Machine Learning in Currency Trading

Fernanda R Bordin

Jingjing “Doris” Zhu

Yihui “Jennifer” Zhou

San Jose State University

Author Note

Fernanda R Bordin, Jingjing Zhu, Yihui Zhou – MS in Data Analytics, San Jose State University

1 Washington Sq, San Jose, CA 95192

Contents

[Executive Summary 3](#_Toc528843517)

[Machine Learning applied to Financial Market - Doris 3](#_Toc528843518)

[Foreign Exchange Market (explain and dynamics) – Fernanda 3](#_Toc528843519)

[Machine learning model selection (why forest) – Fernanda 3](#_Toc528843520)

[Linear Regression 4](#_Toc528843521)

[Logistic Regression 4](#_Toc528843522)

[Classification and Regression Trees 4](#_Toc528843523)

[Support Vector Machines 4](#_Toc528843524)

[Bagging and Random Forest 4](#_Toc528843525)

[Data origin – Jen / Doris 4](#_Toc528843526)

[Features Selection (and explanation of what they are) – Jen/Doris 4](#_Toc528843527)

[Results from application 4](#_Toc528843528)

[Evaluation of the model 4](#_Toc528843529)

[Conclusion 4](#_Toc528843530)

[References 4](#_Toc528843531)

# Executive Summary

**xxxxxxxxxxxx**

# Machine Learning applied to Financial Market – Doris

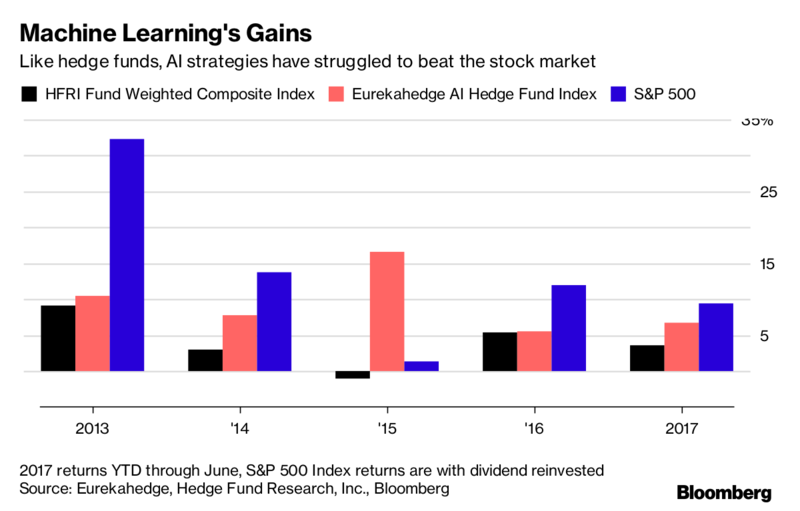
Machine learning, a popular topic we can see everywhere, is not only rolling up a great upsurge in tech giants, but it is also attracting banking and financial sectors to adopt machine learning technologies. Because of the quantitative characteristic of financial industry, it is one of the most suitable industries to apply artificial intelligence and machine learning, a way to achieve artificial intelligence. There are currently at least four areas of applications of machine learning in finance industry, including portfolio management, algorithmic trading, fraud detection and loan/insurance underwriting.

Portfolio management:

It is now common for companies such as Betterment and Wealthfront Robo-advisors algorithms to help users do financial portfolio management. Based on users’ goals and real-time market changes, the system will adjust the portfolio allocation in a timely manner, without the necessity of meeting a physical advisor. (Faggella, 2018)

Algorithmic trading:

Many hedge funds and financial institutions are using complex artificial intelligence systems to do automated trading systems. “High-frequency trading”, which is one kind of algorithmic trading, allows these institutions to make real-time trading decisions. Although the systems that they are using are confidential, there is no doubt that they are using machine learning and deep learning technologies to help them make trading decisions. (Faggella, 2018)



Fraud Detection:

As there was booming online data produced in the past years, the problem of data security risk is drawing people’s attention. Instead of using traditional ways to detect finance fraud by following rules, machine learning can learn and catch potential threats by identifying unique actions and mark them for further security checks. (Faggella, 2018)

Loan/Insurance Underwriting:

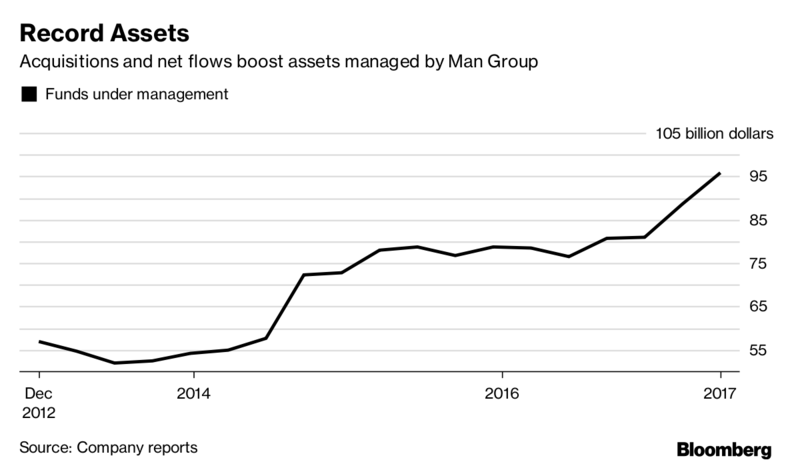
Machine learning algorithms are tending to change the traditional positions of underwriting. By using large amount of customer data, including customers’ general information and other information such as their default rate and car accidents records, large companies can use this large volume of data to train their machine learning algorithms. (Faggella, 2018)

**Examples of companies applying machine learning and their performance – Doris**

Man Group:

Many hedge fund companies were applying computer algorithms to do their quantitative trading decisions, and machine learning, different from the traditional quant, enable computers to find patterns based on large amount of data and trade on their own. Man Group, one of the largest hedge fund companies in Wall Street, was able to apply machine learning algorithms to make autonomous financial trading decisions because of the availability of data and computing resources. (Picker and Funk, 2017)

By 2015, confidence in the technology has been built because artificial intelligence helped to gain almost half of the profit of one of Man’s biggest funds. They are also using machine learning technologies to perform fastest trading, make bets on market momentum, and to scan keywords that might be signals of market change. Total assets under management of Man have raised about 77% and ADL Dimension fund assets has been 5 times since 2014. (Satariano and Kumar, 2017)



JPMorgan Chase:

There have been discussions about banks applying AI to reach automation and the potential to cut traditional jobs. JPMorgan Chase has been invested in a machine learning technology called Contract Intelligence (COiN). It is a platform primarily to “analyze legal documents and extract important data points and clauses.” Generally, it takes about 360,000 hours for humans to review the 120,000 annual commercial credit agreements but applying this machine learning technology will reduce the time to a few seconds. In addition, in 2016, the firm also planned to push out its virtual assistant technology to respond to employees’ service desk requests by integrating a natural language interface, expecting that 1.7 million requested can be addressed each year. The firm is investing a lot in new technologies to focus on new initiatives and to incorporate fintech solutions. (Sennaar, 2018)

# Foreign Exchange Market (explain and dynamics) – Fernanda

Of the many potential trades in the stock market, the foreign exchange market is very well suited for the application of machine learning. It’s mainly constituted by the sale, purchase, exchange and speculation of currencies. Also known as FX, forex or currency market, these operations enable currency conversion for international investments and trades.

Due to its massive trading volume, assets have a high liquidity. FX also operates 24h, 5 days a week and is an over-the-counter market, with most trades taking place on electronic platforms or via telephone. All characteristics that create good data and reinforce its attractiveness for machine learning. Since only 3% of the exchange are transacted as futures or options contract, the scope of this study will be kept focus into spot exchange. A spot FX transaction entails buying one currency and simultaneously another for settlement in a period of two days.

Currencies from every country are traded in the foreign exchange market, but the major currencies are Dollar (USD), Euro (EUR), Yen (JPY), Pound (GBP), Canadian Dollar (CAD), Franc (CHF), Australian Dollar (AUD). Usually, currencies are traded in pair, for the study, we’ll focus in EUR/USD: Euro Zone / United States.

