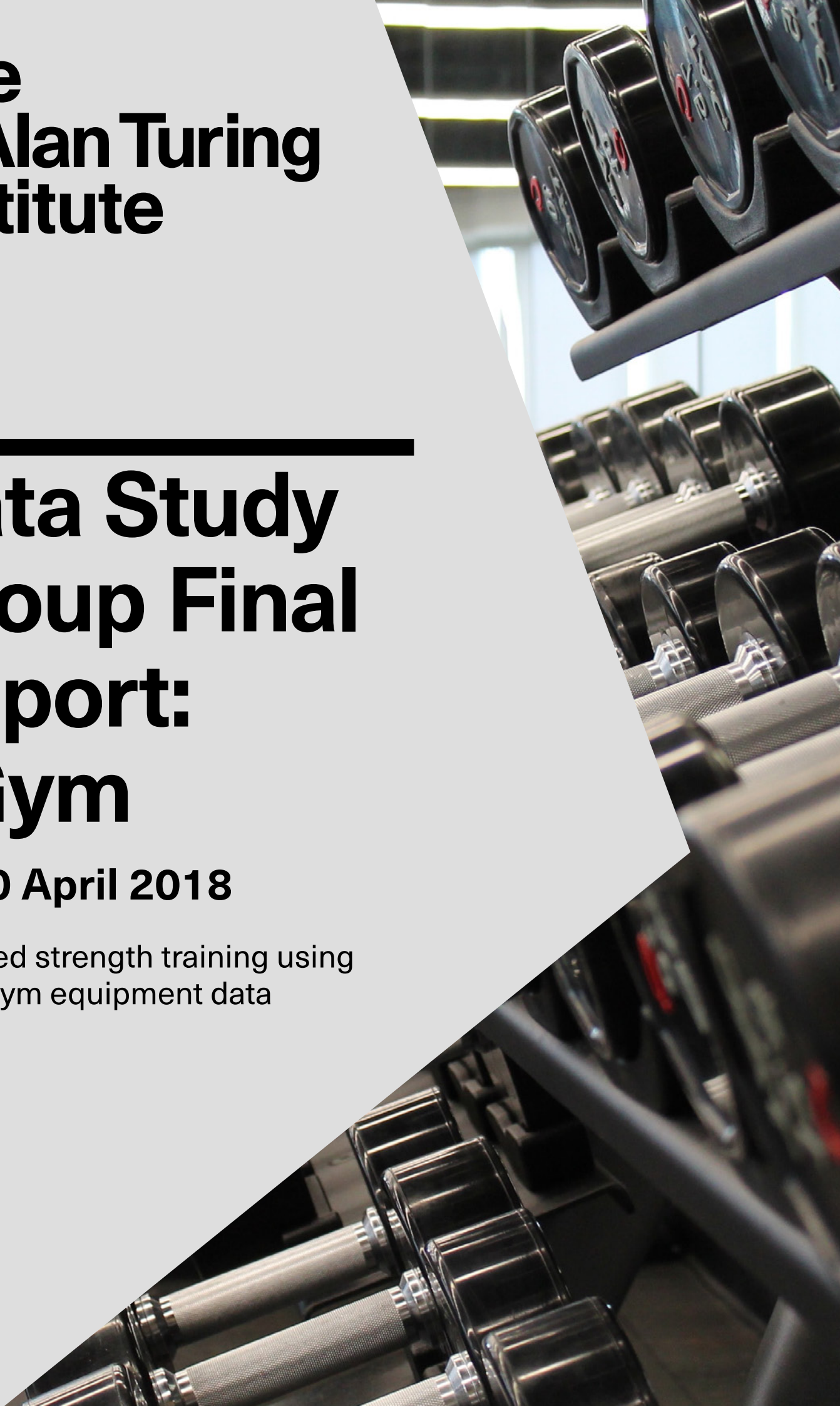
The Alan Turing Institute

Data Study Group Final Report: eGym

16-20 April 2018

Improved strength training using smart gym equipment data



https://doi.org/10.5281/zenodo.3755606

https://www.overleaf.com/project/5bbf54c584b05b2ee364cef2

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Contents

1 Executive Summary

1.1 Challenge Overview

We were provided with information about users of specialised smart gym equipment and usage data over a period of 2 years. The brief was to explore the data and provide insight into how best to help users improve performance. Of particular interest was the link between training and increased strength (as measured by the machines). Also of interest were longitudinal usage of gym-goers, with the view of aiding adherence to the program, and the creation of a simple performance metric by which user improvement could be measured.

1.2 Data Overview

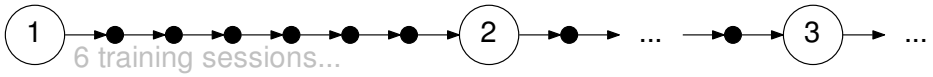


Figure 1: The suggested training schedule for users. Training sessions are represented as points; they typically occur on separate days. The measurement sessions are represented as numbered circles. Each measurement is preceded by 6 training sessions (after the initial measurement).

The data were taken from specialised smart gym equipment. Users undertake a regular training schedule. Each gym has a suite of approximately 9 of the total 18 machines available. When users begin the training program, a personal trainer takes them round the circuit of machines to take their initial strength measurements. After this, the user is encouraged to undertake the training program of 6 sessions. Each session should involve the user using each of the machines in the circuit at least once. After these training sessions are complete, the user undergoes another strength measurement session. See the Figure 1 for an illustration of this process.

1.3 Main Objectives

The main objectives were to:

- Provide insight into factors underlying strength improvement
- Identify trends in user behaviour
- Uncover factors which lead to increased adherence to the training program and improvement
- Create a metric by which users can understand their improvement

These insights would be provided to the gyms, personal trainers, or the users themselves. The problem of measuring user improvement is non-trivial because, aside from each machine having a different range and variance of recorded values (both with respect to different users and the same user in subsequent sessions), different users use a different subset of machines (either by choice or due to the selection of machines their gym has).

1.4 Approach

We performed an extensive exploration of the data, and formulated a prediction problem to understand what factors influence a user's strength measurement for a given machine. We used user demographic information to predict their first ever recorded measurement values for the machines.

1.5 Main Conclusions

The results of the initial strength measurement prediction look very promising. We used a simple latent variable model which represents the user strength as a single number dependent on the user demographic information.

We find that the distribution of strength measurements for each machine is approximately log normal *when broken down by gender*. We can thus normalise strength values for each machine (making them somewhat

comparable) by simply log transforming the values, fitting a single normal distribution to each gender, then subtracting the mean and dividing by the standard deviation of the fitted normal.

There is a clear pattern between strength measurements on different machines, roughly corresponding to main muscle groups (arm, legs, core, hips). Therefore, measurements within the same muscle group should be good predictors of the strength on other machines.

Autocorrelation between subsequent measurements is very high (Pearson coefficient above 0.9) and roughly linear. This implies that simple autoregressive models, or even predicting the previous measurement should give a good estimate of the next measurement.

Improvement between successive measurements is linked to the intensity of the training between the measurements, but the effect saturates after 2-3 trainings per week. This supports the general knowledge that regular training with resting time in-between is ideal for strength improvement.

When modelling the strength data, the 'base torque' issue (discussed in the Data Quality Issues section), must be accounted for. These values could be discarded, but we recommend modelling it as opposed to cleaning it out of the data – the issue could be fixed if machine software is updated.

1.6 Limitations

We found a pervasive issue with strength measurement data: many users record a very similar and low value for each machine. It is postulated that this is related to 'base torque' a user must attain to have their measurement recorded. However, we also find many measurements below this value.

The training data was aggregated at a weekly level, and therefore lacked the granularity to undertake detailed analysis and modelling of the impact of individual training sessions.

1.7 Recommendations and Further Work

We mainly addressed the task of creating a model for user strength, by which users could track their progress. There remains work to do on other areas of interest: relationship between training and strength, a model of user adherence to the training program, and a model of user churn.

