**Medical School Admissions Forecasting**

**Competition Rules**

*March 21, 2018*

Daniel Egger

**Link for the Competition**

[https://www.kaggle.com/t/1972040b48474485a9cf9efbe8211f03](https://www.kaggle.com/t/1972040b48474485a9cf9efbe8211f03" \t "_blank)

Click on the link. Go to the rules tab and select “Understand and Accept” to join the competition.

You must rename your team name (a team of one) for the competition under your actual **first and last name.**

**Training and Test Sets**

Students will use their Kaggle IDs to access the online competition and download a Training set with the following variables:

Unique Applicant ID  
Undergraduate Major  
GPA in Biology, Chemistry, Physics and Mathematics (required pre-med courses)  
GPA for All Other (AO) courses  
Combined GPA – all GPAs are standardized to range from [0, 4.00]   
VR – Verbal Reasoning portion of the MCAT Exam [max 15]  
Physics portion of the MCAT [max 15]  
Written Section of the MCAT [J lowest, T highest]   
L Biology portion of the MCAT [max 15]  
M Combined MCAT score [max 45, and repeat of Written Section letter score]

Each Applicant also has an outcome:

0 (applied and rejected),

1 (applied and accepted) or

3 (did not apply)

The Test Set has the same variables as the Training set but no outcomes.

Students will forecast, for each student the probability of acceptance in the form of a value **b**, where **b** > 0 and < 1.

*Students submitting one or more probability forecast less than or equal to zero, or greater than one, can be disqualified.*

A sample of the format used for submitting forecasts is also provided on Kaggle.

Note that the **base rate** of acceptance varies for each of the Training, Test and Competition data sets. This varying base rate **r** is provided for each set and may be taken into account when calculating individual forecasts.

**Schedule for in-class competition**

Students will be given the Training Set with outcomes, and the Test set without outcomes, on March 22.

Students can submit up to three (3) attempts per day, and their scores will be shown on the leader board.   
  
Submissions for the test set will have a deadline of Wednesday, March 28 at midnight.

On Thursday, March 29, students will be given a third, Competition dataset, as well as the correct answers for the Test Set.

Note that the Competition set has as many students as the Training and Test sets combined. Students will have the whole class time to submit one solution for the Competition Set. It will be a separate Kaggle competition on the site, for which the link will be shared during class.

**Scoring method**

Kaggle scoring uses the “log loss” method.

The “log loss” score **L** assigned to each outcome is as follows, for a set of **n** applicants with base rate **r** = **m/n:**

1 – for acceptance, (-1/n )\*log(b)

0 – for rejection (-1/n)\*log (1-b)

3 – no application – result ignored – no impact on score

Kaggle will generate this score automatically.

Note that for a set with m acceptances and (n-m) rejections, [assuming for the moment the same value **b** was used in all cases] the log loss **L** would be

**L** = -[(m/n)\*log(b) + ((n-m)/n)\*log(1-b)]

The goal is to have the smallest possible score.

Since L must be > = 0, a perfect result would be L = 0.

Note that there are two benchmark scores on the leader board: one with every entry equal to .5, and one with every entry equal to the base rate.

**Relationship of Log Loss L to Information Gain I(X;Y)**

Scoring by log loss will give **exactly the same** rank-ordering of competitors as scoring by information gain.

The information gain *over betting the base rate* ***r***, averaged over n events, would of course be:

1 – for acceptance, (1/n )\*log[(b)/**r**]

0 – for rejection (1/n)\*log [(1-b)/(**1-r**)]

3 – no application – result ignored

For a set with m acceptances and (n-m) rejections, [assuming for the moment the same value b was used in all cases] the information gain I(X;Y) would be:

I(X:Y)

= (m/n )((log b) – (log r)) + ((n-m)/n) (log(1-b) – log(1-r))

= [(m/n)\*log(b) + ((n-m)/n)\*log(1-b)] - [(m/n)\*log(r) + (n-m)\*log(1-r)]

but since **m, n** and **r** are constants for any given dataset,

-[(m/n)\*log(r) + (n-m)\*log(1-r)] is a constant **C** > = 0.

and

[(m/n)\*log(b) + ((n-m)/n)\*log(1-b)] = **-L.**

Therefore,

**C -L** =I(X:Y).

**END**