## Building a Random Forest Regression model for Forex trading using price indicators and a sentiment indicator

## Project Abstract

The project is about building a machine learning model that could predict next day’s currency close price based on previous days’ OHCL data, EMA, RSI, OBV indicators, and a Twitter sentiment indicator. It is based on a Random forests Regressor because it combines the benefits of trees’ predictive power and avoidance of overfitting. Also, the introduction of a sentiment indicator balances the limitations of technical analysis and draws information from the domain of behavioral economics.

## Introduction/Project Motivation

I think ML regression can give away insights for trading strategy setups that otherwise would remain hidden and on the same time, it isn’t so much a black box as you can evaluate the predictors and decide which ones to use. Also, I Day trade Forex currency pairs and hopefully I could use it on real trade setups. The following steps have been taken:

* Apply Machine Learning to predict the trend using as predictors technical indicators and a sentiment indicator so to create a more robust strategy that would consider both technical and fundamental aspects.
* Getting the data from a reliable forex data provider.
* Download historical data of 2 years as a CSV file.
* Create features.
* Use a Random forest model for the problem.
* Use Cross-Validation.
* Train the model.
* Predict on the test.
* Based on tests and accuracy score make some alternations into the predictors.
* Evaluate the final model.

## Data Mining

Forex currency pair of EUR/USD of a 2 years period.

Source: historical data from mt4 trading platform from 2016-12-31 to 2018-12-31, daily timeframe, total 524 trading days.

## Data Analysis

The TA - library was used to compute four technical indicators, the EMA short and long, the RSI and OBV.

The next day’s close price was used as the prediction target.

The GetOldTweets3 library was used to get historical tweets from the same period of 2016-12-31 to 2018-12-31 with max of 5000 tweets due to time limits.

The textblob library was used to get sentiment polarity, and the data resampled using the daily timeframe for the mean calculation.

As predictors the following were used: ['open', 'high', 'low', 'close', 'volume', 'ema\_short', 'ema\_long', 'rsi', 'obv', 'tweet\_sentiment']

Also, we made use of two splits of data. The first split the data to a training test and an untouched set. The second cross-validated the train set 5 times.

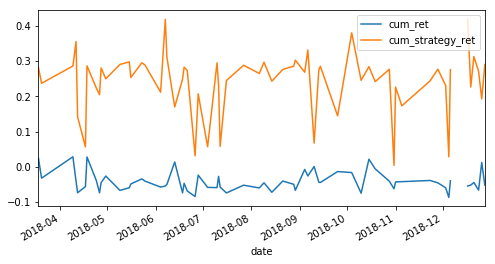
The predictive power of the model tested on the initial untouched set

## Key Findings

It is possible to construct a fairly useful trading model by using ML and particularly Random Forests Regression, using as predictors a mix of price data, technical indicators, and a sentiment indicator.

ML lifts the weight from the shoulder of the trader of finding optimal combinations of various factors and components of trading.

The return of the ML strategy most of the times seems better from the simple buy and hold benchmark strategy.



The Sharpe Ratio of strategy is approximately two most of the times of back-testing it.

Some indicative measures are:

Model score= [0.96615782505226, 0.9499368283378423, 0.9793830955075918, 0.946102298022616, 0.9707968097158307]

Importance of predictors: [0.11186647 0.12960734 0.22891741 0.20857127 0.00230334 0.13688185 0.16457009 0.01260118 0.00233765 0.00234338]

Mean Absolute Error: 0.0045143218795093976

Mean Squared Error: 2.7342703703148714e-05

Root Mean Squared Error: 0.005229025119766467

Visualizing a tree of the model

A picture containing map, text

Description automatically generated

### Challenges/Limitations

The following were a major source of concern:

The gathering of Twitter data was problematic. The 5000 tweets were a compromise between twitter API limitation, time constraints, and local computational force.

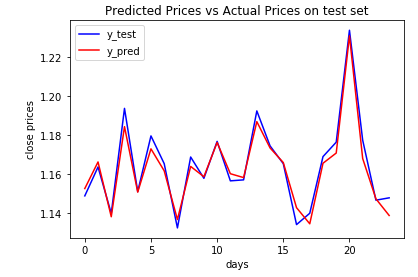
Thus, the final amount of useful data was small to extract robust conclusions.

The need for splitting the data into train, test, validate, and out of sample data further worsened the statistical value of findings.

It is hard to train and implement the model in real time trading due to the constant need for not so readily available twitter data.

### Conclusion:

The model has the potential to be used in practical projects. We can see below how close the predictions on the training and the untouched set are and that it produces a better return from the benchmark with a satisfying Sharpe Ratio. Of course, optimization could and should be done before going live, and that includes the ema periods, the picking of the most useful predictors, the use of further predictors like crossovers, the scaling of some predictors and the use of more in-depth historical data.