Features created in the Exploratory Data Analysis section to encapsulate user consistency were relatively rudimentary (the mean and standard deviation of the user's number of gym visits per week). Future work may seek to better define concept of consistency, perhaps by looking at the number of consecutive non-zero weeks for each user (i.e. streaks). We recommend looking to other products, for instance language learning applications such as duolingo[1] for inspiration. We also did not make use of these features for modelling.

The model of latent strength can be expanded in several different ways: increase the number of latent variables, at the very least we think that two or three are justified and could be designed to affect upper body, lower body, and core body strength machines; provide more data to the model, for example customer height, weight, and body fat percentage; expand the model to predict not only the current strength measurements, but predict the next measurement. This may require the estimation of another latent variable, perhaps representing 'training effort' (a larger value implying a larger improvement).

The connection between training intensity and strength improvement could be further explored. Smoothing the erratic individual strength improvement profiles would clarify the temporal pattern and help highlight where the training effect saturates improvement. Machine-specific training data would be needed to study differences in the connection between different machines – we would like to investigate the cross-effects and causality between training on one machine and improvement on the other.

No work has been undertaken focusing on the type of training, i.e. positive or negative. Future work should look to explore the impact of different training types on strength improvement.

[1] https://www.duolingo.com/research

2 Quantitative Problem Formulation

Multiple quantitative problems were proposed, we mainly address the first and proved preliminary findings for the latter:

1. Prediction of a given user's initial strength based on their demographics
2. Prediction of the next strength measurement given the previous measurements and training information
3. Identifying exercise machines for which a given user's strength measurements will correlate (i.e. clustering exercise machines)
4. Longitudinal analysis of users' recorded strength values

It is reasonable to assume that a gym member's measured strength on each machine is related to how frequent and intense each training session is, as well as inherent attributes such as age and gender. Therefore, it is probable that, given the data we have available about their training history, we can predict a random gym member's performance on a particular machine type. We investigate the accuracy with which we can do this.

In the Exploratory Data Analysis section below, we find that strength growth plateaus against age and number of training sessions. We therefore formulate a regression problem with the dependent variable of a user's recorded strength on a given machine having a non-linear relationship with independent variables of demographic information (and other aggregated training metrics).

Details of feature selection, model selection, and error metrics are given in the respective subsection within the Experiments section below.

3 Data Overview

eGym provided three main datasets providing information about the demography of users, their strength measurements, and their weekly training activity. eGym advise that users perform a strength measurement session after every six training sessions, but the user has the freedom to do this whenever they like.

There were about 1 million users in the database, and 18 machines in total. We created a subset to use for preliminary analysis and modelling. We randomly selected 100 000 users who were both:

1. present in the training and the measurement data and
2. present in the demographic dataset with a non-missing value for gender

Unless otherwise stated, all plots in this section and in the following Experiments section use this subset of the data.

3.1 Dataset Description

DEMOGRAPHICS

user_id	gender	date_of_birth

TRAINING

user_id	weekName	dtype	year	weekNumber	machineLocation_id	trainedMachines	trainingsPerWeek

STRENGTH

user_id	strength_id	value	timestamp	machineTypeProduction	machineLocation_id	week_no	year_no	weekName

Figure 2: A schema of the data. The **TRAINING** dataset summarises the training information per user up to the week level. The **STRENGTH** dataset records the values attained by each user from each strength measurement on each machine. See Figure 1 for an illustration of how strength and training sessions are conducted. Details about variables are given in the Dataset Description section.

We received three datasets (illustrated in Figure 2); they are linked by a gym user identifier, `user_id`. Below we detail the other variable information:

1. Demographic Information:
    - `gender` : Male or female - either self reported or filled in by the gym
    - `date_of_birth` : User's self reported DOB

2. Training:
    - `weekName` : The week number within the year i.e. between 1 and 52 inclusive
    - `dtype` : Type of training (positive or negative)
    - `year` : The year of the training week
    - `weekNumber` : The number of weeks since the user joined *N.B. not to be confused with* `weekName`
    - `machineLocation_id` : The identifier of the gym being used
    - `trainedMachines` : Number of machines used this week
    - `trainingsPerWeek` : Number of training sessions undertaken this week
3. Measurement:
    - `strength_id` : An identifier for this specific measurement
    - `value` : The strength value recorded
    - `timestamp` : The time, date, and year of the measurement
    - `machineTypeProduction` : Machine identifier, a number between 1 and 18
    - `machineLocation_id` : The identifier of the gym being used
    - `week_no` : The number of weeks since the user joined *N.B. not to be confused with* `weekName`
    - `year_no` : The year of the training week
    - `weekName` : The week number within the year i.e. between 1 and 52

The demographic information is recorded by different methods: gender is recorded either by the gym that the user is a member of, or directly on one of the machines. The date of birth is recorded by the application the user uses to interact with their data e.g. the iPhone app, or website.

The training information is aggregated up to a week level. This means we have absolute counts of the number of training sessions each user conducted per week, how many different machines they used, and whether the training was positive or negative. Positive training involves working muscles in the normal way (e.g. like lifting a weight), whereas negative training applies resistance in the opposite direction (e.g. lifting weights in anti-gravity). Machine-specific information on training activity is available from eGym, but is a considerably larger dataset - it was not available for this project.

The strength measurements data is not aggregated, we have a value for a given user and for each machine they used at every point in time they used it. Time for this data was given as a full datetime stamp - the concept of a 'measurement session' was left to us to define, but it was suggested that we aggregated up to the week level and took the maximum.

N.B. When we refer to 'strength data' or 'strength measurements' throughout the report, we are referring to data from the Measurement table. We make very little reference to the training data.

3.2 Data Summary

Figure 3 shows the distribution of user ages in the subset of data. Figure 4 shows a histogram of all the strength values for all machines, and Figure 5 breaks this down by gender.

See Figure 9 for an illustration of how strength measurements change over time.

See Figure 15 for the names of the machines used and an illustration of how they are related.

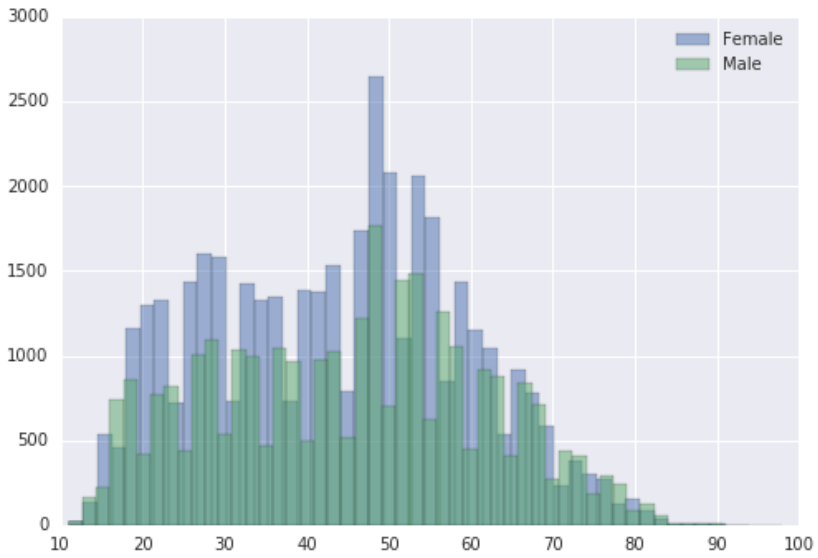


Figure 3: This plot is a histogram of user ages. Age is shown on the x axis, and the count of the bin on the y axis. Counts for males are shown in green, females in blue.

