Subjective Questions

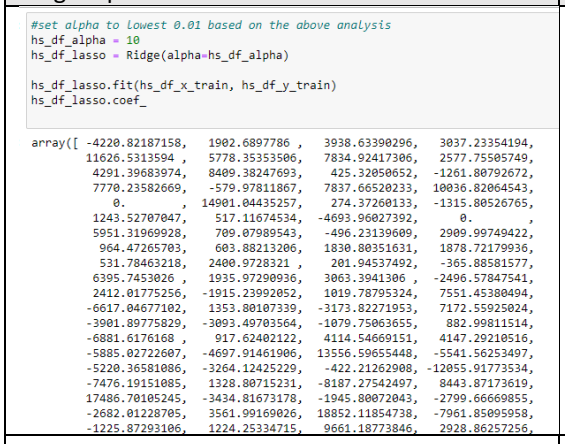
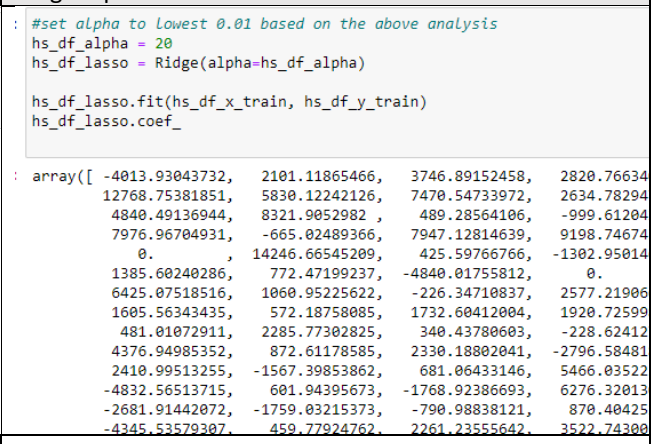
1. **What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

Answer:
The optimal value of alpha for ridge and lasso regression as below:

- **Ridge Alpha**

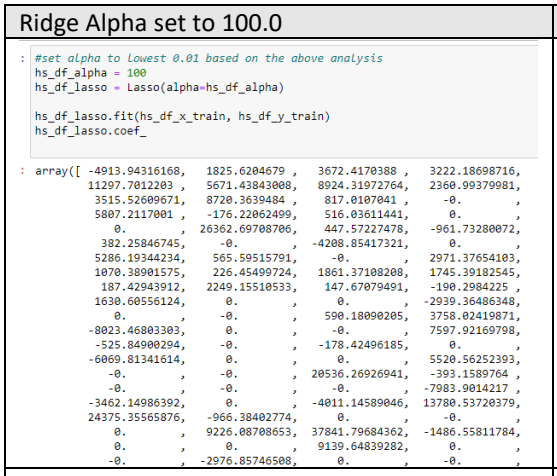
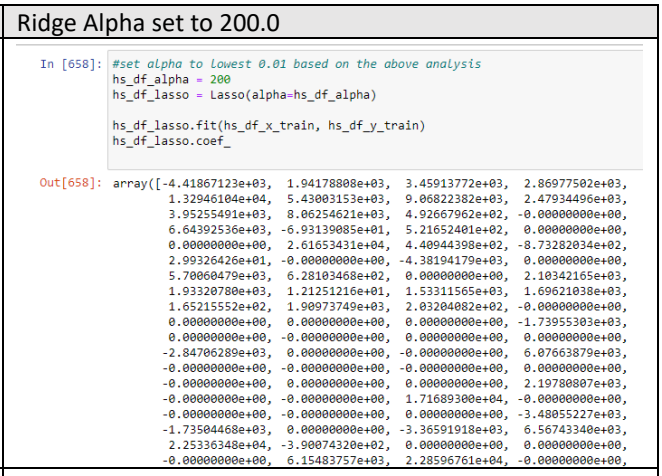
Best optimum • Ridge Alpha 10.0

When Ridge Alpha value is doubles (20) then coefficient **does not** vary much.

Ridge Alpha set to 10.0	Ridge Alpha set to 20.0
	

- **lasso Alpha**

Best optimum lasso Alpha 100.0
When Lasso Alpha value is doubles (200) then coefficient values **drastically** changed.

