Predictive Analytics in the Context of Competitive Powerlifting

OPMA 419 - Winter 2021
Submitted to Professor Sabouri on April 16, 2021

Group 03
Shubham Gehlot (30029458)
Benjamin LeBlanc (30037877)
Nevin Sangha (30041970)

Table of Contents

1. Problem Overview

Powerlifting competitions have existed for centuries dating back to ancient Greek and Persian times and weightlifting has been a perennial part of the Olympics since 1896 (Powerlifting, 2021). Official powerlifting competitions have existed for decades and take place all across the world. A majority of these competitions are sanctioned by the International Powerlifting Federation (IPF) and feature competitors having three attempts at maximal weights in three events, the squat, bench press, and deadlift. Our group's objective is to explore which factors, algorithms, and evaluation methods are the most effective in predicting an athletes deadlift weight. The insights from this project will be useful to determine the difference between the average human and official powerlifting competitors. Additionally, this information and predictive model will be useful for any fitness enthusiast seeking to improve upon or to optimize their personal deadlift record.

2. Data Mining Task

We are going to create a predictive model that predicts the maximum amount of weight able to be lifted in a deadlift by a trained individual. The predictors we will use are sex, age, squat weight, bench weight, bodyweight, and tested for performance enhancing drugs ("PEDs"), among others. The Y-Variable will be a predicted numerical value.

3. Data Explanation & Cleaning

Discuss the Source

This dataset is derived from the OpenPowerlifting archive. The OperPowerlifting archive is a volunteer-run service with a team that consists of eight people. Their goal is to remember every lifter at every level of the sport across every federation. It is a public-domain of powerlifting history and events. The data set is obtained from Kaggle.com and consists of over 1.4 million unique rows and 37 columns of competitor data from the OpenPowerlifting database from April 2019. The dataset can be accessed here: https://www.kaggle.com/open-powerlifting/powerlifting-database.

Exploratory Data Analysis

Prior to executing any predictive analysis, our group wanted to contextualize and visualize the variables relevant to our models.

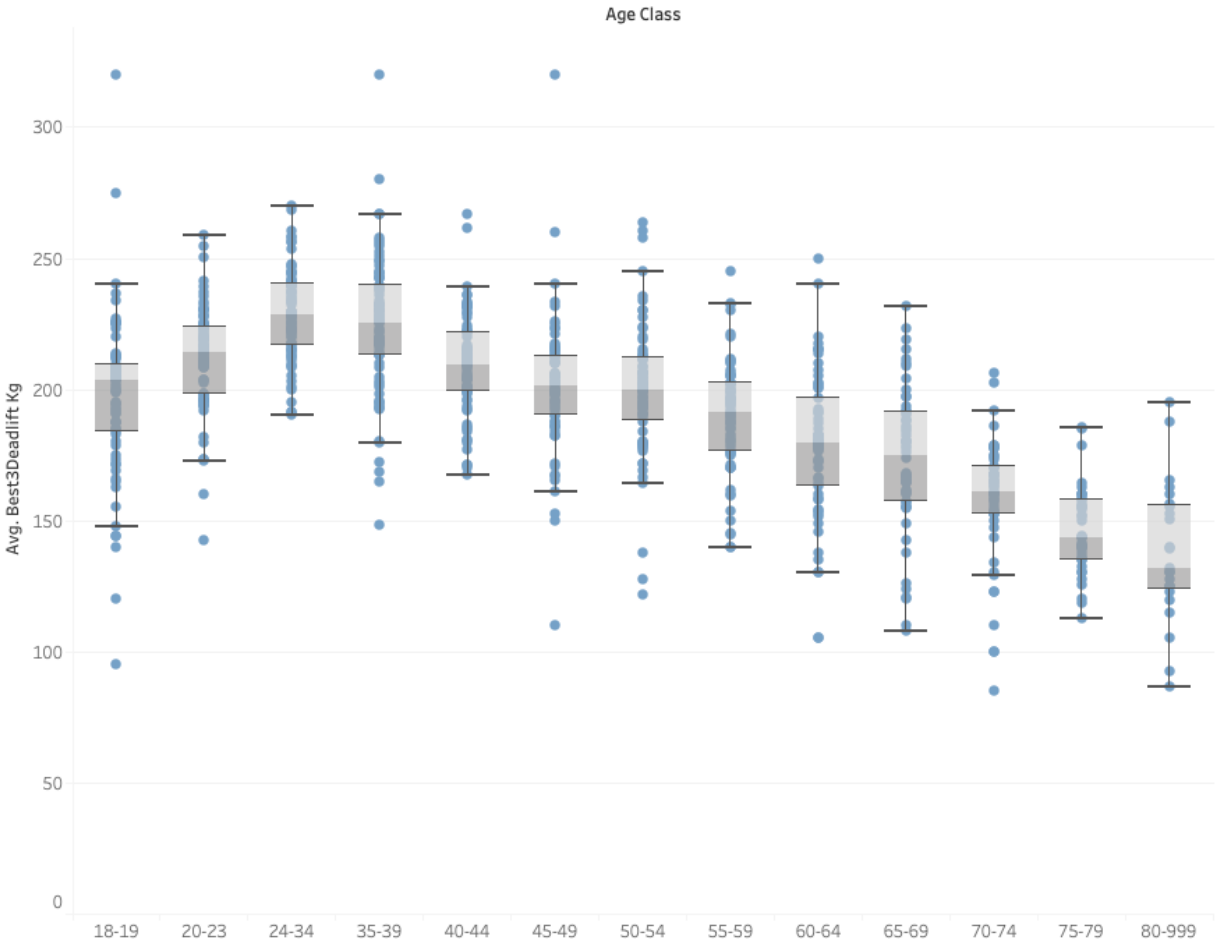


Figure 1: Boxplots of Average Deadlift (kg) for each Meet Country over Age Class

The boxplots above provide a broad scope on the expectations that each competitor has of their respective deadlift attempt in kilograms. Each data point represents the average deadlift for a specific country where the competition was held, this information is then further categorized by age class of each competitor. By further analysing the data, it is clear that the optimal age class for the highest deadlift is between the ages of 24-34 (median of 228.3 kg) & 35-39 (median of 225.4 kg).