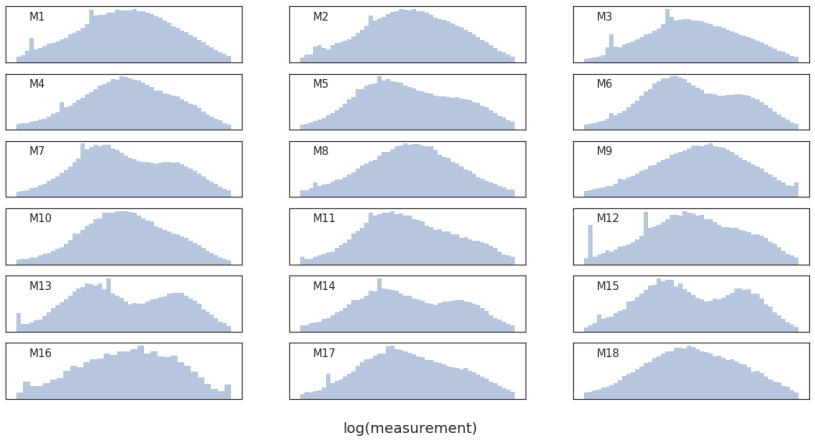


Figure 4: Histograms of all strength measurement value data (having undergone a log transform) for each machine. We observe that some distributions are clearly bimodal. In Figure 5 we see this is explained by gender.

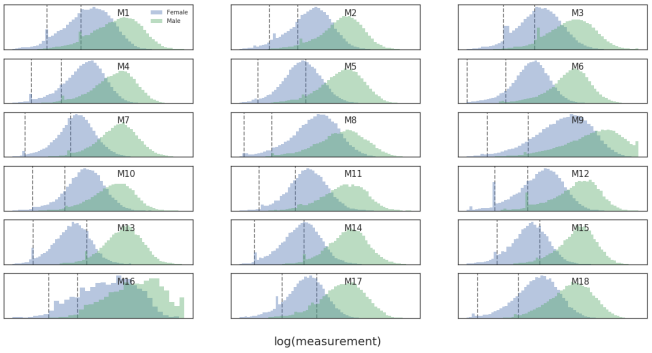


Figure 5: The distribution of of all strength measurements (having undergone a log transform) on each machine, split by gender (men are shown in green, women in blue). In all machines, the distribution of each gender is unimodal, and in general men produce larger measurements than women. Dotted lines indicate the 'base torque' a user must produce in order to receive a reading. For further discussion please see the Data Quality Issues section. The plots are numbered left-to-right, top-to-bottom from machine M1 to M18.

3.3 Data Quality Issues

From the 989 022 users, 9 294 had a missing gender and 312 052 a missing date of birth. There were also users with dates of birth which were clearly incorrect – there were instances of users less than 16, or older than 100, and many with the birth date 01/01/1970 - a default value (it is unconfirmed whether the user is mandated to provide a data of birth and whether the process is the same using the mobile application or the website).

There were duplicated rows in the training data: in our initial subset of 100 000 users out of the roughly 1 million, 1 186 370 out of 4 872 902 rows are duplicates(roughly 24%). We removed them.

One of our goals was to model the distribution of strength measurements. However, when we looked at the distribution of these on each machine type, we saw some unusual features. The overall distribution was smooth, but with some sharp peaks. This was common to virtually all of the machine types.

We verified the reality of these peaks (it could have been a function of the

histogram bin size) by fitting a Gaussian mixture model to the data, which assigned a distribution with a large mixture component at the locations of the peaks. Once we were confident that the effect was real, we informed eGym. They looked into it, and suggested that the measurements could be related to gender-specific '**base torques**' on each machine. We verified that the peaks were indeed related to these values. This can be seen in Figure 5 where the 'base torques' are plotted with dashed lines. There is a value for both males and females, the male value being larger in all cases. Figure 6 provides a clearer picture of the problem.

This was an unexpected problem from eGym's perspective: users should not be able to produce values lower than these base torques. In cases where the dashed lines do not coincide with the peaks, eGym currently suspect that this is due to incorrectly recorded base torque values.

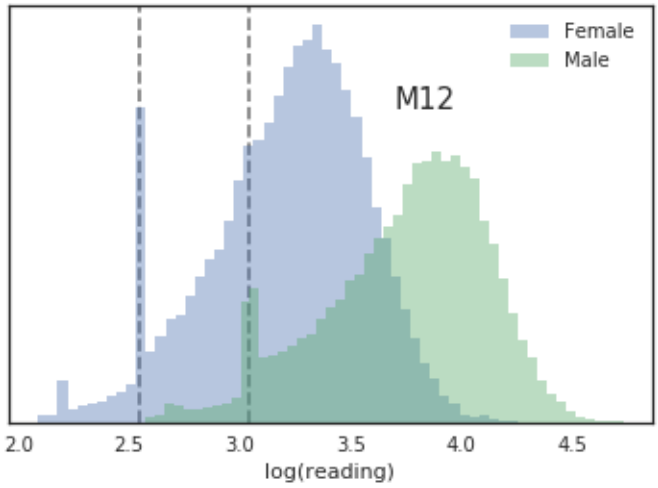


Figure 6: A histogram of the log strength values recorded for machine M12. We see large peaks on the left of the distribution for both genders. The 'base torque' value for each gender is plotted with a dotted line. These seem clearly related. Note also that there appears to be one other peak to the left of this line. We propose this issue may be related to resonance of the measurement equipment

Finally, in Figure 5, machines M9 and M16 exhibit a 'cutoff' issue - a particularly high count for the maximal value suggests some users are hitting a maximal value that machines can record. Users are not supposed

to be able to reach this value.

3.4 Exploratory Data Analysis

In our initial exploration of the data we investigated attributes of users, such as their starting strength and start date, and then clustered the machines. We were interested in exploring user visiting and joining patterns, and whether certain machines were related by recorded strength data (and thus could be predictive of one another).

3.4.1 Analysis of Gym Users

Features were created for the unique users aiming to encapsulate the following factors:

- The period of the year in which the user joined
- The user's strength when they first join
- How user strength changes over time
- How training affects user strength

3.4.1.1 Start period The period of the year in which the user first joined the gym was characterised by week number of their first training week. This was found by querying the training dataset for each of users, each time identifying the rows that pertained to the earliest record of the user and (if this record refers to their first training week) gathering the week number.

Figure 7 shows the number of users who joined in each of the 52 weeks of the year. It shows, rather unsurprisingly, that there is a large increase in new users in the new year, corresponding to people making a New Year's resolution to start exercising more.

There is a noticeable slump in frequency of user first training weeks between weeks 15 and 22; this is a result of the timeframe that the dataset spans. The data provided spans from June 2016 up until the present, resulting

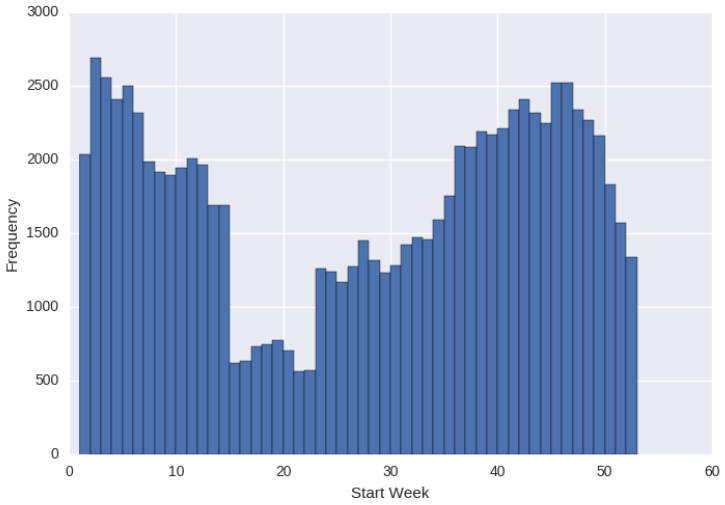


Figure 7: A histogram showing the counts of users who started for each week within the year. There is a dip at week 15 which continues to week 22: this is simply because our data runs from June 2016 to April 2018.

in 2 years of data coverage for weeks 1-14 and 23-52, but only 1 year of coverage for weeks 15-22.

3.4.1.2 Start strength

N.B. If more time had been available, this analysis would have been re-run, transforming the strength measurements with a logarithm, and normalising by gender separately.

