Team Name:

Nilkanth

Tools Used:

- R-studio
- H2O

Installation:

Follow below links for installation guides.

- R-Studio: https://medium.com/@GalarnykMichael/install-r-and-rstudio-on-windows-5f503f708027
- H2O : http://h2o-release.s3.amazonaws.com/h2o/master/1735/docs-website/Ruser/Rinstall.html

Run H20 from R-Studio:

1. Execute following commands in R-Studio to start H2O, here "7g" attribute is used to tell H2O server to use 7GB of available RAM. Itt can be set according to system configurations.
    - `library(h2o)`
    - `h2o.init(max_mem_size = "7g")`
2. Once H2O server is started go to browser and enter default path to access H2O server.

    http://localhost:54321/

3. Open the .flow file to re-execute scenario. Make changes in path for Train and Test data files.
4. Run each cell using un button similar to Jupyter Notebook.

Approach:

1. Ignore Serial number column as it's not going to contribute towards prediction.
2. Convert "date" as categorical variable using "convert to enum" option once data frame is parsed.
3. Select option to Balance training data class counts via over/under-sampling (for imbalanced data).
4. Set parameters like,
    - ntrees = 63300 (Number of trees.)
    - max_depth = 10 (Maximum tree depth.)
    - min_rows = 10 (Fewest allowed (weighted) observations in a leaf.)
    - nbins = 10 (For numerical columns (real/int), build a histogram of (at least) this many bins, then split at the best point)
    - learn_rate = 0.01 (Learning rate from 0.0 to 1.0)
5. Build Gradient Boosting Model (GBM) in H2O
6. Predict result using Test data and download result in csv.