Predicting Sepsis on the PredictEM Masked Blutkulturen Dataset

Bachelor Thesis by Yannick Müller

Preprocessing Dataset

1. Sorted by Hospital Entrance Date
2. Factorized all text entries in the dataset
3. Deleted rows where target value was NaN
4. More than 3 Std delete on train set (Input Tobia: Look at Average outliner?)
5. Filled missing values with -1 (Meeting with Julia)
6. Applied Z-Score on dataset
7. Split Set into Train and Test set: Train set the first 80% time-ordered (delete patients in different sets)

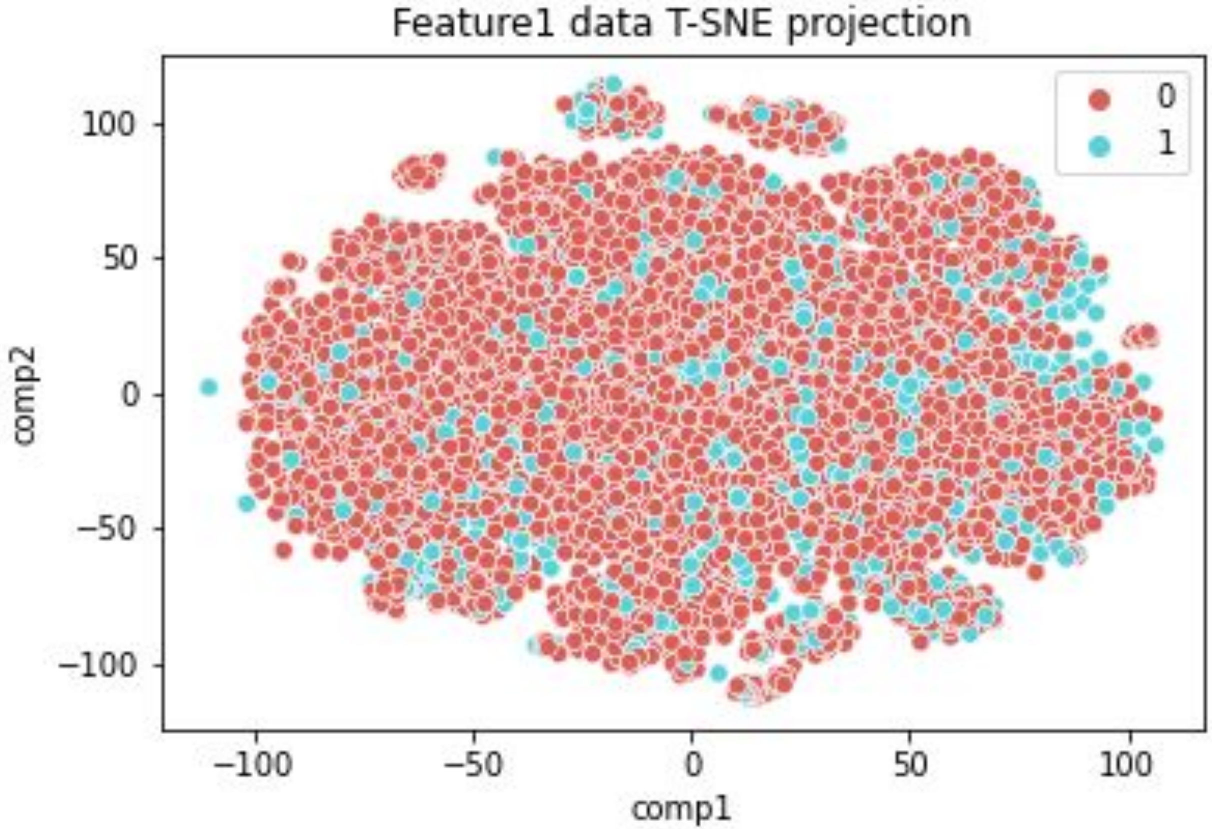
Finding most relevant features

1. Using the correlation matrix I found the features which are most relevant for prediction
2. Using Random Forest I deleted features which made the prediction better or the least worst, then added a feature again which made it the most better
3. Discussions with Oliver what features make sense from hospital point of view
4. Recursive Feature Deletion/Addition with Random Forest

