Fall 2025 15.572 A-Lab Project Report: Forecasting NAV for Private Equity Funds

César Dori, Giuseppe Iannone, Shuyan Feng, Yash Gupta
{cesar_d, pepegius, shuyanf, ygupta}@mit.edu
Massachusetts Institute of Technology, Cambridge, MA

1 Executive Summary

The Challenge: *Bridging the Information Gap in Retail Private Equity*
Fidelity Investments aims to democratize access to Private Equity (PE) by offering buyout funds to retail investors via an exchange-traded fund (ETF). However, a critical regulatory disconnect poses a challenge for this initiative: the ETF must report its Net Asset Value (NAV) within 20 days of the quarter's end, while underlying PE funds often take 45 days or more to report their valuations. The objective of this project is to develop a forecasting engine capable of predicting NAVs during this "blind" period, ensuring regulatory compliance and fair pricing for retail investors.

Methodology
Leveraging historical performance and cashflow data from StepStone - which covers US Buyout funds - and macroeconomic indicators from FRED, the team benchmarked an array of predictive models: traditional econometric (OLS) and ML (Random Forest, XGBoost, and Neural Networks) techniques to determine the optimal balance between accuracy, robustness, and explainability. The chosen OLS model is refined with domain-specific data cleaning and feature transformations.

Key Insights and Results
The log-transformed OLS model achieves an MAPE of 10.6% for 90% of the out-of-sample datapoints, meeting Fidelity's standards for the ETF launch. Additionally, as shown in Figure 1, the tail (large mistakes in fund valuation) is bounded by 30% for 90% of the instances. We note that the simple OLS model outperforms the more complex ML models and shows signficantly lower susceptibility to overfiting.

Business Impact
We propose deploying a log-transformed OLS model, incorporating lagged NAVs, current quarter cash flows, fund age and size, and macro-economic variables. The model not only meets the required performance thresholds but also provides the economic interpretability required by regulators — a transparency advantage that "black box" deep learning seldom offers.

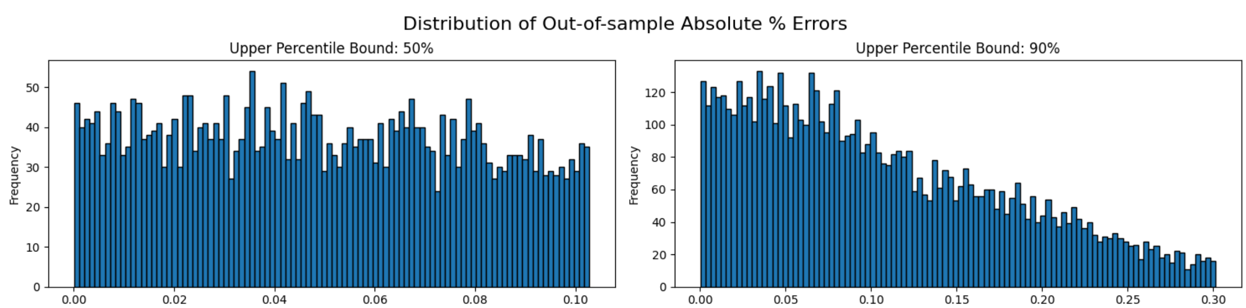

Figure 1: Distribution of Absolute Out-of-Sample (Test) Relative Errors

2 Background and Motivation

2.1 About Fidelity

Founded by Edward C Johnson II in 1946 and headquartered in Boston, Fidelity Investments is among the top 3 asset managers across the globe (trailing only Blackrock and Vanguard), with USD 5.8 trillion in discretionary assets under management (AUM) and USD 15.1 trillion in assets under administration [1]. Since its establishment, Fidelity has expanded to offer a wide array of financial services, including retirement and health benefit programs for employers, employee stock options, brokerage services, investment products (including stocks, bonds, mutual funds, ETFs, options, annuities, and money market funds), robo-investing services, life event planning, and student debt management.

For the present study, Fidelity has partnered with MIT A-Lab to devise an appropriate forecasting technique to predict NAV for its latest private equity fund offerings for retail investers.

2.2 Private Equity Funds and NAV

Private equity refers to investment in privately owned companies or in public companies intended to be taken private. Estimates by the Apollo suggest US private companies vastly outnumber the public ones, with about 90% of the businesses being private. Even for businesses with over $100 million in revenue, more than 85% choose to remain private [2]. Private equity markets are fundamentally different from public equity markets in that they generally have lower liquidity and high transaction costs (due in large part to the absence of stock exchanges), require the participants to be sophisticated or have a minimum critical net worth (much like hedge funds), and are categorized by an exit strategy (for example, initial public offering or IPO, trade sale, or secondary sales) than an infinite holding period.

There are three primary categories of private equity funds [3]:

- **Leverage Buyouts** (LBOs): These funds acquire public companies or established private companies with a significant percentage of the purchase price financed through debt, which becomes part of the target company's balance sheet after the acquisition. LBO managers seek to add value by improving the company operations, boosting revenues, and increasing earnings and cash flows. Some major fund managers include Blackstone, KKR, and Bain Capital.

- **Venture Capital** (VC): VC funds provide financing to private companies, usually startups, with high growth potential. They inject money at the concept creation or mezzanine stage to exit via the initial public offering or IPO. Some major players include Andreesen Horowitz (a16z), Sequoia Capital, and Bessemer Venture Partners.

- **Growth Capital**: The funds make minor equity investments (less than the controlling interest) in more mature companies looking for capital to expand, restructure operations, enter new markets, or finance major acquisitions. Some major players are Viking Global Investors, Silver Lake, and Ares Management.

