Project Report

Steps and Insights:-

1. Initially the given data had 448 features and one output each data has correlation of maximum 0.31 with target and the target was highly skewed . We visualized through heatmap and line plot.
2. We needed to do dimensional reduction of features for better visualization and prediction so we firstly standardized the input for better decomposition and then decomposed it in such a way it covered atleast 95 percent of variance
3. Now the decomposed features obtained (pc1 pc2 pc3) had some outliers . PC1, PC2 were almost normal so used z score on them and pc3 being skewed we used IQR and aggregated them . Further we used DBscan and Isolation forest for outlier detection and finally aggregated their result to get 33 outliers all in all.
4. .We used scatter plot and box plot for outlier visualization in the features. To fix outliers we simply used capping , we used IQR for pc3 since it was skewed and percentile method pc1 and pc2 since they were normal.
5. The target variable was heavily skewed so we applied log transformation afterwards we capped those having value above 99 percentile and below 1

percentile and finally standardizing the target and decomposed input

6. Now we divided the data into train test 80 percent data in train data . Since previous data was not related to current data so ANN or boosting technique could be the ideal choice ,though we applied LSTM ,ANN and xgboost with ANN model outperforming all the models.
7. Since their were less number of data and with large number of outliers during prediction we faced flucating r2 score to fix this we capped the value applied log transformation and also applied data augmentation , used 3 models averaging their output to capture maximum variance managing to reach r2 score of 0.74(for ANN) . Yet the problem persists to some extents due to large number of outliers.
8. Finally once predicted we applied inverse transform and undo the log transform on predicted and actual output for visualization and better understanding
9. We saved model as well as scaler used on target, scaler used on input feature, pca components now these were loaded to give output for any one value . where we choose one input which will be standardized with saved scaler used on input ,decomposed using pca component used on input, predicted ,then output was inverse transformed and inverse log transformed to give actual output.This part of code has been used in app along with the model formed , input scaler, output scaler, pca component

along with few styling for better visualization. App code.ipynb can be used for ui visualization just copy the hyperlink on chrome.

Scope for improvement:-

1. We can use multiple models and combine their result to capture maximum variance
2. Apply data augmentation
3. Having more number datasets