Climbing Logbook Analysis

Alex Friedrichsen
University of Vermont
Burlington, VT
apfriedr@uvm.edu
alex.p.friedrichsen@gmail.com

Cal Warshaw
University of Vermont
Burlington, VT
Cal.Warshaw@uvm.edu
calwarshaw@gmail.com

Tomas Georgsson
University of Vermont
Burlington, VT
tomas.georgsson@uvm.edu

ABSTRACT

In the past several years, climbing has been increasingly popular as a sport around the globe. And as straightforward data input becomes more available to the average person, these hundreds of thousands of climbers have logged climbs to online databases to share their climbs with peers around the world. Climbers should have the ability to analyze their climbs with other climbers as well as get a more in-depth analysis of climbers around the world making different types of climbs. This analysis uses logbook data to group the type of climb being made. The weight, height, experience, age and sex of the climber is examined with the purpose of comparing the characteristics and experiences of the climber to the type climb. We find that a classification tree is the most accurate classification model for our data.

KEYWORDS

PCA, Scree-plot, K-Means, Elbow-plot, LDA, QDA, Classification Tree

1. INTRODUCTION AND RESEARCH QUESTION

The sport of climbing has existed in some form throughout history. However, it was not until recently that many metrics of climbing began becoming quantifiable in a relatively standardized way. Climbs now receive grades through either the US, French, or less commonly other grading systems, and climbers can more accurately know the types of climbs and difficulty of climbs that they are getting into. Indeed, climbers may record their personal information and route information on online logbook databases. With more climbs being recorded than ever before, analyzing these databases has become increasingly interesting with the amount of data available.

Through this analysis, we hope to use data about the climber's physical measurements and experience to classify the type of climb they are most likely engaging in. This information can be used by new climbers and experienced climbers alike to benchmark themselves against other climbers that have made similar ascents to the climber's current goal climbs. This information could make climbing safer, as it might encourage climbers to tackle routes more in their own range or skill level. Moreover, this analysis contributes to a general understanding of what the committed population of climbers looks like currently in the world.

2. RELATED WORK

Not much other work has been done to classify sports shorthands (climb type) based on climber attributes. Mark Dodd did do some basic experimental data analysis (EDA) on this dataset looking at how bmi and affects the ability to send (complete a climb) concluding that there is a downward trend in BMI as the max grade of the climber increases [2].

3. METHODS AND RESULTS

For each part of this project, different methods were used. Below is a discussion on how each step was achieved.

3.1 DATA COLLECTION

We obtained this data from a Kaggle containing a European database of over four million climbers. This data was collected from the world's largest rock-climbing logbook, 8a. The database was imported from a SQLite database into R using the RSQlite package [1].

3.2 DATA CLEANING

The statistics given included a number of variables that we will not be using such as the name of the climber and the competition that they were in. All *deactivated* accounts were removed, *sex* was restricted to "male" or "female" and *birth* date to be after 1930. Any observations with a *weight* < 0 was removed and observations whose *hight* was between 122 cm (4'0") and < 213 cm (7'0") and where bmi was between 12 and 40 were kept. Furthermore, we filtered "out anyone who had a calculated BMI > 28 and had also climbed 5.12a+ as this does not feel realistic and is probably a data entry error" stated by Mark Dodd who also looked at this data [2]. Additionally, we filtered for only observations where 0 < *started* <= 2017, *year* > *started*, *started* > *birth* and *year* > *1900* to deal with other data entry errors that were found.

Shorthand: The type of climb. Three types of climbs were possible: redpoint, flash, onsight, and top-rope.