The primary focus of this study will be buyout funds. A key metric to measure the fund health is its net asset value, or NAV, define as

$$NAV = CF_0 + \frac{CF_1}{1+r} + \frac{CF_2}{(1+r)^2} + ... + \frac{CF_N}{(1+r)^N}$$

Here, CF_i denotes a cashflow i periods (typically quarters or years) from now and r denotes the appropriate discount rate (typically cost of capital). The initial and periodic contributions by the fund investors are examples of cash inflows ($CF_t < 0$), while distributions in form of periodic dividends or the final disbursement are examples of cash outflows ($CF_t > 0$). In general, an accurate estimation of these (future) cash flows as well as the discount rate is quite difficult and subjective, making it hard to forecast NAV of a fund for an external investor. Finally, much akin to the market capitalization of a publicly traded firm ($MC = SharePrice \times \#(Stocks)$), the NAV provides a point estimate of the fund's current value.

2.3 Motivation and Problem Statement

Traditionally, only sophisticated market participants (individuals or institutions) with a certain minimum net worth are eligible to invest in private equity funds, resulting in lesser regulatory oversight and restrictions by the government. However, Fidelity has recently taken up the challenge to make available these private equity funds available to the retail investor through pooled ETFs (exchange-traded funds) as a "40 Act Fund". Under the US Investment Company Act of 1940, mutual funds, ETFs, and closed-end funds are subject to certain disclosures, governance guidelines, and diversification restrictions to protect the investors by providing them transparency and minimizing the conflicts of interest [4].

Such funds must make their valuation disclosures within 20 days of each fiscal quarter. However, the private equity funds underlying the ETF are not subject to such restrictions, and hence may report their valuations (such as NAV) several weeks (typically 45 days or more) into the next quarter. This necessitates the use of a forecasting model to predict the NAV of the underlying funds and hence calculate the ETF value to be reported (an average of the underlying fund NAVs weighted by the amounts invested in each).

More specifically, the problem is to forecast the NAV of a particular US buyout fund at the end of a specific quarter given the data on the prior history of the fund (its previous NAVs and cash flows, for example), static information about it (the fund manager, the fund target size, and its industry or sector focus, for example), and public financial and economic data.

3 Data

3.1 Data Sources

Unlike publicly traded markets such as public equities, commodities, and options, private equity markets suffer from the lack of readily available performance and operations data. Most of the data is typically self-reported and aggregated by data vendors. We draw the data for this study from two such sources:

- ***Preqin:*** Available through WRDS, Preqin (part of Blackrock) provides data on VC and Buyout funds (besides data on real estate, infrastructure, natural resources, and hedge funds) [5]. We use the Fund Performance, Cash Flow, and Fund Details data, available from 1983 Q1 to 2024 Q4. Fund performance reports the key performance indicates like NAV, internal rate of return (IRR), benchmarks used, RVPI, and valuation multiple for each quarter (relevant features discussed in more detail in the following section). Fund Details provides static information like vintage, fund type (for example, Buyout), local currency, target size, as well as the fund manager. Finally, Cash Flows provides information on ongoing transactions (cash inflows and outflows along with the associated amount) throughout the fund's recorded

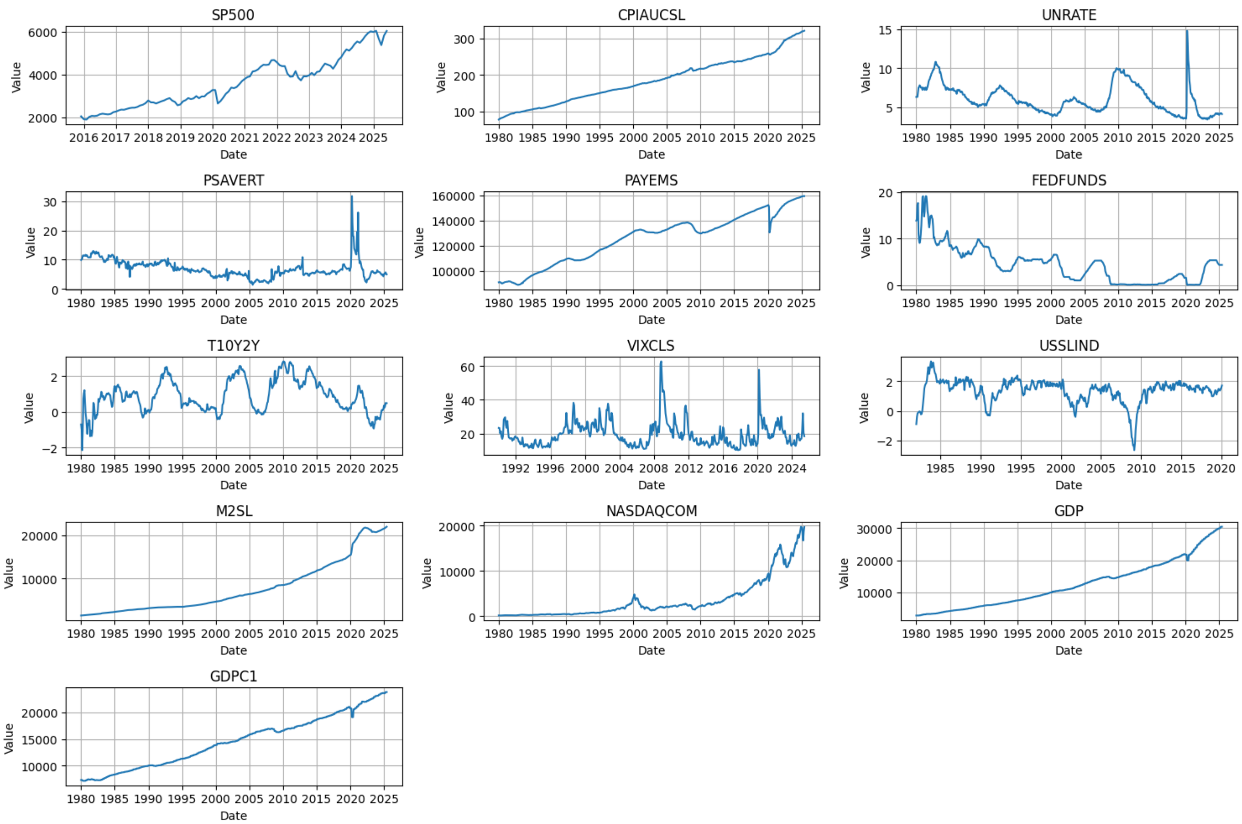

Figure 2: Time Series of Macro-economic Data

history. Preqin data has been used for initial experiments before moving to the more accurate Stepstone data, as instructed by the sponsor.

