Synthetic Financial Datasets For Fraud Detection

University of Colorado Boulder

Contents

1. Problem Statement
2. Related Work
3. Proposes Work
4. Evaluation
5. Conclusion

1. Problem Statement

- Digital Fraud
- The dilemma making financial datasets publicly available
- How Synthetic data can help overcome these barriers

2. Related Work

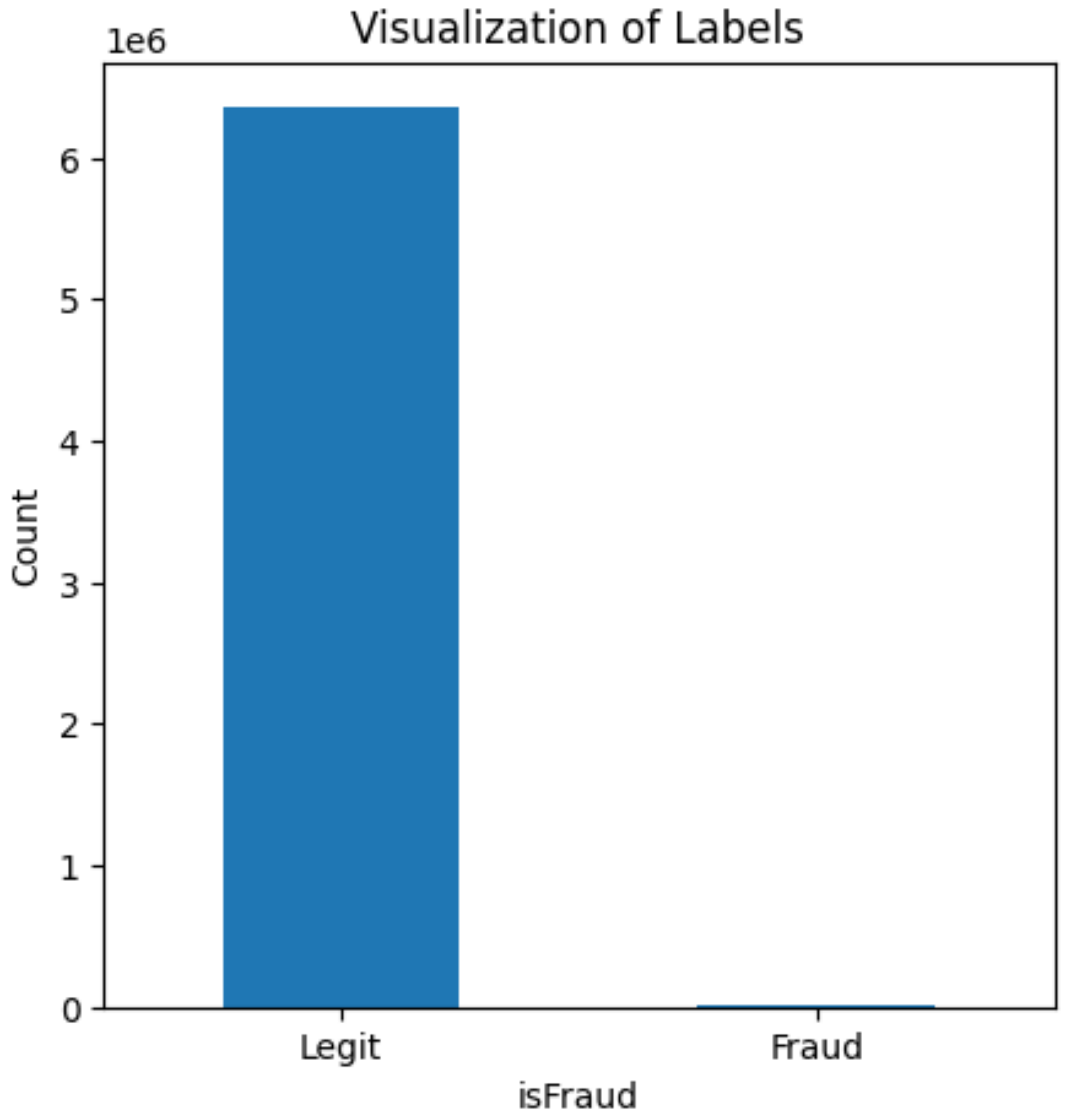
- Synthetic Logs Generator for Fraud Detection in Mobile Transfer Services
- PaySim: A Financial Mobile Money Simulator For Fraud Detection
- Analysis of fraud controls using the PaySim financial simulator
- Advantages of the PaySim Simulator for Improving Financial Fraud Controls
- Fraud Detection in Mobile Payment Utilizing Process Behaviour Analysis

