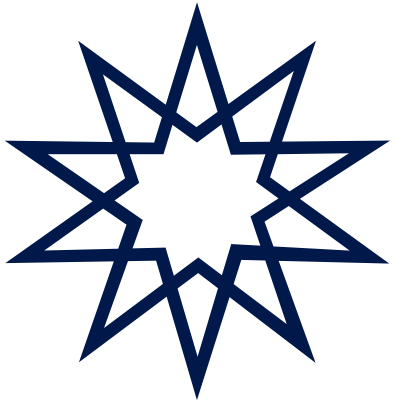
PROJECT PRESENTATION - CREDIT CARD FRAUD DETECTION

Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

BLM5116 Data Mining and Knowledge Discovery

ÖNDER GÖRMEZ

21501035



YILDIZ TEKNİK ÜNİVERSİTESİ

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

IV. Conclusions

V. References

VI. Q & A

Background

Dolandırıcılık Tespit Yöntemleri:

- Geleneksel Dolandırıcılık Tespit Yöntemleri

- Geleneksel Makine Öğrenmesi Yöntemleri
  - Decision Trees
  - K-Nearest Neighbors (KNNs)
  - Support Vector Machine (SVM)

- Derin Öğrenme Yöntemleri
  - Yapay Sinir Ağları (ANN)
  - Konvolüsyonel Sinir Ağları (CNN)
  - Recurrent Neural Networks (RNN)

Background

Dataset Features:

- A research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles)
- Transactions, September 2013 by European cardholders
- In two days, 492 frauds out of 284,807 transactions
- Input variables which are the result of a PCA transformation
- +5K Users studied in Kaggle

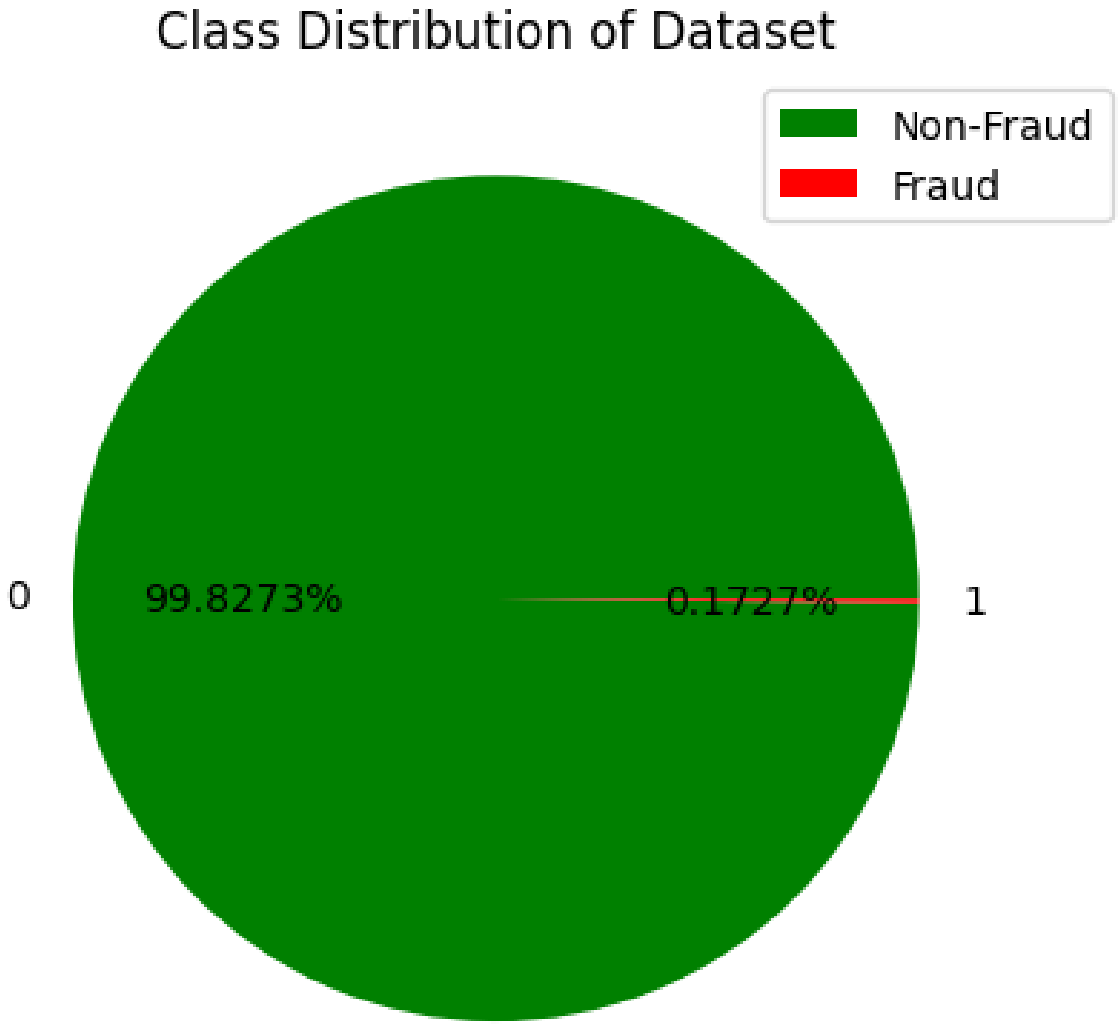
Background

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...

5 rows × 31 columns

...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Background


Class Distribution of Dataset
Non-Fraud
Fraud
0
99.8273%
0.1727%
1

Background

Splitting the Dataset:

%80 train, %20 Test set olarak ayrıldı

- Train set size: 227845
  - Non-Fraud transactions in the training set: 227451 samples, 99.8271%
  - Fraud transactions in the training set: 394 samples, 0.1729%

- Test set size: 56962
  - Non-Fraud transactions in the test set: 56864 samples, 99.8280%
  - Fraud transactions in the test set: 98 samples, 0.1720%

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

IV. Conclusions

V. References

VI. Q & A

Proposed Methods

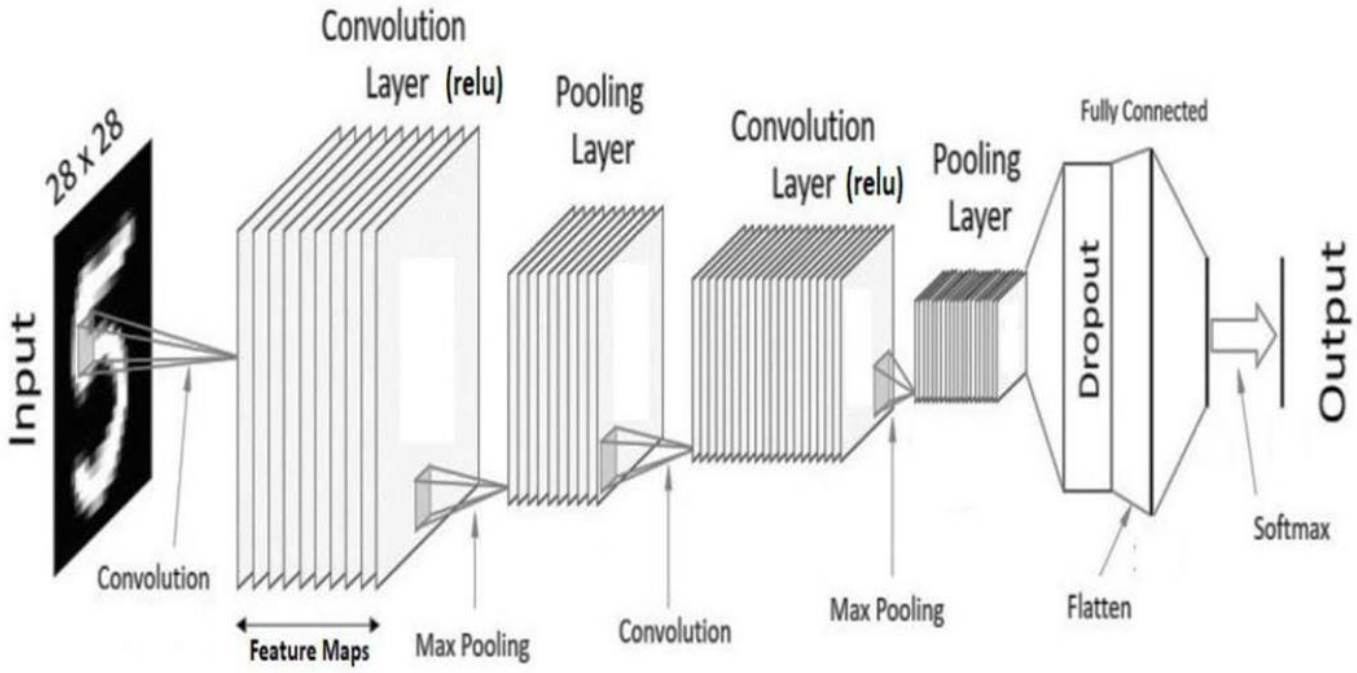


FIGURE 3. CNN output layer.

