Forecasting Implied Volatility of Advanced Micro Devices, Inc. Stock Options

Lam Lay, Vlasios Mastrantonis, Ramachandra Rahul Taduri, Nha Tran, Nigel Tucker

Background

An **Option Contract** is a financial contract between two traders granting the buy or sell right of a specific underlying asset at a predetermined price from (or to) the trader who sold the option, by a specific date, regardless of the current market value of the stock. The seller of the option is obliged to fulfill the requested order at the agreed-upon price.

Why Trade Options?

Trading options allows for great leverage potential. This means that a small amount of initial capital can see large returns, sometimes many times the initial capital investment.

Also, when purchasing options your risk is limited to your initial investment. When selling options, you get paid as soon as the the trade is put on and you make money if the stock moves in your predicted direction, doesn't move at all, or even if it moves slightly against you.

Lastly, traders involved in the stock market will often hedge their trades by buying or selling options on the same stock, just in case the stock moves against their prediction.

CALL

A call option grants the purchaser of the option the right to buy the underlying at a predetermined price from the seller of the option. The seller of the option has an obligation to fill the order at the agreed upon price.

PUT

A put option grants the purchaser of the option the right to sell the underlying at a predetermined price to the seller of the option. The seller of the option has an obligation to buy the underlying at the agreed upon price.

The most important factors to account for when trading an option contract are the price of the underlying asset, the implied volatility of that asset, and the number of days until expiration for the option. These will all play a role in accurately identifying a price for that option. We chose to focus on forecasting the implied volatility of the underlying asset, in our case AMD stock, using historical AMD options data due to the fact that days to expiration cannot be predicted, and predicting the price of the underlying asset effectively goes beyond the scope of this course.

Data Cleaning

Data Sources : Wharton Research Data Services -> Historical Option Data (Option Pricing), Yahoo Finance

Steps:

- Create a data dictionary for all parameters in downloaded data.
- Remove columns with all Null/Nan Values.
- Remove all columns with unnecessary categorical variables.
- Parameters with some null/Nan values are realized. Total count of null/Nan values within some of the columns is only a small fraction of total data. Therefore these rows are just deleted.
- Review remaining columns for unique elements within them (non Nan/non Null). These columns are also removed as they do not add value to data processing.
- All columns with dates are parsed upon reading the input data and converted to datetime objects.
- Remove all duplicate data in dataframe.
- Dependent date columns are removed to avoid confusion.
- Next step is data enhancement.

Data Cleaning # 2

Data Enhancement:

- Merging historical stock data (yahoo finance).
- Merge risk free interest rates (from ten year treasury bond data)
- Add derived variables to dataframe, useful for premium price calculations

Derived Variables:

- Spread = Bid best price - Bid Offer
- Days_to_expiration = Option expiry date - Option Purchase Date
- Rolling Volatility (for 30, 60 & 90 days) = σ (Daily Percentage change in stock price for 'x' days) *$\sqrt{(x)}$
- Option Pricing :
  - Calls : $N(d_1)*S - N(d_2)*K*e^{-rt}$
  - Puts : $N(-d_2)*K*e^{-rt} - N(-d_1)*S$

  Where $d_1 = [\ln(S/K) + (r + \sigma^2/2)*t]/ (\sigma \sqrt{t})$ & $d_2 = d_1 - \sigma \sqrt{t}$

