Descriptive Analysis of Powerlifting Competition Data and Predicting Max Deadlift of Competitors

Project Report

April 2023

Kevin Cho

1 ABSTRACT

Powerlifting meets are held worldwide where competitors attempt their 1 repetition max in the squat, bench, and deadlift. I explored if variables such as gender, age, bodyweight, equipment, and max squat/bench could predict a competitor's max deadlift. An R-squared value of 0.89 was achieved with linear regression, suggesting reasonable fit and predictive power. Variable selection and Ridge did not improve the prediction performance. The best prediction performance was achieved using Random Forest, a non-linear method. Interesting inferences were made by interpreting the coefficients of the linear regression model. Biological males possibly have a have an advantage over females in deadlift (when all other variables are equal), but this would need to be confirmed by further biomechanical and statistical research.

2 INTRODUCTION

Powerlifting is a sport where competitors attempt their 1 repetition maximum strength for three barbell lifts: the squat, bench, and deadlift. Powerlifting meets (competitions) are held around the world by different federations and results of most meets are available online. The data includes information such as a competitor's gender, age, bodyweight, equipment used, and max successful attempt for each lift. I am interested in exploring if these variables can be used to predict a competitor's max deadlift. If a model with good fit is achieved, a competitor at a given gender and age could use it to roughly estimate what they need to do reach their goal deadlift (e.g. increase their bodyweight, increase their squat, etc.).

My approach is to compare various linear regression models using Monte Carlo Cross Validation. I will attempt some variable selection techniques such as LASSO, stepwise, and best subsets. This could help answer the question of which variables are most important for predicting a competitor's max deadlift. The regression coefficients could be used to perform a descriptive analysis (e.g. how much does a 10kg increase in bodyweight affect a person's deadlift, holding all other variables constant). I also plan to explore other techniques such as Ridge and Decision Tree Regression to see if they can improve the predicting power.

I hypothesize that a competitor's max squat is going to be the most important variable for predicting their max deadlift. I believe a competitor's squat would be a reasonable indication of their strength/training level and the squat movement uses many of the same muscles as the deadlift; therefore, I expect a strong correlation between a competitor's squat and deadlift. I expect that age may not be a significant predictor when used in a full model, i.e, if a 20 and 40-year-old can squat and bench the same amount, I would expect them to deadlift a similar amount as well. Younger competitors are likely stronger on average, but age may not have much effect on prediction in the presence of the other variables.

3 DATA SOURCE AND EXPLORATORY ANALYSIS

A public-domain archive of powerlifting history is maintained by openpowerlifting.org. The raw data from 1964-2019 is available via Kaggle, the link can be found in section 8 (Credits) of this report. It includes data from over 22,000 meets and 412,000 lifters worldwide. The original dataset contains 1,048,575 rows x 37 columns. Each row corresponds to a competitor (lifter); the columns correspond to lifter performance and logistical event information. The following processing steps were performed to reduce the dataset before analysis:

1. The variable *Event* was filtered for only "SBD." This means only competitors who performed all 3 lifts (Squat, Bench, Deadlift) on the same day will be included.
2. The variable *Federation* was filtered for only "USAPL." Different federations have different rules for weigh-in time, drug testing, lift judging, etc. Restricting to a single federation make sure all competitors in the dataset are lifting under similar conditions. USAPL is the largest federation in the United States that drug tests.
3. Rows for duplicate performances were removed. This can occur if a competitor enters multiple divisions at the same event; a different row exists for each division they entered but contain identical data for the lifts they performed at that event.
4. Variables of interest were selected: *Sex, Equipment, Age, BodyweightKg, Best3SquatKg, Best3BenchKg, Best3DeadliftKg.*
5. Any rows with missing data were deleted. This can occur if a competitor fails all 3 attempts for a lift, then they will not have data in the "*Best…*" column.

