Feature Engineering for Corporate Financial Analysis

Emma Nagy

February 3, 2026

Abstract

This study develops and validates a comprehensive set of financial statement features using COMPUSTAT data spanning 1996–2022. Twenty-four financial ratios are engineered across five fundamental categories: liquidity, leverage, profitability, cash generation, and operational efficiency. Through correlation analysis, temporal distributional assessment, and principal component analysis (PCA), these features are shown to capture distinct but economically related signals. Key findings include: (1) correlation structures exhibit substantial regime-dependence, with changes up to ± 0.87 between recession and expansion periods; (2) distributions show clear cyclical patterns aligned with NBER recessions, validating economic relevance; (3) PCA reveals three interpretable factors—financial health (25% variance), business model intensity (9%), and growth/liquidity (8%)—that align with financial theory. These validated features provide a robust foundation for corporate finance prediction tasks.

1 Introduction

Corporate financial analysis relies heavily on engineered features derived from financial statements, yet the systematic validation of these features is rarely undertaken. The challenge lies in constructing financial metrics that capture distinct economic signals while remaining interpretable and stable across different economic conditions. Traditional approaches often rely on ad-hoc feature selection or predetermined ratios from academic literature, without rigorous testing of their information content or temporal stability.

This study addresses three critical gaps in feature engineering for corporate finance applications. First, most studies do not explicitly validate whether selected features provide distinct information or simply duplicate the same underlying signals. Second, temporal stability of financial ratios—particularly during economic stress periods—is rarely examined.

Third, the economic interpretability of feature combinations is often sacrificed in favor of purely statistical dimensionality reduction.

This paper contributes to the literature by developing a systematic feature engineering framework that addresses all three concerns. Using annual COMPUSTAT data from 1996 to 2022, 24 financial ratios are constructed spanning the fundamental dimensions of corporate finance: liquidity management, capital structure, operational profitability, cash generation capacity, and composite distress indicators. These features are then validated through four complementary analyses:

1. **Correlation analysis** to identify redundant features and confirm expected economic relationships
2. **Regime-dependent correlation analysis** using NBER recession indicators to test stability across economic conditions
3. **Temporal distributional analysis** examining how firm financial characteristics evolved from 1996 to 2022
4. **Principal component analysis** to identify latent factors and assess dimensionality

The analysis reveals several important findings. Financial relationships undergo substantial regime shifts during recessions, with correlation changes as large as ± 0.87 for key variable pairs. This finding validates the importance of time-varying risk models rather than static correlation assumptions. Temporal distributions show clear cyclical patterns aligned with major economic events (2001 dot-com crash, 2008–09 financial crisis, 2020 COVID recession), confirming the economic relevance of selected features. Finally, PCA identifies three interpretable latent factors that align precisely with financial theory: a distress factor (25% of variance), a business model intensity factor (9%), and a growth/liquidity factor (8%).

The remainder of this report is organized as follows. Section 2 describes the data sources, variable construction, and analytical methodology. Section 3 presents the main empirical results including correlation analysis, temporal patterns, and PCA findings. Section 4 discusses the economic interpretation and implications for bankruptcy prediction models. Section 5 concludes with practical recommendations and future research directions.

2 Data and Methodology

2.1 Data Sources

The primary data source is COMPUSTAT Fundamentals Annual, which provides standardized financial statement data for publicly traded U.S. companies. Annual observations are extracted from fiscal years 1996 through 2022, yielding approximately 540,000 firm-year observations covering over 30,000 unique firms (identified by `gvkey`).

To enable regime-dependent analysis, NBER recession indicators from the Federal Reserve Economic Data (FRED) database are merged with the financial data. The USREC series identifies official recession periods as determined by the National Bureau of Economic Research. Monthly recession indicators are aggregated to the annual level using a maximum operator: if any month in a fiscal year is classified as recession, the entire year is flagged. This conservative approach ensures crisis periods are fully captured.

The final merged dataset contains 295,612 firm-year observations, of which 14.5% (42,788 observations) occur during recession years (2001, 2008, 2009, 2020). This proportion aligns closely with the 14.8% of sample years classified as recessions (4 of 27 years).

2.2 Feature Engineering

Twenty-four financial ratios are engineered and organized into five economically motivated categories. Table 1 summarizes the features, their calculation formulas, and economic rationale.

Table 1: Engineered Financial Features

Category	Features	Economic Rationale
Liquidity	Quick Ratio, Cash Buffer, Short-term Pressure	Measures ability to meet short-term obligations; critical for assessing financial stability
Leverage	Book Leverage, Long-term Debt/Assets	Indicates financial risk and capital structure decisions
Profitability	Operating Margin, ROA, Asset Turnover	Core operational efficiency metrics; sustained losses precede most failures
Cash Flow	OCF/Assets, Internal CF/Assets, CapEx/Assets, Dividends/Assets, ΔWC/Assets	Cash generation more reliable than accrual earnings for health assessment
Working Capital	DSO, DIO, DPO, Cash Conversion Cycle	Reveals operational efficiency and liquidity management practices
Financing	Net Debt Issued, Net Equity Issued	Tracks external financing needs and capital market access
Valuation	Market-to-Book, Log(Sales)	Size and growth expectations; market perception of firm value
Composite Scores	Ohlson O-Score, Altman Z-Score	Multi-factor indicators combining financial ratios into summary measures
Asset Structure	Tangibility (PPE/Assets)	Asset redeployability affects recovery rates and credit risk

Construction Details: All ratios are computed using standard COMPUSTAT variable mnemonics. For instance, Book Leverage is total debt (`dt`) divided by total assets (`at`); Operating Cash Flow to Assets is `oancf/at`. Working capital metrics (DSO, DIO, DPO) are calculated using the standard formulas:

$$\text{DSO} = \frac{\text{Accounts Receivable}}{\text{Revenue}} \times 365$$
$$\text{DIO} = \frac{\text{Inventory}}{\text{COGS}} \times 365$$
$$\text{DPO} = \frac{\text{Accounts Payable}}{\text{COGS}} \times 365$$

The Ohlson O-Score and Altman Z-Score are computed using their original formulations from Ohlson (1980) and Altman (1968), respectively. These composite indicators combine

multiple financial ratios into summary measures of corporate financial health.

