CS-3410 - 1 - Spring 2025– Mid-Term Project

Kudakwashe Chakanyuka, kudakwashe.chakanyuka_ug25

Contents

1 Introduction

Predictive modeling has revolutionized multiple fields by enabling data-driven insights and enhanced decision-making, and the domain of sports science and fitness is no exception. Machine learning techniques provide an exceptional opportunity to forecast performance metrics, tailor training regimens, and ultimately optimize fitness outcomes. Among these applications, strength training stands out as a foundational element of physical well-being, emphasizing compound lifts like the squat, bench press, and deadlift. These exercises are not only critical for building muscle and endurance but also for improving overall functional fitness and resilience.

Achieving success in strength training depends heavily on understanding the interplay of various factors—demographic data, personal attributes, and training history—that influence performance. For instance, variables like age, body weight, experience level, and prior lifting records all contribute to shaping an individual's strength potential. This study aims to bridge the gap between traditional training practices and data-driven insights by developing advanced predictive models.

By employing decision trees and complementing them with bagging and ensemble methods, this research seeks to create a robust tool that goes beyond mere predictions. It aims to empower gym-goers and trainers with actionable recommendations, minimizing guesswork and reducing risks of injuries caused by overloading or undertraining. The interpretability of decision trees makes them especially suitable for this task, offering not just accurate predictions but also transparency in understanding the underlying relationships between key features.

2 Literature Review

The approach used in this study is rooted in insights from Professor Tamara Boderic's notes, emphasizing fundamental machine learning concepts that are highly applicable to predictive modeling tasks. The techniques employed—decision trees, bagging, and ensemble methods like Random Forests—are chosen for their ability to handle complex datasets, provide interpretability, and enhance prediction accuracy. Each method contributes uniquely to the modeling process, and their combined application ensures a robust framework for predicting strength performance.

Decision Tree Splitting Criterion

Decision trees form the foundation of this study's methodology due to their intuitive structure and effectiveness in capturing relationships between input features. The splitting criterion, governed by Gini impurity, measures the quality of splits in classification tasks. Defined as:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2, \tag{1}$$

where p_i represents the proportion of samples in class i. This formula helps identify splits that minimize impurity, ensuring that the resulting subsets of data are more homogeneous in their outcomes. Decision trees are particularly well-suited for this project because they can effectively model non-linear relationships and interactions between features, such as age, body weight, and training frequency, which are critical predictors of strength performance.

Bagging (Bootstrap Aggregating)

Bagging is employed as a technique to reduce variance and enhance the stability of predictive models. By training multiple decision trees on different bootstrapped subsets of the data and averaging their predictions, bagging mitigates the risk of overfitting that often occurs with single decision trees. The final prediction is calculated as:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x), \tag{2}$$

where B is the number of models, and $f_b(x)$ is the prediction from the b-th model. In the context of this study, bagging ensures that the model generalizes well across diverse gym-goer profiles, including novices, intermediates, and advanced lifters. By leveraging ensemble predictions, the model becomes more robust and consistent in estimating one-rep max (1RM) values.

Random Forest Estimation

Random Forest further extends bagging by introducing randomness in the selection of features during tree construction, which reduces correlation between trees and improves overall prediction accuracy. The final prediction for Random Forest is given by:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i(x), \tag{3}$$

where n is the number of decision trees, and $T_i(x)$ represents the prediction from the i-th tree. This technique is particularly beneficial for this project because it provides reliable predictions even when certain features have varying importance. For example, the model can account for demographic attributes like age and gender, which may influence strength performance differently across individuals.

Why These Methods Matter

By combining decision trees, bagging, and Random Forests, the approach adopted in this study ensures a balance between accuracy, interpretability, and robustness. Decision trees provide a clear understanding of how each feature contributes to predictions, while bagging and Random Forests enhance model stability and prevent overfitting. These methods are especially valuable for handling the diverse attributes present in the datasets, such as training frequency, body composition, and prior performance metrics.

Furthermore, the use of these techniques aligns with the objectives of this project—to develop a predictive model that not only estimates strength performance with precision but also offers gym-goers actionable insights to optimize their training. By employing these proven machine learning methods, the study bridges the gap between theoretical concepts and practical applications, making it a cornerstone of evidence-based fitness recommendations.

