Term Project Proposal

Towards Intelligent Financial Fraud Detection: Combating Evolving Threats in Real-Time

Professor: Chantri Polprasert

Presented by:
Team Duo
st124997 - Suryansh Srivastava
st125457 - Ulugbek Shernazarov

Overview

01 Introduction

02 Problem Statement

03 Related Work

04 Methodology

05 Preliminary Results

06 Conclusion

4.1 Datasets and Description

4.2 Features

5.1 Early Model Testing

5.2 Expected Challenges & Future Perspectives

Introduction

What is Fraud? What is Fraud Detection?

Motivation for the Project

Business Understanding and Impacts

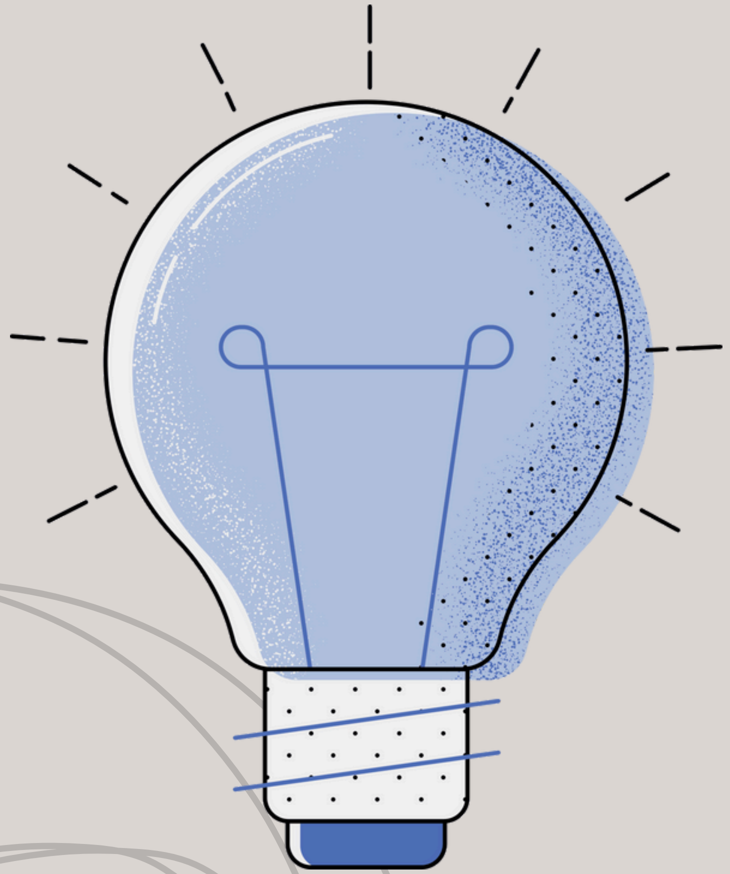
Delivered Outcomes

Problem Statement

Problem Challenges

- Financial Systems Under Threat
- Limitations of Traditional System
- Complexity in Real-Time Detection

Problem Objectives



Our goals:

- Intelligent Fraud Detection System
- Reduce Financial Losses
- Protect Institutions and Consumers

