PEDIGREE ANALYSIS by Byron Rogers

SHORTLISTING
yearling sales with machine learning

Watching the purchasing process that buyers undertake when they are at yearling sales is as much an observation of human prejudice and paranoia as it is of true risk/loss aversion. A lot of time and effort at a yearling sale is spent by yearling buyers discussing the qualitative merit of one yearling, comparing one yearling to another or one yearling by one sire compared to another, with little quantitative basis for such comparison to be undertaken in the first place.

Think about it for a minute, it is rather bizarre, that for all the claims made by vendors that x-rays, vetting, cardiovascular and genetic testing of yearlings are ways of 'failing' their horses, the reality is that the most subjective assessment, and the one that has the greatest impact, has occurred well before any of that takes place. Breeders will let almost any individual, whether they are qualified to do so or not, look at a yearling, and make a pass/fail assessment on the horse, with a 'failure' rate significantly much higher than that of any x-ray, clinical vet or genetic test. This person that has 'failed' the horse, talks to another person , and they talk to another…..

The current subjective nature of yearling sale purchasing is, in part, a byproduct of the way the yearling sales series are set up in the first place. Because of the non-sequential nature (in terms of quality) of the sales series, many of the buying patterns at yearling sales are driven by a "what have you done for me lately" mentality, with the stallion that has most recently produce stakes winners more popular than those that have produced stakes winners in the past, but have been cooler recently. For example, in January in Australia the Widden stallion Sebring was seen by many as being on the edge of a "failure", and yet by Easter he was the hottest young sire on the planet with his yearlings selling accordingly.

This is in some ways, a derivative of loss aversion, where yearling buyers or their representatives much prefer to be seen to be avoiding losses by selecting yearlings by the 'hot' sire, than finding gains by selecting the best yearling by a less popular sire. You can't be seen to be making a 'mistake', even though buying a yearling by the 'hot' sire for what it will cost might actually be one.

Yearling buyers also often have to make a comparative assessment at a yearling sale when armed with incomplete information. Is the 'best' horse at a second tier sale, one that might bring $200,000, as good as the 30th best horse at a premier sale which will bring a similar price? It is an on-going challenge for the yearling buyer to make this assessment, as for each sales season, the race quality of yearling sold varies considerably at each sale and the yearling buyer is often not in a position to know what quality of yearling is coming up at subsequent sales.

To a marked extent this is a result of a selection process undertaken by the vendors. They are naturally tending to take their "best" horses to Easter, Magic Millions and Karaka, as they know these venues are where the strongest buying benches are to be found. One might argue that due to the high influence of trainer effect on outcomes, it is a self-fulfilling prophecy, the "best" horses found at these sales get the best trainers/vets/jockeys.

This, however, still doesn't result in a solution to the challenge that is faced annually by yearling buyers. How do you consistently find the elite racehorse and avoid overestimating the value of the "best" yearling at a lesser sale and underestimating the value of a solid yearling surrounded by horses of similar quality at a strong sale? Is it possible to develop a consistent short-listing technique that at each sale would help identify the pool of horses most likely to produce stakes winners?

One possible answer comes in the form of Machine Learning. Machine learning is a type of artificial intelligence that provides computers with the ability to learn from data with known outcomes, without being explicitly programmed. The most popular type of machine learning algorithms is the use of Ensemble Methods. The phrase "Ensemble Methods" generally refers to building a large number of somewhat independent predictive models and then combining them by voting or averaging to yield a very high performance model – a form of crowd sourcing if you like.

There are a couple of options to consider with machine learning, which can be demonstrated using a data-set we have chosen to learn from. The sale we selected was the 2010 Keeneland September Yearling Sale, which can be downloaded at www.Keeneland.com. From that list of yearlings, the group/graded stakes winners were extracted (Keeneland have a list of past graduates on the site also). As of today's date these horses are five-year-olds so one can be fairly confident that if they are going to become a graded stakes winner, or at least a graded stakes winner that one would like to own, they would have achieved that goal by now. The process resulted in a list of 78 horses that are graded/group stakes winners. To that list of horses, another 214 horses from the catalogue were added – the first 72 horses in the catalogue by hip/lot number that were not stakes winners, 71 horses in the middle of the catalogue, and the last 71 horses in the catalogue. This is a random sample of non stakes winners used to compare to the graded stakes winning population. From there, for each of the yearlings 26 variables for analysis were created:

- Days – this is how many days from the 1st of January 2009 to the date of birth of the horse. This gives us a comparative date of birth parameter with horses born earlier in the year generating lower numbers.
- Generational Interval – this is a calculation of the average age between the sire and dam and the yearling. It is the sire and dam's combined age at conception divided by two.
- Sire Age at Conception – the age of the stallion in years when the yearling was conceived
- Crop Number – what crop number of the sire that the yearling came from
- Log of Conception Fee– this the natural log of the Sire's Service fee at the time of conception. We used a natural log, not the service fee itself as this number is handled more consistently in machine learning.
- Log of Sale Year Fee – this is the natural log of the Sire's Service Fee at the time the yearling is sold.
- Foals of Racing Age – How many foals of racing age the Sire had at the time the yearling was sold.

