1 Overview

We did a visualization of data using PCA and also a performance evaluation between PCA with RBF and simply RBF. Both their codes can be found in the Github Milestone 4 folder. Our performance evaluation used a 10 time 10-Fold Cross Validation on MSE and R^2. We then used a sign test to determine which method was better.

2 Building the Model

Feature transformation can be seen in our CV file. We used the exact same feature transformation for both models and evaluated both of them using MSE and R^2. Our process was feature transformation to PCA to RBF, and feature transformation to RBF. Obviously, we also used the same data sets. Note, RBF was the same as Milestone 1.

For PCA, we visualize the model using histogram models and dependencies for 11 dimensions, using whiten. The variance retained from 11 components is around 96 percent. We wanted to reduce to 2 or 3 dimensions in order to visualize, but when we did that we obtained much worse results - 11 dimesnions proved to be the most accurate. Essentially, we lost too much variance/data by condensing our features into too few dimensions, especially for RBF.

3 Visualization

In order, we have the following images: 1) A figure of a reduction to 2 dimensions. We can see that while there are some discrepancies, there is too much overlap for an RBF to give a proper prediction. 2) A figure of a reduction to 3 dimensions. A little more clarity, but also clear that it is not good enough. 3) A histogram of how often each component shows up. 4) A dependency chart that shows how much each feature from the original impacts each component. 5) A comparison between each component that shows the correlation between them visually.

