METAVERSE TRANSACTION RISK CALCULATION

WQD7006 - Machine Learning

Team Members:

MUHAMMAD ZAHRIEL BIN ISMAIL - 22085509

HAN XIANG - 22093085

TASSLIM BIN MANSOOR ALI - 23056322

IZZUL ILHAM BIN YUSOF - 22107573

ZHANG LEPING - 22098083

CONTENT

Introduction

Project Context

Blockchain technology has emerged as a transformative force across various industries, offering unparalleled transparency, security, and decentralization. With the proliferation of cryptocurrencies and decentralized applications, blockchain networks have become a focal point for financial activities, digital asset transfers, and smart contract executions.

This report aims to provide insights into risk analysis within blockchain transactions through the examination of a comprehensive dataset of blockchain transactions.

Problem Statement

The unregulated nature of the crypto currency blockchains has shown that in the past few years, large amounts of scams and fraud can easily occur on its services as the market is largely separated from conventional regulatory bodies (Banks, Governments, etc.). This is a large issue as risk identification of potential fraud transactions are difficult to identify.

Project Objective

1) Analyze and highlight key characteristics in fraudulent transactions.
2) Create anomaly detection models to identify fraudulent transactions.
3) Deploy a product to assist cryptocurrency blockchain users in making future transactions.

Project Scope

The scope of this project encompasses transactions and data that is recorded on the Ethereum (ETH) blockchain. The data studied will contain a list of anonymized user transactions and is not exclusive to the Ethereum coin, but all cryptocurrencies registered under the ETH blockchain.

Literature Review

Phillips, R., & Wilder, H. (2020).	It is the analysis of various characteristics of cryptocurrency scams, focusing on methodologies, typologies, and trends observed in the field, such as Social Media Exploitation, Website Clustering and Geographic Distribution.
Bartoletti, M., Lande, S., Loddo, A., Pompianu, L., & Serusi, S. (2021)	This insight suggests that scams may emerge on other blockchains as their native cryptocurrencies gain popularity. Standardizing and moderate scam reporting, developing a browser extension to alert users to scams and assist those affected, and enhancing detection with machine learning, expand features, and analyze scam lifetimes and impact.

Data Pipeline

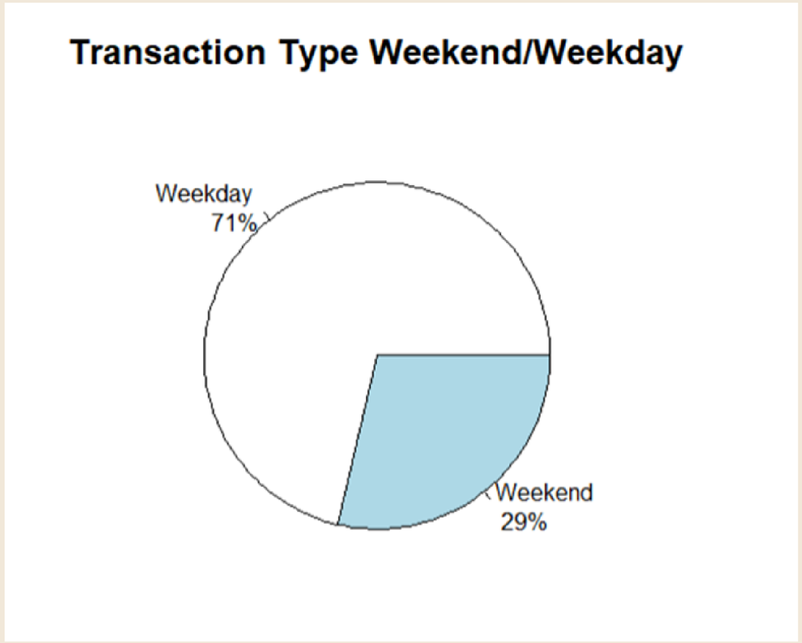
Dataset

Data Structure	Property
Data Source	https://www.kaggle.com/datasets/faizaniftikharjanjua/metaverse-financial-transactions-dataset
Data Name	Metaverse Financial Transactions Dataset
File Type	.csv
File Size	13.5 MB
Year	2024
Data Dimensions	14 Columns, 78601 Rows