- ***Stepstone:*** Made available by Fidelity, Stepstone tracks nearly \$ 389 B of private equity funds for the period 1991 Q1 to 2025 Q2 [6]. Like Preqin, static details on the fund as well as the fund operations (cash transactions) and performance metrics are available through Stepstone. This forms the basis for our final reported model.

The private markets data is supplemented by information on macro-economic variables to gauge the overall health of the economy as well as the investors' appetite for and general performance of new investments. The macro data comes from the repository, ***Federal Reserve Economic Data (FRED)***, maintained by Federal Reserve Bank of St. Louis for providing better insights into Fed's policies [7]. We draw quarterly data from 1991 to 2025 for the S&P 500 index (SP500), consumer price index (CPIAUCSL), unemployment rate (UNRATE), personal savings rate (PSAVERT), population on payroll (PAYEMS), federal funds effective rate (FEDFUNDS), 10-year versus 2-year interest rate (T10Y2Y), forward looking volatility of S&P 500 (VIXCLS), a leading indicator of the economy (USSLIND), money supply (M2SL), NASDAQ stock index (NASDAQCOM), nominal GDP (GDP), and real GDP (GPDC1). Figure 2 shows the time series for the macro-economic features over the relevant period.

3.2 Preparing and Partitioning the Data

We focus here on the processing of the StepStone data, which constituted the primary dataset for this analysis. A similar logic was applied to the Preqin data for our preliminary analysis and experiments.

The raw data consisted of two distinct dataframes. The first contained static **Fund Details**, such as the Fund Manager, Vintage Year (launch year), Fund Country, Fund Size Range, and SPI Fund Sector (e.g., Growth Equity, Buyout). The second dataframe was a **Cashflow** ledger, reporting time-series data for every contribution, distribution, or valuation event (Net Asset Value, or NAV) recorded for each fund.

To construct the final dataset for modeling, we performed the following preprocessing steps:

1. Cumulative Aggregation

To capture the historical performance context of a fund at any given point in time, we transformed the raw transactional data into cumulative metrics. The data was sorted by *Fund Name* and *Date*, and we calculated the running cumulative sum for contributions, distributions, and net cash flows for each fund. This transformation ensures that every data point reflects the total capital called or returned from the fund's inception up to that specific date. Cash flows for the quarter can be estimated as the change in these cumulative metrics since the last quarter.

2. Filtering for Valuation Events

The dataset was subsequently filtered to retain only the rows where a valid *Valuation* (NAV) was reported. This reduced the dataset to a series of quarterly snapshots, where each row represents the state of a fund (cumulative cash flows and current NAV) at a specific valuation date.

3. Integration of Fund Details and Macroeconomic Data

To construct the final analytical dataset, we performed a two-stage enrichment process. First, we merged the dynamic valuation data with static fund characteristics—specifically *Vintage Year*, *Fund Size*, *Sector* (if available), and *Country*—using the *Fund Name* as the linkage key. To align the dataset with the project's specific scope, we applied a strict filter to retain only funds classified as **Buyout** sector funds located in the **United States**. Following this selection, the categorical columns used for filtering were removed, along with the *Net Cash Flow* column, to reduce dimensionality and eliminate redundancy.

Subsequently, we integrated macroeconomic indicators into the dataset. Given that fund valuations and macroeconomic metrics operate on different reporting frequencies, we utilized a "backward" merge strategy. For every valuation date in the fund dataset, we matched the most recent available macroeconomic data point. This approach prevents look-ahead bias by ensuring the model utilizes only the macro information available at the precise moment of the valuation. Any remaining temporal gaps were filled using forward propagation to ensure consistency.

4. Train-Validation-Test Partitioning

Due to the temporal nature of the fund performance and cash flows, the data can not be partitioned randomly. Instead, it has to be segmented along the time axis, which aligns with the ultimate goal of using the patterns learned in the past data for model use in the future. Therefore, we keep everything before 31st December, 2018 in our training partition, datapoints located afterwards but still before 31st December, 2021 as our validation or development partition, and everything that follows as our final out-of-sample testing partition. This achieves an approximate 3 : 1 : 1 split in terms of the number of datapoints, considering the fact that the data is more sparse in the earliest years (attributable to fewer funds reporting and smaller size of the PE investments industry).

3.3 Data Characteristics

The data characteristics in Table 1 provide context for many of the decisions taken in this project. We had 30 years of historical data for nearly 1,000 funds managed by around 2,000 managers. The highly skewed NAV distribution, with a small number of very large funds ($\$27.7bn$), is an observation that motivated the use of log-transformations and scaled error metrics discussed in section 4.3. Although NAV should in principle be non-negative, the raw dataset contains a small number of non-positive entries. These observations are treated as data anomalies. Finally, the time-based train/validation/test split ensures that reported performance reflects true out-of-sample forecasting ability.

