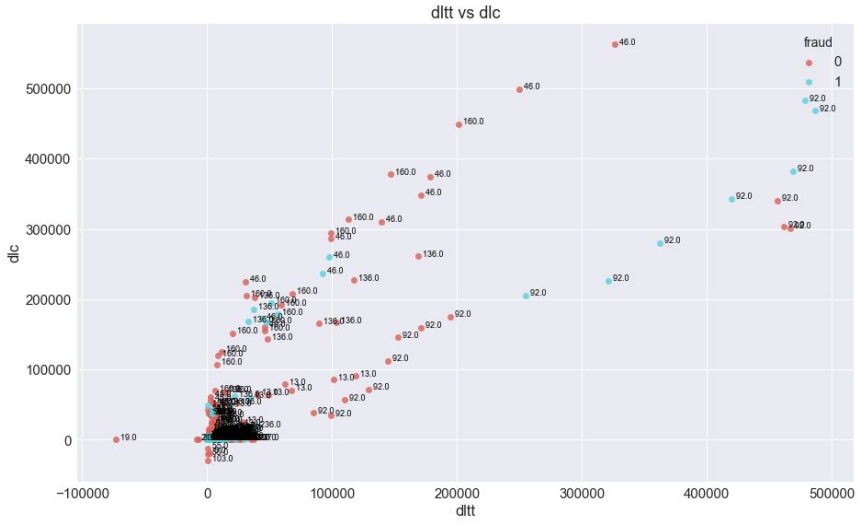
Dataset A, Week 11

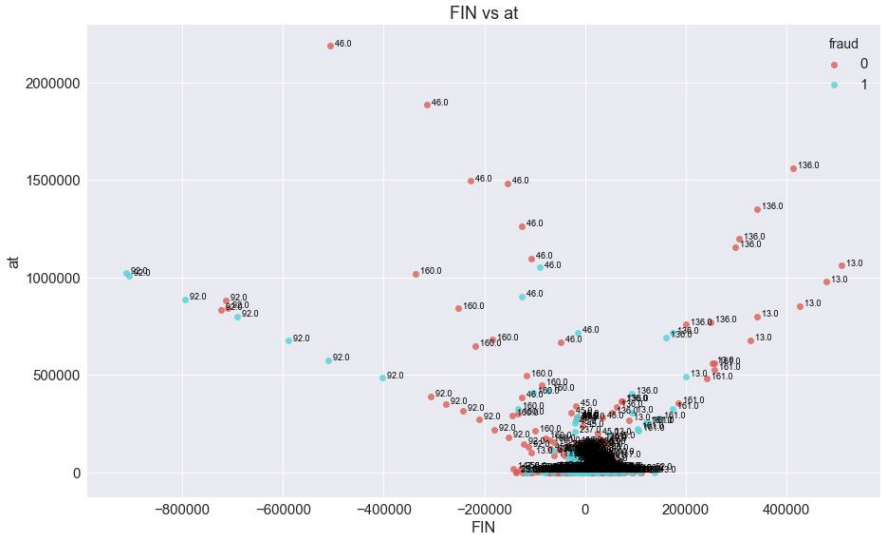
CISC 873 Hannah LeBlanc

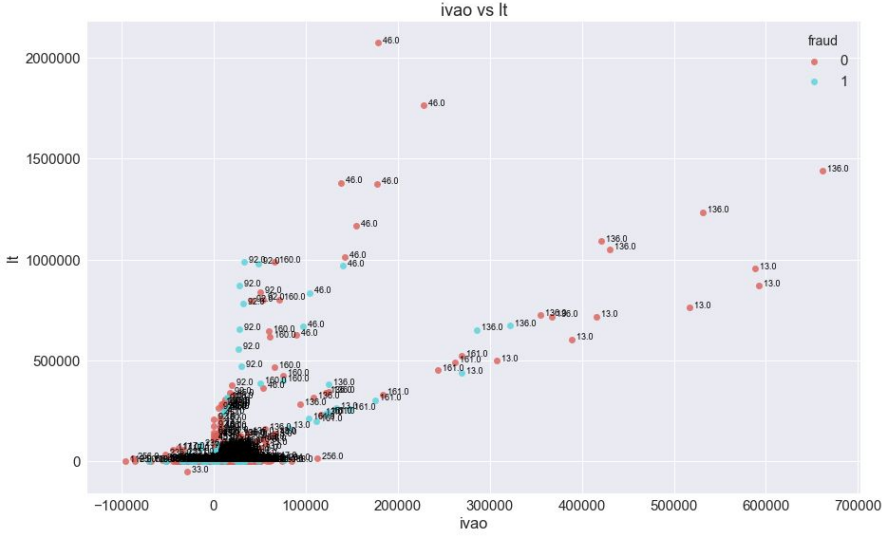
Data Preprocessing

- Dropped ['rsst-acc', 'r-vwretd', 'vwretd', 'ret-t', 'r-1-vwretd', 'ret-t1', 'vwretd-t1'] columns
- 8/144 dropped rows labeled as fraud
- Expectation Maximization with Impyute Python package

```
X shape  (3013, 37)
y shape  (3013, 1)
fraud  621
normal  2392
fraud/all 0.21
```


dltt vs dlc
fraud
0
1
dlc
dltt
500000
400000
300000
200000
100000
0
−100000
0
100000
200000
300000
400000
500000

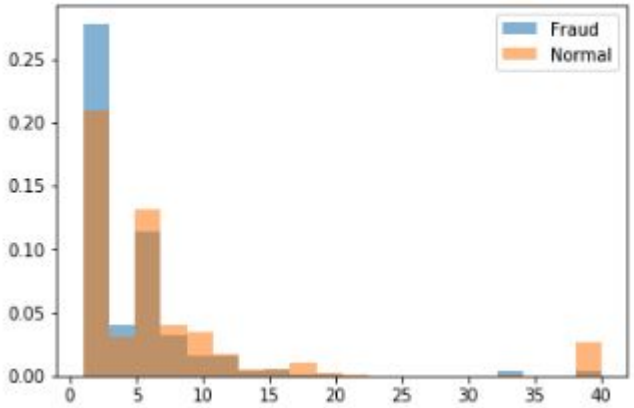

FIN vs at
fraud
0
1
at
FIN
2000000
1500000
1000000
500000
0
−800000
−600000
−400000
−200000
0
200000
400000


ivao vs lt
fraud
0
1
lt
ivao
2000000
1500000
1000000
500000
0
−100000
0
100000
200000
300000
400000
500000
600000
700000


rect vs sale
fraud
0
1
sale
rect
150000
125000
100000