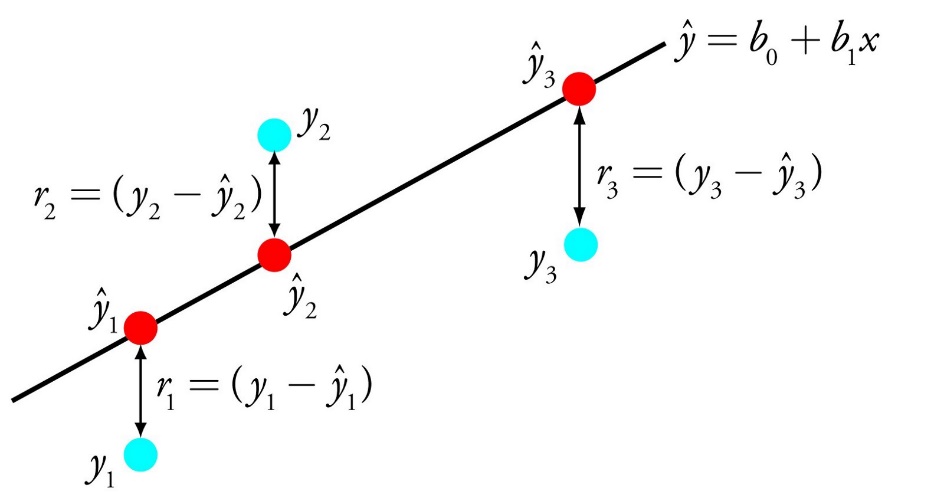
Currency exchanges are commonly heavily leveraged, that means that an investor can borrow and control more than his initial investments. For example, he can invest at a 1:50 leverage, which means that for each dollar deposited, 50 dollars can be committed in investments. On the data section it will be explained how we treated the data for this effect.

# Machine learning model selection (why forest) – Fernanda

In machine learning there isn’t one algorithm that always work best for all problems. It is common to try several different solutions using a test set and comparing performance. Based on the data and problem characteristics it’s possible to select the model that is most likely to succeed and start from there. Some of the most common models are:

## Linear Regression

The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B). (Le, 2018)