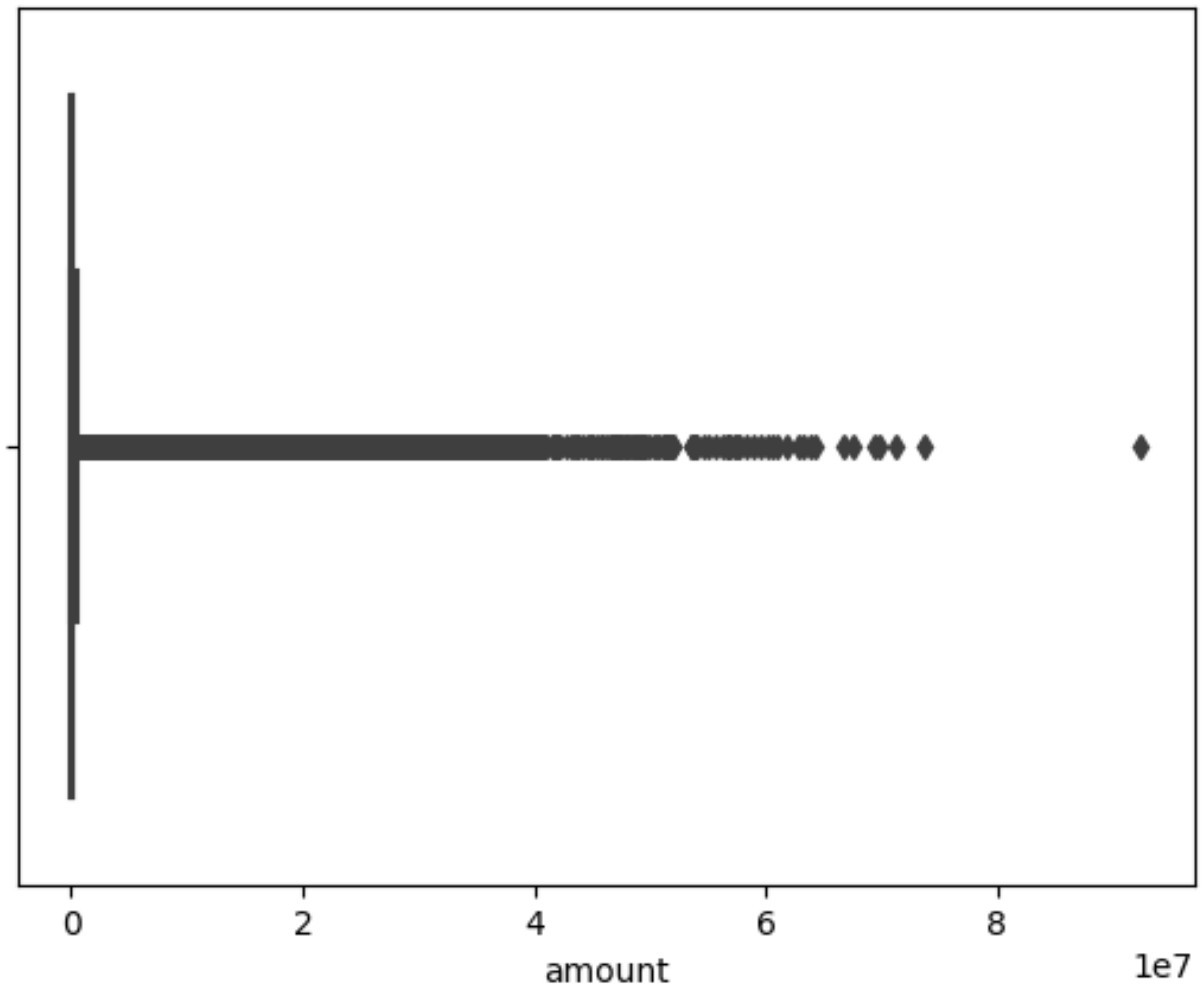
4. Proposed Work

- Analysis of the Synthetic Financial Dataset for Fraud Detection
- Data Splitting
- Data Modeling