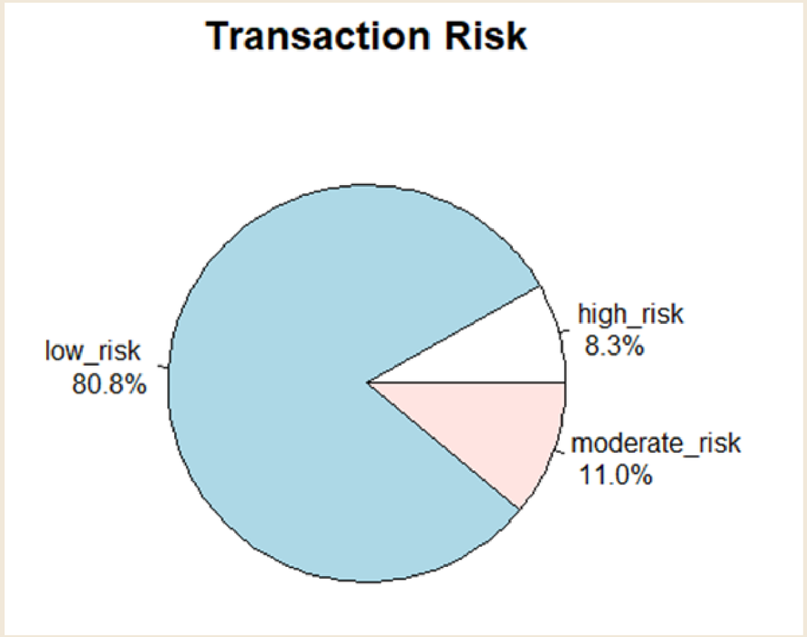
- Phishing and scams are group into similar context
- This transaction is based on Ethereum blockchain
- Transaction day context is binned into weekday and weekend
- Transaction time context is split into morning, afternoon, evening, late night
- IP prefix is used as proxy for network size