Top 42 Features

Age_admission_unz 0.08, sex 0.01, triage 0.05, referral_unz 0.03, ort_vor_aufnahme_spital 0.06, time_unz 0.0, Admission_choice 0.04, pulse_first 0.05, frequency_first 0.04, systolic_bp_first 0.02, systolic_bp_first_inf_100 0.05, diastolic_bp_first 0.04, respiratory_rate_first 0.08, gcs_inf_15_first 0.02, lvl_consc_alert 0.07, temperature_highest 0.0, temperature_lowest 0.01, spo2_first 0.02, o2_gabe 0.05, THZ (Thrombozyten) 0.09, NA 0.01, ASAT 0.05, UREA = Harnstoff 0.13, GFR 0.04, PCThs 0.02, EOS =eosinophile 0.0, CRP 0.15, KA = Kalium 0.0, Hb 0.05, ALAT 0.04, INRiH 0.07, CR 0.10, pH 0.13, LACT =lactat 0.17, GGT 0.13, BIC_st 0.13, BIC_VB 0.11, Leuk 0.02, leuk_sup_10 0.02, leuk_inf_4 0.05, leuk_inf_0_5 0.06, entrance_unz_date 0.01

Top 42 Features



Top 42 Features on predicting blood culture positive

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.21	0.34	0.16	0.83
XG Boost	0.31	0.33	0.28	0.81
Linear Reg.	0.33	0.22	0.57	0.65
Decision Tree	0.31	0.32	0.30	0.80
SVC RBF	0.29	0.24	0.36	0.74
Random Forest	0.26	0.38	0.20	0.83
MLP Classifier	0.31	0.34	0.28	0.81
Lasso CV	0.08	0.41	0.04	0.85
Ridge Class	0.33	0.23	0.57	0.65

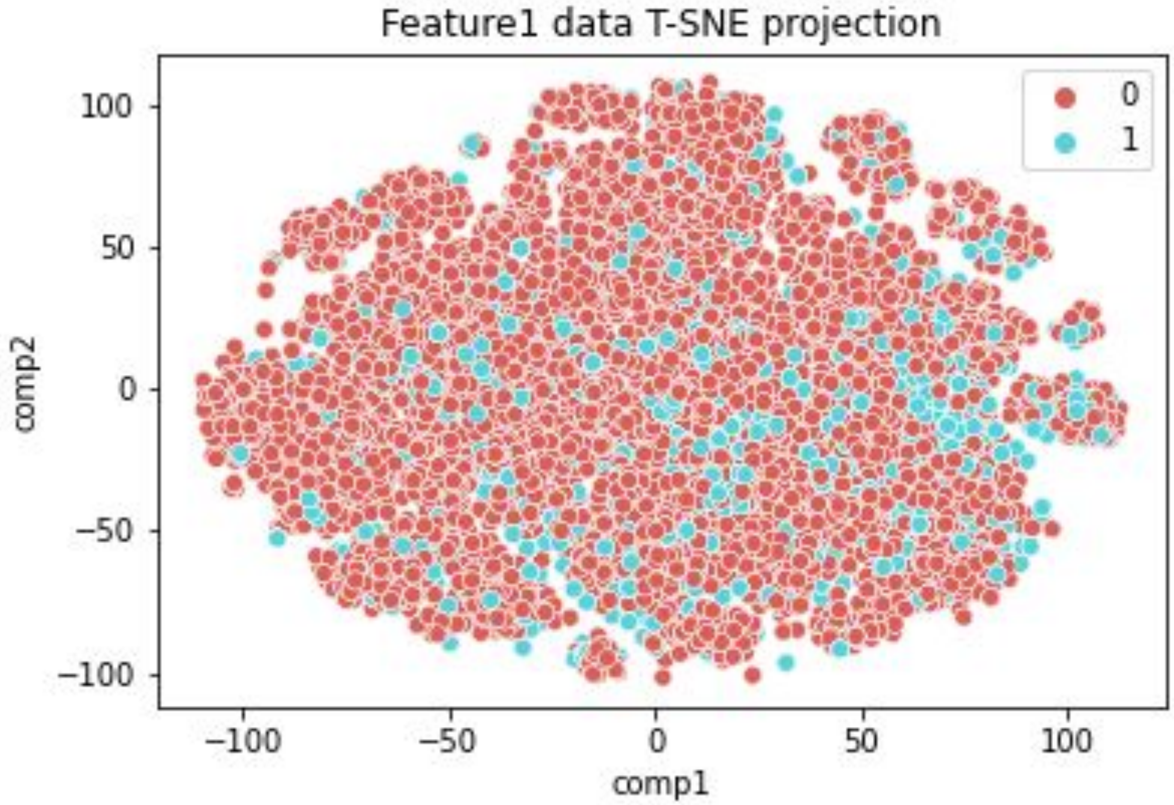
Top 42 Features on predicting 28 days mortality

