**Project Report**

**Data Collection Method**

* **Wallet Input:** The process starts with wallets imported from an Excel file, targeting addresses of interest.
* **On-chain Data:** For each wallet and four major Compound cTokens (cDAI, cUSDC, cETH, cWBTC), token transfer history is fetched via the Covalent API.
* **Scope:** Only transactions relevant to Compound V2 cTokens are considered, focusing the analysis on protocol-specific activity.
* **Data Pipeline:**
  + Fetch transfer events per (wallet, cToken) pair.
  + Save all retrieved transactions in a consolidated CSV for further processing.

**Feature Selection Rationale**

| **Feature** | **Rationale** |
| --- | --- |
| total\_value\_transferred | Captures magnitude and financial engagement level. |
| avg\_tx\_value | Flags outliers (e.g., single large transfers vs. regular use). |
| max\_tx\_value | Detects high-risk/whale activity. |
| tx\_count | Measures transaction frequency—can indicate bot/nefarious activity if anomalously high/low. |
| unique\_txns | Helps filter duplicate or repeat transfer patterns. |
| activity\_span\_days | Captures tenure; longer span may indicate legitimacy. |
| tx\_per\_day | Normalizes activity volume over wallet history. |

Features are engineered to capture activity, magnitude, and diversity of wallet interactions, aiming for properties that relate to both protocol use and typical DeFi risk markers:

**Scoring Method**

* **Feature Normalization:** All key features are scaled using Min-Max normalization to balance scales and avoid feature dominance.
* **Clustering for Risk Proxying:** KMeans clustering is leveraged (n=2, “safe” vs. “risky”), assigning proxy risk labels to the dataset. The cluster with higher values in a selected “safer” feature (e.g., average transfer or total transferred) is assigned as “safe.”
* **Supervised Learning (Auto-Labelling):** A Random Forest classifier is then trained on the normalized features with the KMeans cluster assignments as labels, learning to predict “safe” vs. “risky” status by wallet feature patterns.
* **Risk Score Generation:** Output probabilities from the “safe” class prediction are scaled to a 0–1000 range, resulting in a continuous, interpretable risk/credit score.

**Justification of Risk Indicators Used**

* **Transfer Magnitude (total\_value\_transferred, max\_tx\_value):** Large or highly variable values can suggest high-engagement actors, who may be less risky if active over time, but also may represent whales/bots if their activity is spiky.
* **Transaction Frequency (tx\_count, tx\_per\_day):** High-frequency, short-burst activity can sometimes indicate bot usage or Sybil attacks, while steady, long-term patterns are signals of authentic protocol use.
* **Wallet Tenure and Diversity (activity\_span\_days, unique\_txns):** Wallets active over longer periods and with a diversity of transactions are typically less likely to be newly spun up for attack or manipulation.

**Approach Clarity, Justification, and Scalability**

* **Clear Logic Chain:** Features are directly tied to observable patterns in on-chain DeFi risk and utility. Risk scoring uses a fully explainable pipeline: on-chain extraction → feature engineering → unsupervised label assignment → probability-based scoring.
* **Automated Labeling:** No manual labeling is required, making the pipeline scalable to any size of address/tokens list.
* **Extensibility:** The workflow can be expanded with more features (e.g., protocol-specific actions, peer wallet behavior metrics) or adapted to other DeFi protocols by altering the cToken list.
* **Compatibility:** Runs entirely on open-source tools and public APIs; relies on deterministic processing and is suitable for batch or periodic risk scoring in institutional or research settings.