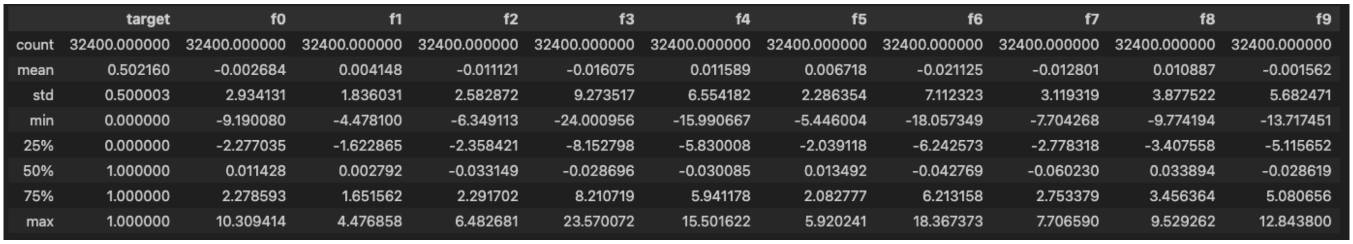
Vincenzo Brigandì - 07/02/2024

Data Scientist Technical Assignment Solution

The primary objective of this project is to develop a classifier using the provided dataset. Given that this is a binary classification task, my initial step was to assess the balance between the two classes, which turned out to be perfectly balanced with 27000 items in each class with no missing values. Subsequently, I proceeded to split the dataset into three sets: the Training set (60%), the Validation set (20%), and the Test set (20%). This division allowed me to conduct Exploratory Data Analysis (EDA) specifically on the Training set. During this phase, I computed descriptive statistics for both the features and the target variable, and no significant outliers were identified.

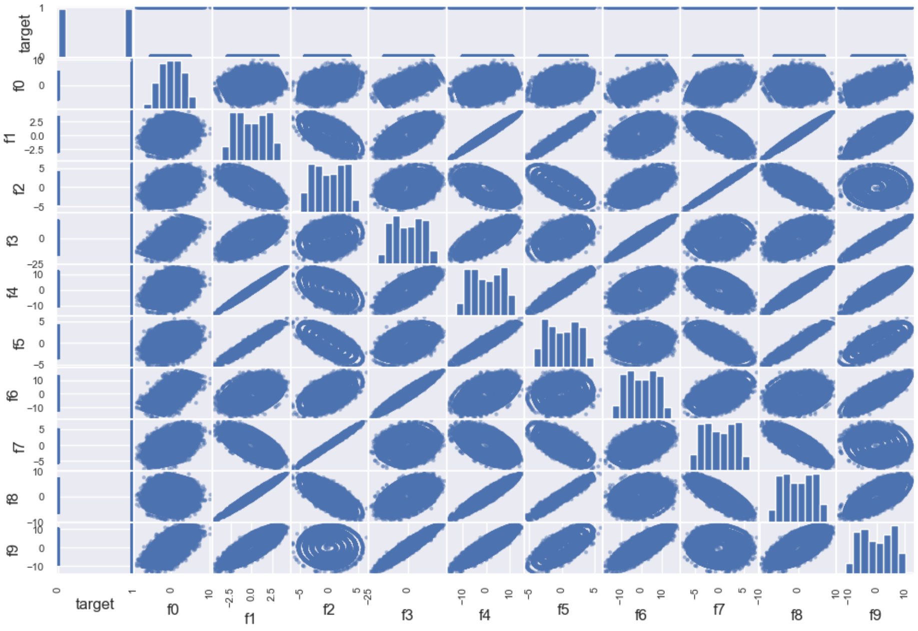


	target	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9
count	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000	32400.000000
mean	0.502160	-0.002684	0.004148	-0.011121	-0.016075	0.011589	0.006718	-0.021125	-0.012801	0.010887	-0.001562
std	0.500003	2.934131	1.836031	2.582872	9.273517	6.554182	2.286354	7.112323	3.119319	3.877522	5.682471
min	0.000000	-9.190080	-4.478100	-6.349113	-24.000956	-15.990667	-5.446004	-18.057349	-7.704268	-9.774194	-13.717451
25%	0.000000	-2.277035	-1.622865	-2.358421	-8.152798	-5.830008	-2.039118	-6.242573	-2.778318	-3.407558	-5.115652
50%	1.000000	0.011428	0.002792	-0.033149	-0.028696	-0.030085	0.013492	-0.042769	-0.060230	0.033894	-0.028619
75%	1.000000	2.278593	1.651562	2.291702	8.210719	5.941178	2.082777	6.213158	2.753379	3.456364	5.080656
max	1.000000	10.309414	4.476858	6.482681	23.570072	15.501622	5.920241	18.367373	7.706590	9.529262	12.843800

