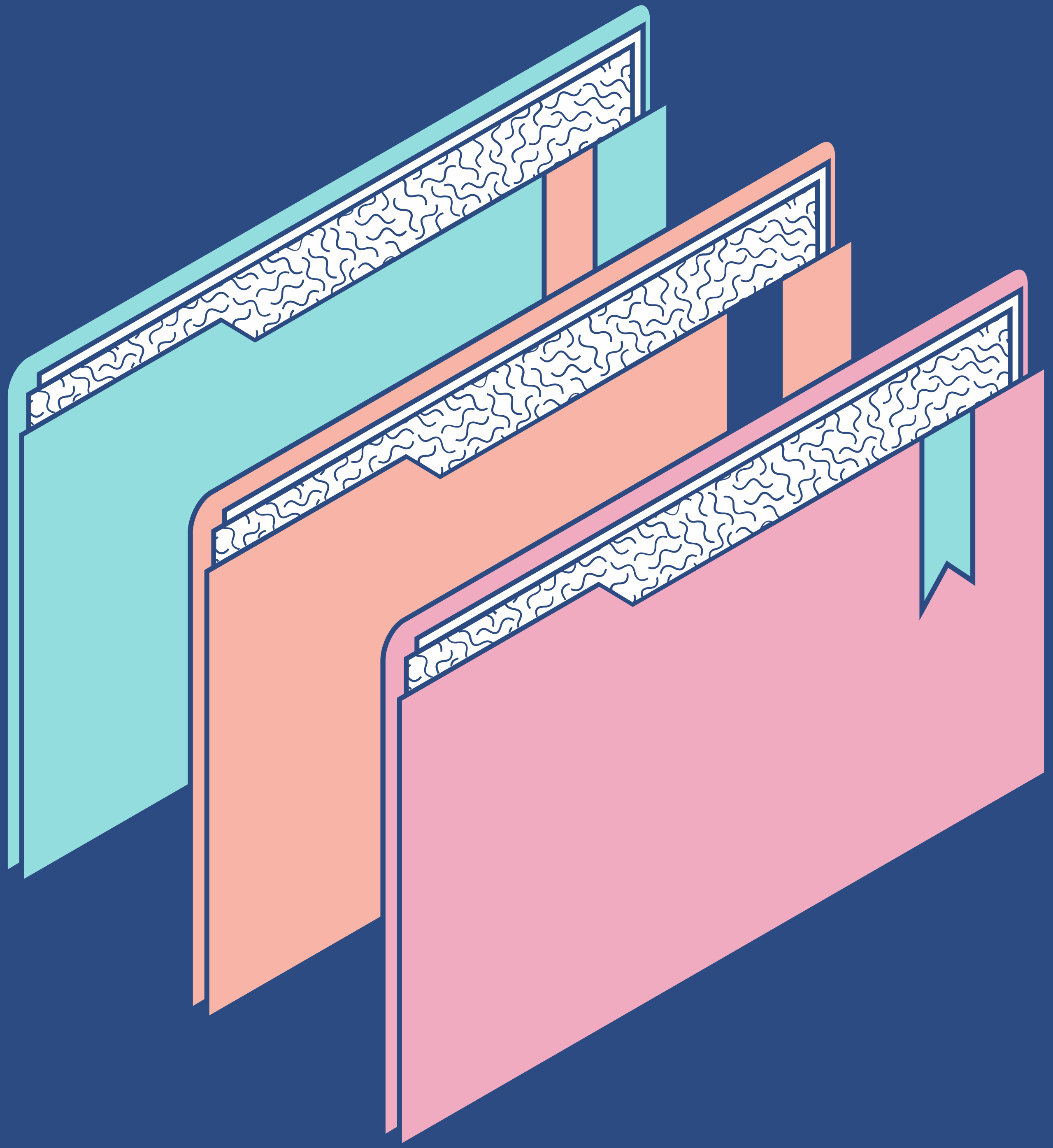
Advanced Data Science Capstone Project by IBM

CAO CHÁNH TRÍ



FRAUD DETECTION

- Dataset Overview
- Technology applied
- Descriptive and Exploratory Analysis
- Modeling
- Evaluating
- Deployment

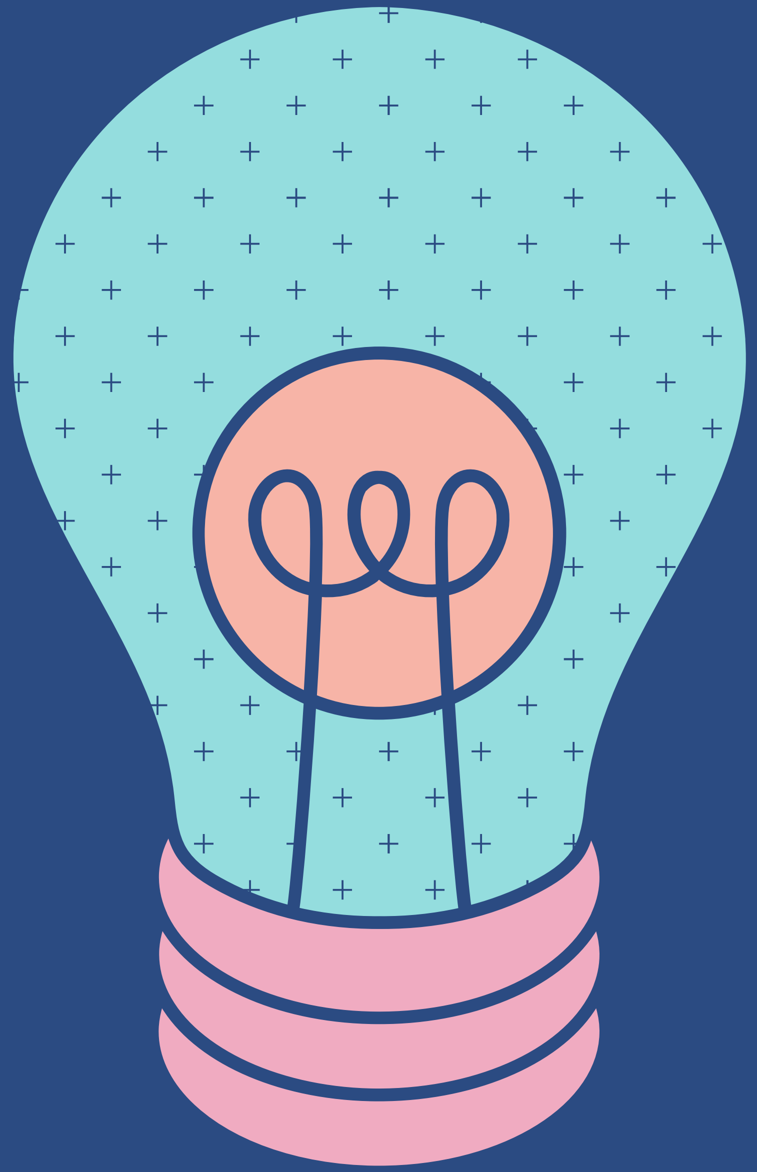
FRAUD DETECTION

Dataset Overview

THIS DATASET IS FICTIONAL AND IS TRYING TO SIMULATE REAL LIFE DETAILS. ANY SIMILARITY TO REAL LIFE CASES IS PURELY COINCIDENTAL.

The data is separated into 2 csv files:

- fraudTrain with 1296675 records
- fraudTest with 555719 records

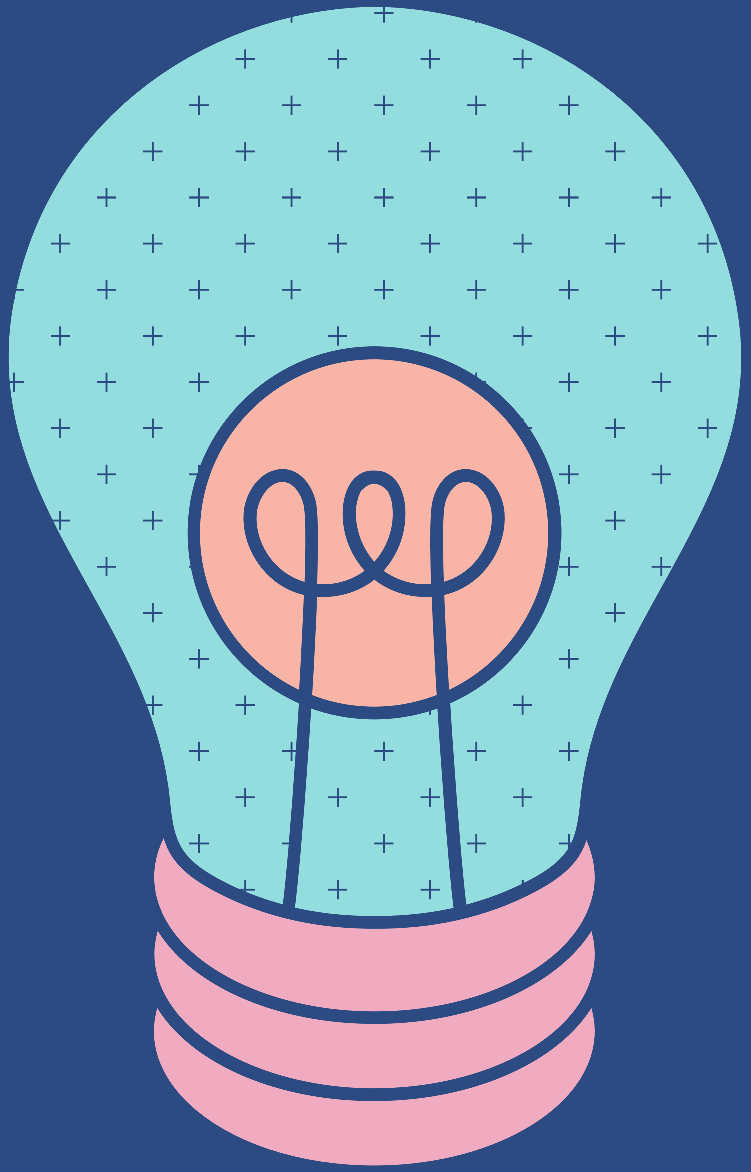


FRAUD DETECTION

Dataset Overview

Originaly, the data comes with 22 columns:

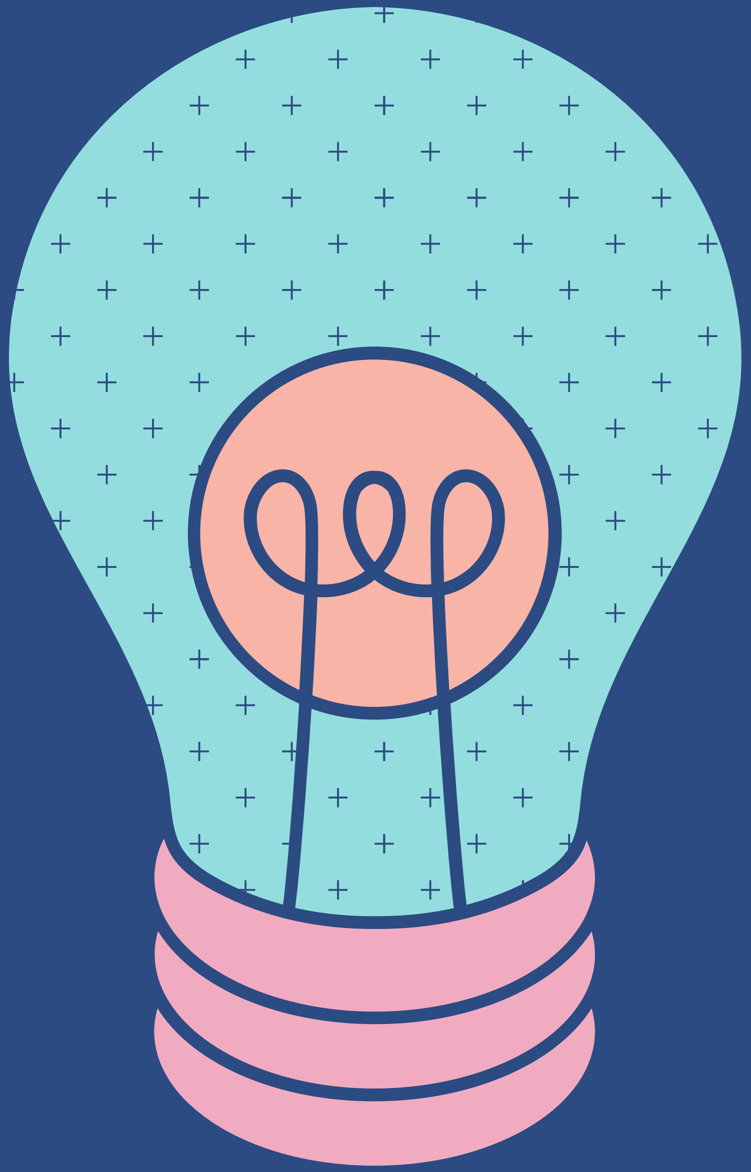
- trans_date_trans_time: The date and time of the transaction.
- cc_num: credit card number.
- merchant: Merchant who was getting paid.
- category: In what area does that merchant deal.
- amt: Amount of money in American Dollars.
- first: first name of the card holder.
- last: last name of the card holder.
- gender: Male or Female



FRAUD DETECTION

Dataset Overview

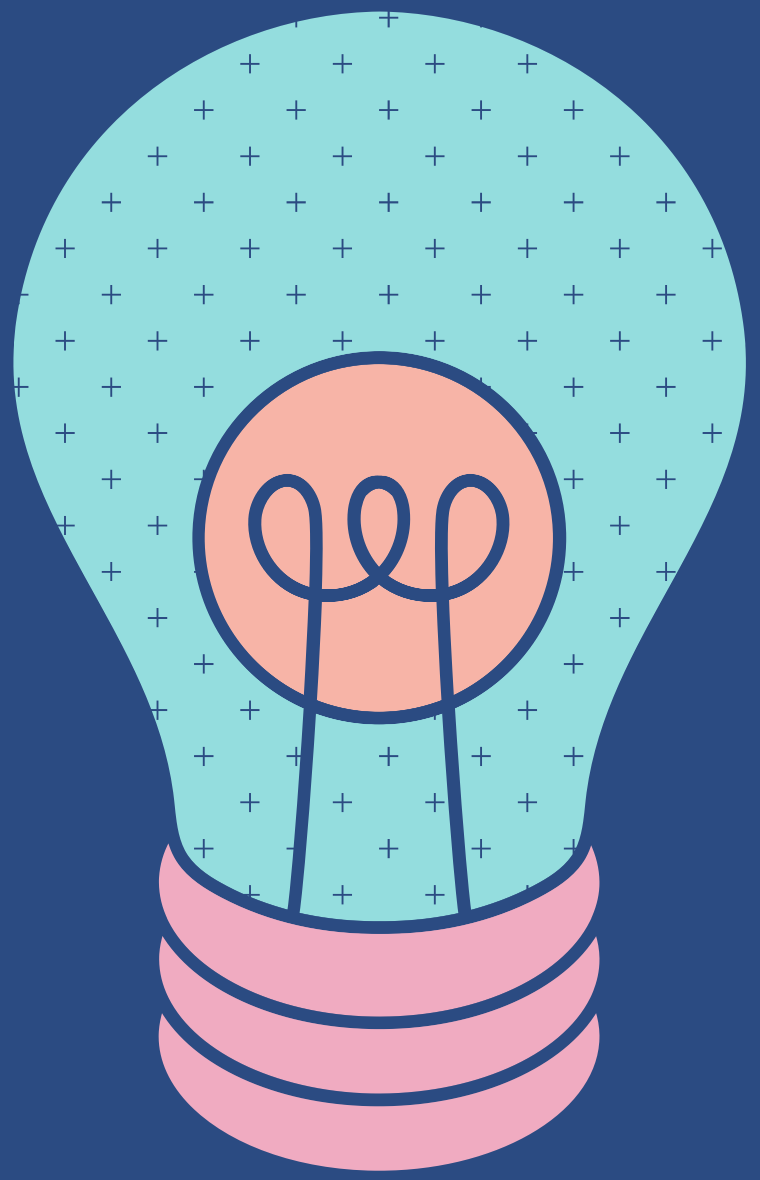
- street: Street of card holder residence
- city: city of card holder residence
- state: state of card holder residence
- zip: ZIP code of card holder residence
- lat: latitude of card holder
- long: longitude of card holder
- city_pop: Population of the city
- job: trade of the card holder



