

# Stock Price Forecasting

time series forecasting in python leveraging machine learning concepts  
from the sktime library for financial analysis


February 2024

 by Alex Ferrer



# The Stock Market



Investopedia

[🔗](#)

**What Is the Stock Market, What Does It Do, and How Doe...**

The stock market consists of exchanges in which stock shares and other financial securities of publicly held companies are bought and sold.

## Overview

This project investigates time series forecasting, applying classic statistical methods to predict daily stock closing prices and discern future stock market trends.

Programming and machine learning concepts are integrated to reveal the stock market data and assess the effectiveness of various forecasting methodologies.

The goal is to identify the most effective model(s) that balance prediction accuracy and interpretability for application in real-life business use-cases by financial analysts.

## Objectives

Gain understanding of diverse forecasting models leveraging machine learning techniques and tools in python.

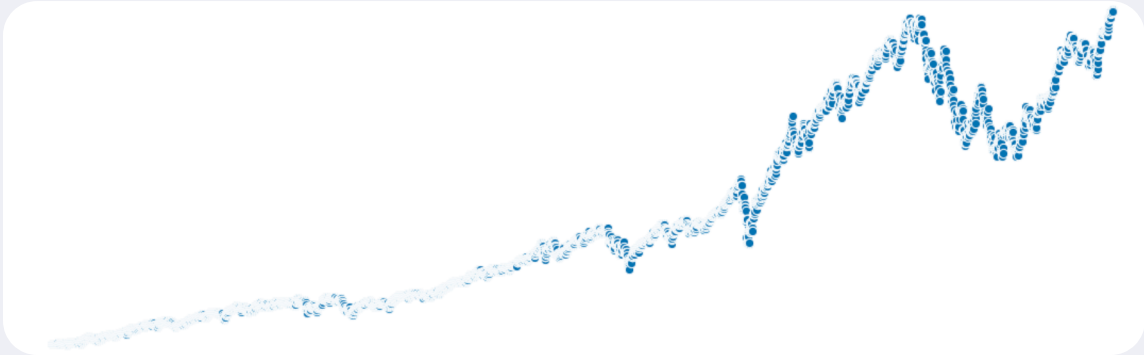
- Explore classical statistical forecasting models
- Optimization applying of machine learning techniques and tools in python
- Assess value-added of using these forecasting models for Financial Analysis

## Programming Languages & Libraries

Programming language: python. Main Libraries used: pandas, numpy, sktime, statsmodels, yfinance

## Data

The dataset used for this project consists of daily stock closing prices from 2014 to 2023 of selected companies listed on Nasdaq, focussing on IBM for forecasting. The data transformation technique employed includes differencing to handle non-stationarity in the time series data.

