A Weightlifter’s Maximum Squat Load Capacity Predicted Using Polynomial Regression

1. Introduction

For many weightlifters, having a goal to work towards is an important part of the sport. However, a weightlifter’s maximum lifting capacity often depends on factors such as age or body weight (a higher muscle mass allows for lifting heavier weights) [Clark, 2021], so there is no universal maximum load goal that applies to everyone. Hence, it would be very useful to get a realistic estimate for a maximum load goal that takes into account such factors. The squat is one of the most important basic lifts in nearly all serious weightlifters’ routines, so it is the exercise that this project will focus on. Hence, in this report, we attempt to predict the maximum amount of weight a person can squat based on their body weight and age.

Section 2 discusses the specifics of the data points used in the project, and section 3 elaborates on the datasets and the models used. Section 4 presents the results of the project and provides discussion about them, and section 5 concludes and summarizes the main points of the project.

2. Problem formulation

The data points in this machine learning problem represent individual weightlifters. The features that characterize the data points are the weightlifter’s body weight and age. The label, so the quantity of interest, is the maximum weight that the weightlifter can squat for one repetition (one squat). This is also called the maximum “load” of the squat.

3. Method

The dataset used in this project is from openpowerlifting.org, which collects data from powerlifting meets (powerlifting competitions). For this project, only data concerning female powerlifters of ages 20 to 40 were used. This choice was made because there are some differences between sexes when it comes to building muscle in strength training, so it is sensible to evaluate sexes individually [Abelsson, 2020]. The age range was chosen because the maximum squat for lifters outside the range is more unpredictable: females typically reach their full strength by 20, so they may not be fully developed before it [Haff and Triplett, 2015], and lifters over 40 may start declining in strength and being more prone to injuries from heavier weights [Expert Advice for the Over-40 Lifter | T Nation, 2018].

The filtered data was split into two sets, SquatData1 consisting of 25 126 data points, and SquatData2 consisting of 17 587 data points. The SquatData1 set was further split into a training set and validation set, consisting of 70% and 30% of the total set, respectively. The SquatData1 set was used to train and validate the models [Jung, 2021]. The SquatData2 set was used as a test set to assess the quality of the model which was chosen through the validation [Jung, 2021].

The data was split into the training, validation, and test sets with the following method: first, the data was put into a random order, then it was divided into SquatData1 (training and validation set) and SquatData2 (test set) by a single split. SquatData1 was further split into separate training and validation sets by another single split.

The information in the following paragraph is based on [Jung, 2021]. The data used in the project is numeric and is characterized by a single feature and a label, so a sensible choice of model is either linear regression or polynomial regression. When looking at the data in a scatterplot it appears to be non-linear, so polynomial regression should yield better results. Polynomial regression uses a hypothesis space consisting of non-linear (polynomial) maps (h(x), ŷ). These polynomials have a maximum degree r, which needs to be chosen according to the data and the desired level of accuracy. The polynomials also have weights which are chosen by minimizing the loss function used in the model.

Since the label is numeric and polynomial regression is used, average squared error loss was chosen as the loss function for the model [Jung, 2021]. The loss function determines the quality of the hypothesis h(x) (also called ŷ) [Jung, 2021]:

$$L=\left(y_{true}-\hat{y}\right)^{2}=\left(y_{true}-h(x)\right)^{2}$$

Squared error loss also has properties (convexity and differentiability) which help find optimal weights for the model efficiently [Jung, 2021]. In addition, polynomial regression and squared error loss can be implemented comfortably in excel, which was the chosen environment for the modelling of this project.

Six different models are compared in this project. Three of them use the weightlifter's bodyweight as the feature to predict maximum squat load, and each has a different polynomial degree: 1, 2 and 3. The three remaining ones use the weightlifter's age as the feature and have different polynomial degrees: 1, 2, and 3 (See Table 1 for a clearer visual of the differences between the models).

4. Results

Table 1: Training error and validation error for each model used. "r" refers to the degree of the polynomial.

