Katarina Delos Santos

PREDICTING FIRST-DAY MARKET CAPITALIZATION OF TECH STARTUP IPOs:

A DATA-DRIVEN APPROACH TO VALUING INNOVATION

In the dynamic landscape of technology startups, initial public offerings (IPOs) represent a critical juncture where market perception meets company potential. This study leverages comprehensive financial and market data to predict the first day market capitalization of tech startups at their IPO.

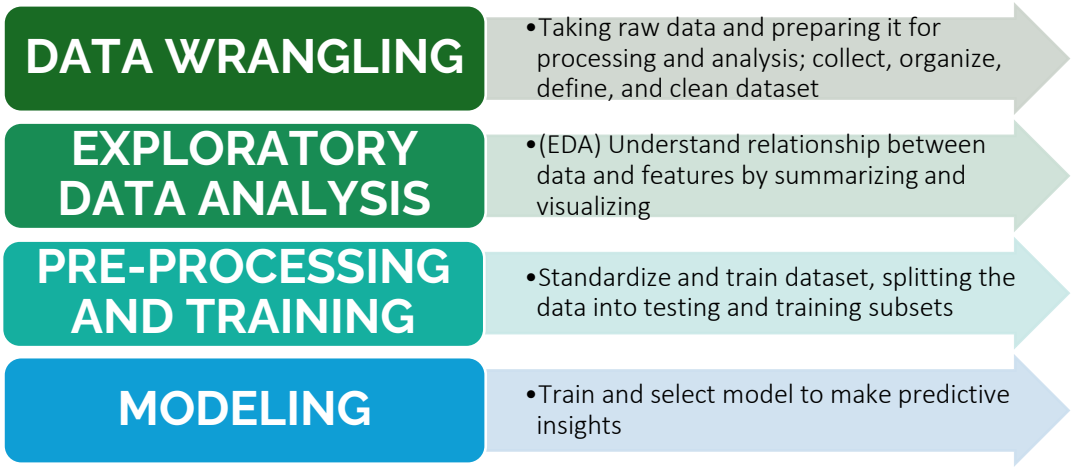
Conclusion and Insights

I. Tech startup IPOs from the last decade provide a rich dataset for analyzing market valuation trends in the technology sector.
II. Efficient data processing and storage using Parquet files enable handling of large-scale financial data for comprehensive analysis.
III. Advanced machine learning models, including Random Forest, XGBoost, LightGBM, and Ensemble Stacking demonstrate strong predictive capabilities for first-day market capitalization.
IV. The best-performing model, selected through rigorous evaluation, offers valuable insights for investors, analysts, and entrepreneurs in assessing potential market reception of tech IPOs.
V. This predictive approach provides a data-driven framework for understanding the factors that influence initial market valuation of tech startups, potentially informing investment strategies and startup valuation methodologies.

PRACTICAL IMPLEMENTATION

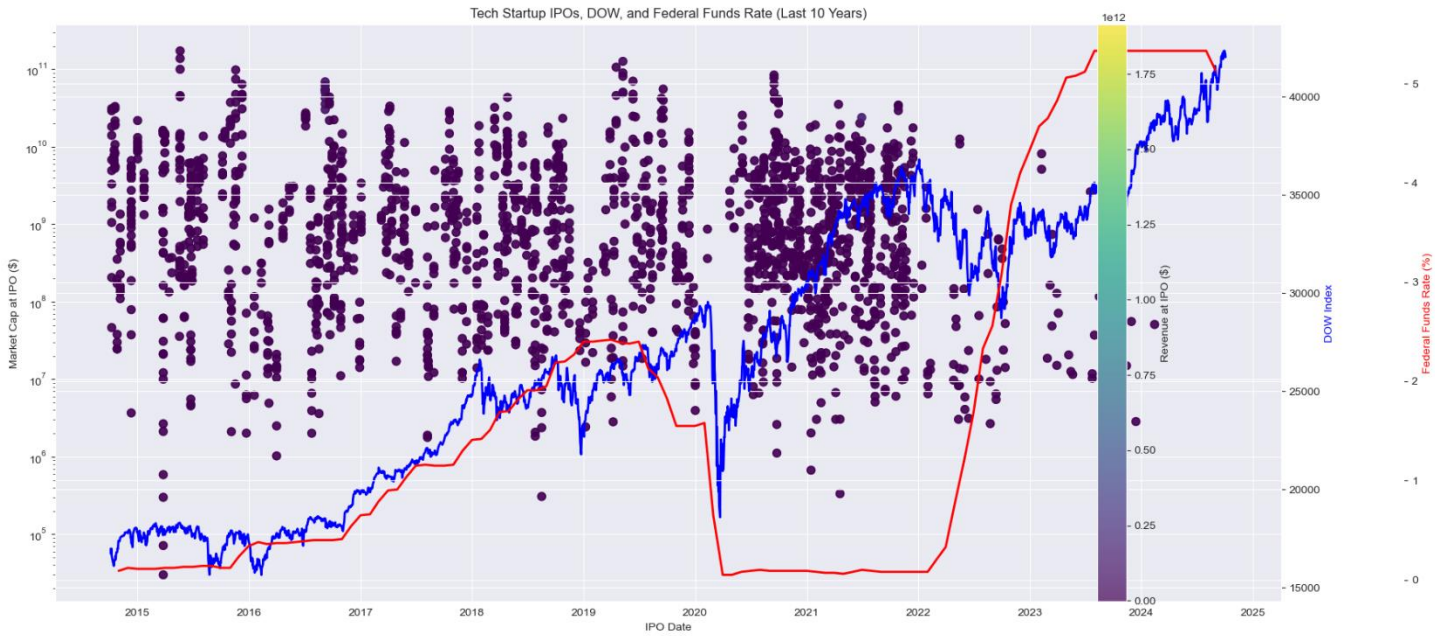
How to Use Prediction Results

i. Portfolio managers can leverage the model's insights to identify high-potential tech startups before their public debut, maximizing investment returns.
ii. The model serves as a predictive tool for startup valuations, helping founders and investors set realistic expectations and negotiate fair terms during pre-IPO stages.



<Figure 1: Data Science Method (DSM)>

I. Introduction



<Figure 2: Tech Startup IPOs, DOW, and Federal Funds Rate (Last 10 Years)>

1) Problem Statement

The technology sector has witnessed unprecedented growth in Initial Public Offerings (IPOs), making accurate market capitalization predictions crucial for investors, underwriters, and companies going public, but Tech IPO valuations face unique challenges of high growth rates, complex business models, intangible assets, market volatility, and network effects. This project presents a machine learning approach to predict first-day market capitalizations of tech IPOs using comprehensive financial metrics. The valuation of tech IPOs has evolved significantly from traditional financial metrics to modern machine learning approaches. Key developments include:

- Market-based valuation methods
- Comparative analysis techniques

- AI-driven prediction models
- Financial ratio analysis

Keywords: Tech IPO valuation, Machine Learning, Random Forest Regression, Financial Prediction, Market Capitalization

2) Goal

This project's primary goal is to develop an accurate and reliable model for predicting the first-day market capitalization of tech startup IPOs using comprehensive financial and market data. In order to achieve this, the following specific objectives have been identified:

Objective	
Data Extraction and Preprocessing	Efficiently extract raw data from Nasdaq and convert it to Parquet format for optimized storage and processing
Feature Selection and Engineering	Identify and prepare relevant features from the SHARADAR datasets, focusing on tech companies that had their IPO in the last 10 years
Model Development and Evaluation	Implement and compare advanced machine learning models, including Linear Regression, Random Forest algorithms, Gradient Boosting algorithms (XGBoost and LightGBM), and ensemble stacking to predict first-day market capitalization
Performance Assessment	Evaluate model performance using appropriate metrics to select the best-performing model for predicting IPO market capitalization
Model Preservation	Save the best-performing model for future use and deployment

<Table 1: Project Objectives>

By accomplishing these goals, the project aims to provide a data-driven tool for predicting the initial market valuation of tech startups at IPO. This tool will offer valuable insights for investors, analysts, and entrepreneurs, potentially informing investment strategies and startup valuation methodologies in the dynamic tech sector.

II. Dataset Source and Financial Ratios

This analysis utilizes data from Nasdaq premium datasets, specifically the SH1 Core US Fundamentals Data database[1]. The comprehensive database provides a rich source of financial information for publicly traded companies in the United States, offering a solid foundation for our study on tech startup IPOs: "updated daily, this database provides up to 24 years of history, for 150 essential fundamental indicators and financial ratios, for more than 16,000 US public companies. Includes detailed corporate actions data. This is reference grade fundamental stock data".

SH1 encompasses a wide range of financial metrics for companies that have gone public in the technology sector over the past decade. To enhance the analysis, financial ratios that provide deeper

[1] https://data.nasdaq.com/databases/SF1

insights into the companies' financial health and potential market performance were also created. The dataset's target variable is first day market cap.

1) Data Cleaning Procedures

Cleaning Procedure	
1. Data Filtering	The code calculates a date 10 years ago from the current date using datetime operations. This allows for filtering companies based on their IPO date, ensuring only recent tech startups are included in the analysis.
2. Sector and IPO Date Filtering	The tickers_df is filtered to include only companies in the Technology sector that had their IPO within the last 10 years. This step narrows down the dataset to focus on relevant companies.
3. Dimension Filtering	When retrieving financial data from the SF1 dataset, the code filters for the 'MRY' (Most Recent Yearly) dimension. This ensures that only the most up-to-date yearly financial data is used for each company.
4. Data Merging	The filtered tech startups data is merged with the latest financial data. This step combines company information with their financial metrics, creating a more comprehensive dataset.
5. Column Selection	The code defines a list of desired columns (ticker, name, sector, industry, firstpricedate, revenue, netinc, assets, marketcap) and then checks which of these columns are available in the merged dataset. This approach ensures that only existing columns are selected, preventing errors due to missing columns.
6. Final Dataset Creation	The final dataset is created by selecting only the columns that exist in both the desired columns list and the available columns in the merged dataset. This step further refines the data by including only the most relevant information.

<Table 2: Data Cleaning Procedures>

Four CSV datasets: SHARADAR_SF1.csv (contains financial statement data), SHARADAR_TICKERS.csv (includes company information and metadata), SHARADAR_SP500.csv (S&P 500 index data), and SHARADAR_INDICATORS.csv. From the SF1 Dataset, the latest financial data for technology companies were retried then merged to combine company information with financial data. Time frame selection was implemented to calculate for the past decade. This filter was used for companies with IPOs in the last ten years. From here, by sector filtering, only companies from the Technology sector were chosen.

A function to cap outliers for various financial columns was utilized. This function used the 1st and 99th percentiles as lower and upper bounds respectively to clip extreme values. StandardScaler was used to normalize all numeric features to make sure they are on the same scale. For numeric columns, missing values were imputed with the median of the respective column. For categorical columns, missing values were imputed with the most frequent value (mode) of the respective column. And the 'date' column was converted to datetime format, with any parsing errors resulting in NaT (Not a Time) values.

After cleaning, the dataset contained a mix of float64 and object (string) columns. After the cleaning process, there were no null values in the dataset. In addition to this, the process was done in chunks to handle the large dataset efficiently and reduce memory storage. These cleaning steps result in a focused, relevant, and up-to-date dataset of tech startup IPOs from the last 10 years, along with their key

financial metrics. The process effectively removed irrelevant data, handles potential missing columns, and ensures data consistency for further analysis (free of outliers and missing values, normalized and scaled, enriched with engineered features and interaction terms, optimized for machine learning model training).

