Cooking Style Project Report

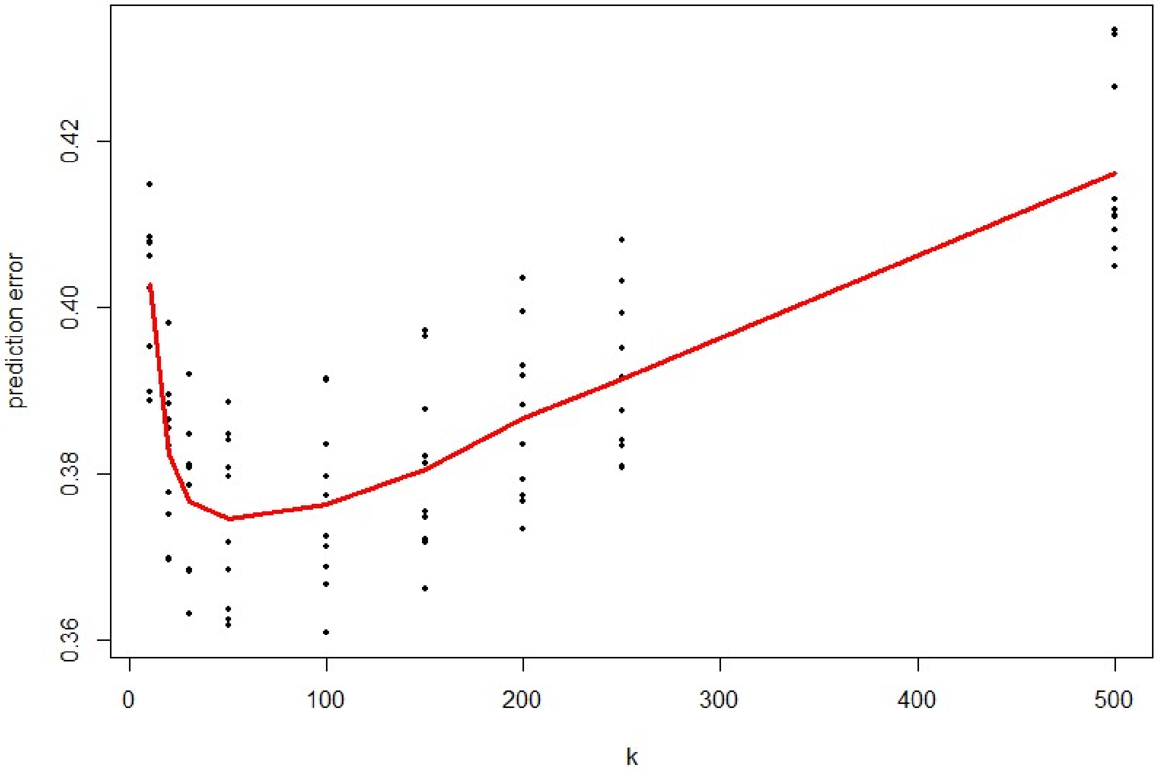
Diwen Li

This project needs to predict the cooking type by what is in it. It contains 39,774 training data with 442 independent variables. Each variable has only two level, which is 1 -> yes and 0->no, means is this food contain in the cooking.

