Do the firm characteristics, financial performance and macroeconomic variables can predict future returns of stock exchange ?

# Summary

* Abstract
* Introduction
* Literature review
* Methodology (question, hypothesis, benchmark, data, variables)
* Results
* Analysis / Discussion / interpretation
* Conclusion
* References
* Appendix (glossary, data collection, program documentation)

# Abstract

In this study, the firm characteristics, financial performance and macroeconomic variables were investigated to predict the future returns of several stock exchanges between the years 2017 to 2021.

**Results indicated a significant relationship among the daily volume, the MACD and the stock prices.**

# Introduction

The purpose of this project was to analyse the relationship between firm characteristics, financial performance and macroeconomic variables (1-2-3) and future returns on several stock exchanges between the years 2017 to 2021.

Many researchers have analyzed the impact of firm characteristics (1), financial performance (2) and macroeconomic variables (3) on stock returns and found that many variables can predict stock returns like ‘Firm size’ or ‘Inflation’.

I tried to verify these affirmations on several stock exchanges with recent data.

My research question is «Do the firm characteristics, financial performance and macroeconomic variables can predict future returns of stock exchange ?»

I tried to answer the research question by analysing 23 stock exchanges, 6 variables in the period 2017-2021.

I analysed 1 **firm characteristic** :

* **Total assets** of the last 5 years, which estimates the **firm size**

I analysed 2 financial performance variables :

* **Cash flow** of the last 5 years
* **Daily volume** exchanged for the stock : which estimates the illiquidity

I also analyzed 3 **macroeconomic variables**:

* **Inflation**
* **Volatility** : represented by the VIX index
* **EPU** : Economic policy uncertainty

I also added 1 technical data **Moving Average Convergence Divergence (MACD) and EMA.**

**MACD** is a trend and momentum indicator that calculates the change and the speed of the evolution of the asset price. It was developed by Gerald Appeal in the late 1970s.

The **EMA** is the moving average that gives more weight to the most recent price points. This allows this type of moving average to react more strongly to recent price changes.

Based on prior empirical studies an economic theories, my hypothesis were :

H1 : Common stock returns is inversely related to firms size

H2 : Book to market ratio has a positive impact on stock returns

H3 : Price earnings ratio has a negative impact on stock returns

H4 : Stocks with low multiples P/B, P/C, P/D, P/E (value stocks) have higher returns than stocks with high multiples P/B, P/C, P/D, P/E (growth stocks).

H5 : Expected returns were positively correlated with illiquidity

H6 : Cash flow has an impact on stock returns.

H7 : Volatility indicators : VIX and EPU have an impact on stock returns

I tried to answer my research question in the following 5 chapters:

Chapter 1 ‘Literature review’ : presents the historical research on this subject based on academic resources.

Chapter 2 ‘Methodology’ : presents the hypothesis, the data, the variables and the methods.

Chapter 3 ‘Results’ : describes my main findings.

Chapter 4 ‘Analysis / Discussion / interpretation’ : explains the results and the limitations of the solution.

Chapter 5 ‘Conclusion’ : summarizes the project and concludes.

1. **Firm characteristics** = **firm size**, leverage, sales growth, asset growth and turnover, age of the firm, dividend pay-out, profitability, access to capital markets and growth opportunities.
2. **Financial performance** = financial performance is usually measured using financial ratios, the categories of ratios include: **liquidity**, activity, profitability, debt or solvency, **cash flow**
3. **Macroeconomic variables** = related to a country : **inflation**, gross domestic product, growth rate, **volatility**, ...

# Literature Review

Many researchers have analyzed the impact of firm characteristics (1), financial performance (2) and macroeconomic variables (3) on stock returns.

Regarding the firm characteristics, the size effect has been well analyzed.

Gordon (2010) found that firms size has a negative impact on common stock returns.

Rutledge et al (2008) studied this effect on Chinese stock market and demonstrated that small firms have higher returns.

Drew et al (2003) indicated that small firms with high growth have higher returns than bigger companies.

Other studies demonstrated the same result : Davis & Desai (1998); Rouwenhorst (1998); Fama & French (1992); Banz (1981) and Ringanom (1981); ...

In the field of financial performance, the ratios have been well analyzed.

The book to market ratio and earnings to price ratio are well known to be able to predict stock returns.

Many studies indicated that high book to market ratio results to high stock returns : Hoang et al (2015); Drew and Veeraraghavan (2003); Lam (2002); Ashiq & Hwang (2002); Rouwenhorst (1998); Fama & French (1996); Maroney (1995); Rosenberg et al (1985).

Other ratios (multiple) have also a significant impact on stock returns : P/B (4), P/C (5), P/D (6), P/E (7).

In several studies the conclusion was that stocks with low multiples (value stocks) had higher returns than stocks with high multiples (growth stocks) :

Fama & French (1992a, 1993, 1996, 1998); Chan, Hamaoa & Lakonishok (1991); Reid & Lanstein (1985); Stattman (1980) and Rosenberg .

In the same field of financial performance, the illiquidity effect has also been demonstrated.

For many researchers the conclusion was that expected returns were positively correlated with illiquidity.

Amihud (2002); Alaraini and Stephens (1999); Brennan and Subrahmanyam (1996); Amihud and Mendelson (1986)

Fiinally the macroeconomic variables like inflation and volatility have also been highligted as good predictors for stock returns.

A famous volatility indicator VIX which is sometimes called the 'Fear factor' has been introduced by Brenner et al (1989).

Another index related to volatility is EPU (economic policy uncertainty) created by Baker et al (2012).

Several studies were conducted to analyze the relationship between VIX or EPU and stock returns.

On VIX : Giot (2002); Brenner et al (1989); Chen et al (1986);

On EPU : Gao et al. 2019; Brogaard and Detzel (2015); Antonakakis et al. (2013)

(1) **Firm characteristics** = **firm size**, leverage, sales growth, asset growth and turnover, age of the firm, dividend pay-out, profitability, access to capital markets and growth opportunities.

(2) **Financial performance** = financial performance is usually measured using financial ratios, the categories of ratios include: **liquidity**, activity, profitability, debt or solvency, **cash flow**

(3) **Macroeconomic variables** = related to a country : **inflation**, gross domestic product, growth rate, **volatility**, ...

(4) **P/B** = Price to book value per share

(5) **P/C** = Price-to-cash flow

(6) **P/D** = Price-to-dividend

(7) **P/E** = Price to earnings per share.

# Methodology

## What is the question/goal ?

Define a model which can predict a local stock market behavior and then define an efficient trading strategie.

