Auto-Encoder Reconstruction Cost for Anomalies Detection scenarios

Comparing Auto-Encoder approach for Anomalies Detection to traditional Upsampling + Classification techniques

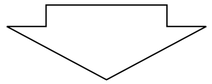
Dataset

Credit Card Fraud Detection
Anonymized credit card transactions labeled as fraudulent or genuine
https://www.kaggle.com/mlg-ulb/creditcardfraud

2 days real European credit cards transactions: **492 frauds out of 284,807** (**0.172%**)

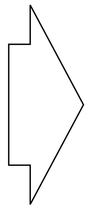
RAW

- 28 Numerical input variables which are the result of a PCA transformation to anonymize customer info
- Transaction Amount
- Time in seconds from T0
- Label Class: 0/1 (1 is the anomaly)



Transform

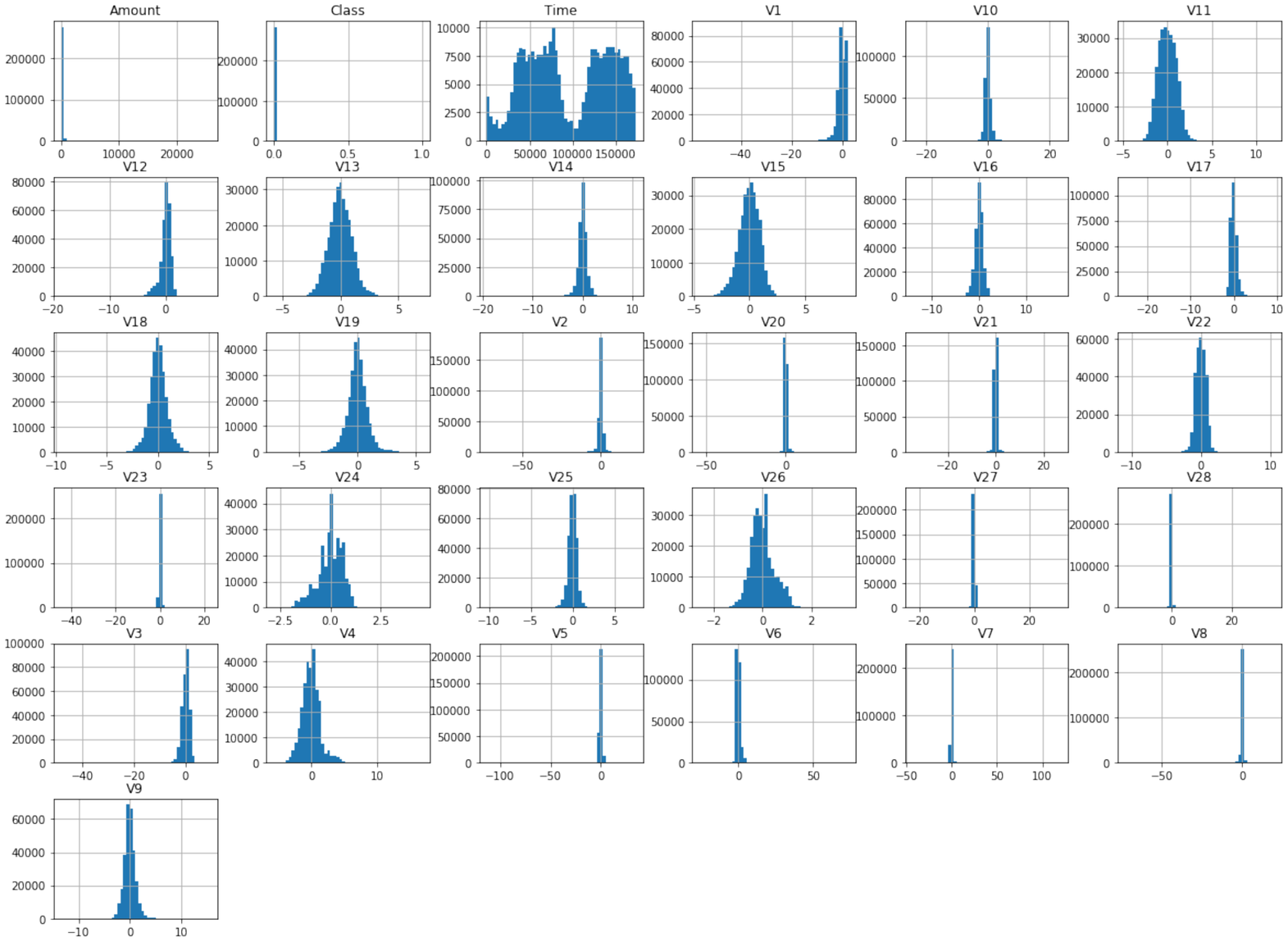
- 29 Numerical Features
  - Original 28 from PCA
  - Transaction Amount (StandardScaler)
- 1 Categorical Feature
  - Time of the transaction using 24 Hour beens