EDA

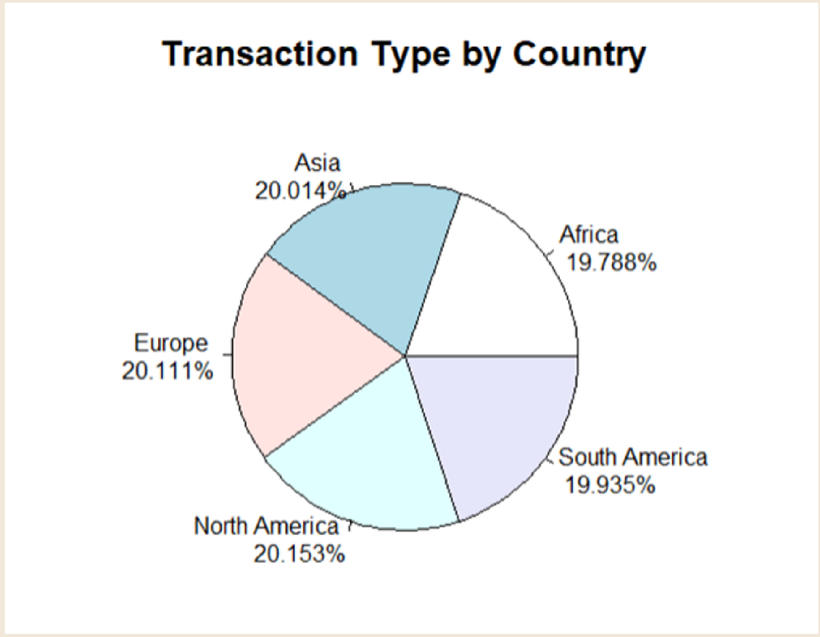
Univariate Analysis



More transaction in weekday

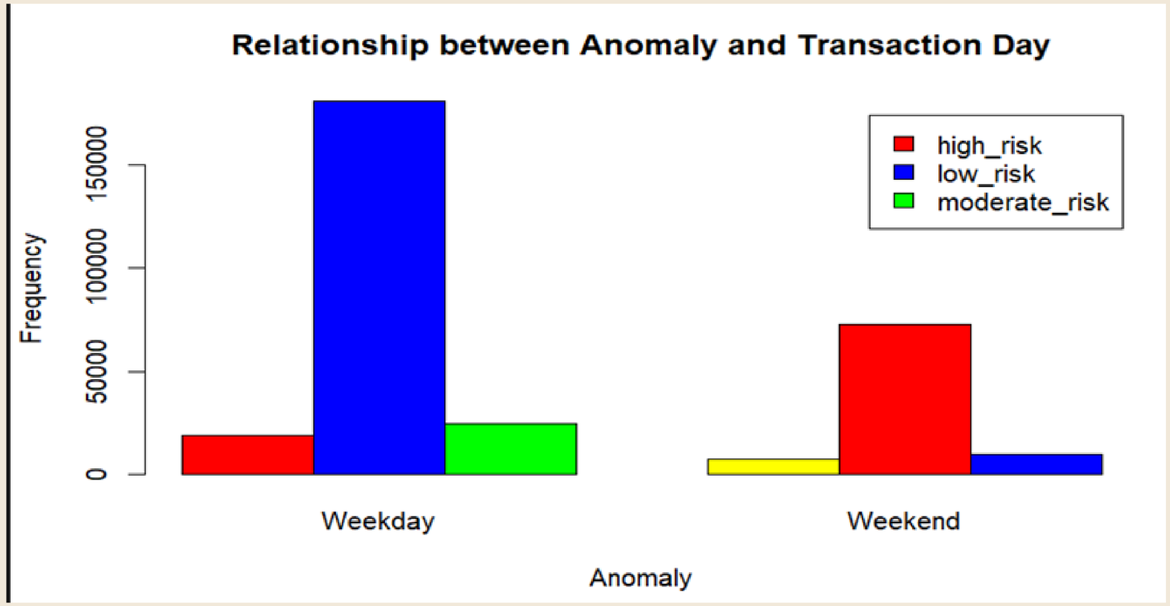


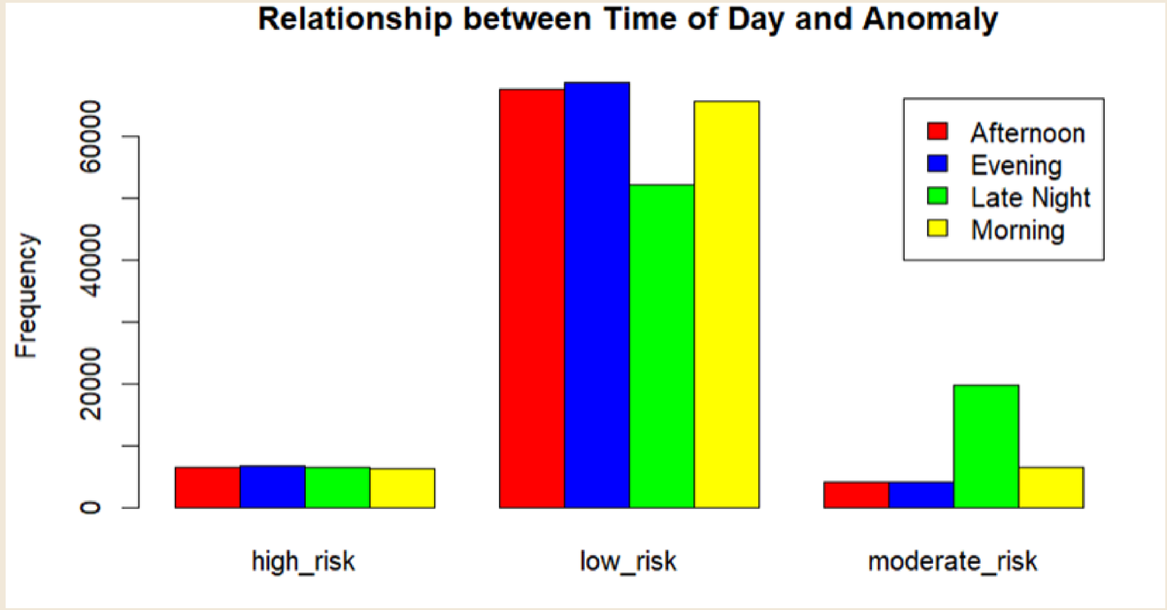
Transaction in metaverse are generally legit