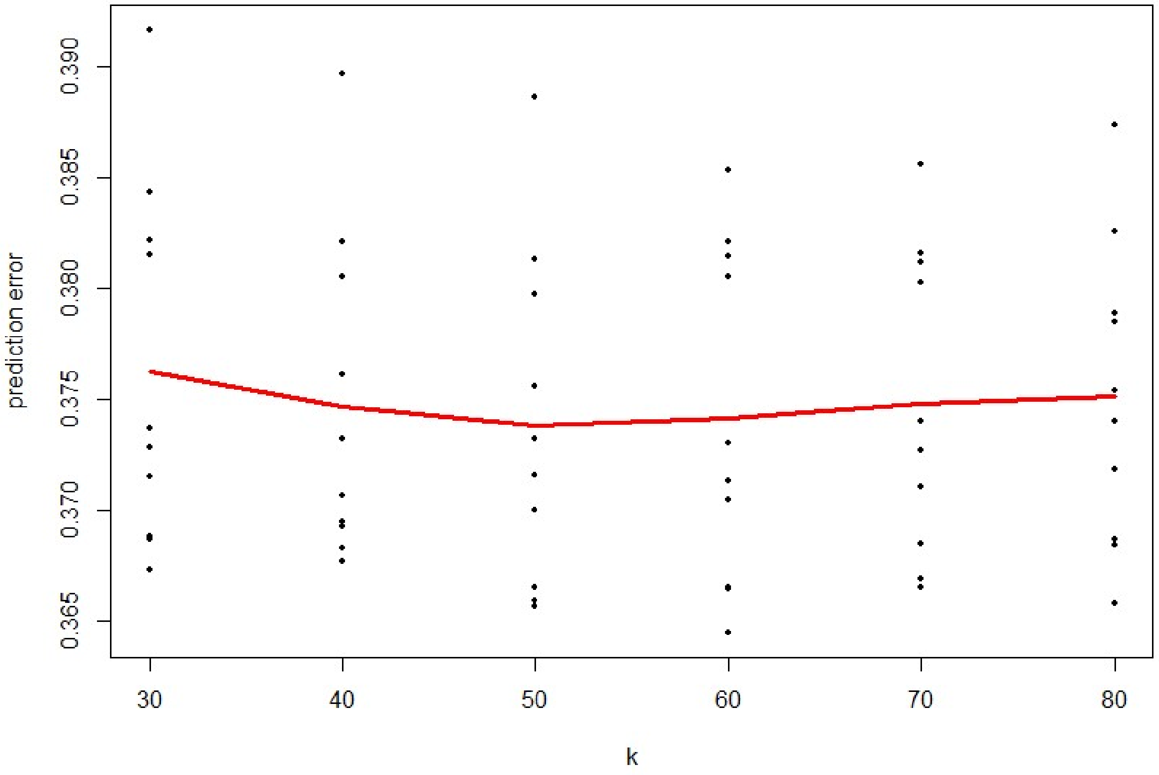
1. Knn Model

Considering it's a classification mission, I pick Knn algorithm as my first model because it's the simplest model.

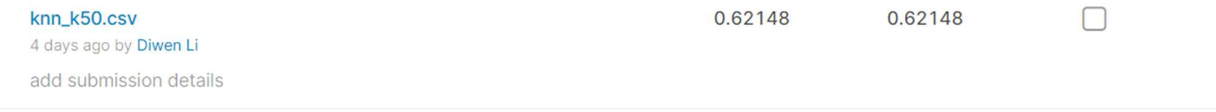
I use cv-10 to find the best parameter. At first, I though with that large data, I may need some large k, so I test 0~500, with step = 50. Result shows below



It shows the prediction res min around 30 to 80 somewhere. So I make try again with range 30~80 with step 10. Result shows below.



It shows the residual is close for what ever k choice in this range. So I pick the lowest one k=50 as my Knn Model parameter.



