

Data modification

Data after modification: a small step for modification of data is done where number of features are reduced manually as per our convenience for better performance. Concatenation of wifi_ascii number and ip_ascii is done with '00' integer in between to differentiate .

```
example
IPAddress                192.168.1.8
ip_ascii                 558
wifi_name                 "B5_1201"
wifi_ascii               478
wifiip                   47800558
```

Concatenation of "start time" number and "usage Time" , first 6 digits represent time

```
StartTime                105247
UsageTime                1
start_usagetime          1052471
```

Battery , batterystatus and batterychargingsource are added

Battery is converted to tens i.e.

battery value	after conversion
34%	30
68%	70
Adapter/AC	2
USB	4
No Charging	0

Battery status :

Charging	1
Not Charging	0

Battery charging source

```
Battery                50
BatteryChargingSource  2
BatteryStatus          1
BBB                    53
```

To find out the little details about the behaviour of the battery, more constants are put to use. Using the constants BATTERY_STATUS_CHARGING and BATTERY_STATUS_CHARGING, we can find out if the battery is charging or not. If the battery is charging, *BStatus* is set to 1 which is otherwise 0.

To find the power source, if the battery is charging, two more constants are used. They are BATTERY_PLUGGED_AC and BATTERY_PLUGGED_USB. So, if the device is being charged through an

adapter, *BattPowerSource* is set to 2 and if it's being charged through USB, *BattPowerSource* is set to 4. If the battery isn't being charged, then the default value of *BattPowerSource* is 0. The end results helps us In easy break down to figure out all the 3 features .

Data Representation

Data read from json file: Data has been read from the json file and data frames created this is a small example of how the data looks under following circumstances .

	Acc_x	Acc_y	Acc_z	Battery	BatteryChargingSource
0	0.693868	0.726671	0.720240	17	0
1	0.693868	0.726671	0.720240	17	0
2	0.693868	0.726671	0.720240	17	0
3	0.693868	0.726671	0.720240	17	0
4	0.693868	0.726671	0.720240	17	0
5	0.693868	0.726671	0.720240	17	0
6	0.693868	0.726671	0.720240	17	0
7	0.223644	1.520575	3.491794	93	4
8	0.184751	1.218227	4.805725	93	4
9	0.184751	1.218227	4.805725	93	4
10	0.351918	0.965198	3.758030	93	4

BatteryStatus	ClickEventID0
0	[Play video]
1	[Play video]
2	[Play video]
3	[Play video]
4	[Play video]
5	[Play video]
6	[Play video]
7	[1, Play Store]
8	NaN
9	[My Files]
10	[Internal storage, 35.42 GB / 64.00 GB]

	RKB	StartTime	TKB	TapCount	NoOfClickEvents	UTC
0	6513506	00:10:15	461335	0	1	1544208226110
1	6513506	00:10:15	461335	0	1	1544208226110
2	6513506	00:10:15	461335	0	1	1544208226110
3	6513506	00:10:15	461335	0	1	1544208226110
4	6513506	00:10:15	461335	0	1	1544208226110
5	6513506	00:10:15	461335	0	1	1544208226110
6	6513506	00:10:15	461335	0	1	1544208226110
7	1	16:52:25	0	4	1	1544181754889
8	6132948	16:52:34	446096	3	0	1544181758184
9	2	16:52:38	0	1	1	1544181759915
10	6132950	16:52:39	446098	10	3	1544181776183

UsageTime	latitude	longitude	meanAcc_x	meanAcc_y	meanAcc_z	\
0	25	12.922172	77.492920	0.007397	-0.008721	9.811530
1	25	12.922172	77.492920	0.007397	-0.008721	9.811530
2	25	12.922172	77.492920	0.007397	-0.008721	9.811530
3	25	12.922172	77.492920	0.007397	-0.008721	9.811530
4	25	12.922172	77.492920	0.007397	-0.008721	9.811530
5	25	12.922172	77.492920	0.007397	-0.008721	9.811530
6	25	12.922172	77.492920	0.007397	-0.008721	9.811530
7	0	12.920037	77.683395	0.158140	-1.075209	2.469071
8	0	12.920037	77.683395	0.102990	-0.744553	4.849455
9	1	12.920037	77.683395	0.102990	-0.744553	4.849455

Data Preprocessing

Data after modification: a small step for modification of data is done where number of features are reduced manually as per our convenience for better performance.

