Predictive Power Comparison between Technology and Healthcare Sectors Over SPY ETF

By:

Abhishek Patil

Introduction

Financial markets exhibit complex dynamics driven by a multitude of factors including economic conditions, investor psychology, political events, corporate earnings, and more. Understanding and accurately forecasting market movements is an extremely challenging task that has significant value for investment planning and risk management.

Advances in machine learning and the availability of extensive historical financial data have enabled new data-driven approaches for modeling and predicting future price movements of various securities and broader market indexes. In particular, deep neural networks have recently shown promise in discovering subtle patterns in financial time series data.

This project aims to develop machine learning models for a specific financial forecasting task - predicting the Standard & Poor's 500 index (S&P 500) Daily Return using the historical daily returns of technology and healthcare stocks. The S&P 500, which consists of 500 large US companies, is among the most commonly used benchmarks for overall US stock market performance.

Accurately forecasting the S&P 500 can provide key signals on forthcoming market regime shifts, allowing for timely portfolio adjustments. This project tries to capture such predictive relationships using machine learning algorithms.

We train a number of regression models on historical daily returns data for several major technology and healthcare stocks to forecasting performance of S&P 500 exchange-traded-fund (SPY ETF). Comparing model performance over multiple time horizons can reveal which sectors exhibit the strongest signals and lead-lag effects for broader markets. The best performing models can serve as valuable predictive indicators for S&P 500 movements.

Effective financial forecasting has been an active domain of fintech innovation and research. This project aims to further that initiative by demonstrating machine learning capabilities for an important predictive modeling task - forecasting market index moves using sector-specific stock price data. The techniques developed could be extended to predict other asset prices and optimized for operational investment strategies.

DATA

The historical price data powering our analysis and models consists of daily closing prices for four technology stocks, four healthcare stocks, and the S&P 500 ETF (ticker: SPY) representing broad US equity market performance.

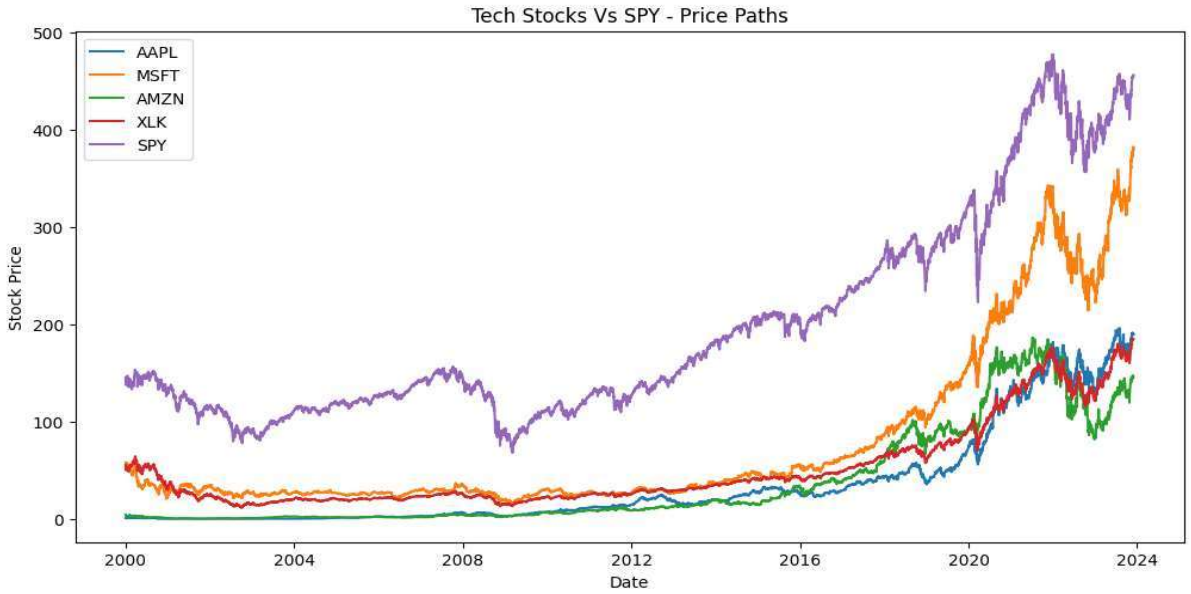
Date	AAPL	AMZN	JNJ	LLY	MSFT	SPY	UNH	XLK	XLV
2000-01-03	0.999442	4.468750	46.09375	65.5625	58.28125	145.4375	6.718750	55.43750	30.296875
2000-01-04	0.915179	4.096875	44.40625	63.5000	56.31250	139.7500	6.632813	52.62500	29.609375
2000-01-05	0.928571	3.487500	44.87500	64.3125	56.90625	140.0000	6.617188	51.84375	29.343750
2000-01-06	0.848214	3.278125	46.28125	66.1875	55.00000	137.7500	6.859375	50.12500	29.437500
2000-01-07	0.888393	3.478125	48.25000	71.0625	55.71875	145.7500	7.664063	51.00000	29.796875

Specifically, we retrieve a time series of daily adjusted closing prices from January 2000 to December 2023 for the following securities:

Technology Stocks:

- Apple (AAPL)
- Microsoft (MSFT)
- Amazon (AMZN)
- Technology Select Sector SPDR Fund (XLK) - ETF benchmarking tech sector

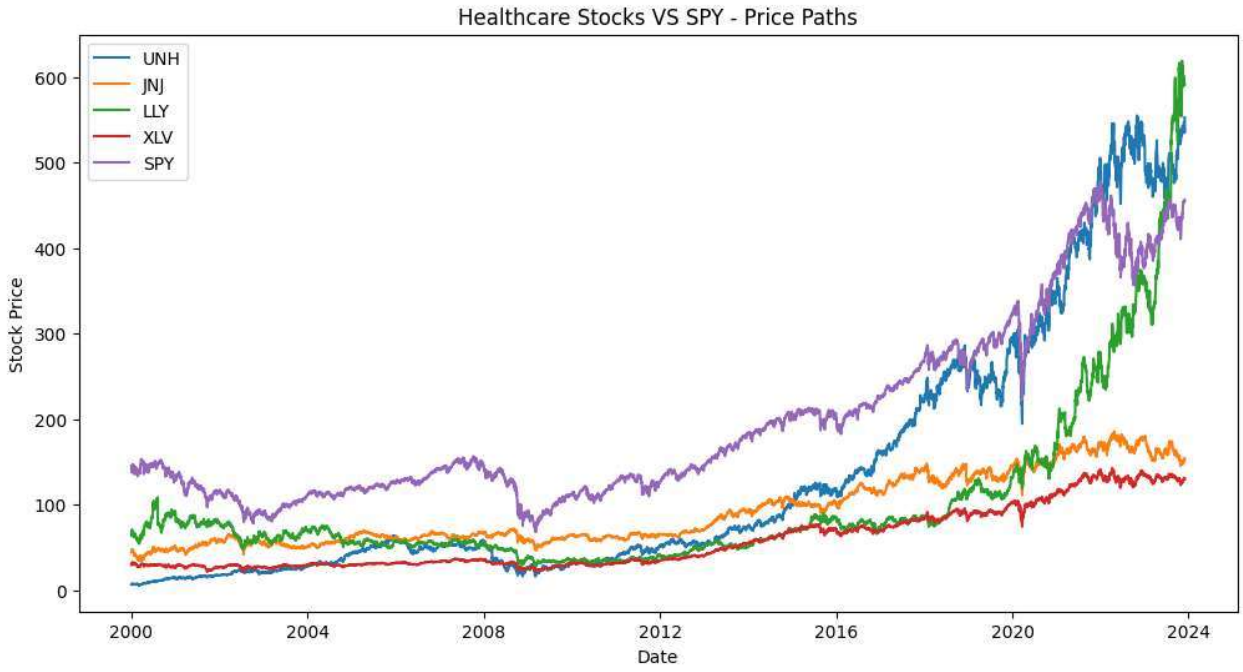
Price path of Tech Stocks Vs SPY:



Healthcare Stocks:

- UnitedHealth Group (UNH)
- Johnson & Johnson (JNJ)
- Eli Lilly and Company (LLY)
- Health Care Select Sector SPDR Fund (XLV) - ETF benchmarking healthcare sector

Price path of Tech Stocks Vs SPY:



S&P 500 ETF:

- SPDR S&P 500 Trust ETF (SPY)

This collection of technology and healthcare stocks represents a diverse mix of influential companies and benchmarks tracking the broader performance within each market sector. The time range encompasses varying market environments and economic cycles.

The prices for these stocks and SPY ETF across 2000-2023 serves as the feature dataset (X matrix) to train our machine learning models. The corresponding SPY prices represent the target variable (y vector) we aim to predict.

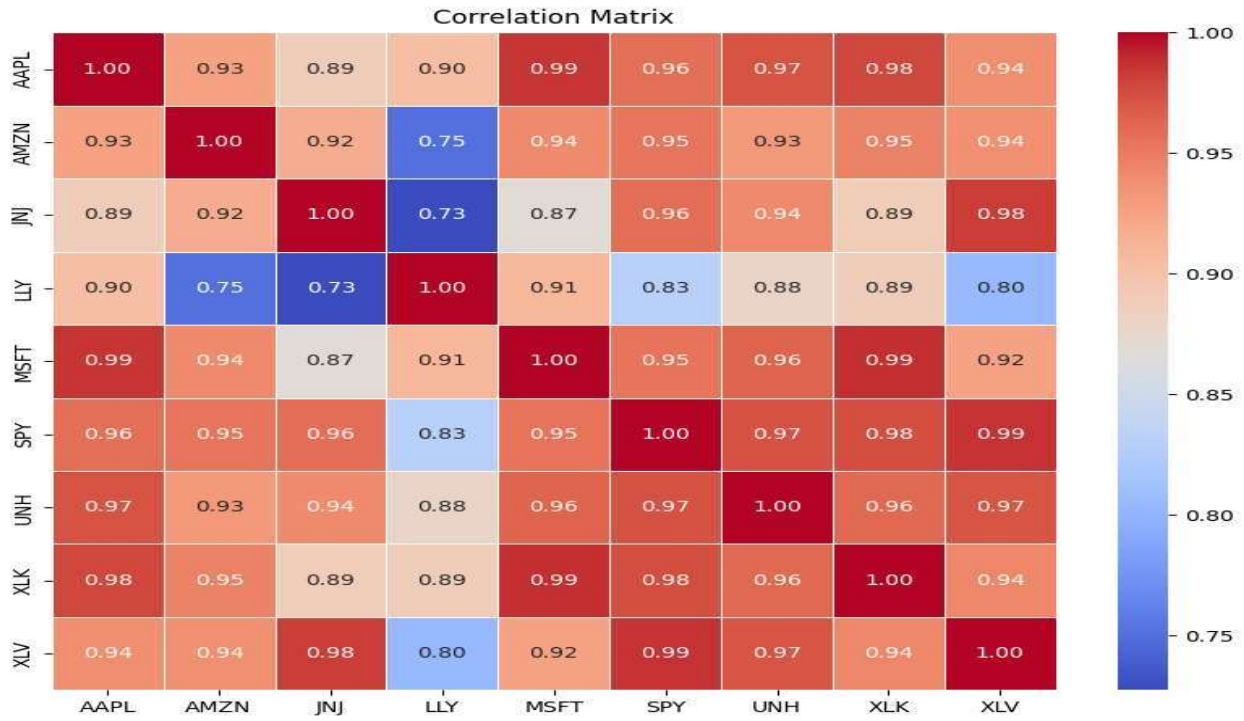
As standard practice, we convert the raw closing prices to percentage changes to stationaries the time series before fitting our models. We set the preprocessed sequences of daily returns for the technology and healthcare stocks as inputs to predict next day SPY returns.

In total, after preprocessing we have over 6000 training examples spanning a 20+ year period. This extensive historical dataset allows testing model performance over long time horizons and evaluating how predictive relationships evolve over market cycles. Feature engineering and model optimization is focused on extracting predictive signals from the sector stocks that reliably anticipate movements in broader market gauges like SPY.

Summary Statistics of the dataset:

	AAPL	AMZN	JNJ	LLY	MSFT	SPY	UNH	XLK	XLV
count	6017.000000	6017.000000	6017.000000	6017.000000	6017.000000	6017.000000	6017.000000	6017.000000	6017.000000
mean	35.334492	37.537064	91.899985	99.183103	78.037451	196.410529	134.113016	50.819410	58.113442
std	50.310012	51.221854	40.797379	96.379315	87.433142	104.740077	149.531636	42.784285	35.258414
min	0.234286	0.298500	34.250000	27.469999	15.150000	68.110001	5.953125	11.580000	21.879999
25%	2.146786	2.261000	60.009998	52.480000	27.049999	119.290001	30.920000	21.860001	30.125000
50%	14.395357	10.135000	67.599998	66.809998	32.689999	145.750000	55.000000	30.250000	36.189999
75%	40.580002	58.417999	129.110001	86.389999	84.559998	256.769989	215.479996	63.480000	82.089996
max	196.449997	186.570496	186.009995	619.130005	382.700012	477.709991	555.150024	185.449997	142.830002

Correlation Analysis:



Timeframes

Daily Return predictions were tailored to four distinct timeframes:

- **Monthly**

  Predicting on a monthly basis caters to short-term investors seeking swift insights into market dynamics and potential opportunities

- **Quarterly**

  A quarterly timeframe accommodates investors with a slightly longer horizon, allowing for a more strategic approach while mitigating short-term volatility.

- **Semi-Annually**

  This timeframe provides a balanced outlook for investors seeking a middle ground between short-term and long-term perspectives.

- **Annually**

  The annual prediction horizon is geared towards long-term investors, offering insights into broader market trends and potential for sustained growth.

Models Employed

In our pursuit of accurate price forecasting, four distinct machine learning models were employed:

Multiple Linear Regression:

- Multiple Linear Regression (MLR) attempts to model the relationship between two or more explanatory variables (x) and a response variable (y) by fitting a linear equation to observed data.
- A key advantage of MLR is its interpretability - the regression coefficients directly indicate the influence of each predictor variable on the target. Additionally, it performs well when there are actually linear relationships and correlations in the data.
- For our analysis, the OHLC prices, volume, and moving averages serve as the predictor variables while the SPY closing price is the response variable. By exploiting the linear interdependencies in the stock data, the MLR model can make reasonably accurate predictions.

Decision Tree Regression:

- Decision Tree Regression (DTR) utilizes a tree representation to map the 'decisions' between target variable splits based on input variable thresholds. It breaks down a dataset into smaller subsets while incrementally developing an associated decision tree.
- Key advantages are handling nonlinear relationships in data as well as being easy to interpret. However, they are prone to overfitting on noise in training data. Random Forest regularization helps mitigate this limitation.
- For stock data, decision trees can model complex combinations of technical indicators and price data that drive movements.