In order to assess the strength of each user when they first joined the gym, a standardised measurement was defined. Each machine had a different distribution of strength values, so in order that we could create a single value summarising the strength for each user (e.g. by taking the mean their strength measurements over different machines that they had used), we first needed to standardise the strength values from the machines. This was achieved by considering each of the values in the measurement dataset, and standardising them against the other values that pertained to the corresponding machine; standardised strength values ($V_{standardised}$)

were defined by

$$V_{standardised} = \frac{V - \mu_{machine}}{\sigma_{machine}}$$

where V is the original strength value for a user for a given machine, $\mu_{machine}$ is the mean value for the machine over all users, and $\sigma_{machine}$ is the standard deviation of the values for the machine over all users.

Having standardised the values, the mean of each user's first measurement session was calculated. We refer to this as the user's mean standardised start strength.

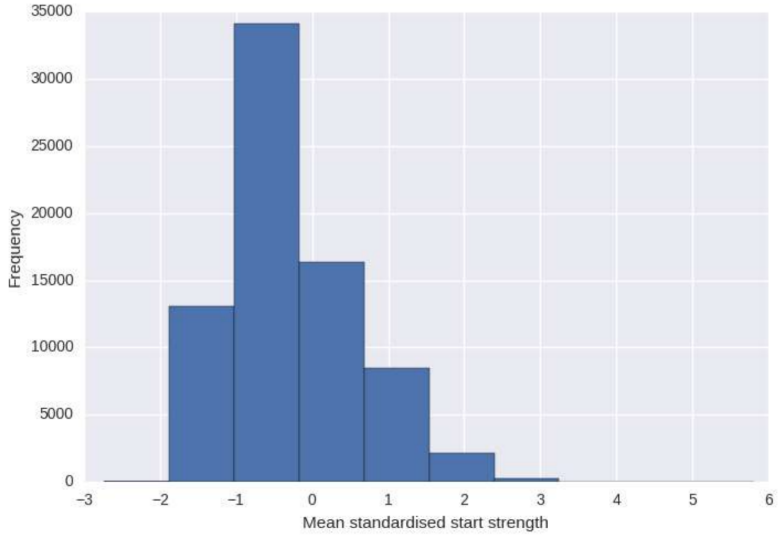


Figure 8: A histogram of the calculated mean standardised start strength of all users

Figure 8 shows the distribution of users' mean standardised start strength. A peak between 0 and -1 could indicate that new users are typically not as strong as the average user; though this conclusion should be revisited if the analysis can be re-run as described at the top of this section.

See Figure 5 for an illustration of the need to log the strength measurements and split by gender before normalising. Once this is done, we would then plot the growth trend: given values of the strength summary statistic (defined for the user's start strength) for each of their measurement sessions, and draw a trend to indicate the rate at which the user's strength is growing. This can be used to indicate when users are plateauing. Whilst

we did not conduct this analysis over *all* machines, some insight is provided in Figure 9 where we plot how the strength data progresses over time for a single machine.

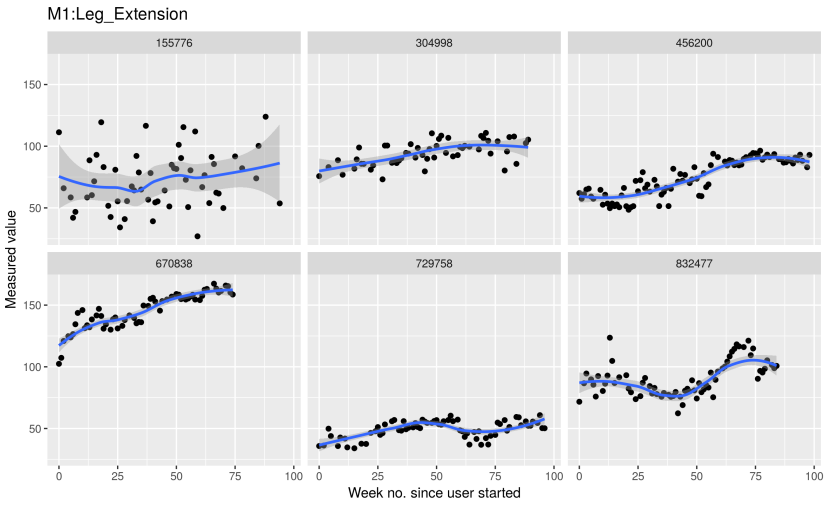


Figure 9: Time-series for a few selected users with long-term data. Measurement data for machine 1 was aggregated by user and week, taking the maximal measurement and the 6 users with the highest number of datapoints were plotted against the week of the measurement, overlaid with LOESS smoothing.

3.4.1.3 User Training Consistency In order to measure a user's consistency, two new variables were derived: the mean number of times that the user visited the gym per week, and the standard deviation in the user's number of weekly visits. In future, we would recommend looking at the number of successive weeks in which the user has visited the gym, and conversely the dead spots between these streaks.

3.4.1.4 Length of Attendance We calculated the length of attendance to the gym for each user by subtracting their first session date from their most recent. The results are shown in Figure 10.

3.4.1.5 Connection between measurement results and training With this investigation our aim was to provide insight as to whether

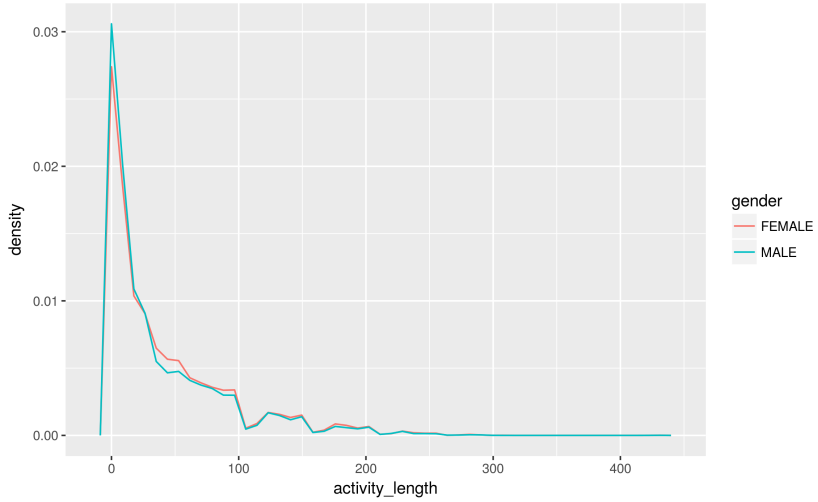


Figure 10: Short term users are very common in the data. This is a density plot of users. The length time from first attendance at the gym to last, i.e. the number of days they were an active member, is shown on the x axis, and the density of users that have attained this commitment level on the y. The majority of users (both male and female) were only short-term users; these users typically stopped attending the gym shortly after joining.

performing training sessions using the gym equipment improves the strength scores collected at measurement sessions. Do the number of training sessions matter, or perhaps it's how frequently users train that matters? Is there an ideal way of training? How important are saturation effects and when do they really kick in?

