



Assignment Declaration

MIS41110: Programming for Analytics

Module Coordinator: Miguel Nicolau

Submission date:27/11/2017
Submission Zip: Group Project for Business Analytics

We confirm that:

- 1. We understand what plagiarism means and accept the plagiarism policy as set out in the Student Handbook.**
- 2. This assignment has been prepared entirely by the members of the group and steps have been taken to ensure that nobody else has been able to copy our work in any form.**
- 3. This assignment does not contain any material taken from unacknowledged sources and that all material has been referenced.**
- 4. We are the original authors of all the work presented in this assignment.**

The assignment has been prepared for assessment in this course and has not been presented as course work in any other course.

Group Members' Details

Group Member's Name	Student Number	Student Signature
Ang Li	17203382	
Shuhong Jiang	15202005	
Suohuijia Wang	17200170	

(Each student must retain a copy of his or her assignment). The UCD web site expands more fully on the nature and consequences of plagiarism and late submission of assignments, these policies must be read prior to submitting this declaration. https://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf and http://www.ucd.ie/t4cms/late_sub.pdf

I. Contribution

We generally discuss the main idea of the program together, then write each specific function separately. After that, each chunk has been reviewed in the group, and designed to generate into an executable loop. The entire program is done through times of face-to-face meetings. In the whole process, we reference lots of online python tutorials resources and also stock analysis tutorials.

Ang Li contributes on finding resources and functions online, plotting the K-line with moving average processing data, the report writing as well as working on some improvements of the project.

Shuhong Jiang contributes on the implementation of data reading from online, plotting the K-line with expected weighted moving average processing data, the UML flow and some review and improvement works.

Suohuijia Wang's contributes on common functions which is in utils.py, such as write company list and read existed data. Write get data function to provide a specified period of dataset and descriptive analysis for users, plot single Kline, and write the final report.

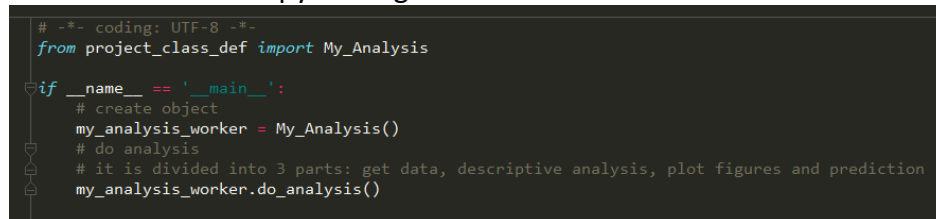
As for prediction part, we discussed together and search online to get algorithms and watch some videos to learn about them. Then, we finished it together.

In the end, we combine all our work together, and encapsulate them as a class for a better use for other people. It is a group work and we discuss every details together.

II. User-manual

This user-manual introduces users about how to run our code and the function of every parts. For convenience, we encapsulate all functions as a class and users can just run main function to do their analysis.

All users should do is to run main.py. see figure 1.



```
# -*- coding: UTF-8 -*-
from project_class_def import My_Analysis

if __name__ == '__main__':
    # create object
    my_analysis_worker = My_Analysis()
    # do analysis
    # it is divided into 3 parts: get data, descriptive analysis, plot figures and prediction
    my_analysis_worker.do_analysis()
```

Figure 1: main.py

1. Get data online

When you run main.py, this program will display existed companies, and it will ask you to enter a company symbol that you like and enter a start time and end time.

1.1 Get new data

If it is your first time to run this program, then there is no data in data folders and no companies in buffer. So you can enter a company you like and this program will download data for you, create a csv file and put this company into company list.

1.2 Get existed data

If you want to get data from company list, then you can enter one of them and this program will call `read_exist_data()` so that you don't have to download data again.

1.3 Get more data

If you want to add more data into existed data, this program can help you do this and you don't have to do anything.

2. Provide descriptive analysis

After getting data, this program goes to next stage, that is, descriptive analysis. We provide 10 choices for users, maximum, minimum, count, mean, variance, standard deviation, idmin, idmax and quantiles. Enter the number you like and it will display statistical data for users. As for quantiles, users should enter the specified percentage as a format of a float. Enter 10 to quit and go to the third stage.

3. Provide visualize analysis

Now, we can plot figures by enter some parameters. We provide two plot functions, K-line with EMA or MA. Users can enter 1 or 2 to choose which figure they want to plot. Choose 3 to quit.

3.1 plot K-line with EMA

Enter 1 to get a start. Firstly, we set a default as 1D to plot the figure, users choose Y to enter other frequency, such as 10D. Then figure is plotted. See figure 2.