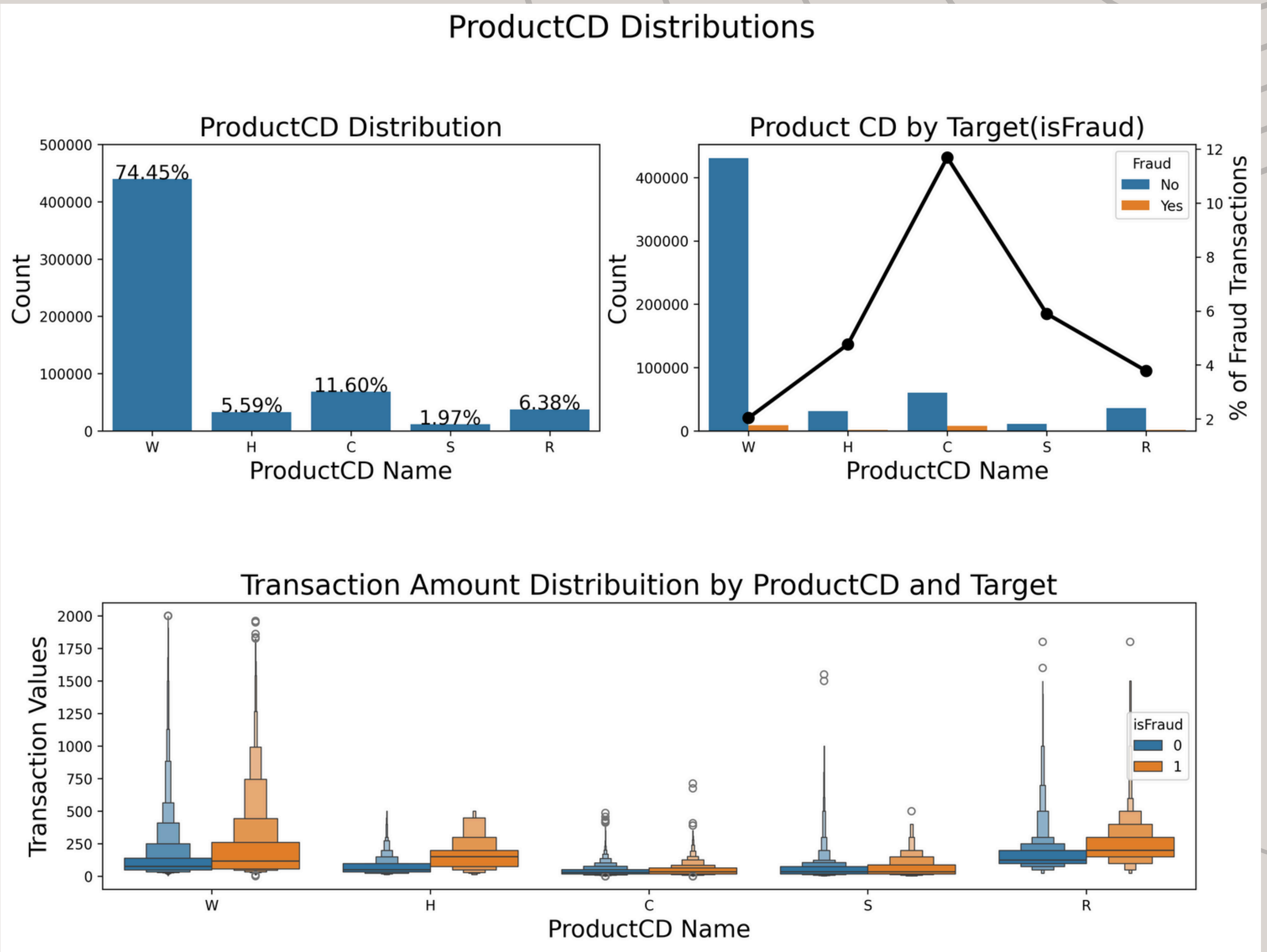
Related Work

- Evolution of Fraud Detection Methods
- Ensemble Methods in Fraud Detection
- Deep Learning Models
- Explainable AI (XAI) in Fraud Detection
- Key Takeway

Methodology

Dataset 1 - Consumer Identity Solutions (CIS) by IEEE

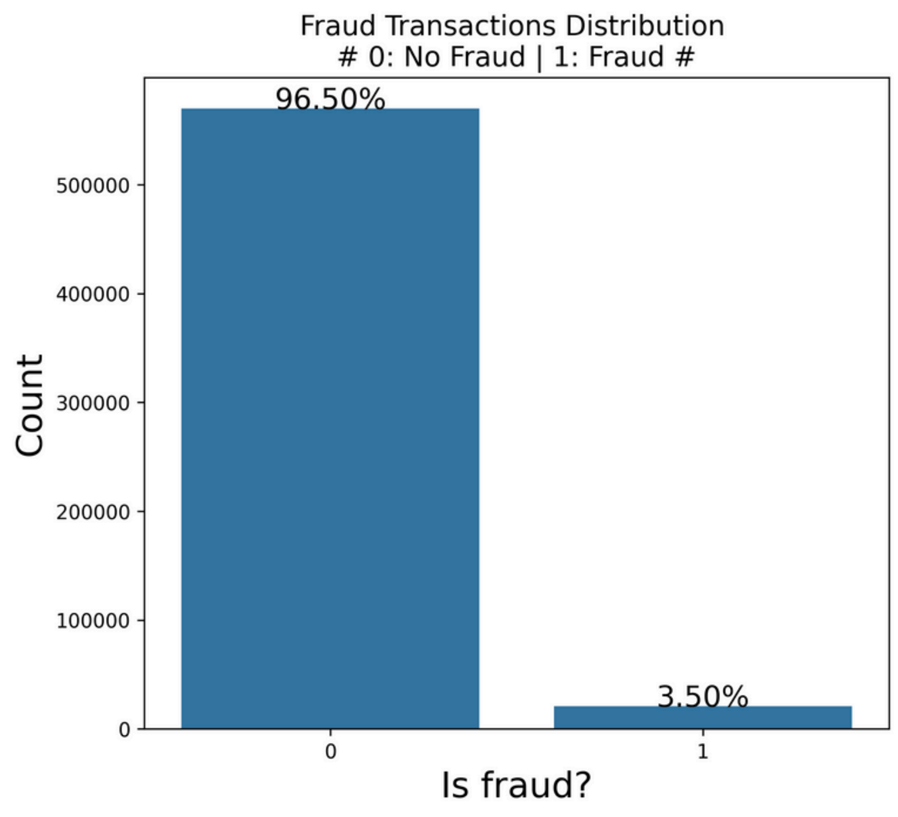
- Set of anonymized transactions from an e-commerce platform
- Approximately 1.1 million transactions
- 433 features (transaction and identity)
- Has imbalance

