Epileptic Prediction

Using data from EEG

By Venkata R Bandaru

Dataset

- The Dataset captures reading from EEG machine for 500 patients.
- Each row in the dataset contains 178 attributes /independent variable.
- Each attribute contains reading from the EEG data at specific time interval.
- Each row contains a response variable.
- The dataset can be found at here
- Prediction on the dataset can be treated as a classification

Motivation for designing a classifier - 1

- A study indicated the annual cost of the estimated 2.3 million epileptic cases was projected to be $12.5 billion.
- According to the study high cost was incurred for diagnostic evaluation and initial treatment.
- Another study found that the annual cost of per patient was about 1000 Euros. The mean cost for the test performed was 174.8 Euros.
- Accurately predicting if the patient has epileptic disorders will help in reducing the costs.
- The classifier obtained by the training on the data will help in accurately diagnosing an epileptic patient .
- The software can especially be embedded in ambulatory EEG devices to indicate possible seizure occurrences.

Approach

- Perform Exploratory data analysis.
- Perform Inferential statistics
- Pre processing steps based on Data analysis.
- Run classification algorithms
- Identify the best classification algorithm
- Create a webservice to make the model available

Exploratory Data Analysis - 1

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
X21.V1.791	135	190	229	223	192	125	55	-9	-33	-38	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
X15.V1.924	386	382	356	331	320	315	307	272	244	232	...	164	150	146	152	157	156	154	143	129	1
X8.V1.1	-32	-39	-47	-37	-32	-36	-57	-73	-85	-94	...	57	64	48	19	-12	-30	-35	-35	-36	5
X16.V1.60	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
X20.V1.54	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

Y is the response or dependent variable
X1..X178 are the readings from the EEG.

