Data Analysis of Powerlifting Competitions

Abstract

We analyze data from various powerlifting competitions around the world from 1974 to present. The data labels competitors by age, weight, and gender and also includes their maximum weight lifted for various powerlifting movements. We apply three classification techniques: SVM, K-NN, and Classification trees to be able to predict a lifters gender or weight class. We were able to accurately identify the gender of powerlifting competitors but struggled to identify the weight class of the competitors.

Contribution

Sam Edds: Data cleaning/Exploration, SVM
Katherine Wilkinson: Background, KNN, Random Forests

1 BACKGROUND

Powerlifting is a strength sport that consists of three lifts; squat, benchpress and deadlift. In competitions athletes have three attempts to lift the maximum amount of weight in each of the different lifts. Their total amount lifted is calculated by summing their best valid weight lifted for each of the three movements. Additionally, competitors are separated into different weight classes based on their gender and their body weight in kilograms. For each weight class, the lifter with the highest total lifted wins. Competitors are also typically judged against weightlifters of similar age. At meets, which take place around the world yearly, lifters begin with the squat, followed by the bench press and then the deadlift. The Open Powerlifting site aggregrates data from multiple power lifting competitions recognized by powerlifting federations around the world from 1974 to present. Using this data we can look at how gender and weight class are affected by lifting capacity.

2 DATA EXPLORATION

Our powerlifting dataset is competitor data from the Open Powerlifting database. We initially are interested to see how the different lifts are correlated with demographic quantitative variables, such as age and body weight.

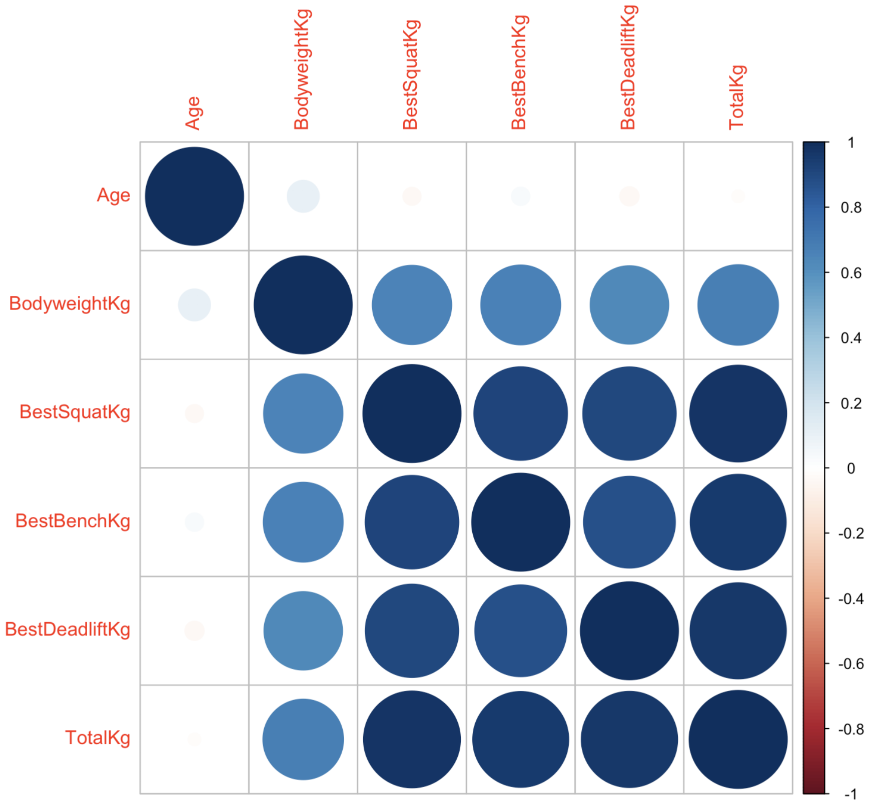


Figure 1: Correlation between quantitiative variables

From our Correlation Plot (Figure 1) we can see that age is, somewhat surprisingly, not very correlated with the amount of weight lifted by the competitors. Bodyweight however is fairly correlated with each of the different lifts. As expected each of the different lifts are highly correlated with each other and with the total amount lifted. That is, each lifter has similar strength across all the three lifts.

While we have information specific to different meets, and among many divisions, we are broadly interestered in using competitor lifting information in predicting gender and weight class (described in detail below).