FRAUD DETECTION

Dataset Overview

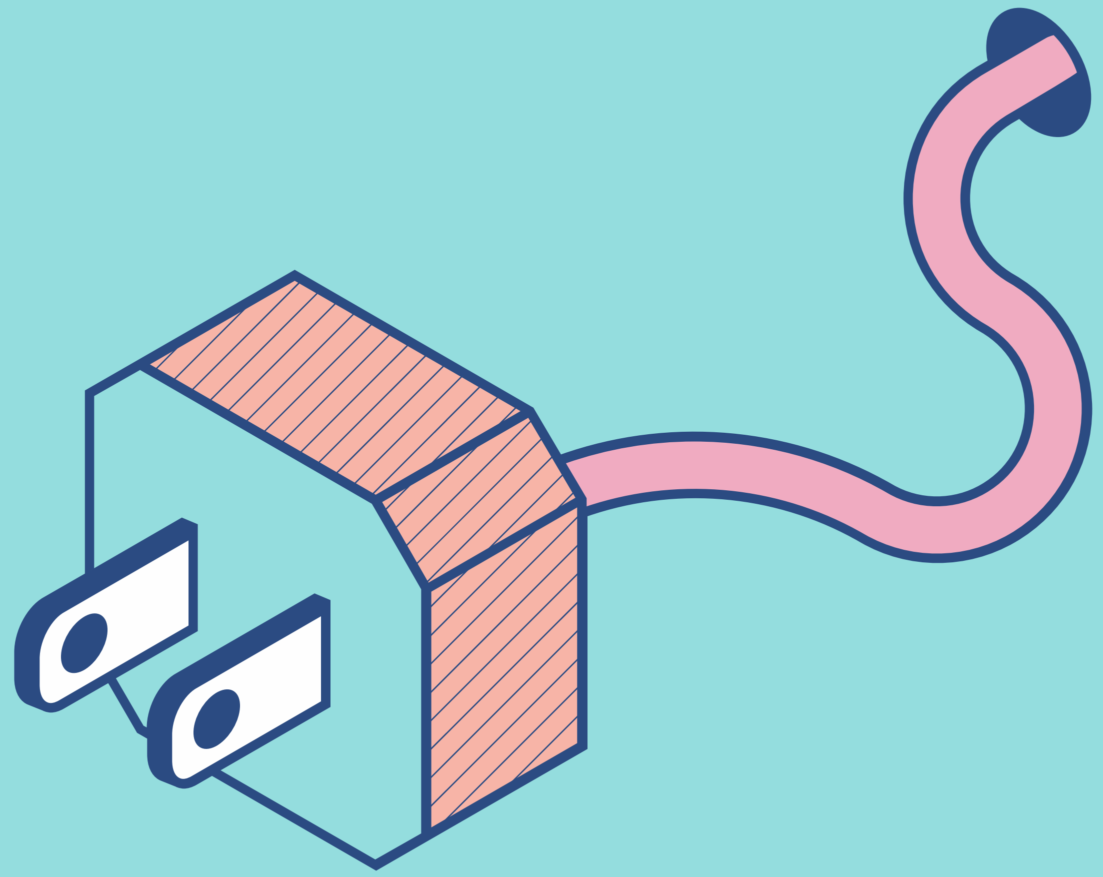
- dob: Date of birth of the card holder
- trans_num: Transaction ID
- unix_time: Unix time which is the time calculated since 1970
- merch_lat: latitude of the merchant
- merch_long:longitude of the merchant
- is_fraud (target): is fraud(1) or not(0)



FRAUD DETECTION

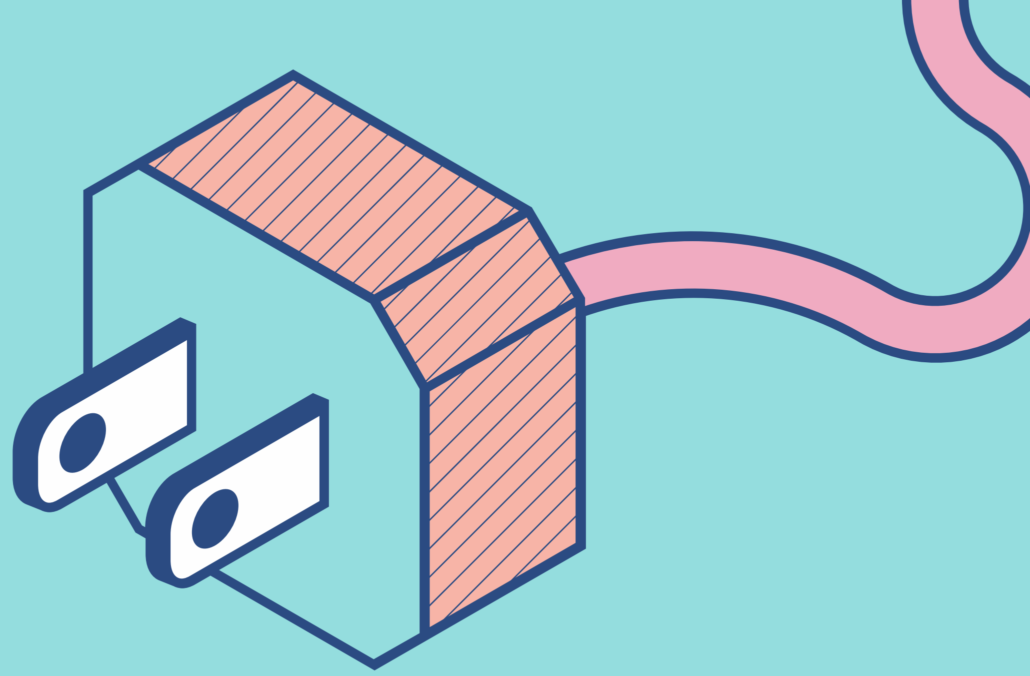
Technology Applied

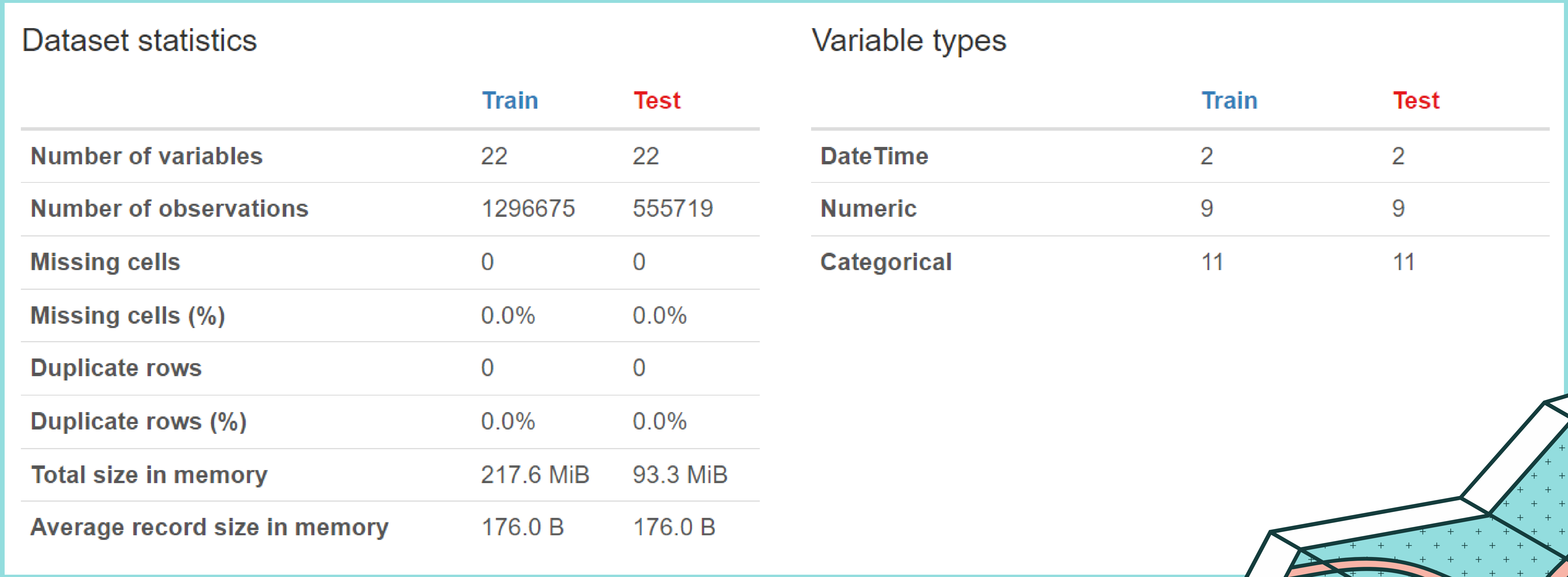
- Auto Descriptive Statistic: pandas profiling
- Visualization: Seaborn, Folium
- Model Creation: Pycarret, Xgboost, Tensorflow
- Deploy with Streamlit



FRAUD DETECTION

Descriptive and Exploratory Analysis





Dataset statistics

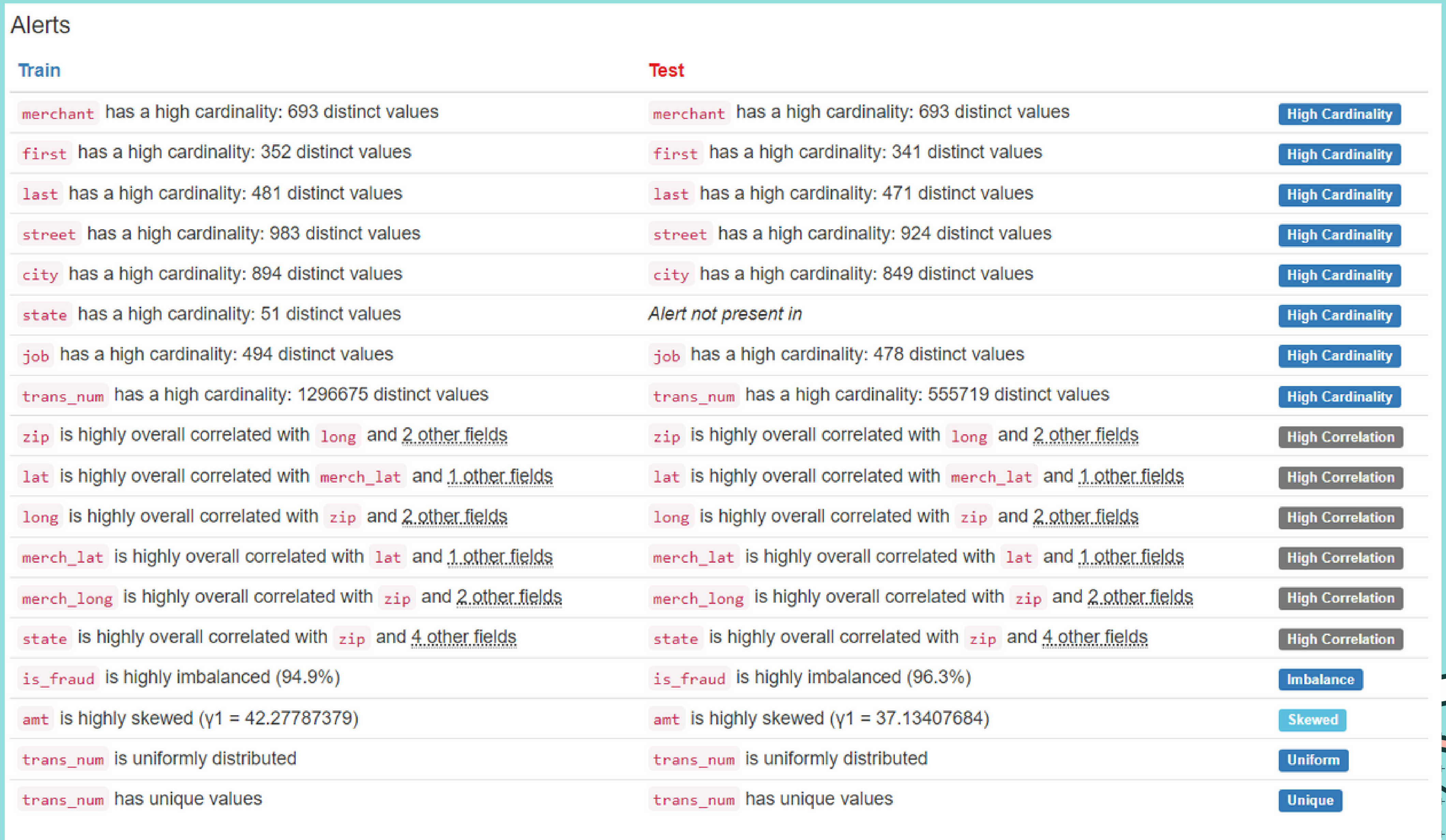
	Train	Test
Number of variables	22	22
Number of observations	1296675	555719
Missing cells	0	0
Missing cells (%)	0.0%	0.0%
Duplicate rows	0	0
Duplicate rows (%)	0.0%	0.0%
Total size in memory	217.6 MiB	93.3 MiB
Average record size in memory	176.0 B	176.0 B