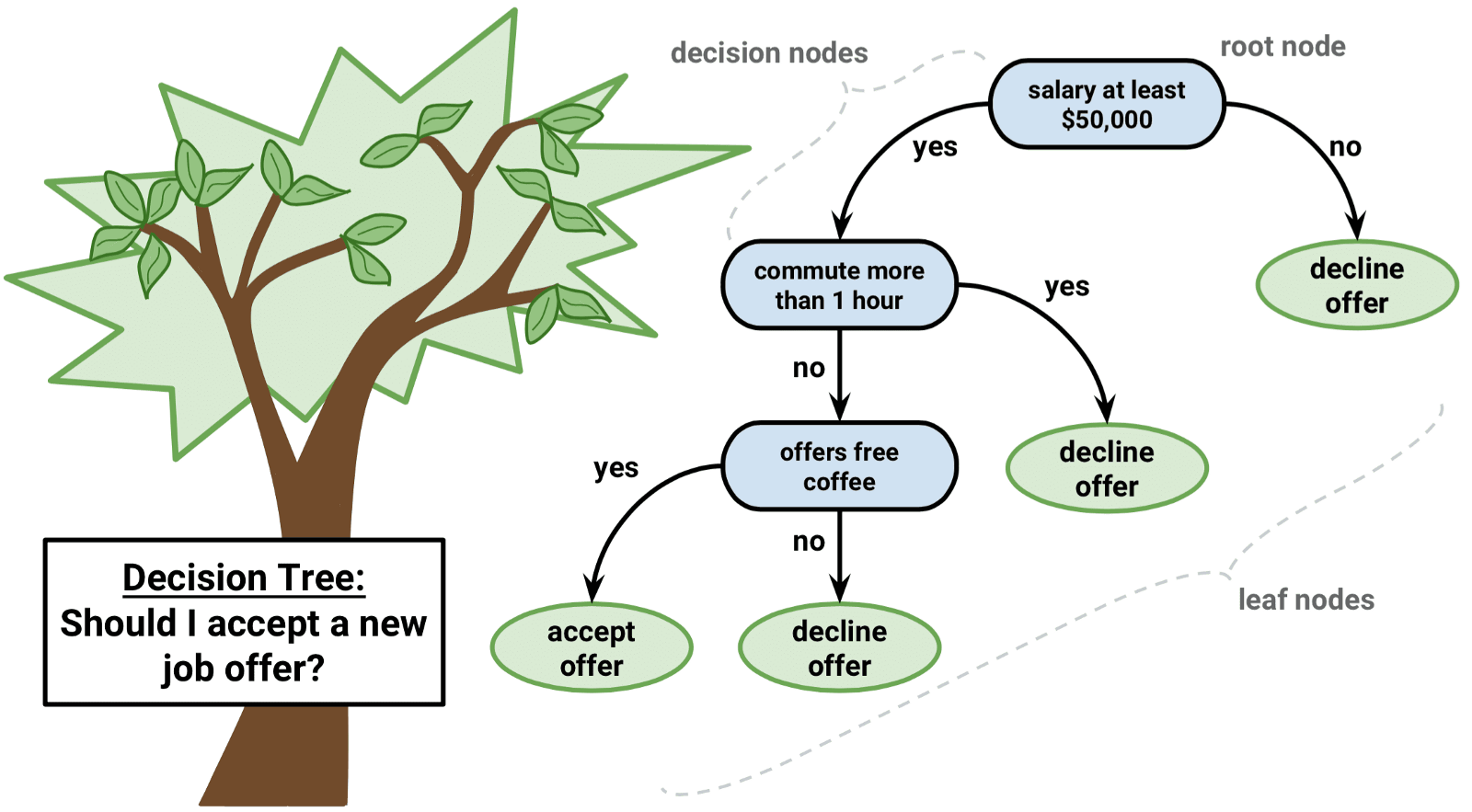
For example: y = B0 + B1 \* x

Through this equation, it predicts y given the input x and the goal of the linear regression learning algorithm is to find the values for the coefficients B0 and B1. Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible. It is a fast and simple technique and good first algorithm to try. (Le, 2018)

Given the volatile nature of the currency market, its variability and the relevance of catching the trends of appreciation or depreciation, this is not the right model.

## Classification and Regression Trees

Traditionally, the decision tree model is a binary tree from algorithms and data structures. Each node is a single input variable (x) and a split point on that variable. The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. (Le, 2018)

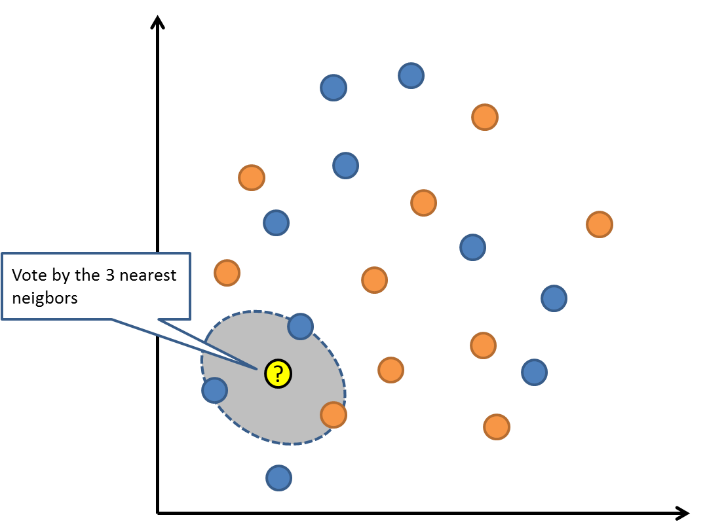


One of the main characteristics of the stock market is the diversity of thinking from different investors. This means that not all investors act the same way on each node and therefore make the decision tree an incomplete model for currency exchange prediction.

