Our Group’s topic is wine quality, we use some features of wine to predict the quality of the wine.

Firstly, I’d like to give an overview of what we did in this project. To begin with, we did some basic data analysis using graphs. Then, we set up models using four different methods and made prediction. Last, we tested the prediction to see which model fits best.

Here are methods we used to set up models. Firstly, we tried linear regression. As there might be some nonlinear relationships, we also tried simple one-layer neural network to see how it works. And then, we used a more complicated ensemble neural network. Last, we wanted to try a method which works in a different way from above. Thus, we did random forests.

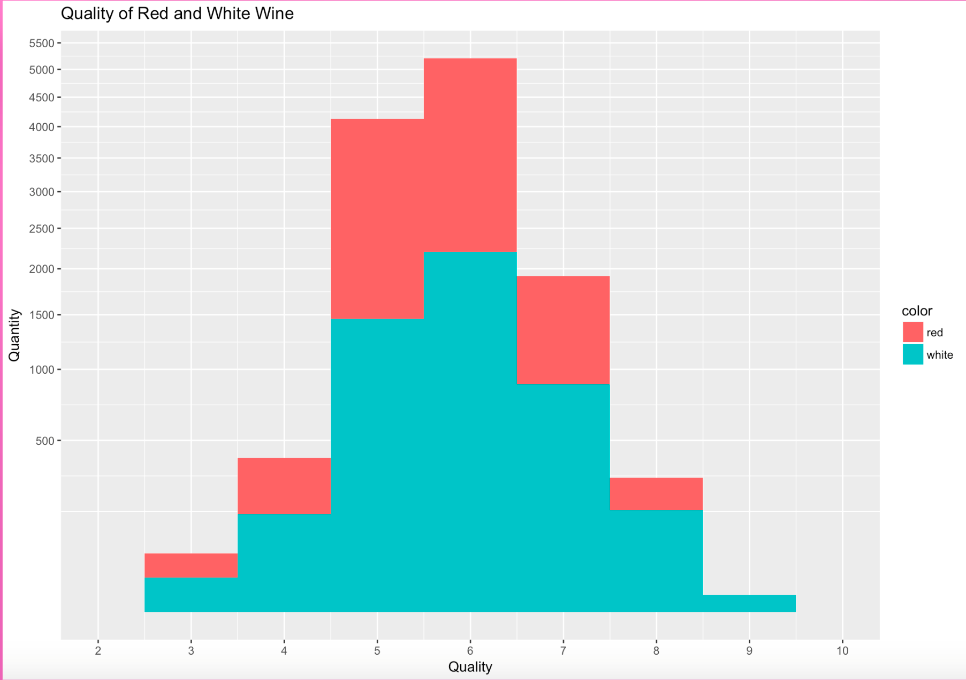
Next, I’d like to show you what our data looks like and our basic analysis of the data.

We drew histogram and correlation graph to analyse the data.

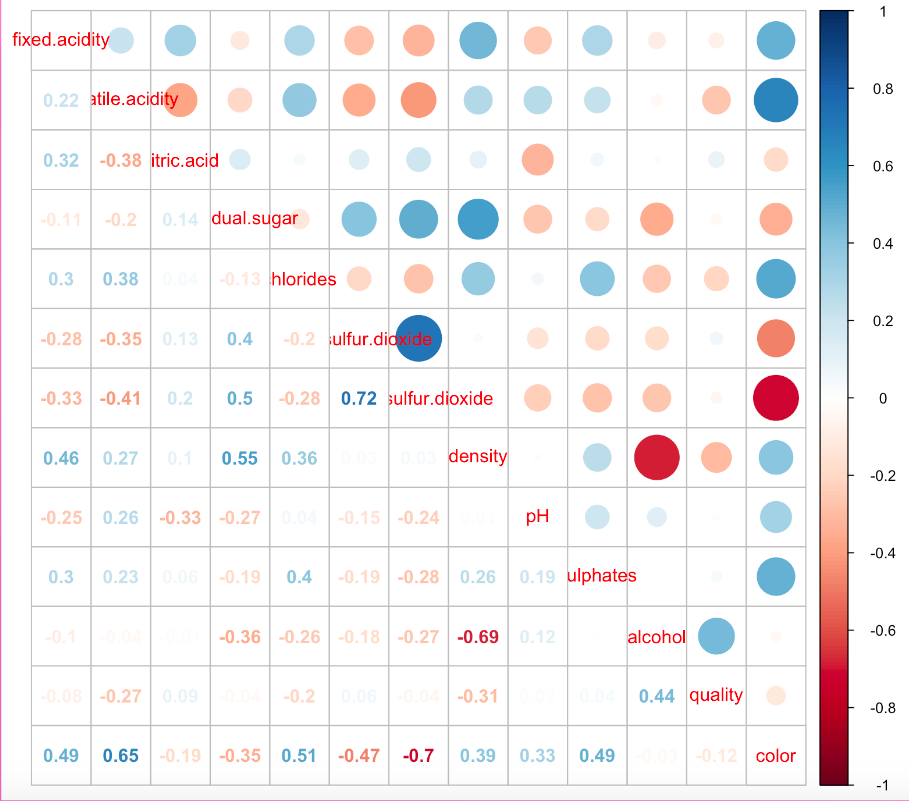
Our data is from UC Irvine Machine Learning Repository. Originally, there are two data sets: red wine and white wine. We combined them and treat the colour of the wine as another independent variable.