Variable types

	Train	Test
DateTime	2	2
Numeric	9	9
Categorical	11	11

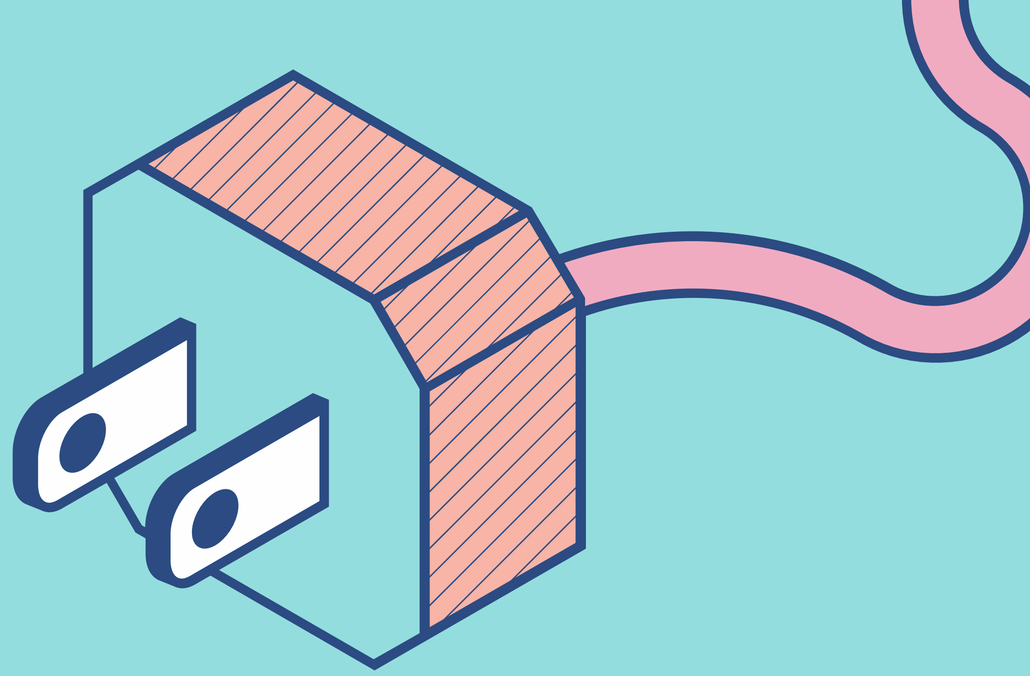
FRAUD DETECTION

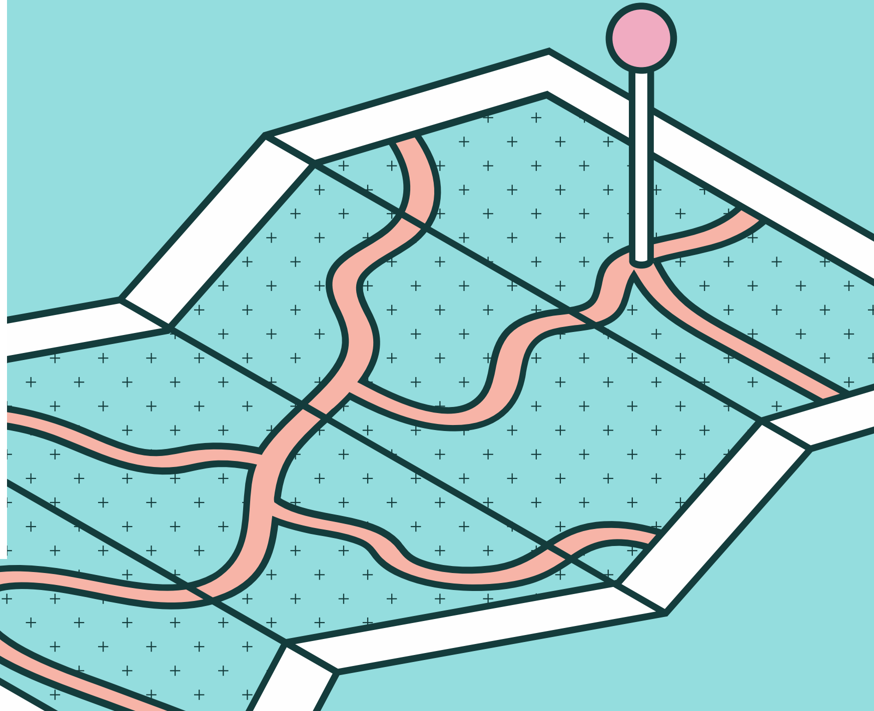
Descriptive and Exploratory Analysis

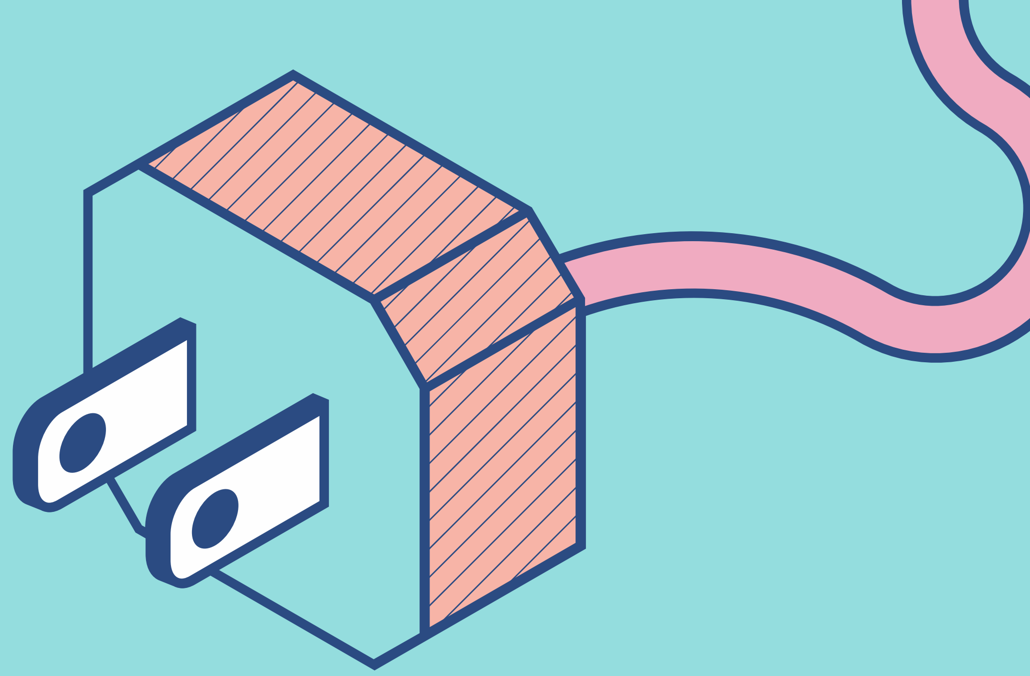


Alerts

Train	Test	
merchant has a high cardinality: 693 distinct values	merchant has a high cardinality: 693 distinct values	High Cardinality
first has a high cardinality: 352 distinct values	first has a high cardinality: 341 distinct values	High Cardinality
last has a high cardinality: 481 distinct values	last has a high cardinality: 471 distinct values	High Cardinality
street has a high cardinality: 983 distinct values	street has a high cardinality: 924 distinct values	High Cardinality
city has a high cardinality: 894 distinct values	city has a high cardinality: 849 distinct values	High Cardinality
state has a high cardinality: 51 distinct values	*Alert not present in*	High Cardinality
job has a high cardinality: 494 distinct values	job has a high cardinality: 478 distinct values	High Cardinality
trans_num has a high cardinality: 1296675 distinct values	trans_num has a high cardinality: 555719 distinct values	High Cardinality
zip is highly overall correlated with long and 2 other fields	zip is highly overall correlated with long and 2 other fields	High Correlation
lat is highly overall correlated with merch_lat and 1 other fields	lat is highly overall correlated with merch_lat and 1 other fields	High Correlation
long is highly overall correlated with zip and 2 other fields	long is highly overall correlated with zip and 2 other fields	High Correlation
merch_lat is highly overall correlated with lat and 1 other fields	merch_lat is highly overall correlated with lat and 1 other fields	High Correlation
merch_long is highly overall correlated with zip and 2 other fields	merch_long is highly overall correlated with zip and 2 other fields	High Correlation
state is highly overall correlated with zip and 4 other fields	state is highly overall correlated with zip and 4 other fields	High Correlation
is_fraud is highly imbalanced (94.9%)	is_fraud is highly imbalanced (96.3%)	Imbalance
amt is highly skewed (γ1 = 42.27787379)	amt is highly skewed (γ1 = 37.13407684)	Skewed
trans_num is uniformly distributed	trans_num is uniformly distributed	Uniform
trans_num has unique values	trans_num has unique values	Unique