Table 1: Summary of Data Characteristics

Category	**Metric**	**Value**
Funds	Number of funds	998
	Number of managers	2,136
NAV	Range (min–max)	\$0 – \$27.7 bn
	Median	\$354.7 m
	Skewness	4.37
NAV quality	Observations with NAV ≤ 0	1,053 (2.5% of raw)
	Funds with negative NAV	289
	Negative NAV (min / mean)	−\$519.4m/ − \$24.0m
Cash flows	Contributions (mean / median)	\$1.84 bn / \$0.79 bn
	Distributions (mean / median)	\$1.55 bn / \$0.39 bn
Train set	Observations (share; dates)	24,097 (63.9%; 1991-12-31–2018-12-31)
Val set	Observations (share; dates)	5,918 (15.7%; 2019-01-31–2021-12-31)
Test set	Observations (share; dates)	7,474 (19.8%; 2022-01-31–2025-06-30)

4 Modeling

The present section describes the results using the Stepstone data (as required by the sponsors), while alluding to some preliminary experiments with the Preqin data for arriving at the correct form of features and model designs.

4.1 Feature Engineering

The initial experiments with the Preqin data led us to the following set of features:

- **Contribution** and **Distribution**: The NAV of a fund grows with cash inflows (contributions) and shrinks with cash outflows (distributions) linearly (disregarding the time value of money effects). These cash flows over the most recent quarter are included as numeric features in the model.

- **Lagged NAV**: The NAV at the end of any particular quarter is unlikely to change significantly from the NAV of the previous quarter (any cash flows and macroeconomic effects are typically small). In fact, we observe correlation higher than 90% between the present and lagged values of NAV for the same fund.

- **Fund Size**: We assign dummies to bin the target size of the fund into several classes: less than \$ 500 M, between \$ 500 M and \$ 1 B, between \$ 1 B and \$ 2 B, between \$ 2 B and \$ 5 B, between \$ 5 B and \$ 7 B, and over \$ 7 B. AUM or fund size is known to impact its performance in both the hedge fund and private equity sphere, since larger funds are able to harness the economies of scale (such as in attracting the best talent) and make riskier investments with buffer to absorb drawdowns and investor clawbacks.

- **Age**: Newer funds are, in general, riskier and more likely to fluctuate from the NAV in the previous quarter given that they are still in growth phase. We calculate the age in years as: $age = \frac{curr\ period - vintage}{365.25}$.

- **Macroeconomic Variables**: Barring S&P 500 (which was dropped for high correlation with other macro variables as well as data availability issues), All macroeconomic features plotted in Figure 2 have been used in the subsequent models. It should be noted that given the trend in the time series (save for FEDFUNDS, VIXCLS, and T10Y2Y), the the time series of the relative changes ($\frac{x_t - x_{t-1}}{x_{t-1}}$) have been considered instead.

For the Preqin data, indicator (0/1) features for the industry focus (such as, pharmaceuticals, technology, and semiconductors) of the funds were also considered. However, they were dropped for little to negative impact on the model, coupled with unavailability with Stepstone data.

4.2 Preliminary Modeling

Given the approximate linear dependence of NAV on cash flows and past values (for small discount rates) as well as on the fund size fixed effects, the simplest candidate model is an ***ordinary least squares regression*** model:

$$NAV_{t+1} = \alpha + \beta_1 \cdot NAV_t + \beta_2 \cdot Contr_t + \beta_3 \cdot Distr_t + \beta_4 \cdot Age + \vec{\gamma}^T \vec{x} + \theta_{size} + \epsilon$$

where $\vec{x}$ denotes the macro variables and θ_{size} the appropriate fund size effect.

It may be noted that potential candidate features like industry focus, internal rate of return (IRR), fund DPI, fund RVPI, and capital called were eliminated to mitigate the problem of multicollinearity (increasing precision) through Lasso regression - progressively increasing the $L1$ penalty and observing an elbow in validation MSE.

To account for potential non-linear interactions, the following alternative ML models were also considered:

- ***Random Forest***: 200 decision tree estimators with a maximum of $log_2(K)$ features each (K being the total number of features) were used.

- ***XGBoost***: 200 estimators with a maximum depth of 20 each were used.

- ***Neural Network***: A fully-connected network with two hidden layers (64 and 32 ReLU-activated neurons respectively) and an intermediate dropout layer (with dropout probability of 25%) was used. The network was trained to minimize the training MSE with a learning rate of 10^{-3} using Adam optimizer.

We compare our models against the benchmark using the approximate arithmetic relationship (does not hold for significant discount rates and absence of all relevant cash flows): $NAV_{t+1} = NAV_t + Contributions_t - Distributions_t$.

Table 2 shows the performance of the model on training and validation partitions based on the fraction of the variance explained ($R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$) and the mean-squared percentage error ($MSPE = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\bar{y}^2}$). It is evident that only OLS achieves a balance between a strong training and validation performance. Coupled with sponsor's need for an economically interpretable and parsimonious model, OLS becomes our model of choice for subsequent refinements.

