Predicting Empire Collapse through Simulation Data from Age of Empires II

By Rafael Sanchez

CPE 646 Final Project Report

December 12, 2025

Abstract

This project investigates whether signals of systemic decline can be detected within a controlled competitive environment by analyzing high-dimensional match data from Age of Empires II: Definitive Edition. Prior research on the rise and collapse of complex systems has been limited by sparse historical records, retrospective analysis, and high levels of structural noise. In contrast, real-time strategy games provide rule-based economies with clear resource constraints, enforced trade-offs, and observable life-cycles from emergence to defeat. A gradient-boosted decision tree classifier (XGBoost) is trained to predict eventual victory from any given game state, and Shapley Additive Explanations (SHAP) are used to identify the variables most associated with winning or losing. On average, matches exhibit a "point of no return" near the middle of play, after which the losing side rarely recovers despite continuing to field units or infrastructure. Feature attribution confirms that economic degradation, especially villager deficits, precedes military defeat, mirroring theoretical expectations that collapse originates in lost productive capacity rather than final-stage conflict. While the dataset is limited in size and scope, the results suggest that competitive strategy environments generate measurable, interpretable signatures of decline, and that structural irreversibility can be inferred well before visible defeat. These findings demonstrate how stylized simulations and machine-learning methods can complement collapse research by providing controlled, time-resolved observations unavailable in historical data.

Section 1: Background Research

Understanding the rise and fall of complex systems has long been a central concern across disciplines including economics, political science, ecology, and systems dynamics. As modern societies face increasing pressures from resource depletion, technological disruption, climate change, and geopolitical instability, interest in identifying early indicators of systemic decline has intensified. Despite this interest, empirical research on collapse remains constrained by the scarcity, incompleteness, and noisiness of real-world historical data. As a result, much of the existing literature relies on theoretical mathematical models and, although useful, are still limited by the design of their system. One of the most influential early attempts to formally model systemic collapse is the World3 model developed in the early 1970s as part of the Limits to Growth study by the MIT Systems Dynamics Group. The central conclusion of the study was that, in the absence of substantial changes in policy and behavior, exponential growth in a finite-resource system would likely result in eventual collapse within the mid twenty-first century (Meadows et al., 1972).

While World3 demonstrated the power of aggregate, feedback-driven modeling, it operated at a highly abstract global level and did not incorporate micro-level decision-making, spatial competition, or strategic interaction between agents. Later work sought to address these limitations by modeling empires and states as spatially distributed systems subject to internal and external pressures. A notable example is the dynamic macro-spatial model of empire growth and collapse proposed in “Modelling of Growth and Collapse of Empires” by Senior Researcher of Economics Yuri Yegorov. This work uses a relatively simple growth framework that allocates output among consumption, investment, and defense while explicitly modeling territorial expansion. A key insight of this model is that empire growth is mainly driven by structural constraints and strategic pressures. Unlike standard economic models, it deliberately avoids assumptions of rational optimization, arguing that pre-industrial rulers were often motivated by non-economic objectives such as territorial expansion or population control. (Yegorov, 2018) The model demonstrates that empires tend to expand until they reach an upper spatial limit imposed by rising transport costs and external resistance, at which point the system becomes fragile. Collapse can then be triggered by relatively small shocks, such as sudden increases in transport costs or coordinated external pressure. (Yegorov, 2018) While this work provides valuable theoretical insight into mechanisms of growth and collapse, it remains largely conceptual and relies on simplified assumptions that are difficult to validate empirically, particularly given the limited availability of fine-grained historical data.

Parallel lines of research in sustainability science and complex systems theory further reinforce the idea that collapse is often systemic rather than localized. Models based on

Lotka-Volterra dynamics, for example, have been used to study sustainability risks arising from interacting economic, environmental, and social subsystems. (Wang et al., 2025) This literature emphasizes that failure in one component of a tightly coupled system can propagate rapidly, leading to cascading breakdowns. Despite the widespread adoption of sustainability metrics such as ESG reporting and global policy initiatives like the UN Sustainable Development Goals, empirical evidence suggests that many large-scale risks, particularly those related to climate change and technological disruption, continue to intensify. (Wang et al., 2025)

Taken together, existing research highlights two persistent challenges in the study of collapse. First, real-world data on the rise and fall of complex systems is sparse, noisy, and often only available after collapse has already occurred. Second, many models operate either at a highly abstract level or within narrow empirical contexts, making it difficult to study how micro-level decisions at each time step throughout the lifecycle of an empire affect it. These limitations motivate the exploration of alternative data sources and modeling environments that allow for controlled observation of growth, competition, and collapse dynamics at scale.

