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 questions:

Q1: What factors are the most highly correlated with length of a racing greyhound's career?

Q2: Which 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.

Raw data: 100 top dogs from each year and their statistics -> 500 dogs with 11 variables

Data cleaning:

- Removing duplicate dogs across years based on names
  - 40 duplicates removed -> 460 observations left
- Variables of interest:
  - **CareerRaces, Wins (that year), Win Percentage (that year), Win Distance (avg across races), Top 3 Placements (that year), and Sex**
  - Variables relating to name, lineage, and variables that would be duplicate information (# 1st, # 2nd, # 3rd place, etc.) removed
- Win Percentage: decimal instead of percent
- Win Distance: numerical instead of with "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).

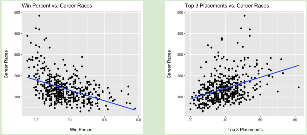

Figure 1. Scatterplots of predictor variables Win Percent and Top 3 Placements vs. Career Races response variable.

Individual linear models fitted -> linear model assumptions met

Quantile-quantile plots created to assess normality: follow 45 deg. line, normality assumption met

Residuals plotted against predictors: no obvious trends, slight curve in plots for WinPercent and Top3 (Fig 2)

- Box-Cox transformation done on both variables to examine this further
- Log transformation did not appreciably improve the adjusted R^2 values of either model -> transformation discarded in favor of comparability and interpretability

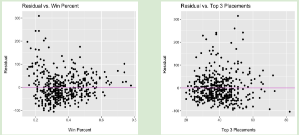

Figure 2. Residuals plotted against Win Percent and Top 3 Placements to assess homoscedasticity, plots show slight curve pattern.

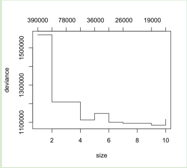

Figure 3. Plot of CV results for full classical tree model.

Model	SSR	F-scores
Full Linear	282775.9	6.980
AIC Linear	294516.2	5.835
Pruned Tree	253818.1	10.258
Random Forest	239812.6	12.127

Table 1. SSR and F-score Values for all compared models.

Linear Data Models

Factors for model assessment: adjusted R-squared value, F-score on training data, sum of squared residuals (SSR), F-score on testing data

Partitioning: 80% training, 20% testing

Models: Full model, AIC, BIC

Ful model: had an adjusted R^2 value of 0.31 and an F-score (training) value of 32.32 (p-value < 0.0001).

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

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

$$Caree\hat{r}Races = 140.11 - 229.5 * WinPercent + 2.1 * Top3$$

On the testing data, the full model had an SSR value of 282775.9 and F-score of 6.98. The AIC (and BIC) model had an SSR value of 294516.2 and F-score of 5.835.

-> Full model was slightly better, but difference is minimal(Table 1).

Ensemble Tree-based Models

Full tree built -> 10-fold cross-validation used to prune the tree and identify the optimal number of terminal nodes = 9 (Fig. 3) -> pruned tree was then built

Variables in pruned tree: Wins, WinPercent, Top3, and WinDist

The pruned tree gave an SSR value of 253818.1 and F-score of 10.258 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 and F-score of 12.127 on the testing data (Table 1).

Results/Discussion

Q1: **Win Percentage (yearly)** and placement in **Top 3** of each race are most significantly correlated to response variable, Career Races (Fig 1)

- 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.

Q2: Full linear model is the best *linear* model, but **Random Forest** is best model overall based on lowest SSR and highest F-score on testing data

Future directions:

- Testing more variation of models, with larger dataset to have more robust findings
- Using RF model to predict on newly collected data

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.