Random Forest Regression:

- Random Forest Regression (RFR) constructs an ensemble of decision trees during training and then averages the predictions from all the trees during prediction. It introduces randomness into tree creation and aims to create de-correlated trees that can make independent errors.
- This ensemble approach provides more accurate and robust predictions compared to a single decision tree. It also avoids overfitting problems and handles non-linear data well. Hence it is well suited to noisy stock market data.

- By aggregating signals from a diverse set of decision trees it can capture intricate patterns driving stock fluctuations.

K-Nearest Neighbors Regression:

- The KNN algorithm is based on finding the k most similar instances in feature space and having those neighbors vote on the outcome. KNN Regression takes the mean outcomes of the k nearest data points to determine its prediction.
- A key advantage of KNN is that it makes no assumptions about structure of data, rather relying on proximity to make predictions. This also makes it adaptable to different data distributions. However, performance depends heavily on distance metric used.
- For stock data, patterns emerging in recent history are likely to influence short term future values. KNN exploits this to make its predictions.

These models were chosen as they can handle both linear and complex non-linear relationships between variables, which is expected in financial data. The models take past price data for the technology and healthcare stocks as input features and are trained to predict the next day's SPY return percentage.

Separate models are trained on the technology and healthcare company price data to compare how well each sector can anticipate market moves. We use expanding windows of historical data to make 1-month, 3-month, 6-month and 1-year out-of-sample SPY return forecasts and evaluate directional accuracy.

Performance Comparison:

Key performance metrics calculated on the out-of-sample predictions include mean absolute error, R-squared, and mean absolute percentage error. Additionally, directional accuracy evaluates if the model correctly predicts up or down market movements.

These metrics allow comparing predictive capabilities across the different sector models and regression algorithms over various time horizons. In general, the technology stocks demonstrate much stronger predictive signals compared to the healthcare companies across all evaluation criteria. This indicates tech stock movements tend to precede shifts in broader markets.

Feature Selection and Tuning:

In order to improve the performance, we focus on improving predictions of the tech sector and healthcare sector models. Recursive feature elimination (RFE) is applied to select key subset of input tech stocks with best predictive power. Principal component analysis is also tested to denoise and reduce model complexity.

Hyperparameter tuning through grid search methods optimizes model configurations to minimize out-of-sample errors. Overall, feature selection and parameter tuning further bolsters tech model performance significantly across all evaluation metrics and time horizons.

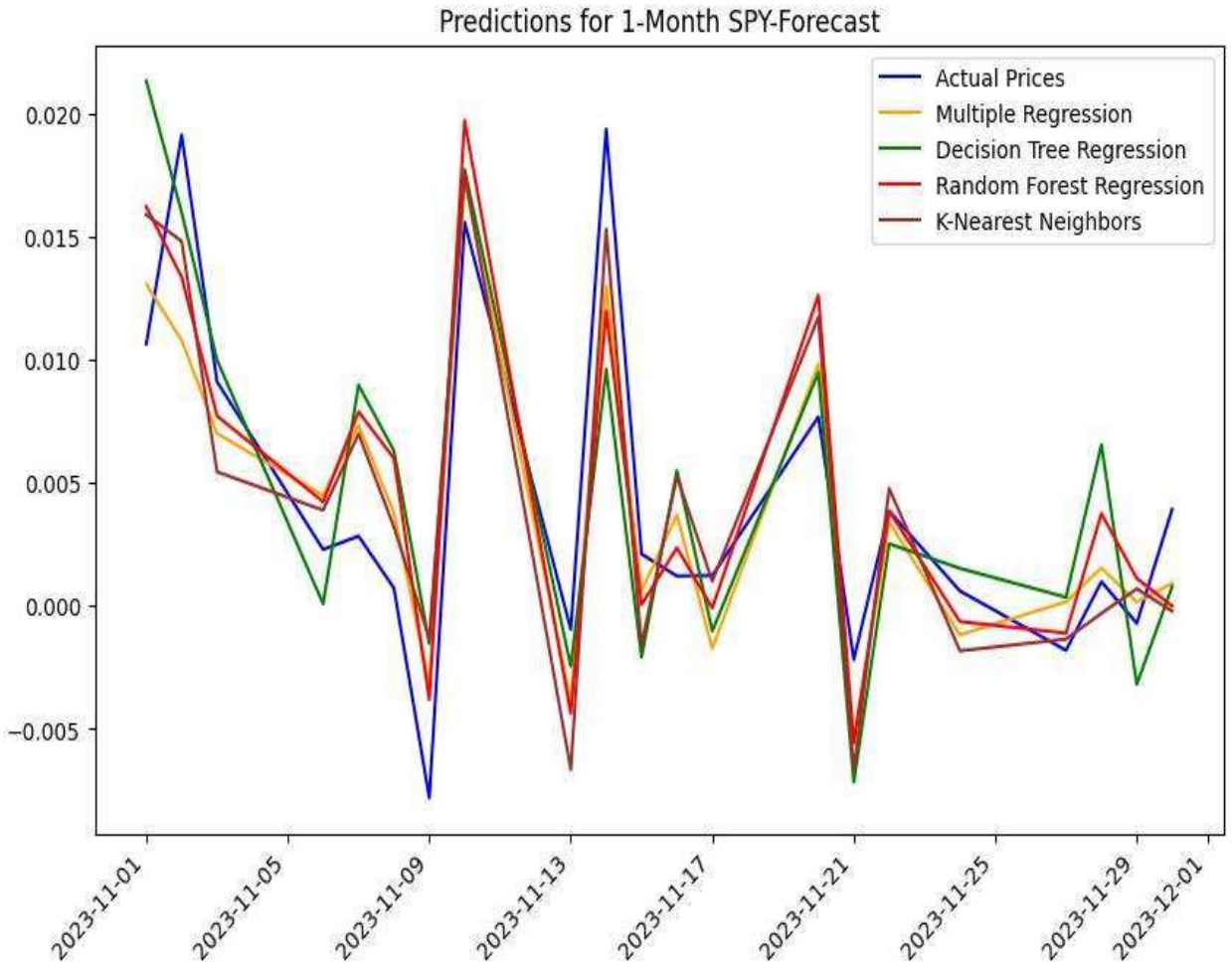
In summary, comparing sector-specific models and applying optimization techniques improves insights into leading indicators and ideal model architectures for predicting market movements using stock price data.

RESULTS:

Tech sector

Results:

1-Month Tech Sector (Just Returns):

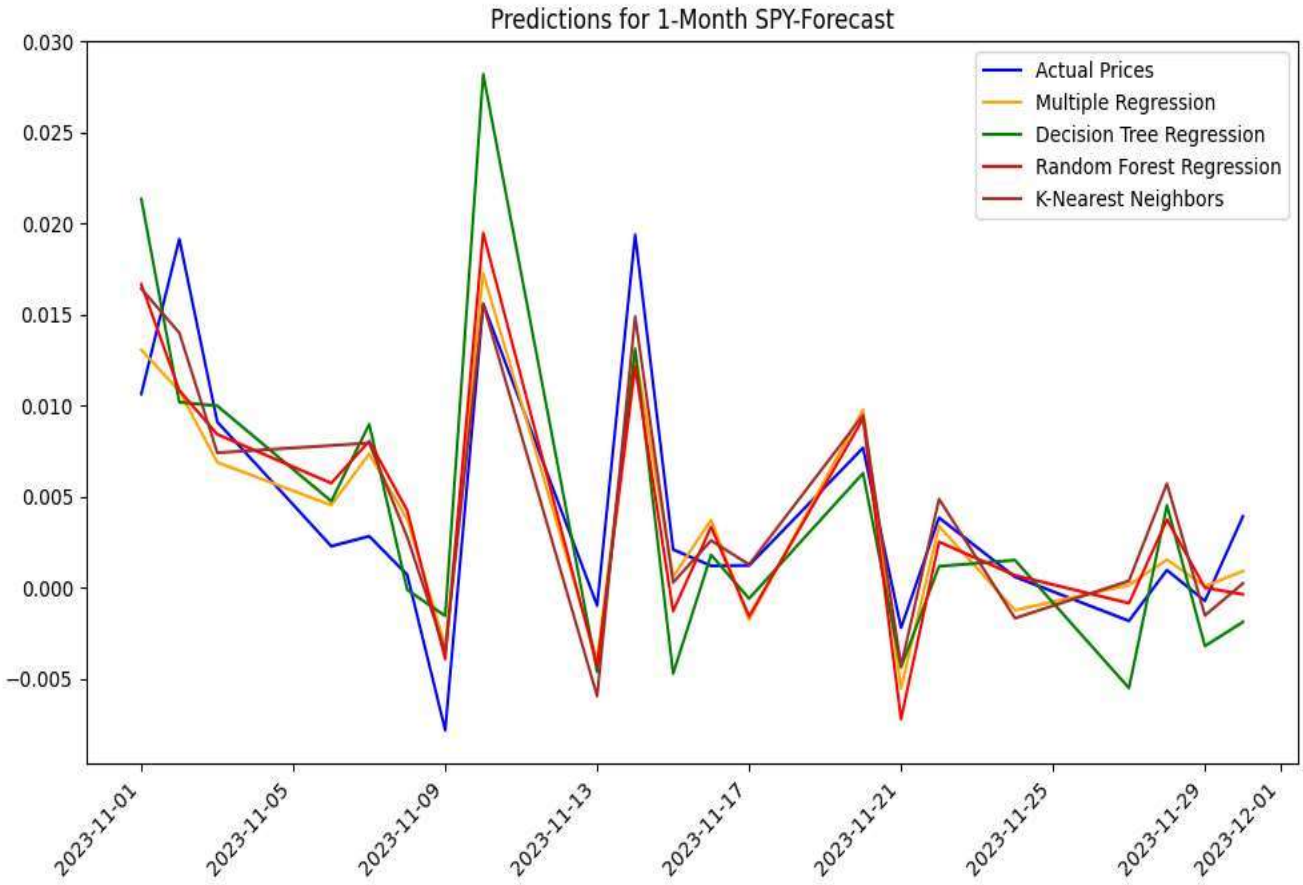


Performance Metric:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.002829	0.000011	0.003373	0.760470	120.681993
Decision Tree	0.003880	0.000022	0.004700	0.534828	180.164541
Random Forest	0.003207	0.000014	0.003749	0.704082	140.617547
K-NN	0.003187	0.000013	0.003622	0.723769	145.045110

1-Month Tech Sector(With RFE):

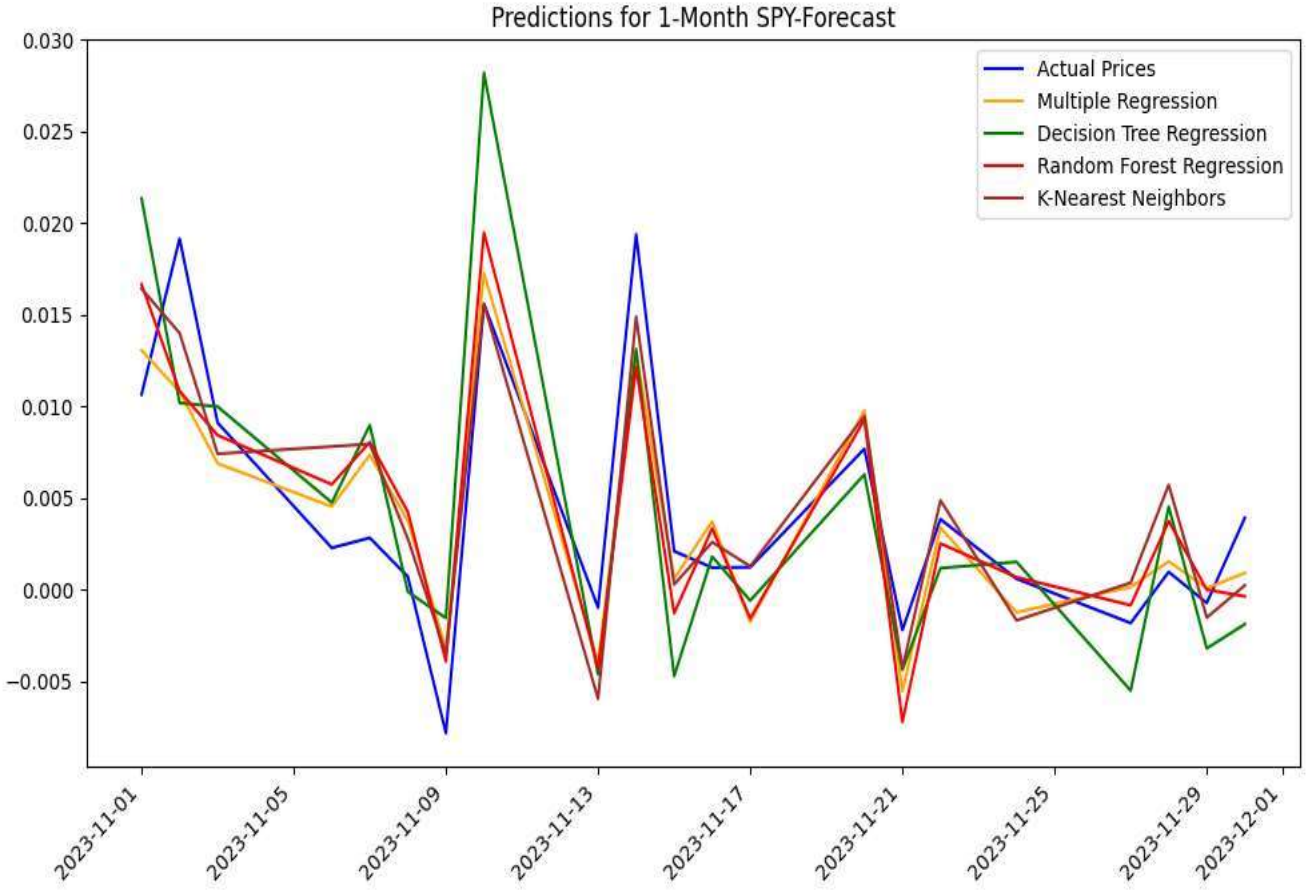


Performance metrics:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.002837	0.000011	0.003382	0.759192	121.041025
Decision Tree	0.004316	0.000030	0.005438	0.377264	146.859904
Random Forest	0.003365	0.000016	0.003976	0.667071	132.717000
K-NN	0.002904	0.000012	0.003436	0.751364	139.436615

1-Month Tech Sector (with PCA):