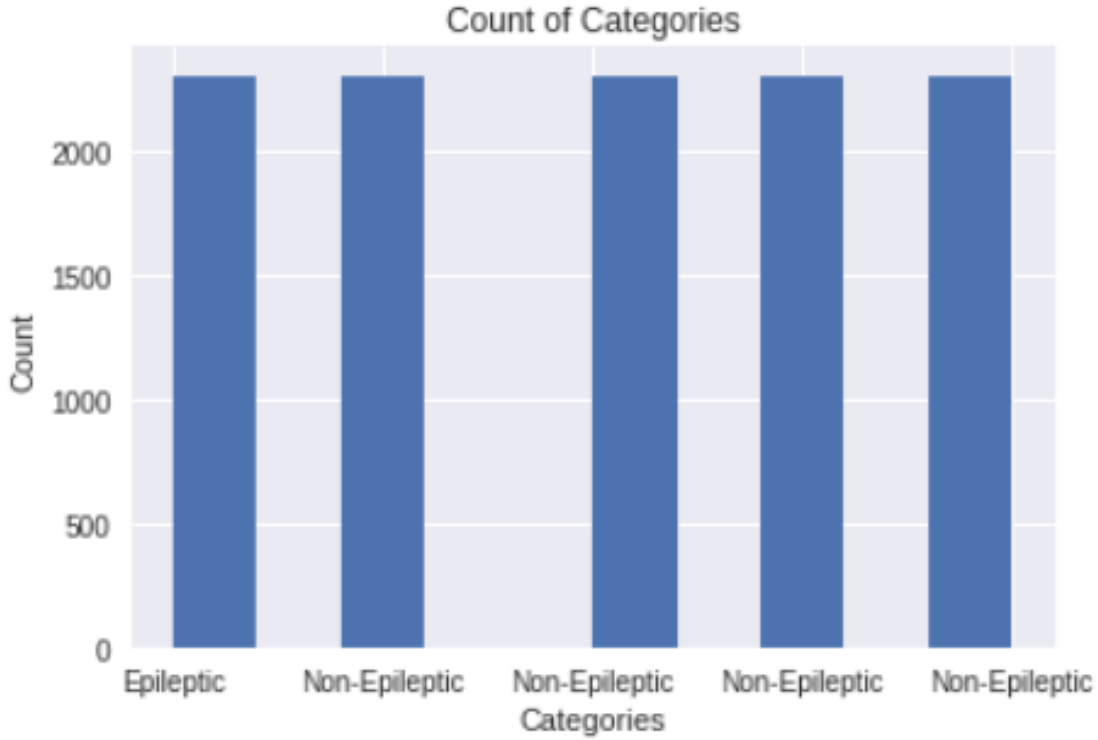
Exploratory Data Analysis - 2

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...
count	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.00000	11500.00000	11500.000000	...
mean	-11.581391	-10.911565	-10.187130	-9.143043	-8.009739	-7.003478	-6.502087	-6.68713	-6.55800	-6.168435	...
std	165.626284	166.059609	163.524317	161.269041	160.998007	161.328725	161.467837	162.11912	162.03336	160.436352	...
min	-1839.000000	-1838.000000	-1835.000000	-1845.000000	-1791.000000	-1757.000000	-1832.000000	-1778.00000	-1840.00000	-1867.000000	...
25%	-54.000000	-55.000000	-54.000000	-54.000000	-54.000000	-54.000000	-54.000000	-55.00000	-55.00000	-54.000000	...
50%	-8.000000	-8.000000	-7.000000	-8.000000	-8.000000	-8.000000	-8.000000	-8.00000	-7.00000	-7.000000	...
75%	34.000000	35.000000	36.000000	36.000000	35.000000	36.000000	35.000000	36.00000	36.00000	35.250000	...
max	1726.000000	1713.000000	1697.000000	1612.000000	1518.000000	1816.000000	2047.000000	2047.00000	2047.00000	2047.000000	...

A cursory glance at the Summary statistics shows

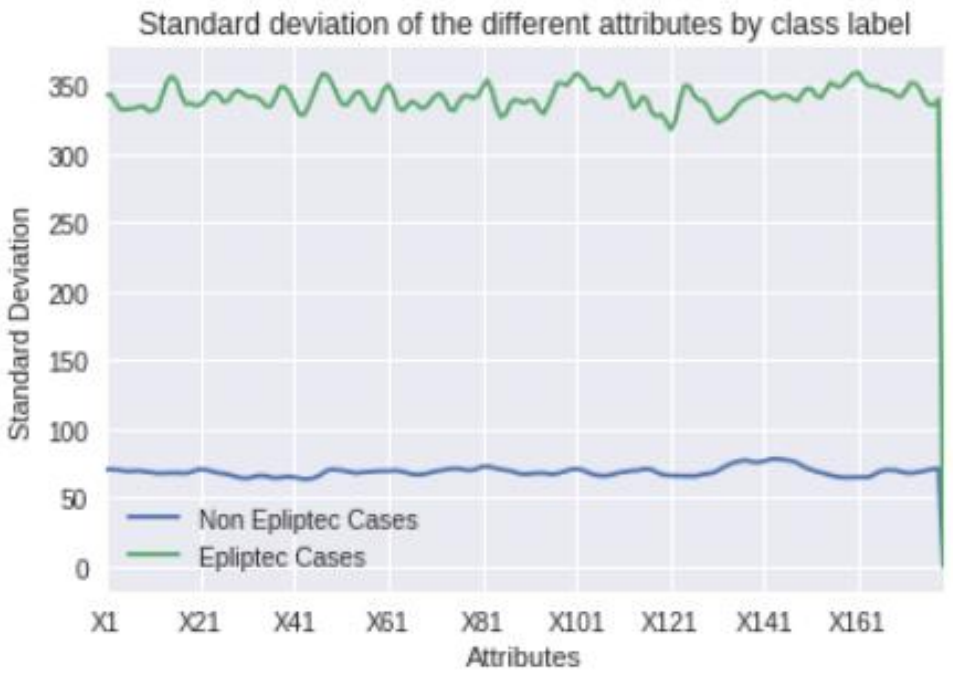
- No missing values
- There is no need to normalize the dataset, since all variables have the same scale.