Partition	**Metric**	OLS	Random Forest	XGBoost	NN	Benchmark
Training	R^2	0.984	0.994	0.973	0.988	0.576
	MSPE (%)	1.305	1.131	2.131	1.293	27.153
Validation	R^2	0.971	0.920	0.887	0.905	0.731
	MSPE (%)	2.395	6.452	9.047	8.203	18.987

Table 2: Comparison of Performance of Preliminary Models on Training and Validation Data

4.3 Model Refinements

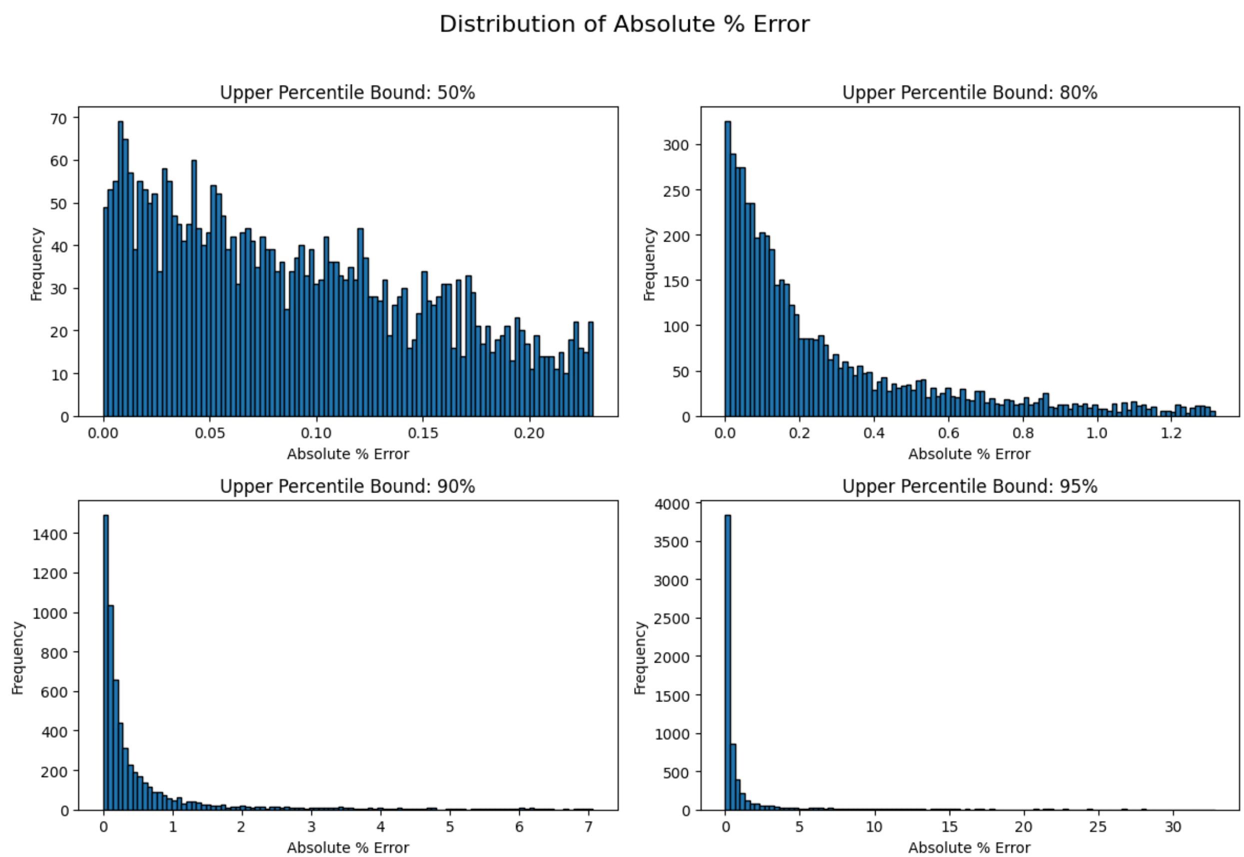

Figure 3: Preliminary Model: Distribution of Absolute Validation Errors

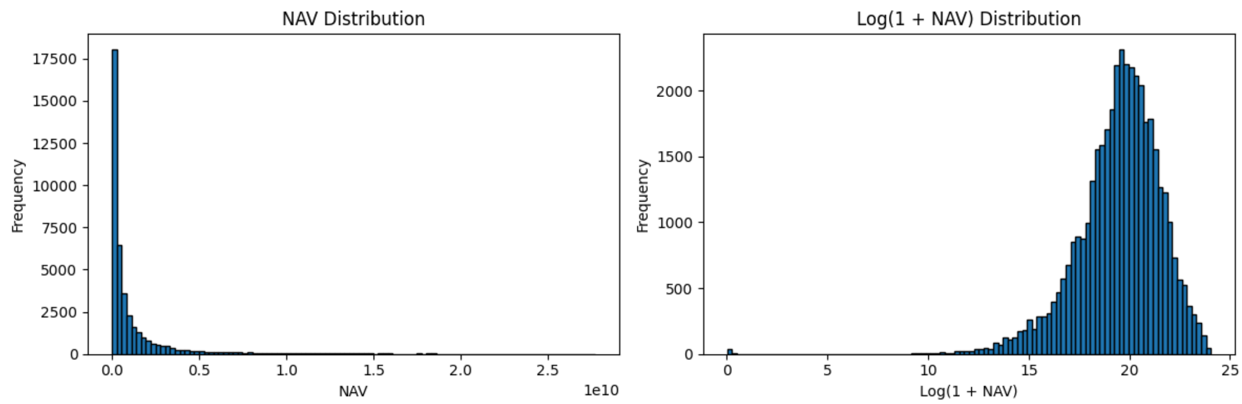

Figure 4: Distribution of Raw and Transformed NAVs

Figure 3 illustrates how a "high" validation R^2 of more than 0.97 can actually be very misleading. It plots the bottom 50, 80, 90 and 95 percentiles of the absolute percentage (or relative) errors ($APE = \left|\frac{y_i - \hat{y}_i}{y_i}\right|$) on the validation partition for the OLS predictions. It is easy to observe that:

- The lower 90% or 95% of the absolute errors (lower-left and lower-right subplots) have the highly right-skewed distribution; that is, there are cases where the model makes very significant mistakes and hence should not be trusted or used.
- Generally a mean absolute percentage error upto 10% in forecasting the valuation of approximately 90% of the constituents is considered acceptable in the ETF industry. Here, only the bottom 50%ile of the errors satisfy that criterion ($MAPE = 9.04\%$). For the entire bottom 90%ile of the errors, we observe a large 108.4% MAPE.