2) Financial Ratios

• Profitability Ratios: Assess a company's ability to generate profits relative to its revenue, assets, or equity. For tech startups, these ratios can indicate the efficiency of their business models and their potential for future growth.
- roa (Return on Assets), roe (Return on Equity), roic (Return on Invested Capital), ros (Return on Sales), gross_margin, net_profit_margin, operating_margin, earnings_yield

• Liquidity Ratios: Measure a company's ability to meet short-term obligations. For tech startups going public, strong liquidity can be a positive signal to potential investors.
- current_ratio, quick_ratio, cash_to_assets

• Solvency Ratios: Evaluate a company's long-term financial stability and its ability to meet debt obligations. These are particularly important for assessing the sustainability of tech startups' business models.
- debt_to_equity, interest_coverage, price_to_tangible_book

• Growth Ratios: Measure the rate at which a company is expanding its revenue, user base, or market share. For tech startups, high growth rates are often key drivers of market valuation.
- revenue, netinc (Net Income), ebitda (Earnings Before Interest, Taxes, Depreciation, and Amortization)

• Efficiency Ratios: Assess how effectively a company uses its assets and manages its liabilities. For tech startups, these can provide insights into operational efficiency and scalability.
- asset_turnover, inventory_turnover, asset_utilization, cash_conversion_cycle, operating_expense_ratio, fcf_to_revenue, capex_to_revenue, fcf_to_assets, gross_profit_to_assets

• Valuation Metrics: Evaluate a company's worth and market value relative to various financial metrics. For tech startups, these metrics help investors assess whether a company's stock is fairly priced and compare valuations across similar companies.
- pe (Price to Earnings) measures the market value per share relative to earnings per share, indicating how much investors are willing to pay for each dollar of earnings; pb (Price to Book) compares a company's market value to its book value, particularly useful for asset-heavy companies; ps (Price to Sales) relates market capitalization to revenue, especially valuable for high-growth companies not yet profitable; ev (Enterprise Value) represents the total value of a company, including debt and excluding cash; evebit (Enterprise Value to EBIT) shows how many years of operating earnings it would take to pay for the business; evebitda (Enterprise Value to EBITDA) indicates company value relative to operating performance before accounting for capital structure; fcfps (Free Cash Flow Per Share) measures available cash flow per share for investors; dividend_yield_to_price shows the dividend return relative to share price, payout_ratio indicates the proportion of earnings paid out as dividends

Ratios	General	Tech Startups	Examples
Profitability	Assess a company's ability to generate profits relative to its revenue, assets, or equity	Can indicate the efficiency of their business models and their potential for future growth	roa (Return on Assets), roe (Return on Equity), roic (Return on Invested Capital), ros (Return on Sales), gross_margin, net_profit_margin, operating_margin, earnings_yield
Liquidity	Measure a company's ability to meet short-term obligations	For tech startups going public, strong liquidity can be a positive signal to potential investors	current_ratio, quick_ratio, cash_to_assets
Solvency	Evaluate a company's long-term financial stability and its ability to meet debt obligations	Particularly important for assessing the sustainability of tech startups' business models	debt_to_equity, interest_coverage, price_to_tangible_book
Growth	Measure the rate at which a company is expanding its revenue, user base, or market share	High growth rates often key drivers of market valuation	revenue, netinc (Net Income), ebitda (Earnings Before Interest, Taxes, Depreciation, and Amortization)
Efficiency	Assess how effectively a company uses its assets and manages its liabilities	Can provide insights into operational efficiency and scalability	asset_turnover, inventory_turnover, asset_utilization, cash_conversion_cycle, operating_expense_ratio, fcf_to_revenue, capex_to_revenue, fcf_to_assets, gross_profit_to_assets
Valuation [Metrics]	Evaluate a company's worth and market value relative to various financial metrics	Help investors assess whether a company's stock is fairly-priced and compare valuations across similar companies	pe (Price to Earnings), pb (Price to Book), ps (Price to Sales), ev (Enterprise Value), evebit (Enterprise Value to EBIT), evebitda (Enterprise Value to EBITDA), fcps (Free Cash Flow Per Share), dividend_yield_to_price, payout_ratio

<Table 3: Financial Ratios Summary>

By combining standard financial ratios with custom-developed metrics, the aim is to capture both the fundamental financial health of tech startups and the unique factors that drive their market valuation at IPO. This comprehensive approach allows for a more nuanced and accurate prediction of first-day market capitalization, considering both traditional financial metrics and factors specific to the tech startup ecosystem. The final dataset consists of 61 features.

III. Exploratory Data Analysis (EDA)

1) Univariate Analysis

1.1) Target Variable: The target variable, market cap on the first day, showed a significant range of values across the dataset. The distribution of this variable revealed important insights about the

companies in our sample. A right-skewed distribution is observed, indicating that while most companies had relatively lower market caps on their first day of trading, there were a few outliers with exceptionally high market caps. This skewness is typical in financial data and reflects the reality of the stock market where a small number of companies often dominate in terms of market capitalization.

1.2) Numerical Features: There are several key observations.

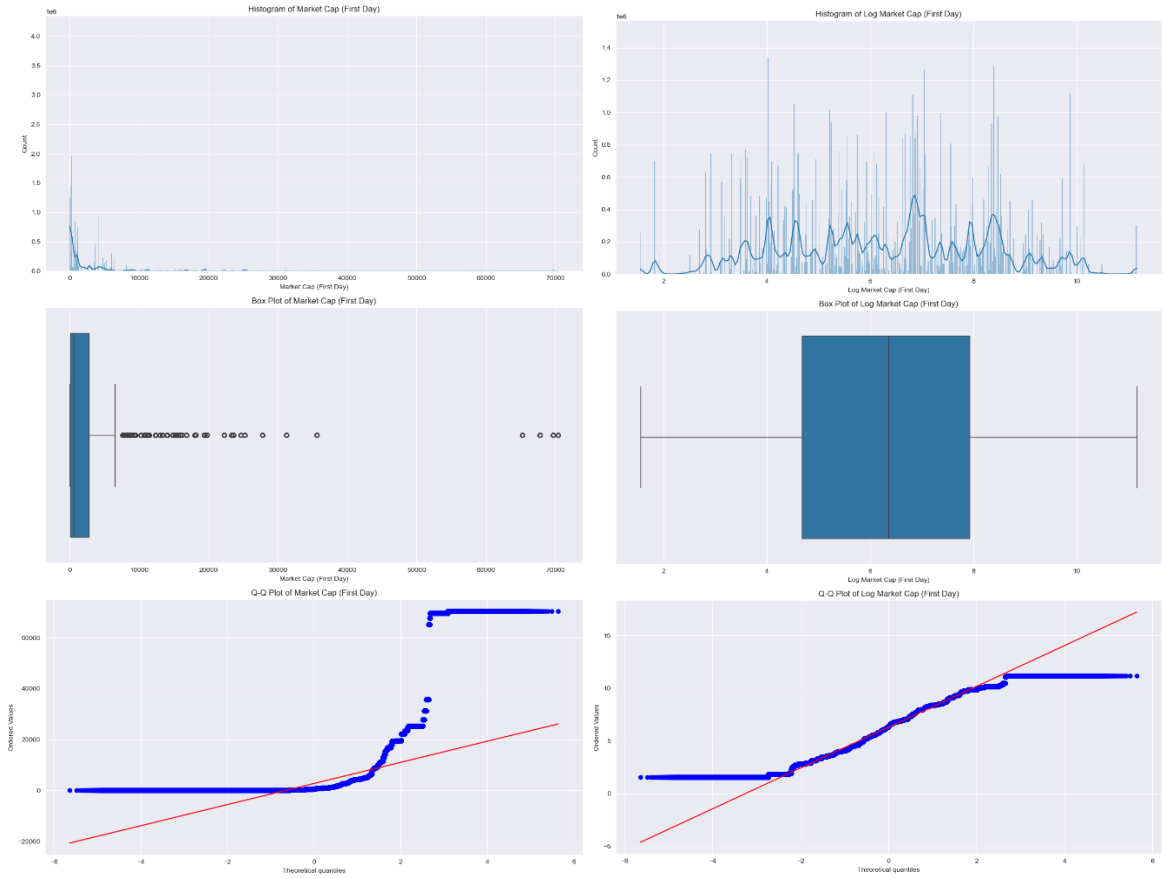
a) Distribution Characteristics: Most numerical variables exhibited non-normal distributions. Many showed right-skewed patterns, like our target variable. This asymmetry is common in financial and economic data, often reflecting natural limits on the lower end of scales (e.g., prices or volumes can't go below zero) but no theoretical upper limit.

b) Outliers: The presence of outliers in multiple features is identified. These extreme values, while potentially representing genuine data points, had a significant impact on the overall distribution of the variables. In some cases, removing these outliers resulted in more symmetrical distributions, allowing for clearer analysis of the central tendencies and typical ranges of our features.

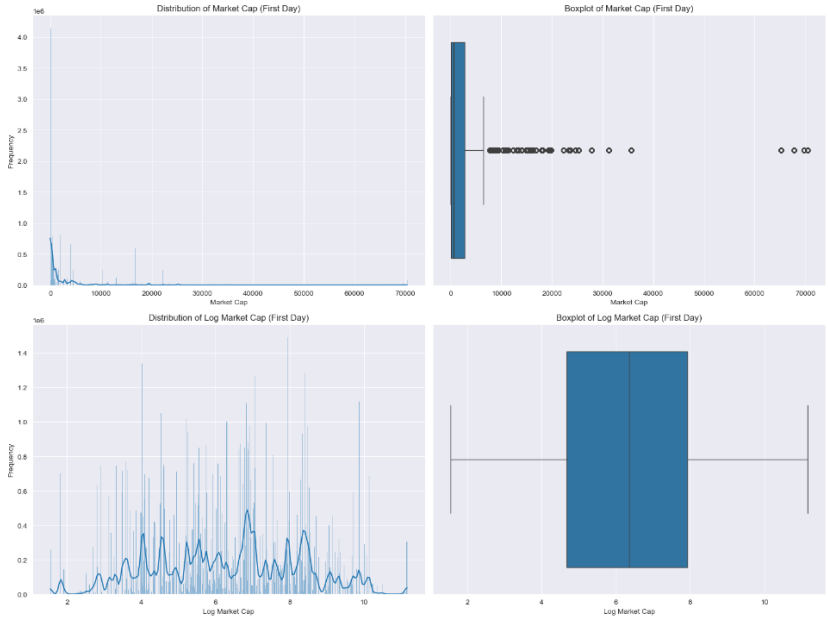
c) Transformation Attempts: To address the non-normal distributions, log transformations were experimented with. However, the effect of these transformations was minimal for most variables. This resistance to normalization through standard transformations suggests that the underlying data structures may have inherent characteristics that deviate from normal distributions.

1.3) Correlation with Target: Initial univariate analysis also provided insights into potential relationships between individual features and our target variable. Some variables showed stronger correlations with the first-day market cap, indicating their potential predictive power in our subsequent modeling efforts.

This univariate analysis laid the groundwork for further investigations, helping understand the individual characteristics of the variables and informing the approach to feature engineering and model selection in the later stages of our project. From the univariate analysis, the rest of the analysis uses the log transformation of the target variable ‘marketcap_first_day’.



<Figure 3: Before (left) and After (right) Distribution of Taking Logarithm of Target Variable 'marketcap_first_day'>



<Figure 4: Before log (above) and After log (below) Distribution and Boxplot of Target Variable>

	Original Target: 'marketcap_first_day'	Transformed Target: log 'marketcap_first_day'
Range	$3.7 million to $70.4 billion	1.55 to 11.16
Mean vs. Median	Mean ($2.71 billion), Median ($573.4 million); right-skewed	Mean (6.30), Median (6.35); more symmetrical
Standard Deviation	$6.36 billion	1.95; more normalized spread
Quartiles	75% of the IPOs have market cap below $2.76 billion with 25% below $106.6 million	

<Table 4: Target Variable Log Transformation Comparison>

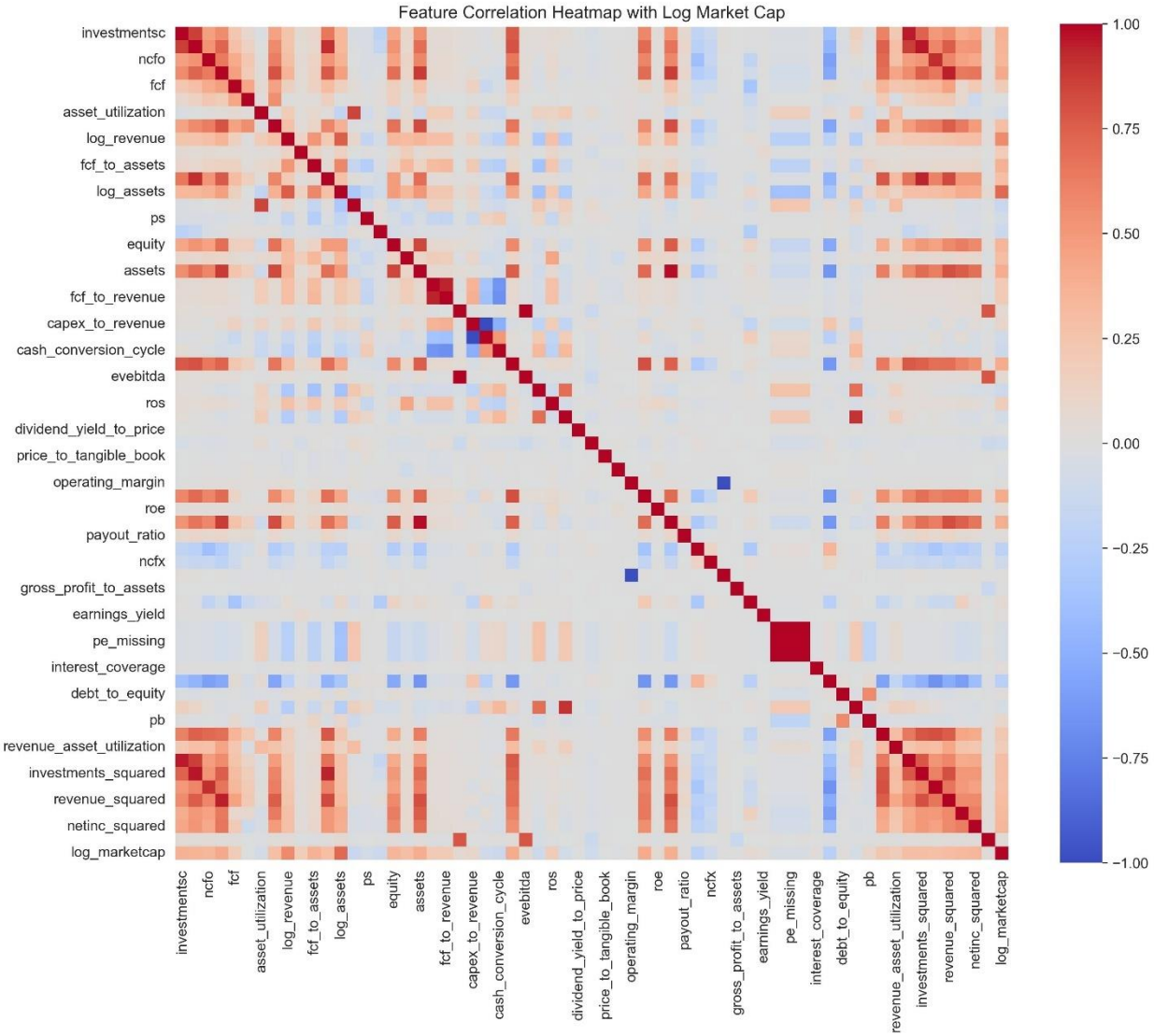
Transformation Key Observations:

1) Right-skewed Original Distribution: The original market cap data is heavily right-skewed, with a long tail towards higher values. This is typical for financial data, especially in the tech startup world where a few "unicorns" can have extremely high valuations.
2) Normalization Effect of Log Transformation: The log transformation has made the distribution more symmetrical and reduced the impact of extreme values. This is evident from the closer mean and median values in the log-transformed data.
3) Compression of Range: The log transformation has compressed the wide range of the original data (from 3.7 to 70,412.3) into a more manageable range (1.55 to 11.16), which is beneficial for statistical analyses and machine learning algorithms.
4) Outlier Handling: The log transformation has likely reduced the impact of outliers, which were prominent in the original data given the high standard deviation.