We started with the measurement data per user, machine and week (created as described below in the Clustering of Machines section). We consider only machine 1 (for no reason other than time constraints), then additionally calculate for each user the time and result of the previous measurement and the elapsed time, number of trainings, and number of machines trained between the two measurements. Finally, we show the:

- distribution of improvement (strength - previous strength)
- improvement as a function of the # of the measurement for the user
- improvement as a function of number of trainings or the number of machines trained between measurements
- improvement as a function of intensity of trainings or the number of

machines trained between measurements (number / gap between trainings)

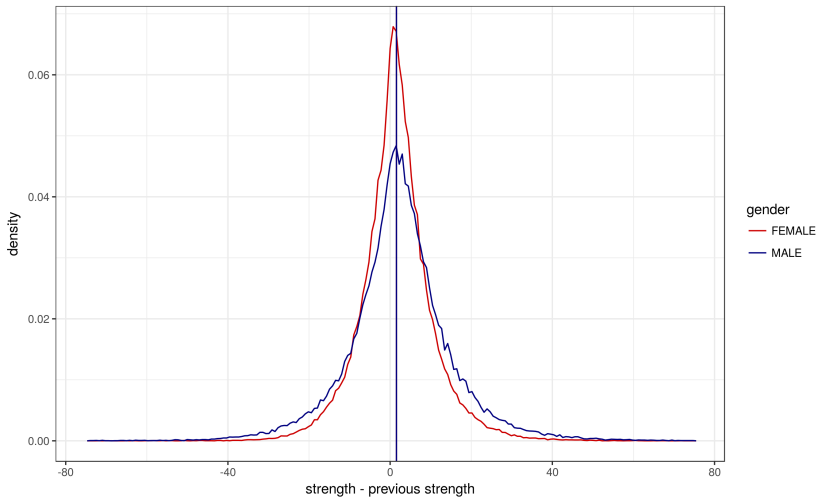


Figure 11: This figure shows a density plot of the difference between subsequent strength measurements for machine M1. The density for females is show in red, and for males in blue. The vertical line marks the overall mean values.

We found that:

- On average, improvements are small. The change in strength is larger for males, but the change in strength relative to the original strength is the same (around 2.63%) for both genders, see Figure 11.
- Improvement is weakly, but negatively correlated with the number of trainings (-0.0249) and the number of machines trained (-0.0089), but positively with *intensity*, where *intensity* is the number of trainings or machines used *per unit time*. Strength improvement has a correlation of 0.0978 with the intensity of trainings, and 0.0855 with the intensity of machines used. All correlations are highly significant (p-value $<$ 0.001).
- Improvement reaches a plateau: it's not worth going to the gym more than approximately 3 times a week. This is shown by Figure 12. This analysis is clearly affected by outliers. These should be cleaned or otherwise handled by a fitting method resilient to outliers, for example RANSAC [1].

The next steps we would recommend for this line of analysis would be to

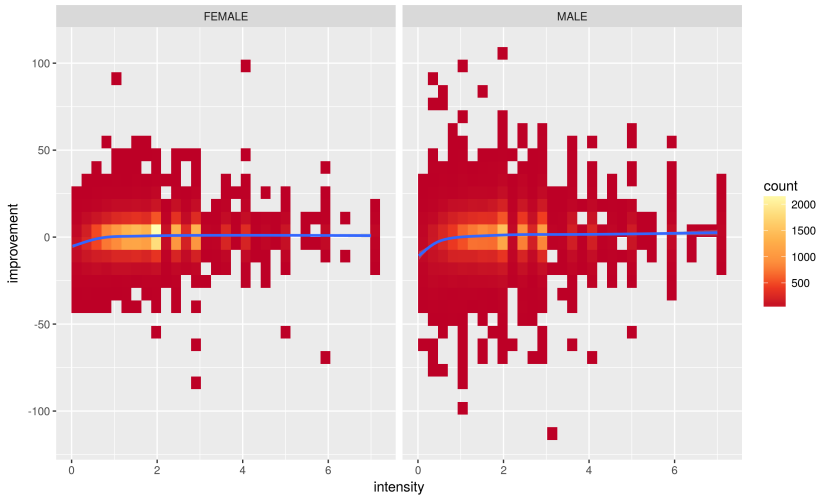


Figure 12: This figure shows the density of improvement (difference between a user's subsequent measurements on machine M1) as a function of intensity (the average number of times the user visited the gym between the subsequent measurements). The blue line is a smoothed fit to the data using the LOESS method.

clean the data (some differences are extremely large - see Figure 12, and report the *relative* improvement per user instead. Given the irregularity of measurement results, we would suggest first smoothing the measurement data per user.

3.4.1.6 Strength change with age We also investigated trends with age. Strength measurements on each machine tended to show a weak trend with age. People become stronger until they are in their 30s, and then gradually become weaker.

We show two plots illustrating how strength varies with age:

- In Figure 13 we show the distribution of all strength values recorded for machine M2 broken down by age
- In Figure 14 we show how maximal strength attained by each user form machine M1 varies with respect to user age.

We found that people in their 30s on aggregate appear to attain the highest values for these machines. The 'sweet spot' is around 25–30 years old,

and the decline in strength with age is more pronounced for males. These general observations transfer across machines.

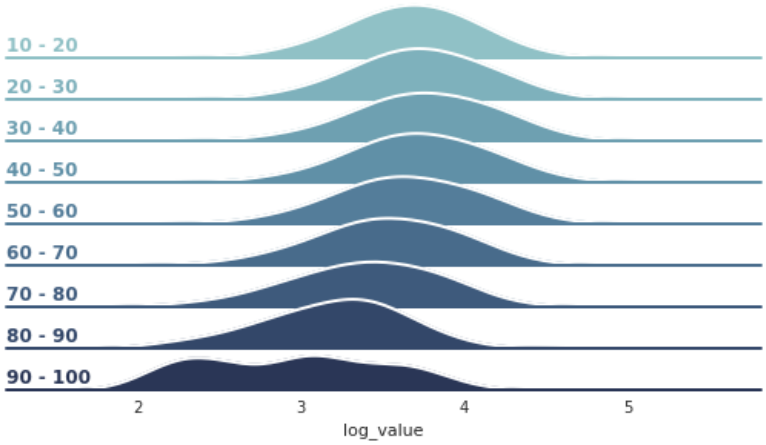


Figure 13: Variation in log strength with age. Age ranges are shown on the left and a kernel density estimation of the log strength distribution is shown. The data shown here correspond to strength measurements from machine M2. To emphasize, a value of 2 corresponds to a recorded strength of $e^2 \approx 7.4$ and a value of 3 is $e^3 \approx 20$, an order of magnitude larger.

3.4.1.7 Reproducing Results All results in this section can be reproduced by running:

- `users/ksuchak/notebooks/build_features.ipynb` to build and visualise all features
- `users/vsiska/training_measurements.R` for the connection between measurement results and training
- `users/vsiska/measurement_lag.R` and `users/awilliams/notebooks/*.ipynb` for the strength versus age plots

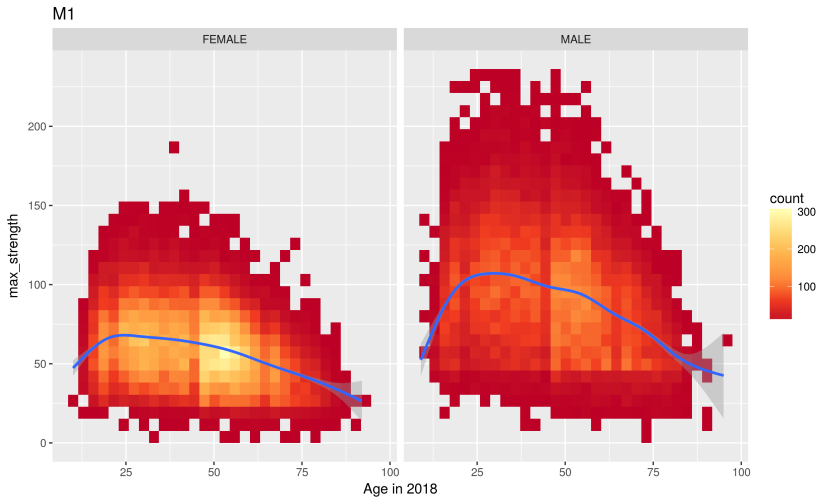


Figure 14: The above is a 2D histogram of maximal strength values recorded from machine 1 (LegExtension) for each user against their age. The figures for the other machines look similar, showing lower strength for young and old users, with a plateau for intermediate ages. The blue curve is a smoothing using a built-in generalized additive model in ggplot2.

