Final Project Report

Dillan Johnson

Survivor Dataset

Project

My family and I have watched every season of the reality TV show 'Survivor'. We have watched week by week since season 1. Currently season 40 is being aired on TV, and it has just gotten to the point in the game where it has become interesting. Therefore, I searched for a dataset that would allow me to train a model on the game, with my end goal being that I could accurately predict the winner by the time of the 'merge' in the game.

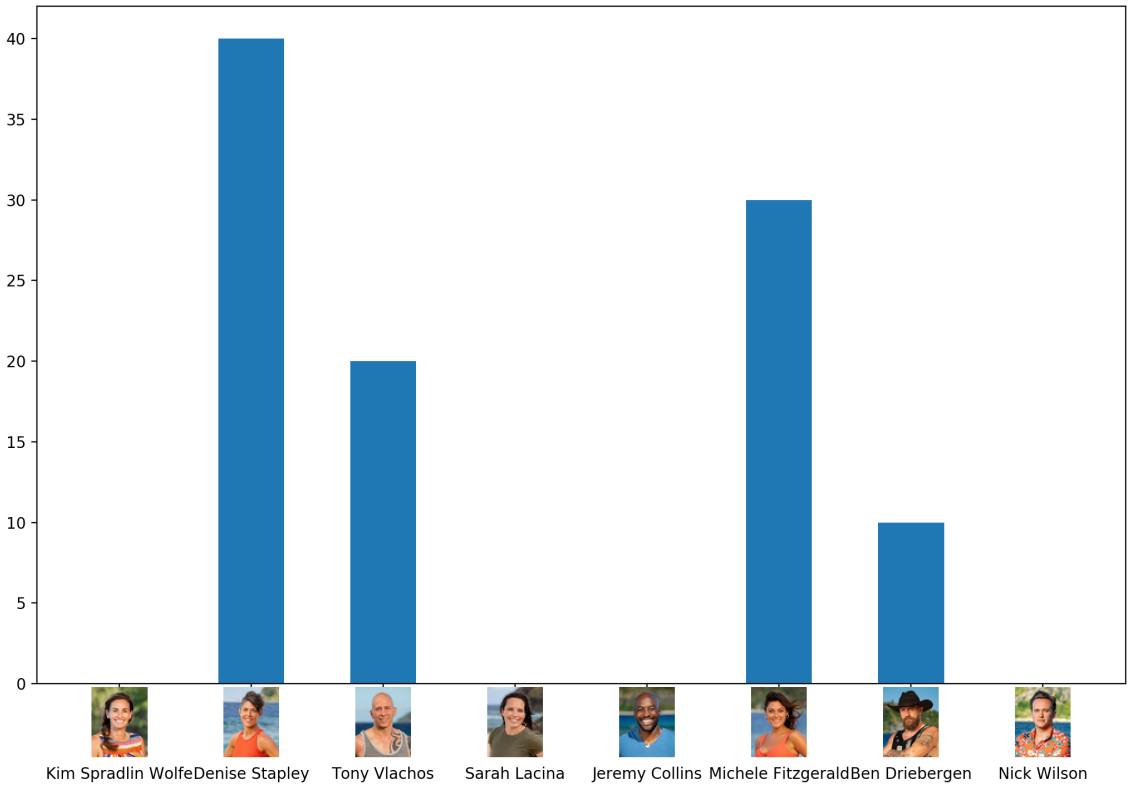
Work

I felt that my weakest project of the semester came from our decision tree ensemble project (NCAA tournament). Which irritated me because I thought using that learning technique had a lot of potential. So throughout the training I was using the different kinds of decision tree ensembles. I trained the model on features that are obtainable throughout the course of the game, that way I can update my predictions week by week as any season is airing.

As with the NCAA dataset, there was much manipulation + formatting that needed to be done with this dataset. With all other projects to this point I was able to access a dataset that was specifically prepped for machine learning. The Survivor dataset that I found (the only detailed one that I could find) was created with no machine learning in mind. This means there were a lot of fields that had random, unique values. So I'd say roughly 40% of my time was spent manipulating data to get in a correct format, and researching the features finding out which ones to keep, etc...

Results

A while back I had read how displaying data is just as important as the data itself, and I thought that was especially true when my first predictions output and I had to go look at the order of the csv file to see which predictions aligned with which contestant. So I made a little program to display the results in a meaningful way. Throughout the different combinations of parameters, and ensemble types, my models liked Tony, Ben, and Denise the most. My best 'Score' came from a regular decision tree (Best Score: 0.5038461538461538)- in which Tony was number 1. But the model that trained and tested the best was a bagging tree classifier; while performing an exhaustive grid search (5 iterations on 3024 candidates – 15120 fits). The predictions from that model for Season40 (current season) are displayed in the photo below.



And the output of that training:

python3 fit_data.py --file-name train-full.csv --model bagging-tree --splitter k-fold --search grid --folds 5 --iterations 5000

Fitting 5 folds for each of 3024 candidates, totalling 15120 fits
[Parallel(n_jobs=-1)]: Done 15120 out of 15120 | elapsed: 11.3min finished

Best Score: 0.40888888888888897
Best Params: {'features__numerical__missing-values__strategy': 'most_frequent', 'features__numerical__numerical-predictors-only__do_numerical': True, 'features__numerical__numerical-predictors-only__do_predictors': True, 'model__bootstrap': True, 'model__bootstrap_features': False, 'model__max_features': 1.0, 'model__max_samples': 0.6, 'model__n_estimators': 10}

Training Labels Correct: 0.9788732394366197
Test Labels Correct: 0.965034965034965