Test 1

The workflow:

- Exploratory data analysis
    - Missing values
    - Distribution
    - FFT
    - Autocorrelations
- Modeling
    - Train/Test Split
    - Baseline: Linear Regression
    - Neural Network
    - Clustering

No missing values are observed in the synthetic data set.

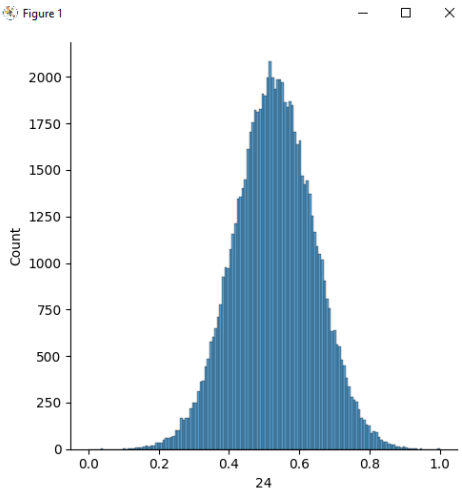
The distributions of a target function and features are normal. Cross correlation heatmap as well as pair plots demonstrate no linear or any obvious correlation between the variables. The plot of a limited amount of samples doesn't show any linear or polynomial relationship between the variables.

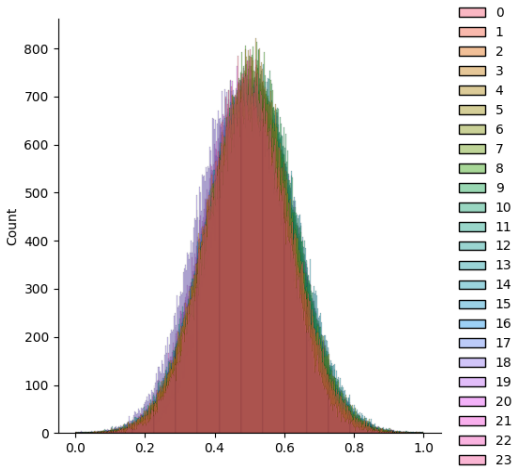
To check if the data could be a time series the autocorrelation plot and pairplot are investigated and show no time series behaviour. FFT transformation also didn't give any indication of a signal. The runs of LR, RF and GB models demonstrated the low model accuracy (R2 for validation dataset ~ 0.015), but neither one step ahead prediction gave a reasonable accuracy.

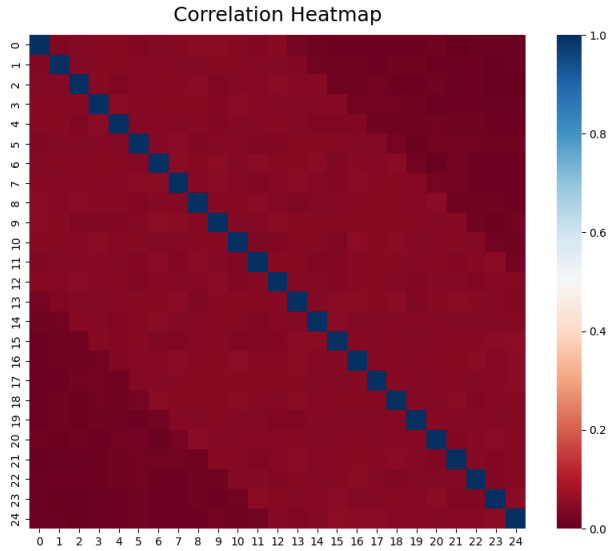
In this case the neural network is proposed as a model that could find the nonlinear dependencies of the variables that cannot be detected from EDA in a multivariate regression problem.