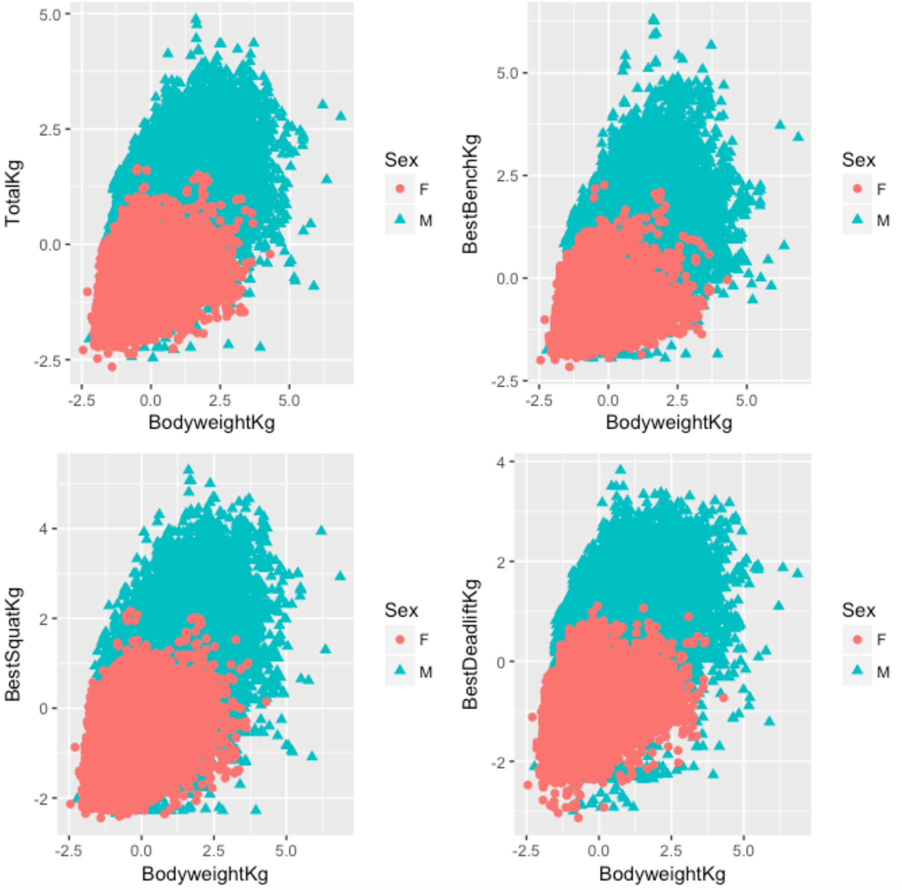
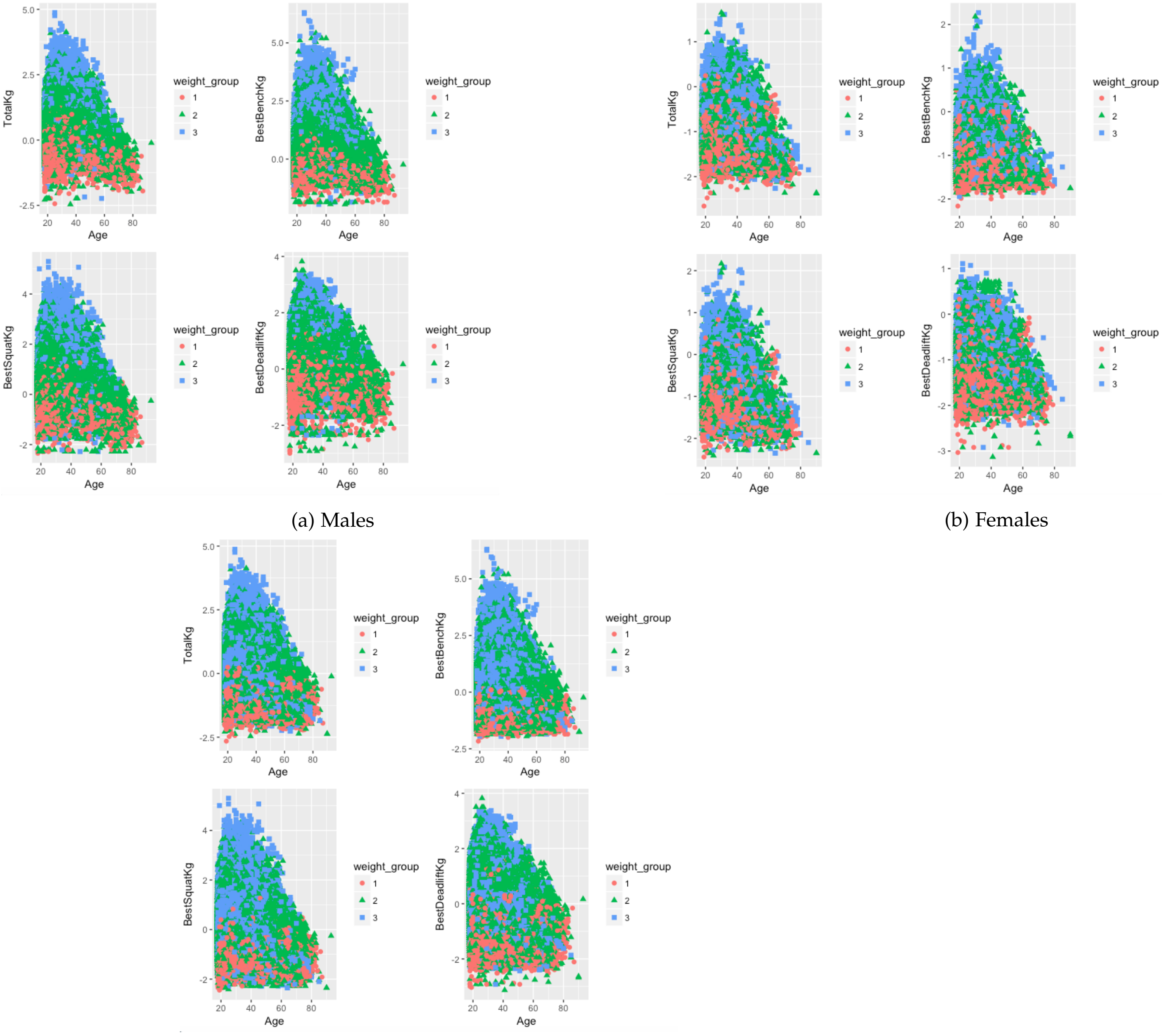