## K Nearest Neighbors

Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value. (Le, 2018)

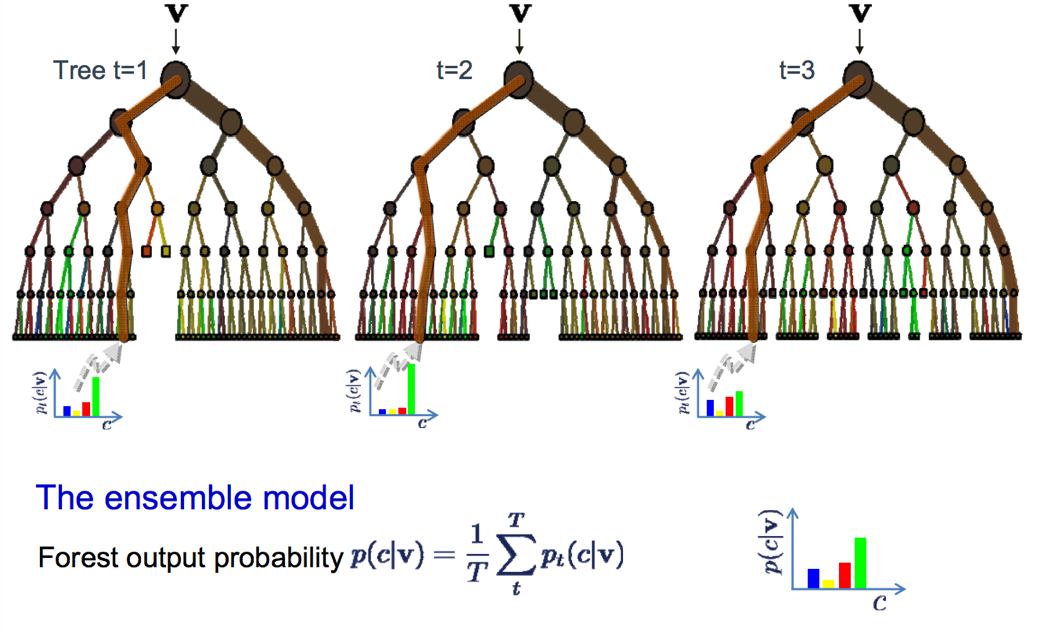
The trick is in how to determine the similarity between the data instances. The simplest technique if your attributes are all of the same scale (all in inches for example) is to use the Euclidean distance, a number you can calculate directly based on the differences between each input variable. (Le, 2018)



Despite being a good model to predict behavior of different investors, considering the timing of each action and the trend might prove to be a challenge.

## Bagging and Random Forest

In Random Forest several decision trees are used to predict the output. Multiple samples of your training data are taken then models are constructed for each data sample. When you need to make a prediction for new data, each model makes a prediction and the predictions are averaged to give a better estimate of the true output value. (Le, 2018)



The decision trees are created so that suboptimal splits are made by introducing randomness. The models created for each sample of the data are therefore more different than they otherwise would be, but still accurate in their unique and different ways. Combining their predictions results in a better estimate of the true underlying output value.

# Data origin – Jen / Doris

We collected all financial data from Bloomberg terminal, including fundamental market indicators and technical indicators which are calculated and based on time series. We have included a detailed description on the indicators in the features selection part.

We selected EUR/USD exchange rate prediction dataset. Specifically, the paper aims to utilize machine learning model to predict whether the currency pair will travel sideways, meaning appreciate or depreciate above certain range or remain range-bound. Thus we decide to categorize EUR/USD price change into three groups: if EUR/USD appreciates more than 0.5%, price change will be labeled ‘1’; if EUR/USD depreciates more than 0.5%, price change will be labeled ‘-1’; if EUR/USD price change maintains within the range -0.5% to 0.5%, price change will be labeled ‘0’.

We also selected our frequency as weekly, over a period of 4 years. We selected weekly frequency as our study wanted to look over a longer term and try to find patterns over longer period instead of bringing out instant trading solutions.

# Features Selection (and explanation of what they are) – Jen/Doris

Back in the days in Economics class, we all learned Interest Rate Parity, Purchasing Power Parity and learned the fancy, sometimes confusing curves trying to explain fluctuations in the foreign exchange market. However, in real work, foreign exchange rates are not easily determined by several formulas. Foreign exchange rates can be impacted by a myriad of features, from fundamental economy, to economic events and emotions. Feature selection is an important integral part to machine learning models and in this paper we decide to include both fundamental and technical indicators to our feature.

In the fundamental features, indicators derived from the Interest Rate Parity are included, for example interest rates in the US over 2year and 10year. As foreign exchange trades in pair, and our focus is EUR/USD, so for the base rate we decided to choose US interest rate which has reflected a pickup since late 2015 when interest rate was at historical low. In this year, the consistent rise in US interest rate has led to market turmoil since July 2018. A clear comparison is that with EUR base rate which has been lingering from negative to zero for the past three years to boost European economy.