In our analysis, we removed top-rope as we were interested in only sports climbing.
Usa_routes: The usa standard difficulty of a climb
Height: height of the climber in centimeters
Weight: weight of the climber in kilos
Sex: the gender of the climber

3.3 FEATURE ENGINEERING

We created the variable BMI from the weight and the height of the climber. We also created the age at ascent from the birth date of the climber and the date of the climb. Finally, we created the number of years climbed variable from the year the climber started climbing and the date of the climb. In full, the variables we selected from the set to use during our analysis are:
Age_at_ascent: The age the climber was at ascent in years
Years_climbed_at_ascent: years the climber has been climbing at ascent
BMI: the body mass index of the climber

3.4 EXPLORATORY DATA ANALYSIS / DESCRIPTIVE STATISTICS

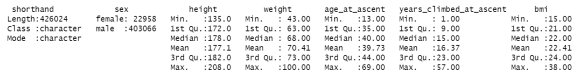

Figure 1: Summary Statistics

The average height is 177.1 cm, the average weight is 70.41 kg, the average age at ascent is 39.73 years, the average number of years climbed at ascent is 16.37 years, and the average BMI is 22.41 kg/m^2.

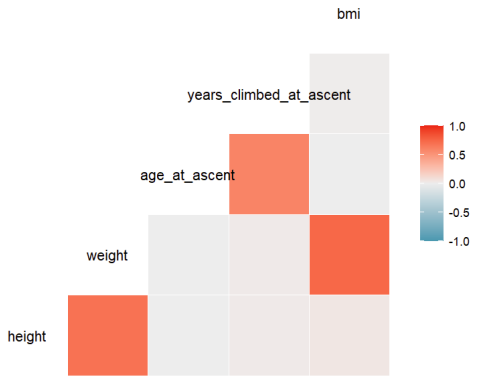

Figure 2: Correlation matrix

Figure 2 describes the correlation matrix of the quantitative variables being analyzed. The weight and the height, and the age at ascent and years climbed are the most correlated variables in the dataset. Both highly correlated pairs are not unexpected; the older one is the more likely one has more years of climbing experience, and the biological correlation between height and weight is trivial. In addition, weight is highly correlated with BMI and seems to outweigh height in the relationship between height and weight in creating BMI.

```
              Df   Pillai approx F num Df den Df    Pr(>F)
shorthand      2 0.015758   676.63     10 852036 < 2.2e-16 ***
Residuals 426021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Manova

Another of the first basic steps we took in analyzing our dataset was to perform a manova test on the data. We rejected the null hypothesis with a p-value of 2.2 * 10^-16. Thus, there is a difference in means in at least one of the variables.

3.5 PRINCIPLE COMPONENT ANALYSIS (PCA)

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     10.8780 10.5061 4.97238 4.8875 0.33890
Proportion of Variance  0.4265  0.3978 0.08912 0.0861 0.00041
Cumulative Proportion   0.4265  0.8244 0.91348 0.9996 1.00000
[1] 55.48758
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     1.4258 1.2623 0.9802 0.63673 0.08807
Proportion of Variance 0.4066 0.3187 0.1921 0.08108 0.00155
Cumulative Proportion  0.4066 0.7252 0.9174 0.99845 1.00000
```

Figure 4: Principle Components Summary

The first two PCs have the highest proportions of variance with 0.4265 and 0.3978 respectively and represent 82.4% of the variation in the data.

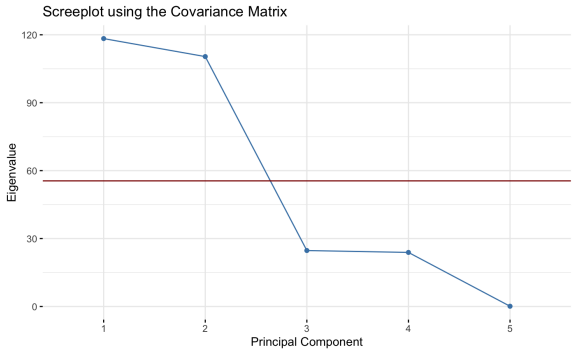

Figure 5: Scree-plot for PCs

The scree-plot above shows that the PC cutoff is between 2 and 3 PCs. Both the summary of the PCs and the scree-plot agree that the first 2 PCs are sufficient.

