

CSCI3320 Programming Project

Groupmate:

1.Tang Ying Kin 1155079801

2.Cheung Ka Chai 1155078606

Part2. The dataset and preprocessing

✍ Write down the number of horses, the number of jockeys, and the number of trainers in `prjreport.pdf`.

Ans: Number of horses: 2155 Number of jockeys 105 Number of trainers 93

Part3. Classification

3.4 Writing A Report

After you obtain all the results, you are required to write a brief report in the file `prjreport.pdf` to answer the following questions. You can write freely as long as you answer the questions clearly since there is no strict format requirement for the report.

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

✍ Q: Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples. (2 pts)

✍ Q: How do the cross validation techniques help in avoiding overfitting? (2 pts)

✍ Q: In addition to the Precision-Recall metric, there are many other metrics can be derived according to the confusion matrix, e.g., the true negative rate $TNR = \frac{TN}{TN+FP}$, the negative predictive value $NPV = \frac{TN}{TN+FN}$ and so fourth. How do you choose evaluation metrics for imbalanced datasets according to the class distribution? Please give your understanding and provide some examples. (2 pts)

3.4

Q1)

Logistic regression:

It is an appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Gaussian Naive Bayes:

It can perform online updates to model parameters via `partial_fit` method.

Multinomial models:

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

Multivariate Bernoulli models:

Like MultinomialNB, this classifier is suitable for discrete data. The difference is that while MultinomialNB works with occurrence counts, BernoulliNB is designed for binary/boolean features.

SVM:

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Random Forest:

It is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

Q2)

First, we would consider the size of the training set. If the size of the training set is small, high bias classifiers such as Naïve Bayes will have an advantage over low bias classifiers, as the latter one will be overfitted.

Also, we would consider the advantages (and disadvantages) of each classifiers.

Advantages of Naive Bayes:

It is simple as its algorithm is doing counting. If the Naive Bayes conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, and thus we need less training data. And even if the Naive Bayes assumption doesn't hold, a Naive Bayes classifier still often does a great job in practice.

Disadvantage of Naive Bayes:

It can't learn interactions between features

Advantages of Logistic Regression:

Lots of ways to regularize the model, and we don't have to worry as much about our features being correlated, like we do in Naive Bayes. We also have a nice probabilistic interpretation, unlike SVMs, and we can easily update our model to take in new data (using an online gradient descent method), again unlike SVMs.

Advantages of SVMs:

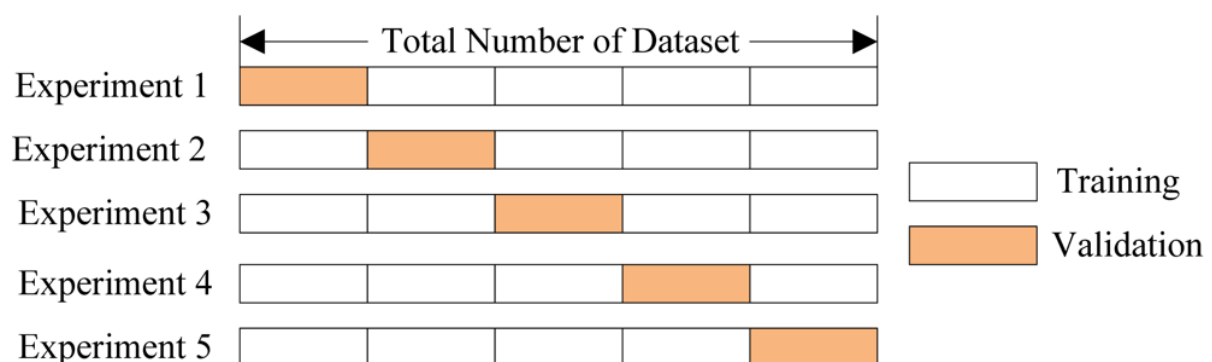
High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if the data isn't linearly separable in the base feature space.

Advantages of Random Forests:

It is fast and scalable, and we don't need to worry about tuning a bunch of parameters like we do with SVMs

Q3)

In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality. For example, we could have 5 folds or experiments. We divide the data into 5 pieces, each being 20% of the full dataset.



We run an experiment called experiment 1 which uses the first fold as a holdout set, and everything else as training data. This gives us a measure of model quality based on a 20% holdout set, much as we got from using the simple train-test split. We then run a second experiment, where we hold out data from the second fold (using everything except the 2nd fold for training the model.) This gives us a second estimate of model quality. We repeat this process, using every fold once as the holdout. Putting this together, 100% of the data is used as a holdout at some point. Returning to our example above from train-test split, if we have 5000 rows of data, we end up with a measure of model quality based on 5000 rows of holdout (even if we don't use all 5000 rows simultaneously).

