Genetic parameters of sport horses. Evidence from show-jumping results.

Vojtěch Mišák

2020
May

Introduction

All horse breeders would like to breed as best horses as possible. One way, how to improve sport performance of foals is to focus on physical properties and sport results of their parents.

The aim of this project is to contribute to the topic of genetic parameters of show-jumping horses with machine learnings (ML) methods. Machine learning frameworks can be used for sport result prediction (most notably recent study by Bunker and Thabtah, 2019). There is a reasonable assumption that ML tools might help to improve current methods that are used for evaluating future performance of foals.

Results of this project might be used in future research, because the market with horse sperm is rapidly growing (Hellsten et al., 2006).

1 Review of study

The background study for this project is *Review of genetic parameters estimated at stallion and young horse performance tests and their correlations with later results in dressage and show-jumping competition* (Hellsten et al., 2006).

In this article authors review studies from seven European countries - Belgium, Denmark, France, Germany, Ireland, The Netherlands and Sweden.

Hellsten et al., 2006 conclude that average correlation between stallion performance and future sport results of foals is **0.39**.

2 Replication of study

To replicate the study by Hellsten et al., 2006, I have collected show-jumping results from Czech equestrian federation of 7 years old horses. Because there is no way, how to match these data with ancestors‘ sport results, I had to match them manually, which was very time demanding.

To summarize, my dataset contains 179 seven years old horses registered in Czech equestrian federation.

Each horse has its competition level which is the maximal height of obstacles of the competition on that the horse has ever competed. In other words, top horses compete on 160 cm high obstacles, so their level is 160. On the other hand, beginner horses typically have higher level up to 120, because they do not jump over higher obstacles than 120cm.

For each seven years old horse I tried to find its competition level and competition level of its sire, dam and sire of dam. Success rate of finding competition levels of horse pedigrees was about 70 %. Summary statistics of my dataset is provided in Table 1.

To replicate the article by Hellsten et al., 2006, I computed correlation between competition levels of 7 years old horses and their sires. The correlation is **0.29**. Moreover, the correlation between between competition levels of 7 years old horses and competition levels of their dams, and sires of dams is 0.19, and 0.25 respectively.

Table 1: Data summary statistics

	Min	**1Q**	**Median**	**Mean**	**3Q**	**Max**	**NAs**
7 yo horses level	80	110	115	114.3	125	135	0
Sire level	110	145	150	146.5	150	160	47
Dam level	100	115	120	124.5	135	150	67
Sire of dam level	110	140	150	144.5	150	160	49

3 Extension of study by ML methods

I have extended the study by Hellsten et al., 2006 by following methods. Firstly, I have imputed missing values using Random forests.

Secondly, using Truncated regression and Random forests regression I foretasted competition levels of 7 years old horses.

Lastly, I divided young horses into two groups: premium and non-premium. Using Logistic regression and SVMs I tried to cluster young horses into premium/non-premium groups.

All steps are described in subsections bellow.

3.1 Missing data imputation

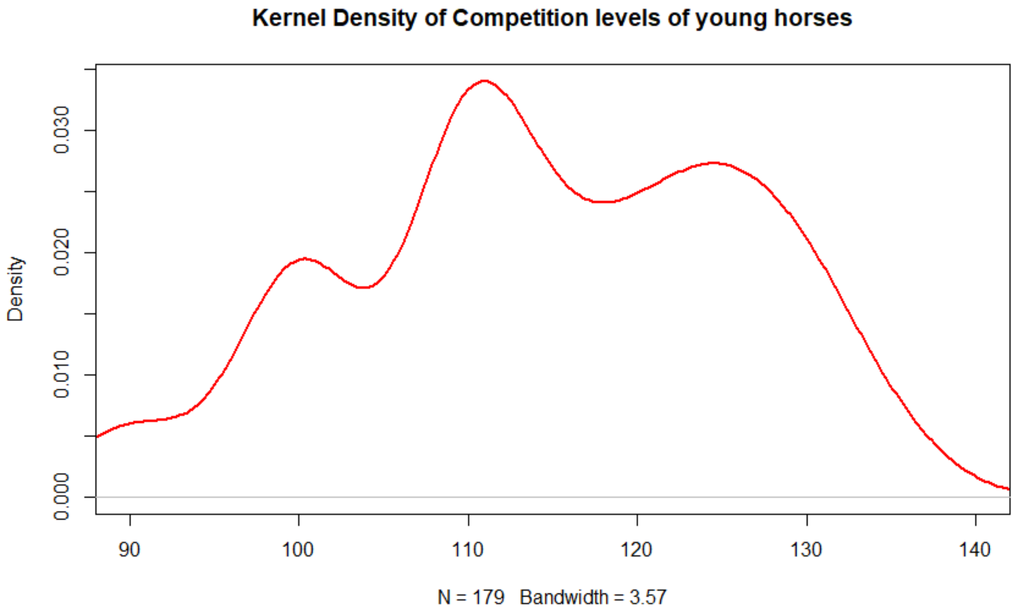
I have imputed missing data using *rfImputed* R function which uses random forests to predict the missing values. After that, I have split data set into training dataset and testing dataset. The proportion between Train and Test dataset is set to 0.8.

Now, I am able to replicate correlation measures again. The correlation between competition levels of 7 years old horses and their sires is **0.24**. Moreover, the correlation between between competition levels of 7 years old horses and competition levels of their dams, and sires of dams is 0.24, and 0.23 respectively.

We can see that the correlations have settled around 0.24. The reason why is probably because Random forest measure, how often two data points end in the same leaf for different trees. Then it imputes. Therefore it seems that computing correlations on datasets with missing values, as I did in Section 2, is more trustworthy approach.