Participants in metaverse are quite diverse and generally balance

Multivariate Analysis

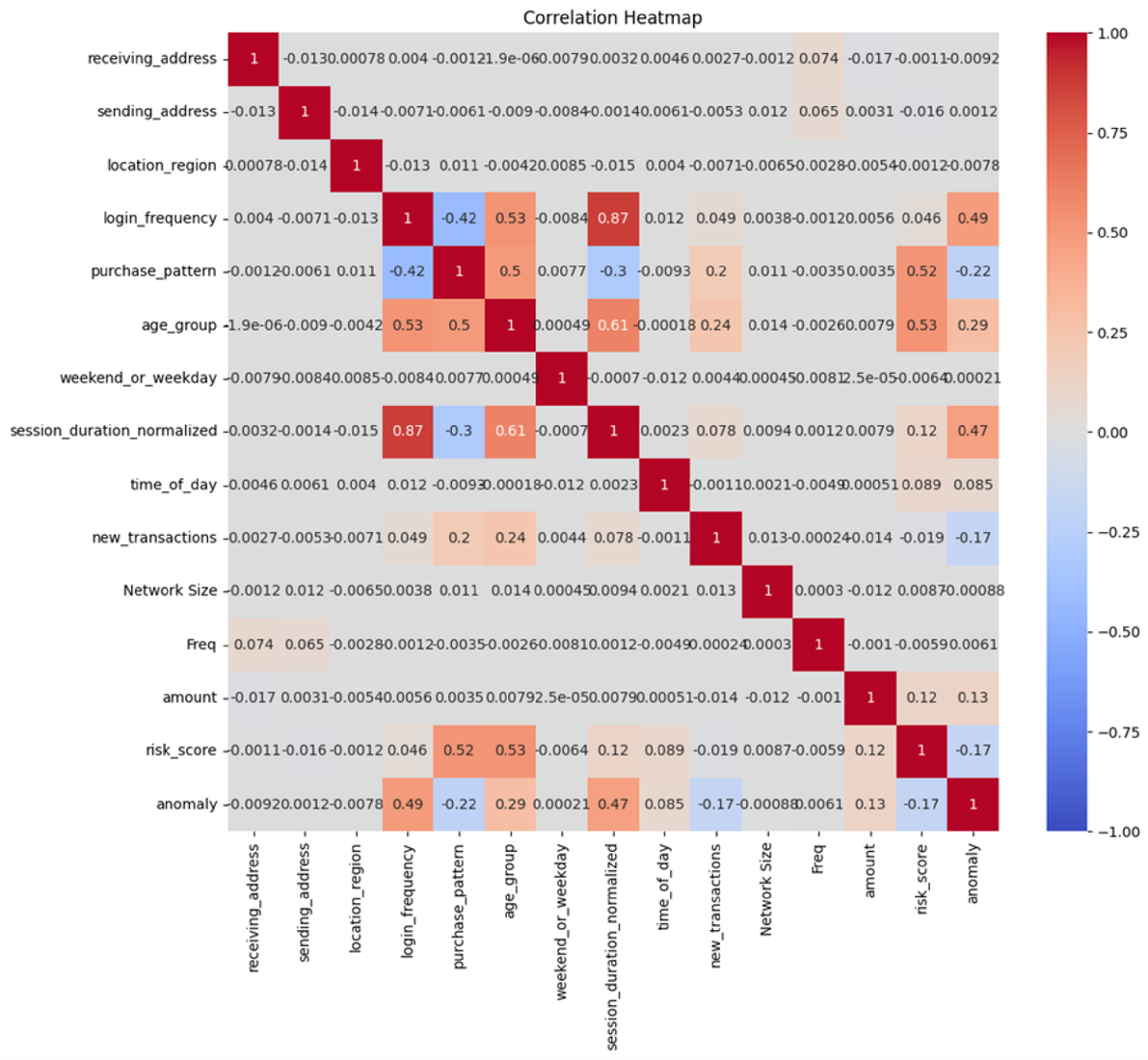




Mapping two features against each other, we can see that most of the high-risk transactions occur over weekends while low risk transactions occur more towards the weekday. This shows that transactions over the weekend have higher odds of being a scam.