3 Dataset Source and Description

To build a robust predictive model for strength performance, three distinct datasets were utilized. Each dataset provided unique insights and captured a broad spectrum of gym-goer demographics, ranging from elite athletes to beginners. The combination of these datasets facilitated the development of an inclusive and generalizable model, ensuring comprehensive coverage of varying fitness levels and attributes.

i) Powerlifting Rankings Dataset

This dataset, obtained from `https://www.powerliftingrankings.com`, provides detailed performance statistics from professional powerlifting competitions. It is particularly valuable for its focus on elite athletes and their accomplishments in competitive strength sports. The dataset includes:

- **Demographic Data:** Information such as age, gender, body weight, and height, essential for understanding how these attributes correlate with strength performance.
- **Performance Metrics:** Results of major lifts, including the squat, bench press, and deadlift, recorded in competition settings. These metrics represent the maximum weights achieved under standardized conditions.
- **Training Trends:** Details on progression over time, offering insights into how athletes improve their performance with structured training regimens.

This dataset serves as a benchmark for the upper limit of human strength potential, providing critical insights into factors contributing to peak performance.

ii) Kaggle Dataset

The Kaggle dataset offered a broader perspective by focusing on individuals with varying levels of fitness expertise. The data, sourced from multiple CSV files, captures insights into performance metrics across intermediate and recreational gym-goers. This dataset is particularly useful for understanding trends among non-elite participants and bridging the gap between professional athletes and general fitness enthusiasts. Key features extracted include:

- **Training Habits:** The dataset includes information on training frequency (e.g., days per week) and session duration, offering insights into workout consistency and types of exercises performed.
- **Performance Data:** Historical records of performance are present, such as estimated one-rep max (1RM) values for major lifts (squat, bench press, and deadlift) and detailed logs of sets, reps, and weights lifted.
- **Self-Reported Metrics:** Data points such as the Rate of Perceived Exertion (RPE) are included, providing a subjective measure of an individual's effort levels and fatigue during training sessions.
- **Demographics:** Features like age, gender, body weight, and height allow for further segmentation and analysis of strength performance trends across diverse groups.

The data from Kaggle enhances the breadth of the study by presenting a wide range of participants, ensuring that the model is applicable to audiences with varying levels of fitness expertise.

iii) General Fitness Metrics Dataset

The third dataset included individuals with little to no prior strength training experience, sourced from files such as `openpowerlifting.csv`. This dataset broadens the diversity of the training data and ensures that the model generalizes well across multiple experience levels. Key features of this dataset are:

- **General Health Metrics:** Data points like BMI, muscle mass percentage, resting heart rate, and blood pressure help in understanding the broader health context that may affect strength performance.
- **Cardiovascular Performance:** Metrics such as VO2 max, which offer insights into baseline fitness levels, are included to evaluate participants' overall physical readiness.
- **Baseline Strength Data:** Information on light lifting exercises or assessments, such as bodyweight squats and grip strength, provides starting points for measuring progress and modeling performance.
- **Demographics:** Attributes such as age, height, gender, and weight ensure that the model accounts for variability among participants.

This dataset introduced variability in the data, allowing the model to account for diverse profiles and predict outcomes for novice gym-goers.

4 Data Exploration and Feature Selection

The datasets collectively contained approximately 24,000 entries, representing individuals with varying levels of experience and fitness. During preprocessing, several steps were undertaken to ensure the quality and consistency of the data:

- Common features such as age, gender, body weight, and strength performance metrics were standardized across the datasets to maintain consistency.
- Missing values, particularly in attributes such as age and height, were imputed to reduce data sparsity and enable effective analysis.
- Categorical variables, such as workout intensity levels, were encoded to ensure compatibility with machine learning algorithms.
- Outliers, especially from the Powerlifting Rankings Dataset, were carefully identified and handled to prevent skewing the dataset towards elite-level performance exclusively.

Through this preprocessing, a unified dataset was created, offering a comprehensive view of strength performance across different demographics and training levels. Key attributes retained for modeling include:

- **Demographics:** Age, gender, body weight, height, and BMI (Body Mass Index).
- **Training History:** Experience level, training frequency, workout intensity, and exercise types.
- **Performance Metrics:** One-rep max (1RM) values for major lifts such as squat, bench press, and deadlift, along with the number of sets and reps performed, and self-reported Rate of Perceived Exertion (RPE).

