Machine Learning Analysis of Horse-Rider Dyad Performance in Show Jumping

John Hutchinson

A Senior Thesis submitted in partial fulfillment
of the requirements for graduation
in the Honors Scholars
Liberty University
Spring 2025

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial
fulfillment of the requirements for graduation from the
Honors College of Liberty University.

_____
Diana Schwerha, Ph.D.
Thesis Chair

_____
Michael Zamperini, M.S.
Committee Member

_____
Cindy Goodrich, Ed.D.
Assistant Honors Director

_____
Date

Abstract

This study utilized machine learning techniques to analyze factors that contribute to the performance of horse and rider combinations (dyads) in Fédération Équestre Internationale (FEI) jumping as ranked by the FEI Combination in Jumping ranking list. Data on horses and riders were gathered from 6000 FEI webpages, and machine learning models were applied on both a priori data such as sire and color and a posteriori data such as competition statistics to study not only the factors that contribute to the success of a horse and rider combination but also to examine if any of these already known factors can provide insight into future performance. Data were analyzed using the CatBoost algorithm and Shapley Additive Explanations (SHAP). All machine learning models were trained and evaluated with K-fold cross validation on these different combinations of factors to produce a statistically significant value from a sample in a regression of the predicted rankings and the actual rankings, with the model trained on all data leading to a R^2 of 0.66. These results vastly outperformed regression analysis of individual variables against the target. Analysis of key features revealed that the features that contribute most to the models' predictions are those related to the amount of practice and experience of both the horse and rider.

Keywords: Machine learning, CatBoost, SHAP, Horse-Rider performance

Machine Learning Analysis of Horse-Rider Dyad Performance in Show Jumping

Every year, the equestrian sport and industry beneficially impact the global economy. In Europe alone, the sector contributes over 100 billion euros a year to the economy and provides more than 400,000 full-time job equivalents, with over six million horses permanently utilizing six million hectares of grassland for grazing (Cooke, 2013). With that being said, the horse industry is incredibly important economically, but is mostly beyond the reach of academia, except in breeding and veterinary studies. This study seeks to introduce modern data analysis and machine learning techniques on some of the most readily available data in the industry: the FEI Combination in Jumping Rankings and the horses and riders that comprise it. The goals of this study are twofold: to evaluate the feasibility of machine learning in evaluating the quality and athleticism of a horse and to glean any insights into factors that contribute to good performances from horses and riders.

Literature Review

Machine Learning

Background. Machine learning is a branch of artificial intelligence that uses data and algorithms to mimic how humans learn. Broadly speaking, machine learning trains a machine to perform a task without explicit programming through the use of data and algorithms. Usually, machine learning falls into three categories: supervised, unsupervised, and reinforcement learning. For the purposes of this study, only supervised learning is considered. Supervised learning is typically used for making predictions. The process of supervised learning involves feature variables and target variables. The data that comprises these variables is typically divided into train data and test data. The train data are used for training the model, which involves applying the model's algorithm to the data to find any relations between the features

and the target. Then, the model's performance is evaluated out of sample on the test data. If a model fits the train data well, but its performance on the test data is significantly worse, this means that the model is overfit on the train data. This implies that the model only learned spurious and temporary relations between the features and the targets. Supervised learning can further be broken down into two more categories based on the target. If task requires predicting discrete or categorical variables, it is a classification problem, but if the variables are continuous, it is a regression problem. This study utilizes a model called CatBoost for predicting a horse and rider pair's FEI rank, which is inherently a regression problem. CatBoost is a SOTA machine learning algorithm that is based on gradient boosting, which in turn relies on one of the simplest machine learning models: the decision tree (*What is machine learning?*. n.d.).

Decision Trees. A decision tree, more specifically known as a classification and regression tree (CART), is one of the earliest supervised machine learning algorithms. Decision trees were formally published in 1984 by Breiman et al. in their book *Classification and Regression Trees*; however, the applications of similar models date to even earlier times. A decision tree consists of nodes and branches and is built through a process called recursive binary splitting. Basically, a classification decision tree starts with a root node that contains all the targets and then seeks to separate the targets into their respective classes by reducing the impurity at each decision node. This is achieved by using some function, such as Gini impurity or information gain, to find what feature and value of that feature separates the target variable into its respective classes the most. This process is repeated until the classes are completely separated or some other criterion, such as a maximum depth of the tree, is satisfied. For a regression tree, the process is modified slightly to handle continuous target variables. Instead of trying to separate the samples into different classes, the model seeks to split them into two

different points where the mean squared error is minimized. If each target has features that make it identifiable in any way, the greedy application of this process will lead to each target getting a prediction of exactly its value in sample, but the model will likely perform poorly out of sample. Boosting attempts to remedy overfitting by using many weak learners (trees with a limited depth) in series.

Gradient Boosting. Gradient boosting stacks shallow decision trees, a type of weak learner, consecutively. Each consecutive tree seeks to minimize the error of the previous tree by predicting the residual errors from the predictions of the previous trees. In practice, gradient boosting follows this algorithm: Fit the decision tree; calculate the residual error; fit a new tree that predicts and adjusts for the residuals; and then continue the process until some stopping criterion is met.

CatBoost. This study uses a machine learning model called CatBoost (Categorical Boosting), which achieves SOTA performance on tabular data. At its core, CatBoost is a gradient boosting algorithm with additional changes. CatBoost was proposed by Prokhorenova et al. with two major improvements over existing gradient boosting implementations (2019). The first of these improvements is ordered boosting, a permutation-driven algorithm which solves a target leakage called prediction shift. Prediction shift occurs when the gradient boosting, after a few iterations, relies on the entire training sample to create predictions, which shifts the distribution of the target variable and introduces information in the training data that cannot be known in testing and leads to a decrease in performance from training to testing. CatBoost essentially adapts the training to take place sequentially in artificial time, ensuring that during the training the model can only access its observed history of training samples, and not the entire dataset, thus remedying the prediction shift.

SHAP. The feature importance of the several factors is determined using SHAP (SHapley Additive exPlanations), a unified framework for determining why a model makes a certain prediction, formally known as model interpretability. In other words, how each feature contributes to the prediction of model, based not only on that feature's specific value, but also considering its interactions with other features. The SHAP algorithm expands on the classic Shapley value estimation that is commonly used in game theory. On an intuitive level, Shapley value estimation is quite simple. For all possible permutations of features, the model is trained and gives two different predictions: one that includes the feature whose importance is being calculated and another that lacks it. The mean difference between these two values across all possible combinations is the impact that this specific feature has on the predictions. Mathematically, this is represented in (1), which finds the weighted average of the marginal contribution of feature *i* across all possible permutations of features of all subsets *S* of *F,* where *F* represents all the features provided to the machine learning model.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{1}$$

Shapley values are a clever and intuitive solution for finding the contribution of a feature to a prediction, but they suffer a major problem with scaling, as the computation time rises exponentially with the number of features. SHAP seeks to remedy the calculation of vanilla Shapley values through approximations and an understanding of the workings of the underlying model, allowing for a much more efficient, although still computationally intensive algorithm.

Naturally, any correlations found with SHAP do not necessarily imply causation. In correlation, when one variable changes, its correlated variable tends to change. However, in causation, when one variable changes, the other variable changes in reaction to that change.

SHAP only discovers correlations that have been detected by the machine learning model but does not imply that the correlations do in fact drive actual changes in the target variable (Dillon et al., n.d.)

Show Jumping

Background. Show jumping is one of three equestrian sports in the Olympics, but it may be unfamiliar to many people. In show jumping, a rider and a horse try to complete a course as fast as possible without knocking down any of the jumps. Although many different variations of scoring exist (called tables), the most common consists of a first round and a jump off. The first round is a qualifying round for the jump off. To qualify, the horse and rider must go around the course clear, which means completing the course without knocking down any of the jumps and within a certain amount of time (called the time allowed). Knocking down a jump results in a penalty of four faults, and every second commenced above the time allowed results in a penalty of one fault. Every horse and rider pair that goes clear in the first-round proceeds to the jump off. The goal of the jump off is to complete the course clear and as quickly as possible. Places are determined by sorting riders by number of faults incurred ascending and then sorting each group by time ascending (*FEI jumping rules*, 2023).

Ranking Methodology. This study uses the FEI's (Fédération Équestre Internationale) ranking of individual horse and rider combinations in show jumping. The FEI is the international governing body of equestrian sport and seeks "to develop equestrian sport globally in a modern, sustainable and structured manner with guaranteed athlete welfare, equal opportunity and ethical partnership with the horse" (*About FEI*). The FEI is responsible for calculating international rankings of riders, which as can be imagined, is a complicated and heavily regulated process that must address many different situations. A full summary of the

calculation of the Longines rankings for show jumping can be found in *Rules for the Longines rankings*. In short, riders earn points for each competition based on their final placing and the point group, which is determined based on prize money, height (which at a minimum is 1.45 meters), and special factors such as whether that competition is part of the Olympics, a Nations Cup, or FEI Finals. Higher placings obviously lead to more points than lower placings, but a place in a higher class also earns significantly more points than the same place in a lower class. For example, winning a 1.60-meter class leads to three times more points than winning a 1.45-meter class. These higher point values come from various factors that increase the level of competition and prestige of the event. Higher classes, in addition to the extra challenge of added height, also offer more prize money, which typically attracts more competition.

Past Studies

Many studies have been conducted regarding the statistical impact of horse traits in breeding. However, little research has been done analyzing the factors that contribute to the performance of top horse and athlete pairs. This study seeks to apply modern machine learning techniques to predict horse and rider performance in show jumping and then identify the factors that most contribute to the predictions of their performance. So instead of simply predicting the performance of a horse and rider pair, for some sort of gain, such as predicting the rank of a prospective horse, this study seeks to identify factors that are correlated with higher ranks.

Genetics and the heritability and desirability of different traits in horses have been studied extensively. For example, Welker et al. (2018) use classical statistical models to determine how breeding can be leveraged more effectively to produce desired traits in horses. Many similar studies on breeding have been conducted, particularly in Europe, where breeding is a major industry.

