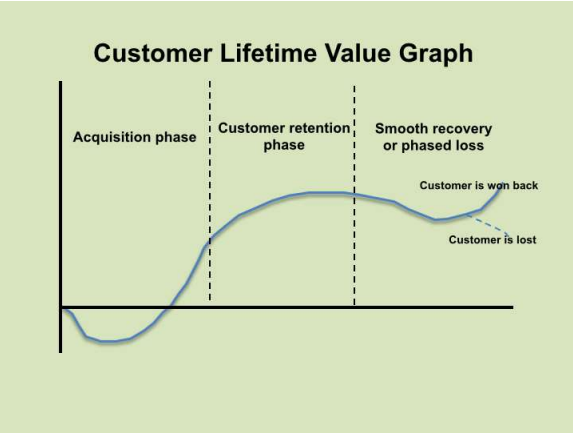
Insurance
Data
Science

Unsupervised Learning applied to the Customer Lifetime Value (CLV)

SPREAD PREDICTIVE ANALYTICS TOWARDS INSURANCE MARKETING FIELD

A key tool: Customer Lifetime Value. It's the value of a customer for a Company over the span of their lifetime relationship



Github repository: https://github.com/claudio1975/Customer_Lifetime_Value

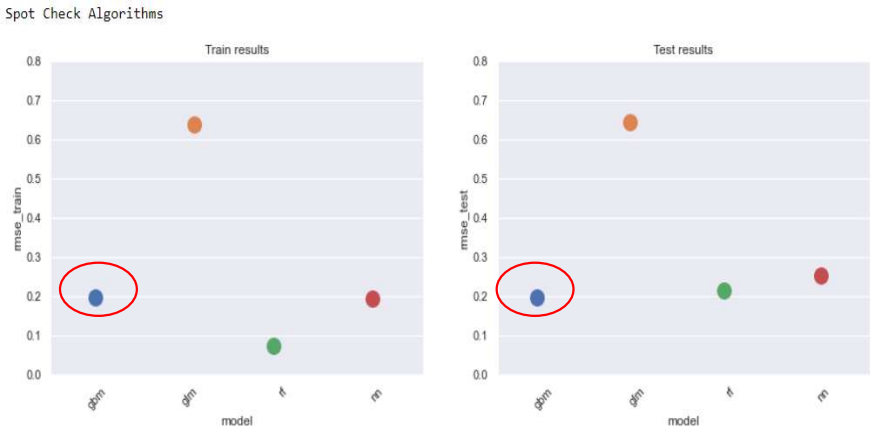
Insurance
Data
Science

Which model can I use to predict CLV?

USE CASE: AUTOINSURANCE DATA SET

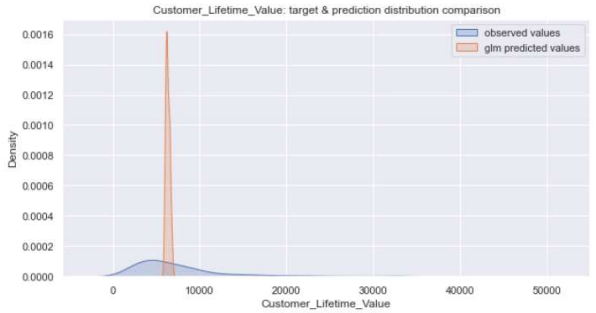
-Generalized Linear Model (GLM); -Random Forest (RF);
-Gradient Boosting Machine (GBM); -Neural Networks (NN)

model	rmse_train	rmse_test
gbm	0.196001	0.197300
glm	0.637301	0.643779
rf	0.073639	0.214578
nn	0.194967	0.252401

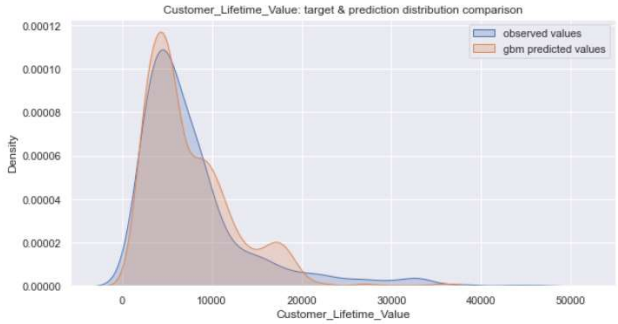


...and the winner is.....GBM!!!...Used as flat (baseline) model for next experiments!!!