As part of the exploratory analysis, a correlation matrix was computed to identify relationships among features and their impact on strength performance. Notable correlations were observed between body weight and one-rep max values for major lifts, as well as between age and training

frequency. The dataset exhibited a near-normal distribution in strength performance metrics, further validating its suitability for predictive modeling.
This comprehensive exploration and feature selection process ensures that the dataset captures a wide range of patterns and trends, facilitating accurate and actionable predictions for individuals across all levels of fitness expertise.

5 Methods

This study focused on modeling strength performance for the three most fundamental compound exercises: squat, bench press, and deadlift. These exercises form the backbone of most strength training regimens and provide a comprehensive measure of overall physical performance. Each exercise contributes uniquely to overall strength:

- **Squat:** Primarily targets the lower body, engaging muscles such as the quadriceps, hamstrings, glutes, and calves. It also requires stabilization from the core and upper body, making it a critical exercise for building overall power and lower-body strength.
- **Bench Press:** Focuses on upper body pushing strength by working the pectoral muscles, triceps, and shoulders. As a primary indicator of upper body strength, it is a common benchmark in strength assessments.
- **Deadlift:** Evaluates total-body strength, emphasizing posterior chain activation, including the glutes, hamstrings, spinal erectors, and grip strength. It is a full-body movement, incorporating both pulling and stabilizing mechanics.

Machine learning techniques were employed to predict expected performance for each exercise based on a variety of user attributes. This predictive modeling process involved the following steps:

i) Feature Engineering

Key features were derived from the unified dataset to serve as inputs for the machine learning models. These features included:

- **Demographic Attributes:** Age, gender, body weight, and height were incorporated to account for physiological differences that influence strength capacity.
- **Training History:** Attributes such as years of experience, training frequency, and workout intensity were utilized to capture the impact of training habits on strength development.
- **Performance Metrics:** Historical one-rep max (1RM) records and self-reported metrics such as Rate of Perceived Exertion (RPE) were critical in determining baseline performance levels and progress over time.

ii) Target Variables

The predictive models were trained to estimate one-rep max (1RM) values for each exercise. This metric was chosen as it represents the maximum weight that can be lifted in a single repetition, serving as a standardized measure of strength performance across diverse populations.

iii) Model Training

Supervised machine learning techniques were applied to train individual models for each exercise. Decision trees were selected as the primary model due to their ability to handle non-linear relationships and interactions between features. Additional techniques such as bagging and Random Forest ensembles were also utilized to improve the robustness and accuracy of predictions. These ensemble methods were particularly effective in reducing overfitting and capturing complex feature interactions.

iv) Model Evaluation

Each model was evaluated using cross-validation techniques to ensure reliability and generalization. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared were computed to assess the accuracy and performance of the models. The evaluation process highlighted the effectiveness of ensemble methods in enhancing predictive accuracy compared to individual decision trees.

v) Insights and Implications

The models provide actionable insights into how demographic, training, and performance attributes influence strength outcomes. For instance, significant correlations were observed between body weight and squat performance, as well as between training frequency and bench press capacity. These insights can be utilized to develop tailored training recommendations for individuals based on their unique attributes.

By engaging in these machine learning techniques, the study achieved robust predictions for strength performance across the three fundamental exercises, offering a data-driven approach to optimizing training regimens.

6 Experimentation

To develop accurate predictive models for strength performance, Random Forest Regression was employed as the primary machine learning technique. This approach was selected for its ability to handle non-linearity and interactions between features, making it well-suited for the complex relationships inherent in strength training data. Each exercise—squat, bench press, and deadlift—was modeled separately to account for their unique characteristics and performance dynamics.

Evaluation Metrics

The models were evaluated using a variety of performance metrics to ensure accuracy and reliability. These metrics include:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions, providing a straightforward interpretation of model performance. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

  where y_i represents the actual value, $\hat{y}_i$ the predicted value, and n the total number of observations.

- **Mean Squared Error (MSE):** Penalizes larger errors more heavily by squaring the difference between predicted and actual values. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5}$$

- **Root Mean Squared Error (RMSE):** Represents the square root of the MSE and provides an error measure in the same units as the target variable:

$$RMSE = \sqrt{MSE}. \tag{6}$$

- R^2 **Score:** Indicates how well the model explains the variance in the data, with values closer to 1 representing better performance. It is calculated as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \tag{7}$$

  where $\bar{y}$ is the mean of the actual values.