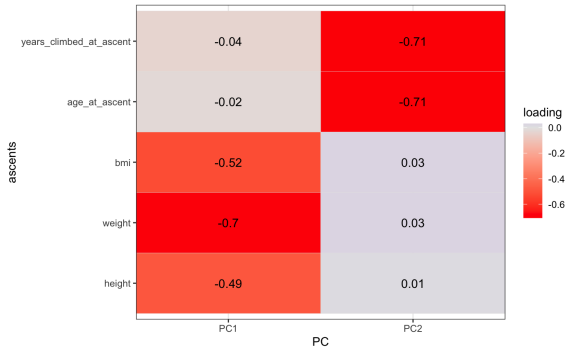

Figure 6: Principle Component Loadings

The loading of Weight is strongest in PC1 and the loading of years climbed at ascent is highest in PC2. It seems that PC1 is dominated by the body attributes (Weight and Height) while PC2 is dominated by time (The amount of experience in years climbing: years_climbed_at_ascent, and the age of the climber: age_at_ascent).

3.6 K-MEANS

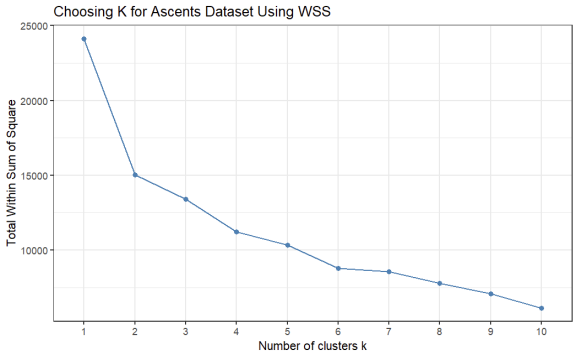

Figure 7: K-means WSS Elbow Plot

A k-means analysis was conducted using the height, weight, age of the climber and the number of years climbed. The first step was to standardize our data. After rescaling, we chose the number of clusters using within cluster sum of squares or "wss". In Figure 7 we can see the elbow plot of the potential number of clusters. While the slope is gradual, we thought the best choice would be two clusters due to the steeper drop off between 2 and 3 clusters.

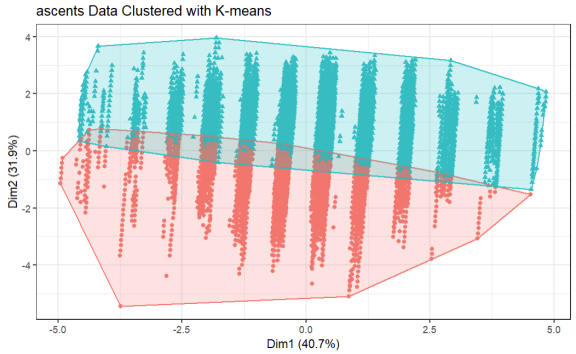

Figure 8: K-means Clustering with K=2

The clusters in the above figure 8 are merged slightly along the first dimensions. This means that while there is grouping, it is not complete. The structure of the plot is interesting: it leads us to believe that there are two distinct clusters, but there are more additional groupings than can be easily captured in two dimensions. We wanted to know more about what comprised these clusters, and judging from the results of our PCA, they are probably composed of different variables means. The observations are not disjoint sets, and never will be, as the essence of our data

demands overlaps in data points. For instance, while the majority of taller climbers might do similar climbs and have similar characteristics in age and weight, there is still going to be wide range in all of these variable within the larger data set.

3.7 QUANTITATIVE DESCRIPTIVE ANALYSIS (QDA) AND LINEAR DISCRIMINANT ANALYSIS (LDA)

```
Confusion Matrix and Statistics

          predicted
actual      flash onsight redpoint
  flash      3661    6968    28761
  onsight    7320   34287   119608
  redpoint  11977   33167   180275

Overall Statistics

               Accuracy : 0.5122
                 95% CI : (0.5107, 0.5137)
    No Information Rate : 0.7714
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0633

 Mcnemar's Test P-Value : <2e-16
```