75000
50000
25000
0
−25000
0
200000
400000
600000
800000
1000000

Missing Values

```
average missing # missing values in fraud:  4.03
average missing # missing values in normal:  6.6
```



	null_normal
vwretd-t1	2528
ret-t1	1054
r-1-vwretd	1054
ret-t	1003
vwretd	1003
r-vwretd	1003
rsst-acc	729
NCO	556
ch-emp	460
WC	392
act	391
FIN	378
lct	373
bm	344
issue	343
prcc-f	332
emp	312
ivao	303
ch-cs	293
ch-earn	281

	null_fraud
vwretd-t1	629
ret-t1	153
r-1-vwretd	153
ret-t	146
vwretd	146
r-vwretd	146
rsst-acc	140
NCO	111
ch-emp	82
lct	72
WC	72
act	72
FIN	60
issue	56
emp	50
ivao	48
dltis	35
ch-cs	35
ch-earn	33
sstk	33

Before

Layer (type)	Output Shape
input_48 (InputLayer)	(None, 37)
dense_381 (Dense)	(None, 32)
dense_382 (Dense)	(None, 16)
dense_383 (Dense)	(None, 8)
dense_384 (Dense)	(None, 2)
dense_385 (Dense)	(None, 8)
dense_386 (Dense)	(None, 16)
dense_387 (Dense)	(None, 32)
dense_388 (Dense)	(None, 37)