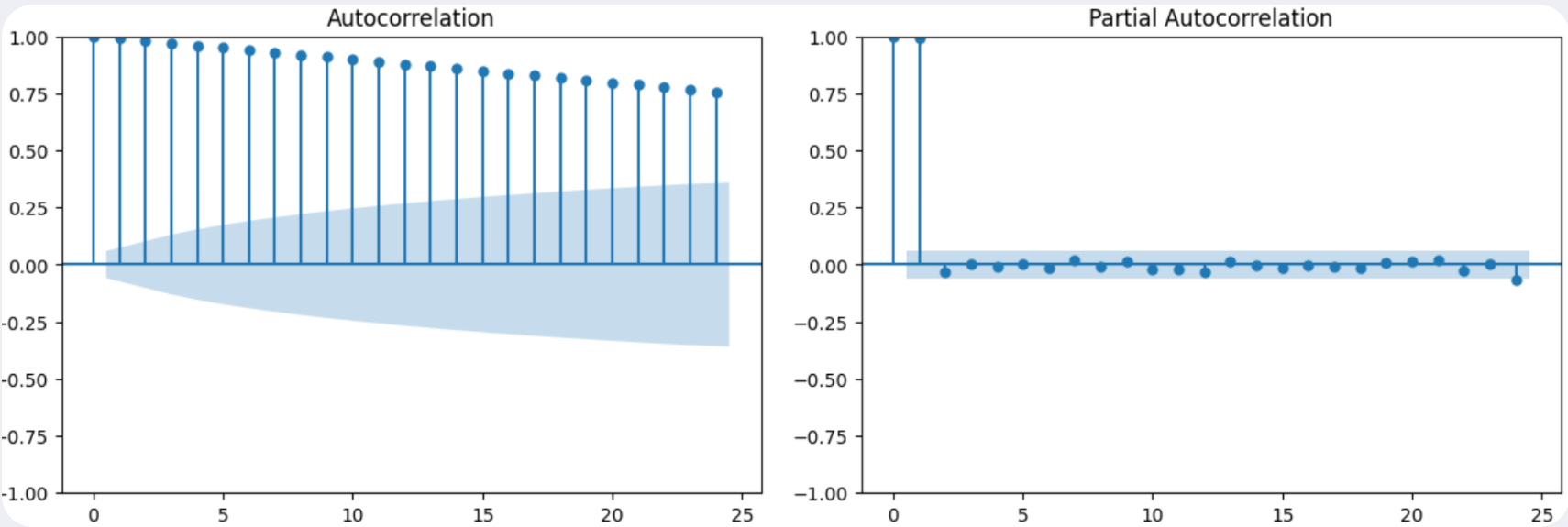


## EDA

The data was imported and cleaned, handling missing dates and formatting.

Visual inspection was executed to gain better understanding of the data.

- Hypothesis testing was conducted to determine if the data was stationary. This involved the Dickey-Fuller Test and KPSS test.
- Autocorrelation was also examined for data before and after transformation (method: differencing), and residuals.



## Models

Several models have been considered for this project, including Simple Moving Average (SMA), Seasonal Exponential Smoothing (SES), AutoRegressive Integrated Moving Average (ARIMA), AutoARIMA, Prophet. Additionally, machine learning models such as Gradient Boosting and K-Nearest Neighbors from the sktime library have been explored.

## Evaluation Metrics & model selection criteria

The performance of these models is evaluated using Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE). Cross-validation techniques to select the best models, specifically GridSearchCV and RandomizedGridSearchCV, are used for hyperparameter tuning.

The aim is to find a balance between these factors to ensure the chosen model is not only accurate but also easy to understand and quick enough for real-time applications.


## Challenges

The main challenges in this project include gaining a deep understanding of key statistical tests and classic statistical time series forecasting techniques. These concepts are fundamental to the field of data science and are crucial for developing accurate and reliable models.

Another significant challenge is finding a balance of accuracy and computing load when performing cross validation to select model parameters or to calculate model error measures. Each model has its strengths and weaknesses, and choosing the right one requires careful consideration of the data characteristics and the specific requirements of the project.