	Acc_x	Acc_y	Acc_z	Battery	BatteryChargingSource
0	0.693868	0.726671	0.720240	20.0	0
1	0.693868	0.726671	0.720240	20.0	0
2	0.693868	0.726671	0.720240	20.0	0
3	0.693868	0.726671	0.720240	20.0	0
4	0.693868	0.726671	0.720240	20.0	0
10	0.351918	0.965198	3.758030	90.0	4
...					
796	245.345547	75.244381	186899.823606	70.0	4
797	88.234374	26.610375	67445.676629	70.0	4
798	184.428848	53.228362	142788.432024	70.0	4
799	192.012064	53.236510	150274.834726	70.0	4

	BatteryStatus	ClickEventID0
0	0	[Play video]
1	0	[Play video]
2	0	[Play video]
3	0	[Play video]
4	0	[Play video]
5	0	[Play video]
9	0	[My Files]
10	0	[Internal storage, 35.42 GB / 64.00 GB]
..
797	0	-3
798	0	-3
799	0	-3

[800 rows x 7 columns]

	RKB	StartTime	TKB	TapCount	TotalNoOfClickEvents	\
0	6513506	1015.0	461335	0	1	
1	6513506	1015.0	461335	0	1	
2	6513506	1015.0	461335	0	1	
10	6132950	165239.0	446098	10	3	
..	
796	9	143253.0	9	4	1	
797	4113329	143303.0	288165	59	1	
798	9	143346.0	9	4	0	
799	9	143402.0	9	10	3	

	UTC	UsageTime	latitude	longitude	meanAcc_x	\
0	1544208226110	25	12.922172	77.492920	0.007397	
1	1544208226110	25	12.922172	77.492920	0.007397	
2	1544208226110	25	12.922172	77.492920	0.007397	
3	1544208226110	25	12.922172	77.492920	0.007397	
..	
798	1544605440041	13	12.920036	77.683361	43.622288	
799	1544605454657	11	12.920036	77.683361	43.994694	

	meanOrien_y	meanOrien_z	package_ascii	ip_ascii	\
0	-17.345741	1.221795	1876.0	653	

1	...	-17.345741	1.221795	1876.0	653
2	...	-17.345741	1.221795	1876.0	653
3	...	-17.345741	1.221795	1876.0	653
4	...	-17.345741	1.221795	1876.0	653
5	...	-17.345741	1.221795	1876.0	653
6	...	-17.345741	1.221795	1876.0	653
7	...	-25.223969	-3.789785	2452.0	648
8	...	-25.782839	-2.346300	1351.0	648
9	...	-25.782839	-2.346300	2581.0	648
10	...	-27.219486	1.669745	1901.0	648
..
796	...	-56011.417969	-8164.063477	2581.0	648
797	...	-12459.999023	-1812.444214	1699.0	648
798	...	-37401.636719	-5432.669434	3015.0	648
799	...	-37427.093750	-5434.396484	1901.0	648

	Network_Provider	NetworkProvider_ascii	BBB	wifiip	UKB
0	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
1	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
2	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
3	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
4	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
5	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
6	"Xiaomi_3C74"	843.0	20.0	843.000653	6052171.0
7	-3	96.0	94.0	96.000648	1.0
8	"SYMC-MyWiFi"	898.0	94.0	898.000648	5686852.0
9	"SYMC-MyWiFi"	898.0	94.0	898.000648	2.0
10	"SYMC-MyWiFi"	898.0	94.0	898.000648	5686852.0
..
796	-3	96.0	75.0	96.000648	0.0
797	-3	96.0	74.0	96.000648	3825164.0
798	-3	96.0	74.0	96.000648	0.0
799	-3	96.0	74.0	96.000648	0.0

	start_usagetime
0	1015.0250
1	1015.0250
2	1015.0250
3	1015.0250
4	1015.0250
5	1015.0250
6	1015.0250
7	165225.0000
8	165234.0000
9	165238.0100
10	165239.0140
..	...
795	143217.0260
796	143253.0900
797	143303.0410
798	143346.0130
799	143402.0110