Encoder

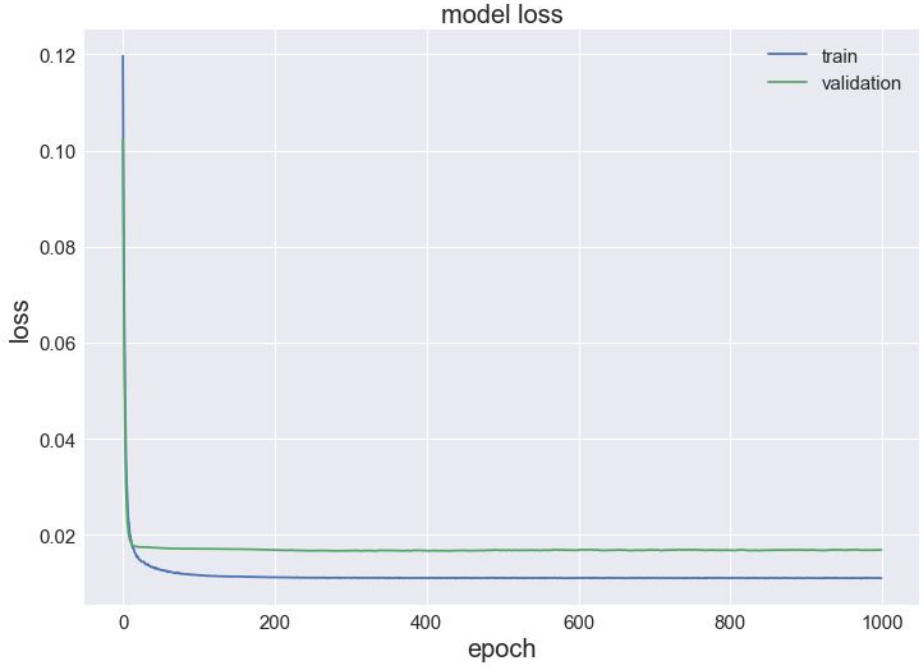
Latent Layer

Decoder

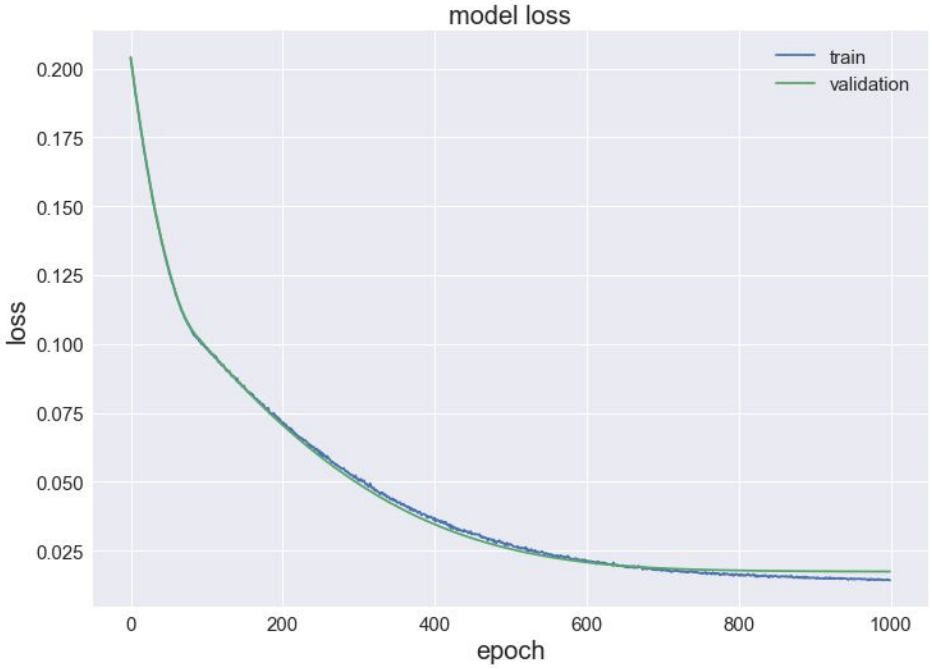
After

Layer (type)	Output Shape
input_26 (InputLayer)	(None, 37)
dropout_127 (Dropout)	(None, 37)
dense_137 (Dense)	(None, 8)
dropout_128 (Dropout)	(None, 8)
dense_138 (Dense)	(None, 3)
dropout_129 (Dropout)	(None, 3)
dense_139 (Dense)	(None, 8)
dropout_130 (Dropout)	(None, 8)
dense_140 (Dense)	(None, 37)