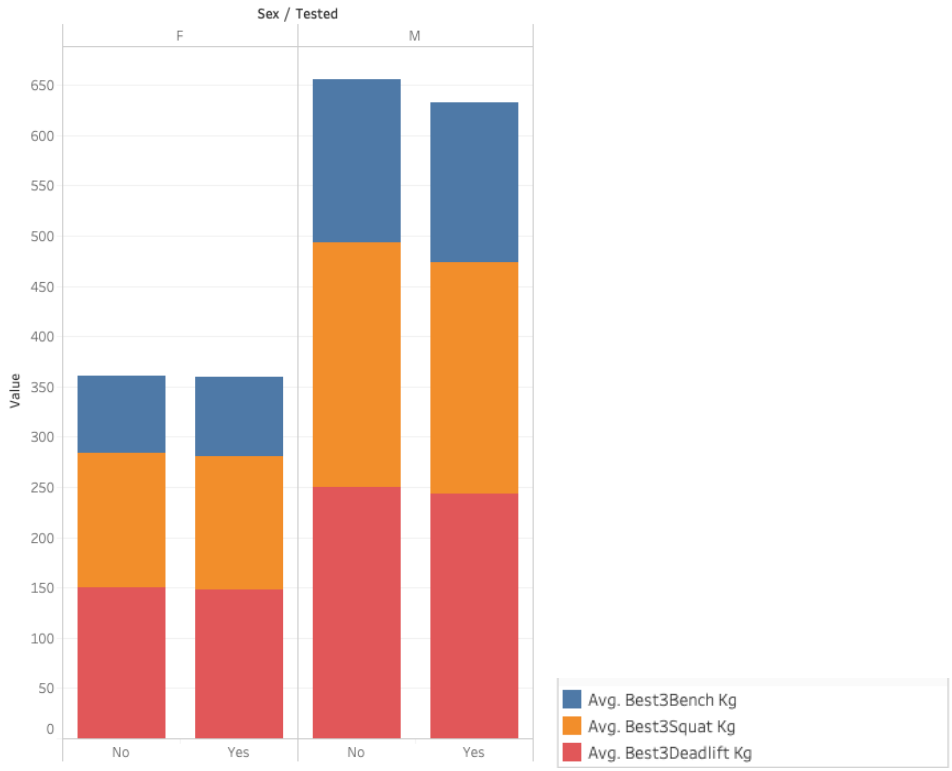


Figure 2: Total kg breakdown by Gender and Tested

The bar chart above displays a breakdown of a powerlifter's total kg (sum of squat, deadlift, and bench), discrepancies between male and female competitors, and competitors who have been tested compared to those who have not. Studies show that women on average have lower appendicular muscle mass than men (Janssen et al., 2000). Thus it is not surprising to see that male competitors have on average approximately 97.2 kg heavier deadlift than female competitors. Additionally on average for both male and female competitors, the deadlift will be the most significant event when calculating total kg. Lastly, getting tested for PEDs dramatically affects males results when compared to females. The average total kg for male competitors not tested is about 23.2 kg higher than for those who have.

Attributes	sex = M	equipment = Single-ply	equipment = Wraps	equipment = Multi-ply	tested = Yes	age	bodyweight_kg	best3squat_kg	best3bench_kg	best3deadlift_kg
sex = M	1	0.011	0.075	0.083	-0.102	0.031	0.513	0.614	0.664	0.699
equipment = Single-ply	0.011	1	-0.343	-0.145	0.354	0.048	-0.034	0.189	0.210	0.053
equipment = Wraps	0.075	-0.343	1	-0.132	-0.580	-0.051	0.088	0.011	-0.028	0.042
equipment = Multi-ply	0.083	-0.145	-0.132	1	-0.272	0.041	0.077	0.218	0.197	0.093
tested = Yes	-0.102	0.354	-0.580	-0.272	1	0.009	-0.139	-0.120	-0.090	-0.113
age	0.031	0.048	-0.051	0.041	0.009	1	0.019	-0.209	-0.143	-0.200
bodyweight_kg	0.513	-0.034	0.088	0.077	-0.139	0.019	1	0.661	0.673	0.652
best3squat_kg	0.614	0.189	0.011	0.218	-0.120	-0.209	0.661	1	0.894	0.891
best3bench_kg	0.664	0.210	-0.028	0.197	-0.090	-0.143	0.673	0.894	1	0.834
best3deadlift_kg	0.699	0.053	0.042	0.093	-0.113	-0.200	0.652	0.891	0.834	1

Figure 3: Confusion Matrix

The confusion matrix looks at the correlations between all of our variables used in our analysis. First thing to note is the high correlations with (Sex = M) variable. It appears that if your sex is male, then your bodyweight in kg, best squat in kg, best bench in kg, and best deadlift in kg goes up on average compared to females. This could be related to men having more muscle mass on average compared to women (Janssen et al., 2000). Another interesting correlation is the negative correlation between (tested = yes) and (equipment = wraps). It appears that if you are tested the less likely you are to use wraps. It appears in general that if you are tested for drugs, your weights in kg across the board go down. The most understandable correlations are the highly positive correlations between your bodyweight, bench, squat, and deadlift. The more mass you have, and the more mass you move for any of the lifts, the more you can lift for the others.

Data Cleaning

Initially our data had 1.4 million rows so our data cleaning process was very time intensive to plan and implement. Our data cleaning and preparation outside of RapidMiner primarily was done using R and Excel. Since we had so many records of data to play with, we were merciless when it came to counting data.

In R, we started with selecting columns from the dataset to keep, which is explicitly shown in our Appendix A. we decided to use complete cases of data, meaning that we got rid of any record that had NA's. After we had complete cases of data, we went further and cleaned up the tested data and the country data, to ensure the columns were properly labeled. After that, we filtered the data to only include records that had ages of 18 plus ($age \geq 18$). We also filtered the records for the year, anything greater than 2010, and our data goes up to 2019, so we would have approximately 9 years of data ($date \geq 2010 - 01 - 01$). After that, we exported the dataset into Excel.