The development of modern finance is a story of international finance with international trade becomes ever more robust. Thus, it is necessary to examine the impacts from other trading currency pair to include in our model features.

With the growth in the capital market, economy fundamental indicators are gradually losing their impact on the currency movement to supply and demand in the capital market and investment fund flows. Currency is the denomination of many asset prices and a significant change in asset prices linked with investment fund flows would certainly impact foreign exchange rate. A good, large volume, and liquid indicator of such capital markets is stock market. So we have included major stock market indices into the feature as well, including S&P 500 index, Dow Jones Index, Euro Stoxx Index, FTSE100 Index, and Topix Index. Another major one denominated in USD is oil. So we included oil price, WTI in the features too.

EUR and JPY are traditionally viewed as hedging currencies, that is, when there is turmoil or heightened volatility in the market investors tend to resort to EUR, JPY for safety. Given this nature, we also included Gold price in the features.

Below is a full list of fundamental indicators that we have included in the feature selection.

|  |  |
| --- | --- |
| Features | Explanation |
| AUD/USD exchange rate | Daily Price of AUD/USD exchange rate |
| CAD/USD exchange rate | Daily price of CAD/USD exchange rate |
| GBP/USD exchange rate | Daily price of GBP/USD exchange rate |
| USD/JPY exchange rate | Daily price of USD/JPY exchange rate |
| SPX Index | S&P500 index daily price |
| CCMP Index | Nasdaq Composite index |
| DJI index | Dow Jones Index |
| SX5E Index | Euro Stoxx 50 Index |
| UKX Index | FTSE100 Index |
| TPX Index | Topix Index |
| CL1 COMB Comdty | WTI oil price |
| XAU BGN | Gold price |
| USGG2YR | US government 2 year rate |
| USGG10YR | US government 10 year rate |

The foundation of technical indicators is that historical price action predicts future price action. It assumes that all the factors that influence a price, whether it be economic, political, social or psychological, have already been factored into the current exchange rate by the market. Technical indicators are calculated using price, trading volume, price change percentage, and similar data to generate time series data to reflect possible price changes in financial asset prices. Common calculators include open, close, low, high, average price, standard deviation etc. Foreign exchange traders have their own favorite toolbox of technical indicators. In this paper, we selected most popular technical indicators from Bloomberg.

A list of detailed descriptions follow:

GPO Bar Chart (Open, High, Low, Close):

Displays the open/high/low/close chart view of the Price Chart.

Simple Moving Average (SMAVG (5) on Close):

SMA is an arithmetic moving average calculated by summing recent close price and then dividing by the number of time periods.

Trading Envelop (TE UB(20,2), TE LB(20,2)):

Envelopes are technical indicators that are typically plotted over a price chart with upper and lower bounds. Traders can interpret envelops to define trading ranges.

Parabolic Studies (PTPS(0.02)):

PTPS graphs stop-and-reversal(SAP) that are suggesting when SAP points are break, the position is closing and entering into an opposite direction.

Bollinger Bands(UBB(2), BollMA(20), LBB(2), BollW (EUR BGN),%B (EUR BGN)):

Bollinger Bands is a study used to identify periods of high and low volatility by using standard deviation around a simple moving average.

Hurst Exponent (Hurst(25) (EUR BGN)):

Hurst Exponent uses historical information to predict future prices. It is a measure of long-term memory of time series relates to autocorrelation.

Stoller Average Range Channels(Moving Average(6,14), Upper STARC Band(2), Lower STARC Band(2), Lower STARC Band(2), STARC%B (EUR BGN)):

Similar to BOLL, STARC also plot two bands around a simple moving average, but instead of standard deviation, it uses adds or subtracts the Average True Range to the moving average.

Commodity Channel Index (CMCI(13) (EUR BGN)):

CMCI measures the variation of a security’s price from its statistical mean. It can identify possible divergences that may indicate a forthcoming trend for a selected security.

Directional Movement Index ('+DMI(14), '-DMI, ADX):

DMI tells the directional movement of a security using today’s high and low prices relative to the previous day’s high and low prices.