A screenshot of a cell phone

Description automatically generated

### Annexure/Codes

Please look at the attached files:

[C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\Source.gv.pdf](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\Source.gv.pdf) [C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\tweets.csv](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\tweets.csv) [C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\Final\_Project\_Christos\_Gklinavos\_v6.html](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\Final_Project_Christos_Gklinavos_v6.html) [C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\Final\_Project\_Christos\_Gklinavos\_v6.ipynb](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\Final_Project_Christos_Gklinavos_v6.ipynb) [C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\forex.csv](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\forex.csv) [C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final Project (Optional)\EPAT-09-Final-Project-Final Submission-Christos-Gklinavos\Source.gv](file:///C:\Users\Christos\SkyDrive\Έγγραφα\QUANTRA\EPAT\EPAT-09-Final%20Project%20(Optional)\EPAT-09-Final-Project-Final%20Submission-Christos-Gklinavos\Source.gv)

### Bibliography

**EPAT-09-Final Project**

[What Does the Individual Option Volatility Smirk Tell Us About Future Equity Returns? | Journal of Financial and Quantitative Analysis | Cambridge Core](https://www.cambridge.org/core/journals/journal-of-financial-and-quantitative-analysis/article/what-does-the-individual-option-volatility-smirk-tell-us-about-future-equity-returns/ECFD16BA9ACBDC8D577D1BD866FBEA72)

[Random Forests Classifiers in Python (article) - DataCamp](https://www.datacamp.com/community/tutorials/random-forests-classifier-python)

[Decision Forest Regression - Azure Machine Learning Studio | Microsoft Docs](https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/decision-forest-regression)

[(50) Azure Machine Learning Studio: Decision Forest Regression - YouTube](https://www.youtube.com/watch?v=5eiE_X6Yr7s)

[Getting started with Neural Network for regression and Tensorflow](https://medium.com/@rajatgupta310198/getting-started-with-neural-network-for-regression-and-tensorflow-58ad3bd75223)

[How (not) to use Machine Learning for time series forecasting: Avoiding the pitfalls](https://towardsdatascience.com/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424)

[Practical Deep Learning for Coders, v3 | fast.ai course v3](https://course.fast.ai/index.html)

[Why Random Forests can’t predict trends and how to overcome this problem?](https://medium.com/datadriveninvestor/why-wont-time-series-data-and-random-forests-work-very-well-together-3c9f7b271631)

[Sentiment Analysis in Trading Explained Using R](https://www.quantinsti.com/blog/sentiment-analysis-in-trading)

[Trading Commodities Using Natural Language Processing](https://knect365.com/quantminds/article/ce6da282-1aad-484f-a0a2-cf704248c792/trading-commodities-using-natural-language-processing)

[Using NLP and Deep Learning to Predict the Stock Market](https://towardsdatascience.com/using-nlp-and-deep-learning-to-predict-the-stock-market-64eb9229e102)

[Sentiment Analysis - When Commodity Trading Meets Deep Learning](http://www.datagenicgroup.com/sentiment-analysis-when-commodity-trading-meets-deep-learning/)

[Cross-Validation in Machine Learning Trading Models](https://www.quantinsti.com/blog/cross-validation-machine-learning-trading-models)

[Machine Learning for Quants and Traders - QuantInsti](https://www.quantinsti.com/blog/overview-machine-learning-trading)

[How Is Machine Learning Used In The Stock Market? [PR]](https://www.quantinsti.com/blog/machine-learning-stock-market)

[Machine Learning For Stock Price Prediction Using Regression](https://www.quantinsti.com/blog/machine-learning-trading-predict-stock-prices-regression)

[Best Algorithmic Trading Blogs of 2017 – QuantInsti® – Medium](https://medium.com/@QuantInsti/best-algorithmic-trading-blogs-of-2017-dc8ce133a9fb)

[Trading Using Machine Learning In Python Part-2 – QuantInsti® – Medium](https://medium.com/@QuantInsti/trading-using-machine-learning-in-python-part-2-8d5fa84c9926)

[Development Of Cloud-Based Automated Trading System With Machine Learning [EPAT PROJECT]](https://www.quantinsti.com/blog/epat-project-automated-trading-maxime-fages-derek-wong)

[Free Forex Historical Data: Forex Currency Pairs](https://www.global-view.com/forex-trading-tools/forex-history/index.html)

[Free Forex Historical Data: Forex Currency Pairs](https://www.global-view.com/forex-trading-tools/forex-history/index.html)

[Forex Trading Resources - Courses, Software, Guides and More!](https://www.tradingheroes.com/resources/)