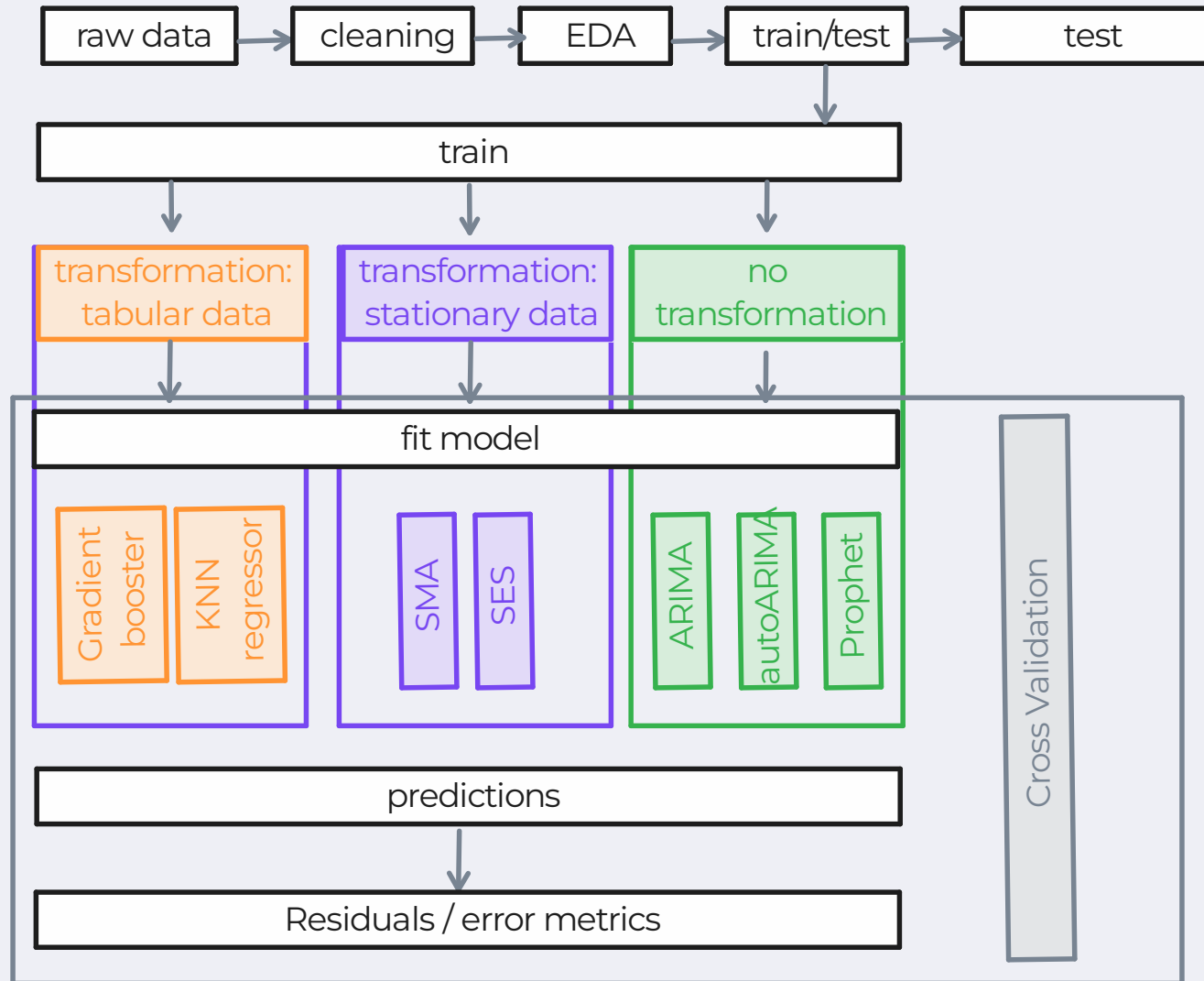
Additionally, it is imperative to clarify that this project is not expected to produce any financial advice and is expected to be considered as a study in different forecasting techniques using data from the stock market.

## Bibliography



- "Introductory Econometrics for Finance" by Chris Brooks. Cambridge University Press, 2019.
- "Statistical forecasting: notes on regression and time series analysis" by Robert Nau, Fuqua School of Business, Duke University.  
<https://people.duke.edu/~rnau/411home.htm>
- Python Library: The SKTime Python library.  
(<https://www.sktime.net/en/stable/index.html>)
- Websites: Medium.com. TowardsDataScience. Analytics Vidhya. Rob J Hyndman.  
(Accessed: January 26, 2024).

# Step-by-Step Data Analysis Workflow



# Classic Statistical Forecasting Models

## SMA (moving average)

Simple. Can smooth out price fluctuations and filter out noise

For irregular data

Responds sluggishly to trends

## SES (exponential smoothing)

Flexible in how quick to respond to recent trends.

For non seasonal data without trends

Responds sluggishly to trends

## ARIMA (autoregressive integrated moving average)

Wide ranging models and parameters.

Performs well on highly aggregated and plentiful data.

Complex parameter determination

## Regression Analysis

Assumption of linear correlation with other variables

For correlated multivariate data

Overfitting risk

# Machine Learning Regression Models

## GradientBoosting regression (ensemble method)

good accuracy and flexibility

idea for complex non-linear data, easy to use with categorical and numerical data

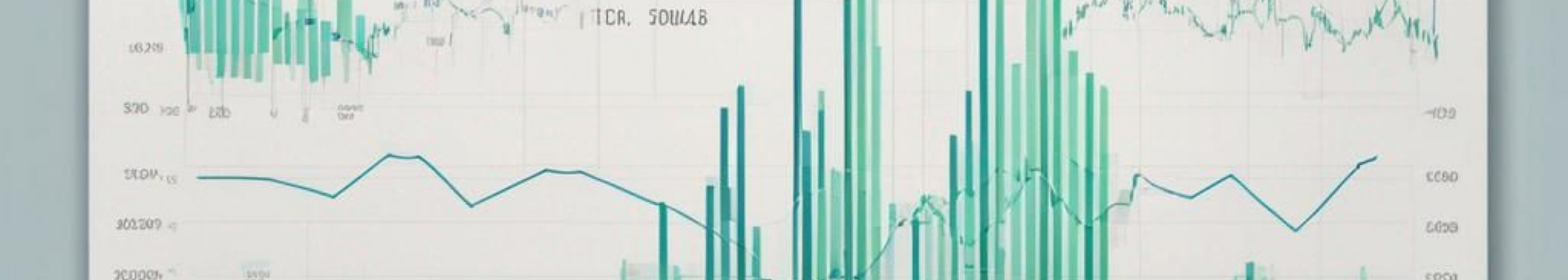
overfitting risk, difficult to interpret, computationally expensive