Figure 2: K-line with EMA (10D)

3.2 plot K-line with MA

Enter 2 to get a start. Also you have to enter Y or N to set a frequency. Then K-line with MA is plotted. See figure 3.

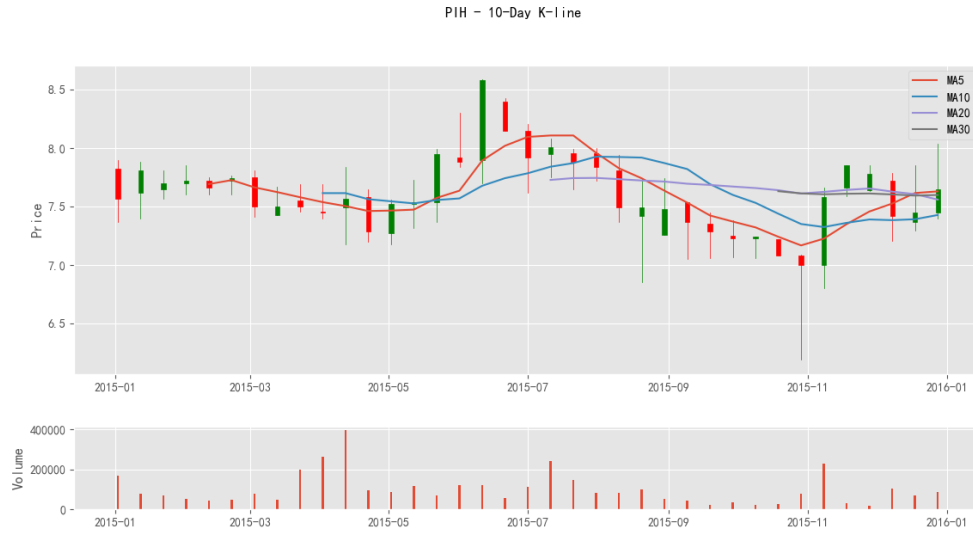


Figure 3: K-line with MA (10D)

4. Model training and prediction

When you enter 3 to quit visualize analysis system, you will enter the next stage, model training. Firstly, like the third stage, you should choose whether to resample to predict a lower frequency data. And then enter the period of training data, start time and end time. Note that end time of training data should be less than original dataset, since if not, we cannot have test data to predict.

We provide two models for users, linear or not. Choose Y/N according to what you like. Figure 4 is a linear model prediction for PIH and we can see that the prediction is not satisfactory. Figure 5 is a non-linear model.