Q4)

Use precision and recall to focus on small positive class

When the positive class is smaller and the ability to detect correctly positive samples is our main focus (correct detection of negatives examples is less important to the problem) we should use precision and recall.

Use ROC when both classes detection is equally important

When we want to give equal weight to both classes prediction ability we should look at the ROC curve.

Use ROC when the positives are the majority or switch the labels and use precision and recall

When the positive class is larger we should probably use the ROC metrics because the precision and recall would reflect mostly the ability of prediction of the positive class and not the negative class which will naturally be harder to detect due to the smaller number of samples. If the negative class (the minority in this case) is more important, we can switch the labels and use precision and recall (As we saw in the examples above — switching the labels can change everything).

Part4. Regression

4.1.1 Support Vector Regression Model(SVR)

✍ Parameter tuning/selection is required. First, SVR accepts different kernel functions. They could be one of linear, poly, rbf, sigmoid, precomputed, select one of them and state your reason in `prjreport.pdf`. Second, `epsilon` and `C` are two critical parameters. Please state what role do they play in the model, what value do you assign and why do you select these values in `prjreport.pdf`

Ans:

First, I would select linear as kernel function of the model. Because some of features are related to the finishing time of a horse linearly. For example, draw of horse and weight of a horse might put high relation on the finishing time of horse on a race because draw of horse would affect the actual distance for a horse to complete a race. (there are some an edgeover the field as a horse would have a shorter distance to the bend if a horse has better draw) So use linear model would have a more accurate result. Other complex model would cause overfitting problem in the training set so linear model is enough. Second, `C` is the constant penalty added to the error function to reduce overfitting. `Epsilon` is the margin of error function to accept small error which is smaller than `epsilon`. In the program I assign `C=2` , `epsilon = 0.2` because high values on `C` and `epsilon` may cause underfit which means the model may not generalization the relationship between features and label. So `C = 2` and `epsilon` is enough.

4.1.2 Gradient Boosting Regression Tree Model(GBRT)

Ans:

✍ Parameter tuning/selection is required. First, `GradientBoostingRegressor` accepts different loss functions. They could be one of `ls`, `lad`, `huber`, `quantile`, select one of them and state your reason in `prjreport.pdf`. Second, `learning_rate`, `n_estimators` and `max_depth` are three critical parameters. Please state what role do they play in the model, what value do you assign and why do you select these values in `prjreport.pdf`

I use `ls` as loss function in GBRT because least square is a simple way to do the curve fitting and could have a better generalization on race time of horse.

`Learning_rate` is used in optimization the loss function through iteration because the loss function will converge finally. So learning rate determine the number of step that can be used to optimize the model. `n_estimators` is the number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance. Max depth is maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree. Too large depth of tree may result in long computation time on building the tree and efficiency of program may be low and the overfitting may happen as there are too many node while boosting as GBRT is robust to reduce bias and variance.

Therefore I set `learning_rate = 0.2` , `n_estimators = 200` and `max_depth = 6` to build a model to generalization the regression as `max_depth` should be keep in low because it can reduces steps for computation.

4.2 Predicting on Test Data (10pts)

- Record your best result in the form (model_name, RMSE, Top_1, Top_3, Average_Rank) for both SVR and GBRT model. Here, you are required to save your best result together with chosen parameters in `prjreport.pdf` and we will score this part by your prediction results.
- Sometimes your result will be dramatically better if you have done some data normalization over both `training_data` and `training_label`. Observe that finish time of different horses is sometimes really close to each other and other features like 'declared_horse_weight' are much larger than others. Please try to normalize them and retrain your model to show whether normalization improves the result. Here, scikit-learn has provided some useful tools [9] like `StandardScaler` in the module of `sklearn.preprocessing`.

Note that you have to use the mean and variance of the **training data** to normalize/rescale the prediction result to get RMSE. Please state your prediction RMSE with and without normalization and analyse the result in `prjreport.pdf` for both SVR and GBRT model.

Evaluation of SVR model with normalization

Mean square error = 1.8965

C:\Users\Tang Ying Kin\Anaconda3\lib\site-packages
r instead, but in the future will perform element

Top1_accuracy: 0.20625

Top3_accuracy: 0.4666666666666667

Average Rank of all predicted top1 horse: 4.6542

Evaluation of SVR model without normalization

Mean square error = 18.6836

Top1_accuracy: 0.08333333333333333

Top3_accuracy: 0.28541666666666665

Average Rank of all predicted top1 horse: 6.3917

Evaluation of GBRT model with normalization

Mean square error = 1.6188

Top1_accuracy: 0.20208333333333334

Top3_accuracy: 0.41875

Average Rank of all predicted top1 horse: 5.0687

Evaluation of GBRT model without normalization

Mean square error = 1.6165

Top1_accuracy: 0.20625

Top3_accuracy: 0.4166666666666667

Average Rank of all predicted top1 horse: 5.0563

(ps:All are rms values , typo mistake)


