## Subjective Questions – Advanced Regression

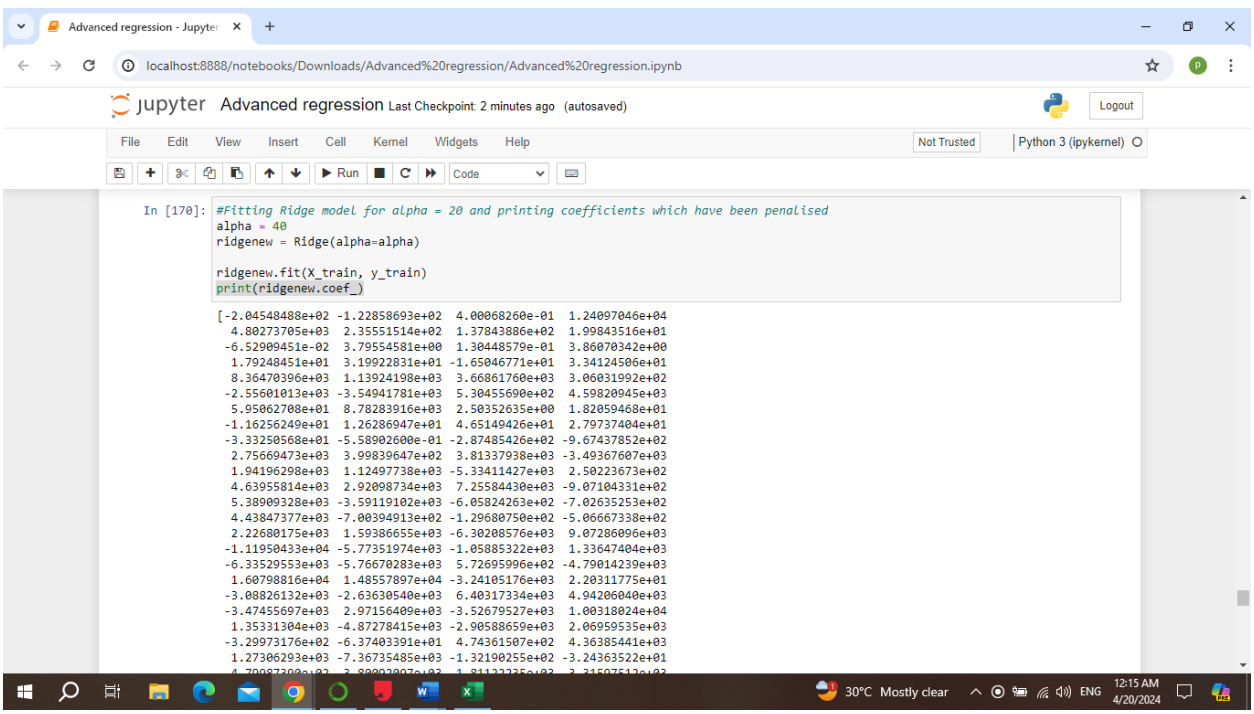
## Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

The optimal value of alpha are,

Ridge {'alpha': 20}
Alpha {'alpha': 20}

After choosing double the value of alpha for ridge and lasso, r2score values are dropped in both t est and train data



Advanced regression - Jupyter

localhost:8888/notebooks/Downloads/Advanced%20regression/Advanced%20regression.ipynb

jupyter Advanced regression Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

```
r2_test_lr = r2_score(y_test, y_pred_test)
print(r2_test_lr)
metric2.append(r2_test_lr)

rss1_lr = np.sum(np.square(y_train - y_pred_train))
print(rss1_lr)
metric2.append(rss1_lr)

rss2_lr = np.sum(np.square(y_test - y_pred_test))
print(rss2_lr)
metric2.append(rss2_lr)

mse_train_lr = mean_squared_error(y_train, y_pred_train)
print(mse_train_lr)
metric2.append(mse_train_lr**0.5)

mse_test_lr = mean_squared_error(y_test, y_pred_test)
print(mse_test_lr)
metric2.append(mse_test_lr**0.5)
```

```
0.8724130948333536
0.8633499180060508
814093612625.8152
385177298389.7433
797349277.7921795
879400224.6341171
```

In [173]:
```
#Fitting Ridge model for alpha = 20 and printing coefficients which have been penalised

alpha = 40
```