In Excel we checked to see if we had complete data, fixed the ordering of the rows, added three new rows and added one new column to the dataset. We added a categorical column titled Nevin_Ben_Shaan with response yes (Y) or no (N). This column was strictly to separate our data

inside of RapidMiner to test our data against the model we created. The three rows of data that we included were sex, age, weightlifting numbers, and more. After all this data cleaning, we were ready to import our data into RapidMiner.

4. Data Partitioning, Optimization, and Dimension Reduction

Data Prep

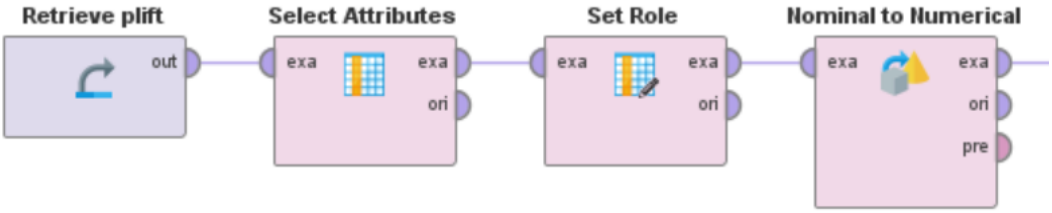


Figure 4: Our Data Prep within RapidMiner

We set the role for best3deadlift_kg as label as this is our dependent variable in our analysis. We also have three categorical variables that we turned into dummy variables in our dataset. Namely, equipment, sex, and tested. Within Nominal to Numerical, we selected those attributes and set our comparison groups as the following:

comparison group attribute	comparison group
equipment	Raw
sex	F
tested	No

Figure 5: Comparison Groups

After our final data preparations, we were ready to move forward with our analysis.

Data Partitioning

For the data partitioning we decided to go with a 80% and 20% split for training and validation respectively. With that, we also used shuffled sampling with local random seed of 70500 to keep samples consistent. The reason we decided to do an 80% and 20% split is due to our sheer amount of data size (99 thousand rows).

Optimize Parameters

k-NN Optimization

We ran an optimize parameter on our k-NN model to find our optimal k. We checked k values from 1 to 20. The optimal k that was received was 12.

iteration	k-NN.k	root... ↑
12	12	22.417
13	13	22.421
11	11	22.430
14	14	22.439
10	10	22.441
15	15	22.468
16	16	22.472
17	17	22.483
19	19	22.486
18	18	22.491
9	9	22.502
20	20	22.507
8	8	22.542
7	7	22.597
6	6	22.676
5	5	22.826
4	4	23.078
3	3	23.445
2	2	24.340
1	1	216.544

Figure 6: k-NN Optimize Parameter

Decision Tree Optimization

We ran an optimize parameter on our Decision Tree model to find our optimal minimal leaf size. The minimal leaf size that was received was 63.

iteration	Decisio...	root... ↑
63	63	23.141
50	50	23.142
61	61	23.146
62	62	23.147
64	64	23.150
65	65	23.150
60	60	23.150
66	66	23.152
39	39	23.155
49	49	23.155
40	40	23.156

Figure 7: Decision Tree Optimize Parameter

Dimension Reduction: Initial

Referring back to our Data Prep, in our select attributes operator we decided to exclude attributes such as GlossBrenner, IPF points, McCulloch, Wilks, place, and total kg due to the fact that they are calculations that incorporate your best deadlift (our dependent variable). We got rid of some excessive categorical attributes, such as country, meet country (almost always the same as country), meet state, federation and division. We also got rid of variables that would not add

any benefit to our modeling, such as name, date, event (all same), and place (since we didn't want to compare scores, but predict deadlifts). Lastly, we got rid of age class and weight class in kg since they were primarily used in exploratory data analysis and we have the numerical data which is more beneficial for our prediction. The full select attributes list can be found here:

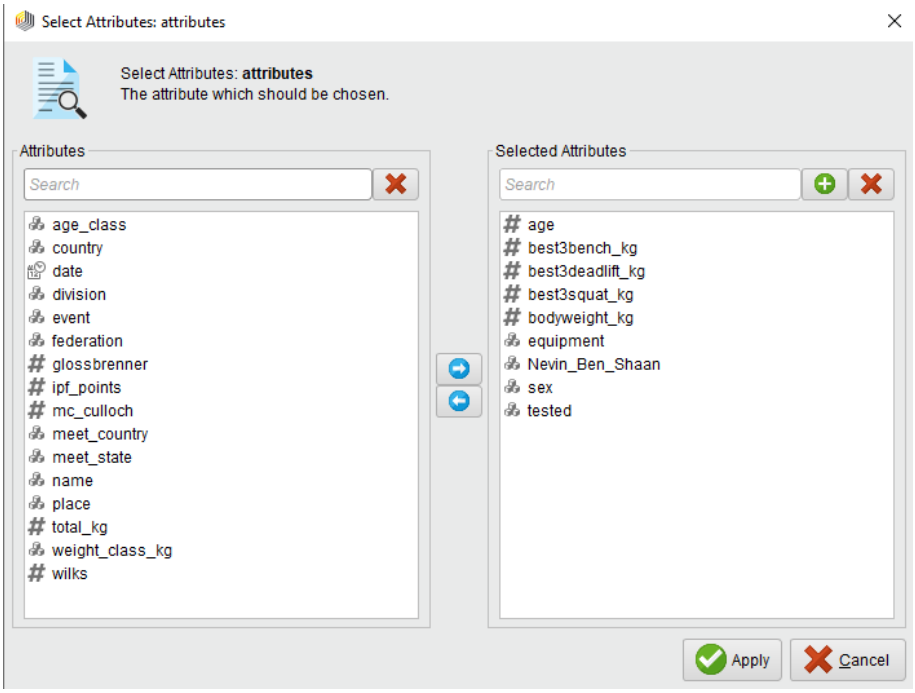


Figure 8: Select Attributes

Dimension Reduction: Backwards Elimination

Model	**Parameters**	**RMSE**	**R-squared**	**Predictors Removed**
Multiple Linear Regression	No Feature Selection and No Eliminating Collinear Features	**24.218 +/- 0.000**	**0.856**	**tested = Yes**
Decision Tree	Negative 1 Maximal Depth and Optimized for Best Minimal Leaf Size of 63	**23.151 +/- 0.000**	**0.869**	**tested = Yes**

Neural Net	200 Training Cycles 1 Hidden Layers size of 2 Local seed of 70500	**23.093 +/- 0.000**	**0.872**	**tested = Yes**

Due to k-NN being a lazy, computational heavy, learner we decided to go against using it for our feature selection process since we have 99 thousand rows.