The resulting processed dataset contains data from 1997-2019 and is 82,183 rows x 7 columns. The variables are described in greater detail below:

- ***Sex (***Binary): Biological gender of competitor (0=Female, 1=Male)
  - Originally categorical but converted to binary so it can be used in regression
- ***Equipment (***Binary): Lifting gear used by competitor (0=Raw, 1=Single-ply)
  - Originally categorical but converted to binary so it can be used in regression
  - Raw: Belt, knee sleeves, and wrist wraps may be used
  - Single-ply: Singlets, shirts, and knee wraps with extra support may be used in addition to Raw gear
- ***Age (***Continuous): Age in years of competitor
- ***BodyweightKg (***Continuous): Bodyweight in kg of competitor
- ***Best3SquatKg (***Continuous): Max squat in kg of competitor
- ***Best3BenchKg*** (Continuous): Max bench in kg of competitor
- ***Best3DeadliftKg*** (Continuous): Max deadlift in kg of competitor **[Response Var.]**

The relationship of the binary variables with the response variable are visualized with the boxplots in Figure 1 below. The left plot shows that the median max deadlift for males (*Sex*=1) is roughly 100kg higher than that of females (*Sex*=0). This is not a surprise since males on average have more absolute strength than females due to biological factors, but it will be interesting to see if this variable is significant in a regression model in the presence of other controlling variables such as bodyweight. The right plot shows that median max deadlift of Single-ply lifters (*Equipment*=1) is just slightly lower than that of Raw lifters (*Equipment*=0). This was a surprise to me since I expected Single-ply lifters to benefit from a gear advantage and have a higher median deadlift. This could be behavior driven, it's possible "stronger" competitors choose to compete Raw instead of Single-ply. It will be interesting to see if there is a conditional relationship that can be uncovered in a regression model with other predictors.

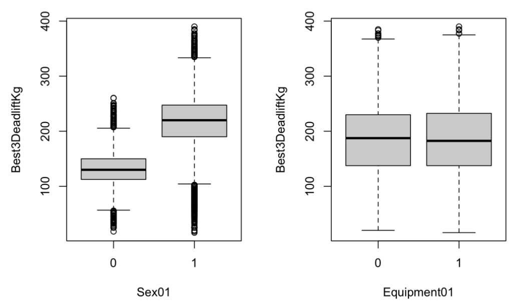


Figure 1 – Boxplots of *Best3DeadliftKg* (response var.) vs binary variables

The correlation coefficients of the quantitative variables and response were visualized into a heat map seen in Figure 2 below; the values of the coefficients can be found in Table A1 of the appendix. Deadlifts show high correlation to squat (0.92) and bench (0.89), but moderate correlation to bodyweight (0.61) and practically no correlation to age (0.01). Based on this I would not expect age to have much effect in the regression model. Squat and bench have high correlation to each other (0.91) indicating possible multicollinearity; therefore, this will be checked later with a VIF (Variance Inflation Factor) analysis.

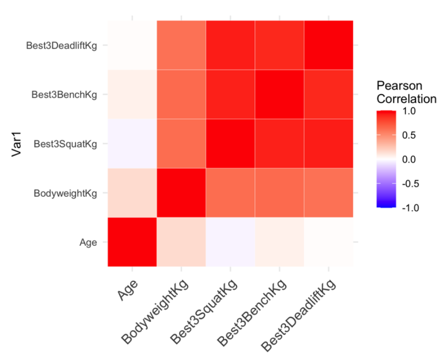
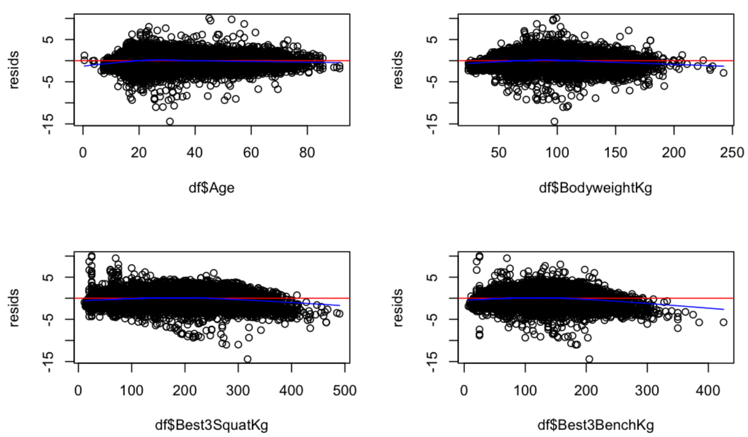
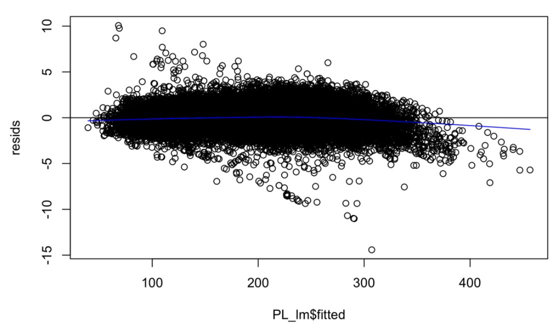