TABLE 8. The summary of CNN sequential model.

Layers (Types)	Output Shape	Param#
conv1d (Conv1D)	(None, 29, 32)	96
batch normalisation (Batch No)	(None, 29, 32)	128
dropout (Dropout)	(None, 29, 32)	0
conv1d_1 (Conv1D)	(None, 28, 64)	4160
batch_normalisation_1 (Batch)	(None, 28, 64)	256
dropout_1 (Dropout)	(None, 28, 64)	0
flatten (Flatten)	(None, 1792)	0
dense (Dense)	(None, 64)	114752
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

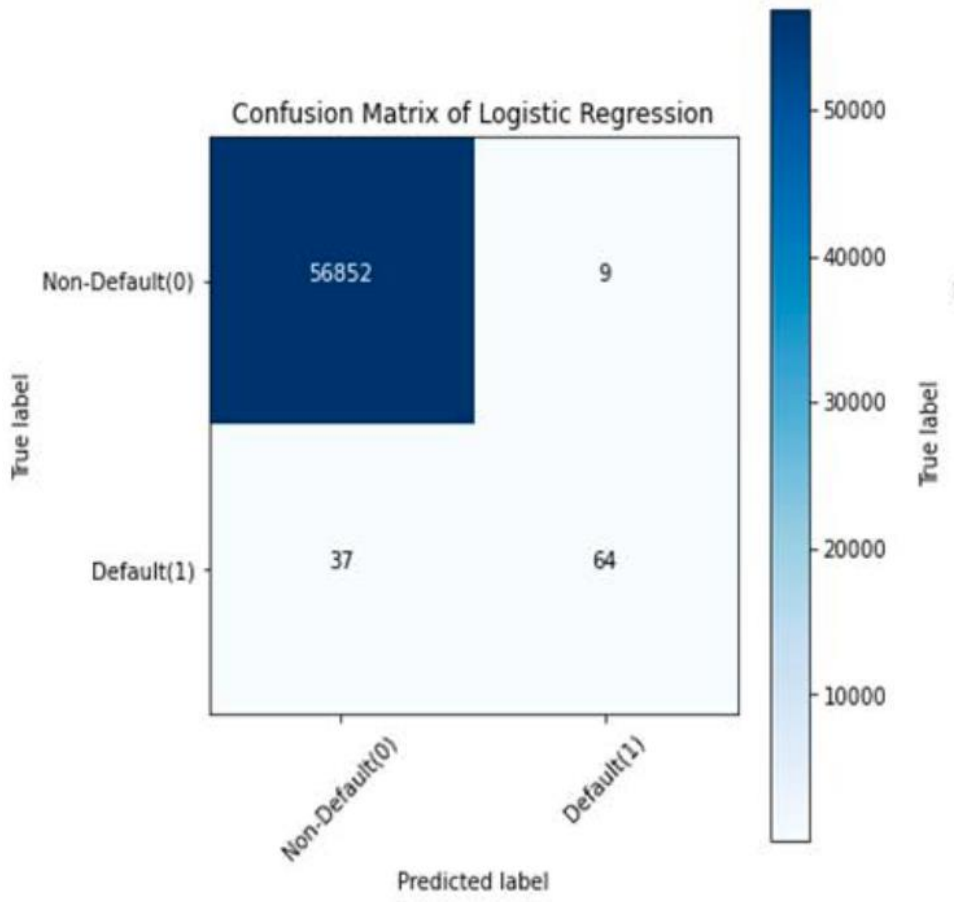
IV. Conclusions

V. References

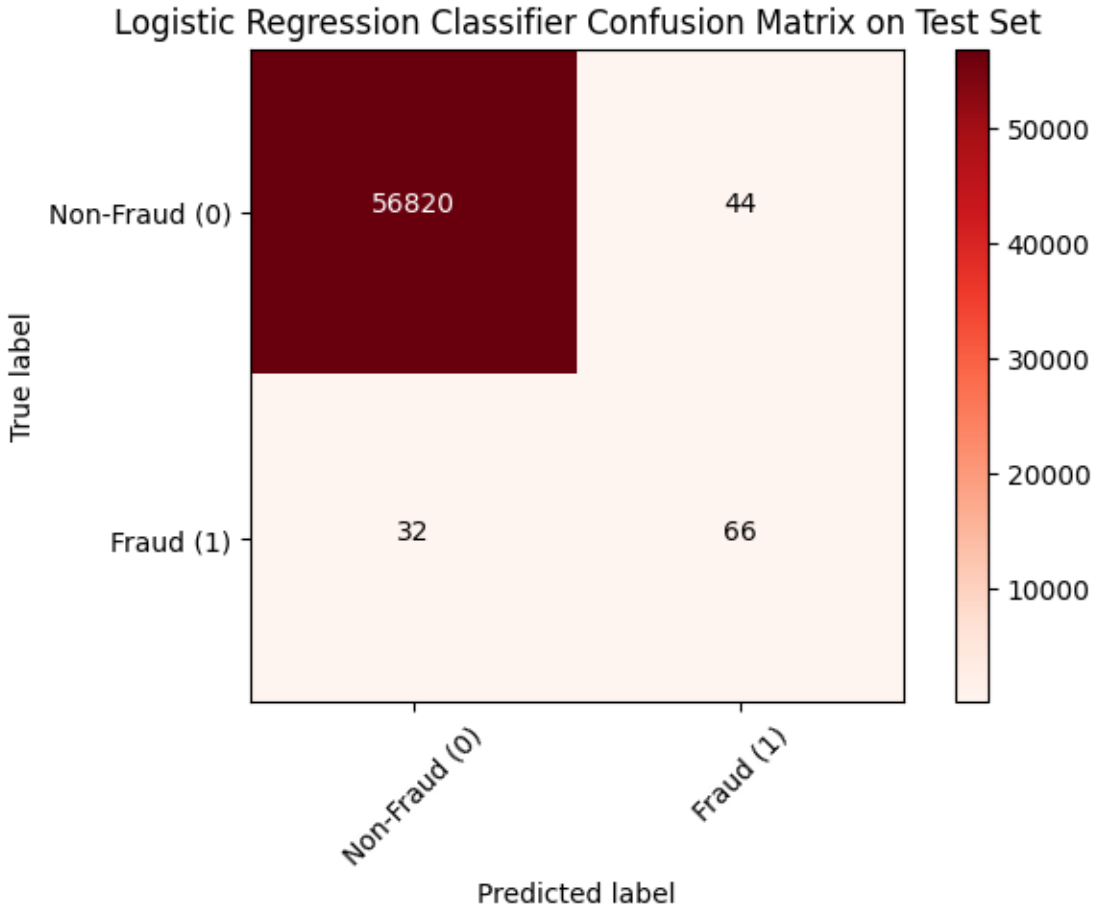
VI. Q & A

Experiments and Results

Result on Paper

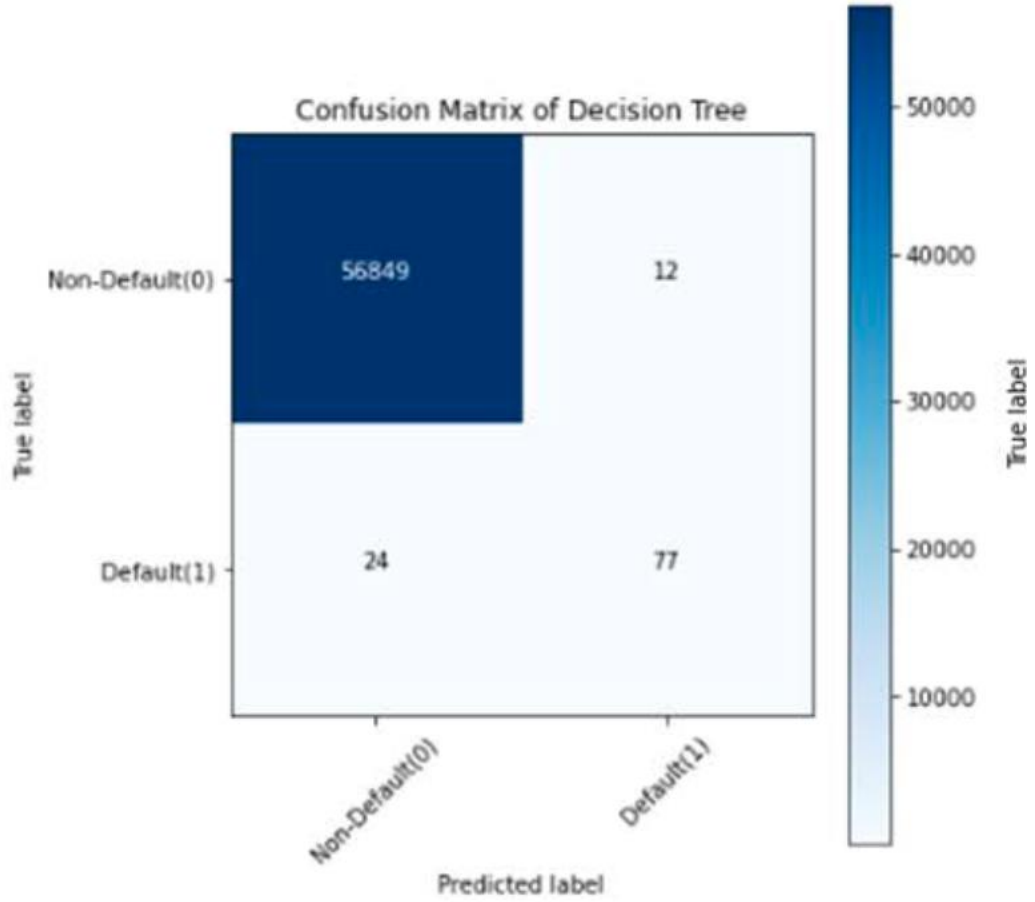


Our Result

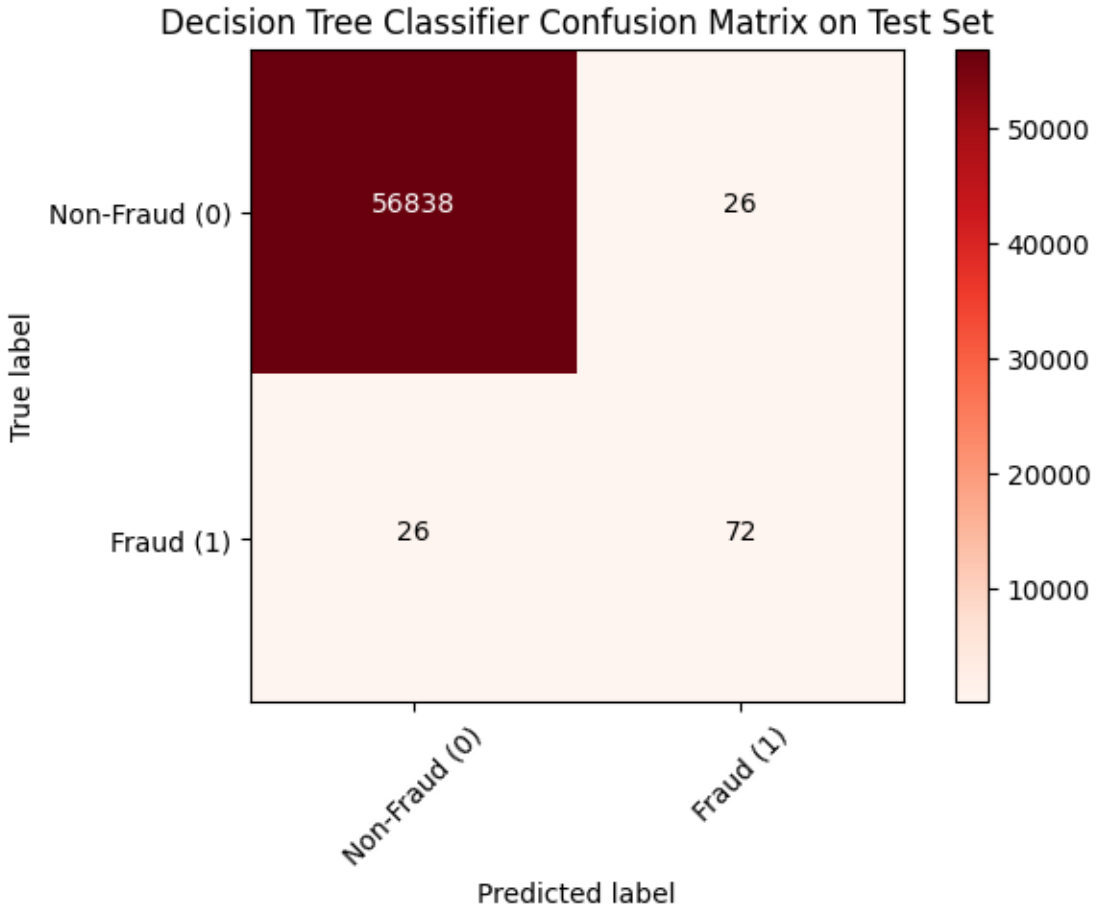


Experiments and Results

Result on Paper

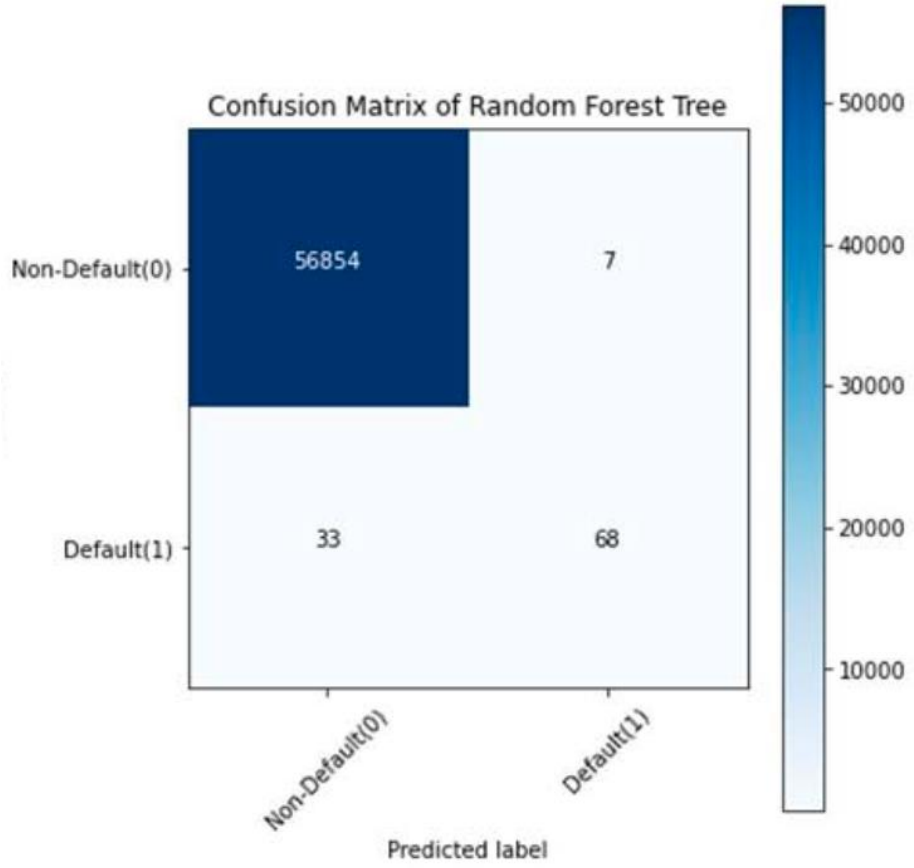


Our Result