	Bodyweight as the feature			Age as the feature		
	r =1	r = 2	r = 3	r =1	r = 2	r = 3
Average training error	1,329.270	1,316.616	1,316.604	1,550.950	1,545.943	1,545.943
Average validation error	1,368.633	1,353.728	1,353.668	1,590.914	1,587.737	1,587.737

From the table it can be seen that the lowest validation error is achieved with bodyweight as the feature and with polynomial degree three (r = 3). This model also resulted in the lowest training error. Hence, the model with bodyweight as a feature and polynomial degree three is the one selected out of all the models. This chosen model was applied on the test set to obtain a performance indicator (test error) [Jung, 2021]. The test error was calculated using average squared error loss again, and it is 1,315.459.

The validation error is higher than the training error, however it is not significantly higher relative to the size of the errors, so overfitting is not implied [Jung, 2021]. This applies to all of the non-selected models as well. In addition, the error on the test set is lower than both the training and validation errors, implying that overfitting is not an issue here.

Overall, each error obtained from the model is quite high so the results could most likely be improved, meaning they are not optimal. The square root of the test set error is 36.26926, and the average true label (maximum squat load of the lifter) is 124.6774. Since average squared error was used, the square root of the error is the difference of the true label and the predicted label. This means that on average, the true label was ±36kg from the predicted label. For a weightlifter, 36kg is not an incredibly big difference, but it is significant.

A specific benchmark for the error was not found, but from personal experience as a weightlifter, it appears that the model is good enough to produce a very rough goal that the lifter can tune according to their knowledge of their own skills. However, the model is not sufficient to produce a concrete goal that the lifter could stick to without making alterations. The results could be improved by using more features that are relevant to lifting, such as protein intake and percentage of muscle mass in the lifter’s body.

In addition, a more powerful solver than the excel one used in this project could be beneficial. The excel solver set all weights to zero for polynomials with degree more than three, which could be due to the solver not being powerful enough for the very large amount of data it had to go through. Hence, a more powerful solver (or computer) could ensure that the optimal solution was reached. Another model, such as deep learning, could also be better suited for the large amount of data used in the project [Jung, 2021].

5. Conclusion

In this project, six different models were compared by their ability to predict a weightlifter’s maximum squat load. The models differed in the features used and the polynomial degree (Table 1). The model chosen through single split validation had body weight as a feature and polynomial degree three. This yielded training error 1,316.604 and validation error 1,353.728, so it does not seem to overfit much if at all [Jung, 2021]. The model was applied to a test set, which yielded a test error of 1,315.459, which is smaller than both the training and validation error.

A benchmark was not found for the error level, but from personal experience as a weightlifter, the model seems to produce results which can be useful, but are not optimal. Future work on this project could include using more relevant features, more powerful tools, and possibly an entirely different model such as an Artificial Neural Network[1]. In addition, as time passes, more data is added to openpowerlifting.org meaning more useful data is available to improve the model.

[1] Jung, A., 2021. *Machine Learning: The Basics*. [ebook] Available at: <http://mlbook.cs.aalto.fi> [Accessed 25 March 2021].

Bibliography

(1) Clark, P., 2021. *Muscle mass and strength, any correlation? | The Training Station Gym*. [online] Training Station Gym. Available at: <https://phillytrainingstation.com/blog/muscle-mass-and-strength-correlation/> [Accessed 25 March 2021].

(2) Abelsson, A., 2020. *Sex Differences in Strength and Muscle Mass: Do Males and Females Gain the Same? – StrengthLog*. [online] StrengthLog. Available at: <https://www.strengthlog.com/do-males-and-females-gain-the-same/#easy-footnote-bottom-8-3161> [Accessed 25 March 2021].

(3) Haff, G. and Triplett, T., 2015. *Essentials of Strength Training and Conditioning*. 4th ed. [ebook] p.Chapter 7: Age- and Sex-Related Differences and Their Implications for Resistance Exercise. Available at: <https://www.open.edu/openlearn/ocw/pluginfile.php/617068/mod_resource/content/1/e217_1_excf223_nsca_chapter7_p144_145.pdf> [Accessed 25 March 2021].

(4) T NATION. 2018. *Expert Advice for the Over-40 Lifter | T Nation*. [online] Available at: <https://www.t-nation.com/training/expert-advice-for-the-over-40-lifter> [Accessed 25 March 2021].

(5) Jung, A., 2021. *Machine Learning: The Basics*. [ebook] Available at: <http://mlbook.cs.aalto.fi> [Accessed 25 March 2021].

(6) Openpowerlifting.org. n.d. *Powerlifting Rankings*. [online] Available at: <https://www.openpowerlifting.org> [Accessed 25 March 2021].