ABSTRACT -

We have added a new column named Classified scores that includes data for classifying the score column ranking from 1 to 5 where 1 lies between [0-0.2), and 5 include[0.8-1.0].
We have used techniques like PCA, RFE, ANOVA, and selected the best features according to their result(we have not hardcoded their value). So, the number of features might vary a little based on data.

Binding the Dataset -

Firstly, we found the true recovery rate in the target table by the average of the Recovery rate of the lowest score and the recovery rate of the second highest score using OVER PARTITION BY and RANKING methods. We have used JOINS and UNIONS to join the data divided in the different datasets to get one dataset.

Preprocessing data -

Removing quasi-constant features -

Quasi-constant features are the features that are almost constant. These features have the same values for a very large subset of the outputs. Such features are not very useful for making predictions. We have set 0.01 as the value for the threshold parameter, which means that if the variance of the values in a column is less than 0.01, remove that column. In other words, remove feature columns where approximately 99% of the values are similar.

Checking Duplicate Features -

In data preprocessing and analysis, you will often need to figure out whether you have duplicate data and how to deal with them. So we created a function to check for the length of the duplicate values.

Removing Correlated Columns -

We have removed correlated columns/ features to increase the performance of the model.

Feature Selection -

Principal Component Analysis(PCA) -

We have used PCA to reduce the number of features used for fitting the model by projecting high dimensional data into a new lower dimensional representation of the data that describes as much of the variance in the data as possible with minimum reconstruction error.

Recursive Feature Elimination(RFE)-

We have used RFE to configure the selected features (columns) in a training dataset in predicting the target variable.

Anova Testing -

We have performed ANOVA testing to understand how each independent variable's mean is different from the others, and to understand their behavior and connection towards the dependent variables.

Scale the dataset -

We have applied the *StandardScaler* dataset directly to standardize the input variables.We have used the default configuration and scale values to subtract the mean to center them on 0.0 and divide by the standard deviation to give the standard deviation of 1.0

Predicting the model-

K-fold Split and XGB Booster-

So, during the splitting of data into training and testing, we need to keep in mind the overfitting of our model. Hence the cross-validation of the data plays a crucial role in determining this. The k-fold technique works by dividing the data into groups, k subsets and k-1 subsets and then determining the performance of the model on each fold. We have divided it into 15 splits and trained our XGBRegressor. This is a model that learns with its own mistakes(we used a learning curve of 0.07 in order to take care of overfitting) and uses ensemble modeling(in our case we have used 5000 models).