Exploratory Data Analysis - 3



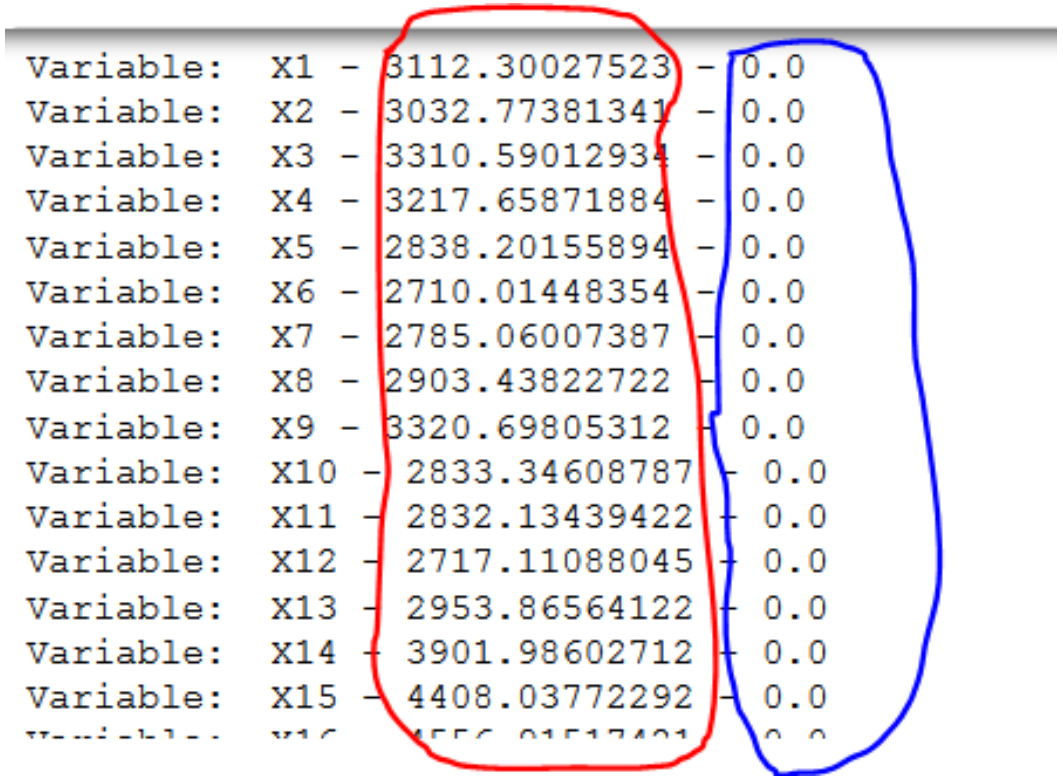
There are 5 different labels/ Dependent variables.
Label 1 indicates an Epileptic patient. All other Labels indicate a Non Epileptic patient.
There is a class imbalance between the Epileptic and a Non Epileptic patient. We will handle this during the pre processing of the data.

Exploratory Data Analysis - 4



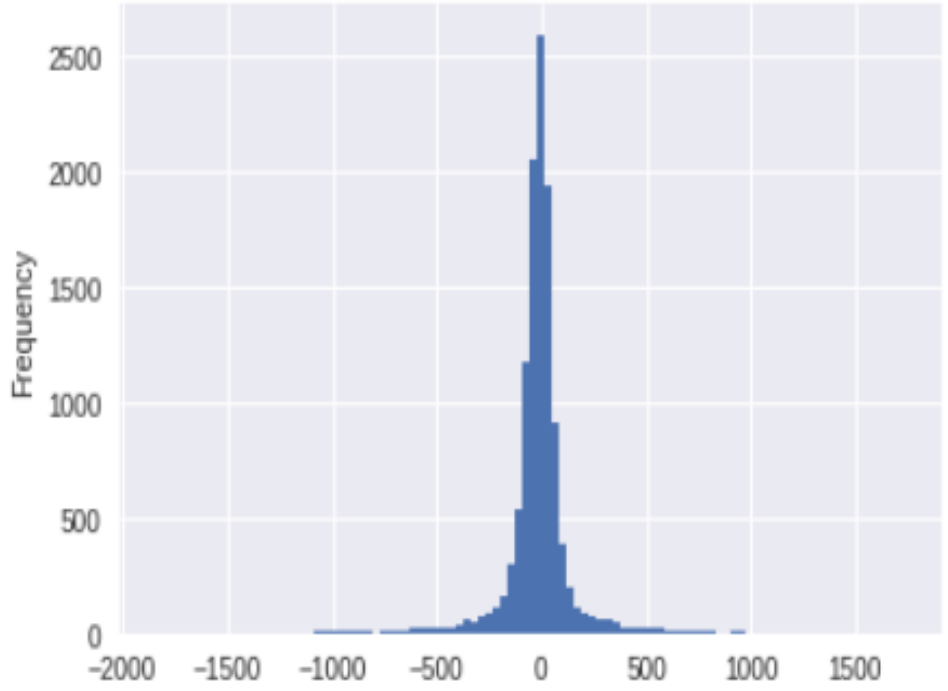
It can be clearly seen that Epileptic cases (Y=1) have a high variation across all attributes

