CSCI3320 Programming Project

Groupmate:

1.Tang Ying Kin 1155079801

2.Cheung Ka Chai 1155078606

Part2. The dataset and preprocessing

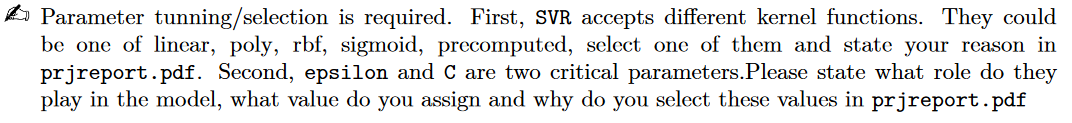


Ans:

Part3. Classification

Part4. Regression

4.1.1 Support Vector Regression Model(SVR)

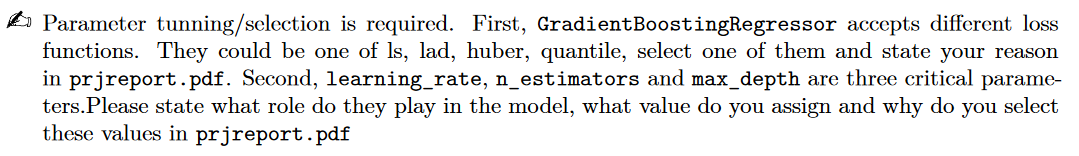


Ans:

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 edge over 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:

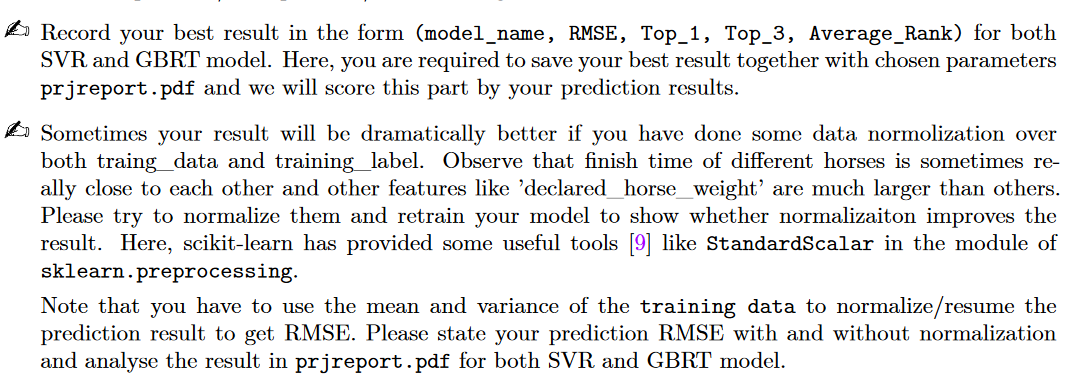


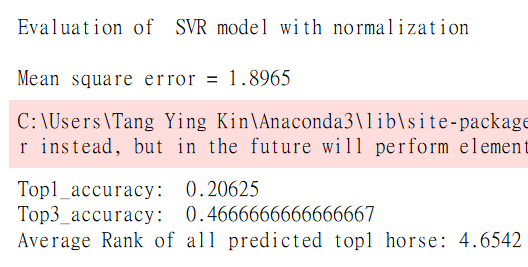
I use ls as loss function in GBRT because least square is a simple way to do the curve fitting and could have a better generalization on race time of horse.

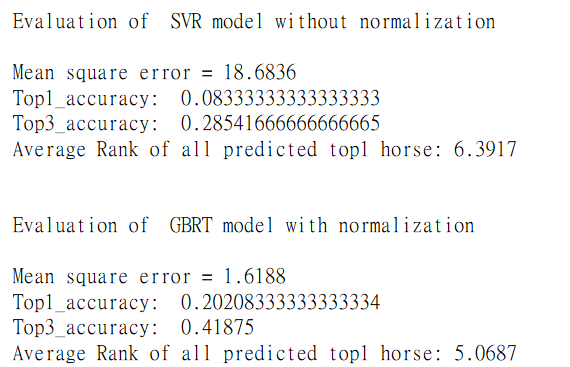
Learning\_rate is used in optimization the loss function through iteration because the loss function will converge finally. So learning rate determine the number of step that can be used to optimize the model. n\_estimators is the number of boosting stages to perform. Gradient boosting is fairly robust 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.