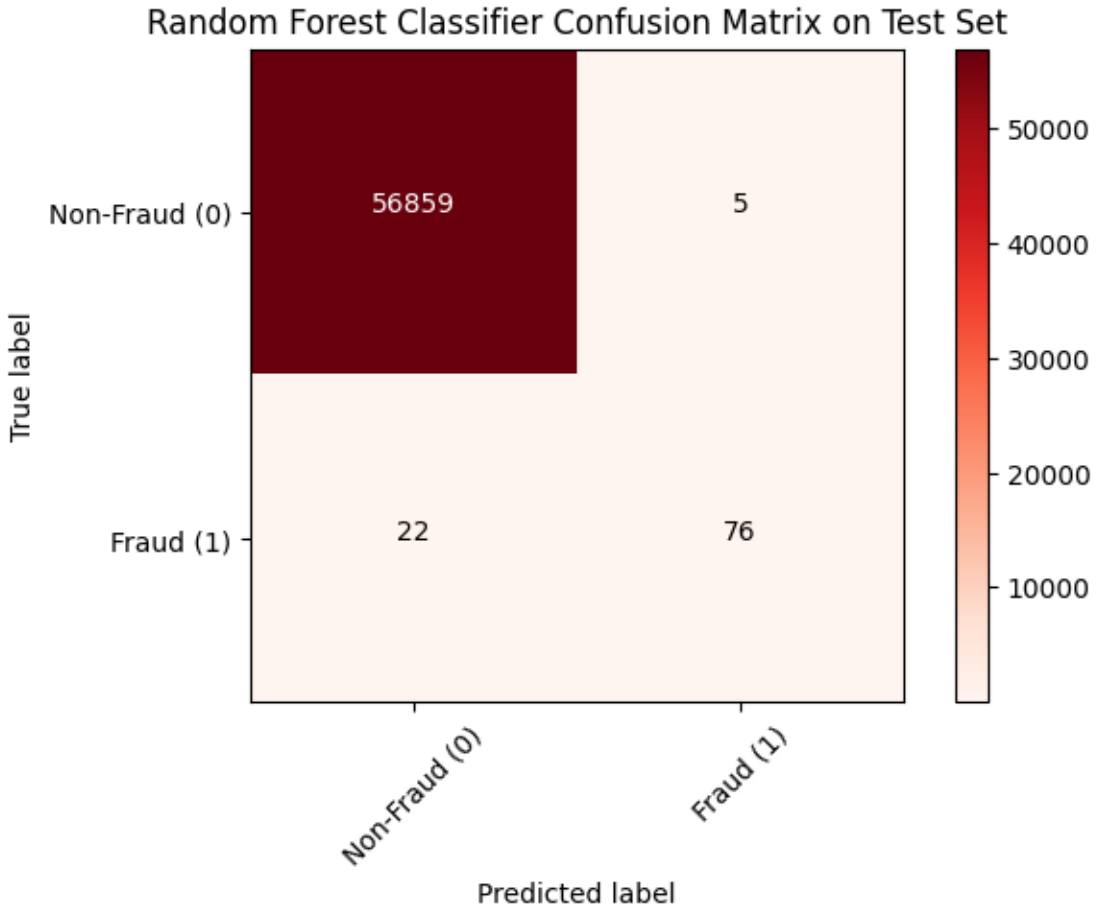


Experiments and Results

Result on Paper

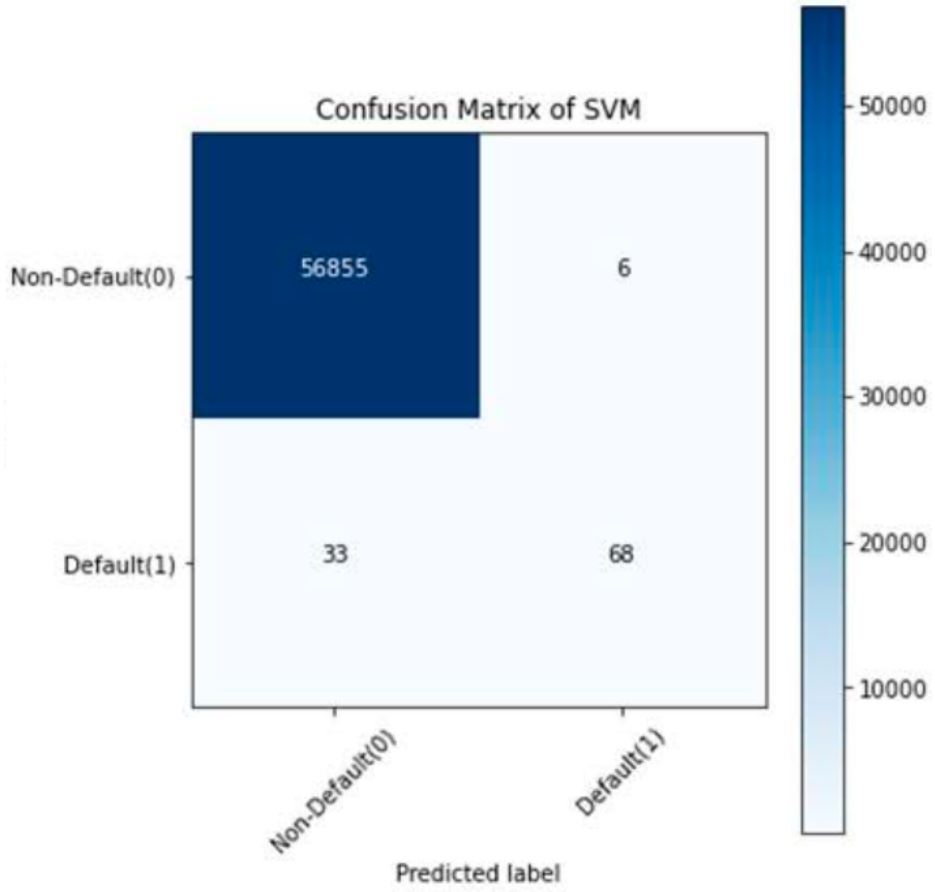


Our Result

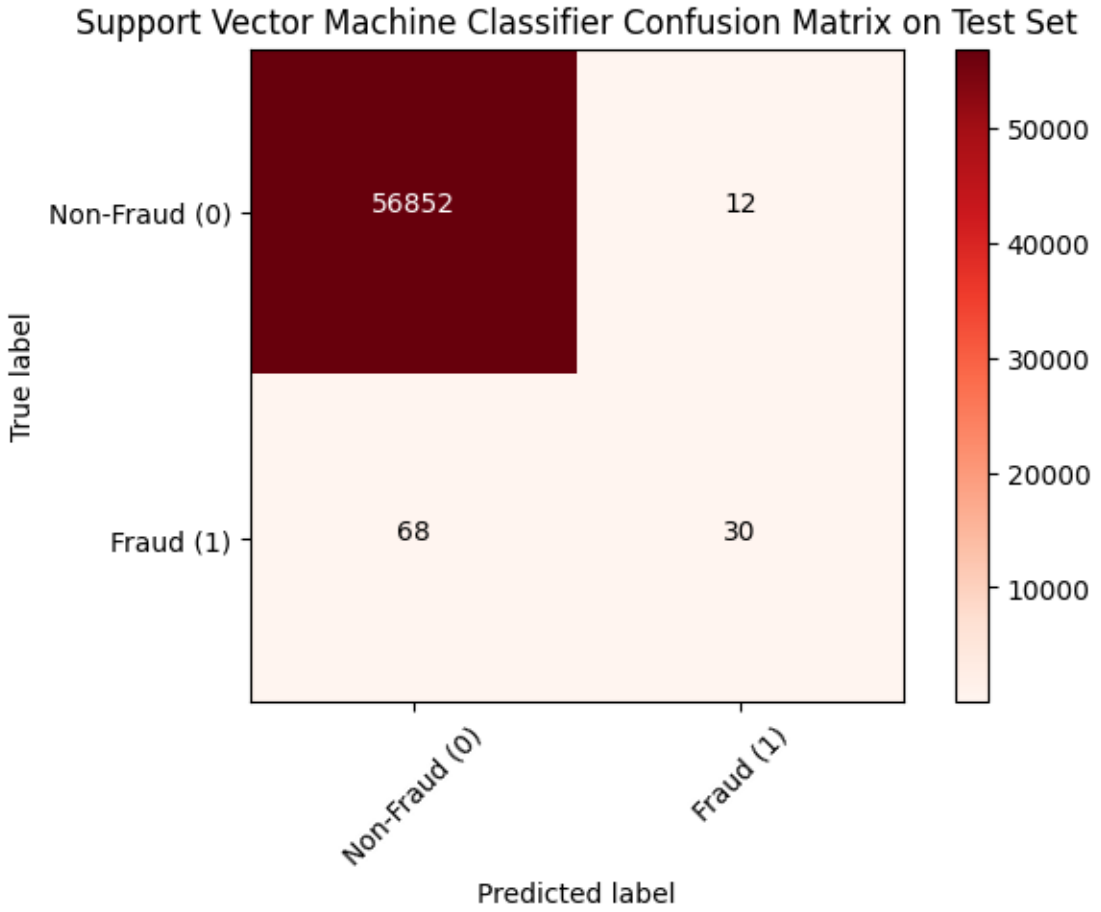


Experiments and Results

Result on Paper

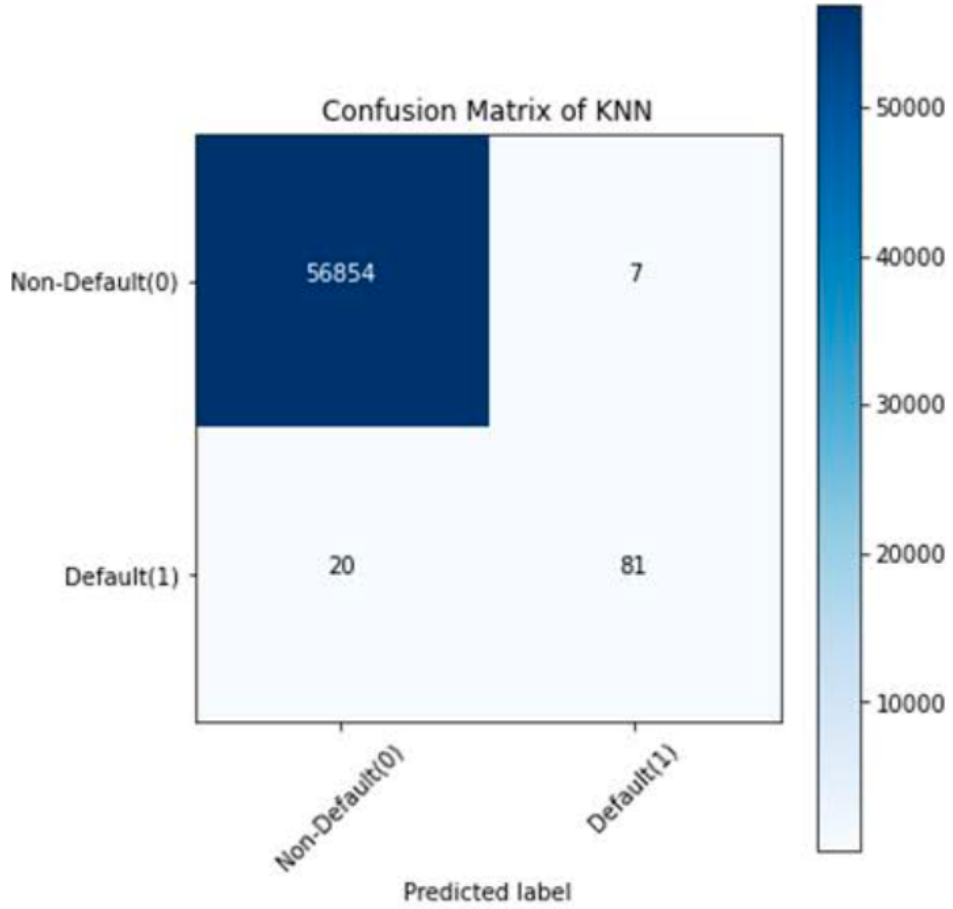


Our Result

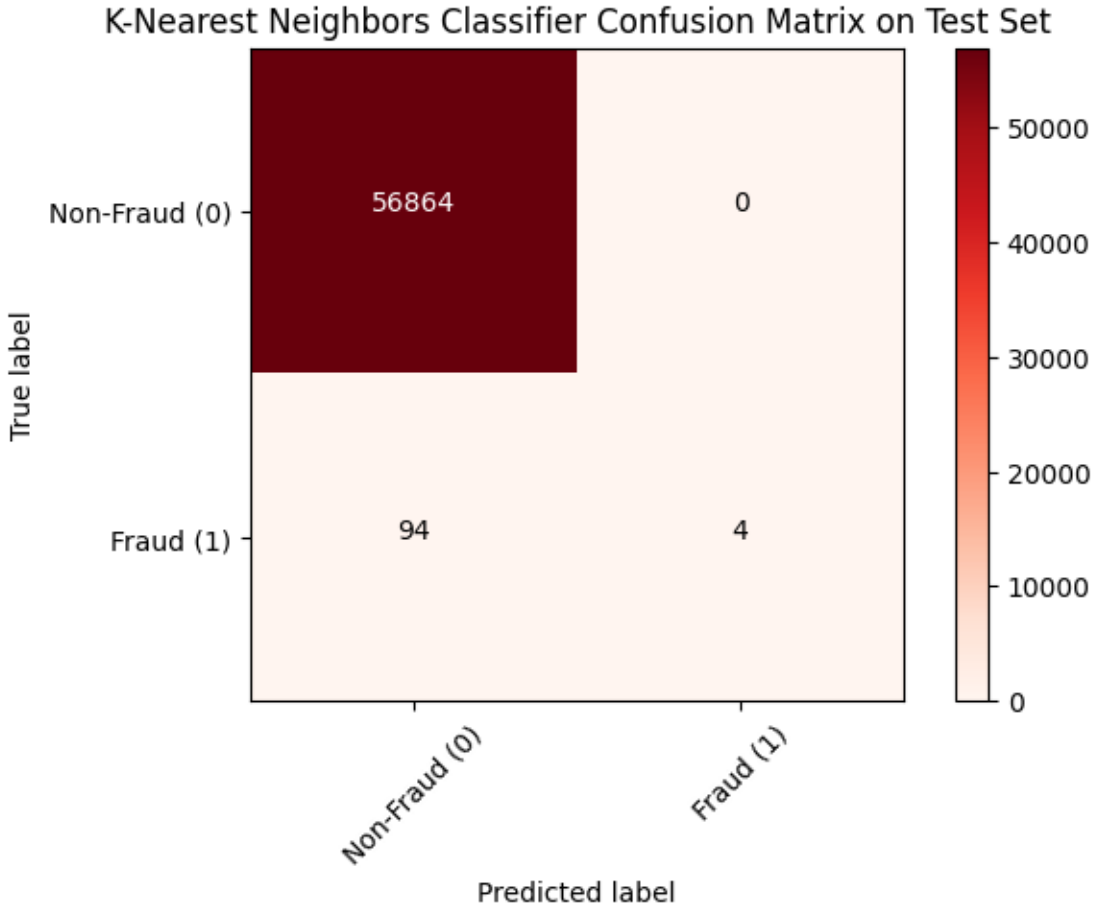


Experiments and Results

Result on Paper

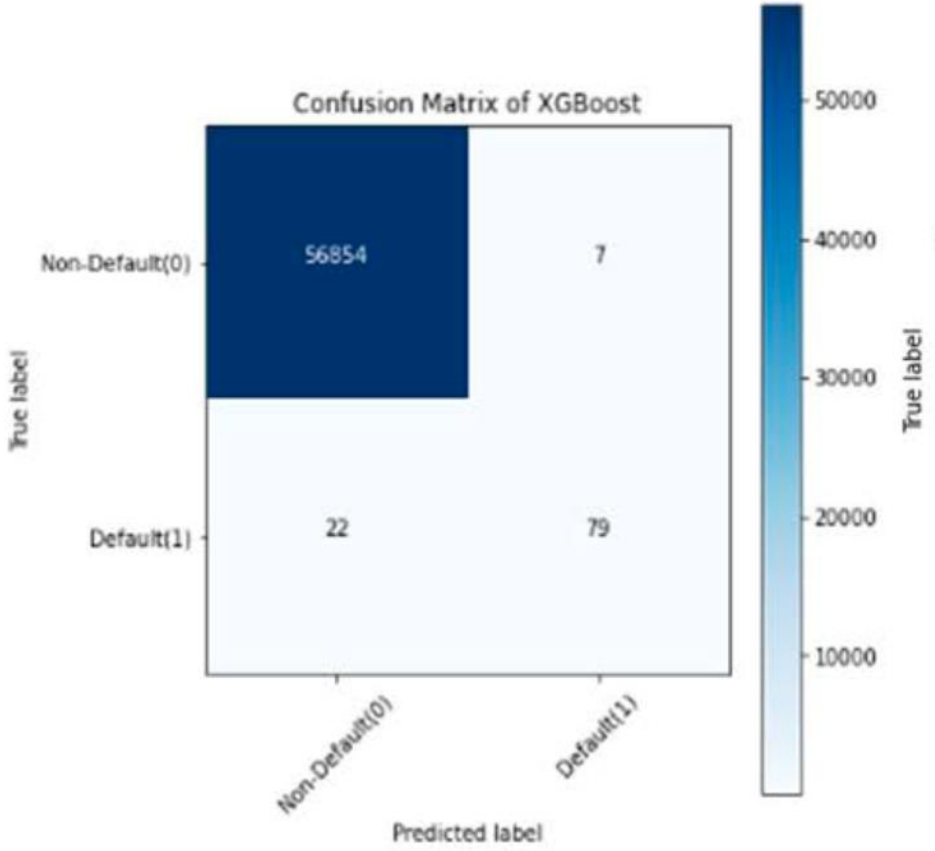