Since no specific model eliminated unique predictors, we have decided to test removing (tested = Yes) from our processes, and compare it to full models as well.

5. Analysis

Naïve Rule

For our Naïve rule we would predict that everyone could do the mean deadlift out of our training set of data. Therefore, based on the image below, we would predict that everyone could deadlift 216.657 kg.

Figure 9: Training set statistics

Our performance on the validation set for the Naïve rule model was quite high, as saying that everyone can deadlift 216.657 kg is quite excessive, and is quite the impressive feat. Overall, the root mean squared error was a reported 63.911 kg.

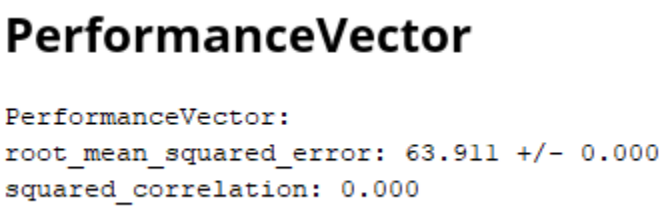


Figure 10: Performance on Naïve Rule

Model Testing

For our model testing we decided to use four different algorithms to predict deadlift weights. The four models that we decided to compare are k-Nearest Neighbours, Multiple Linear Regression, Decision Tree, and Neural Nets. We decided to use Root Mean Square Error and R-squared to compare and contrast the four models. We also did two sets for each model, one set using all predictors, another set removing (tested = yes) as seen in our backward elimination.