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.32	0.84	0.10	0.93
XG Boost	0.27	0.37	0.21	0.91
Linear Reg.	0.0	0.0	0.0	0.92
Decision Tree	0.17	0.17	0.17	0.88
SVC RBF	0.07	0.44	0.03	0.89
Random Forest	0.18	0.37	0.12	0.92
MLP Classifier	0.27	0.31	0.24	0.92
Lasso CV	0.0	0.0	0.0	0.90
Ridge Reg	0.0	0.0	0.0	0.92

Top 32 Features

Age_admission_unz 0.08, sex 0.01, triage 0.05, referral_unz 0.03, ort_vor_aufnahme_spital 0.06, Admission_choice 0.04, pulse_first 0.05, frequency_first 0.04, systolic_bp_first 0.02, diastolic_bp_first 0.04, respiratory_rate_first 0.08, gcs_inf_15_first 0.02, lvl_consc_alert 0.07, temperature_highest 0.0, spo2_first 0.02, o2_gabe 0.05, THZ (Thrombozyten) 0.09, ASAT 0.05, UREA = Harnstoff 0.13, GFR 0.04, PCThs 0.02, CRP 0.15, Hb 0.05, ALAT 0.04, INRiH 0.07, CR 0.10, pH 0.13, LACT =lactat 0.17, GGT 0.13, BIC_st 0.13, BIC_VB 0.11, Leuk 0.02

Top 32 Features



Top 32 Features

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.24	0.40	0.17	0.84
XG Boost	0.33	0.34	0.32	0.81
Linear Reg.	0.29	0.20	0.52	0.63
Decision Tree	0.25	0.27	0.24	0.79
SVC RBF	0.29	0.24	0.37	0.73
Random Forest	0.30	0.43	0.23	0.84
MLP Classifier	0.20	0.24	0.20	0.79
Lasso CV	0.05	0.43	0.03	0.85
Ridge Class	0.29	0.20	0.52	0.63