Inferential Statistics – Test for Normality -1

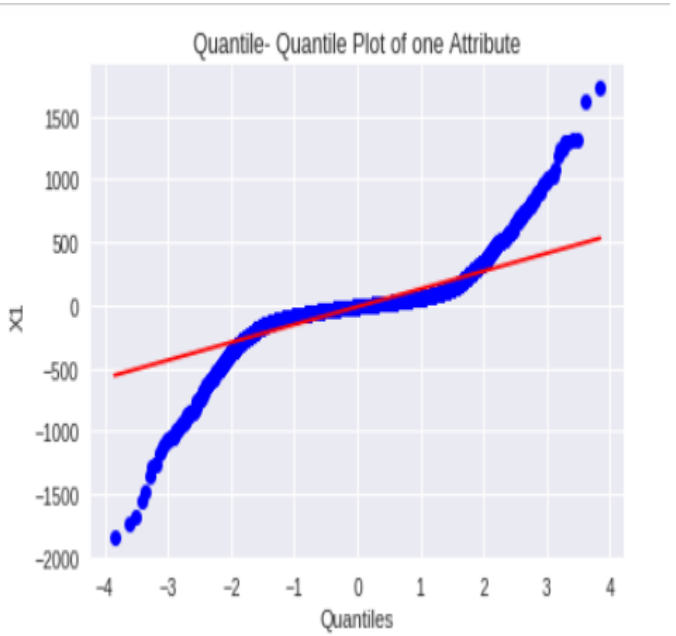


None of the variables are Normally distributed. The above is normal test . The High values (highlighted in red) and p value of 0 (highlighted in blue) indicate we can reject the null Hypothesis (The data is normally distributed)c

Inferential Statistics – Test for Normality -2




```
X1  -  19.0611932036
X2  -  18.2928199553
X3  -  18.4027045409
X4  -  18.3769456378
X5  -  17.737606333
X6  -  18.3765345279
X7  -  19.1466644963
X8  -  20.7219836039
X9  -  22.8007796461
X10  -  20.1583467644
X11  -  17.6858338111
X12  -  17.3692749103
X13  -  19.4135608448
X14  -  22.090855244
X15  -  22.9274403265
```



Plotting a Histogram for one of the attribute indicates that there is a high amount of kurtosis. This is highlighted by kurtosis value of 20 +.
The Q- Q plot on the far left also proves that the distribution of the variables are not normal.

Inferential Statistics – Correlation

```
X1  -  X2  - : 0.947728563382
X1  -  X3  - : 0.808191568857
X2  -  X3  - : 0.944622619007
X3  -  X4  - : 0.939521899995
X4  -  X5  - : 0.938635772132
X5  -  X6  - : 0.941266904074
X6  -  X7  - : 0.942731943499
X7  -  X8  - : 0.943499364474
X7  -  X9  - : 0.804300313349
X8  -  X9  - : 0.947479311107
X8  -  X10  - : 0.810992233444
X9  -  X10  - : 0.946728960851
X9  -  X11  - : 0.802872225449
X10  -  X11  - : 0.944024309668
X11  -  X12  - : 0.940959320036
X12  -  X13  - : 0.940121519252
```