Learning Rate

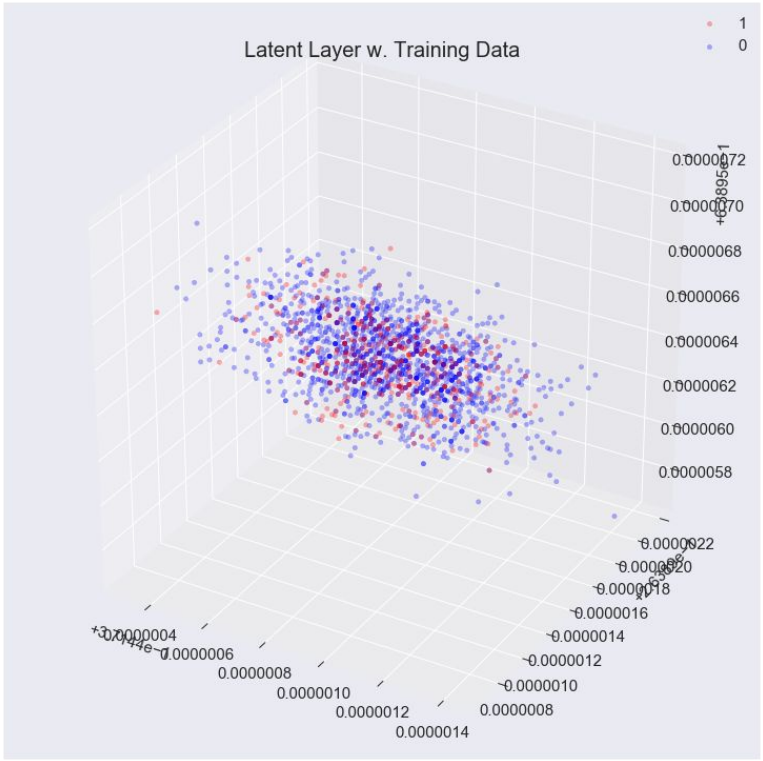


LR = 0.001

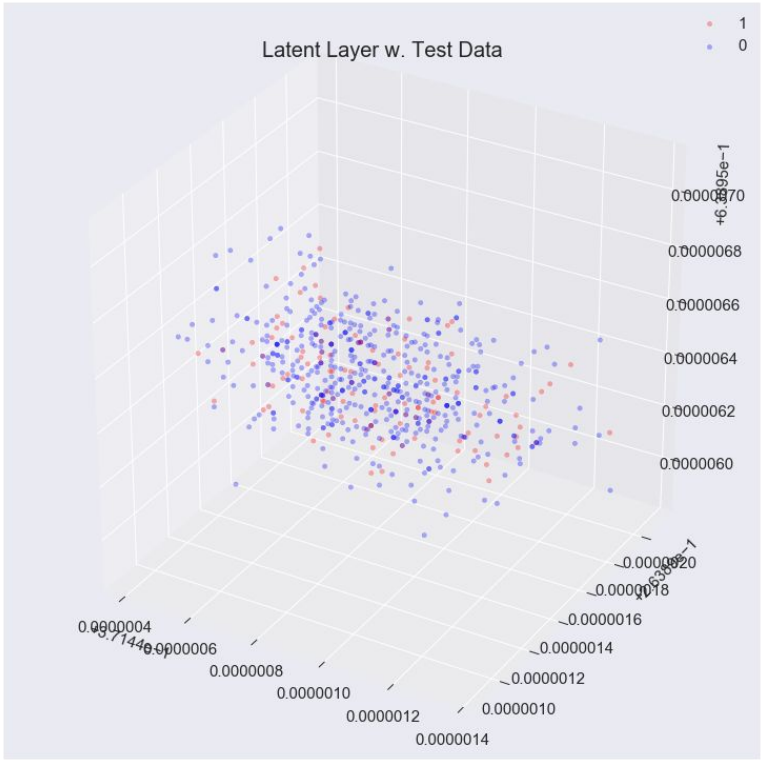


LR = 0.00001
L1 Kernel regularization = 0.01

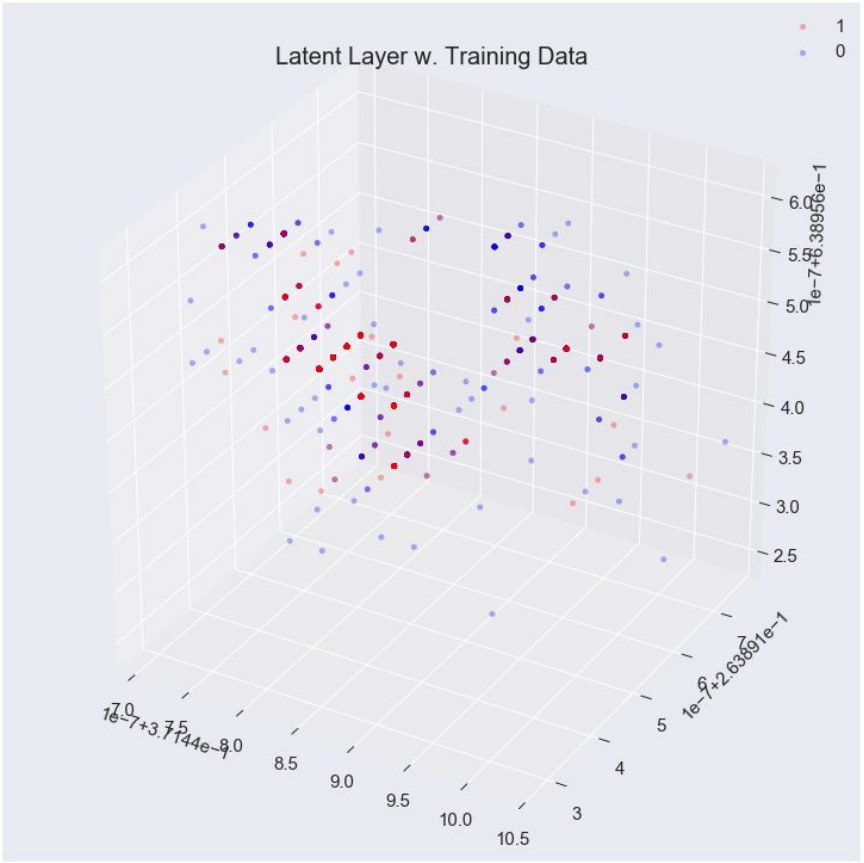
```
Train:  (1807, 37)
Validate:  (603, 37)
Test:  (603, 37)
```