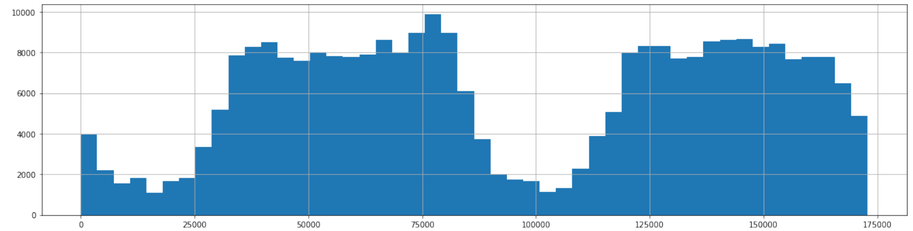
Multi-Input

- 29 Numerical Features -> Tensor Length 29
- 1 Categorical Feature values 0-24 —> Embedding Tensor 12

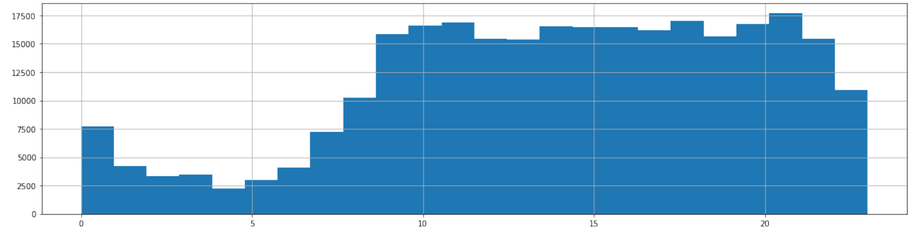
Features Hist



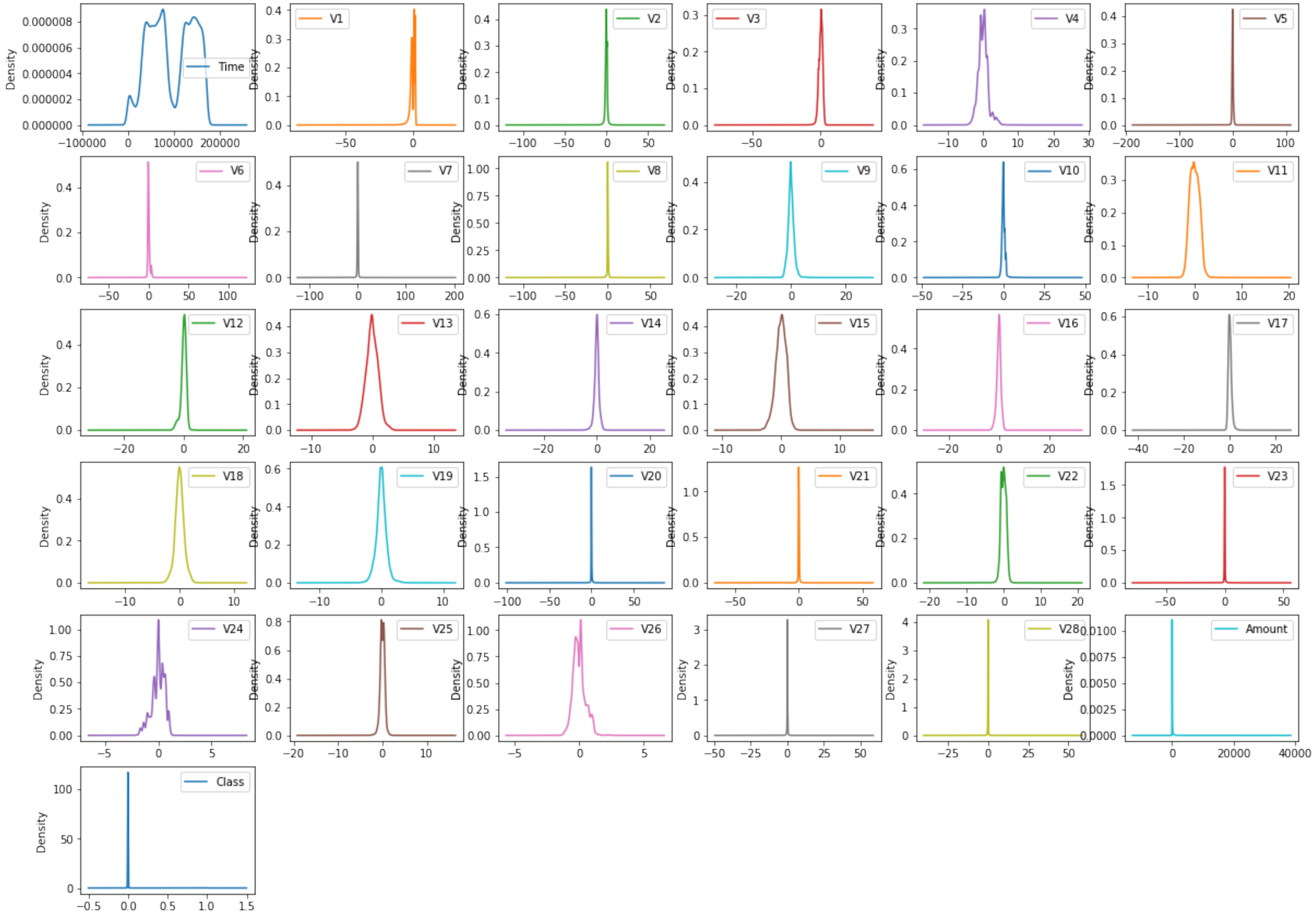
Time in 48 total hours



