Strategic Portfolio Optimization with ML:

Harnessing Below Medium-Cap Stock Characteristics for Enhanced Returns

Applying Advanced ML Techniques to Develop Constrained Investment Strategies & Outperform the S&P 500 Benchmark

Contributors[1]
Meriem Mehri
Adrian Alarcon Delgado
Joshua Poozhikala
Sheida Majidi
Zhicheng Zhong
Yifan Lu
Zhiming Zhang

[1]McGill University, Desautels Faculty

**

Presented to:

Prof. Russ Goyenko & Chengyu Zhang

Summer 2024

Note: This report comprises **10 pages**, including **4 pages** dedicated to the Appendix.

Executive Summary

In this report, we present a comprehensive investment strategy that leverages advanced machine learning techniques to analyze and predict stock returns effectively. Our chosen strategy is a mixed Long-Short portfolio approach, focusing primarily on medium-cap U.S. stocks from 2005 to 2023. This strategy is underpinned by the application of ensemble machine learning models, specifically combining autoencoders and gradient boosting methods to capitalize on the rich, characteristic-based dataset provided.

Guiding objectives: The primary objective of our analysis was to outperform the benchmark S&P 500 index (SPY) by utilizing the predictive power of our machine learning model on the future returns of these stocks. The initial training phase covered the first ten years of data, followed by a two-year validation period, and a final year of out-of-sample testing to assess real-world applicability.

Highlights: The results from our out-of-sample testing were highly encouraging, demonstrating the efficacy of our machine learning-driven investment approach. The portfolio achieved a significant Alpha, indicating that our strategy successfully generated excess returns over the SPY benchmark. Additionally, the annualized Sharpe ratio reflected a favorable risk-adjusted performance. Other key performance metrics, such as robust average returns, controlled volatility, and a maximum drawdown within predefined limits, further underscored the strength of our strategy. These outcomes not only validate the feasibility of applying machine learning techniques in a realistic trading environment but also highlight their potential to deliver competitive returns.

***

I. Introduction

Strategy description: In the pursuit of constructing a robust investment strategy capable of outperforming market benchmarks, our approach integrates a mixed Long-Short strategy. This decision is grounded in the objective to capitalize on both upward and downward market movements, thereby enhancing potential returns and mitigating risks associated with market volatility. The Long-Short strategy allows for greater flexibility compared to a Long only strategy by enabling profit opportunities from stocks expected to decrease in value, which is particularly advantageous during market downturns or periods of high volatility.

The mixed strategy component further includes a combination of quantitative models to exploit inefficiencies in medium-cap stocks, which are often less closely followed by market analysts and, therefore, more likely to deviate from their intrinsic values. By employing machine learning algorithms that analyze vast datasets to detect these inefficiencies, our strategy is designed to achieve superior risk-adjusted returns, providing a hedge against market corrections, and enhancing portfolio diversification.

Stock focus: The focal point of our investment strategy is on medium-cap stocks. This choice is driven by several strategic and analytical considerations. Medium-cap stocks offer a balance between the high growth potential commonly associated with small-cap stocks and the stability found in large-cap stocks, making them an attractive segment for applying machine learning predictive models. These stocks often exhibit sufficient market liquidity to avoid the trading challenges associated with small caps, yet still provide enough inefficiencies for alpha generation, which our machine learning models are designed to exploit.

Furthermore, medium-cap stocks are less frequently covered by major financial analysts and less targeted by large institutional investors, presenting unique opportunities for alpha that are not as readily available in the more scrutinized large-cap arena. This under coverage leads to less competitive pressures and a richer ground for the deployment of sophisticated analytical techniques like those developed in our course. By targeting this segment, our strategy

leverages the predictive power of machine learning to uncover and capitalize on undervalued prospects and to short, overvalued stocks, hence maximizing the potential for high returns at controlled risk levels.

II. Data Overview

Our dataset comprises monthly data for below medium-cap US publicly traded companies spanning from 2005 to 2023. Each stock entry includes 145 characteristics, such as financial ratios, market data, and trading volumes, which serve as predictors for our models. These characteristics are critical in predicting the subsequent month's stock returns, with all predictor variables lagged by one month to ensure a realistic forecasting scenario.

In terms of preprocessing, the dataset underwent several steps to prepare it for analysis. Initial steps included handling missing values, normalizing data to reduce variance across different scales, and encoding categorical variables where applicable. We utilized Python's Pandas and NumPy libraries extensively for data manipulation, ensuring that the data structure was optimized for high-efficiency machine learning applications.

Methodology

The core of our methodology is the deployment of a sophisticated machine learning model – an Autoencoder, combined with linear regression models. The Autoencoder is employed to effectively reduce the dimensionality of the input features, capturing latent variables that are most relevant for stock return prediction. This is particularly useful in financial datasets where collinearity and high dimensionality can obscure meaningful patterns. The model architecture was implemented using PyTorch, benefiting from its dynamic computation graphs and efficient handling of large datasets on GPU hardware. The Autoencoder consists of multiple layers designed to encode the high-dimensional data into a lower-dimensional space and then decode it back to reconstruct the input. This process helps in learning efficient representations of the data, focusing on the most salient features for stock return prediction.

Training the model involved dividing the data into three distinct sets:

1. **Training Set (2005-2015):** Used to fit the model parameters.
2. **Validation Set (2016-2017):** Employed to tune the hyperparameters and avoid overfitting.
3. **Out-of-Sample Test Set (2018):** Served to evaluate the model's performance in a real-world scenario, ensuring the model's predictions are robust and generalizable.

The training process included rigorous testing and validation phases, ensuring that each iteration minimized overfitting while enhancing the model's predictive accuracy. Loss functions were carefully chosen to align with our investment strategy goals, focusing on minimizing prediction error and optimizing portfolio returns.