## K-Nearest Neighbors Regression

simple to use, little data preparation

suitable for problems where instances are close to each other, not good for high-dimensionality data, careful selection of k to avoid overfitting

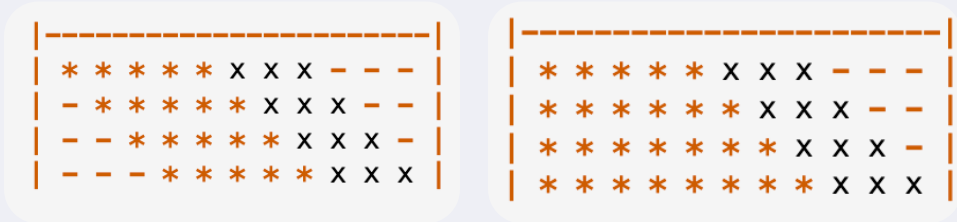
computationally intensive



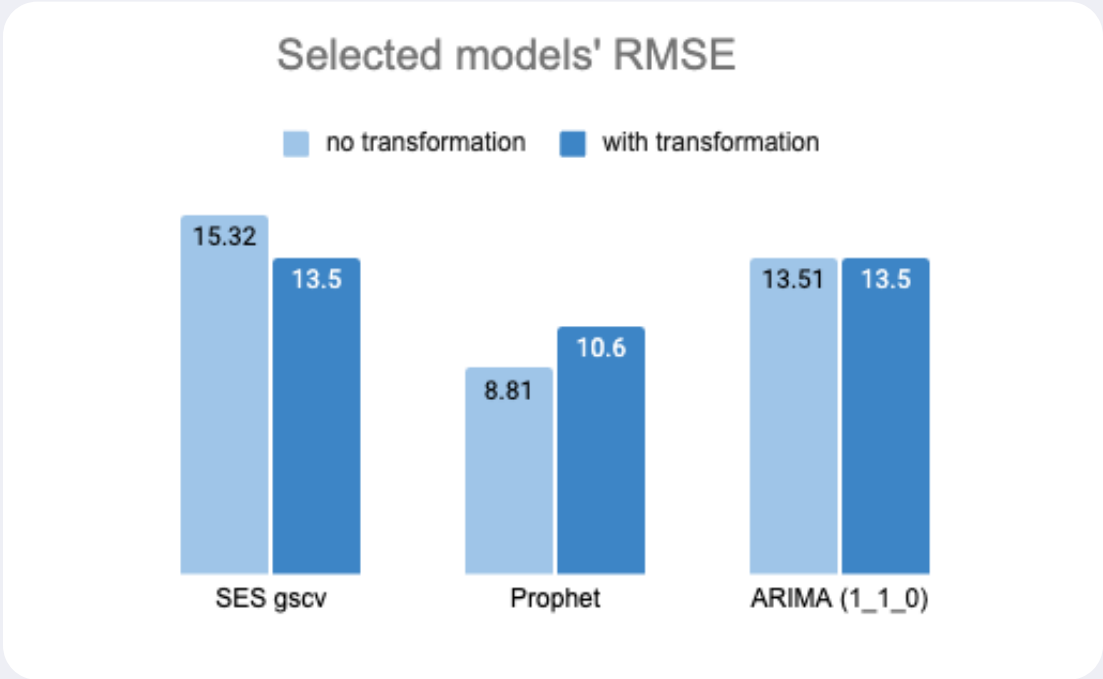
# Application of Cross Validation to Time Series Forecasting



