September 2025

Quantitative Predictive Modeling and Portfolio Risk Optimization on the S&P500

Alonzo R. Diaz Avalos

Abstract

This project investigates predictive modeling and portfolio strategies on the S&P500 using quantitative data analysis techniques. In the first phase, daily and monthly returns are analyzed alongside macroeconomic indicators to predict market movements. Binary classification models (logistic regression, KNN, SVM) are employed to predict the direction of monthly returns, while linear regression estimates their magnitude, applying k-fold cross validation with performance metrics reported for each method. In the second phase, two trading strategies are evaluated: buying the top 50 performing stocks of the previous day and buying the bottom 50. Strategy performance is assessed using historical returns, and the best-performing strategy is further optimized through weight allocation based on the volatility of the previous year. A Pareto frontier is constructed to illustrate the trade-off between risk and return. Finally, portfolio risk is quantified using historical, Monte Carlo, and parametric Value-at-Risk methodologies. This very simplistic and exploratory analysis demonstrates how predictive modeling, optimization, and risk assessment can be integrated to inform quantitative investment decisions.

Contents

1 Introduction

Financial markets are inherently complex and influenced by a multitude of economic, social, and behavioral factors. For investors and portfolio managers, understanding and anticipating market movements is crucial for informed decision-making. The S&P500, as one of the most widely followed equity indices globally, provides a rich environment for quantitative analysis due to its broad market coverage, high liquidity, and availability of historical data. This project aims to explore how predictive modeling and portfolio optimization techniques can be applied to the S&P500 to guide investment strategies while quantifying associated risks.

The motivation for this research arises from the need to integrate multiple analytical approaches into a cohesive framework, leveraging both predictive analytics and optimization to demonstrate how quantitative methods can enhance investment decision-making. The work is organized as follows.

The first phase focuses on predicting market behavior. Daily and monthly returns of S&P500 stocks are analyzed alongside macroeconomic indicators to anticipate whether the next months return will be positive or negative using, binary classification models such as logistic regression, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), within a cross-validation framework, while linear regression estimates their magnitude. The purpose of this phase is to generate predictive insights and to illustrate the effectiveness of various quantitative approaches in capturing market dynamics.

In the second phase, the focus shifts to portfolio strategies. Two simple yet illustrative strategies are considered: buying the top 50 performing stocks of the previous day and buying the bottom 50. The best-performing strategy is then optimized using volatility-based weighting to balance risk and return and is further analyzed using multiple approaches, including historical, Monte Carlo, and parametric Value-at-Risk (VaR) methods. This integration of predictive modeling with portfolio optimization demonstrates the potential of a systematic, quantitative approach to investment management.

Overall, this project presents a structured, data-driven methodology to study financial markets. Although the analysis is exploratory and simplified, it provides a clear demonstration of how predictive modeling, optimization, and risk assessment can be applied in practice, making it a compelling example of a quantitative approach to investment decision-making.

2 Methods and Data acquisition

The primary dataset consists of historical prices of S&P500 constituent stocks from January 2015 to January 2025. The S&P500 index represents 500 of the largest publicly traded companies in the United States and is a widely used benchmark for US markets. A current list was obtained from the Wikipedia page for S&P500 companies.

For each stock and for each trading day, the following price data were retrieved using the Yahoo Finance API in yfinance Python library as a dataframe:

- Open: the price at which the stock started trading on a given day.
- High: the highest price reached during the trading day.
- High: the highest price reached during the trading day.
- Low: the lowest price reached during the trading day.
- Close: the last traded price of the stock during the trading day.
- Adjusted Close (Adj Close): the closing price adjusted for corporate actions
- Volume: the number of shares traded on that day.

In this analysis we will look at Adj Close only instead of Close because it corrects for dividends and stock splits, giving a more accurate picture of real returns. The SPY ETF (exchange-traded fund) that tracks the S&P500 index was also retrieved and provides a convenient single instrument reflecting the overall S&P500 performance in a market-capitalization-weighted way.

Macro variables were collected from the Federal Reserve Economic Data (FRED) database later used for prediciton analysis, and include:

- UNRATE: unemployment rate (% of unemployed workers in the labor force)
- CPIAUCSL: consumer price index for all urban consumers, measuring inflation.
- FEDFUNDS: effective federal funds rate, the overnight lending rate between banks.
- UMCSENT: University of Michigan Consumer Sentiment Index, reflecting consumer confidence.
- RSAFS: retail sales total volume, indicating consumer spending.
- INDPRO: industrial production index, measuring output from manufacturing, mining, and utilities.
- M2SL: money supply, representing cash, checking deposits, and near-money assets.

where all of these indicators are reported on a monthly basis.