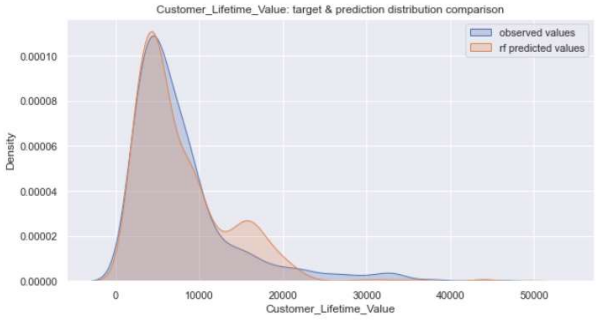
GLM



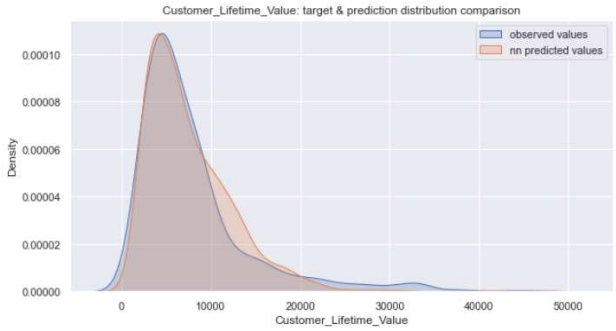
GBM



RF



NN



Insurance

Data

Science

Dimensionality reduction as feature engineering

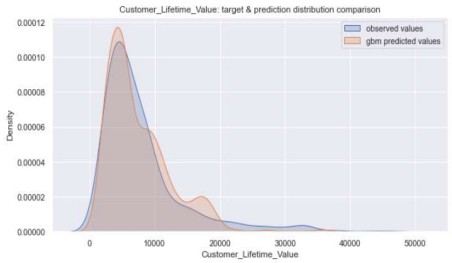
model	rmse_train	rmse_test
gbm_flat	0.196001	0.197300
gbm_pca	0.484353	0.564168
gbm_autoencoder	0.591485	0.627873
gbm_isomap	0.624898	0.665873

CLV prediction using features from dimensionality reduction tools: worst performences compared with flat (baseline) model

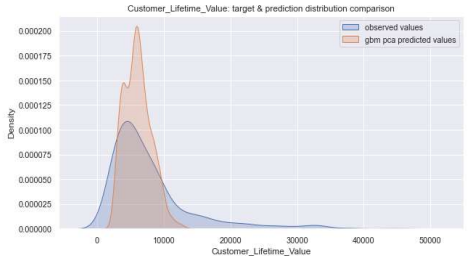
Data augmentation of flat (baseline) model with features extracted by dimensionality reduction tools: improved performences with PCA and Autoencoder features

model	rmse_train	rmse_test
gbm_flat	0.196001	0.197300
gbm_pca_fe	0.192164	0.196431
gbm_autoencoder_fe	0.194582	0.196857
gbm_isomap_fe	0.195837	0.198643

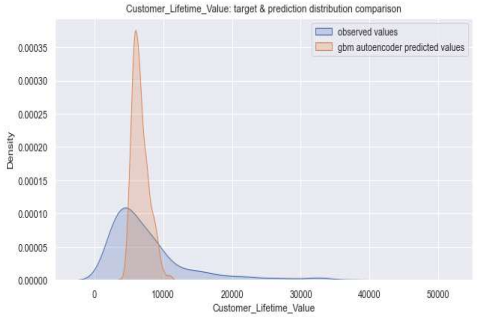
Flat (Baseline)



