**Saleem Huda**

**Some notes on making probabilities**

Here I have some thoughts on :

1. scoring of prob models, rankings
2. converting scores to probabilities
3. **SCORING PROB MODELS WITH ORDERING BEFORE USING PAYOUTS**

**Background**

We need a way to score probability models that is intuitive.  I looked at log-likelihood and weighted rank correlations kendall-tau , spearman , etc.  but its not clear to an outsider how well a probability model is doing in the case of evaluating ordered outcomes.  I came up with 2 metrics that should help.

***A) scoreProbModelPerm*** assigns a unique prime  number to the runners that are in the correct position in the top 4 for a given race.  The score for the race is the product of the prime numbers which makes it easy to tell which runners were scored correctly: ranges from**1 to 210**

B) ***scoreProbModelComb***siimply counts the number of runners in the top 4 of the probModel that were also in the top 4 of the finishers in the race  **range  0 to 4**

Example from 'PHA\_20100608\_6'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **runnerId** | **pmSpeedAC** | **rankSpeedAC** | **officialFinishPosition** | **sPos** |
| PHA\_20100608\_6\_1 | 0.003546 | 7 | 6 | 1 |
| PHA\_20100608\_6\_2 | 0.066799 | 4 | 4 | 2 |
| PHA\_20100608\_6\_3 | 0.050849 | 6 | 2 | 1 |
| PHA\_20100608\_6\_4 | 0.066799 | 4 | 8 | 1 |
| PHA\_20100608\_6\_5 | 0.220293 | 2 | 3 | 1 |
| PHA\_20100608\_6\_6 | 0.478885 | 1 | 1 | 7 |
| PHA\_20100608\_6\_7 | 0.111256 | 3 | 5 | 1 |
| PHA\_20100608\_6\_8 | 0.001574 | 8 | 7 | 1 |
|  |  |  | **scorePerm** | **14** |
|  |  |  | scoreComb | 2 |

***Define:***

***sPos\_1 = 7****,if****probModelRank[runnerFinishPos\_1] ==1****else, 1*

***sPos\_2 = 5****, if****probModelRank[runnerFinishPos\_2] ==2****else, 1*

***sPos\_3 = 3,****if****probModelRank[runnerFinishPos\_3] == 3****else 1*

***sPos\_4 = 2,****if****probModrelRank[runnerFinishPos\_4] == 4****else 1*

*scoreProbModelPerm = sPos\_1\*sPos\_2\*sPos\_3\*sPos\_4*

If the probModel gets all runners in top4 (superfecta) in sequence correctly the score is**7 \* 5 \* 3 \* 2 = 210**

**First, Second, Third = 7\*5\*3 = 105**

**First and Second : 7 \* 5 = 35**

**Third and Fourth : 3 \* 2 = 6**

**First and Fourth : 7 \* 2 = 14**

1. **CONVERTING SCORES INTO PROB MODELS**

STEP 1 - Convert Speed Figures Into Probabilities

In order to convert speed figures to odds without fitting a model to a data history, we will make 2 assumptions which may be oversimplifying but at least gets us a sensible starting point without overfitting:

1) The difference in speedRunner\_i and speedRunner\_j is proportional to the lengths by which runner\_i will outperform runner\_j

2) the expected speed figure for each runner is normally distributed around the speed figure we are using (i.e. Maury Speed, projected HDW Speed, mean last10 , etc..)  with variance equal to the variance of the speed figures in the race

**DD NOTE - I don't see where assumption #1 plays in the calculations below at all.  I have a really hard time buying that #2 is a reasonable assumption.  It just doesn't seem to make sense.**

**SH NOTE - Assumption #1 does not get involved in the calculations , but one needs to accept it in order for speed ratings to make sense in anything other than ordinal space.  If we have numbers that range from 0 - 100 and don't think that a 100 speed figure racing against a 95 speed figure is signifcantly different then a 100 racing against a 50 speed figure then why bother with speed figures that aren't just ordinal integers?  Assumption #2: is the simplest starting probability model that was tractable.  Ideally you would have a better estimate of the variance of each runner.  But intuitively it makes sense that if the variance of the projected speed figures for the horses in a race is very high relative to the top speed horses then we should assume that we have a race where the higher speed figures should be discounted as a horse that is truly running a 110 should never be racing against a 35 .**

We can think of this as each runner rolling a dice with the same number of sides but the numbers on each runners die having a different range.

We want to solve for the probability that runner\_i with his dice rolls a higher number than any of the other runners in the race given that they are all using dice with the same number of sides.

We will look at:

|  |
| --- |
| MNR\_20110405\_7 |

For explanation I will divide the speed figures by 10 and round so, looking at the  **speedFigureInteger**column we will say MNR\_20110405\_7\_1 is rolling a dice that has an expected value of 8 but could be as low as 5 or as high as 11

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Runner | id | speedFigures | stdevForRace | 1Stdv Range | speedFigureInteger | stdevIntegerForRace | range |
| MNR\_20110405\_7\_1 | 1 | 80.375 | 29.430 | (50,110) | 8 | 3 | (5,11) |
| MNR\_20110405\_7\_2 | 2 | 72.375 | 29.430 | (42,102) | 7 | 3 | (4,10) |
| MNR\_20110405\_7\_3 | 3 | 95.75 | 29.430 | (65,125) | 10 | 3 | (7,13) |
| MNR\_20110405\_7\_4 | 4 | 41 | 29.430 | (11,70) | 4 | 3 | (1,7) |
| MNR\_20110405\_7\_5 | 5 | 72 | 29.430 | (42,102) | 7 | 3 | (4,10) |
| MNR\_20110405\_7\_6 | 6 | 89.64 | 29.430 | (60,120) | 9 | 3 | (6,12) |
| MNR\_20110405\_7\_7 | 7 | 80 | 29.430 | (50,110) | 8 | 3 | (5,11) |
| MNR\_20110405\_7\_8 | 8 | 73 | 29.430 | (43,103) | 7 | 3 | (4,10) |

This is where the model is different than using a simple ranking of the runners by speed.   We are generating a probability based on taking into account the speeds of all other horses and the amount of variation in the race.

In order to get the probabilities we need to numerically solve the following to get the***probability that a horse outperforms all other horses conditional on the horses in the race***

**Fi(speed\_i)** is the cumulative Density Function evaluated at**x=speed\_i** where the distribution is**Normal(mean=speed\_i,stdevSpeed\_race)**

**fi(x)** is the *pdf evaluated at x=speed\_i* for the above distribution

**F(x) and f(x)** are evaluated numerically using**ScoreToProbViaIntegral()**

In this case we get the following probabilities.  We can see that runner 3,  **MNR\_20110405\_7\_3**, with a 95.75 speed figure is roughly 28% prob of winning .  So we have been able to establish probabilities of winning in an objective way.

| **Runner** | **SpeedFigure** | **NormProbMaury** |
| --- | --- | --- |
| MNR\_20110405\_7\_1 | 80.375 | 0.129071572 |
| MNR\_20110405\_7\_2 | 72.375 | 0.081925777 |
| **MNR\_20110405\_7\_3** | **95.75** | **0.279410271** |
| MNR\_20110405\_7\_4 | 41 | 0.009159129 |
| MNR\_20110405\_7\_5 | 72 | 0.080122168 |
| MNR\_20110405\_7\_6 | 89.64 | 0.208681824 |
| MNR\_20110405\_7\_7 | 80 | 0.126457645 |
| MNR\_20110405\_7\_8 | 73 | 0.085005972 |