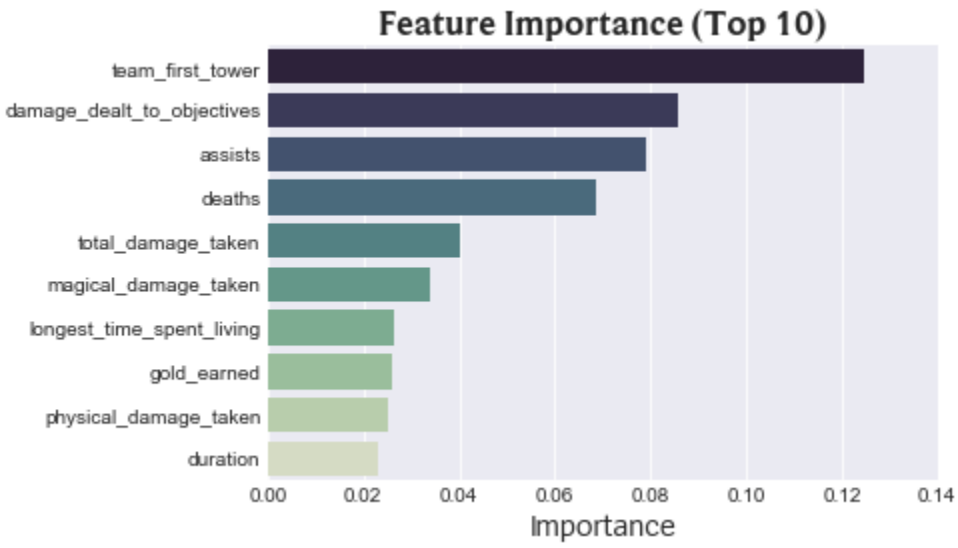
LEAGUE OF LEGENDS

A DATA SCIENCE PROJECT

May 6, 2018

Based on the results of my analysis, I am recommending the implementation of a Random Forest model for use in explaining and predicting match outcomes.

Twelve model classes were tested in total. Although most of these reached a predictive accuracy around 80%, the Random Forest provides some key advantages. First, a Random Forest is based on the Decision Tree model, which provides useful explanatory information like a ranking by importance of the input variables for prediction.



For example, the graph above was created in just a few steps using the "feature importance" function of the Random Forest. It shows which variables had the strongest impact in making predictions more accurate. This is an easy way to

Second, the Random Forest is resource-efficient. As the name suggests, it is simply a collection of hundreds of Decision Tree models, each learning on a random subset of the data. Thus, it can easily be segmented to take advantage of parallel processing and distributed computing. That makes this model especially suited to scale up for much larger projects.

The metrics we used to evaluate and compare each are as follows:

- **Accuracy**—the overall "success rate" of the model predictions, taken as the average of individual accuracy scores from a 10-fold cross-validation.[1]
- **Lift**—the margin by which Accuracy improves upon the Base Accuracy Rate.
    - The Base Accuracy is the score we'd expect with random uninformed guesses. Because the underlying data is split almost perfectly between Wins and Losses, the BAR is (roughly) a convenient 50%.
- **CV Spread**—this is the difference between the largest and smallest Accuracy scores from cross-validation. A spread of more than a few points indicates inconsistency. It suggests that the model suffers from *overfitting*—a condition where the training data is "memorized" so well that the model has difficulty with new data. (Think "studying to remember the dates of historical events, and then being tested over why they were important".)

Model Comparison

Model	Accuracy	Lift	CV Spread
DT	69.36%	18.75%	12.66%
XT	79.73%	29.13%	11.84%
RF	81.82%	31.21%	9.27%
BDT	78.19%	27.59%	12.82%
GB	81.05%	30.44%	7.59%
BAG	80.78%	30.18%	8.27%
LOG (L2)	80.13%	29.53%	8.97%
LOG (L1)	80.66%	30.05%	7.29%
NB-B	74.27%	23.66%	8.47%
KNN-15	73.51%	22.91%	7.09%
KNN-36	75.98%	25.37%	10.53%
STK-5	82.08%	31.47%	8.92%
SVM (Lin)	80.79%	30.19%	9.89%
SVM (RBF)	81.18%	30.57%	7.69%
ANN-1	80.78%	30.18%	9.62%

[1] 10-fold cross-validation involves slicing the data into 10 equal pieces, using nine to train the model and the tenth slice for testing it. This is repeated 10 times, so that each slice is held out once for testing. This is an important practice for ensuring that model performance is consistent—and thus useful for new data.

Results

It is a bit unusual to see several classes of models, with fundamentally different prediction methods, all arrive at virtually the same results. Also concerning is the recurring issue with overfitting. Ideally, a model should perform equally well on every fold of a cross-validation. However, even the best models resulted a 7-point spread in accuracy scores.

The apparent cap on model performance, coupled with the overfitting issue, leads me to conclude that the sample size (1100 matches) may be too small.

As previously mentioned, the Random Forest model is easily scalable. If this project were to be scaled up, even into the hundreds of thousands of observations, this would remain a viable model. Such an increase in scale would be possible—limited only by the rate at which the Riot Games API can be accessed, and by the computing power needed to process a larger dataset.

Going forward, I plan to investigate the feasibility of a larger-scale project, aggregating match data from potentially hundreds of users. As far as implementation, one potential application might be in Bayesian conditional probability—using prior data to estimate the outcome of an ongoing match, as new information becomes known.

Summary:

- **Random Forest** (RF) is the recommended model to implement. It predicts with 81% accuracy, and brings other advantages in interpretability and efficiency.