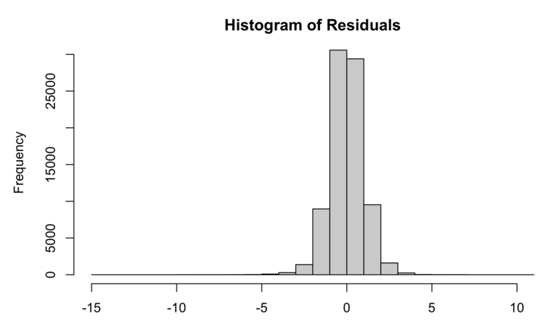
Figure 2 – Heat map of correlation coefficients (quantitative variables and response)

Before proceeding with the analysis plan and comparing models, I built a linear regression model using the full processed dataset (before splitting to training/test). This is to check the linear regression model assumptions and determine if it is reasonable for this problem. Figure 3a below shows the standardized residuals vs quantitative predicting variables to check the linearity assumption. The assumption seems to hold reasonably well for most of the data. However, the model may lose some accuracy for lifters with bodyweight > 200kg, squat > 400kg, and bench > 300kg. For this small population of very heavy/strong lifters the residuals are slightly negatively biased meaning the predictions are overestimated. Figure 3b shows the standardized residuals vs. fitted values to check the constant variance assumption. The same trend is seen here as we saw for linearity; the assumption holds for most the data but starts to overestimate at high fitted values. Figure 3c shows that the histogram of standardized residuals is reasonably normal, so the normality assumption holds. All the assumptions hold reasonably well enough to proceed, it just needs to be kept in mind that the model will start to overestimate for very strong lifters.



(a)





(b) (c)

Figure 3 – Plots to check regression model assumptions (a) Standardized Residuals vs. Predictors (b) Standardized Residuals vs. Fitted Values (c) Histogram of Standardized Residuals

Next, I checked for multicollinearity by calculating the VIF of all predictors. The values can be found in Table A2 of the appendix. All values are less than the threshold of 10, indicating that multicollinearity should not be an issue for the regression model (*Wikipedia*, 2022). Finally, I checked for outliers by calculating Cook's distance, the plot can be found in Figure A1 of the appendix. All data points have a Cook's distance much below the threshold value of 1, indicating there are no outliers or influential points (*Penn State,* 2018). I will proceed with analysis without removing any datapoints from the processed dataset.

4 METHODOLOGY

6 models will be evaluated, they are described and numbered below:

1. Linear regression with all 6 predictors
    a. The lm() function in R was used
    b. This is the baseline model
2. Linear regression with the best subset of *k*=4 predictors
    a. The regsubsets() function in R was used
    b. Purpose is variable selection and prediction
    c. An exhaustive search is performed of all combinations of *k*=4 predictors; the model with the lowest RSS (residual sum of squares) is chosen
    d. This model was compared to Model 1 with partial F-tests to check if the predictors not selected are significant
3. Linear Regression with stepwise variable selection using AIC
    a. The step() function in R was used
    b. Purpose is variable selection and prediction
    c. A combination of forward and backward selection; predictors are added or removed from the model at each step until the AIC reached a minimum
4. LASSO Regression
    a. The lars() function in R was used
    b. Purpose is variable selection and prediction
    c. Tunes the estimated coefficients by accounting for the L1 norm penalty
    d. Penalty parameter (λ) tuned by minimizing Mallow's Cp (0 found to be optimal, meaning there is no penalty term and it is the same as model1)
5. Ridge Regression
    a. The lm.ridge() function in R was used
    b. Purpose is only prediction, no variable selection is performed
    c. Tunes the estimated coefficients by accounting for the L2 norm penalty

    d. Penalty parameter (λ) tuned by generalized cross validation (1.1 found to be optimal)
6. Random Forest Regression
    a. The randomForest() function in R was used
    b. Purpose is only prediction, no variable selection is performed
    c. Predicts the response of a new datapoint based on the average prediction of an ensemble of decision trees
    d. Each tree is trained using a random bootstrap sample of the training data
    e. 3 hyperparameters were tuned using 3-fold CV on the training data (plots for tuning can be found in Figure 2A of the appendix)
        i. *nodesize*: Minimum size of the terminal nodes of each tree (1000 was found to be optimal)
        ii. *ntree*: Number of trees in the forest (500 was found to be optimal)
        iii. *mtry*: Number of random selected variables that can be considered at each split (3 was found to be optimal)