Ans:

Normalization over the training set could have better result on the accuracy because every feature has different variance and means , it may not be accurate to do regression without

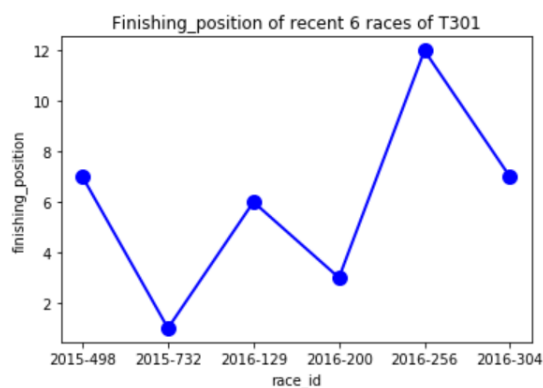
normalization. From the above result we can see that normalization give a great effort on the accuracy of the prediction even in SVR and GBRT. And GBRT get the advantages less than SVR because it is a boosting method. It will eventually converge even it has not been normalized.

6. Visualization

6.1 Line Chart of Recent Racing Result (4 pts)

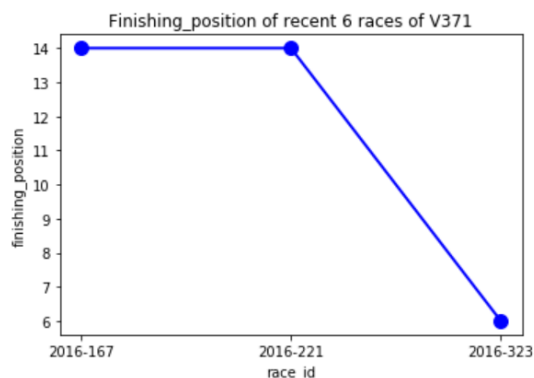
 Select two horses that you are interested in, put down the plots of these two horses in your report, and briefly describe what you observe from the plots.

1. Horse ID: T301



2. Horse ID : V371

Please input a horse ID and the program will generate a line chartV371

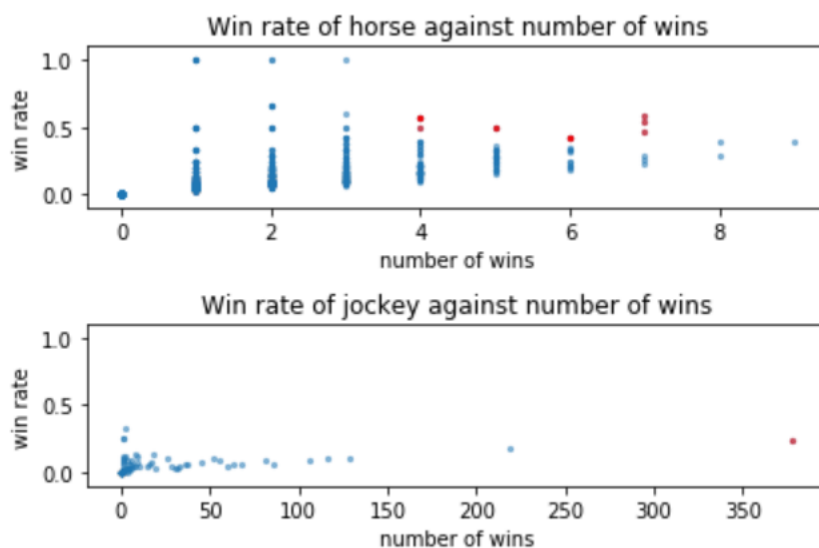


The first horse I interested maintain average result in first 4 race of recent 6 races but it suddenly drop to 12th position in 5th race.

The second horse we interested is that even through it only has 3 recent races , it became the last in 2 of these 3 races!

6.2 Scatter Plot of Win Rate and Number of Wins (4 pts)

- ✎ Put down the plot in your report, and write down the “best” horse and the “best” jockey in your opinion, and briefly explain why.