Data after conversion to numbers : Local outlier factor works with numbers only hence its essential to convert all the features to appropriate numbers :

	start_usagetime	package_ascii	UsageTime	StartTime	DayOfTheWeek	\
0	1015.025	1876.0	25.0	1015.0	7.0	
1	1015.025	1876.0	25.0	1015.0	7.0	
2	1015.025	1876.0	25.0	1015.0	7.0	
3	1015.025	1876.0	25.0	1015.0	7.0	
4	1015.025	1876.0	25.0	1015.0	7.0	
5	1015.025	1876.0	25.0	1015.0	7.0	
6	1015.025	1876.0	25.0	1015.0	7.0	
7	165225.000	2452.0	0.0	165225.0	6.0	
8	165234.000	1351.0	0.0	165234.0	6.0	
9	165238.010	2581.0	1.0	165238.0	6.0	

	wifiip	BBB	latitude	longitude	TapCount	...	Network	\
0	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
1	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
2	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
3	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
4	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
5	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
6	843.000653	20.0	12.922172	77.492920	0.0	...	1.0	
7	96.000648	94.0	12.920037	77.683395	4.0	...	1.0	
8	898.000648	94.0	12.920037	77.683395	3.0	...	1.0	
9	898.000648	94.0	12.920037	77.683395	1.0	...	1.0	

	meanAcc_x	meanAcc_y	meanAcc_z	meanGyro_x	meanGyro_y	meanGryo_z	\
0	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
1	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
2	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
3	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
4	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
5	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
6	0.007397	-0.008721	9.811530	0.991565	-0.016973	0.011061	
7	0.158140	-1.075209	2.469071	1.000000	0.000000	0.000000	
8	0.102990	-0.744553	4.849455	1.000000	0.000000	0.000000	
9	0.102990	-0.744553	4.849455	1.000000	0.000000	0.000000	

	meanOrien_x	meanOrien_y	meanOrien_z
0	1.263105	-17.345741	1.221795
1	1.263105	-17.345741	1.221795
2	1.263105	-17.345741	1.221795
3	1.263105	-17.345741	1.221795
4	1.263105	-17.345741	1.221795
5	1.263105	-17.345741	1.221795
6	1.263105	-17.345741	1.221795
7	-68.047508	-25.223969	-3.789785
8	-77.375458	-25.782839	-2.346300
9	-77.375458	-25.782839	-2.346300

Data segregated :

Data is now being segregated with respect to one characteristic for easy better understanding and efficient working of the algorithm

1506.0

50

	start_usagetime	UsageTime	StartTime	wifiip	BBB	latitude
2	140848.04000	4.0	140848.0	898.000659	55.0	12.920027
10	141424.07036	7036.0	141424.0	898.000659	44.0	12.920036
12	161142.00000	0.0	161142.0	898.000659	44.0	12.920036
...						
170	152834.01000	1.0	152834.0	898.000648	70.0	12.920035
172	153034.09100	91.0	153034.0	898.000648	70.0	12.920035
217	204230.02320	2320.0	204230.0	579.000650	40.0	13.027343
224	62729.03000	3.0	62729.0	579.000650	20.0	13.027338
229	114558.04113	4113.0	114558.0	898.000648	25.0	12.920037

	longitude	TapCount	UKB	Network	meanAcc_x	meanAcc_y	\
2	77.683363	5.0	1277.0	1.0	-3.000000	-3.000000	
10	77.683348	650.0	54000.0	1.0	0.018649	-0.027941	
12	77.683348	0.0	54000.0	1.0	1.646033	-2.319430	
...							
202	77.683363	10.0	288056.0	1.0	30.969275	3.572598	
206	77.622356	4.0	4059554.0	1.0	-3.000000	-3.000000	
214	77.622356	36.0	4064792.0	1.0	-2.412123	-1.355919	
217	77.622377	107.0	47278.0	1.0	-0.190471	-0.012298	
224	77.622355	9.0	15716.0	1.0	-3.000000	-3.000000	
229	77.683421	711230.0	197457.0	1.0	-1.389972	2.836159	

	meanAcc_z	meanGyro_x	meanGyro_y	meanOrien_x	meanOrien_y	\
2	-3.000000	-3.000000	-3.000000	-3.000000	-3.000000	
10	10.453092	0.795230	-0.114234	-104.496223	-24.085508	
12	884.952026	67.031052	-9.726879	-8940.844727	-2066.784912	
...						
202	5416.623535	461.266876	-68.685883	-65242.312500	-4768.722168	
206	-3.000000	-3.000000	-3.000000	-3.000000	-3.000000	
214	718.702942	63.706146	-3.076397	4665.310547	-1299.573853	
217	49.055882	4.408081	-0.205122	357.110840	-82.503586	
224	-3.000000	-3.000000	-3.000000	-3.000000	-3.000000	
229	8.608446	0.762185	-0.112540	-144.143967	-5.640756	

Feature scaling

Feature scaling can vary your results a lot while using certain algorithms and have a minimal or no effect in others. To understand this, let's look why features need to be scaled, varieties of scaling methods and when we should scale our features.

