Approach

Initial Approach:

The preprocessing was initially aimed at trying to achieve good correlation scores between the target and numerical columns but the numeric columns showed little promise with respect to that and the basic linear regression model was peforming better than an equivalent basic multilayered ANN. Using a Random forest regressor gave me marginal improvement but the scores were still low. To change this, I decided to encode all the columns using a Target enoder. The encoding of the User_id column gave a direct correlation of 0.8+ with the target column and the model created after this lifted my score significantly to 0.38 on the public leaderboard. To further improve this, I used models like Huber Regressor and it helped to achieve a score around 0.41. I experimented using outlier removal, different encoders and scaling methods but it did not help. The only thing that seemed to help was removing the age column as it had a very low correlation show. The last step coupled with batch normalization of my ANN lifted the score to 0.42. The standard scaled solution was performing better after encoding and batch normalization. Finally with further hit and trial using a combination of optimizers and activation functions, a score of 0.43 was achieved and submitted as final solution.

The key to achieving this score was using right encoders and optimizers, along with batch normalization and logic based experimentation.

Final Workflow:

Data Preprocessing And EDA:

- The train and test datasets where loaded using Pandas library and descriptive analysis was performed on the train data. The descriptive analysis showed insights related to null values and statistical keys related to numerical columns of the train data.
- Outlier detection and data distribution check were performed using seaborn library. No outlier removal was performed eventually.
- All categorical columns except “Gender” were encoded using Target Encoder after comparing the correlation scores. The Gender column was encoded using one hot encoding method.

- The train and test data were scaled using standard scaler and columns like row_id and age were dropped before using train_test_split to split the data for model building.

Model Building:

- After comparing the ANN results with various ML Models, an ANN was finalised that utilised 8 Hidden Layers, a combination of Relu And linear activation functions, Adamax optimizer, L1 regularization and Batch Normalization.
- The Final Model Gave 0.505 R2 score on local train-test data and eventually gave 0.436 R2 score on Public leaderboard and 0.442 on Private leaderboard. This score was slightly better than the 0.41 score that was achieved using Huber Regressor Model.