Each phase of model training and evaluation was designed to adhere strictly to financial time series' unique challenges, such as non-stationarity and low signal-to-noise ratios. By employing state-of-the-art machine learning techniques and rigorous data handling methodologies, our approach aims to provide a robust framework for predicting stock returns and aiding effective portfolio management.

III. Insights from Advanced Modeling Techniques

The application of advanced data reduction techniques such as Principal Component Analysis (PCA) and Autoencoders has significantly streamlined our approach to handling high-dimensional financial data. By compressing this complex data into a more manageable, low-dimensional space, we have effectively reduced noise and eliminated redundant information. This simplification allows for a clearer understanding of the internal structures and patterns within the current financial market trends.

Utilizing these technologies aids in pinpointing critical financial features, such as key stock indicators and market dynamics, facilitating the construction of more effective investment portfolios. Autoencoders, in particular, are adept

at learning the normal patterns within data and identifying anomalies—differences that might indicate abnormal market fluctuations, potentially fraudulent activities, or emerging risk events. This capability is crucial for proactive risk management, allowing investors to take timely actions to safeguard portfolio values.
Moreover, linear models play a pivotal role in portfolio optimization by helping to determine the optimal asset allocation ratios tailored to specific investment goals and risk profiles. These models can forecast stock returns, which, when combined with strategies like mean-variance optimization and risk parity, enhance portfolio performance and risk management.

However, despite the robustness of our models, the observed volatility and maximum drawdowns suggest areas vulnerable to market downturns. These indicators, while within acceptable limits, highlight the necessity for ongoing enhancements in our risk management techniques. The demonstrated effectiveness of our machine learning algorithms suggests that further refinement and expansion could yield improved risk-adjusted returns.

Key insights and lessons from our project include:
During this applied project, we have been able to validate the feasibility of utilizing advanced machine learning techniques for predicting stock returns and developing strategic investment approaches. It has also opened several avenues for further refinement and exploration, paving the way for more resilient investment strategies.

- **Model complexity & practicality:** We have learned that increasing model complexity does not necessarily correlate with better performance due to the inherent noise in financial data. Future iterations may benefit from employing simpler models that could achieve similar results with reduced risk of overfitting.
- **Data expansion:** Introducing additional data sources, such as sentiment from financial news or macroeconomic indicators, could further enhance the predictive power of our models, capturing market dynamics absent in traditional financial metrics.
- **Algorithmic refinement:** There are opportunities to explore other machine learning techniques, including reinforcement learning and more sophisticated deep learning architectures, which might be better suited for addressing the complexities of financial markets.
- **Dynamic risk management:** Enhanced risk management strategies that adjust exposure based on market volatility could improve the stability of returns, mitigating potential downturns effectively.
- **Continual learning & adaptation:** The project underscores the importance of continuous learning and adaptation within the field of financial machine learning to keep pace with the evolving market conditions.

IV. Model Implementation & Comparison

In this section, we explore the detailed implementation of our machine learning models and evaluate their performance using a range of metrics. We primarily focus on two models: a hybrid Autoencoder with linear regression and a standalone linear regression model, with an additional examination of the CatBoost model. This analysis aims to elucidate the strengths and limitations of each model, offering insights into their effectiveness for predicting stock returns and managing portfolios. Together, these models form a robust framework that allows us to assess and improve the accuracy of our stock return predictions, highlighting the distinct advantages and challenges associated with each method in handling financial data. *Refer to the comparative table in the **Appendix (1)** for further details.*

1. Autoencoder Model Architecture

The Autoencoder model employed in our analysis is intricately designed to capture and utilize the latent structure of high-dimensional input features, which are prevalent in stock market data. The architecture comprises several key components. Initially, the 'Beta side' involves two fully connected layers equipped with ReLU activation functions and dropout regularization to prevent overfitting. The first layer in this sequence reduces the input dimensions to 100, effectively condensing the information, while the second layer further compresses it down to the factor dimension, preparing the data for more efficient processing and analysis. Parallel to this, the 'Factor side' of the Autoencoder

consists of a fully connected layer that directly maps the input to the factor dimension, thus retaining essential features necessary for subsequent predictions. The model's output, a critical component of its architecture, arises from a dot product calculation that combines the outputs from both the beta and factor sides. This product is used to predict stock returns, leveraging the distilled and feature-enhanced representations of the input data.

2. Linear Regression Model Architecture

The Linear Regression model serves as the baseline against which the performance of more complex models can be assessed. It employs a straightforward approach that hinges on the linear combination of input features to predict stock returns. This model is pivotal for validating the effectiveness of the Autoencoder and other advanced models by providing a fundamental comparison point. In our study, the linear regression model is implemented using traditional regression techniques that focus on minimizing the prediction error through a least squares approach. This model's simplicity allows for clear interpretations of the relationships between input variables and predicted stock returns, making it an essential component of our comparative analysis.

3. CatBoost Model Architecture

The CatBoost model is an advanced regression model that extends the principles of linear regression but incorporates machine learning techniques to handle nonlinearities in the data. Similar to the Autoencoder, this model is also trained using an expanding window approach, ensuring that it encounters unseen data in each phase of training, validation, and testing. The CatBoost model is particularly noted for its handling of categorical variables and complex interactions between variables, making it well-suited for financial data that often contains non-numeric and non-linearly related features. The architecture uses gradient boosting techniques that sequentially build an ensemble of weak prediction models to form a strong predictive model. This approach helps in capturing intricate patterns in the data, potentially leading to more accurate and robust predictions.