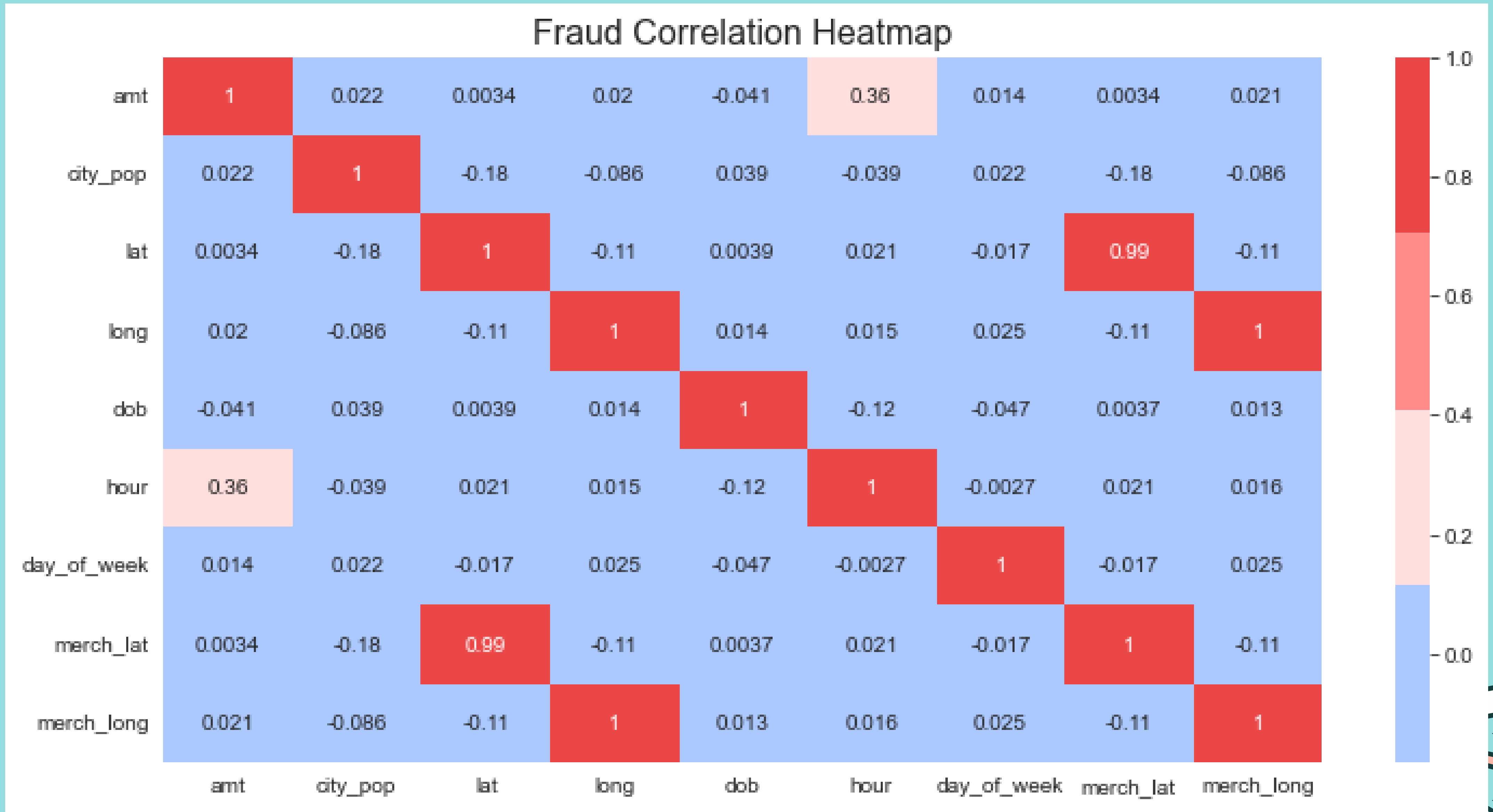






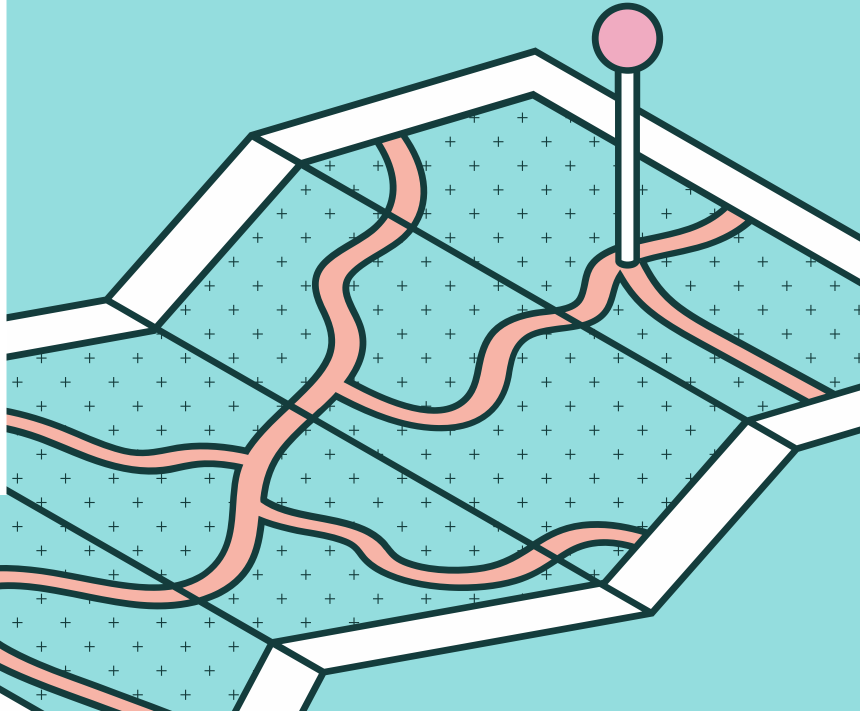
FRAUD DETECTION

Descriptive and Exploratory Analysis



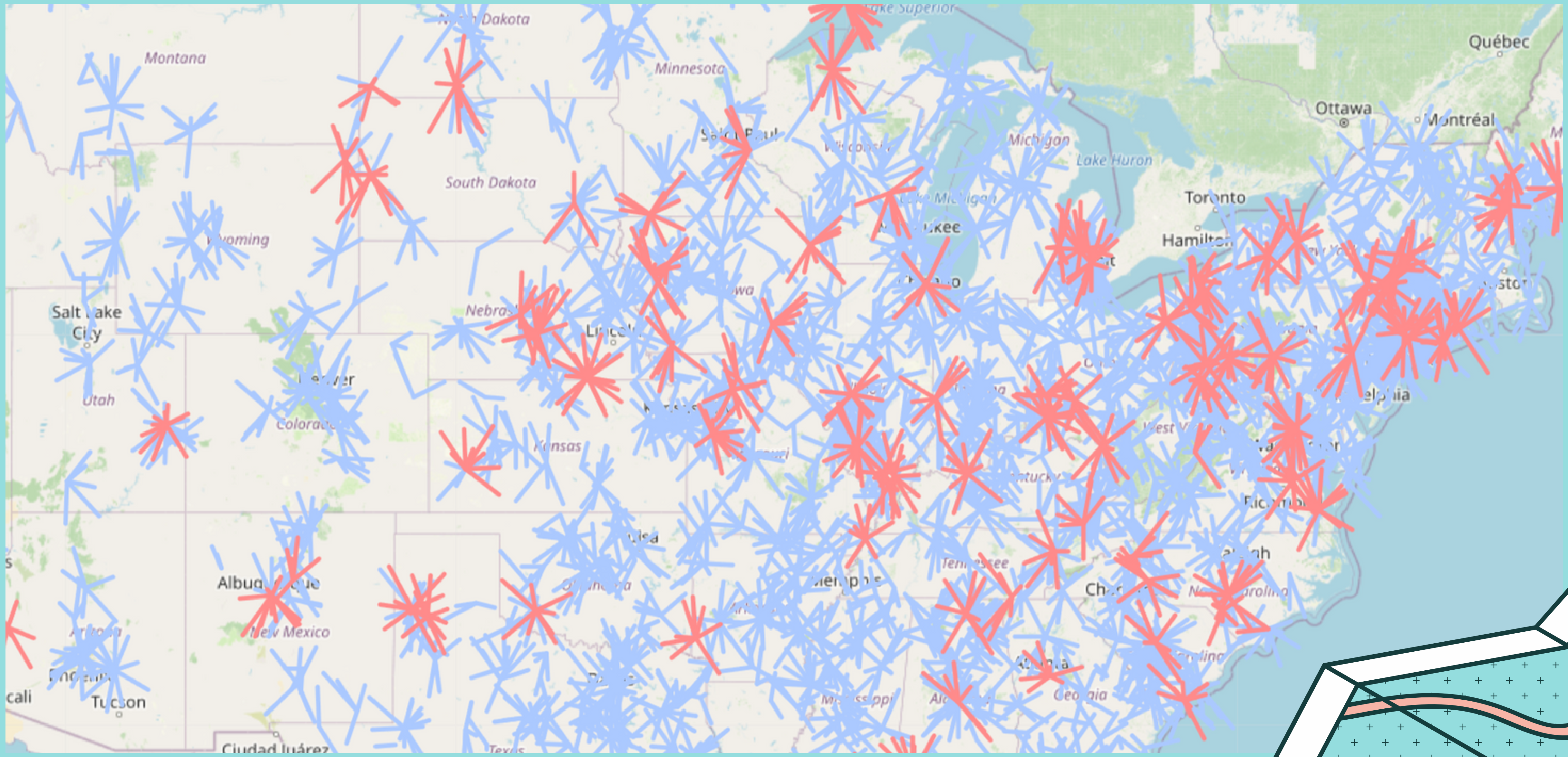
Fraud Correlation Heatmap

	amt	city_pop	lat	long	dob	hour	day_of_week	merch_lat	merch_long
amt	1	0.022	0.0034	0.02	-0.041	0.36	0.014	0.0034	0.021
city_pop	0.022	1	-0.18	-0.086	0.039	-0.039	0.022	-0.18	-0.086
lat	0.0034	-0.18	1	-0.11	0.0039	0.021	-0.017	0.99	-0.11
long	0.02	-0.086	-0.11	1	0.014	0.015	0.025	-0.11	1
dob	-0.041	0.039	0.0039	0.014	1	-0.12	-0.047	0.0037	0.013
hour	0.36	-0.039	0.021	0.015	-0.12	1	-0.0027	0.021	0.016
day_of_week	0.014	0.022	-0.017	0.025	-0.047	-0.0027	1	-0.017	0.025
merch_lat	0.0034	-0.18	0.99	-0.11	0.0037	0.021	-0.017	1	-0.11
merch_long	0.021	-0.086	-0.11	1	0.013	0.016	0.025	-0.11	1



FRAUD DETECTION

Descriptive and Exploratory Analysis



FRAUD DETECTION

Feature Engineering

1 Subset the meaningful columns:
'category', 'amt', 'gender', 'city_pop', 'lat', 'long'

2 Introduce 'late_hour' and 'early_hour'

- Late_hour: transaction after 10pm
- Early_hour: transaction before 3am

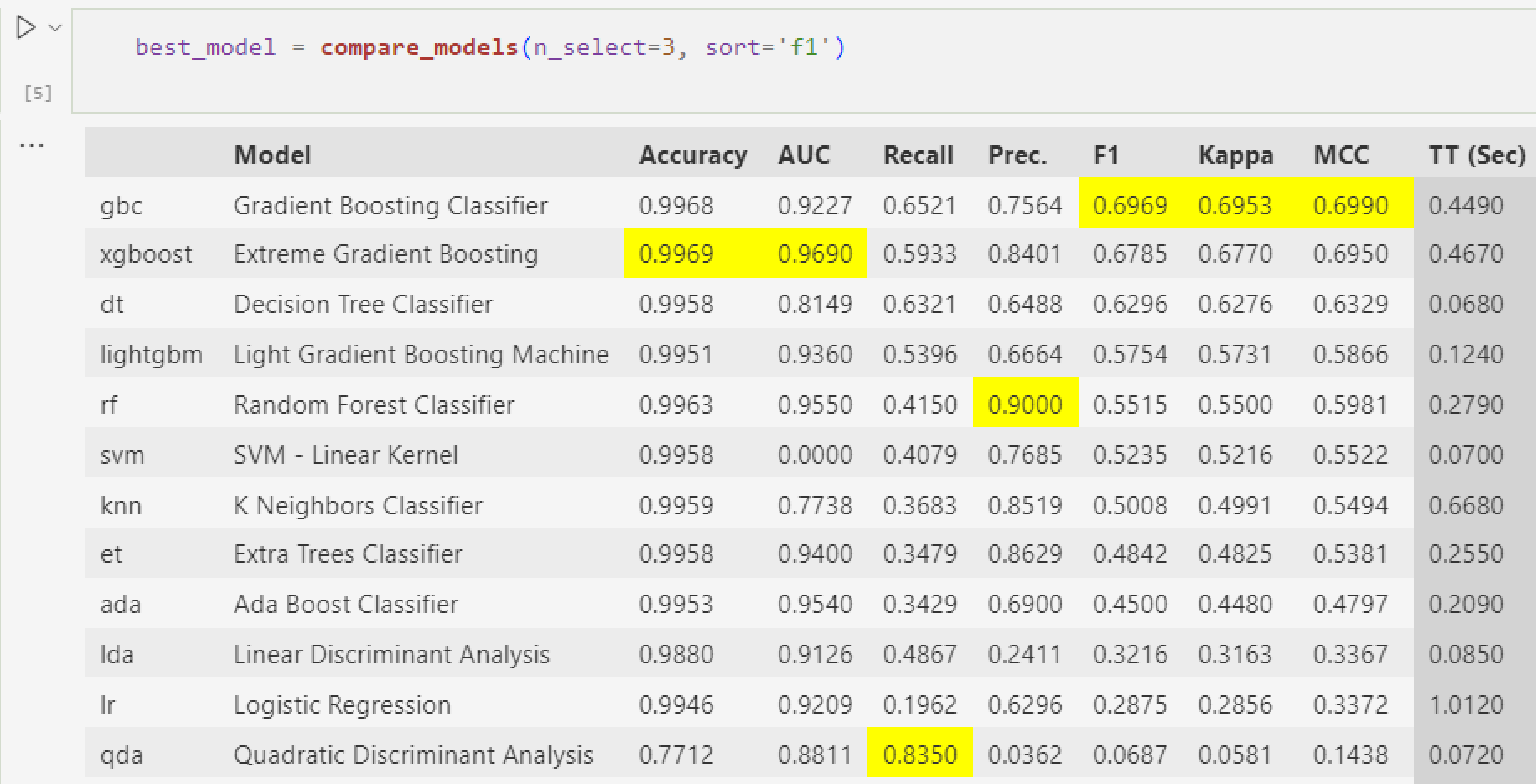
3 Introduce 'elderly' and 'young'

