Homework 3
Nikhil Ambati

Below are the previous results which I have got by performing PCR on the same dataset.

Model	**RSQ**
Principal Component Regression	0.5928

On the same data I have performed support vector regression, lasso, and ridge regressions. Below are the results.

RMSE Train & Test results:

Model	**RMSE Train**	**RMSE Test**	**RSQ**
Support vector regression	0.6408	0.8745323	0.6274
LASSO	6.7485	9.8548	0.3984
Ridge regressions	6.5875	10.2589	0.3714

1. As per the observation I feel these models performed **much better** in the terms of testing accuracy. In addition to this I believe these models can handle outliers in a better way.

These models are better in handling **redundant** features. PCR is not that efficient in terms of handling redundant data. Also, PCR retains only the primary contributing principal components, Lasso, Ridge, and SVR offer the flexibility of **adjusting and fine-tuning** the hyperparameters to achieve a better fit to the data. Hence this will **avoid overfitting** on data.

Also, ridge helps to mitigate overfitting and improve performance. This is done by adding **penalty terms** to the loss function.

2. I have re-trained the SVR model by changing the test and train data split. Below are the results.

I basically split the data in different quantity (Train-test) starting from **95% to 50%** and observed the below values.

SVR Train Data	**SVR Test Data**	**RMSE**
95	5	1.399217
90	10	1.331156
85	15	0.7061562
80	20	0.8745323
75	25	0.9992581
70	30	0.8517461
65	35	1.06493
60	40	1.004373

55	45	1.094843
50	50	1.16755
45	55	1.094843

From the above table I could see that at some **split size** (85-15) I can observe there is change in test accuracy of 15-17% across the SVR regression model. Ideally the split should be 80:20 or 70:30.

I believe randomly dividing the dataset is very crucial to prevent biases. Also, the sensitivity depends on the size & quality of data. In our case the size of the data is not that huge, so random split is **highly sensitive**. This we can observe for the above table.

For SVR random split is very crucial, random split can impact the selection of optimal hyperparameters and if we didn't split the data randomly the model may perform well in the train data but not on test data. In addition to this in these models' sensitivity is more important than overfitting because the goal of regression is to predict the target variable accurately.

Also, when the **data size** is increased the performance of these regression models have increased. This happened due to increase amount of data or information available for regularizing the model. I observed this factor by taking only half of the data first and my testing accuracy reduced significantly. Also sometime the testing accuracy may depend on the data characteristics and its important to study with different data size and evaluate the performance.