Best horse:P303 win rate = 0.5833333333333334 number of wins = 7

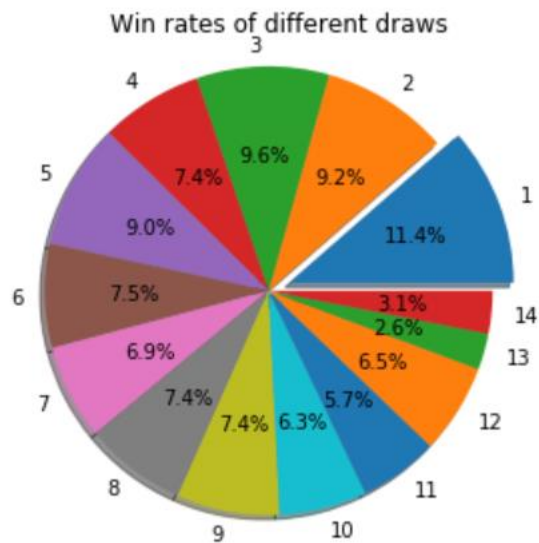
Best jockey:J Moreira win rate = 0.2445019404915912 number of wins = 378

For the best horse , it has 7games in total and it kept almost 0.6 win rate.

For the best jockey it wins 378 games and it has 0.2445 win rate , it is a very high win rate we can see from the graph and he maintains almost the highest win rate among all jockeys in training set.

6.3 Pie Chart of the Draw Bias Effect (4 pts)

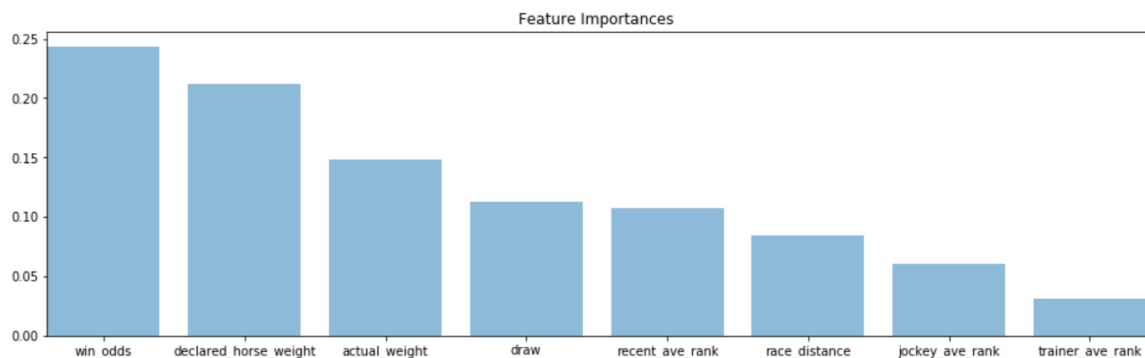
✍ Put down the plot in your report, briefly describe what you observe from the plots, and answer whether low draws really have a considerable advantage?



From the pie chart we can see that a low draws would give a considerable chance to the win rate for a horse as it can reduce the total distance for a horse to run. Hence finish time for a horse would also reduce so it is more easier to win for low draws.

6.4 Bar Chart of the Feature Importances (4 pts)

✍ Put down the plot in your report, and briefly describe what you observe from the plots.

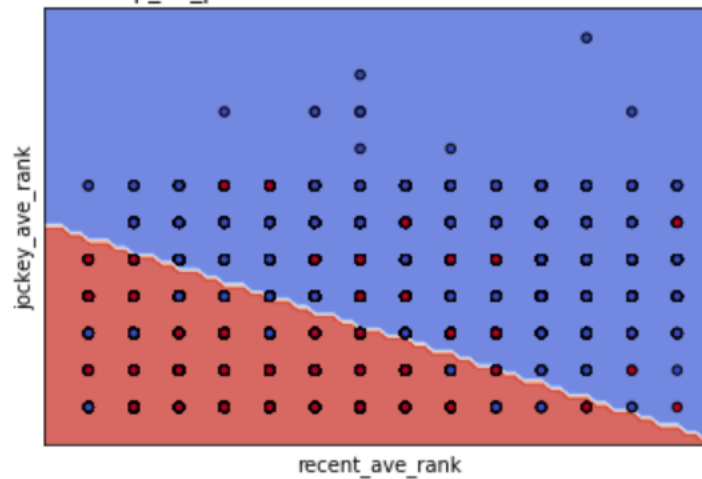


From the graph we would see that win_odds, declared_horse_weight, actual_weight, give most importance on the finishing position of a horse as the finishing position depends on the times, and horse weight affect the speed of horse and hence the finishing position would be affected. win_odds is depends on the past result of the horse so it reflect the performance of a horse so it can reflect the current result of a horse.

6.5 Visualize SVM (4 pts)

✍ Put down the plot in your report, and briefly describe what you observe from the plots.

SVM visualization of top_50_percent classification on recent rank and jockey ave rank



Red colour represent a horse is in top_50_percent in a race with jockey rank and recent rank. Otherwise it is in blue colour. From the graph we can see that the higher rank of jockey rank and recent rank of corresponding horse, it has higher chance to be the top 50 per cent in a race.