The reduced processed dataset (82,183 rows x 7 columns) was split to 70% training (57,528 samples) and 30% test (24,655 samples). Each model was built with the training data then tested with the test data. Mean Squared Error (MSE) was calculated for each model using the test predictions and true values. This single set of training/test data was used mainly to observe regression coefficients and variable selection. For models 1-5 the model performance was further evaluated using Monte Carlo CV, which reduces variance. The dataset is large so 10 runs were used to prevent excessive runtime; the dataset was re-sampled to 70% training/30% test within each run. The random forest model does not require Monte Carlo CV because the bootstrapping algorithm and ensemble of trees already reduces variance. The Monte Carlo CV MSE for Models 1-5 and Test MSE for Model 6 will be used to compare model performance.

5 ANALYSIS AND RESULTS

	Model	Train_MSE	Test_MSE	CV_MSE	CV_variance
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	LM w/ all predictors	382.6141	383.7175	384.9885	18.7409
2	LM with k=4 best predictors	383.7912	384.7560	386.0864	19.6394
3	LM with stepwise AIC	382.6141	383.7175	384.9885	18.7409
4	LASSO Regression	382.6141	383.7175	384.9885	18.7409
5	Ridge Regression	382.6141	383.7167	384.9882	18.7405
6	Random Forest	343.2596	350.6215	350.6215 *	NA

*Imputed from Test MSE

Table 1 – Performance metrics of all models

The performance metrics of all models are summarized in Table 1 above. Monte Carlo CV was not performed on Random Forest as explained in the methodology section, so its Test MSE was imputed for CV MSE to compare the other models. The model summaries can be found in Figures A3-A8 of the appendix. The baseline model (Model 1: linear regression with all predictors) has a CV MSE of 384.99. Model 3 (stepwise AIC), 4 (LASSO), and 5 (Ridge) all have the same training, test, and CV MSE as the baseline model. In Model 3, the stepwise regression algorithm found that the model with all predictors has the lowest AIC on the training data, which makes it equivalent to the baseline model. This behavior was repeated its Monte Carlo runs since the CV MSE is identical to the baseline model. In Model 4, the optimal penalty parameter for LASSO was found to be 0 when minimizing Mallow's Cp on the training data; this reduces the penalty term to 0 and makes the model equivalent to the baseline model. Again, this behavior was repeated in the Monte Carlo runs since the CV MSE is identical. In Model 5, the optimal penalty parameter for Ridge was found to be 1.1 by generalized cross validation on the training data. This made the penalty term very small relative to the sum of squares terms in the ridge estimator equation. As a result, the estimated coefficients are effectively the same as the baseline model up to 3+ significant figures and the CV MSE is effectively the same as well.

Model 2 (k=4 best subset) has a CV MSE of 386.09 which is just ~1 unit higher than the baseline model. The 4 predictors selected were *Sex, Equipment, Best3SquatKg, and Best3BenchKg*. That leaves *Age* and *BodyweightKg* as the two predictors not selected. Since the CV MSE for Model 2 is so close to the baseline model, I wanted to check if the two unselected predictors have any predictive power. I performed a partial F-test comparing Model 2 to Model 1 which resulted in a p-value < 2.2e-16 (Figure A9 in appendix). Comparing to a significance level of 0.05, the p-value is much smaller meaning at least one of the two predictors is non-zero and has predicting power. To check if both predictors were significant, I performed two additional and separate partial F-tests. This time I compared a model without *Age* to the full model, and a model without *BodyweightKg* to the full model. Both partial F-tests results in extremely small p-values, meaning that both *Age* and *BodyweightKg* are non-zero with predicting power (Figure A10 in appendix). It's interesting to see that even though Model 2 with 4 predictors has very similar CV MSE to the baseline model with 6 predictors, it is still statistically significant to include the 2 extra predictors *Age* and *Bodyweight*.