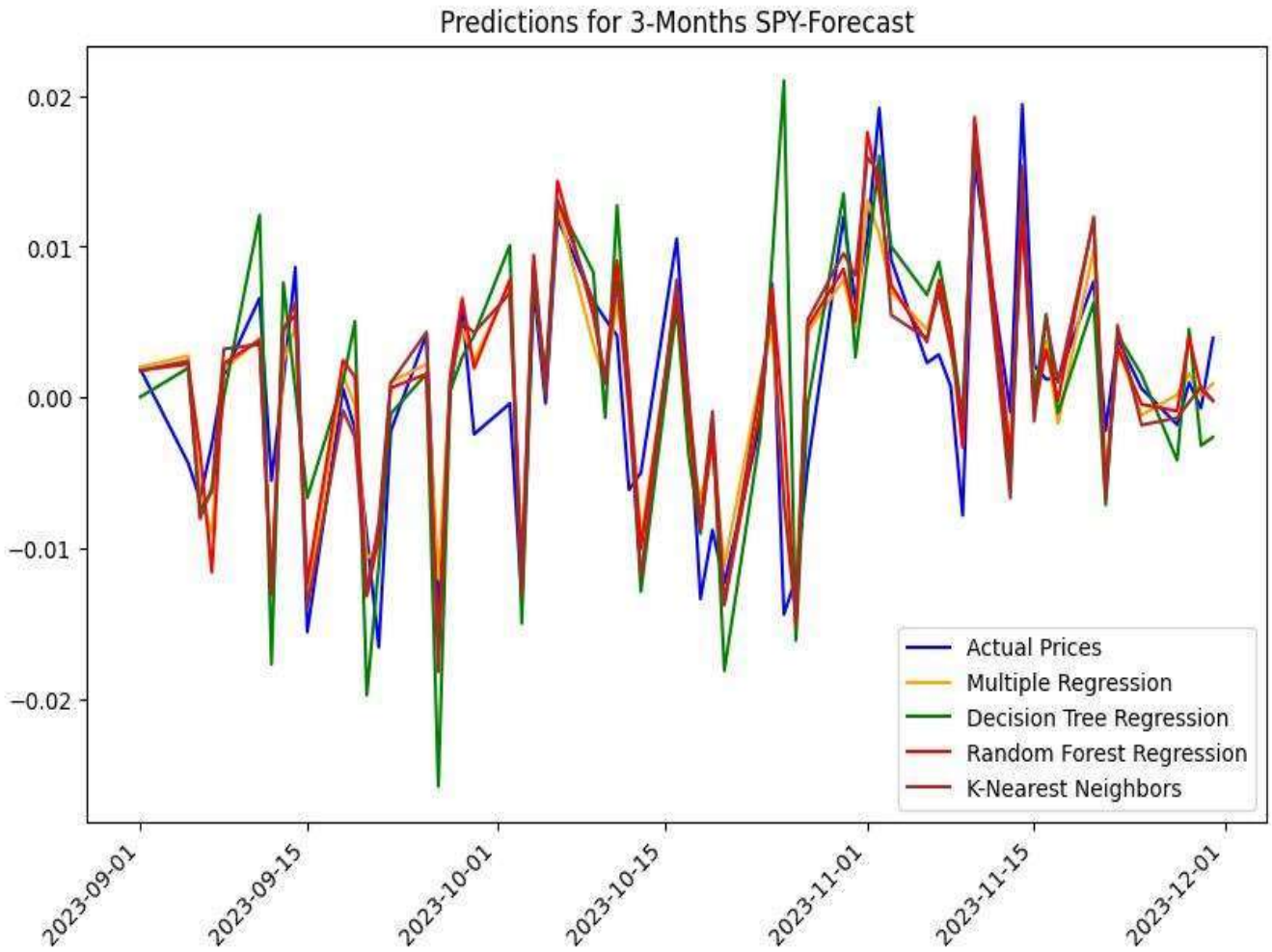


Performance Metrics:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003556	0.000019	0.004388	0.594556	153.226888
Decision Tree	0.004885	0.000037	0.006110	0.213911	213.184795
Random Forest	0.003684	0.000020	0.004475	0.578320	135.273080
K-NN	0.003288	0.000015	0.003922	0.676118	153.908882

3-Month Tech Sector (Just Returns):

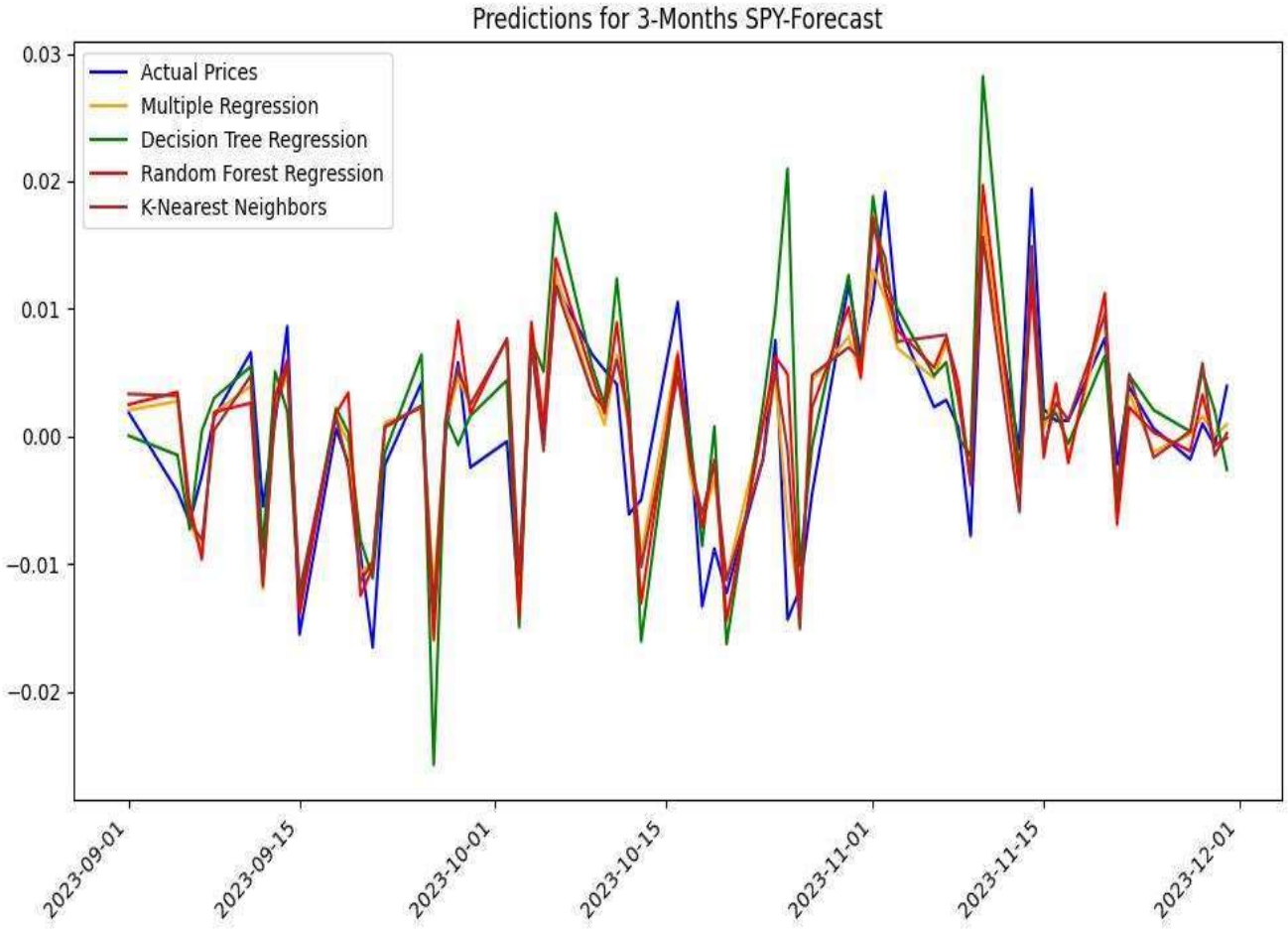


Performance Metrics:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003137	0.000015	0.003843	0.774304	219.477829
Decision Tree	0.004700	0.000046	0.006764	0.300682	298.252994
Random Forest	0.003524	0.000019	0.004322	0.714525	184.241881
K-NN	0.003216	0.000016	0.003998	0.755720	128.647520

3-Month (With RFE):

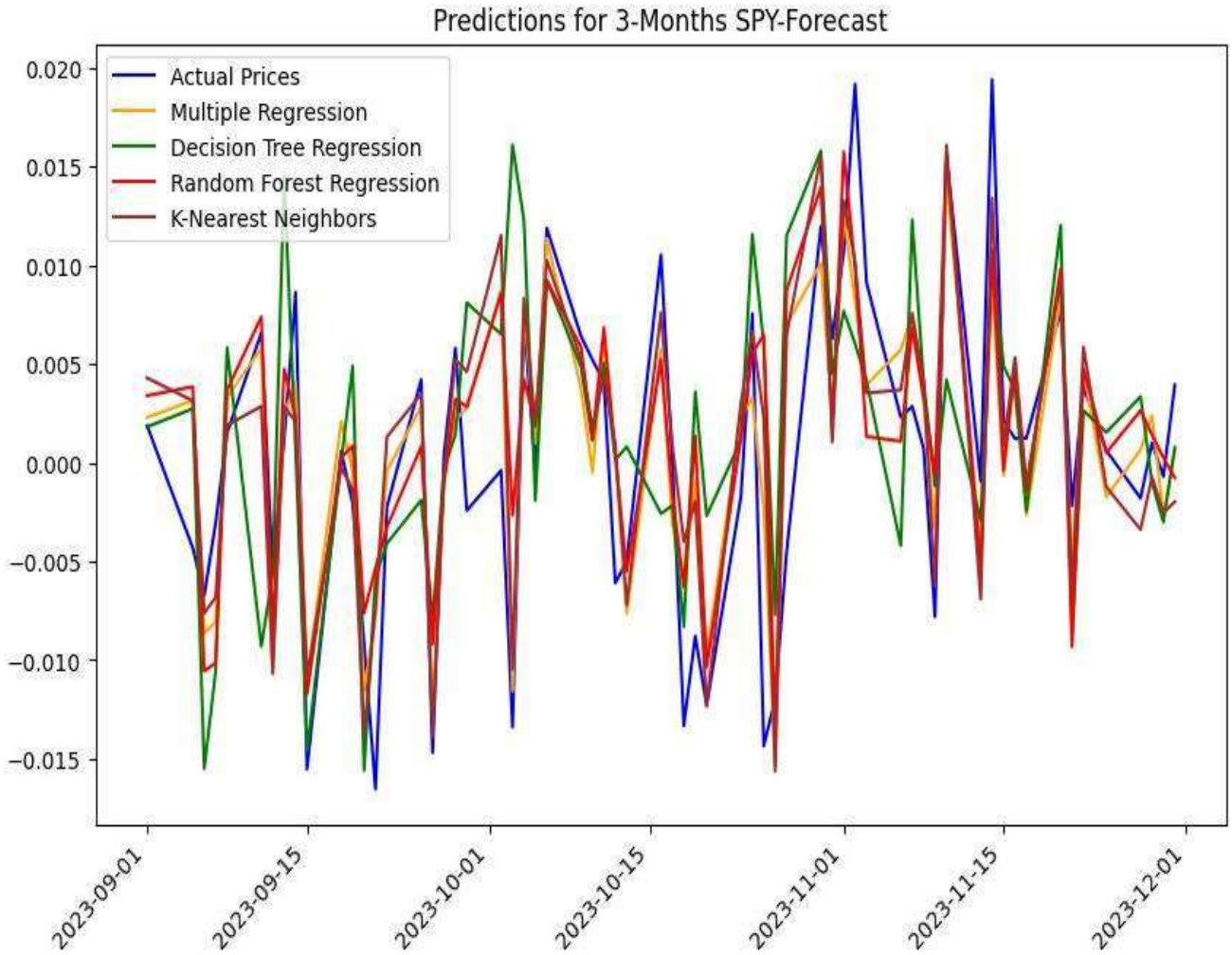


Performance Metrics:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003147	0.000015	0.003855	0.772851	220.891177
Decision Tree	0.004158	0.000041	0.006434	0.367164	144.838748
Random Forest	0.003654	0.000022	0.004724	0.658923	185.342900
K-NN	0.003222	0.000018	0.004189	0.731805	189.824666

3-Month (With PCA):

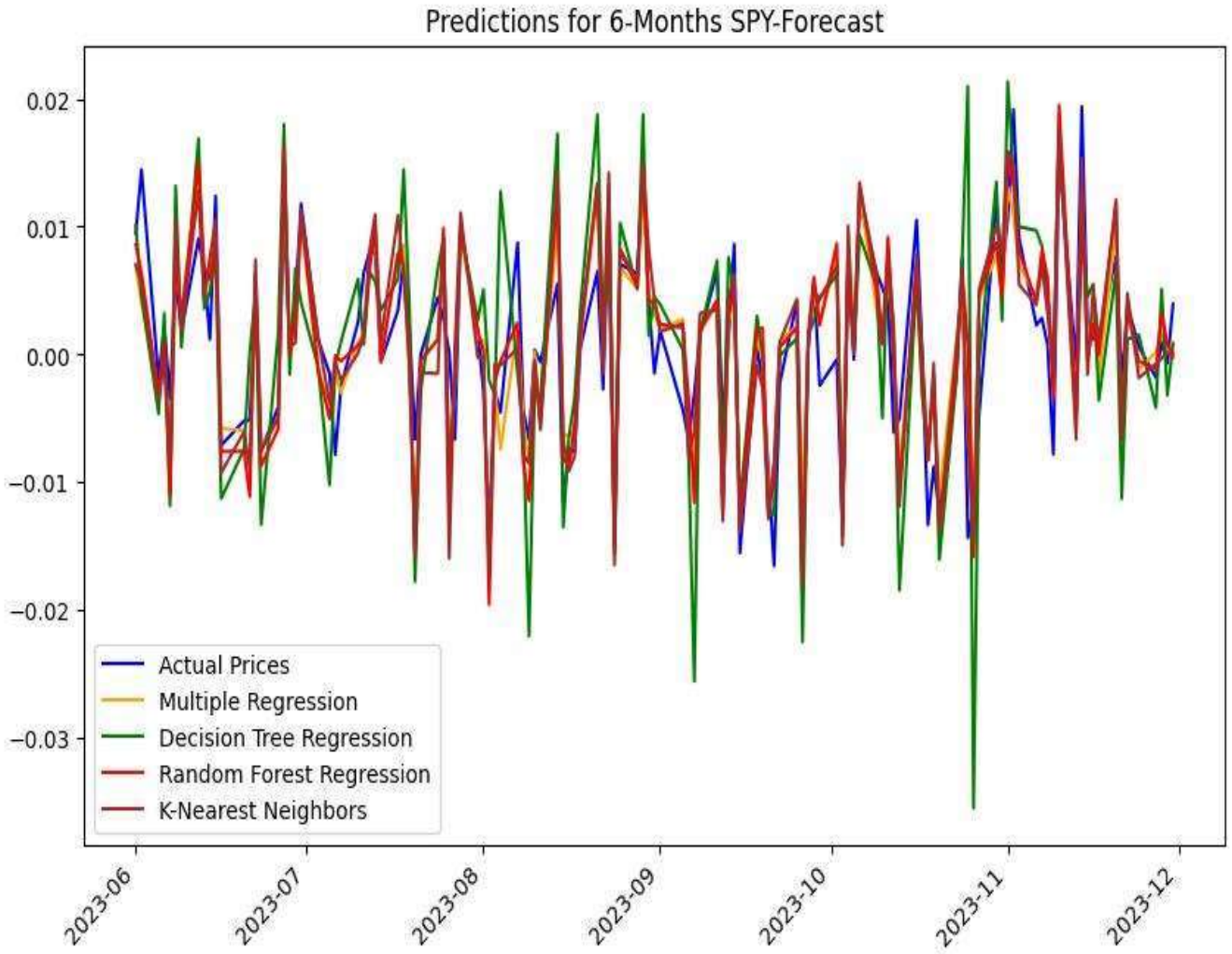


Performance Metrics:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003746	0.000022	0.004671	0.666502	251.063420
Decision Tree	0.006095	0.000065	0.008045	0.010773	256.253046
Random Forest	0.004300	0.000032	0.005661	0.510077	204.185567
K-NN	0.003728	0.000024	0.004924	0.629416	191.339462

