Data Science Advanced

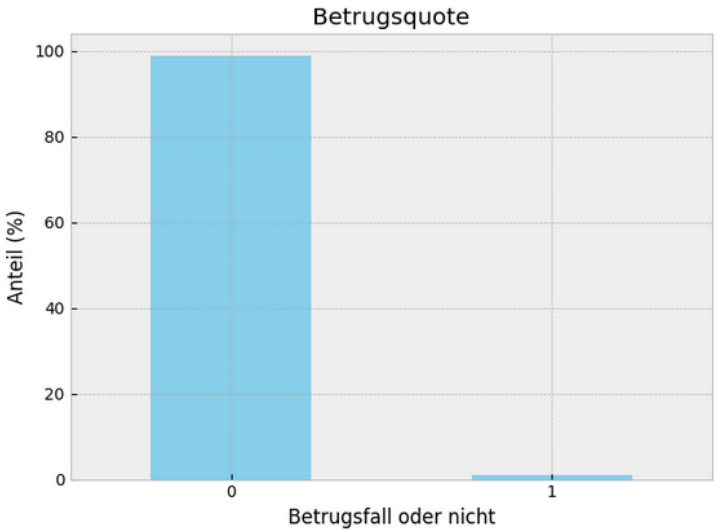
Bank Fraud Classification Project

Marc-Vincent Müller

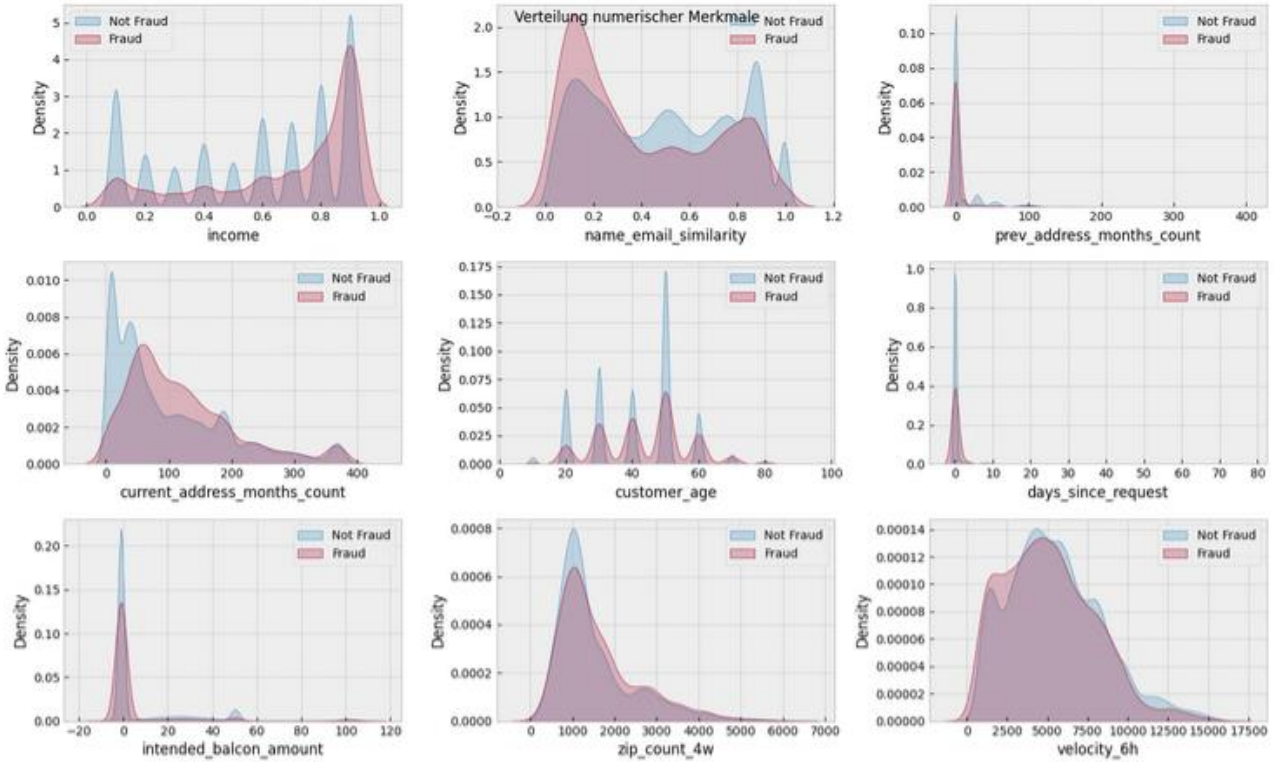
12.12.2025

Data Science Advanced – Bank Fraud

Data set & analysis




```
Prozentuale Verteilung: fraud_bool
0    98.897
1     1.103
Name: proportion, dtype: float64
```



Highly unevenly distributed data, increase of fraud with higher income

Data Science Advanced – Bank Fraud

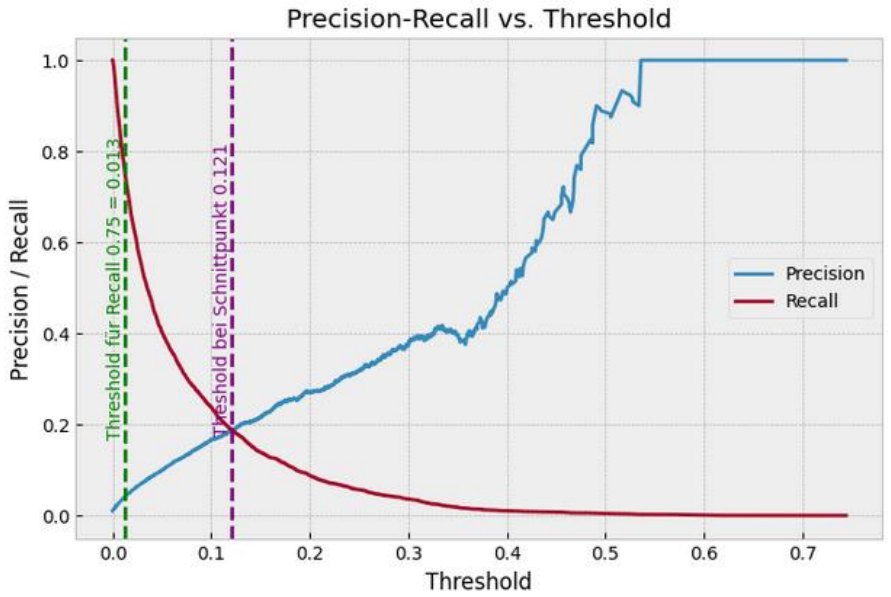
Steps

1. Drop x1 and x2
2. Search missing values
3. Drop columns or refill values with imputation
4. One hot encoding of categorial features