Model 6 (Random Forest) was able to achieve a Test MSE of 350.62. This is a ~35-unit improvement from the baseline model. I expect the Random Forest was able to produce more accurate predictions for the population of heavy/strong lifters where the linearity/constant variance assumptions for linear regression were beginning to not hold. While random forest has the best prediction performance, it cannot be used for a descriptive analysis. Therefore, I proceeded with Model 1 and trained it with the full processed dataset for a descriptive analysis.

```
Call:
lm(formula = Best3DeadliftKg ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-282.161  -11.771   -0.155   11.785  196.898 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   44.777918   0.314838  142.22   <2e-16 ***
Sex01         16.439803   0.208570   78.82   <2e-16 ***
Equipment01  -16.518415   0.187159  -88.26   <2e-16 ***
Age           0.091070   0.006271   14.52   <2e-16 ***
BodyweightKg  -0.035686   0.004289   -8.32   <2e-16 ***
Best3SquatKg   0.673516   0.002890  233.07   <2e-16 ***
Best3BenchKg   0.240393   0.004202   57.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.57 on 82176 degrees of freedom
Multiple R-squared:  0.8878,	Adjusted R-squared:  0.8878 
F-statistic: 1.084e+05 on 6 and 82176 DF,  p-value: < 2.2e-16
```

Figure 4 – Model summary for linear regression using all predictors trained with full processed dataset

Figure 4 above shows the model summary for the linear regression model with all predictors trained with the full processed dataset. The R-squared value is 0.89 which means 89% of the variability in the response variable is explained by the regression model. This suggests the model is a reasonably good fit and we can interpret the relationship of the predictors with the response using the estimated coefficients. Each interpretation below is made with the condition that all other variables are held constant.

Males (*Sex=1)* are expected to have a 16.4kg higher max deadlift than Females (*Sex*=0). This is interesting because it suggests that males and females of equal bodyweight that can squat and bench the same amount will not deadlift the same amount on average. It's possible that differences in male and female anatomy give males a mechanical advantage in deadlift.

Single-ply lifters (*Equipment*=1) are expected to have a 16.5kg lower max deadlift than Raw lifters (*Equipment*=0) on average. This leads me to believe that perhaps the extra lifting gear allowed in Single-ply events is more advantageous to squat and bench than it is to deadlift. I.e., the extra gear can help a generally "weaker" lifter achieve a squat and bench similar to that of a Raw lifter of equivalent bodyweight, biological gender, and age; but the extra gear will not help the "weaker" lifter match the Raw lifter's deadlift.

With every 1 year of age, a competitor's max deadlift is expected to increase by 0.091kg. Although age was proven to be statistically significant earlier with the partial F-test, the coefficient is very small and does not seem to have practical importance. The majority of lifters are in the 20-40 year-old range, which would only amount to a calculated 1.8kg difference in max deadlift. In the real world, I do not think there is much reason to expect competitors with equivalent max squat and bench to have significantly different deadlifts because of their age.

With every 1kg of bodyweight gained, a competitor's max deadlift is expected to decrease by 0.036kg. This is similar to what was seen with variable age; although the coefficient is statistically significant it is very small and not of much practical importance. The majority of competitors are in the 50-100kg range, which would only amount to a calculated 1.8kg difference in max deadlift. This is slightly surprising to me because I had a perception that heavier lifters had a much lower deadlift to squat ratio compared to lighter lifters based on my anecdotal experience. This analysis suggests that the difference in max deadlift of competitors with different bodyweight is extremely small if all other variables are equivalent.

With every 1kg of max squat gained, a competitor's max deadlift is expected to increase by 0.67kg. Squat has the largest coefficient of the quantitative predictors; this is not a surprise since showed the highest correlation during the exploratory analysis. As a competitor trains and increases their max squat, it is reasonable to expect they are also training deadlifts so their max for that lift will increase as well.

With every 1kg of max bench gained, a competitor's max deadlift is expected to increase by 0.24kg. The coefficient is smaller than that for squat but stills significantly larger than those for age and bodyweight. This seems reasonable because squat uses more muscles in common with deadlift than bench does. However, I expect the relationship between squat, bench, and deadlift is more so correlation than causation. In the real world, a lifter increases their max in the 3 lifts by training each lift. It is unlikely that a lifter would increase their max deadlift much if they are only training squat/bench and not deadlift.