- Elderly: whose dob before 1960
- Young: whose dob after 1990

Fraud Detection

Modeling

AutoML with Pycarret for model suggestions



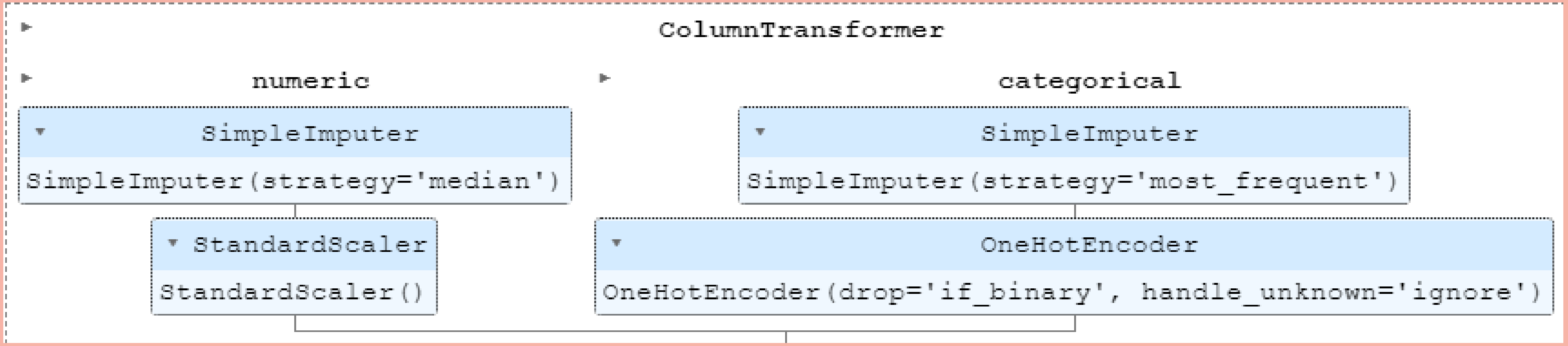
```
best_model = compare_models(n_select=3, sort='f1')
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9968	0.9227	0.6521	0.7564	0.6969	0.6953	0.6990	0.4490
xgboost	Extreme Gradient Boosting	0.9969	0.9690	0.5933	0.8401	0.6785	0.6770	0.6950	0.4670
dt	Decision Tree Classifier	0.9958	0.8149	0.6321	0.6488	0.6296	0.6276	0.6329	0.0680
lightgbm	Light Gradient Boosting Machine	0.9951	0.9360	0.5396	0.6664	0.5754	0.5731	0.5866	0.1240
rf	Random Forest Classifier	0.9963	0.9550	0.4150	0.9000	0.5515	0.5500	0.5981	0.2790
svm	SVM - Linear Kernel	0.9958	0.0000	0.4079	0.7685	0.5235	0.5216	0.5522	0.0700
knn	K Neighbors Classifier	0.9959	0.7738	0.3683	0.8519	0.5008	0.4991	0.5494	0.6680
et	Extra Trees Classifier	0.9958	0.9400	0.3479	0.8629	0.4842	0.4825	0.5381	0.2550
ada	Ada Boost Classifier	0.9953	0.9540	0.3429	0.6900	0.4500	0.4480	0.4797	0.2090