Our Result

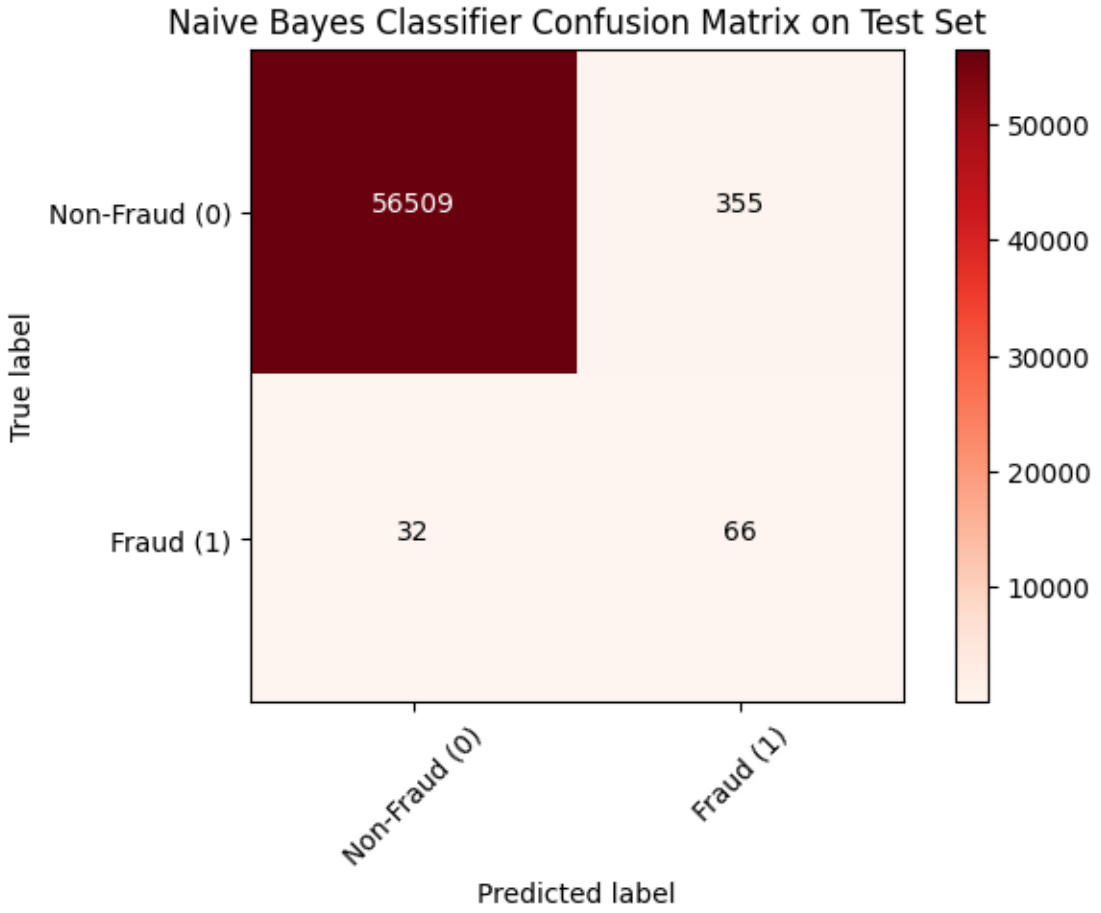


Experiments and Results

Result on Paper



Our Result

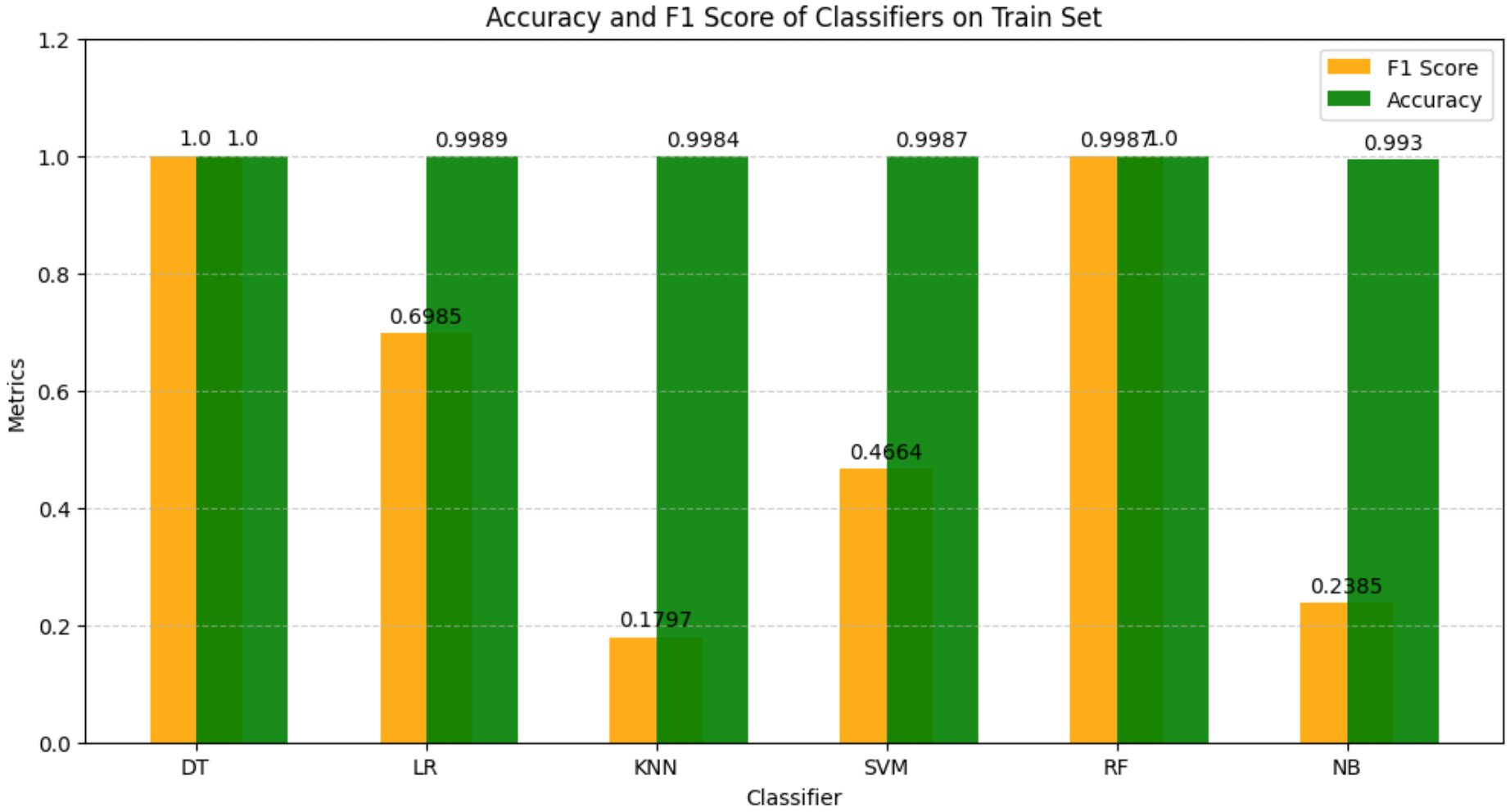


YTÜ YILDIZ TEKNİK ÜNİVERSİTESİ

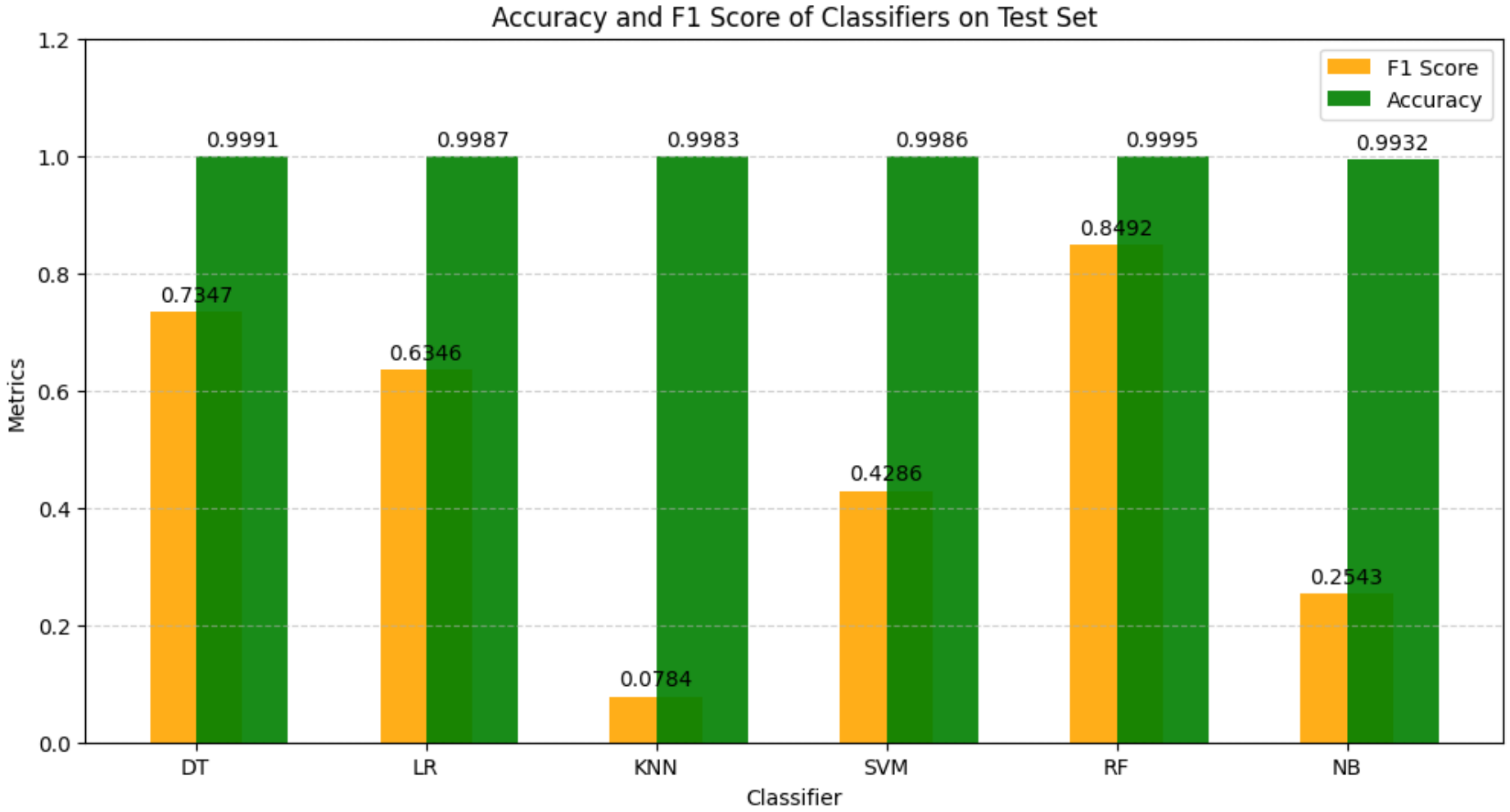
Experiments and Results

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
DT	Train	1.0000	1.0000	1.0000	1.0000	394
DT	Test	0.9991	0.7347	0.7347	0.7347	98
LR	Train	0.9989	0.6915	0.7056	0.6985	394
LR	Test	0.9987	0.6000	0.6735	0.6346	98
KNN	Train	0.9984	0.9750	0.0990	0.1797	394
KNN	Test	0.9983	1.0000	0.0408	0.0784	98
SVM	Train	0.9987	0.7674	0.3350	0.4664	394
SVM	Test	0.9986	0.7143	0.3061	0.4286	98
RF	Train	1.0000	1.0000	0.9975	0.9987	394
RF	Test	0.9995	0.9383	0.7755	0.8492	98
NB	Train	0.9930	0.1470	0.6320	0.2385	394
NB	Test	0.9932	0.1568	0.6735	0.2543	98