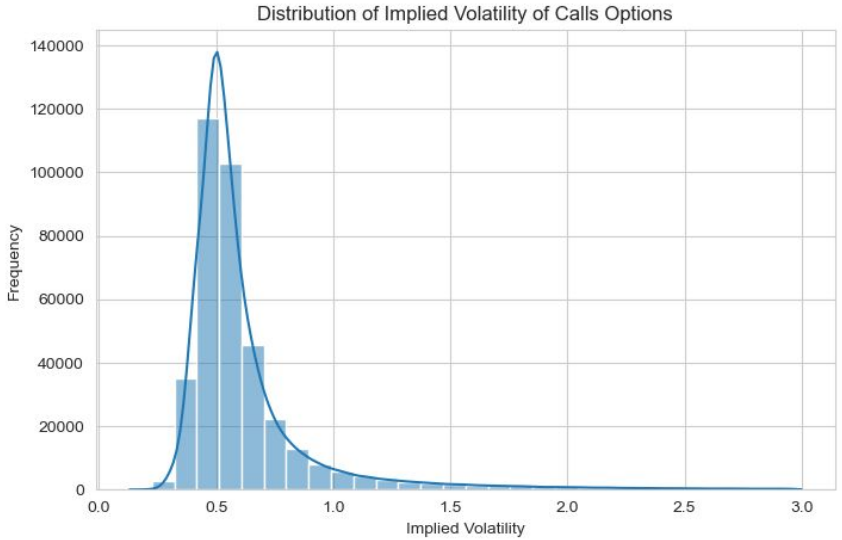
  And, N() is the CDF of the normal distribution function, K is the strike price, σ is volatility, r is risk free interest rate, t is days to expiration, S is the market price of the option

Exploratory Data Analysis

Calls/Puts split

Implied Volatility distribution:

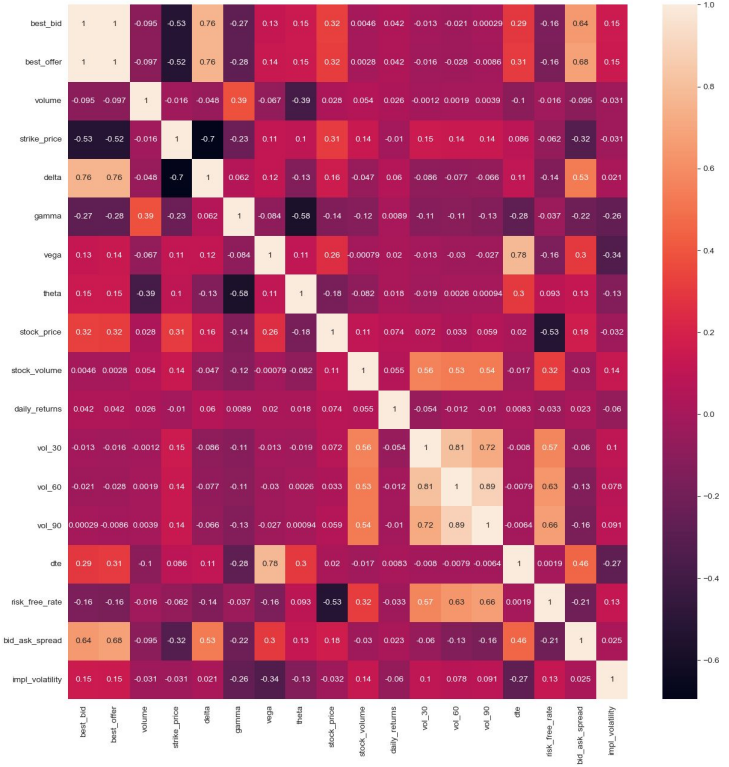
- Right skew distribution, non-linear relationships
- Possible approaches:
  - Random Forest
  - Gradient Boosting
  - Support Vector Regression
  - Neural Networks