3.4.2 Clustering of Machines

We used the strength measurement data to uncover relationships between different machines. We used hierarchical clustering using Ward's method [2] on the correlation matrix of strength measurements across users and machines. We followed the following procedure:

1. From the strength measurement data, calculate the average strength per user, machine and week (for weeks with at least one measurement). This served as a measure of strength at that time and also gets rid of spurious duplicated measurements.
2. Calculate summary statistics of user's strength for each machine. Mean, median, first, last and maximum values were explored.
3. For each pair of machines, calculate correlation between users' strength, considering only users with measurements on both machines.
4. Calculate distance measure between machines x and y as d(x,y) = 1 - cor(x, y), where cor is the Pearson correlation.
5. Run Ward's hierarchical clustering on the distance matrix between

machines

The topology was stable across different measures of strength, roughly corresponding to expected body areas: upper body, lower body, core and the outlier calf strengthening device. Triceps appear clustered with core machines, but it was established this was somewhat expected: users hence use their whole bodies to perform a tricep dip, which is more similar to core exercise machines than others.

In summary, clear groups of machines were discovered, and they seem reasonable given what we know about the machines. These results could be used to estimate strength on unmeasured machines.

Next, the analysis should be redone with the maximal strength per user, machine and week. It is difficult to fake the maximal strength, whereas some users try to game the system by providing weak measurements, so it is a better measure of the user's true strength.

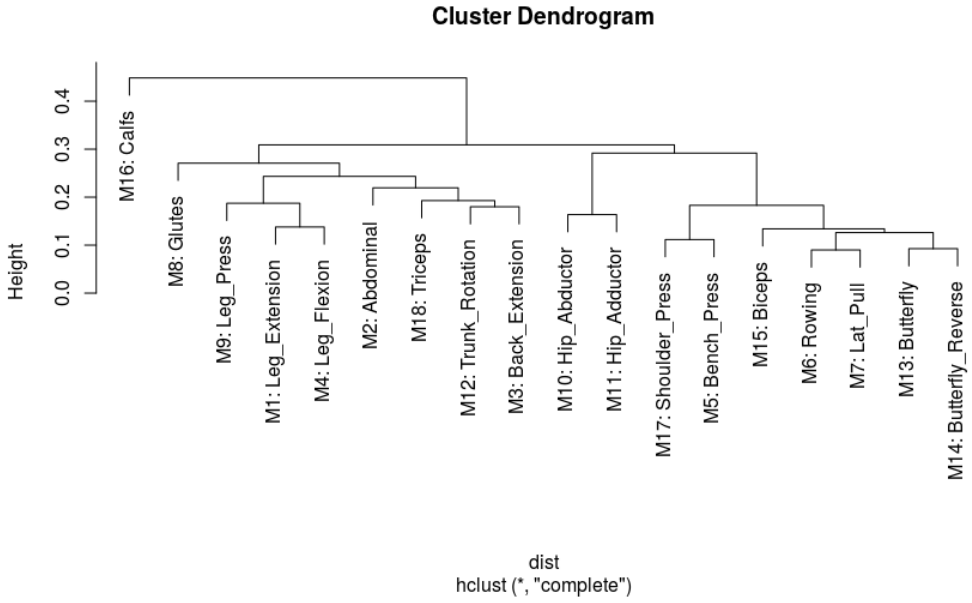


Figure 15: The results of clustering the data about machines. We have additionally provided the name of each machine and see that strength measurement values alone clearly identify like machines. For example, machines focusing on legs, M9, M1, and M4 are grouped, and M16 is (correctly) identified as a very different machine from all the others.

3.4.2.1 Reproducing Results All results in this section can be reproduced by running `users/vsiska/machine_correlations.R`

4 Experiments

4.1 Latent Variable Model of Users' Strength

4.1.1 Task description

It would be useful for eGym to be able to model how much their customers will be able to lift on each of their machines. An example use would be as a standard baseline for users to compare themselves with over time. An initial PCA analysis of users' performance across 12 of the machines implied that a single component was responsible for roughly 80% of the variance. This suggested that a model with a single latent variable, representing the overall strength of the customer, would be able to produce reasonable predictions.

4.1.2 Experimental set-up

We produced a sample dataset of 361 customers from a specific gym, each of whom were first-time visitors to the gym. Each of these users had done the same circuit of 12 machines. We did this to avoid missing value issues - not all gyms use all 18 machines. For each user, we possessed their age at the time that they visited the gym, and their gender. The task was to predict how each of the users would perform on all 12 machines given only their gender and age. We split the data into a training set of 270 users and a test set of 91 users.

Our model is best represented by a graph; this is shown in Figure 16.

The latent strength factor S is distributed as

$$S \sim \mathcal{N}\left(\alpha \times \text{age} + \beta \times \text{gender}\right),$$

where gender is represented as a binary variable (0 for women, 1 for men). The score on a given machine is then distributed as

$$\log M_i \sim \mathcal{N}\left(\gamma_i + \delta_i S\right).$$

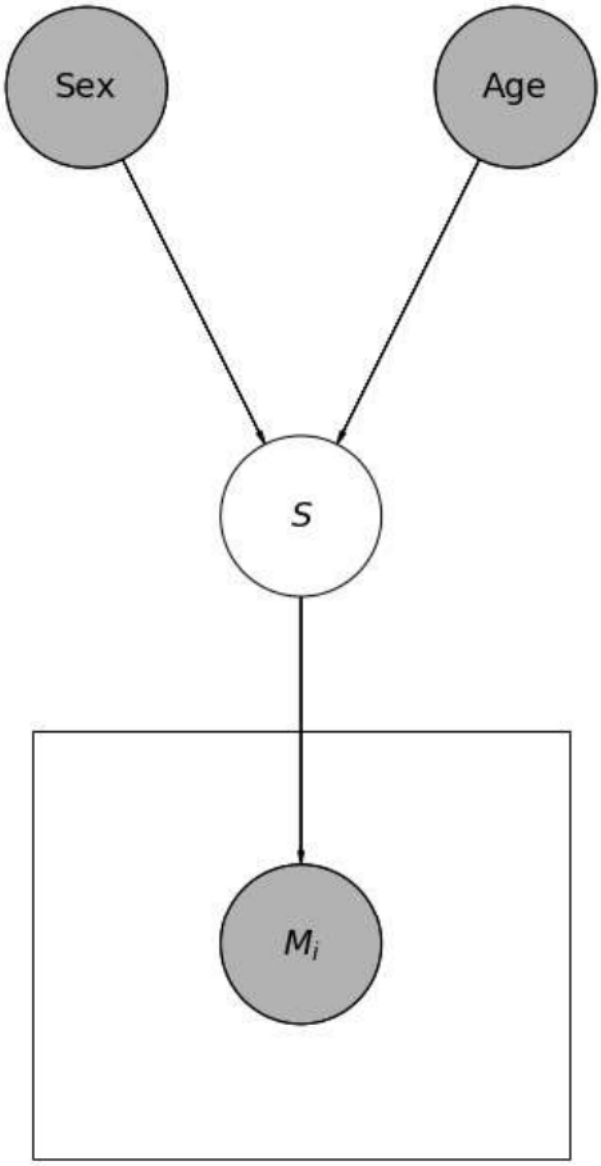


Figure 16: The latent strength factor is influenced by the age and sex of the customer. The customer's score on a machine is in turn influenced by their strength.

Thus, given a gender and an age, we can generate scores for all machines.