Multiple Linear Regression Model

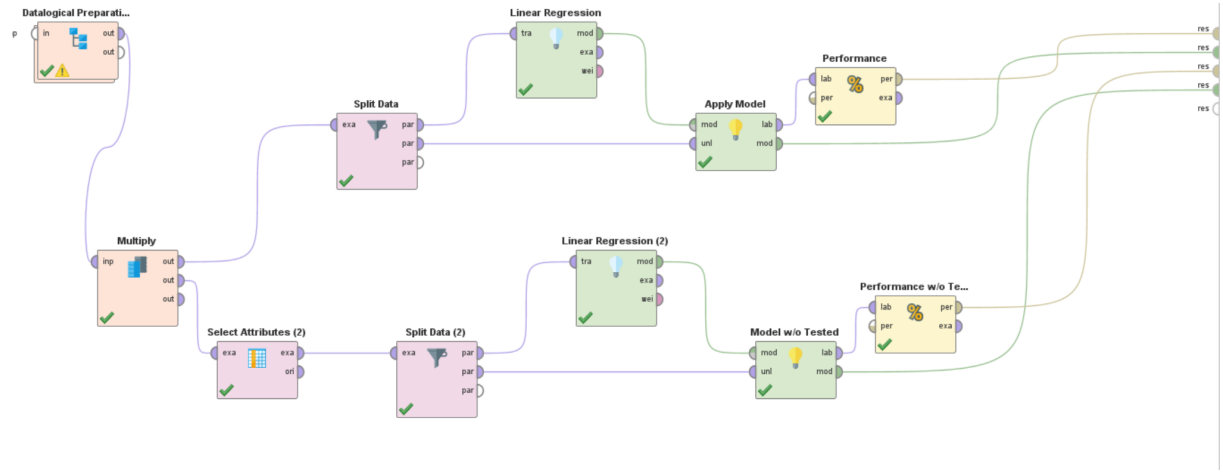


Figure 11: Multiple Linear Regression Process

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 24.218 +/- 0.000
squared_correlation: 0.856
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 24.206 +/- 0.000
squared_correlation: 0.857
```

Figure 12: Multiple Linear Regression Performance
12.1 (left): without (tested = yes)
12.2 (right): with (tested = yes)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
sex = M	28.526	0.261	0.205	0.574	109.426	0	***
equipment = Single-ply	-19.158	0.229	-0.134	0.975	-83.543	0	***
equipment = Wraps	-6.840	0.220	-0.046	0.993	-31.029	0	***
equipment = Multi-ply	-35.841	0.422	-0.125	0.954	-84.880	0	***
age	-0.164	0.007	-0.035	0.966	-24.059	0	***
bodyweight_kg	0.062	0.005	0.023	0.520	11.670	0	***
best3squat_kg	0.597	0.003	0.717	0.237	225.459	0	***
best3bench_kg	0.099	0.004	0.088	0.216	26.649	0	***
(Intercept)	70.939	0.415	?	?	171.040	0	***

Figure 12.1.1: Multiple Linear Regression Model (without tested = yes)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
sex = M	28.591	0.261	0.206	0.574	109.659	0	****
equipment = Single-ply	-18.895	0.232	-0.132	0.976	-81.515	0	****
equipment = Wraps	-8.070	0.272	-0.054	0.994	-29.637	0	****
equipment = Multi-ply	-37.090	0.452	-0.130	0.959	-82.007	0	****
tested = Yes	-1.896	0.247	-0.014	0.987	-7.689	0.000	****
age	-0.166	0.007	-0.036	0.966	-24.355	0	****
bodyweight_kg	0.061	0.005	0.023	0.520	11.566	0	****
best3squat_kg	0.596	0.003	0.716	0.237	225.297	0	****
best3bench_kg	0.098	0.004	0.087	0.215	26.428	0	****
(Intercept)	72.745	0.476	?	?	152.671	0	****

Figure 12.2.1: Multiple Linear Regression Model (with tested = yes)

k-NN Model

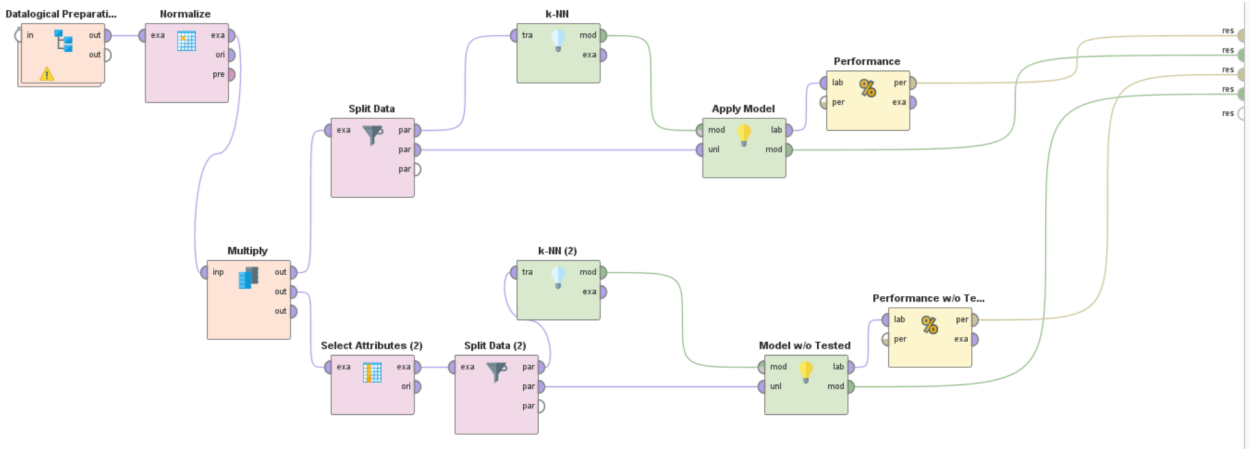


Figure 13: k-NN Process

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 22.416 +/- 0.000
squared_correlation: 0.877
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 22.417 +/- 0.000
squared_correlation: 0.877
```

Figure 14: k-NN Performance
14.1 (left): without (tested = yes)
14.2 (right): with (tested = yes)