For the data cleaning and processing:

- The datetime index was standardized for all datasets.
- Missing stock prices were forward-filled to avoid gaps in return calculations.
- Stocks with more than 5% missing Adj Close values were removed to maintain data quality: it removes stocks with too many gaps that could distort results, while keeping most of the dataset intact.
- Daily stock returns were computed as simple percentage changes of Adj Close, and monthly returns were computed from month-end Adj Close.
- The macroeconomic dataset and S&P500 monthly returns were merged to create a complete dataset for predictive modeling.

This approach ensures the data are clean, aligned, and ready for subsequent analysis.

Methods used in this project are the following:

- Predictive modeling: linear regression, binary classification (KNN, SVM, Logistic Regression), k-fold cross validation.
- Portfolio strategy: optimization techniques.
- Risk analysis: historical, parametric and Montecarlo VaR.

3 Analysis and Results

In this section, we summarize the analysis conducted, highlighting the main results and key observations derived from the predictive modeling and portfolio strategy evaluation.

3.1 Exploratory Analysis

We begin our analysis by examining the SPY ETF as a proxy for the overall S&P500 market. Figure 1 shows the Adj Close of SPY over the 10-year period from 2015 to 2025, providing an overview of the market trend across this timeframe.

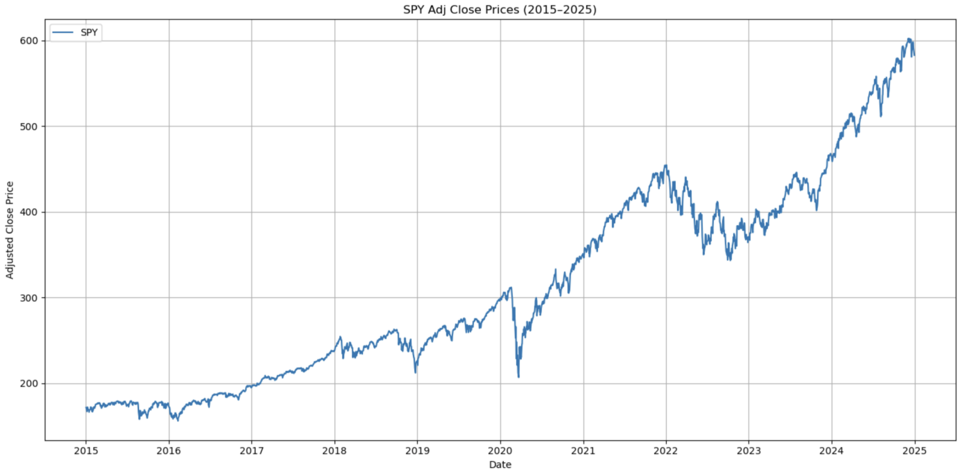

Figure 1: Adjusted closing price of SPY ETF from 2015 to 2025.

To further explore the histogram distribution of market movements, we compute daily returns based on Adj Close and present their frequency distribution in Figure 2. The histogram highlights the variability of daily returns and gives a first insight into the empirical distribution of market fluctuations.

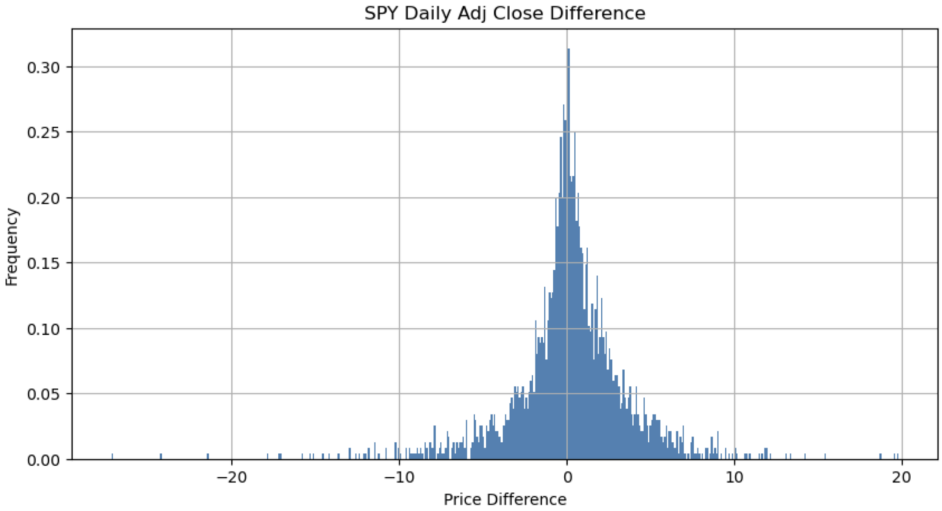

Figure 2: Daily returns frequency for Adj Close of SPY ETF from 2015 to 2025.