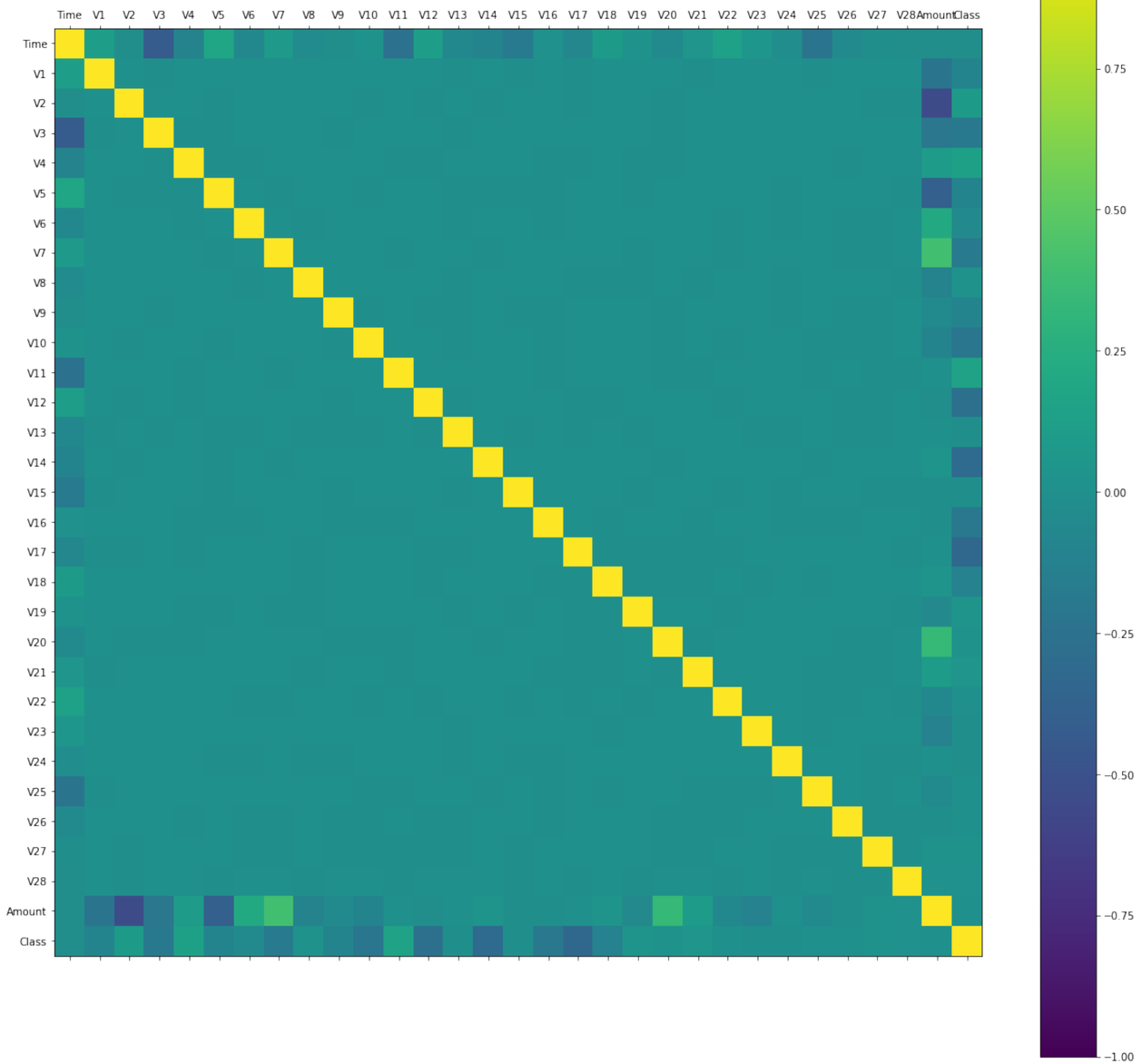
Time in 24 hours bin



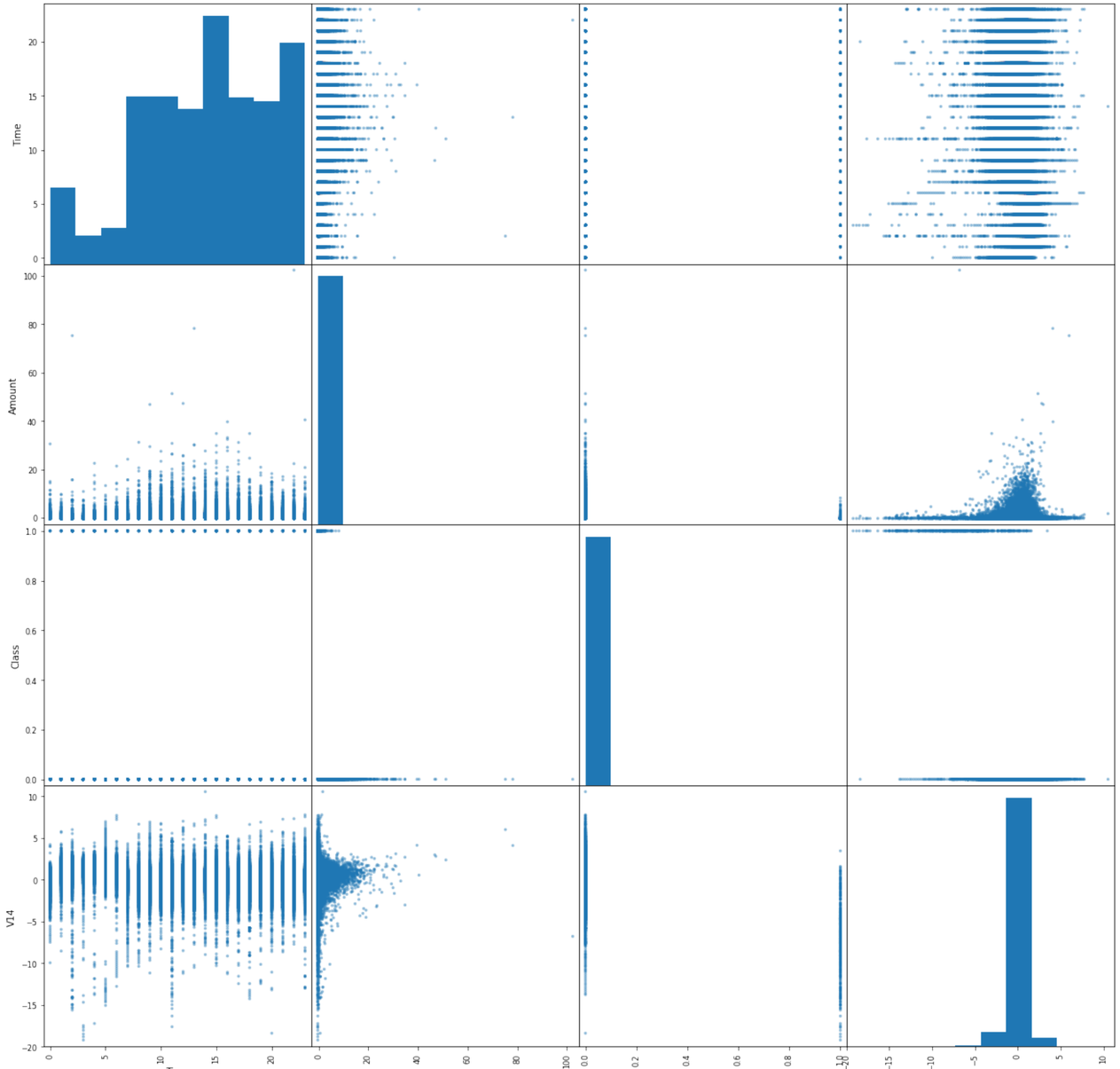
Features Density



Correlation



Correlation Matrix



Training Approach

Classifier

- Feed 80% of both anomaly (+) and non-anomaly (-) transactions