4. Proposed Work

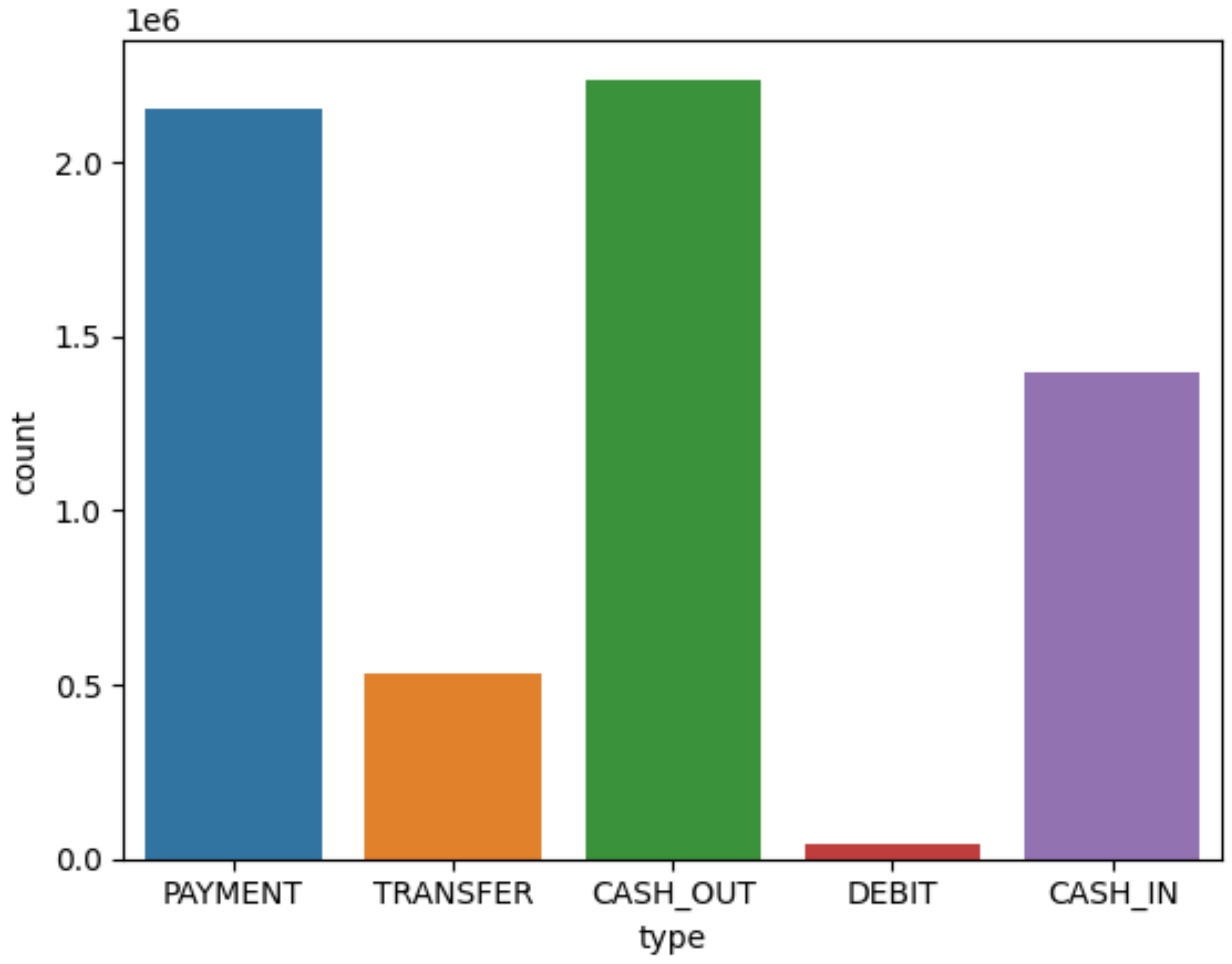
4.1 Analysis of the Synthetic Financial Dataset for Fraud Detection





