Interpretable Model Tests on Synthetic Datasets

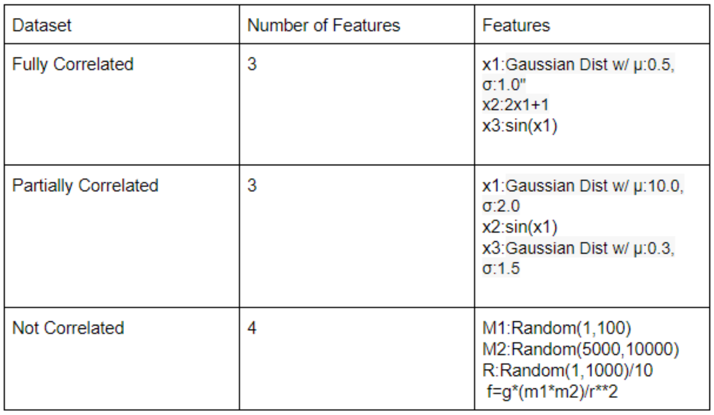
Chetan Singh, Praneeth Chedella

January 2020

Abstract

This report details the work done by our team on the task of understanding of interpretable machine learning models. We considered PDP and ICE models for interpreting the data by using synthetic and non synthetic data sets.

1 Datasets



Dataset	Number of Features	Features
Fully Correlated	3	x1:Gaussian Dist w/ μ:0.5, σ:1.0" x2:2x1+1 x3:sin(x1)
Partially Correlated	3	x1:Gaussian Dist w/ μ:10.0, σ:2.0 x2:sin(x1) x3:Gaussian Dist w/ μ:0.3, σ:1.5

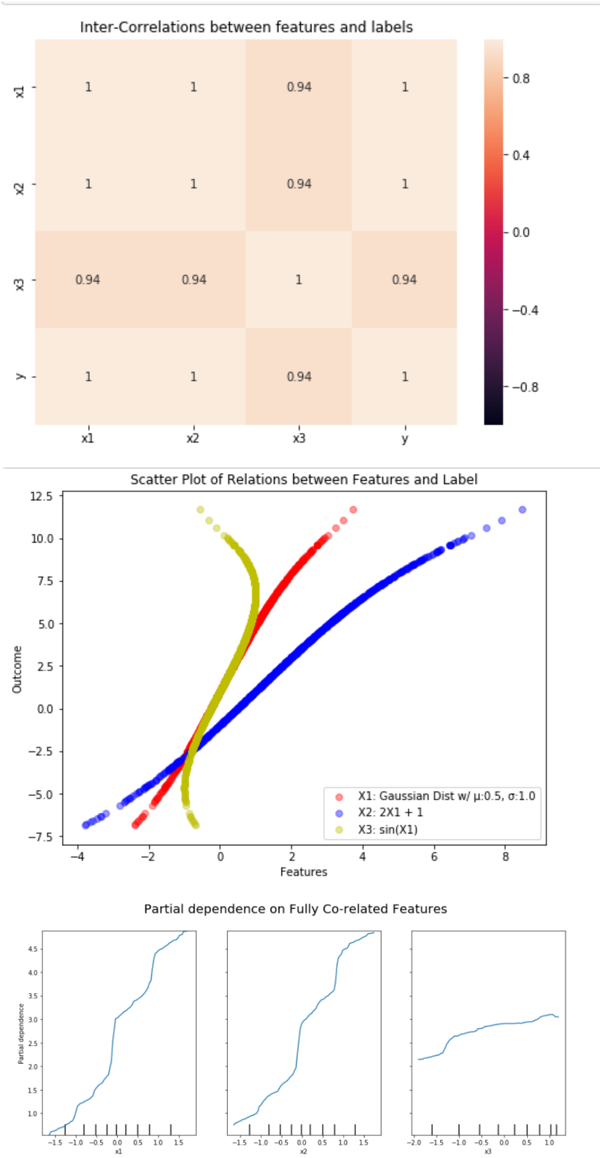
Not Correlated	4	M1:Random(1,100)
		M2:Random(5000,10000)
		R:Random(1,1000)/10
		f=g*(m1*m2)/r**2

Figure 1: Feature Table

2 Results and Observations

2.1 Fully Co-Related Dataset

The Fully Co-Related dataset is a synthetic dataset that has co-related features, the feature *x1* is generated from a Gaussian Distribution with mean and sigma as 0.5 and 1.0 Resp., feature *x2* is a linear combination of *x1* as *2x1 + 1*. The feature *x3* is generated from *sin(x1)*. Therefore *x1 and x2* are linearly co-related with each other, whereas *x1 and x3* are not linearly co-related with each other. Hence when a Co-Relation matrix is plotted it shows the least co-relation value among the three features.