Key Observation	Original vs. log marketcap_first_day
Right-skewed Original Distribution	The original market cap data is heavily right-skewed, with a long tail towards higher values. This is typical for financial data, especially in the tech startup world where a few "unicorns" can have extremely high valuations.
Normalization Effect of Log Transformation	The log transformation has made the distribution more symmetrical and reduced the impact of extreme values. This is evident from the closer mean and median values in the log-transformed data.
Compression of Range	The log transformation has compressed the wide range of the original data (from 3.7 to 70,412.3) into a more manageable range (1.55 to 11.16), which is beneficial for statistical analyses and machine learning algorithms.
Outlier Handling	The log transformation has likely reduced the impact of outliers, which were prominent in the original data given the high standard deviation.

<Table 5: Target Variable Transformation Key Observations Summary>

The log transformation has effectively addressed the high skewness and wide range of the original market cap data, creating a more normally distributed target variable.



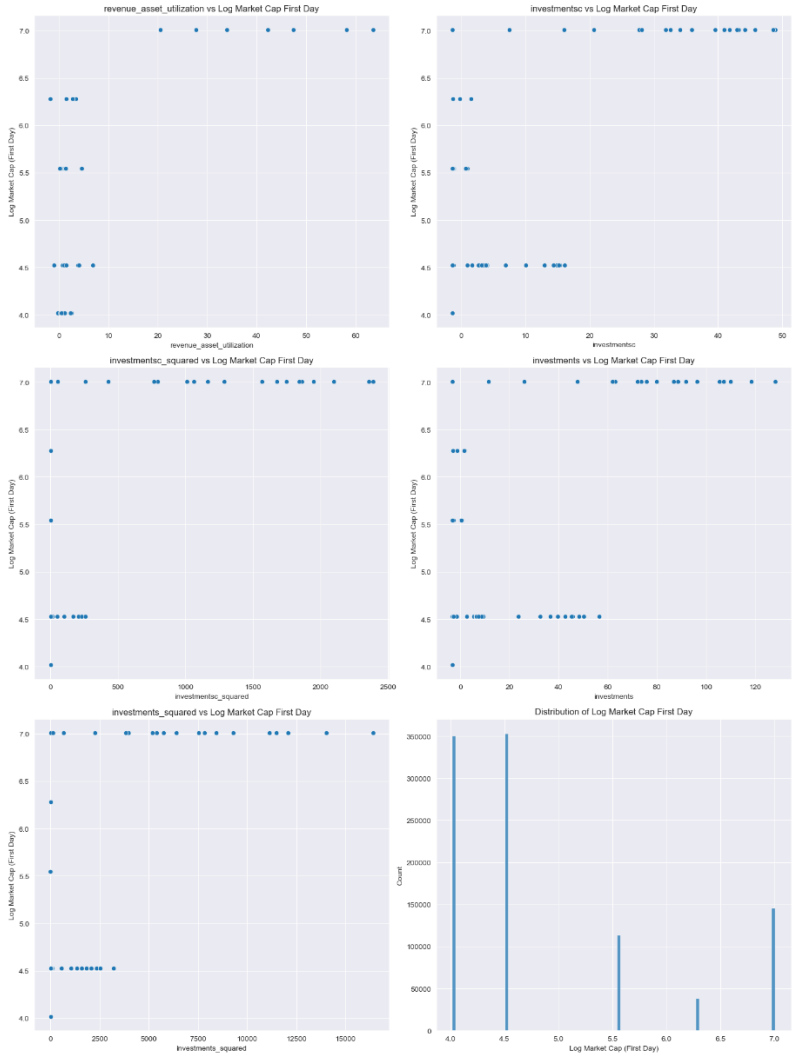
<Figure 5: Feature Correlation Heatmap with log marketcap_first_day>

The correlation analysis of market capitalization predictors reveals several strong and insightful relationships. The strongest positive correlations with log_marketcap are demonstrated by log_assets (0.712) and log_revenue (0.563), indicating that company size metrics are powerful indicators of market capitalization. Revenue and assets in their raw forms also show substantial positive correlations at 0.418 and 0.405 respectively.

On the contrary, the analysis identifies key negative correlations that provide valuable insights. Capital expenditure (capex) shows the strongest negative correlation at -0.241, followed by business-related net cash flow (ncfbus) at -0.219. Asset turnover and investment-related net cash flow (ncfinv) also display notable negative correlations of -0.145 and -0.131 respectively. The cash conversion cycle rounds out the top negative correlations at -0.120.

These correlation patterns highlight the complex interplay between financial metrics and market capitalization, where size-related metrics show strong positive associations while operational and cash flow metrics tend to show negative relationships. This understanding provides valuable direction for feature selection and model development in predicting market capitalization.

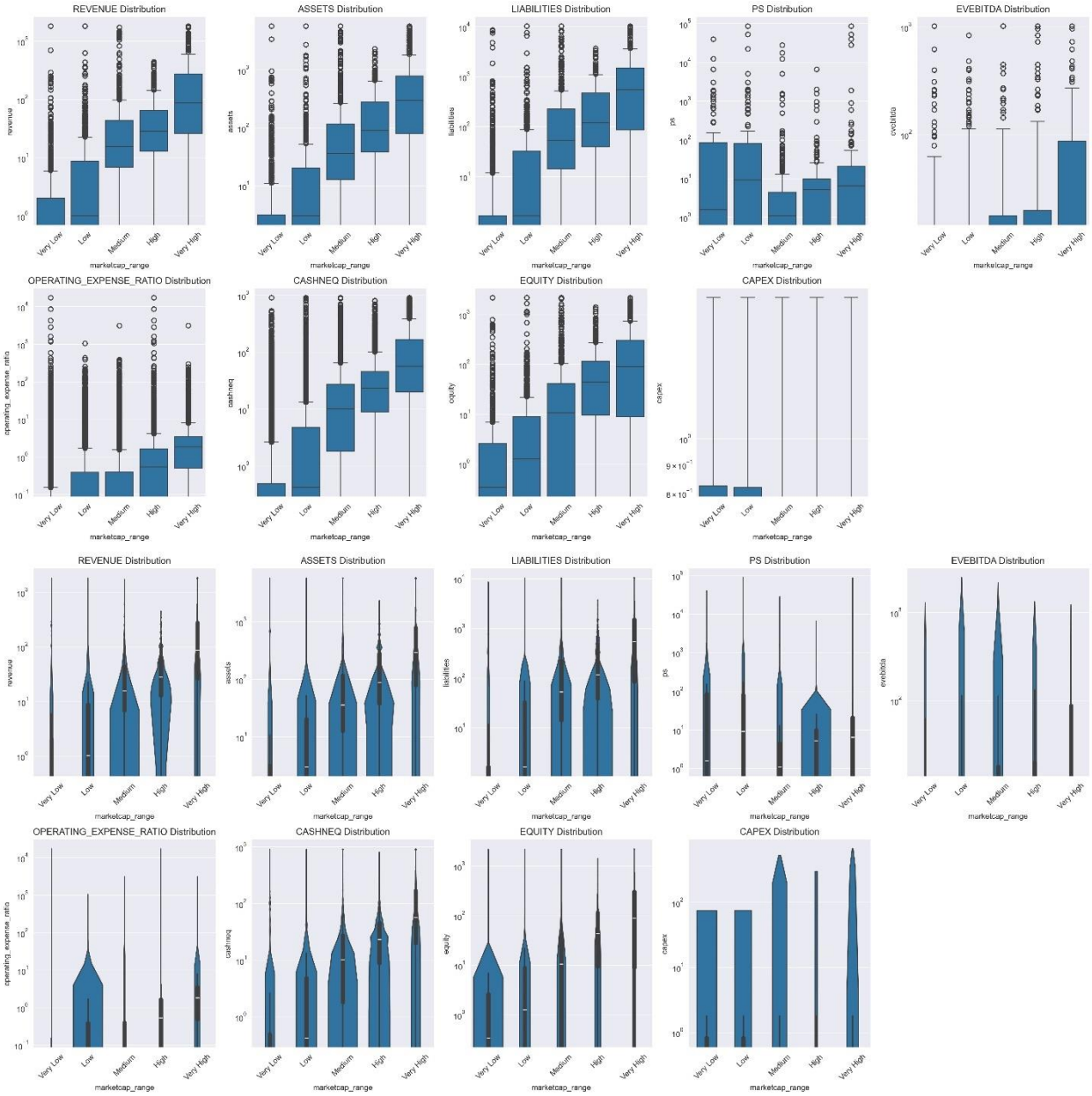
2) Bivariate Analysis



<Figure 6: Features vs. log marketcap_first_day>

By creating the data visualization of scatterplots above using Seaborn and subplot arrangement using Matplotlib, identification of top correlated features and a visual representation of feature relationships can be seen. The distribution plot (Figure 4 bottom right corner) shows the spread of log-transformed market capitalizations, helping identify typical valuation ranges. Revenue asset utilization shows the strongest positive correlation (0.78), so companies efficiently generate revenue from their assets to achieve higher market caps. Investment-related metrics dominate the top correlations: investmentsc, investmentsc^2, investments, and investments^2; this pattern suggests that companies with robust investment activities typically command higher market valuations. Cash flow metrics furthermore demonstrate significant influence: ncfo (0.57), investments_ncfo (0.54), and fcf_to_assets (0.54). These

correlations highlight the importance of strong cash flow management for IPO success. Revenue shows a positive correlation confirming that top-line growth remains a key valuation driver.



<Figure 7: Boxplot (upper 2 rows) and Violin Plots (bottom 2 rows) of Top Features>

The box plots and violin plots show raw data distributions across differing market cap ranges, actual financial metrics (revenue, assets, etc.) and how these metrics naturally cluster into market cap categories. Revenue shows a strong positive correlation with log market cap with a clear progression from low to high. This strong positive correlation between revenue and market cap aligns with Fama's Efficient Market Hypothesis (EMH), where market efficiency prices reflect available revenue information; higher revenues lead to proportionally higher valuations; and market processes revenue data systematically. In EMH, market prices effectively incorporate all publicly available information, including figures of revenue. Skewness decreases as market cap is increasing, which indicates more stable revenue distributions.

Assets and liabilities scale with company size; asset/liability ratios remain relatively consistent across ranges. Higher market caps show lower skewness and kurtosis, suggesting more normal distributions; larger firms' lower skewness in assets/liabilities supports Trade-off Theory. PS ratio exhibits interesting patterns with low cap companies showing the highest mean ps; high volatility in lower market caps and more ps metrics in the high cap range, aligning with Tobin's Q theory of market value vs. replacement cost. Operating metrics reveal operational efficiency with operating ratio patterns supporting Chandler's Scale and Scope theory. Operating expense ratios increase with the market cap; very high cap companies maintain higher but more stable ratios; and lower market caps show extreme kurtosis values. These efficiency metrics demonstrate principles of returns to scale.

Additionally, cash and equity positions strengthen with size. There is a clear progression in cash holdings from Low to Very High. Equity follows similarly in pattern with strongest positions in Very High cap. Distributions become more symmetrical in higher caps, which support Financial Slack theory (how companies maintain excess liquidity and resources beyond immediate operational needs to provide flexibility and buffer against uncertainty). Thus, there should be higher cash reserves in tech companies, conservative leverage ratios, substantial credit line maintenance, and strategic resource allocation. Capital expenditure (CAPEX) in higher caps indicates more investment. More consistent CAPEX patterns are in the higher market caps with lower variability in Very High cap companies, which align with Tobin's Investment Theory and larger cap consistency supports Rational Expectations Theory.

IV. Pre-processing and Training Data

Parquet file storage was utilized after extracting the raw data from Nasdaq for efficient further processing and analysis. This allows for quick access to the processed data, which is particularly beneficial when working with large datasets like multiple tech startup IPOs over the past decade. It streamlines the subsequent steps for the analysis pipeline, making it easier to perform complex computations and model training on the prepared data.

For the data processing setup, an IncrementalPipeline class that handles both preprocessing and model training was created, and chunk-based processing to handle large datasets efficiently using Parquet files was implemented and then separate pipelines were set up for numeric and categorical features.