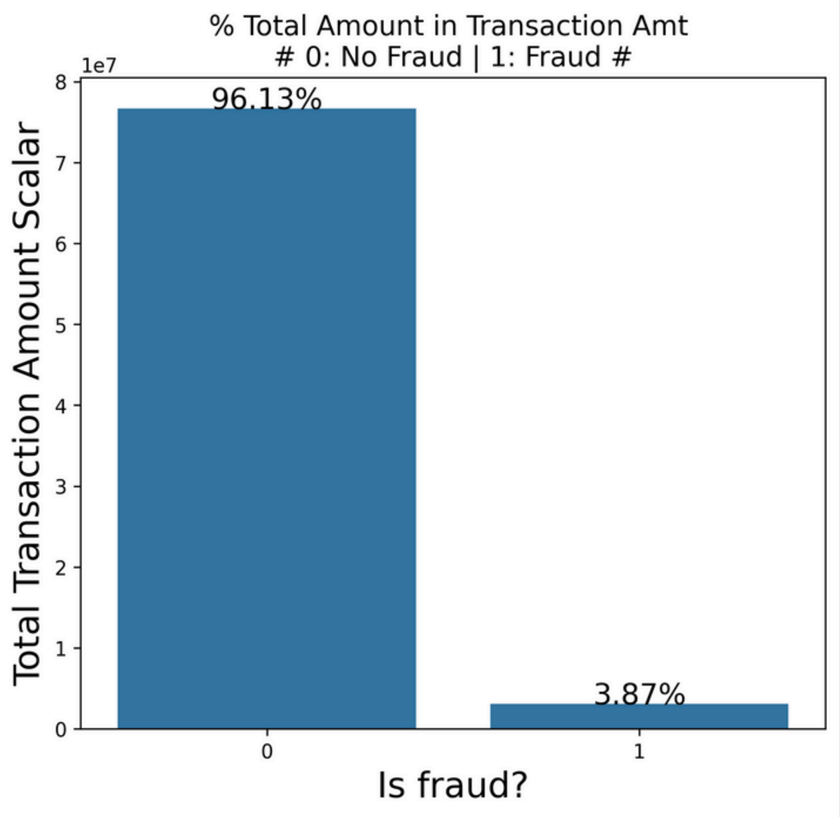


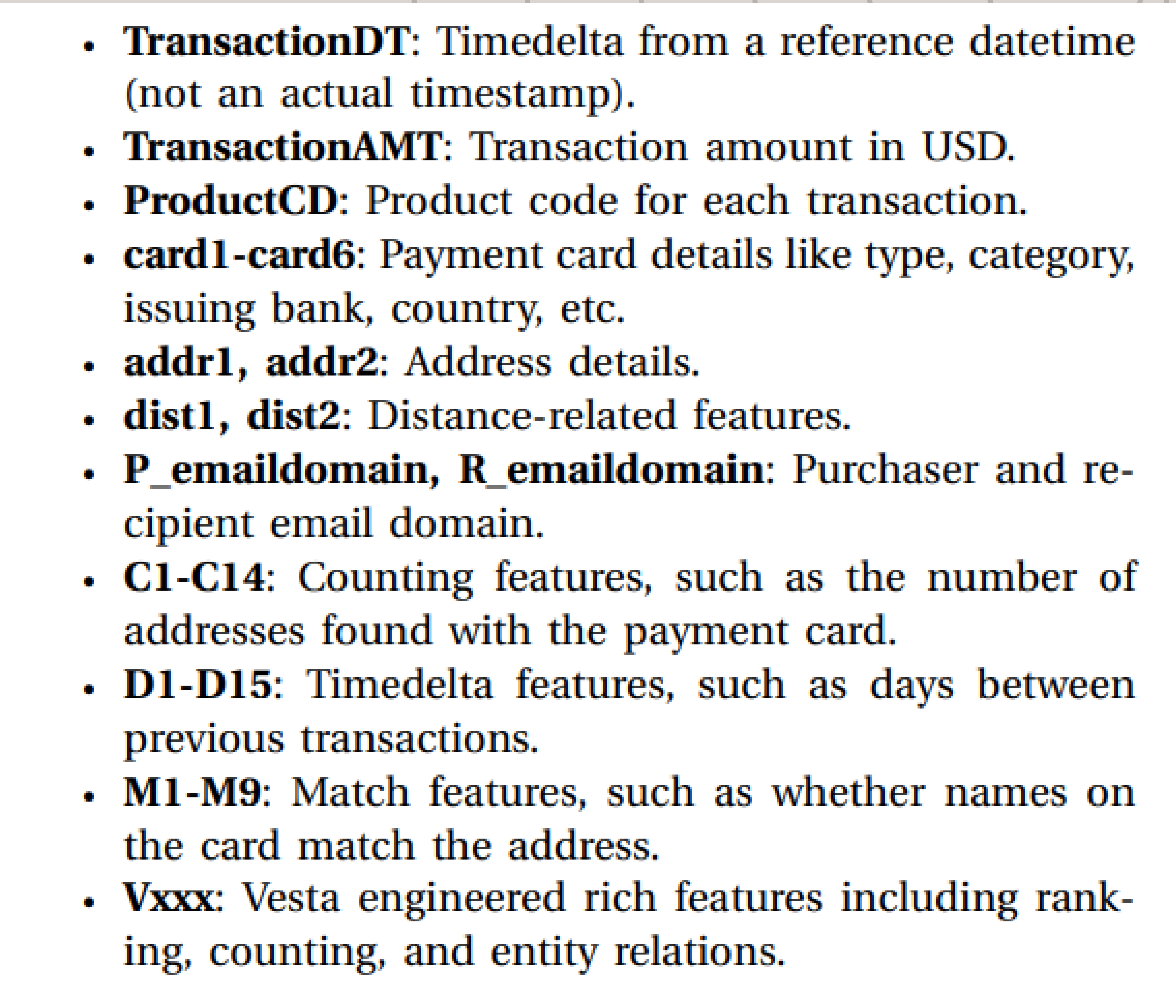
ProductCD Distributions

Methodology

Features - Consumer Identity Solutions (CIS) by IEEE

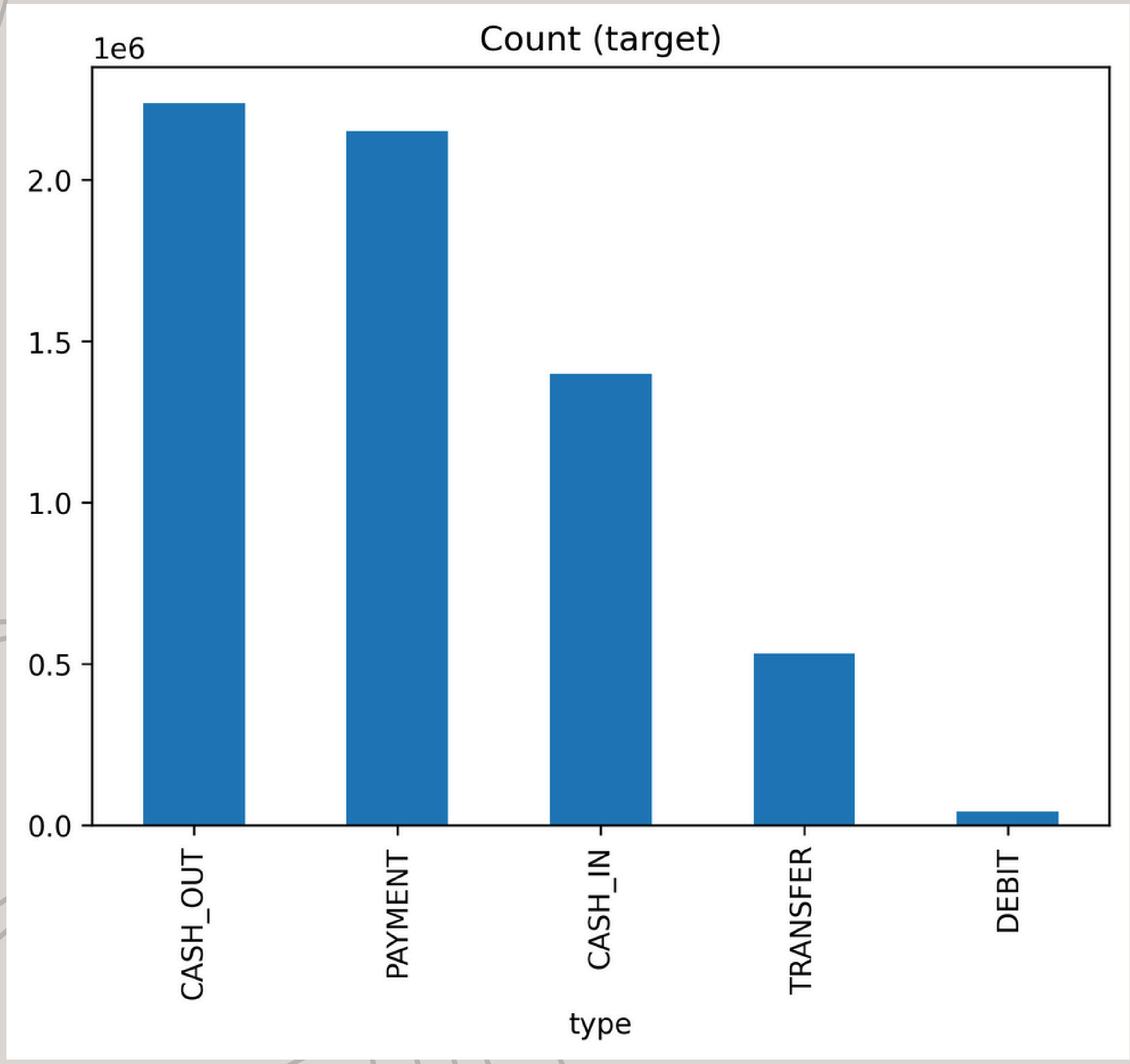






- **TransactionDT**: Timedelta from a reference datetime (not an actual timestamp).
- **TransactionAMT**: Transaction amount in USD.
- **ProductCD**: Product code for each transaction.
- **card1-card6**: Payment card details like type, category, issuing bank, country, etc.
- **addr1, addr2**: Address details.
- **dist1, dist2**: Distance-related features.
- **P_emaildomain, R_emaildomain**: Purchaser and recipient email domain.
- **C1-C14**: Counting features, such as the number of addresses found with the payment card.
- **D1-D15**: Timedelta features, such as days between previous transactions.
- **M1-M9**: Match features, such as whether names on the card match the address.
- **Vxxx**: Vesta engineered rich features including ranking, counting, and entity relations.