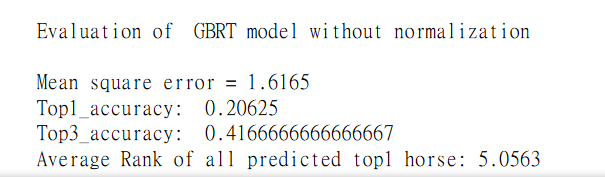
Therefore I set learning\_rate = 0.2 , n\_estimators = 200 and max\_depth = 6 to build a model to generalization the regression as max\_depth should be keep in low because it can reduces steps for computation .

4.2 Predicting on Test Data (10pts)







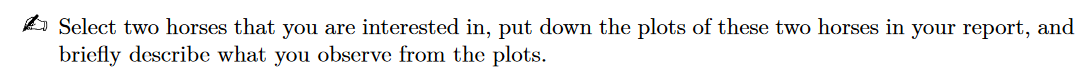


Ans:

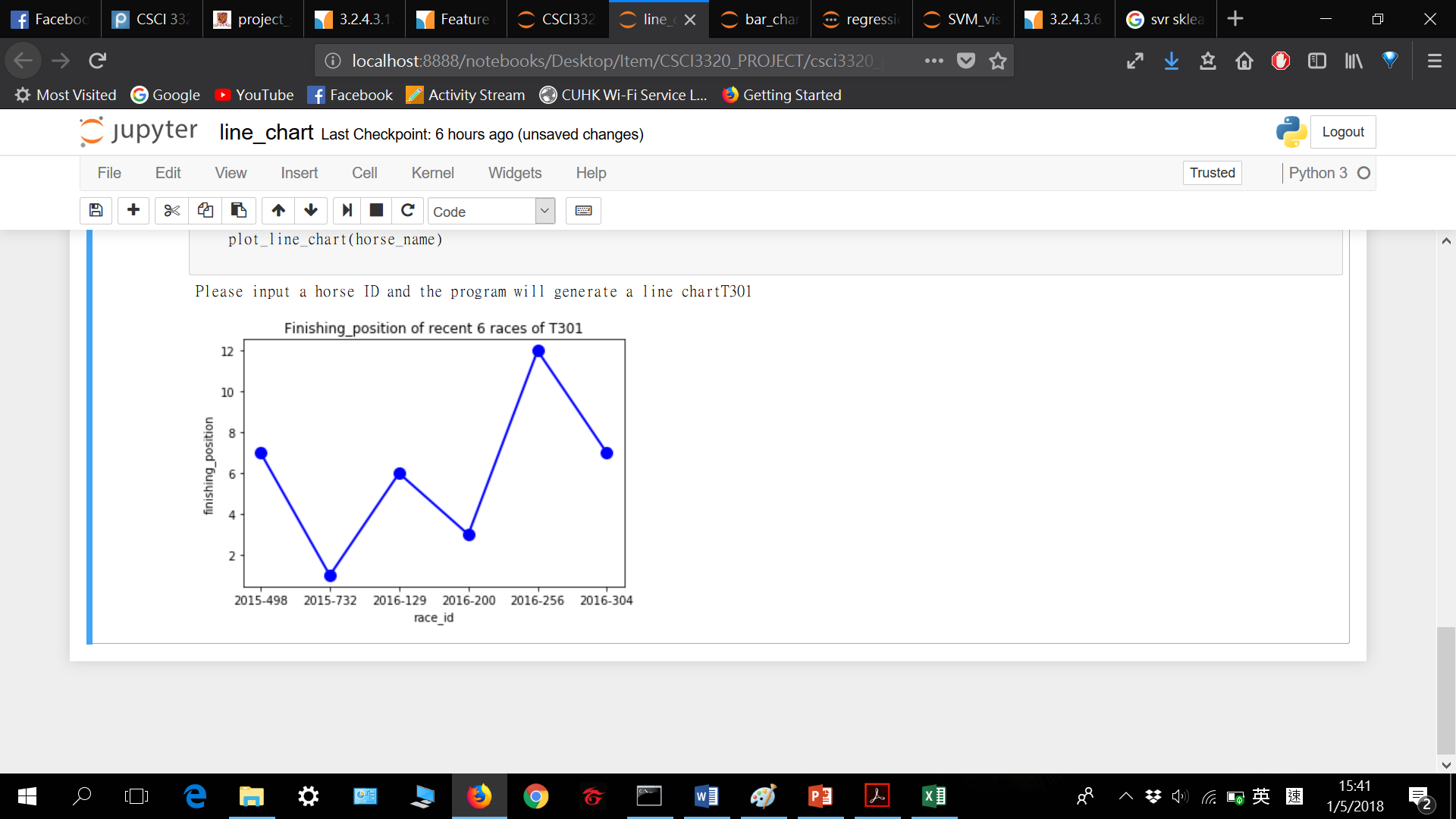
Normalization over the training set could have better result on the accuracy because every feature has different variance and means , it may not be accurate to do regression without normalization. From the above result we can see that normalization give a great effort on the accuracy of the prediction 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