Additionally, statistical analysis of the distribution of faults demonstrates that they are far from random. Marlin and Williams (2020) assess the distribution of faults in Nations Cup competition and determine that faults are much more common in the second half of the round than in the first half. Additionally, horses are much more likely to knock down upright fences and fences that are part of combinations**.**

Other authors have evaluated the impact of different variables on competition. Hanousek et al. (2020) established that horse age, sex, and number of riders impacted the performance of horses in the British Eventing Horse Trials through principal component analysis, and Visser et al. (2003) examine the relationship between show jumping performance and horses' early personality traits, concluding that a horse's early personality and learning ability predict a substantial portion of show-jumping performance. To this author's knowledge, this study is one of the first to apply modern machine learning techniques to predict and analyze show jumping rankings using such a large number of variables including both the horse and the rider.

Methods

Data Collection

This study uses data publicly available on the FEI website. The dataset consists of over the top 3000 horse and rider combination in jumping rankings. For each entry, various data points were collected, which can be seen in Appendix A.

The data can be broken into five clear categories: athlete details, athlete performance, horse details, horse performance, and horse pedigree. Both horse and athlete details are descriptive details about the horse and athlete that were collected on their registration with the FEI. The athlete and horse performance both primarily consist of a list of the 50 most recent competitions for the corresponding horse and athlete. The horse pedigree includes the pedigree

of the horse as far back as is recorded by the FEI, along with some details associated with each horse, such as color and breed.

Preprocessing

The data from the FEI website were not immediately conducive to machine learning due to its dirtiness, presumably just ingesting the raw data entered during registration into a database and then displaying it when it is queried by a webpage. Additionally, the format of the class results varied based on the type of competition. Because of the many variations in categorical data, for example, the same breed may be entered in different languages or all capitalized or even misspelled, efforts were taken to standardize each category so the model can learn associations with that category.

To reduce the data dimensionality due to the classes, several features were created to summarize the athlete and horse performance. These both describe the performance of the athlete or horse in their classes and describe the classes themselves. They include metrics such as the number of entered competitions, the mean height of the fences, the percentage of the classes that are 1.60 meters, and the mean time and faults in both the first round and the jump off.

The data were then separated into different datasets with various purposes. The two initial datasets consist solely of the horse and athlete details and exist to discover if there is any a priori information available for predicting the rankings. The horse details dataset in particular, if any predictability exists, could be used to evaluate potential horses based on their characteristics before purchase to glean insights into their chances of success. It is important to note, however, that for these purposes, this study may suffer from selection bias, as it has already narrowed down the field of millions of horses to the few thousand best in the world, so it would be

advisable to conduct another study with a more comprehensive dataset to fulfill this purpose, unless, of course, the buyer is only evaluating horses that have already reached this field but may have more potential. Additionally, under the Terms and Conditions of the FEI website, any material obtained from it may not be used for commercial purposes (*FEI website terms and conditions, n.d.*). The athlete details dataset was designed not only to isolate the athlete details themselves, but also to compare the amount of predictability with the horse details. In other words, which is a better predictor of success, the rider, or the horse?

Two other datasets, the horse and athlete performances, were created for the purpose of discovering a posteriori the factors that contribute to horse and rider success, as measured by the rankings. As these contain information that is used in calculating the rankings, like the placings and the class heights, these cannot be used to predict the rank in advance, but their use lies in uncovering what factors contribute the most to these rankings.

Finally, the last dataset aggregates all the former data into a single table. As it includes the performance data, it also cannot be used for predicting rank in advance, but it does provide an overview of the most crucial factors in comparison to each other. For a list of all datasets and their respective features, refer to Appendix A.

Exploratory Data Analysis

The data were analyzed and visualized in several ways to uncover any noteworthy characteristics, relationships, and other valuable insights. Different variables were graphed on their relevant charts, and continuous variables were tested for correlation with rank. This process allows a better understanding of the dataset and offers the opportunity to identify any anomalies, patterns, or errors in the data.

Machine Learning

The CatBoost machine learning algorithm was then applied to each dataset, to infer the rank from the features. Each dataset followed the same pipeline for training and evaluating the machine learning model, which started with a splitting technique known as K-fold cross validation. K-fold cross validation splits the data for training and testing in a way that seeks to maximize the statistical significance of the results by providing the most test data. In this process, the data is partitioned into a certain number of sections (k folds). In the first iteration, the first $k-1$ sections are used for training, and the kth section is used for testing. In the second iteration, the first $k-2$ sections and the kth section are used for training, and the $k-1$th section is used for testing. This process continues until each section has been held out once, with the model being trained on all the other sections, which essentially allows testing on the entire dataset.

In each iteration of the K-fold cross validation, the CatBoost regressor was fit on the training data and then tested on the hold out data. The results were then evaluated using R^2—that is, how much of the variance of the rankings is explained by the predicted rankings. These metrics were then aggregated across iterations to find a mean and standard deviation.

Results

Exploratory Data Analysis

The EDA process revealed several interesting facts about the data that might be interesting to the keen observer or horse enthusiast. These range from simple descriptive statistics to tests of important variables that are later used in comparison with the machine learning results. Each feature was analyzed with techniques relevant to its data type. In general,

this implied that discrete data are analyzed with pie charts and percentages, and the continuous data are investigated with scatter plots and t-tests.

Horse Details/Pedigree. The horse details and pedigree consist primarily of categorical data, so the EDA consisted mainly of analysis of percentages and comparative rank based on these categories. Analysis of the horse sex revealed that of the total sample, 47% were geldings, 32% were mares, and 21% were stallions (Figure B1).

Analyzing the breeds revealed that only nine breeds comprised 86% of the horses in the rankings, with the Dutch warmblood being the most frequent at 18.1% of the total sample and the Belgian warmblood having the lowest median rank out of the whole, at 0.443.

The following table displays all the breeds with over 50 horses in the rankings, along with descriptive statistics. A graphical representation is shown in Figure B2.

Despite the uniformity of the breeds, the breeders were quite fragmented and disparate, with most breeders only having one horse in the rankings. The primary exception to this rule is Gestüt Lewitz, a German breeder run by Paul Schockemöhle, which had over 150 horses in the rankings.

The primary continuous variable in the horse details dataset was the age of the horse. This feature was analyzed with both regression against the rank and a scatter plot. As can be seen in Figure B3, horse age and rank have a weak correlation, and the regression results in a R^2 of only 0.007. However, due to the substantial number of samples, this is also highly significant, with a p-value of essentially zero.

Table 1

Descriptive Statistics of the FEI Rankings

Breed	**Percentage**	**Mean Rank/Total**	**Median Rank/Total**	**Std Rank/Total**
Dutch Warmblood	18.1%	0.523	0.528	0.293
Selle Francais	13.3%	0.464	0.434	0.305
Holsteiner	12.4%	0.523	0.523	0.293
Oldenburg	10.9%	0.470	0.454	0.277
Belgian Warmblood	9.5%	0.443	0.406	0.292
Hanoverian	6.5%	0.519	0.511	0.281
Zangersheide	5.8%	0.486	0.469	0.287
Westphalian	5.5%	0.526	0.551	0.303
Irish Sport	3.8%	0.499	0.521	0.287

Attentive readers may also notice clear stripes that seem to group periodically and increased sparsity before the age of 9 and after the age of 17. The groupings come from the typical equine breeding cycle, which results in most foalings occurring between April and June in the Northern Hemisphere, and the sparsity occurs because before the age of eight, horses are still being developed, and most will start retiring from this level of competition, around 17.

Athlete Details. Similarly to the horse details, the athlete details also consist primarily of categorical data. However, one of the numerical variables, the number of horses ridden by the athlete in FEI competition, offers one of the greatest relationships with rank, as seen in Figure B4.

Expectedly, the number of horses ridden by an athlete is negatively correlated with rank—that is, athletes who have ridden more horses are ranked higher, which means their rank decreases in numerical value. The unexpected part of this relationship is its strength. With a R^2 of 0.165 and a p-value that is essentially 0, the number of horses alone explains 16.5% of the variance in athlete rankings. When considering the sheer number of factors that determine athlete performance, this relationship is extremely strong, especially because in actuality, the number of horses ridden conveys no information about the longevity of the pairing of each horse and rider. Did the rider show the horse in one competition, or did they have a long partnership? Obviously, one would have much more of an impact than the other, but even without this information, the number of horses provides a strong correlation with rank. The gender of the athlete is another interesting statistic. 61% of the athletes in the rankings are male and only 39% are female (Figure B5). The median male rank over the maximum rank is 0.47, and the same statistic for females is 0.55. This is just a simple descriptive statistic but is almost certain to produce a machine learning model that penalizes females based on their gender, leading to completely different discussions on fairness and biases in machine learning, and whether or not models should be able to use such features, even if they are statistical facts, that are beyond the scope of this paper.

Although the athletes in the rankings compete for 87 different countries, 75% of the athletes come from only the most represented 16. France is the most represented country,

comprising 11% of the riders in the rankings, followed by Germany, which has 9% of the riders. Only three of the 16 countries: the United States, Mexico, and Brazil, are not located in Europe (Figure B6).

Horse Performance. As opposed to the details datasets, the performance datasets consist largely of continuous data. Additionally, these will be more highly correlated with the target variable, because the placings highly depend on these numbers, and since the heights of the classes are included, the features include a brief summary of all the data used for calculating the rankings. Therefore, these datasets are included to glean insight into best practices that may lead to better results and the model cannot be used as any sort of predictor, except of course through varying the parameters to find the optimal result to decide on the best course of action. In other words, answering questions such as how many and what kind of classes should be done, how long a break should be taken, or what area needs to be improved to gain the greatest increase in the rankings. Due to the considerable number of variables, only the most salient relationships are featured here.

The first of these is the number of competitions in which a horse or athlete has competed (Figures B7, B8). For both the horse and athlete, the number of competitions is negatively correlated with rank, meaning that competing in more competitions is associated with a numerically lower, and therefore better, rank. One interesting detail is the riders typically compete in many more competitions than horses. As seen in the graphs, the maximum number of competitions horses compete in is just below 500, but the athletes that have competed in the most competitions have competed in over 5000, which is more than 10 times more.