Exercise-Specific Model Training

Separate Random Forest models were trained for each exercise: squat, bench press, and deadlift. This division allowed the models to focus on the unique factors influencing performance for each lift. The squat model, for example, prioritized variables such as body weight and lower-body training frequency, while the bench press model emphasized upper-body attributes like shoulder strength and chest muscle activation. The deadlift model, on the other hand, analyzed total-body strength and posterior chain efficiency.

Cross-Validation and Generalization

To ensure robust predictions, cross-validation techniques were utilized during model training. This involved splitting the data into multiple folds, training the model on one subset, and validating it on another. By rotating through all subsets, this approach minimized bias and variance, ensuring the models could generalize effectively to unseen data.

Overall, the performance evaluation revealed that Random Forest models excelled at capturing complex feature interactions across all three exercises. For example:

- Strong correlations were observed between body weight and squat 1RM values, emphasizing the role of mass in lower-body strength.
- Training frequency and experience levels emerged as significant predictors for bench press performance, showcasing the impact of consistency on upper-body strength.
- The deadlift model highlighted posterior chain development and grip strength as critical contributors to lifting capacity.

These findings underscore the importance of tailored training regimens and demonstrate the potential of machine learning to offer personalized performance predictions. By addressing the unique demands of each lift, the models provide actionable insights that can enhance training outcomes across varying fitness levels.

7 Results and Analysis

The strength performance multipliers presented in the table were derived based on calculations and existing research on strength training benchmarks. These multipliers provide a framework for estimating expected strength performance across different exercises, serving as a standardized metric for comparing individuals at varying skill levels.

Exercise	Beginner Multiplier	Advanced Multiplier
Squat	0.75	1.25
Bench Press	0.80	1.20
Deadlift	0.70	1.30

Table 1: Strength Multipliers for Prediction

Implications of Strength Multipliers

The multipliers represent ratios between a lifter's body weight and the weight they are expected to lift. These values are categorized for beginners and advanced lifters to account for differences in training experience, physiological adaptation, and neuromuscular efficiency.

- **Squat:** For beginner lifters, a multiplier of 0.75 indicates that they are expected to lift 75% of their body weight. As skill level advances, the multiplier increases to 1.25, reflecting the enhanced strength and lower-body power gained through consistent training and improved technique.
- **Bench Press:** The bench press values demonstrate a slightly higher baseline multiplier for beginners (0.80), as the movement primarily involves the upper body and often requires less stabilization compared to the squat or deadlift. Advanced lifters achieve a multiplier of 1.20, signifying significant strength development in the chest, shoulders, and triceps.
- **Deadlift:** With a starting multiplier of 0.70, the deadlift begins at a lower baseline for beginners, as it engages a complex combination of muscle groups and requires mastery of form to optimize lifting capacity. Advanced lifters, however, can achieve a multiplier of 1.30, demonstrating the profound impact of posterior chain training and increased proficiency in the movement.

Application of Results

These multipliers provide a practical tool for evaluating strength performance in real-world scenarios. Fitness practitioners and trainers can utilize these benchmarks to establish realistic performance goals for gym-goers based on their body weight and experience level. For example:

- A beginner weighing 70 kg should target approximately 52.5 kg for squats (70 kg $\times$ 0.75), while an advanced lifter of the same weight might aim for 87.5 kg (70 kg $\times$ 1.25).
- Similarly, for the bench press, the same beginner might aim for 56 kg (70 kg $\times$ 0.80), whereas an advanced lifter could strive for 84 kg (70 kg $\times$ 1.20).

Overall, the differences in multipliers reflect the unique demands of each exercise:

- The squat involves a significant reliance on leg strength and core stability, requiring progressive overload and technique refinement for improvement.
- The bench press emphasizes upper-body pushing strength, making it relatively more accessible for beginners compared to the squat or deadlift.
- The deadlift, despite its high potential for strength gains, presents initial challenges due to its requirement for whole-body coordination and posterior chain engagement.

These insights emphasize the importance of tailoring training programs to individual strengths and weaknesses. By setting achievable targets based on these multipliers, lifters can track progress effectively while minimizing the risk of injury. This structured approach to performance evaluation ensures that individuals at all fitness levels can develop their strength systematically and sustainably.