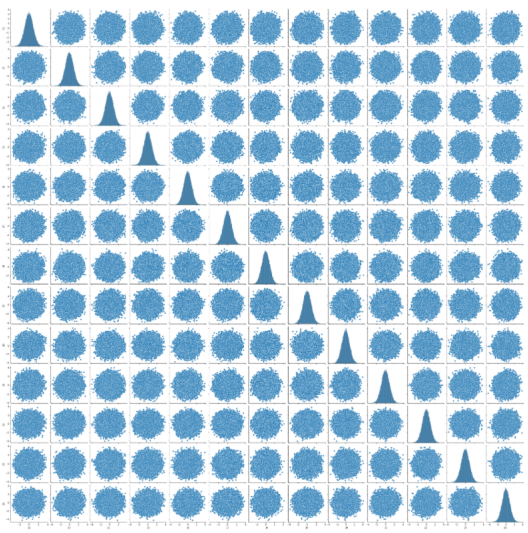
Although a neural network converges for a training data set, it doesn't generalize to the unseen validation dataset.

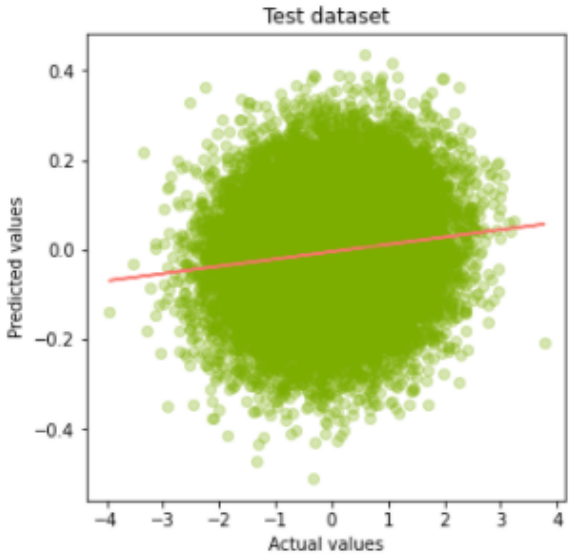
The hyperparameters of the neural network could be optimized (regularization, dropouts, number of neurons and layers) for the model to not overfit. Also PCA can be run to see if there is a higher dimensional structure. The current comprehensive analysis was not able to detect the relationship within the data. The dataset has some structure and not completely random, but after conducting comprehensive analysis (cross-correlations, autocorrelation, fft, validation of different models etc) making accurate predictions out of dataset was not achieved.