Figure 2: Gender Diagnostics

We first examine data summaries and create graphs to understand what variables best separate our data, exploring different possible weight categories (considering the number of competitors that would fall into each). Our graphs in Figure 2 show gender can likely be best classified using k-nearest neighbors (KNN), support vector machines (SVM), or decision trees (Random Forest). We noticed while weight class is not as cleanly separated, male and female plotted individually are fairly separated (See Figure 3). This is not particularly surprising because we chose different weight classes for men and women since on average many more women weigh less than men. Again, our graphs show KNN, SVM, and Random Forest as models that may classify well. Based on our graphs and since none of these methods have strong underlying assumptions, we will proceed with KNN, SVM, and Random Forest.



(a) Males

(b) Females

(c) Total

Figure 3: Weight Diagnostics

As a result of our data exploration we will conduct analyses on 4 different variables: Gender (M / F), Total weight (male and female both have 3 groups), and finally male weight and female weight separately. Specifically we will test how well a competitor's best bench press (kg), best deadlift (kg), best squat (kg), age, and weight predict the competitor's gender. For our weight class variables, we will run one analysis that groups male and female together, predicted by a competitor's best bench press (kg), best deadlift (kg), best squat (kg) and age. We will then run a separate analysis for male and female weight classes because we expect while there are three weight classes for male and females, because these are different and based on the differences in actual amount of weight lifted between men and women in these groups that our separate analyses will predict results with much lower error rates.

3 DATA DESCRIPTION

The following is a list of our variables and a description of each:

Variables Used in Analysis

Variable Name	Description
Weight	Body weight in Kg
Squat	Best back-squat in Kg
Bench	Best bench press in Kg
Deadlift	Best Deadlift in Kg
Sex	Male or Female
Age	18+
Weight Class	See below

We use our data exploration to help guide our data cleaning procedures. During our data exploration we tested different weight groupings for men and women upon which we could classify our data. Because the weight ranges for men and women differed, we split men and women into the following classes:

For Men:

1. Under 80 kg
2. 80-120 kg
3. 120 + kg

For women:

1. Under 60 kg
2. 60-80 kg
3. 80 + kg

4 DATA CLEANING

The following section describes our preprocessing data cleaning techniques applied prior to analysis.

We require our observations to have data for heaviest squat, heaviest deadlift, heaviest bench, and total score because total score is a sum of these. If someone is missing any component their overall score will be much lower, so we exlude them.

Because we have data from 1974-2018 powerlifting meets and people participate in anywhere from 1 to 300+ meets we randomly sample each person to pick one obersvation (competitors must be 18 years or older). There is no aprior reason to choose a specific age or meet, and we only want one observation per person so our observations are independent.

We break our train and test set, sampling evenly from men and women 75% each. Finally we center and scale our continuous variables, which is particularly important for KNN and SVM so our predicted values of our SVM/KNN/random forest results an be accurately compared to that of the actual data.

We will examine our results through 5-fold cross validation for KNN, SVM, and random forests. After choosing our optimal model we apply the final models for each and view the train and test error to judge our accuracy.