The next few graphs present findings that are slightly counterintuitive. The first of these displays the horse and athlete mean time in the first round (Figures B9, B10). The

counterintuitive part, though, is a strong negative correlation between the rank and the time, implying that horses and athletes with slower times perform better. How can this phenomenon exist when the classes are placed on time and faults? In this situation, one must be careful in interpreting correlation as causation. The answer is likely that the higher courses which result in more points are longer and harder enough than the lower courses, that even with increases in experience and skill, the times also increase. An interesting note on these regressions is that the horse first round mean time has a much higher R^2 than the athlete. This more significant result likely stems from the fact that the athlete rides many different horses, most of which do not contribute to the ranking with the particular horse in this horse-rider combination, but the horse is usually ridden only by one athlete, so all its results will count towards the rankings. A situation similar to the analysis of the mean time for the first round occurs for the jump off statistics, including both mean time and faults for the horse and the rider (Figures B11-B14). All of these variables have a counterintuitive correlation that likely results from increased difficulty in higher and longer courses.

The last features analyzed here are the break time and the standard deviation of break time for both the horse and the rider (Figures B15-B18). These show that both features have a strong positive correlation with rank—that is horses and athletes that compete often and consistently typically have better ranks.

Model Results

Machine learning models were trained and evaluated on six separate groups of features with k-fold cross validation with five folds. For each, metrics such as R^2 and p-value of the correlation between the predicted rankings and the actual rankings, the top-mean decile spread, the bottom-mean decile spread, and the inter-decile spread were calculated.

The decile spread metrics were formulated to capture any improvements that can be made by utilizing the models in their best and worst predictions. The predicted rankings were sorted into deciles, and then for the top and bottom deciles, the mean of the actual rankings was found. These were then used in calculations relative to each other and the means to find the percentage of improvement.

The top-mean decile spread, as shown in (6), measures how much better the actual rankings of the top predicted rankings are than random guessing, which has an expected value of the mean rank. The bottom-mean decile spread (7) calculates the opposite metric for the mean versus the worst ten percent of predictions. The inter-decile spread (8) calculates the percent improvement of the actual rankings corresponding to the top ten percent of predictions over those rankings corresponding to the bottom decile of predictions.

Let the indices of the highest and lowest predictions be as follows, with vectors y and $\hat{y}$, representing the actual rankings and predictions, respectively.

$$T_p = Quantile_p(\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}) \tag{2}$$

$$I_{0.9} = \{i | \hat{y}_i \geq T_{0.9}\} \tag{3}$$

$$I_{0.1} = \{i | \hat{y}_i \leq T_{0.1}\} \tag{4}$$

Then the mean of the actual rankings corresponding to the top and bottom predictions is written as:

$$\mu_{0.1} = \frac{1}{|I_{0.1}|} \sum_{i \epsilon I_{0.1}} y_i \tag{5}$$

$$\mu_{0.9} = \frac{1}{|I_{0.9}|} \sum_{i \epsilon I_{0.9}} y_i \tag{6}$$

These equations are used to calculate the decile spreads:

$$D_{0.1} = \frac{\mu - \mu_{0.1}}{\mu} \tag{7}$$

$$D_{0.9} = \frac{\mu - \mu_{0.9}}{\mu} \tag{8}$$

$$D_{0.9-0.1} = \frac{\mu_{0.9} - \mu_{0.1}}{\mu_{0.9}} \tag{9}$$

Table 2

Machine Learning Results

Dataset	R^2	p-value	Top-mean decile spread	Bottom-mean decile spread	Inter-decile spread
Horse Details	0.053± 0.012	2.71e-11± 4.66e-11	0.223± 0.43	-0.264± 0.034	0.385± 0.035
All Horse	0.536± 0.012	8.16e-142± 1.26e-141	0.782± 0.043	-0.678± 0.017	0.870± 0.027
Athlete Details	0.193± 0.039	7.46e-23± 1.49e-22	0.499± 0.046	-0.424± 0.034	0.648± 0.034
All Athlete	0.494± 0.022	2.52e-75± 5.04e-75	0.770± 0.042	-0.547± 0.046	0.851± 0.030
A Priori	0.461± 0.014	3.23e-61± 6.26e-61	0.718± 0.037	-0.588± 0.046	0.822± 0.021
All	0.663± 0.018	2.39e-105± 4.16e-105	0.827± 0.022	-0.703± 0.025	0.898± 0.013

Table 2 shows the out of sample regression results of the predictions and the actual rankings. All the results are all statistically significant, even when the effect size is small, due to the sample size. As expected, the horse and athlete details models had the worst performance because they possessed only a priori information that was only slightly correlated with ranking. However, despite only having R^2 of 0.053 and 0.193, respectively, the predictions of the top and bottom deciles can lead to impressive improvements in rank, with the horse details model having a 22.3% improvement in rank over the mean, and the athlete details having an improvement of 49.9%.

The performance of these models gives a significant improvement over random guessing, but better results were achieved by considering all the features that can be known a priori. These include both the horse and athlete details, and the mean and standard deviation of the break time for the horse and athlete. Although the latter variables are not technically known before competition, they can be determined by the athletes, assuming they have the time and resources. The performance of the model trained on this data is quite impressive, with an R^2 of 0.461 and an improvement in rank of 71.8% over the mean.

For the remaining datasets, the efficacy of the predictions matters less, except insomuch as they need to possess some explanatory power to provide a valid analysis of feature importance. In this case, they all have relatively high R^2, all above 0.494. With more feature engineering, the R^2 could likely be raised significantly, but the current features are posed in a way that allows them to be actionable and insightful goals for an athlete to achieve to increase their place in the rankings.

Feature Analysis

Although it is beneficial to train models to predict the rank of a horse and rider, these models are nothing more than black boxes that output numbers. It is helpful to interpret these models with feature analysis through SHAP. Graphs displaying the SHAP values of at most the top ten features can be found in Appendix D. These graphs display the SHAP values, or the additive value of that feature on model output, by feature, along with a color scale for high and low values of numerical features to show variations in the value of that feature.

For the horse details model, the feature that contributed the most to the model's predictions was the number of ownerships (Figure D1). Higher numbers of ownerships were rewarded by the model, resulting in a numerically lower rank, and fewer ownerships were

penalized. The next most important variable was the age of the horse. Unsurprisingly, horses that were too young and inexperienced were penalized highly, horses that were too old were slightly penalized, and horses that were the perfect age for a combination of experience and athleticism were slightly rewarded. The remaining numerical variable, sire age, was rewarded for younger ages. As the recorded age is the sire age currently, not at the birth of the horse, this result does not lead to any conclusive results as to the best time for a stud to sire a horse, but it could imply younger sires result in more athletic horses. As the other variables are categorical, their relationships with the predictions are more difficult to interpret, and a more thorough analysis will have to be done in future work.

The mean height of the fences in the classes jumped contributed the most to the predictions made by the model trained on all the horse features, implying that the horses who compete consistently on the higher level have higher rankings (Figure D2). The surprising part of this result is that the mean height is even more important than the placing related features, so, for the purpose of climbing the rankings, it would be better for athletes to sacrifice their placings in favor of competing in higher classes. The model also found the counterintuitive correlations between jump off mean time and first round mean time discussed earlier. Additionally, break time is the third most important feature, also displacing other features related to placings in classes and emphasizing the importance of frequent competition.

The athlete details dataset consists of only four features, which in order of importance are: number of horses, competing for, age and gender (Figure D3). Higher numbers of horses result in a significant improvement in the predicted ranking, but there is a less clear pattern for the athlete's age, so its impact varies significantly based on its interactions with other features. The model has also learned strong patterns relating to the country the athlete is representing, but

more detailed analysis would have to be done to determine these patterns. The gender is divided into two clear categories, which, as discussed earlier, is likely male and female.

Notably, in the model trained on all the athlete features, the two most important variables, standard deviation of break time and number of competitions, are not related to the placings and therefore introduce no data leakage (Figure D4). Analysis of these two variables reveals only the expected results: lower standard deviations of the break time and a higher number of competitions are rewarded. In fact, all the remaining relationships follow an expected pattern or have been discussed previously.

The model trained on the a priori data is noteworthy because it is the first model to incorporate information about both the horse and the athlete (Figure D5). As in the previous model, the standard deviation of the break time for the athlete is once again the most influential feature. However, the next four features are related to the horse: break time, standard deviation of break time, administering national federation, and age. The most conspicuous fact about these features is that the impact of the horse's age has changed drastically from the model trained on all horse results, with its impact on the model value almost reversing. The remaining features are all related to the athlete and have similar contributions to those described earlier. This fact means that six of the top ten features are related to the athlete and four are related to the horse. However, it does not imply that the athlete is more important to the success of the pair than the horse or vice versa. Because the athlete chooses the horse, factors represented in the athlete's data, such as experience measured by the number of horses that athlete has ridden, may account for unknown details about horses, such as carefulness and speed.

The model trained on all the features follows a similar pattern to those described above, except the mean height of the classes rises to new prominence as the top feature. These results

include a large amount of data, but what does it all mean? The data verifies that practice and experience are key features in determining skill. Riders that want to increase their place in the rankings should focus on competing in higher classes, even at the expense of more time and faults, and competing frequently and consistently. Additionally, it may be more beneficial to focus on placing consistently, instead of risking everything to earn a place on the podium, and ending up with no points, as the percentage of times in the top eight is a much more important feature than the percentage of times in the top three.

Partial Dependence Plots

SHAP not only outputs the mean absolute value of the contribution of a feature to the predictions but also determines how much each value of that feature contributes to the prediction and records feature interactions. These relationships can be viewed with SHAP partial dependence plots.

Examples of these relationships are included in Appendix E. The SHAP dependence plot for age shows a distinct U-curve with a minimum around 4000 days or about 11 years old. This shape reflects the tradeoff between athleticism and experience for horses in show jumping (Figure E1). The combination of both these factors peaks when a horse is about 11 years old, leading the model to place horses at this age higher in the rankings. The feature that most influences predictions based on horse age is the number of ownerships. The relationship is not easily apparent, and is likely not statistically significant, but it appears that for older horses, horses that have switched hands more times are more likely to be predicted lower in the ranks. Intuitively, this relationship could occur because more athletic horses are more likely to generate more demand and have more people willing to buy as they demonstrate their athleticism, and more people willing to sell as they increase in value.

