Compound V2 Wallet Risk Scoring: Detailed Project Report

Data Collection Method

Schema Exploration and Sourcing:

To ensure complete and accurate coverage of all relevant wallet behaviors, the project began with systematic exploration of Compound protocol subgraphs using The Graph's APIs.

- Introspection queries were run (via custom scripts) on several candidate Compound V2 and V3 subgraphs to extract and analyze the full schema of available data structures in JSON format.
- This step was critical for confirming which on-chain data fields (balances, events, market activity, liquidations, etc.) could be reliably queried for the full list of provided wallet addresses.

Subgraph Selection and Query Development:

- Dedicated Python scripts were written to construct and execute GraphQL queries for each wallet, gathering all protocol-relevant on-chain data (e.g., deposits, borrows, repays, withdrawals, market participation).
- Four separate Compound V2 and V3 subgraphs on Arbitrum were evaluated. Only one V2 subgraph yielded valid, complete data for the challenge wallet set; no V3 subgraph yielded matching data for the wallet set at all. Thus, all downstream analysis and modeling were V2-based, though the pipeline was designed to flexibly accept V3 data in the future.

Raw Dataset Creation:

Queried data for all target wallets was aggregated and stored in a single raw transaction dataset (compound_wallets_raw_data.csv) that formed the backbone for all subsequent feature extraction and modeling.

Feature Selection Rationale

Theory- and Data-Driven Design:

Drawing from extensive literature on DeFi lending risk and practical experience, features were engineered to capture both direct on-chain risk factors and broader behavioral patterns. Key considerations in selection included direct risk explainability, empirical support from EDA, and non-redundancy.

Key Features Used:

- Monetary_flows: total_supplied_usd, total_borrowed_usd, total_withdrawn_usd, total_repaid_usd
- Ratios: collateralization_ratio (protocol-level risk margin), repayment_rate (user reliability), withdraw_to_supply_ratio (draining/abandonment signal)
- Engagement: number_markets_used, num_borrow_events, num_deposit_events, total_protocol_tx (acts as activity proxy)
- Risk incident flags: liquidations_suffered, has_been_liquidated, high_withdraw_flag
- Non-linear and log features: total_supplied_usd_log, total_withdrawn_usd_log, total_protocol_tx_log captured scale while mitigating outlier/whale effects.

Feature Pruning and Robustness:

EDA revealed several features were highly collinear (e.g., protocol event counts and USD totals), so only one representative from each group was retained to avoid double-counting. Features dominated by missing or zeroed values (e.g., V3 specifics, empty fields) were dropped or carefully imputed to preserve variance and modeling stability.

Scoring Method

Heuristic Rule-Based Scoring:

- A domain-informed, expert-weighted heuristic scoring function was implemented first.
- Features were normalized (0=safest, 1=riskiest), clipped to prevent outlier-dominance, and combined as a weighted sum.
    - Key weights: collateralization ratio (0.25), repayment rate (0.20), liquidations suffered (0.15), withdraw-to-supply ratio (0.15), diversification (0.10), protocol activity (0.10), borrow risk (0.05).
- The formula generated a composite risk value for each wallet, linearly rescaled to a 1–1000 range (1 = safest, 1000 = riskiest) for clarity in reporting.
- Upon close EDA review, normalization and scaling choices were tuned to handle the challenge dataset's real skew: most wallets have low activity, while a few "whales" and risky users define the variance.

Machine Learning Enhancement:

- The engineered features and heuristic score together formed the basis for ML training.
- XGBoost, LightGBM, and Random Forest regressors were compared for predictive capability against the heuristic risk score as the target.
- Model efficacy was assessed primarily with RMSE, MAE, and R^2 on validation splits. XGBoost was selected as the winning model for its best RMSE and interpretability.
- The final model, trained on the complete feature set, generated refined, learnable risk predictions for all wallets, output in final_predictions.csv.

Justification of Risk Indicators Used

a. **Collateralization Ratio**

   Strongly predictive of liquidation risk; EDA revealed that most “safe” wallets cluster above 1, with a minority near threshold or below. This indicator was heavily weighted and rigorously normalized.

b. **Repayment Rate**

   Demonstrates user reliability; EDA confirmed near-binary separation between non-repayers (high risk) and consistent repayers (safe).

c. **Liquidations Suffered**

   Historical liquidations, though rare, are a powerful risk signal. EDA showed these events cluster in a small wallet subset, supporting strong penalty calibration.

d. **Withdraw-to-Supply Ratio**

   Observed in EDA as a strong separator between wallets still exposed to protocol risk and those that have “drained” assets—thus, critical for risk of exit behavior or strategic non-participation.

e. **Activity & Diversification**

   Moderate protocol activity and broader asset use both correlated with safety. Extremely low or high activity, or lack of diversification, was associated with edge-case risky behaviors.

f. **Handling of Outliers and Redundancy**

Deep EDA revealed “whale” outliers and strong feature correlations, driving clipping, log transformation, and careful feature selection to avoid overweighting or spurious model sensitivity.

Conclusion

This project delivered a robust, end-to-end on-chain wallet risk scoring system. The workflow integrated schema introspection, precision subgraph querying, EDA-driven feature engineering, rigorous cleaning, interpretability-focused heuristic scoring, and state-of-the-art ML regression (XGBoost). Each risk signal was empirically vetted, weighted, and validated via both rule-based and machine learning methods, ensuring clarity, scalability, and predictive strength for operational or compliance DeFi risk tools.