ADX is a trend strength indicator. the ADX value is a measure of the strength of the trend regardless of the trend direction; the higher the value of ADX, the stronger the trend.

Moving Average Convergence/Divergence(MACD(12,26), Sig(9), Diff):

Moving average convergence divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of prices.

Elder Impulse System (Moving Average(13), MACD(12,26), Signal(9), Difference):

The Elder Impulse System is a paint bar study identifies bullish and bearish phases in any market and timeframe by combining two indicators.

Keltner Bands (KLTN UB(10,100), KLTN MA(10), KLTN LB(10,100)):

Keltner Bands are bands drawn above and below a center line to illustrate bullish and bearish breakouts. The center line is a simple moving average of the 'Typical Price'.

Trend, momentum, volume and volatility (Upper Band(HLCAverage,20), Average(HLCAverage,20), Lower Band(HLCAverage,20), CCI(HLCAverage,20) (EUR BGN), ):

TMV combines trend, momentum, volume and volatility to present an analysis tool that will provide a more in-depth view of the state of a trend.

TrendStall (ADX ROC(5), ADX MA(5)):

TrendStall identifies points at which a trend is losing momentum and is likely to stall or consolidate. The determination is based on a Rate of Change of the ADX.

Volatility Based Envelops (Upper Band, Lower Band):

The Volatility Based Envelopes (VBE) are two envelopes placed above and below the price action that dynamically adapt to volatility changes in prices without affecting the smoothness integrity of the envelopes.

WLPR William % R

According to Investopedia, Williams %R, also known as the Williams Percent Range, is a type of momentum indicator that moves between 0 and -100 and measures overbought and oversold levels. The Williams %R is commonly used to find entry and exit points in the market. It is, however, prone to false signals as it moves between overbought and oversold. For that reason, using the indicator alongside other price and trend methodologies can help mitigate some of the false signals.

# 

# Results from application (Jen)

The project is conducted entirely on jupyter notebook with many Python packages.

The project can be divided into three parts: data preprocessing, model training, and model evaluation.

In the first part, we examined the columns of both fundamental and technical indicators using correlations between each other. With regard to the dependent variable, which is the direction of EUR/USD exchange rate, we categorized them into three types: if EUR/USD appreciates more than 0.5%, label is 1; if EUR/USD depreciates more than 0.5%, label is -1; and if EUR/USD fluctuates between the range -0.5% and 0.5%, label is 0. We used lambda function to achieve labeling. To check whether the grouping makes sense, we also checked the numbers of data points in three categories and all three categories contain approximately the same number of data points. We also did standardization on the features to be in the same range as some of the features, such as SPX index, the number could be in thousands whereas exchange rates could be less than 1 for example AUD/USD exchange rate. Lastly, we did a Pearson correlation study to check the correlation between features and labeled data. This step is quite meaning as we spotted one technical indicator, the ROC (Rate of Change) is exactly the same as current EUR/USD exchange rate as the correlation is 1 between ROC and EUR/USD exchange rate. So we decided to remove it from features.

In the second part, we chose four supervised learning models to train and test in order to predict which direction EUR/USD exchange rate will go among the three categories. Four models include: Random Forest, K Nearest Neighbors, Gradient Boosting, and Logistic Regression. The four models are selected due to their strengths. For example, within classification models, if we are looking for accuracy, we could consider Kernel SVM, Random Forest, Neural network and Gradient Boosting Tree. If we are looking for speed and explainability, we can choose decision tree or logistic regression. Revisiting our goals here suggested that we adopt ensemble models, so we choose random forest, gradient boosting; and we choose logistic regression as a traditional method to handle classified labeled data; lastly we choose KNN, a clustering method as comparison in performance knowing that clustering method might not be the most suitable method here.

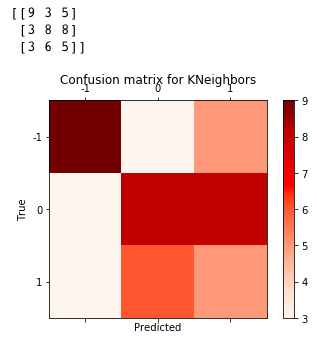
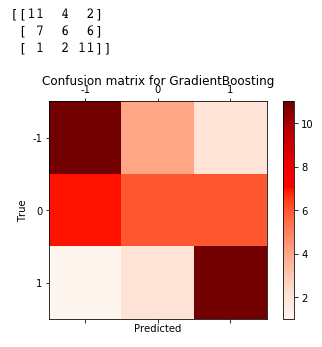
We chose 80% data as training data, and 20% data as test data. We also applied cross validation here to solve for hyperparameter under regularization before working on the coefficients. According to Wikipedia, In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once.