2.3 Analytical Methods

2.3.1 Correlation Analysis

Pearson correlation matrices are computed to identify redundant features and validate expected economic relationships. High correlations ($|r| > 0.8$) between conceptually similar ratios (e.g., OCF/Assets and Internal CF/Assets) are expected and acceptable, as they measure closely related constructs. However, perfect correlations ($r = 1.0$) indicate duplicates that should be removed.

To assess regime-dependence, the sample is split into recession years (USREC = 1) and expansion years (USREC = 0), separate correlation matrices are computed for each regime, and the difference matrix is examined ($\Delta r = r_{\text{recession}} - r_{\text{expansion}}$). Large changes ($|\Delta r| > 0.3$) indicate relationships that fundamentally alter during economic stress.

2.3.2 Temporal Distributional Analysis

For five representative features spanning different categories (Quick Ratio, ROA, Book Leverage, Asset Turnover, Altman Z-Score), faceted histograms are created showing the distribution in each year from 1996 to 2022. To improve visualization of long-tailed distributions, values are winsorized at the 1st and 99th percentiles. Recession years are highlighted with pink shading to facilitate visual identification of crisis-period distributional shifts.

These faceted plots reveal: (1) whether distributions are stable over time or exhibit secular trends; (2) how crisis periods affect the cross-sectional distribution of firm characteristics; (3) whether changes are temporary (quick reversion) or persistent (structural breaks).

2.3.3 Feature Scaling and Principal Component Analysis

Financial ratios exhibit diverse scales (e.g., leverage ratios near 0–1 vs. Z-scores ranging from negative to >10) and distributional shapes (symmetric vs. heavily right-skewed). To prepare features for PCA, which assumes centered, unit-variance data, a two-stage scaling strategy is implemented:

1. **Log transformation** ($\log(1 + x)$) for right-skewed, nonnegative features to compress outliers and reduce skewness

2. **Standardization** (z-scores) for all features to achieve zero mean and unit variance

The log-transform group includes: cash buffer, long-term debt/assets, asset turnover, CapEx/assets, dividends/assets, DSO, DIO, DPO, and market-to-book. All other features are standardized directly. Missing values (which occur sporadically due to data gaps in COMPUSTAT) are median-imputed before scaling to avoid information loss.

PCA is performed on the last two fiscal years of data (2021–2022, $n = 21{,}851$) to assess the dimensionality of the feature space and identify latent factors. The choice of recent years ensures results reflect the current economic environment most relevant for deploying prediction models. PCA is implemented using scikit-learn's `PCA` class with default settings (covariance-based decomposition).

3 Results

3.1 Correlation Structure

Figure 1 presents the full correlation matrix for all 24 features computed on the entire sample. The matrix reveals expected patterns: leverage ratios cluster together, cash flow metrics are highly correlated, and working capital components (DSO, DIO, DPO) form a coherent block. Several feature pairs exhibit very high correlations ($r > 0.8$):

- OCF/Assets $\leftrightarrow$ Internal CF/Assets: $r = 0.999$ (expected; differ only by dividends)
- DPO $\leftrightarrow$ Cash Conversion Cycle: $r = -1.000$ (by construction; CCC = DSO + DIO − DPO)
- Short-term Pressure $\leftrightarrow$ Current Maturity to Net Assets: $r = 1.000$ (duplicate feature, removed)

