Compound V2 Wallet Risk Scoring: Originalized Deliverable Report

Data Collection Approach

Schema Discovery and Data Sources

The project began by extensively investigating the Compound protocol's subgraphs via The Graph's APIs to capture all relevant wallet behaviours.

- Custom scripts were developed to perform schema introspection and analyze several Compound V2 and V3 subgraphs, mapping out every accessible JSON data structure.
- This analysis confirmed which on-chain fields—such as balances, transaction events, market participation, and historical liquidations—were accessible and reliable for each wallet in scope.

Subgraph Selection and Querying

- We wrote Python programs to generate and execute GraphQL queries for each target address, pulling detailed information about deposits, loans, repayments, withdrawals, and market engagement.
- After evaluating four different subgraphs spanning Compound V2 and V3 on Arbitrum, only one V2 subgraph returned complete and accurate records for all challenge wallets. No V3 subgraph was suitable. As a result, the downstream processing focused exclusively on V2 data, although the tooling is designed for future V3 integration.

Raw Dataset Compilation

The gathered protocol data covering all target wallets was merged and stored as a raw transactions dataset (compound_wallets_raw_data.csv). This dataset served as the primary input for all additional analysis and model building.

Feature Engineering and Selection Rationale

Theory- and Data-Centric Feature Design

Drawing on research literature in DeFi risk and practical protocol experience, features were constructed to reflect both core blockchain risk factors and broader behavioral trends. Selection prioritized risk interpretability, strong empirical backing from exploratory data analysis (EDA), and lack of overlap between features.

Key Feature Types:

- **Monetary flows:** Sums of supplied, borrowed, withdrawn, and repaid funds (in USD)
- **Ratios:**

    - *Collateralization ratio*: Gauges wallet's risk buffer on the protocol
    - *Repayment rate*: Assesses user reliability
    - *Withdraw-to-supply ratio*: Signals asset draining or neglect
- **Participation metrics:** Number of markets accessed, transaction counts for borrowing and depositing, total protocol interactions as a gauge of engagement
- **Risk events:** Records of historical liquidations, Boolean flags for being liquidated and high withdrawal activity
- **Log transformations:** Key financial and activity metrics were also captured on a logarithmic scale to minimize the influence of large outliers ("whales")

Feature Reduction and Validation

Exploratory data analysis highlighted strong correlations among some features, such as between event counts and dollar values. To maintain clarity and avoid double-counting, only one representative from each correlated group was included. Fields dominated by missing or zero values (e.g., certain V3 fields) were either dropped outright or imputed with care to ensure a stable modeling base.

Risk Scoring Methodology

Expert-Informed Heuristic Scoring

- An initial risk assessment used a weighted, rule-based scoring system grounded in domain expertise.
- All features were normalized between 0 (low risk) and 1 (high risk), with extreme values clipped to prevent skew from outliers.
- Weight assignments included: collateralization ratio (0.25), repayment rate (0.20), past liquidations (0.15), withdraw-to-supply ratio (0.15), diversification (0.10), activity (0.10), borrow risk (0.05).
- The composite score was linearly rescaled to a 1–1000 scale, clearly distinguishing between the safest and riskiest wallets.
- Tuning of normalization and scaling accounted for the actual dataset shape, which included many low-activity users and a handful of highly active or risky accounts.

Machine Learning Refinement

- Using the engineered features along with the initial heuristic score, several regression models were trained, including XGBoost, LightGBM, and Random Forest.

- Model comparisons based on RMSE, MAE, and R^2 on held-out validation data favored XGBoost for both accuracy and explainability.
- The winning model was trained on the full dataset to produce refined, learnable risk scores for all reviewed wallets, results of which were saved in final_predictions.csv.

Explanation for Risk Factors

Collateralization Ratio:
A direct indicator of liquidation risk. The safer wallets consistently maintained ratios above 1, while riskier ones hovered near or below this value. This metric directly influenced risk weighting and normalization.

Repayment Rate:
A clear sign of user reliability. Consistent repayments indicated a low-risk wallet, as evidenced by the strong separation seen during analysis.

Historical Liquidations:
Although infrequent, any record of liquidation strongly signaled high risk and required stringent penalization in the final score.

Withdraw-to-Supply Ratio:
Provided insight into whether a wallet had removed a significant share of its funds, suggesting potential churn or risk-minimizing behavior.

Activity and Diversification:
Balanced activity and a wide distribution of assets aligned with safer behavior. Risk stood out among either minimally engaged or overly active/distinctly undiversified users.

Managing Outliers and Redundancy:
Analysis uncovered a small number of outsized “whale” wallets and evident feature dependencies. This led to solution strategies like log transformation, feature clipping, and selective inclusion to ensure the system was neither unfairly skewed nor redundant.

Conclusion

This project produced a comprehensive, scalable risk scoring system for Compound V2 wallets. The approach combined detailed subgraph analysis, streamlined data querying, principled feature design, diligent data processing, an interpretable heuristic scoring baseline, and state-of-the-art regression modeling with XGBoost. Each risk signal was selected and justified based on robust empirical evidence, resulting in a tool that is transparent, robust, and ready for use in compliance or DeFi analytics.