Here is what our data looks like. There are twelve independent variables including colour, alcohol, density, fixed acidity, chlorides, and so on. We used these features to predict the quality of the wine. The quality is scored from one to nine. One means worst quality and nine represent the best.

We drew this histogram using red and white data set separately. From the graph, we could see that the distributions of them are basically the same and scores of quality concentrates on 5, 6 and 7.



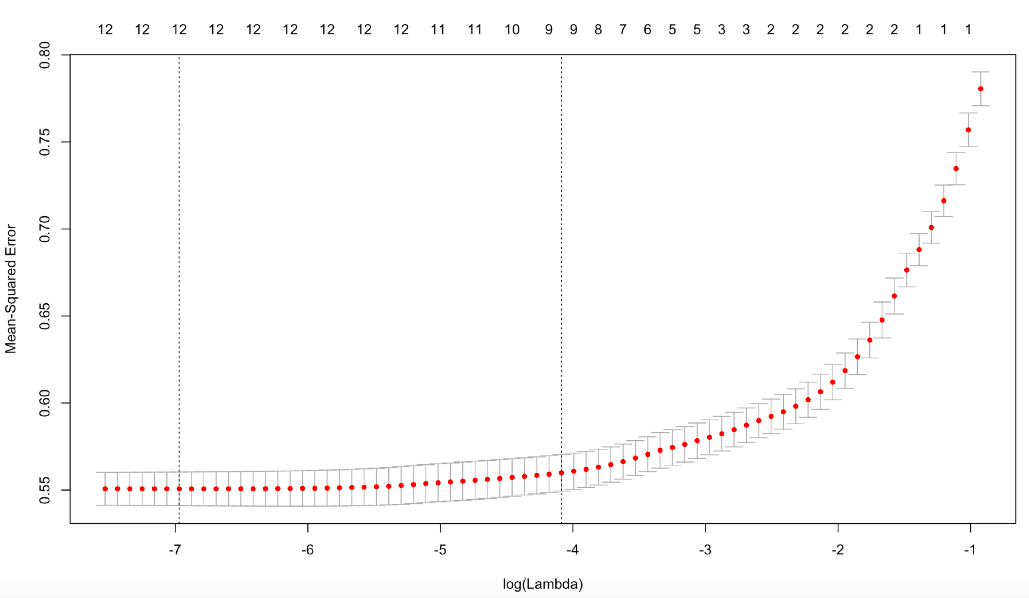
From this correlation plot, we could see that alcohol has a strongly positive relationship with quality. Volatile acidity, chlorides and density tend to relate to quality negatively but still strongly impact it.



Before setting up models, we scale the data, and divide data randomly to two parts. Two thirds of them are treated as training data, and the rest one third are testing data.