The transformation is given by:

```
class sklearn.preprocessing.MinMaxScaler(feature_range=(0, 1), copy=True)
```

$$X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))$$

$$X_scaled = X_std * (max - min) + min$$

where min, max = feature_range.

This transformation is often used as an alternative to zero mean, unit variance scaling. we have scaled it from 0 to 1 this example for 50 entries, some of which is illustrated below

```

[[5.41659147 1.00454804 5.4165907 6.7737557 4.83561644 8.93409781
8.99999402 1.00005624 1.09347315 9. 1. 8.66075404
1. 1. 8.92304205 8.85383069 9. 1.1560556
8.8419982 ]
[5.44915833 9. 5.44915584 6.7737557 3.63013699 8.9341031
8.99999243 1.00731128 1.19605818 9. 1.01832578 8.68338479
1.00519997 1.01912341 8.99974755 8.85065275 8.99251198 1.15617168
8.9959794 ]
[6.56394512 1. 6.56394654 6.7737557 3.63013699 8.9341031
8.99999243 1. 1.19605818 9. 1.02820539 8.66593625
1.34321647 1.35287258 8.74423733 8.57397814 8.26709572 1.14991412
8.84116564]

.....
[5.99767572 1.45252985 5.99767493 6.7737555 7.57534247 8.93410279
8.99999407 1.00465672 8.04071136 9. 1.01880678 8.68489764
1.02034874 1.03549193 8.99149485 8.83552223 8.96776704 1.1547206
8.97591185]
[6.05986453 1.13757817 6.0598653 6.7737555 6.47945205 8.93410279
8.99999407 1.00064114 1.30231093 9. 1.02171211 8.68811959
1.07712339 1.0959217 8.95765089 8.7792554 8.88873639 1.15091103
8.90456104]
[8.89263605 1.00113701 8.8926368 1. 3.19178082 8.9999914
8.99371141 1.00004499 8.98980824 9. 1. 8.66075404
1. 1. 8.92304205 8.85383069 9. 1.1560556
8.8419982 ]
[9. 3.63786242 9. 1. 3.19178082 8.99999951
8.99371355 1.00120355 1.18297895 9. 1.01705624 8.6835039
1.02012095 1.03732785 8.99733168 8.86510611 8.97176619 1.15356721
8.99983307]

```

Featuring Extraction :

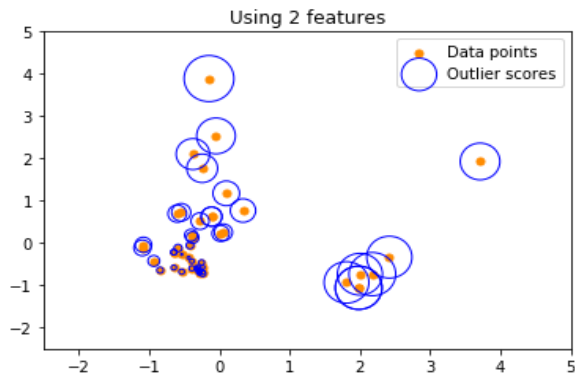
PCA for Data Visualization

For a lot of machine learning applications it helps to be able to visualize your data. Visualizing 2 or 3 dimensional data is not that challenging. However, even the Iris dataset used in this part of the tutorial is 4 dimensional. You can use PCA to reduce that 4 dimensional data into 2 or 3 dimensions so that you can plot and hopefully understand the data better.

```

1506.0
50
[[-0.39091245 -0.59880565]
[-0.53553262 -0.67791003]
[-0.38956116 -0.42892372]
[ 1.80910667 -0.93189361]
[ 2.17668607 -0.7356885 ]
[ 1.9818238 -1.05886989]
[ 1.98182425 -1.0588727 ]
[ 2.41818572 -0.33480851]
[ 1.99992724 -0.7383986 ]
[ 3.70877667 1.92837814]
[-0.24357631 1.76721875]
[-0.59810462 0.69472954]
[-0.54254008 0.72577948]
[-0.64609398 -0.21630954]
[-0.3760541 2.10260188]