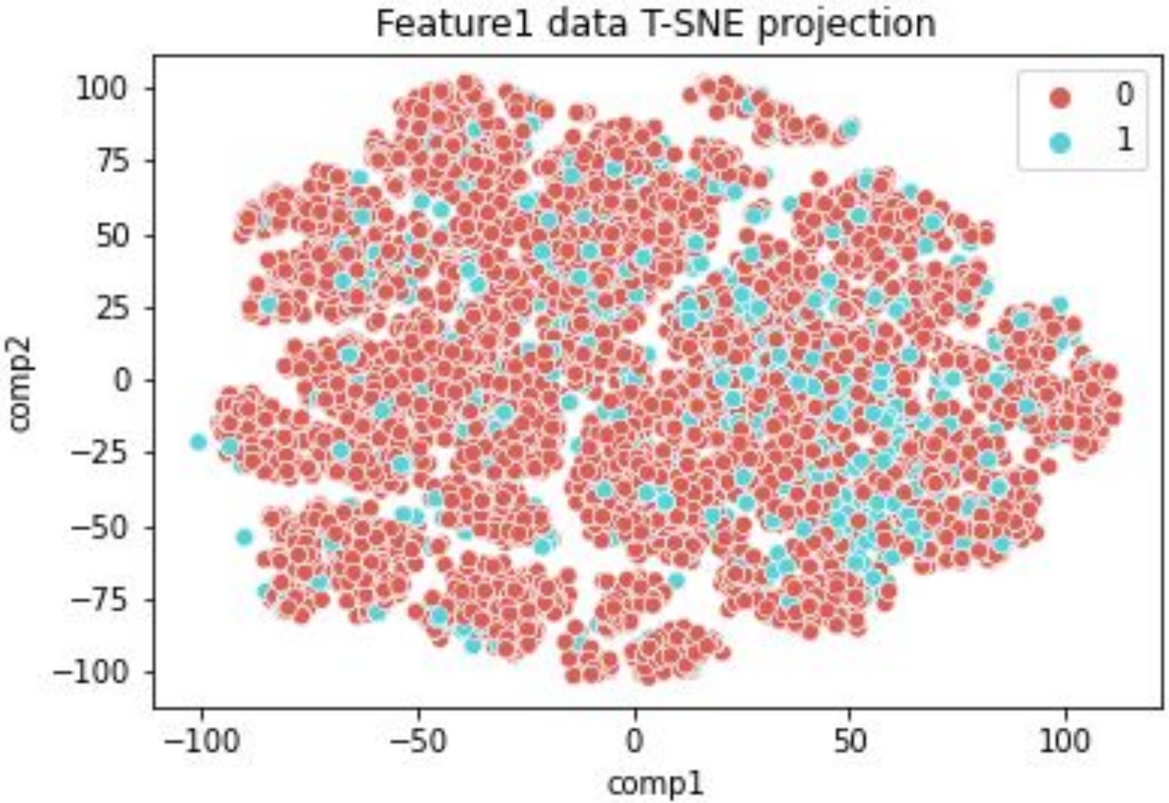
Top 32 Features on predicting 28 days mortality

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.17	0.45	0.11	0.92
XG Boost	0.23	0.33	0.18	0.91
Linear Reg.	0.0	0.0	0.0	0.92
Decision Tree	0.17	0.17	0.16	0.88
SVC RBF	0.11	0.67	0.06	0.93
Random Forest	0.20	0.44	0.13	0.92
MLP Classifier	0.20	0.24	0.17	0.90
Lasso CV	0.0	0.0	0.0	0.92
Ridge Reg	0.0	0.0	0.0	0.92

Top 16 Features

Age_admission_unz	0.08	Hb	0.05
sex	0.01	INRiH	0.07
respiratory_rate_first	0.08	CR	0.10
lvl_consc_alert	0.07	pH	0.13
o2_gabe	0.05	LACT =lactat	0.17
THZ (Thrombozyten)	0.09	GGT	0.13
UREA = Harnstoff	0.13	BIC_st	0.13
CRP	0.15	BIC_VB	0.11

Top 16 Features



Top 16 Features

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.22	0.45	0.15	0.84
XG Boost	0.27	0.29	0.26	0.79
Linear Reg.	0.30	0.21	0.50	0.65
Decision Tree	0.18	0.18	0.18	0.76
SVC RBF	0.31	0.25	0.40	0.73
Random Forest	0.26	0.34	0.21	0.82
MLP Classifier	0.22	0.25	0.19	0.79
Lasso CV	0.05	0.43	0.03	0.85
Ridge Class	0.30	0.21	0.50	0.65