Section 2: Methodology

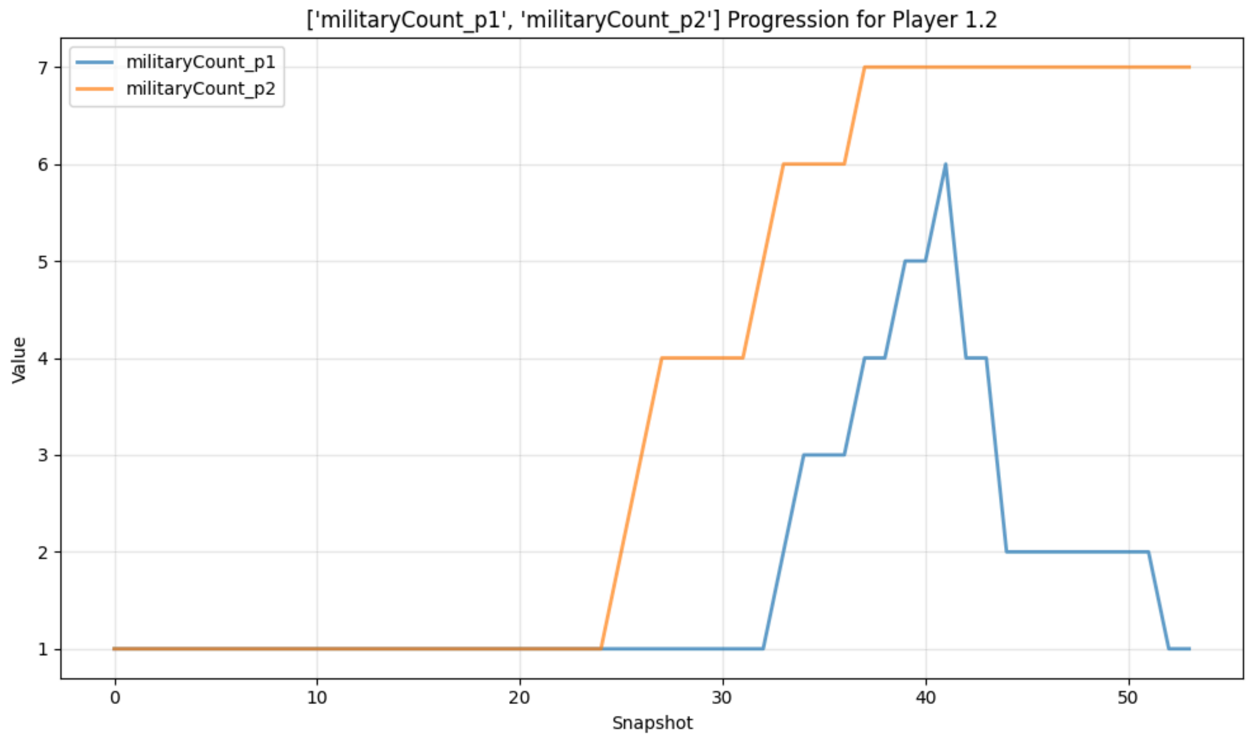
Age of Empires II: The Age of Kings is a real-time strategy (RTS) game originally developed by Ensemble Studios and released in 1999. Set broadly within a medieval historical framework, the game places players in control of a civilization tasked with managing resources, developing technology, expanding territory, and engaging in military conflict against rival empires. (“Age of Empires II”) Over two decades after its original release, Age of Empires II has remained highly active, with the release of Age of Empires II: Definitive Edition in 2019. Today, the game is widely regarded as one of the most influential and enduring titles in the RTS genre.

From a modeling perspective, Age of Empires II provides a compelling proxy for studying empire growth and collapse because it explicitly encodes many of the structural constraints faced by real-world civilizations. Players must balance finite resources like food, wood, gold, and stone while allocating labor, investing in technology, maintaining military strength, and defending territory. Economic expansion, technological advancement, and military power are tightly coupled, and misallocation in any one dimension can lead to rapid decline. Importantly, the game enforces trade-offs: investing heavily in military power can starve economic growth, while excessive economic expansion without defense can leave an empire vulnerable to an attack from the opponent. The game also unfolds over distinct developmental stages, referred to as Ages, which loosely correspond to historical periods of societal advancement. Progression between Ages requires meeting specific economic and infrastructural prerequisites (“Age of Empires II”), mirroring the notion that civilizations cannot advance technologically without sufficient economic foundations. Failure to transition effectively between these stages often places players at a structural disadvantage that is difficult to reverse.

Unlike abstract theoretical models, Age of Empires II offers an unusually rich empirical advantage: scale. Competitive multiplayer matches generate large volumes of structured, time-resolved data capturing economic performance, military activity, technological choices, resource management, and population dynamics throughout the course of a match at each time step. Because all players operate under identical rules and constraints, the resulting dataset enables controlled comparative analysis across thousands of independent “civilizations,” each representing a complete life cycle from emergence to victory or collapse. This level of granularity and consistency is rarely available in historical or contemporary geopolitical data. The goal of this study is not to claim that in-game collapse directly mirrors real-world societal collapse, but rather to investigate whether consistent, interpretable patterns of decline emerge within a constrained competitive system, and to explore how those patterns compare to established theories of systemic failure.

The dataset used in this project consists of recorded Age of Empires II: Definitive Edition multiplayer matches, stored in a custom JSON-based “record” format. Each record corresponds to a complete match and contains a sequence of time-indexed snapshots capturing the evolving game state. Each match record includes: a list of game state snapshots, the final winner of the game, and a structured representation of each player’s economic, military, technological, and population state at each snapshot.

For example, the diagram below records the military unit counts for Player 1 and Player 2 during a match, with Player 1 losing their units at the end sharply which symbolize a major battle.



For each match, the two players are identified and tracked consistently across all snapshots. For each snapshot, raw state features are extracted separately for each player, including:

- Economic indicators (resources, villagers, total resource accumulation)
- Military indicators (unit counts, kills, deaths)
- Infrastructure indicators (building counts by type)
- Technological indicators (current Age, research counts)
- Population and map exploration metrics

Building types are not aggregated indiscriminately. Instead, strategically significant structures such as Town Centers, Castles, and military production buildings are preserved as separate features.

