**Project Report**

尤凱YOU Kai 1155076734,

**Part 2. The Dataset and pre-processing**

2.2.3

the number of horses: 2155

the number of jockeys: 105

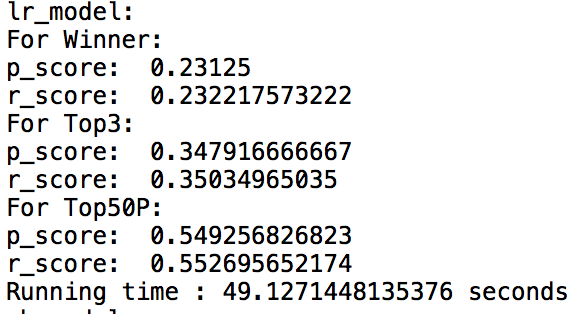
the number of trainers: 93

**Part 3. Classification**

The prediction results are stored as csv file in the “predictions” folder.

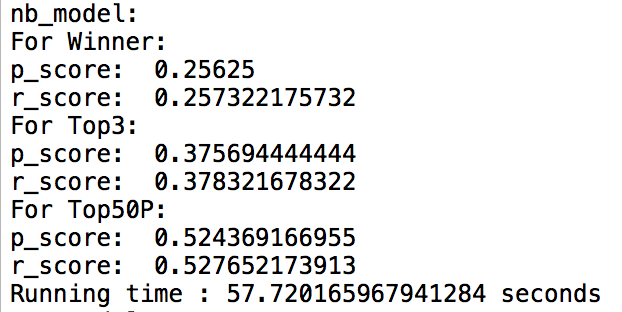
3.1.1 Logistic Regression

The running time and prediction evaluation with parameters that I chosen (after adjustment and comparison):



3.1.2 Naïve Bayes

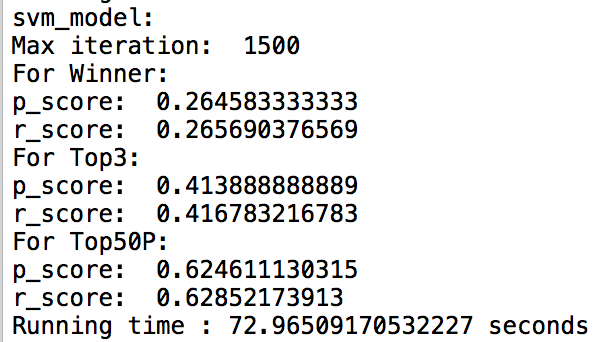
The running time and prediction evaluation with parameters that I chosen (after adjustment and comparison):



For the problem “which Naïve Bayes classifier to choose”, I choose the MultinomialNB classifier. The reason is that 1) this classifier get the largest score after comparison between others. (Although it depends on the parameters chosen. It’s the general result.) 2) The MultinomialNB classifier is the most suitable model because it fits the condition in Horse Racing.

3.1.3 SVM

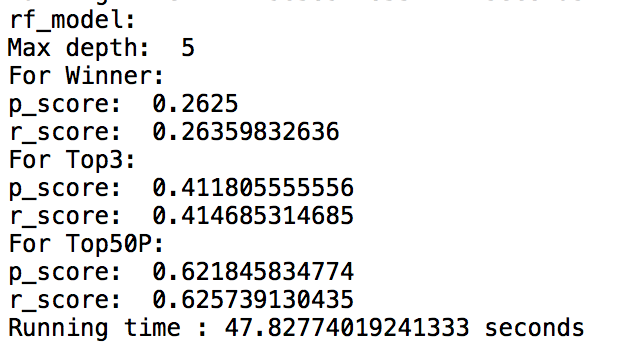
The running time and prediction evaluation with parameters that I chosen (after adjustment and comparison):



For the problem “which Kernel function to choose”, I choose the “rbf” function in this project. Because the function get the best prediction result between others and it’s the default one.

3.1.4 Random Forest

The running time and prediction evaluation with parameters that I chosen (after adjustment and comparison):



3.4 Additional questions

Q: What are the characteristics of each of the four classifiers? (2 pts)

A: Logistic regression is a regression model where the dependent variable (DV) is categorical.

Naive Bayes classifiers are a family of simple "probabilistic classifiers "based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Q: Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples.

A: The different classification models fit different conditions. It mostly depends on the training data types. In general, for one specific scenario provided, we could first select some certain classification models depends on the training data type. Then, train it and evaluate the prediction results to find the most suitable one.

For example, the logistic regression model fit the condition where the output can take only two values ("0" and "1"). And the Bernoulli naive Bayes model in Naive Bayes family also fits well. We could choose several of models and compare them.

Q: How do the cross validation techniques help in avoiding overfitting?

A: Cross Validation reserves a particular sample of a dataset on which it does not train the model. Later, it tests the model on this sample before finalizing it. In this way, it limits problems like overfitting, giving an insight on how the model will generalize to an independent dataset.

Q: In addition to the Precision-Recall metric, there are many other metrics can be derived according to the confusion matrix.

How do you choose evaluation metrics for imbalanced datasets according to the class distribution? Please give your understanding and provide some examples.

A: In my opinion, the choice of evaluation metrics should depend on the target of our training and prediction. For example, in this project, we care about how to bet and win money using our prediction model. So the prediction accuracy of positive prediction and positive class is obviously what we care more. So the Precision-Recall metric is better than other evaluation metrics here.