5. Scaling all features
6. Feature selection (only log regression)
7. Define models
8. Data reduction to 10 % and train-test-split
9. Gridsearch (only log regression and decision trees) → Optimization Score: ROC-AUC
10. Fit best model with best hyperparameters on all train data
11. Model evaluation → Precision-Recall-Curve: Threshold optimization on at least 75 % Recall on test data
12. Visualisation with Confusion Matrix
13. Comparison of results on train and test data

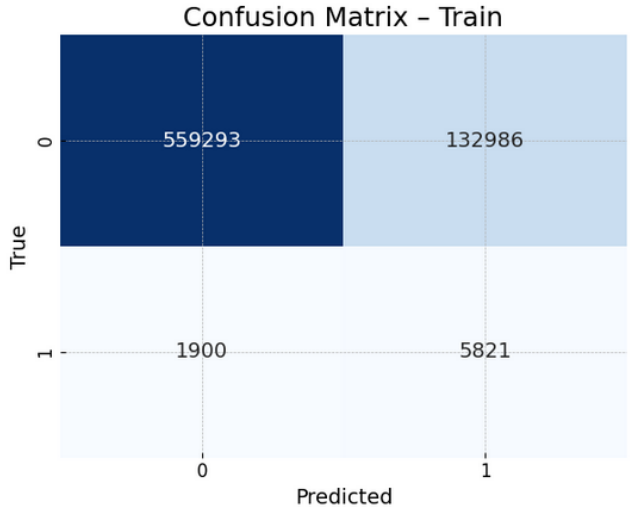
First optimization on ROC-AUC score, then threshold adjustment on at least 75 % Recall