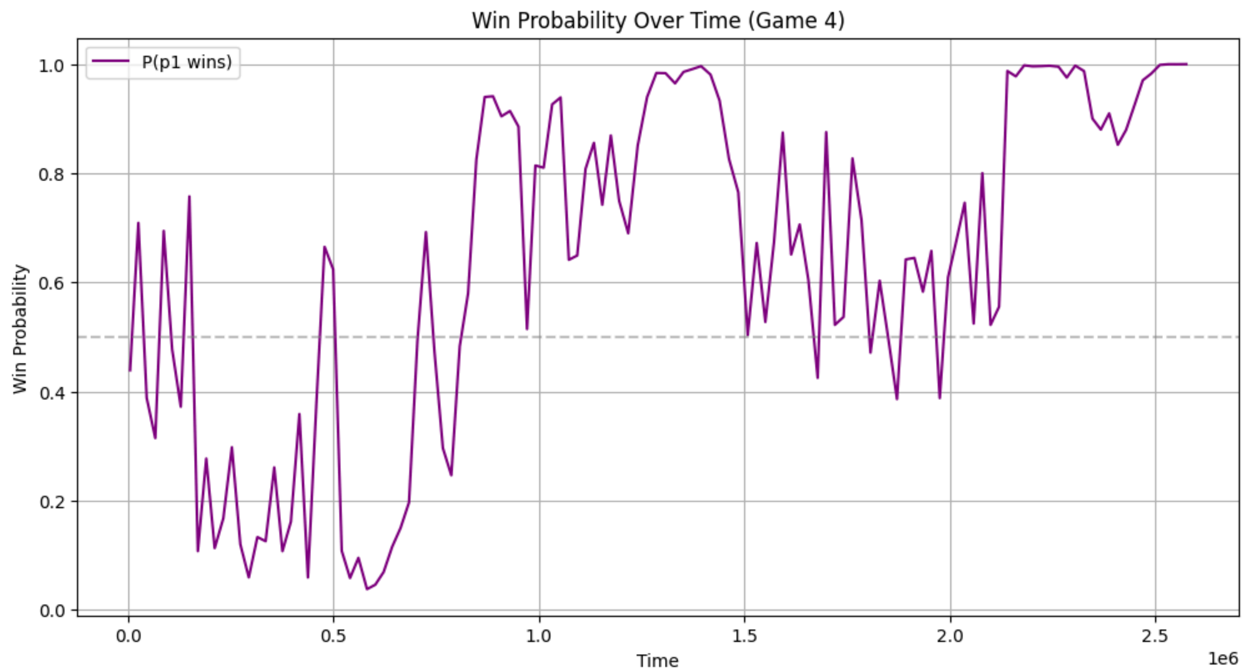
In addition to absolute player states, different features are computed at each snapshot to capture relative advantage, since just feeding this data at each step doesn't contain information about the match up until that moment. So we also include calculated differences between Player 1 and Player 2 for each relevant variable (e.g., villager difference, military difference, building difference), and momentum features, measuring average change over a fixed window of multiple previous snapshots. These features are computed independently for each player and capture growth rates, acceleration, stagnation, and decline. This allows each snapshot to contain a compact representation of recent history, enabling the model to reason about trends rather than just static values alone.

The predictive task is formulated as a binary classification problem: given a snapshot of the game state, predict which of the two players will eventually win the match.
A gradient-boosted decision tree model (XGBoost) is used for this task. This model was selected for several reasons:

- Strong performance on structured, tabular data
- Ability to capture nonlinear interactions between variables
- Compatibility with probabilistic outputs
- Native support for interpretable feature attribution methods

To avoid data leakage, matches are split at the game level, not the snapshot level. All snapshots from a given match are assigned exclusively to either the training set or the test set. Approximately 80% of matches are used for training and 20% for evaluation. The model is trained using log-loss optimization, producing probabilistic predictions that represent the estimated likelihood of one player's eventual victory at each snapshot.

Rather than relying solely on categorical predictions, the trained model produces win probability estimates for each snapshot. This enables time-resolved analysis of how confidence in a player's victory evolves throughout the match. Standard evaluation metrics such as accuracy and ROC-AUC are computed on matches to assess overall predictive performance, but this can be misleading because of the time-specific nature of the data. As such, the primary analytical focus is on within-match probability trajectories, which allow identification of turning points, collapse phases, and periods of uncertainty. Here is an example of what the output of the model's confidence percentage of Player 1 winning is in the following game:



Finally, To analyze which features contribute most strongly to predictions, SHAP (Shapley Additive Explanations) is applied to the trained model. SHAP provides a principled, game-theoretic method for attributing each prediction to individual features. By aggregating SHAP values across a representative subset of training snapshots, global feature importance profiles are obtained. These profiles reveal which economic, military, technological, and temporal factors most consistently influence predicted outcomes. This interpretability framework enables direct comparison between empirical model behavior and theoretical expectations from systems collapse literature.