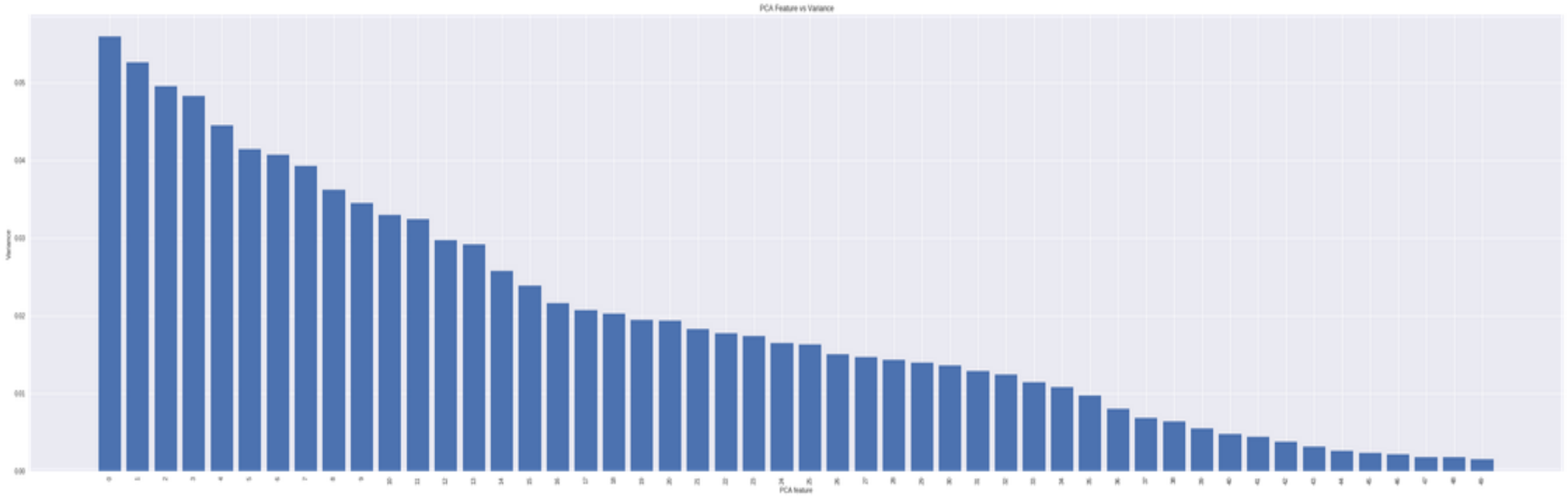
Variables are highly correlated with each other. We will have to handle this during the pre-processing step.

Pre processing the data - 1

Pre Processing steps include

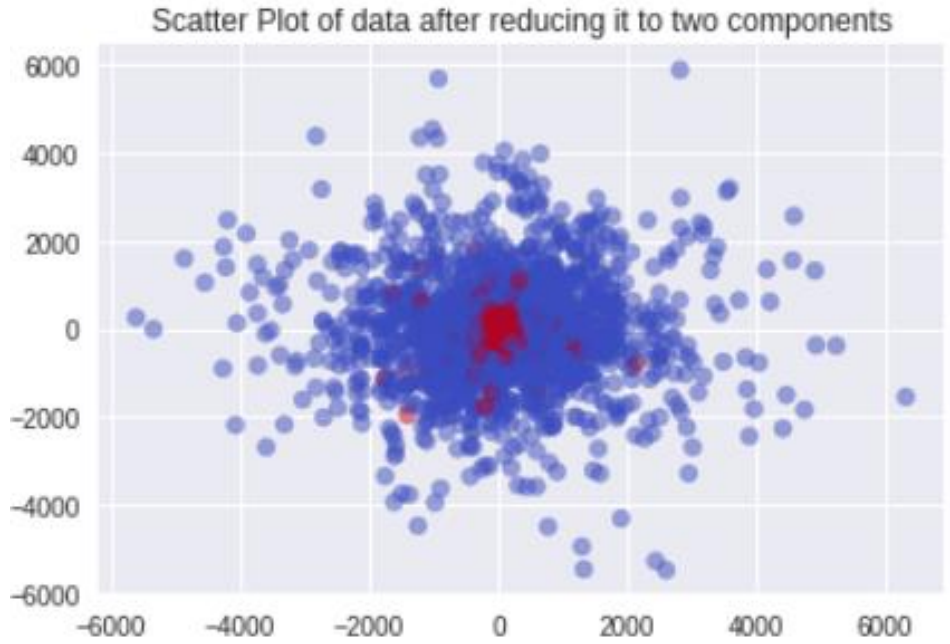
1) PCA to reduce the number of attributes and eliminate correlated variables
2) Remove class imbalance