The color of the horse also produces an interesting dependence plot (Figure E2). Although logically, the color of the horse has no impact on performance, the machine learning model which explains a significant amount of variation of out-of-sample learned a relationship between color and performance, so perhaps it could be a proxy for other genetic traits. The model identifies black horses as the best performing and bay horses as the worst. This fact, although on its own not a solid conclusion that irrefutably improves the superiority of certain colors, is an interesting observation. The color of the horse most interacts with the age of the horse, probably because age is one of the most important predictors in the horse details performance, but it is hypothetically possible that different colored horses tend to come from different breeds which mature at different speeds or tend to come from countries that have different horse development timelines.

The dependence plot for the mean height jumped by the horse shows a distinct reverse S-curve that clearly captures the importance of the different heights to the rankings (Figure E3). The lower classes result in worse predictions, but the higher classes result in better predictions. The most important feature interaction occurs with the round one mean faults, implying that the model learned first to sort horses into their division and then learned to separate horses within their division by one of the most important variables related to their scores.

Analysis of SHAP Predictions

It is constructive to consider in detail how SHAP values interact in interpreting the predicted values of the model. An analysis of the SHAP values for the entry with the lowest predicted rank fulfills this objective while also offering insights into the type of features to consider when evaluating and training a horse (Appendix F). The model starts at the bottom by predicting the expected value of the target for the dataset and then modifies its prediction

according to each feature in an additive manner. For example, the athlete standard deviation of break time is 3.01, leading the model to subtract -360.29 from its prediction. These waterfall plots allow model interpretability for individual predictions. This example clearly demonstrates the dominating effect of certain features when making predictions.

Future Work

This research found that even with data that seems to have little predictive power using linear models, the rank of a horse can be predicted to such a degree that the actual rank improves significantly. In the future, the performance of this model could be improved by using more data. Naturally, the first step would be to include the performance of any horses in the horse's pedigree, but further steps may even require the collection of a new dataset with different variables that are not collected by the FEI, such as temperament, demeanor, training program, and even factors like diet. However, the simplest way to extend the study would be further analysis of the feature interactions to determine what other factors contribute either positively or negatively to the rankings. This problem can also be scaled up or down, depending on the use case. For example, the most serious competitors have much less separation from each other than someone who competed in a couple of 1.45-meter classes and is at the bottom of the rankings. These riders also are the ones who would likely be interested in gaining any edge they can in any way possible, including machine learning. It would therefore be advisable to train a model on only these athletes and include more personally relevant data for each sample. Additionally, the problem could be scaled into a classification problem of whether a horse competes at this high of a level to determine the factors that lead to such a high-level of performance in classes. In addition to these reformulations of the problem, machine learning also has other applications in

the horse industry, including prediction of falls and injuries, optimal pairings of horses and riders, and even applications of computer vision to analyze a horse's jump.

References

About FEI. (n.d.). FEI. https://inside.fei.org/fei/about-fei

Cooke, G. (2013). *International movements of competition horses* [PowerPoint slides]. FEI. https://inside.fei.org/system/files/1_Trends_in_Growth_of_Equestrian_Sport_GCO.pdf

Dillon et al. (n.d.) *Be careful when interpreting predictive models in search of causal insights — SHAP latest documentation*. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%C2%A0insights.html

FEI jumping rules. (2023). FEI. https://inside.fei.org/sites/default/files/Jumping_Rules_2023_ clean_ BoardResolutionJan.pdf

FEI website terms and conditions. (n.d.). FEI. https://www.fei.org/terms-conditions

Hanousek, K., Salavati, M., & Dunkel, B. (2020). The Impact of Horse Age, Sex, and Number of Riders on Horse Performance in British Eventing Horse Trials. *Journal of Equine Veterinary Science,* 94, 103250. https://doi.org/10.1016/j.jevs.2020.103250

Marlin, D., & Williams, J. (2020). Faults in international showjumping are not random. *Comparative Exercise Physiology, 16*(3), 235–241. https://doi.org/10.3920/CEP190069

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features* (arXiv:1706.09516). arXiv. https://doi.org/10.48550/arXiv.1706.09516

Rules for the Longines rankings. (2020). FEI. https://inside.fei.org/system/files/Longines Ranking_Rules_2020.pdf

Visser, E. K., Van Reenen, C. G., Engel, B., Schilder, M. B. H., Barneveld, A., & Blokhuis, H. J. (2003). The association between performance in show-jumping and personality traits

earlier in life. *Applied Animal Behaviour Science, 82(4)*, 279–295. https://doi.org/10.1016/S0168-1591(03)00083-2

Welker, V., Stock, K. F., Schöpke, K., & Swalve, H. H. (2018). Genetic parameters of new comprehensive performance traits for dressage and show jumping competitions performance of German riding horses. *Livestock Science*, *212,* 93–98. https://doi.org/10.1016/j.livsci.2018.04.002

What is machine learning? (n.d.). IBM. https://www.ibm.com/topics/machine-learning

Appendix A

Description of Datasets

Table A1

Description of Horse Details

Feature	Description
Sex	The sex of the horse. Stallion, gelding, or mare
Castrated/Sterilized	Whether the horse is castrated or sterilized. Yes, no, or unknown
Breed	The breed of the horse
Number of Ownerships	The number of owners the horse has had
Administering NF	Administering National Federation
Issuing NF	Issuing National Federation
Age	The age of the horse in days
Birth Month	A discretized numbering of the horse's birth month
Color	The color of the horse
Sire	The sire of the horse
Sire Age	The age of the horse's sire
Sire Color	The color of the horse's sire
Dam	The dam of the horse
Dam Age	The age of the horse's dam
Dam Color	The color of the horse's dam
Sire of Dam	The sire of the horse's dam
Dam's Sire Age	The age of the dam's sire
Dam's Sire Color	The color of the dam's sire
Sire's Sire Age	The age of the sire's sire
Sire's Sire Color	The color of the sire's sire
Dam's Dam Age	The age of the dam's dam
Dam's Dam Color	The color of the dam's dam
Sire's Dam Age	The age of the sire's dam
Sire's Dam Color	The color of the sire's dam

Table A2

Description of Athlete Details

Feature	Description
Gender	The gender of the athlete
Competing for	The country the athlete represents
Number of horses	The number of horses the athlete has ridden in FEI competition
Age	The age of the athlete in days

Table A3

Description of Horse Performance and Athlete Performance

Feature	Description
Round 1 Mean Time	The average time taken to complete the first round
Round 1 Mean Faults	The average number of faults incurred in the first round
Jump Off Mean Time	The average time taken to complete the jump off
Jump Off Mean Faults	The average number of faults incurred in the jump off
Elimination Percent	The percentage of the last 50 classes that were not completed
Break Time	The average time in days between competitions
Std Break Time	The standard deviation of the time in days between competitions
Mean Place	The average place earned in competition
Median Place	The median place earned in competition
Max Place	The place earned in the worst performance
Min Place	The place earned in the best performance
Percent in Top 3	The percentage of classes that end with a place in the top 3
Percent in Top 8	The percentage of classes that end with a place in the top 8
Mean Height	The mean height of jumps in the last 50 classes
Percent 1.60	The percentage of the last 50 classes that are 1.60 meters