- Upscale the anomaly transactions in the train dataset
- Handle Multi-Input for Numeric features and Categories
- Use Embedding for Categories
- Build a classic DNN classifier

Auto-Encoder

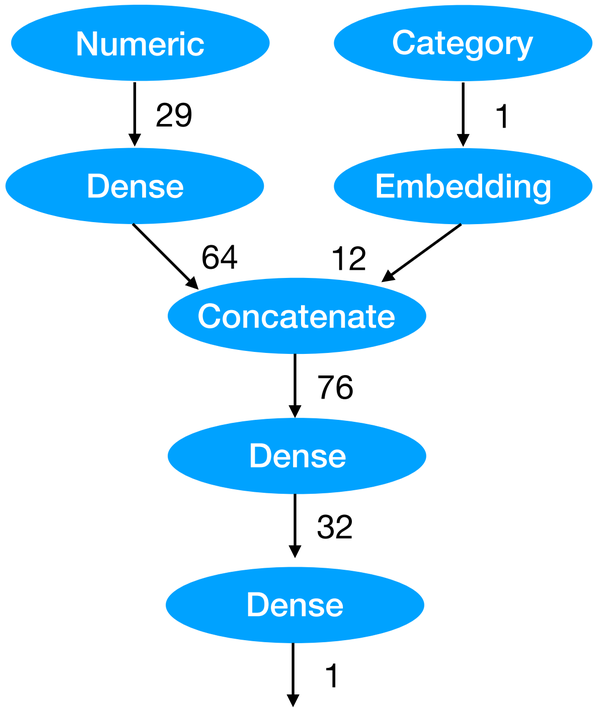
- Feed 80% of only the non-anomaly (-) transactions
- Handle Multi-Input for Numeric features and Categories
- Use Embedding for Categories
- Use an Auto-Encoder to learn compression/ decompression on non-anomaly transactions
- Use a custom layer to calculate the reconstruction cost
- Train to reduce this cost to 0