6 CONCLUSIONS

The baseline model with all 6 predictors resulted in the lowest CV MSE out of the regression models. Stepwise AIC and LASSO selected all predictors, so they were identical the baseline model. When forcing variable selection by searching for the subset with the 4 best predictors, the CV MSE was only slightly worse. However, partial F-tests found that it was still statistically significant to include all 6 predictors. Ridge regression had a miniscule effect on coefficient estimates and was effectively the same as the baseline model. The data is fit reasonably well to the assumptions of linear regression but not perfectly, therefore I expect the Random Forest model was able to improve prediction accuracy by fitting better to the non-linear regions.

Although all 6 predictors were found to be statistically significant, the descriptive analysis showed that age and bodyweight had very small coefficients and were not of much practical importance in predicting max deadlift. It was not surprising to see that squat and bench have the most predictive power since they are correlated to how much a lifter trains and their general strength; but it was insightful to see it confirmed quantitatively. The most unexpected observations came from the binary variables for biological gender and equipment. The results led me to hypothesize that males may have a biomechanical advantage over females in deadlift, and that the extra lifting gear allowed in Single-ply events is relatively more advantageous to squat and bench than it is to deadlift. It would be interesting to conduct additional research to confirm if these observations are indeed true in the real world.

An idea for future work would be to repeat this analysis but use squat or bench as the response variable instead of deadlift. It would be interesting to see if the same conclusions are confirmed or if new observations are made. This analysis could also be extended to a larger population of lifters. We could explore if a model trained with data from the USAPL federation would have similar error when tested on lifters of another federation or country. It may be discovered that other federations or countries require a new model to be built; then it would be interesting to see what observations a descriptive analysis could uncover.

6.1 LESSONS LEARNED

In hindsight, the descriptive analysis proved more interesting than the prediction aspect of the project. By interpreting the coefficients of the regression model, I was able to make observations and hypotheses about powerlifting that otherwise would have never crossed my mind. The prediction aspect of the project showcased the power of Random Forest and how its non-linear nature can improve prediction accuracy. However, I cannot think of a compelling real-world use for a powerlifter to predict their max deadlift with one of these models since they ended up being highly correlated to squat and bench. The takeaway ends up being simply that you should train more and become stronger to increase your max deadlift. Prediction could possibly be useful to impute missing data in the openpowerlifting dataset for other projects. I still found this project to be productive and informative, but in the future I would try to construct more rigorous research questions.

7 APPENDIX

See supplemental file for R code and output.

	Age	BodyweightKg	Best3SquatKg	Best3BenchKg	Best3DeadliftKg
Age	1.00	0.15	-0.04	0.06	0.01
BodyweightKg	0.15	1.00	0.63	0.64	0.61
Best3SquatKg	-0.04	0.63	1.00	0.91	0.92
Best3BenchKg	0.06	0.64	0.91	1.00	0.89
Best3DeadliftKg	0.01	0.61	0.92	0.89	1.00

Table A1 – Correlation matrix of quantitative variables and response

Sex01	Equipment01	Age	BodyweightKg	Best3SquatKg	Best3BenchKg
2.162875	1.097860	1.105878	1.827695	6.444261	7.528158

The VIF threshold is: 10

Table A2 – VIF values for regression model with all predictors

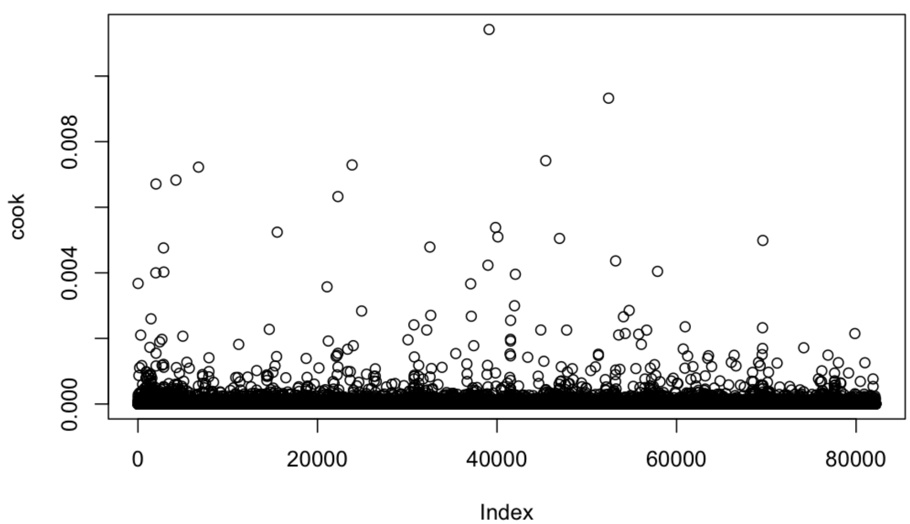
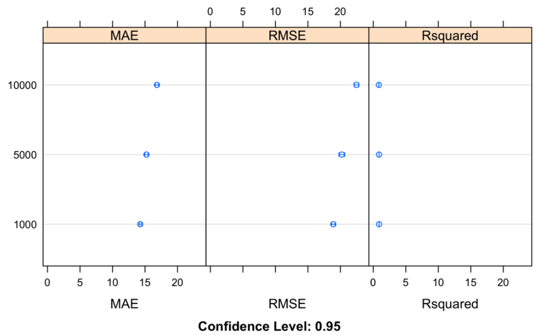
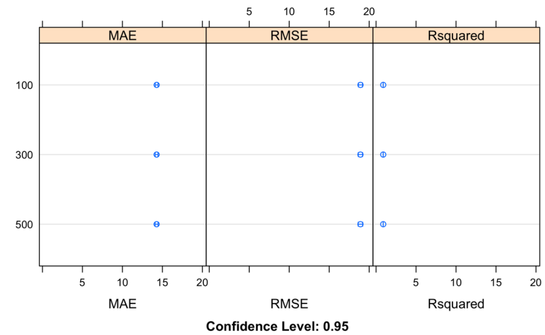