Exploratory Data Analysis

- Correlation between the features

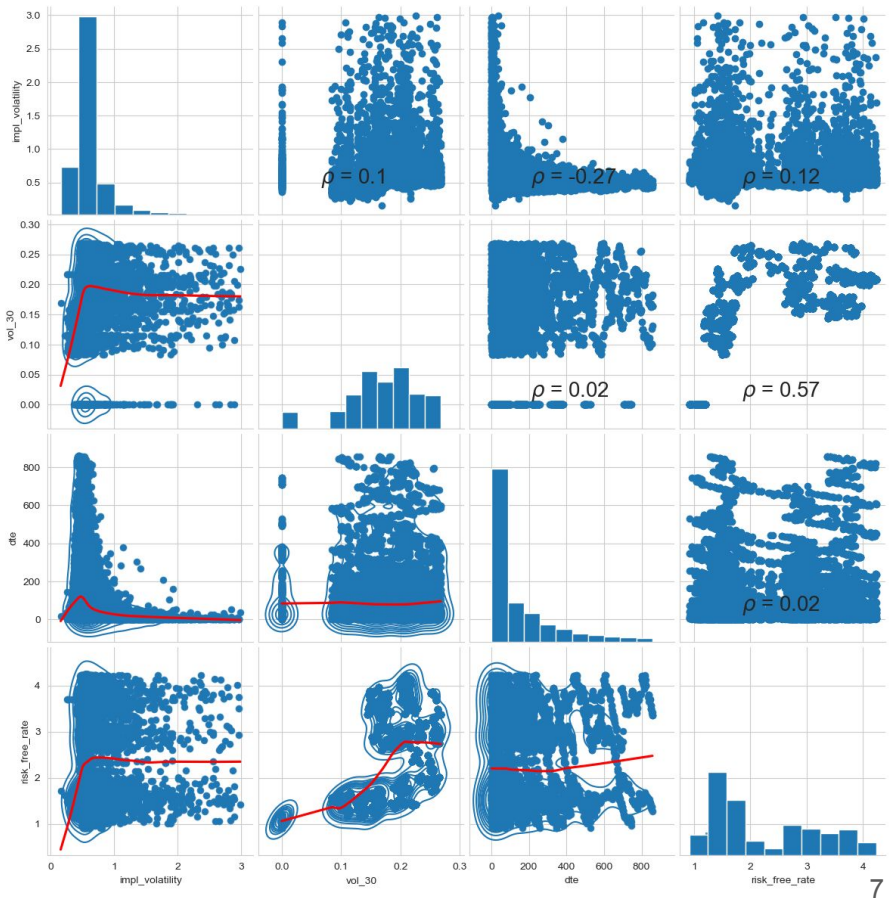
Call options	Best bid	Best offer	Stock volume	…
Implied volatility	0.15	0.15	0.14	…



Call options correlation heat map

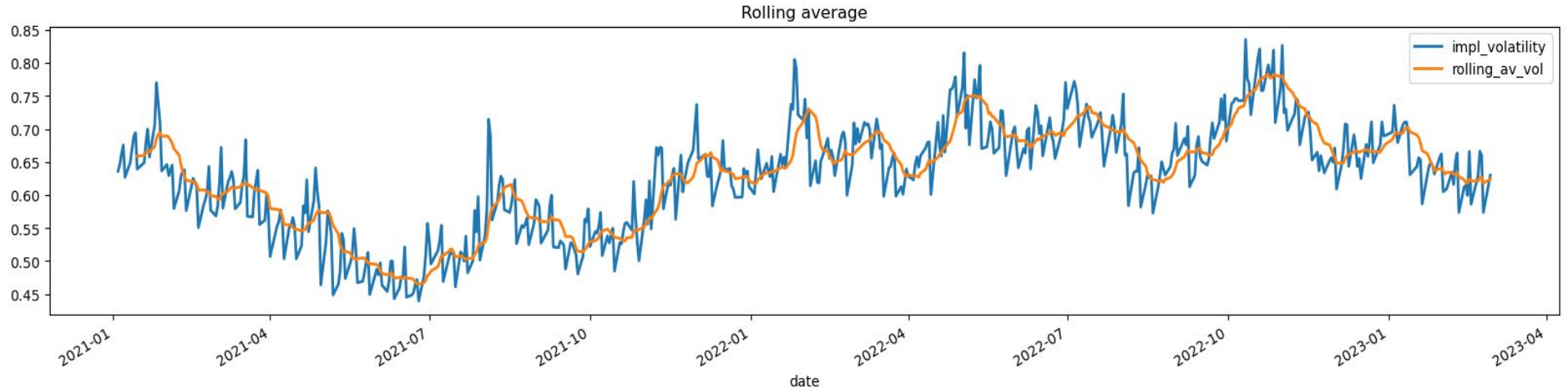
Exploratory Data Analysis

- Call options data
- Kernel density estimate (KDE) plot and Pairplot of some features
  - Non-linear relationship



Base models: Rolling Average and ARIMA

For our base models we select models that are easy to implement and simple to understand. To this end, the simplest model is a simple moving average of step 10 trained on the whole data set:

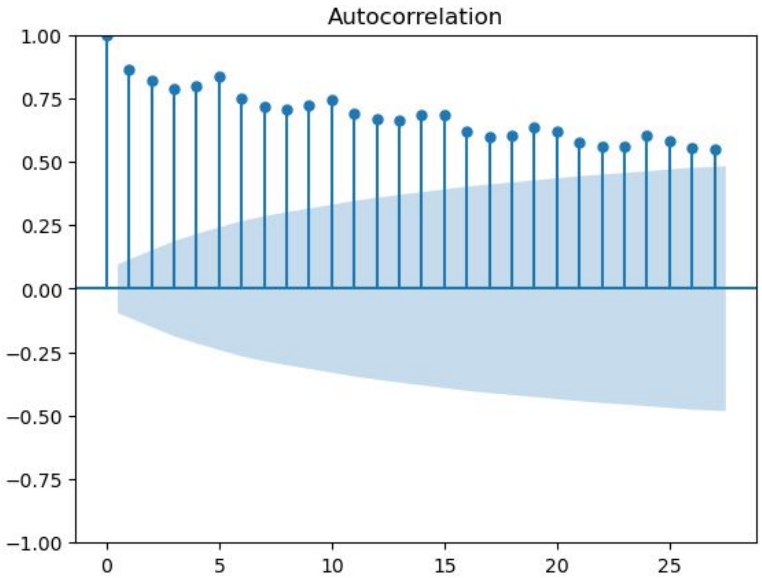


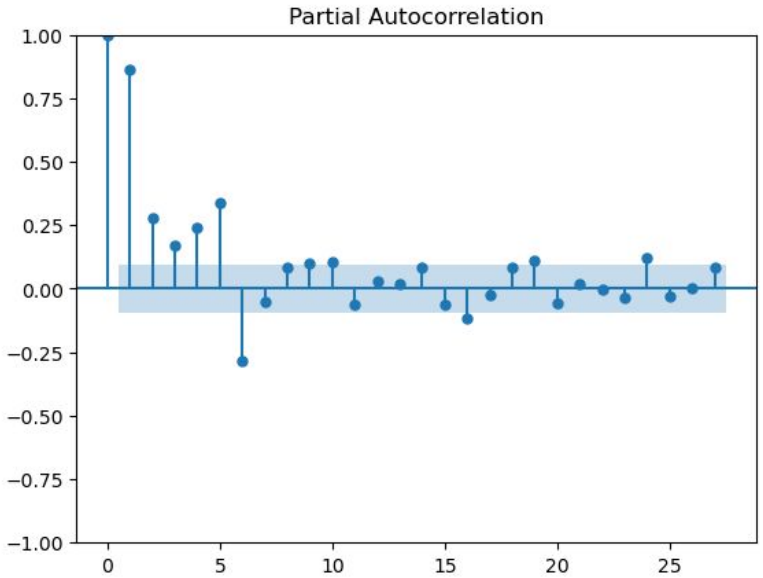
In addition, we select an AutoRegressive Moving Average (ARIMA) model, which is widely used in times series forecasting. The ARIMA model consists of three terms:

- An AutoRegressive (AR) term, measuring the dependence on lagged observations
- An Integrated (I) term, measuring how stationary the time series is (i.e., if the statistical properties of the series remain constant in time)
- And a Moving Average (MA) term, which measures the dependency between an observation and a residual error from a moving average model applied to lagged observations.

We set AR = 0 and MA = 5.