Appendix B

Exploratory Data Analysis

Figure B1

Horses by Sex

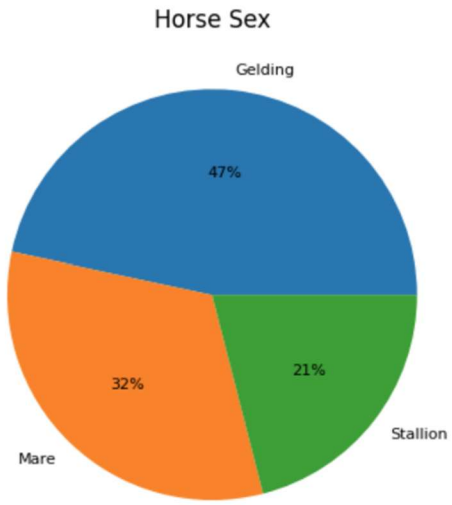


Figure B2

Horses by Breed

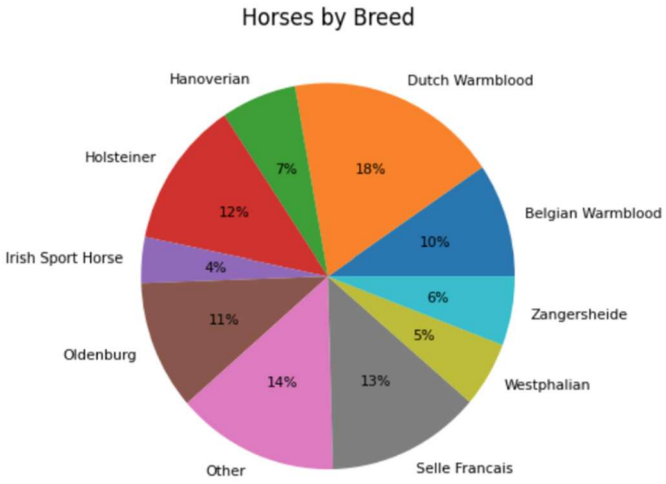


Figure B3

Horse Age vs. Rank

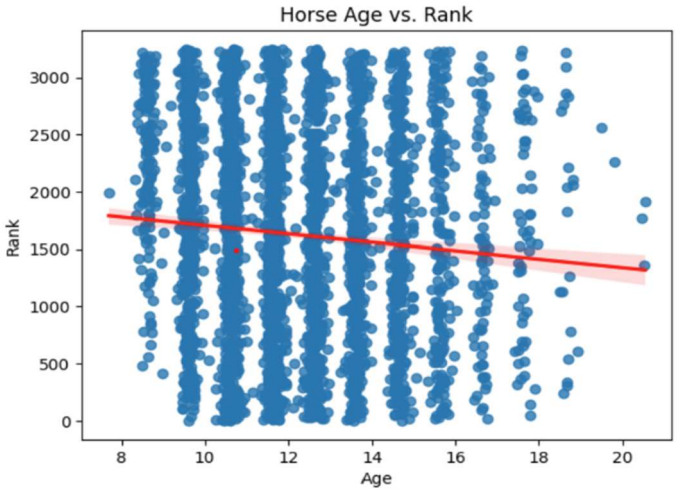


Figure B4

Number of Horses vs. Rank

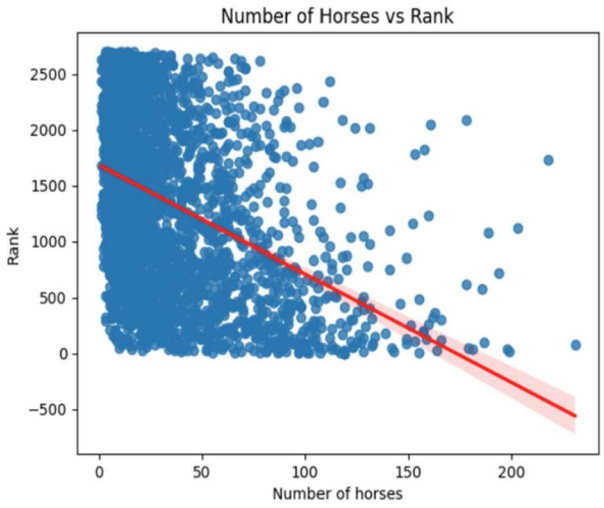


Figure B5

Athletes by Gender

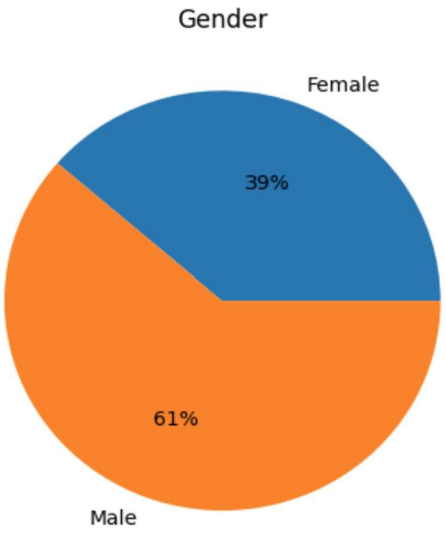


Figure B6

Athletes by Country

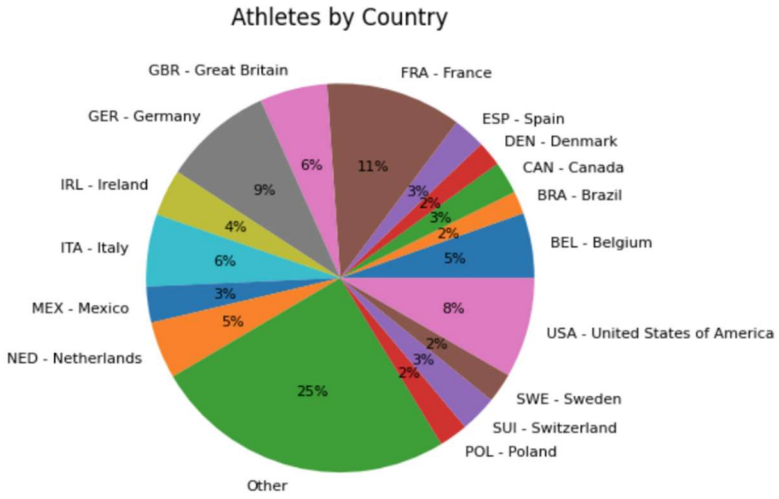


Figure B7

Horse Number of Competitions vs. Rank

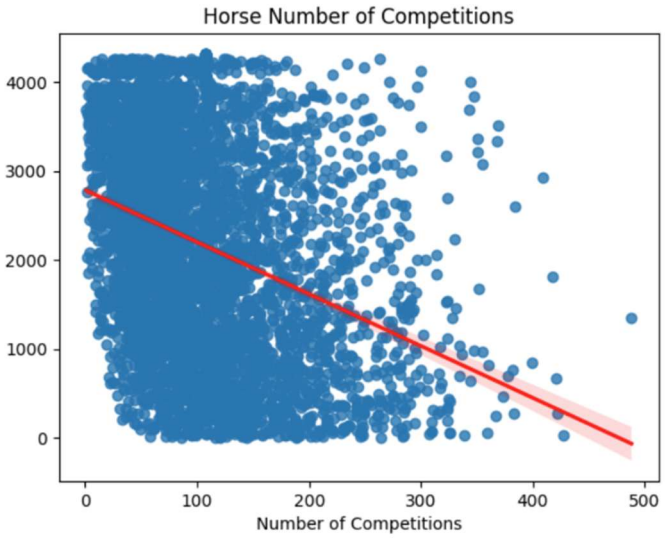


Figure B8

Athlete Number of Competitions vs. Rank

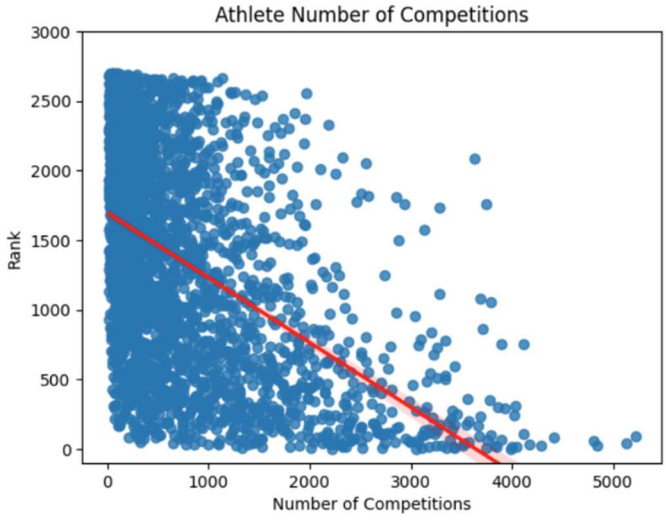


Figure B9

Horse First Round Mean Time vs. Rank

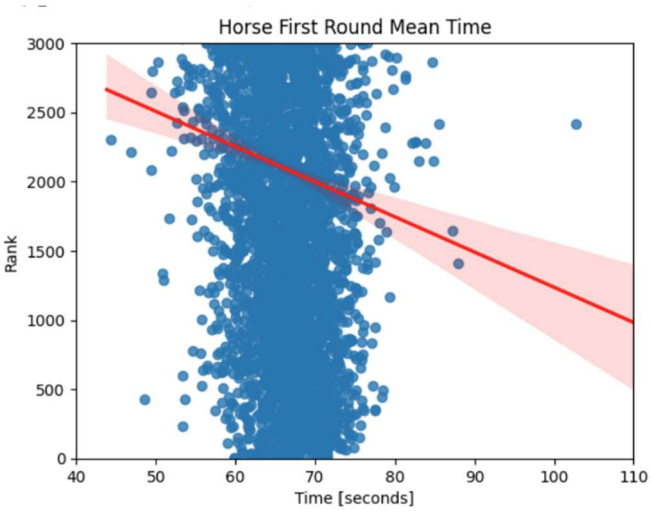