4.1.3 Results

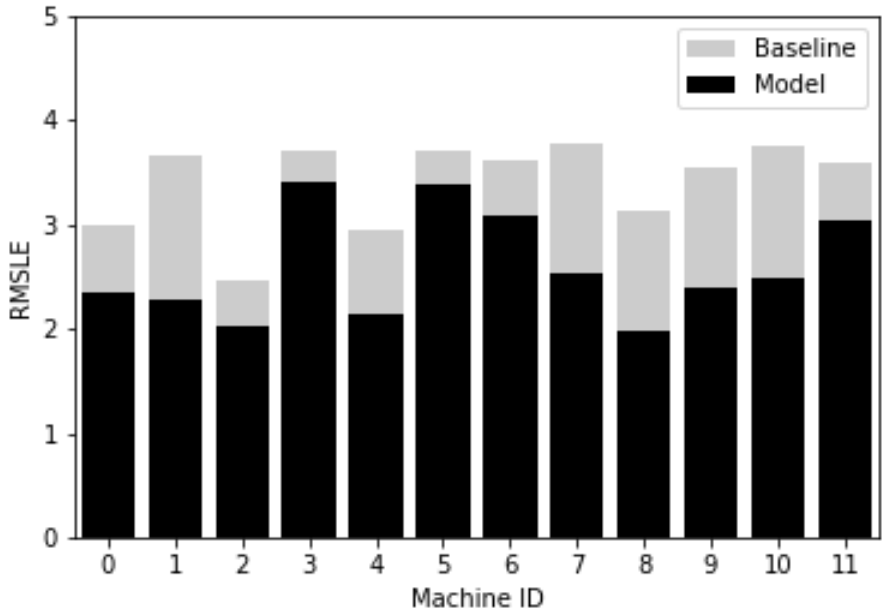


Figure 17: Our model beats the baseline on all machines. If we changed the baseline to predict the mean of the customer's gender, rather than the overall mean, it is likely that the model would perform comparably well to the baseline.

We computed the root mean squared logarithmic error of the model's predictions for the test set and compared them with a benchmark. The benchmark was simply to predict the mean on each machine as the value for each customer (after log-transforming). Our model achieves a better RMSLE than the baseline on all of the machines, this is likely due to the fact that men and women form quite different distributions on all machines. Had we used the mean for the customer's gender, we would likely see that the model performed comparatively well to the baseline.

4.1.4 Reproducing results

All below notebooks are contained within users/awilliams/notebooks/. The notebook `data_preprocessing.ipynb` outputs a file `initial_visit_data.csv` (the path at which this file will be saved should be changed by the user). The notebook `pca.ipynb` produces the PCA components. The notebook `second_model.ipynb` fits the model described above to the data and produces the graph of rmsle values. The anaconda environment used to produce the notebooks can be reproduced using the file `users/awilliams/requirements.txt`.

4.1.5 Conclusions

The model described here is simple, but it works adequately well. Obvious extensions could be

- Including further demographic data, e.g. body weight and body fat percentage.
- Using more than one latent factor to produce more accurate results - two are certainly justified as they could represent upper and lower body strength.
- The model can predict a customer's value on a new machine given their age and gender *and* their scores on other machines. It would be interesting to see how the accuracy improves when customers have used more machines.
- Extending the model to include time, so that users' improvements over time can be modelled.

4.2 Autocorrelation Structure in Measurements

4.2.1 Task description

The aim was to explore the autocorrelation structure in measurements, to find out if a simple autoregressive model could be used to predict measurements for the future.

4.2.2 Experimental set-up

We used the measurement values per user, machine and week as described in the Clustering of Machines section. We examined the one-lag autocorrelation and the general autocorrelation profile per machine (up to 10 lags, subsetting for users with at least 10 measurements on the corresponding machine).

4.2.3 Results

There was a very strong autocorrelation and a roughly linear, Gaussian pattern for the first lag. An exception was machine 9, where there was a cutoff in strength (discussed further in the Data Quality Issues section). Autocorrelation decreased with lag, but remained high for all machines (above $\sim$70% by 10 lags). Examples are shown in Figures 18 and 20, with their autocorrelations shown in Figures 19 and 21 respectively.

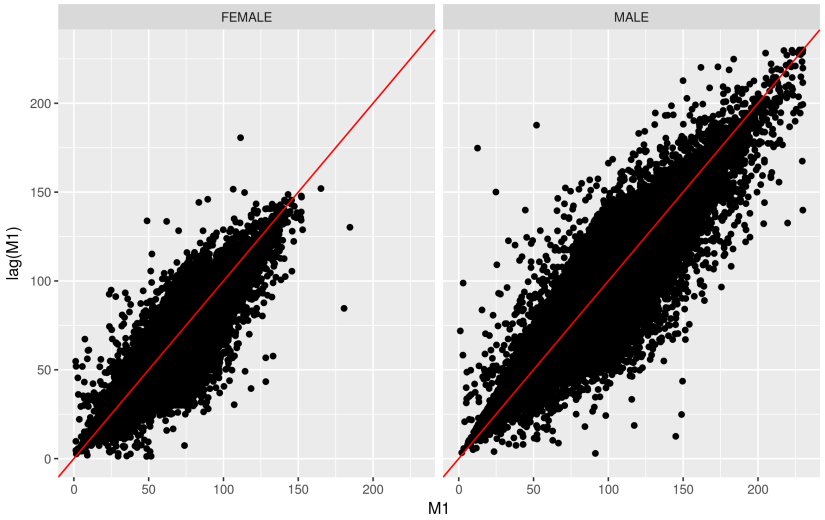


Figure 18: This figure shows the result of the previous measurement as a function of the current measurement on machine 1. Red line marks the identity.

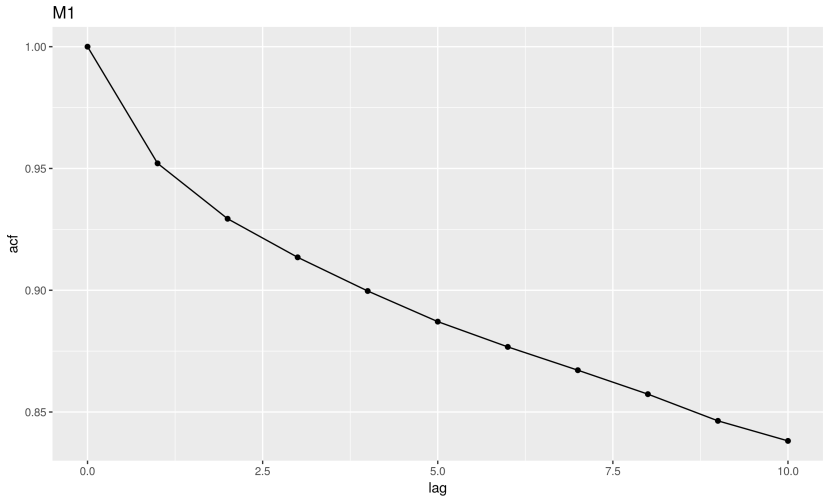


Figure 19: This figure shows the autocorrelation function of measurements on machine 1.

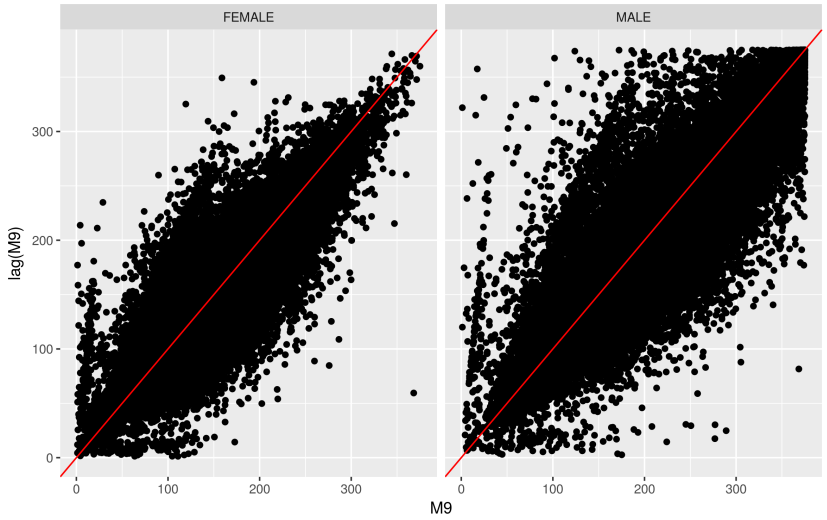


Figure 20: This figure shows the result of the previous measurement as a function of the current measurement on machine 9. Red line marks the identity. Note the pronounced 'corner' for males – this is a result of the 'cutoff issue' discussed in the Data Quality Issues section