We have evidence to suggest that once you get past about 1200 foals in a top class sire, and about 600 foals in a lower level sire, your chance of becoming a group/graded winner is significantly diminished.

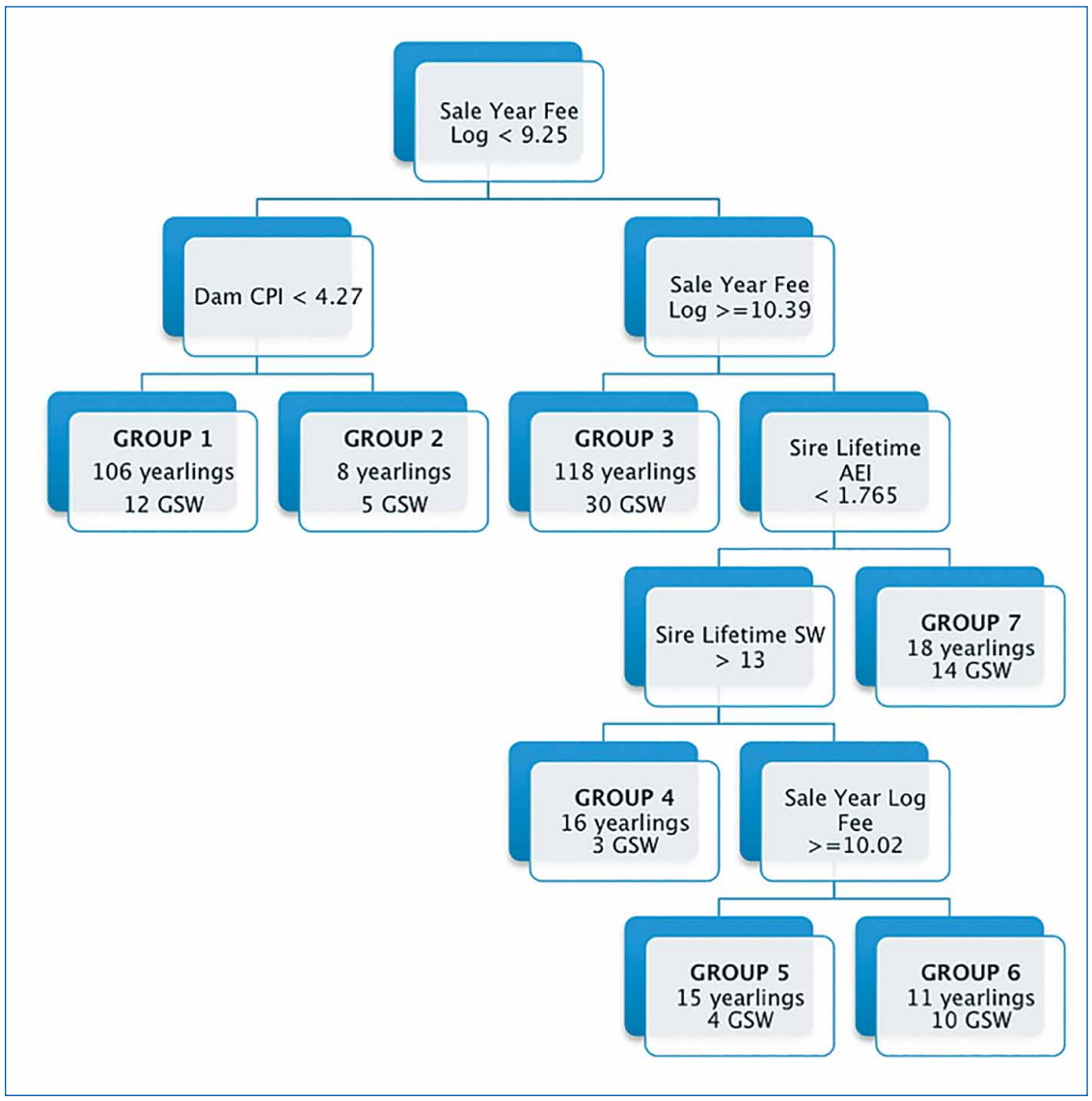
- Sire Stakes Winners – How many lifetime Stakes winners the Sire had at the time the yearling was sold.
- Sire Lifetime AEI – The Lifetime Average Earnings Index (AEI) of the Sire at the time the yearling was sold.
- Dam Age at Conception – the age of the dam in years when the yearling was conceived.
- Foal Number – The foal rank of the yearling out of the mare. Was it the first foal, etc.
- Dam Raced? – Was the dam of the yearling raced (yes/no)
- Dam Winner? – Was the dam of the yearling a winner (yes/no)
- Dam Wins – How many wins did the dam of the yearling have.
- Dosage Index – the Dosage Index of the yearling.
- Center of Distribution – the Center of Distribution of the yearling.
- Dam CPI – the Class Performance Index of the dam. The CPI is a measure generated by the Jockey Club Information systems and is a guide to the racing class of the mare.
- Dam SW? – Is the dam a Stakes Winner (Yes/No).
- Dam GSW? – Is the dam a Graded Stakes Winner (Yes/No).
- Dam G1SW? – Is the dam a Grade One winner (Yes/No).
- Dam Foals to Race – How many foals the dam had to race when the yearling was sold.
- Dam Winners – How many winners the dam had when the yearling was sold.
- Dam # of SW – how many stakes winners the dam had when the yearling was sold.
- Dam # of GSW – how many graded stakes winners the dam had when the yearling was sold.
- Granddam SW? – is the Granddam of the yearling a stakes winner?
- SW Relations to Dam – how many stakes winners are full or half relations to the dam.

