00	1.00	2944



Future Work

- התייחסות לעליית/ירידת מקרי הונאה במהלך תקופות בהן הקריפטו יותר תנודתי מבחינת המחיר שלו
- יצירת מאגר פתוח לקהל הרחב, שכולל רשימה של כתובות שחשודות בטרנזקציות הונאה, כך שחברות יוכלו להימנע מלבצע עבורן טרנזקציות

תודה שהקשבתם!

נדב ישר
דניאל רביב