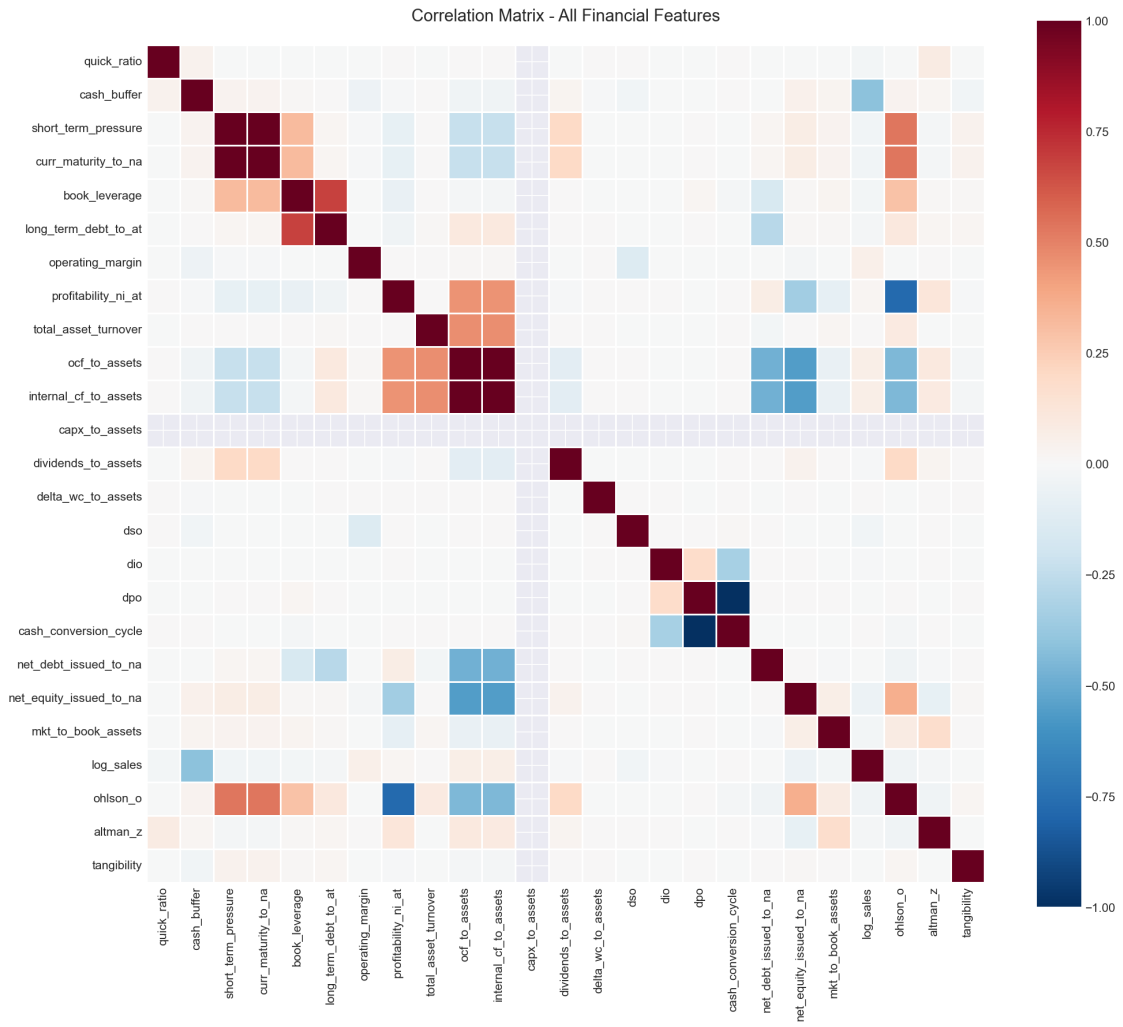

Figure 1: Correlation Matrix - All Financial Features. Pearson correlations computed on full sample (1996–2022). Red indicates positive correlation, blue negative. Features cluster into expected groups: leverage (top-left), profitability (center), working capital (middle), and valuation/distress (bottom-right).

Despite some high correlations, most features provide distinct information. For instance, while operating margin and ROA both measure profitability, their correlation is only moderate ($r \approx 0.5$), suggesting they capture different aspects of operational performance. This balance—features are related but not redundant—is ideal for comprehensive predictive modeling.

3.2 Regime-Dependent Correlations

Figure 2 shows the change in correlations between recession and expansion periods ($\Delta r = r_{\text{recession}} - r_{\text{expansion}}$). Red cells indicate relationships that become more positively correlated

during recessions; blue cells indicate relationships that weaken or reverse.