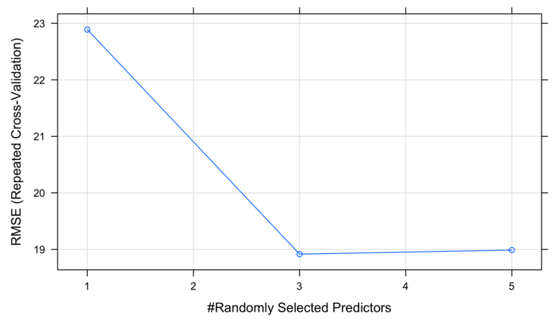
Figure A1 – Cook's distance for regression model with all predictors



(a)



(b)



(c)

Figure A2 – Random Forest 3-fold CV error when varying (a) *nodesize* (b) *ntree* (c) *mtry*

```
Call:
lm(formula = Best3DeadliftKg ~ ., data = pl1train)

Residuals:
     Min       1Q   Median       3Q      Max
-282.332  -11.780   -0.137   11.877  186.239

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.942985   0.374491 120.011  < 2e-16 ***
Sex01         16.411436   0.249039  65.899  < 2e-16 ***
Equipment01  -16.783273   0.224133 -74.881  < 2e-16 ***
Age            0.087555   0.007466  11.728  < 2e-16 ***
BodyweightKg  -0.041875   0.005128  -8.167 3.23e-16 ***
Best3SquatKg   0.675626   0.003452 195.724  < 2e-16 ***
Best3BenchKg   0.242062   0.005015  48.271  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.56 on 57521 degrees of freedom
Multiple R-squared:  0.8883,    Adjusted R-squared:  0.8883
F-statistic: 7.625e+04 on 6 and 57521 DF,  p-value: < 2.2e-16
```

Figure A3 – Model summary for linear regression with all predictors on training data (Model 1)

```
Call:
lm(formula = as.formula(mod2form), data = pl1train)

Residuals:
     Min       1Q   Median       3Q      Max 
-281.212  -11.731   -0.111   11.904  186.288 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   45.532190   0.239681  189.97   <2e-16 ***
Sex01         16.155129   0.247237   65.34   <2e-16 ***
Equipment01  -16.733311   0.220697  -75.82   <2e-16 ***
Best3SquatKg   0.662570   0.003309  200.21   <2e-16 ***
Best3BenchKg   0.248254   0.004823   51.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.59 on 57523 degrees of freedom
Multiple R-squared:  0.888,	Adjusted R-squared:  0.888 
F-statistic: 1.14e+05 on 4 and 57523 DF,  p-value: < 2.2e-16
```