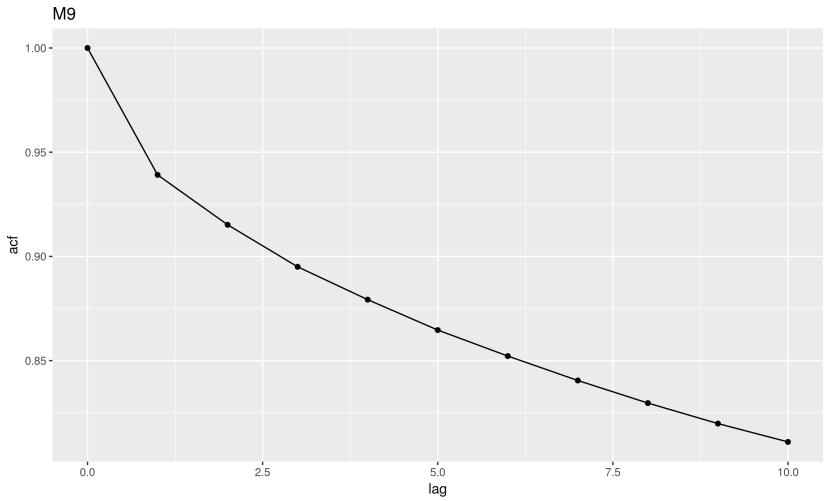


Figure 21: This figure shows the autocorrelation function of measurements on machine 9.

4.2.4 Reproducing results

Results can be reproduced by running the following R script: `users/vsiska/measurement_lag.R`

4.2.5 Conclusions

The strong autocorrelation pattern implies that an autoregressive model would be accurate in predicting future measurements. Furthermore, a few data issues were again illustrated:

1. Cutoff on machine 9
2. 'Whiskers' on the 1-lag plots i.e. a sudden large increase/decrease between measurements. These could be related to the 'base torque' issues discussed in the Data Quality Issues section.

4.3 Other Preliminary Work

In this section we list some other work for which results were only preliminary. The reader may be interested in reviewing the code and continuing these analyses.

- Using linear mixed model to predict strength in machines targeting the core muscle groups to optimize machine workout plan. This work could be continued to optimize a machine workout plan for eGym users using all machines. Longitudinal Analysis of the Core Strength Circuit: louis/data exploration.nb.html and Louis/lmm.nb.html
- Predicting next strength value with Bayesian Linear Regression, Random Forest, or Neural Network using Edward and Tensorflow: pfulop/Regression_Analysis.ipynb. This approach could be very useful to pursue when thinking of scaling the modelling to the full data.

5 Future Work and Research Avenues

We mainly addressed the task of creating a model for user strength, by which users could track their progress. There remains work to do on other areas of interest: relationship between training and strength, a model of user adherence to the training program, and a model of user churn.

For example: we did not investigate the level to which gym goers (or indeed gym instructors) actually adhere to training guidelines, such as the recommendation of 6 training sessions followed by a measurement session (alternating positive and negative training).

We identified a parallel between strength training and language learning: both require continual and regular practice; the higher the level of strength/fluency is attained, the longer it takes to deteriorate (or more importantly new concepts or strength acquired quickly deteriorates rapidly). We recommend looking to language products, for instance language learning applications such as duolingo[2] for inspiration for how to improve adherence and how to measure improvement.

[2] https://www.duolingo.com/research

Additionally, abstract concept of ‘skill’ (in our context, some abstract concept of overall strength) has been well studied and applied to modelling the outcome of games such as chess [3], football [4], and online games [5]. It is easy to see how the concept of strength could be ‘gamified’ and measured in a similar way.

The model of latent strength can be improved and expanded in several different ways:

- Firstly retrain the model of strength given age and gender on full data. This can now be used to plot strength as time progresses: given the age and gender of a user, and a set of strength their measurements, we can marginalise out any missing machine values and get a distribution of their strength variable. This would work much like the ELO rating system, recently adapted by Microsoft as the TrueSkill ranking system [5].
- When thinking of scaling this model, consider use of Tensorflow and Edward. See the Other Preliminary Work section for more information.
- Although a single latent variable was sufficient to produce reasonable predictions, we would intuitively expect a model with multiple latent factors to perform better. For example, a person’s lower body strength can be quite different to their upper body strength (e.g. cyclists vs. climbers). Modifying the model to include more latent variables is not difficult.
- More demographic information would likely boost the performance of the model significantly. For example, the customer’s height, weight and body fat percentage would be strong indicators of performance.
- We used the model to predict 12 machines given only demographic information. However, the *same model* is also capable of predicting performance on a new machine, given that the customer has recorded measurements on other machines. It would be interesting to see how much more accurate the model becomes in such a scenario.
- The model could be expanded to make predictions about the next session, given the previous session and demographic information. This would be a simple expansion of the current model, and would allow eGym to produce training plans for their customers. Duolingo have investigated this problem in the context of learning language, and their findings could transfer into the present domain [6].
- For further modelling ideas, our recommended text is [7]

The connection between training intensity and strength improvement could be further explored. Smoothing the erratic individual strength improvement profiles would clarify the temporal pattern and help highlight where the training effect saturates improvement. Machine-specific training data would be needed to study differences in the connection between different machines – we would like to investigate the cross-effects and causality between training on one machine and improvement on the other.

No work has been undertaken focusing on the type of training, i.e. positive or negative. Future work should look to explore the impact of different training types on strength improvement.

6 References

[1] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. [Cited in section 3.4.1.5.]

[2] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. [Cited in section 3.4.2.]

[3] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978. [Cited in section 5.]

[4] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010. [Cited in section 5.]

[5] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill(tm): A bayesian skill rating system. pages 569–576. MIT Press, January 2007. [Cited in section 5.]

[6] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1848–1858, 2016. [Cited in section 5.]

[7] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014. [Cited in section 5.]

7 Team members

7.1 James Owers

James is a 2nd year PhD student at the University of Edinburgh working on music generation at a signal level, and at a MIDI level.

He was the facilitator on this project.

7.2 Patric Fulop

Patric is a 2nd year PhD student at the University of Edinburgh working on machine learning methods for metric learning in high-dimensional data.

In this project, he worked on predicting a new machines strength measurements using regression analysis.

7.3 Ming Li

Ming is a Data Scientist in London and independent researcher. He is interested in Computer Vision and contributes to Python scientific stack.

He produced a regression model predicting gym member's strength on an existing machine.

7.4 Chimdimma Noelyn Onah

Noelyn is a PhD student at the Data Analytics and Society Centre for Doctoral Training in the University of Manchester. Her research is focused on the application of data science on burn care, with a focus on clustering methods. She was responsible for the clustering of users.

7.5 Veronika Siska

Veronika is a 4th year PhD student at the University of Cambridge, about to submit. She is a physicist turned complex systems scientist turned bioinformatician, working on various questions using data analysis and modelling.

She was responsible for analysing machines: she clustered machines, analysed the autocorrelation structure between measurements and the connection between training and strength improvement.

7.6 Louis Soussand

Louis is a biostatistician at the neurology department in Beth Israel Deaconess Medical Center. He is working on neuroinformatics and neuroimaging statistics.

In this project, he worked on the longitudinal analysis.

7.7 Keiran Suchak

Keiran is a PhD student at the Leeds Institute for Data Analytics. Having started his career as a physicist, he has a diverse academic and professional background. He now finds himself developing new methods for the realtime simulation of how pedestrians move around urban spaces.

In this investigation, he was responsible for the development of new user features.

7.8 Angus Williams

Angus is a data scientist at the Alan Turing Institute. He has a PhD in Astronomy and has worked as a data scientist for the last year, with projects

in a variety of sectors. He is most interested in Bayesian methods and probabilistic programming.

He is responsible for the latent strength model (implemented in Stan) and finding the ‘spikes’ in the data.

turing.ac.uk
@turinginst