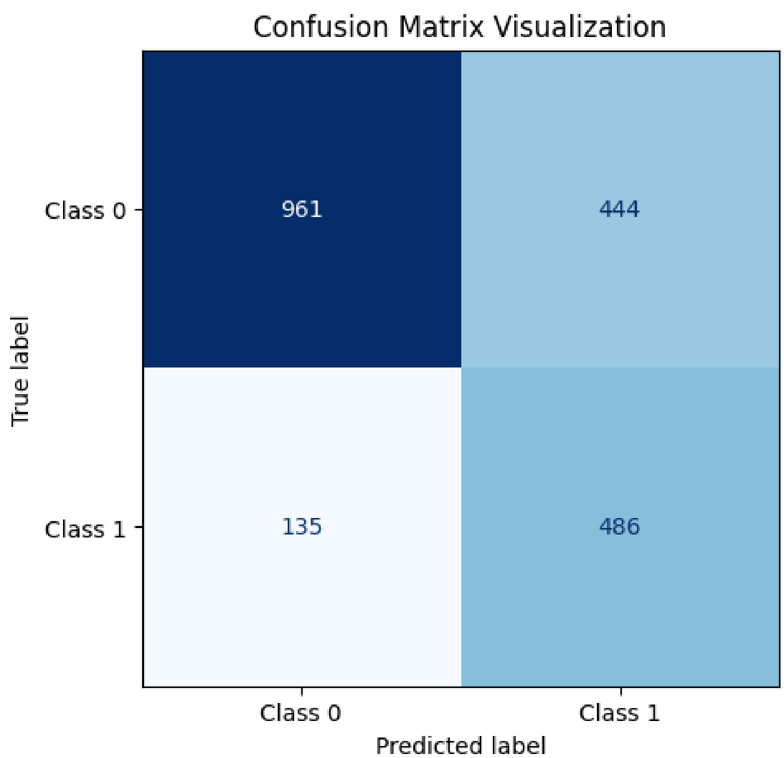
All data processing, feature engineering, model training, and analysis are implemented in Python Jupyter Notebook using NumPy, Pandas, Scikit-learn, XGBoost, and SHAP. The trained model is saved to disk after training and can be reloaded for evaluation for further analysis without retraining.

Section 3: Results and Analysis

The predictive model was trained using snapshot-level features from each recorded match, treating each moment in time as an independent probability estimate of the final outcome. After training, we evaluated the model using standard classification metrics. The overall confusion matrix for the held-out test set is shown below:

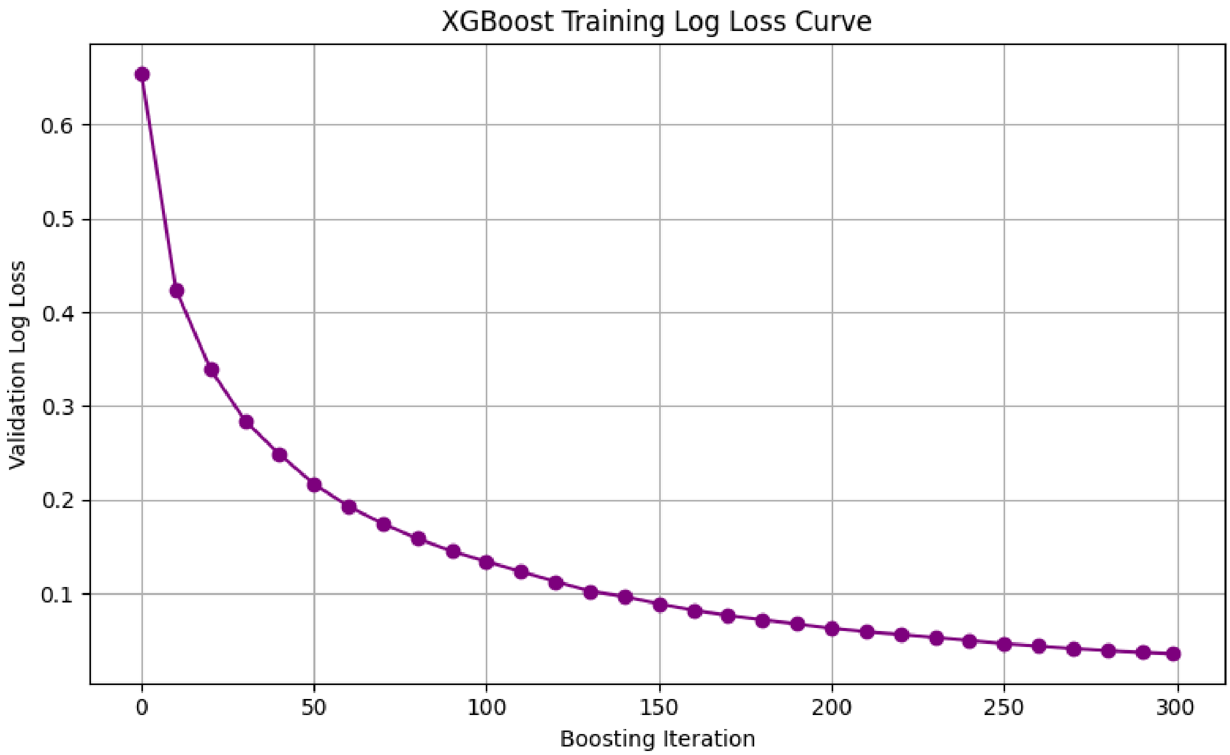
CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.88	0.68	0.77	1405
1	0.52	0.78	0.63	621
accuracy			0.71	2026
macro avg	0.70	0.73	0.70	2026
weighted avg	0.77	0.71	0.73	2026