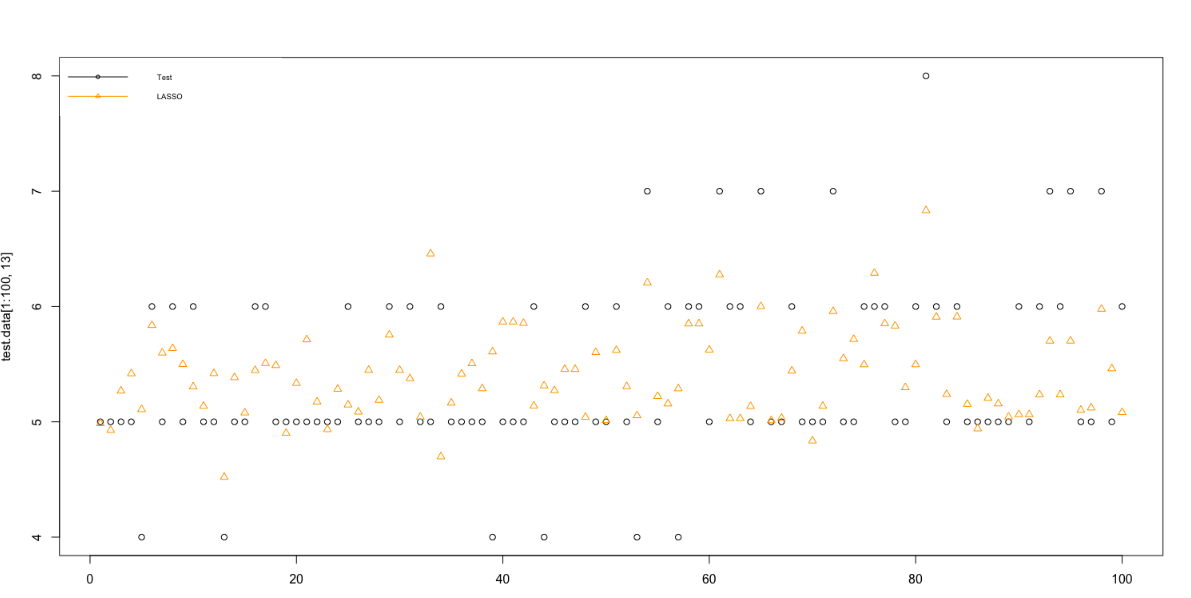
Now, you should have a brief idea of how our data looks like. I am going to introduce methods we used to make models. Firstly, I am going to talk about linear regression.

We used LASSO to select best variables. To decide the optimal lambda for penalty term, cross validation is adopted.

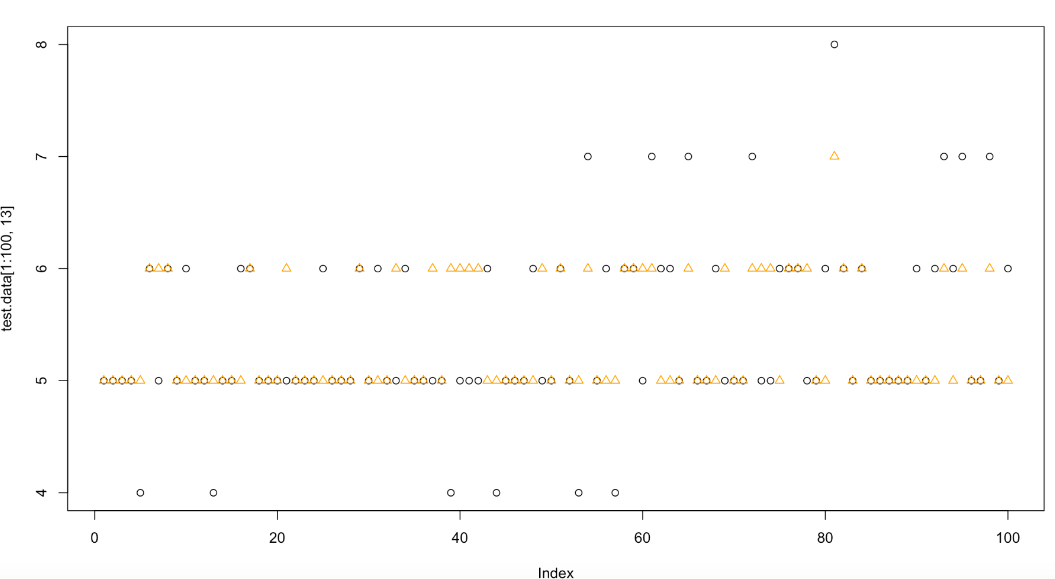


This is the graph drawn while finding the best lambda. The mean square error minimized when log lambda is equal to -6.79. Thus, we choose 0.00094 as lambda.

Here is the result of our perdition. The black circles are testing data, and yellow triangles are our prediction. We could see scores predicted are mainly between 5 and 6. As the mean square error is 0.52, this model actually works well.



Because the quality scores are integers, we also tried the round prediction. We round the prediction to integers. Visually, the result seems better than the later one, and more than half of the prediction points are exactly overlapped with the testing data. However, the mean square error actually increased 0.03 compared to the previous one.