These issues prompt us to make the following refinements to the modeling pipeline:

- A few records with negative NAV are observed; however, a fund value can not dip below zero (it will be dissolved before reaching that point, in practice). After thorough domain review and discussion with the Fidelity team, it was concluded that those records are likely erroneous values and are dropped from the analysis.
- As shown in Figure 4, the distribution of NAVs is very right-skewed, prompting the application of a log-transformation - $L_t = ln(L_t + 1)$ - resulting in the following model:

  $$L_{t+1} = \alpha + \beta_1 \cdot L_t + \beta_2 \cdot Contr_t + \beta_3 \cdot Distr_t + \beta_4 \cdot Age + \vec{\gamma}^T \vec{x} + \theta_{size} + \epsilon$$

  It may be noted that the subsequent absolute error values are still on the correct scale since they are calculated as $\left|\frac{e^{L_t} - e^{\hat{L}_t}}{e^{L_t} - 1}\right| = \left|\frac{y_t - \hat{y}_t}{y_t}\right|$.
- There may be longer term dependencies present in the data besides just the immediate past lag. In fact, the regression $L_t = \alpha + \beta_1 L_{t-1} + \beta_2 L_{t-2} + \epsilon$ results in a statistically significant non-zero coefficient on L_{t-2} ($|t| = 11.9 > 2$). To incorporate the effect of past lags, we reformulate the OLS model as

  $$L_{t+1} = \alpha + \beta_1 \cdot L_t + \beta_2 \cdot L_{t-1} + \beta_3 \cdot \bar{L}_{t-1:\,t-4}$$
  $$+\beta_4 \cdot Contr_t + \beta_5 \cdot Distr_t + \beta_6 \cdot Age + \vec{\gamma}^T \vec{x} + \theta_{size} + \epsilon$$

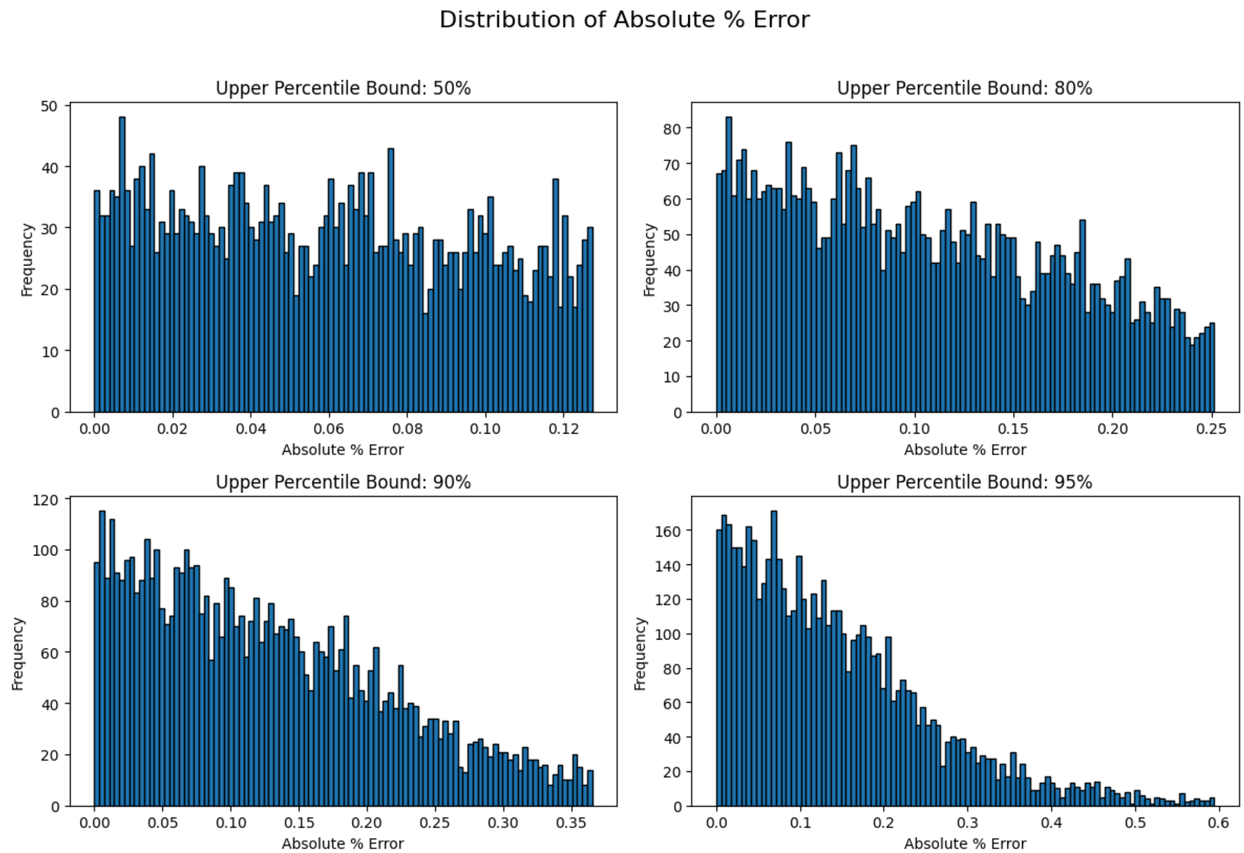

Figure 5: Refined Model: Distribution of Absolute Validation Errors

Here, $\bar{L}_{t-1:\,t-4}$ represents the moving average over four lags. We avoid too many null values by allowing a minimum rolling window of just one lag.