Methodology

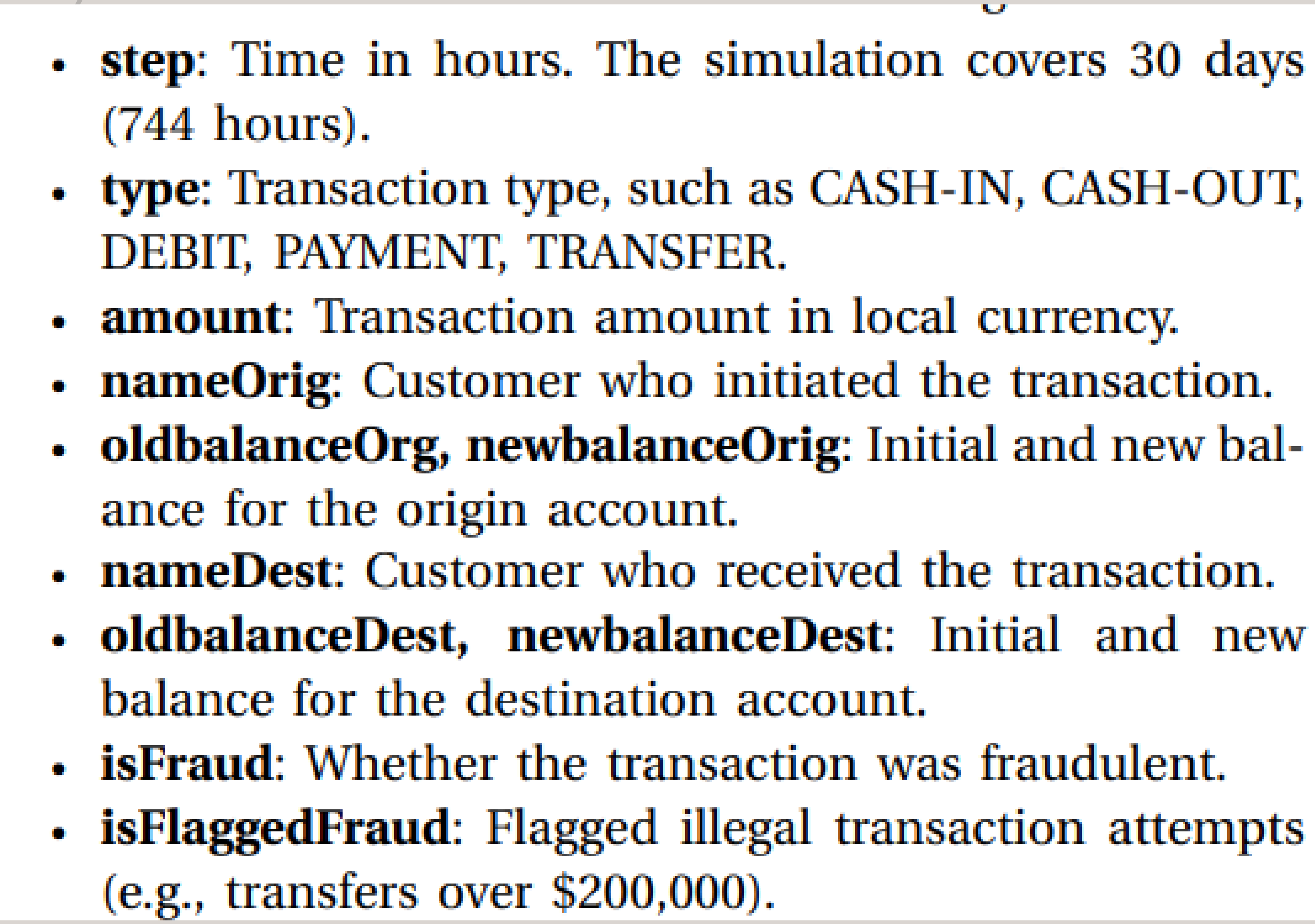


Dataset 2 - PaySim Dataset

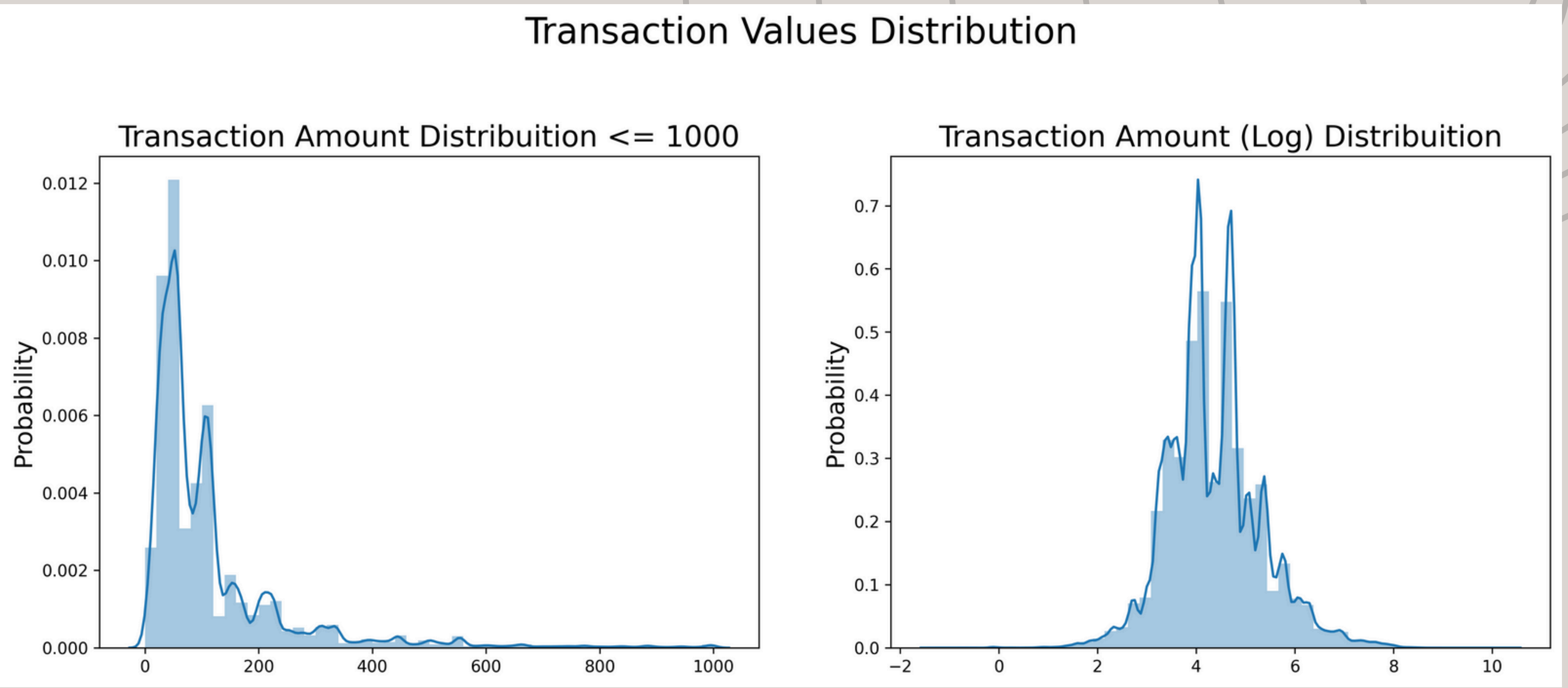
- Synthetic Fraud Detection Dataset
- Models mobile financial transactions based on real-world financial behavior
- 11 features
- 6.3 million transactions
- Small fraction of transactions labeled as fraudulent

Methodology

Features - Synthetic PaySim Dataset



- **step**: Time in hours. The simulation covers 30 days (744 hours).
- **type**: Transaction type, such as CASH-IN, CASH-OUT, DEBIT, PAYMENT, TRANSFER.
- **amount**: Transaction amount in local currency.
- **nameOrig**: Customer who initiated the transaction.
- **oldbalanceOrg, newbalanceOrig**: Initial and new balance for the origin account.
- **nameDest**: Customer who received the transaction.
- **oldbalanceDest, newbalanceDest**: Initial and new balance for the destination account.
- **isFraud**: Whether the transaction was fraudulent.
- **isFlaggedFraud**: Flagged illegal transaction attempts (e.g., transfers over $200,000).



Transaction Values Distribution

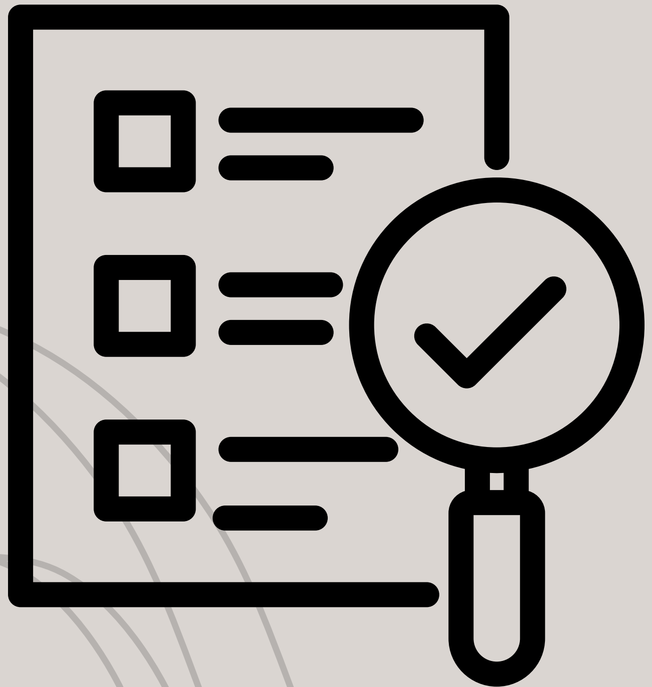
Transaction Amount Distribuition <= 1000