Figure B10

Athlete First Round Mean Time vs. Rank

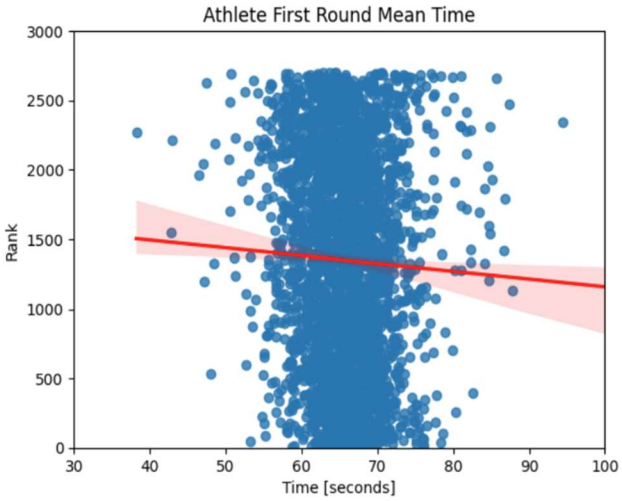


Figure B11

Horse Jump Off Mean Faults vs. Rank

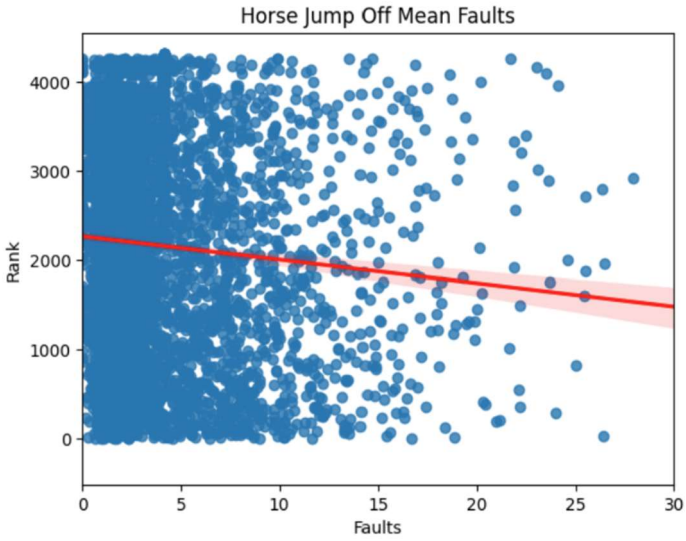


Figure B12

Athlete Jump Off Mean Time vs. Rank

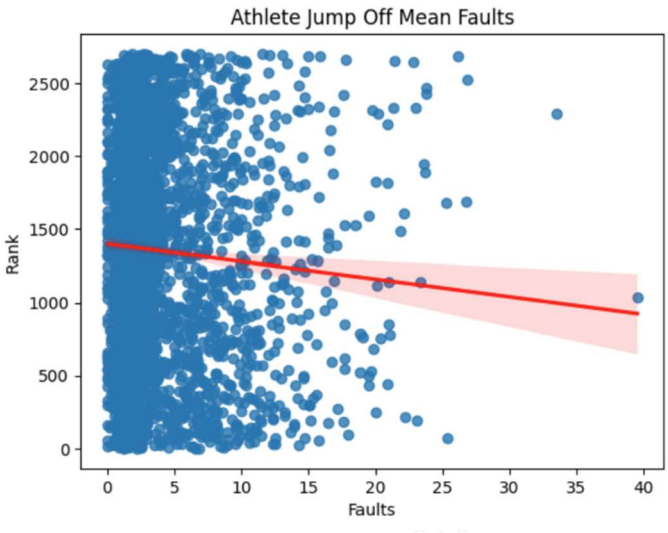


Figure B13

Horse Jump Off Mean Time vs. Rank

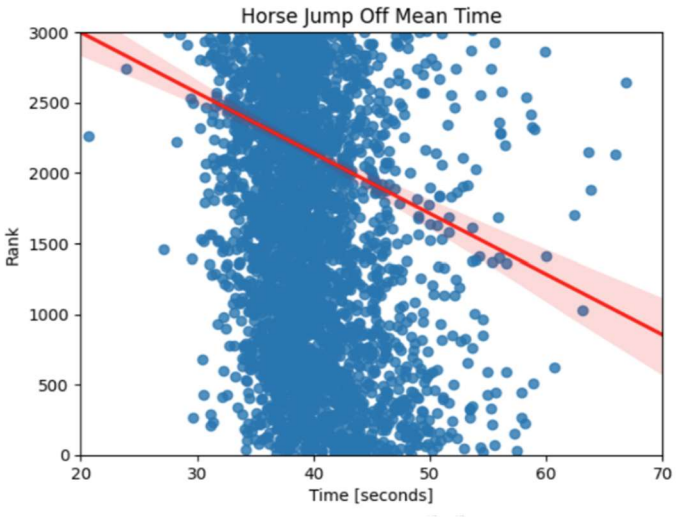


Figure B14

Athlete Jump Off Mean Time vs. Rank

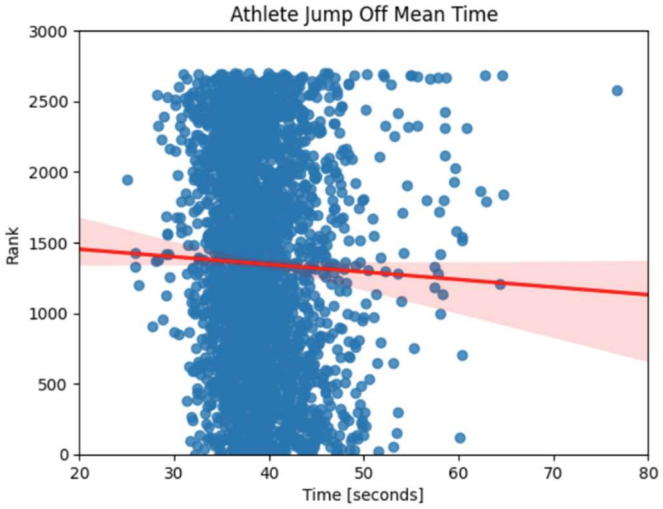


Figure B15

Horse Break Time vs. Rank

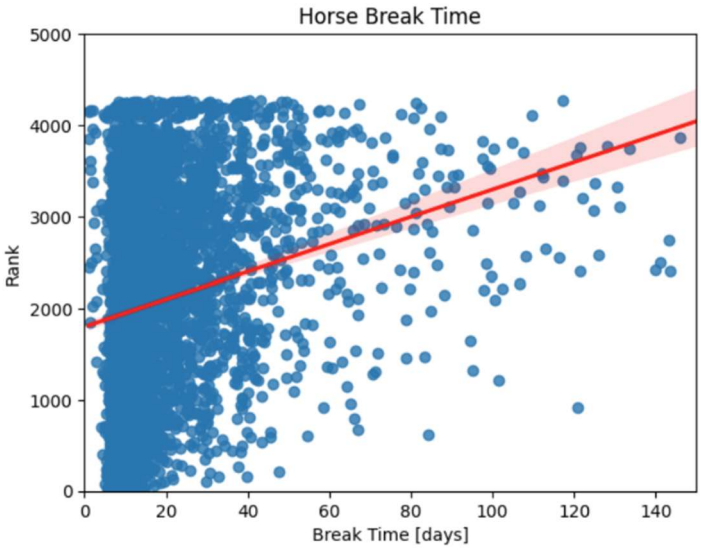


Figure B16

Athlete Break Time vs. Rank

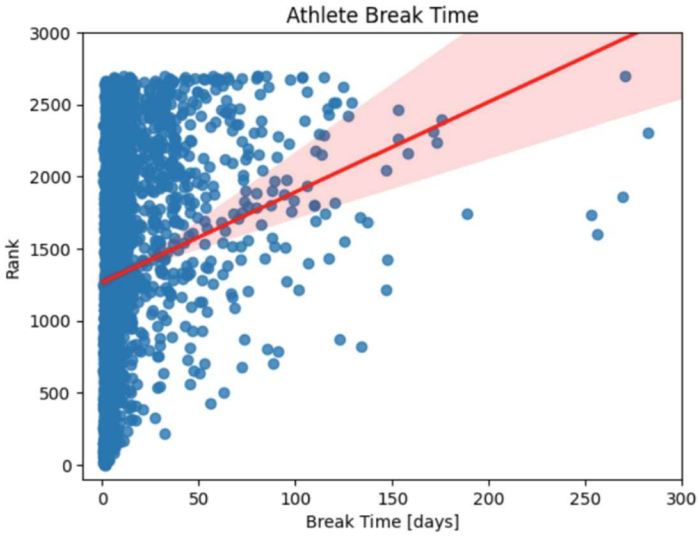


Figure B17

Horse Standard Deviation of Break Time vs. Rank

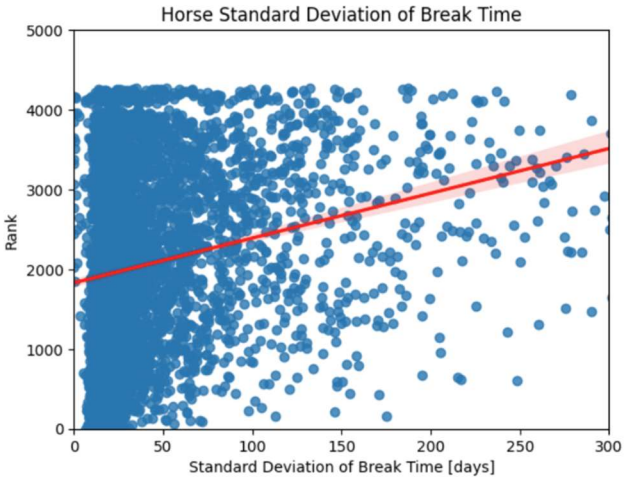
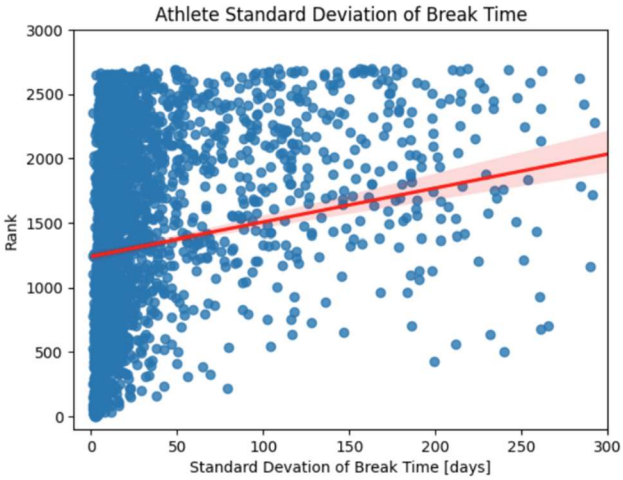