Ridge Alpha set to 100.0	Ridge Alpha set to 200.0
	

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?

Answer:

r2_score computed using Lasso regression is lower (i.e. 0.8610351617897898) compare to Ridge regression (i.e. 0.9289551150785119) hence I would choose Lasso regression.

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?

Answer:

```
hs_df_train_cols=hs_df_x_train.columns[hs_df_rfe.support_]


hs_df_x_train2=hs_df_x_train[hs_df_train_cols]
hs_df_x_train_cons = smapi.add_constant(hs_df_x_train2)

#create first liner regression model
hs_df_lr_mdl =smapi.OLS(hs_df_y_train,hs_df_x_train_cons)

#fit the model
hs_df_lr_mdl=hs_df_lr_mdl.fit()

#Print the summary
hs_df_lr_mdl.summary()
```

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.797
Model:	OLS	**Adj. R-squared:**	0.793
Method:	Least Squares	**F-statistic:**	215.2
Date:	Tue, 14 Jun 2022	**Prob (F-statistic):**	4.64e-290
Time:	03:24:29	**Log-Likelihood:**	-10582.
No. Observations:	893	**AIC:**	2.120e+04
Df Residuals:	876	**BIC:**	2.128e+04
Df Model:	16		
Covariance Type:	nonrobust		

Columns highlighted in the below table shall be excluded.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.911e+05	1.1e+04	17.332	0.000	1.69e+05	2.13e+05
BsmtFinSF1	1.431e+04	985.475	14.519	0.000	1.24e+04	1.62e+04
BsmtFinSF2	-823.6565	1138.967	-0.723	0.470	-3059.080	1411.767
BsmtUnfSF	1875.0789	984.265	1.905	0.057	-56.714	3806.872
TotalBsmtSF	1.696e+04	1561.980	10.859	0.000	1.39e+04	2e+04
1stFlrSF	1.264e+04	1637.211	7.721	0.000	9428.338	1.59e+04
2ndFlrSF	1.384e+04	1019.403	13.576	0.000	1.18e+04	1.58e+04
GrLivArea	2.196e+04	897.665	24.467	0.000	2.02e+04	2.37e+04
RoofStyle_Gable	2.285e+04	7257.858	3.148	0.002	8603.689	3.71e+04
RoofStyle_Gambrel	2.316e+04	1.34e+04	1.726	0.085	-3172.183	4.95e+04
RoofStyle_Hip	3.065e+04	7512.534	4.080	0.000	1.59e+04	4.54e+04
RoofStyle_Mansard	-2556.5676	1.67e+04	-0.153	0.878	-3.53e+04	3.02e+04
RoofStyle_Shed	3.767e+04	3.06e+04	1.232	0.218	-2.23e+04	9.77e+04
RoofMatl_Metal	6.885e+04	3.06e+04	2.253	0.024	8876.114	1.29e+05
RoofMatl_Tar&Grv	1.047e+04	1.65e+04	0.633	0.527	-2.2e+04	4.29e+04
Exterior1st_CBlock	-9727.4452	1.88e+04	-0.518	0.604	-4.66e+04	2.71e+04
Exterior2nd_CBlock	-9727.4452	1.88e+04	-0.518	0.604	-4.66e+04	2.71e+04
ExterQual_Fa	-5.718e+04	1.57e+04	-3.646	0.000	-8.8e+04	-2.64e+04
ExterQual_TA	-4.084e+04	2859.455	-14.281	0.000	-4.64e+04	-3.52e+04
GarageQual_Fa	-1.758e+04	1.19e+04	-1.473	0.141	-4.1e+04	5850.066
GarageQual_TA	-8986.9745	1.01e+04	-0.893	0.372	-2.87e+04	1.08e+04

Omnibus:	238.369	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1777.064
Skew:	1.005	Prob(JB):	0.00
Kurtosis:	9.612	Cond. No.	2.45e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.58e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

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?

Answer:

Simple models with test accuracy is not lesser than the training score is the best models. Null values, outliers shall be examined properly as they impact the assessment. Unwanted null values and outlier records shall be remoted or imputed. Data Dictionary shall be closely followed to fill null values.

If required, Bias-Variance trade-off shall be used.