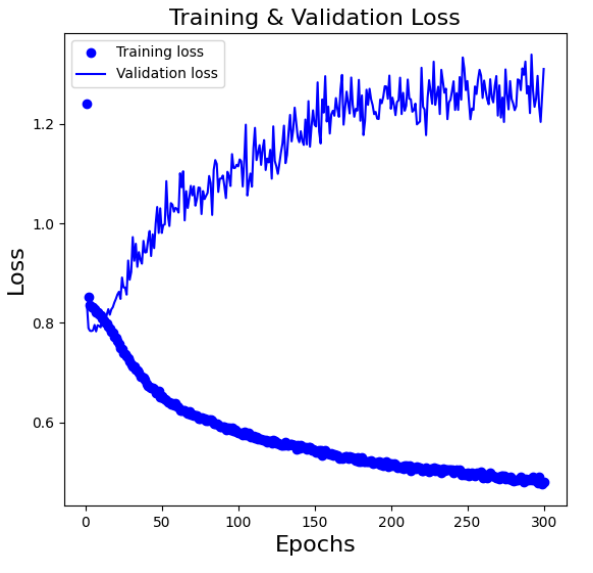


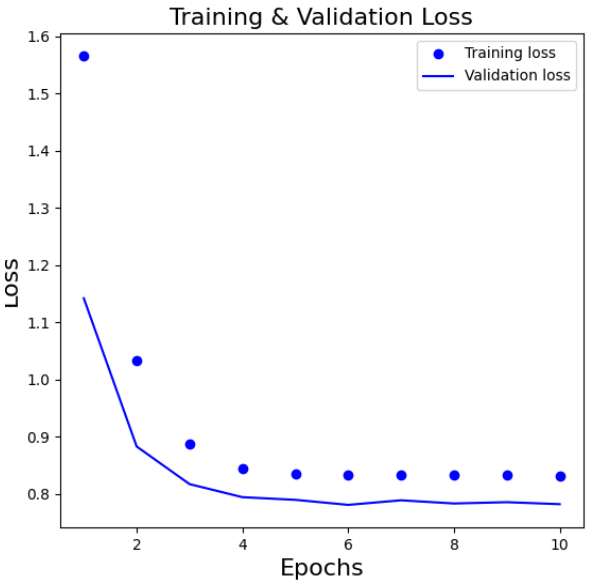




Linear Regression Results

Neural Network Results





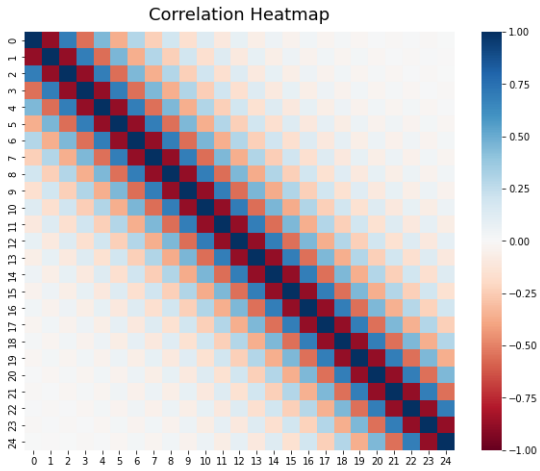
Test 4

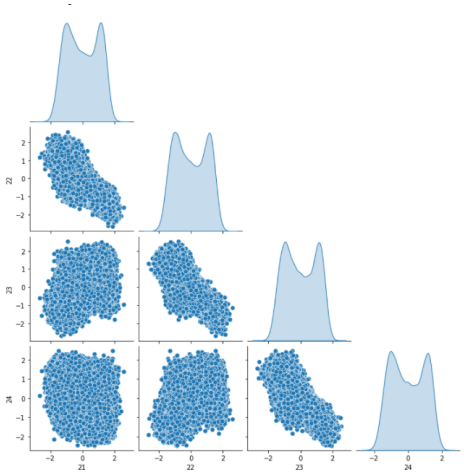
The workflow:

- Exploratory data analysis
    - Missing values
    - Distribution
    - Check if the data is time series
- Modeling
    - Baseline: Linear Regression
    - Random Forest
    - Gradient Boosting
- Hyperparameters Optimization

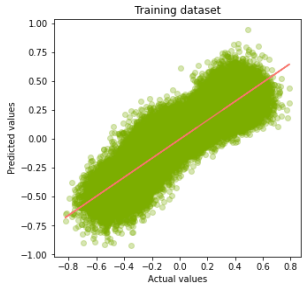
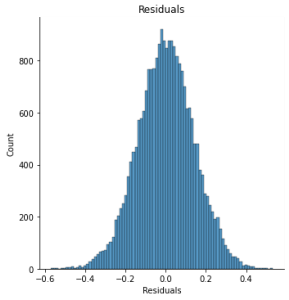
No missing or null values are observed in the data. The data represents synthetic data with bimodal distribution. The correlation coefficients demonstrate the strong linear relationship between the target variables. As well as multicollinearity among the predictors is observed.The pairplots of the variables with highest correlation coefficients illustrate the segmented linear dependency with the target variable.
Linear regression is chosen as the baseline model. In order to take into account the non-linear dependency, random forest and boosting methods were applied.
The autocorrelations were built to see if the data should be treated as time-series and not cross-sectional.

The errors are presented in the Table below. All the models provide reasonable accuracy in terms of R2, mse metrics for the test data set. All methods can be used as a forecaster for this data set. RF and boosting algorithms are not affected by multicollinearity. Number of features can be reduced to 2 (23d and 24th) out of 24 without loss in the predictive accuracy. Hyperparameters optimization for xgboost doesn't provide any significant improvement to the model





	train_mse	train_r2	test_mse	test_r2
Linear Regression	0.019	0.82	0.019	0.82
Random Forest	0.013	0.879	0.013	0.876
Gradient Boosting	0.012	0.887	0.012	0.88
XGB XGB with Optimized Hyperparameters		0.9		0.891 0.892
Feature reduction to 2 features		0.887		0.884

Linear Regression		
		
Gradient Boosting