We identify that there is no clear relationship between high-risk trades at any specific time while moderate risk sees more activity at late night timings. Low risk transactions are mostly evenly distributed.

Correlation Analysis



- There exists high positive correlation between “risk_score” and “purchase_pattern” and “age_group”
- High negative correlation between “login_frequency” and “purchase_pattern”
- Scam transactions or more likely to be influenced by the age of the account holders and whether the purchases were random or specific purpose

Modelling

Model development

The study is divided into two parts: regression modeling and classification modeling. Regression modeling aims to calculate transaction risk scores using linear regression, MARS, and XGBoost models; classification modeling identifies fraudulent transactions using decision tree, random forest, and MARS. This combination of methods leverages the strengths of each model to ensure a thorough analysis and reliable evaluation of the dataset. By employing a multi-model and multi-algorithm approach, the goal is to minimize biases and robustly handle complex data relationships.

Model Practical Implementation and Comparison

Classification

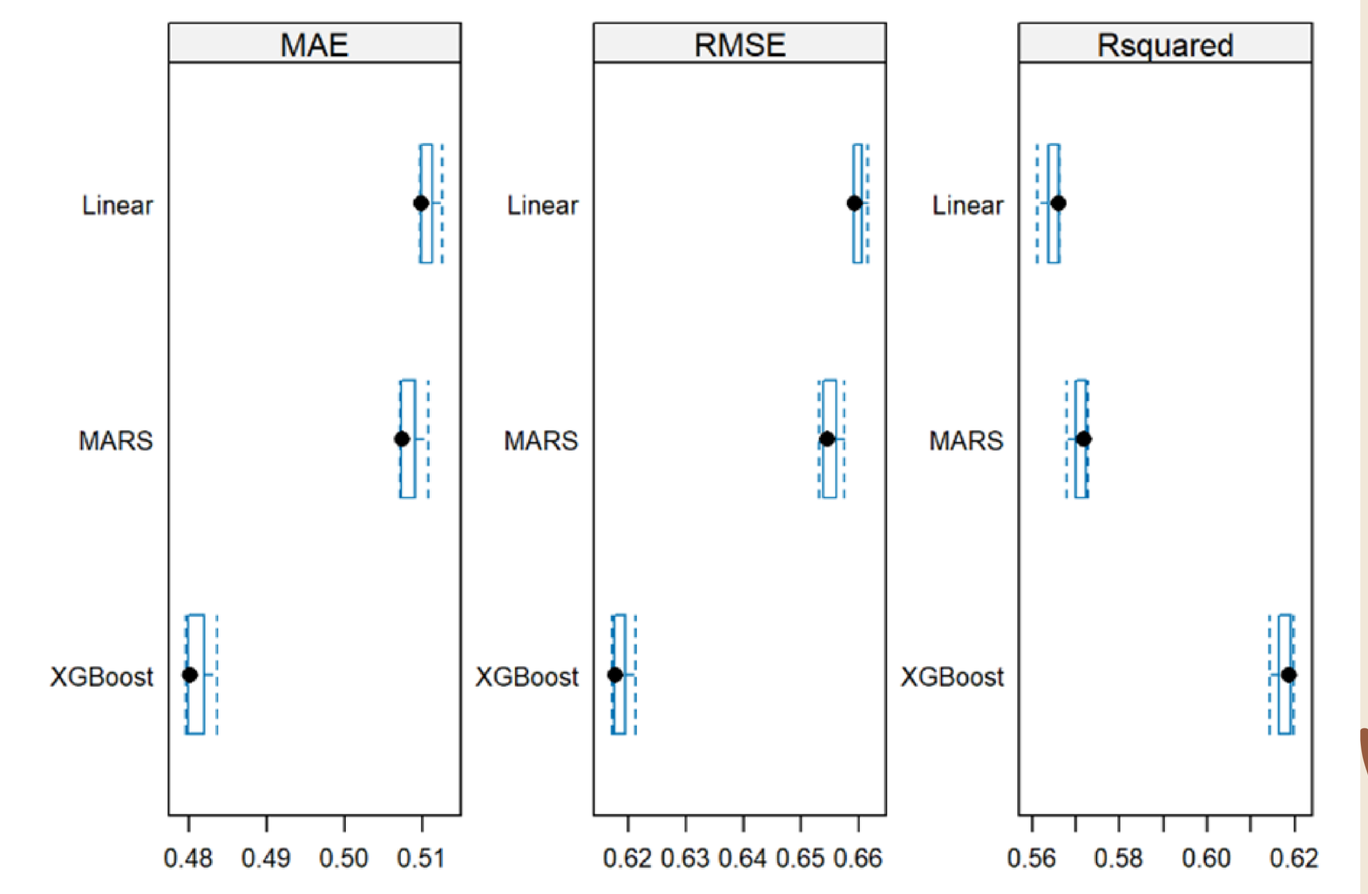
1) Decision Tree
2) Random Forest
3) MARS
   (multivariate adaptive regression spline)