Figure 9: QDA Confusion Matrix Summary

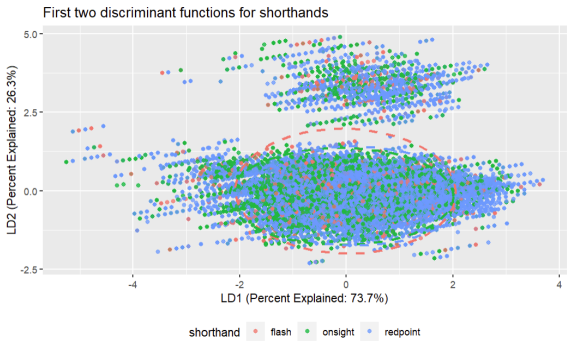

Figure 10: Discriminants for Shorthand

QDA analysis was conducted on climb type (shorthand) using height, weight, bmi, age_at_ascent, years_climbed_at_ascent. The Box's M test rejected the null hypothesis for homogeneity of variance, so we knew QDA was the preferred method of discriminant analysis. However, we decided to verify our results and use both QDA and LDA for comparison. Our results for both were equally inconsequential, as the predictive accuracy for the type of climb was lower than the "No information rate"; guessing "redpoint" for every climb was a better classifier than QDA (or LDA). The accuracy of our QDA was 51.22%, whereas the " No information rate" was 77.14% (77 percent of climbs in the data were redpoint). In Figure 10, we can see there is heavy overlap in the plotted climbs when attempting to determine the linear discriminants.

3.8 CLASSIFICATION TREE

```
Confusion Matrix and Statistics

          predicted
actual     flash onsight redpoint
  flash      176     204      438
  onsight     33    2224      688
  redpoint    33     371     5833

Overall Statistics

               Accuracy : 0.8233
                 95% CI : (0.8157, 0.8307)
    No Information Rate : 0.6959
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6331

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: flash Class: onsight Class: redpoint
Sensitivity                0.7273         0.7946          0.8382
Specificity                0.9342         0.8999          0.8671
Pos Pred Value             0.2152         0.7552          0.9352
Neg Pred Value             0.9928         0.9185          0.7008
Prevalence                 0.0242         0.2799          0.6959
Detection Rate             0.0176         0.2224          0.5833
Detection Prevalence       0.0818         0.2945          0.6237
Balanced Accuracy          0.8307         0.8472          0.8527
```

Figure 11: Classification Tree Confusion Matrix

The Classification Tree had an accuracy of 82.3%, which was higher than the No Information Rate of 69.6% which is the accuracy if redpoint was randomly chosen each time.

4. DISCUSSION

Overall, our results seem to provide a clear choice of which model is best at classifying the shorthand of a climb based on the attributes of a climber. After performing all of our analyses on the data, we were left with one model that worked far and away the best at classifying. The accuracy of each of our main models were: QDA: 51.2%, Classification Tree: 82.3%, showing that the classification tree was the best performing model of those tested.

Future work and limitations:

- Factor analysis because the two PCs are individually represented very well by differentiating categories.
- Analyzing the data based on the difficulty of climbs with similar methods (classification, etc.) would likely be a fruitful analysis.

- This dataset struggles with self-selection bias; strong climbers are more likely to be logging their climbs in the first place. Additionally, users who complete climbs may be more likely to log only the ones the feel represent the level they should be at, rather than perhaps the reality of the situation [2].
- We would also like to perform time-series analysis with this dataset to see how fast people increase in climb difficulty over time, and which factor influence this, particularly concerning years climbed.

5. ACKNOWLEDGMENTS

A special thanks to 8a.nu for creating and maintaining the climbing logbook and to Kaggle who continues to provide excellent access for users to find and explore datasets that allow for opportunities to advance research in a variety of fields!

6. REFERENCES

[1] 8a.nu climbing logbook: 2018. *https://www.kaggle.com/datasets/dcohen21/8anu-climbing-logbook*. Accessed: 2022-05-09.

[2] Climb through the data with me: 2020. *https://medium.com/@buckthecanuck/climb-through-the-data-with-me-80fb144ea408*. Accessed: 2022-05-09.