3.2 Truncated regression

Competition levels follows truncated distribution, because minimal level of show jumping competition is 80. See Figure 1 bellow.



Therefore truncated regression seems to be a better approach than OLS regression.

I estimated 7 yo horse competition level based on competition levels of their sires, dams and sires of dams. Output of truncated regression made on training dataset is listed bellow. The RMSE of the model is 11.23[1] . The RMSE of the same model that was run on testing dataset is 11.47.

```
Estimate Std. Error t-value  Pr(>|t|)
(Intercept)     30.322454  21.342626  1.4207  0.155391
sireLevel        0.364762   0.117480  3.1049  0.001904 **
damLevel         0.110286   0.083206  1.3255  0.185021
sireOfdamLevel   0.115477   0.097736  1.1815  0.237394
sigma           11.232005   0.685602 16.3827 < 2.2e-16 ***
```

[1] Note that RMSE is equals to sigma in the model, which comes from the specification of *truncreg* function in R.

3.3 Random forests regression

Random forest regression were used to "upgrade" estimation approach above. Similarly to the subsection 3.2, I estimate and validate random forests model in two steps.

Firstly, I estimated and calibrated Random forests model on training dataset. Secondly, I run the model from previous task on test dataset.

RMSE were used as a accuracy measurement. Random forests model on training dataset had RMSE = 7.24. Model RMSE on testing dataset was 11.53.

In real life, horse breeders do not really care about the exact competition level of their foal, but they distinguish between super horse and others. For this purposes I extended the analysis in subsections bellow.

3.4 Premium horses detection using Logit

For needs of this analysis I divided 7 yo horses into two groups. Premium horses and non-premium horses. Horse is considered to be premium, if its competition level is above or equal to 125. Recall Table 1 to see that 125 corresponds to the third quantile of competition levels of young horses. In other words, 25 % of young horses are considered as premium, 75 % are non-premium horses.

Logit model was used as a simple method to cluster 7 yo horses into premium/non-premium groups.

On the logit model output bellow, we can see that model were able to detect premium horses in 4 cases out of 9. The accuracy was 45 %. In the case of non-premium horses, Logit model correctly detected 22 horses out of 27 (accuracy was 81 %).

```
predikceLogit <- print(with(fulldataTest,
          table(y=Premium, glPred=pred>=0.4)))

# glPred
# y     FALSE  TRUE
# 0       22     5
# 1        5     4
```

3.5 Premium horses detection using SVMs

I provided three types of SVMs models.[2] First model uses radial basis kernel function, the second model uses linear kernel function, the third model uses polynomial kernel function. All outputs and their discussion are listed on the new page.

We can see that the best SVM model is the last one, i.e. the model that uses polynomial kernel function. It is the best not only on testing data, but also on training data.

The accuracy of the best SVM model is 66 % both when clustering non-premium horses and premium horses.

[2] I provided accuracy measures both on training and testing dataset (see attached R code). In this paper I listed only outputs made on testing data, because I do not want to overwhelm the reader with too many models.

```
mSVMV<-ksvm(Premium ~ sireLevel + damLevel + sireOfdamLevel ,
         data=fulldataTrain , kernel=”rbfdot”, C=10)
fulldataTest$Kernel<- predict(mSVMV, newdata=fulldataTest ,type=”response”)
predikceSVM <- print(with(fulldataTest ,
         table(y=Premium,glPred=Kernel>=0.30)))

glPred
y    FALSE TRUE
0      20    7
1       6    3


mSVMV<-ksvm(Premium ~ sireLevel + damLevel + sireOfdamLevel ,
         data=fulldataTrain ,      kernel=”vanilladot”, C=10)
fulldataTest$Kernel<- predict(mSVMV, newdata=fulldataTest ,type=”response”)
TableSVM <- with(fulldataTest ,table(y=Premium,Kernel=Kernel))
predikceSVM <- print(with(fulldataTest ,
         table(y=Premium,glPred=Kernel>=0.0466)))

    glPred
y    FALSE TRUE
0      18    9
1       5    4


mSVMV<-ksvm(Premium ~ sireLevel + damLevel + sireOfdamLevel,
         data=fulldataTrain , kernel=”polydot”, C=10)
fulldataTest$Kernel<- predict(mSVMV, newdata=fulldataTest ,type=”response”)
TableSVM <- with(fulldataTest ,table(y=Premium,Kernel=Kernel))
predikceSVM <- print(with(fulldataTest ,
         table(y=Premium,glPred=Kernel>=0.0466)))

    glPred
y    FALSE TRUE
0      18    9
1       3    6
```

4 Discussion

I can summary this project as follows. When I replicated the background study (see correlation measurements in Section 2), I can see that higher level of correlation was found on the the simple dataset. After the missing data imputation, using Random forests, the level of correlation decreased.

Moreover, it is hard to decide, whether is better Truncated regression model or Random forests model. RMSE in both cases were almost equal.

In the final part of the project, I clustered horses using Logit model and SVMs. I can say, that SVM model with polynomial kernel function is a better approach than Logit model, but not by much.

To summarize, I can say that ML methods, I used in this project, do not give us a significant improvement when compared to standard measurements as linear regression or even correlation measurements. On the other hand, there might still be ways, how can ML and AI in general, change the way of horse breeding.

Bibliography

Bunker, R.P. and Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), pp.27-33.

Hellsten, E.T., Viklund, Å., Koenen, E.P.C., Ricard, A., Bruns, E. and Philipsson, J., 2006. Review of genetic parameters estimated at stallion and young horse performance tests and their correlations with later results in dressage and show-jumping competition. Livestock Science, 103(1-2), pp.1-12.