	Decision Tree	Random Forest	MARS
Accuracy	0.863	0.955	0.859
Recall	0.608	0.849	0.593
F1	0.849	0.887	0.837

Model Practical Implementation and Comparison

Regression

1) Linear
2) XGBoost
   (eXtreme Gradient Boosting)
3) MARS
   (multivariate adaptive regression spline)



Model Evaluation

- The **classification model** is responsible for anomaly detection. This model helps identify potential risks or anomalies by recognizing abnormal or unusual behaviors within the data.
- The **regression model** is used to predict risk scores. This model assesses potential risk levels by analyzing and calculating the risk grade or score based on given inputs.
- The analysis results indicate that XGBoost and Random Forest perform well due to their ability to handle complex data relationships and feature interactions, making them particularly suitable for dealing with nonlinear and highly variable datasets.

Approach and Innovation

Phase	Steps Taken
Data Preprocessing	Handled outliers and missing data Validate ETH sending and receiving addresses Normalizing timestamps Calculate IP network size
Modelling	Ensemble machine learning algorithms such as XGBoost and Random Forest for both regression and classification predictions
Feature Engineering	Analyzed information such as dates, purchase patterns and login frequency to spot abnormal trends that might indicate fraud
Deployment	Utilized API from cryptocurrency website **Etherscan.io** to monitor transaction records and obtain information from the blockchain

Commercialization

- ✓ Offering a Software as a Service (SaaS) model
- ✓ Licensing Model for Enterprise Users
- ✓ Partnerships with companies that utilize cryptocurrency

Market Validation

Milestones	Success Metrics	Engagement Strategy
Conduct comprehensive market research	Adoption rate from the pilot program users	Collaborate with top blockchain developers and metaverse platform providers
Implement pilot / beta programs	Quantitative and qualitative input from users	Provide training and certification programs to universities and corporations
Gather and examine user feedback	Increase in revenue	Collaborate with industry experts to publish whitepapers and case studies.

Cost Analysis

Category	Description	Cost
Technology	Machine Setup	RM5000 / PC
Infrastructure	Amazon Web Services (AWS) Instances, Storage, Bandwidth, Security and API Scrapper	~RM3600 / month
Personnel	Data Scientists, ML Engineers, Sofftware Developers, IT Support and Training	RM6000 / month per headcount
Other Expenses	R&D, Marketing Campaigns and Insurance	~RM60,000 / month

Conclusion

- In this study, we proposed a framework to assess transactional risk within the metaverse.
- Our results show that the XGBoost model performs best by producing the risk score for a particular metaverse transaction.
- The Random Forest model best identifies anomalies in metaverse transactions with the best accuracy and F1 score.
- This study would be able to assist the metaverse participants and developers in identifying transactional risk and help them to take preventive measures in performing future transactions within the metaverse.

Deployment (DEMO)