4. Proposed Work

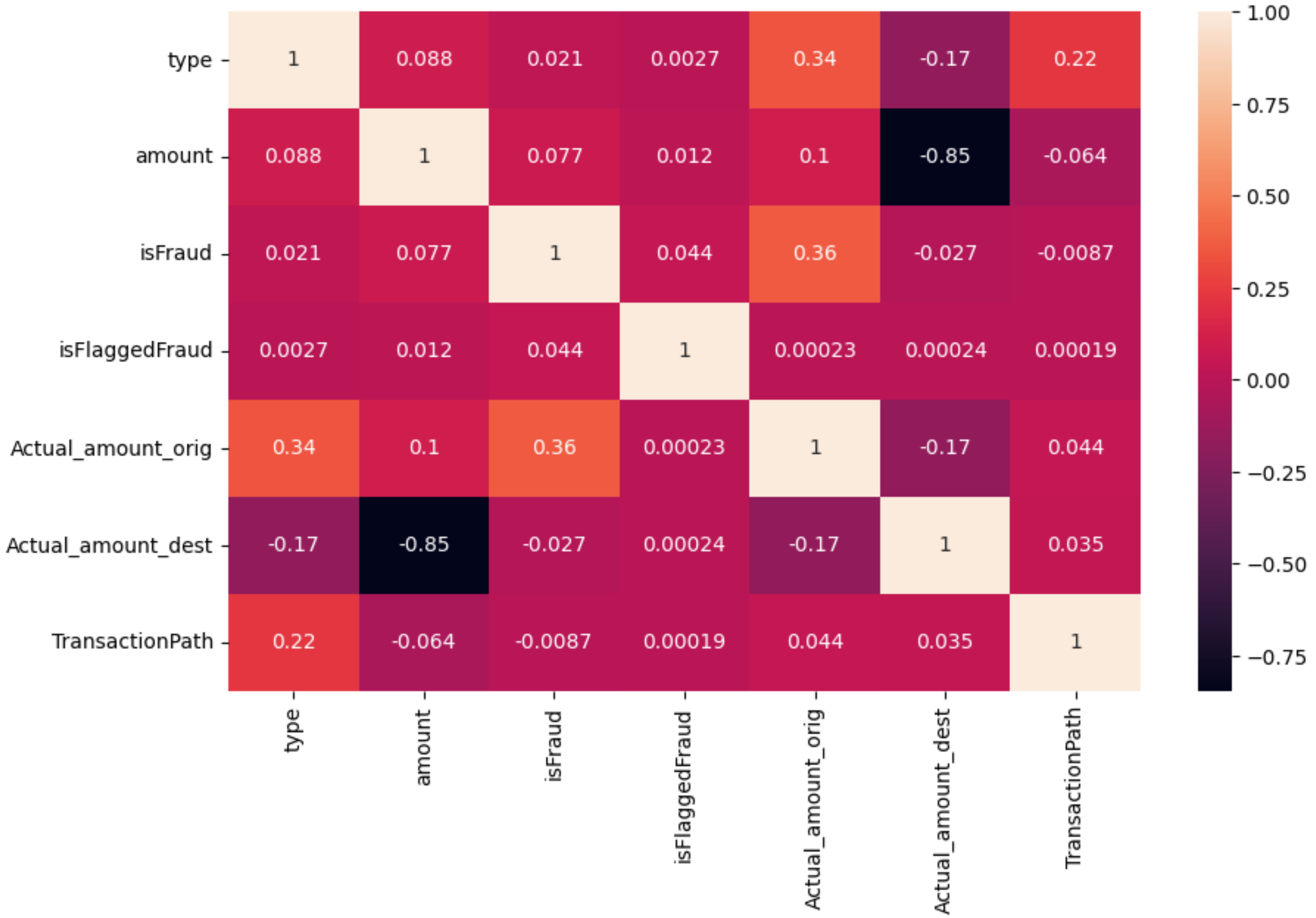
4.1 Analysis of the Synthetic Financial Dataset for Fraud Detection



4. Proposed Work

4.1 Analysis of the Synthetic Financial Dataset for Fraud Detection

	variables	VIF
0	type	2.687803
1	amount	3.818902
2	isFraud	1.184479
3	isFlaggedFraud	1.002546
4	Actual_amount_orig	1.307910
5	Actual_amount_dest	3.754335
6	TransactionPath	2.677167



4. Proposed Work

4.2 Data Splitting

A)
```
Shape of X_train:  (4453834, 6)
Shape of X_test:  (1908786, 6)
```

B)
```
Counts before oversampling minority class:

0    4448056
1       5778
Name: isFraud, dtype: int64
```

C)
```
Counts after oversampling:

0    4448056
1    4448056
Name: isFraud, dtype: int64
```