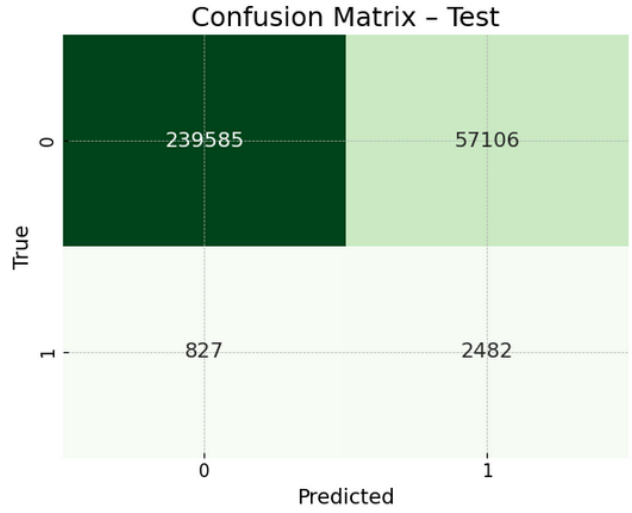
Data Science Advanced – Bank Fraud

Model – Logistic Regression



Threshold bei Recall ~75 %: 0.012567435199216 46
Threshold bei Schnittpunkt: 0.12129834878162572





--- TRAIN SCORES ---
ROC-AUC: 0.8619624836621834

	precision	recall	f1-score	support
0	1.00	0.81	0.89	692279
1	0.04	0.75	0.08	7721
accuracy			0.81	700000
macro avg	0.52	0.78	0.49	700000
weighted avg	0.99	0.81	0.88	700000

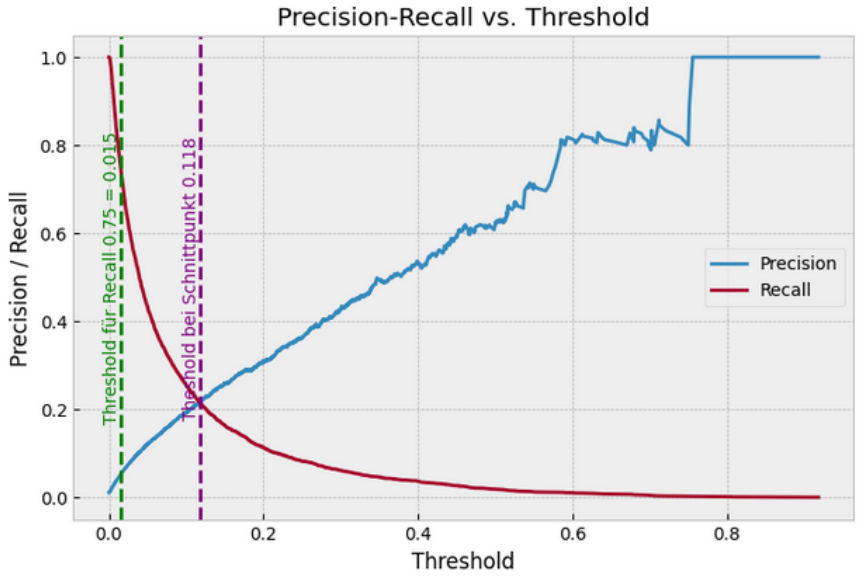
--- TEST SCORES ---
ROC-AUC: 0.8579801606399762

	precision	recall	f1-score	support
0	1.00	0.81	0.89	296691
1	0.04	0.75	0.08	3309
accuracy			0.81	300000
macro avg	0.52	0.78	0.49	300000
weighted avg	0.99	0.81	0.88	300000