It’s a supervised learning problem.

The performance should be higher than random prediction : more than 50%

«Do the firm characteristics, financial performance and macroeconomic variables can predict future returns of stock exchange ?»

## Hypotheses of this research

Based on the concepts and findings of previous studies, the hypotheses of this research are as follows:

H1 : Common stock returns is inversely related to firms size

H2 : Book to market ratio has a positive impact on stock returns

H3 : Price earnings ratio has a negative impact on stock returns

H4 : Stocks with low multiples P/B, P/C, P/D, P/E (value stocks) have higher returns than stocks with high multiples P/B, P/C, P/D, P/E (growth stocks).

H5 : Expected returns were positively correlated with illiquidity

H6 : Cash flow has an impact on stock returns.

H7 : Volatility indicators : VIX and EPU have an impact on stock returns

## Existing solution

**I noted them for information, the aim is for me to train my new skills on ML not to do a benchmark !**

Some products (**SentimentTrader, SentiTrade**) offer their services for a time-limited subscription fee, other websites (**MarketWatch, DataMinr**) allow free registration but their approach still remains a secret.

A little more transparent solution is the online sentiment analysis tool **Sentdex.com** maintained by Harrison Kinsley who also runs a YouTube channel providing tutorials on data analysis.

Although the Sentdex product has not been described by any academic paper, the tutorial videos and the product homepage provide basic information about the used data and processes.

The algorithm running behind is written using Python’s Natural Language Toolkit and crawls input data from over 20 of the most famous American journals (Reuters, Bloomberg, WSJ, etc.) [Kin15].

[Kin15] Harrison Kinsley. Sentiment Analysis. 2015.

## Methods

### Overview

I’ll go through the KDD process and follow the necessary steps until I find a useful and understandable pattern which makes sense to answer to the question.

### Developing an understanding of the application domain

Done previously by reading and analysing documentation (about 80 hours). See references.

## Data

### Dataset

Our dataset consists of all available stocks listed on below Stock Exchanges in the period

2017-2021.

|  |  |  |
| --- | --- | --- |
| **Market** | **MIC** | **Market place** |
| Australian Securities Exchange | XASX | Sydney |
| Istanbul Stock Exchange | BIST | Istanbul |
| Brazil Stock Market | BS | São Paulo |
| Chile Stock Market | BVS | Santiago |
| European Stock Exchange | XAMS | Amsterdam |
| European Stock Exchange | XBRU | Brussels |
| European Stock Exchange | XMSM | Dublin |
| European Stock Exchange | XLIS | Lisbon |
| European Stock Exchange | XOSL | Oslo |
| European Stock Exchange | XPAR | Paris |
| Frankfurt Stock Exchange | FWB | Frankfurt |
| Warsaw Stock Exchange | WSE | Warsaw |
| Indonesia Stock Exchange | IDX | Jakarta |
| Johannesburg Stock Exchange | XJSE | Johannesburg |
| London Stock Exchange | XLON | London |
| Nasdaq Stock Market | XNAS | New York City |
| New-York Stock Exchange | XNYS | New York City |
| NYSE American | AMEX | New York City |
| Stockholm Stock Exchange | XSTO | Stockholm |
| Helsinki Stock Exchange | XHEL | Helsinki |
| Russian Trading System | RTS | Moscow |
| SIX Swiss Exchange | XSWX | Zurich |
| Shanghai Stock Exchange | XSHG | Shanghai |
| Shenzhen Stock Exchange | XSHE | Shenzhen |
| Tel-Aviv Stock Exchange | TASE | Tel-Aviv |
| Saudi Stock Exchange | XSAU | Riyadh |
| Toronto Stock Exchange | XTSE | Toronto |

Data were obtained from Yahoo Finance for historical of price and firm characteristics, financial performance data.

The macroeconomic data ‘Inflation’ and ‘Volatility’ were obtained from the below Web site :

<https://fred.stlouisfed.org/>

## Variables

What I want to determine at the end it’s to know if I have to buy, sell or keep a stock.

So my dependent variable needs to be categorical.

The **dependent variable** is based on the close price. It’s a **trend**. Therefore, I’ll use classification methods.

I calculate the percentage of increase or decrease of the close price of the day versus the close price of the day – 1.

If this percentage is greater or equal than 0 then the **trend** is equal to 1.

Otherwise it’s equal to -1.

if ((Close price of the day - close price of the day-1)/close price of the day-1) >= 0