6-Months(Just Returns):

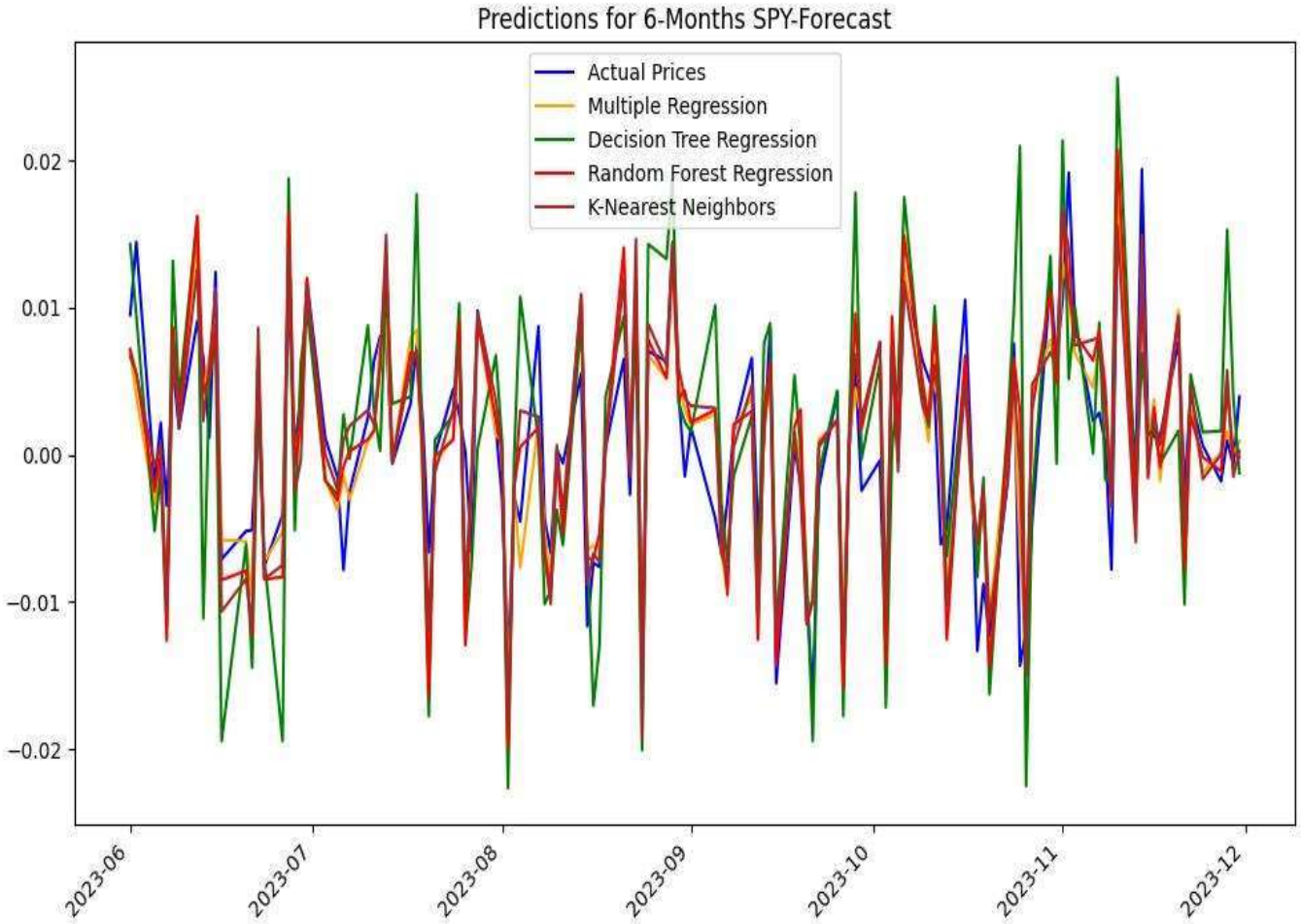


Performance Metrics:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.002932	0.000014	0.003732	0.748391	inf
Decision Tree	0.004939	0.000048	0.006953	0.126779	inf
Random Forest	0.003319	0.000018	0.004264	0.671566	inf
K-NN	0.003326	0.000018	0.004258	0.672582	inf

6-Month (With RFE):

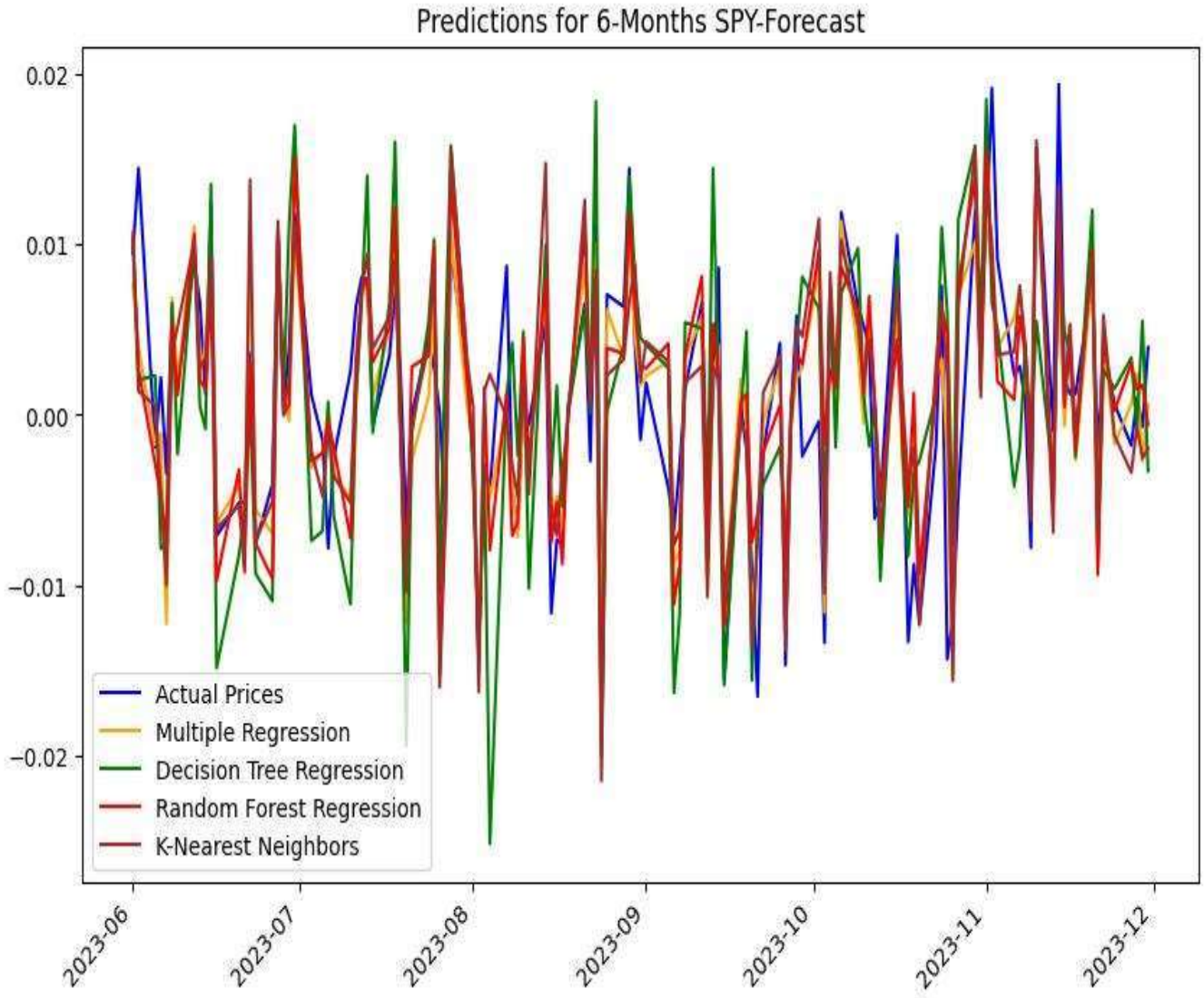


Performance Metrics:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.002936	0.000014	0.003735	0.748056	inf
Decision Tree	0.005322	0.000050	0.007044	0.103796	inf
Random Forest	0.003456	0.000020	0.004485	0.636594	inf
K-NN	0.003180	0.000017	0.004075	0.699990	inf

6-Month (With PCA):

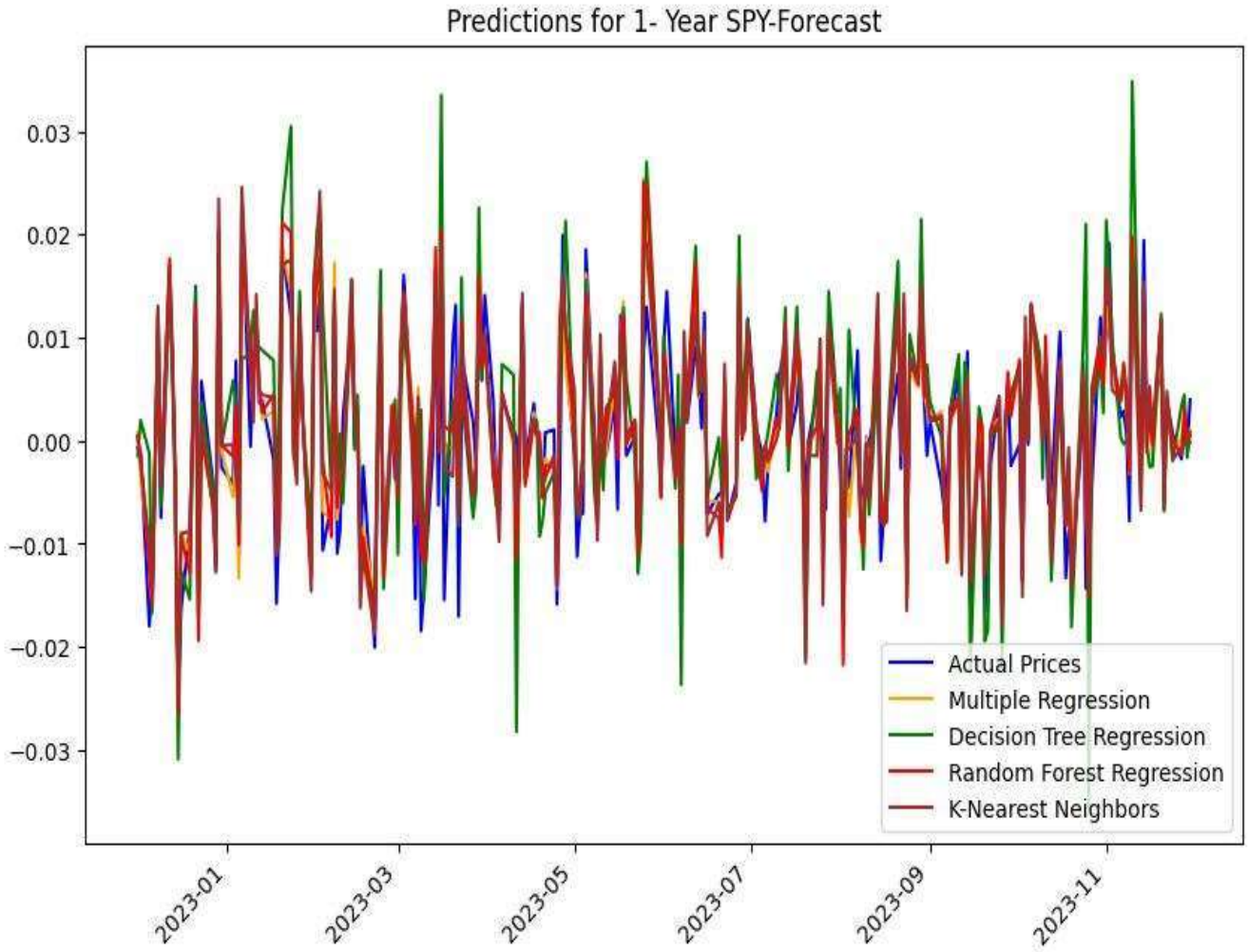


Performance Metrics:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003177	0.000017	0.004175	0.685153	inf
Decision Tree	0.005153	0.000043	0.006524	0.231249	inf
Random Forest	0.003722	0.000025	0.004959	0.555737	inf
K-NN	0.003729	0.000024	0.004921	0.562661	inf

1-year (Just Returns):