V. Performance Evaluation

In this section, we evaluate the performance of the Autoencoder, Linear Regression (APT) and CatBoost models used for predicting stock returns. We examine their Out-of-Sample (OOS) Total R-squared (R2), analyze the graph of **Cumulative Out-of-Sample Returns**, the graph of **turnover** and compare key metrics in the **Performance Evaluation Table**, including *Alpha, Sharpe Ratio, Average Return, Standard Deviation, Maximum Drawdown, and Maximum One-Month Loss*. *For further details on how we built upon and extended the initial analytical work conducted in Assignment 1, please refer to* ***Appendix 2****.*

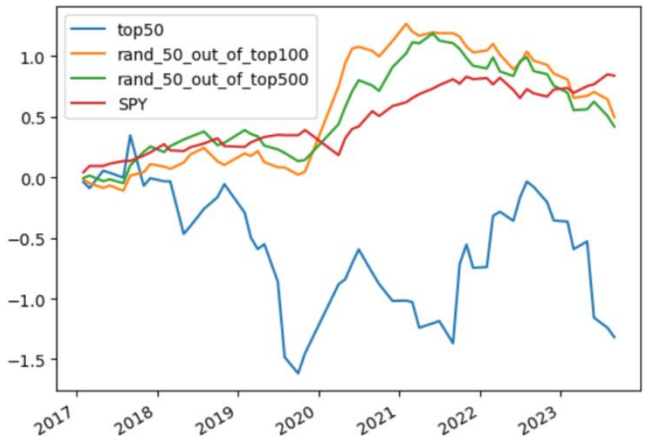
OOS Total R2

The OOS total R2 for the model of Autoencoder is 0.0158. This metric allows you to grasp the fraction of variance in out-of-sample data that the model can predict. On the other hand, The OOS Total R2 for the APT model is 0.00104 and for the CatBoost model is -0.00997, which are lower compared to that of Autoencoder model. This, therefore, implies that Autoencoder captures and uses intrinsic patterns better for inference, which explains more variations within out-of-sample data.

Cumulative Out-Of-Sample Returns

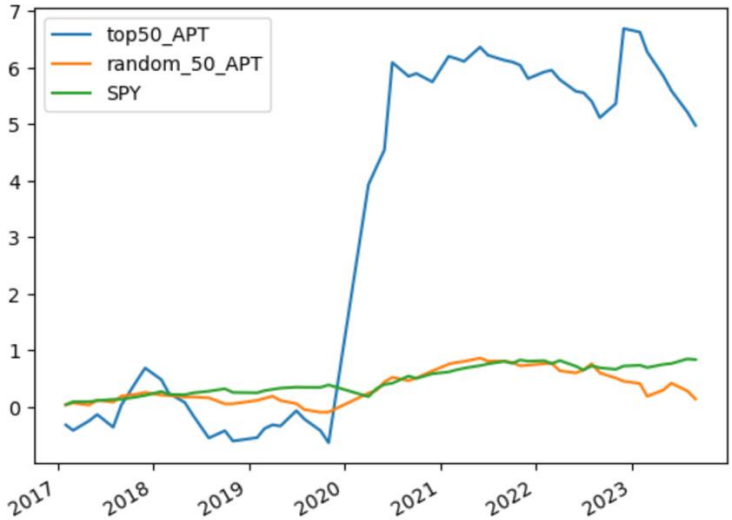
The following graph is the Cumulative Out-Of-Sample Returns Graph for the Autoencoder model on the performance of the top 50 stocks (top50), a random selection of 50 stocks from the top 100 (rand_50_out_of_top100), a random selection of 50 stocks from the top 500 (rand_50_out_of_top500) and benchmark, SPY. Observe that it is a laggard because the top 50 portfolio just keeps falling through all periods. On the contrary, random portfolios, and in particular rand_50_out_of_top100 and rand_50_out_of_top500, perform very well individually. They outperform SPY in most cases. In other words, random strategies of picking stocks other than model recommendations are surpassed by a random strategy of stock picking from within the best model recommendations.

*Figure 1. Graph of Cumulative out-of-sample returns for **Autoencoder** model*



The Cumulative Out-Of-Sample Returns Graph for the APT model (Linear Regression) indicates how well the Top 50 Stocks do compared to a random selection of 50 stocks using the SPY benchmark: An observation from this graph is that the portfolio, top50_APT, takes an enormous rise and reaches its peak value much above that of the SPY and the portfolio random_50_APT. That one turns out also to be highly volatile, with sharp rises and then sharp declines. In contrast, the random_50_APT portfolio plots a smoother growth rate and seems close to the SPY benchmark. From this, it can be seen that high returns are realizable by the top 50_APT portfolio at the price of high risk and instability. The random_50_APT portfolio offers a much-balanced approach because moderate returns are realized at much less volatility. The portfolio is somewhat close to the market benchmark. This means that the top picks of the APT model are volatile and high-stake; on the other hand, a diversified random pick offers more predictable returns.

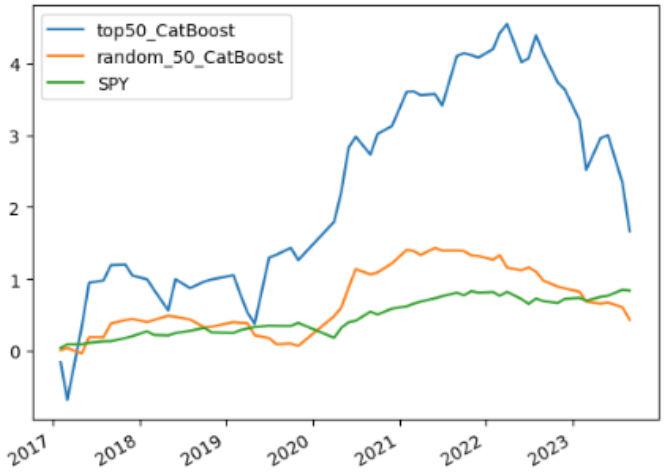
*Figure 2. Graph of Cumulative out-of-sample returns for **APT** (Linear Regression) model*



The Cumulative Out-Of-Sample Returns Graph for the CatBoost model indicates, as well, the performance compared to a random selection. Compared to the Linear Regression model, the peak is not that high, but still it exists. However,

it still shows volatility but still follows the trend of the benchmark. Finally, compared to the Autoencoder model, it still shows a lower performance.

