Name : Pratiksha Agrawal

15111029

IITR

Approach:

1. Since there were a lot of categorical, therefore they label encoded.
2. Miasing values were looked for, there were none
3. PCA was used for dimensionality reduction
4. 4 ensembles with different seeding to reduce variance of LightGBM was used; so that it is fast and gives the most accurate result
5. **Dart ALGORITHM** IN LIGHTGBM was used, so that it uses dropout and doesn’t overfit. Dart is dropout in trees
6. Plot importance was plotted, which is a bit a skewed. It shows that better features cud have been made or features cud have been modified. The prediction ios depending a lot on top features, which is bad.

Things that cud have been done:

1. The correlation cud have been evaluated and redundant features removed
2. Stacked ensembling could have been done instead of simple ensemble
3. Since a lot features are categorical, the one hot encoding cud have been done.
4. Other methods like linear regression could have tried and ensemble with them could have been taken
5. Data could have been visualized better and then modelling should have been done