knn_k50.csv 0.62148 0.62148 □

4 days ago by Diwen Li

add submission details

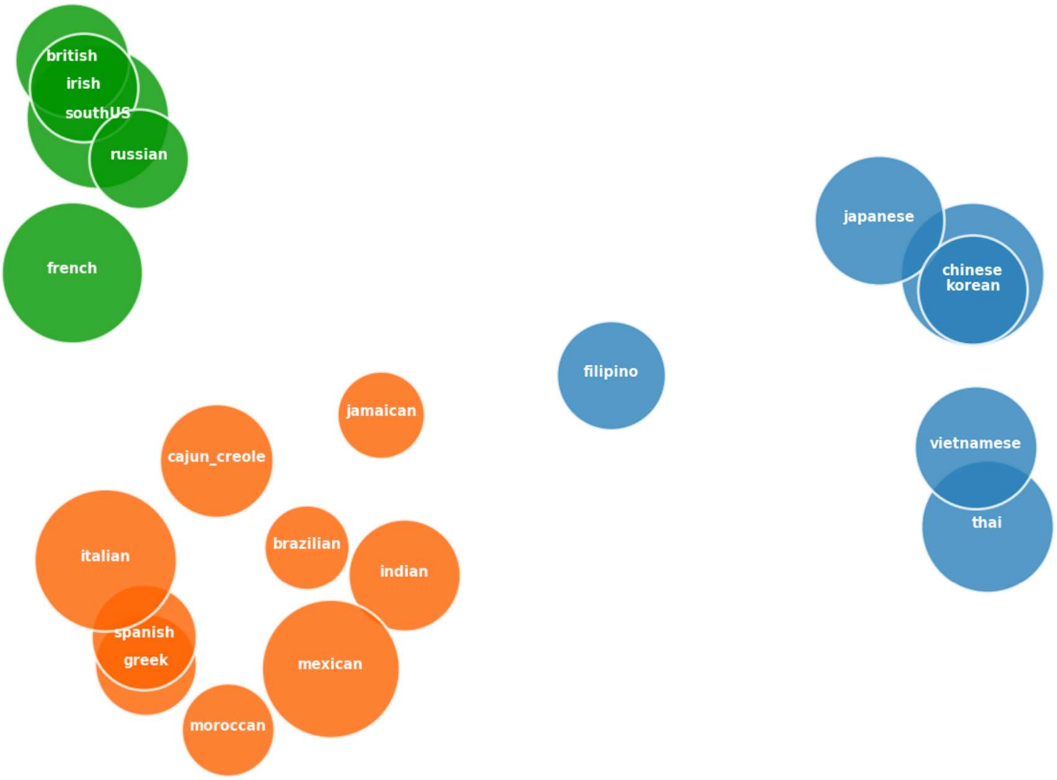
Considering it has 0.373 prediction err in cv10 training data set. I don't surprise it don't have a good score.

Conclusion:

Knn Model is not a good algorithm for this data. It has some reason.

For this data has 400+ variables, which means it plot the dataset into a 400+ degree space, and find the 50 closest data point to decide what type it is. But some data will mess up in a small space while other not, so if using this algorithm, I think i need to set up different k for different type. Like if typeA's and typeB's data point are close to each other, then when I find the test point in this area, I may need to pick a smaller k. And if a test point is quite far from other points, it may be hard to find what type is it.

And I found a data picture that also prove my though.



I found this from a Notebook on Kaggle (https://www.kaggle.com/alonalevy/cultural-diffusion-by-recipes) it shows some type of cooking are so close to others. For example, Chinese and Korean are totally repeated. Using Knn algorithm can't classify them clearely.

2.NNet Model

So I give up testing knn model, and I try to find a model with some logical or tree like so that it should be a better model for this data.

I pick NNet Model as my second model (Artificial Neural Network).

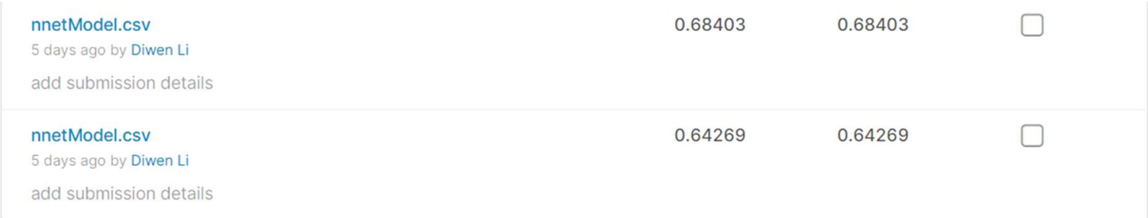
NNet model need size, rang, decay, maxit parameter, but it takes a long time to run the model, so its kind of impossible to use cv10 to find the best parameter.