30°C Mostly clear ENG 12:15 AM 4/20/2024

Advanced regression - Jupyter

localhost:8888/notebooks/Downloads/Advanced%20regression/Advanced%20regression.ipynb

jupyter Advanced regression Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

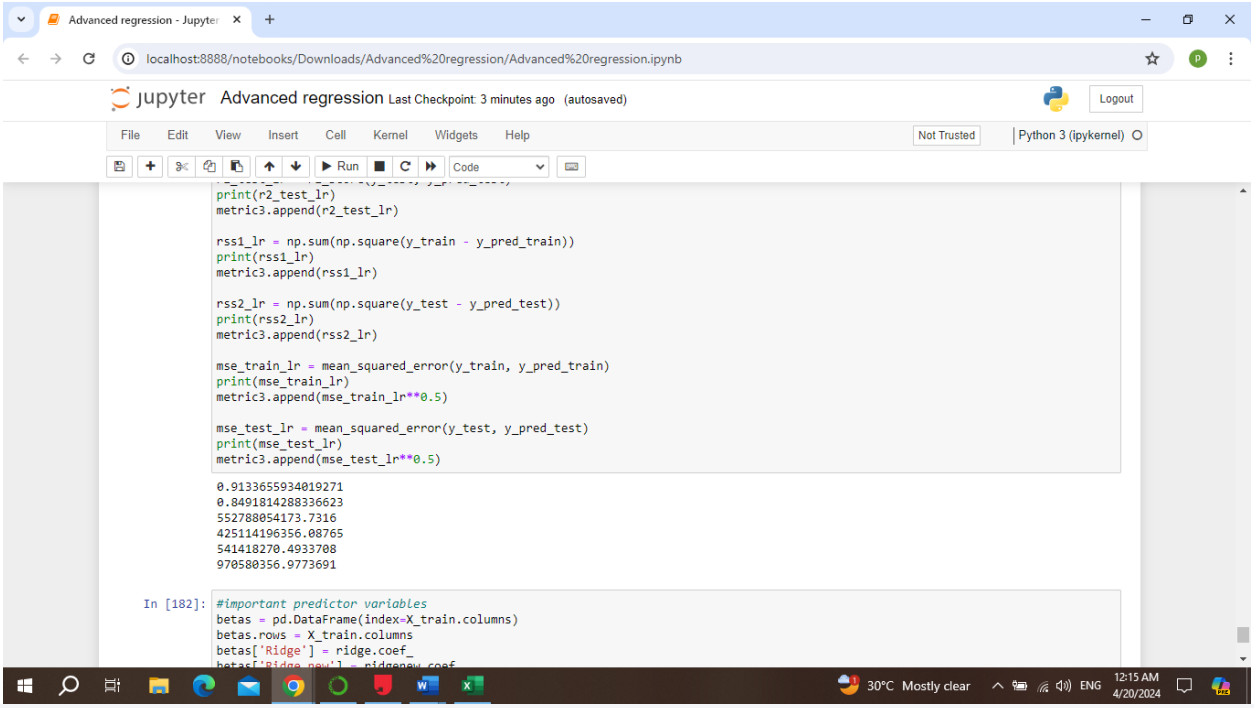
Not Trusted | Python 3 (ipykernel)

```
797349277.7921795
879400224.6341171
```

In [173]:
```
#Fitting Ridge model for alpha = 20 and printing coefficients which have been penalised

alpha = 40

lassonew = Lasso(alpha=alpha)

lassonew.fit(X_train, y_train)
print(lassonew.coef_)
```

```
[-1.39444508e+02 -8.62318804e+01  4.60551133e-01  8.94436498e+03
  4.84175280e+03  1.74597804e+02  1.15612288e+02  1.29803608e+01
  1.11785604e+00  6.55988576e+00  1.70581368e+00  1.12567349e+00
  4.55758888e+01  6.55649632e+01  1.01015682e+01  7.88976806e+00
  8.33778447e+03  1.84960658e+03  5.65309479e+03 -7.27853191e+02
 -2.22388789e+03 -1.26999968e+04 -0.00000000e+00  2.24190389e+03
  4.31128001e+01  1.22103803e+04 -1.69091308e+01  9.88752685e+00
  6.50488965e+00  7.60446296e+00  4.07380510e+01  2.10109657e+01
 -1.33553453e+01 -3.77439432e-01 -1.90133963e+02 -6.65936104e+02
  0.00000000e+00  2.04833477e+03  5.44757159e+03  0.00000000e+00
  1.45991952e+04  0.00000000e+00 -2.33065902e+04  9.22050216e+02
  1.09212175e+04  9.78988234e+03  1.17896559e+04 -0.00000000e+00
  1.00259449e+04 -9.69311954e+03 -0.00000000e+00 -4.28834917e+02
  1.38352166e+03  0.00000000e+00  0.00000000e+00  0.00000000e+00
  2.35600873e+02  2.91359588e+03 -0.00000000e+00  2.30354199e+04
 -1.38981287e+04 -1.40978194e+03 -3.34517109e+03 -0.00000000e+00
 -1.29678623e+04 -5.92024740e+03  0.00000000e+00 -4.14971097e+03
  4.39719272e+04  2.89349903e+04 -5.53841727e+03  2.07307761e+02
 -3.92683807e+03  1.36388777e+03  2.31002540e+04  2.66301056e+04
```