## SlidingWindow and ExpandingWindow examples from sktime's documentation:



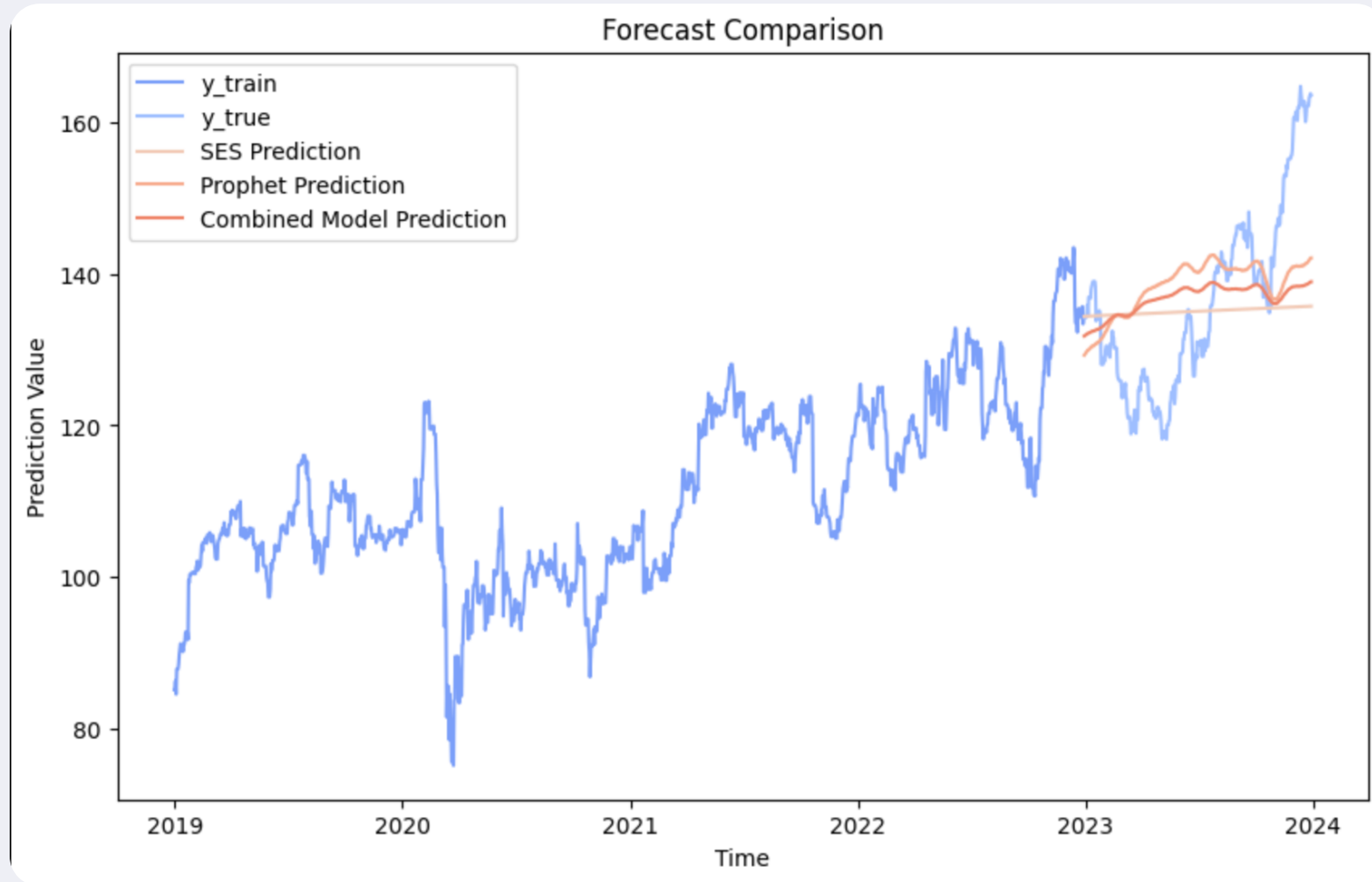
## Selected Models' RMSE comparison



# Phase i: Results and Analysis of the Forecasting Models

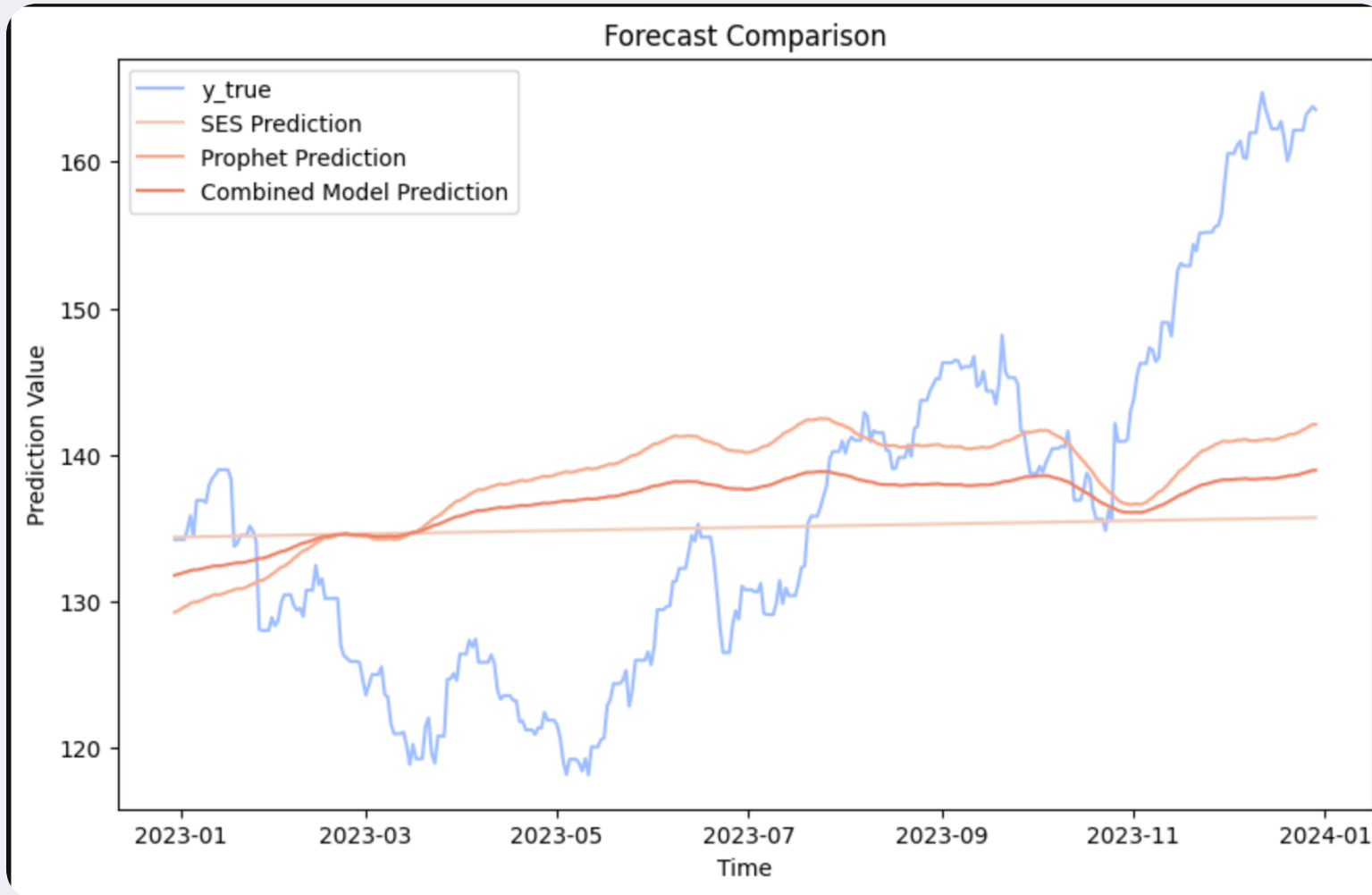
	Model	RMSE	MAE	MAPE
0	SMA_gscv_win3_t0	15.000758	12.049256	0.076594
1	SMA_gscv_win30_t1	26.394814	21.503861	0.136863
2	SES_gscv_damp0_trend0_s5_t0	14.848146	11.916442	0.075752
3	SES_gscv_damp0_trend0_t0	15.320774	12.332927	0.078400
4	SES_gscv_damp0_trend0_sp365_t0	15.320774	12.332927	0.078400
5	SES_gscv_damp0_trend0_t1	13.510300	10.900504	0.069356
6	SES_gscv_damp0_trend0_sp365_s0_t1	13.510300	10.900504	0.069356
7	SES_gscv_s0_t1	13.510300	10.900504	0.069356
8	Prophet_param0_t0	8.811093	7.643583	0.050422
9	Prophet_param0_t1	10.598033	8.732641	0.055728
10	ARIMA_1_0_0_t(0)	19.866409	16.017380	0.101775
11	ARIMA_1_1_0_t(0)	13.506661	10.897783	0.069339
12	ARIMA_1_1_1_t(0)	13.500144	10.891824	0.069301
13	ARIMA_1_0_0_t(1)	13.511943	10.901975	0.069366
14	ARIMA_2_0_0_t(1)	13.496555	10.888162	0.069278
15	autoARIMA_p1_10_sp1_t(0)	15.317440	12.329825	0.078381