lda	Linear Discriminant Analysis	0.9880	0.9126	0.4867	0.2411	0.3216	0.3163	0.3367	0.0850
lr	Logistic Regression	0.9946	0.9209	0.1962	0.6296	0.2875	0.2856	0.3372	1.0120
qda	Quadratic Discriminant Analysis	0.7712	0.8811	0.8350	0.0362	0.0687	0.0581	0.1438	0.0720

Fraud Detection

Modeling

Build my own preprocess pipeline



ColumnTransformer

numeric

SimpleImputer

SimpleImputer(strategy='median')

StandardScaler

StandardScaler()

categorical

SimpleImputer

SimpleImputer(strategy='most_frequent')

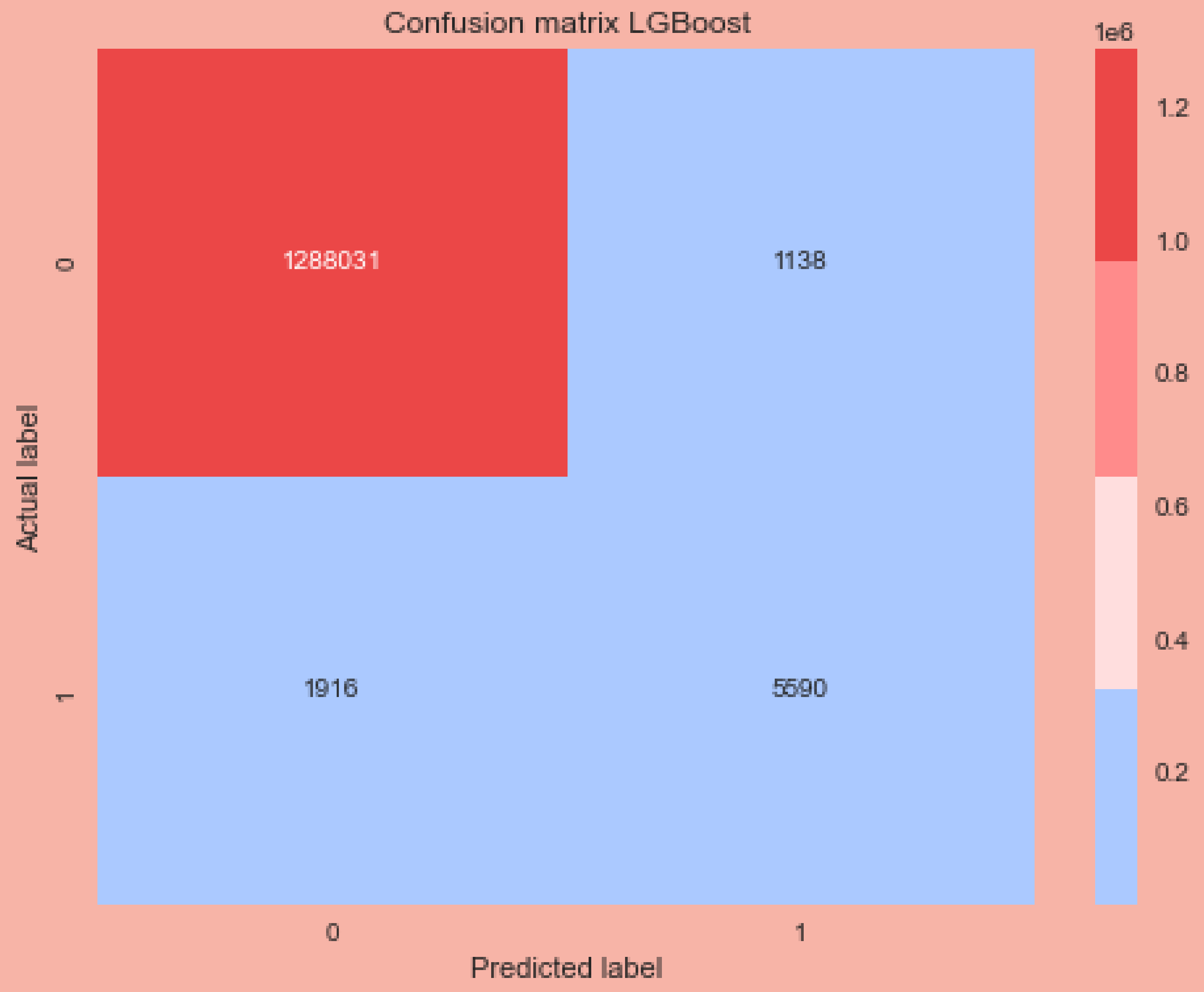
OneHotEncoder

OneHotEncoder(drop='if_binary', handle_unknown='ignore')

Fraud Detection

Modeling

Light Gradient Boosting Classifier

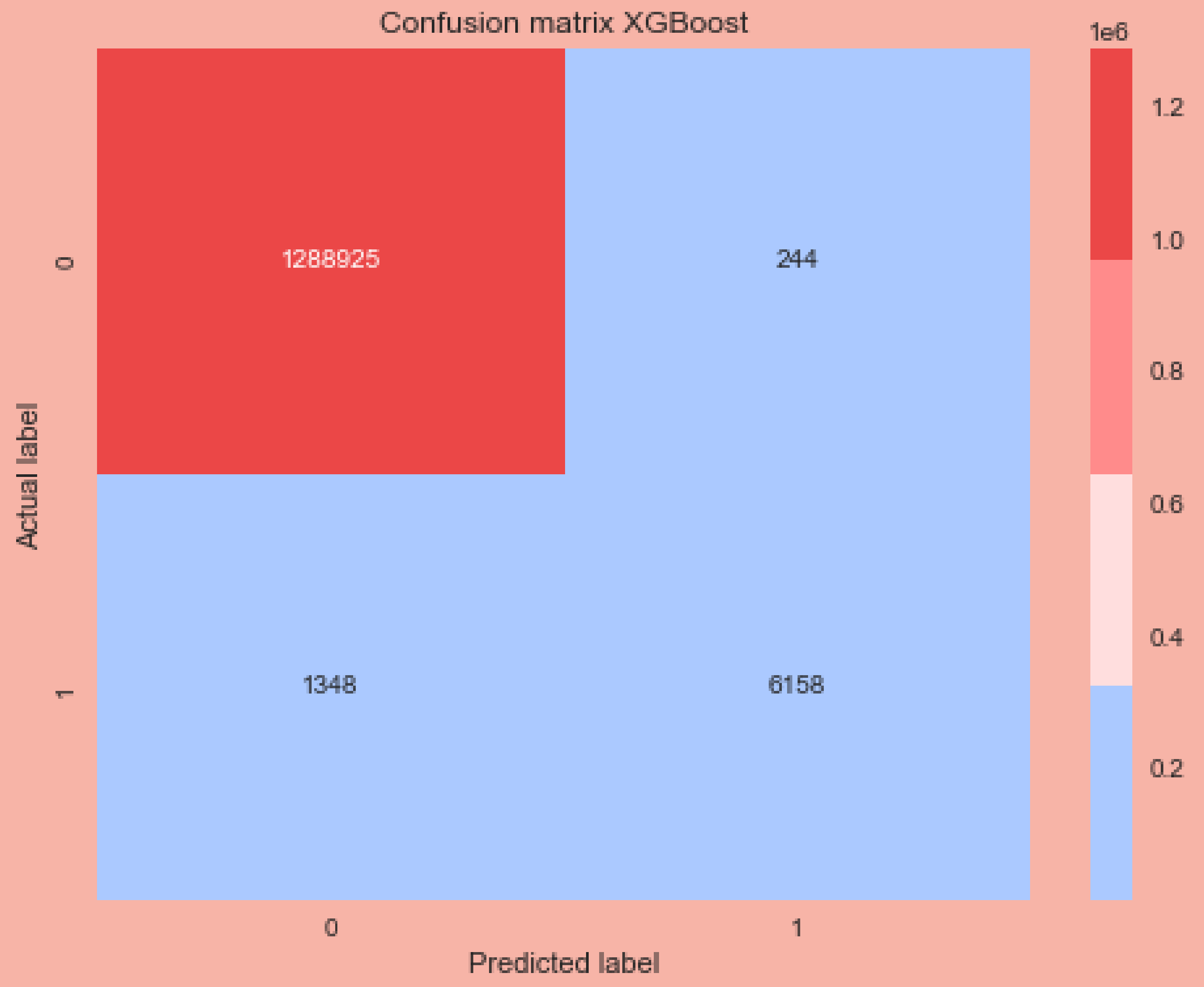


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.83	0.74	0.79

Fraud Detection

Modeling

XGBoost Classifier

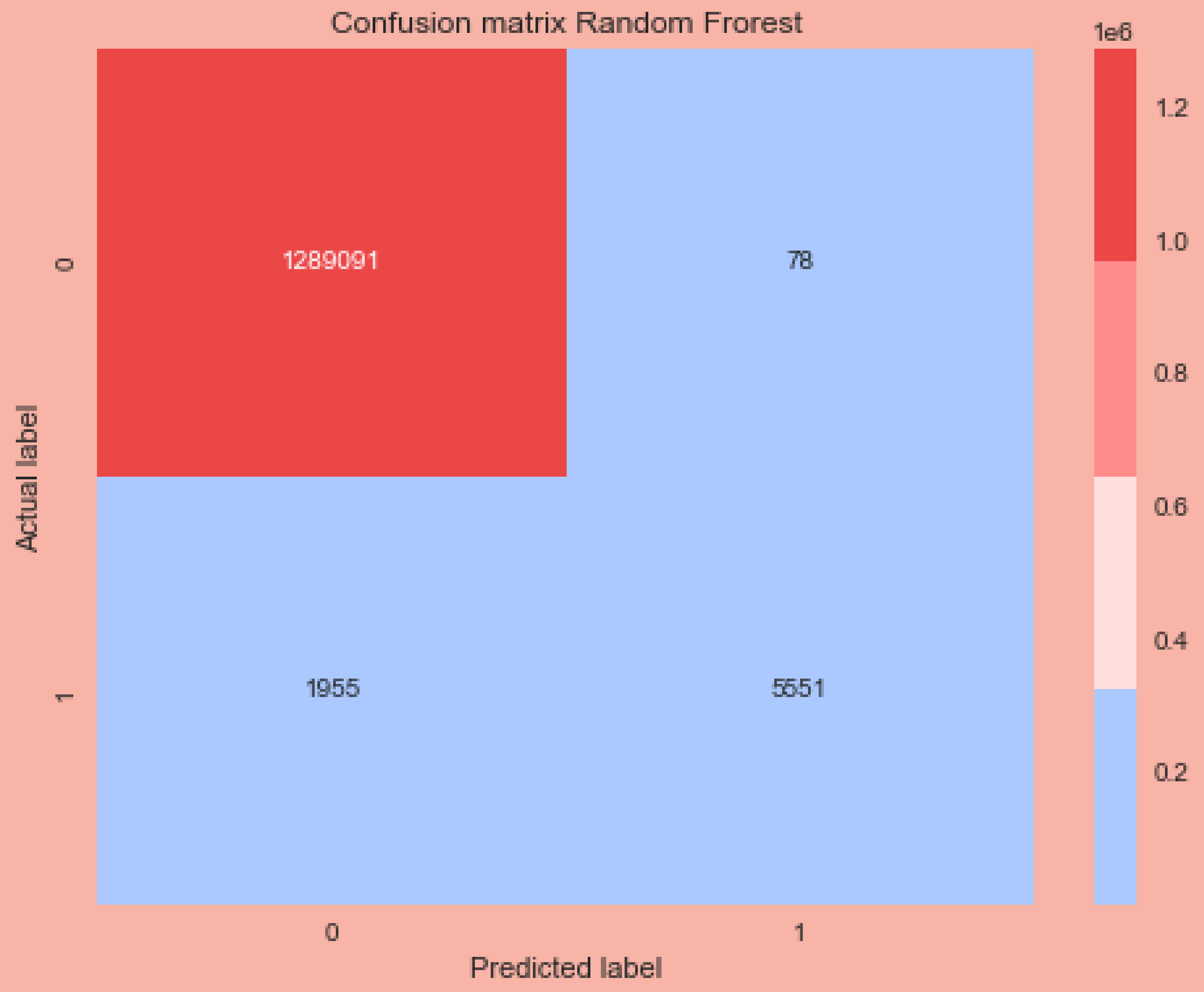


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.96	0.82	0.89

Fraud Detection

Modeling

Random Forrest Classifier