I search online to find is there some hint for using this model. I found most situations, parameter like rang, decay just use its default parameter is good enough. So I only need to deal with size.

Still using cv to find best parameter, but this time I only set cv5. But sadly, I wait for whole day and still cant run it all. So I stop.

I decide to just set .1 of the the dataset as the test dataset and just calculate different size's res.

The errMatrix shows when size increases , the res will get lower, my computer takes 6hour to run size 10, so I finally decide us size 10 as my final parameter.



nnetModel.csv 0.68403 0.68403 □
5 days ago by Diwen Li
add submission details

nnetModel.csv 0.64269 0.64269 □
5 days ago by Diwen Li
add submission details

(the first submission is using default parameter, and the second is using size = 10)

Conclusion:

NNet should be quite a good model, but I can't do much of it. I'm not familiar with how it works. All I can do is try to find a better parameter that fits the model better. And it really require a large computation so I have to stop.

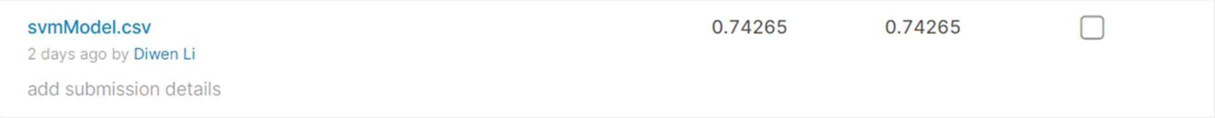
But obviously, this model works much better than Knn, it has higher score than KNN when using the default parameter. And when I pick a better size parameter, it has a high score with .68.

So I think if I can understand how this works and find a better parameter, it can have a much better score.

3.SVM Model

SVM should be the best model because of how it works.

The default parameter already gave a .74 score. Which is really high.



svmModel.csv 2 days ago by Diwen Li add submission details	0.74265	0.74265	□

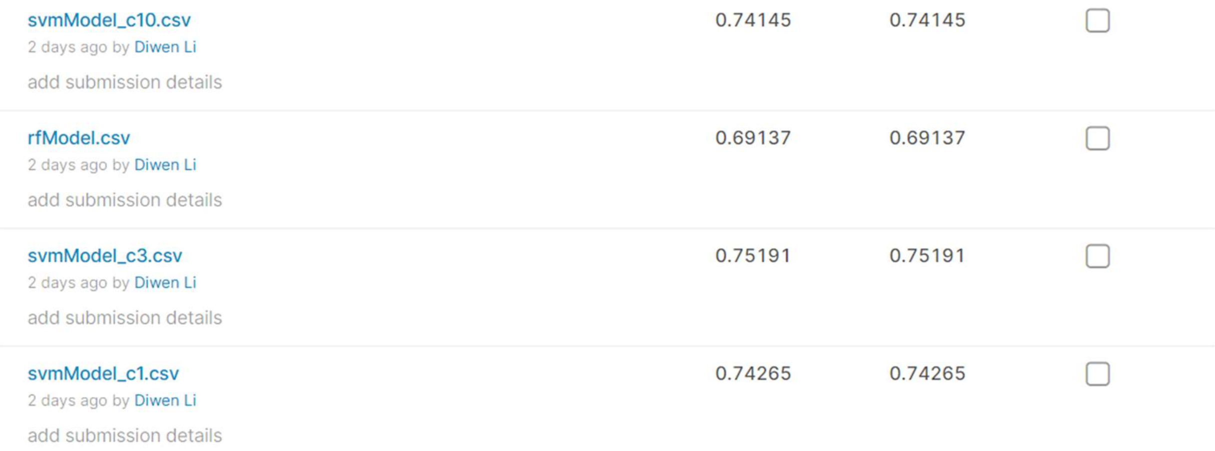
But in this algorithm, there’s only one parameter I can change when chosing “C-classification” kernal, that is c.