- While the relationship $NAV_{t+1} = NAV_t + Contribution_t - Distribution_t$ does not hold exactly, severe deviations from it represent anomalies rather than noise. Examples include huge market shifts, failed investments, last-minute (unreported yet) contributions and distributions, and recording errors. Therefore, we remove datapoints with such large deviation from the training partition to avoid learning from them, while still retaining them in the validation and test partitions since the real data would have such deviations. Specifically, if $Divergence = \left|\frac{NAV_t + Contr_t - Distr_t - NAV_{t+1}}{NAV_{t+1}}\right|$ denotes the relative divergence from the linear relationship for a particular data point, the records with the top 1% divergence are dropped from the training data.

Figure 5 shows the distribution of absolute errors on the validation partition after making these changes (incrementally). It is easy to note that both the number and magnitude of extreme right outliers have been reduced substantially. Furthermore, now 50% of the errors are within approximately $10 - 12\%$ and 90% of the errors are within 35%. The following section reports further details on the performance of the refined model (proposed) across the data.

4.4 Results with the Proposed Model

Table 3 provides the MSPE, R^2 and MAPE (for different percentiles of the absolute errors) across the three partitions of the data. The similar performance (fit and errors) across the the partitions indicates that overfit, if any, is not severe. Additionally, it is encouraging that the MAPE for 90% of the test datapoints is close to 10% (desired).

Metric	**Train**	**Validation**	**Test**
Support	24341	5918	7474
$R^2(\%)$	98.00	96.90	96.17
MSPE (%)	1.504	2.453	2.875
MAPE (%): Bottom 50%ile	5.228	5.972	4.927
MAPE (%): Bottom 80%ile	9.193	10.546	8.811
MAPE (%): Bottom 90%ile	11.181	12.709	10.609
MAPE (%): Bottom 95%ile	12.736	14.398	11.946

Table 3: Perforamnce of the Proposed OLS Model

Figure 6 shows predictions for a sample fund (ID: 1-1197) throughout its available history (2019-25). It may be observed that the fit remains close to the actual values, trailing its shape (attributable to reliance on the past NAV) and the errors remain close to zero, exceeding 10% only in exceptional quarters (like the drawdown in 2023 Q2).

Figure 6: Predictions across time for a sample fund

5 Further Discussion

5.1 Fund Manager Fixed Effects

Through discussions with our sponsor and a closer look at our data, it became apparent that MAPE varied systematically between fund managers. In particular, some managers exhibited consistently higher (or lower) prediction errors across their funds, suggesting that part of the residual variation might be driven by manager-specific effects not captured by our existing features.

The model was extended to account for these effects via a set of manager dummy variables (with one manager omitted to act as the intercept: $NAV_{it} = \alpha + X_{it}^{\top}\beta + \gamma_{m(i)}$ where α is a common intercept, β are coefficients on fund and market level features, $\gamma_{m(i)}$ is a manager specific fixed effect.

Given that our dataset contains more than 350 managers, many of whom oversee only one or two funds, adding the dummy variables increases the model's complexity without providing additional predictive power. Figure 7 shows that this extension to the model mostly captures noise and increases the absolute percentage error compared to the proposed model.