5 ANALYSIS

5.1 *Determining Gender*

5.1.1 *KNN*

Nearest neighber can help us determine the gender of a lifter by looking at the those who lift similar weight. As we saw in our initial scatter plots, it does appear that lifters of the same gender have similar individual lifts and similar total kilograms lifted.

Using 5 fold-cross validation, we determine that the optimal number of neighbors is 7. That is, when looking a single lifter, we then use their nearest 7 neighbors to determine what their gender is. From Figure 4 it can seen that the error rate begins to level off at about 7 neighbors for the CV error, Train Error, and Test Error. As expected, our train error is the lowest and the CV and Test Errors are similar. Using the optimal 7 neighbors, we get the error rates seen in Table 1. With Error Rates for the Test Data and Cross Validation at around 5%, we can pretty well predict the gender of a power lifter by their 7 nearest neighbors.

	k=7
Train	0.04566
Test	0.05413
Cross Validation	0.0513

Table 1: Summary of KNN Errors

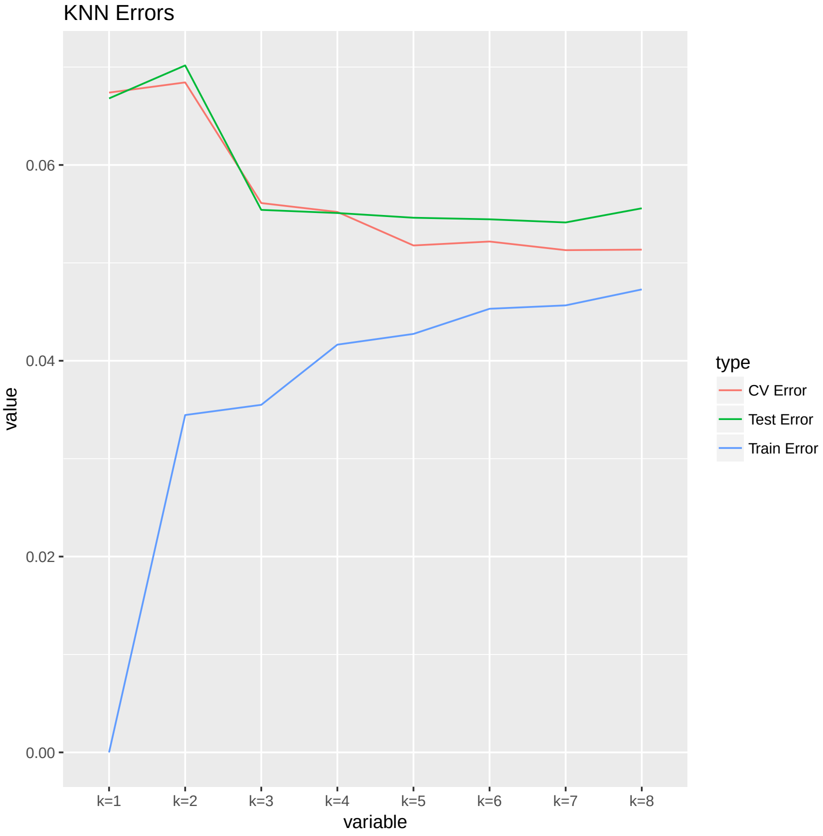


Figure 4: KNN Error Plot

5.1.2 *Random Forests*

For our decision trees we examine how sensitive our results are to tree size. We examine between 0-6 maximum depth, with no restrictions on tree size being 0. Based on our optimal cross-validated model we compute the train and test errors to determine gender based on their maximum lifts, total weight lifted, and their age.

We see that no restrictions on our model produce the best cross-validated error, although one then immediately suspects overfitting the training data. All of our graphs also show adding some kind of tree size restrictions creates much higher error, but as the restrictions increase, the range of errors drops steadily.