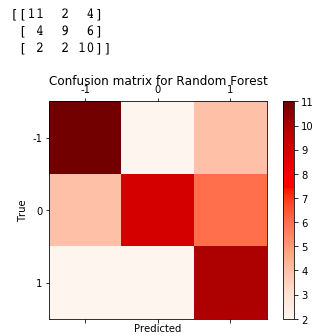
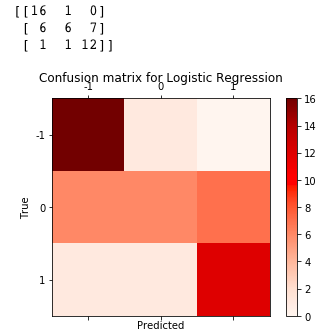
The training data essentially helped us determine the lambda. In python package, we ran the GridSearch package embedded in python to select the most suited parameters for teach model. We also included 5-fold cross validation. Then we get our parameters and fitted the models to test them.

In the third part, we evaluated the models using accuracy rate on training, confusion matrix, and feature importance on the random forest. Among the accuracy rates on training data, all three models except KNN shows accuracy around 53% with logistic regression the highest 53.6%. This result of 53.6% is better off than a random guess. As we have discussed earlier, we have approximately the same number of data points in three classified buckets so the probability of being right for each category would be around 33%.

We then used confusion matrix to compare and analyze all models. The confusion matrix is used to describe performance of a classification model. The data sets are evaluated on actual class and predicted class matrix and has four types: True Positive (TP), False Negative (FN), False Positives(FP), and True Negatives (TN). As we now face three categories in our classified outcome, we decided to show the exact numbers and color range in the confusion matrix as per below. The color block on the diagonal should be close to crimson to have a higher accuracy rate.







Lastly, for the random forest model, we tested feature importance, a method to see how the branching is done for the random forest. Feature importance shows us how important each feature is to the tree. In our model, there is no dominating feature on the random forest thus the random forest model is acceptable from this perspective. In fact, we tried to include ROC (rate of change) in the features and ran the modeling. The result shows mistake in the modeling at the feature importance stage too: ROC has a much larger feature importance of over 30% of all features. After removing ROC, the largest feature importance now is around 6%.

# Conclusion

XXXXXX

# References

Faggella, D. (2018). *Machine Learning in Finance – Present and Future Applications.* Retrieved from

TechEmergence

https://www.techemergence.com/machine-learning-in-finance/

Picker, L. and Funk, J. (2017). One of the world's largest hedge funds is now letting computers trade

completely on their own. Retrieved from CNBC

<https://www.cnbc.com/2017/09/28/man-group-one-of-worlds-largest-funds-moves-into-machine-> learning.html

Satariano, A. and Kumar, N. (2017). The Massive Hedge Fund Betting on AI. Retrieved from

Bloomberg

https://www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai

Sennaar, K. (2018). *AI in Banking – An Analysis of America’s 7 Top Banks*. Retrieved from

TechEmergence

https://www.techemergence.com/ai-in-banking-analysis/

Investopedia. (2017). *Foreign Exchange Market*. Retrieved from Investopedia: https://www.investopedia.com/terms/forex/f/foreign-exchange-markets.asp

Le, J. (2018). *10 Algorithms for Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

NASDAQ. (2018). *NASDAQ - FOREX Education* . Retrieved from NASDAQ: https://www.nasdaq.com/forex/education/advantages-of-trading-forex.aspx

Prado, M. L. (2018). Advances in Financial Machine Learning. In M. L. Prado, *Advances in Financial Machine Learning.* New Jersey: John Wiley & Sons, Inc.

Williams %R technical indicator <https://www.investopedia.com/terms/w/williamsr.asp>

Confusion Matrix, <http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/>

Cross Validation, <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>