The histogram of daily returns was fitted with several candidate distributions. The best fit was obtained using the Generalized Normal Distribution with parameters $\beta = 0.8606$, $\mu = 0.00054$, and $\alpha = 0.00571$ with a p-value of 0.147.

The parameters were estimated using scipy.stats.gennorm.fit, and the goodness-of-fit was assessed via the Kolmogorov-Smirnov test in scipy.stats.kstest. The resulting p-value of 0.147 suggests that the distribution provides an acceptable fit for the SPY daily returns, and represents the best fit among the tested distributions, including Normal, Student's t, Laplace, Logistic and Cauchy distributions.

From the S&P500 data, we can extract several informative summaries of extreme movements and top performers. We show some examples in the following tables:

Date	Freq
2020-03	19
2020-04	15
2020-05	10
2020-06	9
2023-03	9

Most frequent left-tail months for JPM (q=0.05)

Stock	Return
COIN	1.176
CRWD	0.749
ERIE	0.615
WSM	0.608
KEY	0.560

Top 5 gainers (2023-07 – 2023-12)

Stock	Return
ALB	-0.367
PAYC	-0.351
EL	-0.254
HSY	-0.251
UAL	-0.251

Top 5 losers (2023-07 – 2023-12)

Stock	Return
APA	2.138
PLTR	1.676
MRNA	1.264
SMCI	1.124
ENPH	1.102

Top 5 single-month gainers

Stock	Return
APA	-0.832
TRGP	-0.787
CZR	-0.713
NCLH	-0.706
OKE	-0.673

Top 5 single-month losers

Date	Count
2022-06-13	492
2022-09-13	492
2021-11-30	489
2022-05-18	489
2020-06-11	489

Dates with most simultaneous losing tickers

Stock	Count
ALGN	126
ADSK	126
BG	126
UNP	126
ES	126

Stocks with most left-tail events (q=0.05)

Stock	Count
ALGN	126
ADSK	126
BG	126
UNP	126
ES	126

Stocks with most right-tail events (q=0.05)

Here, q = 0.05 refers to the 5% quantile of the return distribution. Events in the left tail (below the 5th percentile) represent extreme negative returns (large losses), while events in the right tail (above the 95th percentile) represent extreme positive returns (large gains).

Note that the dates associated with these extreme events may correspond to significant socio-economic or geopolitical events that affected the market during that period.

3.2 Predictive Modeling

We will now use the macroeconomic variables to predict the Adj Close monthly movements of the SPY. However, before modeling, it is important to examine the correlations among the variables. Some of the macro variables are highly correlated, which could potentially cause multicollinearity issues. To illustrate this, we show the correlations of the three most correlated variables along with SPY monthly returns:

	AdjClose	CPIAUCSL	RSAFS	M2SL
AdjClose	1.000	0.943	0.949	0.924
CPIAUCSL	0.943	1.000	0.978	0.907
RSAFS	0.949	0.978	1.000	0.943
M2SL	0.924	0.907	0.943	1.000

Table 1: Correlation matrix for SPY monthly returns (AdjClose) and the most correlated macroeconomic variables. Only the most correlated variables are shown for clarity.

Despite the potential for multicollinearity, we proceed with these variables in our analysis given the limited size of the dataset.

Monthly returns for SPY were computed from month-end Adj Close to align directly with the monthly frequency of the macroeconomic data.

We then model the monthly adjusted closing price of SPY using a multiple linear regression on the selected macroeconomic variables. Linear regression is chosen as a first step because it provides a transparent and interpretable framework, allowing us to directly assess the contribution of each macroeconomic factor through its coefficient. The regression is estimated using Ordinary Least Squares (OLS), and the model fit is evaluated with standard performance metrics (MAE, RMSE, Rš), which indicate how well the model explains variations in SPYs monthly price. The model is formulated as:

$$\begin{aligned}\text{AdjClose} =& \beta_0 + \beta_1 \cdot \text{UNRATE} + \beta_2 \cdot \text{CPIAUCSL} + \beta_3 \cdot \text{FEDFUNDS} + \beta_4 \cdot \text{UMCSENT} + \\ & \beta_5 \cdot \text{RSAFS} + \beta_6 \cdot \text{INDPRO} + \beta_7 \cdot \text{M2SL}.\end{aligned} \tag{3.1}$$

The estimated coefficients from the regression are:

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
-1907.0658	10.3626	5.7259	-23.8414	3.8023	0.0004	1.3072	0.0017

Table 2: Linear regression coefficients for SPY monthly AdjClose vs macro variables.

and the plot is shown in Figure 3.