trend = 1

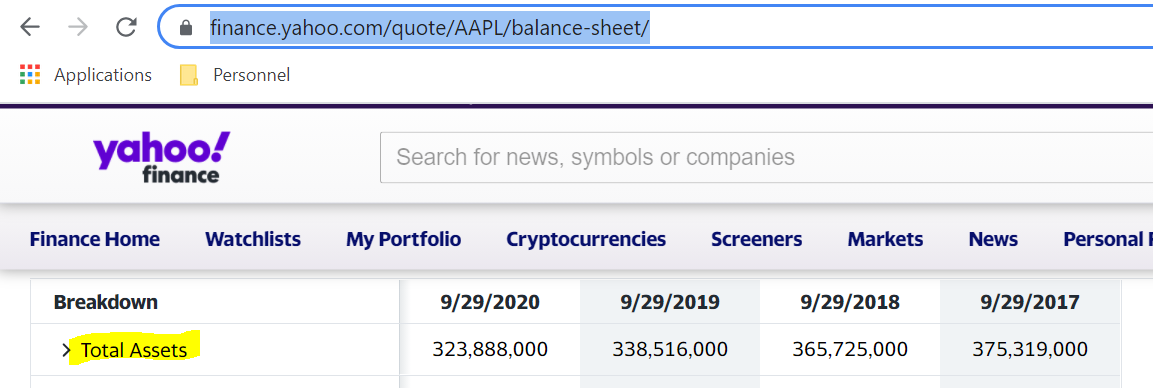
elif

trend = -1

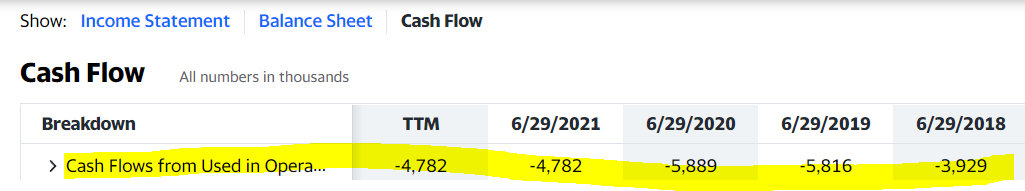
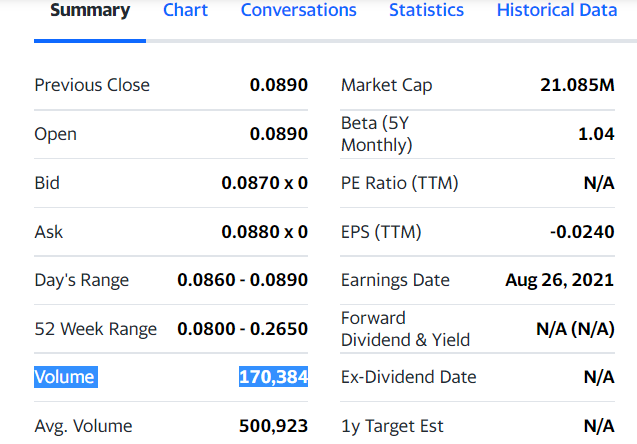
Below are the **independent variables**:

### Firm characteristics data

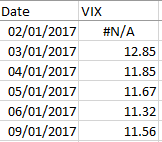
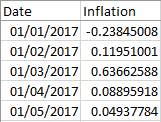
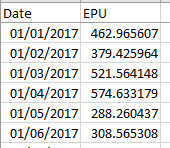
The data has been extracted from Yahoo Finance. Only date available on this website has been taken into account.

* **Firm size :** it can be estimated from the **total assets** of the firm.
  + Total assets is available in Yahoo here :
  + <https://finance.yahoo.com/quote/AAPL/balance-sheet/>
  + 

### Financial performance data

* **Cash flow** : a simple definition of a cash flow statement is how money, that is cash and cash equivalents, enters and exits a company
  + See Financials – Cash flow / In Yahoo Financials get field « totalCashFromOperatingActivities »
  + 
* **Illiquidity** **:** it can be estimated with the daily **volume** of the stock.
  + aapl.get\_price\_data()
  + 
* **Close price :** Can be check on Yahoo finance on ‘Historical data’.

### Macroeconomic variables data

* **Volatility :** how fast prices change, is often seen as a way to gauge market sentiment, and in particular the degree of fear among market participants.
  + VIX is the [ticker symbol](https://en.wikipedia.org/wiki/Ticker_symbol" \o "Ticker symbol) and the popular name for the [Chicago Board Options Exchange](https://en.wikipedia.org/wiki/Chicago_Board_Options_Exchange" \o "Chicago Board Options Exchange)'s CBOE Volatility Index, it’s the most famous volatility index
  + VIX, is a real-time market index representing the **market's expectations** for volatility **over the coming 30 days**.
  + There is a VIX for specific stocks market, for emerging markets, VIX based on gold, silver, ... : VIX CAC40, …
  + It’s a **daily** data
  + The data were obtained from : <https://fred.stlouisfed.org/>
  + 
* **Inflation :** it’s the loss of the purchasing power of the currency which results in a general and lasting increase in price
  + It’s calculated **monthly** by **country**
  + The data were obtained from : <https://fred.stlouisfed.org/>
  + 
* **EPU** : is the Economic policy uncertainty
  + The data were obtained from : <http://www.policyuncertainty.com/>
  + It’s calculated **monthly** by **country**
  + 

### Technical variable

**Moving Average Convergence Divergence (MACD) and EMA.**

MACD is a trend and momentum indicator that calculates the change and the speed of the evolution of the asset price. It was developed by Gerald Appeal in the late 1970s.

I calculated it from the price historic.

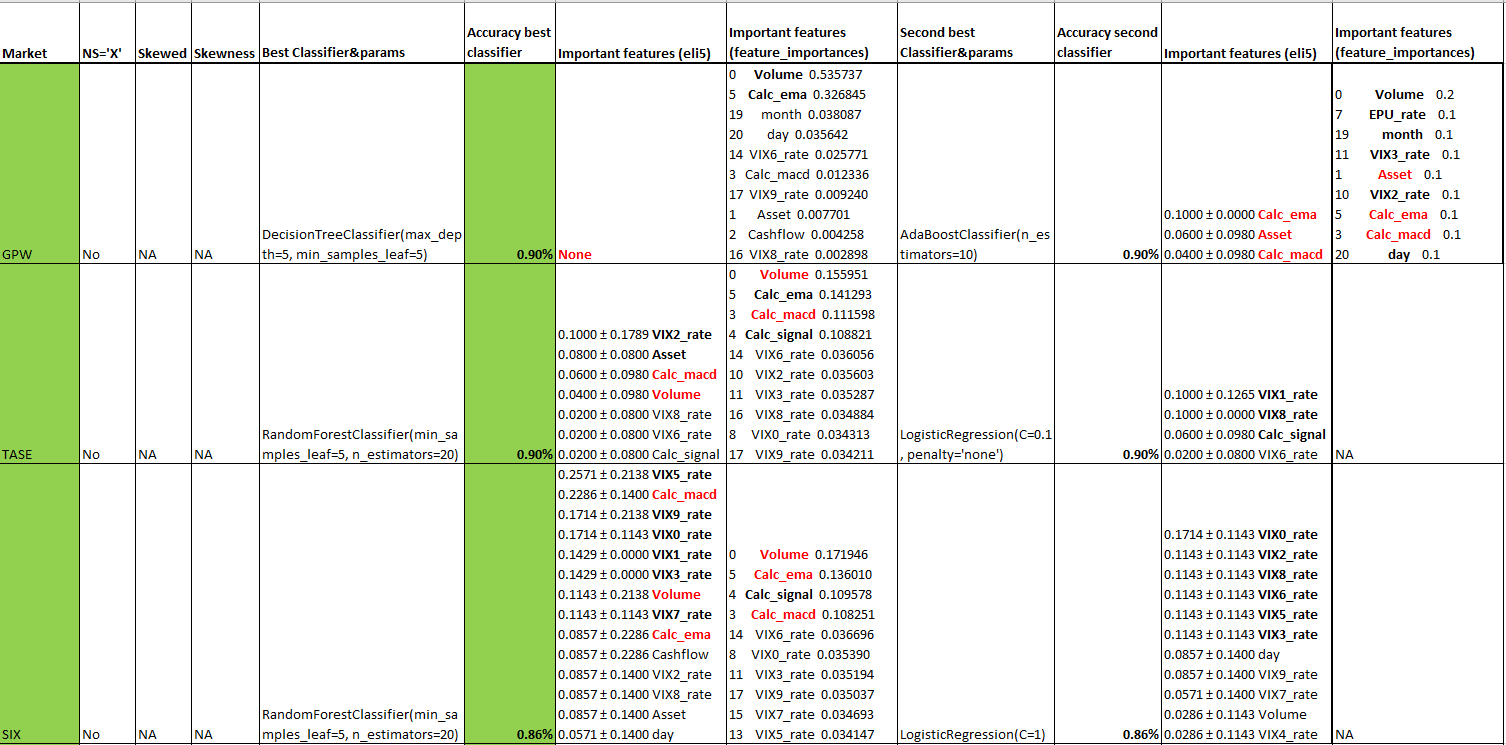
The **EMA** is the moving average that gives more weight to the most recent price points. This allows this type of moving average to react more strongly to recent price changes.