Figure A4 – Model summary for linear regression with *k*=4 best predictors on training data (Model 2)

```
Call:
lm(formula = Best3DeadliftKg ~ Sex01 + Equipment01 + Age + BodyweightKg + 
    Best3SquatKg + Best3BenchKg, data = pl1train)

Residuals:
     Min       1Q   Median       3Q      Max 
-282.332  -11.780   -0.137   11.877  186.239 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   44.942985   0.374491 120.011  < 2e-16 ***
Sex01         16.411436   0.249039  65.899  < 2e-16 ***
Equipment01  -16.783273   0.224133 -74.881  < 2e-16 ***
Age            0.087555   0.007466  11.728  < 2e-16 ***
BodyweightKg  -0.041875   0.005128  -8.167 3.23e-16 ***
Best3SquatKg   0.675626   0.003452 195.724  < 2e-16 ***
Best3BenchKg   0.242062   0.005015  48.271  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.56 on 57521 degrees of freedom
Multiple R-squared:  0.8883,	Adjusted R-squared:  0.8883 
F-statistic: 7.625e+04 on 6 and 57521 DF,  p-value: < 2.2e-16
```

Figure A5 – Model summary for linear regression with stepwise AIC on training data (Model 3)

```
                   Sex01  Equipment01          Age BodyweightKg Best3SquatKg Best3BenchKg 
 44.94298504  16.41143633 -16.78327313   0.08755549  -0.04187455   0.67562591   0.24206228 
```

Figure A6 – Model coefficients for LASSO regression on training data (Model 4)

```
                   Sex01  Equipment01          Age BodyweightKg Best3SquatKg Best3BenchKg 
44.94482610  16.41145112 -16.78204779   0.08752172  -0.04183990   0.67555962   0.24212737 
```

Figure A7 – Model coefficients for Ridge regression on training data (Model 5)

```
##              %IncMSE IncNodePurity
## Sex01        30.55188    16897872.5
## Equipment01  72.09496      891743.5
## Age          35.66265     1187055.4
## BodyweightKg 21.95011     4458373.5
## Best3SquatKg 64.94309    98507143.7
## Best3BenchKg 26.37299    52921762.0
```

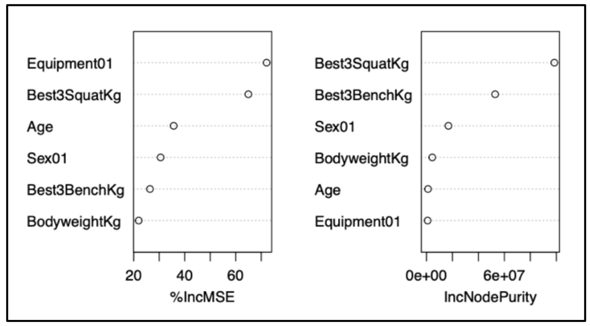


Figure A8 – Importance measurements for Random Forest on training data (Model 6)

```
Analysis of Variance Table

Model 1: Best3DeadliftKg ~ Sex01 + Equipment01 + Best3SquatKg + Best3BenchKg
Model 2: Best3DeadliftKg ~ Sex01 + Equipment01 + Age + BodyweightKg +
    Best3SquatKg + Best3BenchKg
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  57523 22078738
2  57521 22011025  2     67713 88.476 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure A9 – ANOVA Partial F-Test on *Age* + *BodyweightKg* together

```
Analysis of Variance Table

Model 1: Best3DeadliftKg ~ (Sex01 + Equipment01 + Age + BodyweightKg +
    Best3SquatKg + Best3BenchKg) - Age
Model 2: Best3DeadliftKg ~ Sex01 + Equipment01 + Age + BodyweightKg +
    Best3SquatKg + Best3BenchKg
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  57522 22063657
2  57521 22011025  1     52631 137.54 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Model 1: Best3DeadliftKg ~ (Sex01 + Equipment01 + Age + BodyweightKg +
    Best3SquatKg + Best3BenchKg) - BodyweightKg
Model 2: Best3DeadliftKg ~ Sex01 + Equipment01 + Age + BodyweightKg +
    Best3SquatKg + Best3BenchKg
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  57522 22036546
2  57521 22011025  1     25521 66.694 3.235e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure A10 – ANOVA Partial F-Test separately on (a) *Age* (b) *BodyweightKg*

8 CREDITS

1. R Software
2. OpenPowerlifting Database. (2019, April). In *Kaggle*.
   https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database
3. Penn State (2018). Identifying Influential Data Points. In *STAT 462: Applied Regression Analysis*
   https://online.stat.psu.edu/stat462/node/173/
4. Variance Inflation Factor. (2022, December 2022). In *Wikipedia*.
   https://en.wikipedia.org/w/index.php?title=Variance_inflation_factor&oldid=1125859986