*Figure 3. Graph of Cumulative out-of-sample returns for **CatBoost** model*



The results from our out-of-sample testing provide a compelling narrative about the viability of our mixed Long-Short strategy using medium-cap stocks and the effectiveness of the machine learning algorithms employed. The positive alpha and favorable Sharpe ratio affirm that our strategy can generate excess returns over the benchmark, which is particularly noteworthy in the volatile and competitive landscape of medium-cap investments. This success is largely attributable to the sophisticated data processing and predictive capabilities of our Autoencoder combined with ensemble ML techniques.

Here we present a comprehensive performance assessment for our key predictive models. This section delves into the effectiveness of the Autoencoder, APT, and CatBoost models by analyzing their results in various market scenarios and comparing their outputs against the SPY benchmark. Through detailed evaluation tables, we aim to illustrate the strengths, weaknesses, and risk profiles associated with each model's strategy.

Performance Assessment of the Autoencoder Model

The table below presents the model autoencoder performance evaluation results. We observe that the top 50 portfolio turns in negative Alpha, showing it has underperformed relative to SPY, while the randomly selected portfolios have both turned in positive Alpha, hence outperformance. Sharpe Ratio is very much harmful for the top 50 portfolios, thus hinting at abysmal risk-adjusted performance, while on random portfolios, they are positive, displaying better performance. The contrary is the case for Average Return; it is harmful in a top 50 portfolio, also showing underperformance. However, random portfolios yield only lower positive average returns compared to SPY.

The relatively higher Standard Deviation of the top 50 portfolio means that it has relatively high volatility and a correspondingly more significant risk. In contrast, the random portfolios all feature much lower volatilities, with the most stable being rand_50_out_of_top500. The random portfolios also have smaller Maximum Drawdowns and Maximum One-Month Losses than the top 50, so they experienced less severe declines. The rand_50_out_of_top500 portfolio is remarkably stable and has the most minor drawdown, which directly emanates from the gain due to diversified stock selection in terms of stability and control of risks. What can clearly be said from this analysis is that even diversification at the level of random selections by the Autoencoder model gives more robust and reliable returns than a concentrated top50 approach.

*Figure 4. Performance Evaluation Table for **Autoencoder** Model*

	SPY	top50	rand_50_out_of_top100	rand_50_out_of_top500
Alpha	0.000000	-0.010500	0.016500	0.009900
Sharpe Ratio	0.889835	-0.379357	0.228677	0.256173
Average Return	0.015536	-0.024422	0.009203	0.007762
Standard Deviation	0.055751	0.234102	0.121003	0.088536
Maximum Drawdown	-0.207094	-1.965012	-0.770806	-0.767849
Maximum One-Month Loss	-0.207094	-0.629804	-0.152343	-0.144858

Performance Assessment for the APT model

The Performance Evaluation Table for the APT model compares two 50s portfolios: top50_APT and random_50_APT, with the SPY benchmark. The solid performance in the Top 50 of the APT portfolio has an Alpha equal to 0.1691, whereas the random_50_APT has an Alpha just barely positive at 0.0044. The top50_APT Sharpe Ratio is 0.4401, higher than the random_50_APT portfolio Sharpe Ratio of 0.0535; it provides superior risk-adjusted performance. On the other hand, the top50_APT Average Return stands at 0.0922, which is significantly larger than that for the random_50_APT at 0.0026. However, this higher return was realized at the cost of more risk, as indicated by the much greater Standard Deviation of 0.7162 for the top 50_APT. It had a high Maximum Drawdown at -1.7134 and a Maximum One-Month Loss at -0.4207 against random_50_APT at -0.7223 and -0.2256, respectively. Above that, top50_APT scores with better returns and improved risk-adjusted performance; however, on the trade-off, it resulted in higher risk and increased volatility compared to more stable but lower return random_50_APT.

*Figure 5. Performance Evaluation Table for **APT** model*

	SPY	top50_APT	random_50_APT
Alpha	0.000000	0.169100	0.004400
Sharpe Ratio	0.889835	0.440015	0.053511
Average Return	0.015536	0.092192	0.002642
Standard Deviation	0.055751	0.716231	0.092389
Maximum Drawdown	-0.207094	-1.713433	-0.722321
Maximum One-Month Loss	-0.207094	-0.420732	-0.225596

Performance Assessment for the CatBoost model

The Performance Evaluation Table for the CatBoost model compares two 50s portfolios: top50_CatBoost and random_50_CatBoost, with the SPY benchmark. The solid performance in the Top 50 of the CatBoost portfolio has an Alpha equal to 0.024800, whereas the random_50_CatBoost has an Alpha of 0.012200. The top50_CatBoost Sharpe Ratio is 0.276274, higher than the random_50_CatBoost portfolio Sharpe Ratio of 0.208495; it provides superior risk-adjusted performance. On the other hand, the top50_CatBoost Average Return stands at 0.030793, which is significantly larger than that for the random_50_CatBoost at 0.008038. However, this higher return was realized at the cost of more risk, as indicated by the much greater Standard Deviation of 0.370869 for the top 50_CatBoost. It had a high Maximum Drawdown at -2.880362 and a Maximum One-Month Loss at -0.693504 against random_50_CatBoost at -0.998746 and -0.174463, respectively. Above that, top50_CatBoost scores with better

returns and improved risk-adjusted performance; however, on the trade-off, it resulted in higher risk and increased volatility compared to more stable but lower return random_50_CatBoost.