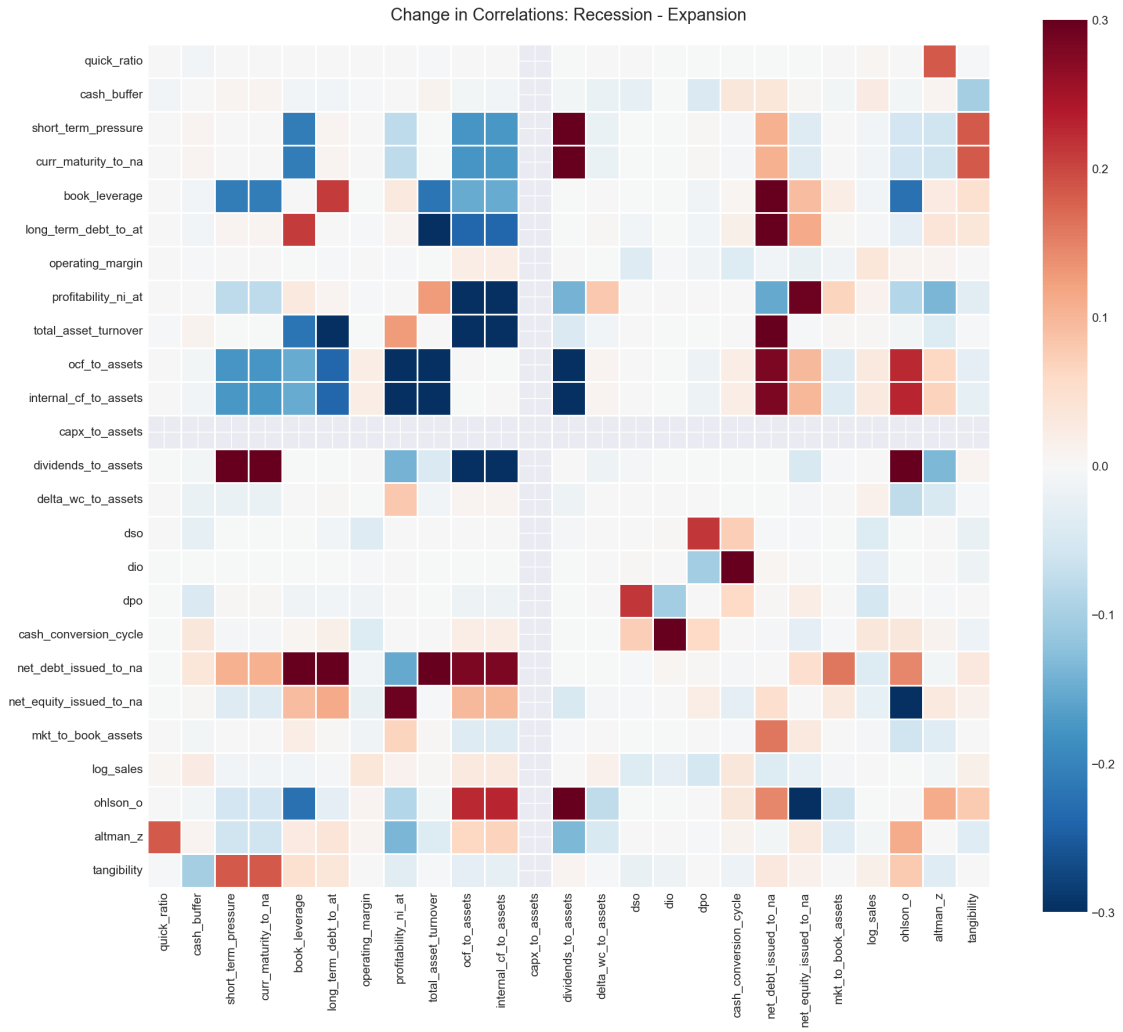

Figure 2: Change in Correlations: Recession - Expansion. Shows Δr for each feature pair. Large positive changes (red) indicate relationships that strengthen during economic stress; large negative changes (blue) indicate relationships that weaken or reverse. Most changes are moderate ($|\Delta r| < 0.2$), but several key pairs exhibit substantial regime-dependence.

The most striking regime-dependent changes are:

- **Asset Turnover ↔ Net Debt Issued**: $\Delta r = +0.87$
  During recessions, firms that issue debt exhibit dramatically lower operational efficiency. This likely reflects distressed firms tapping credit markets to survive operational shortfalls.

- **Asset Turnover ↔ OCF**: $\Delta r = -0.78$
  The normal positive relationship between efficiency and cash generation *inverts* during

crises. Efficient operations no longer guarantee cash flow when demand collapses or working capital requirements spike.

- **Short-term Pressure ↔ Dividends**: $\Delta r = +0.74$
  Firms under debt pressure synchronize dividend policies during recessions, either cutting aggressively or maintaining to signal stability. This coordination does not occur during normal times.

- **Leverage ↔ Debt Issuance**: $\Delta r = +0.50$
  High-leverage firms are more likely to issue additional debt during recessions (refinancing or distressed borrowing), whereas in expansions, debt issuance is uncorrelated with existing leverage levels.

These large regime shifts validate the importance of time-varying models. A predictive model trained on full-sample correlations will systematically mischaracterize how relationships change during the stress periods when prediction accuracy matters most.

3.3 Temporal Evolution of Financial Characteristics

Five representative features spanning different categories (Quick Ratio, ROA, Book Leverage, Asset Turnover, Altman Z-Score) are examined through faceted histograms showing distributions from 1996 to 2022. Recession years (2001, 2008, 2009, 2020) are highlighted. Detailed histograms are presented in Appendix A; key temporal patterns are summarized below.

Key Temporal Patterns:

1. **Cyclical Sensitivity:** Profitability (ROA) and distress (Altman Z) distributions exhibit clear cyclical patterns aligned with NBER recessions. During crisis years, distributions shift toward worse outcomes (left for ROA and Z-score), then revert during recoveries. This validates the economic relevance of these features.

2. **Secular Trends:** Asset turnover shows a steady long-term decline from 1996 to 2022. This structural shift reflects the economy's movement toward more capital-intensive industries (technology with intangible assets, capital-heavy manufacturing) and away from asset-light service businesses.

3. **Crisis Heterogeneity:** The 2020 COVID recession exhibits a distinct V-shaped pattern (sharp shock, quick recovery) compared to the prolonged 2008–09 adjustment. This suggests that crisis-specific features (e.g., policy response, shock origin) matter for prediction models.

4. **Stability vs. Volatility:** Some features (Quick Ratio, Leverage) show remarkable distributional stability over time, while others (ROA, Z-Score) are highly cyclical. This heterogeneity suggests that ensemble models combining stable and dynamic features may outperform single-feature approaches.

3.4 Principal Component Analysis

PCA on the 2021–2022 data reveals that **14 of 24 components** (58%) are needed to capture 90% of the variance. Figure 3 shows the scree plot with individual and cumulative explained variance ratios.