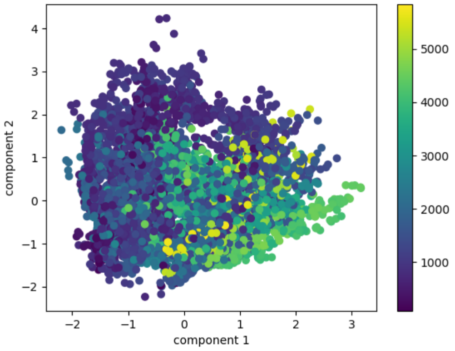

Figure 1: 2 Dimension PCA

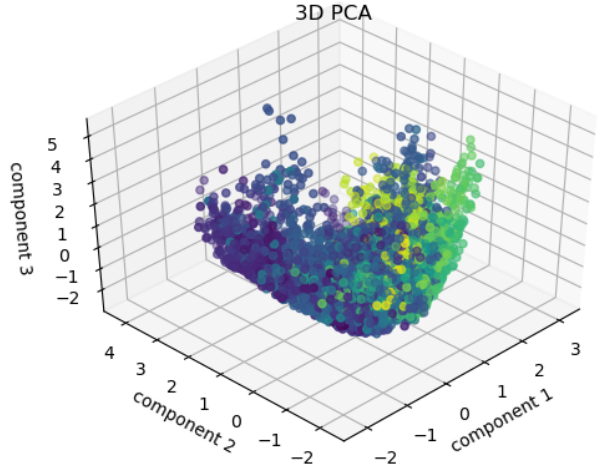

Figure 2: 3 Dimension PCA

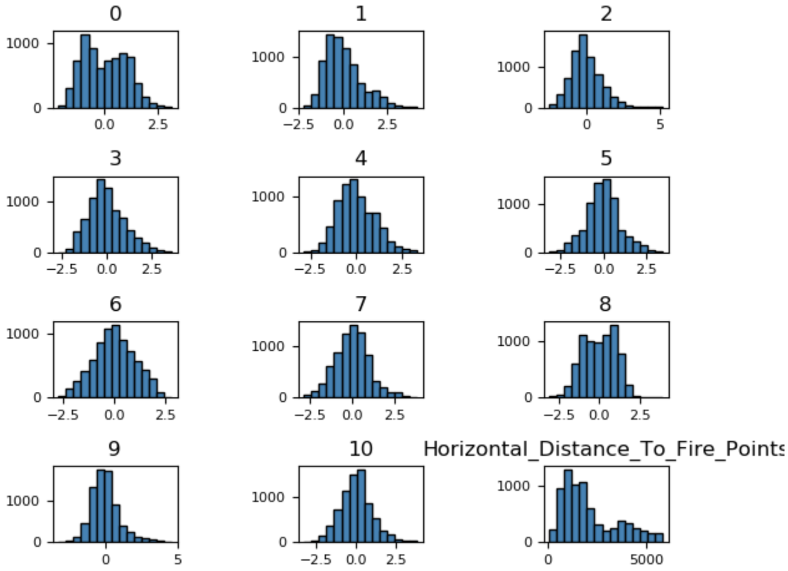

Figure 3: Histograms

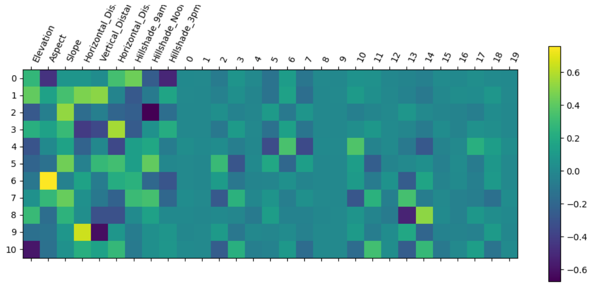

Figure 4: Dependencies

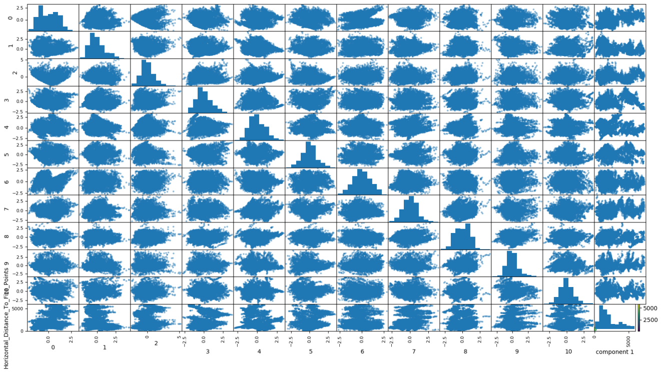

Figure 5: 2D Comparisons

4 Performance Comparison

Shown in the chart below of MSE and R^2 results of each 10-Fold CV run. PCA had an average MSE of 264.637 and R^2 of 0.9624 while RBF had an average MSE of 266.301 and R^2 of 0.9619

PCA	RBF	MSE Winner
266.4561003	268.9720765	PCA
263.2960901	268.1444154	PCA
266.803272	265.8051147	RBF
257.8432446	260.9941679	PCA
266.2363569	270.4000389	PCA
268.7418446	263.7731692	RBF
270.9569755	274.6385528	PCA
268.1189536	266.8380138	RBF
258.9771179	261.4804447	PCA
258.9426846	261.9596406	PCA

PCA	RBF	R^2 Winner
0.96195218	0.96122566	RBF
0.96255366	0.9611656	RBF
0.96177291	0.96212258	PCA
0.9642735	0.96331164	RBF
0.96199314	0.96068929	RBF
0.96148387	0.96305067	PCA
0.96069516	0.95958824	RBF
0.96112886	0.96171172	PCA
0.9639274	0.96328531	RBF
0.96409666	0.9631571	RBF

5 Statistical Sign-Test

We use a non-parametric test because RF and RBF are both non-parametric methods. The null hypothesis for the sign-test is $H_0 : p(RBF_{win} > RF_{win}) = 0.5$.

For sign-test would reject the null hypothesis. For $\alpha = 0.05$ and $n = 10$, we need $RBF_{win} or RF_{win} > 8$. For MSE, PCA only does better 7 times, so we fail to reject the null hypothesis. For R^2, RBF only does better 7 times, so we also fail to reject the null hypothesis. Thus, with a 95 percent confidence, we fail to reject the null hypothesis.