1. EDA done found missing data.
2. Initial data set (17521, 2395)
3. Remove some columns using thumb rule of 70-75% available so that at least you can impute it.
4. Transform shape 1 (17521, 2134)
5. Impute the data
   * 1. Median – 72.8 693 feature non-zero
     2. Mean – 72.9 890 features non-zero
6. There were only 693 features which have some contribution to result.
7. Use LGB model giving 72.8% results &

There is class imbalance in the data

1. 15760
2. 1761

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

Need to short down these variables to low number using LDA or some decomposition approach, there was high correlation between few features too.

Why need feature reduction?

<https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e>

Random Forest Classifier

Median Impute 84.31 % (2134)

1. 20410
2. 32

Mean Impute 84.26% (2134)

1. 20406
2. 36

2165 features

Mean Impute 84.24% (2166)

0 20415

1 27

Median Impute 84.

0 20414

1 28

Median Impute Less feature 2109 84.35%

0 20411

1 31

Median Impute 2114

0 20409

1 33