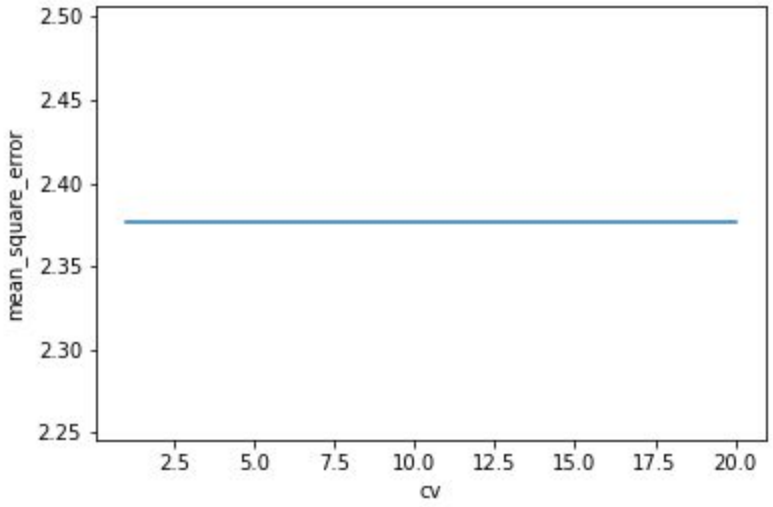
Covid-drug-discovery-for-covid19

Rdkit with Ridge regression

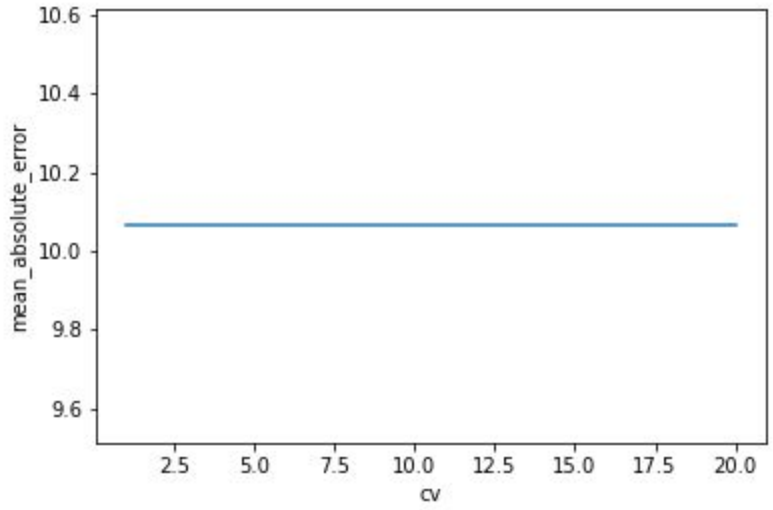
Reason : - ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.

Observations:-

Cv vs mse:-
Here cv is cross-validation. Checking whether it effects the training or not



Cv vs mae:-
Here as we can see that it does not making any difference in MSE and MAE



Different Alphas : -
Its is not giving any better mse and mae on any choice of alphas.

Using this method, I got MAE SCORE = 2.3763
MSE SCORE = 10.0644

SO this method is giving Comparatively large amount of mse and mae so I tried other methods.

Mol2vec embedding with ridge regression

For the Mol2vec method I used the embedding that was given in the resource itself.
Embedding data file used = `model_300dim_kaggle.pkl`

Observations:-

Cv vs mse, Cv vs mae:-

In this method changing the cv does not result in changing the mse and mae.
MAE score: 1.7649
MSE score: 5.6485

This method is giving way better performance with respect to using rdkit with ridge.regression Because here we are using pre trained data . and was also observed that choosing the right pretrained data is very important. When I used the Different pre trained data With this method I was getting bad mse and mae.
It was mae = 1.9437
Mae = 7.6347
So choosing right pretrained data is very important for this process.

ridkit with mol2vec embedding with ridge regression

Observations:-

Cv vs mse:-

Again mse with respect to different cv was constant.

Cv vs mae:-
Again mae with respect to different cv was constant.

MSE SCORE : 5.4649

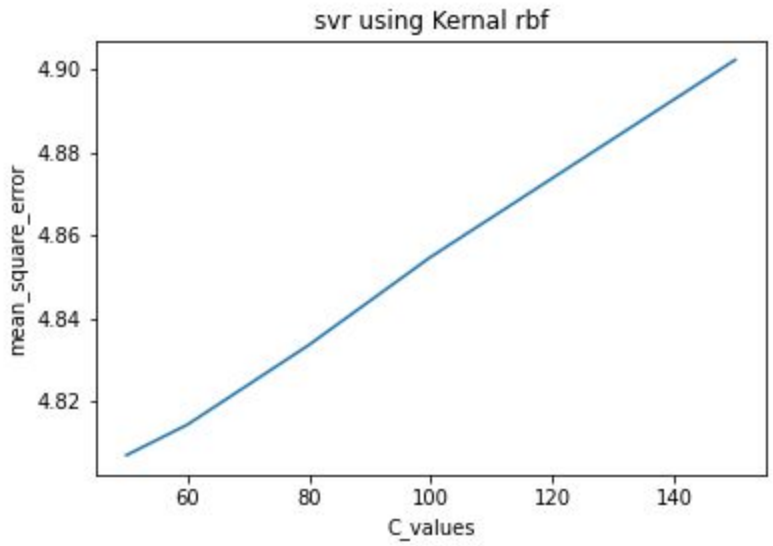
MAE SCORE : 1.6649

Here we can see the because of using rdkit and embedding and concatenate the features to the trained data with rdkit and perform the regression using then i got improved mse and mae with respect to the

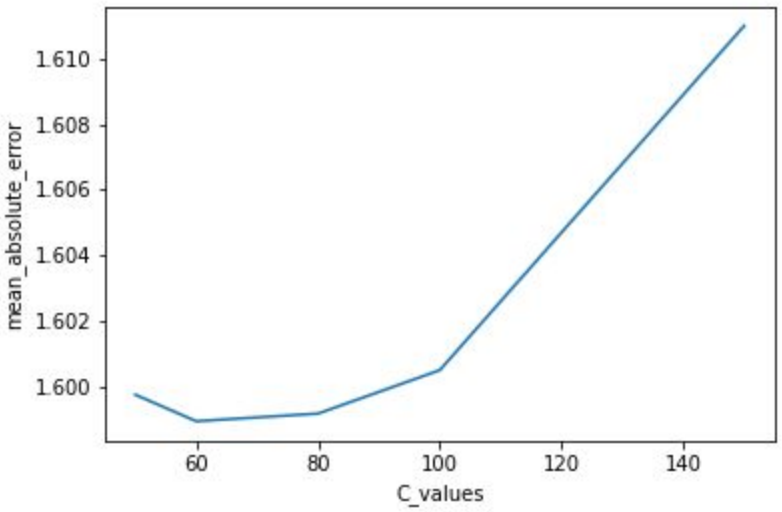
SVR regression(Best method)

Observation:

hyper parameter C vs mse: - using kernel : - rbf



hyper parameter C vs mae : - using kernel :-rbf



Conclusion :-

SVR with Molvec embedding gave me the best mse and mae when tried different c coefficients.
SVR MSE : 4.8071
SVR MAE : 1.5997
How to run the svr code : - hardcoded the path of the train and test files just change that only.