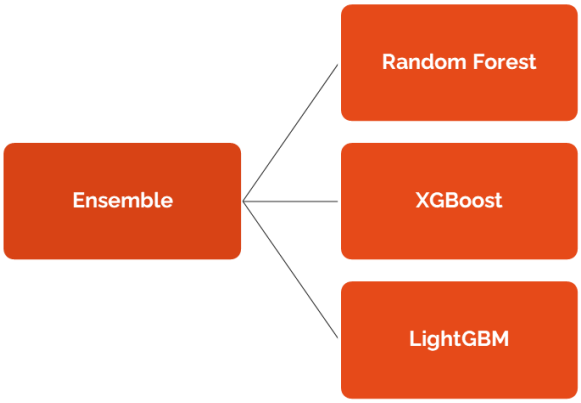
The preprocessing pipeline for numeric features has a median imputation for missing values applied and standard scaling (mean=0, std=1) was performed. For categorical features, constant imputation with 'missing' as fill value was used and then one-hot encoding with unknown category handling was applied.

Training implementation was carried out by processing data in manageable chunks of 10,000 rows. For each chunk, data was split into training (80%) and test (20%) sets; incrementally fitted the SGD regressor; applied log transformation to the target variable (marketcap_first_day); and model performance was evaluated using MSE and R^2 metrics. The approach uses scikit-learn's Pipeline and ColumnTransformer for robust preprocessing, while the incremental learning allows handling large datasets well without loading everything into the memory at once. This implementation provides a

scalable solution for processing and analyzing prediction data while maintaining good memory efficiency through chunked processing.

Dask DataFrames was implemented to efficiently handle our large-scale financial dataset stored in Parquet format. By using dd.read_parquet() to load our preprocessed features from the preprocessed parquet files, a memory-efficient data processing pipeline was established. Dask's capabilities were leveraged to examine our dataset structure through features_ddf.info() and features_ddf.dtypes, giving valuable insights into our data characteristics while maintaining optimal memory usage.

V. Modeling and Results



<Figure 8: Ensemble Stacking Concept- Comprised of 3 Models>

A description of the implemented models and their characteristics are as below:

Random Forest Model:
- Configured with 100 trees, max depth of 10, optimized parameters for split and leaf samples
- Uses sqrt feature selection strategy for optimal feature subset selection
- Leverages parallel processing with n_jobs= -1 for faster computation
- Provides robust feature importance rankings through mean decrease in impurity

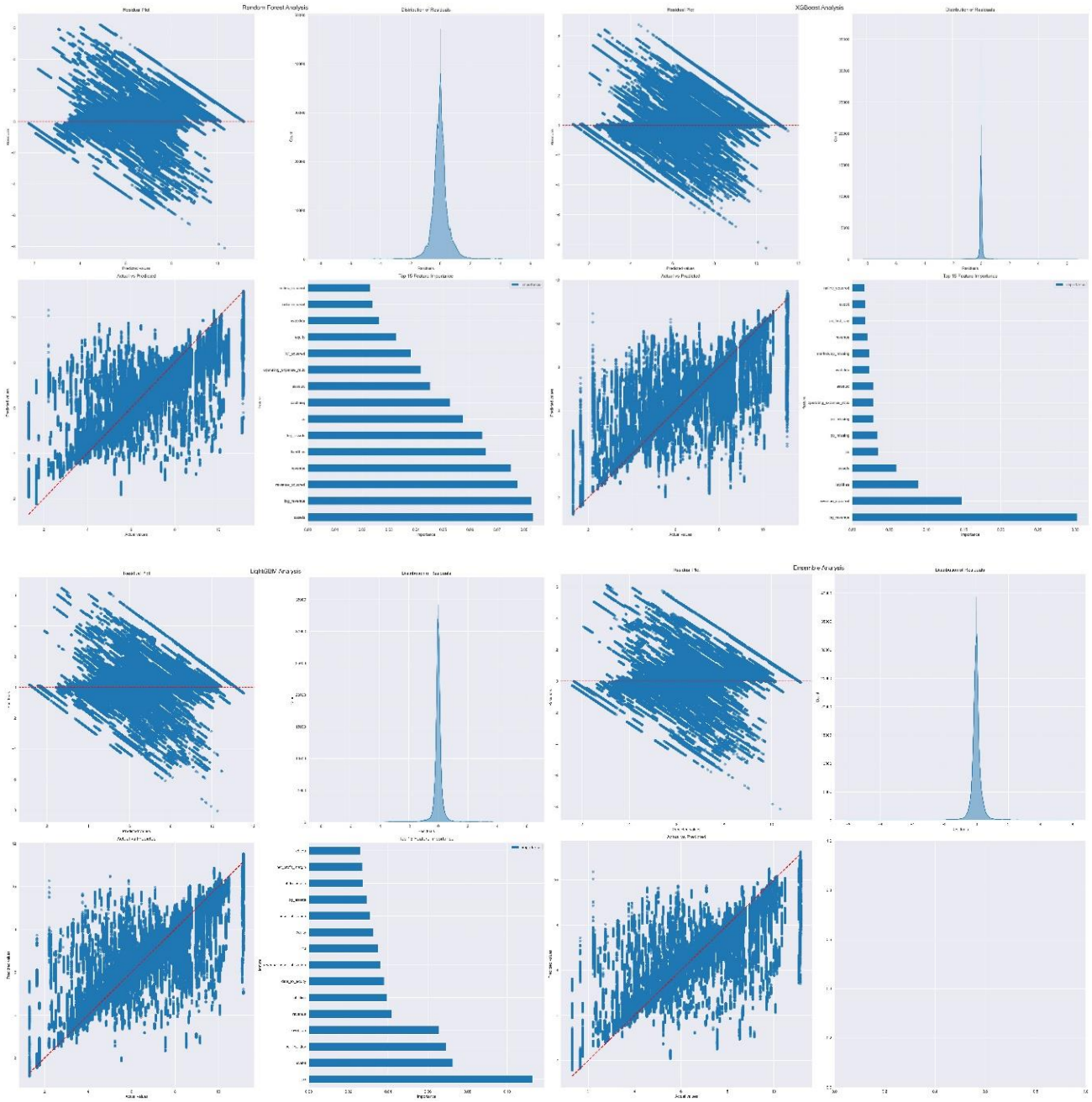
XGBoost Model:
- Implements a highly tuned configuration with 733 trees and a learning rate of 0.027
- Uses column and row sampling (colsample_bytree=0.636, subsample=0.777) to prevent overfitting
- Incorporates L1 and L2 regularization (reg_alpha=0.1, reg_lambda=1.0)
- Optimized gamma parameter of 2.472 for tree construction

LightGBM Model:
- Employs 500 trees with a 0.05 learning rate for balanced speed and accuracy
- Uses leaf-wise growth strategy with num_leaves=31
- Regularization parameters (reg_alpha=0.1, reg_lambda=1.0)
- Features automatic depth control with max_depth= -1

Stacking Ensemble Strategy:
- Combines predictions (Figure 8) using weighted averaging
- Assigns weights of 0.4 to both Random Forest and XGBoost and 0.2 to LightGBM

- Leverages the strengths of each model while minimizing individual weaknesses
- Demonstrates improved prediction stability and reduced variance.

The models were trained on a dataset of 25 million rows with performance evaluated using R^2, MSE, RMSE, and residual statistics. Each model contributes unique strengths to the ensemble.

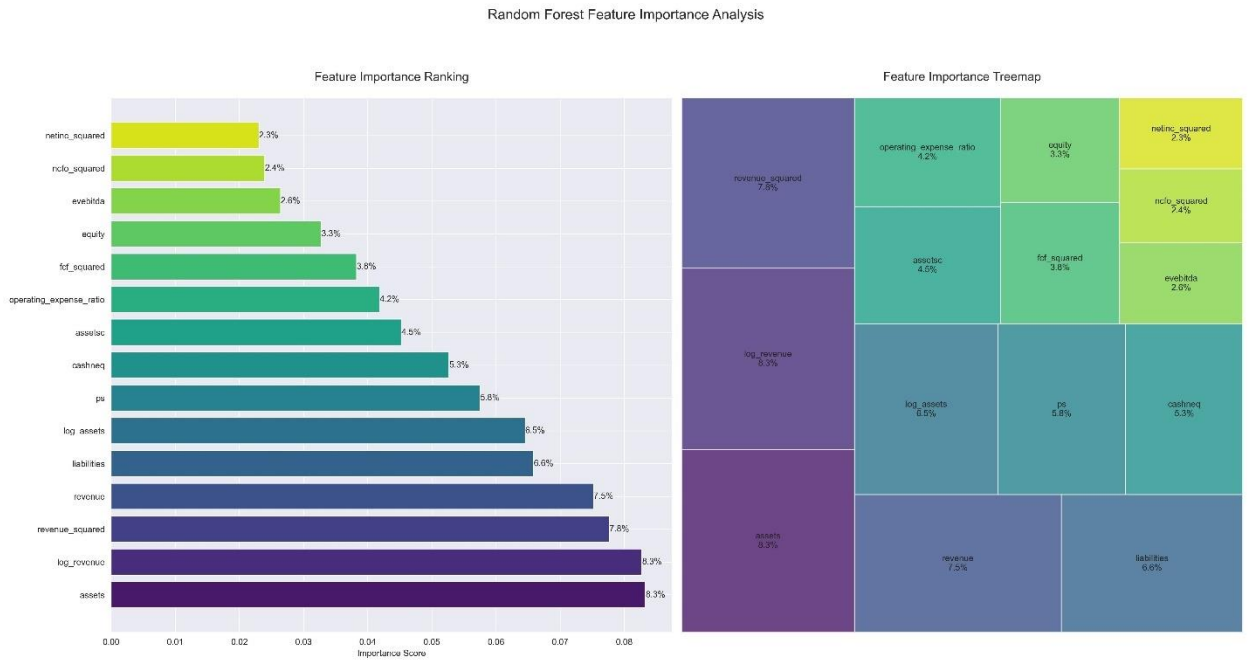


<Figure 9: RF (upper left), XGBoost (upper right), LightGBM (bottom left), Ensemble (bottom right) Comparison Model Performance>

Statistic	Random Forest	XGBoost	LightGBM	Ensemble
R^2	0.8946	0.9559	0.9436	0.9416
MSE	0.4059	0.1697	0.2173	0.2249
RMSE (original scale)	3423.8690	2303.9076	2686.7740	2797.0218
Residuals Skewness	0.2887	0.3762	0.1309	0.3112
Residuals Kurtosis	13.1899	38.6206	33.4788	28.0920
Top Feature 1	assets 0.083306	log_revenue 0.302904	ps 0.113000	N/A
Top Feature 2	log_revenue 0.082760	revenue_squared 0.147632	evebit 0.072667	N/A

<Table 6: Comparison Model Statistics>

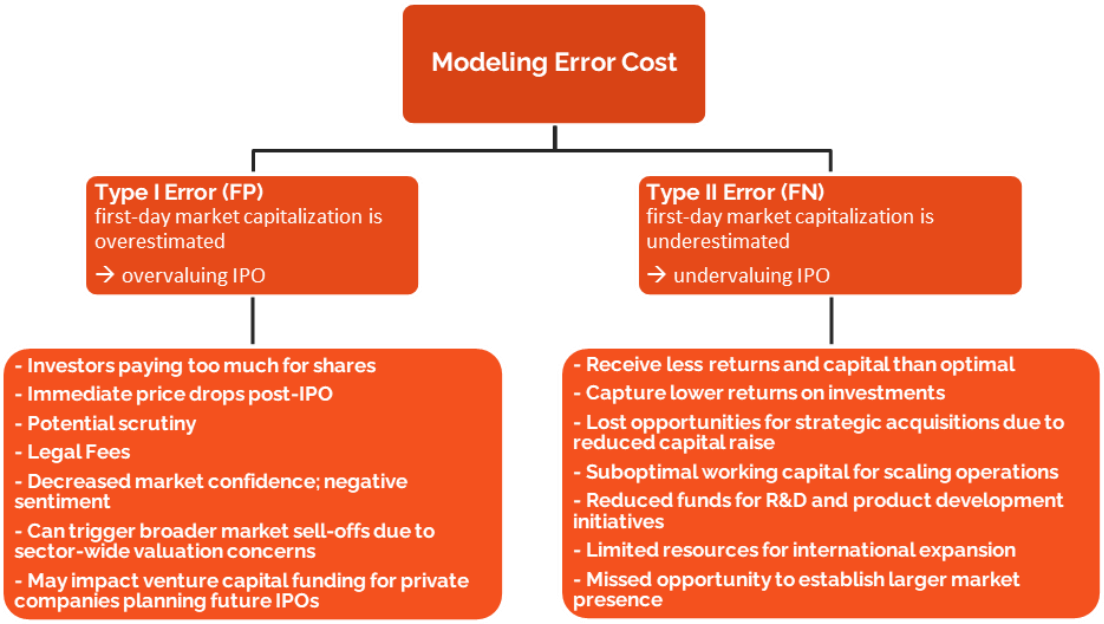
The results with 25 million samples show robust model performance. The Random Forest model demonstrates excellent stability with a very balanced feature distribution of 0.083 to 0.023. It has the lowest kurtosis (13.18) amongst all models, most balanced skewness (0.288), and a strong R^2 of 0.8946. XGBoost shows high accuracy with best R^2 (0.9559) and lowest RMSE (2303.90). But there is high feature concentration on log_revenue (0.302). LightGBM offers a middle ground with a strong R^2 of 0.9436, lowest skewness (0.131), and puts focus on ps as the top feature (0.113). Throughout the models, feature importance patterns remain consistent with assets, revenue, and liabilities as key predictors. Moreover, financial ratios maintain significant importance, and market metrics (ps, evebit) show strong predictive power.