Models

Feed 80% (+) (-)

TRAINING

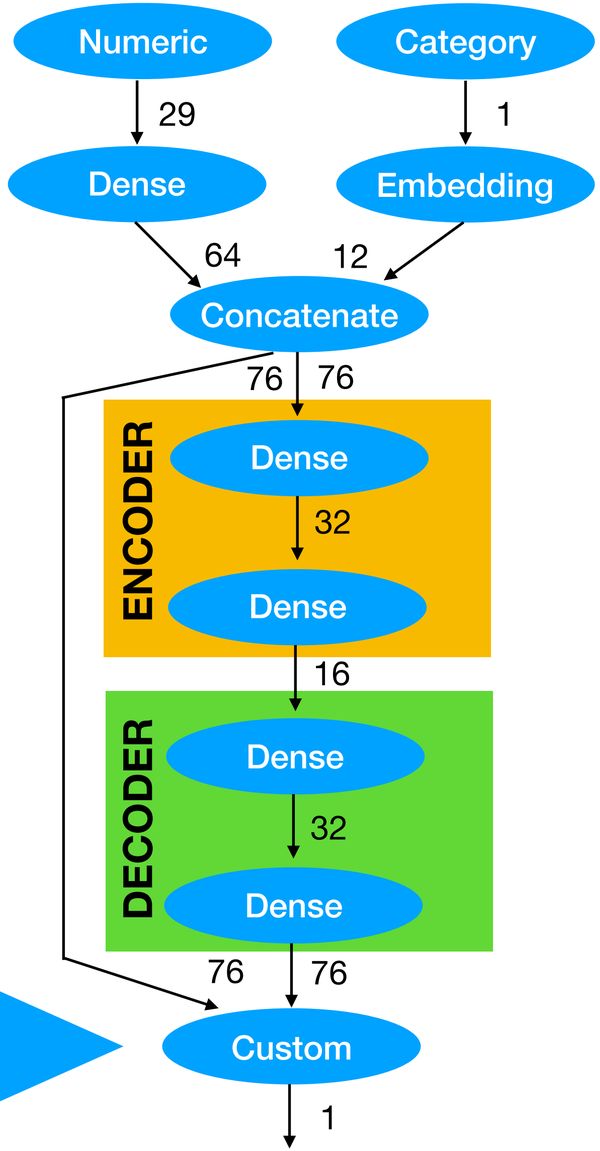
Classifier



Feed 80% only (-)