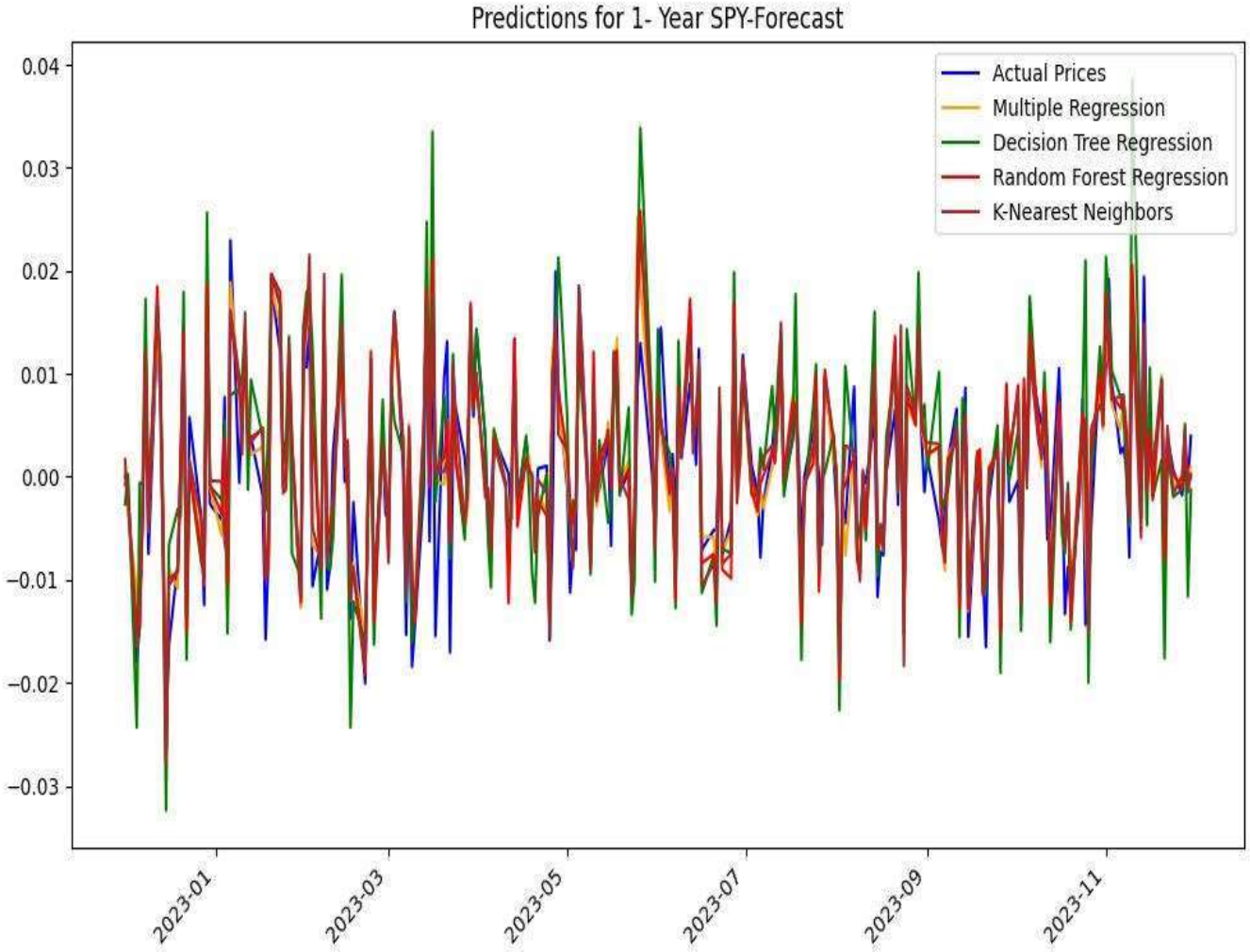


Performance Metrics:

Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003116	0.000017	0.004182	0.773524	inf
Decision Tree	0.005229	0.000051	0.007162	0.335816	inf
Random Forest	0.003395	0.000020	0.004505	0.737188	inf
K-NN	0.003546	0.000021	0.004636	0.721675	inf

1-year (With RFE):

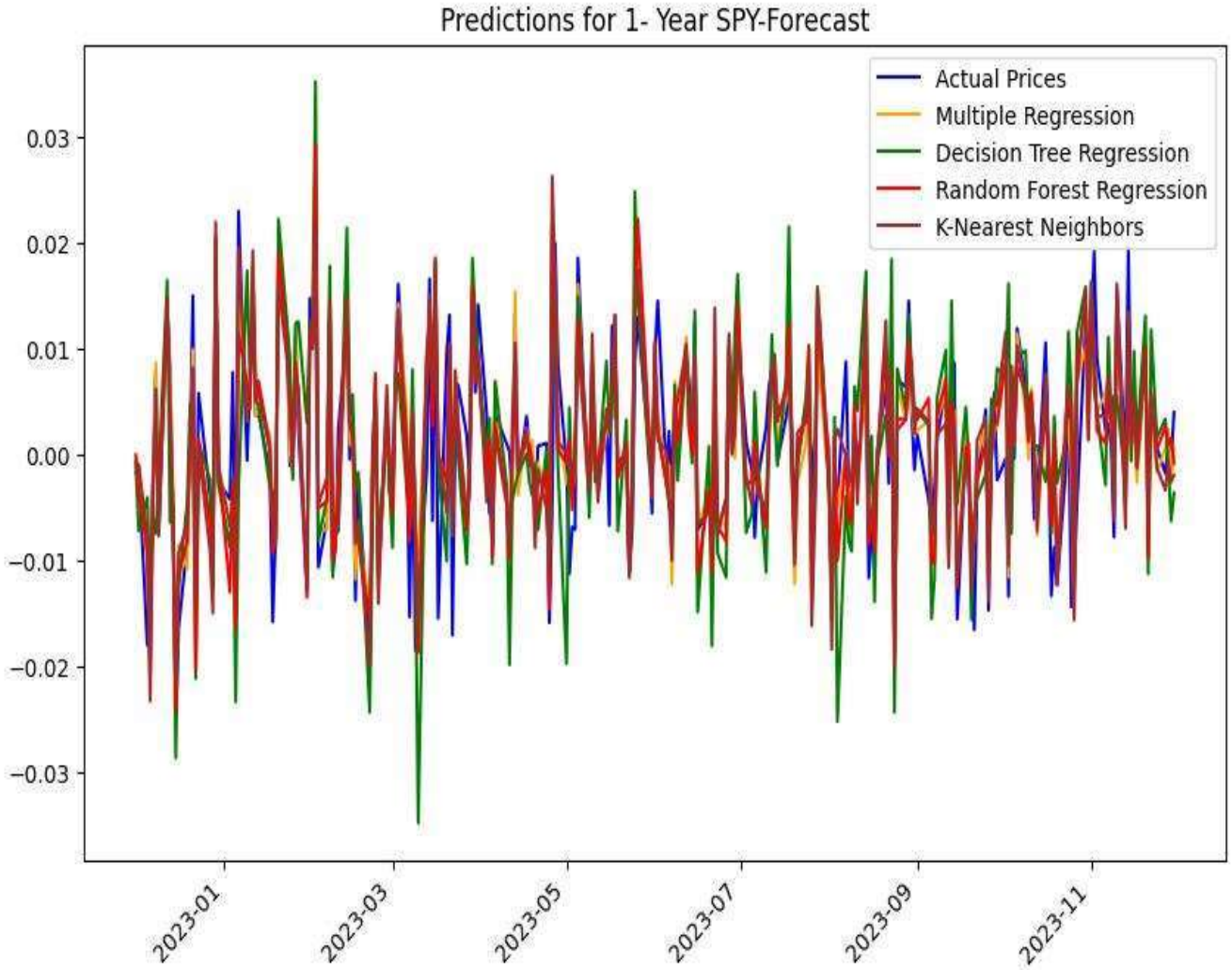


Performance Metrics:

Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003116	0.000017	0.004174	0.774349	inf
Decision Tree	0.005018	0.000046	0.006754	0.409222	inf
Random Forest	0.003513	0.000022	0.004692	0.714882	inf
K-NN	0.003370	0.000020	0.004428	0.746087	inf

1-year (With PCA):



Performance Metrics:

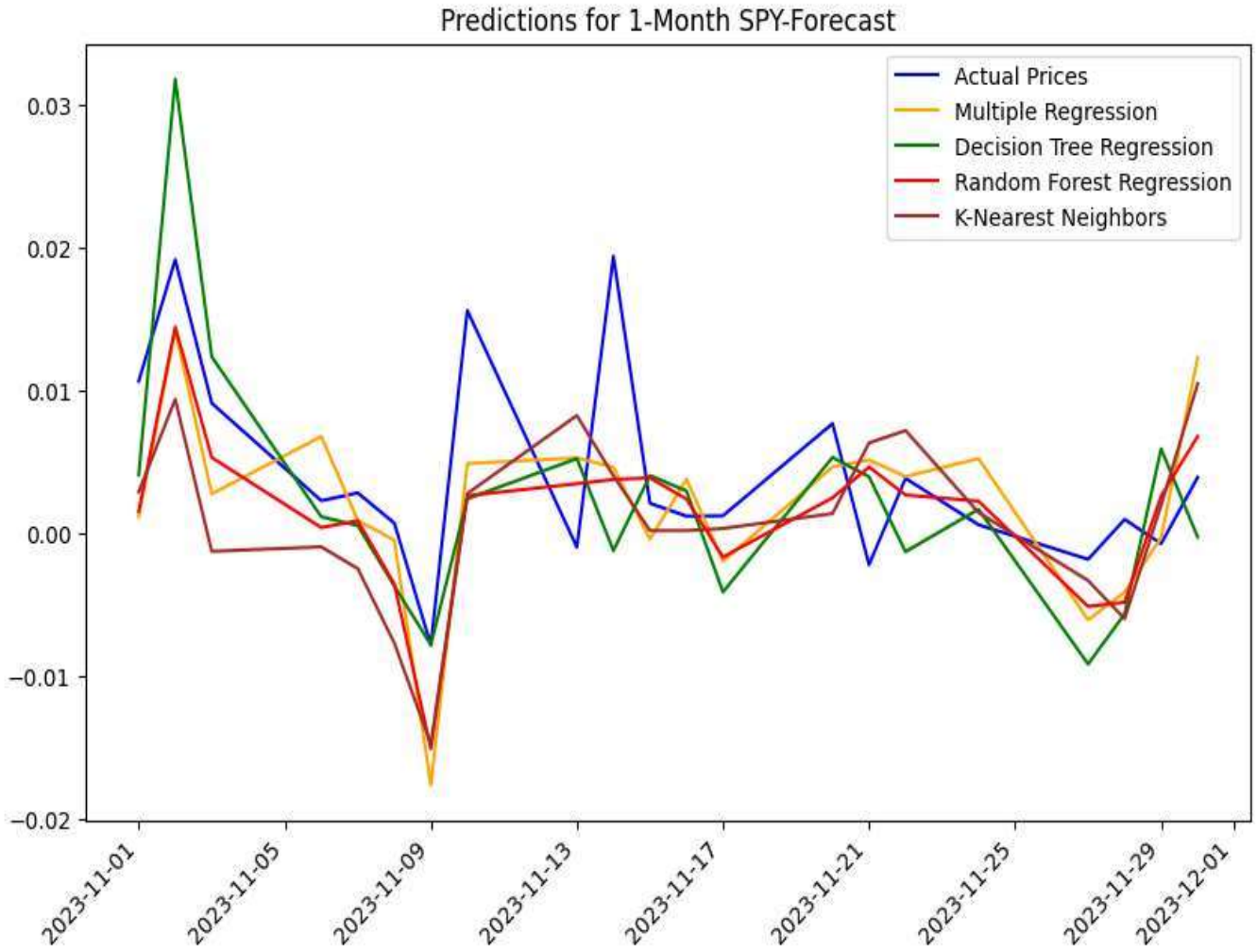
Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.003381	0.000021	0.004622	0.723359	inf
Decision Tree	0.006043	0.000065	0.008075	0.155504	inf
Random Forest	0.003955	0.000029	0.005383	0.624715	inf
K-NN	0.003988	0.000030	0.005454	0.614741	inf

Results:

Healthcare Sector

1-Month (Just Returns):

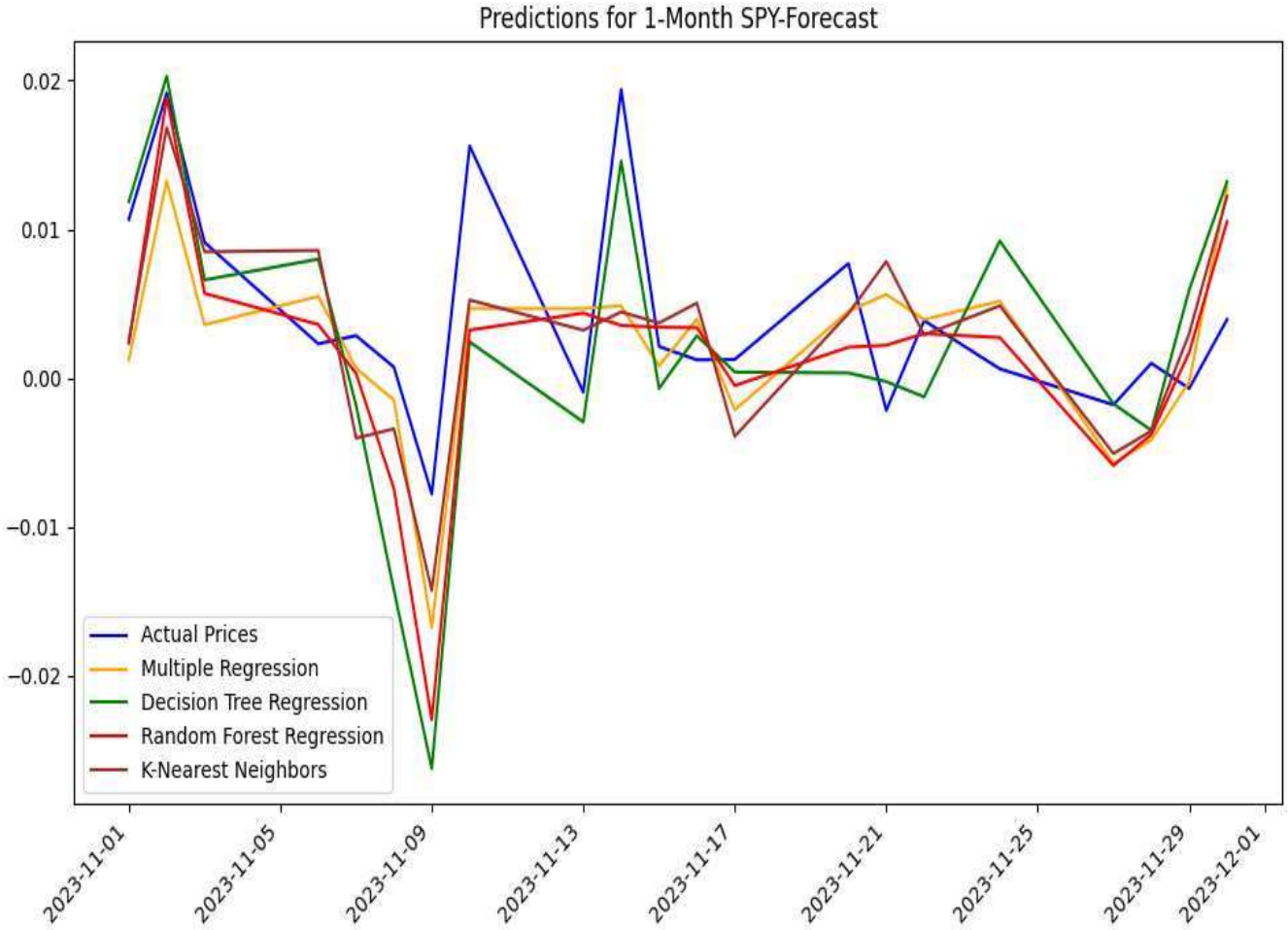


Performance metric:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005302	0.000041	0.006430	0.129360	203.413468
Decision Tree	0.005656	0.000054	0.007368	-0.143121	245.004626
Random Forest	0.004857	0.000037	0.006111	0.213769	191.257413
K-NN	0.006172	0.000054	0.007351	-0.137821	247.703722