# Selected Models' Forecasts





# Selected Models' Forecasts: Zoom-in





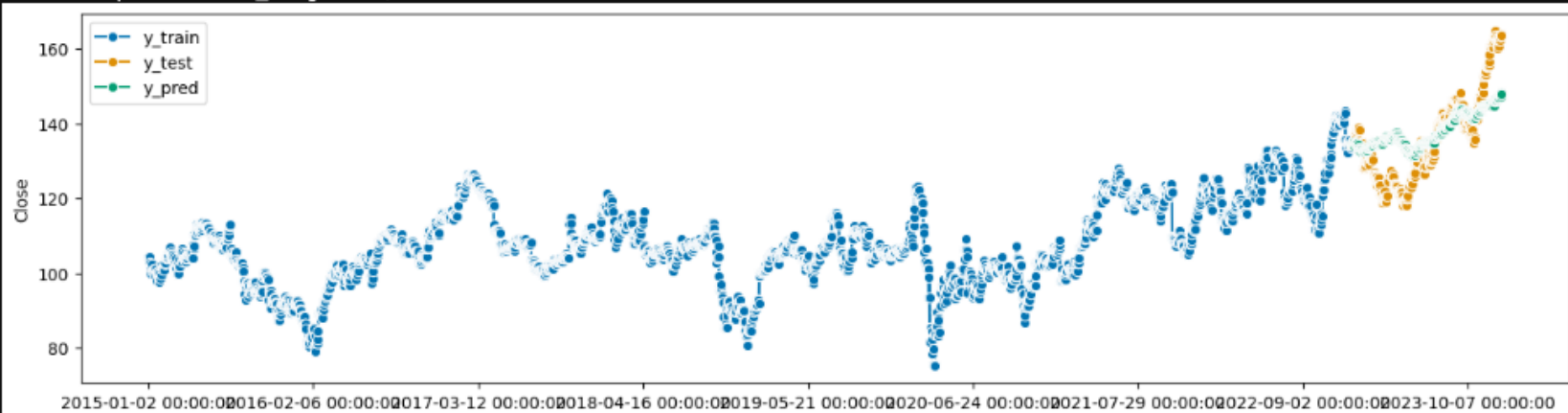
# Phase ii: Results and Analysis of the Forecasting Models – KNN

index			Model	RMSE	MAE	MAPE
0	20	KNN params: win_length = 3; n= 9		8.567190	6.812043	0.050860
1	16	KNN params: win_length = 10; n= 5		9.748522	7.821240	0.057677
2	11	KNN params: win_length = 45; n= 4		9.753191	8.305673	0.063176
3	3	KNN params: win_length = 4; n= 1		10.516394	8.087595	0.059058
4	5	KNN params: win_length = 45; n= 1		10.653323	7.112785	0.049394
5	15	KNN params: win_length = 4; n= 5		10.855545	8.433417	0.059904
6	17	KNN params: win_length = 45; n= 5		11.085755	9.479507	0.071800
7	18	KNN params: win_length = 1; n= 9		11.800161	10.174421	0.078043
8	1	KNN params: win_length = 2; n= 1		11.894990	9.753387	0.071456
9	23	KNN params: win_length = 45; n= 9		12.041478	10.192305	0.078289
10	10	KNN params: win_length = 10; n= 4		12.246120	10.059587	0.077418
11	4	KNN params: win_length = 10; n= 1		12.272942	10.375575	0.079075
12	14	KNN params: win_length = 3; n= 5		12.294335	10.053989	0.074064