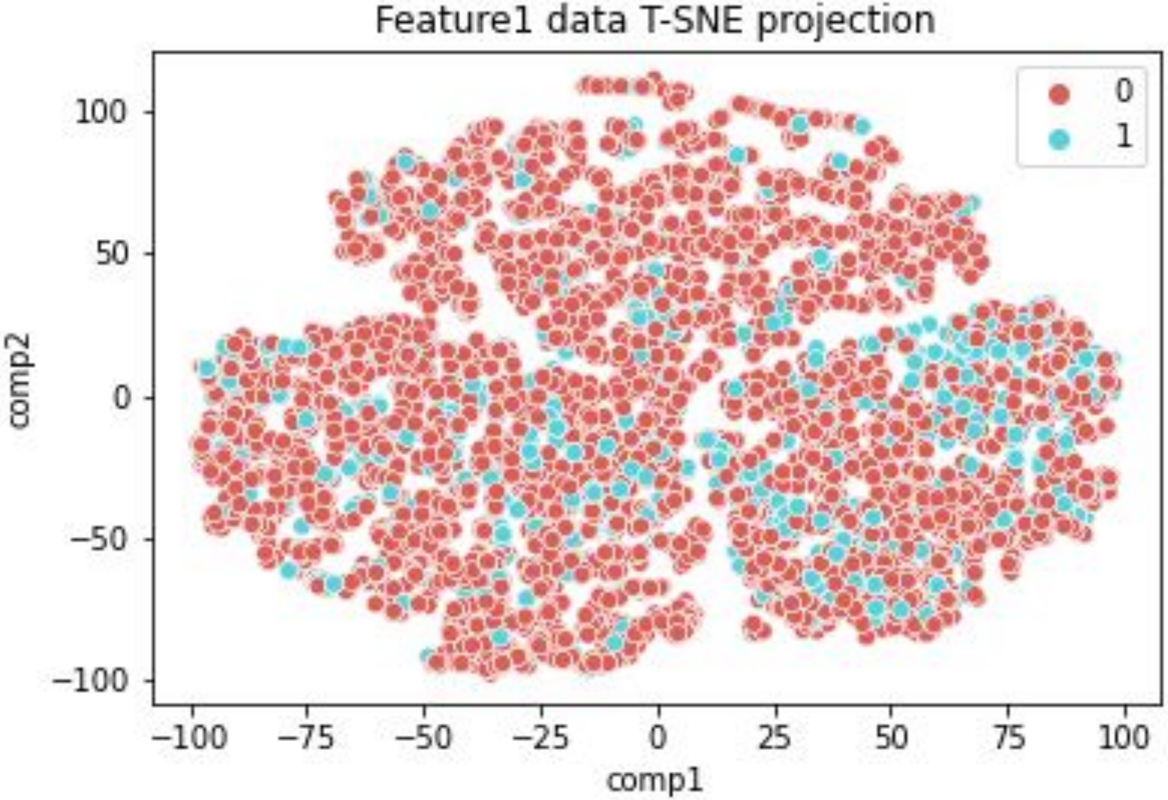
Top 16 Features on predicting 28 days mortality

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.19	0.44	0.13	0.92
XG Boost	0.28	0.36	0.24	0.91
Linear Reg.	0.0	0.0	0.0	0.92
Decision Tree	0.22	0.24	0.21	0.89
SVC RBF	0.05	0.33	0.03	0.92
Random Forest	0.22	0.47	0.15	0.92
MLP Classifier	0.24	0.29	0.22	0.90
Lasso CV	0.0	0.0	0.0	0.92
Ridge Reg	0.0	0.0	0.0	0.92

Top 6

Age_admission_unz	0.08
THZ (Thrombozyten)	0.09
CRP	0.15
LACT =lactat	0.17
BIC_st	0.13
ASAT	0.05

Top 6 Features



Top 6 Features

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.18	0.40	0.12	0.84
XG Boost	0.25	0.27	0.23	0.79
Linear Reg.	0.30	0.21	0.51	0.64
Decision Tree	0.18	0.19	0.17	0.77
SVC RBF	0.30	0.23	0.45	0.69
Random Forest	0.23	0.26	0.21	0.80
MLP Classifier	0.21	0.37	0.15	0.84
Lasso CV	0.05	0.68	0.03	0.85
Ridge Class	0.30	0.21	0.51	0.64