	Error	Type
1	0.0535	Train Error
2	0.05501	Test Error

Table 2: Summary of Tree Errors

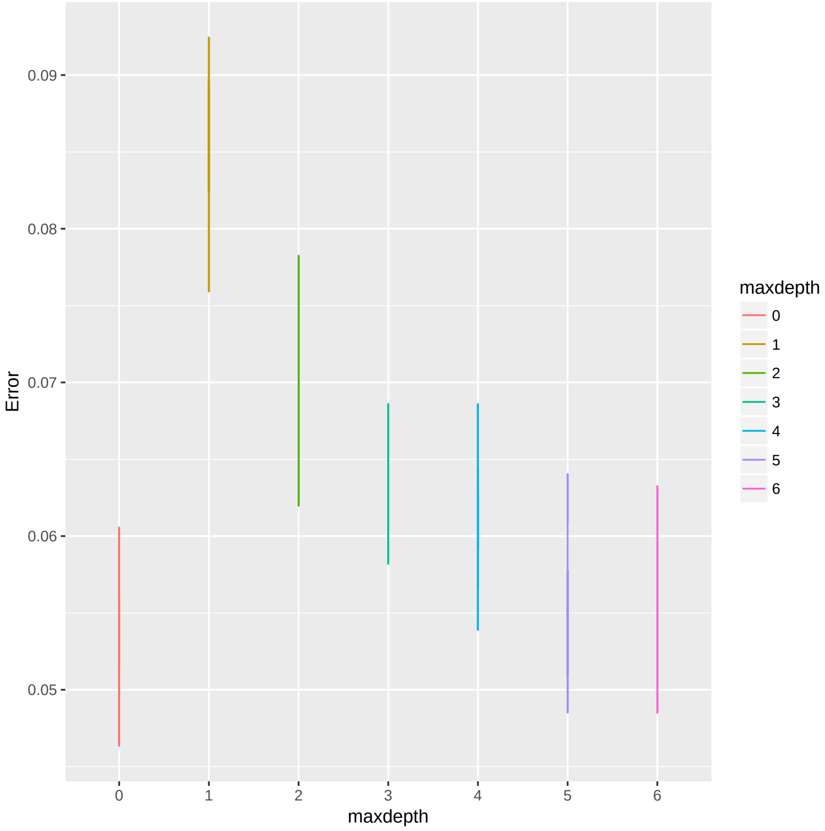


Figure 5: Decision Tree Depth Plot

We find the prediction errors using our optimal 6 depth tree, which should give us the lowest error without overfitting, as demonstrated in Figure 5. We have a similar error rate to our KNN results for the Test Data set at 5.5%. We can see that our Train Error is slightly higher, but not significantly so (See Table 2).

KNN and Random Forests give us similar results when attempting to predict gender.

5.1.3 *SVM*

We examine support vector machines as a method of classification. For SVM we use the tuning parameter cost which is the weight for penalizing the soft margin. We test costs from e^{-2} to e^{2}, which acts as a control over the total amount of slack allowed.

We also examine the tuning parameters for certain kernel choices, of which we use Linear and Gaussian. Our tuning parameters for Linear are just costs. For Gaussian we vary cost and gamma (.01, .05, .1 respectively). A small gamma acts as a distribution with large variance. Therefore the support vector has an influence on deciding the class of a given point even if the distance between them is large. Conversely, for a large gamma, the support vector does not have wide spread influence on determining the class of a given point.

We optimize over these different parameters, cross-validating for each combination, and see the averaged following results in the SVM graphs in Figure 6.

We choose our train/test based on our optimal model which produced the lowest cross-validated error rate, by kernel type. We then compute final train and test errors to see which of our optimal model produces the best results compared to the other kernels. From these results we can see that the Gaussian kernel gives us the lowest error for predicting Gender with a 4.71% test error. This gives us a slightly better correct prediction rate than compared to the KNN and random forests method.

	Test Error	Train Error	Type
1	0.0471	0.0465	Gaussian
2	0.0589	0.0592	Linear

Table 3: Train/Test Best Error Gender

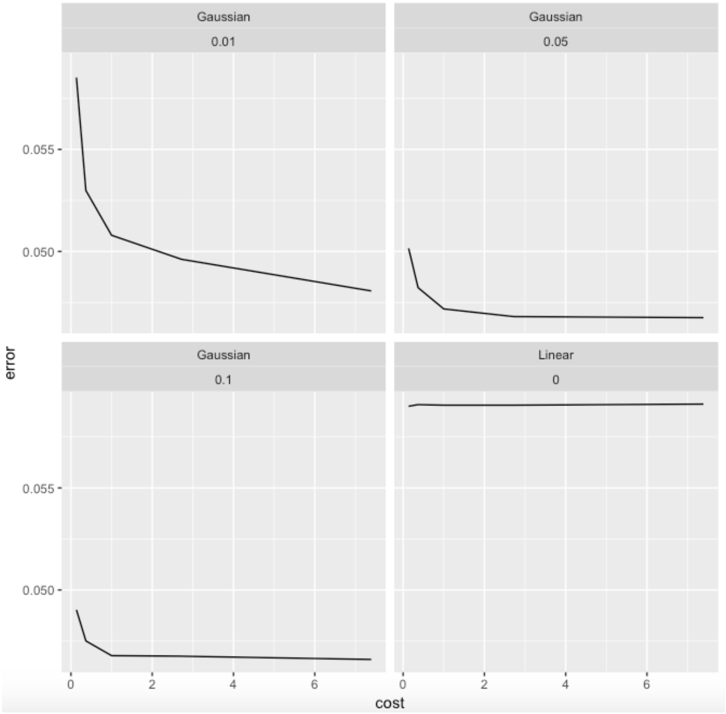


Figure 6: Gender Cross-Validated SVM plots

5.2 *Determining Weight Class*

As we saw in our exploratory analysis, there may be some separation by weight class. For our analysis to predict weight class, we will look at three different subsetted data sets: all adult lifters, just male lifters, and just female lifters.