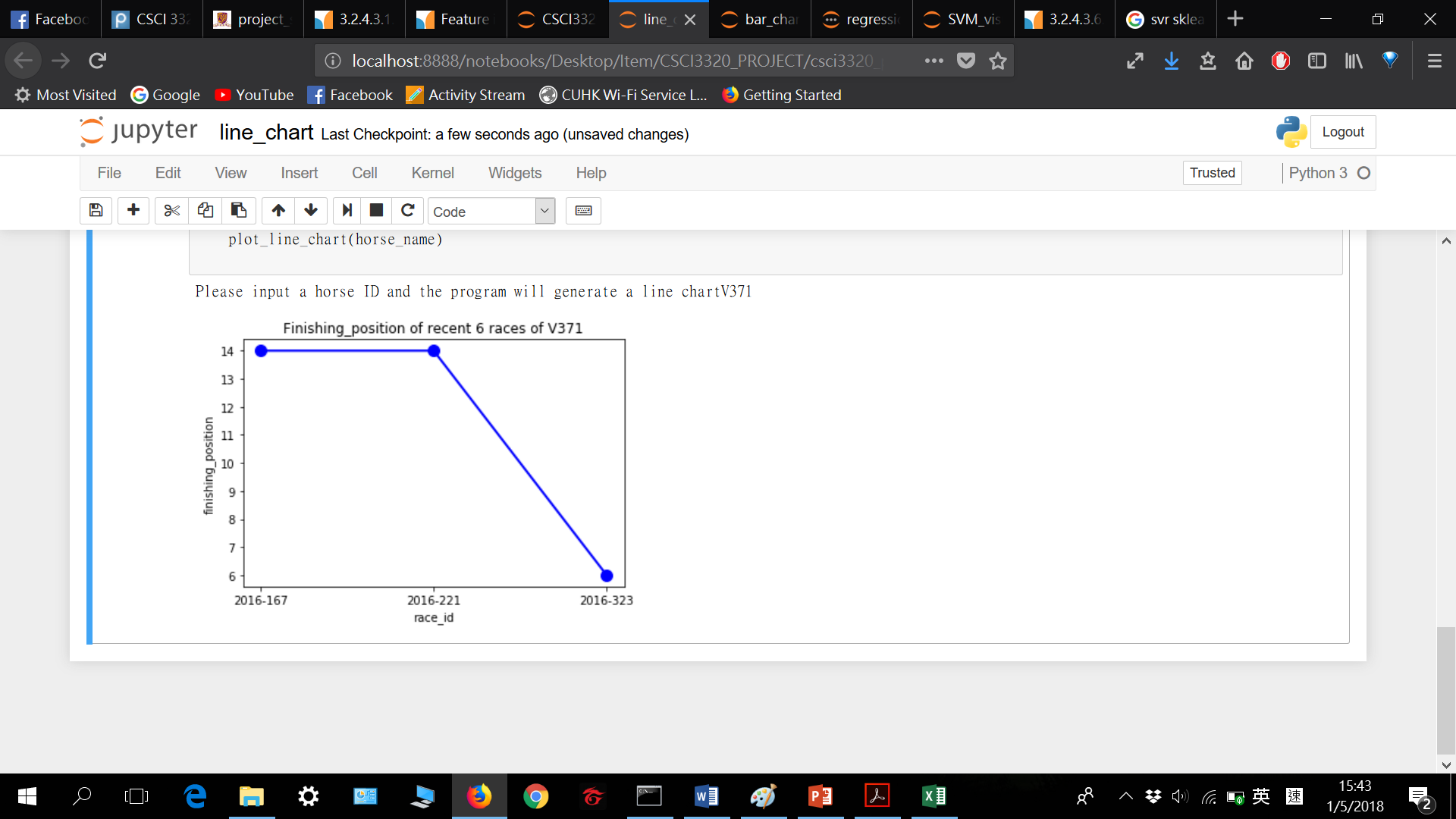
6.1 Line Chart of Recent Racing Result (4 pts)