1-Month (With RFE):

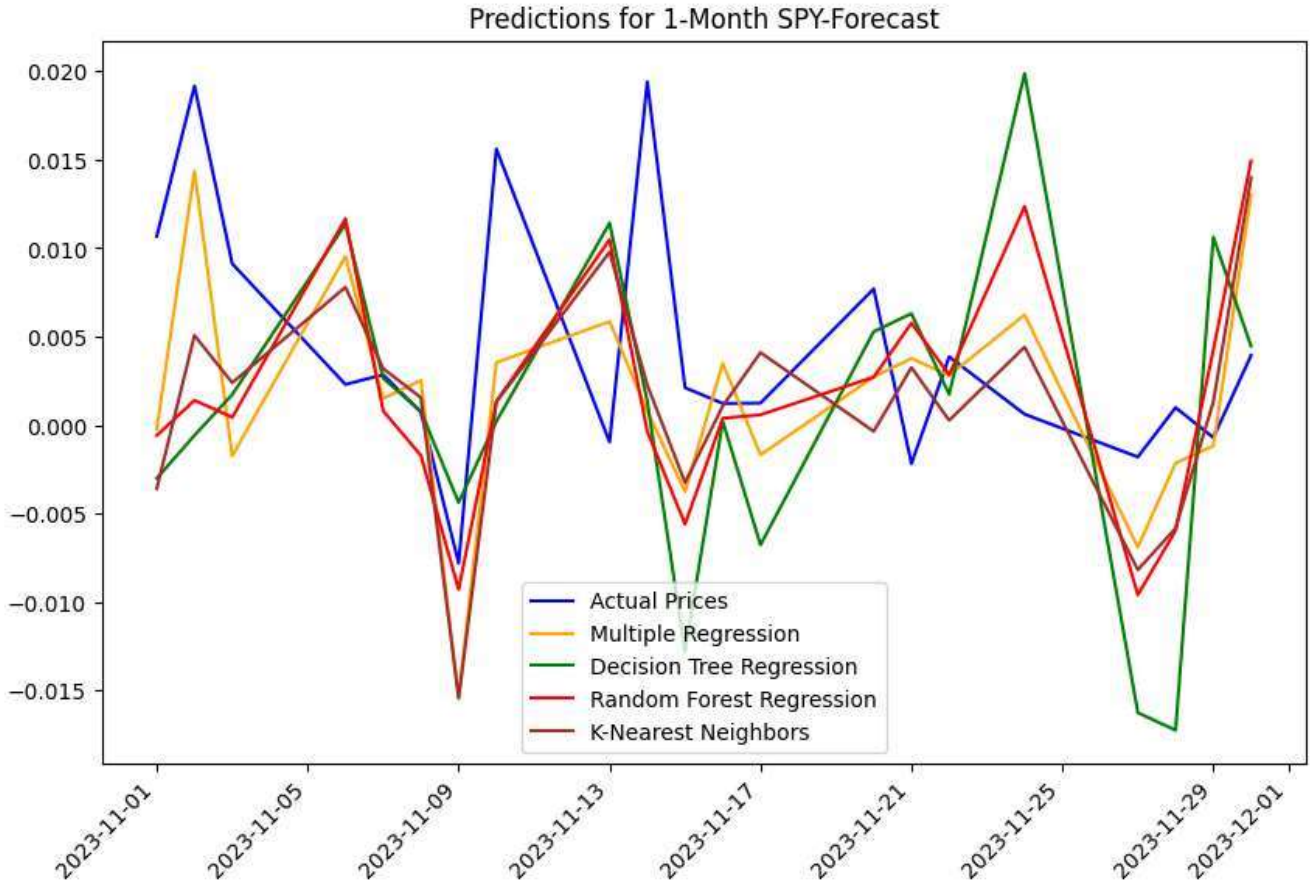


Performance metric:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005241	0.000041	0.006368	0.146252	203.580154
Decision Tree	0.005594	0.000055	0.007400	-0.153014	321.310876
Random Forest	0.005188	0.000046	0.006812	0.022932	215.970037
K-NN	0.005399	0.000041	0.006381	0.142572	249.479883

1-Month (With PCA):

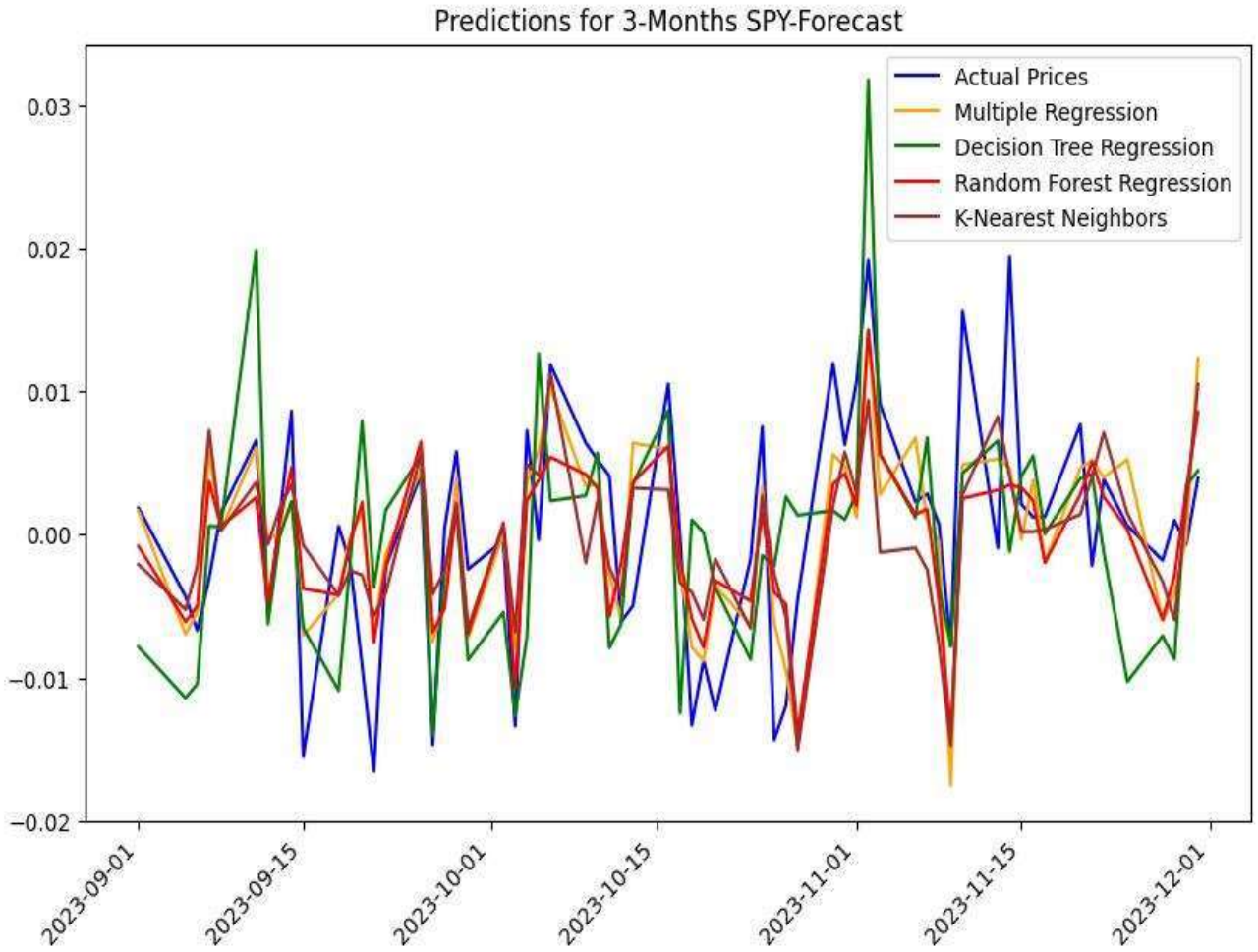


Performance metric:

Overall Performance Metrics for 1-Month SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.006119	0.000056	0.007475	-0.176540	224.413784
Decision Tree	0.009530	0.000136	0.011654	-1.859939	550.357460
Random Forest	0.007814	0.000090	0.009473	-0.889394	356.480960
K-NN	0.006955	0.000071	0.008453	-0.504558	247.231096

3-Month (Just Returns):

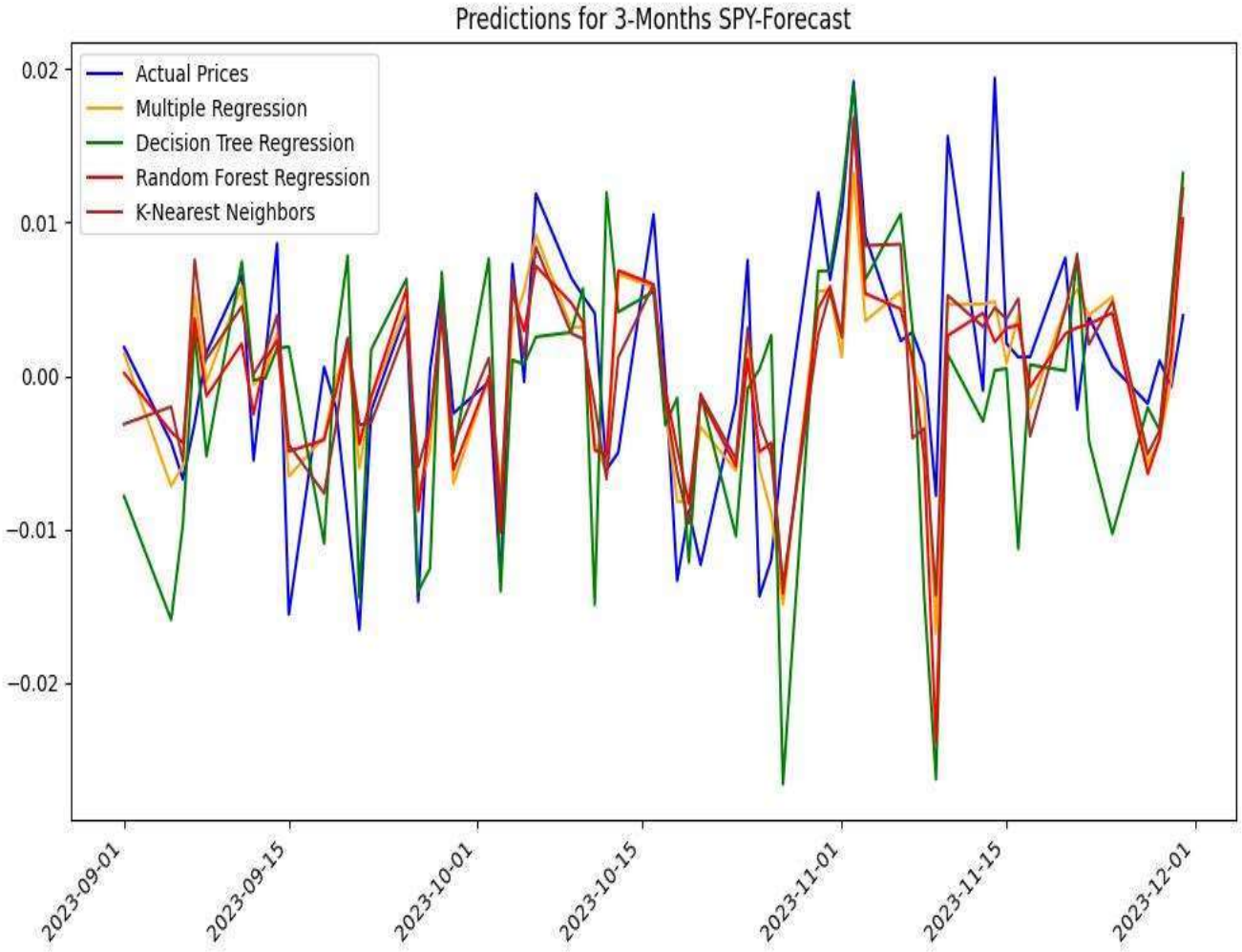


Performance metric:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.004651	0.000033	0.005768	0.491381	246.425017
Decision Tree	0.006678	0.000068	0.008249	-0.040091	720.765329
Random Forest	0.004771	0.000035	0.005903	0.467402	222.864009
K-NN	0.005528	0.000045	0.006693	0.315206	296.628677

3-Month (With RFE):