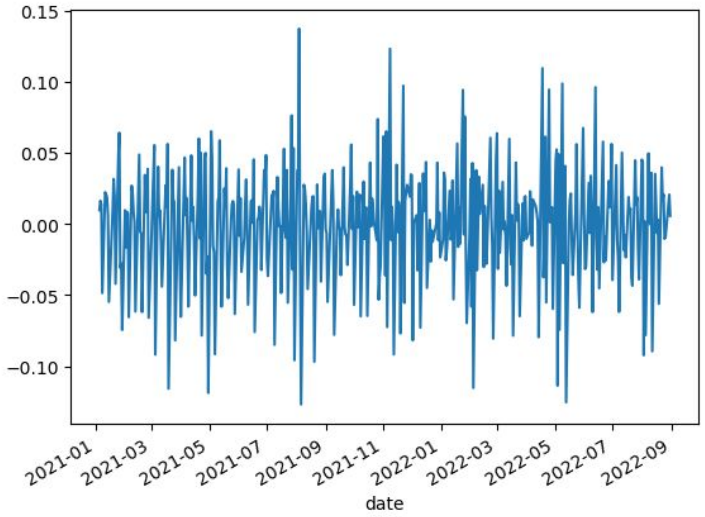
To select the I term in the model we compute the autocorrelation and partial autocorrelation of the time series:

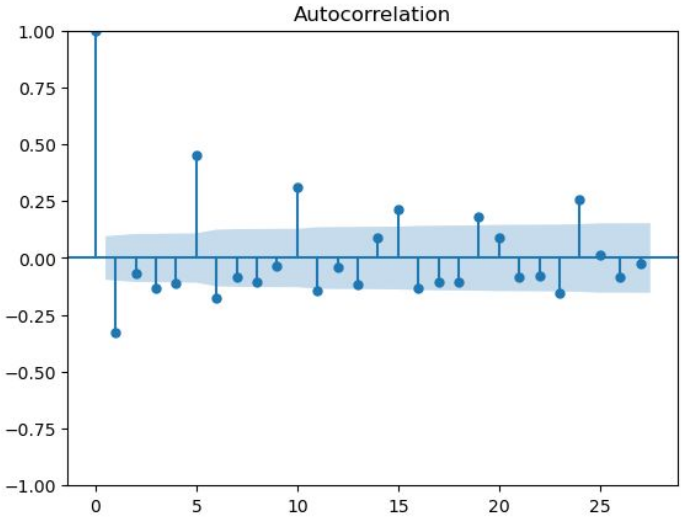




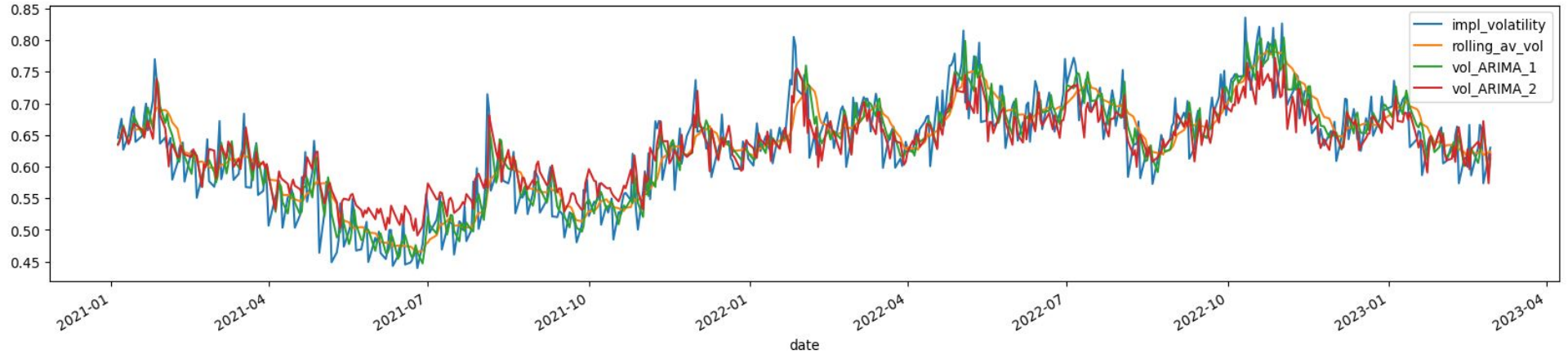
In addition, the p-value from the ADF test nears .5, indicating highly non-stationary data. So we transform to stationary by finding the first differences time series.

This looks much more like stationary, so we select I = 1. In addition, the p-value from the ADF test in near 0, indicating stationary data.