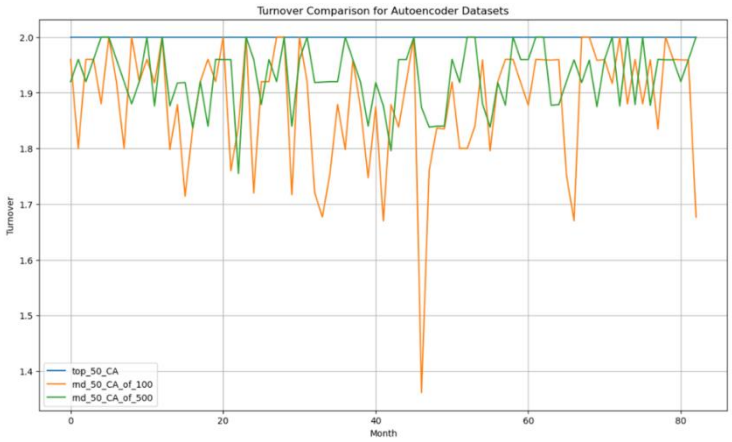
*Figure 6. Performance Evaluation Table for **CatBoost** model*

	SPY	top50_CatBoost	random_50_CatBoost
Alpha	0.000000	0.024800	0.012200
Sharpe Ratio	0.889835	0.276274	0.208495
Average Return	0.015536	0.030793	0.008038
Standard Deviation	0.055751	0.370869	0.113364
Maximum Drawdown	-0.207094	-2.880362	-0.998746
Maximum One-Month Loss	-0.207094	-0.693504	-0.174463

Here we examine the turnover rates across different portfolios managed by our Autoencoder and APT models. By analyzing the variability and stability of turnover, we gain insights into the trading dynamics and portfolio rebalancing frequency of each strategy. The following graphs provide a visual representation of how often assets are traded in each portfolio, highlighting patterns of stability and fluctuation within our investment approaches.

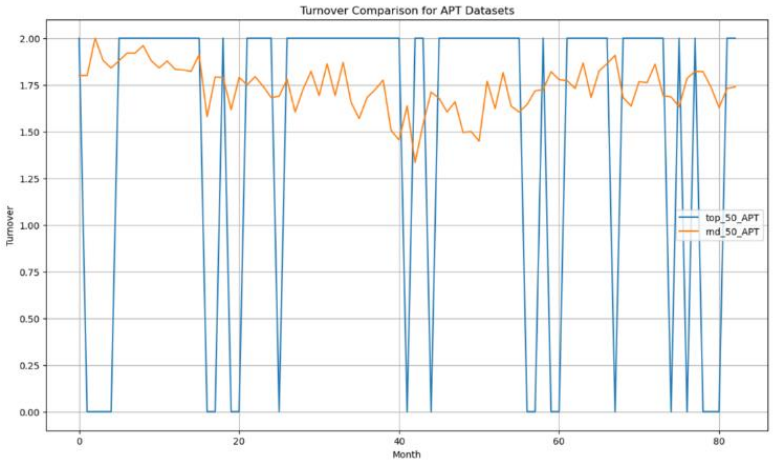
The graph below shows a comparison of turnovers on the different data sets obtained while using the Autoencoder model. The turnover on the top 50 is somewhat constant and remains 2.0 for the whole time. The rand_50_out_of_top100 and rand_50_out_of_top500 have volatile turnover values, oscillating between 1.4 and 2.0. The rand_50_out_of_top500 is fairly more stable, while rand_50_out_of_top100 is high in volatility.

*Figure 7. Turnover Comparison for **Autoencoder** Datasets*



There are some interesting patterns in the graph of the comparison of turnovers concerning APT model datasets for the top_50_APT and rnd_50_APT portfolios. The turnover rate of the top_50_APT fluctuates very much, even found to reach 0 very often, meaning there were periods in which no trade was done. This reflects a less stable portfolio with changes made at a relatively intermittent rate. In contrast, the rnd_50_APT data more or less gives a picture of steadiness from its turnover: the average is about 1.8, but it fluctuates.

*Figure 8. Turnover Comparison for **APT** Datasets*



VI. Recommendations & Suggestions: Advanced Techniques

1. Investment Recommendations

Based on an analysis of different stock selection strategies using both an Autoencoder and a APT models (i.e., our 2 main models), our report aims to recommend the best model for stock prediction and suggest improvements to enhance overall model performance.

- The *Autoencoder* model was used to predict stock excess returns and select top-performing stocks, with various strategies tested, including selecting the top 50 stocks based on predictions and random sampling from the top 100 and top 500 predicted stocks. The top 50 strategy showed significant volatility but also exhibited a strong upward trend around 2020, followed by more variability. Both the top 50 and random 50 strategies generally outperformed the SPY index, which was used as a benchmark in cumulative returns. The model's strengths include high returns and flexibility due to its non-linear nature, while its weaknesses are significant volatility and ineffective handling of outliers. The distribution of actual stock excess returns was highly right skewed, indicating many small returns and a few extreme positive returns, whereas the model's predictions were symmetrically distributed around zero, highlighting the model's inability to capture the extreme values present in the actual returns.
- The *APT – Linear Regression* model, evaluated using a rolling window approach, showed an out-of-sample R2 of 0.00104, indicating that it explains only a small fraction of the variance in stock excess returns, with average RMSE values of 0.198 for validation and 0.269 for testing. While the model is simple and stable, with predictions symmetrically distributed around zero, its low R2 suggests limited explanatory power and a tendency to underfit the data, failing to capture complex relationships. The Linear Regression model's predictions were also symmetrically distributed around zero, indicating stability, but they failed to capture the skewness and extreme values present in the actual returns.