4. Proposed Work

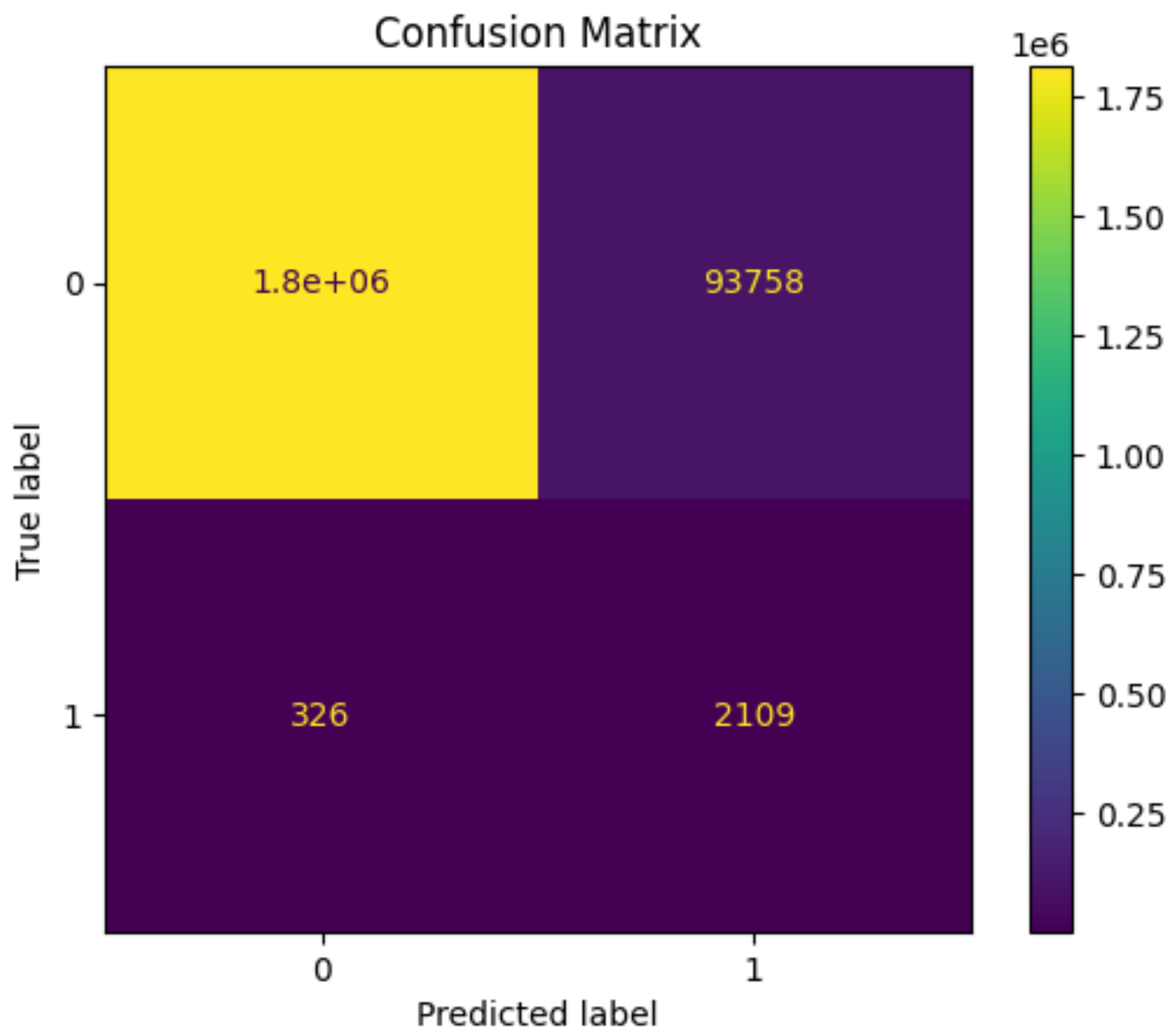
4.3 Data Modeling

```
# LOGISTIC REGRESSOR
y_train = upsampled.isFraud
X_train = upsampled.drop('isFraud', axis=1)

logistic_regressor = LogisticRegression(solver='liblinear', random_state=0)
logistic_regressor.fit(X_train, y_train)

y_pred_lr = logistic_regressor.predict(X_test)
logistic_regressor_score = logistic_regressor.score(X_test, y_test) * 100
```

