Emilie LEDOUSSAL

Advanced Data Analysis and Machine Learning

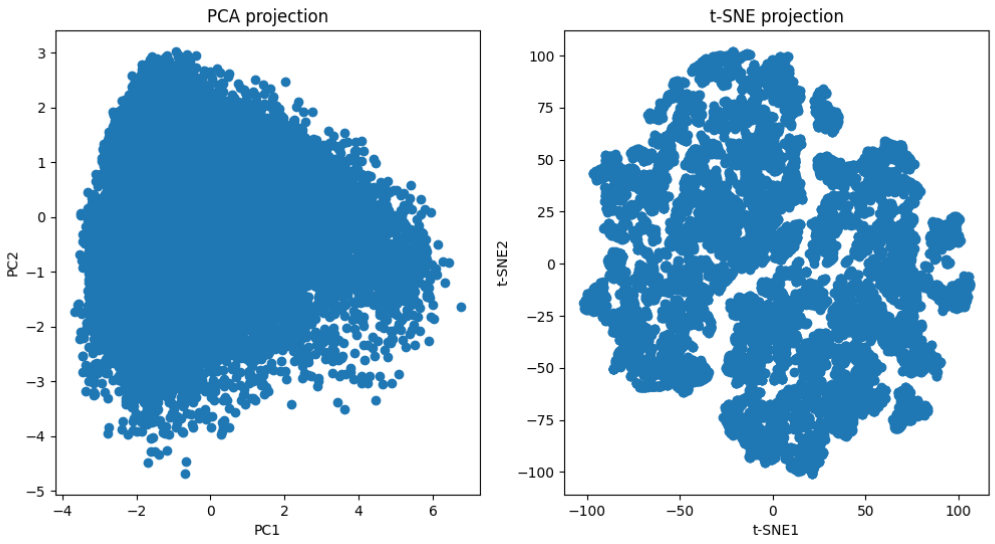
Task 1. Comparison between linear and non-linear Data Reduction

The dataset we use in this task is an hourly bike rental dataset from 2011–2012, combining the time, weather, if the day is a working day, if the user is casual or registered, to explain how season, date, weather, and user type and how it impacts the bike rent.

We can assume that the number of bikes rented in not a linear situation, it is affected by seasons, weather, holidays, days-off, and plenty of non-characterizable reasons why someone would prefer to take the bus or stay at home instead of renting a bike to go to work. It is the perfect dataset to compare linear and non-linear techniques.

1)Visualization of PCA vs t-SNE

After standardizing the data with a standard scaler, we use both PCA and t-SNE on the exact same data.



The first plot is the Principal Component Analysis, we can see that there's only one big cluster, which does not help much to distinguish non-linear, it is useful to see that the

Principal Components captured most of the variance, but it fails to capture the non-linear ways of the dataset

On the other hand, on the second plot, there is the t-SNE projection, which is a non-linear technique. It preserve local neighbor relationships and form small well-separated clusters and highlight some tendency. This is a good way to investigate and identify local tendency, for example, how bike rental goes on rainy days or holidays.

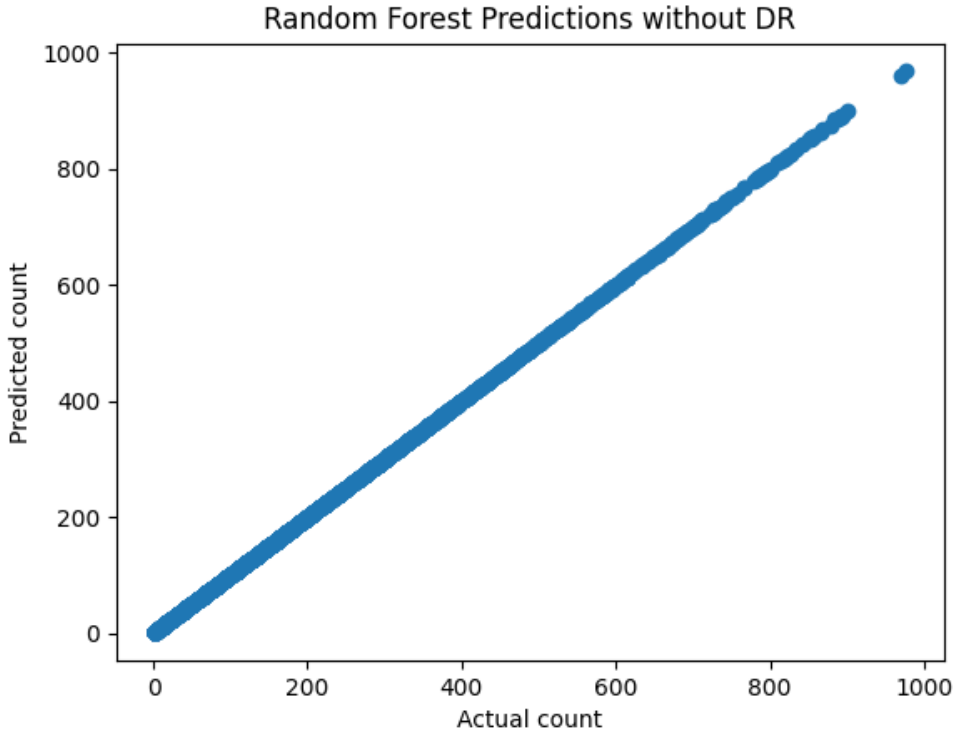
PCA is good for a global overview of the situation while t-SNE is better for a more precise use and identify clusters.