```

In the above example its clear we have one point that is away from the clusters formed at (2,4) which is an outlier in this case

MORE EXAMPLES

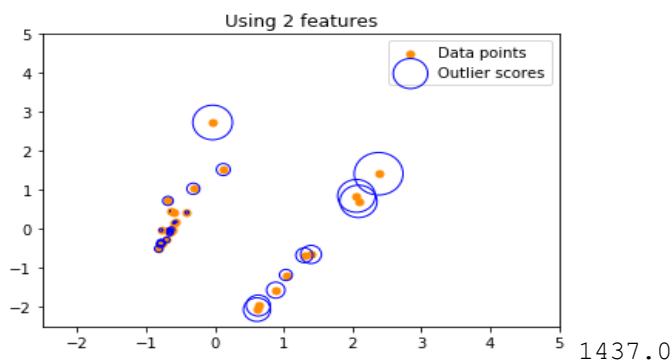
Below are some more examples :

1676.0

30

```
[ -0.99262362 -0.99352984 -0.99219981 -0.99299604 -0.9927245  -1.01513507
  -1.03607362 -1.22182052 -1.11173622 -1.04456439 -1.0612055  -1.08588694
  -1.39147425 -1.22740646 -0.98927265 -0.99180798 -1.25468828 -0.98715462
  -1.01565254 -0.98917631 -1.01954713 -0.98843273 -0.98715462 -0.98819037
  -1.00586043 -0.99834542 -0.99702134 -0.99695785 -0.99693728 -0.99727592]
```

```
[ 1  1  1  1  1  1  1  1  1  1  1  1  1 -1  1  1  1  1  1  1  1  1  1  1
  1  1  1  1  1]
```

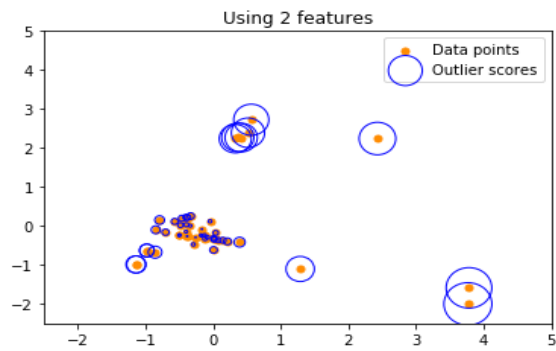


1437.0

48

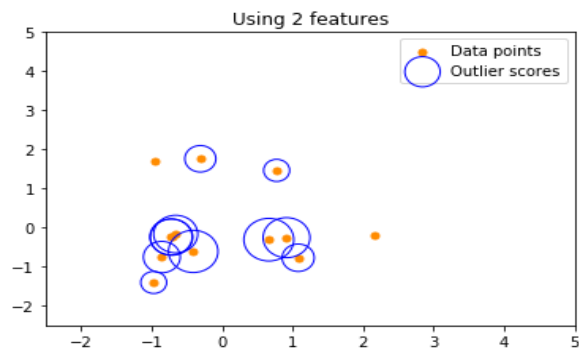
```
[ -4.18525931 -4.19639006 -4.13221772 -4.37477346 -4.76678212 -5.10709416
  -1.11478634 -0.98275663 -1.02906017 -1.00651811 -1.02800661 -0.94282049
  -0.96962272 -1.08005781 -3.42231994 -1.00136803 -0.9922148  -1.03354717
  -1.04442243 -0.96017617 -0.9682363  -0.97008511 -1.12824256 -1.07013565
  -1.03393289 -0.96183809 -7.41271652 -8.07582049 -1.10643655 -0.94085214
  -0.96544425 -0.9368061  -1.0817918  -1.07566655 -1.03226517 -0.94392047
  -0.96512387 -1.32489227 -1.08330646 -0.97908361 -1.18712314 -1.59694368
  -1.59691947 -1.09712881 -1.51715745 -2.1049614  -2.09669238 -1.23306459]
```

```
[ 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  1  1  1 -1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1]
```



```
1467.0
13
[-1.01456096 -1.00589109 -1.00568361 -1.01086961 -0.98345281 -0.99186262
-0.99547572 -0.98345281 -1.01394016 -0.99632194 -0.99179927 -1.00032105
-1.00763052]

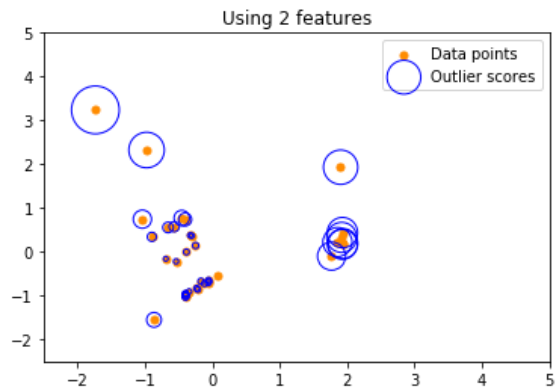
[-1  1  1  1  1  1  1  1  1  1  1  1  1]
```



```

1604.0
36
[-1.59688766 -1.61441973 -1.59207572 -1.56649541 -1.79853341 -1.6049325
-1.52504853 -0.99282872 -1.04680351 -1.2010669 -0.9902614 -0.982243
-1.02540672 -1.04112186 -0.98502389 -1.13502463 -0.98473675 -0.9941025
-2.57722725 -1.08806702 -1.85707811 -0.96678827 -0.98913972 -0.98619318
-0.98970255 -0.99063679 -0.98900619 -0.99411555 -0.99745667 -0.99101345
-0.9910651 -1.00260642 -0.9930226 -0.9856916 -1.12092636 -0.98784328]

[ 1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1  -1   1   1   1   1   1
 1   1   1   1   1   1   1   1   1   1   1]
```