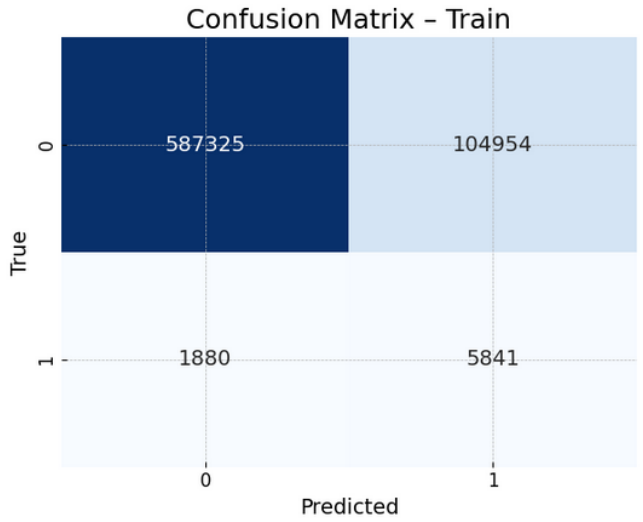
Without SMOTE, ridge regression, class weight: None, C: 3

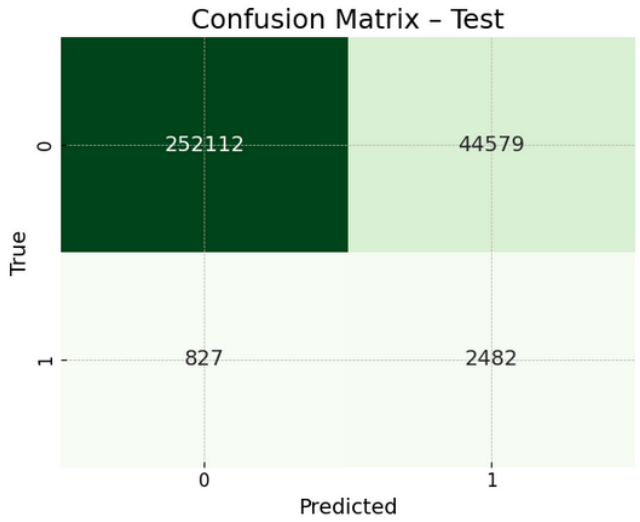
Data Science Advanced – Bank Fraud

Model – Decision Tree



Threshold bei Recall ~75 %: 0.01521433988097543
Threshold bei Schnittpunkt: 0.11760161799233325





--- TRAIN SCORES ---
ROC-AUC: 0.885528925297161

	precision	recall	f1-score	support
0	1.00	0.85	0.92	692279
1	0.05	0.76	0.10	7721
accuracy			0.85	700000
macro avg	0.52	0.80	0.51	700000
weighted avg	0.99	0.85	0.91	700000

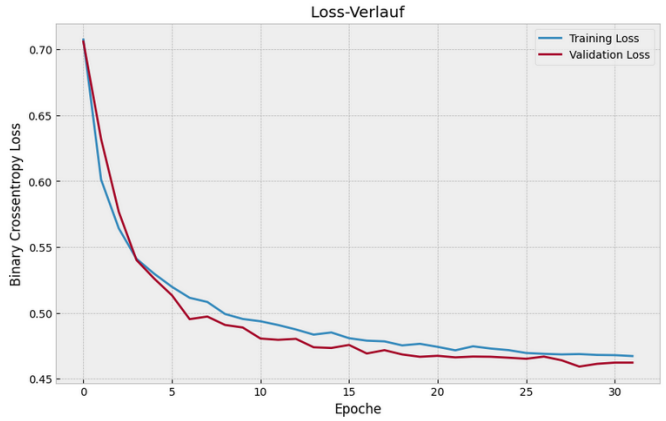
--- TEST SCORES ---
ROC-AUC: 0.8841483464497155

