Assignment Overview

**Forecasting Model Week 1 Assignment**

In week 1 we introduced new types of regression models in handling the categorical outcome variables (e.g., win or lose). By using the subsets of data from the NHL database, we show that the logistic regression model fits the data better than the LPM model when it comes to the binary dependent variable (i.e., Win vs. Lose). Later this week, we lay down basics for the forecasting model by using NHL dataset as well. In essence, it is important to use the predictor variable available ahead of the time so that the prediction becomes meaningful in the practical sense.

The assignment for this week is to generate the similar results using NBA dataset. Specifically, you will be required to accomplish the following three analytic tasks.

1. Fitting the logistic regression in handling the categorical outcome variables
2. Evaluating the performance of different models
3. Building a forecasting model

To complete the assignment, you will need to repeat the same kinds of steps we took to generate the results.

# Assignment Instructions - Part 1

These are the steps you need to take to complete the assignment for Week 1. **Step 1. Building the Dataset – NBA Data set**

1. Load the “NBA\_Game2” file, extract 2017 regular season records, named “NBA \_2017”. The regular season games can be defined with a Season ID equal to 22017 and a Game ID less than 1000000000 (these filters will exclude preseason games).
2. Subset the NBA data including 2017 regular season records.
3. Explore the dataset using some useful commands.
4. Manipulate the dataset and define the variables. This proceeds in several steps:

* Filter 2017 regular season records and select columns to be used for analysis.
* Manipulate the MATCHUP column to obtain columns for each team.
* Create the home dummy variable.

**Step 2. Running Regression – Part I**

1. Fit LPM of win (W vs. L) on Pythagorean Winning Percentage.
2. Plot the LPM to articulate problems.
3. Fit a logistic regression on win on Pythagorean Winning Percentage.
4. Print the results including coefficients, p-values, and dependent variable.
5. Obtain the fitted results using the logistic regression model.
6. Calculate the success rate.
7. Fit multiple logistic regression by incorporating the home team advantage.
8. Obtain the fitted results using the logistic regression model.
9. Obtain the success rate.

**Step 3. Basics of Forecasting (NBA Example) – Part II**

1. Load the “NBA\_Game2” file, extract 2017 regular season records, named “NBA17”.
2. Explore the data set.
3. Subset the data for forecasting.
4. Manipulate the dataset and define the newly added variables.
5. Aggregate the resulting data frame into the team level data.
6. Manipulate the team level data and define the variables for forecasting.
7. Plot the variables used for forecasting.
8. Obtain the correlation between Pythagorean winning % for the 1st half of the regular season (less than or equal to game ID 21700615 in the pre period and all game ids after this in the post period).
9. Manipulate the 2nd half of the regular season data to be used as DV for forecasting.
10. Fit two regressions using 1) Pythagorean winning % and 2) 1st half winning % as I.V respectively.
11. Compare the performance between two models

**Beware**: Even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.