Figure B18

Athlete Standard Devation of Break Time vs. Rank



Appendix C

Machine Learning Results

Figure C1

Rank vs. Predicted Rank Horse Details

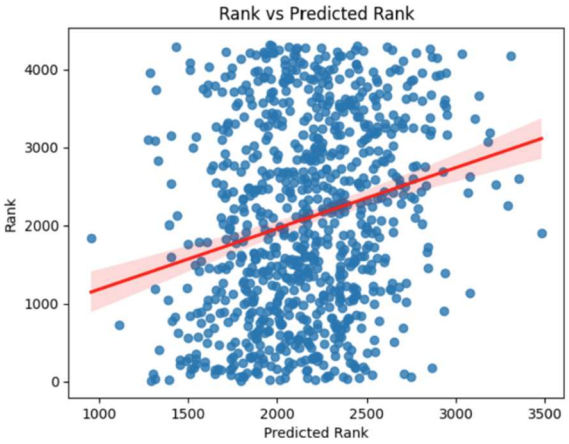


Figure C2

Rank by Predicted Decile Horse Details

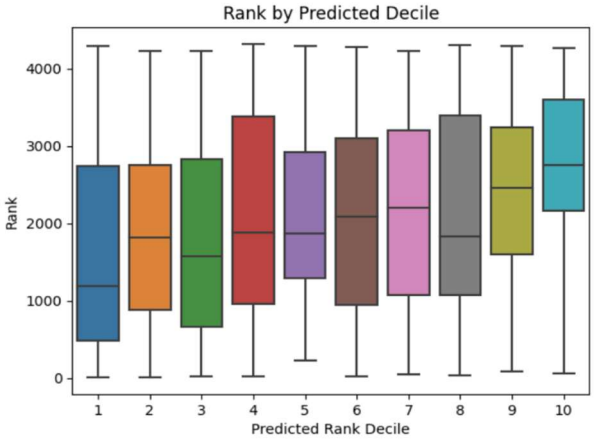


Figure C3

Rank vs. Predicted Rank All Horse Features

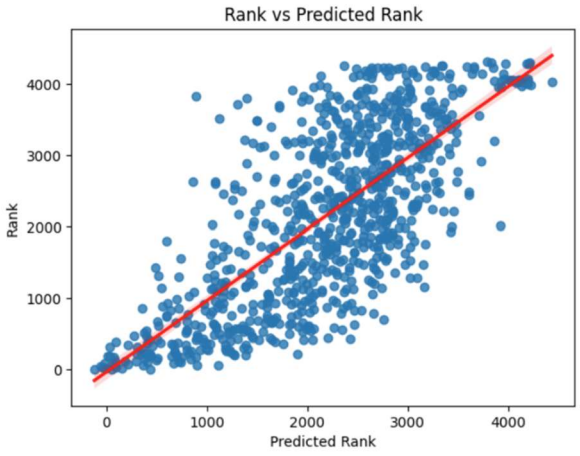


Figure C4

Rank by Predicted Decile All Horse Features

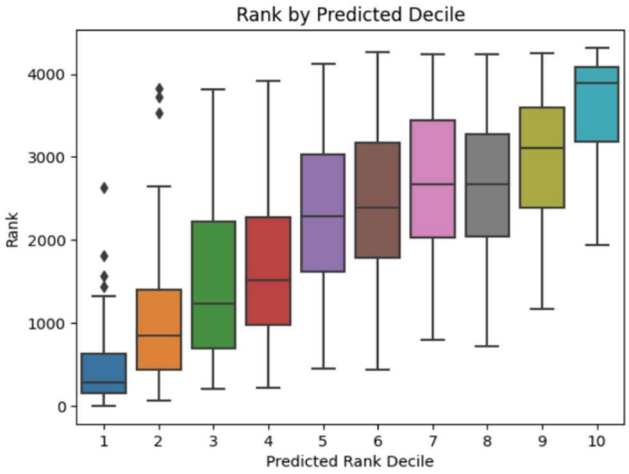


Figure C5

Rank vs. Predicted Rank Athlete Details

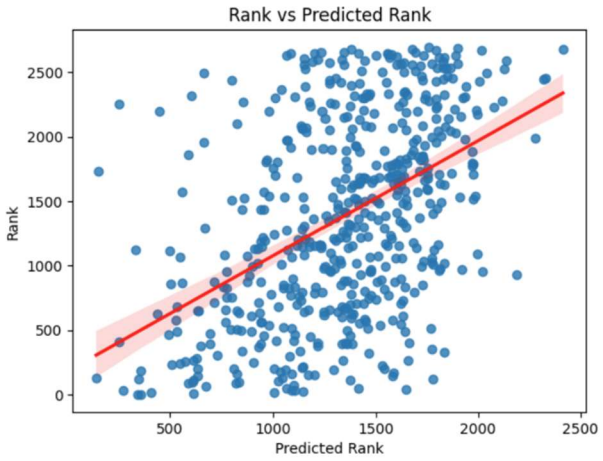


Figure C6

Rank by Predicted Decile Athlete Details

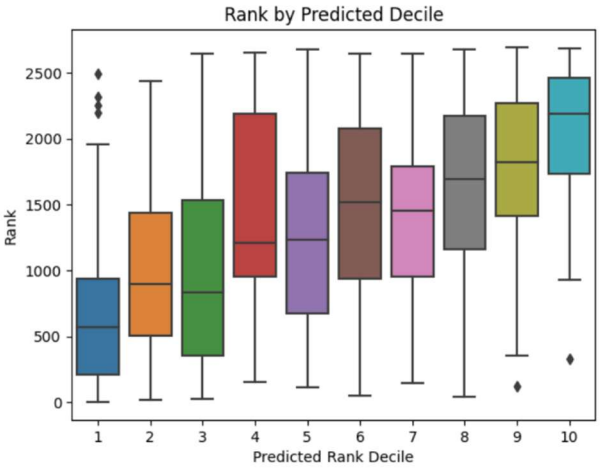


Figure C7

Rank vs. Predicted Rank All Athlete Features

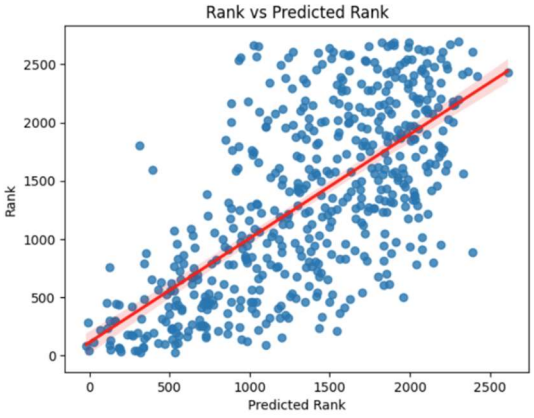


Figure C8

Rank by Predicted Decile All Athlete Features

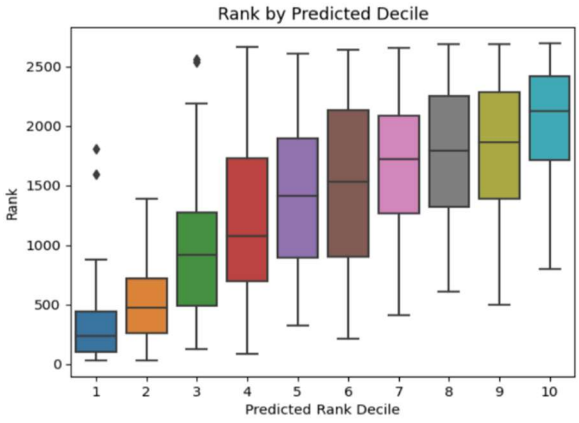


Figure C9

Rank vs. Predicted Rank A Priori Features

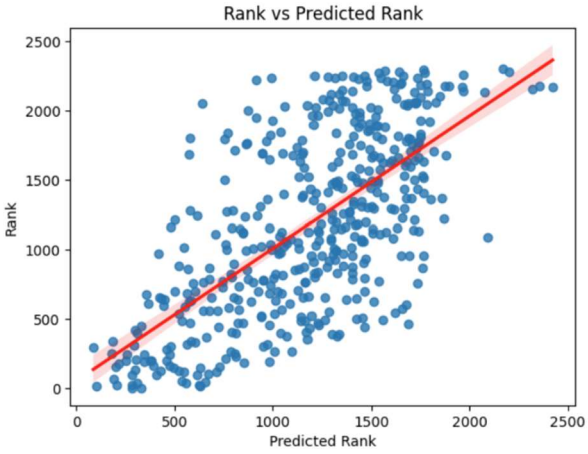


Figure C10

Rank by Predicted Decile A Priori Features

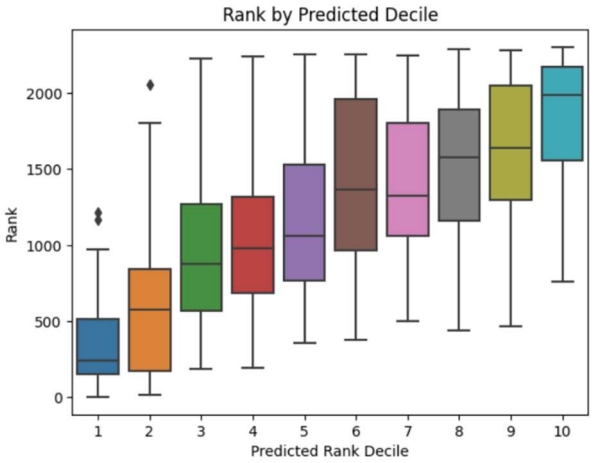


Figure C11

Rank vs. Predicted Rank All Features

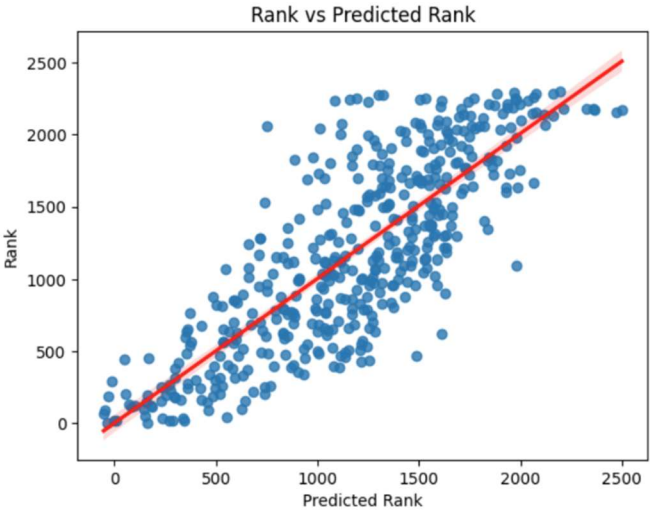
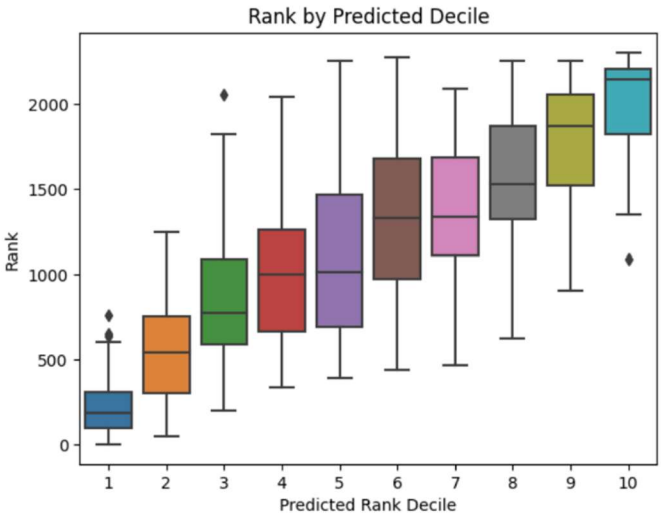