<Figure 10: Random Forest Feature Importance Analysis>

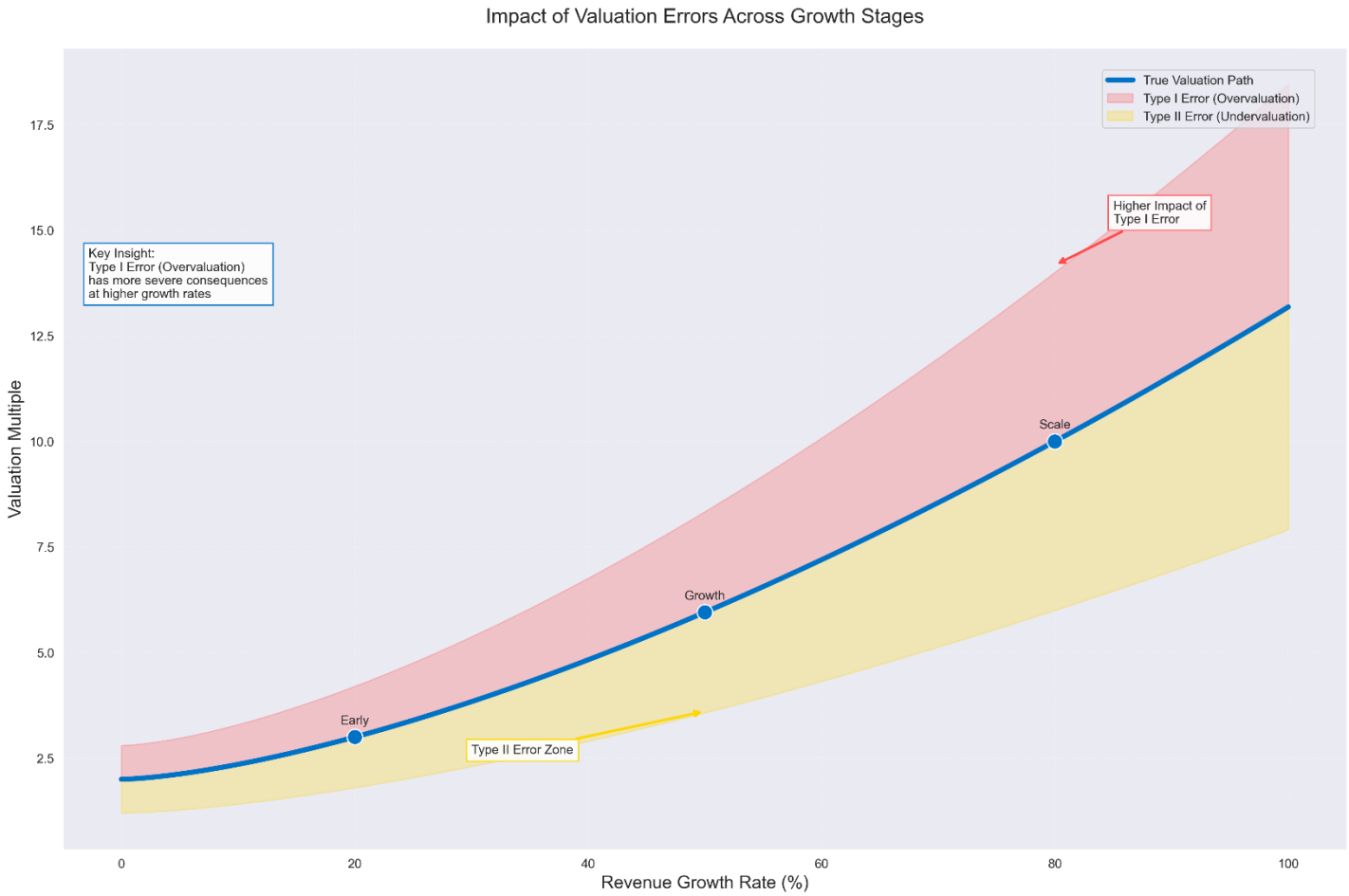
Type I Error (False Positive- Overestimation): A FP occurs when first-day market capitalization is overestimated, which leads to overvaluing the IPO. This leads to financial consequences such as investors paying too much for the shares; a company might face immediate price drops and price corrections post-IPO; potential scrutiny, legal, and reputational risks for underwriters; and a market that might lose confidence in future IPO valuations, leading to decreased market confidence.

Type II Error (False Negative- Underestimation): A FN occurs when the first-day market capitalization is underestimated. This leads to undervaluing the IPO and financial consequences, where a company leaves money on the table; early investors and founders receive less returns and capital than optimal; shares might dramatically spike on the first day, indicating poor pricing and inefficiency; and there is a lost opportunity for the company to raise maximum capital for expansion and growth.



<Figure 11: Type I Error vs. Type II Error Modeling Error Cost>

Contrasting Type I and II Error Costs: In terms of market dynamics, a FP creates artificial market expectations; leads to volatility in early trading days; affects institutional investor participation rates; and influences future IPO pricing strategies. On the other hand, a FN creates opportunities for short-term traders; generates high first-day returns for initial investors; affects company market positioning; and influences competitive advantage in the sector.



<Figure 12: Impact of Type I and Type II Valuation Errors Across Company Growth Stages>

Therefore, a FP (overestimation) is generally considered more severe in IPO market cap prediction. For example, in terms of market impact, overvaluation creates systemic risks in the market and affects broader investor confidence. This can trigger regulatory investigations, creating negative ripple effects for future tech IPOs. The financial consequences of immediate post-IPO drops hurt investor portfolios. There is a higher likelihood of shareholder lawsuits, leading to investment bank and underwriter reputation damages. These consequences can furthermore freeze market appetite for subsequent tech IPOs, having long-term effects of making future IPO pricing more challenging; reducing institutional investor participation; creating skepticism in tech sector valuations; and impacting overall market efficiency. Unlike a FP, a FN could create positive market momentum, which generates excitement around the stock while building confidence in the tech sector, allowing for natural price discovery.

Historical examples powerfully demonstrate the impact of IPO overvaluation. The dot-com bubble of the late 1990s serves as a prime case study, where excessive IPO valuations led to widespread market damage, affecting not just individual companies but entire tech sectors. Recent tech IPO corrections have similarly triggered extended periods of valuation reassessment across the industry. From a regulatory standpoint, the Securities Exchange Commission (SEC) maintains heightened scrutiny on overvalued IPOs, implementing more rigorous oversight mechanisms. This increased regulatory attention often results in a higher frequency of class action lawsuits and demands for more comprehensive post-

overvaluation disclosure requirements. The long-term market structure impacts are equally significant. Market making capabilities undergo substantial changes as firms adjust their risk parameters, while investment banks must evolve their risk management practices to adapt to these lessons learned. These factors frequently catalyze fundamental structural changes in IPO processes, including enhanced due diligence procedures, more conservative pricing strategies, and refined allocation methodologies. Together, these historical patterns, regulatory responses, and market structure adaptations underscore the critical importance of accurate IPO valuations in maintaining market stability and investor confidence.

The Random Forest model's more balanced error distribution makes it valuable for risk-averse stakeholders, even though it has lower overall accuracy compared to XGBoost, LightGBM, or the Ensemble model. The Random Forests's feature importance ranking captures important IPO valuation drivers:

1) Balance Sheet Strength- Assets (0.083) and liabilities (0.065) as top indicators, strong emphasis on working capital through assetsc (0.045), and cashneq (0.052) highlighting liquidity importance
2) Revenue Metrics- Multiple revenue transformations (log, squared) capturing non-linear scaling effects, operating expense ratio (0.041) indicating operation efficiency, asset utilization metrics showing capital efficiency
3) Market Multiples- ps (price-to-sales) at 0.057 reflecting tech sector valuation norms, evebitda (0.026) capturing operational value, equity value components (0.032) for ownership structure impact

Random Forest is simpler than XGBoost, LightGBM, and Ensemble models in several key ways:

1) Algorithm Structure: Random Forest uses a straightforward averaging of multiple decision trees. Each tree is independent and built using simple random sampling
   The voting/averaging mechanism is straightforward and intuitive
2) Hyperparameter Tuning: There are fewer hyperparameters to tune. The model is more robust to parameter settings and less sensitive to small parameter changes
3) Training Process: Random Forest has naturally parallel processing. There is no sequential building unlike boosting methods, and so there is a simpler optimization objective.
4) Interpretability: Random Forest is easier to explain to non-technical stake holders since the feature importance is more straightforward to interpret, and decision paths can be traced through individual trees.
5) Implementation: Random Forest requires less computational resources. It is more stable across different implementations and easier to deploy and maintain in production.

This simplicity makes Random Forest an excellent baseline model and a reliable choice when model interpretability and ease of implementation are priorities. Its performance (R-squared: 0.89) is strong considering its simpler approach, demonstrating the value of straightforward, robust modeling techniques in financial applications. Random Forest's balanced approach helps identify both unicorn

potential and downside risks, making it particularly valuable for IPO pricing committees, institutional investors, market makers, and investment banking underwriters.

The emphasis on balance sheet metrics aligns with fundamental financial theory, modern IPO valuation frameworks, and a hierarchical importance aligning with Pecking Order Theory of financing (information asymmetry, financing hierarchy, market timing), Growth Option Valuation (real options analysis, embedded options) models, and Modern Tech Sector valuation (Rule of 40 Integration, TAM Expansion Potential, Platform Economics) frameworks:

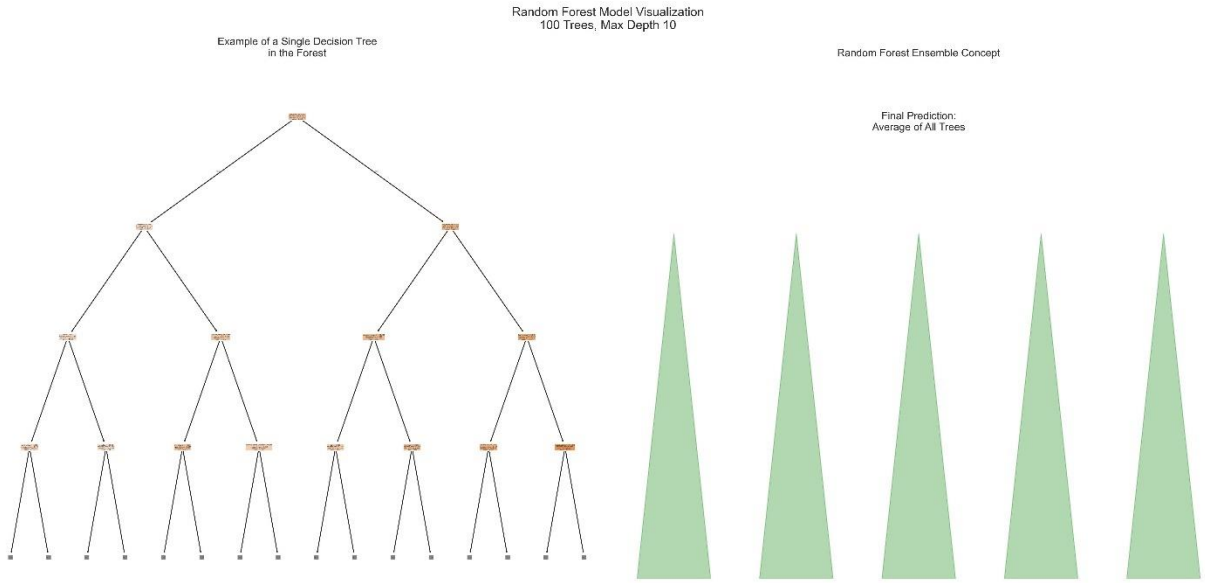
1) Assets (0.083 importance)
- Working Capital Efficiency: The high weighting of current assets (assetsc: 0.045) reflects the tech sector's need for operational liquidity
- Asset Light Model: Tech startups typically have higher intangible asset values, making total assets a key differentiator for scalability potential
- Growth Options Theory: Total assets serve as a proxy for real options in expansion and scaling

2) Liabilities (0.065 importance)
- Capital Structure Signals: Debt levels indicate both financial risk and tax shield benefits per Modigliani-Miller theory
- Agency Cost Theory: Liability structure reflects management's risk appetite and governance quality
- Financial Flexibility: Lower liability levels typical in tech IPOs indicate greater future financing options

3) Cash Position (cashneq: 0.052)
- Burn Rate Buffer: Essential for tech startups' runway calculations
- Real Options Theory: Cash reserves represent embedded options for future investments (Asset base (0.083) represents platform for future expansion; cash reserves (0.052) quantify strategic flexibility; and operating leverage captured through expense ratios (0.041).
- Signaling Theory: High cash levels signal strong pre-IPO investor confidence

