💼 Wallet Risk Scoring Methodology Explanation

🔍 DATA COLLECTION METHOD
Primary Data Source: **Etherscan API**

Using **Etherscan API** (a tool to see everything happening on the Ethereum blockchain) to check what each wallet has been doing.

- We downloaded all transactions (normal + internal) for each wallet.

- But we **only looked at transactions related to Compound V2** — a popular DeFi (decentralized finance) platform where people lend and borrow crypto.

- To avoid overloading the API, we added a short pause between requests.

- We also added a timeout in case any request took too long or failed.

🎯 Compound V2 Protocol Focus:

Because Compound is a major platform for borrowing and lending crypto, analyzing it gives us a solid idea of each wallet's **financial behavior** — like borrowing habits, repayments, or risky actions like getting liquidated.

🏛️ Smart Contracts We Tracked

Each of these is like a "department" inside the Compound bank:

Contract	What It Does
Comptroller	Manages the whole system
cDAI	Handles DAI loans
cETH	Handles Ethereum loans
cUSDC / cUSDT	Handles USDC and USDT loans
cUNI / cLEND	UNI & LEND tokens
cWBTC	Wrapped Bitcoin

Only analyzed activity that interacted with **these smart contracts** just like we'd only check someone's **bank activity**, not their Amazon shopping.

📊 How We Chose the Risk Features

We selected **10 things** about each wallet that can show how risky or safe it is in terms of DeFi activity.

🔟 Key Features (Explained Simply):

1. **Total Borrowed**

   o Number of times the wallet borrowed crypto

   o Rationale: Higher borrowing frequency indicates greater exposure to liquidation risk

   o Calculation: Count of high-gas transactions (>100k gas) with zero ETH value (typical borrow pattern)

   o Risk Impact: More borrowing = **potentially more risk**

2. **Total Supplied**

   o Number of times they added money as collateral

   o Rationale: Supply activity demonstrates responsible lending behaviour and collateral provision

   o Calculation: Count of transactions with ETH value > 0 and moderate gas usage (>50k gas)

   o Risk Impact : More supplying = **lower risk**

3. **Repay Ratio**

   o How much they supplied vs. how much they borrowed

   o Rationale: Most critical indicator of repayment behaviour and financial responsibility

   o Calculation: total_supplied / total_borrowed (defaults to 1.0 if no borrowing)

   o Risk Impact **: Higher ratio = lower risk** (good repayment behaviour)

4. **Liquidation Count**

   o How often they got liquidated (means their borrowed loan became too risky)

   o Rationale: Direct indicator of past financial distress and high-risk behaviour

   o Calculation: Count of high-gas internal transactions (>200k gas, typical of liquidations)

   o Risk Impact: Any liquidations = significantly higher risk

   o Even 1 liquidation = **big red flag**

5. **Days Active**

   o How long this wallet has been active on Compound

   o Rationale: Age of wallet shows longevity and experience

   o Calculation: (max_timestamp - min_timestamp) / 86400 seconds

   o Older = **more experienced**, safer

6. **Net Position**

   o Did they pull out more money than they put in?

- Rationale: High outflows could signal extraction or unsustainable borrowing

- Calculation: (total_value_out - total_value_in) / 1e18 (converted to ETH)

- Large outflows = **risky**

7. **Transaction Count**

    - Total number of relevant Compound transactions

    - Rationale: Measures engagement with Compound protocol

    - Calculation: Total number of Compound-related transactions (normal + internal)

    - Active users = **less risky** (they know the system)

8. **Average Gas Used**

    - Complex operations cost more gas

    - Rationale: Reflects complexity and purpose of wallet activity

    - Rationale: Reflects complexity and purpose of wallet activity

    - Very high or low gas usage can mean **odd behavior**

9. **Unique Markets Used**

    - Number of different assets interacted with

    - Rationale: Diversification across assets spreads risk

    - More = **diversification**, less risky

10. **Failed Transaction Ratio** (Ratio: 0-1)

- How many of their transactions failed

- Rationale: Transaction failures suggest bad planning or underfunded actions

- Calculation: failed_tx_count / total_tx_count

- High failure rate = **Higher failure rate = higher risk (poor execution)**

🧠 SCORING METHOD OF WALLETS

📍 **What We Did:**

We didn't use labeled data (no "this wallet is risky" tag). Instead, we used **K-Means Clustering**, an unsupervised machine learning method, to group wallets into 5 risk groups based on their behavior.

🔧 **Preprocessing:**

- Missing values (like no activity) were treated as zeros

- All values scaled between 0 and 1 (so that big numbers don't dominate)

🔧 **Clustering**

- **Applied K-Means with k=5**

- **Used random_state=42 and n_init=10 for reproducibility**

🗓️ Risk Formula:

Risk Factor Calculation

Risk Factor = (liquidation_count × 0.30) + ((1 - repay_ratio) × 0.25) + (failed_tx_ratio × 0.20) + (net_position × 0.15) + ((1 - normalized_days_active) × 0.10)

We calculated a **risk factor** using these weights:

Feature	Weight	Why?
Liquidations	30%	Shows financial collapse happened
Repay behavior	25%	Bad repayment = risky
Failed transactions	20%	Shows poor planning
Net outflow	15%	May mean over-leveraging
Account age	10%	New users = less trustworthy

The **higher the risk factor**, the **lower the score**.

🎛️ Score Mapping (0 to 1000 scale):

Cluster	Risk Level	Score
Lowest Risk	Very Safe	900
Low Risk	Responsible	750
Medium Risk	Some concern	600
High Risk	Risky behavior	400
Very High Risk	Dangerous	200

RISK INDICATORS JUSTIFICATION

High Risk Indicators (Lower Scores 200-400):

- Multiple liquidations: Direct evidence of inability to manage collateral ratios

- Low repay ratio (<0.5): Poor repayment behavior indicates default risk

- High failed transaction ratio (>10%): Suggests insufficient funds or poor planning

- Large net outflows: May indicate over-leveraging or capital flight

- New accounts (<30 days active): Lack of proven track record in DeFi

Medium Risk Indicators (Scores 500-700):

- Occasional failed transactions (5-10%): Some operational issues but manageable

- Moderate borrowing activity: Active but not excessive leverage

- Limited market diversification: Concentration risk in few assets

- Moderate account age (30-365 days): Some experience but still developing

Low Risk Indicators (Higher Scores 750-900):

- Zero liquidation history: No evidence of financial distress

- High repay ratio (>0.8): Demonstrates excellent repayment behavior

- Low failed transaction ratio (<5%): Good operational management

- Balanced net position: Sustainable financial management

- Long account history (>365 days): Proven experience through market cycles

- High market diversification: Risk spread across multiple assets

- Consistent activity patterns: Regular, predictable usage

Model Validation & Scalability:

- Unsupervised approach: No need for labeled training data, works with any wallet set

- Feature engineering: Based on established DeFi risk management principles

- Scalable architecture: Can process thousands of wallets with same methodology

- Interpretable results: Clear mapping between features and risk scores

- Robust to outliers: Clustering approach handles extreme values well

Limitations & Considerations:

- Historical focus: Based on past behavior, may not predict future changes

- Protocol-specific: Only analyzes Compound V2 activity, not other DeFi protocols

- Market conditions: Risk patterns may change during different market cycles

- Data availability: Limited by Etherscan API rate limits and data completeness

📃 **What the Scores Mean**

Score Range Meaning

Score Range	Meaning
800–1000	Very Low Risk
600–799	Low Risk
400–599	Medium Risk
200–399	High Risk
0–199	Very High Risk