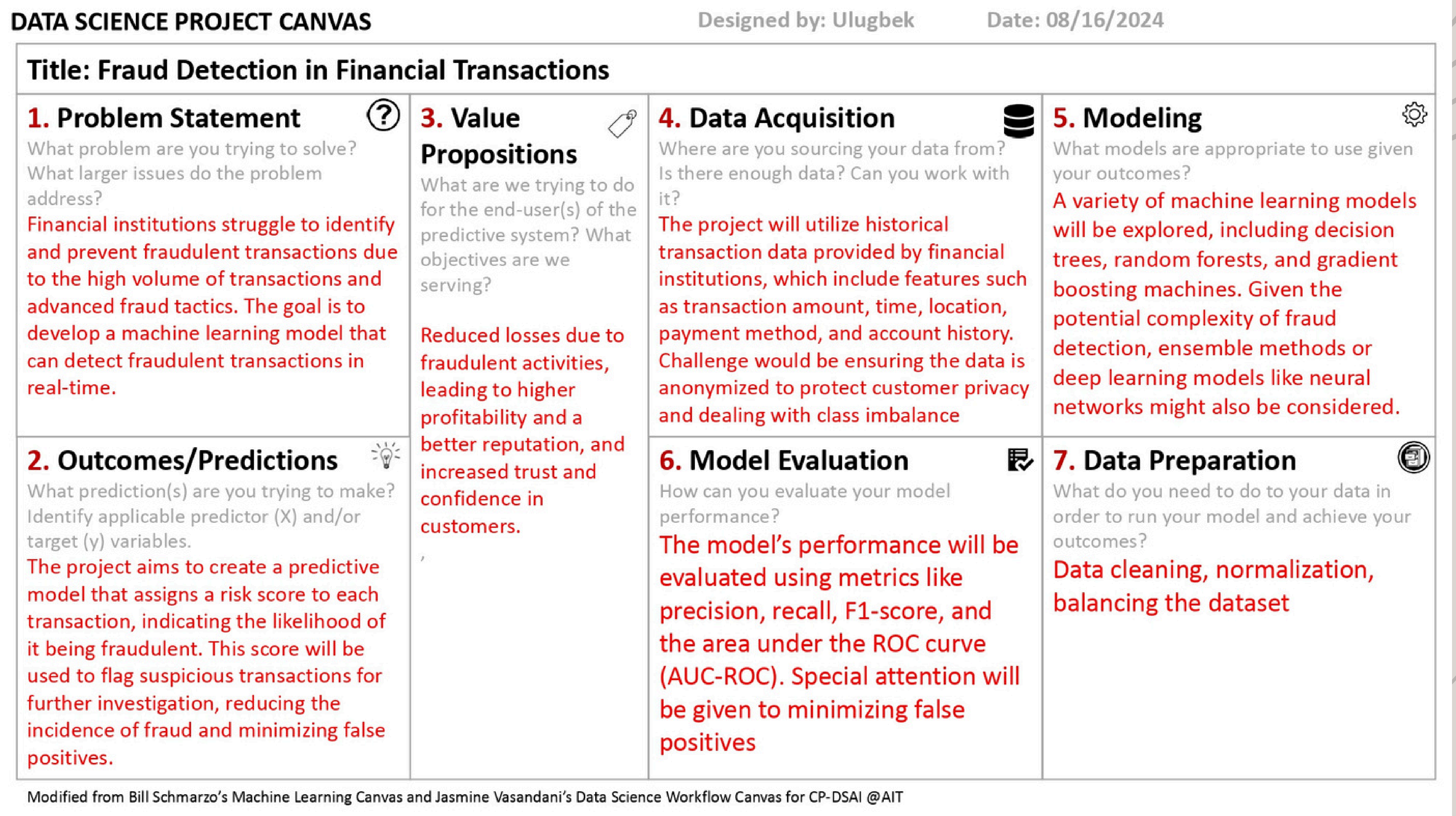
Transaction Amount (Log) Distribuition

Methodology

Evaluation

- Precision, Recall, F1-score
- AUR-ROC
- Confusion Matrix





DATA SCIENCE PROJECT CANVAS

Designed by: Ulugbek

Date: 08/16/2024

Title: Fraud Detection in Financial Transactions

1. Problem Statement

What problem are you trying to solve? What larger issues do the problem address?

Financial institutions struggle to identify and prevent fraudulent transactions due to the high volume of transactions and advanced fraud tactics. The goal is to develop a machine learning model that can detect fraudulent transactions in real-time.

2. Outcomes/Predictions

What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.

The project aims to create a predictive model that assigns a risk score to each transaction, indicating the likelihood of it being fraudulent. This score will be used to flag suspicious transactions for further investigation, reducing the incidence of fraud and minimizing false positives.

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

Reduced losses due to fraudulent activities, leading to higher profitability and a better reputation, and increased trust and confidence in customers.

,

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

The project will utilize historical transaction data provided by financial institutions, which include features such as transaction amount, time, location, payment method, and account history. Challenge would be ensuring the data is anonymized to protect customer privacy and dealing with class imbalance

6. Model Evaluation

How can you evaluate your model performance?

The model's performance will be evaluated using metrics like precision, recall, F1-score, and the area under the ROC curve (AUC-ROC). Special attention will be given to minimizing false positives

5. Modeling

What models are appropriate to use given your outcomes?

A variety of machine learning models will be explored, including decision trees, random forests, and gradient boosting machines. Given the potential complexity of fraud detection, ensemble methods or deep learning models like neural networks might also be considered.

