Hackathon Generic Approach:

1. Load Data (Both Train & Test Separately)
2. Understand the Data (check each of the following in both the train & test)
   1. Check Head & Tail
   2. Info, Describe
   3. Null Values
   4. Bad data like ‘$’ or ‘#’ in numerical column or any other unwanted character
3. Clean the data
   1. Treat Missing Values in both the train & test
   2. Remove bad data values in both the train & test
   3. Encode the object variables in both the train & test
   4. Feature Engineer (if needed)
   5. Scale/Normalise (if needed)
4. Make a simple model using any algorithm
   1. Fit the model on Train
   2. Predict on the Test
   3. Store the predicted values in an array
5. Submission (Approach 1)
   1. Import the sample submission file
   2. Check if your test data and sample submission file has same ID column sequence (if not, then sort them such that each ID’s individual predicted value is placed on corresponding ID in submission file)
   3. Replace the “target” column in Submission File with predicted array
   4. Export the file to a CSV
   5. Make sure it has the same headers as the sample submission file
   6. Upload on the platform and check your score
6. Submission (Approach 2)
   1. From your Original Test Set take the ID
   2. Create a new data frame with ID and corresponding predicted values
   3. Export that data frame to CSV
   4. The number of rows (including headers) & columns should match with the sample submission else the platform will not accept it
   5. Upload on the platform and check your score
7. Now go back to the step 3 (this is an iterative process)
   1. Check if Scaling approach change helps
   2. Check if feature engineering helps
   3. Try removing unnecessary variables (feature importances) & check if it helps
   4. Try grid search
   5. Try advanced model tuning techniques (like non-parametric ensemble methods)

Few Pointers to take care of:

1. Do not drop null values from Test Set
2. Whatever preprocessing step you perform on Training, it must also be performed on the Test set
3. Try using “n\_jobs = -1” while fitting the model for parallel processing to decrease the time taken for fitting the model. This can take up all your computational resources and your PC might start working slow for any other task you perform on the PC.
4. Recommended to make copies of datasets at every checkpoint so you don’t have to restart from first. You can directly read the latest checkpoint dataset and start from there.