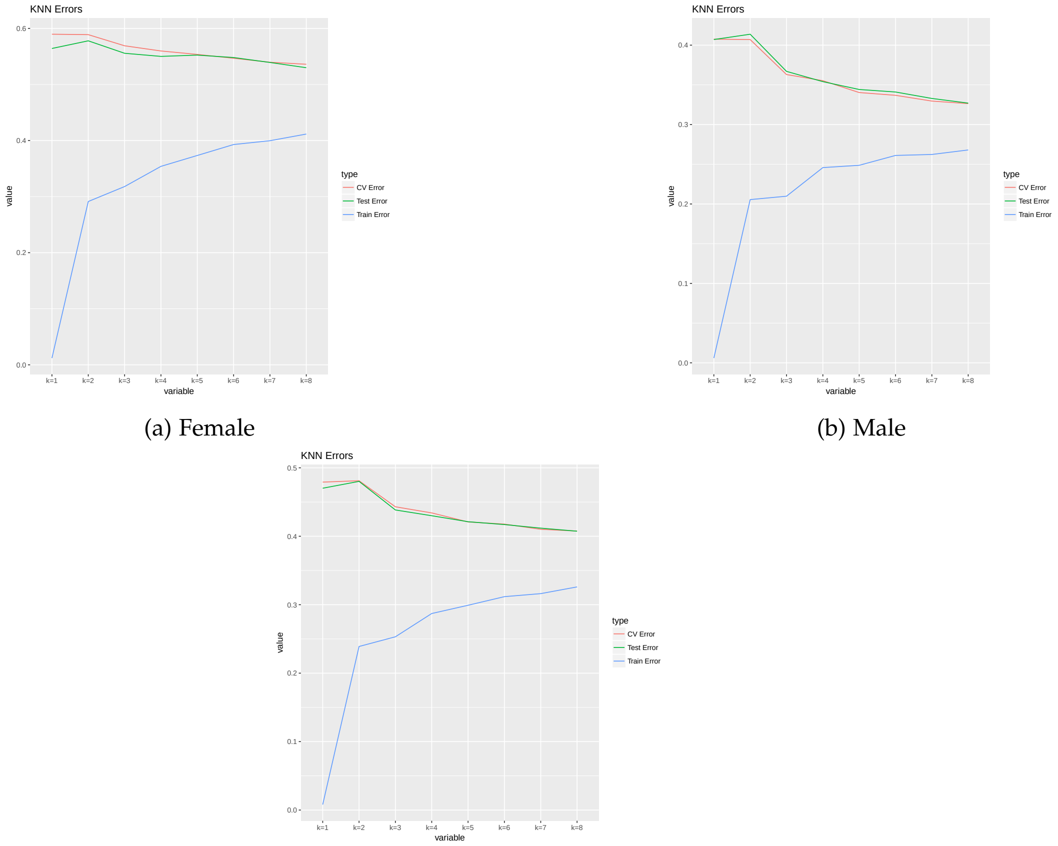
5.2.1 *KNN*

For each of the three subsetted data sets, we found in Figure 7 through cross validation that 8 is the optimal number of neighbors for classification purposes. The error rates however do differ fairly significantly (See Table 4). Most notably, the Test error for the male only data set is nearly 10% less than the when using all competitors and is about 20% less than the error rates for the female only data set.

	Male: k=8	Female: k = 8	Total: k = 8
Train	0.268	0.4116	0.326
Test	0.327	0.523	0.4074
Cross Validation	0.3264	0.536	0.4076

Table 4: Summary of KNN Errors

From these results, only about 70% of male weightlifters and fewer than 50% of female weightlifters will be classified in the right weight class by their closest neighbors.