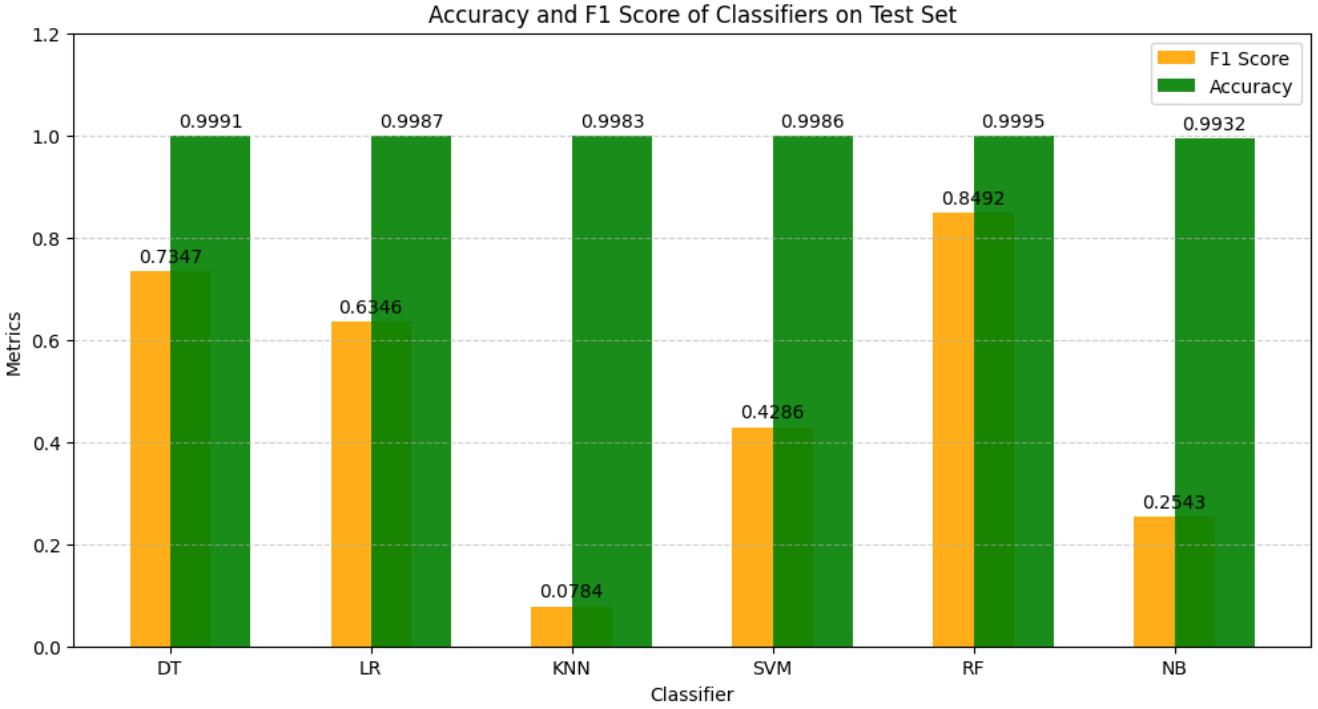
Experiments and Results

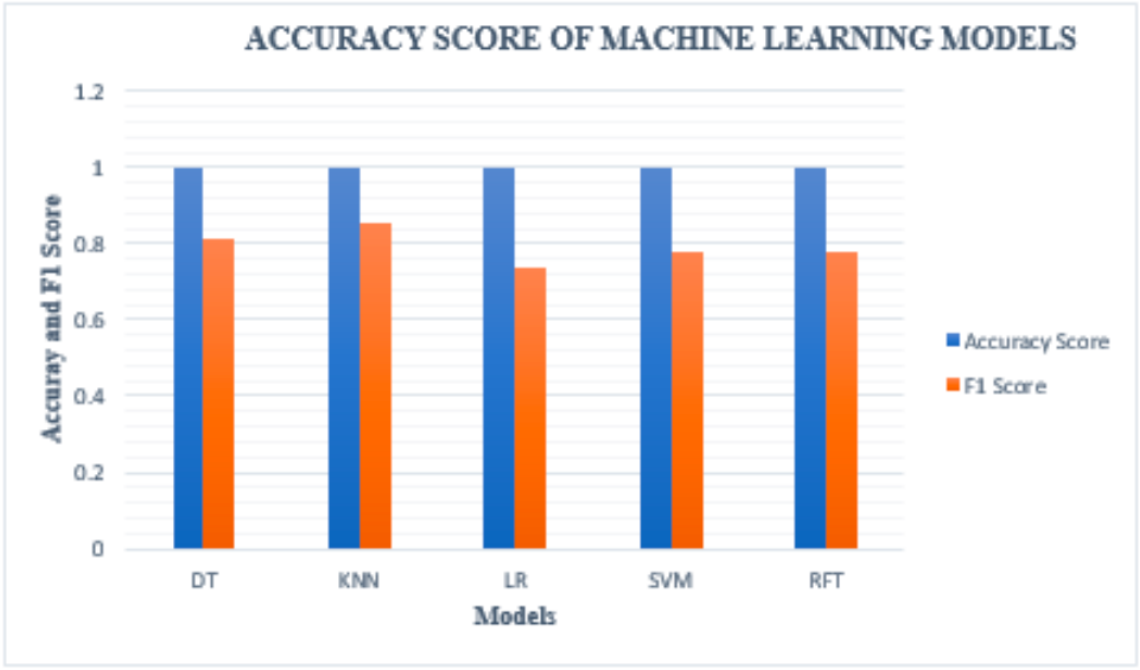


Experiments and Results



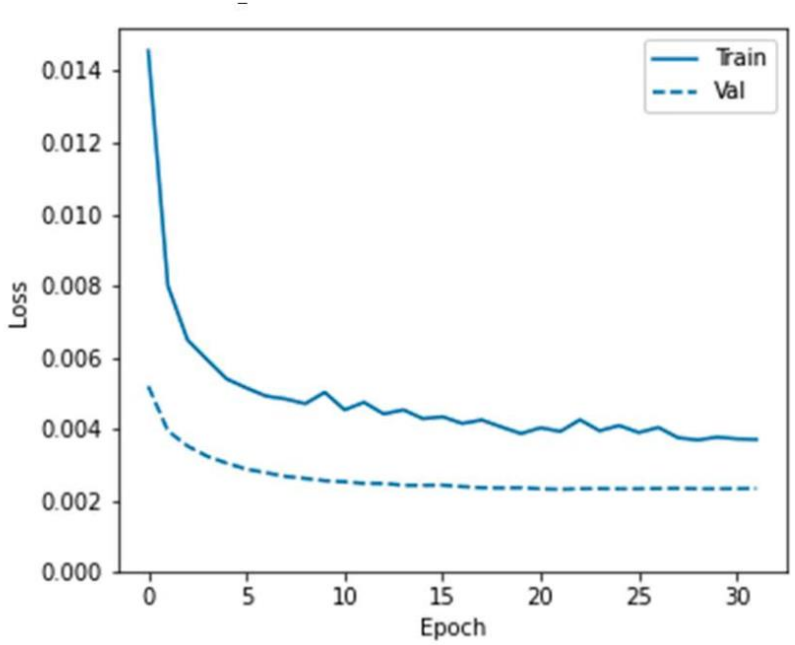
Experiments and Results

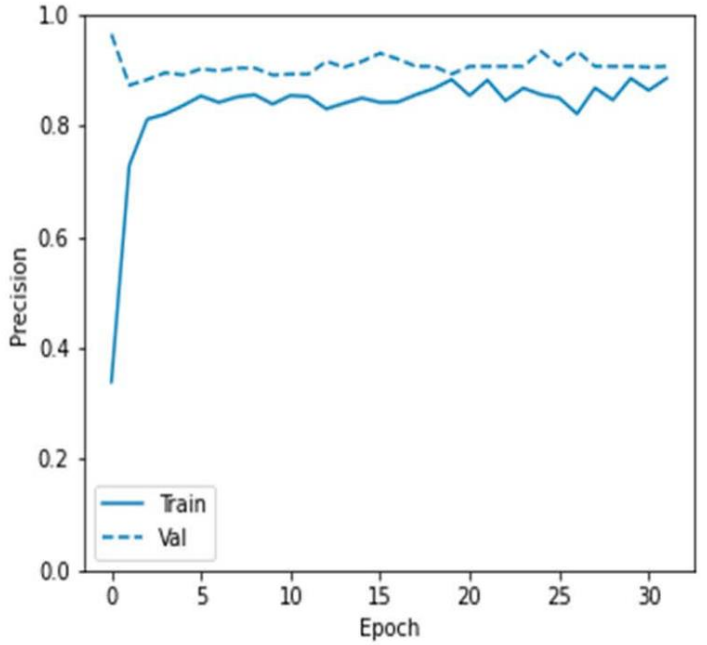


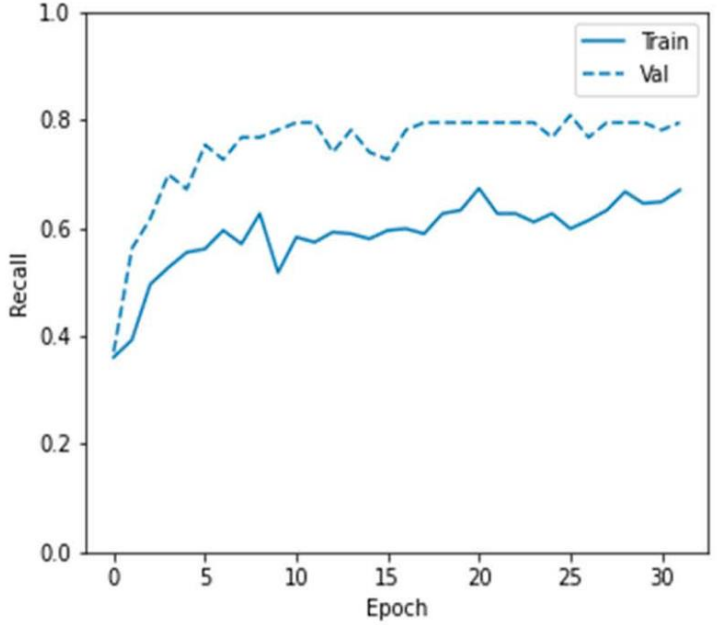


Experiments and Results

Result on Paper







Experiments and Results

TABLE 6. The result of CNN model using epoch size as 35 and 14.

Metrics	Epoch Size 35	Epoch Size 14
Loss	0.004	0.014
TP	83.0	87
FP	6.0	66
TN	56849.0	56789
FN	24.0	20
Accuracy	0.999	0.998
Precision	0.932	0.568
Recall	0.775	0.813
AUC	0.929	0.942
PRC	0.816	0.741
Total Fraudulent Transactions	**107**	**107**

Result on Paper

TABLE 7. The accuracy of deep learning models using different epochs.

Name of Algorithms	Layers	Epoch Size	Training Accuracy	Validation Accuracy
CNN with imbalanced	11	20	0.932	0.913
	13	50	0.912	0.913
	14	100	0.963	0.943
	17	100	0.955	0.947
	20	100	0.949	0.935
CNN with balanced dataset	14	100	0.946	0.958
Baseline	05	20	0.907	0.883

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

IV. Conclusions

V. References

VI. Q & A

Conclusions

- Başarı Sırası: RF > DT > LR > SVM > NB > KNN
- Her makine öğrenmesi yöntemi her problem seti için uygun değildir.
- Veri seti ön işleme modelin çalışabilmesi / performansı için önemlidir

Conclusions

Future Works

- Derin öğrenme yöntemleri üzerinden başarım ölçümü yapılabilir.
- Imbalanced veri setini balanced bir veri seti haline getirerek sınıflandırma performansları ölçülebilir.
  - Random Oversampling
  - Random Undersampling
  - SMOTE (Synthetic Minority Over-sampling Technique)

Conclusions

Future Works

- KNN’de yaşanan başarısızlığın düzeltilmesi için çalışma yapılabilir.
- SVM’de farklı kernel parametreleri ve Grid Search ile parametre optimizayonu yapılabilir.

Paper Future Works

- Farklı Deep Learning algoritmaları kullanılabilir.

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

IV. Conclusions

V. References

VI. Q & A

References

[1] Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

[2] Machine Learning Group - ULB - Credit Card Fraud Detection Dataset

[3] Credit Card Fraud Detection Using Lightgbm Model

Agenda

I. Background

II. Proposed Methods

III. Experiments and Results

IV. Conclusions

V. References

VI. Q & A

Q & A

THANK YOU FOR LISTENING…