Neural Net Model

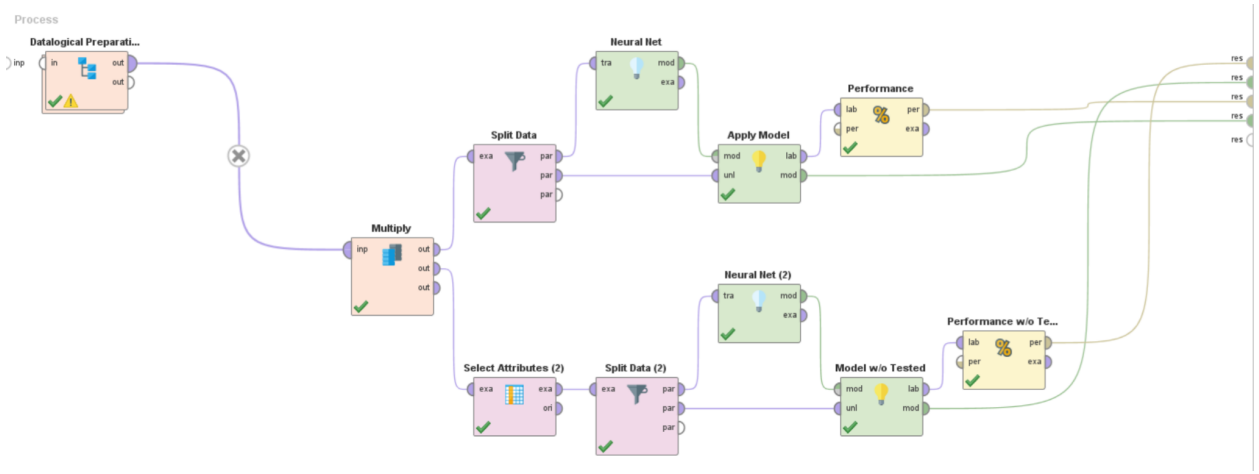


Figure 15: Neural Net Process

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 22.894 +/- 0.000
squared_correlation: 0.872
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 23.502 +/- 0.000
squared_correlation: 0.872
```

Figure 16: Neural Net Performance
16.1 (left): without (tested = yes)
16.2 (right): with (tested = yes)

Decision Tree Model

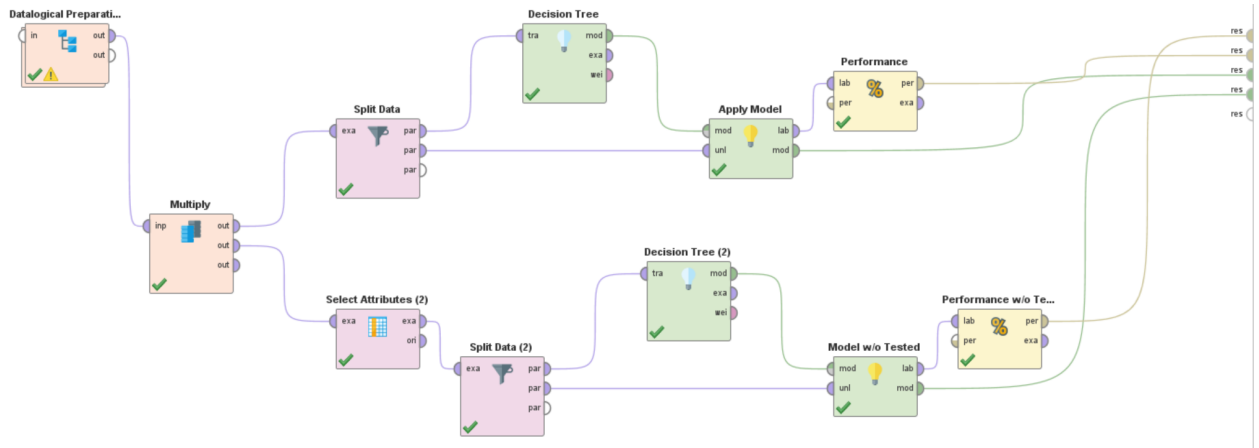
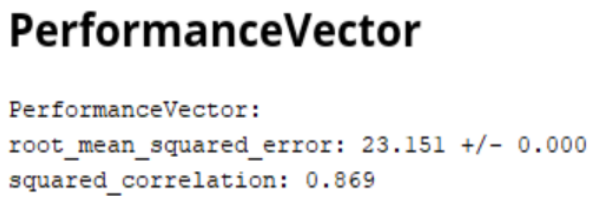


Figure 17: Decision Tree Process



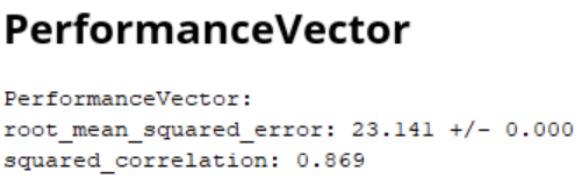


Figure 18: Decision Tree Performance
18.1 (left): without (tested = yes)
18.2 (right): with (tested = yes)

Model Results Table

Model	Parameters	RMSE	R-squared	Predictors Removed
Naïve Rule	N/A	**63.911 +/- 0.000**	**0.000**	**N/A**
Multiple Linear Regression	No Feature Selection and No Eliminating Collinear Features	**24.206 +/- 0.000**	**0.857**	**N/A**
Multiple Linear Regression with Tested Removed	No Feature Selection and No Eliminating Collinear Features	**24.218 +/- 0.000**	**0.856**	**tested = yes**
k-NN	Normalized, Optimized for Best K of 12, Weighted vote on, and Numerical Measures with Euclidean Distances	**22.417 +/- 0.000**	**0.877**	**N/A**
k-NN with Tested Removed	Normalized, Optimized for Best K of 12,	**22.416 +/- 0.000**	**0.877**	**tested = yes**

	Weighted vote on, and Numerical Measures with Euclidean Distances			
Neural Net	200 Training Cycles 1 Hidden Layers size of 5	**22.894 +/- 0.000**	**0.872**	**N/A**
Neural Net with Linear with Tested Removed	200 Training Cycles 1 Hidden Layers size of 5	**23.502 +/- 0.000**	**0.872**	**tested = yes**
Decision Tree	Negative 1 Maximal Depth and Optimized for Best Minimal Leaf Size of 63	**23.141 +/- 0.000**	**0.869**	**N/A**
Decision Tree with Tested Removed	Negative 1 Maximal Depth and Optimized for Best Minimal Leaf Size of 63	**23.151 +/- 0.000**	**0.869**	**tested = yes**

Outcome

Comparing all of our models to our Naïve rule, all of them exceeded the performance (root mean squared error) by nearly 40 kg. After analyzing all the models, with the backwards eliminations result of (tested = yes) excluded and included, we have selected a model that will be used going onwards with the report. That model is our k-NN (without tested, normalized, optimal K of 12). We chose k-NN due to it having the lowest reported RMSE and highest R-squared value on our validation set.

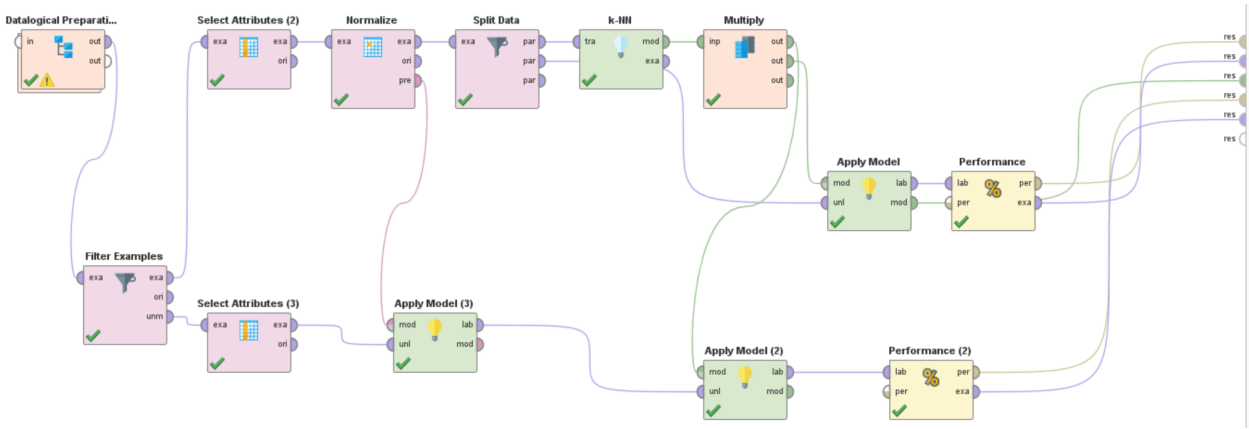


Figure 19: Final k-NN Process, our data is processed separately at the bottom

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 22.416 +/- 0.000
squared_correlation: 0.877
```

Figure 20: k-NN performance, best performance

Row No.	best3deadlif...	prediction(b...	sex = M	equipment =...	equipment =...	equipment =...	age	bodyweight...	best3squat_...	best3bench...
1	125	132.116	-1.516	-0.615	-0.557	-0.236	1.404	-0.155	-1.161	-1.147
2	162.500	148.081	-1.516	-0.615	1.794	-0.236	-0.751	-0.193	-0.803	-1.235
3	180	195.019	-1.516	-0.615	1.794	-0.236	-0.824	-0.260	-0.185	-0.795
4	155	136.101	-1.516	-0.615	1.794	-0.236	-1.117	-0.251	-0.966	-1.367
5	135	160.940	-1.516	-0.615	1.794	-0.236	-0.824	-1.193	-0.705	-0.707
6	145	146.972	-1.516	-0.615	1.794	-0.236	-1.117	-0.201	-0.835	-1.235
7	232.500	237.275	0.660	-0.615	1.794	-0.236	-0.678	-0.803	0.141	0.305
8	215	223.213	0.660	-0.615	1.794	-0.236	-0.678	0.866	0.076	-0.223
9	290	249.430	0.660	-0.615	1.794	-0.236	-0.459	-0.176	0.727	0.877
10	255	246.559	0.660	-0.615	1.794	-0.236	-0.240	0.565	0.304	0.129
11	260	273.962	0.660	-0.615	1.794	-0.236	-1.190	0.736	0.857	0.613