(a) Female

(b) Male

(c) All Lifters

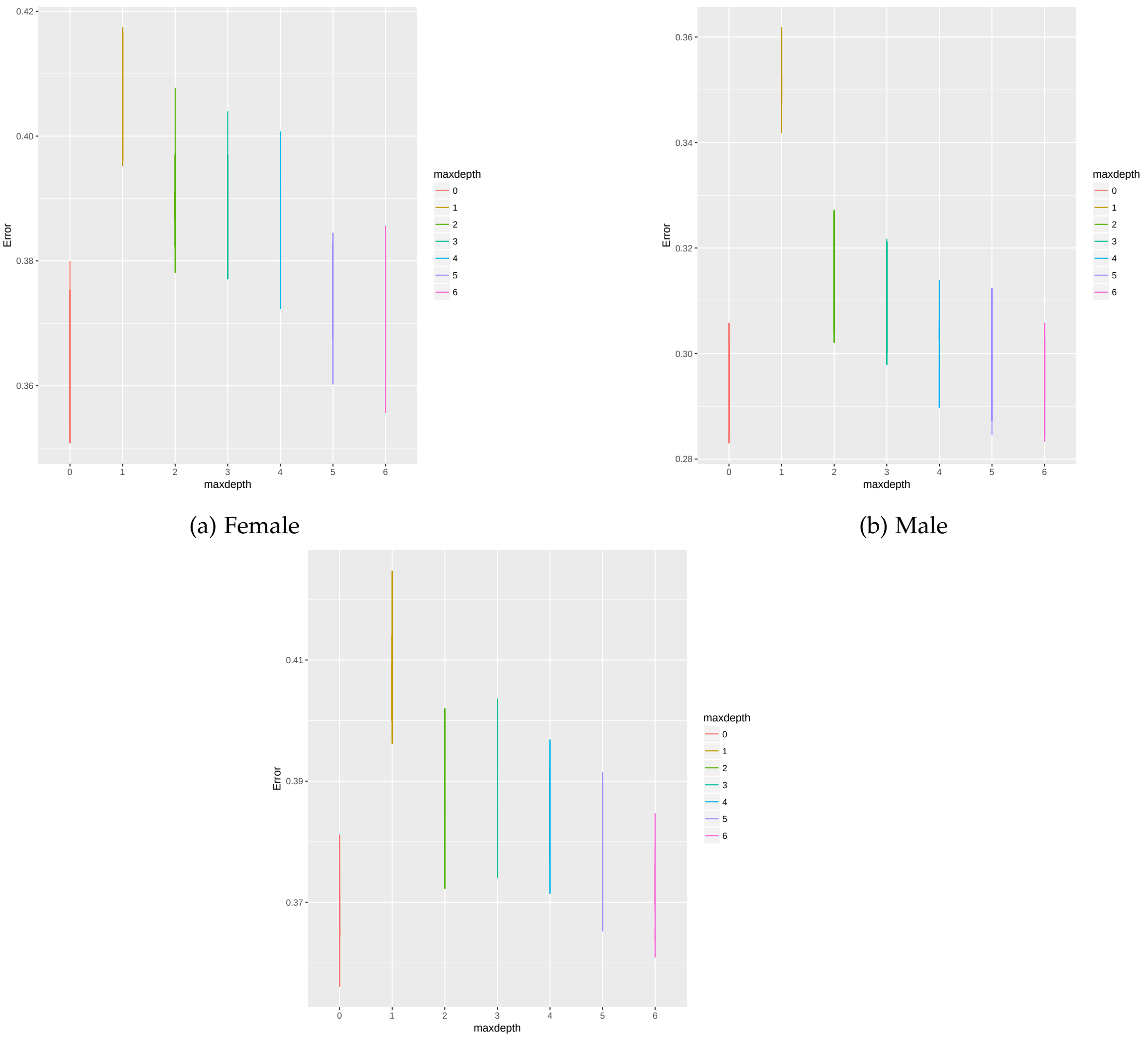
Figure 7: KNN Errors by Number of Neighbors

5.2.2 *Random Forests*

Overall, we see the tree size sensitivity is consistent among our different samples, as the shapes of our graphs in Figure 8 (showing all of the cross-validated error for each of the 10 runs) are similar, with just varying levels of error. We find 6 to be the optimal tree depth without overfitting for our classification tree for all three datasets. Similarly to the KNN results, the Male only data set gives the lowest prediction error rate at 35.86% (See Table 5). We can see that both methods so far are poor predictors of weight class.