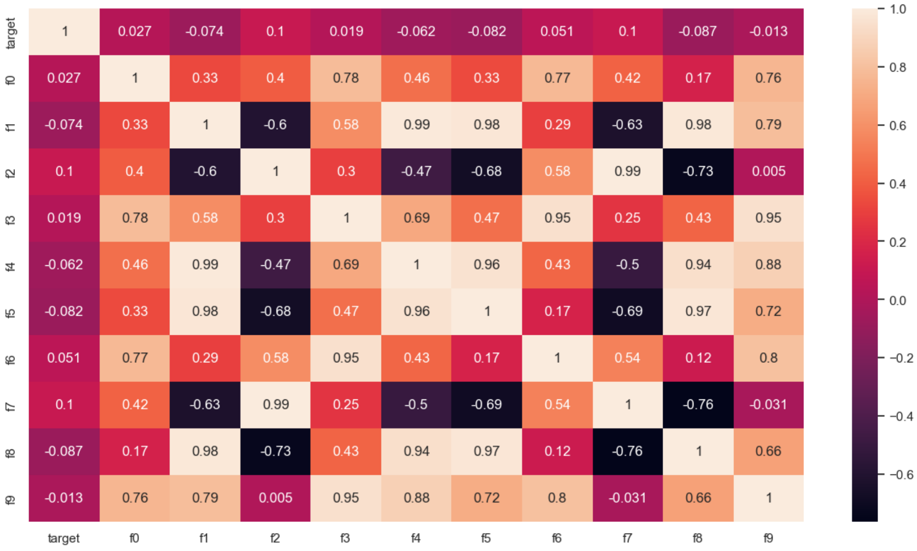
Descriptive Statistics

Further exploration involved generating scatter plots for each feature in relation to every other feature as well as in relation to the target variable. These visualizations revealed that the target variable exhibited limited correlation with any specific feature, suggesting the potential necessity for feature engineering. Notably, certain pairs of features displayed a high degree of correlation, such as f1 and f8, as well as f2 and f7. Additionally, the correlation patterns between some variables displayed a distinctive spiral shape, which hinted at the underlying structure of the dataset, as seen in the correlation between f2 and f9. Regarding the distribution of variables, some exhibited a bimodal pattern, implying a possible polarization towards the -1 and +1 values of the target variable.

Furthermore, I computed the Spearman correlation, which is a rank correlation effective at capturing monotonic relationship within the data, both among variables and with the target variable. The Spearman correlation with the target variable remained close to zero, indicating that a simple linear model may struggle to accurately predict the correct class. Additionally, some variables confirmed to be highly correlated with others with coefficients reaching as high as 94%, 96%, and 97%, suggesting a possible redundancy.



