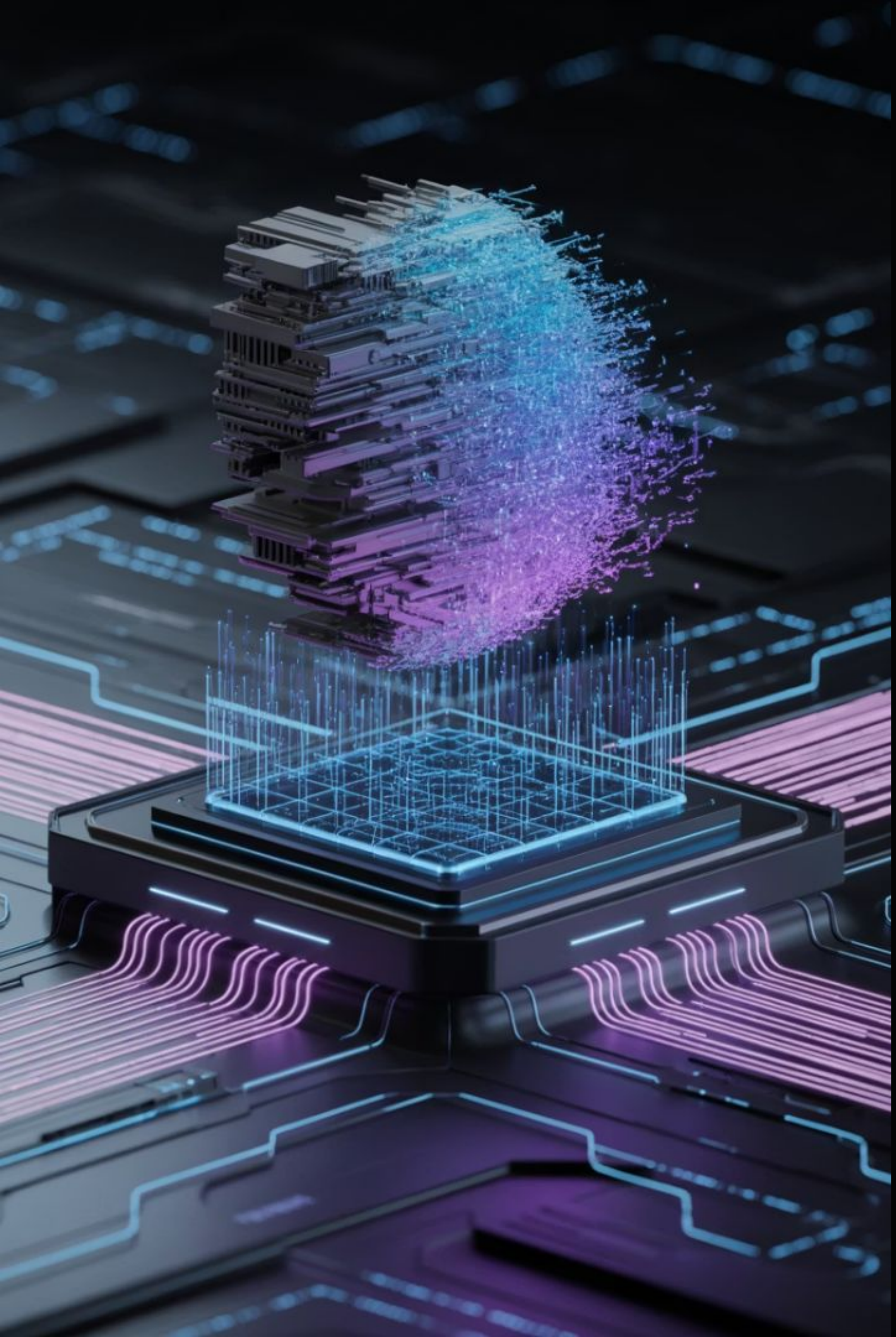
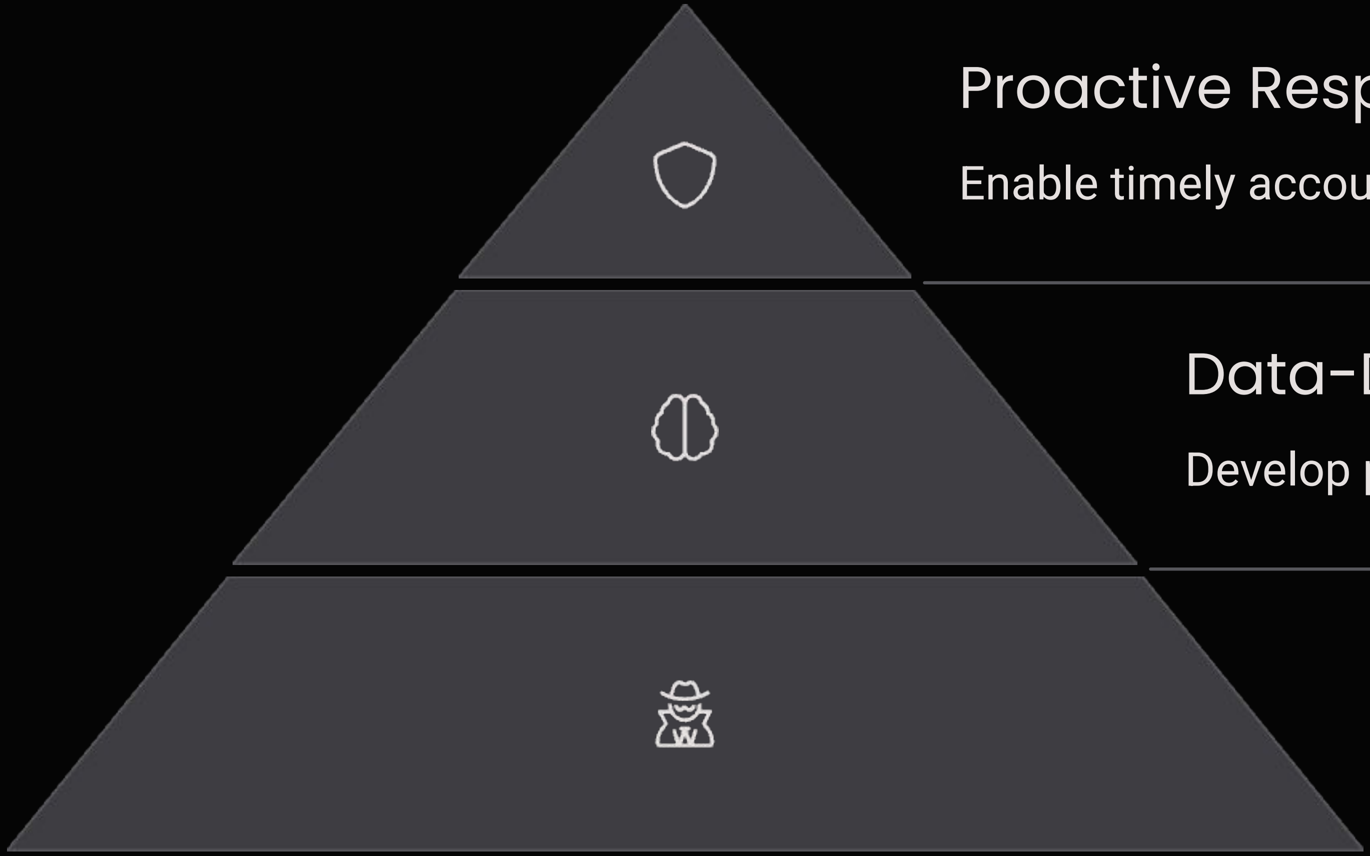