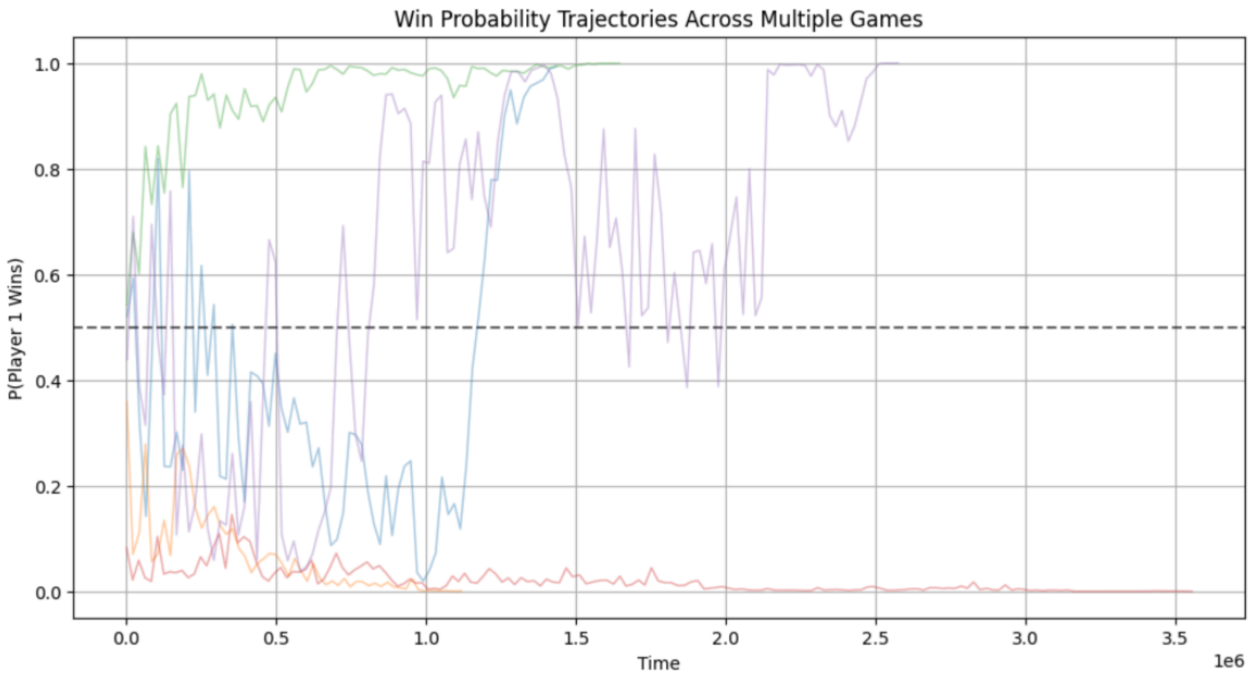
Class 0 means Player 1 is selected as the one who won a match, and Class 1 as Player 2.

In addition, the model's learning progression was tracked through log-loss on the training set, indicating stable convergence:



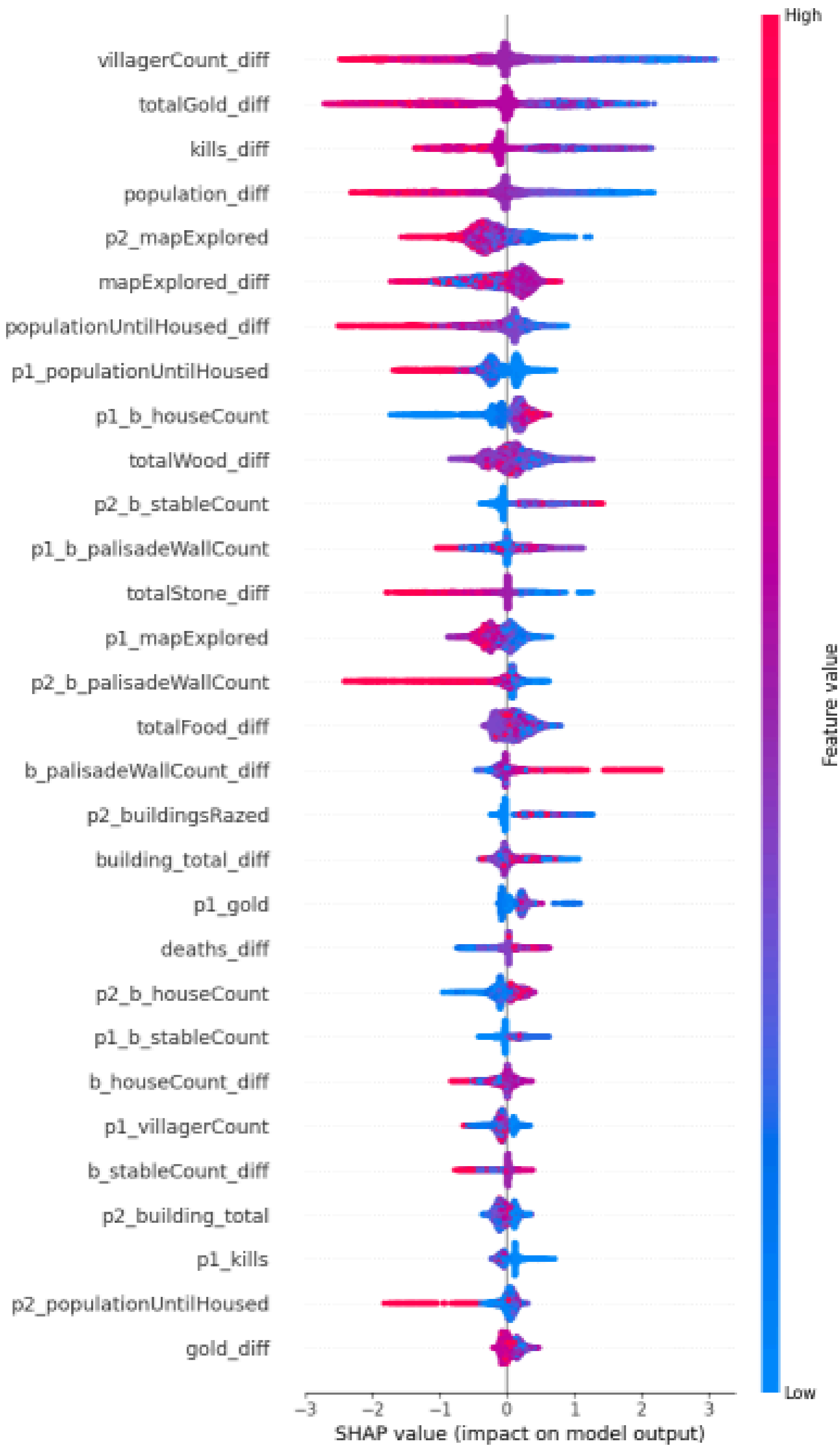
However, confusion matrices are not the real story because the data is inherently temporal in nature. For a real time strategy type of predictor, we have to map the predictions over time for a particular game to see if there is some meaningful progression over time. Ideally, the probability of winning (or losing) should go up or down as you get closer and closer towards the end of the match when one of the players has clearly achieved victory.