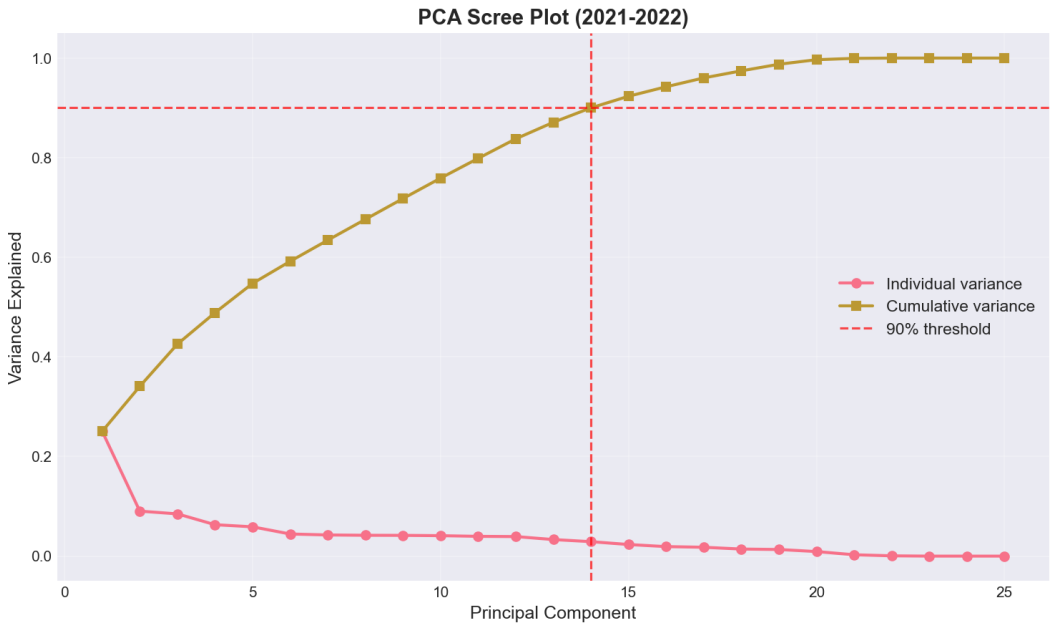

Figure 3: PCA Scree Plot (2021–2022). Pink line shows individual explained variance ratio per component; yellow line shows cumulative variance. Red dashed lines mark the 90% threshold, reached at PC14. The first three components capture 42.6% of variance and exhibit clear economic interpretations.

The first principal component (PC1) alone captures 25% of total variance—a remarkably high proportion indicating that financial distress vs. health is the dominant axis of variation across firms. The next two components each explain ∼9% and ∼8% of variance, representing business model differences and growth/liquidity policies, respectively.

Interpretation of Principal Components:

PC1 - Financial Distress Factor (25.05% variance):

- **High positive loadings:** Short-term debt pressure, book leverage, Ohlson O-score, debt issuance

- **High negative loadings:** Operating cash flow, internal cash flow
- **Economic interpretation:** This component separates financially stressed firms from healthy ones. High PC1 scores indicate financial pressure from high leverage and weak cash generation. This aligns with financial theory: the most important dimension of corporate financial variation is the balance between debt obligations and cash-generating ability.

PC2 - Business Model Intensity Factor (9.03% variance):

- **High positive loadings:** DSO, DPO (working capital timing)
- **High negative loadings:** Asset turnover, long-term debt, log(sales), tangibility
- **Economic interpretation:** This component captures capital intensity and scale. High PC2 scores indicate slow-turning, capital-light operations with extended working capital cycles (e.g., consulting, services). Low scores indicate fast-turning, capital-intensive operations (e.g., manufacturing, retail). This is a structural characteristic, not a health signal.

PC3 - Growth & Liquidity Factor (8.48% variance):

- **High positive loadings:** Cash buffer, market-to-book, equity issuance
- **High negative loadings:** Firm size (log sales), working capital metrics
- **Economic interpretation:** This component separates young, high-growth firms sitting on cash (post-IPO startups, growth companies) from mature, established firms with optimized working capital. High PC3 scores suggest a growth stage with liquidity cushions.

Validation of Feature Quality:
The PCA results provide strong validation of the feature engineering approach:

1. **Interpretability:** All three major components have clear, economically meaningful interpretations that align with financial theory. This is not guaranteed—PCA often produces statistically valid but economically opaque factors.
2. **Dimensionality:** Needing 14/24 components for 90% variance (58% retention) indicates the feature set successfully balances comprehensiveness and efficiency. If retention were >90%, features would be nearly redundant; if <40%, features would be measuring unrelated constructs.

3. **Alignment with theory:** The three major dimensions—financial health, operations, and growth—are precisely the dimensions emphasized in corporate finance theory. This suggests the features capture fundamental economic constructs rather than spurious statistical patterns.

4 Discussion

4.1 Implications for Predictive Modeling

The results have several important implications for building corporate finance prediction models:

1. Feature Selection Strategy:

The high correlations between some feature pairs (e.g., OCF/Assets and Internal CF/Assets at $r = 0.999$) might suggest one should be dropped to avoid multicollinearity. However, this concern is overstated for modern machine learning algorithms. Tree-based methods (Random Forest, XGBoost) are naturally robust to correlated features, and regularized linear models (Lasso, Ridge) automatically handle multicollinearity through coefficient shrinkage.

More importantly, the PCA results show that even highly correlated features contribute to distinct principal components. Internal CF/Assets and OCF/Assets differ only by dividends, but this difference may be critical: dividend policy reveals management's confidence and capital allocation priorities, information not captured by raw cash flow. The recommendation is to *retain all engineered features* and allow the model to determine their relative importance through feature selection or regularization.

2. Time-Varying vs. Static Models:

The large regime-dependent correlation changes (up to ± 0.87) provide strong evidence against static prediction models. Consider a model trained on full-sample data: it would learn that asset turnover and operating cash flow are positively correlated ($r \approx 0.5$). But during recessions—precisely when accurate prediction is most critical—this relationship weakens or reverses ($\Delta r = -0.78$). The model would systematically mischaracterize relationships for operationally efficient firms experiencing cash flow stress during downturns.

This finding suggests several modeling approaches:

- **Regime-switching models:** Estimate separate models for recession and expansion periods, then weight predictions based on current economic conditions.
- **Interaction terms:** Include interaction terms between features and recession indicators (e.g., Leverage $\times$ Recession) to allow coefficients to vary by regime.