Fraudulent Users Detection

Welcome to our presentation on using advanced machine learning techniques to identify and prevent fraudulent user activity.

by Marek Bajdík

The Challenge: Identifying Fraudulent Users



Proactive Response

Enable timely account reviews and security measures

Data-Driven Approach

Develop predictive models beyond static rule systems

Fraud Risk

Combat financial losses and reputational damage

Traditional detection methods often fail against sophisticated fraud schemes. Our goal is to leverage machine learning for accurate, real-time identification.

Leveraging User and Transaction Data

User Profiles

- Sign-up details
- Country information
- KYC verification status

Transaction Data

- Transaction amounts
- Currencies used
- Merchant information
- Timestamps

We merged these datasets to create comprehensive user activity profiles. Data quality issues were addressed through cleaning and standardization.

Understanding the Data

Geographic Patterns

Suspicious location and phone country combinations identified

Transaction Behaviors

Unusual amount patterns and volatility flagged

Timing Analysis

Sign-up and activity timing revealed suspicious patterns

Our exploratory analysis revealed distinct differences between legitimate and fraudulent user behaviors. These insights guided our feature engineering.

Building Predictive Features

Categorical Encoding

Transformed categories using dummy and WoE encoding

Time-Based Features

Extracted year, month, hour patterns from timestamps

Aggregation Features

Summarized transaction history over 7 and 30-day windows

User Behavior Metrics

Created transaction frequency and pattern indicators

Selecting the Most Informative Features

Final Feature Set

Diverse combination of WoE, one-hot, and engineered features

Feature Clustering

Addressed multicollinearity between related features

Univariate Analysis

Used SelectKBest to identify predictive power

We optimized our feature set by focusing on the 9th transaction as our prediction point. This balanced data availability with early fraud detection.

Building the Predictive Model

1. **Model Selection**

   Chose XGBoost Classifier for superior performance with complex data

2. **Imbalance Handling**

   Implemented RandomUnderSampler within the pipeline

3. **Data Splitting**

   Created time-based train, validation, and test sets

4. **Hyperparameter Tuning**

   Optimized model with RandomizedSearchCV

Conclusion & Next Steps

Model Success

Achieved high AUC score on test data

Real-time Implementation

Deploy for continuous monitoring and protection

Ensemble Methods

Explore additional model combinations

External Data

Incorporate additional data sources

Our model successfully identifies fraudulent users before significant damage occurs. We'll continue refining our approach for even better protection.