Figure 20.1: Sample of our predictions on our best model

Testing Data: Our Own Data

After finding the best model for predicting our data, we have also decided to use this model to help predict what some of our deadlift numbers should be. Therefore, we entered our data into the dataset, separated ourselves via filter, and predicted our deadlifts. The results are found below:

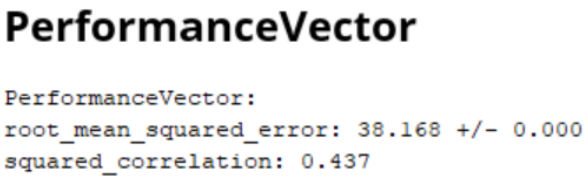


Figure 21: Performance on our test data

Row No.	best3deadlif...	prediction(b...	sex = M	equipment =...	equipment =...	equipment =...	age	bodyweight...	best3squat_...	best3bench...
1	130	150.002	0.660	-0.615	-0.557	-0.236	-0.971	-0.979	-1.174	-1.024
2	166	193.982	0.660	-0.615	-0.557	-0.236	-1.044	0.255	-0.796	-0.179
3	120	176.455	0.660	-0.615	-0.557	-0.236	-1.044	-0.352	-1.070	-0.795

Figure 21.1: Predictions on our test data

As we can see, the predictions made on our data compared to our training and validation data was a lot worse. The RMSE was 38.168 and R-squared was 0.437. Contrasting to our validation set, RMSE was nearly double our model performance, and R-squared was nearly half to the one seen in our validation set. This can be accounted for many reasons, but the simplest being that we are not professional powerlifters but rather we do it as a hobby to stay healthy. We can take these predictions as an idea of what our deadlifts should be, and potentially a goal in the future.

6. Findings, Insights, Challenges & Limitations

In our modeling and dataset we made quite a few interesting discoveries. Firstly, all of our models greatly exceeded our naïve rule, almost all by 40 kg RMSE. Our best model was k-NN (without tested, normalized, optimal k of 12). Our worst model was Multiple Linear Regression with tested removed. An interesting observation between the best and worst models was that the RMSE only had a difference of 1.8 kg, and the R-squared difference was 2.1%. Overall, all of our models had fairly high accuracy and predictive power, and we believe this is due to our dataset having 99 thousand rows and all of our predictors being somewhat relevant to the prediction. Some trends and insights we found within the data include that men's results are more skewed when tested is taken into consideration. Additionally, factors such as gender (male having a high correlation), squat and bench records, and overall bodyweight have a significant impact on an individuals deadlift potential.

Our project was subject to various challenges, beginning with the overall size of the chosen dataset. The raw extract featured over 1.4 million rows of data and required extensive data cleaning through R, Excel, and RapidMiner. The total data prep process became especially tedious when developing our predictive models. The dataset size in combination of certain RapidMiner operators resulted in heavy computational intensity when executing some processes (see Appendix C). Due to k-NN being a lazy learner this operator, along with neural net and

backwards elimination resulted in some processes taking nearly 3-5 minutes to execute. When testing operators, our highest runtime for a single process was 22 minutes. This time can quickly add up, as everytime the models were adjusted they needed to be executed to view the changed results.

Some limitations of our analysis include the absence of potential important predictors. The data set features many essential predictors for deadlift, however for a more complete analysis it would have been useful to have additional information about the competitors themselves. This includes potential predictors such as height, body fat %, diet/nutritional information, and other key factors in the powerlifting and bodybuilding world. Another limitation was that in the raw data we had was the country but not its continent, or another type of binning. We would have used country as part of our analysis, however it would have required 100's of dummy variables in order to fit into our model.

7. Recommendations

We recommend that fitness enthusiasts input their own statistics into the data and utilize this model to predict their optimal deadlift target. The model which produced the lowest RMSE was k-NN with the variable tested removed at 22.416 and an R-squared of 0.877. The best predictors for a competitors deadlift are sex, equipment, age, bodyweight, best3squat, and best3bench.

While sex and age are out of the control of an individual, we suggest that your optimal weightlifitng is maximzed in the between the ages of 24-39. Using either multi-ply or single ply equipment may help add a few extra kilograms to your deadlift maximum. We know that mass moves mass, thus the heavier an individual is the more they will be able to lift, at a healthy weight. Our last suggestion is that individuals train all aspects of their body through exercises such as squat and bench press, as this will help grow muscles which are ultimately used when deadlifting.

References

Janssen, I., Heymsfield, S. B., Wang, Z., & Ross, R. (2000). Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr. *Journal of Applied Physiology*, *89*(1), 81–88. https://doi.org/10.1152/jappl.2000.89.1.81

Powerlifting. (2021, March 24). *In Wikipedia.* https://en.wikipedia.org/w/index.php?title=Powerlifting&oldid=1013972062

Appendix

Appendix A: Data Dictionary

This is our explicit dataset, including all the columns and definitions from the initial dataset from kaggle.