7. Data Preparation

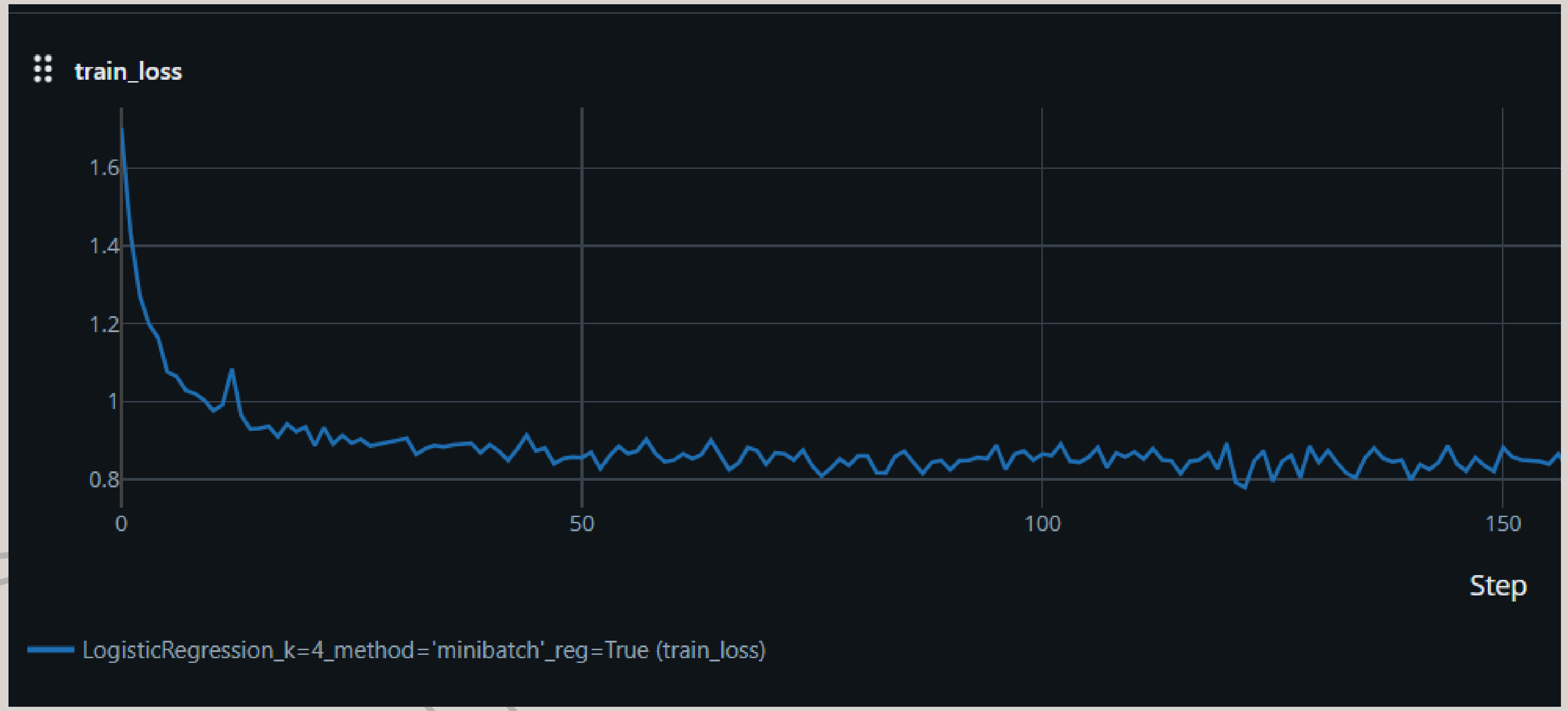
What do you need to do to your data in order to run your model and achieve your outcomes?

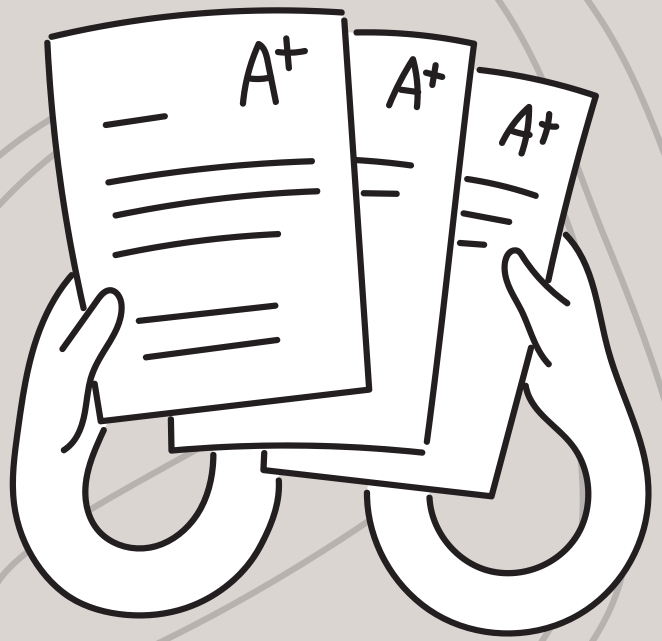
Data cleaning, normalization, balancing the dataset