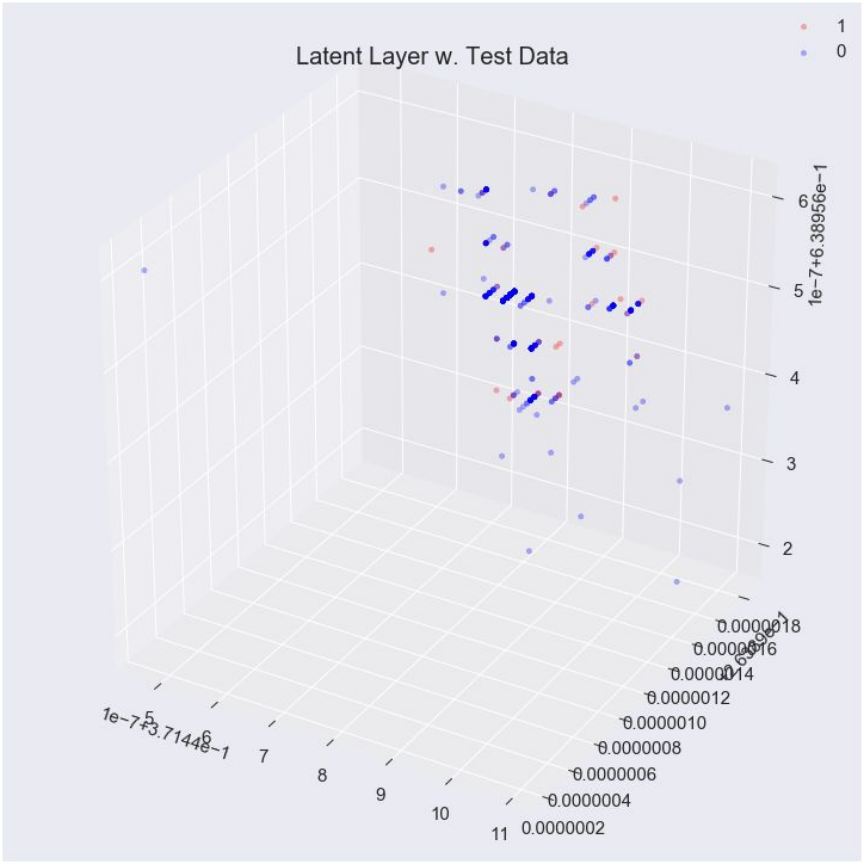
X_train_noisy



X_test_noisy



X_train (uncorrupted)