30°C Mostly clear ENG 12:15 AM 4/20/2024




Most important predictor variables after the change is implemented are,

| | Lasso new |
|---|---|
| **Neighborhood_Crawfor** | 23035.41995 |
| **Street_Pave** | 14599.19525 |
| **GarageCars** | 12210.38031 |
| **LandContour_Lvl** | 11789.65589 |
| **LandContour_HLS** | 10921.21748 |

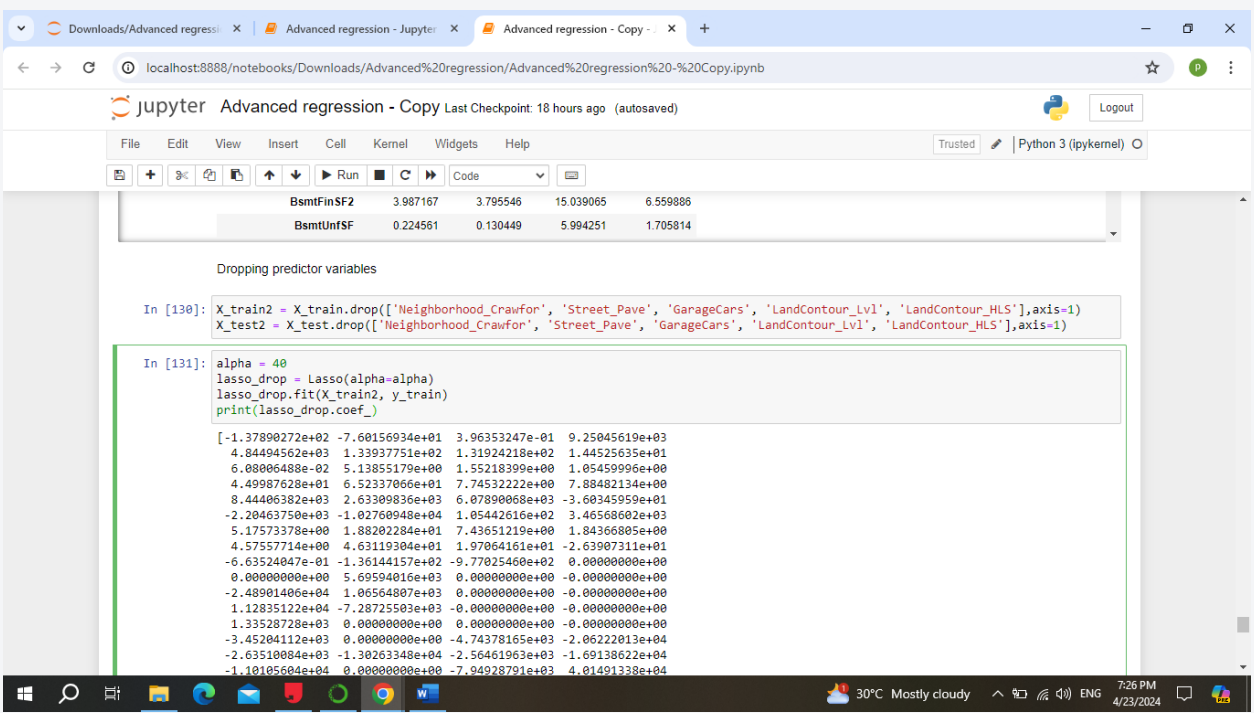
## Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

The r2_train value of lasso increases whereas the r2_test data of lasso decreases. So, we choose r2_train value.

## Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?



```
#important predictor variables
betas = pd.DataFrame(index=X_train2.columns)
betas.rows = X_train2.columns
betas['Lasso_drop'] = lasso_drop.coef_
pd.set_option('display.max_rows', None)
betas.sort_values(by=['Lasso_drop'], ascending=False)
```

| | Lasso_drop |
|---|---|
| RoofMatl_WdShngl | 135314.884998 |
| RoofMatl_CompShg | 71679.918309 |
| RoofMatl_WdShake | 47642.114719 |
| Neighborhood_NoRidge | 40149.133817 |
| RoofMatl_Tar&Grv | 35804.161375 |

## Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

To ensure the robustness and generalizability of predictive models, we must make sure that the quality of the data is clean. Data must be validated to remove errors, outliers, missing values, and biases. Then using techniques like normalization, scaling, aggregation we refine the variables. Based on the nature and complexity of data available, we have to choose the appropriate model. Also, ensure that overfitting and underfitting is taken care. Regularization techniques are used to reduce the complexity and variance of the model.

If the model is not robust, it cannot be trusted for future analysis.