WELLESLEY

Factors Associated with a Racing Greyhound's Career Length

Adelle Wang '24, Data Science Major Capstone

Background and Research Question

Greyhound racing has been on the decline in the United States for decades, now being banned in 43 states, but is still a hotly competitive arena for sports-type gambling. Tens of thousands of greyhounds are bred for their speed every year, and many organizations such as Grey2K USA or GreySave are in a constant race to adopt out retired greyhounds as they come in from the track [1].

A greyhound's typical career length can range from 1.5 to 3.5 years. They begin rigorous training a few months after their first birthday, and are raced until they are injured or decline in performance [2]. This capstone project seeks to identify potential factors and fit a model that may predict the length of a racing greyhound's career, information that could be used to minimize greyhound injuries and retire racers earlier in favor of their health and longevity.

This study utilizes data collected from the racing greyhound database Greyhound-Data to fit first order models with the following research question:

What factors/model is the most accurate for predicting the length of a racing greyhound's career?

Data Collection and Cleaning

The data was extracted from the simple statistics section of the Greyhound-Data.com database [3]. A range of years was chosen to bypass the sharp decline in racing created by the shutdown of many racing tracks due to COVID-19: the total dataset was taken from 2016-2020. The data consists of the 100 top dogs from each year and their statistics, resulting in a raw dataset of 500 dogs with 15 variables.

Once the data was collected, the dataset was cleaned by removing duplicate dogs based on their names, where certain dogs may have been in the top 100 across multiple years. A total of 40 rows were duplicates and removed, resulting in 460 observations left over. The variables of interest identified were **CareerRaces, Wins (that year), Win Percentage (that year), Win Distance (avg across races), Top 3 Placements (that year), and Sex.**

In addition, the Win Percentage variable was adjusted to contain the decimal value instead of a percentage, and the Win Distance was converted to a numerical value by removing the "m" units.

Data Exploration

The quantitative predictor variables were first checked for multicollinearity; when no multicollinearity was detected, a scatterplot was created with each predictor plotted against the response, CareerRaces. Win percentage and Top 3 had the strongest linear relationship with career races, with correlation coefficients of -0.43 and 0.38 respectively (Fig 1).

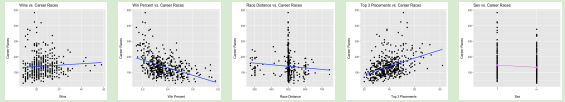

Figure 1. Scatterplots of predictor variables vs. Career Races response variable.

The individual linear models were also fitted and linear model assumptions were verified. Quantile-quantile plots were created to assess normality of the random errors, and the residuals were plotted against each predictor to assess homoscedasticity. The qqplots show that the residuals generally follow a 45 degree line, suggesting that the normality assumption is met. The plot of residuals vs. predictors shows no obvious trends, apart from a slight curve in the plots for WinPercent and Top3; a Box-Cox transformation was done on both variables to examine this further (Fig 2). The log transformation did not appreciably improve the adjusted R^2 values of either model, so the transformation was discarded in favor of comparability and interpretability.

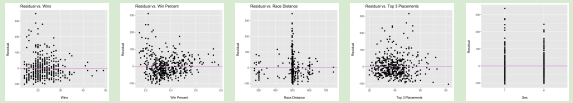

Figure 2. Residuals plotted against predictor variables to assess homoscedasticity.

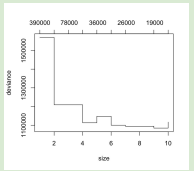

Figure 3. Plot of CV results for full classical tree model.

Model	SSR
Full Linear	282775.9
AIC Linear	294516.2
Pruned Tree	253818.1
Random Forest	239812.6

Table 1. SSR Values for all compared models.

Linear Data Models

The factors used to assess model performance and determine the optimal linear regression model include: adjusted R-squared value, F-statistic, and the sum of squared residuals. The dataset was randomly partitioned into two subsets, a training set and a validation set, with 80% of the observations to fit a model, and then validate the fitted model on the remaining 20% of the observations. The full model was fitted, and AIC and BIC criterion were used to find the AIC model and BIC model.

The full model had an adjusted R^2 value of 0.31 and an F-statistic value of 32.32 (p-value < 0.0001).

$$\hat{CareerRaces} = 162.46 - 7 * Sex(m) + 0.9 * Wins - 251.9 * WinPercent - 0.04 * WinDist + 1.7 * Top3$$

The AIC and BIC model returned the same model, with an adjusted R^2 value of 0.3 and an F-statistic value of 79.25 (p-value < 0.0001).

$$\hat{CareerRaces} = 140.11 - 229.5 * WinPercent + 2.1 * Top3$$

On the testing data, the full model had an SSR value of 282775.9. The AIC (and BIC) model had an SSR value of 294516.2, meaning that the full model was slightly better, although the difference of 11740.3 is hardly significant, being less than 5% of either SSR value (Table 1).

Ensemble Tree-based Models

A full tree was first built and 10-fold cross-validation was used to prune the tree and identify the optimal number of terminal nodes, which in this case is 9 terminal nodes according to the plot of the CV tree results. The pruned tree was then built with the variables Wins, WinPercent, Top3, and WinDist. The pruned tree gave an SSR value of 253818.1 on the testing data (Table 1).

A random forest model was also fitted in the hopes that it would be robust against overfitting to the data due to its combination of multiple trees. The random forest model gave an SSR value of 239812.6 on the testing data (Table 1).

Results/Discussion

Based on the data exploration and scatterplots plotted for each variable, it appears that the variables for Win Percentage by year and placement in the Top 3 of each race are the most significantly correlated to the response variable, Career Races. Based on the SSR values and adjusted R^2 values of the models tested, the full linear regression was the best linear model for predicting the data and the random forest model had the best sum of squared residuals over all the models.

The results of Win Percentage and Top 3 being highly correlated may demonstrate that a generally successful dog will have a longer career, but a dog pushed to place 1st may need to retire earlier due to health decline.

References

[1] "Greyhound Racing in the United States." GREY2K USA Worldwide, www.grey2kusa.org/about/states.php.
[2] "How Long Is the Average Greyhound Racing Career?" Towcester Racecourse & Leisure, 16 Oct. 2023.
[3] Statistics - Greyhound-Data, www.greyhound-data.com/statistics.htm.