My Name (myNetID)

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!

A picture containing text, monitor, screen, television

Description automatically generatedA picture containing text, monitor, screen, flat

Description automatically generated

Chart, bar chart

Description automatically generated

*Statistical descriptions of all features is too large of a table to place here, so it can be found in my printed PDF of my code attached at the end of this document*

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 3 components: 0.994405917309303*

**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).

**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.

*Overall, the linear regression model did the best in accuracy on the untransformed data. However, most likely, this model, along with the SVR model for untransformed data, are most likely not generalizable since it is dependent on 30 features. When the data was transformed using PCA, the accuracy was very comparable to what was shown for the full 30 feature model, but the model will most likely be more generalizable as a result of reducing the data to 3 principal components.*



**Part 5: Appendix**

Link to github repo