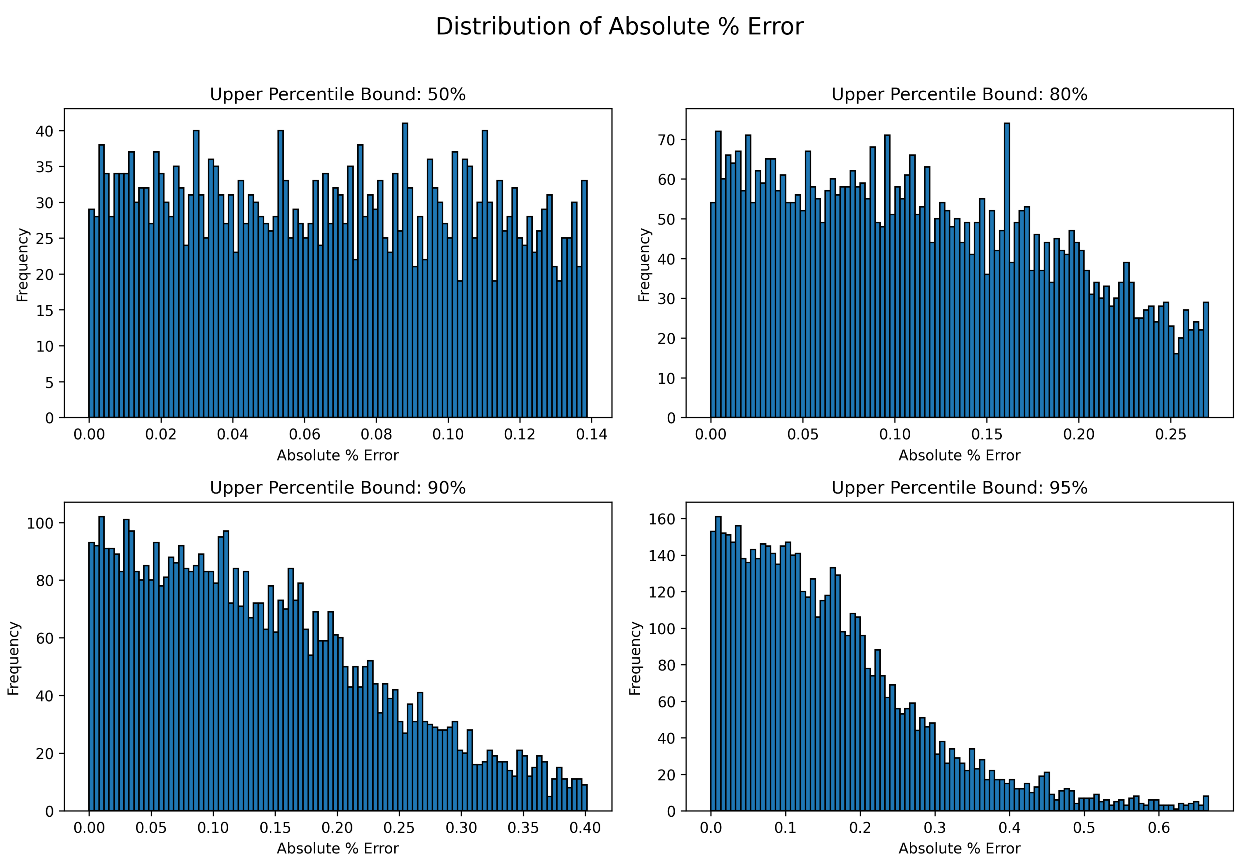

Figure 7: Refined Model with Fund Manager Effects: Distribution of Absolute Validation Errors

5.2 Dependence on Public Markets

For publicly traded equities markets, it is well-known that stock returns exhibit high correlation with their corresponding sector ETF returns. For example, the stock for Google, Inc (ticker:

GOOG) is known to be 76% correlated with the technology ETF Invesco QQQ (ticker: QQQ) [8]. Hence, if the ETF returns are known in advance (not the case for public markets), they can serve as a strong predictor of the returns of the related stocks.

Since the investments made by private equity funds are also affected by the performance of the companies invested in (for example, optimism about AI and technology can yield positive returns for Sequoia's technology-focused funds), it is plausible to consider the related sector ETFs as a candidate feature. To get an idea of the predictive power of the ETF returns, we measure the correlation of the private equity fund returns (defined as percentage change in NAV) for a particular sector (pharmaceuticals, for instance) against the corresponding ETFs (in this case, VanEck PPH).

Private Equity Industry	Technology	Pharmaceuticals	Finance
ETF	VanEck SMH	VanEck PPH	VanEck BLZD
Correlation (%)	1.37	0.33	-2.58

Table 4: Correlation between Private Equity Fund Returns and Related ETF Returns

Table 4 reports the sample correlations for three major ETF-covered industries based on the Preqin data. The observed correlations have significantly lower magnitudes than those generally found for publicly-traded stocks [8]. Coupled with the fact that Stepstone does not provide the required information to associate a sector with a fund, sector-wise ETF returns are dropped as a candidate feature.

6 Conclusion and Future Work

The study covered how past data on private equity fund valuation and cash flow operations, coupled with fund characteristics and information on macroeconomic climate, may be leveraged to forecast future fund valuations, or NAV. An interesting theme highlighted in the study was that a state-of-the-art, complex model might not always be the optimal choice. We saw how a parsimonious ordinary least-squares regression model delivers consistent performance across in- and out-of-sample data, whereas more powerful models like XGBoost and Neural Networks overfit to noise.

Additionally, the study underscored the importance of detailed error analysis. Aggregate performance metrics like R^2 are often highly misleading. Therefore, we shifted our focus towards minimizing the tail risk than maximizing the average fit. We saw the importance of thorough data cleaning in this exercise - removing noisy and anomalous data points from the training data affected the final performance more than the model choice.

Our proposed regression model achieved an out-of-sample MAPE of 10.6% for 90% of the datapoints, indicating strong utility for Fidelity's upcoming 40 Act fund. However, we note the following limitations: (1) the model is significantly reliant on the past values of NAV and loses predictive power if the past history is not available in time, (2) since Stepstone does not map fund IDs to actual fund names, the model is unable to leverage manager-specific information (such as skill or investment-style of KKR), and (3) the model is still unable to forecast one-off events like a heavy cash inflow due to a new investor joining the fund.

The private equity industry is constantly evolving, necessitating the perpetual model improvements. We note the following potential avenues for improving the forecasting performance and reliability: (1) NLP information about fund managers and from fund disclosures may be used to extract sentiment on future fund outcomes as well as about fund's general plan on investments and cash distribution, (2) a general model may be adapted to Fidelity's specific family of invested funds through fine-tuning (similar to transfer learning), and (3) it may be valuable to consider alternative non-linear models such as CNNs for automatic feature engineering and LSTMs to capture temporal dependencies in the data.

7 Acknowledgements

The team is sincerely thankful to Ben Song and Jin Jin from Fidelity for their inputs and guidance throughout the project. We really appreciate the team mentor, Sandhya Bhaskar's, assistance around the technical queries and best practice in data science. Finally, the team is grateful for the amazing opportunity arranged by the A-lab team - including Prof. Almaatouq, Prof. Aral, and the course TAs - as well as the insightful lecture series and class discussions.

References

[1] "Britannica Money," Britannica.com, Aug. 14, 2025, https://www.britannica.com/money/Fidelity-Investments [Accessed: December, 2025].

[2] Torsten Sløk, "Most of the US Economy Is in Private Markets - Apollo Academy," Apollo Academy, Apr. 30, 2024. https://www.apolloacademy.com/most-of-the-us-economy-is-in-private-markets/ [Accessed: December, 2025].

[3] *Fixed Income, Derivatives, Alternative Investments, Portfolio Management*, CFA Level 1 Curriculum, Wiley, 2023.

[4] "Introduction and Overview of 40 Act Liquid Alternative Funds," *Citi Prime Finance*, 2013.

[5] "WRDS Overview of Preqin," https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/preqin/ [Accessed: December 2025].

[6] "Stepstone: What We Do," https://www.stepstonegroup.com/what-we-do/asset-classes/private-equity/ [Accessed: December 2025].

[7] "What is FRED," https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/ [Accessed: December 2025].

[8] "GOOG vs. QQQ," https://portfolioslab.com/tools/stock-comparison/GOOG/QQQ/ [Accessed: December 2025].