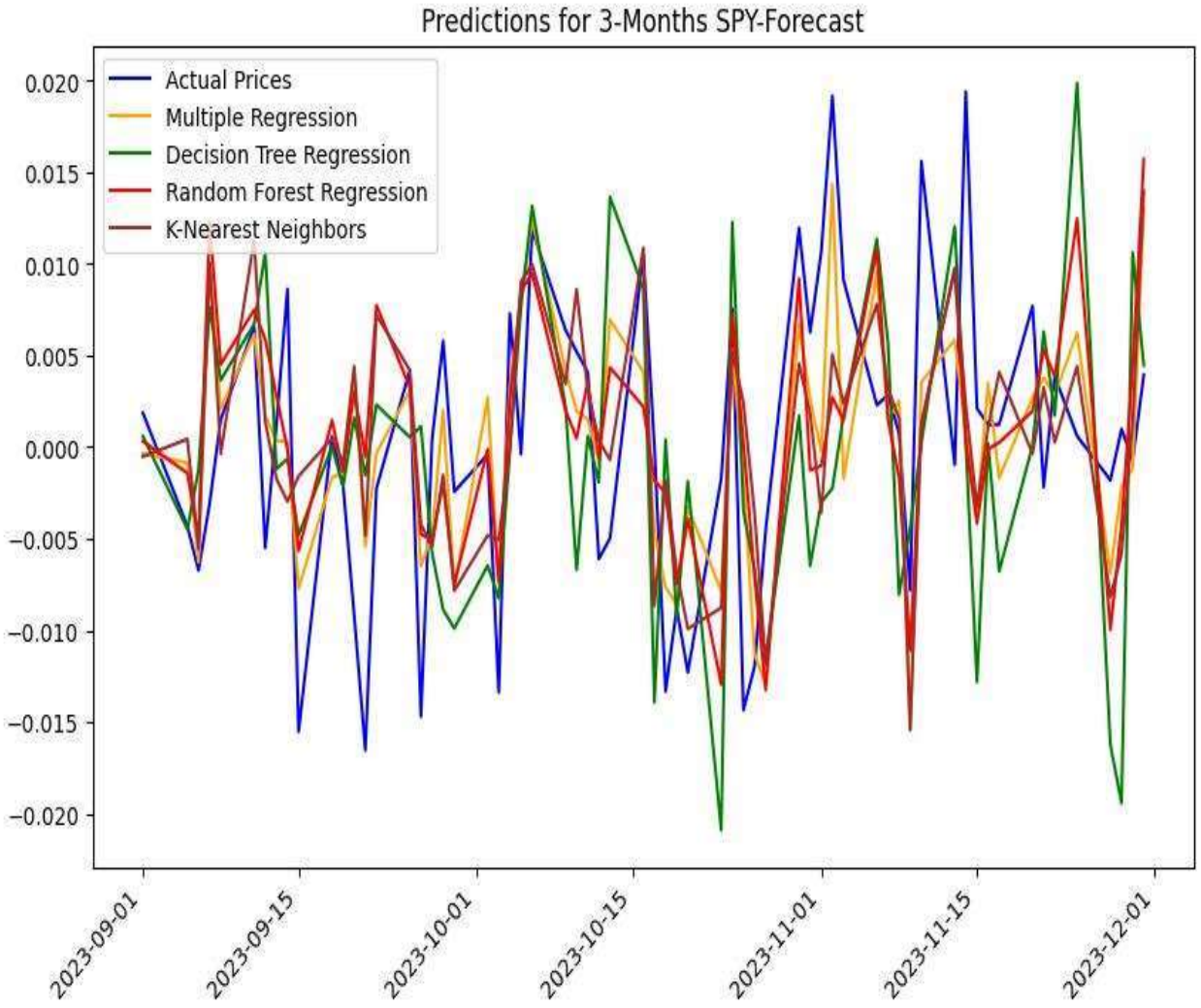


Performance metric:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.004717	0.000034	0.005845	0.477820	236.513738
Decision Tree	0.007351	0.000088	0.009396	-0.349526	446.338684
Random Forest	0.004882	0.000040	0.006301	0.393065	189.817775
K-NN	0.004954	0.000038	0.006143	0.423168	235.043929

3-Month (With PCA):

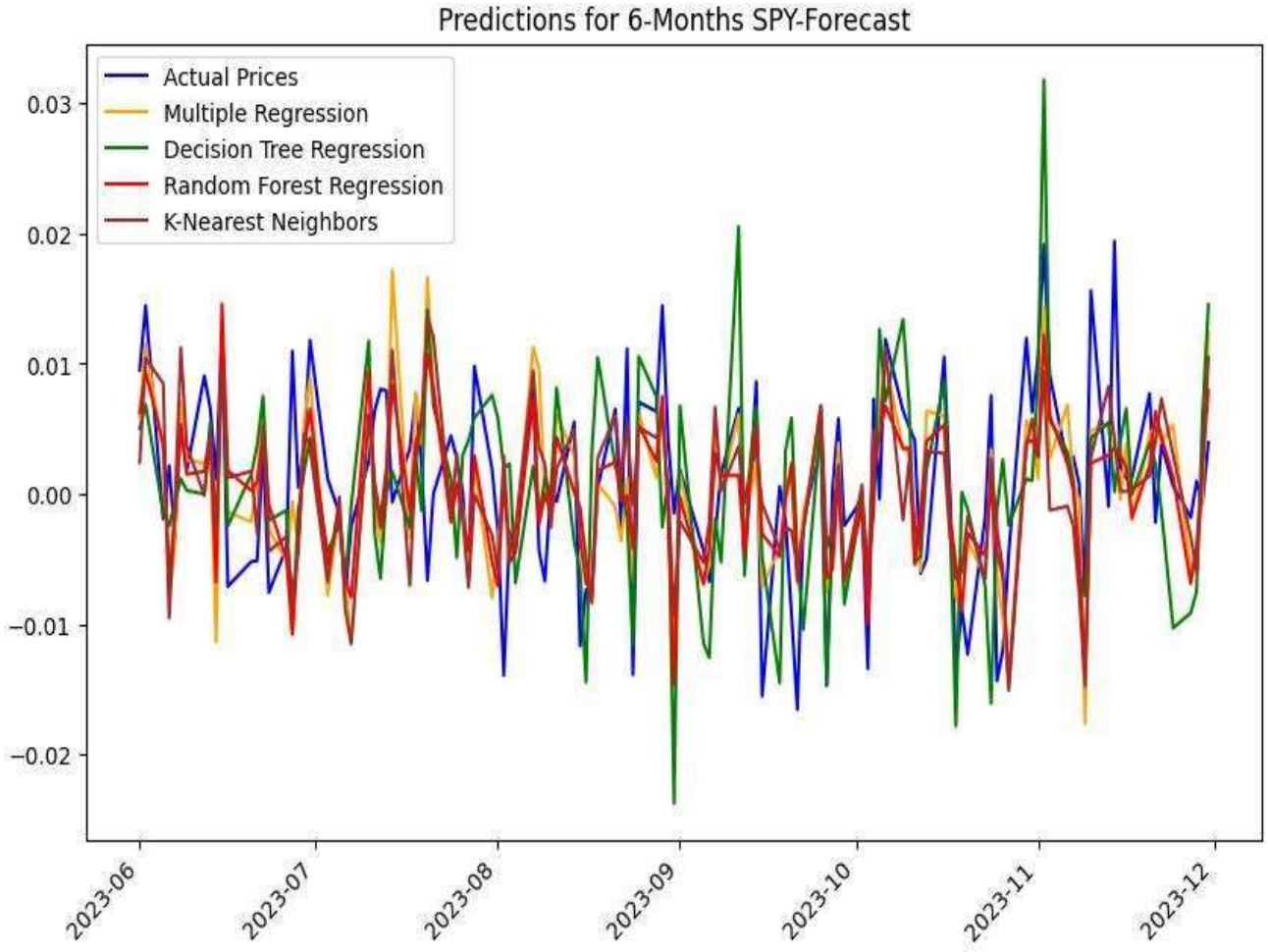


Performance metric:

Overall Performance Metrics for 3-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005358	0.000044	0.006670	0.319981	361.714180
Decision Tree	0.008575	0.000108	0.010404	-0.654433	859.366539
Random Forest	0.006556	0.000066	0.008094	-0.001406	310.595870
K-NN	0.006391	0.000061	0.007788	0.072783	526.752684

6-Month (Just Returns):

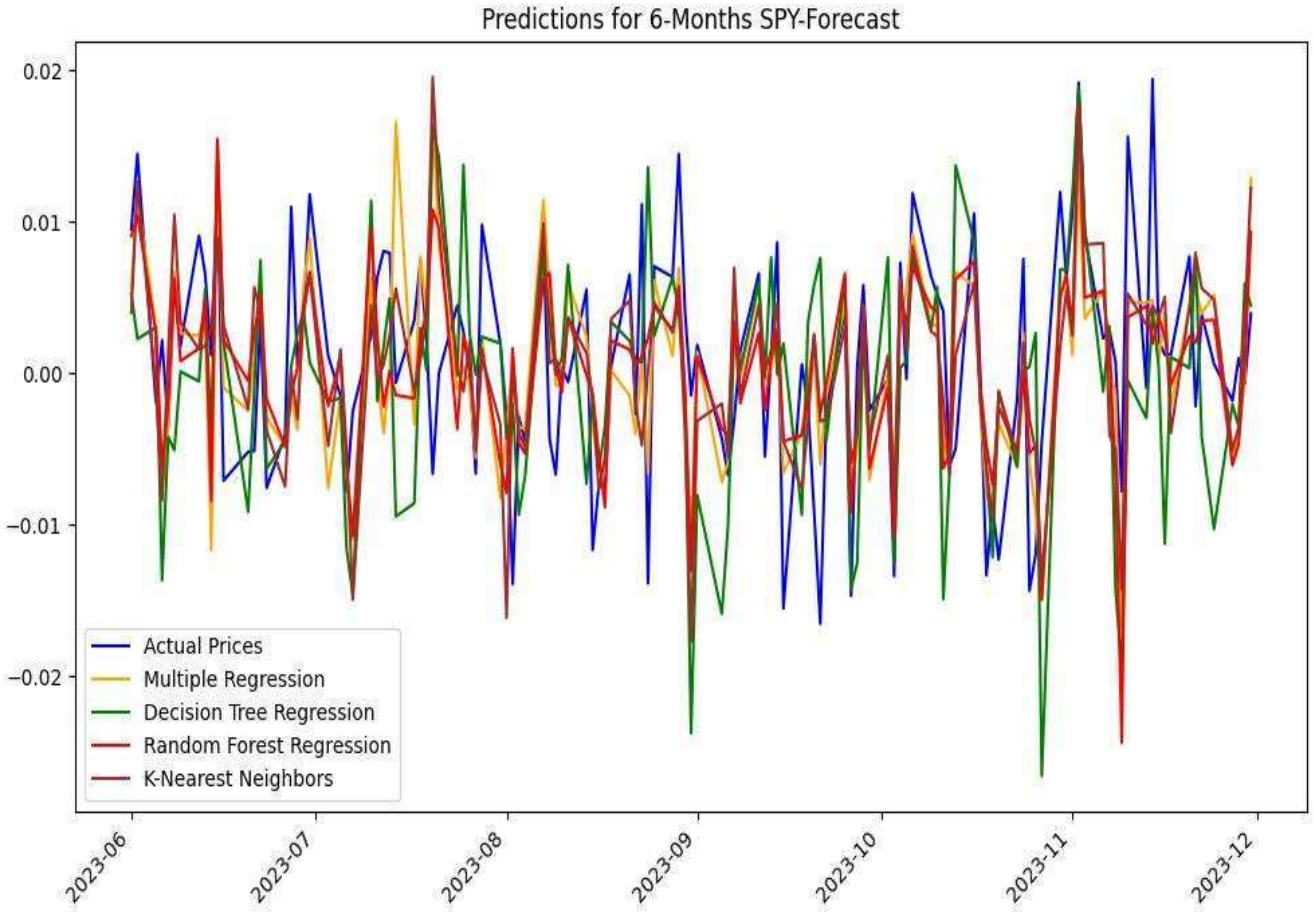


Performance metric:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005013	0.000042	0.006492	0.238711	inf
Decision Tree	0.006496	0.000067	0.008197	-0.213562	inf
Random Forest	0.005117	0.000042	0.006443	0.250189	inf
K-NN	0.005497	0.000047	0.006882	0.144542	inf

6-Month (With RFE):

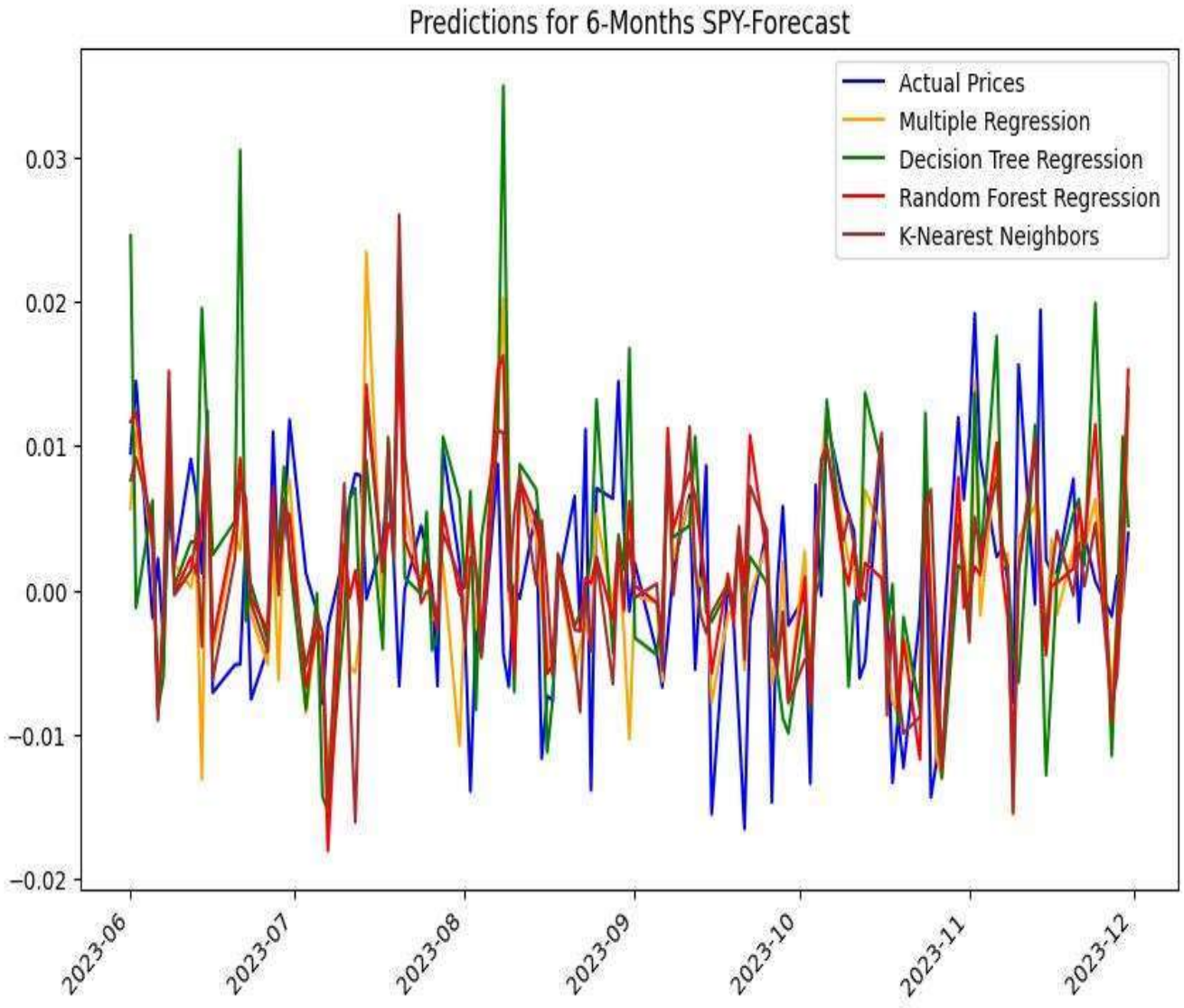


Performance metric:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.004996	0.000042	0.006472	0.243465	inf
Decision Tree	0.006757	0.000081	0.008973	-0.454232	inf
Random Forest	0.005067	0.000043	0.006537	0.228130	inf
K-NN	0.005385	0.000047	0.006889	0.142864	inf

6-Month (With PCA):