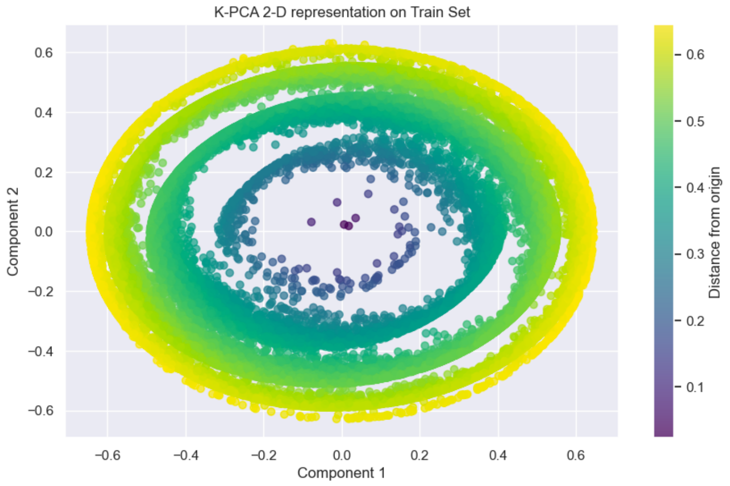
Features/Target Scatter plot

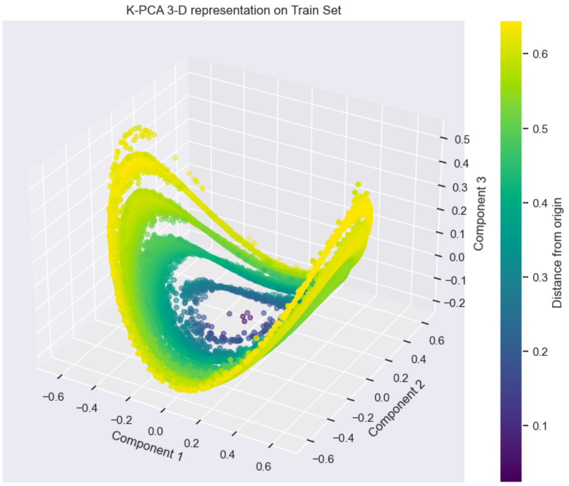


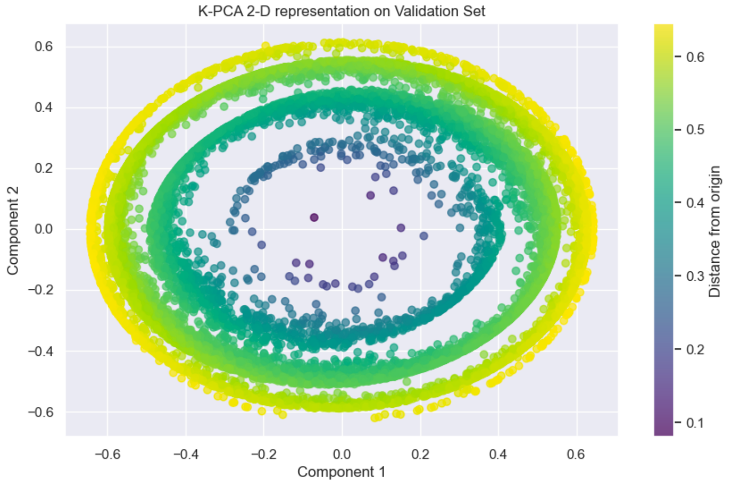
	target	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9
target	1	0.027	-0.074	0.1	0.019	-0.062	-0.082	0.051	0.1	-0.087	-0.013
f0	0.027	1	0.33	0.4	0.78	0.46	0.33	0.77	0.42	0.17	0.76
f1	-0.074	0.33	1	-0.6	0.58	0.99	0.98	0.29	-0.63	0.98	0.79
f2	0.1	0.4	-0.6	1	0.3	-0.47	-0.68	0.58	0.99	-0.73	0.005
f3	0.019	0.78	0.58	0.3	1	0.69	0.47	0.95	0.25	0.43	0.95
f4	-0.062	0.46	0.99	-0.47	0.69	1	0.96	0.43	-0.5	0.94	0.88
f5	-0.082	0.33	0.98	-0.68	0.47	0.96	1	0.17	-0.69	0.97	0.72
f6	0.051	0.77	0.29	0.58	0.95	0.43	0.17	1	0.54	0.12	0.8
f7	0.1	0.42	-0.63	0.99	0.25	-0.5	-0.69	0.54	1	-0.76	-0.031
f8	-0.087	0.17	0.98	-0.73	0.43	0.94	0.97	0.12	-0.76	1	0.66
f9	-0.013	0.76	0.79	0.005	0.95	0.88	0.72	0.8	-0.031	0.66	1

