Dataset:

The quality of dataset was moderately good. Though, initially it was ambiguous as how to use the dataset for training, it became clear after some analysis. Though, even till the end I could not get the revenue score positive. Also, it was not clear to me how to utilize the monthly revenue in prediction

Errors in the dataset:

1) Duplicate Hospital_District Entries in Hospital Profile data.

2) Number of employees in some Hospital_District were more than 10 lakhs which is not possible

3) Some Hospital_District Combinations present in other files were not in Hospital Profile data

4) It was not clear as how to utilize the monthly revenue in prediction

Pre-Processing:

The negative entries were not present in the dataset and hence they had to be created.

- All the hospital district combinations in all 4 files were taken into account. There were nearly 30000 such combinations.
- Corresponding to these combinations, 15 instrument id data was merged to get nearly 450000 rows.
- Out of these 200000 rows that were in the solution, were removed leaving 250000 rows.
- If the combination of Hosp_Dist_Inst was there in either hospital revenue file or projected revenue file, it was marked as a positive, else it was marked negative.
- Several features were calculated. Nearly 50 of them. Some of these features were – mean revenue for a hospital, mean revenue for a district, mean revenue for an instrument, total number of instruments of each kind a hospital has leased, total number of instruments of each kind a district has leased, mean revenue for a hospital district combination, mean revenue for a hospital instrument combination, mean revenue for a district instrument combination.
- Missing values were imputed using median of the available values.
- Highly correlated features were removed. Normalization and Scaling was done wherever required.

Training:

Several training approaches were tried. The best of them were ensembled for a robust model.

- Initially, the dataset was converted into a matrix with rows as hospital district combination and columns as instrument. Binary Collaborative Filtering Techniques were used. This model gave a score of nearly 0.11 on the leader board.
- It was then approached as a binary classification problem. Several boosting and bagging algorithms were trained on the dataset. Neural Networks were also used. Extreme Gradient Boosting and Gradient Boosting Machine Performed well on the

validation set and the leader board. These models were then ensembled. This approach yielded a leader board score of 0.15.

- The third approach was to train individual models for each instrument. The same models i.e. gradient boosting machine and extreme gradient boosting were ensembled individually on each part of the dataset and the prediction was combined. This was the best approach and gave a leader board score of 0.175.

Problems Faced:

- The models gave an f-score of as high as 0.95 in the repeated cross validation that was performed on the dataset. However, the f-score on the leader board never went past 0.35.
- It was very difficult and challenging to choose one final solution, considering that the self-validation set I was using was not a very good indicator of the final score.

Overall, it was an amazing Experience. Had never worked on such kind of dataset before. This documentation is very incomplete but the amount of time left was very less. In case any clarifications are required from my side in regards to the source code or solution approach please mail me at siddhantaditya01@gmail.com.