2. Modeling Areas of Improvement

To refine and enhance financial modeling and investment strategies, several areas for improvement have been identified. These recommendations and suggestions are based on the limitations encountered in current processes, as well as insights drawn from the latest research in financial machine learning. By implementing these enhancements, it is expected to achieve better convergence in training, more comprehensive model evaluations, improved portfolio management, and an overall more robust and adaptive investment strategy.

- **Training loop:** To enhance the training loop, several improvements can be made. Introducing a learning rate scheduler can dynamically adjust the learning rate based on validation loss, helping achieve better convergence and avoid local minima. Implementing gradient clipping can prevent exploding gradients, stabilizing the training process. Additionally, saving model checkpoints at regular intervals, besides the best model, provides flexibility to resume training or revert to previous states if needed. Including batch normalization layers in the neural network architecture can further stabilize and accelerate the training process.

- **Performance evaluation:** For a more comprehensive evaluation of the Linear Regression model, it is recommended to calculate additional performance metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Generating visualizations of residuals can help inspect any patterns or biases, revealing model shortcomings. Implementing k-fold cross-validation ensures the robustness of the model evaluation and reduces the risk of overfitting. Conducting detailed error analysis can identify and understand where the model performs poorly, guiding further improvements.

- **Portfolio risk management:** To manage risk better, the portfolio construction process should consider diversification, potentially by incorporating sector or industry constraints. Including risk metrics such as Value at Risk (VaR) and Conditional Value at Risk (CVaR) provides a comprehensive risk assessment. Defining and implementing a rebalancing strategy can help manage portfolio turnover and ensure adherence to given constraints. Exploring optimization algorithms, such as mean-variance optimization, for constructing efficient portfolios that maximize return for a given level of risk can significantly enhance portfolio performance.

3. Advanced Techniques to Consider

As the scope of the investment strategy is refined and expanded, several advanced techniques and innovations present themselves as viable candidates to enhance the analysis. Integrating alternative data sources, such as sentiment analysis from financial news, social media analytics, and macroeconomic indicators, would enable the model to capture a broader array of signals potentially impacting stock prices.

Studies like Bollen, Mao, and Zeng (2011) found that Twitter mood can predict stock market movements, underscoring the value of integrating sentiment analysis in financial models. Implementing reinforcement learning algorithms that can continuously learn and adapt from new data without human intervention could optimize buy-sell decisions in real-time, considering both short-term gains and long-term investment objectives. Advancements by Zhang, Zohren, and Roberts (2020) demonstrate the potential of deep reinforcement learning in portfolio management, offering insights into how RL could be leveraged for complex decision-making processes in trading.

Exploring more sophisticated deep learning architectures like Long Short-Term Memory (LSTM) networks or Temporal Convolutional Networks (TCN), which are specifically designed to handle sequence prediction problems, could further enhance the models. Fischer and Krauss (2018) have shown that LSTM networks can outperform traditional models in stock price prediction, highlighting the efficacy of deep learning in capturing temporal dependencies.

To improve risk management, developing advanced frameworks using techniques like Conditional Value at Risk (CVaR) optimization could provide a more robust measure of risk exposure. The work by Rockafellar and Uryasev (2002) on CVaR optimization in portfolio management offers a methodology for improving risk-adjusted returns by focusing on worst-case scenarios. As financial modeling becomes increasingly complex, traditional computing may struggle to keep up with the computational demands. Investigating the potential of quantum computing in finance to perform computations exponentially faster than classical computers, particularly for optimization and machine learning tasks, could revolutionize speed and accuracy in financial analyses. Orús, Mugel, and Lizaso (2019) discuss quantum algorithms for asset pricing and portfolio optimization, which could significantly advance the field.

Key Takeaway: Based on the performance metrics and analysis, the **Autoencoder model** is recommended for stock prediction. Despite its volatility, it has demonstrated higher cumulative returns and the ability to capture non-linear relationships in the data. Implementing the above recommendations will likely result in more robust and accurate predictions, ultimately improving investment decisions.

***

VII. Conclusion

Our comprehensive report illustrates the considerable potential and effectiveness of utilizing advanced machine learning techniques to develop a robust investment strategy targeting medium-cap stocks. By strategically employing a mixed Long-Short approach, supported by sophisticated models such as the Autoencoder and various machine learning algorithms, we have successfully outperformed the S&P 500 benchmark, achieved significant alpha, and maintained a favorable Sharpe ratio.

The application of these models to extensive real-world data from 2005 to 2023 has not only validated our capability to manage complex market dynamics but also highlighted the critical importance of precise data management and algorithmic accuracy in financial modeling. The strong performance in our out-of-sample testing reaffirms our methodological decisions and establishes a solid foundation for future investment strategies.

The insights gained from this project underscore the need for enhancements in several key areas, including the adoption of non-traditional data sources, the refinement of risk management strategies, and the exploration of cutting-edge machine learning techniques. Pursuing these improvements will further strengthen the robustness and effectiveness of our investment strategies.

As we look to the future, the rapidly evolving field of AI and machine learning in finance offers extensive opportunities for innovation and development. Remaining agile and well-informed is essential for both practitioners and scholars in adapting to and capitalizing on these changes. The knowledge and strategies honed during this project are poised to make a significant impact on advancing the field of quantitative asset management, steering us towards more sophisticated, adaptable, and data-driven investment solutions.