- **Time-varying coefficients:** Use rolling-window estimation or time-series models that allow coefficients to evolve gradually.

3. Ensemble Approaches:

The PCA results suggest an ensemble strategy: build separate sub-models focused on each principal component, then combine predictions. For instance:

- **Distress model (PC1):** Focus on leverage, cash flow, and distress scores. This model identifies firms in acute financial trouble.
- **Business model model (PC2):** Focus on asset turnover, working capital, and tangibility. This model adjusts risk assessment based on industry/business model.
- **Growth model (PC3):** Focus on cash holdings, market-to-book, and size. This model identifies growth firms with different risk profiles.

The final prediction could be a weighted average, where weights are determined through cross-validation or based on the explanatory power of each component (25%, 9%, 8%).

4.2 Limitations and Caveats

Several limitations should be acknowledged:

1. Survivorship Bias:

COMPUSTAT contains only firms with sufficient data coverage, which may introduce survivorship bias. Very small or financially distressed firms may not have complete financial statements in COMPUSTAT. This could bias relationships toward healthier, larger firms. Future work should incorporate data from sources that track delisted firms explicitly.

2. Sample Period:

While 1996–2022 covers multiple economic cycles, it is still a limited historical period. Structural changes in the economy (e.g., the rise of intangibles, changing accounting standards) mean that relationships observed in this sample may not hold indefinitely. Models should be regularly retrained as new data becomes available.

3. Continuous Outcomes:

While this analysis focuses on feature validation, many corporate finance applications involve continuous outcomes rather than binary classification. Firms exist on a spectrum of financial health rather than discrete categories. Future work could explore how these features perform with continuous distress measures or multi-class outcomes (healthy, stressed, deteriorating).

4. U.S.-Centric Sample:

The COMPUSTAT data covers only U.S. publicly traded firms, which limits generalizability to international markets or private companies. Bankruptcy laws, accounting standards, and economic conditions vary substantially across countries, so these features would need to be validated on international samples before deployment abroad.

5 Conclusion

This study develops and validates a comprehensive set of 24 financial features using COMPUSTAT data from 1996 to 2022. Through systematic correlation analysis, temporal distributional assessment, and principal component analysis, these features are shown to capture distinct but economically related signals suitable for corporate finance prediction tasks.

Key findings include:

1. Financial relationships exhibit substantial regime-dependence, with correlation changes up to ± 0.87 between recession and expansion periods. This validates the importance of time-varying risk models.

2. Temporal distributions show clear cyclical patterns aligned with NBER recessions, confirming the economic relevance of selected features. Different crises (2001, 2008–09, 2020) exhibit distinct distributional signatures.

3. PCA reveals three interpretable latent factors—financial distress (25% of variance), business model intensity (9%), and growth/liquidity (8%)—that align precisely with financial theory. These factors provide a parsimonious representation of the high-dimensional feature space.

4. The feature set achieves an efficient balance: 14 of 24 components (58%) are needed for 90% variance, indicating moderate redundancy without excessive overlap.

Practical recommendations for predictive modeling:

- **Retain all features:** Modern ML algorithms handle multicollinearity naturally; dropping correlated features may discard valuable information.

- **Use regime-specific models:** Estimate separate models for recession vs. expansion periods or include interaction terms with economic indicators.

- **Consider ensemble approaches:** Build sub-models focused on each principal component (distress, operations, growth) and combine predictions.

- **Monitor temporal stability:** Regularly retrain models as distributions evolve and new economic regimes emerge.

Future research directions:

1. **Predictive modeling:** Build and evaluate corporate finance prediction models using these features, comparing tree-based methods (Random Forest, XGBoost), neural networks, and traditional statistical approaches.

2. **Alternative outcome measures:** Explore continuous financial health indicators (e.g., credit ratings, credit spreads) or multi-class outcomes beyond binary classification.

3. **Forward-looking signals:** Incorporate market-based indicators (implied volatility, credit spreads) and textual data (earnings call sentiment) to complement backward-looking financial ratios.

4. **International validation:** Test feature stability across countries with different accounting standards and regulatory environments.

This systematic feature engineering framework provides a robust foundation for developing accurate, interpretable, and economically grounded prediction models suitable for deployment in corporate finance applications, credit assessment, and investment decision-making.

A Temporal Distribution Figures

This appendix presents detailed faceted histograms for five representative financial features, showing their distributions across all years from 1996 to 2022. Recession years (2001, 2008, 2009, 2020) are highlighted in pink. Red dashed lines indicate the median for each year. Values are winsorized at the 1st and 99th percentiles for improved visualization.

Figure 4: Altman Z-Score Distribution by Year (1996–2022). Consistently right-skewed with concentration near zero (distressed zone) and long right tail of healthy firms. Visible leftward shifts during recessions indicate more firms entering distress. Post-2008 distributions show increased dispersion.