TRAINING

Auto-Encoder



Reconstruction Cost

$\sum (Encoder\_Input - Decoder\_Output)^2$

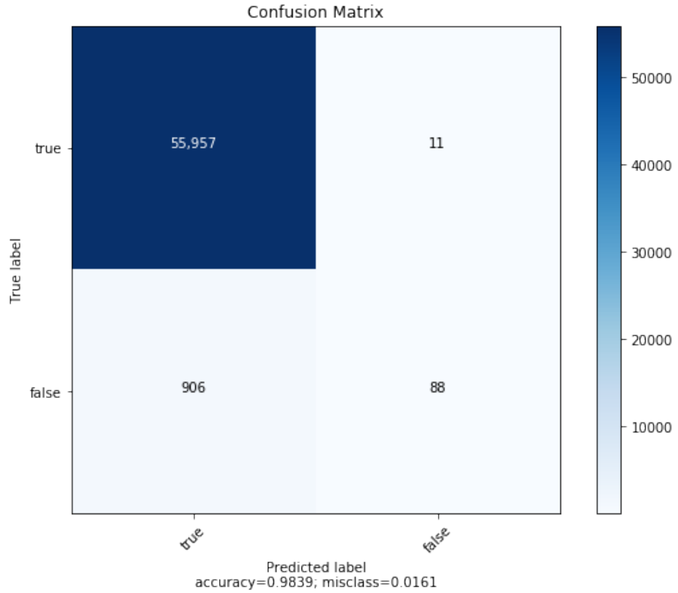
Results

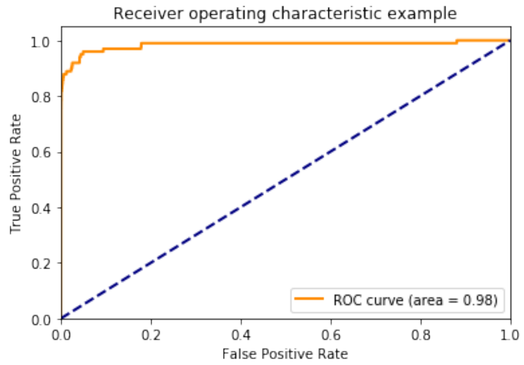
Feed 20% (+) (-)

SCORING

Classifier

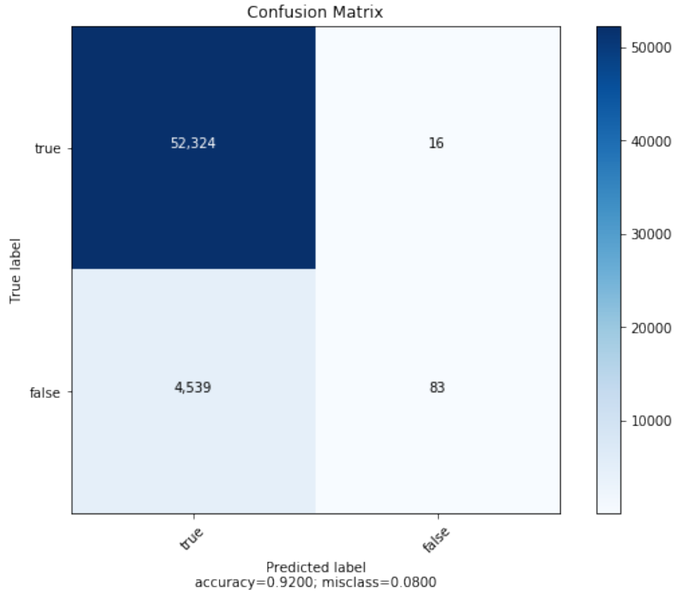
accuracy: 0.9839015484006882
precision: [0.99980346 0.08853119]
recall: [0.98406697 **0.88888889**]
fscore: [0.9918728 0.1610247]
support: [56863 99]

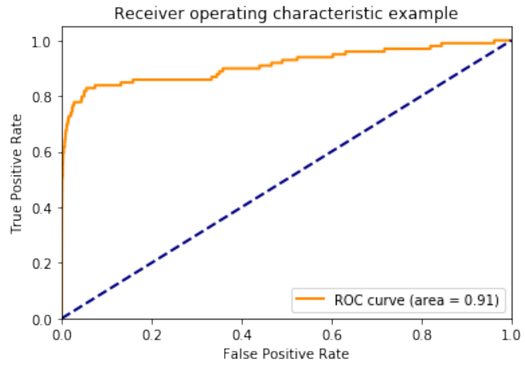




Auto-Encoder (THRESHOLD = 0.17)

accuracy: 0.9200344089041818
precision: [0.99969431 0.01795759]
recall: [0.92017656 **0.83838384**]
fscore: [0.95828869 0.03516204]
support: [56863 99]





Interactivity test

THRESHOLD 0.17

Feed 80% -
SCORING

TRAINED NEGATIVE

Greater Than Threeshold: 13932 over 227452
Accuracy: 0.94
Mean: 0.05370073765516281
Std: 0.09088834375143051
Min: 5.205785419093445e-05
Max: 3.160778045654297