Pre processing the data – PCA



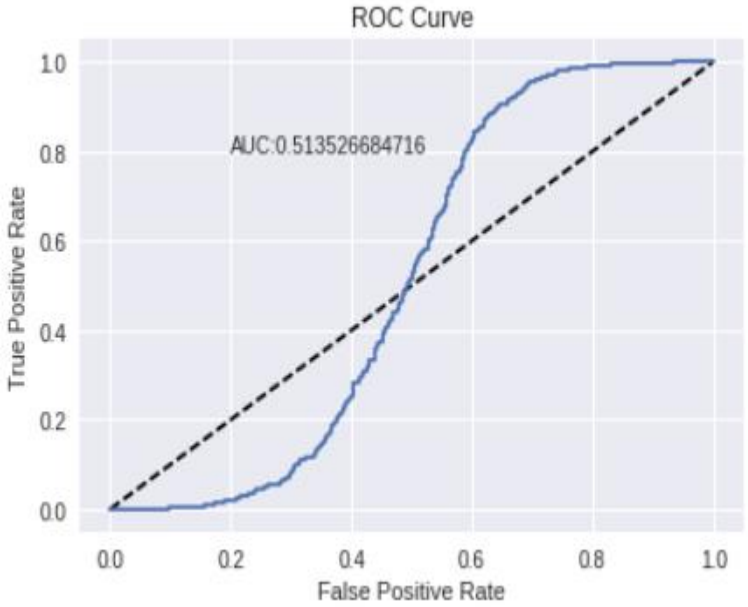
The Y axis indicates the Variance. X indicates the number of components. The value tapers down at about 80 components.

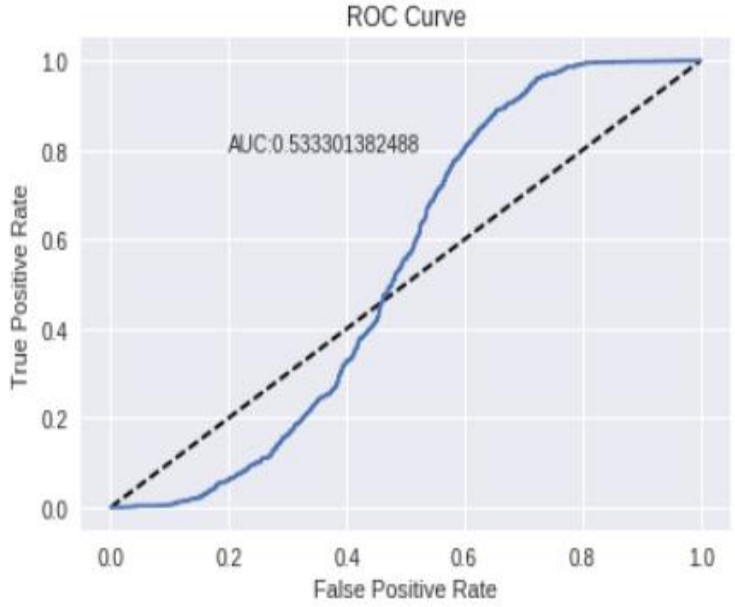
Plot of two components derived using PCA



Blue does indicate non epileptic patients. Red indicates epileptic patients.