	Male	Female	Total
Train Error	0.3518	0.48902	0.4076
Test Error	0.3586	0.4948	0.4126

Table 5: Summary of Classification Tree Errors



(a) Female

(b) Male

(c) All Lifters

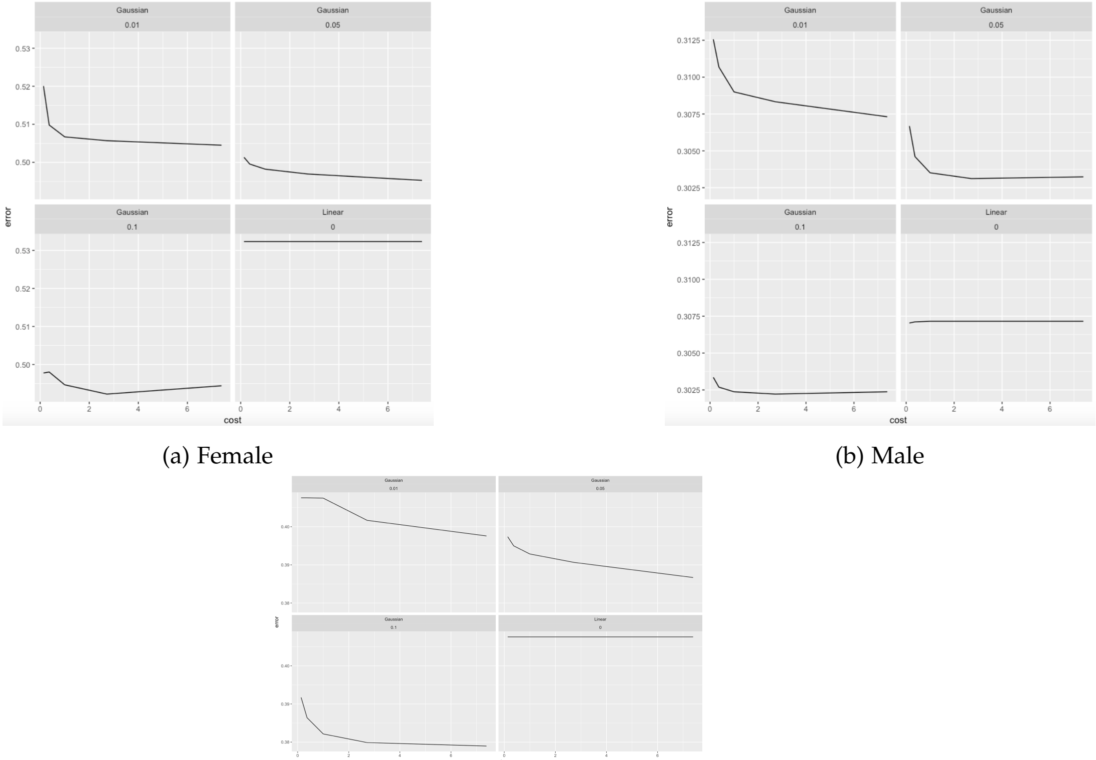
Figure 8: Classification Tree Errors by Depth

5.2.3 *SVM*

When looking at SVM models to predict weight class, we found our optimal cost for the male dataset was at 2.7 and at 7.39 for the female and all lifters datasets, while the gamma parameter is at 0.1 for all three datasets (See Figure 9). Similar to when predicting gender, the Gaussian model consistently performs better. Again, in Table 6, we can see that the male data set has the lowest error compared to the other datasets and nearly 50% of female lifters will be classified in the wrong weight class with this method.

	Males: Gaussian	Females: Gaussian	Total: Gaussian
Train Error	0.2997	0.4875	0.3775
Test Error	0.2914	0.4803	0.3813
	Males: Linear	Females: Linear	Total: Linear
Train Error	0.3075	0.5323	0.4076
Test Error	0.3009	0.5169	0.4126

Table 6: Summary of SVM Errors



(a) Female

(b) Male

(c) All Lifters

Figure 9: SVM Errors by Model

6 CONCLUSION

From these results, we find that we can more accurately predict gender than weight class when looking at age and kilograms lifted in squat, benchpress, deadlift, and total weight. It is much more difficult to correctly predict weight class even when separated by male and female lifters. This is consistent with what we saw in our intial data exploration. Additionally these results are not unexpected as weight classes are arbitrarily selected whereas gender is not. We also explored other weight class groupings and our intial data explorations showed relatively no change in separation. Overall, SVM methods, particularly with a Guassian kernel, gave us the best results while nearest neighbor and random forest methods were very similar to each other.

Our study has limitations in that we are randomly selecting only a single result for each individual and are not looking at how an individual lifters body weight and lifting ability changes over time. The dataset also uses powerlifting competitions over many years and does not take into account improvement in technique, equipment and training methods. It may be interesting to also note that there has been an increase in females in the sport in more recent years. If we focused on more recent years, it may change the composition of lifters and weight lifted and thus the performance of these classification methods.

7 REFERENCES

1. Open Powerlifing, Web *http://www.openpowerlifting.org/*

2. Powerlifting, Web *https://en.wikipedia.org/wiki/Powerlifting*

3. Arnold Schwarzenegger, Image *https://i.ytimg.com/vi/6EH4WXCiFho/hqdefault.jpg*