Note that since the variables CPIAUCSL, RSAFS, M2SL were very correlated, the model, as expected, puts some of their coefficients close to zero because the predictors carry almost the same information. In such cases, the regression cannot clearly separate their individual effects on the target, so it assigns weight to just one or two of them while suppressing the others. However, we decide not to proceed with PCA since very little data is available, and we verified that the same analysis with Lasso does not improve performance significantly.

The performance of the model is evaluated using the Coefficient of Determination R^2: 0.9612, which measures the proportion of variance in the SPY monthly Adj Close that is explained by the macroeconomic variables. Overall, this indicates a very good fit to the data. The high R^2 suggests that the macro variables capture most of the variability in SPY monthly returns, approximately 96% of the variance is explained by the regression.

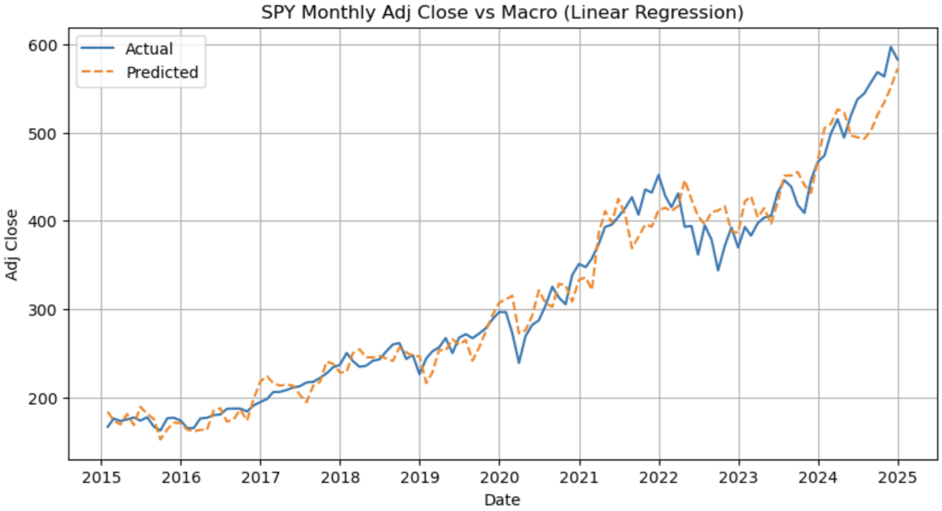

Figure 3: Predicted and Actual monthly adjusted closing prices of SPY using the multiple linear regression model with macroeconomic variables over the 10 year range.

In a second approach, in Figure 4, we implement linear regression using a 75/25 train/test split to evaluate the model on unseen data with 5-fold cross validation. This approach is used to assess the models ability to generalize to new, unseen data and to obtain a more reliable estimate of its predictive performance. In addition we standardize the variables so that each macroeconomic feature has mean 0 and standard deviation 1, ensuring they are on a comparable scale; this improves numerical stability, makes coefficients easier to interpret, and is especially important when using cross-validation or regularized regression, although for plain OLS predictions it does not change the fitted values. In this graph we plot instead Predicted vs Actual Adj Close. The performance metrics indicate a good fit, with training and test Rš values of 0.9569, showing that the model generalizes well, and as expected the R^2 is lower in this case since cross validation is testing the model in unseen data. The estimated coefficients from the regression are:

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
313.1465	17.8787	150.5944	-44.4759	54.1041	36.4391	3.9327	6.1053

Table 3: Linear regression coefficients for SPY monthly AdjClose Predicted.

Note how the biggest contribution comes from the CPIAUCSL variable.