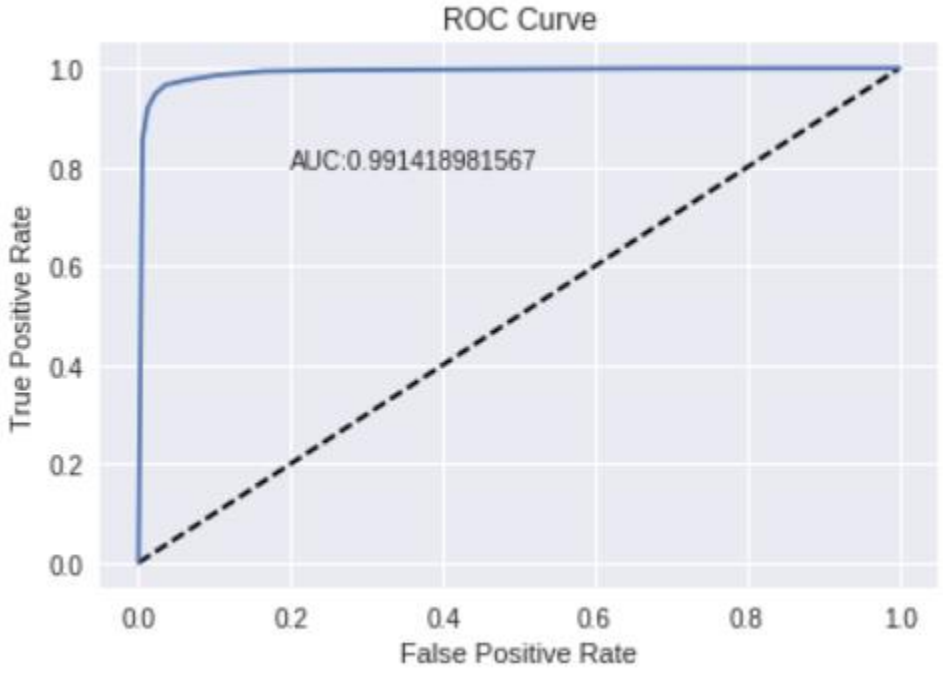
Classification Algorithms Logistic Regression

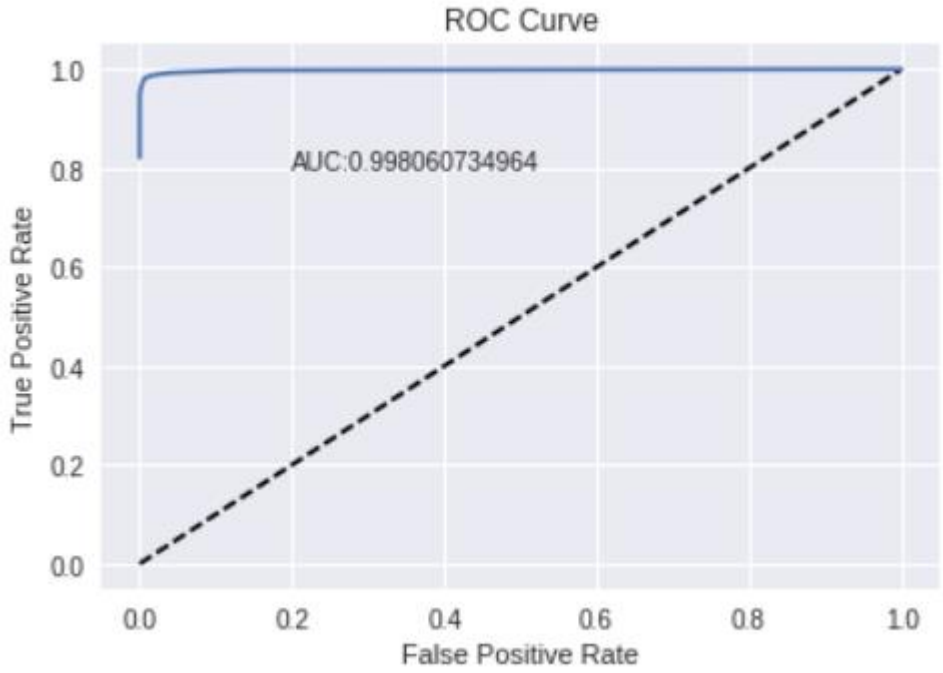




ROC Curve using Logistic regression without any preprocessing step. The AUC value is 0.51
ROC Curve using Logistic regression with preprocessing step. The AUC value is 0.53