# Results

These are the complete results :



This is an extract :



We can see the accuracy of the 2 best classifiers by market, with details on which classifier and the important variables.

# Analysis / Discussion / interpretation

## Analysis

I tested 23 markets.

We can disctinct 3 groups :

* 9 markets over 23 had an accuracy superior to 0.78%
* 7 markets over 23 had an accuracy equal to 0.70%
* 7 markets over 23 had an accuracy between 0.50% and 0.60%

For the first group with the highest accuracy, classifiers appear as the best classifier on several markets:

* RF 4 times
* DT 3 times
* LR 2 times

For the first group with the highest accuracy, classifiers appear as the second best classifier on several markets:

* Adaboost 5 times
* LR 4 times
* DT 2 times
* Bernouilli 2 times

The important features for RF are in this order :

* Volume 4 times
* Calc\_macd 3 times
* Calc\_ema 3 times

The important features for DT are in this order :

* Calc\_ema 4 times
* Volume 3 times

The important features for Adaboost are in this order :

* Calc\_ema 3 times

For the second group with the accuracy equal to 0.70%, classifiers appear as the best classifier on several markets:

* LR 4 times

For the second group with the accuracy equal to 0.70%, classifiers appear as the second best classifier on several markets:

* Bernouilli 5 times
* Adaboost 2 times
* RF 2 times

The important features for RF are in this order :

* Calc\_ema 1 time

The important features for DT are in this order :

* Calc\_ema 2 times
* Volume 2 times

The important features for Adaboost are in this order :

* Calc\_ema 1 time
* Volume 1 time

**Based on these results, we can say that the best classifiers are LR, DT, Adaboost and RF.**

**The important features for all these classifiers are calc\_ema, volume, calc\_macd.**

Thes results confirm the fith hypothesis «Expected returns were positively correlated with illiquidity».

Unfortunately I was not able to confirm the others hypotheses :

H1 : Common stock returns is inversely related to firms size

H2 : Book to market ratio has a positive impact on stock returns

H3 : Price earnings ratio has a negative impact on stock returns

H4 : Stocks with low multiples P/B, P/C, P/D, P/E (value stocks) have higher returns than stocks with high multiples P/B, P/C, P/D, P/E (growth stocks).

H6 : Cash flow has an impact on stock returns.

H7 : Volatility indicators : VIX and EPU have an impact on stock returns

I can partially reply to the research question :

«Do the firm characteristics, financial performance and macroeconomic variables can predict future returns of stock exchange ?»

Financial performance data (Volume) and technical indicators (calc\_ema, calc\_macd) based on close price, have a significant impact on the future prices prediction.

## Discussion

To validate these results it could be valuable to analyse the same markets with **more recent data**.

Also to extend the predictions to others fields of research already highlighted by many academic resources it could be good to investigate the impact on the below subjects on the future stock price:

* **Calendar**
* **Twitter**
* **Internet search**
* **Google trend**

## Acting on the discovered knowledge

If we want to use the discovered knowledge we have to adapt the program to extract day to day data and to focus only on variables which have an explanation power.

# Conclusion

The main objective of this study was to define a model which can predict a local stock market behavior and then define an efficient trading strategie.

The performance should be higher than random prediction : more than 50%

The model trained has an accuracy superior to 50% on all of the 23 markets and 9 of them have an accuracy superior to 78%.

The goal is achieved.

# References

## Firm characteristics and financial performance data

* An investigation of pricing multiples’ ability to predict abnormal returns on the Oslo Stock Exchange
* Defining and Designing a Model to Predict the Performance of Mutual Funds by Using Macroeconomic Variables in Tehran Stock
* Does the fear gauge predict downside risk more accurately than econometric models Evidence from the US stock market
* Stocks as Lotteries Can Extreme Positive Returns Predict Future Returns
  + firm size, cash flows, illiquidity
* The Ability Of Earnings, Cash Flow To Predict Future Earnings, Cash Flow And Stock Price

## Macroeconomic variables data

* Can uncertainty predict stock market returns
  + volatility, inflation
* Do Based-Market Data Predict Stock Return Better Than Accounting Data The Case of Tehran

# Appendix

## Glossary

**Return** is the percentage change in the asset value. I used the ~~adjusted~~ close prices to calculate returns.

**~~Adjusted close price~~** ~~is the official closing price adjusted for capital actions and dividends.~~

~~Nevertheless returns provide useful information about the probability distribution of asset prices.~~

~~This is essential for investors and portfolio managers as they use this information to value assets and manage their risk exposure.~~

~~The one of models that is used to explain the stock returns is the Capital Asset Pricing Model (CAPM).~~

~~Capital Asset Pricing Model definite the returns of stock as function of the systematic risk of a stock.~~

**Cash flow** : a simple definition of a cash flow statement is how money, that is cash and cash equivalents, enters and exits a company.

**Volatility** is a statistical measure of the dispersion of returns for a given security or market index.

In most cases, the higher the volatility, the riskier the security.

Volatility is often measured as either the standard deviation or variance between returns from that same security or market index.

In the securities markets, volatility is often associated with big swings in either direction.

For example, when the stock market rises and falls more than one percent over a sustained period of time, it is called a "volatile" market. An asset's volatility is a key factor when pricing options contracts.

**Trading volume** is the number of shares traded in each day during a trading session. Volume can be used to measure stock liquidity, which in turn has been shown to be useful in asset pricing as several theoretical and empirical studies have identified a liquidity premium.

Liquidity can help to explain the cross-section of expected returns.

