Prajwal kumar chinthoju (pkc3)

IE598 MLF F18

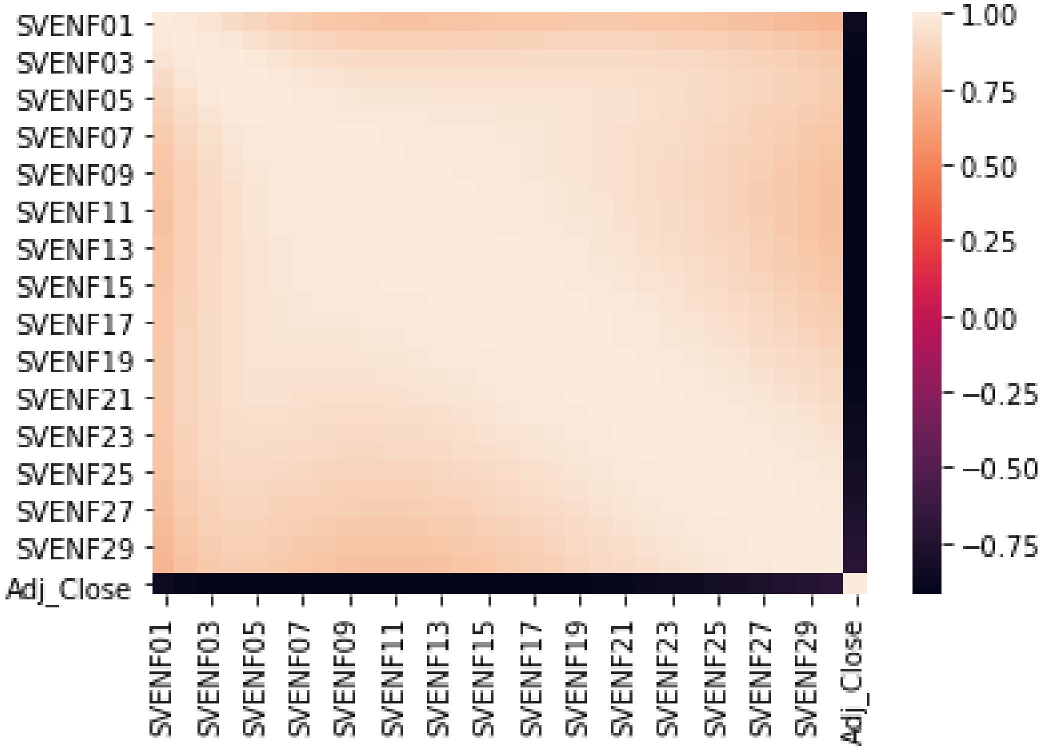
Module 5 Homework (Dimensionality Reduction)

Use the Treasury Yield Curve dataset

Part 1: Exploratory Data Analysis

Describe the data set sufficiently using the methods and visualizations that we used previously. Include any output, graphs, tables, that you think is necessary to represent the data. Label your figures and axes. DO NOT INCLUDE CODE, only output figures!

Heat map correlation:

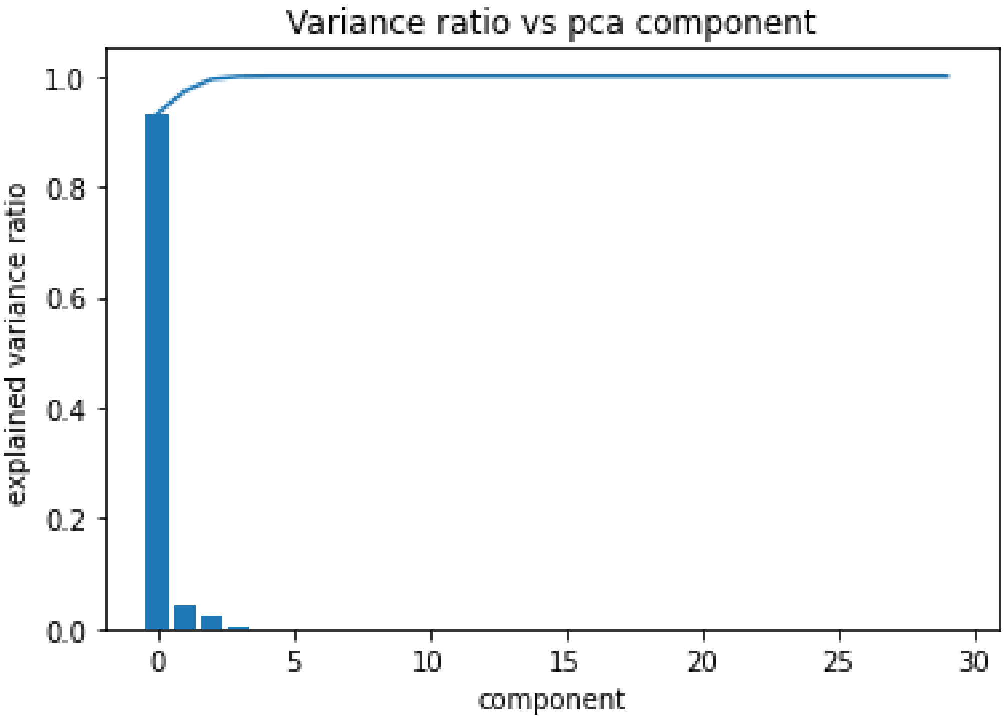


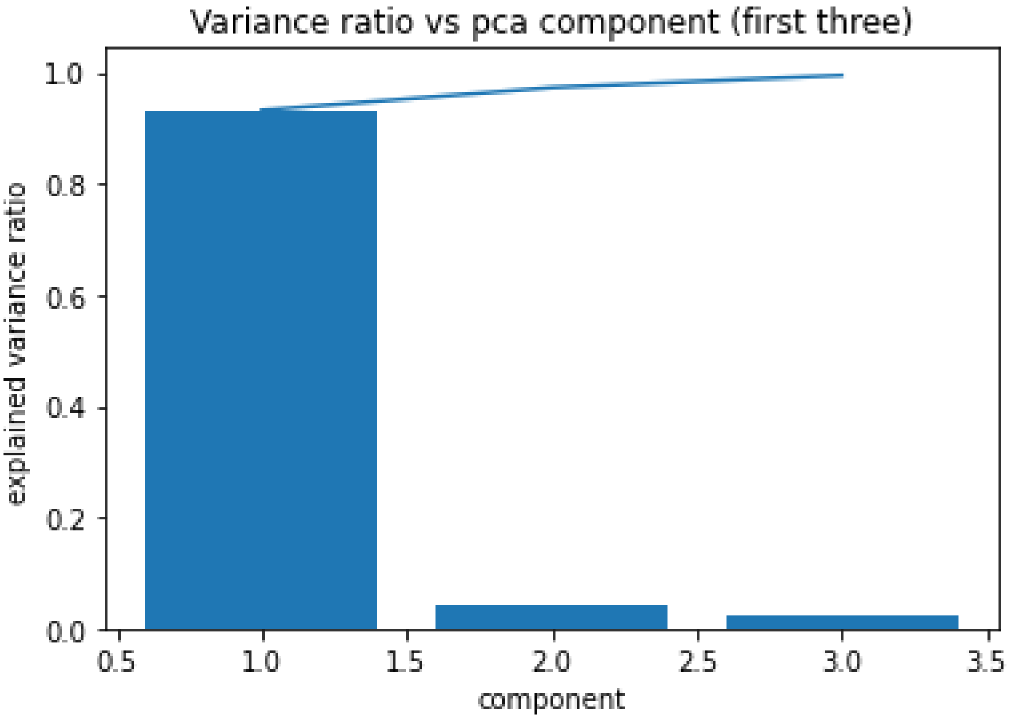
As we see from the heart map every feature is highly correlated with other features as variation in yield rate is very less across different maturities.

Split data into training and test sets. Use random_state = 42. Use 85% of the data for the training set. Use the same split for all experiments.

Part 2: Perform a PCA on the Treasury Yield dataset

Compute and display the explained variance ratio for all components, then recalculate and display on n_components=3.





What is the cumulative explained variance of the 3 component version.

Cumulative explained variance for first three components is 0.9944

Part 3: Linear regression v. SVM regressor - baseline

Fit a linear regression model to both datasets (the original dataset with 30 attributes and the PCA transformed dataset with 3 PCs.) using SKlearn. Calculate its accuracy R2 score and RMSE for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

Fit a SVM regressor model to both datasets using SKlearn. Calculate its accuracy R2 score and RMSE for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

R2 Scores:

R2 Score (train)		
	Linear Regression	**SVM**
without PCA	0.902273035	0.989409472
with PCA	0.867268386	0.979038942

R2 Score (test)		
	Linear Regression	**SVM**
without PCA	0.904130954	0.989572643
with PCA	0.863706809	0.977038429

RMSE scores:

RMSE (train)		
	Linear Regression	**SVM**
without PCA	0.311789241	0.102639076
with PCA	0.363363338	0.144397782

RMSE (test)		
	Linear Regression	**SVM**
without PCA	0.314084055	0.103584268
with PCA	0.374493052	0.153711898

Part 4: Conclusions

Write a short paragraph summarizing your findings. Which model performs best on the untransformed data? Which transformation leads to the best performance increases? How does training time change for the two models. Report your results using the Results worksheet format. Embed the completed table in your report.

Training time (seconds)		
	Linear Regression	**SVM**
without PCA	0.005990505	0.645667315
with PCA	0.002002478	0.878219366

For linear regression, the time taken to fit train data is less for PCA transformed data with first three components. This is expected because the model will be working on 3 columns as opposed to 30. However, for SVM this time is less without PCA (repeated for multiple random states). With both the models, the R2 accuracy slightly decreases with PCA transformed data (in both train and test data) and RMSE increases slightly but the model with PCA is more parsimonious and takes less time to train and hence it is better to use PCA transformation when the data is highly correlated.

Part 5: Appendix

Link to github repo: IE517/IE517_FY21_HW5_prajwal.ipynb at main · chinthojuprajwal/IE517 (github.com)