Figure C12

Rank by Predicted Decile All Features



Appendix D

Figure D1

SHAP Values for Horse Details Model

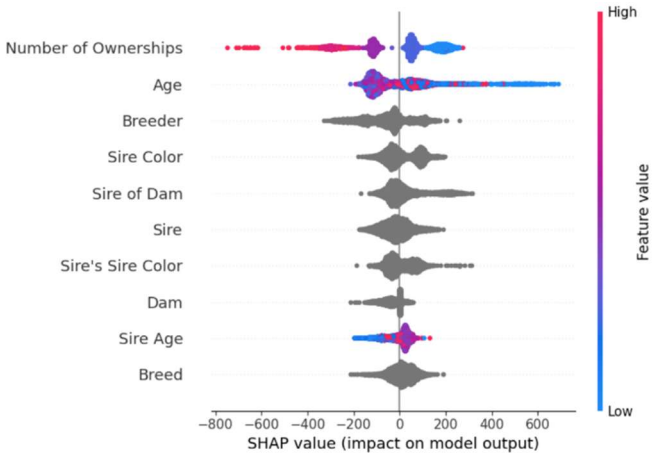


Figure D2

SHAP Values for All Horse Features Model

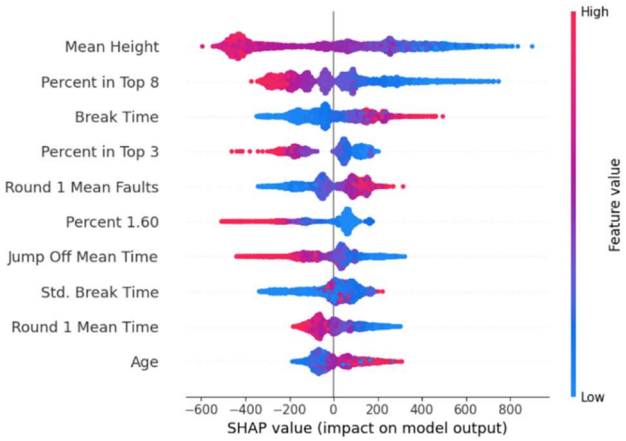


Figure D3

SHAP Values for Athlete Details Model

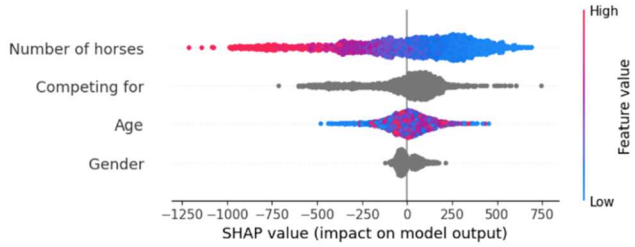


Figure D4

SHAP Values for All Athlete Features Model

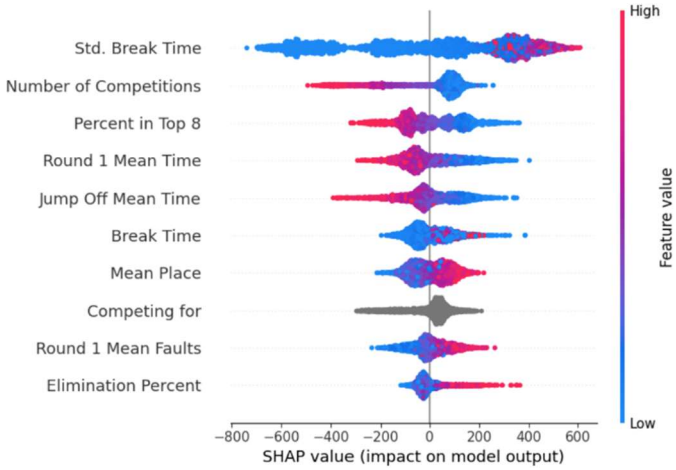


Figure D5

SHAP Values for A Priori Features Model

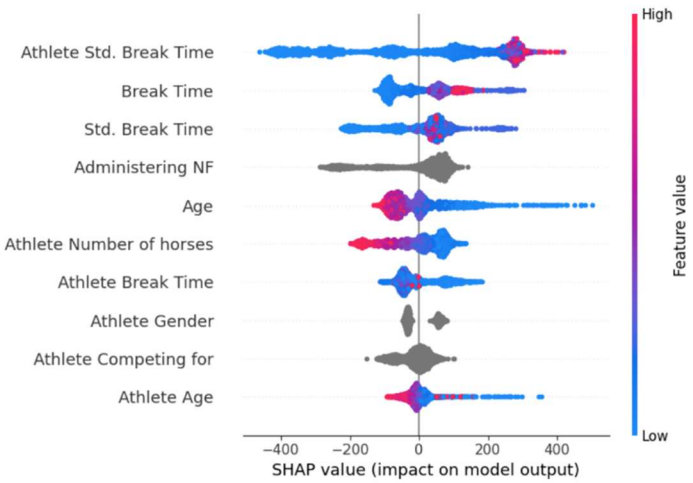
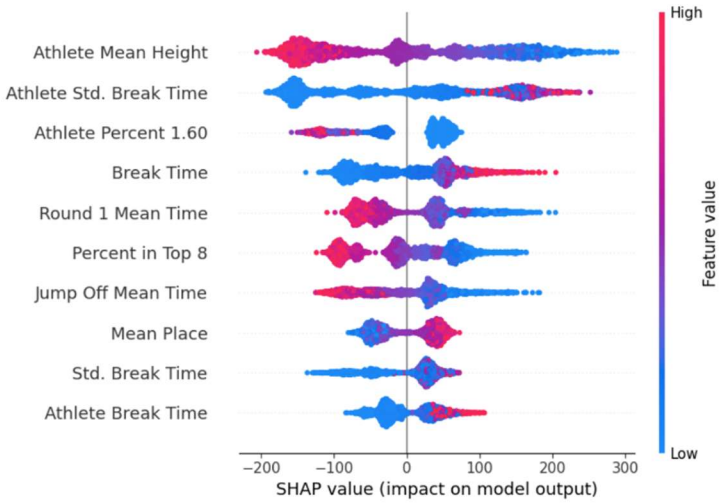


Figure D6

SHAP Values for All Features Model



Appendix E

SHAP Dependence Plots

Figure E1

SHAP Dependence Age [days] and Number of Ownerships

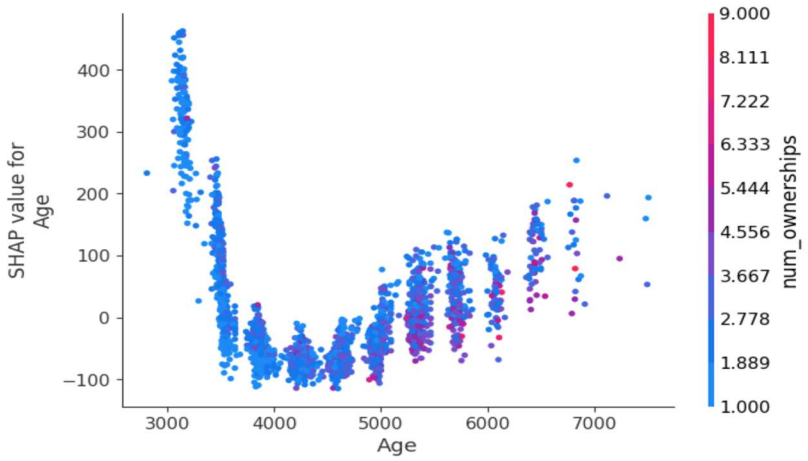


Figure E2

SHAP Dependence Plot for Color and Age [days]

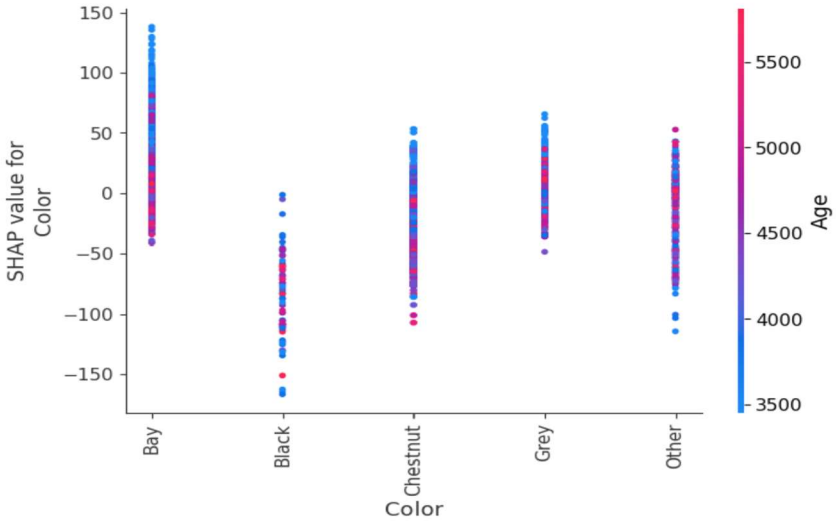
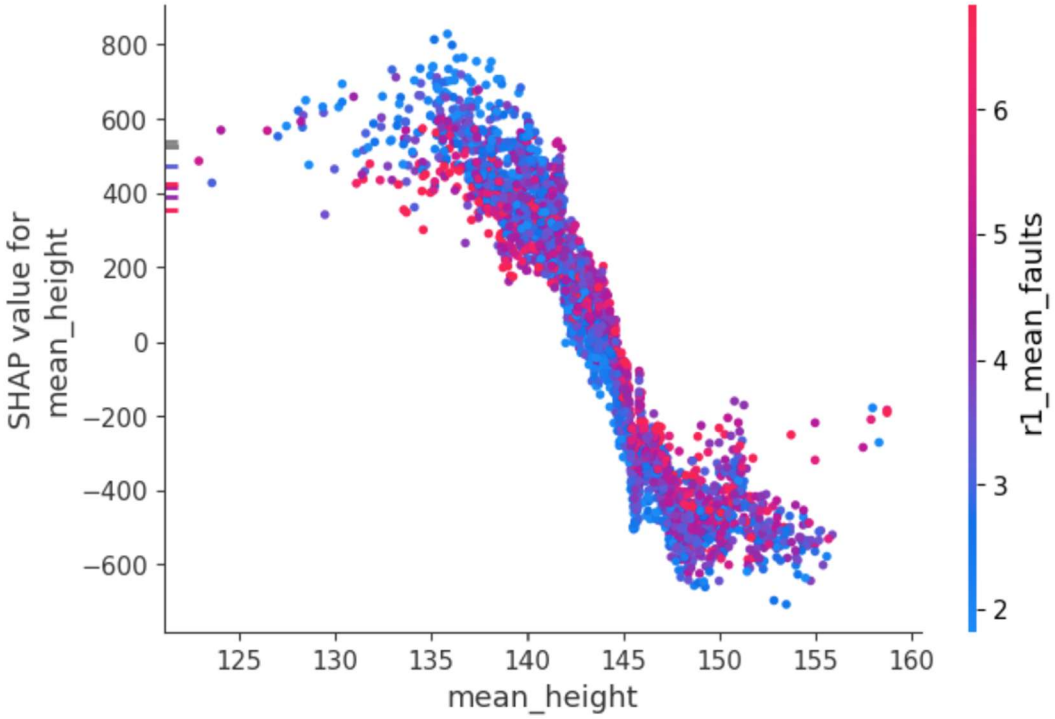


Figure E3

SHAP Dependence Plot for Mean Height [cm] and Round 1 Mean Faults



Appendix F

SHAP Waterfall Plot

Figure F1

Waterfall Plot of Lowest Prediction