We next perform binary classification to predict whether the monthly return of SPY will be positive (1) or non-positive (0). Four methods are applied: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Logistic Regression. Hyperparameters are tuned for each model: the best K for KNN, the best sensitivity parameter C for SVM, and k-fold cross-validation for Random Forest and Logistic Regression, this in order to find the best prediction model. These classification methods were selected to explore a mix of simple, interpretable models (Logistic Regression, KNN) and more complex, flexible models capable of capturing nonlinear patterns (SVM with RBF kernel, Random Forest), allowing us to compare performance and robustness across different approaches. The table below summarizes the train and test accuracies for each method, including the

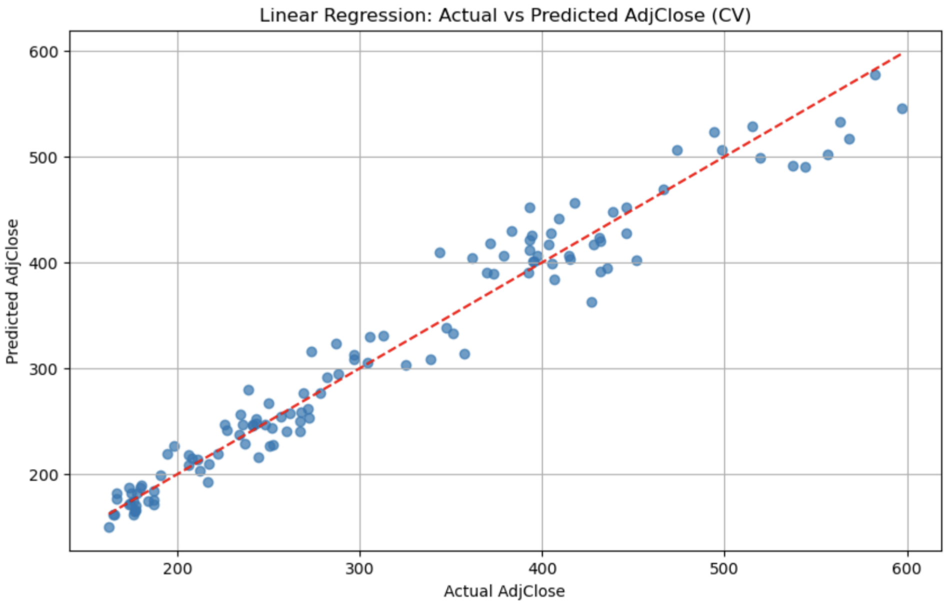

Figure 4: Predicted vs Actual SPY monthly adjusted closing prices using linear regression with a 75/25 train/test split and 5-fold cross-validation.

selected hyperparameters.

	KNN	SVM	RandomForest	Logistic Regression
Parameter	$K = 10$	$C = 10^{-6}$	6-fold CV	4-fold CV
Train Accuracy	0.7111	0.6778	0.6778	0.6779
Test Accuracy	0.7667	0.7333	0.7333	0.7000

Table 4: Binary classification results for predicting positive monthly SPY returns, including the main model parameters.

The classification results are worse than linear regression because predicting only the direction of monthly returns (up or down) discards information about the actual magnitude of price changes, making the problem inherently harder and less precise.

3.3 Portfolio Strategy and Optimization

We perform a portfolio optimization comparing SPY with two strategies: Top 50 and Bottom 50 daily performers. Each day, the strategies select the 50 best- and worst-performing stocks based on the previous days returns. Once the stocks are selected, their weights for the portfolio are optimized using the mean and covariance of daily returns from the previous year, in order to maximize the Sharpe ratio, which is a measure of risk-adjusted return calculated as the ratio of the portfolios excess return over the risk-free rate to its standard deviation

$$\text{Sharpe Ratio} = \frac{E[R_p] - R_f}{\sigma_p}$$

where $E[R_p]$ is the expected portfolio return computed from historical data as the mean of daily portfolio returns over a year, R_f is the risk-free rate which we chose to fix at 0.2, and σ_p is the portfolio standard deviation which is computed from the covariance matrix of historical returns of the selected assets, combined with the portfolio weights:

$$\sigma_p = \sqrt{\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}}$$

where $\mathbf{w}$ is the vector of portfolio weights of previous year and $\mathbf{\Sigma}$ is the covariance matrix ofreturns. For the first year, no previous year data is available, so all selected stocks are assigned equal weights.

The optimization is subject to the following constraints:

- Initial capital: \$1,000,000.
- Number of trading days per year: 252.
- Long-only positions: weights are bounded between 0.0 and 0.40 per stock woth total weights sum to 1.

This approach ensures that the portfolio both reflects historical risk-return characteristics (through previous year statistics) and dynamically reacts to short-term stock performance (through daily selection). We show now the yearly returns % of the two strategies with respect to the SPY:

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Top50 (%)	2.71	9.23	8.55	-12.81	49.82	-1.21	40.44	-9.16	46.74	10.68
Bottom50 (%)	7.84	28.61	41.95	3.92	52.38	43.51	75.10	-6.68	27.37	70.45

Table 5: Yearly returns for Top 50 and Bottom 50 daily stock selection strategies.

