**Modeling & Evaluation**

* Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?

We basically want to predict the next pitch type. The heuristic baseline is:

Pitch type: FF with percentage: 35.593%

Pitch type: SL with percentage: 14.594%

Pitch type: FT with percentage: 12.922%

Pitch type: CU with percentage: 10.342%

Pitch type: CH with percentage: 8.181%

Pitch type: SI with percentage: 8.119%

Pitch type: FC with percentage: 5.438%

Pitch type: KC with percentage: 2.173%

Pitch type: other with percentage: 2.638%

So our problem is a classification problem, and for classification we decided to use logistic regression, decision tree and random forest.

We use l2-regularized logistic regression with normalized data .

Pros:

1. Logistic Regression is pretty efficient dealing with our data(since it has more than 2M rows), actually, LR is the fastest algorithm we used.
2. Logistic Regression has less parameters to be tuned which make it easier to improve.
3. LR has low variance and so is less prone to over-fitting

Cons:

1. Since we have a mix of multi-class and numerical features and our target variable is also multi-class, we need to create several dummy variables in order to get better result.
2. LR requires that each data point be independent of all other data points. If observations are related to one another, then the model will tend to overweight the significance of those observations.
3. LR is highly biased

Then we tried decision tree to improve the performance:

Pros:

1. DT is unbiased
2. DT can handle multi-class data natively
3. DT is almost as fast as LR

Cons:

1. DT is prone to over fitting
2. DT has high variance
3. Hard to visualize
4. More parameters to be tuned

Random forest is the best algorithm we find:

Pros:

1. RF has independent trees to the effect of correlation is minimized
2. RF, like DT, can handle very well high dimensional spaces as well as large number of training examples
3. Always performs better than DT

Cons:

1. RF is the most slow method among the three we used
2. Even more parameters than DT to be tuned

* Identify an appropriate baseline model and report its performance.

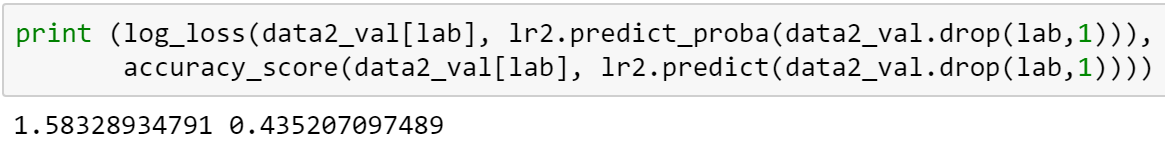
We use l2-regularized logistic regression with normalized data as our baseline model. We used accuracy and Log-loss to assess the performance of this model:

The Log-loss is1.58328934791 and the accuracy is 0.435207097489

* Describe an evaluation framework you will use to improve upon the baseline.

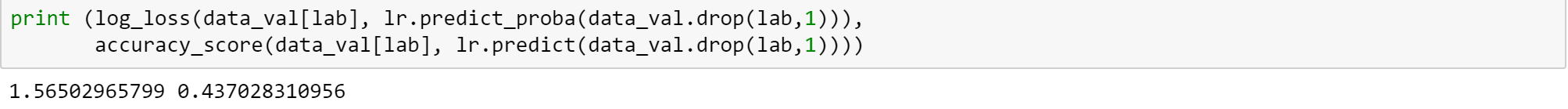
We want to improve the accuracy and decrease the Log-loss of our model. To improve the baseline model, we will try different models with several combinations of hyper parameter settings. The model with lowest Log-loss and highest accuracy will be our optimal choice.

* Perform an analysis of possible algorithms and use the data science experimental framework to choose an optimal candidate.
* Demonstrate how you were able to improve upon the baseline and document the process of doing so.



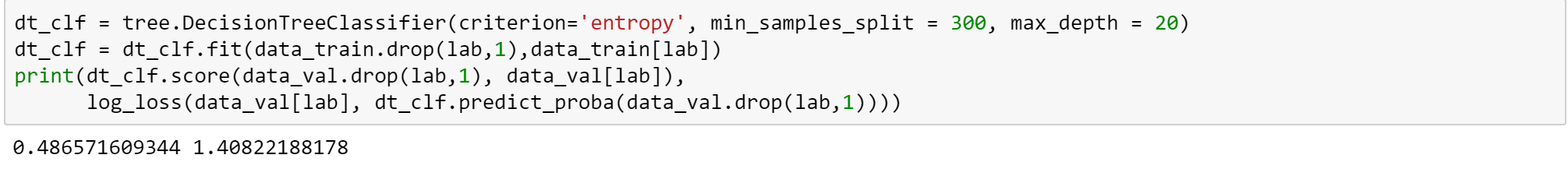
We first created dummy variables for multi-class features and then ran the l2-regularized logistic regression from which we have following result:

Log-loss decreased to 1.56502965799 and accuracy increased to 0.437028310956



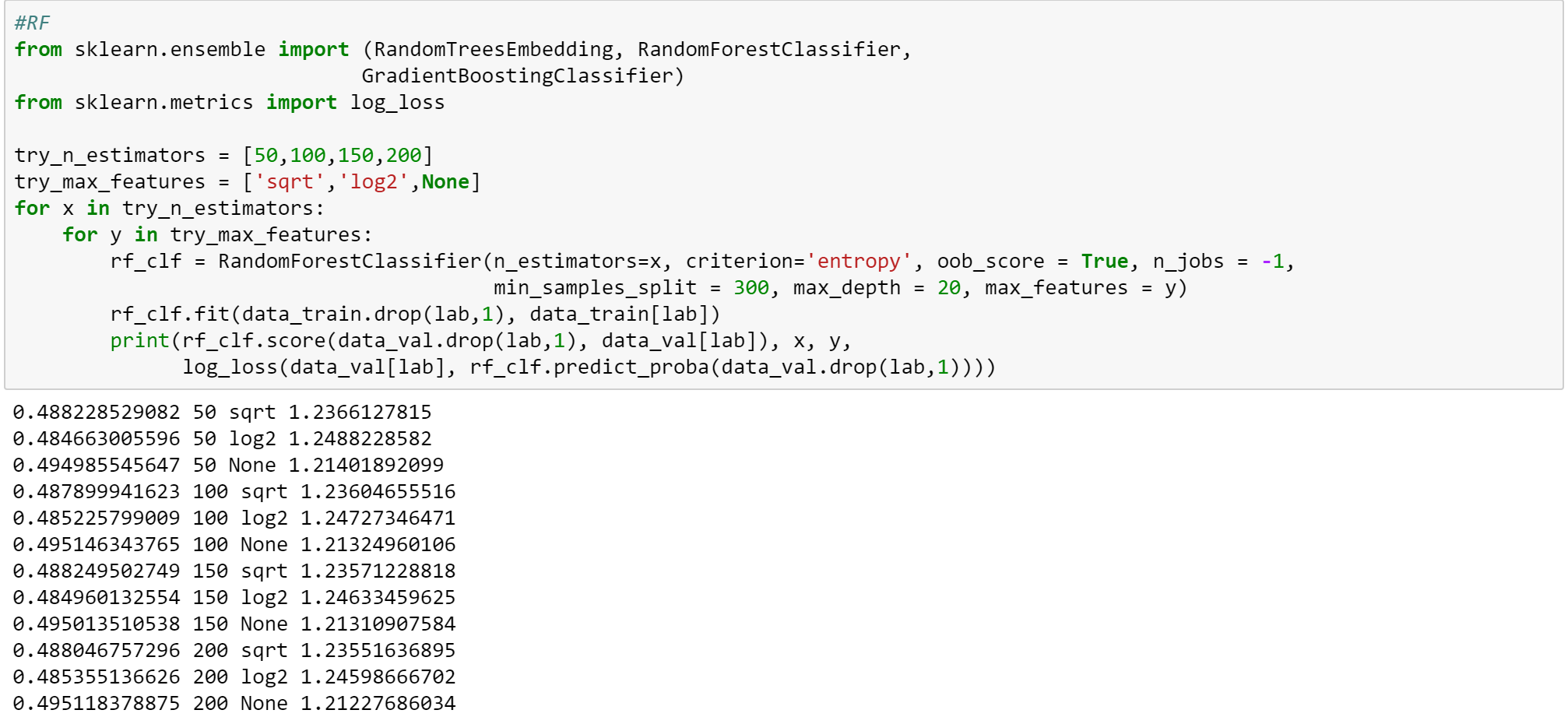
We then used DT to see if we can lower the Log-loss and improve accuracy. After trying out different hyper parameters settings, our best DT returns result:

Log-loss decreased to 1.40822188178 and accuracy increased to 0.486571609344



We eventually used RF to improve the baseline model and the best result we got is:

Log-loss decreased to1.21324960106 and accuracy increased to 0.495146343765



We tried to do feature selection and used various subsets of feature to run our models. Without exception, using all the features we have performances better than using part of features. Since our dependent variables have small correlation, most of the pairwise correlations are less than0.02, this could mean that we can improve our model by adding more useful features.

* Discuss why and how this model should “solve” the business problem (i.e., improve along some dimension of interest to the firm).

Deployment

* Discuss the type of evaluation metric that should be used to choose the best algorithm. How does this metric relate to the business problem?

Net Gain metric