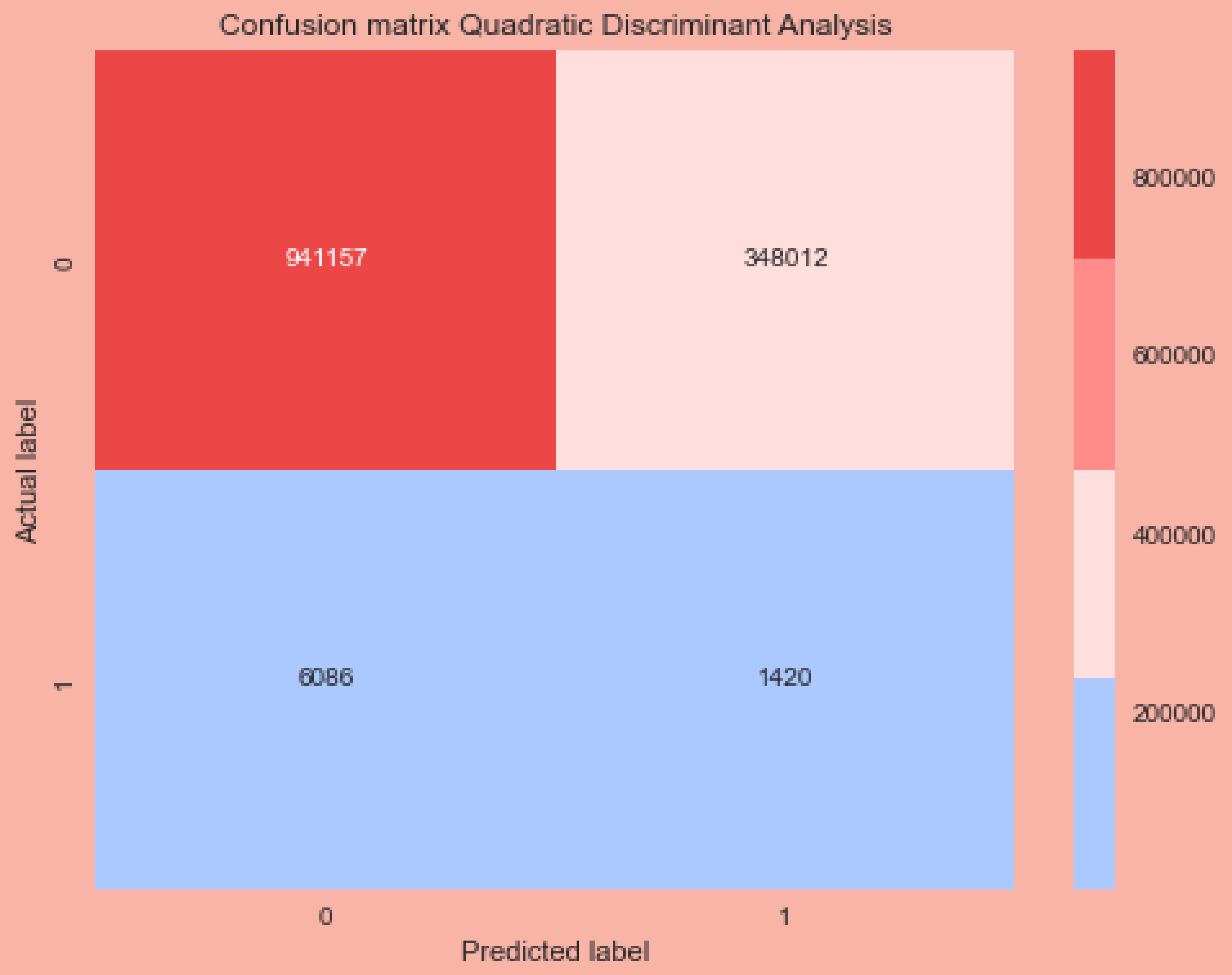


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.99	0.74	0.85

Fraud Detection

Modeling

Quadratic Discriminant Analysis

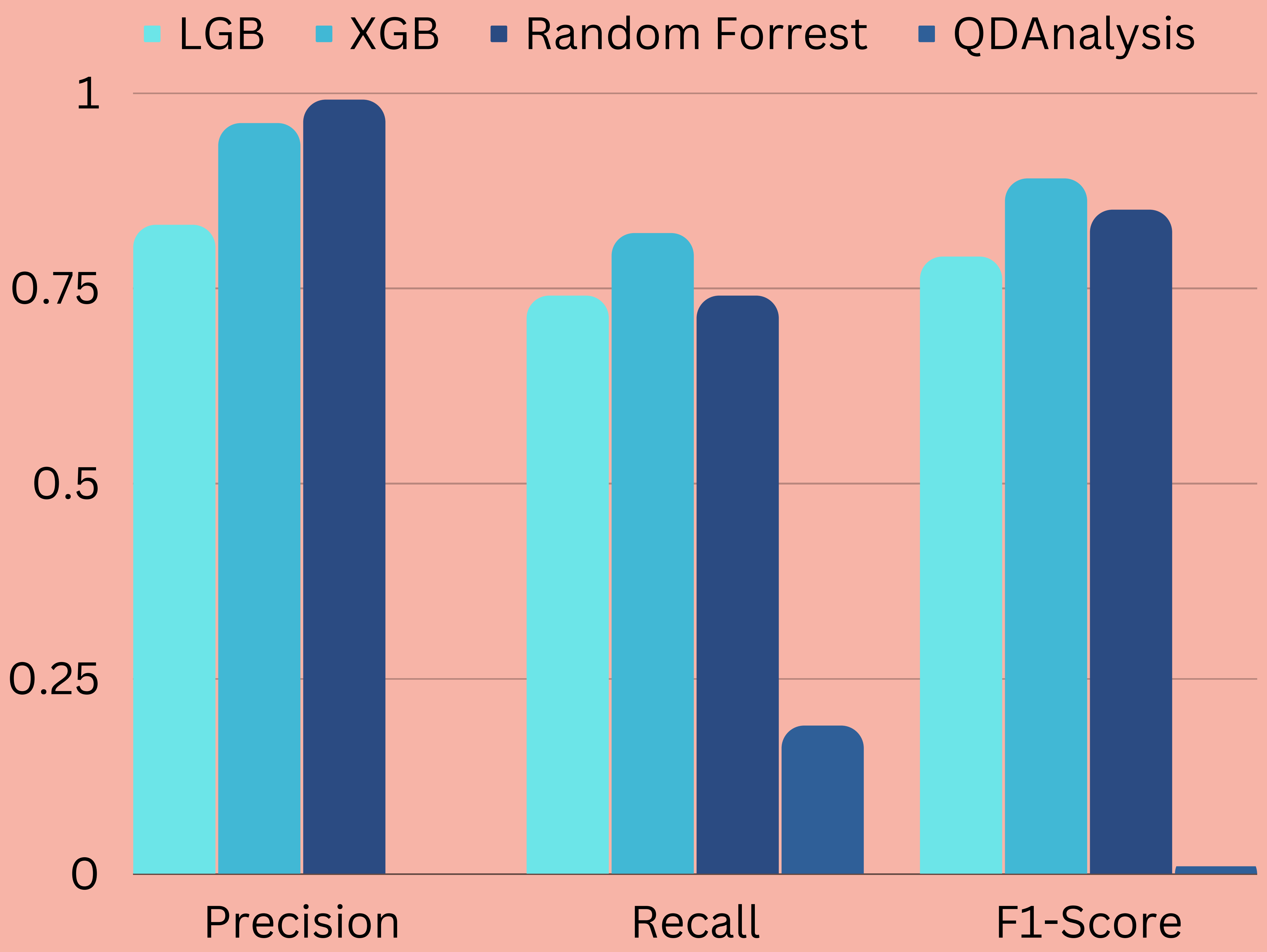


	Precision	Recall	F1 score
Normal	0.99	0.73	0.84
Fraud	0.00	0.19	0.01

Fraud Detection

Modeling

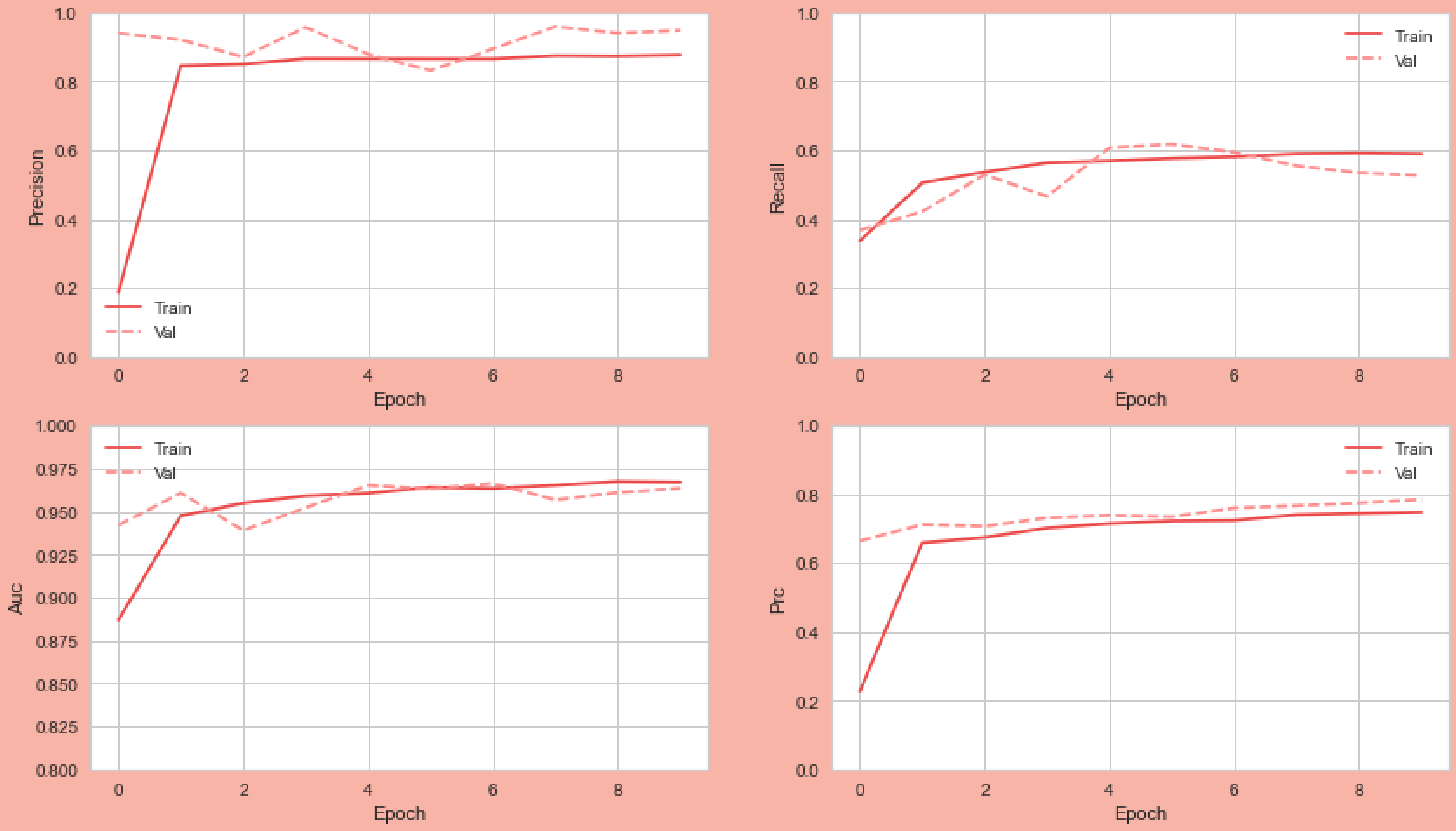
Basic ML Models



Fraud Detection

Modeling

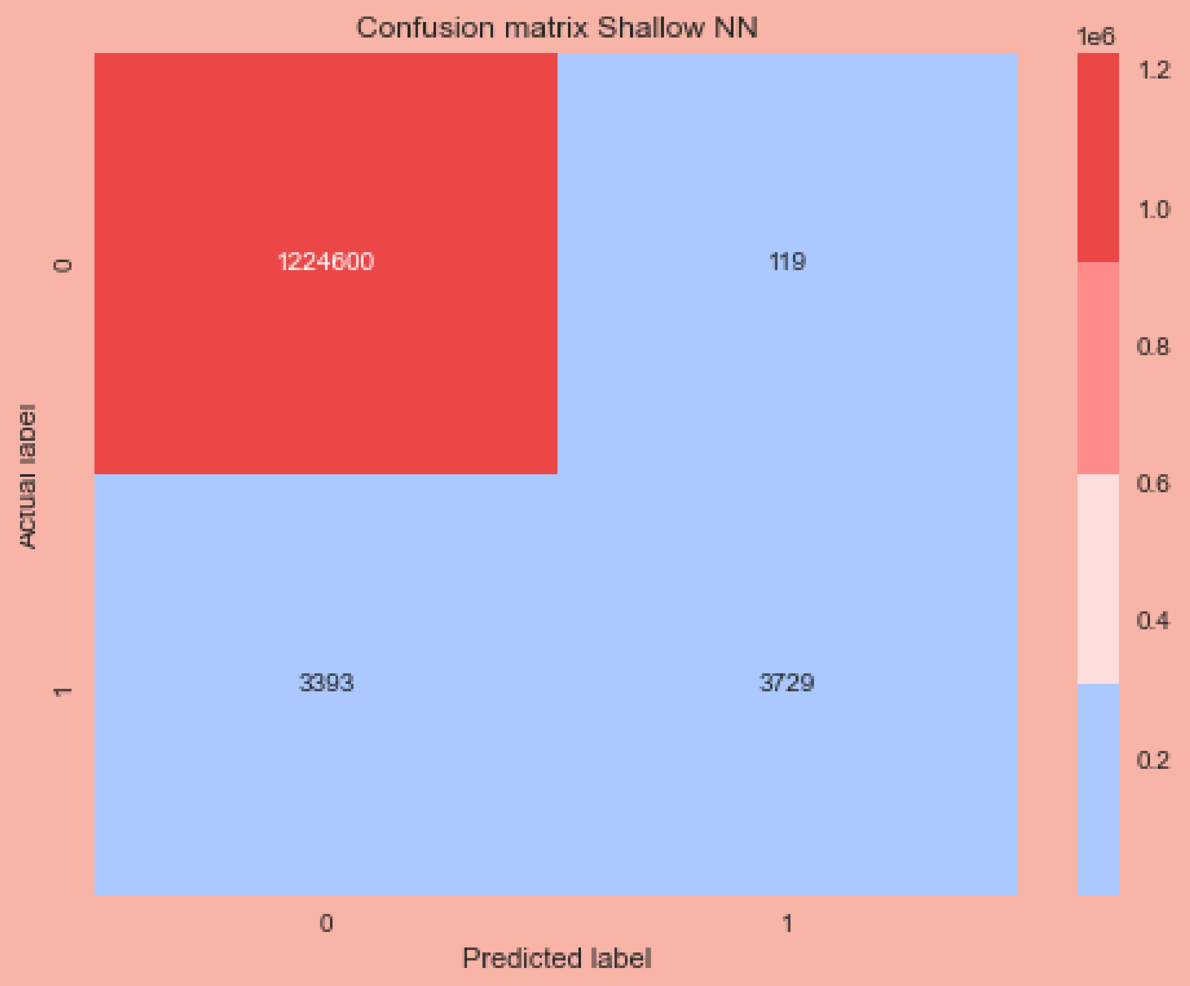
Shallow Neural Net (289 params)



Fraud Detection

Modeling

Shallow Neural Net (289 params)

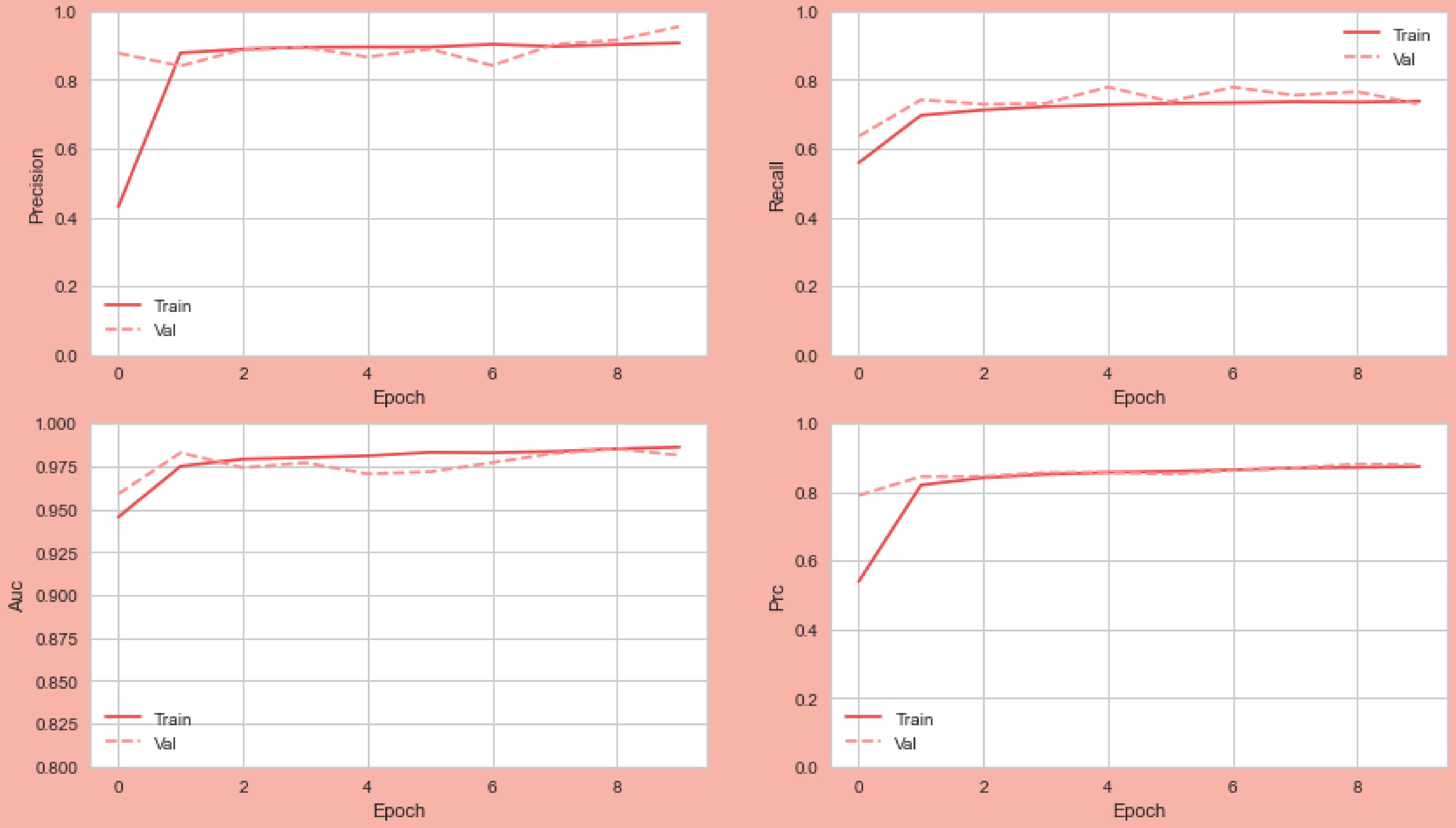


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.97	0.52	0.68

Fraud Detection

Modeling

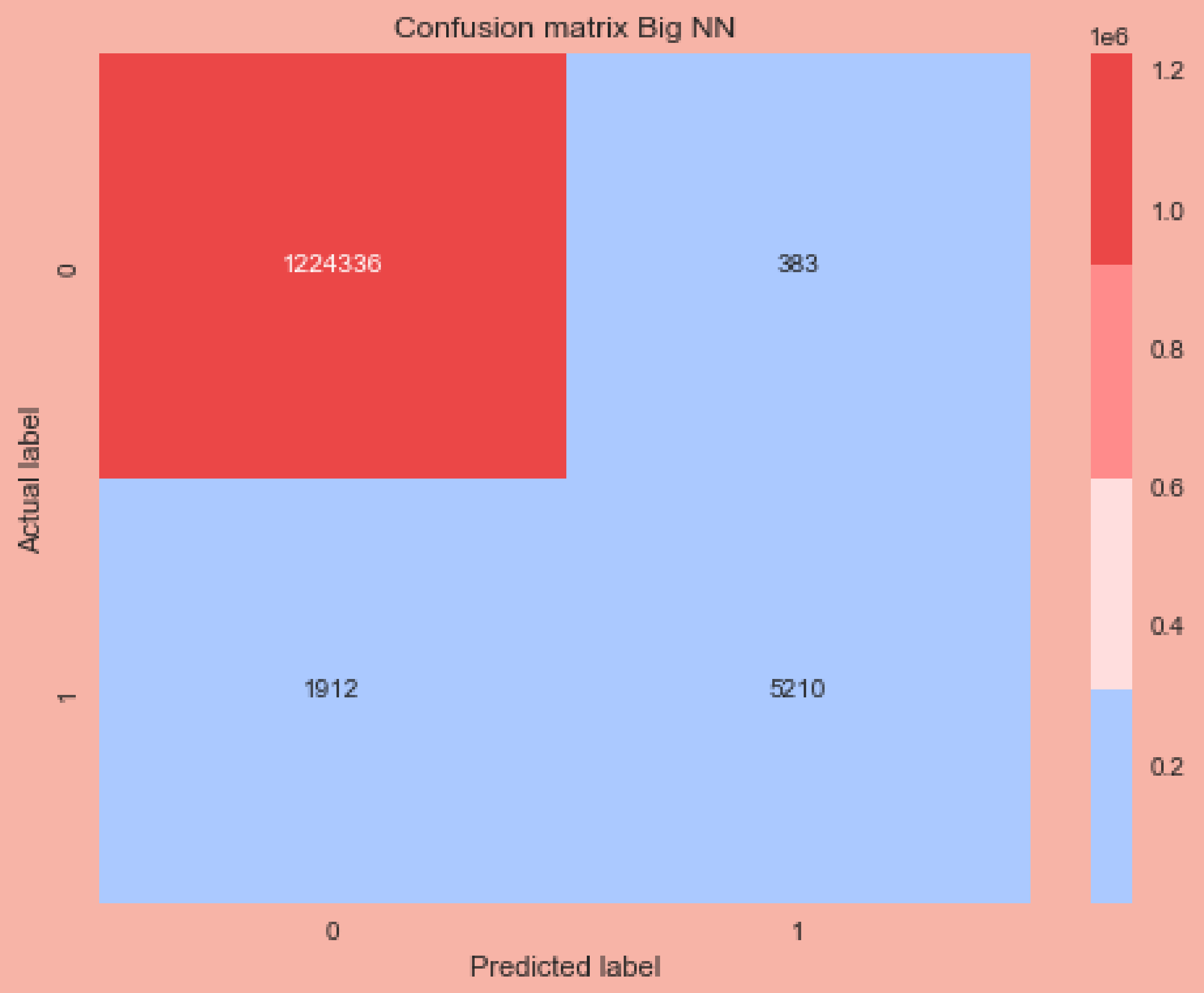
Bigger Neural Net (6,721 params)



Fraud Detection

Modeling

Bigger Neural Net (6,721 params)

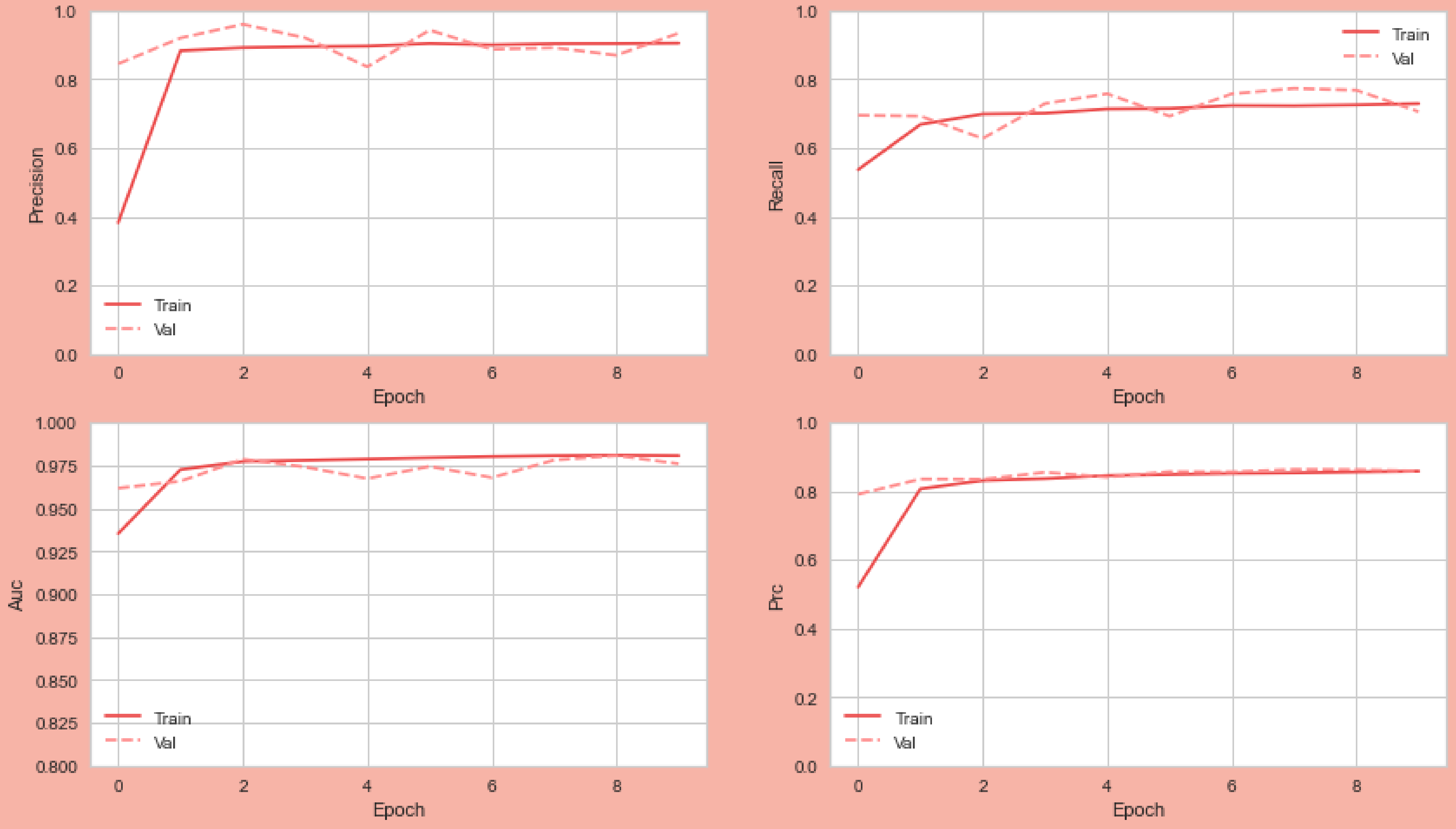


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.93	0.73	0.82

Fraud Detection

Modeling

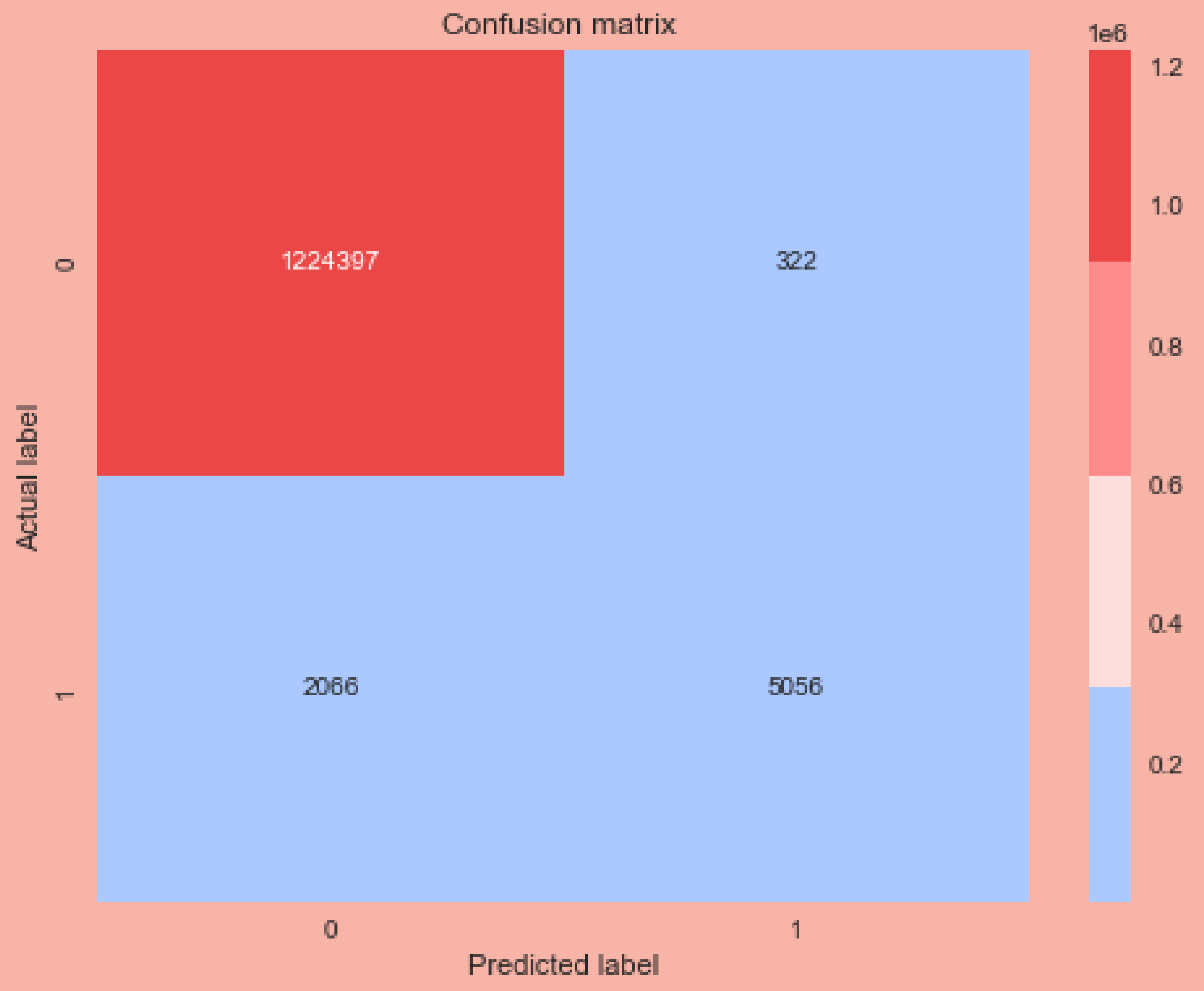