Inter-Correlations between features and labels

Scatter Plot of Relations between Features and Label

Partial dependence on Fully Co-related Features

Figure 2: Correlation Matrix, Scatter Plot and PDP for Fully Co-Related Dataset

Since it is a co-related dataset we can see that the PDP is confirming the information that is shown in the scatter plot there by saying that the features x1 and x2 are the features that are impacting the outcomes as they are linearly co-related, and giving a completely different interpretation for *x3* as it do not have much of an impact as it is generated from the *sin(x1)*.

2.2 Non Co-Related Dataset

The Non Co-Related dataset is a synthetic dataset has features from the Newton's Law of Gravitation which are completely not co-related to each other, the features *m1, m2, r* are random values that follows different distributions and the F value is given by

$$F = G(m1 * m2)/r^2$$

where the value of G is gravitational constant as 0.667 for only this test case, and m1, m2 are the masses of two different objects, and finally r is the distance between the two masses. While computing the co-relation matrix and the scatter plot, we can confirm the non-co-relations. and from the above PDP we can see that the values *m1, m2* are impacting a small fraction of F, whereas the r seems to affect all the ranges of F proving that the Radius between two Masses is inversely proportional to the Force.

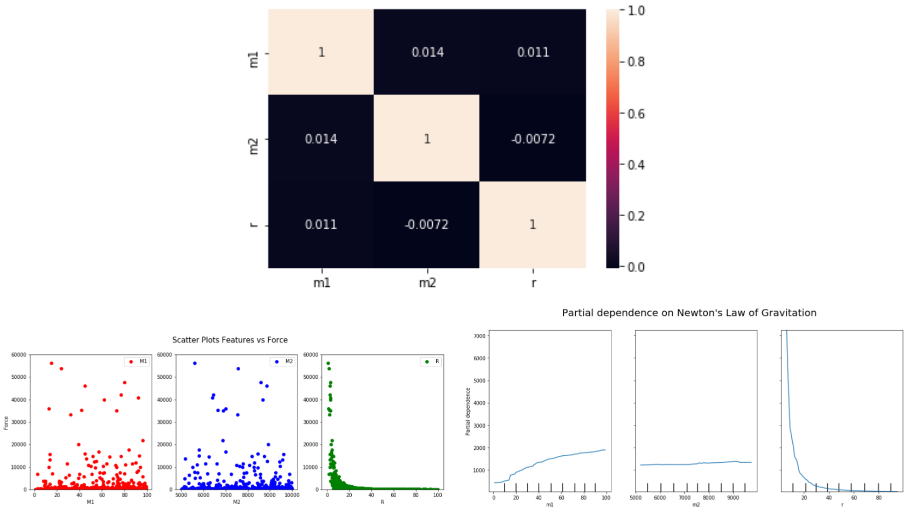


Figure 3: Correlation Matrix, Scatter Plot and PDP for Non Co-Related Dataset

2.3 Partially Co-Related Dataset

The Partially Co-Related dataset is a synthetic data set that has few features that are co-related to each other and few that are not at all co-related. Here the feature *x1* follows a Gaussian distribution with mean and standard deviation as

10.0 and 2.0 Resp., the feature *x2* is the *sin(x1)*. Feature *x3* follows a Gaussian distribution with mean and standard deviation of 0.3 and 1.5 Resp.

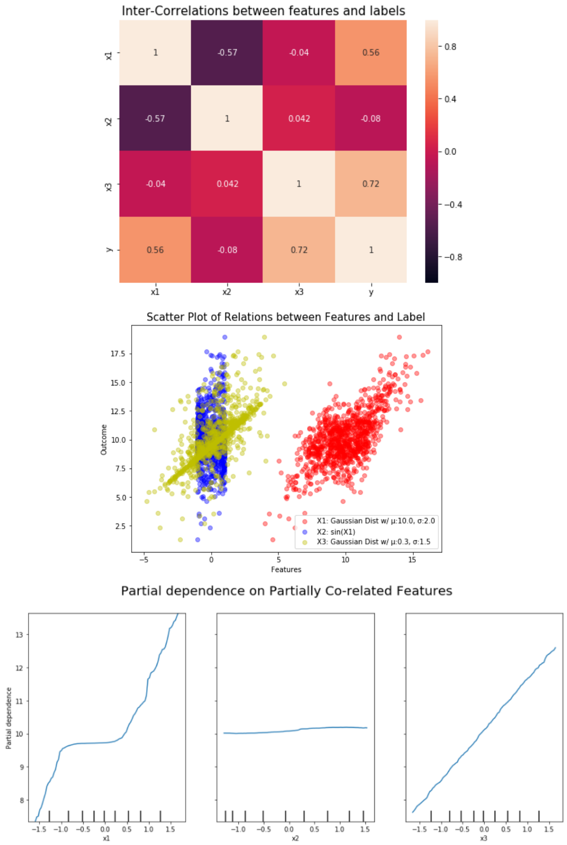


Figure 4: Correlation Matrix, Scatter Plot and PDP for Partially Co-Related Dataset