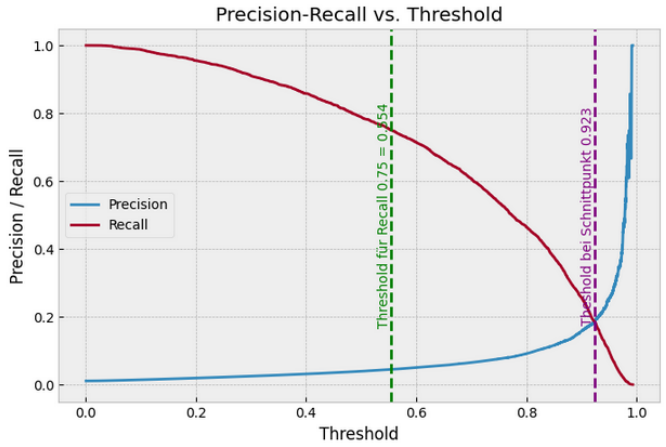
	precision	recall	f1-score	support
0	1.00	0.85	0.92	296691
1	0.05	0.75	0.10	3309
accuracy			0.85	300000
macro avg	0.52	0.80	0.51	300000
weighted avg	0.99	0.85	0.91	300000

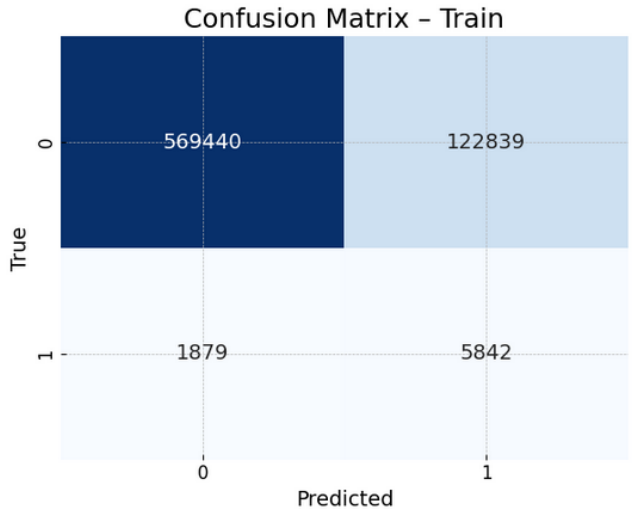
Best tree: LightGBM, max depth: 3, n estimators: 300, num leaves: 3, scale pos weight: 1, learning rate: 1

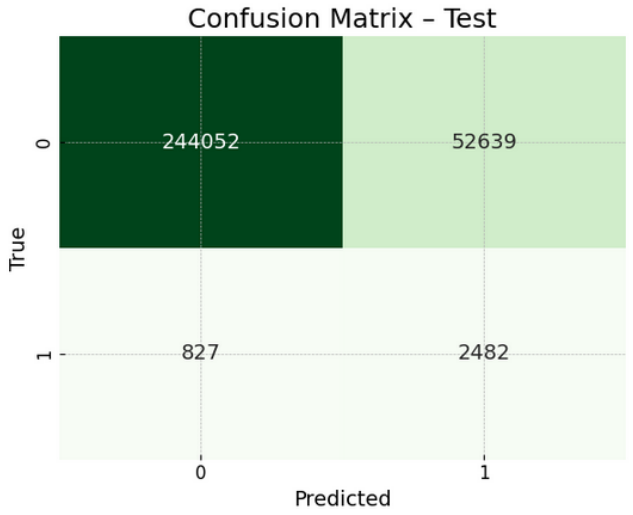
Data Science Advanced – Bank Fraud

Model – Neural Network









--- TRAIN SCORES ---
ROC-AUC: 0.8703571476891511

	precision	recall	f1-score	support
0	1.00	0.82	0.90	692279
1	0.05	0.76	0.09	7721
accuracy			0.82	700000
macro avg	0.52	0.79	0.49	700000
weighted avg	0.99	0.82	0.89	700000