To visualize this, we plotted the model's predicted probability of Player 1 winning across time for many different matches. In every case, the final predicted probability aligns perfectly with the actual outcome. This means the probability converges toward 1.0 for Player 1 victories and toward 0.0 for Player 1 defeats (Player 2 wins):



These trajectories are informative. Early-game probabilities are noisy and often oscillate. Mid-game probabilities reflect economic tempo, military engagements, and map control. Late-game probabilities often stabilize sharply, indicating that once the losing side has lost production capacity or economic base, the outcome becomes structurally irreversible.

We next examined feature importance using SHAP, which quantifies each variable's marginal contribution to the win-probability model across thousands of decision trees. The SHAP summary diagram ranks features by global influence:



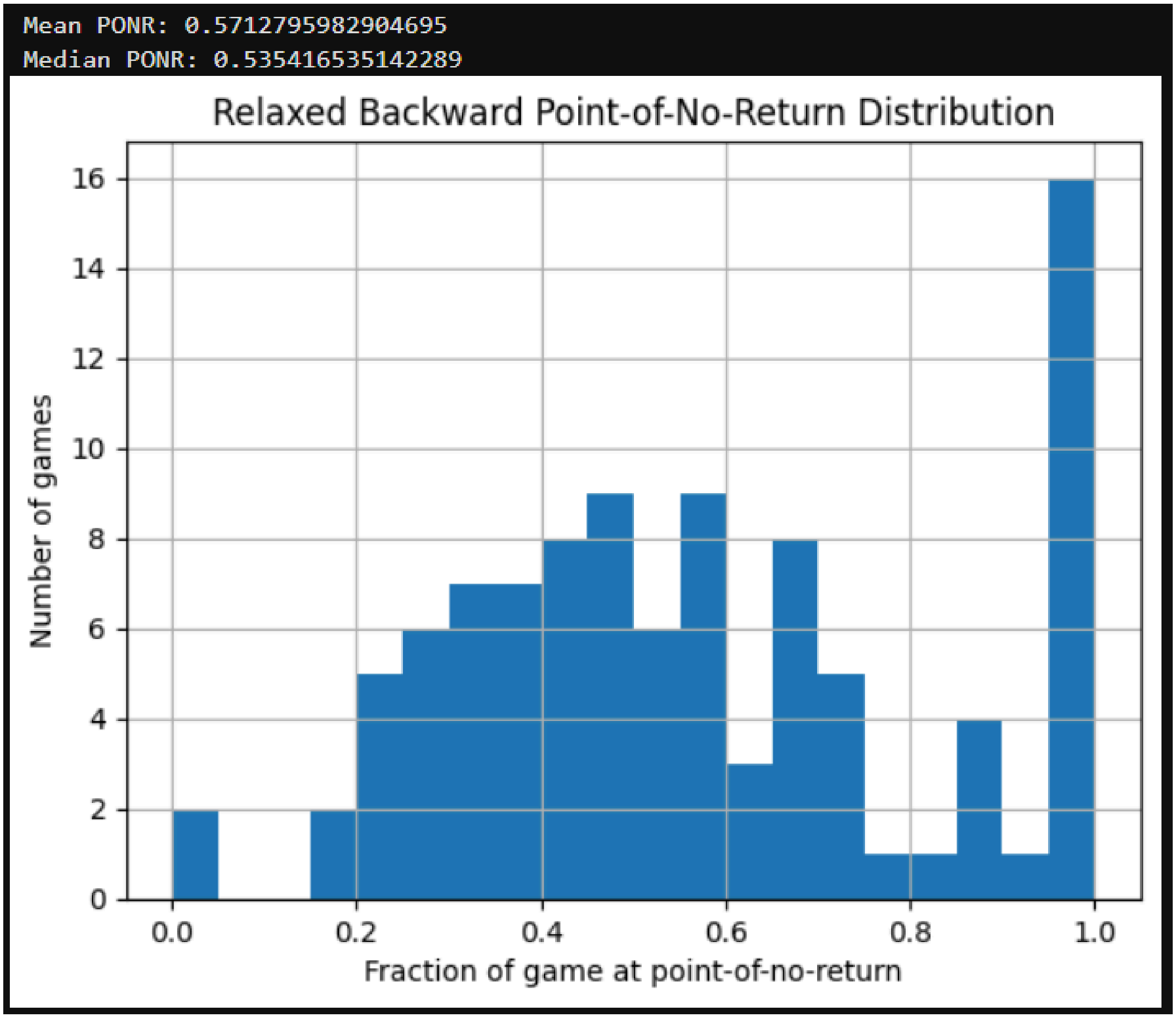
High
Feature value
Low
villagerCount_diff
totalGold_diff
kills_diff
population_diff
p2_mapExplored
mapExplored_diff
populationUntilHoused_diff
p1_populationUntilHoused
p1_b_houseCount
totalWood_diff
p2_b_stableCount
p1_b_palisadeWallCount
totalStone_diff
p1_mapExplored
p2_b_palisadeWallCount
totalFood_diff
b_palisadeWallCount_diff
p2_buildingsRazed
building_total_diff
p1_gold
deaths_diff
p2_b_houseCount
p1_b_stableCount
b_houseCount_diff
p1_villagerCount
b_stableCount_diff
p2_building_total
p1_kills
p2_populationUntilHoused
gold_diff
−3 −2 −1 0 1 2 3
SHAP value (impact on model output)