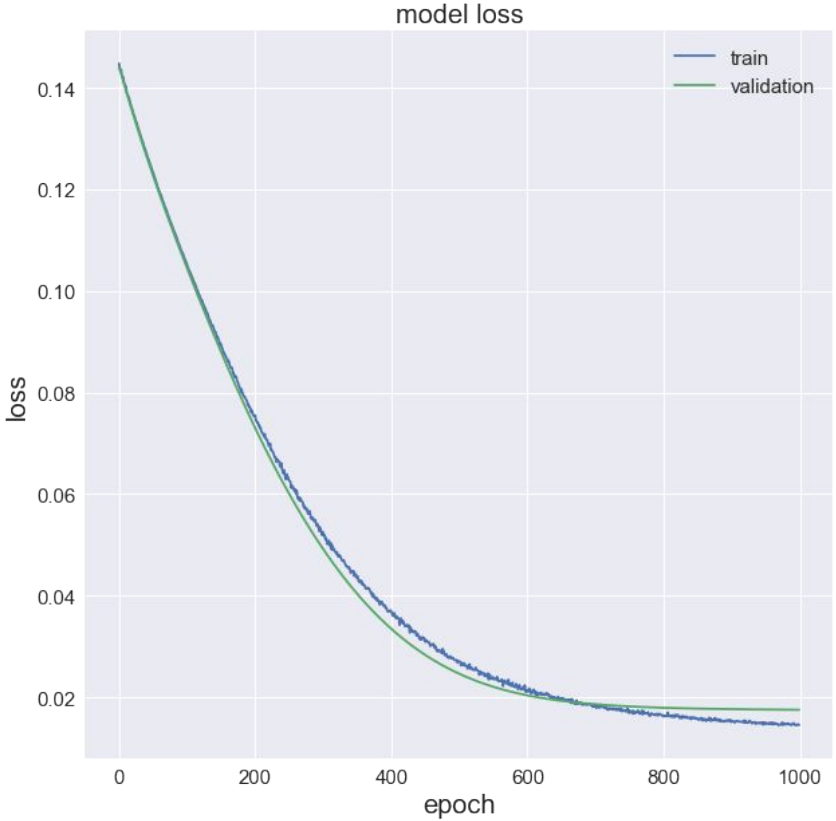


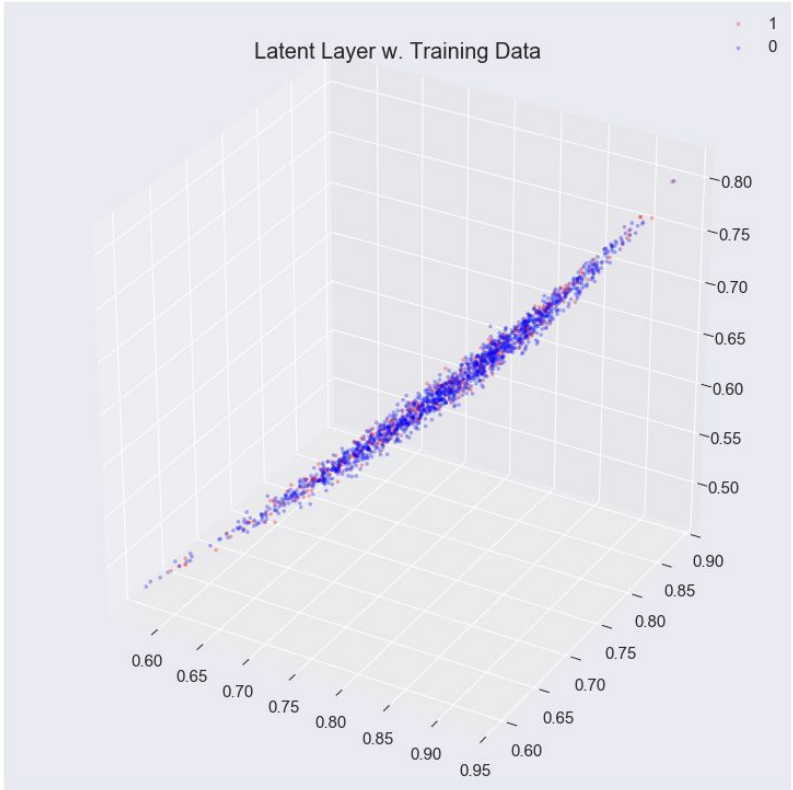
X_test (uncorrupted)

Stacked denoising autoencoder using greedy layerwise unsupervised pre-training

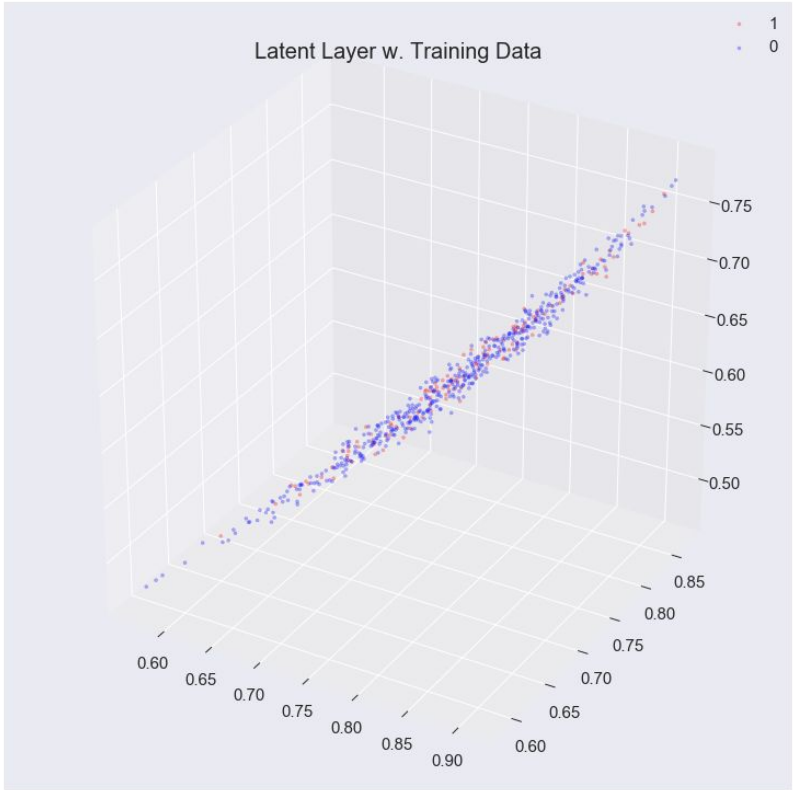
```
inputs =Input(shape=(X_all.shape[1],))
x = DenseLayerAutoencoder([8, 3], activation='sigmoid',dropout=0.5)(inputs)
model = Model(inputs=inputs, outputs=x)
print(model.summary())
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_40 (InputLayer)        (None, 37)                0
_________________________________________________________________
dense_layer_autoencoder_4 (D (None, 37)                376
=================================================================
```