1. Horse ID: T301



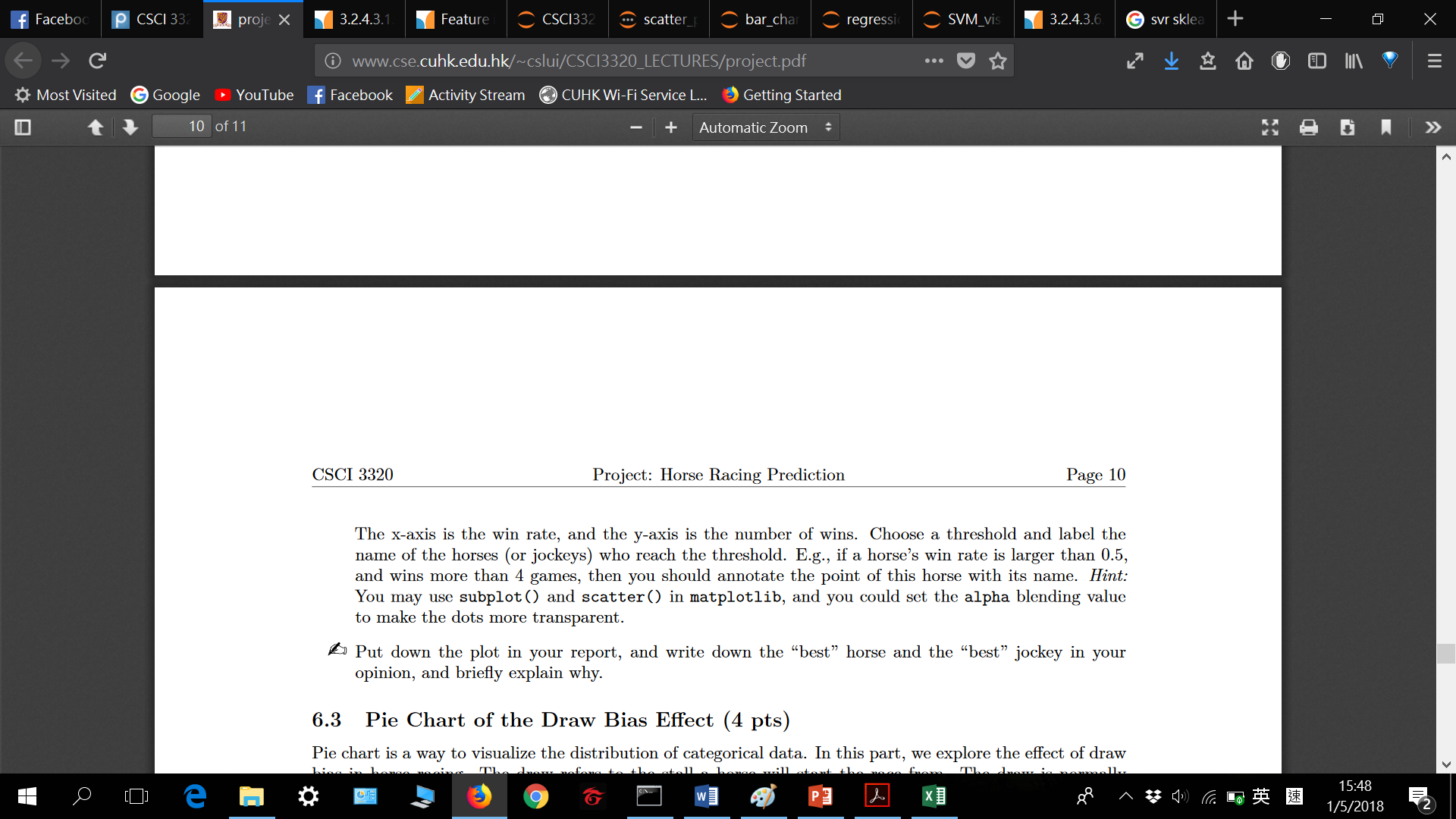
1. Horse ID : V371

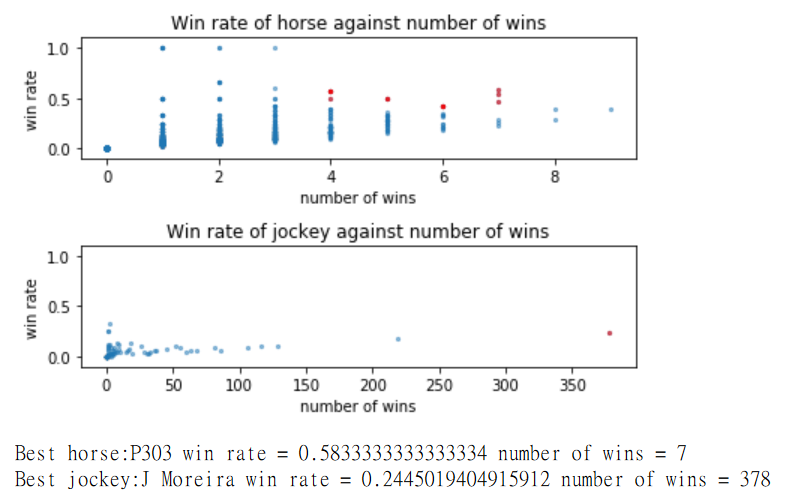


The first horse I interested maintain average result in first 4 race of recent 6 races but it suddenly drop to 12th position in 5th race.

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)

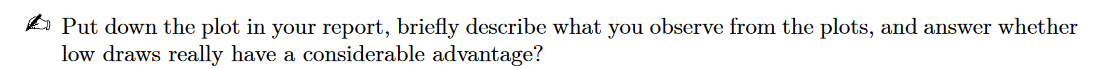


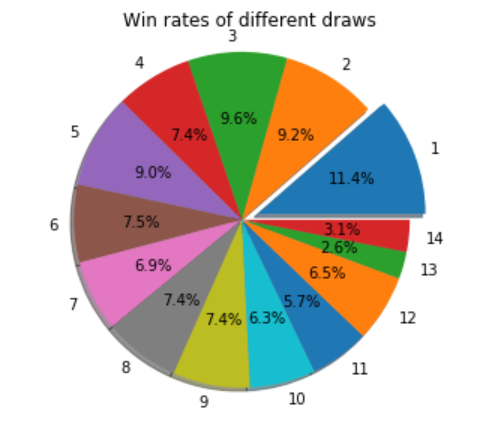


For the best horse , it has 7games in total and it kept almost 0.6 win rate.

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.