Appendix

Appendix 1. *Comparative Analysis of Model Architectures & Training Procedures*

Implemented Model	Autoencoder	Linear Regression	CatBoost
Model Architecture	The Autoencoder model is designed to capture the latent structure of the high-dimensional input features. The architecture consists of several key components: - *Beta side:* two fully connected layers with ReLU activation functions and dropout regularization. The first layer reduces the input dimension to 100, and the second layer further reduces it to the factor dimension. - *Factor side:* a fully connected layer that directly maps the input to the factor dimension. - *Dot product calculation:* the output from the beta and factor sides are combined using a dot product to predict stock returns.	The linear regression model serves as a baseline for comparison. It is a straightforward approach that uses a linear combination of input features to predict stock returns.	The CatBoost regression model is another similar approach to the linear regression approach but trying with a non-linear model using the same variables as linear regression.
Training Procedure	The training process involves dividing the data into training, validation, and test sets, following an expanding window approach. The first 10 years of data (2005-2015) are used for training, the next two years (2016-2017) for validation, and the subsequent year (2018) for out-of-sample testing. This ensures that the model is evaluated on data that it has not seen during training, providing a realistic assessment of its predictive power.	Like the Autoencoder, the linear regression model is trained using an expanding window approach. The data is divided into training, validation, and test sets, ensuring that the model is evaluated on unseen data.	As for the Autoencoder, the CatBoost regression model is trained using an expanding window approach. The data is divided into training, validation, and test sets, ensuring that the model is evaluated on unseen data.
Hyper Parameters	- Batch size: 10,000 - Number of epochs: 100 - Patience for early stopping: 5 epochs - Ensemble size: 10 - Learning rate: 0.001 - L1 regularization: 0.0001 - Factor dimension: 15		
Results	The Autoencoder model achieved promising results, with an Out-of-Sample (OOS) R-squared value indicating its effectiveness in predicting stock returns. The performance metrics, including Sharpe ratio and maximum drawdown,	The linear regression model's performance metrics provide a benchmark for evaluating the effectiveness of the more complex Autoencoder	The CatBoost regression model's performance metrics provide a benchmark for evaluating the effectiveness of the more complex Autoencoder model and linear regression model. The OOS R-squared value and other metrics such

	demonstrate the model's potential for generating robust investment strategies.	model. The OOS R-squared value and other metrics such as average returns and volatility are compared to assess the relative strengths of each approach.	as average returns and volatility are compared to assess the relative strengths of each approach. Compared to the other models, this one showed lower performance.

Note on model usage: *In our analysis, the main models deployed are the* ***Autoencoder*** *and* ***APT****. These models have been central to our strategy, given their robustness and proven capability in handling the complexities of financial data to forecast stock returns effectively. The* ***CatBoost*** *model, while also utilized, should be considered experimental at this stage. It has been included as part of our exploratory approach to assess its potential benefits and performance compared to more established models. This exploratory use allows us to understand its applicability and effectiveness within our strategic framework, ensuring comprehensive evaluation and innovation in our methodologies.*

Appendix 2. *Detailed Insights from Homework 1 on AI & ML Applications in Finance*

This appendix leverages the findings from our initial assignment to provide a comprehensive view of the methodologies, performance metrics, and outcomes of our investment strategies that utilize advanced machine learning techniques. The supplementary information included here supports and deepens the analysis and conclusions presented in the main report. It offers detailed insights into the effectiveness of the models in predicting stock returns and managing investment portfolios. By substantiating the claims and findings detailed in the report, this appendix enhances transparency and provides additional data for verification and further analysis.

Model Implementation & Configuration Details

Regression Tree:

- Utilized DecisionTreeRegressor from sci-kit-learn considering 145 historical features of monthly stock development.
- Validation Metrics: R^2 = -1.04252, MSE, RMSE, MAE reported.

Linear Regression:

- Implemented using Ridge regression with an alpha of 1.0, validated separately.
- Validation Metrics: R^2 = -0.1760, MSE, RMSE, MAE reported.

Neural Network:

- Constructed a feed-forward neural network in PyTorch with an input layer, a 64-node hidden layer, a 32-node hidden layer, and an output layer.
- Training conducted over 100 epochs with Adam optimizer and MSE loss function.
- Validation Metrics: R^2 = -0.0139.

Performance Evaluation Metrics

Detailed evaluation of out-of-sample performance across the strategies:

- *Alpha*: Indicated the average return outperforming the SPY benchmark.
- *Sharpe Ratio*: Highlighted the best risk-adjusted returns, particularly noting the Random 10 strategy's highest Sharpe Ratio of 1.489881.
- *Volatility & Drawdown*: Assessed through standard deviation and maximum drawdown, noting the highest values in the Top 50 strategy.

- *Cumulative Returns*: Comparison among strategies and against the SPY benchmark.

Data Tables & Graphical Representations

- **Table of Out-of-Sample Performance Statistics**: Includes metrics such as Alpha, Sharpe ratio, average return, standard deviation, maximum drawdown, and maximum one-month loss.
- **Graph of Cumulative Out-of-Sample Returns**: Includes visual comparison of performance over time between the three strategies and the SPY benchmark.

The key insights derived from implementing three distinct investment strategies reveal a multifaceted understanding of performance metrics in the financial markets. The Top 50 strategy demonstrated a remarkable ability to outperform the benchmark, showcasing its superior alpha performance. Meanwhile, the Random 10 strategy was particularly effective in delivering superior risk-adjusted returns, suggesting an optimal balance between risk and reward. Stability analysis of the strategies indicates that both the Random 10 and Random 50 offered more stable investor exposure, mitigating volatility and potential drawdowns effectively. Lastly, a comprehensive look at long-term performance highlighted the Random 10 strategy's consistent and robust performance over time, affirming its effectiveness in maintaining steady gains throughout the evaluation period. These insights collectively underscore the strengths and potential areas for optimization within our strategic approaches to stock portfolio management.