**Accounting data** (Firm Size, Return on Equity, Return on Assets, profit margin ratio, Financial Leverage ratio).

**Based-market data** (Price to Earnings ratio, book to market ratio and Dividend yield).

**Size** : The natural log of total asset (actifs) at the end of the year

**Book to Market value (BM)** : Book value of stock over stock price at the end of the year.

Valeur comptable du stock par rapport au prix du stock à la fin de l’année

**Book-to-market ratio** : The book-to-market ratio compares a company's book value to its market value. The book value is the value of assets minus the value of the liabilities. The market value of a company is the market price of one of its shares multiplied by the number of shares outstanding.

**Return on Asset (ROA) :** Earning after tax over total asset at the end of the year

**Return on Equity (ROE)** : Earning after tax over equity at the end of the year.

The **coefficient of return on equity** reflects how many turns it takes to pay bills for the date of analysis.

**Margin Profit (MP)** : Earning after tax over Sales.

**Price to earnings ratio (PE)** : Stock price over earnings per share.

**Financial Leverage (LEV) :** Total debt over total asset.

The **financial leverage** is the amount of debt a firm uses to finance assets.

**Dividend Yield (DY)** : Dividend per share over stock price.

**Abnormal Return:** A term used to describe the returns generated by a given security or portfolio over a period of time that is different from the expected rate of return. The expected rate of return is the estimated return based on an asset pricing model(…).

**Earnings for price ratio (EP)** : A measure indicating the rate at which investors will capitalize a firm's expected earnings in the coming period. This ratio is calculated by dividing the projected earnings per share by the current market price of the stock. A relatively low E/P ratio anticipates higher-than-average growth in earnings. Earnings-price ratio is the inverse of the price-earnings ratio. Also called earnings capitalization rate, earnings yield.

The **idiosyncratic volatility** is the difference between total risk and the systematic risk of a stock,

I define idiosyncratic volatility as the standard deviation of the regression residual of the Fama and French three-factor model.

The **Short-Term Reversal** Effect : The short-term reversal anomaly, the phenomenon that stocks with relatively low returns over the past month or week earn positive abnormal returns in the following month or week, and stocks with high returns earn negative abnormal returns, is well-researched.

**Illiquidity** : illiquidity refers to assets that cannot be easily exchanged for money.

This may be due to the fact that there are not enough investors willing to buy them.

**MACD Moving Average Convergence Divergence :** MACD is a trend and momentum indicator that calculates the change and the speed of the evolution of the asset price. It was developed by Gerald Appeal in the late 1970s.

## Program documentation

### Data collection

Data collection is the ultimate first step, it is not part of the KDD process itself.

First I get a list of stock markets :



Here are the markets I retained :

|  |  |  |
| --- | --- | --- |
| **Market** | **MIC** | **Market place** |
| Australian Securities Exchange | XASX | Sydney |
| Istanbul Stock Exchange | BIST | Istanbul |
| Brazil Stock Market | BS | São Paulo |
| Chile Stock Market | BVS | Santiago |
| European Stock Exchange | XAMS | Amsterdam |
| European Stock Exchange | XBRU | Brussels |
| European Stock Exchange | XMSM | Dublin |
| European Stock Exchange | XLIS | Lisbon |
| European Stock Exchange | XOSL | Oslo |
| European Stock Exchange | XPAR | Paris |
| Frankfurt Stock Exchange | FWB | Frankfurt |
| Warsaw Stock Exchange | WSE | Warsaw |
| Indonesia Stock Exchange | IDX | Jakarta |
| Johannesburg Stock Exchange | XJSE | Johannesburg |
| London Stock Exchange | XLON | London |
| Nasdaq Stock Market | XNAS | New York City |
| New-York Stock Exchange | XNYS | New York City |
| NYSE American | AMEX | New York City |
| Stockholm Stock Exchange | XSTO | Stockholm |
| Helsinki Stock Exchange | XHEL | Helsinki |
| Russian Trading System | RTS | Moscow |
| SIX Swiss Exchange | XSWX | Zurich |
| Shanghai Stock Exchange | XSHG | Shanghai |
| Shenzhen Stock Exchange | XSHE | Shenzhen |
| Tel-Aviv Stock Exchange | TASE | Tel-Aviv |
| Saudi Stock Exchange | XSAU | Riyadh |
| Toronto Stock Exchange | XTSE | Toronto |

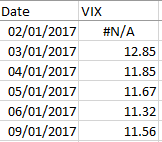
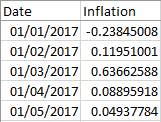
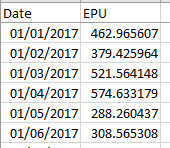
Second I found a list of tickers for each market by using different website listed in the below file.



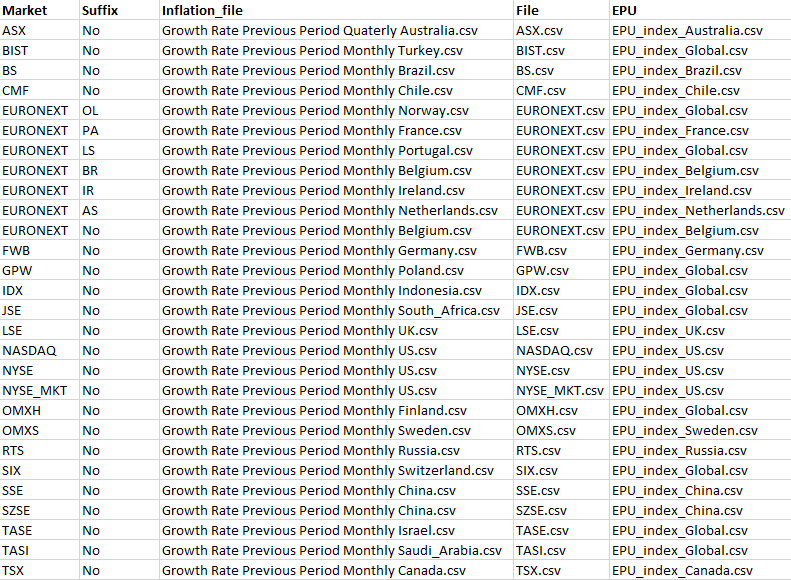
Here is an example of the ticker file by market :