and from the above PDP we can see that it clearly interprets that the feature *x2* has almost no impact on the outcome as it is non-linearly co-related and generated from the feature *x1*. Following through the ICE plots demonstrates similar behaviour. One final note that for the feature *x3* plotted in ICE, the impact on the outcome is little dampened as compared to PDP.

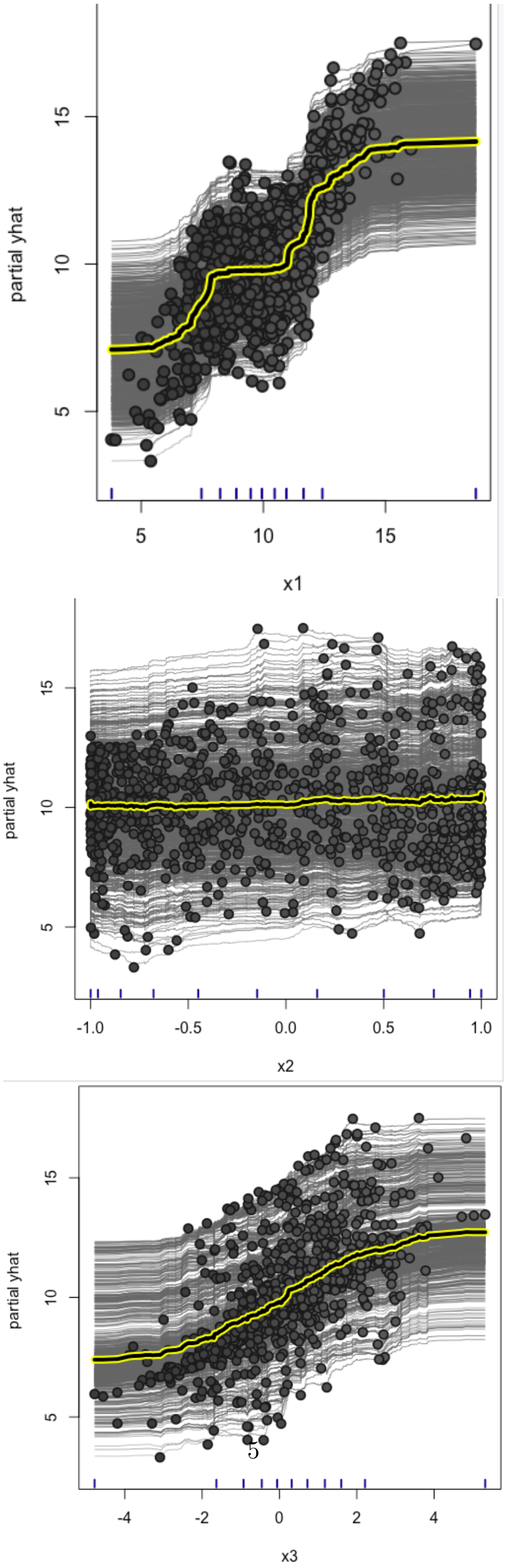


Figure 5: ICE Plots for Partially Co-Related Dataset

3 Conclusions and Findings

After verifying the results of both synthetic (mentioned in the report) and non synthetic datasets (Boston Housing) using PDP and ICE plots we observed the following points:

1.Even though plotting a Scatter plot or Co-Relation matrix shows the highly co-related features, but they tend to fail when dealing with the features that are not linearly co-related which can be seen using PDP. Ex: Fully Correlated Dataset *(features x1 and x3)*.

2. PDP fails to distinguish the impact of the features which are linearly co-related to each other and ends up demonstrating the same effect on the outcome for those features. Ex: Fully Correlated Dataset *(features x1 and x2)*.

3. In many cases, the interpretations done by PDP can be also interpreted by Scatter plots. Ex: Non Correlated Dataset *(feature r)*.

4. As in some cases through PDP we can manage to interpret results, whereas Scatter plots can give confusing results. Ex: Non Correlated Dataset *(features m1 and m2)*.

5. ICE plots give us more detailed interpretation on an instance level, demonstrating the trend of the feature *w.r.t* densities of instances. Ex: Partially Correlated Dataset *(feature x2)*.

4 References

`https://github.com/csngh/interpretable_ml_tests`