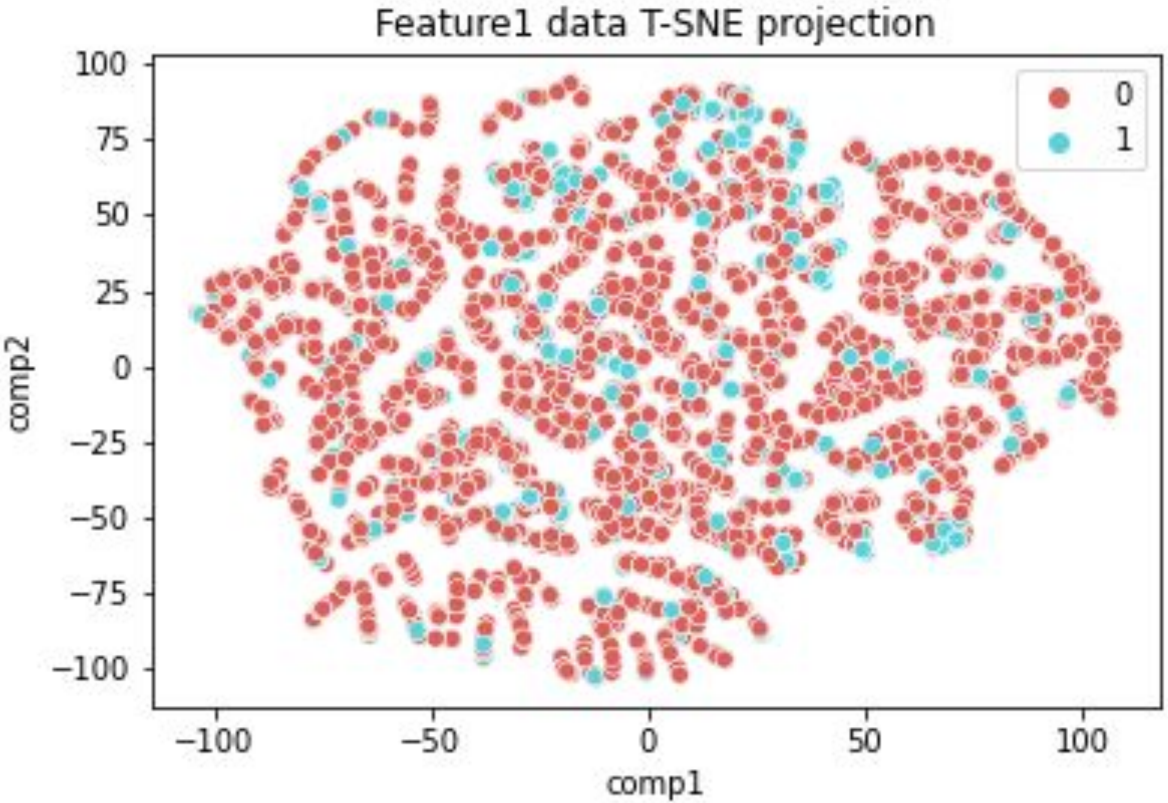
Top 6 Features on predicting 28 days mortality

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.08	0.31	0.05	0.92
XG Boost	0.15	0.24	0.11	0.91
Linear Reg.	0.0	0.0	0.0	0.92
Decision Tree	0.26	0.25	0.27	0.88
SVC RBF	0.0	0.0	0.0	0.92
Random Forest	0.23	0.37	0.16	0.92
MLP Classifier	0.15	0.51	0.09	0.92
Lasso CV	0.0	0.0	0.0	0.92
Ridge Reg	0.0	0.0	0.0	0.92

Current Features Doctors use

Age_admission_unz	0.08
CRP	0.15
PCThs	0.02

Features: Age, CRP, PCT



Features: Age, CRP, PCT

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.01	1.0	0.01	0.85
XG Boost	0.16	0.26	0.11	0.82
Linear Reg.	0.26	0.18	0.52	0.71
Decision Tree	0.17	0.19	0.15	0.78
SVC RBF	0.27	0.18	0.52	0.58
Random Forest	0.20	0.19	0.22	0.75
MLP Classifier	0.02	0.20	0.01	0.85
Lasso CV	0.0	0.0	0.0	0.85
Ridge Class	0.26	0.18	0.52	0.57

