Model Approach:

1. Looked at the dataset and noticed that there are numerical columns, categorical columns and timestamp columns. Also Noticed that there are NA (np.nan) values for some of the columns of the dataset.
2. Looked at the size of the data before and after dropping of the records that has NA (np.nan). The percentage of records that were dropped was more than 7 % which implies we need to impute the data. Although we can use various approaches for imputation. I chose mean imputation for numerical data, and most frequent occurrence for the categorical data, depending on the number of unique values of the columns I have chosen one hot encoding and hash encoding for categorical data.
3. Scaling of the data makes the model converge faster to start with I chose min-max scaler as the scaling technique.
4. Now that the data is preprocessed I have used a sklearn cross validator which has kfold validation to choose the best model form Linear Regressor, Neural Network, Decision Tree Regressor, Random Forest Regressor.
5. For the given data set Linear Regressor has the lowest cross validation error. (error is RMSE)
6. Now that the model is chosen I have trained the model with Train Test Approach and noticed that not always the validation error is more than the training error. Then I tried Train / Validation / Test Approach and found the same. One of the reasons for this to happen is the existence of skewness in the target variable. Although the skewness in the target variable can be handled by transforming the target variable (log transformation/ square / square root/ cube / cube root) depending on if the variable is left skewed / right skewed linear regression doesn’t assume/ require the data to be normal.
7. To avoid transformations, I have implemented a stratification model named fractional stratification which distributes the skew among train validation and test data sets
8. Once the best fit model (lowest validation error RMSE) is produced by fractional stratification technique regularization is applied to fine tune the model. On the best produced model residual plot is produced to prove the assumption of linear regression that residual errors follow normal distribution and predictions are made on the best model produced.

Improvements:

1. I have used mean imputation to start with. We can look at other imputation techniques like median / mode / similar value / KNN imputation which might improve the model predictions
2. I have used one hot / hash encoding to start with. We can look at other encoding techniques like binary / sum / polynomial / backward difference / helmert encoding or fine tune already used encoding to reduce the number of features
3. I have used min-max scalar to start with. We can look at other scaling techniques like

standard scalar/ robust scalar / normalizer which might improve the model predictions

1. I have used regularization to fine tune the model. Regularization doesn’t have much impact when there is noise in the data. We can look at Feature Importance and Feature selection algorithms such as Shapiro / Pearson / ExtraTressClassifier / Recursive Feature to figure out the feature importance’s and feature ranking to choose only important features that can improve the model predictions.