[TA-Lib](https://mrjbq7.github.io/ta-lib/func_groups/momentum_indicators.html)

[python - How to run Ta-Lib on multiple columns of a Pandas dataframe? - Stack Overflow](https://stackoverflow.com/questions/51712269/how-to-run-ta-lib-on-multiple-columns-of-a-pandas-dataframe)

[Using DataFrames and TA-Lib in python · Issue #120 · mrjbq7/ta-lib · GitHub](https://github.com/mrjbq7/ta-lib/issues/120)

[talib.SMA Python Example](https://www.programcreek.com/python/example/92310/talib.SMA)

[TA-Lib](https://mrjbq7.github.io/ta-lib/func.html)

[How to find Forex historical data that will help you better backtesting](https://forextester.com/data/datasources)

[Trading Strategy: Technical Analysis with Python TA-Lib](https://towardsdatascience.com/trading-strategy-technical-analysis-with-python-ta-lib-3ce9d6ce5614)

[3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)

[Twitter Sentiment Analysis using Python - GeeksforGeeks](https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/)

[(51) Twitter Sentiment Analysis | Sentiment Analysis In Python Using Tweepy and Textblob | Edureka - YouTube](https://www.youtube.com/watch?v=27P268Q7pE0)

[Sentiment Analysis of Twitter Timelines – Dmitry Rastorguev – Medium](https://medium.com/@dmitryrastorguev/sentiment-analysis-of-twitter-timelines-61c73eeacedf)

[This Is How Twitter Sees The World : Sentiment Analysis Part One](https://towardsdatascience.com/the-real-world-as-seen-on-twitter-sentiment-analysis-part-one-5ac2d06b63fb)

[How to get twitter historical data?](https://www.researchgate.net/post/how_to_get_twitter_historical_data)

[GitHub - Jefferson-Henrique/GetOldTweets-python: A project written in Python to get old tweets, it bypass some limitations of Twitter Official API.](https://github.com/Jefferson-Henrique/GetOldTweets-python)

[- Orgneat](http://www.orgneat.com/Request)

[GitHub - taspinar/twitterscraper: Scrape Twitter for Tweets](https://github.com/taspinar/twitterscraper/)

[How to use it? · Issue #52 · Jefferson-Henrique/GetOldTweets-python · GitHub](https://github.com/Jefferson-Henrique/GetOldTweets-python/issues/52)

[GetOldTweets3 · PyPI](https://pypi.org/project/GetOldTweets3/)

[python - Get data from Twitter using Tweepy and store in csv file - Stack Overflow](https://stackoverflow.com/questions/21865413/get-data-from-twitter-using-tweepy-and-store-in-csv-file)

[Reading and Writing CSV Files in Python – Real Python](https://realpython.com/python-csv/)

[string - Python write to CSV line by line - Stack Overflow](https://stackoverflow.com/questions/37289951/python-write-to-csv-line-by-line)

[TextBlob: Simplified Text Processing — TextBlob 0.15.2 documentation](https://textblob.readthedocs.io/en/dev/)

[Random Forest Algorithm In Trading Using Python](https://www.quantinsti.com/blog/random-forest-algorithm-in-python)

[RandomForest Algorithmic Trading Strategy With 85% Winrate - Binary Options Ninja](http://binary.tradingninja.com/randomforest-algorithmic-trading-strategy-with-85-winrate/)

[Gold Price Prediction Using Machine Learning In Python](https://www.quantinsti.com/blog/gold-price-prediction-using-machine-learning-python)

[Twitter Sentiment Gather Data](https://vicmora.github.io/blog/2017/02/20/twitter-sentiment-gather-data)

[Decision Trees](https://blog.quantinsti.com/decision-tree/)

[Ensemble Methods - Bagging And Boosting](https://blog.quantinsti.com/ensemble-methods-bagging-boosting/?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_may2019)

[Gini Index For Decision Trees](https://blog.quantinsti.com/gini-index/?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_may2019)

[Detecting Bots On Twitter Using Botometer](https://blog.quantinsti.com/detecting-bots-twitter-botometer/?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_may2019)

[Cross Validation in Machine Learning Trading Models](https://blog.quantinsti.com/cross-validation-machine-learning-trading-models/)

[sklearn.model\_selection.TimeSeriesSplit — scikit-learn 0.21.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html)

[How To Backtest Machine Learning Models for Time Series Forecasting](https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/)

[Random Forest Algorithm with Python and Scikit-Learn](https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/)

[Train/Test Split and Cross Validation in Python – Towards Data Science](https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6)

[Predicting Stock Trends Using Technical Analysis And Random Forests](https://blog.quantinsti.com/predicting-stock-trends-technical-analysis-random-forests/?utm_source=attendee&utm_medium=email&utm_campaign=11jun2019web)