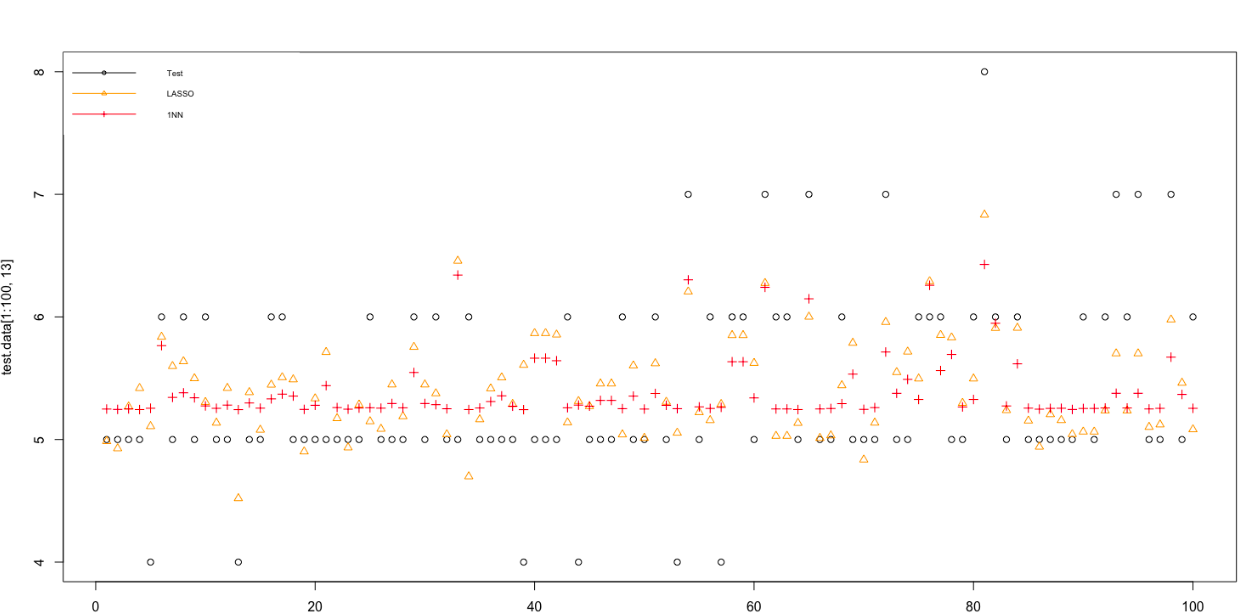


That’s what we did using linear regression. To explore nonlinear relationships, we tried neural network.

Firstly, I want to give you a brief idea of neural network. There is an input layer which in our case are 12 features of wine or say, 12 independent variables. They go through a hidden layer and give an output which is wine quality. In this picture, there is only one hidden layer and you could see there are four nodes in it which are called neurons.

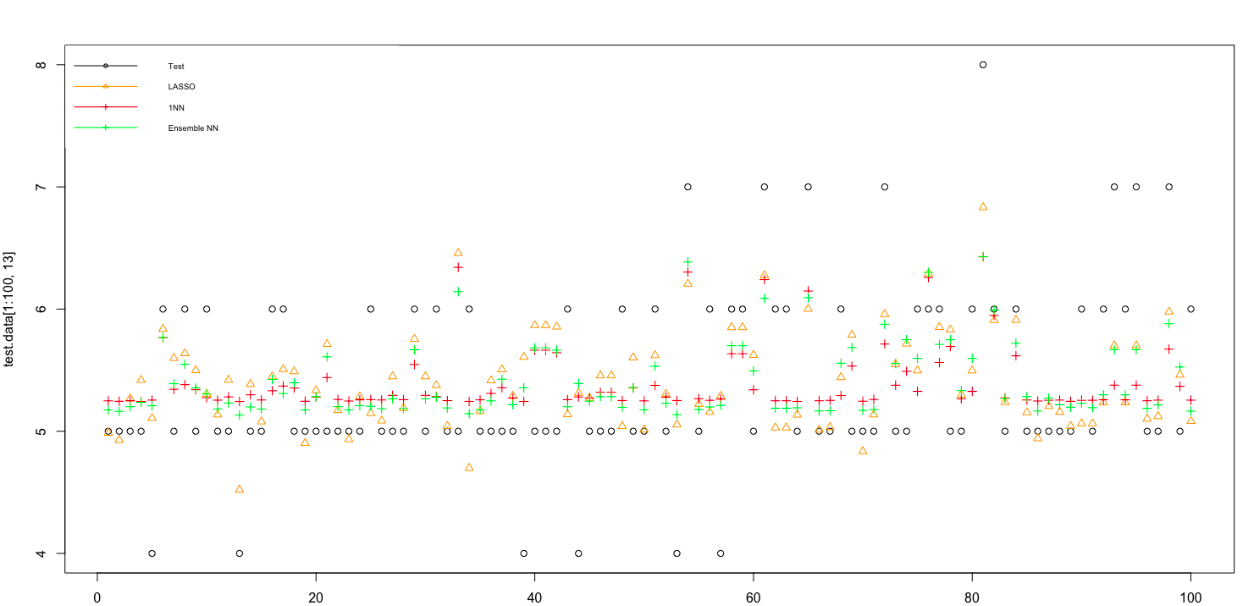
Here is a neural network with two hidden layers and four neurons each.