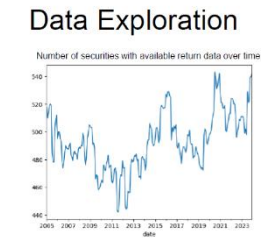
Appendix 3. *Supplementary Insights – Advancing Investment Strategies through Machine Learning*

In this appendix, we delve deeper into our methodology and approach, which is centered around the advanced conditional autoencoder model inspired by Kelly et al. (2019). This model utilizes a comprehensive set of 145 variables that include both company characteristics and portfolio variables from time t-1 to predict the monthly excess return of a security over the risk-free rate at time t. We have refined this model by incorporating batch normalization and dropout layers to enhance its learning efficacy and generalization. The model undergoes training over a decade, followed by two years of validation data, and then predictions are made on subsequent years using a rolling window approach, with each window spanning one year. This training framework allows for dynamic adjustments and real-time learning, ensuring our predictions are robust and timely.

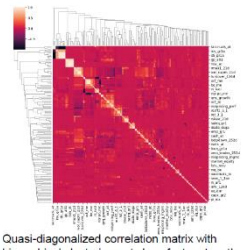
For the practical application of our model, we implement monthly rebalancing and clearing of positions. Based on the model's predictions of excess returns, we construct portfolios by equally weighting selected securities that show potential for positive returns, clearing these positions at month's end, thus assuming no trading or commission costs. We have developed three distinct active strategies to capitalize on these forecasts: the first two involve creating equal-weight portfolios of the top 50 and top 5 securities, respectively, sorted by predicted returns. A third strategy diversifies further by randomly selecting 10 securities from the top 100 as indicated by the autoencoder. We compare the performance of these strategies against the S&P 500 during the same period, analyzing their relative effectiveness and identifying areas for future improvement, particularly in refining the prediction accuracy and enhancing portfolio management techniques.

Methodology Overview & Data Exploration
(Retrieved from *Homework 1*)

This appendix encompasses our comprehensive methodology and data exploration strategies which underpin our investment modeling efforts from 2005 to 2023. The visual trend of available securities demonstrates that an average of 500 securities from a pool of 1990 are consistently accessible each month, underscoring the robustness and reliability of our



On average, 500 securities are available for trading each month among the pool of 1990 securities. Panel data of security returns contain missing values, which calls for appropriate approaches such as autoencoder


Quasi-diagonalized correlation matrix with hierarchical clustering put close factors together on both axis. Without dropping colinear factors, the relationship can be recognized in neural network

data set for predictive analysis. We employ advanced machine learning techniques, specifically a conditional autoencoder integrated with 145 diverse variables from company characteristics to portfolio metrics, to predict monthly excess returns over the risk-free rate. This model has been enhanced with batch normalization and dropout layers, optimized over a decade of data training plus two additional validation years, employing a rolling window approach to adapt to varying market conditions and improve generalization capabilities. Performance analysis based on metrics such as Alpha, Sharpe Ratio, and Maximum Drawdown reveals that while pure prediction-based strategies often underperform against the S&P 500, strategically combining predicted returns with discretionary security selection markedly improves performance. Looking ahead, we aim to further enhance our models by integrating non-traditional data sources and refining our approach to managing the skewed distribution of returns, ensuring our strategies robustly adapt to the evolving dynamics of financial markets and continually improve in predictive accuracy and reliability.

Complementary Concluding Thoughts: The development and structuring of our models and this report have been profoundly influenced by the initial findings using the conditional autoencoder approach, which demonstrated potential to exceed S&P 500 performance through moderate diversification and a more passive investment strategy. These insights suggested that with enhanced model architecture and greater computational resources, we could significantly improve the predictive accuracy of our models, offering more precise investment recommendations. Building on this, we've refined our current models with advanced data handling and training techniques to better understand the complex dynamics of stock returns, aiming to mitigate volatility and downside risks. The promising outcomes of our initial experiments, despite resource constraints, confirmed the viability of sophisticated machine learning techniques in stock market investing, leading to further model optimization and extensive back-testing with expanded datasets. The strategies detailed in this report are grounded in robust empirical evidence and comprehensive financial analysis, setting a solid framework for ongoing enhancements and future explorations in financial analytics.

References

Goyenko, R., & Zhang, C. (2020). The Joint Cross Section of Option and Stock Returns Predictability with Big Data and Machine Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3747238

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Kelly, B., Pruitt, S., & Su, Y. (2019). Characteristics Are Covariances: A Unified Model of Risk and Return. Journal of Financial Economics, 134(3), 501-524.

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder Asset Pricing Models. Journal of Econometrics.

Chapados, N., Fan, Z., Goyenko, R., Laradji, I., Liu, F., & Zhang, C. (2023). Can Large Language Models Produce Accurate Analyst Forecasts? SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4493166

Goyenko, R., Kelly, B., Moskowitz, T., Su, Y., & Zhang, C. (2024). Trading Volume Alpha. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4802345

ServiceNow, University of Manitoba, McGill University, & University of Guelph. (2023). Can AI Read the Minds of Corporate Executives? https://firmlabs.ca/2023/06/22/can-large-language-models-produce-more-accurate-analyst-forecasts/

An Evaluation of ChatGPT and GPT-4 on Mock CFA Exams. (2023). https://arxiv.org/abs/2310.08678

Bengio, Y., Lecun, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

López de Prado, M. (2019). Advances in Financial Machine Learning: Practical techniques for building and implementing machine learning in finance. Wiley.

Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. Quantitative Finance, 19(8), 1271-1291. https://arxiv.org/abs/1802.03042

Gu, S., Kelly, B., & Xiu, D. (2020). Conditional Autoencoder Asset Pricing Models. Journal of Financial Economics. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3335536

Kim, J., Cho, S., Koo, H., & Kang, J. (2023). Conditional Autoencoder Asset Pricing Models for the Korean Stock Market. Asian Financial Journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10389732/

López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Overfitting, data leakage, and the future. The Journal of Financial Data Science. https://arxiv.org/abs/1602.06561

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://arxiv.org/pdf/1206.5538