**What is Logarithmic Loss (Log Loss)?**

The accuracy of each group’s prediction will be measured by Log-Loss in the **FBAS March Data Crunch** **Madness Competition**. An understanding of how this metric is calculated and how it should be interpreted is important for competitors to optimize their chances of winning.

**Log-Loss** measures the accuracy of a classifier.The loss function quantifies the amount by which the prediction deviates from the actual value. In the case of the March Madness competition, Log-Loss will put a number to how well - on average - a group is able to **predict the probability** of Team A winning a game over Team B. Thus, Log-Loss is useful when your research objective is not only to say if an object belongs to class A or class B, but to also provide the probability of each outcome. **The ultimate goal is to *minimize* Log-Loss**: a perfect classifier would have a Log-Loss of exactly **zero**.

If there are only 2 classes (which is true in the case of March Madness - one of two teams can which each game), Log Loss is defined as:

Log Loss = - frac{1}{N} sum_{i=1}^N [y_{i} log , p_{i} + (1 - y_{i}) log , (1 - p_{i})].

Where:

* **N** is the number of games played
* **Pi** is a group’s predicted probability of Team 1 beating Team 2 (in a specific game)
* **Yi** is **1** if *Team 1* wins and **0** if *Team 2* wins

Again, the smaller the Log-Loss the better. Games which are not played are ignored in the scoring. The use of the logarithm provides extreme punishments for being both **confident** and **wrong**. In the worst possible case, a prediction that something is true when it is actually false will add infinite to your error score. In order to prevent this, ***predictions are bounded away from the extremes by a small value.***

A group’s Log-Loss scores for each game are averaged together to get a final Log-Loss.

**Calculating Log Loss using example seed dataset:**

If you have gone through the SPSS modeler tutorial and developed a model for predicting Team 1’s outcome given the two teams’ seeds, you should be able to calculate your model’s log-loss for each game and for your overall model. For example, for game ID 2011-1433-1199, the model assigned a 19.9% probability that team 1 would win the game. In actuality, team 1 lost. Given this outcome and the log-loss equation given above, the log-loss for this particular game will be:

-1\*(0\*LOG(.199) + (1-0)\*LOG(1-.199) = .**0964**

To give you a better understanding of how log-loss rewards a model for getting as close to accuracy as possible, suppose the model actually assigned a 1% probability that team 1 would win the game. Since team 1 lost, this is much closer to the actual probability of team 1 winning (0%). Now, the log-loss equation becomes:

-1\*(0\*LOG(.01) + (1-0)\*LOG(1-.01) = .**0044**

To calculate your overall model’s log-loss, simply average all game log-loss scores together.