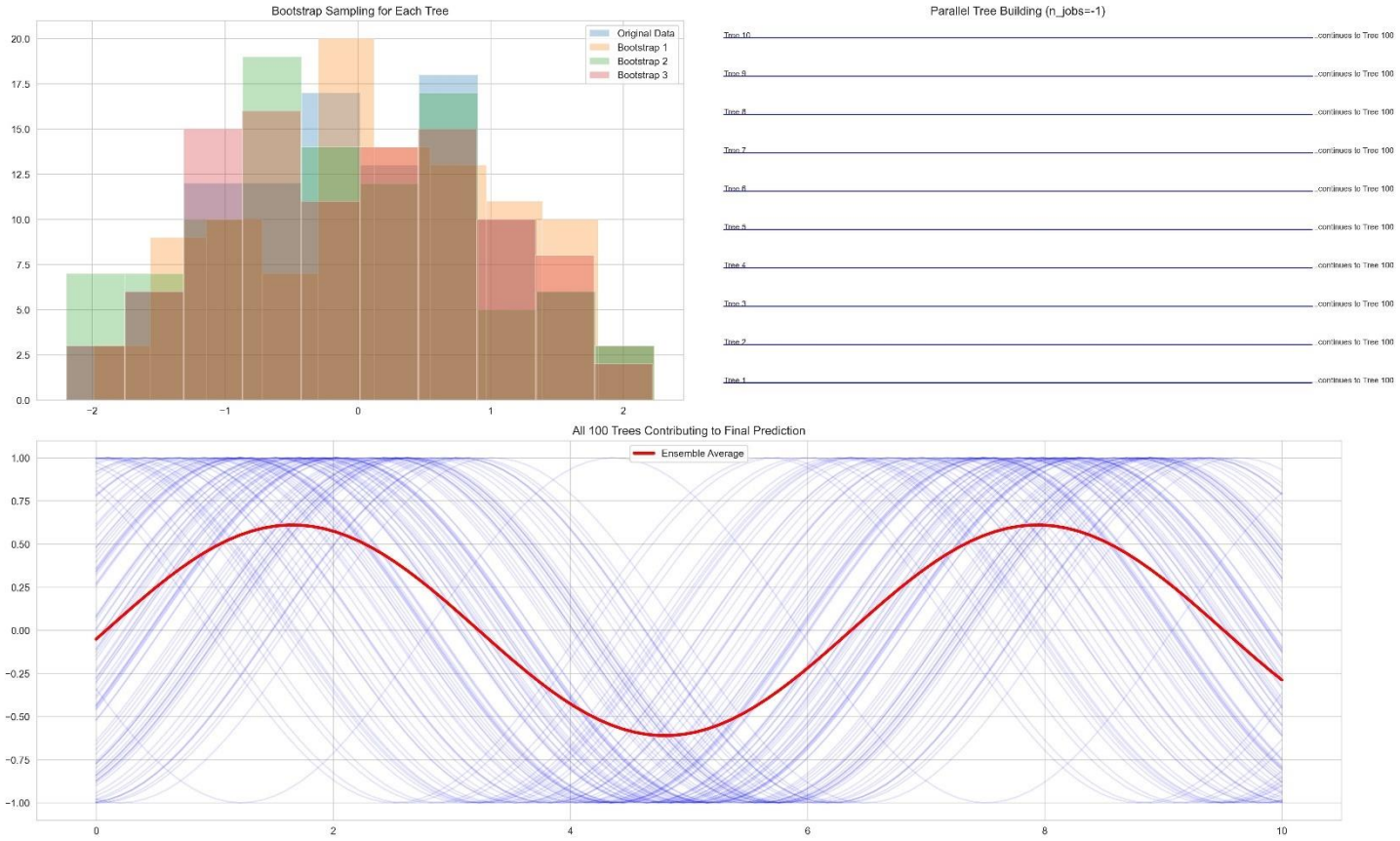
4) Operating Metrics Integration
- Asset Utilization: Operating expense ratio (0.041) shows efficiency in converting assets to revenue
- Scale Economics: Revenue squared (0.077) captures exponential growth potential
- Network Effects: Log transformations (log_assets: 0.064) capture diminishing returns in scale

5) Market Value Drivers
- Enterprise Value Components: evebitda (0.026) captures operation efficiency
- Equity Value Drivers: ps ratio (0.057) reflects revenue multiple expansion potential
- FCF Metrics: FCF^2 (0.038) indications exponential value creation potential



<Figure 13: Concept Visual for Random Forest Model>

Figure 13 illustrates the actual decision rules uses, feature names and thresholds at each split, and the first three levels of decision-making (max_depth=3) on the left. The right panel illustrates the ensemble concept of how multiple trees work together, which are shown as green triangles, the ensemble nature of the model (in this case, 100 trees), and the concept of averaging predictions across all trees. The illustration represents square root feature selection at each split, parallel processing (n_jobs= -1), and minimum requirements for node splitting (min_samples_split=5, min_samples_leaf=2).

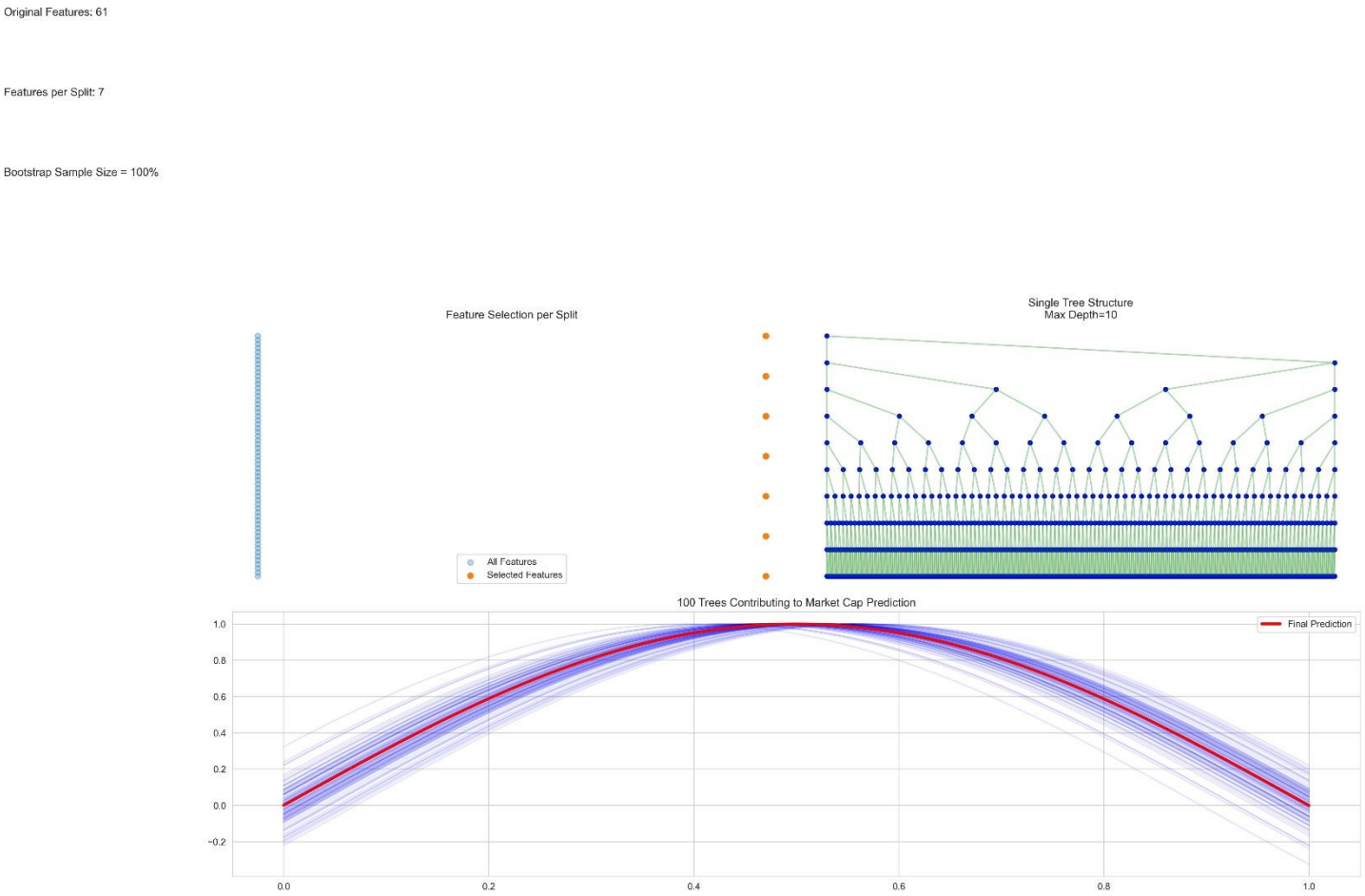


<Figure 14: Concept Visual for Mechanisms of the Random Forest Model>

The visualizations made in Figure 14 illustrates three core concepts of the Random Forest Model:

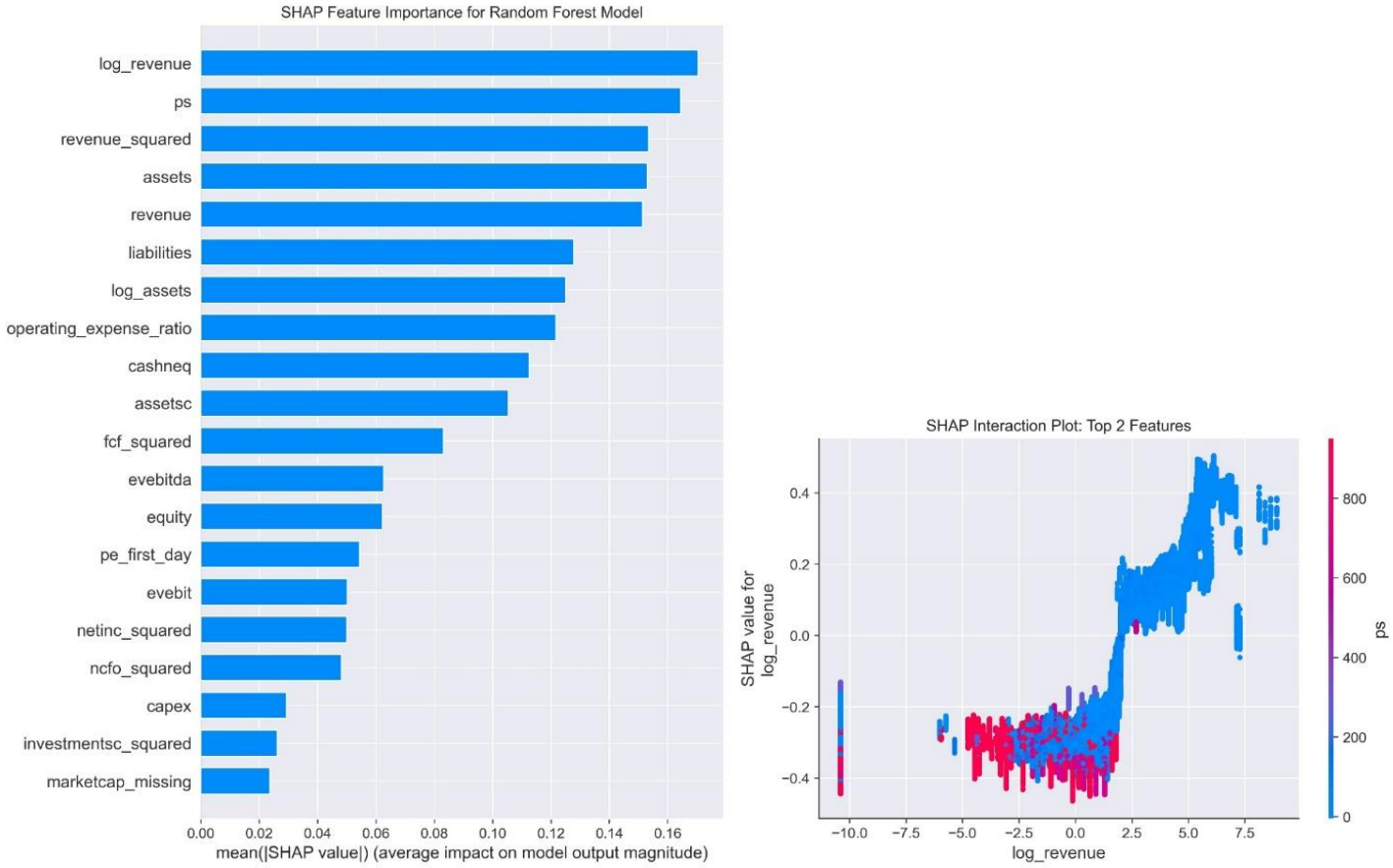
1) The top-left plot shows how the model creates bootstrap samples for each of the 100 trees, where each tree gets its own random sample from the financial dataset.
2) The top-right plot demonstrates how the model utilizes parallel processing (n_jobs= -1) to build all 100 trees simultaneously, maximizing computational efficiency.
3) The bottom plot reveals how the model combines predictions from all 100 trees (light blue lines) to create a final ensemble prediction (red line) for market capitalization.

This visualization effectively captures the fundamental process that the model uses to make its predictions.



<Figure 15: Concept Visual for Random Forest Architecture>

Figure 15 illustrates the data boostrap process (top left), which shows how it handles the 61 features with feature selection per split (approximately 8 features based on square root) and visualizes the bootstrap sampling process at 100% sample size. The top right panel displays a decision tree with ten levels of depth; shows the binary splitting process; illustrates how decisions branch out exponentially ($2^{\text{depth nodes}}$); and the green connections show path relationships between decision nodes. The bottom panel represents the ensemble prediction process. It demonstrates how 100 individual trees contribute to the final prediction with blue lines representing individual tree predictions and a red line to show the final aggregated prediction, showing how ensemble averaging reduces prediction variance. This comprehensive visualization could help stakeholders understand how this Random Forest model combines multiple decision trees to generate robust market capitalization predictions, while maintaining interpretability through systematic feature selection and structured decision paths.