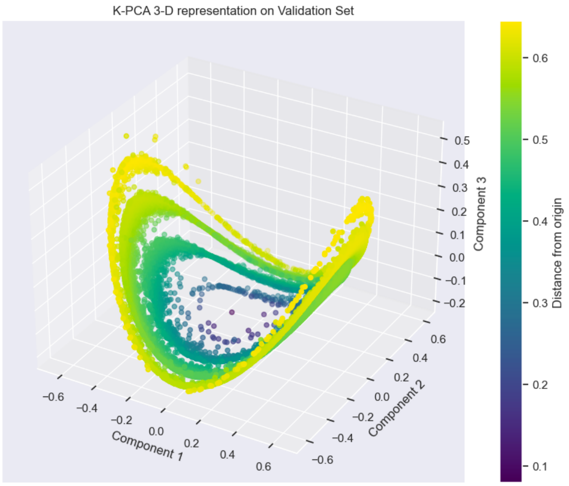
Spearman Correlation Matrix

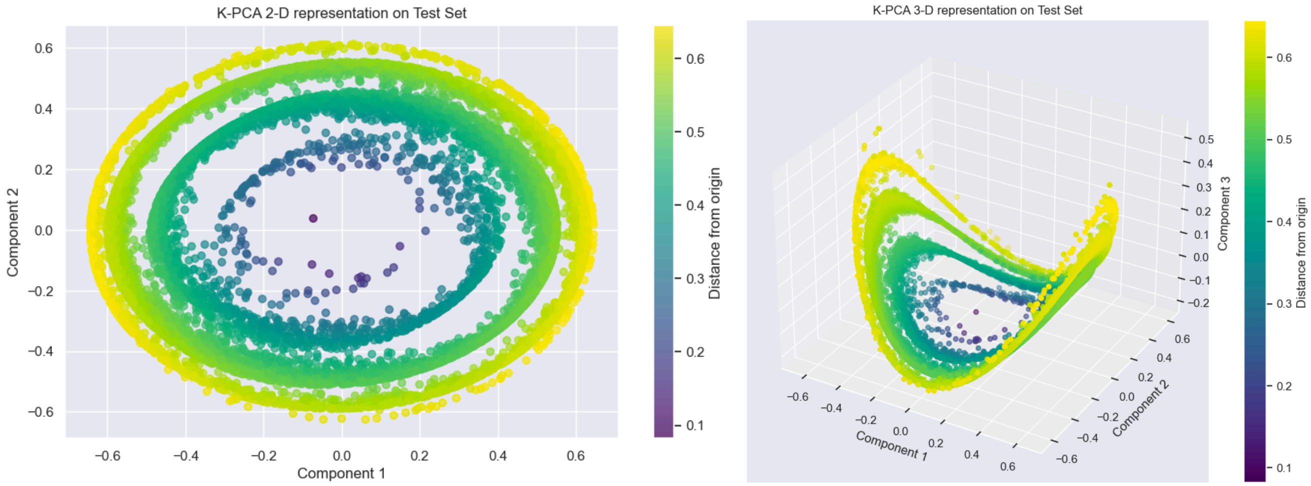
Building upon the insights gained from our previous observations and the hints provided regarding the dataset's generation, I decided to explore dimensionality reduction techniques. Specifically, I applied T-SNE, Locally Linear Embedding, Multi-dimensional scaling, Isomaps, and Kernel PCA to the dataset after standardizing it using a standard scaler. Among these techniques, Kernel PCA stood out as the only one capable of revealing a clear pattern, consistent with the allusion made in point 2) of the text. It required fine-tuning, particularly with a gamma parameter of 0.025 and employing an RBF kernel. By reducing the dataset to either 2 or 3 components, a striking pattern emerged, resembling a bullseye in 2D and a conic spiral in 3D. Notably, this pattern was consistently observable in both the Train, Validation and Test sets.


K-PCA 2-D representation on Train Set


K-PCA 3-D representation on Train Set

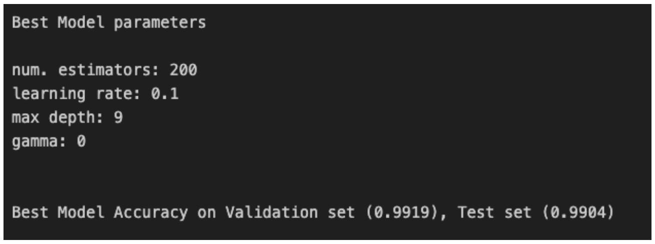

K-PCA 2-D representation on Validation Set


K-PCA 3-D representation on Validation Set



Kernel PCA Extracted features (first 2 LHS, and first 3 components RHS)

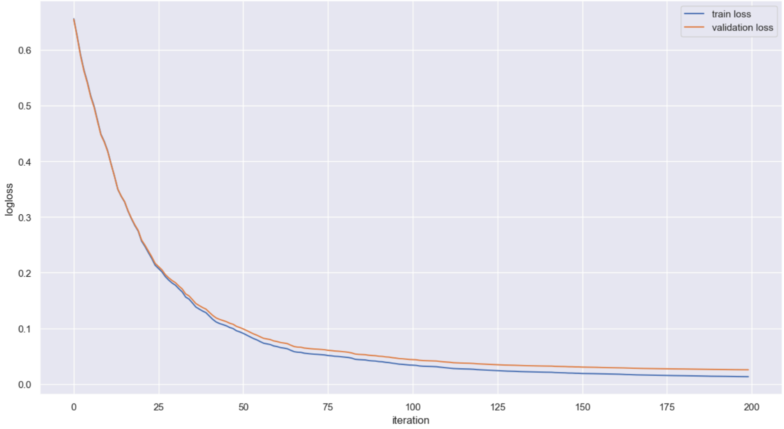
In the final phase of my analysis, I enriched the original dataset with the newly derived features from Kernel PCA and proceeded to implement an XGBoost Classifier model for target prediction. To optimize the model's performance, I conducted a grid search focusing on key hyperparameters such as the number of estimators, learning rate, maximum depth, and gamma. The model that performed best on the validation set achieved an impressive accuracy of nearly 99%. Using the parameters of the “Best Model” I did the prediction that I submit as solution of this exercise.



```
Best Model parameters

num. estimators: 200
learning rate: 0.1
max depth: 9
gamma: 0


Best Model Accuracy on Validation set (0.9919), Test set (0.9904)
```



Train and Validation Loss convergence for Best Model