The target variable for the model was GSW with a "1" for those yearlings that were graded stakes winners and a "0" for those that were not. It must be made clear that we are undertaking "supervised learning" in that we know the outcomes of the data – who is a stakes winner and who is not. It may be just as informative to undertake "unsupervised" learning, and see if the algorithms can effectively cluster the yearlings into similar groups that have disproportionate numbers of graded/group stakes winners, but for the purpose of this exercise we will undergo supervised learning.

It must also be said that the sample group, while random, is not a perfect group for analysis, the ideal being to analyze multiple sales over a number of years. The purpose of this study, however, is to gain clues as to what variables are important and can then be used as a starting point to discriminate in a larger study.

There are a couple of ways that a machine learning classification technique can be utilized with data. For those that really want to follow along, I have used the statistical computing programming language R in RStudio. For most of you reading along this will mean nothing, but suffice to say it is a program that allows you to use various methods to manage the data. With R there are a couple of packages that you can load that are the basis of machine learning and binary classification which is effectively what we are doing here – is the horse a graded stakes winner (1) or not (0).

Recursive partitioning is a fundamental tool in data mining. It helps us explore the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical outcome. In R, the package rpart allows us to implement the recursive partitioning function that results in what is commonly known as a decision tree. This decision tree results in the question being asked of the population if it satisfied a condition that was generated from the data itself, and multiple questions being asked until it forms a group where discrimination can no longer be made. The downside to using rpart is that it can overfit the data. To avoid this we pruned the tree size to one that minimizes the cross-validated error. The resulting tree from the data is below.



The decision tree has created 7 groups of varying size. The fact that there are two groups (1 & 3) with over 100 each and the other 5 having less than 20 in them shows that the data can create some distinctions between the yearlings, but that further data is needed to truly validate this. Three Groups, while small, have a disproportionate number of graded stakes winners in them.

Group 2 – these are the group of yearlings that are sired by stallions whose service fee at the time the yearling was sold was less than $15,000, but whose dam's were high class performers (CPI greater than 4.27). It is a small group of only 8 yearlings in the population, 5 of which are graded stakes winners, but it may stand up to greater scrutiny with larger numbers as the sires of these yearlings are what would be termed as "former commercial sires", i.e older sires who have proven they can get a good horse (they had sired an average of 30 Lifetime SW) but just not consistently with their lifetime AEI's all below 1.60. Importantly, they were out of mares that could really run and were all early in their stud career. The fact that the latter (age of mare) was not a variable that this group was further subdivided on indicates that there is a level of homogeneity in this group.

Group 6 –These yearlings are by a group of sires whose service fees at the time their yearlings were sold were sitting at between $15,000 to $20,000. They were either first season sires, or sires who had only a few crops but hadn't yet broken out to the top echelon. There was one proven sire in Proud Citizen. A close look at this group sees that two stallions – Scat Daddy and Discreet Cat, both first crop sires who had a strong start to their career feature as the sire of 5 of the 10 stakes winners in the group so it is probable that this group might be found wanting when larger numbers are examined across multiple years.