# KNN Models Forecasts

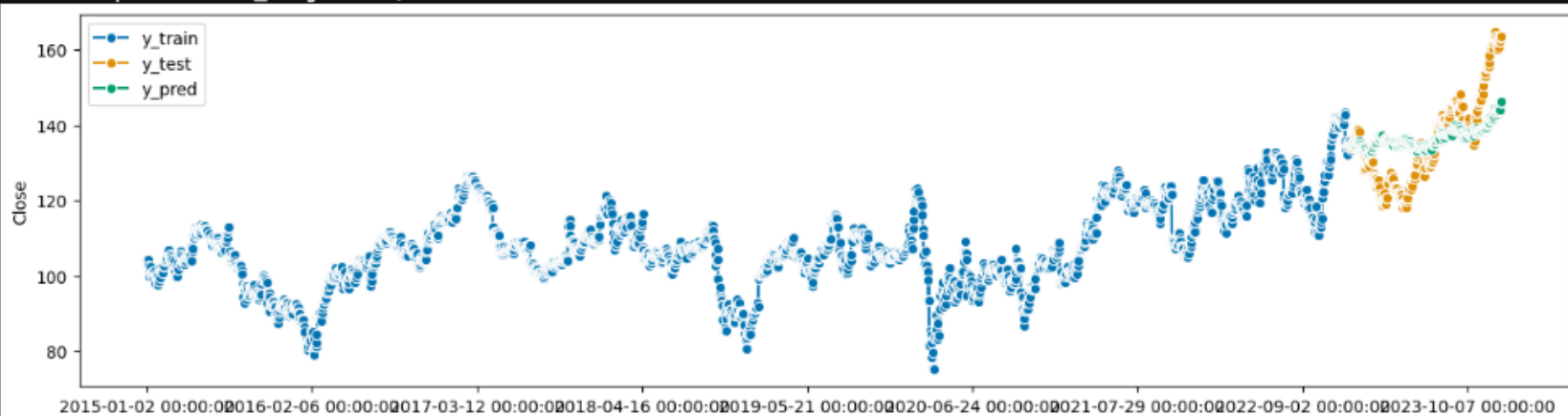
KNN params: win\_length = 3; n= 9

	Model	RMSE	MAE	MAPE
0	KNN params: win_length = 3; n= 9	8.56719	6.812043	0.05086



KNN params: win\_length = 10; n= 5

	Model	RMSE	MAE	MAPE
0	KNN params: win_length = 10; n= 5	9.748522	7.82124	0.057677



# Conclusion and Future Directions

Summary	<p>Time series forecasting can be computed in both Excel and programming languages. In the latter, there are additional tools available that make it easier to process large amounts of data, automate tasks, and more importantly, efficient ways to perform model tuning</p> <p>The recommendation of time series model usage will inevitably depend on the problem to solve and data. The ultimate suggestion for Financial Analysis is to unerstand the data and the statistics first, set up a benchmark forecaster, and opmitize using advanced techniques if possible.</p>
Selected Models	<ul style="list-style-type: none"><li>Model 1. t(1) simple exponential smoothing (SES) * gridsearch : params: smoothing_level': 0.5, 'sp': 365, 'trend': None</li></ul> <p>this model is recommended as the go-to model for its simple methodology and ease-of use while providing overall good results in accuracy</p> <ul style="list-style-type: none"><li>Model 2 t(0) Prophet: should be evaluated and understood further before utilization</li><li>Model 3. t(0) ARIMA 1_1_0: similar results to SES, and as this model is more complex, the usage of SES would be recommended instead</li><li>Model 4. Combi Model: combined model of SES &amp; Prophet: could be used to find a balance between the flat forecast of the SES model and the prophet forecast.</li><li>Model 5. t(1) KNN Regressor: params: win_lengh = 10; n= 5 This model should be evaluated further to understand how the distances between the different points affect the forecasted values.</li></ul>
Future Research	<p>Exploration of advanced machine learning techniques and AI-powered predictive models for enhanced stock price forecasting:</p> <ul style="list-style-type: none"><li>increasing the stocks analyzed to an hierarchical panel (stock index + companies indexes) with stock clustering, and global model training to capture cross-market relationships and improve the overall accuracy of the forecasts.</li><li>Use-case analysis for different financial analysis tasks as well application of these forecasting techniques to different organization types</li><li>advanced ensemble methods to optimize predictions</li></ul>