For the Top 50 Strategy: Final Capital = \$3,255,858.18, Total Return = 225.59%, while for the Bottom 50 Strategy: Final Capital = \$15,872,552.64, Total Return = 1487.26%. We finally show the total performance over the 10 years of analysis. The daily returns of each strategy were computed year by year: for the first year, equal weights were assigned to all stocks, while for subsequent years, the Top 50 and Bottom 50 stocks were selected based on the previous years mean returns, and their portfolio weights were optimized using the historical covariance matrix. The daily portfolio returns were then aggregated across all years, and annualized performance metrics, including annual return, annualized volatility, and Sharpe ratio, were calculated for each strategy and for SPY to provide a comprehensive comparison:

Metric	Top 50 Gainers	Bottom 50 Losers	SPY
Annual Return	0.2393	0.3478	0.1481
Annual Std Dev	0.2181	0.2754	0.1762
Sharpe Ratio	1.0051	1.1902	0.7268

Table 6: Performance metrics for the Top 50 Gainers, Bottom 50 Losers, and SPY.

The strategy with highest Sharpe Ratio is the Bottom 50 Losers. The plot of the capital growth for the SPY and the two strategies is given in Figure 5.

Finally, for the Bottom 50 strategy, we illustrate the Pareto frontier in Figure 6.

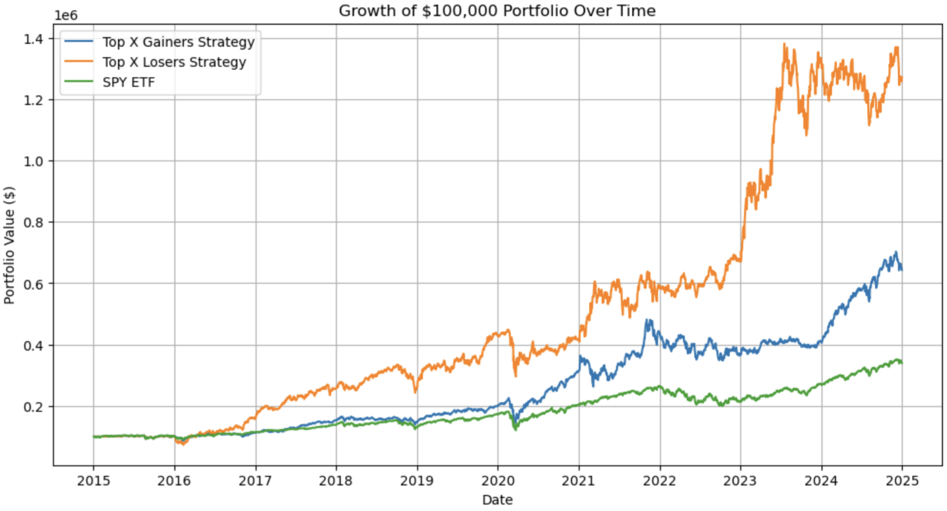

Figure 5: Capital growth for the SPY, Top 50 gainers and Top 50 Losers startegy.

To explore the robustness of the Bottom 50 Losers strategy, we simulate a large number of perturbed portfolios by slightly modifying the original portfolio weights each year. For each perturbation level ϵ (ranging from 0.001 to 1), we generate portfolios where the yearly weights are multiplied by random factors within $[1 - \epsilon, 1 + \epsilon]$, then clipped to a minimum and maximum allowed weight and normalized to sum to 1. Daily returns for each portfolio are computed as the weighted sum of the daily returns of the selected stocks. These simulated daily returns are aggregated over all years to calculate the annualized return, annualized volatility, and Sharpe ratio for each perturbed portfolio, in order to compare with the one we found before. Limiting the perturbation avoids exploring the full weight space, which would require an excessive number of simulations. In this example, we provided 500000 simulations.

We can appreciate how the optimized Sharpe ratio we found lies on the efficient frontier, the boundary of best optimized risk-return combinations. The center of the Pareto frontier is more densely populated because small random perturbations around the optimized weights tend to generate portfolios with moderate risk-return tradeoffs, while reaching the extreme high-return or low-volatility regions requires highly specific weight allocations that are much less likely to occur by chance.