Group 7 – With 14 graded stakes winners from 18 horses, possibly this may be the more interesting group to further investigate. These yearlings were sired by stallions whose service fee at the time the yearling was sold fitted between $20,000 to $30,000 and whose Lifetime AEI at the time the yearling was sold was above 1.765. This is a solid group of yearlings by proven sires such as Ghostzapper, More Than Ready, Macho Uno and Arch. This group was also thoroughly buyable with an average sale price of US$94,500 for the 14 graded stakes winners (with $475,000 for Wine Princess, out of the Champion Azeri, skewing the numbers a little). It is also probable that with a larger data set that there would be further partitioning of this group by the Dam's Racing Class (CPI). ↗

PEDIGREE ANALYSIS Shortlisting yearling sales with machine learning

Rpart results in a single classification tree. Random Forests, another machine learning technique, grows many classification trees and decides a final predicted outcome by combining the results across all of the trees in an ensemble method. The Random Forest algorithm is implemented by using the obviously named randomForest package in R. One of the great advantages that randomForest offers is that besides being less prone to overfitting, it also allows us to rank the importance of each variable. After running randomForest on our dataset, below is a list of all of the variables ranked by their importance in determining the best model.

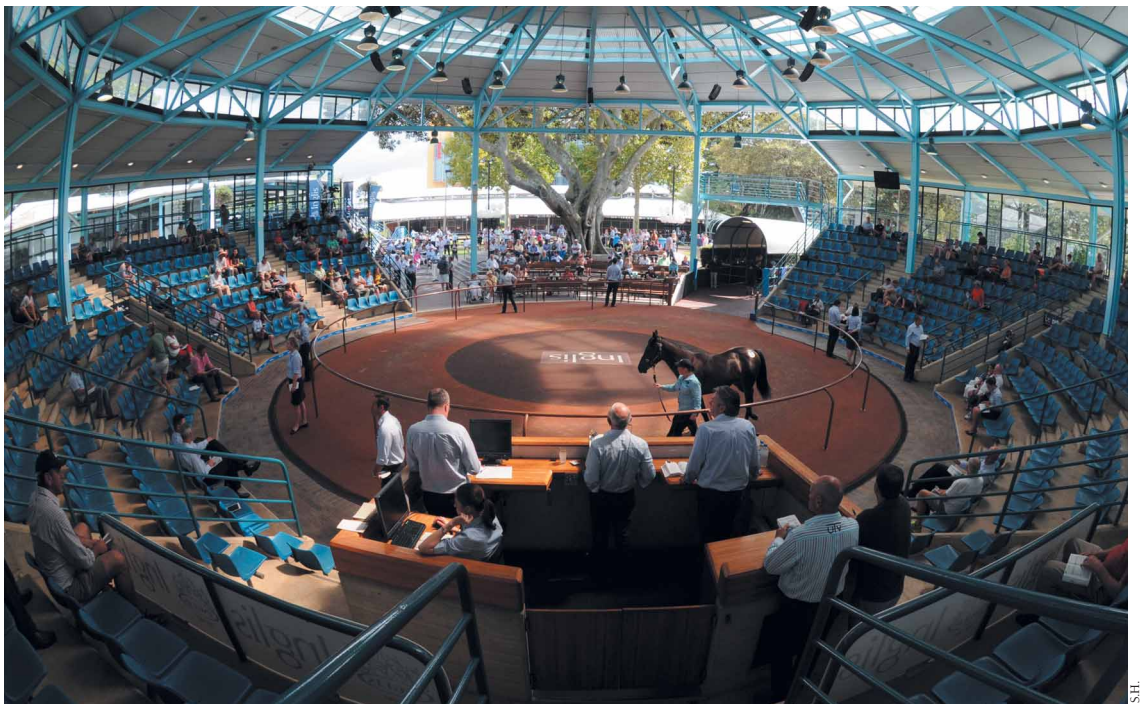
Variable	%IncMSE
Log of Sale Year Fee	21.60
Sire Lifetime AEI	16.53
Log of Conception Fee	14.08
Dam CPI	10.89
Dam is SW?	9.03
Sire Lifetime SW	8.20
Dam No of GSW	6.30
Dam No of SW	5.01
Sire Foals of Racing Age	4.92
Dam Age at Conception	4.80
GI	4.27
Dam Foals to Race	3.99
Dam Raced	3.28
Sire Age At Conception	2.88
Crop No	2.63
Dam is GSW	1.68
Foal Rank	1.63
Dam Foals to Win	1.54
Dam Wins	1.32
Dam is a G1SW	0.35
CoD	-0.98
Dam is a Winner	-1.79
SW Relations	-2.62
Dosage Index	-2.87
Grand Dam SW	-3.84
Days	-4.94

The above table is sorted by the percentage decrease in Mean Squared Error of each variable where higher values indicate more important variables in determining the status of the yearling as a graded stakes winner or not.