This figure should be read from top to bottom in order of global importance. The features at the top influenced the model's predictions more than those at the bottom. Each horizontal strip corresponds to a single feature, and every point on that strip represents one training example. The horizontal axis shows the SHAP value, which is how much that feature pushed the prediction toward Player 1 winning (positive direction) or toward Player 1 losing (negative direction). The color encodes the raw feature value for that example: red means a high value of that feature; blue means a low value. So a cluster of red points on the right tells that high values for that feature tend to push the model toward predicting a Player 1 win. Conversely, blue points on the left indicate that low values tend to push toward predicting Player 1 losing.

Interpreting SHAP requires care. The model is not asserting that "more villagers automatically wins the game" or "high gold causes victory." Rather, the model captures the historical regularities in our dataset: a strong villager population historically correlates with superior economic throughput; economic throughput funds military investment; military investment drives kills, razes buildings, and captures map control. These cascades move SHAP contributions. The model therefore detects the consequences of strategic choices.

Another central conceptual metric in this study is what we termed the Point of No Return (PONR), which is the point in a match at which the model becomes ≥ x% confident in the correct eventual winner and never again reverses that judgment for the remainder of the match. This measures irreversibility. Operationally, for each game, the model produced win-probabilities at every snapshot. We conditioned on the actual winner and converted all probabilities into "confidence in the true winner." Beginning at the final snapshot, we moved backward in time until the model's confidence fell below a given certainty threshold, which we decided to be 97%. The earliest time from which confidence never again dipped below that threshold was recorded. Finally, this time was normalized by the total game length.

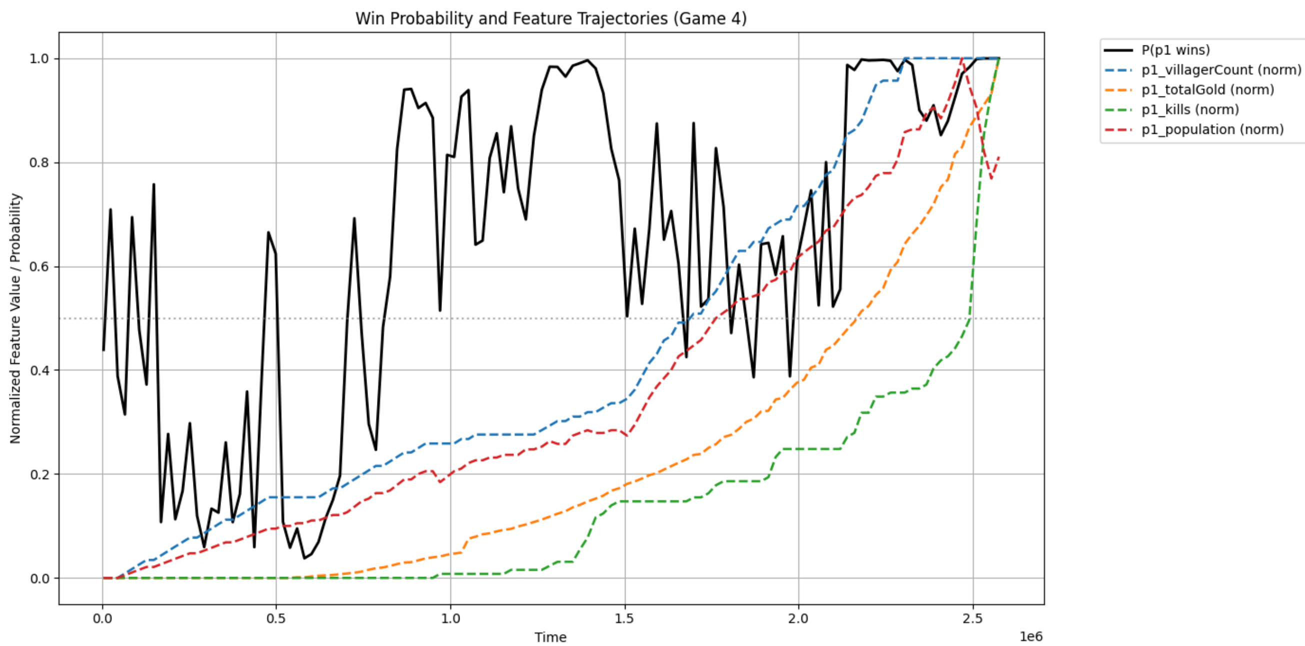
This statistic asks at what fraction of historical play was the model 97% certain this player would not change their perceived outcome? The median and mean PONR across games was recorded below as well as each individual's game PONR:

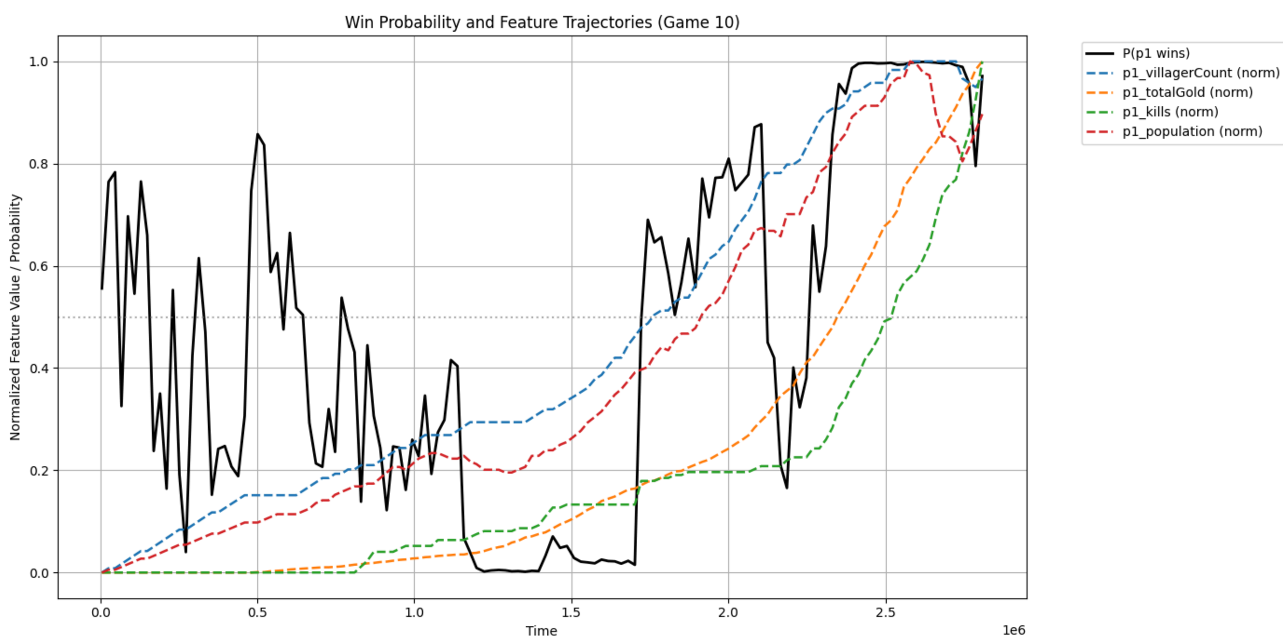


This means that on average around the 57% mark of a match, the losing player was generally already in unrecoverable decline. This aligns with strategic intuition. Empires generally collapse when: production stops at large scale, resources dry up, villagers die, military cannot be replaced or outcompete the enemy's military.

What follows (and the reason why this certainty occurs around what appears to be the mid-point mark rather than later in the game, with some games having that outcome decided only close to the end due to high uncertainty) is because of "garbage time": a visible shell persists while the interior system has already died. Many games continue physically for a while while the outcome is statistically over and there is no more meaningful play going on. Interestingly, a minority of matches register extremely early certainty (near 0-10%). These outliers often correspond to: mismatched skill levels, failed openings, snowballing rushes, etc.

Finally, we also plotted key features versus probability for individual matches. For example, villager differential often tracks closely with win-probability:





When villager production collapses, the probability curve mirrors the economic crash. Similar behavior is seen for: gold income, military count, kill/death differential. Keep in mind, this doesn't mean they are the only ones that affect the outcome as each game has its own set of important features, but overall this mattered when each player was achieving victory because of their implications in the game.

Section 4: Conclusion

This dataset is a stylized economy, in a bounded map, interpreted through a probability model, and unfortunately there wasn't as much data as we would have hoped to gather (around 100 different matches). However, several clear parallels emerge: collapse is path dependent, and mistakes propagate unless corrected before it's too late. In both AOE2 and historical systems, misallocated labor or premature militarization produces compounding deficits. Collapse is economic before it is military. Military defeat is usually a symptom of economic exhaustion, not the initiating cause. When villagers die faster than they can be replaced, decline accelerates. Another conclusion that can be drawn is that irreversibility emerges before visible failure. The losing empire still has buildings. It still has units. But probably it has already lost the game because for whatever reason, they failed to outsmart their opponent. This mirrors real empires where collapse is not announced when the capital burns. History has shown that collapse usually has begun long before the visible collapse occurs. Once odds cross irreversible thresholds, recovery is statistically rare.

Overall, the study shows that even in a stylized competitive environment, collapse is best understood not as a dramatic final moment, but as a gradual loss of adaptive capacity. The "dead empire" persists visually long after its systemic viability has ended. This analytic structure establishes a framework for mapping collapse dynamics in environments where traditional empirical data is sparse or unavailable. For better predictions and future study, more data should be collected and more variability. This data was only limited to 2 players with equally spread resources throughout, but more unequal systems should be studied more as that mirrors the real world. This study does not claim predictive power for geopolitics. Instead, it demonstrates with certainty how competitive systems produce detectable signatures of decline, and how statistical irreversibility precedes visible collapse.

Works Cited

aoerecinfo. (n.d.). aoe2dataset. GitHub repository. Retrieved from aoerecinfo/aoe2dataset: Age of Empires II Definitive Edition rec analysis dataset

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. III. (1972). The limits to growth: A report for the Club of Rome's project on the predicament of mankind. Retrieved from World3 (Limits to Growth) - The Study of Progress

Microsoft. (n.d.). Age of Empires II. In Wikipedia. Retrieved December 12, 2025, from Age of Empires II - Wikipedia

rsanche4. (n.d.). AoE2-RTS-XGBoostPredictorModel. GitHub repository. Retrieved from rsanche4/AoE2-RTS-XGBoostPredictorModel

Wang, Y., & Zhi, T. (2025). Sustainability risks under Lotka-Volterra dynamics (arXiv:2509.04780). Retrieved from Sustainability Risks under Lotka-Volterra Dynamics

Yegorov, Y. (2018). Modelling of growth and collapse of empires [Conference paper]. 14th Viennese Conference on Optimal Control and Dynamic Games. Retrieved from (PDF) Modelling of Growth and Collapse of Empires