- (*) indicates removed in R
- (**) indicates not selected in RapidMiner
- (***) indicates used for filtering (we added this column in Excel)
- Bolded indicates used in predictive model

Column	Definition
Name**	Name of the contestant
Sex	Either of the two major forms of individuals that occur in many species and that are distinguished respectively as female or male
Event**	The lifts competed in. SBD is the standard three powerlifts, but competitors may do only some of the lifts
Equipment	The type of equipment that was worn.
Age	The age of the lifter in Years
Age_class**	The age class in which the lifter competed. Federations may break competition down by different age groups.
Division**	Defines who the lifter is competing against
Bodyweight_kg	The lifter's body weight in kilograms.
Weight_class_kg**	The weight class in which the lifter competed. Federations may break competition down by different weight classes.
squat1kg*	The lifter's first squat attempt in kilograms. Negative means the attempt was failed.
squat2kg*	The lifter's second squat attempt in kilograms. Negative means the attempt was failed.
squat3kg*	The lifter's third squat attempt in kilograms. Negative means the attempt was failed.

squat4kg*	The lifter's fourth squat attempt in kilograms. Negative means the attempt was failed. Not commonly used.
best3squat_kg	The best squat of [Squat1Kg, Squat2Kg, Squat3Kg]. This value will count towards TotalKg
bench1kg*	The lifter's first bench attempt in kilograms. Negative means the attempt was failed.
bench2kg*	The lifter's second bench attempt in kilograms. Negative means the attempt was failed.
bench3kg*	The lifter's third bench attempt in kilograms. Negative means the attempt was failed.
bench4kg*	The lifter's fourth bench attempt in kilograms. Negative means the attempt was failed. Not commonly used.
best3bench_kg	The best squat of [Bench1Kg, Bench2Kg, Bench3Kg]. This value will count towards TotalKg
deadlift1kg*	The lifter's first deadlift attempt in kilograms. Negative means the attempt was failed.
deadlift2kg*	The lifter's second deadlift attempt in kilograms. Negative means the attempt was failed.
deadlift3kg*	The lifter's third deadlift attempt in kilograms. Negative means the attempt was failed.
deadlift4kg*	The lifter's fourth deadlift attempt in kilograms. Negative means the attempt was failed. Not commonly used.
Best3deadlift_kg - Prediction	The best squat of [Deadlift1Kg, Deadlift2Kg, Deadlift3Kg]. This value will count towards TotalKg
total_kg**	The sum of Best3SquatKg, Best3BenchKg, Best3DeadliftKg
place**	The placement of the lifter in the competition
wilks**	Wilks points, often used as a Best Lifter formula. See https://en.wikipedia.org/wiki/Wilks_Coefficient

mc_culloch**	McCulloch points, another formula often used as a Best Lifter formula.
glossbrenner**	Glossbrenner points, another formula often used as a Best Lifter formula.
ipf_points**	IPFPoints, a formula used as a Best Lifter formula for the IPF and its affiliates. See https://www.powerlifting.sport/rulescodesinfo/ipf-formula.html
tested	If the lifter was drug tested or not.
country**	The country of origin for the lifter
federation**	The Powerlifting Federation the meet was hosted by.
date**	Date of the meet.
meet_country**	Country the meet was hosted in
meet_state**	The state the meet was hosted in
meet_name**	The name of the meet.
nevin_ben_shaan	A categorical record to indicate if it is our data, strictly used for filtering and then disposed of.

Appendix B: R-Script

```
# Importing Data ------------------------------------------------------------
plift <- read.csv("openpowerlifting.csv", stringsAsFactors = F) %>%
  clean_names() %>%
  select(name, sex, event, equipment, age, age_class, division, bodyweight_kg, weight_class_kg,
         best3squat_kg, best3bench_kg, best3deadlift_kg, total_kg, place, wilks, mc_culloch, glossbrenner, ipf_points,
         tested, country, federation, date, meet_country, meet_state, meet_name)


# Manipulating Powerlifting Data --------------------------
bad_data <- plift[!complete.cases(plift), ]
plift_1 <- plift[complete.cases(plift), ]
plift_2 <- plift_1 %>%
  mutate(date = ymd(date),
         country = gsub(" ", "", country, fixed = TRUE),
         tested = case_when(
           str_length(tested) > 1 ~ "Yes",
           TRUE ~ "No"
         )) %>%
  filter(date >= ymd("2010-01-01"),
         age > 18,
         country > 1) %>%
  mutate(country =
           case_when(
    str_detect(country, "USSR") ~ "Russia",
    str_detect(country, "WestGermany") ~ "WestGermany",
    TRUE ~ country
  ))



# Write to Excel Sheet ------------------------------------------------------
writexl::write_xlsx(plift_2, "plift.xlsx")
```

OPMA 419 Group Project 2
Benjamin LeBlanc, Shaan Gehlot, and Nevin Sangha
Powerlifting - Deadlift Prediction
Press 'Alt' + 'o' keys (not zero) to view (close) all sections
Press 'Shift' + 'Alt' + 'o' to open all sections

# Libraries --
library(tidyverse)
library(janitor)
library(writexl)
library(Hmisc)
library(lubridate)
library(stringr)

# Set working directory --
setwd("C:/Users/benro/OneDrive/Notes/Winter 2021/OPMA 419/Group Project 2")

# Importing Data --
plift <- read.csv("openpowerlifting.csv", stringsAsFactors = F) %>%
  clean_names() %>%
  select(name, sex, event, equipment, age, age_class, division, bodyweight_kg, weight_class_kg,
         best3squat_kg, best3bench_kg, best3deadlift_kg, total_kg, place, wilks, mc_culloch, glossbrenner, ipf_points,
         tested, country, federation, date, meet_country, meet_state, meet_name)

```
# Manipulating Powerlifting Data --------------------------
bad_data <- plift[!complete.cases(plift), ]
plift_1 <- plift[complete.cases(plift), ]
plift_2 <- plift_1 %>%
  mutate(date = ymd(date),
         country = gsub(" ", "", country, fixed = TRUE),
         tested = case_when(
           str_length(tested) > 1 ~ "Yes",
           TRUE ~ "No"
         )) %>%
  filter(date >= ymd("2010-01-01"),
         age > 18,
         country > 1) %>%
  mutate(country =
           case_when(
    str_detect(country, "USSR") ~ "Russia",
    str_detect(country, "WestGermany") ~ "WestGermany",
    TRUE ~ country
  ))

# Write to Excel Sheet -----------------------------------------------------
writexl::write_xlsx(plift_2, "plift.xlsx")
```

Appendix C: Computational Intensity