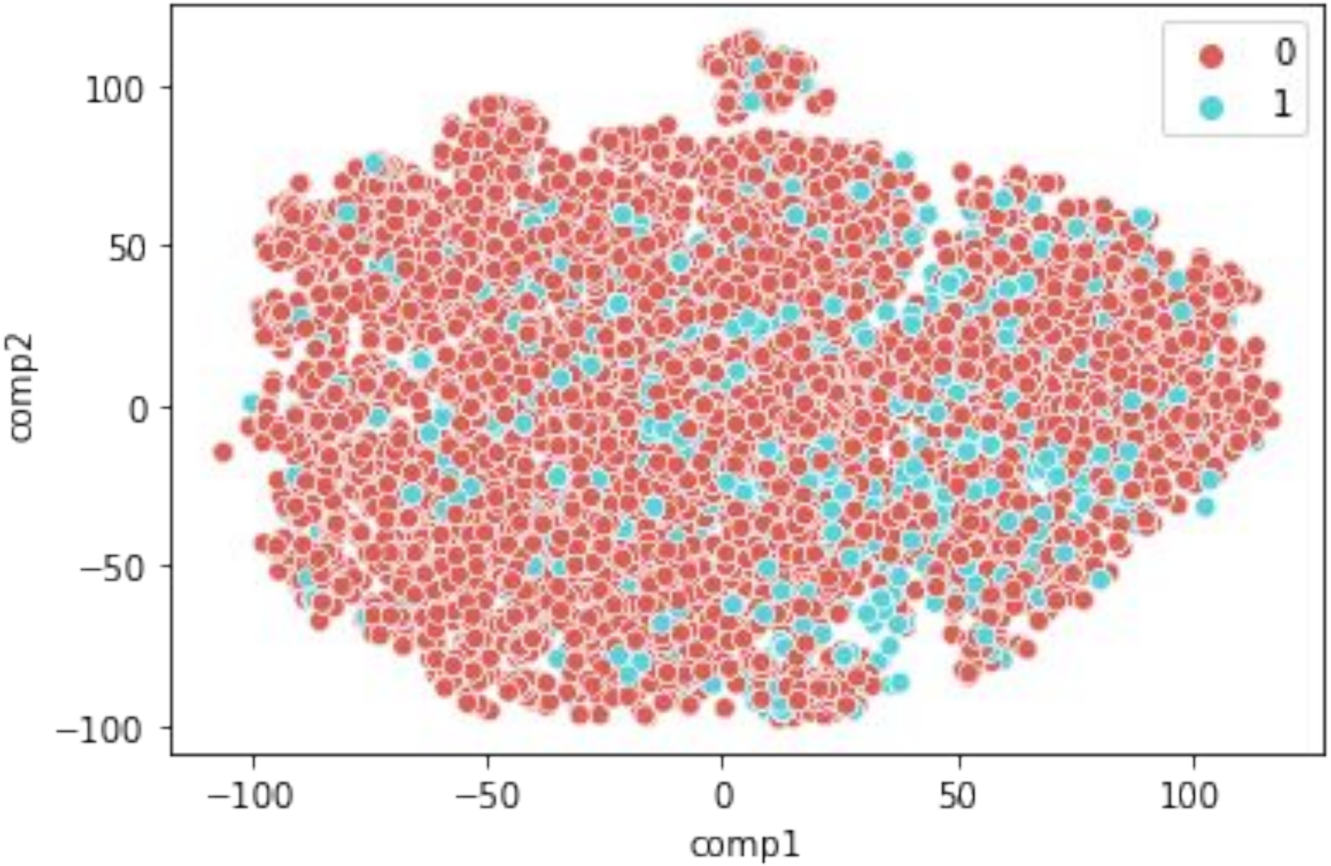
Top 3 Features on predicting 28 days mortality

