Team A
Emily Miller and Bianca Gonzalez
Spring 2017 Statistical Learning
Professor: Albert Kim.

Final Project Write-Up:

Our final project used the data set: Ghouls, Goblins, and Ghosts. This dataset uses continuous covariates to classify different types of monsters (categorical outcomes). When initially looking at this dataset, we did an exploratory analysis to visualize predictors and classifications in our training dataset. We found correlations between several of the predictors and monster type (Figures 1). We also found clustering of the data by predictors so we used a clustering method (Figure 2).

Our first categorical outcome prediction method and most effective method was K-Nearest Neighbors. We decided to try KNN because after plotting predictors against type of monsters, we could see some clustering of the data by monster type (Figure 2). To make the model more interpretable, we kept all predictors when making our models. In KNN, we performed a cross-validation and found the optimal nearest neighbors at 173. This CV proportion correct score was .38. We think this CV score, although also our highest, was only .38 correct because it cross-validated on one fifth of the dataset (our generated test set). When we predicted with our test data set, we scored .73. After finding the optimal K we submitted to Kaggle. The Kaggle score was close to ours at .74 correct. We initially felt this was a low proportion correct, however after looking at the leaderboard we noticed we were on the higher range of scores that were not overfit. The upper range of the kaggle leaderboard scores was 1-.75 with only two scores above .77. These two scores were overfit as acknowledged in team names. This was the most effective method but we continued to search for a method that would generate a higher score.

The 'extras' and next two methods we used were Categorical Regression Trees (CART) and K-Means Methods. CART is a method used for categorical outcomes to identify classifications given a series of conditions or questions. We re-coded our categorical outcomes to numerical ones to use the CART R method. After cross-validating, we found the optimal maximum depth was four. We then used this knob to generate our predictions and submit to Kaggle. Our CV-score was .66 and Kaggle score was .64. After cross-validating and reviewing our Kaggle score, we decided to move on with another method. Finally, the K-Means Clustering Method was the only unsupervised learning method we tried. K-means is generally used when we have data without defined grouping. The K amount creates the groups in the dataset to classify our predictions. The score given to the model was .73 correct, although we found this model difficult to implement due to the need to identify which cluster belonged to which monster type. Therefore, we decided to finish the KNN analysis. Our exploratory analysis showed clustering in our model and it was this that allowed us to perform a method that utilized the predictor clustering to our benefit. Our best model was K Nearest Neighbors.

Team A
Emily Miller and Bianca Gonzalez
Spring 2017 Statistical Learning
Professor: Albert Kim.

Final Project Write-Up:

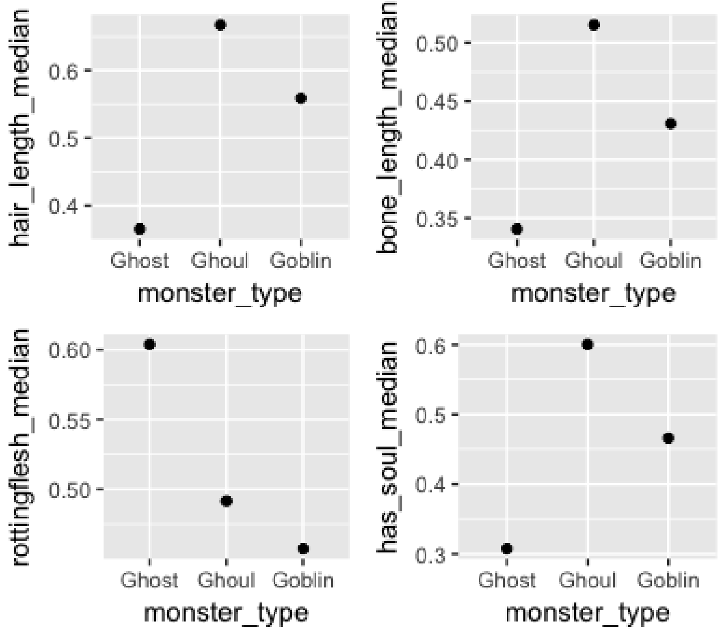

Figure 1.

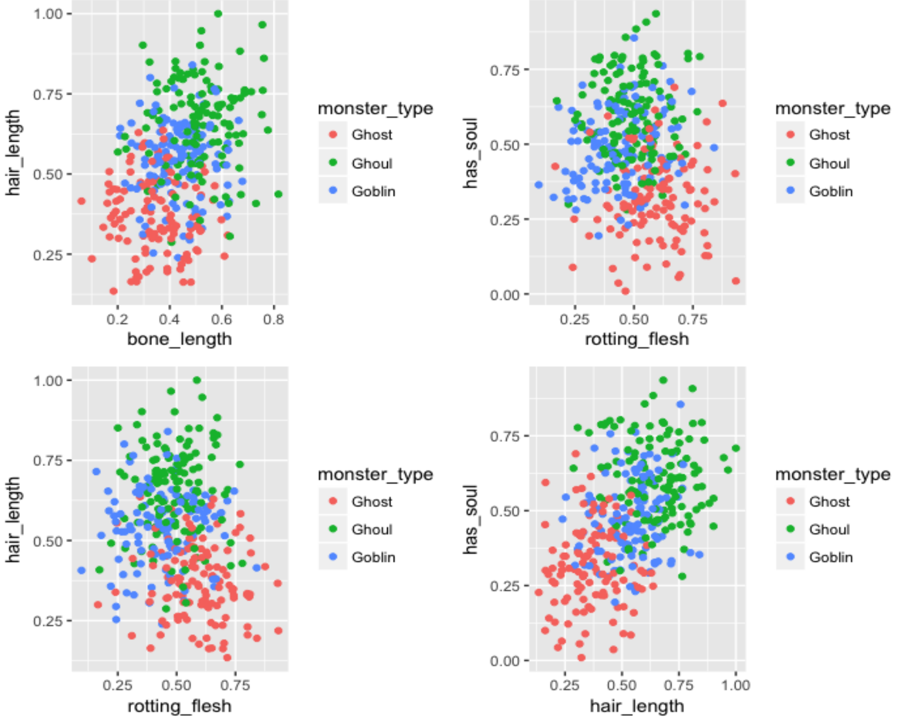

Figure 2.