3.4 Risk Analysis

We now evaluate the 10-day Value-at-Risk (VaR) of the Top 50 Losers Strategy using three approaches: historical simulation, parametric estimation, and Monte Carlo, in order to compare different risk measurement methods and assess the potential losses under both empirical and model-based assumptions.

Value-at-Risk (VaR) is a standard risk measure that estimates the potential loss of a portfolio over a given time horizon at a specified confidence level. In simpler terms, a 95% 10-day VaR of 10,000 USD means that there is a 5% probability that the portfolio will lose more than 10,000 USD over the next 10 days. For the three methods methods:

- Historical simulation: uses the empirical distribution of past returns.

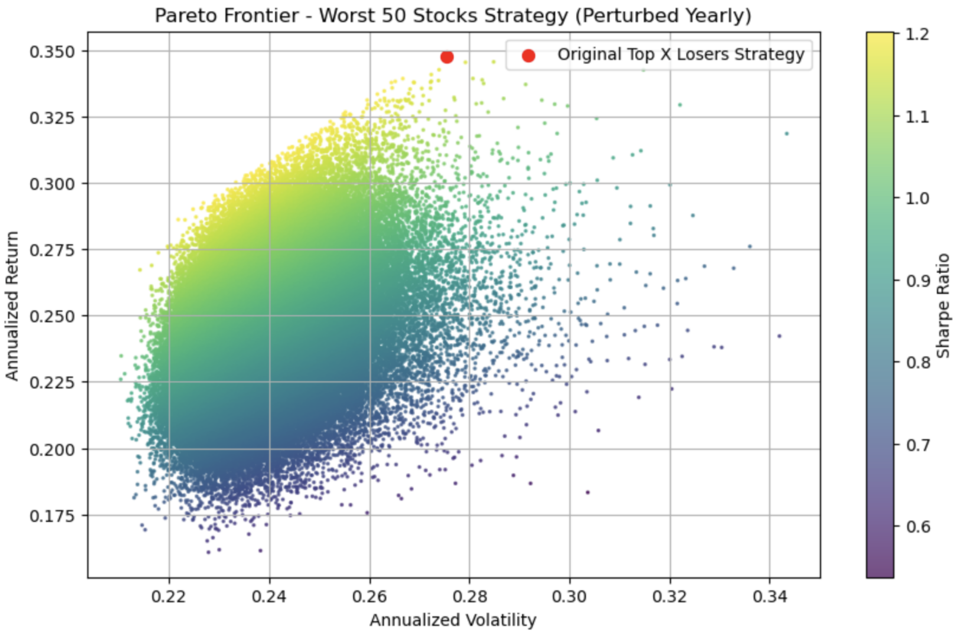

Figure 6: Pareto frontier of annualized return versus volatility for the Bottom 50 strategy, showing the effect of yearly weight perturbations and highlighting the original strategys performance.

- Parametric estimation: assumes a specific distribution (generalized normal) and computes VaR analytically.
- Monte Carlo simulation: generates random portfolio return scenarios based on the assumed distribution to estimate the loss quantile.

Given the daily returns for the Top 50 losers startegy, we can easily fit the best distribution and find the optimal parameters for it. The portfolio value is set to $100,000, with a 10-day risk horizon, a two-year (252 Œ 2 days) rolling calibration window, and a 95% confidence level. At each step, we compute the 10-day cumulative returns from the calibration window, derive the VaR, and then verify whether the realized portfolio loss exceeds the predicted threshold.

The backtesting results are summarized in Table 7. For the historical method, 106 exceedances were observed out of 2002 checks, corresponding to 5.29%, very close to the expected 5% level. The parametric approach, based on fitting a generalized distribution, produced only 47 exceedances (2.35%), suggesting that the model is too conservative. The Monte Carlo method, also using the fitted generalized distribution, gave 61 exceedances (3.05%), again under the expected level but closer to the theoretical 5% target.

Overall, the historical method aligns most closely with the nominal exceedance rate, while the parametric and Monte Carlo methods tend to underestimate tail risk. The reason is that the parametric and Monte Carlo methods are limited by how well the fitted distribution matches reality. Even if the generalized distribution is flexible, the fit may not fully capture the true skewness, kurtosis, or clustering of extreme losses in financial returns. As a result, both methods smooth over tail risks and systematically underestimate exceedances compared to the historical method.

	Historical	Parametric (Gen. Dist.)	Monte Carlo (Gen. Dist.)
Total 10-day VaR checks	2002	2002	2002
Number of exceedances	106	47	61
Percentage exceedances	5.29%	2.35%	3.05%

Table 7: 10-day VaR backtesting results at 95% confidence level for the Top X Losers Strategy.