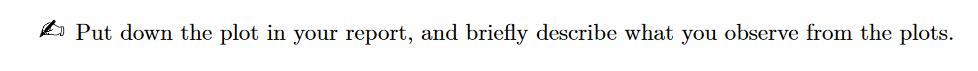
6.3 Pie Chart of the Draw Bias Effect (4 pts)

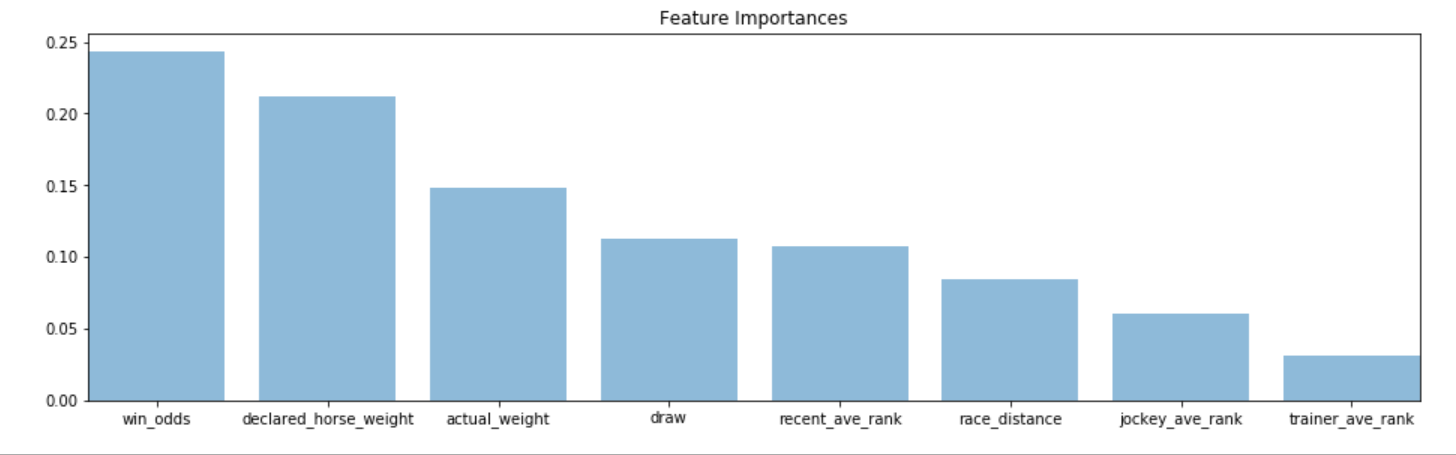




From the pie chart we can see that a low draws would give a considerable chance to the win rate for a horse as it can reduce the total distance for a horse to run. Hence finish time for a horse would also reduce so it is more easier to win for low draws.

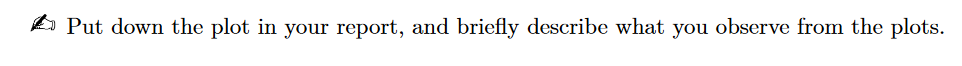
6.4 Bar Chart of the Feature Importances (4 pts)

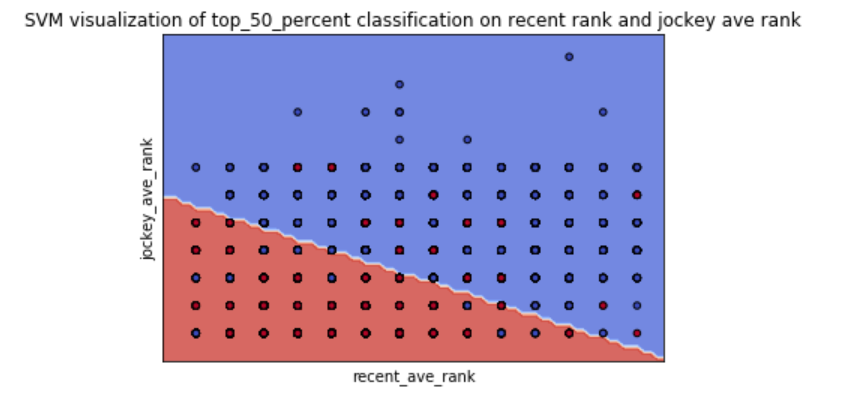




From the graph we would see that win\_odds declared\_horse\_weight , actual\_weight , give most importance on the finishing position of a horse as the finishing position depends on the times , and horse weight affect the speed of horse and hence the finishing position would be affected win\_odds is depends 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)





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.