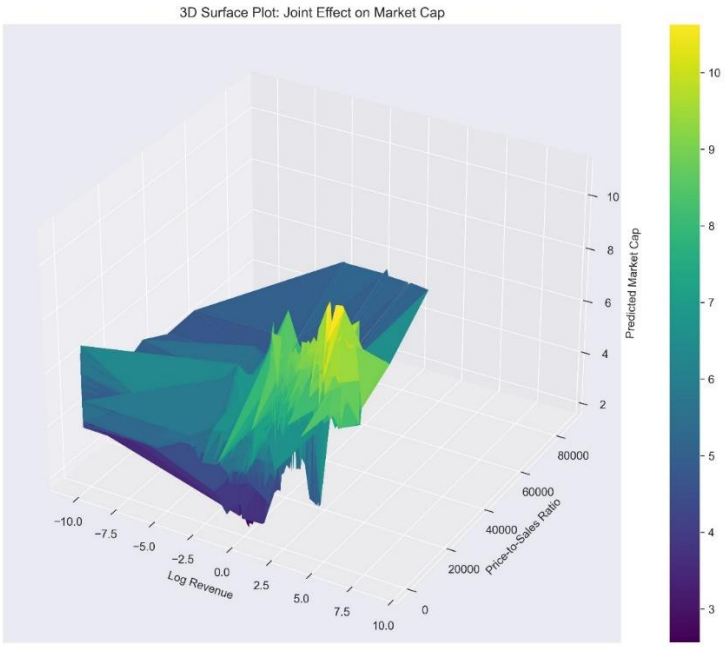


<Figure 16: SHAP Feature Importance for Random Forest Model>

SHAP (Shapley Additive exPlanations) analysis provide interpretability for the Random Forest model predicting IPO market capitalization. The SHAP summary plot visualizes the impact of each feature on model predictions. Log_revenue emerges as the most influential feature (0.170 mean absolute SHAP value). Price-to-sales (ps) and revenue_squared follow as second and third most important features, respectively.

This aligns with market logic where revenue growth and scale are primary valuation drivers for tech companies. The significant impact of assets and liabilities indicates that balance sheet strength remains a key consideration for investors. Operating expense ratio's high ranking (8th place) demonstrates that operational efficiency matters substantially in valuation outcomes. The presence of both raw and transformed variables (log, squared terms) in the top features shows that the relationship between financial metrics and market cap is often non-linear.

These findings provide valuable guidance for investors and analysts in prioritizing which financial metrics deserve the most attention when evaluating potential tech IPOs. The model effectively captures both traditional valuation metrics (like EVEBITDA) and tech-specific indicators (like revenue multiples), reflecting the unique characteristics of technology company valuations.



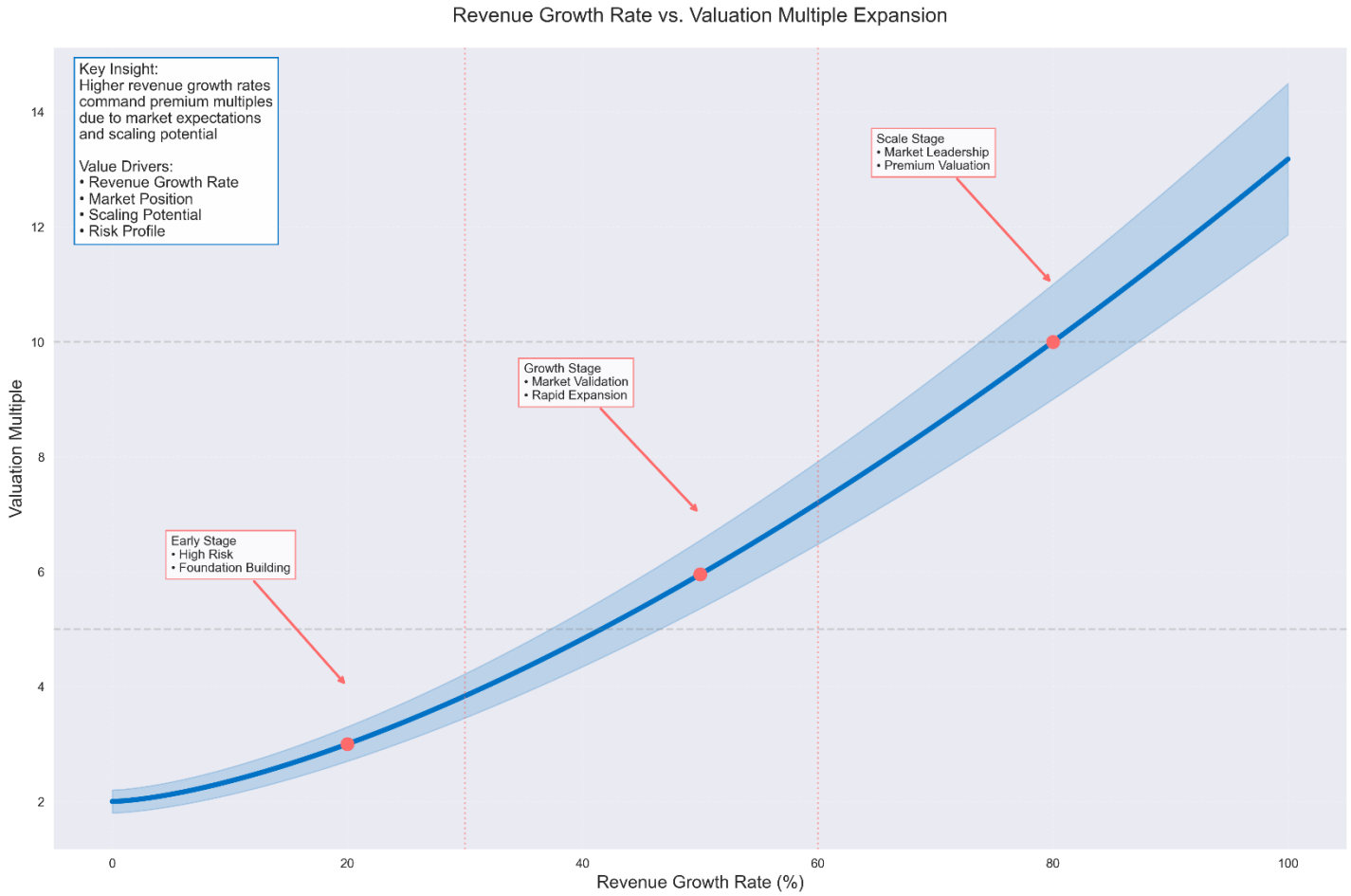
<Figure 17: 3D Surface Plot Log Revenue vs. ps>

The 3D visualization and statistical analysis of the relationship between the two most influential features (log_revenue and ps ratio) and their combined effect on predicted market capitalization accomplishes analysis for a matrix between features and predicted values, generating comprehensive distribution statistics (mean, median, standard deviation, quartiles), and creating quintile-based analysis showing hos market cap varies across various ranges of both features. This aids in understanding the interaction effects between these key metrics.

There are strong correlation patterns with log revenue showing positive correlation (0.61) with market cap; ps ratio has slight negative correlation (-0.09) with market cap; and log revenue and ps have weak negative correlation (-0.18). Distribution characteristics show that log revenue ranges from -10.39 to 8.92, centered at 1.95. PS ratio shows wide range (-294.3 to 86395.1), median at 3.7. Market cap predictions span 1.76 to 11.15, mean at 6.29. Quintile analysis of log revenue impact show a clear ascending pattern from 4.28 to 8.00 with the strongest jump between second and third quintiles (5.24 to 6.74). The statistics reveal patterns in how log revenue and ps ratio interact to predict market capitalization. Revenue has dominance with log revenue showing powerful predictive strength; each revenue quintile increases drives substantial market cap growth with the most dramatic impact occurring in middle revenue ranges (2.26 to 3.02). There are nuances in the ps ratio with optimal ps range ixisting in the middle quintiles, peak market cap predictions at $0.3 < ps < 24.6$. High PS ratios (>24.6 suggest potential overvaluation. As for the distribution, market cap predictions cluster around a mean of 6.29, 75% of predictions fall below 7.64, and exceptional cases reach up to 11.15.

Revenue growth consistently drives valuation increases; ps ratio acts as a moderating factor; and their combined effects explain valuation variations. Revenue growth shows clear step-up pattern, and ps ratio shows an inverted U-shape effect. The sweet spot exists at moderate ps ratios with high revenue. From a financial perspective, log revenue's 0.61 correlation confirms the fundamental principle that revenue scale directly drives market capitalization; each revenue quintile jump, especially 2.25 to 3.02,

represents key valuation thresholds investors recognize; and log transformation captures the diminishing marginal returns of revenue growth. 0.3 < ps < 24.6 indications market efficiency in pricing. Peak valuations in middle ps quintiles (6.79) suggest rational pricing zones, and declining impact in highest ps quintile (5.41) signals market skepticism of extreme multiples. The investment implications thus are revenue growth remains the most reliable value creator, ps ratios serve as valuation guardrails; and the sweet spot is when high revenue growth is met with sustainable ps multiples.

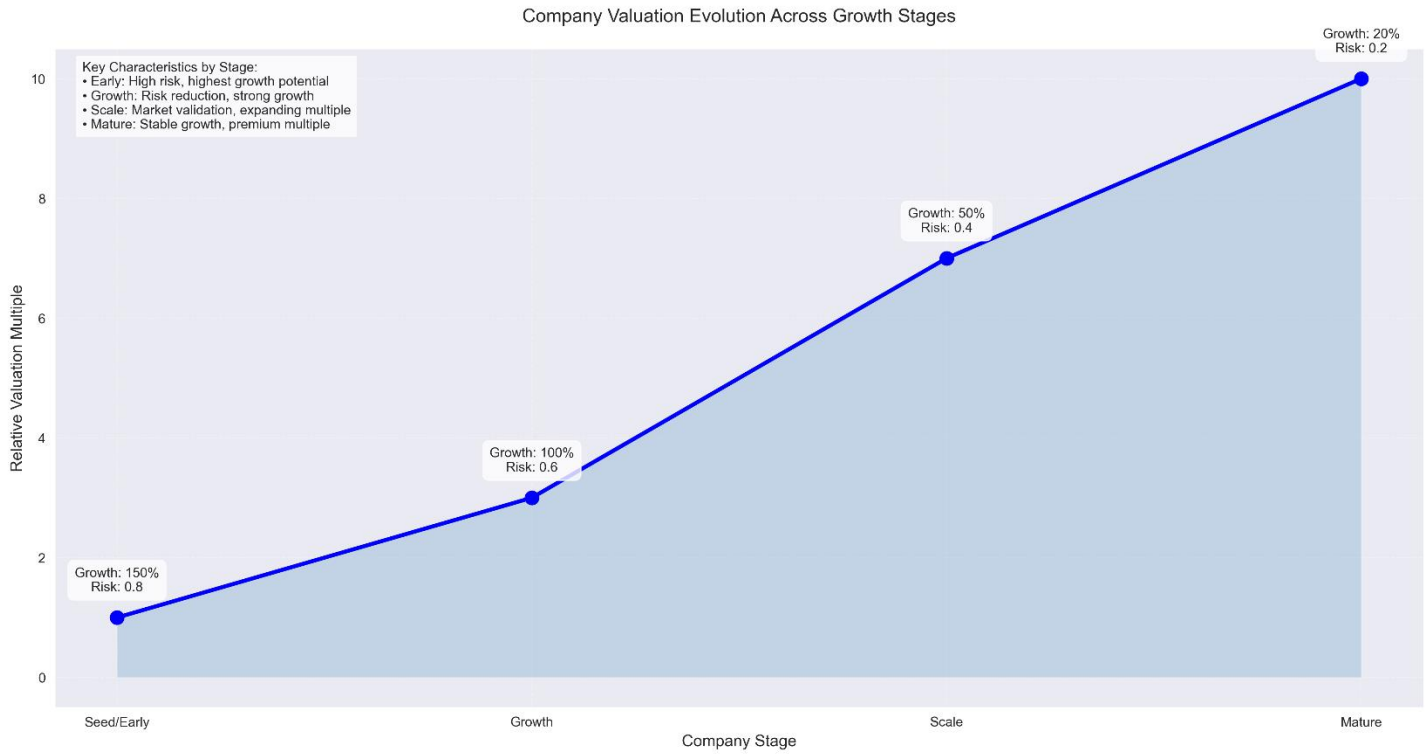


<Figure 18: Revenue Growth Rate Drives Multiple Expansion”>

This is useful for valuation framework applications to set clear benchmarks for different company sizes, multiple ranges for different growth stages, and seeing quantifiable impact of scale on valuation. It is important to note that if one is able to scale revenue, then the multiple will expand. Meaning at lower growth rate, the valuation multiple is lower vs. one can achieve a premium multiple at higher revenue growth rate. Figure 18 is a visualization created to show a clear curve of the non-linear relationship between revenue growth and valuation multiples with three distinct stages: Early Stage (lower multiples), Growth Stage (expanding multiples), and Scale Stage (premium multiples).

The visualization of Figure 18 demonstrates the powerful relationship between a company's growth trajectory and its market valuation. The graph illustrates how companies progress through three distinct stages—Early, Growth, and Scale—with each stage commanding progressively higher valuation multiples. Starting in the Early Stage with lower multiples, companies that successfully execute their

growth strategies move into the Growth Stage where multiple expansion begins to accelerate. As companies reach the Scale Stage, they achieve premium multiples, reflecting market confidence in their proven business model and sustained growth potential. The confidence bands surrounding the curve indicate the typical valuation ranges at each growth rate, while the milestone markers highlight key transition points. This framework effectively shows how higher revenue growth rates command premium multiples due to market expectations and scaling potential, making it a valuable tool for understanding valuation dynamics across different growth phases.