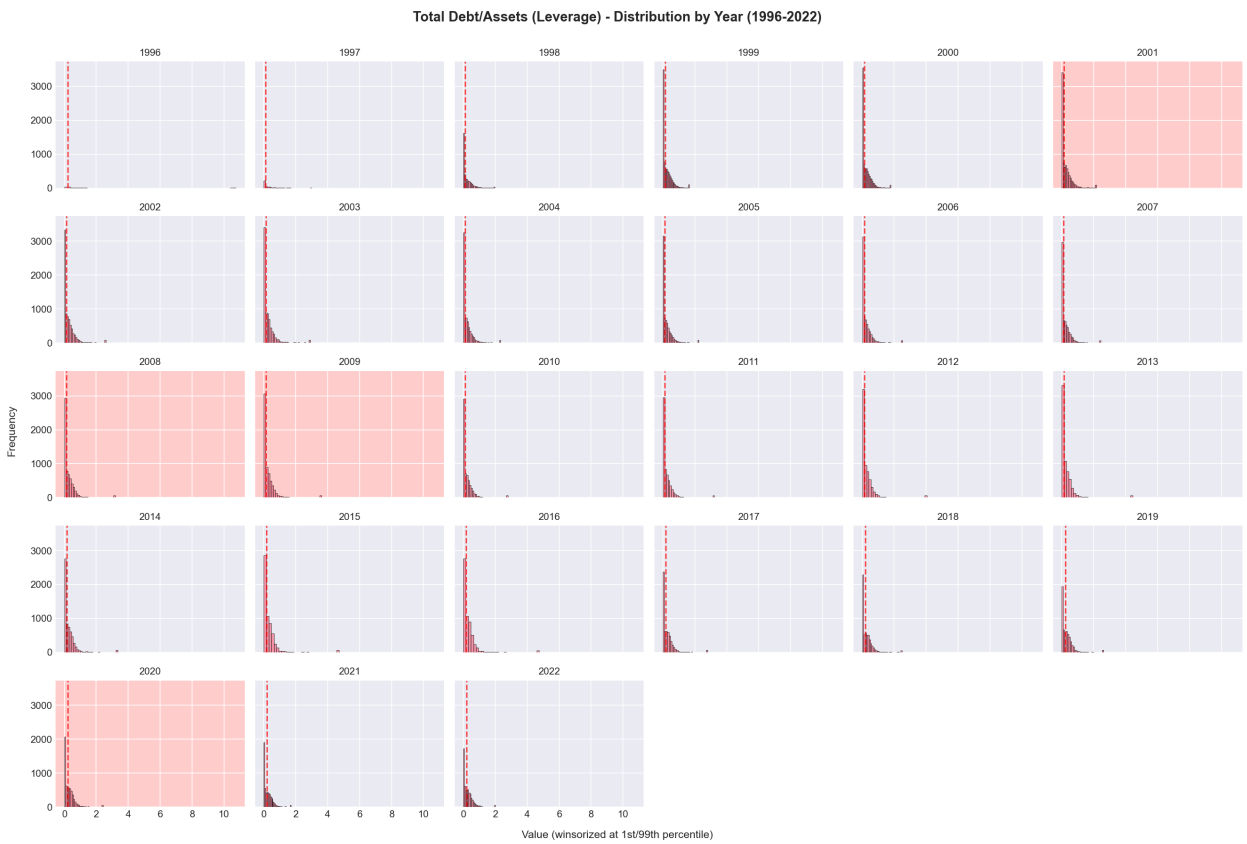

Figure 5: Book Leverage Distribution by Year (1996–2022). Highly right-skewed with strong mode near zero. Distribution remarkably stable over 27 years, suggesting fundamental capital structure decisions change slowly. Slight median increase during 2008–09 reflects either increased borrowing or equity value compression.