Deeper Neural Net (1,249 params)



Fraud Detection

Modeling

Deeper Neural Net (1,249 params)

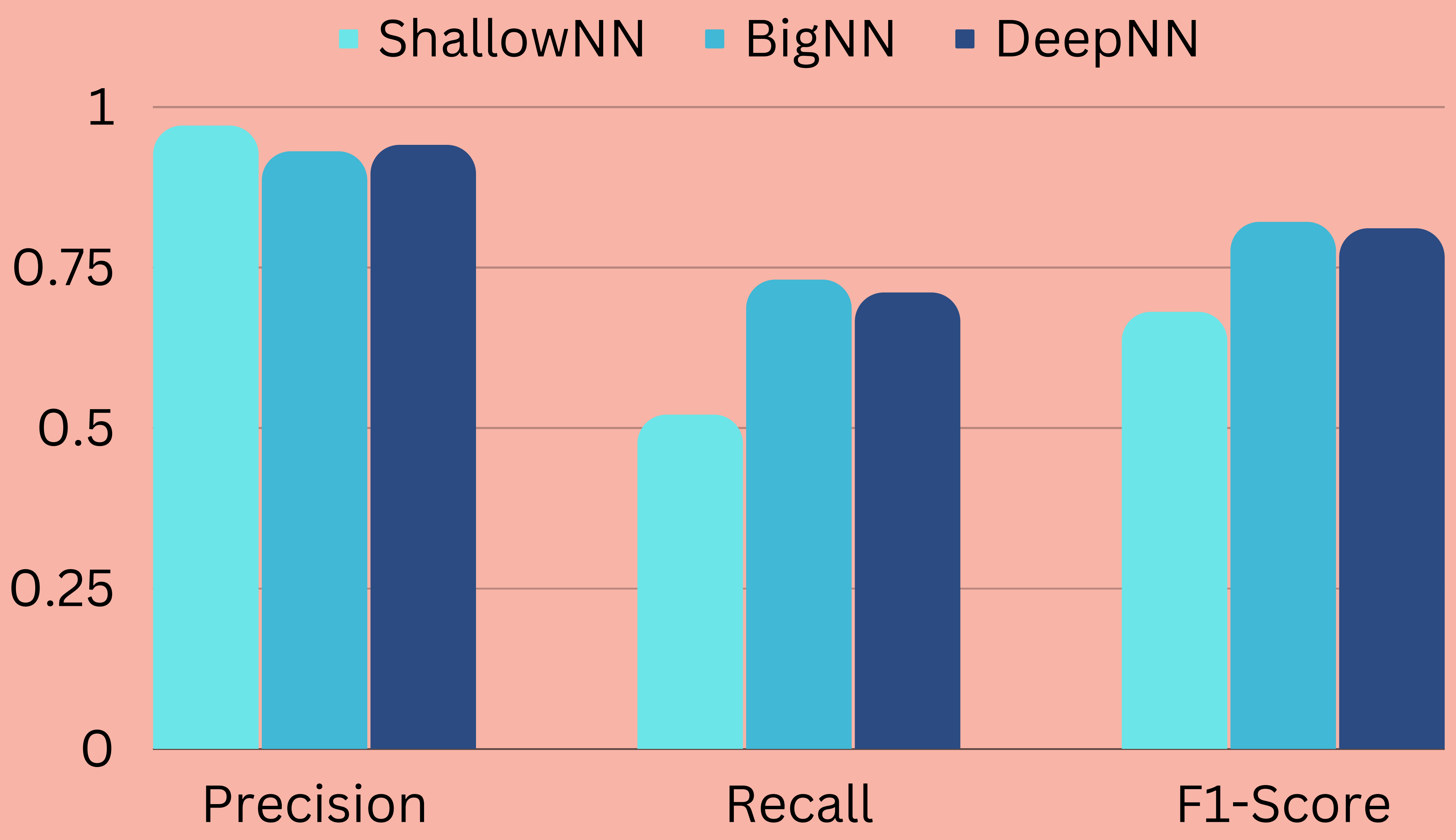


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.94	0.71	0.81

Fraud Detection

Modeling

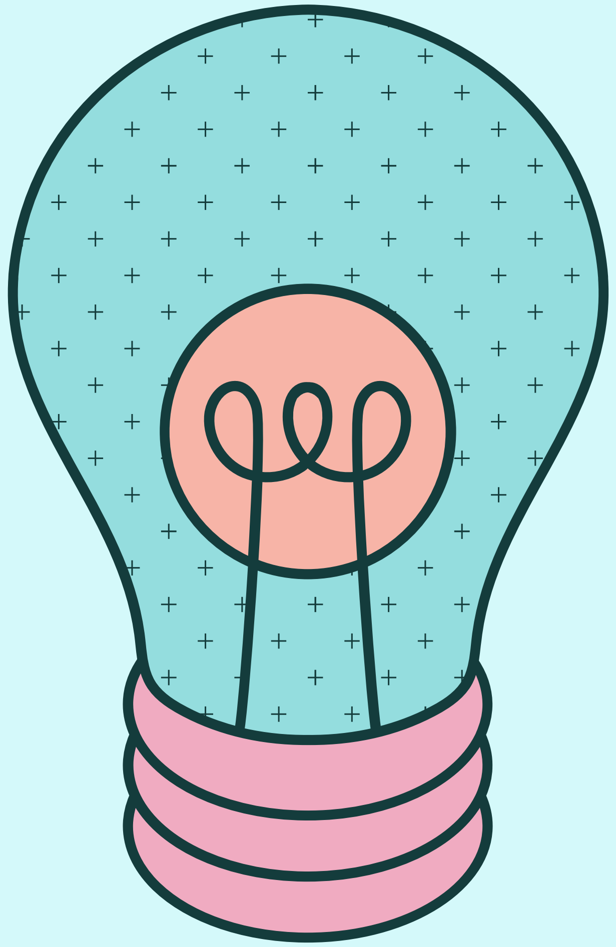
Neural Networks

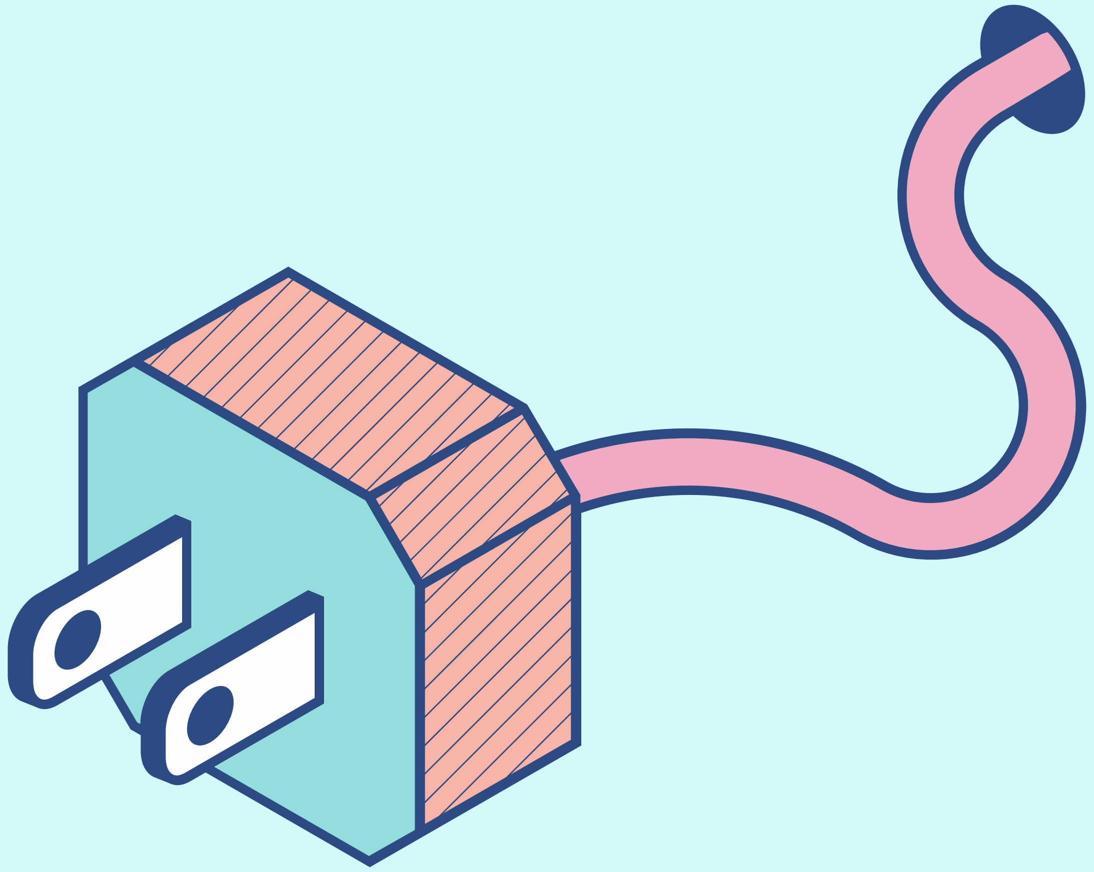


Total Parameters		
289	6,721	1,249

Fraud Detection

Tuning and Evaluating

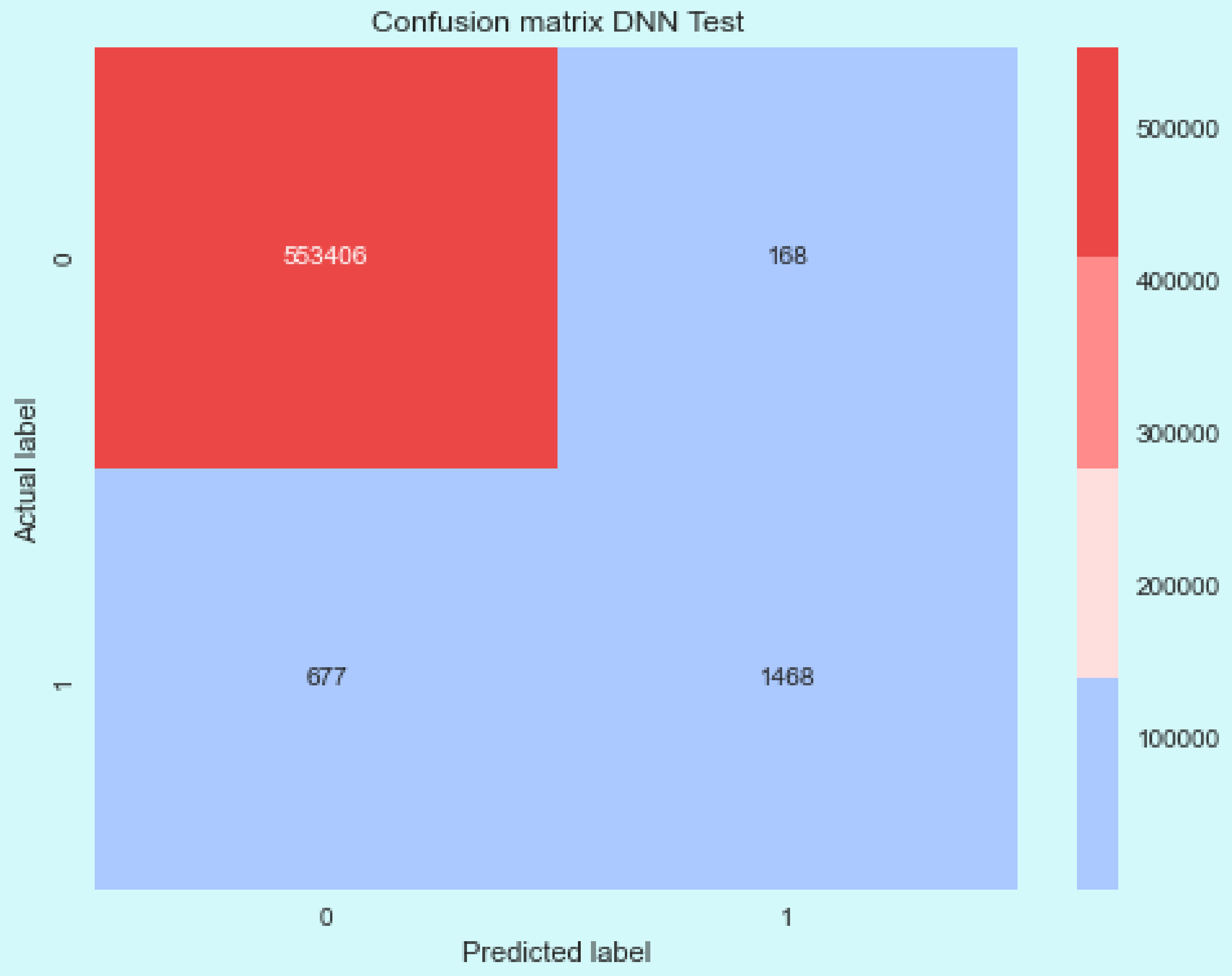




Fraud Detection

Tuning and Evaluating

DeepNN:
insignifficant improvement

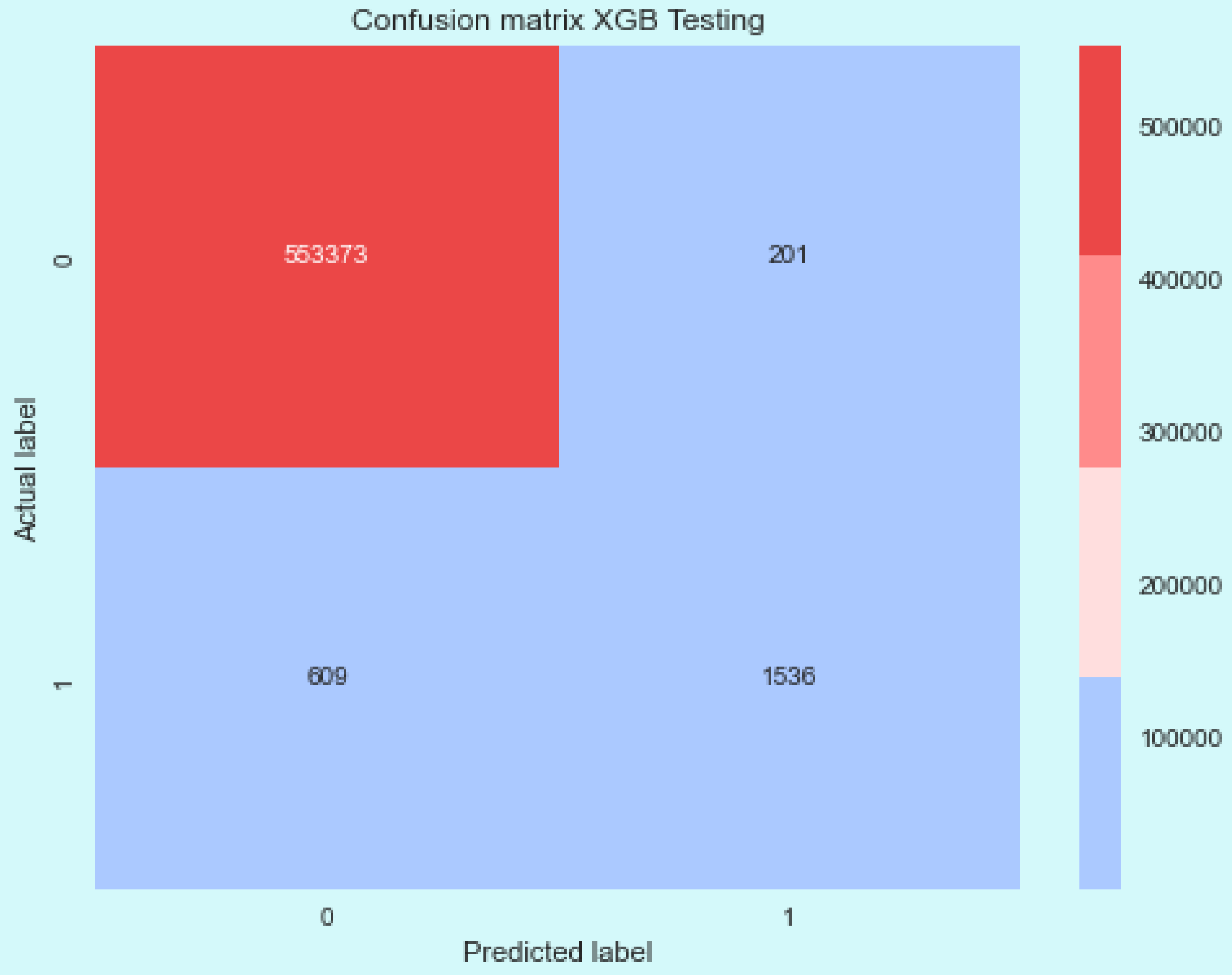


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.90	0.68	0.78

Fraud Detection

Tuning and Evaluating

XGBoost:
Avoiding overfitting

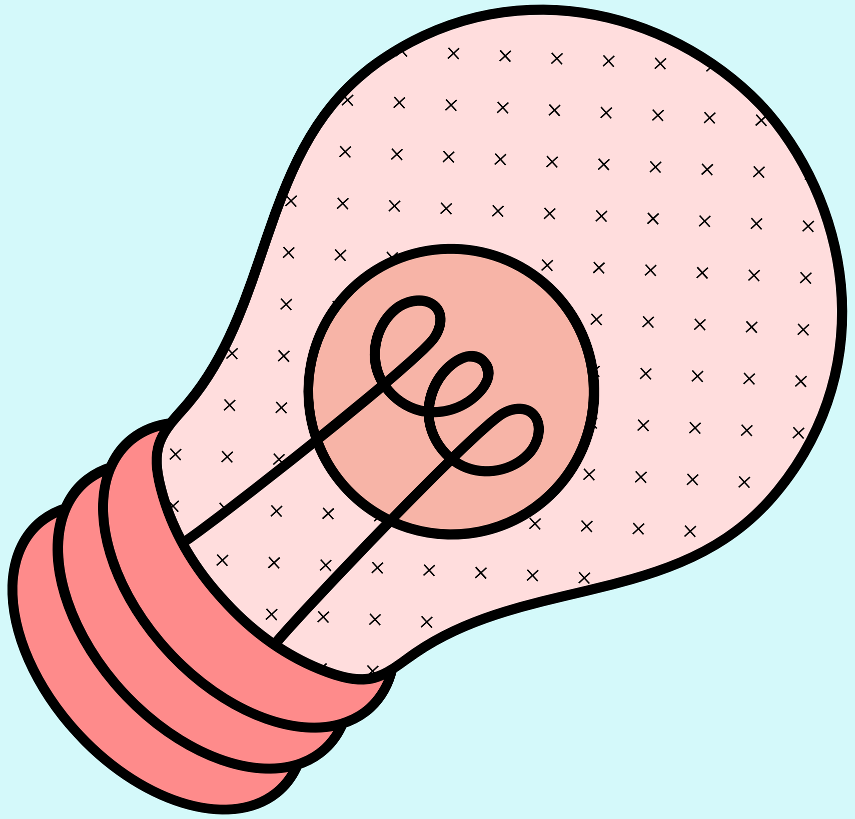


	Precision	Recall	F1 score
Normal	1.00	1.00	1.00
Fraud	0.88	0.72	0.79

Fraud Detection

Tuning and Evaluating

OVERALL RESULT



MACRO AVG	Precision	Recall	F1 score
DeepNN	0.95	0.84	0.89
XGBoost	0.94	0.86	0.90

Fraud Detection

Deployment with Streamlit