2) Comparison of PCA and t-SNE for Data Reduction on a simple model like Random Forest

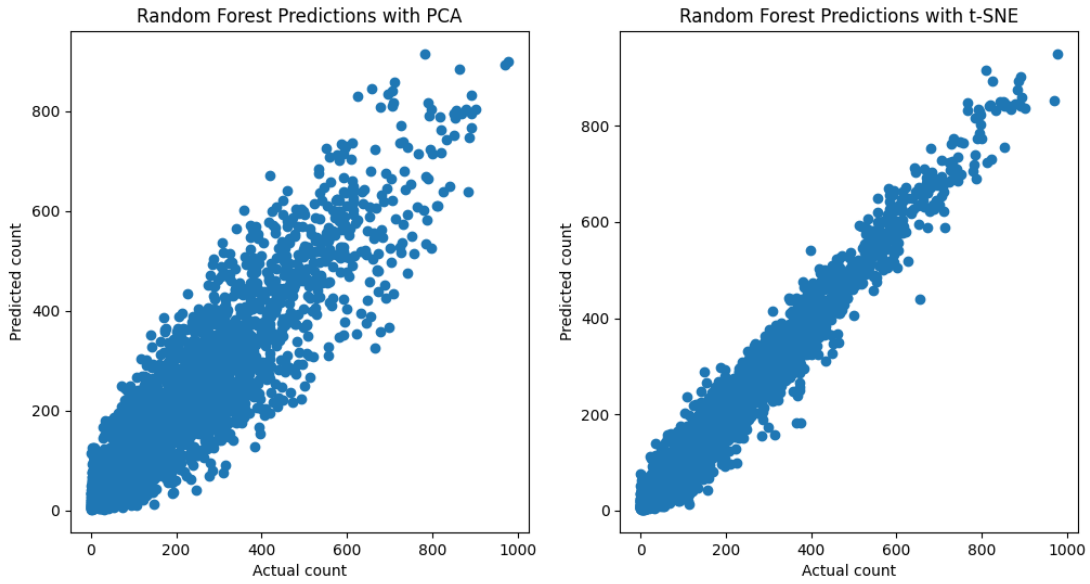
As usual, I will split the data in train and test set, for this application, the validation one is not needed, let's keep it simple. We use the RandomForestRegressor model.

I will use Mean Square Errors and Predicted vs Actual plots to characterize the performance of these models.

First in this figure, is the model without any data reduction, it's doing very well with a MSE of 0.0596. It fit very well to the predicted vs actual plot.



Now, let's see with Data Reduction, here is the comparison of the the PCA DR and the t-SNE data reduction.



Because of the data reduction, the MSE increased greatly, but it's normal. The Mean Squared Error with PCA is of 4916.938839384349 and the Mean Squared Error with t-SNE is 1083.920867721519

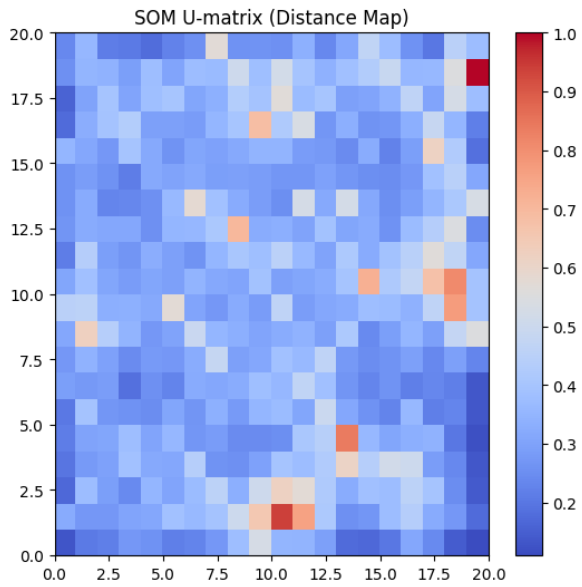
We can see that on these plots and MSE, that for a non-linear Dataset, t-SNE Data-Reduction perform the best, since it degraded less the trend, and that on the Predicted Vs Actual, the points are closer to the middle line.

Task 2. Visualizing with SOM

SOM means Self Organizing Map, it's an unsupervised neural network technique that maps high dimensioned data on a lower dimension grid(2D), while preserving topological relationships of the input. Here we will use it on a dataset of containing the value of pixels of multiple images of handwritten numbers, it's a non- linear situation and we can note that some numbers when written can be very close to another numbers like 4 that can look like 9, 5 like 8 ect...

To use SOM in Python, we're going to use the library Minisom available on Github: https://github.com/JustGlowing/minisom

We have here the U-Matrix or Distance Maps that shows clusters of data that shows when some data is distinct from the other (red) or ambiguous (blue) and could be mixed



Here we have the SOM Visualization of the Dataset, and how it was sorted without any supervision. We can see clusters of numbers that are usually most likely to be mixed like 4,9 and 7. 5 and 9, and more. Each person has their way of handwriting their numbers and here the clusters shows how the same numbers can be mixed with other numbers in a different way. 4 and 7 can be mixed if you put a too big bar on your 7, but if you put no bar and make your 7 too flat, it can be taken as a 1. This representation shows these different clusters