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for CP-DSAI @AIT

Preliminary Results

Dataset 1 - Consumer Identity Solutions (CIS) by IEEE



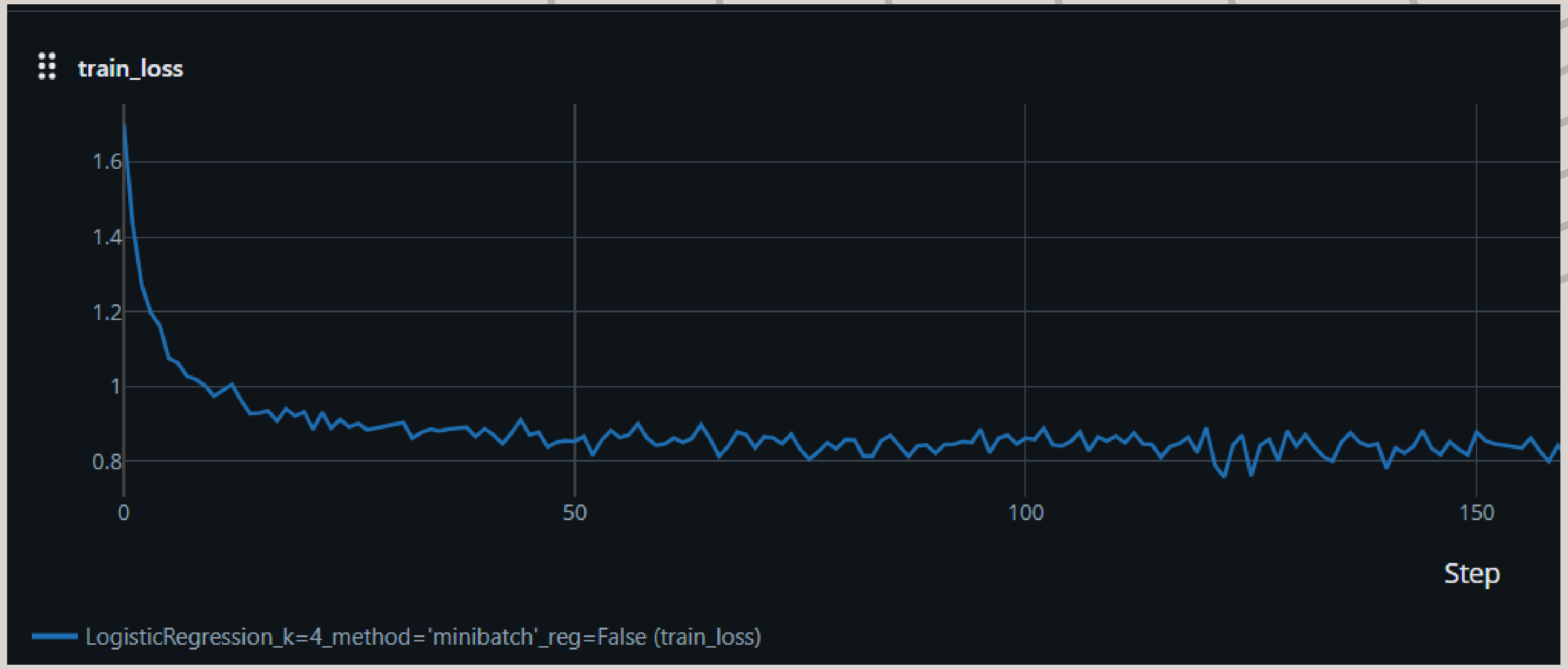


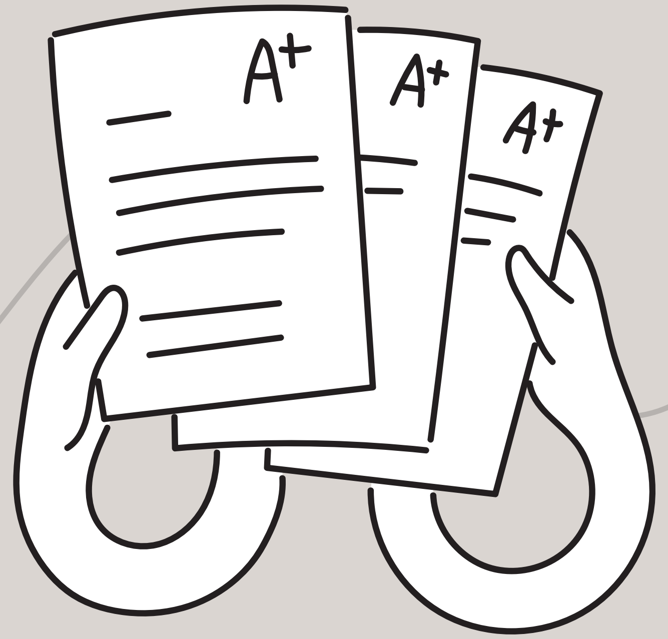
- High Cardinality
- Requires Feature Engineering
- Class Imbalance (SMOTE?)
- Logistic Regression achieved f1_score 0.35 and AUR-ROC score - 0.62
- More advanced models and feature engineering

Preliminary Results

Dataset 2 - Synthetic PaySim Dataset

- Fraud Rate is Low (0.12% flagged as fraud)
- Strong indicator' features
- Legitimate transactions are categorized under PAYMENT
- Logistic Regression showed a relatively high precision but lower recall, with f1_score 0.28 and AUC-ROC - 0.65
- More balancing techniques or advanced decision trees/ensemble methods





Preliminary Results

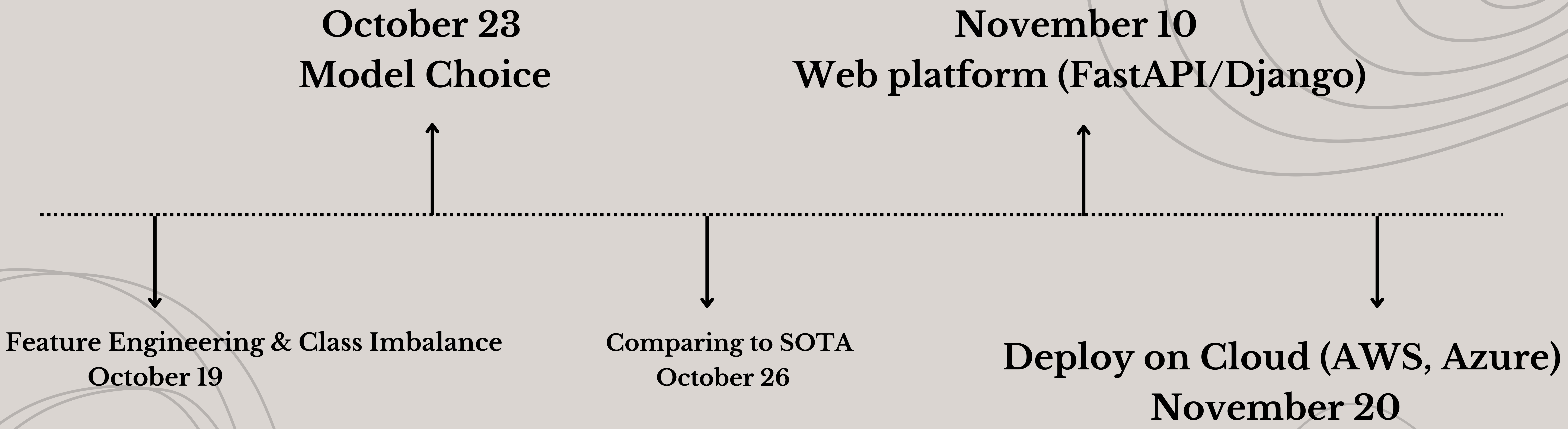
Expected Challenges and Next Step

Several challenges emerged:

- Class imbalance
- Feature Engineering
- Data Volume
- Choose advanced models (XGBoost, Random Forest, and Neural Networks)
- Hyperparameter optimization
- Comparing to the SOTA results in kaggle/huggingface as well as comparing the output models from each dataset

EXPEC
TATION

Proposed TimeScope



Feature Engineering & Class Imbalance
October 19

October 23
Model Choice

Comparing to SOTA
October 26

November 10
Web platform (FastAPI/Django)

Deploy on Cloud (AWS, Azure)
November 20

Thank You

Presented by Suryansh & Ulugbek