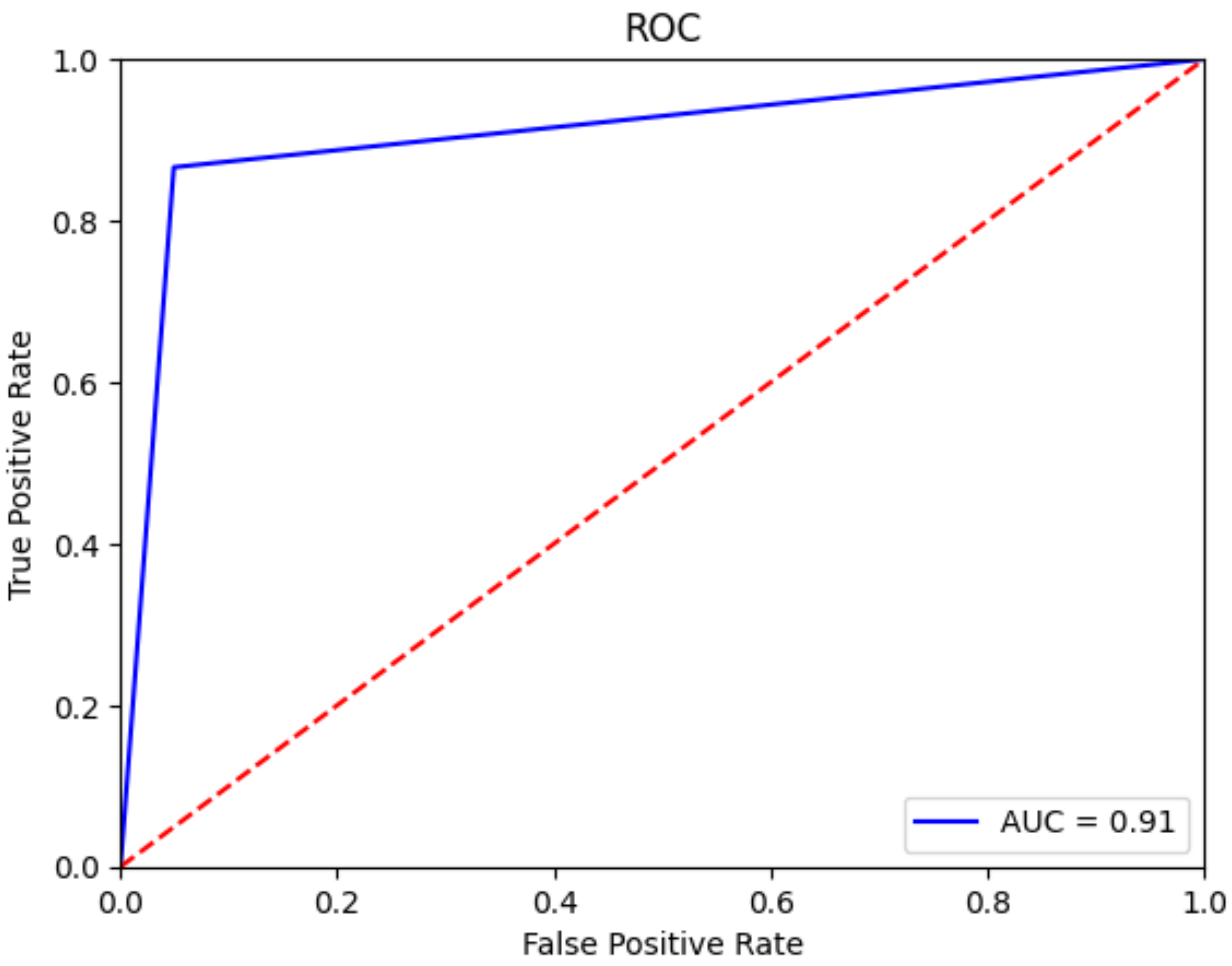
5. Evaluation

Classification Report

	precision	recall	f1-score	support
0	1.00	0.95	0.97	1906351
1	0.02	0.87	0.04	2435
accuracy			0.95	1908786
macro avg	0.51	0.91	0.51	1908786
weighted avg	1.00	0.95	0.97	1908786



5. Evaluation



6. Conclusion

- Digital Fraud Threats
- Synthetic Data Alternatives
- Modelling and Results
- Opportunities
- Further Research