As mentioned before, firstly, we did a simple one-layer neural network with two neurons in the hidden layer. Red crosses represent our prediction using this method. From the graph, it is actually hard to tell which methods make better prediction as at some points one-layer neural network performs more accurate, but for others, linear regression acts better. Mean square error tells us one-layer neural network does not works as good as linear regression, as this model’s MSE is 0.538 which is nearly 0.02 larger than linear regression’s.



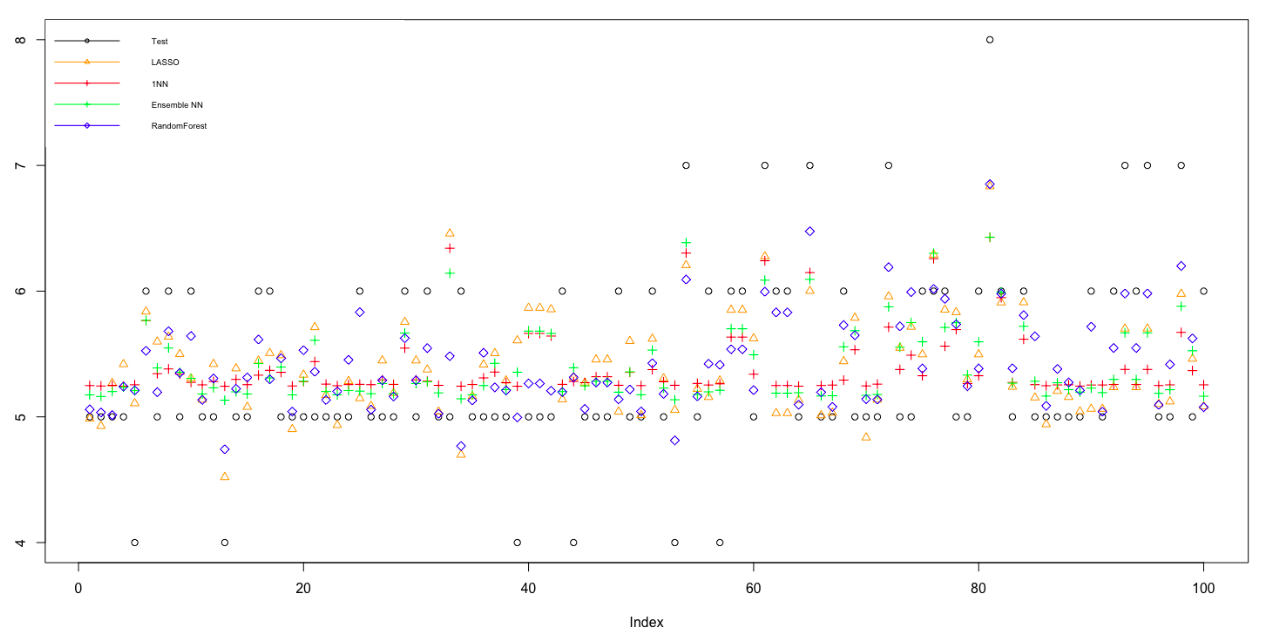
As one-layer neural network does not work well, we tried the ensemble neural network. We did ten models including one layer two neurons, one layer three neurons and up to four layers with four neurons each. Aside from different construction of hidden layers, we ran each model 5 times with different initial parameters in order to avoid local optimum.

Green crosses are our prediction using ensemble neural network. Clearly, compare to the red crosses, result of one-layer neural network, green crosses are closer to the testing data. Overall, it also seems to be better than linear regression. The mean square error does confirm our conclusion from the graph, it is 0.497, smallest MSE we have got.



Last but not least, we want to tried random forest, as it works in a different way from linear regression and neural network. We want to see how it works for this data set.

Blue squares represent the prediction result of random forest. If we look at the testing data one by one, we could see that random forest did better than other methods on majority points. The MSE of it turns out to be only 0.34, which is the smallest among four methods.



In the end, I want to compare the MSE of four methods.One-layer neural network performs worst with biggest MSE of 0.54. Linear regression is only slightly better than it. Ensemble neural network already works well, but random forest performs best with the smallest MSE.