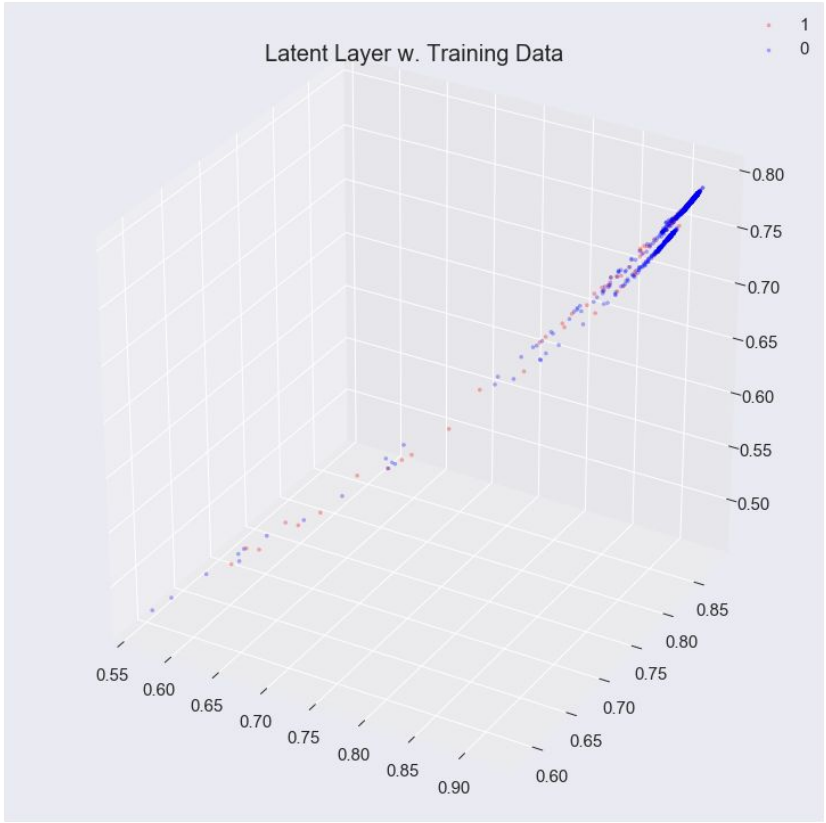




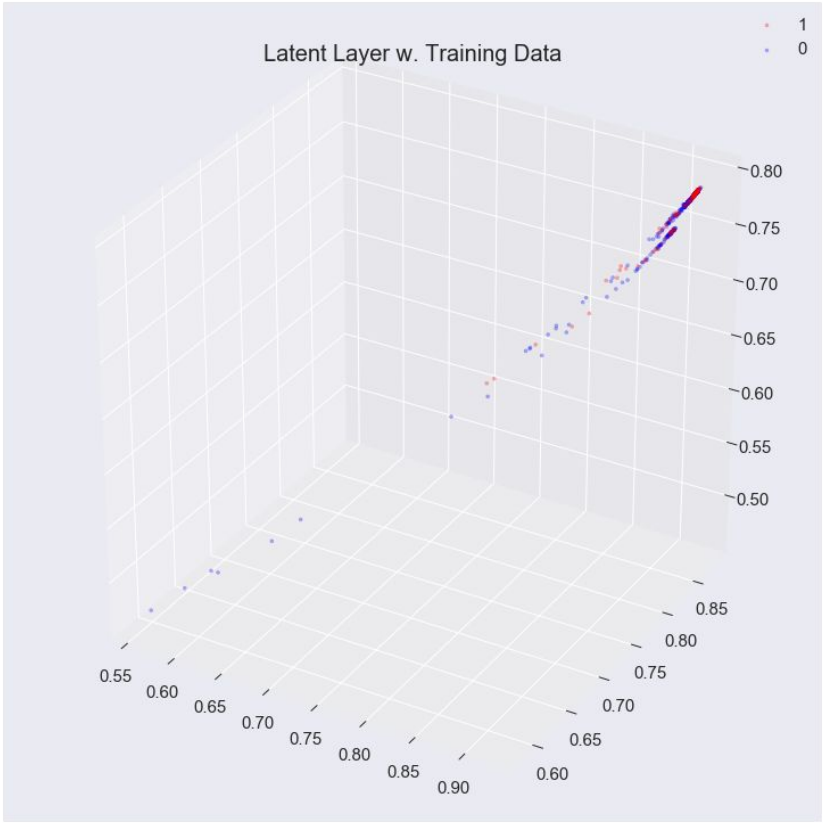
X_train_noisy



X_test_noisy



X_train (uncorrupted)



X_test (uncorrupted)