2019 NBA Hackathon Application – Business Prompt

Task: We have provided a sample dataset of 1,000 (real) Instagram posts by @nba since October 1, 2017 (211 individual photos, 109 photo albums, and 680 videos) for which your task is to predict total "engagements." Note that these "engagements" are ***not*** real – i.e., we've artificially generated the `Engagements` column such that there's no (intentional) correlation with the real-life engagement totals belonging to posts by @nba.

To assist with your model, we have provided an identical dataset of 7,766 (real) Instagram posts by @nba in the same timeframe (1,595 individual photos, 713 photo albums, and 5,458 videos). That is, with the `Engagements` column filled in. **Using these inputs, we would like you to predict `Engagements` for each of the 1,000 posts in the holdout set.**

- `training_set.csv`
  - This dataset includes data related to 7,766 (real) Instagram posts by @nba since 10/1/2017. From left to right, the columns are:
    - `Engagements`: Artificially generated "engagements"
    - `Followers at Posting`: # of followers at the time of posting
    - `Created`: Datetime stamp of post (Eastern time)
    - `Type`: Classification of post as Individual Photo, Photo Album, or Video
    - `Description`: @nba's post caption / description
- `holdout_set.csv`
  - This is a random holdout set of 1,000 (real) Instagram posts by @nba since 10/1/2017
  - All columns are the same as those described above, but *Engagements* has been removed

You will be graded on **Mean Absolute Percentage Error (MAPE)** on "engagements". We selected this metric due to scaling in the "engagements" response variable. This metric is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right|,$$

where n = *1,000* is the total observations in the holdout set, and A_i and P_i are the ith actual and predicted "engagements." Some tips to help you in your modeling:

- Consider all factors that may drive "engagements" in the real world. While the true relationships between these features and the true response variable will be different, a large subset of such features likely have a (significant) relationship with our artificially generated "engagements."
- Consider temporal effects such as the day of week or time of day for a given post.
- Consider how "engagements" are distributed across different post Types.

Please submit a file named `holdout_set_[Individual_or_Team_Name].csv` with the `Engagements` column filled in with your response variable. Please also return all code or relevant working files *separately* (i.e., not zipped up with your `.csv`). Thank you!