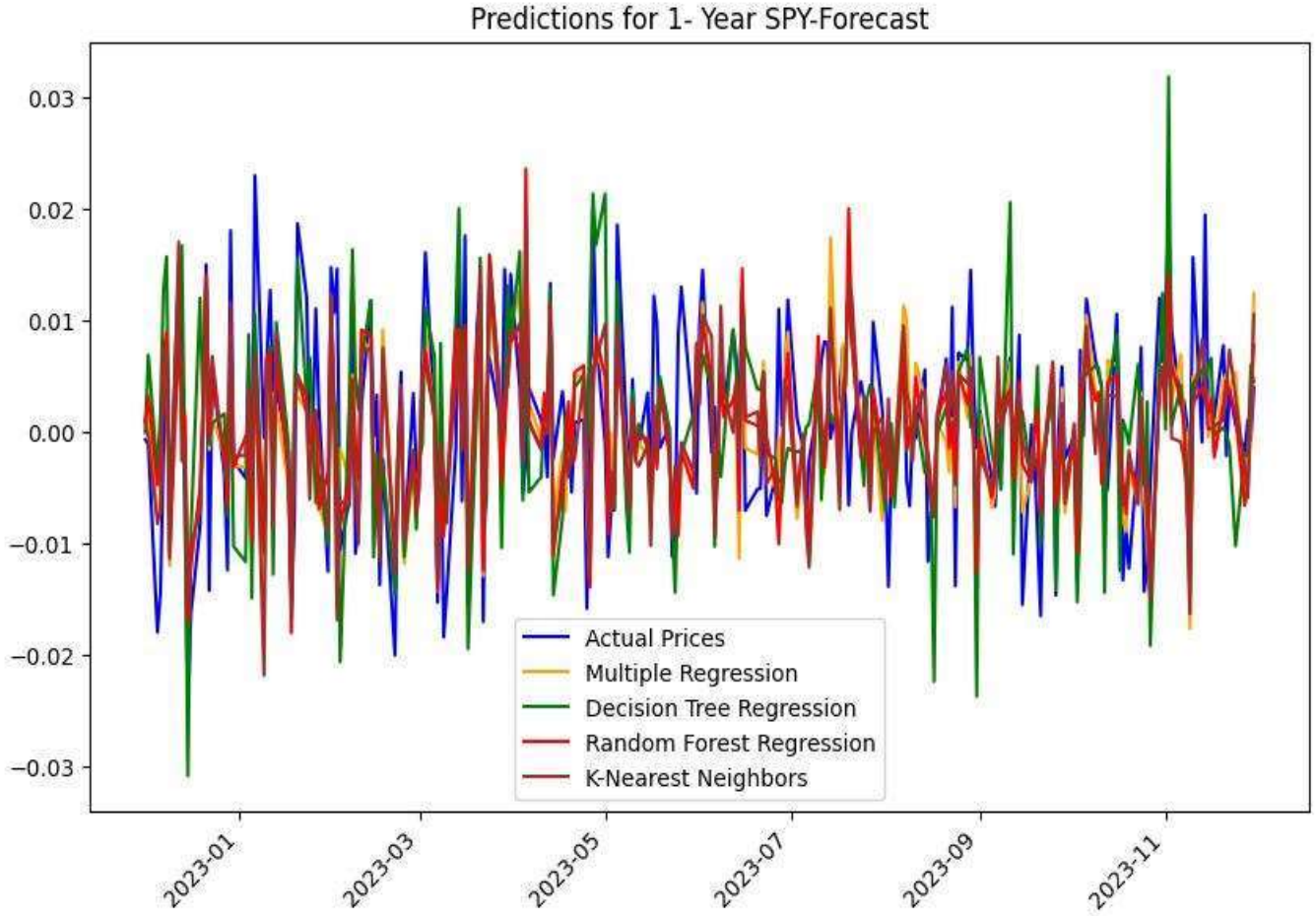


Performance metric:

Overall Performance Metrics for 6-Months SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005948	0.000061	0.007801	-0.099320	inf
Decision Tree	0.007698	0.000106	0.010303	-0.917528	inf
Random Forest	0.006329	0.000065	0.008054	-0.171724	inf
K-NN	0.006555	0.000071	0.008445	-0.288158	inf

1-Year (Just Returns):

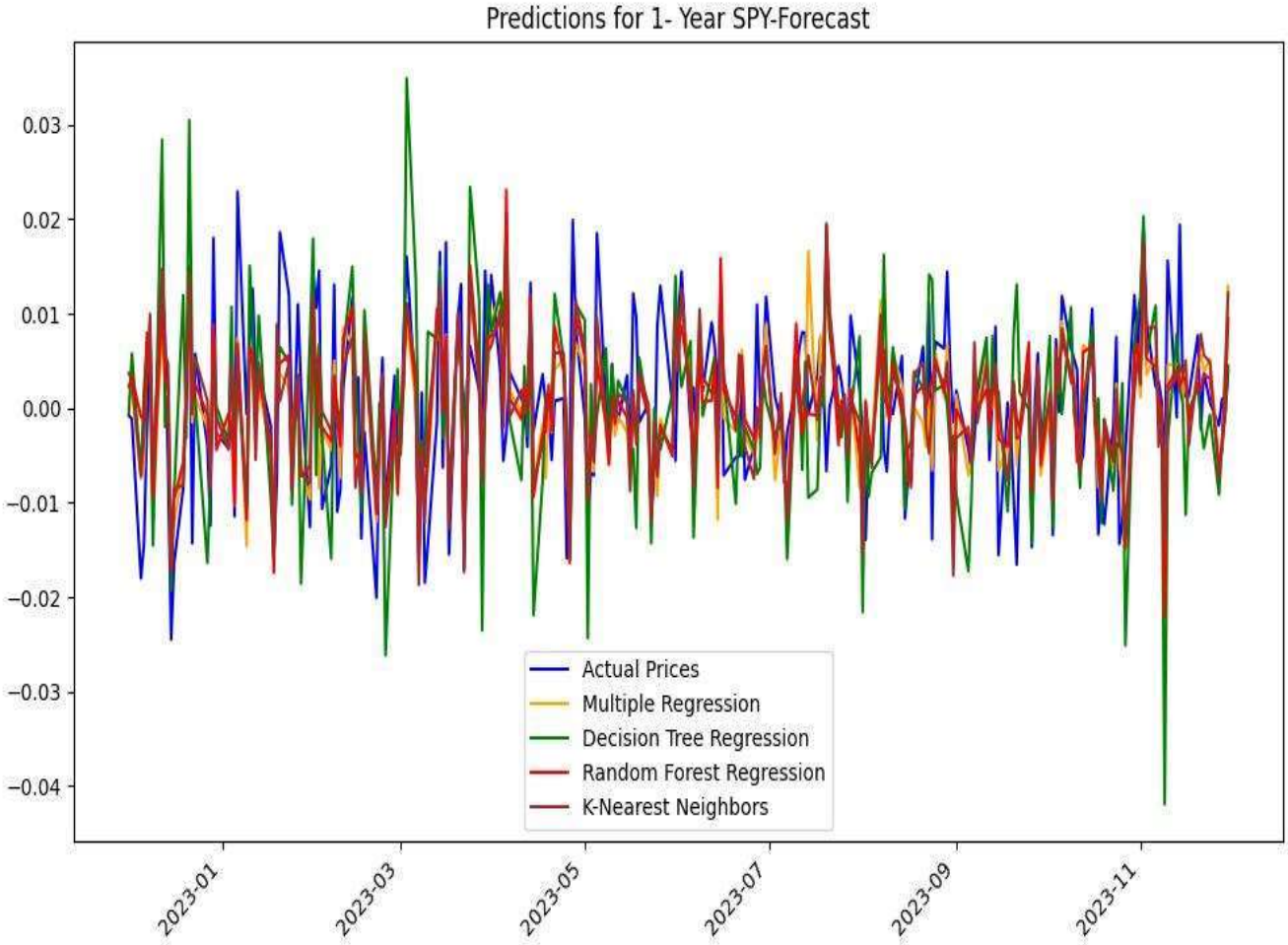


Performance metric:

Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005470	0.000049	0.006991	0.367065	inf
Decision Tree	0.006477	0.000065	0.008089	0.152747	inf
Random Forest	0.005530	0.000050	0.007106	0.346046	inf
K-NN	0.005603	0.000053	0.007257	0.317961	inf

1-Year (With RFE):

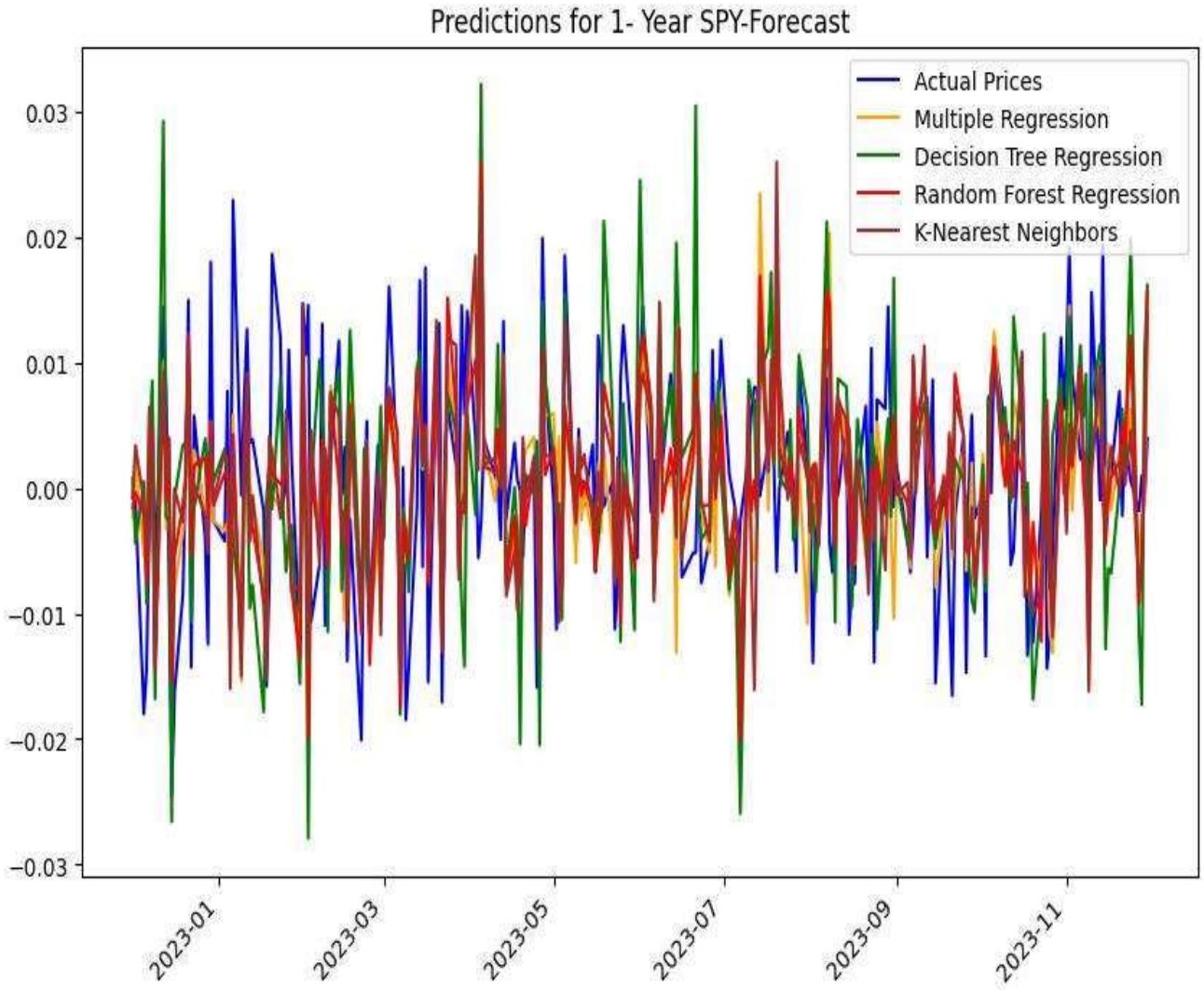


Performance metric:

Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.005421	0.000048	0.006925	0.378905	inf
Decision Tree	0.007886	0.000101	0.010046	-0.306852	inf
Random Forest	0.005563	0.000052	0.007179	0.332507	inf
K-NN	0.005659	0.000052	0.007240	0.321294	inf

1-Year (With PCA):



Performance metric:

Overall Performance Metrics for 1- Year SPY-Forecast:

	MAE	MSE	RMSE	R-Squared	MAPE
Multiple Regression	0.006511	0.000069	0.008323	0.103012	inf
Decision Tree	0.007953	0.000107	0.010335	-0.383290	inf
Random Forest	0.006724	0.000071	0.008446	0.076180	inf
K-NN	0.006875	0.000075	0.008675	0.025541	inf

Conclusion:

This project demonstrated the capability of machine learning models to uncover valuable signals in sector-specific stock data for predicting movements in broader market indexes. Comparing performance across regression algorithms and input feature sets revealed that technology stocks, especially after select feature filtering, provide the strongest leading indicators of S&P 500 returns across monthly, quarterly, semi-annual and annual time horizons.

The optimized Multiple Regression and Random Forest model using technology stocks as inputs significantly outperformed baseline models and accurately forecasting both the direction and magnitude of market shifts over long time periods. This indicates a consistent lead-lag relationship where tech stock fluctuations tend to precede analogous moves in subsequent broader index returns.

On the other hand, healthcare stocks exhibited much weaker predictive power, implying market-leading signals are specific to certain sectors. The project results align with the notion that technology firms, with their growth orientation and sensitivity to economic conditions, provide useful indicators on investor sentiment and future capital flows across the wider marketplace.

Overall, the analysis demonstrates the viability of data-driven predictive modeling for tactical market timing using sector stocks. The methodology could be productionized through an algorithmic trading system that continuously scrapes data, retrains models, and automates trade orders based on forecasted index returns.

Future Applications:

There are several promising avenues to build on this work and derive further value from the predictive modeling approach:

- Operationalize models for algorithmic trading strategies that dynamically adjust market exposure based on predicted returns. Entry and exit signals would be automated via API connections.
- Incorporate sentiment, fundamentals, alternative data as additional features for potentially better forecasts. Apply neural networks to model complex nonlinear relationships.
- Expand modeling to a broader set of sector ETFs and global indexes to forecast both domestic and international markets.
- Research optimal portfolios that maximize risk-adjusted returns derived from the forecasted market directions. Evaluate performance against traditional passive indexing.
- Provide predictions as a service to institutional investors for making data-driven tactical tilts in their portfolios. Offer customized modeling based on client priorities.

In summary, this project serves as a valuable proof-of-concept for leveraging machine learning on financial data to forecast market moves. Significant opportunities exist to extend this core approach towards building cutting-edge predictive analytics systems for powering automated, insight-driven investment strategies that can consistently outperform the market.