Therefore, we train the ARIMA(0,1,5) in the whole dataset. For comparison, we also train a second ARIMA model with parameters (0,0,5).

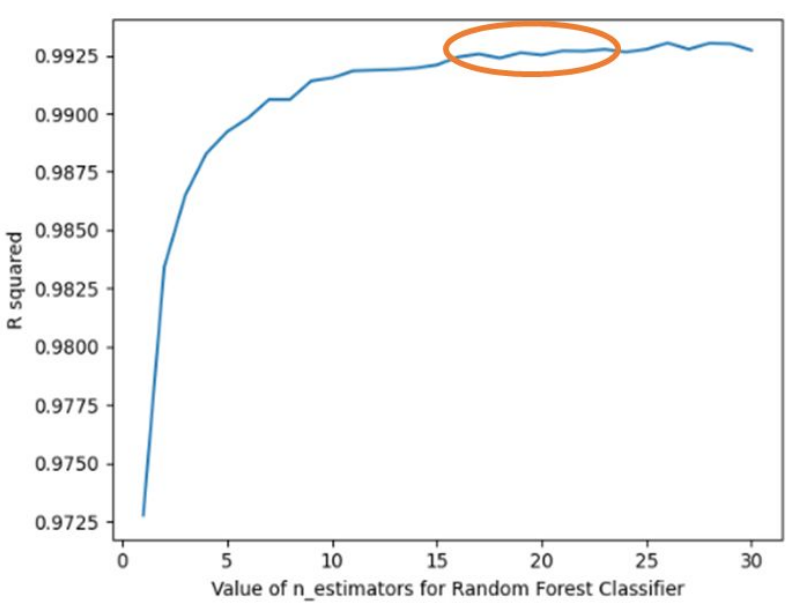


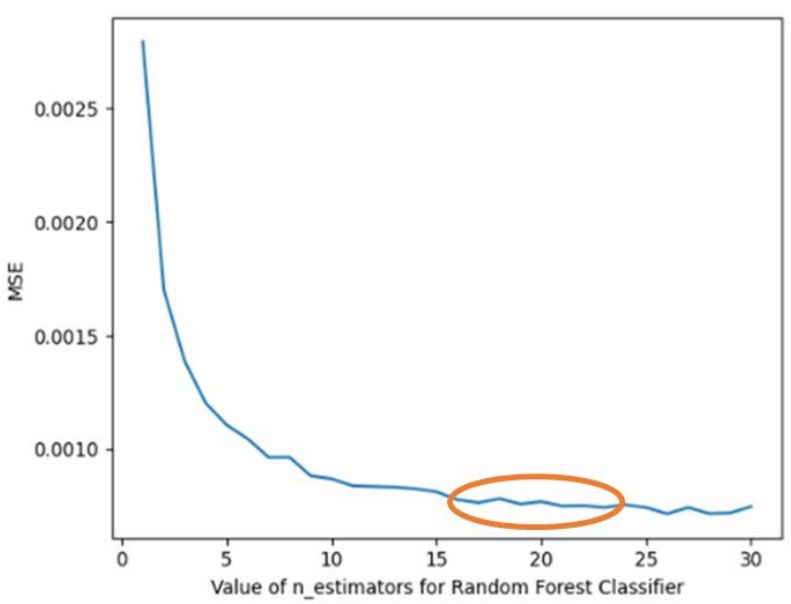
We evaluate the models on the following metrics:

Implied Volatility	Rolling Average	ARIMA(0,1,5)	ARIMA(0,0,5)
MAE	0.0280	0.0279	0.0323
MSE	0.0279	0.0019	0.0016
RSE	0.0323	0.0444	0.0407

Random Forest

- Set the **maximum depth** of each decision tree to be **20**.
- Find the **optimal n_estimators**, or number of trees in the forest, by training simple Random Forest models with the same dataset and n_estimator ranging from 1 to 30.
  - The optimal n_estimator seem to be **around 20**.





Random Forest

- Retrain the model with optimal n_estimator = 20.
- Performance metrics:
  - MAE = 0.0141
  - MSE = 0.0008
  - RSE = 0.0277
  - R squared = 0.9925