Test Set	F1 Score	Precision	Recall	Accuracy
Dense NN	0.0	0.0	0.0	0.92
XG Boost	0.12	0.29	0.08	0.92
Linear Reg.	0.0	0.0	0.0	0.92
Decision Tree	0.16	0.16	0.15	0.88
SVC RBF	0.0	0.0	0.0	0.92
Random Forest	0.16	0.17	0.16	0.88
MLP Classifier	0.0	0.0	0.0	0.92
Lasso CV	0.0	0.0	0.0	0.92
Ridge Reg	0.0	0.0	0.0	0.92

Improved 26 Features

LACT =lactat, LACT =lactat inf quant 0.7, CRP, GGT, GGT inf median, INRiH inf median, INRiH, THZ (Thrombozyten), THZ (Thrombozyten) inf quant 0.3, ASAT inf quant 0.3, UREA = Harnstoff, UREA = Harnstoff inf quant 0.3, ort_vor_aufnahme_spital: anderes Spital, BIC_VB, BIC_st, BIC_st inf quant 0.1, lvl_consc_alert, pH inf quant 0.1, CR inf quant 0.8, EOS =eosinophile inf quant 0.6, age_admission_unz, diastolic_bp_first inf quant 0.4, respiratory_rate_first inf quant 0.4, systolic_bp_first inf quant 0.4, Hb, PCThs

T-SNE



Test Set	F1 Score	Precision	Recall	ROC	Accuracy
Dense NN	0.32	0.26	0.39	0.60	0.75
XG Boost	0.22	0.26	0.19	0.55	0.8
Linear Reg.	0.34	0.25	0.57	0.63	0.68
Decision Tree	0.26	0.26	0.27	0.57	0.78
SVC RBF	0.24	0.21	0.28	0.55	0.74
Random Forest	0.11	0.27	0.07	0.52	0.84
MLP Classifier	0.22	0.25	0.18	0.55	0.80
Lasso CV	0.01	0.1	0.0	0.5	0.85
Ridge Classifier	0.34	0.24	0.57	0.63	0.68
Gaussian Process Aprox.	0.35	0.23	0.69	0.64	0.61
Naive Bayes	0.32	0.23	0.52	0.61	0.67
SGD Classifier	0.25	0.16	0.50	0.53	0.55
Bayesian Ridge	0.35	0.25	0.58	0.64	0.68

Learnings

1. Once accidently had shuffle=True in the preprocessing, so data was not time-ordered for the train/test sets

   => RandomForest had F1 Score 0.8, Precision 0.8, Recall 0.8 and Accuracy 0.93

   If shuffle=False and data is time-ordered

   => RandomForest has F1 Score 0.23, Precision 0.26, Recall 0.21 and Accuracy 0.80

   Shows how fast models can learn from similar predictions.

Up Next

1. Introduce more boolean values
2. Polynomial functions on features
3. More advanced neural network
4. Feature Selection
5. Fourier Transform
6. ROC-curve / Area under the Curve
7. Pretrained Net for Medical Data and then fine tune on this dataset
8. Hyperparameter optimization
9. Ensemble machen
10. Sklearn GridSearchCV
11. Difference One-Hot-Encoding and Factorization

Ideas

1. Autoencoder. Give Output to Linear Reg
2. PCA. Give output to Linear Reg
3. Concentrate and optimize Dense NN, XGBoost, LinearReg, SVC, Ridge Classifier, Gaussian Process, Bayesian Ridge

   Don’t put anymore work into the other methods and just use them as validation

Questions

1. Why are the inf_0_5 and sup_0_7 beneficial for data analytics. Is this not just a simplification and valuable information gets lost?
2. Any more ideas for feature tuning?
3. Is the Z-Score better than normalization for medical data?
4. Similar Project/connections to exchange more knowledge?
5. Why does the class weight adjustments lower the F1-score?
6. Did this presentation fulfil your expectations? What should be improved next time?