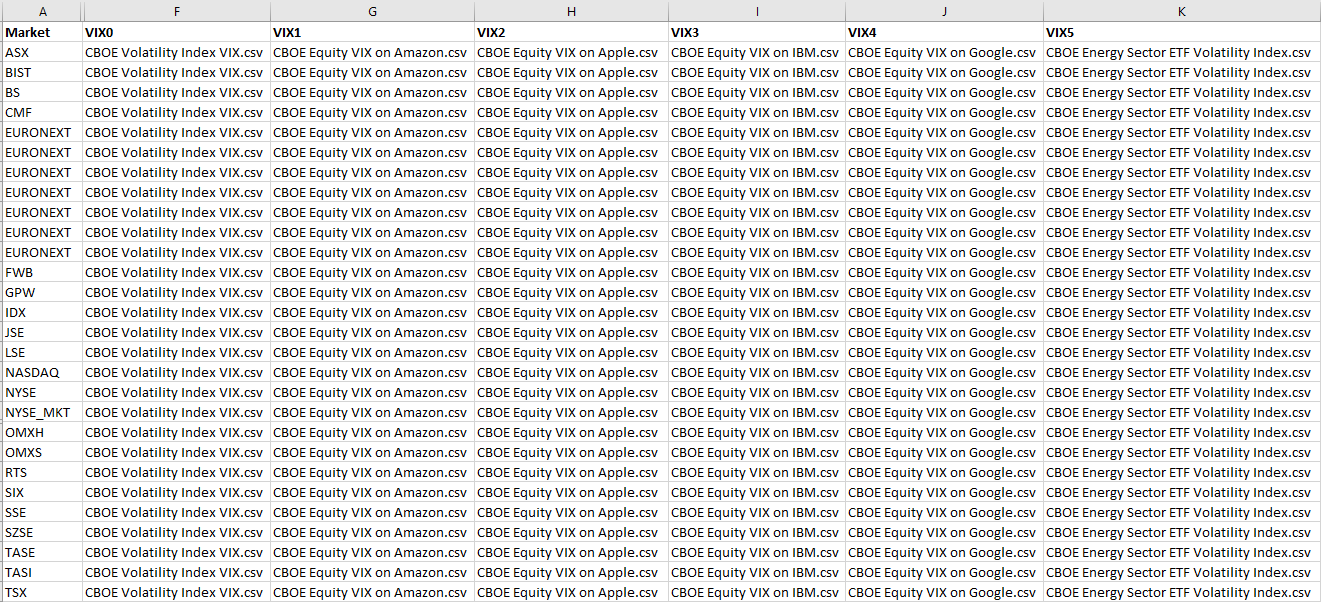
I get macroeconomic variables data from the below sources :

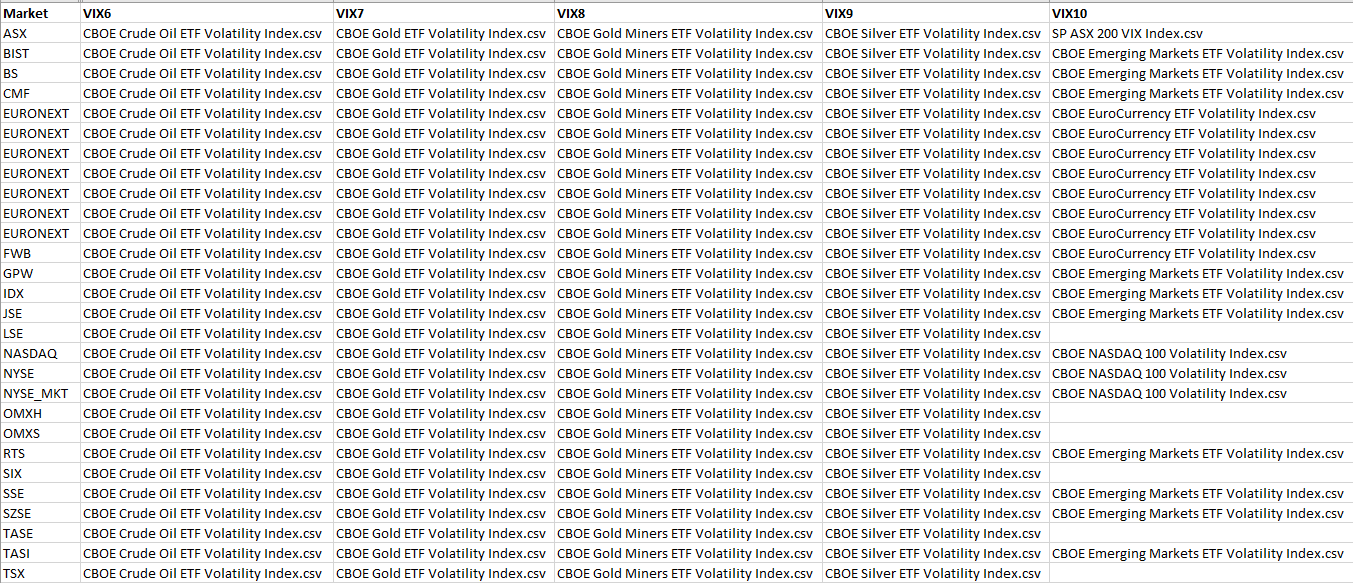
* **Volatility :** how fast prices change, is often seen as a way to gauge market sentiment, and in particular the degree of fear among market participants.
  + VIX is the [ticker symbol](https://en.wikipedia.org/wiki/Ticker_symbol" \o "Ticker symbol) and the popular name for the [Chicago Board Options Exchange](https://en.wikipedia.org/wiki/Chicago_Board_Options_Exchange" \o "Chicago Board Options Exchange)'s CBOE Volatility Index, it’s the most famous volatility index
  + VIX, is a real-time market index representing the **market's expectations** for volatility **over the coming 30 days**.
  + There is a VIX for specific stocks market, for emerging markets, VIX based on gold, silver, ... : VIX CAC40, …
  + It’s a **daily** data
  + The data were obtained from : <https://fred.stlouisfed.org/>
  + 
* **Inflation :** it’s the loss of the purchasing power of the currency which results in a general and lasting increase in price
  + It’s calculated **monthly** by **country**
  + The data were obtained from : <https://fred.stlouisfed.org/>
  + 
* **EPU** : is the Economic policy uncertainty
  + The data were obtained from : <http://www.policyuncertainty.com/>
  + It’s calculated **monthly** by **country**
  + 