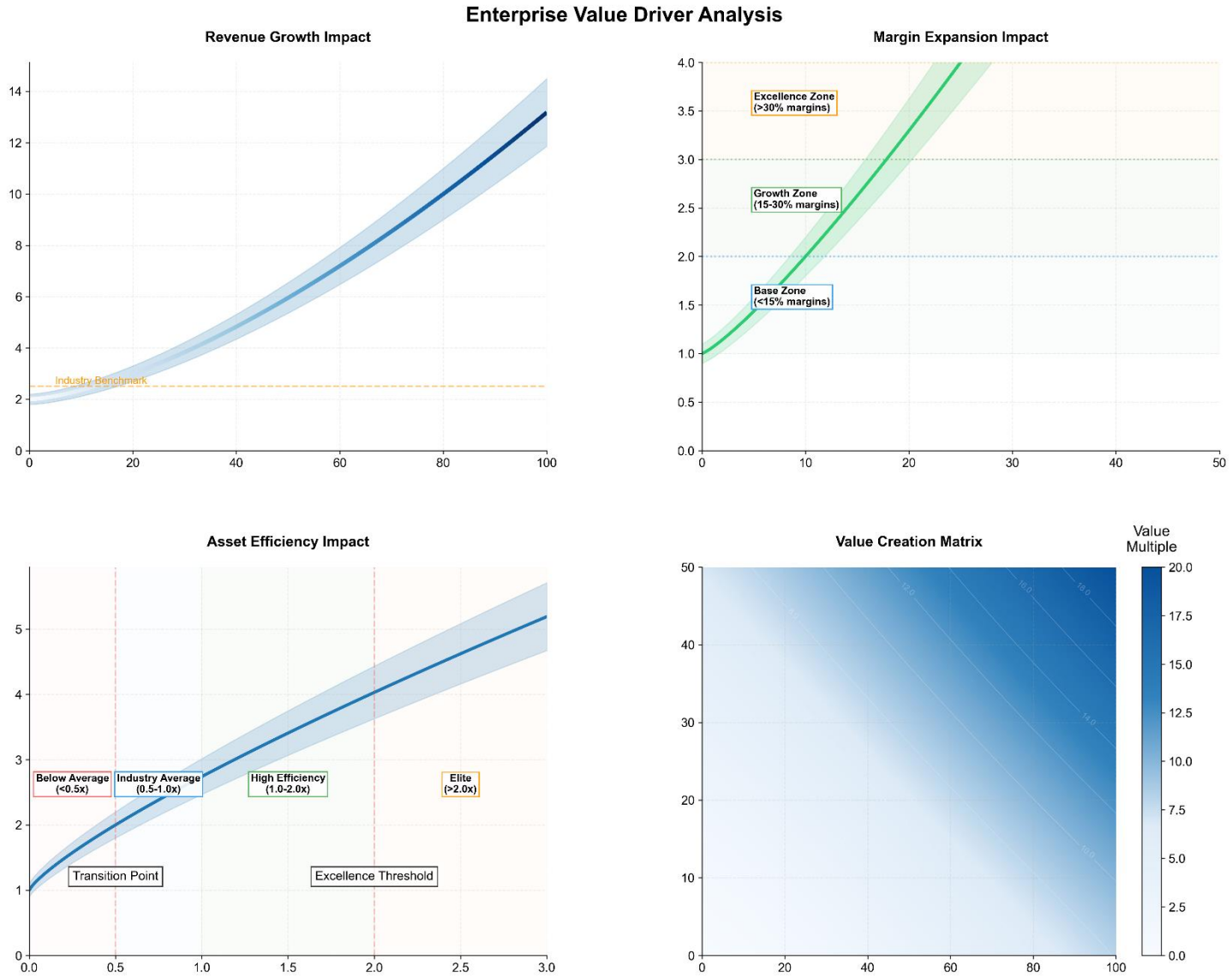


<Figure 19: Stage-Based Company Valuation Framework Concept>

Figure 19 illustrates the dynamic relationship between company growth phases and their corresponding valuations.

- Starting from the Seed/Early stage, companies typically experience the highest growth rates (150%) paired with elevated risk levels (0.8), commanding initial valuation multiples.
- As companies progress through the Growth stage, they demonstrate strong expansion while beginning to reduce risk factors.
- The Scale stage marks a significant milestone where market validation drives multiple expansion, despite moderating growth rates.
- Finally, at the Mature stage, companies achieve premium multiples backed by stable growth and significantly reduced risk profiles (0.2).

This progression highlights how successful companies can create substantial value by executing through each stage effectively, with valuation multiples expanding from 1x to 10x through the journey. The framework serves as a valuable tool for understanding how market participants value companies differently at each stage of their development.

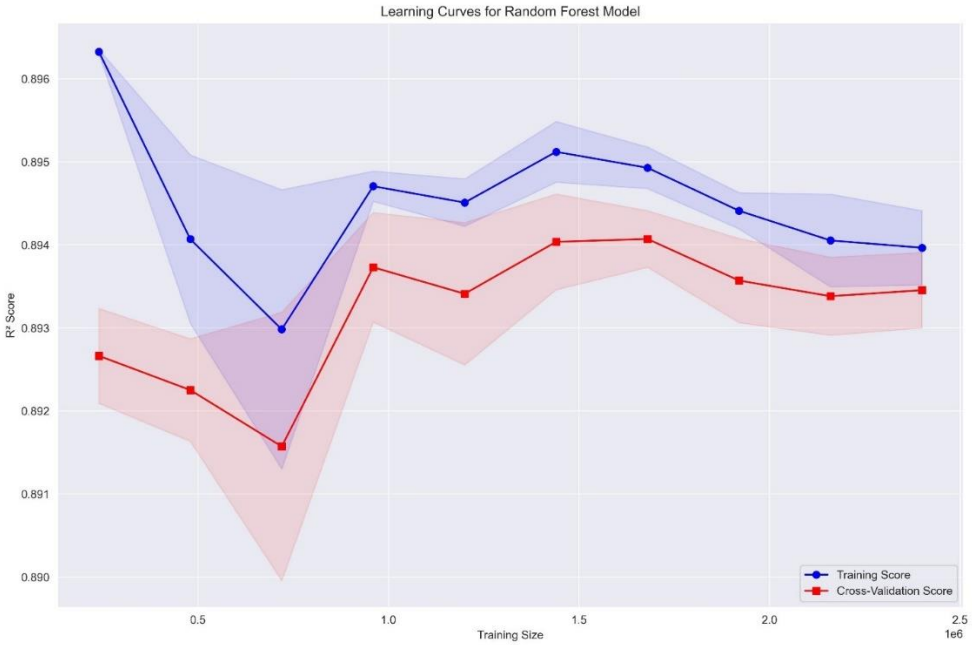


<Figure 20: Revenue Growth, Margin Expansion, Asset Efficiency>

The Enterprise Value Driver Analysis reveals three critical factors that significantly impact company valuations. Revenue growth demonstrates a strong correlation with multiple expansion, where growth rates exceeding 20% drive exponential value creation above the industry benchmark of 2.5x. The margin analysis identifies three performance tiers: Base Zone (<15% margins), Growth Zone (15-30%), and Excellence Zone (>30%), with each 10% improvement in margins contributing approximately 1.2x to the valuation multiple. Asset efficiency metrics establish four distinct categories of operational excellence, ranging from Below Average (<0.5x turnover) to Elite Performance (>2.0x), with optimal value creation beginning at the 1.0x threshold. The Value Creation Matrix illustrates the powerful synergistic effects between growth and margins, highlighting accelerating returns when companies achieve simultaneous improvements in both metrics. This comprehensive analysis indicates that companies maximizing all three drivers – particularly those maintaining high growth rates while operating in the margin Excellence Zone and Elite asset efficiency category – achieve the highest valuation multiples in the market. Nevertheless, however, growth is the most important.

	Random Forest	SHAP
Calculation Method	Features ranked based on total decrease in impurity across all trees	Features ranked based on average absolute impact on model output
Temporal vs. Global	Shows which features most frequently split trees	Shows actual magnitude on impact on predictions
Feature Interactions	Might undervalue features that work strongly in combination	Captures interaction effects explicitly in its calculations
Scale Sensitivity	Can be biased towards high-cardinality features	More robust to feature scaling and cardinality
Prediction Impact	Importance reflects tree structure decisions	Measures direct impact on the log market cap predictions
Features	Shows which features most useful for splitting decisions	Shows which features have largest impact on final predictions

<Table 7: Random Forest vs. SHAP Feature Importance Rank Differences>



<Figure 21: Learning Curves for Random Forest Model>

The learning curve statistics reveal excellent model performance and stability. High performance metrics with training scores consistently above 89% and validation scores closely tracking training scores. The figure shows remarkable stability with very low standard deviations (<0.002) across all sample sizes. Training and validation scores maintain consistency, and there is minimal variance between different training sizes. Learning progression has initial strong performance at 240,000 samples. There is stable performance through increasing sample sizes with optimal performance around 1.4-1.7 million samples. The model has no overfitting (training and validation scores are closely aligned); good generalization across sample sizes, and robust performance with larger datasets. In addition, the consistent R^2 scores indicate reliable log market cap predictions. Low standard deviations suggest highly stable predictions.

VI. Findings and Actionable Insights

The model's balanced feature importance distribution excels at capturing vertical-specific value drivers, modeling cross-vertical scaling patterns, and evaluating hybrid business models. This sophisticated framework diversity explains why the Random Forest model effectively captures multiple transformation types of key metrics, reflecting the complex value creation mechanisms across different tech verticals.

The model's ability to incorporate these varied frameworks demonstrates its robustness in predicting market capitalizations across the diverse tech sector landscape and can be a good tool for IPO valuation teams, institutional investors (data-driven entry points for IPO participation, portfolio optimization using first-day performance predictions, sector-specific investment strategies), tech company leadership (strategic timing of IPO based on market conditions, valuation benchmarking against comparable companies, capital structure optimization pre-IPO), market analysts (quantitative validation of IPO valuations, sector-wide trend analysis, performance forecasting with high confidence), and investment strategists (portfolio allocation decisions, risk management frameworks, arbitrage opportunity identification).

The diverse frameworks across tech verticals reveal distinct valuation patterns and metrics that drive market capitalization predictions. In SaaS/Cloud companies, revenue metrics capture recurring revenue stability (log revenue 0.082), cohort expansion rates (revenue^2 0.077), and revenue retention dynamics (base revenue 0.075 core growth metrics). Their operating expense ratio (0.041) reveals sales efficiency metrics, R&D investment effectiveness, and customer success scalability. In E-commerce platforms, asset utilization metrics track platform transaction velocity, marketplace liquidity, and inventory turnover efficiency (Total assets 0.083 platform scale, current assets 0.045 inventory management, cash position 0.052 growth funding). Revenue^2 (0.077) models network effect acceleration (GMV growth), cross-border scaling, and category expansion dynamics; FCF^2 0.038 market liquidity, net income^2 0.023 profitability scaling. For FinTech Ventures, liability metrics evaluate capital adequacy ratios, risk-weighted assets, and regulatory compliance capacity; cash position indicates float management efficiency, working capital optimization and treasury operations scale.

Asset importance in Deep Tech/AI companies captures IP portfolio value, data asset accumulation, and infrastructure scaling efficiency with R&D capitalization shown through operating expense ratios, asset utilization metrics, and technology stack value. Marketplace platforms use revenue transformations model for network effect scaling, take rate optimization, and cross-side network values, while operating metrics track their user acquisition efficiency, platform contribution margins, and marketplace health metrics.

VII. Further Research and Future Applications

The developed Random Forest regression model for predicting Tech Startup IPO market capitalization presents numerous opportunities for expansion and practical applications. The model can be enhanced by implementing real-time valuation updates that incorporate dynamic market data, allowing for more responsive and accurate predictions. Integration of market sentiment analysis through social media

feeds, news articles, and investor forums would add valuable psychological dimensions to the predictions.

The model's framework can be extended beyond tech startups to cover other sectors, creating a comprehensive IPO valuation tool across industries. Further feature engineering opportunities exist in incorporating additional variables such as founder experience metrics, patent portfolios, and competitive landscape indicators. These advancements would deliver strategic benefits including more informed investment decisions, sophisticated risk management capabilities, optimized market timing, and deeper competitive intelligence. Financial institutions, venture capital firms, and investment banks can leverage this enhanced model to gain a competitive edge in the IPO market, ultimately leading to better-calibrated valuations and investment strategies.

The current Random Forest model can be significantly enhanced by incorporating Bayesian optimization techniques, which would enable more sophisticated parameter tuning and uncertainty quantification in IPO valuations. Bayesian methods excel at handling complex probability distributions and can provide valuable confidence intervals for predictions, making them particularly suitable for the volatile nature of IPO markets. Future iterations could explore ensemble approaches combining Neural Networks, capturing different aspects of the IPO valuation dynamics.

Deep learning architectures, particularly transformers with attention mechanisms, could be implemented to detect subtle patterns in historical IPO data and market conditions. These advanced modeling techniques would work synergistically with the existing Random Forest framework to create a more robust and comprehensive prediction system. The integration of these sophisticated models would enable better capture of non-linear relationships and temporal dependencies in IPO valuations, leading to more nuanced and accurate market capitalization predictions. This multi-model approach would be especially valuable for stakeholders requiring different levels of prediction granularity and risk assessment in their investment strategies.