And same as nnet model, this algorithm takes a long time to run, so I cant use cv to find the best c.

So I still randomly pick .1 of the train dataset as test dataset, and calculate its res.

I pick c in range(1~5) with step 1.

The errMat shows when c = 3, it has the lowest res. Which is alse same as my test via kaggle

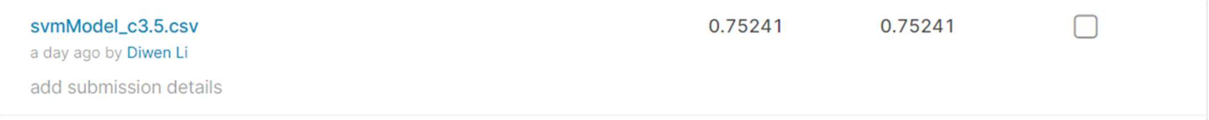


svmModel_c10.csv 2 days ago by Diwen Li add submission details	0.74145	0.74145	□
rfModel.csv 2 days ago by Diwen Li add submission details	0.69137	0.69137	□
svmModel_c3.csv 2 days ago by Diwen Li add submission details	0.75191	0.75191	□
svmModel_c1.csv 2 days ago by Diwen Li add submission details	0.74265	0.74265	□

(ignore the rfModel in middle)

So it should have the lowest res in some where around 3.

So my next range is 2.5~3.5 step .2 and the result shows c=3.5 has the lowest res.



svmModel_c3.5.csv a day ago by Diwen Li add submission details	0.75241	0.75241	□

Conclusion:

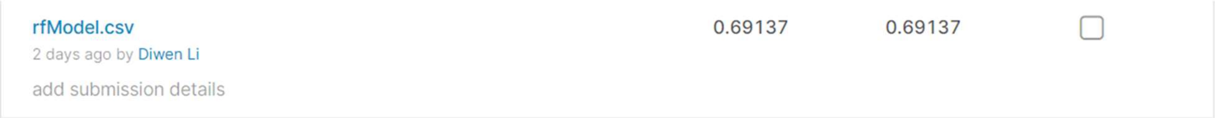
SVM should be the best model. But still, I don’t know how its kernel works and I cant improve my model more because my computer needs couple days to calculate which takes really a long time.

4.RandomForest Model

Like I though, using tree algorithm should works good. Like if it has bulgogi (some food that only shows in Korean), then it must be Korean dish.

Other algorithm focus in space distance or data only(e.g. Knn), but some classification need logical like the example above. So I pick RF Model as my final model.

My first try with random pick of parameter shows this algorithm is good as I think.



rfModel.csv 0.69137 0.69137 ☐

2 days ago by Diwen Li

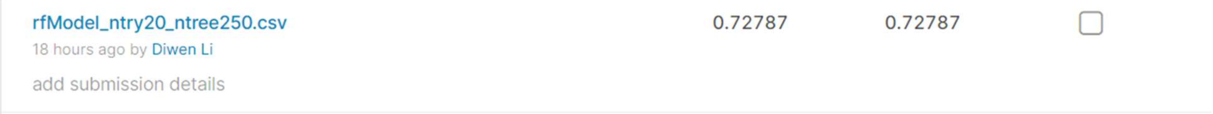
add submission details

And the time for runing the program is also much less then Model2 and 3. But still needs hours to complete each Model.

I decide not to try CV because I have 2 parameter to decide now, and if using CV, it may takes me weeks of time to find the best model.

So I just using same way as other model. Just randomly pick .1 as my testing dataset.

And I found the best parameter ntry = 5, ntree = 250.



rfModel_ntry20_ntree250.csv 0.72787 0.72787 ☐

18 hours ago by Diwen Li

add submission details

Conclusion:

RF model should be a good model too, but the best parameter I can find still not good enough. Compare to default parameter in SVM model, it still need to improve more.