For illustrative purposes only, we show in Figure 7 the histogram of the daily returns for the first day of analysis, overlaying the 95% 10-day Monte Carlo VaR and the fitted generalized normal distribution.

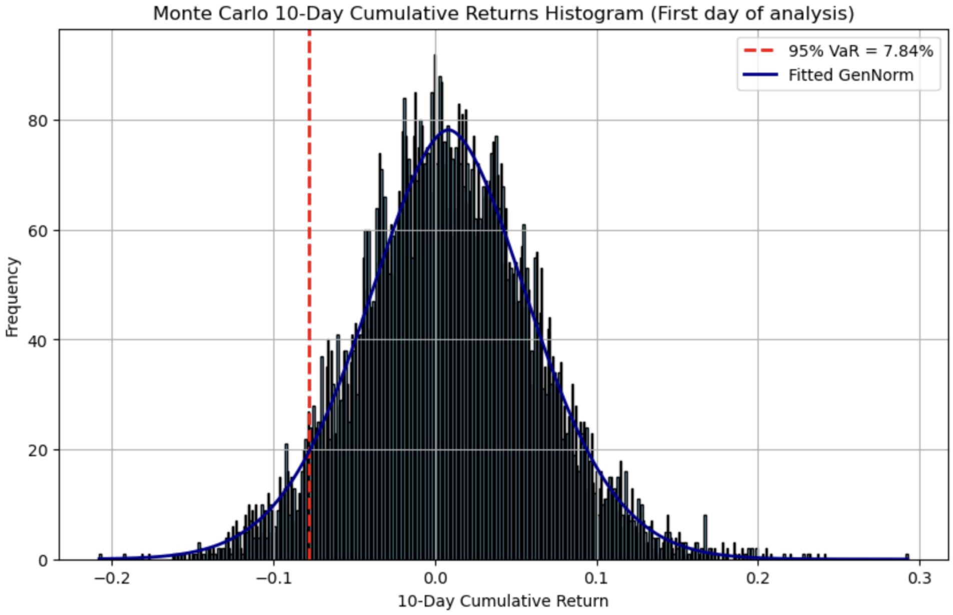

Figure 7: Histogram of daily returns of first day of analysis, with 95% 10-day Monte Carlo VaR (red dashed line) and the fitted generalized normal distribution (blue curve).

The VaR for this single day is higher than the long-term average because it is based solely on the initial calibration window, which may capture slightly more extreme past returns, and because Monte Carlo simulation emphasizes the tail outcomes in this limited sample. Since we have sufficient historical data, using the empirical distribution avoids potential biases from distributional assumptions and gives a more robust measure of risk. As a result, the first-day estimate reflects a conservative risk measure compared to the full historical period.

4 Conclusions

This project provided a comprehensive quantitative analysis of SPY returns and alternative stock selection strategies, combining regression, classification, portfolio optimization, and risk assessment.

Linear regression using macroeconomic indicators showed strong predictive power for actual SPY prices, whereas binary classification of positive monthly returns performed worse, highlighting the information loss inherent in discretizing returns.

The portfolio optimization demonstrated that weighting stocks based on previous-year variance allows the construction of strategies with favorable risk-adjusted performance. Notably, the Bottom 50 losers strategy achieved the highest Sharpe ratio, outperforming both SPY and the Top 50 gainers in terms of risk-adjusted returns. The Pareto frontier analysis confirmed that the optimized allocations lie near the efficient edge, and random perturbations showed that extreme combinations are harder to reach, concentrating outcomes in the central region.

Value-at-Risk analysis using historical, parametric, and Monte Carlo methods revealed that historical VaR closely aligns with the expected 5% exceedance, while parametric and Monte Carlo approaches slightly underestimate risk due to distributional assumptions. First-day Monte Carlo VaR was higher than average because early samples may reflect less smoothed return variability.

Overall, the project highlights the importance of careful strategy design, risk measurement, and the limitations of model assumptions. The combination of statistical modeling and portfolio optimization provides actionable insights into both predictive and risk-aware investment strategies.

The analysis could be improved by incorporating additional macroeconomic or market features to enhance predictive accuracy and by using dynamic, time-varying portfolio weights to better capture changing market conditions. Additionally, employing alternative distributional assumptions or stress-testing scenarios in the VaR calculations could provide a more robust assessment of tail risk.

A Python Source Code

Please refer to the Python Source Code file in the GitHub webpage:
https://github.com/AlonzoDiazAvalos/Data-Analytics