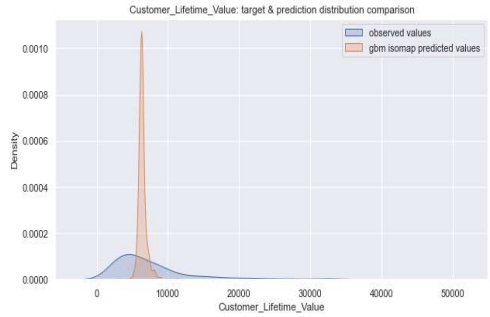
PCA



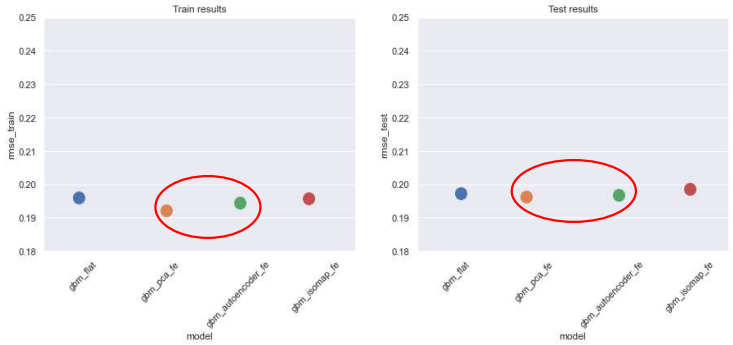
Autoencoder

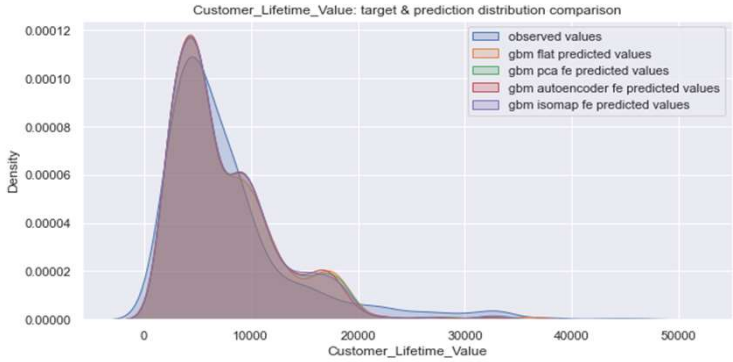


Isomap



Spot Check Algorithms





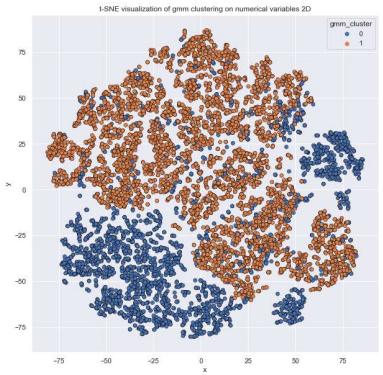
Claudio G. Giancaterino: «Unsupervised Learning applied to the Customer Lifetime Value (CLV)»

Insurance Data Science

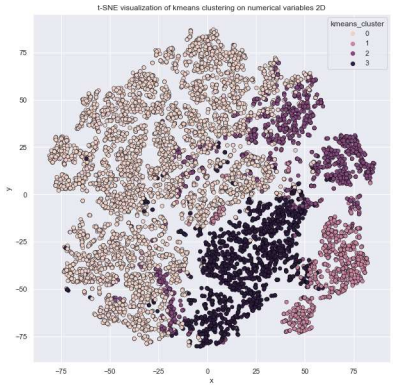
Clustering as portfolio analysis

Improve portfolio analysis by the combination of t-SNE and Clustering tools on numerical features

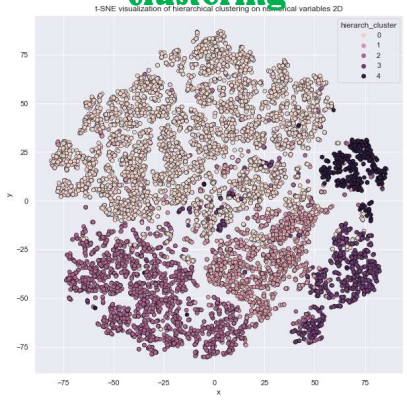
t-SNE with GMM



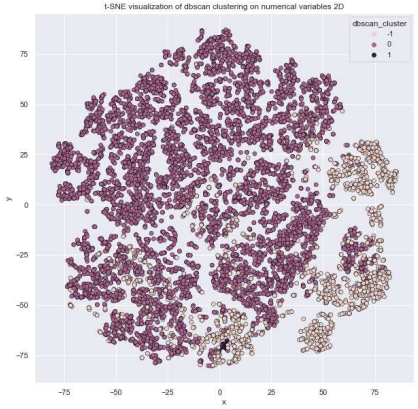
t-SNE with K-Means



t-SNE with Hierarchical Clustering



t-SNE with DBSCAN



Visualization as helpful tool driving porftolio queries

K-Means cluster analysis shows «0» cluster as region with greater density and with greater percentage both of Claim Amount and Monthly Car Premium (around 50%)

kmeans_cluster	Monthly_Premium_Auto	PERCENTAGE
0	454,824.00	53.42
2	187,154.00	21.98
3	141,777.00	16.65
1	67,710.00	7.95

kmeans_cluster	Total_Claim_Amount	PERCENTAGE
0	1,891,867.85	47.71
2	1,151,870.66	29.05
3	621,495.02	15.67
1	299,733.52	7.56

From silhouette score comparison of clustering tools, K-Means seems to be more suitable

	labels	silhouette_score
kmeans	4	0.175795
dbscan	3	0.127336
gmm	2	0.138758
hierarchical_clustering	5	0.159866