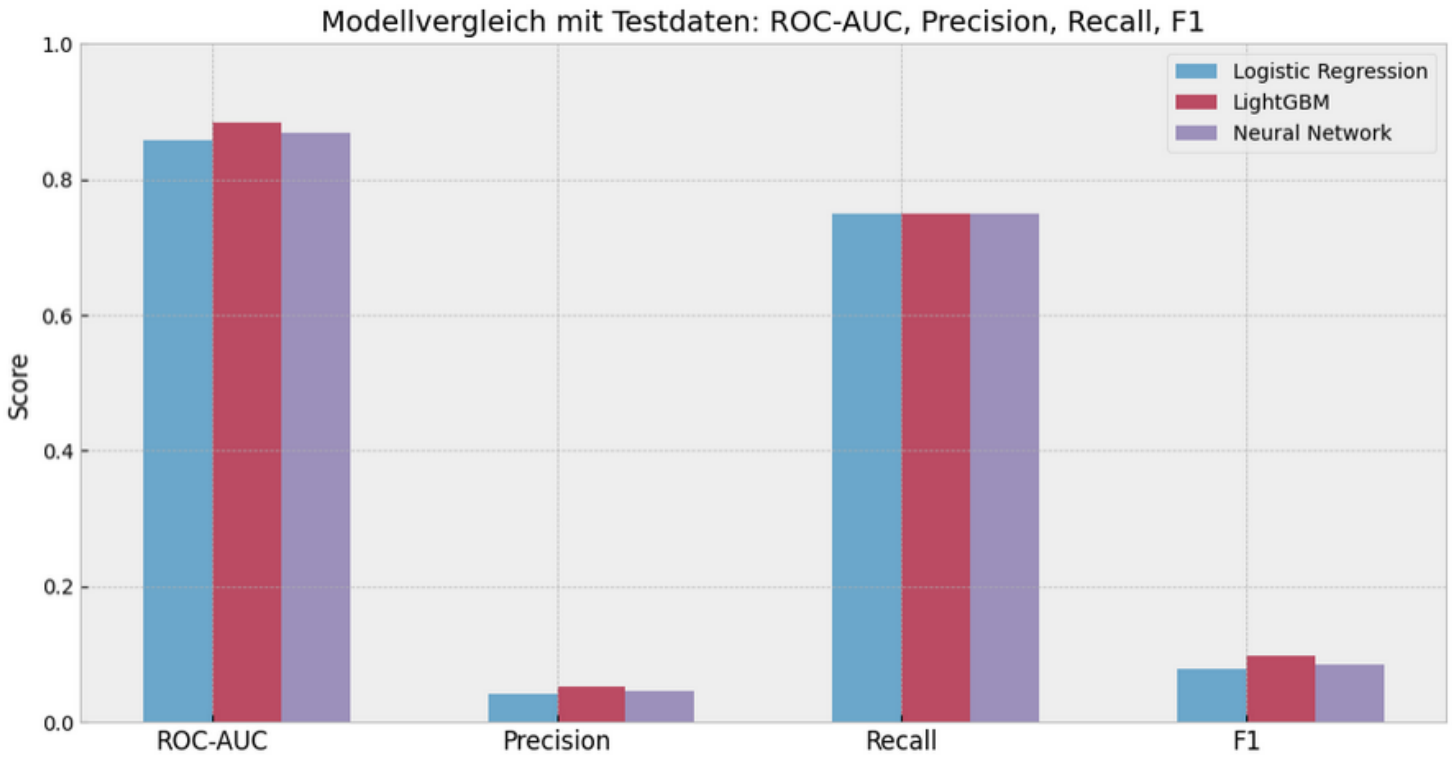
--- TEST SCORES ---
ROC-AUC: 0.8687170492853082

	precision	recall	f1-score	support
0	1.00	0.82	0.90	296691
1	0.05	0.75	0.08	3309
accuracy			0.82	300000
macro avg	0.52	0.79	0.49	300000
weighted avg	0.99	0.82	0.89	300000

2 intermediate layers, ReLU activation , regularisation: Dropout against overfitting, early stopping

Data Science Advanced – Bank Fraud

Model – Comparison and summary



Very similar scores regardless of the model, best results with LightGBM, difficult fraud detection due to uneven distrubtion, Additional features required to improve model (for example x1 and x2)