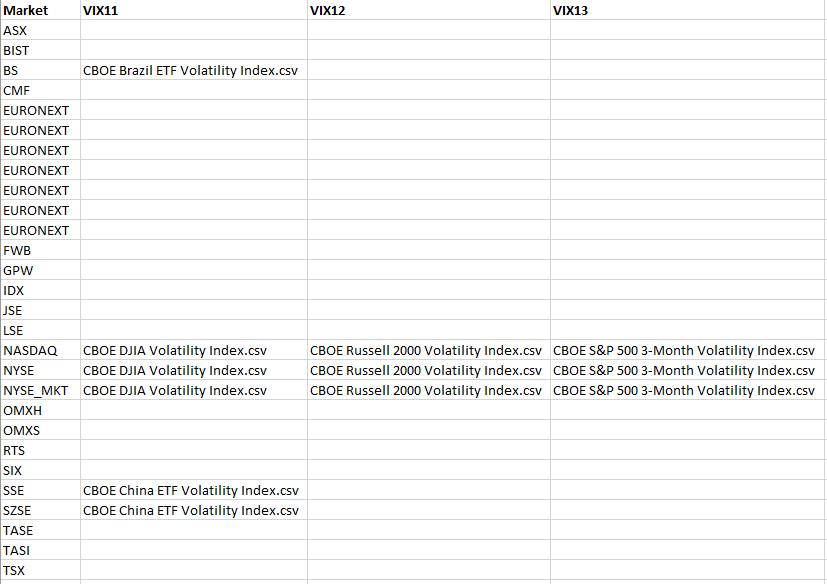
I specified for each market which inflation file and EPU file has to be used :



Same for VIX files, I specified several VIX file per market :







I used python to develop extractors to get financial performance data related to each tickers (total asset, cash flows, historical of prices, …) from Yahoo Finance using YahooFinancials library.

I was not able to use below data for a lack of history in Yahoo Finance :

* Book-to-Market Ratio (priceToBook)
* Price to earnings ratio
* Book-to-Market Ratio
* Price to earnings ratio
* Earnings-to-price ratio (E/P ratio)

Below is the code used to extract data :

Get\_stock\_data.py extracts data related to the stock (assets and cash flows of the last 4 years).



Get\_historical\_stock\_data.py extracts the historic of price and volume for each stock.



Python programs were launched from **scripts** located on a Linux server and scheduled through **crontab**.







I duplicated scripts in order to parallelize the load; each script loaded a specific market.

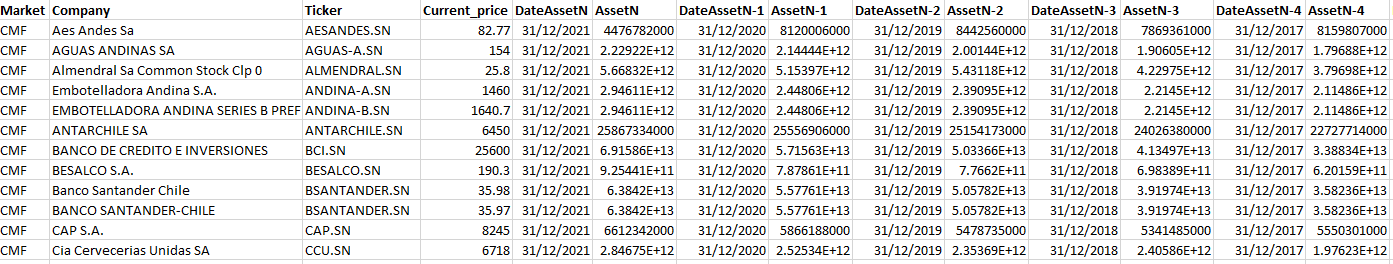
But it was generating some error due to heavy workload on the server.

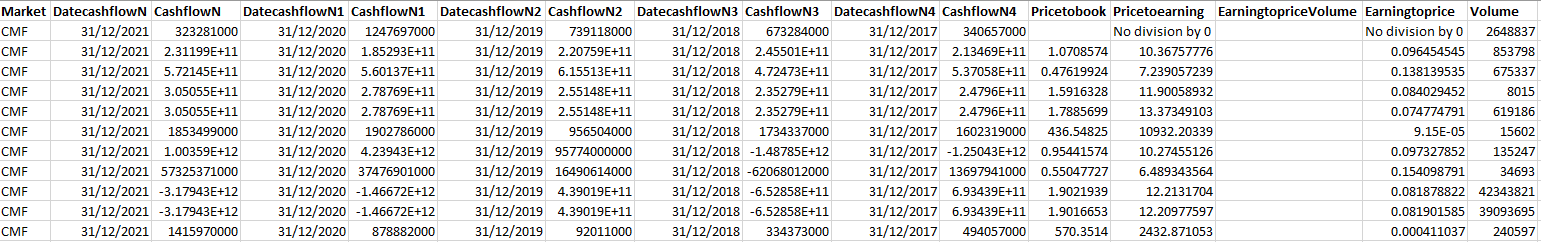
So it’s better to handle the load sequentially; 1 script for all markets.

I had also some errors due to http request which is not responding. I deal with this error by grepping the error in the python program.

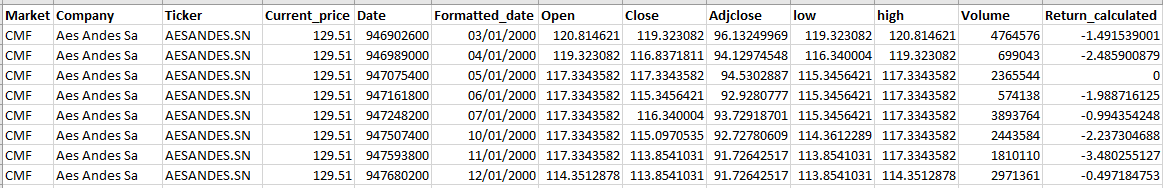
These are the data get from Yahoo Finance :

**Stock data :**





**Historical prices :**



### Preparation and prediction program

I created a program to prepare the data and predict the dependent variable.

Python program : **Stock\_market\_predictions.ipynb**

**Mettre en objet la dernière version du pgm**

#### Main program

##### Overview

The main program executes the below macro steps which will be detailed in the next parts.

The macro steps belong to the KDD process.

1. Firstly, I’ll create a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. Secondly, I’ll clean and preprocess the data.
3. Thirdly, I’ll try to find useful features to represent the data depending on the goal of the task
   1. This is data reduction and projection.
4. Fourthly, I’ll explore the data trying to find data patterns with different data-mining methods.

Data have been collected :

* Financial performance data : historical of price data for many tickers of different markets.
* Macroeconomic variables data : Volatility and Inflation data. For volatility I get EPU, (Volatility by country) and VIX indexes
* Technical data : I’ll calculate MACD and EMA from historical of price data

##### Macro steps

I loop on markets.csv file which list all markets and files related to each market.