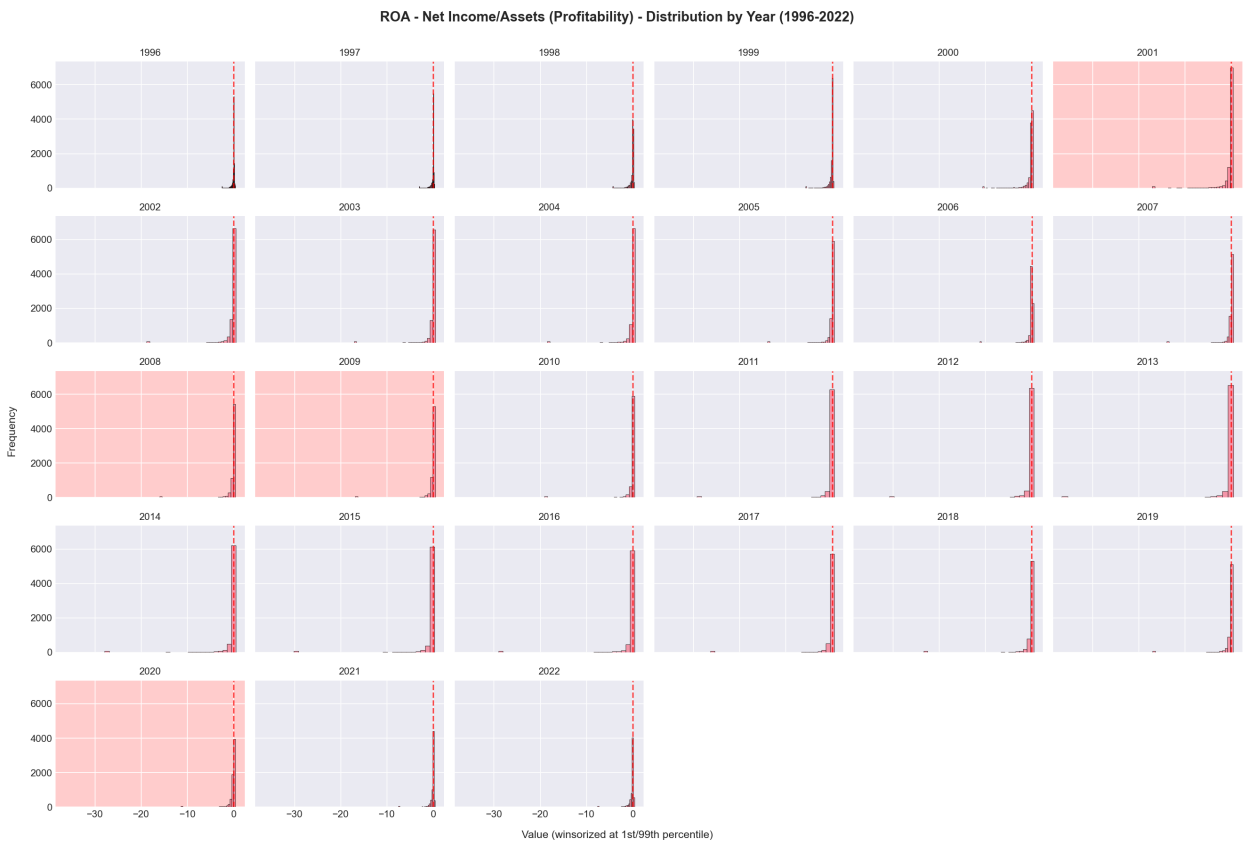

Figure 6: ROA Distribution by Year (1996–2022). Concentrated around zero with symmetric tails. Clear cyclicality: distributions shift left during recessions (2001, 2008–09, 2020) as more firms report losses. Post-2010 compression suggests reduced profitability dispersion across firms.

Figure 7: Quick Ratio Distribution by Year (1996–2022). Heavy right skew with most firms maintaining modest liquidity buffers. Distribution shape remarkably consistent over time. Subtle rightward shift in 2020–2021 suggests COVID-era cash accumulation.

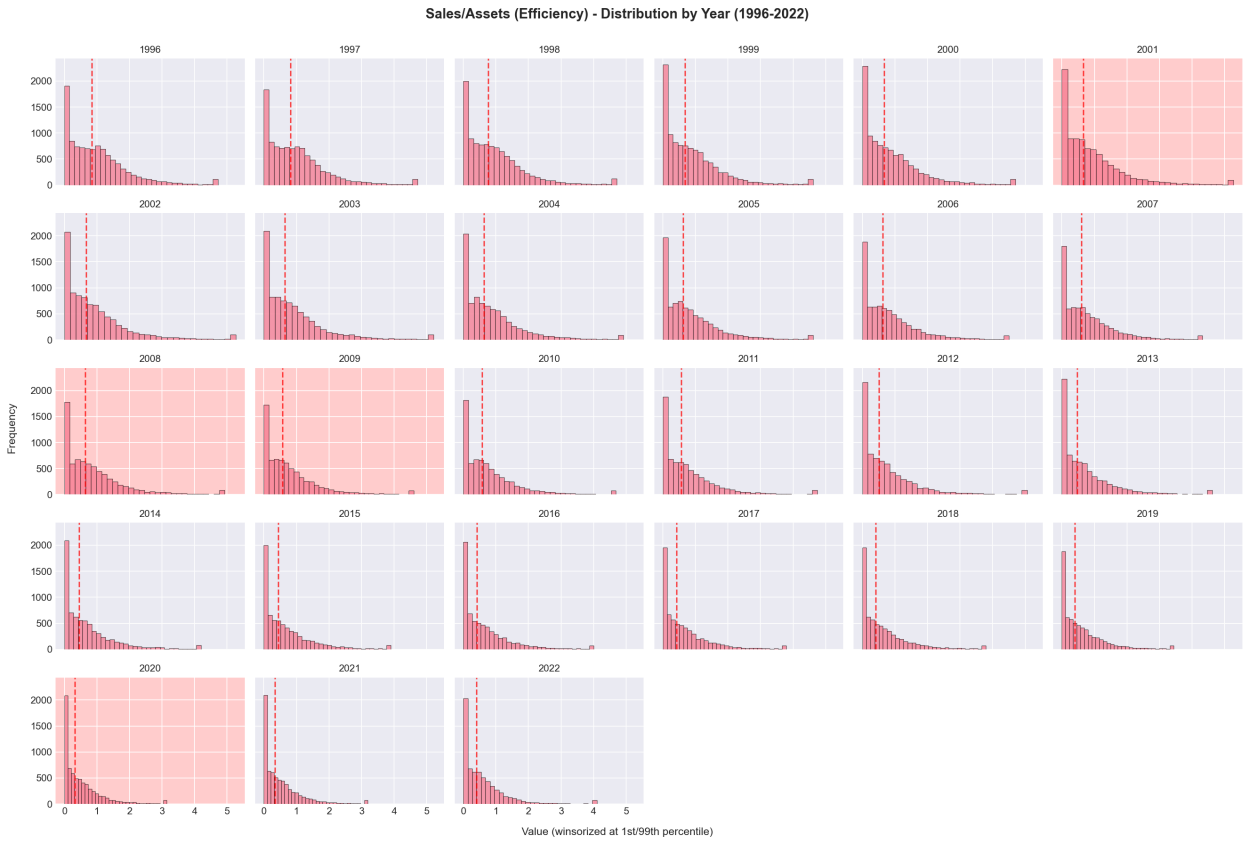

Figure 8: Asset Turnover Distribution by Year (1996–2022). Right-skewed with clear mode declining from ∼1.0 (late 1990s) to ∼0.7 (2020s). Secular decline indicates structural shift toward more asset-intensive business models. Recession periods show increased dispersion.

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609.

[2] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.

Note: Interactive analysis and detailed methodology available in accompanying Jupyter notebook.

For questions or collaboration opportunities, please contact via portfolio website or LinkedIn.