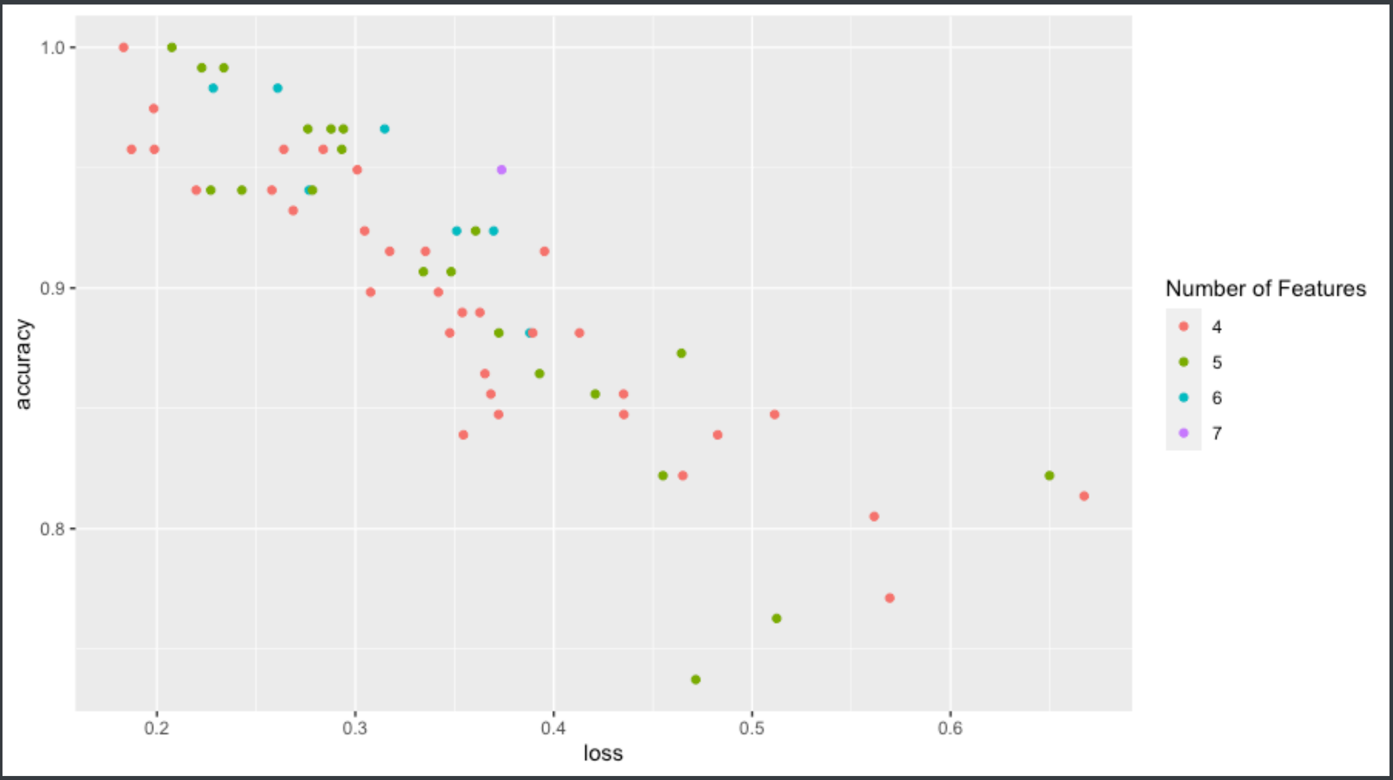
Random Forest Results

Classification of Shape using Summary Statistics

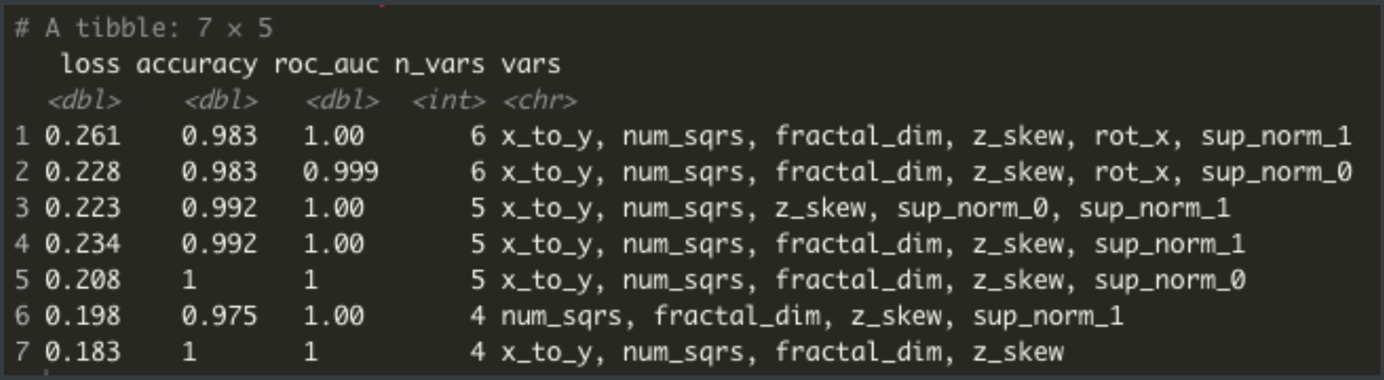
We fit a random forest to classify observations into one of five classes:

- comet
- compact
- fan
- stream
- other

Using all of the summary statistics which I have investigated so far, the random forest had perfect accuracy. I then refit the forst using a number of different subsets of these features, and computed the loss function and accuracy in each case. The following plot shows loss against accuracy, coloured by the number of predictors used:



The best performing combinations of parameters are shown in the following table:



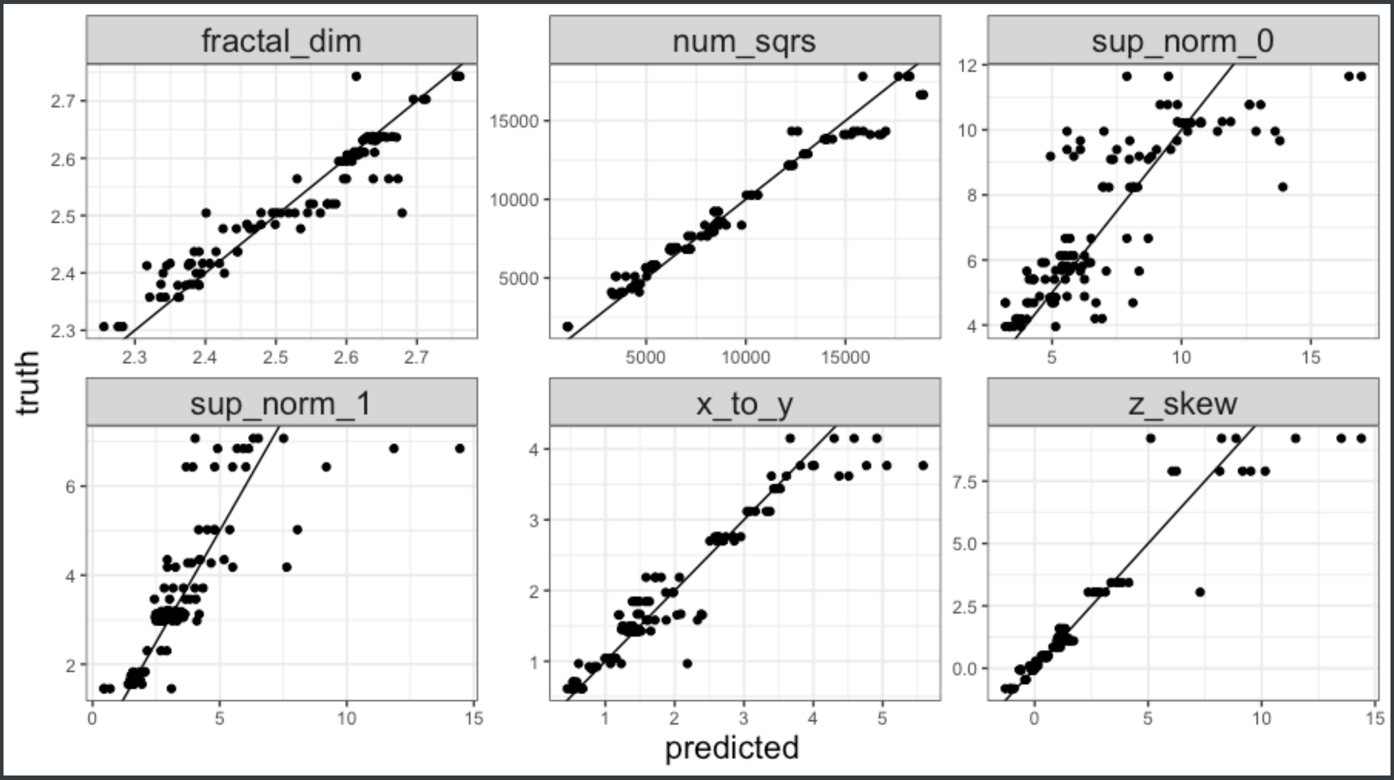
```
# A tibble: 7 × 5
   loss accuracy roc_auc n_vars vars
  <dbl>    <dbl>   <dbl>  <int> <chr>
1 0.261    0.983   1.00       6 x_to_y, num_sqrs, fractal_dim, z_skew, rot_x, sup_norm_1
2 0.228    0.983   0.999      6 x_to_y, num_sqrs, fractal_dim, z_skew, rot_x, sup_norm_0
3 0.223    0.992   1.00       5 x_to_y, num_sqrs, z_skew, sup_norm_0, sup_norm_1
4 0.234    0.992   1.00       5 x_to_y, num_sqrs, fractal_dim, z_skew, sup_norm_1
5 0.208    1       1          5 x_to_y, num_sqrs, fractal_dim, z_skew, sup_norm_0
6 0.198    0.975   1.00       4 num_sqrs, fractal_dim, z_skew, sup_norm_1
7 0.183    1       1          4 x_to_y, num_sqrs, fractal_dim, z_skew
```

These results come from models which were fit to training data with no resampling procedure applied to balance the classes. If SMOTE upsampling is used, the model performs even better. I present the results of this model though as it shows greater differentiation between combinations of parameters.

Given the very good performance of this model, it could be worth investigating a larger number of classes, and perhaps entertain the idea of subclasses.

Regression of Summary Statistics on Parameters

Having established that I was able to predict shape based on the data, the next step is to go from parameter to shape. I'm not sure of the best way to do this, but I have investigated using regression random forests, to predict each of the summary statistics based on parameter values. The following plot shows the true values of each parameter against those predicted by random forest. An y=x line has been superimposed to demonstrate how far the predictions deviate from the truth. We see in the case of fractal dimension, number of squares and x-to-y ratio that the models predict quite well, but in the other three that there is some heteroscedasticity in the prediction.



A two step prediction process

I then investigated using these two models in a two step process. I fed simulation parameters into the regression random forests to predict summary statistic values. I then used the classifaction random forest, trained on the actual data, to classify these predicted statistics, and found that it predicted with 100% accuracy.

Concerns

- The two models are trained on the same data - I'm not sure if this is an issue
- Something feels a bit off about this procedure
- Could go straight from parameter values to predicting class, but this doesn't feel quite right