Classification Algorithms Random Forest





ROC Curve using Random forest before and after Pre-Processing and after . There is an increase in the AUC value from 0991 to 0.998

Classification Algorithms Random Forest

Score: 0.967652173913

Confusion Matrix:

	0	1
0	553	37
1	56	2229

Classification Report:

	precision	recall	f1-score	support
1	0.91	0.94	0.92	590
2	0.98	0.98	0.98	2285
avg / total	0.97	0.97	0.97	2875

Score: 0.979565217391

Confusion Matrix:

	0	1
0	2267	8
1	86	2239

Classification Report:

	precision	recall	f1-score	support
1	0.96	1.00	0.98	2275
2	1.00	0.96	0.98	2325
avg / total	0.98	0.98	0.98	4600

Accuracy, Precision , Recall before and after pre-processing. There is a improvement in the values post processing.

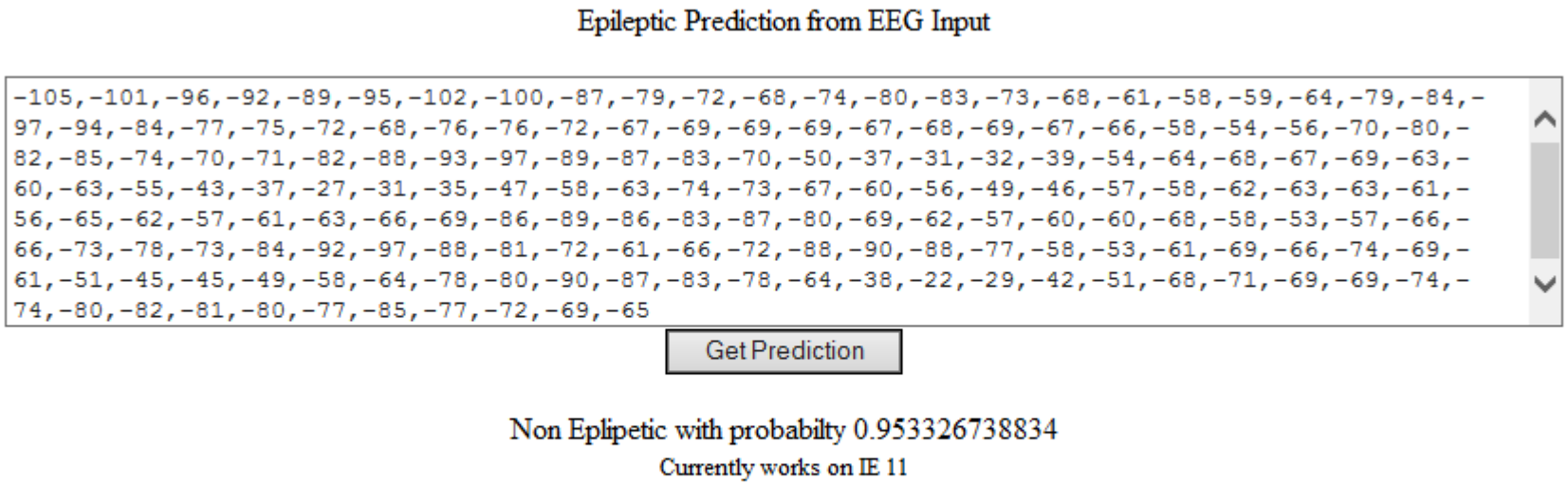
Conclusion

- Random forest model was a far better predictor than the Logistic regression classifier.
- The is due to the fact that there is no clear boundary separating the classes (Slide 14).
- Normalization of the data was not needed since the attributes were distributed in a similar manner (Slide 6).
- Reducing the number of attributes using PCA and Oversampling the classes which were uneven helped in increasing the accuracy metric.

Create a webservice

- The Model was exported using pickle.
- Using Python Flask module and gunicorn http webserver the model was exposed through a webservice.
- A web page to input the values was created.

Web interface to call the model



Credits

- Many thanks to my mentor Amir Ziai
- Thanks to UCI for hosting the Dataset
- Thanks to the authors of the Dataset (Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE (2001) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907)