From this point we can start to look at the variables that were of importance in both rpart and randomForest and bring them together for a more complete model over a greater data set.

There is a high correlation between the variable log of the Conception Fee and the log of the Sale Fee so in a larger study we can drop the former in preference for the more powerful latter. It is also noteworthy that the foaling date of the yearling (days), if the granddam is a stakes winner and how many stakes winning relations the dam has isn't relevant to classifying graded stakes winners. The same goes for Dosage which has no relation to racing class.

The variable importance generated by randomForest indicates that there is really only 6 variables that don't belong (the six negative MSE variables), but when pruning the trees generated by rpart, it is clear that using more than 10 variables will only create a decision tree that overfits the data and partitions the groups down to meaningless numbers. Thus, the 10 most important variables seem to be:



S.H.

1 **Log of Sale Year Fee –** this is unsurprising. Previous hedonic pricing models by Robbins & Kennedy (2010) indicated that the Service Fee of the stallion at the time the yearling was sold was generally a good indication of the merit of the sire.

2 **Sire Lifetime AEI –** a variable to discriminate for sire performance (or lack therof if they are first crop sires).

3 **Sire Lifetime SW –** another variable to discriminate sire performance. This variable may drop in importance in a larger study where there is more variability of stallions.

4 **Sire's Foals of Racing Age –** this is probably a negative predictor. As previously discussed in Bluebloods, we have data to suggest that when an elite stallion has more than 1200 foals his production of graded stakes winners drops off considerably.

5 **Dam CPI –** a variable to discriminate for the racing class of the mare. The appearance of this variable is unsurprising given the proven correlation between the race performance of the mare and subsequent performance of her offspring.

6 **Dam is SW? –** this is a yes/no variable so we were surprised that it rated so highly. It seems that there are situations where you have mares with high CPI's that are not stakes winners and Stakes winners with low CPI's.

7 **Dam's # of GSW –** a measure of the production of the dam at the time the yearling is sold. Our data suggests that if the mare has a graded stakes winner and she is young enough, she's more likely to have another.

8 **Dam's # of SW –** similar metric to above comment on the dam's production of stakes winners.

9 **Dam Age at Conception –** this is probably to do with the declining production of stakes winners from old mares. We are a little surprised that it picked this variable over Foal Rank (i.e how many foals out of the mare) but it has.

10 **Generational Interval –** this is the difference between the yearling and the age of the sire and dam when they were mated to produce the yearling. It seems that there is something to be said for young mares going to older, presumably proven, sires. There were 20 graded stakes winners in our study group where the sire was 14 years or older at the time of conception. The average dam age for these 20 graded stakes winners was 8.5 years.

The above shows how we can start to use machine learning techniques to classify yearlings for selection. As previously stated the limitation of this small study needs to be considered. We have only looked at one sale in one year and a random subset of that sale. That said, we do now have some variables to use in a larger study which will give us greater insight and allow us to create a decision tree model that can be used to classify yearlings across all sales and ensure that the 'shortlisting' of yearlings is consistent.

Groups 2, 6 and 7 have an extraordinary high percentage of graded stakes winners in them when compared to other groups. This is not to say that there will not be stakes winners in other groups of yearlings. There will be. The point of this exercise is to find a group of yearlings that a yearling buyer can firstly be sure will be consistent in their quality regardless of the timing/venue of the sale, but secondly, and more importantly, prove to be a group that has a higher percentage of graded/group stakes winners than other groups giving significant advantage. As mega-horsehandicapper Dana Parham once famously said…"I strike where I have significant advantage, where I have advantage I have winners." We are using machine learning to firstly find where advantage can be found and then apply consistency and ignore all the other noise that is found at yearling sales.

To a certain extent the classification of these horses into groups relies on the widespread assortative matings that occur in this industry. Mares of similar racing quality are generally bred to stallions of similar service fee and as we have seen in other studies by in large the service fee is a reflection of horses of similar genetic merit. This results in yearlings being able to be classified into similar groups with the hope being that as we have shown with groups 2, 6 and 7 above, one group has significantly more graded stakes winners in it than others. If this group is a manageable portion of the population, this would then allow for all the horses in the group to undergo further cardiovascular and genetic testing resulting in a further shortlist with a significantly high proportion of stakes winners found in it. ■