Feed 20% -
SCORING

NEW NEGATIVE

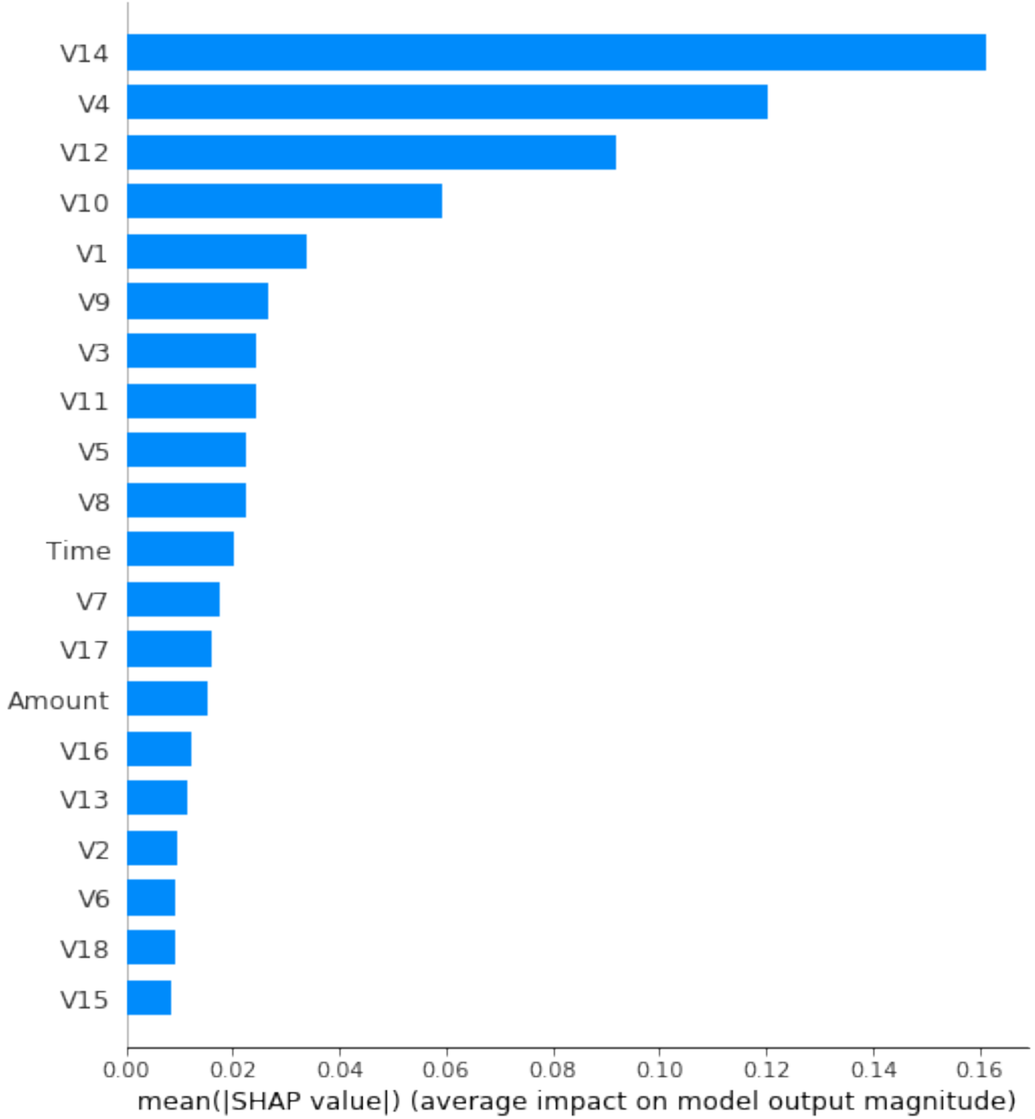
Greater Than Threeshold: 4809 over 56863
Accuracy: 0.92
Mean: 0.07014649361371994
Std: 0.1701817512512207
Min: 8.863704715622589e-05
Max: 26.871849060058594

Feed 100% +
SCORING

NEW POSITIVE

Greater Than Threeshold: 416 over 492
Accuracy: 0.85
Mean: 3.5444180965423584
Std: 7.813464164733887
Min: 0.001192888943478465
Max: 55.32634735107422

Features Interpretability



Source Code

https://github.com/JacopoMangiavacchi/AutoEncoderForAnomalyDetection