Roc curve :

ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

one such graph is shown which has :12 entries having

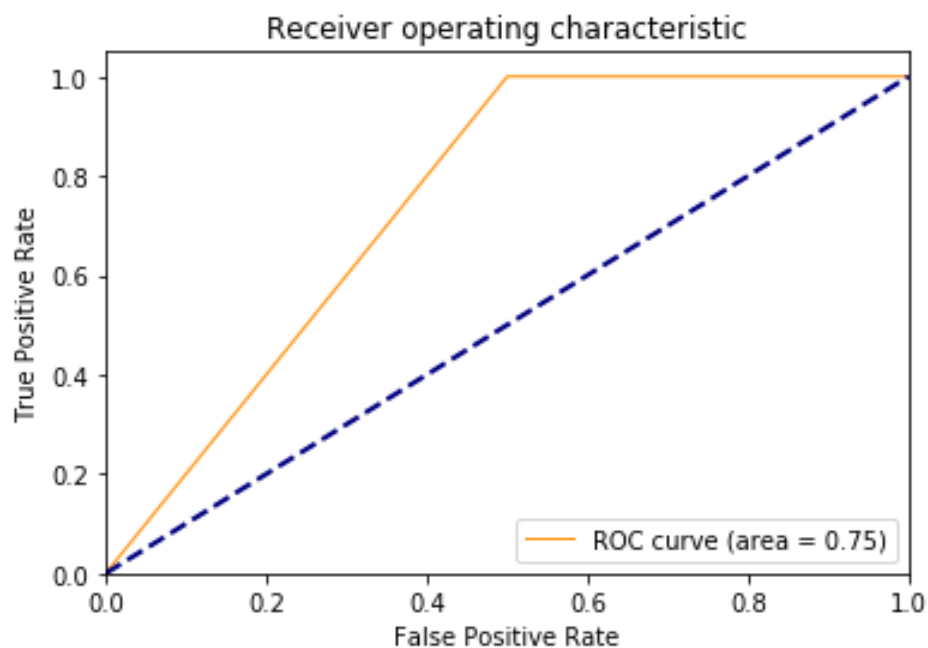
1 True positive (last entry was used by a different user)

1 True negative (last but one)

0 False positive

1506

[1 1 1 1 1 1 1 1 1 1 -1 -1]



Conclusion:

Once the data was retrieved from the local storage or the cloud, the algorithm was put to use, and outliers were identified. The precision and accuracy of the algorithm depended on the usage of the applications in question. Below is an statistical study of the results .

How was It done: Application was used by the owner for two days and later was used by us about few apps for sometime and then data was projected as below .

USER1 :

Analysis True Positive : 97.69%

Fasle positive : 3.68%

USER2:

Analysis True Positive : 99.43%

Fasle positive : 0.56%

True Negative: 33%

Disclaimer : Data collected and used is limited and the percentage is w.r.t only this data.