1. **Preparation** step is then run, see ‘**Data preparation’** for details.
2. **Clean and preprocess** step is run, see ‘**Data cleaning and preprocessing**’ for details.
3. **Data reduction and projection** is then run, see ‘**Data reduction and projection’** for details.
4. **Data Mining,** I’ll explore the data trying to find data patterns with different data-mining methods, see ‘**Data Mining’** for details.

\* Variable V for Verbose will allow different level of details to be displayed. Level 1 is the more synthetic as level 3 is the more detailed.

### Data preparation

|  |  |
| --- | --- |
| **Task** | **Function(arguments)** |
| **First part : "Creating a target data set"** | |
| **Drop unuseful columns** | |
| Delete not useful columns :  **Dataset\_1** : Market, Company, Current\_price, Pricetobook, Pricetoearning, Earningtoprice, Volume  **Dataset\_2** : Market, Company, Current\_price, Formatted\_date, Open, Adjclose, low, high, Return\_calculated | drop\_unusefull\_columns(dataset\_1, dataset\_2, market)  dataset\_1 = stock data  dataset\_2 = historical prices/volumes |
| **Merge datasets** | |
| Merge stock data with historical prices/volumes based on ticker (stock) | merge\_datasets\_by\_ticker(cleanuped\_dataset1, cleanuped\_dataset2, ticker) |
| **Cleanup merged datasets** | |
| **1/ Drop rows when asset or cash flows were not found**.  **2/ Keep historical data related to asset and cash flows dates** (only the last 4 years were available for asset and cash flows so I keep only historical prices of the last 4 years).  **3/ Calculation of the return which will be used later to calculate the dependent variable (trend)**  I calculate the **percentage** of increase or decrease of the close price of the day versus the close price of the day – 1.  **‘Calc\_return’** : «close price of the day» - «close price of the day-1» / «close price of the day-1» | cleanup\_merged\_dataset(merged\_set) |
| **MACD, signal and EMA calculation** | |
| MACD = Price average at 12 days - Price average at 26 days  Signal = MACD average at 9 days  EMA = MACD - signal | MACD\_calculated\_dataset(merged\_set, market) |
| **Adding macroeconomic variables data (VIX, inflation, EPU)** | |
| Adding macroeconomic variables data for side files specified in the markets.csv file. | Infos\_added\_set(ticker\_with\_MACD\_set, market, suffix) |

### Data Cleaning and preprocessing

|  |  |
| --- | --- |
| **Task** | **Function(arguments)** |
| **Second part : "Cleaning the dataset"** | |
| **Split dataset into training and test** | |
| Split dataset into chronological sets (training and test), a variable horizon will define the test size. | split\_dataset(dataset) |
| **Replace missing data** | |
| I will use the interpolate method to fill the missing inflation rate, volume, VIX rate.  I do this for **training and test set** **separately**. | repl\_missing\_data(dataset, settype) |
| **Identify and remove noise (error and residuals = outliers)** | |
| Convert alphanumeric into numeric variables | convert\_to\_numeric(dataset) |
| Identify outliers using LocalOutlierFactor.  Add a column called 'outlier' to the dataset containing a -1/1 flag for outliers | identify\_outliers(dataset, no\_neigh, contam, typ) |
| Now remove residuals = outliers  Apply separately on train and test dataset. | remove\_outliers(dataset) |
| **Check and address skewness** | |
| Check skewness with kurtosis. If the value is not between -0.5 and 0.5 then I address skewness with RandomOverSampler. | Functions :  check\_and\_address\_skewness(dataset, typ)  check\_skewness(dataset, typ)  address\_skewness(dataset, typ) |

### Data reduction and projection

|  |  |
| --- | --- |
| **Task** | **Function(arguments)** |
| **Third part : "Data reduction and projection"** | |
| **Normalize or standardize variables** | |
| Normalize dataset using MinMaxScaler.  Standardize dataset using StandardScaler. | norm\_std\_variables(dataset) |
| **Check correlation** | |
| Check correlation between independent variables and dependent one ==> **line plot on all variables** | correlated\_variables(dataset) |
| Check important variables according to Mutual information. Chi-squared, | return\_variables\_to\_be\_retained\_or\_removed(dataset) |
| Display a **heatmap** for all variables. | find\_correlation(X) |
| Display scree-plots, loadings on principal components | find\_pca\_variables(X) |

### Data mining

I fit models on validation set with all the data related to a market. All stocks of this market will be used for fiitting.

Then I predict on the test set for each stock independently.

I keep building models until I find a suitable one that **works with the test set**.

|  |  |
| --- | --- |
| **Task** | **Function(arguments)** |
| **Fourth part : "Data mining"** | |
| First I add a **trend** variable based on the return calculated.  It will be our **dependent variable**.  If the return calculated is greater or equal than 0 then the **trend** is equal to 1.  Otherwise it’s equal to -1.  As the dependent variable is a trend which can take 2 values, I’ll use **classification methods**. | |
|  |  |
| Fit and predict with **Simple Linear Regression and Multiple Linear Regression** | Functions :   * evaluation\_process1(dataset, variable) * evaluation\_process2(dataset\_train, dataset\_test, variable) * calculate\_vif(variables) * evaluation\_process3(dataset\_train)   **Accuracy** evaluated with **R-squared**. |
| Fit and predict with **Clustering data** |  |
| Fit and predict with **Decision trees** | **Accuracy** evaluated with **confusion matrix** : function calculate\_cm(predicted, actual) |
| Fit and predict with **Random Forest and AdaBoost model** | **Accuracy** evaluated with **confusion** matrix : function calculate\_cm(predicted, actual) |
| Fit and predict by **comparing different tree-based models** | **Accuracy** evaluated with **confusion** matrix : function calculate\_cm(predicted, actual) |
| Fit and predict with **support vector machines** | **Accuracy** evaluated with confusion matrix : function calculate\_cm(predicted, actual) |
| Fit and predict with **support vector regression model** | **Accuracy** evaluated with **confusion matrix** : function calculate\_cm(predicted, actual) |
| Fit and predict with **MLP in Keras** | **Accuracy** evaluated with **R-squared**. |
| Fit and predict with **neural networks** | **Accuracy** evaluated with **confusion matrix** : function calculate\_cm(predicted, actual) |
| Fit and predict with **ANN Hyperparameter exploration** |  |
| Fit and predict with **Microsoft Light Gradient Boosting Machine model** | **Accuracy** evaluated with **confusion matrix** : function calculate\_cm(predicted, actual) |
| Fit and predict with **Logistic Regression** | **Accuracy** evaluated with **R-squared**. |
| Fit and predict with **Naive Bayes** | **Accuracy** evaluated with **R-squared**. |