The yarn ratings on Ravelry are provided by a range of artists and hobbyists, all with different demands for a yarn. What one user considers a good yarn (for example low cost but low quality) may be considered poor by another. In addition, since many yarns are produced by other hobbyists, there is a tendency for ratings to be artificially high INSERT REFERENCE. These factors made prediction of a yarn’s rating difficult and manifest as noise in the data.

The Random Forest method significantly outperformed a tuned Neural Network and SVM. Even without parameter tuning, a Random Forest would have outperformed the best performing models from the other two methods. Therefore, it can be said that for the given parameter search space and computational time, a Random Forest algorithm provides the best means to predict the rating for a given yarn. This does not necessarily mean it is the best of the three methods since a more extensive or optimised parameter search may allow either of the other two methods to exhibit an improvement in performance. REFERENCE

DISCUSS VARIABLE IMPORTANCE

Although the Random Forest method gave the best accuracy, the improvement in RMSE over simply guessing the mean rating for each yarn was only XX, from XX. Inspection of the plots in figure XX suggests that the all models had the tendency to overfit on the ‘id’ attribute. As discussed, noise in the data will always impose a limit on the performance of predictive models for this dataset. As this is the first known study of this dataset, any significant improvement over guessing the mean was considered a success; however there is little doubt that other models could provide improved performance and the minimum practically achievable RMSE remains an open challenge.