User Interaction Model

The prediction model offers an interactive experience where users can input specific demographic and physical attributes to receive tailored predictions. The required inputs include:

- **Age:** Essential for accounting for physiological differences and strength adaptation over time.
- **Body Weight:** A critical factor in calculating strength performance, as it directly influences the expected load capacity.
- **Height:** Provides context for body mechanics and leverage, which can impact performance in compound lifts.
- **Gender:** Acknowledges physiological variations that may affect strength outcomes and training adaptations.
- **Disability Status (if applicable):** Ensures inclusivity by adapting predictions to accommodate individuals with unique physical conditions or constraints.

Once these inputs are provided, the model processes the data using trained regression algorithms to predict expected performance levels for squat, bench press, and deadlift. These predictions leverage the comprehensive dataset and machine learning models to deliver accurate, personalized insights for users at various fitness levels. The interactive functionality emphasizes usability, allowing gym-goers and trainers to make informed decisions about strength training targets based on individual attributes.

8 Conclusion

This study has effectively demonstrated the capability of machine learning techniques in enhancing strength performance prediction, offering a data-driven approach to understanding and optimizing powerlifting outcomes. By employing methods such as decision trees, bagging, and Random Forest regression, the analysis revealed significant relationships between demographic attributes, training history, and performance metrics. The results confirm that variables such as age, body weight, BMI, and gender serve as pivotal factors influencing powerlifting performance. These insights align with established principles in sports science while also providing new opportunities for predictive modeling in the fitness domain.

The predictive models built for squat, bench press, and deadlift captured the unique physiological and mechanical demands of each exercise. A comprehensive dataset, formed through the integration of multiple sources, enabled the inclusion of diverse user profiles, from beginners to elite athletes. This holistic representation ensured the models' applicability across a broad spectrum of fitness levels.

The research also emphasized the importance of interactive and user-friendly tools, as demonstrated through the user interaction model. By allowing users to input attributes such as age, body weight, and training frequency, the model delivers personalized predictions tailored to individual fitness profiles. This functionality provides practical benefits, enabling gym-goers and trainers to make informed decisions about setting realistic strength training goals and monitoring progress over time.

Future Directions

While the current study showcases the potential of machine learning techniques for predicting strength performance, there are opportunities for further refinement and expansion:

- **Incorporation of Biomechanical Factors:** Future models could integrate additional variables such as joint angles, force production metrics, and muscle activation patterns to enhance prediction accuracy and provide deeper insights into the biomechanics of strength training.
- **Larger and More Diverse Datasets:** Expanding the dataset to include more participants from varied backgrounds and fitness levels could improve model generalizability and robustness.
- **Real-Time Predictive Systems:** Developing real-time applications that incorporate wearable technology data could offer immediate feedback and recommendations during training sessions.
- **Advanced Machine Learning Techniques:** Exploring deep learning or hybrid approaches that combine heuristics with machine learning algorithms may further optimize predictive performance.

Overall, this document has outlined a comprehensive approach to modeling strength performance through machine learning techniques. From data exploration and feature selection to model training and user interaction, each component contributed to the development of a robust and practical system for predicting strength outcomes. These findings hold promise for enhancing strength training practices, advancing the integration of data-driven methodologies in fitness, and paving the way for innovative solutions in sports science and personalized fitness coaching.

9 References

References

[1] Andrew Ng. *CS229 Lecture Notes*. Stanford University. Available at: https://cs229.stanford.edu/lectures-spring2022/main_notes.pdf

[2] Prof. Tamara Broderick. *6.036/6.862: Introduction to Machine Learning*. Massachusetts Institute of Technology. Course website: https://introml.odl.mit.edu

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. *OpenAI Gym*. 2016. Available at: https://github.com/openai/gym

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd Edition. MIT Press, 2018.

[5] The New York Times Company. *Wordle*. 2022. Available at: https://www.nytimes.com/games/wordle/index.html

[6] Alex Newgent. *Wordle Environment for Reinforcement Learning*. Available at: https://github.com/alex-nooj/wordleenv

[7] Pranay Agarwal. *Wordle-AI*. Available at: https://github.com/pranay1208/wordle-ai

[8] K.A. Brown. *Model, Guess, Check: Wordle as a Primer on Active Learning for Materials Research*. npj Computational Materials.

[9] Michael Bonthron. *Rank One Approximation as a Strategy for Wordle*. arXiv preprint arXiv:2204.06324, 2022.