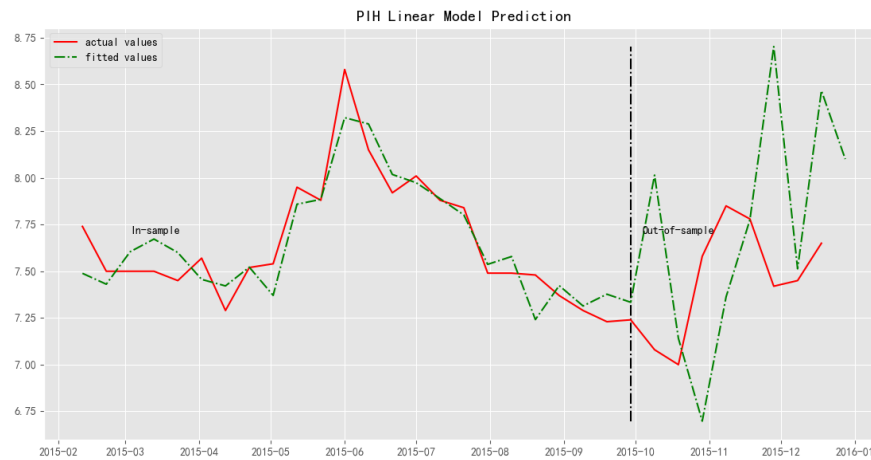


Figure 4: linear model prediction for PIH

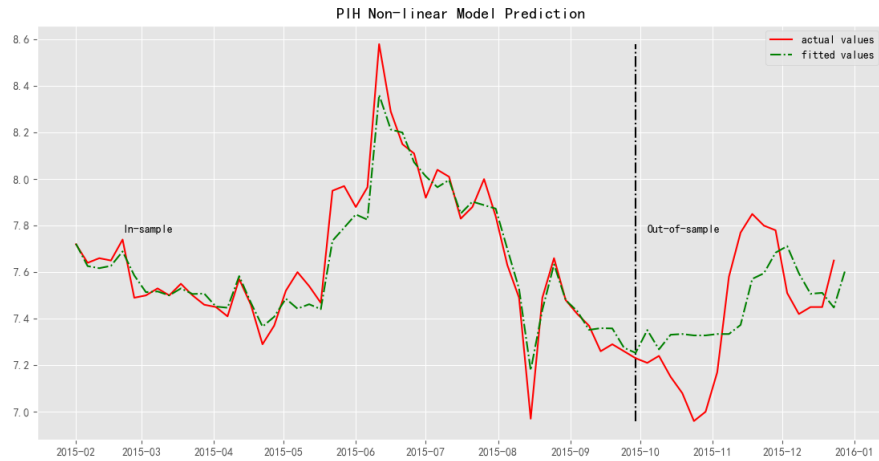


Figure 5: non-linear model prediction for PIH

5. Model prediction for given days

In this stage, users can choose given days to predict, and this program will return predicted value and true value. If the date is not in dataset, then return predicted value and the true value is none. Users can have 5 times to make mistakes.

III. UML Activity Diagram

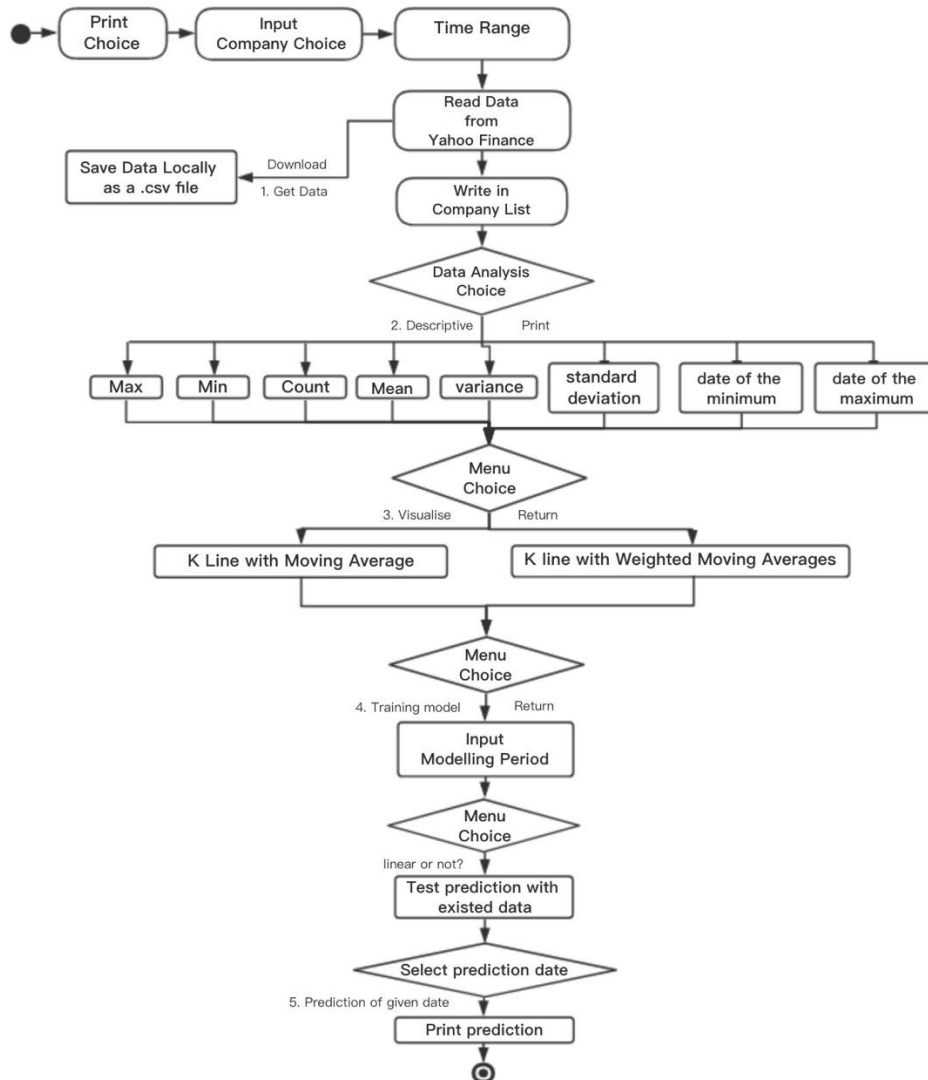


Figure 6: UML Activity Diagram

IV. Improvements

There are several shortages that can be improved in future, and we also use a class in order to let those advanced programmers to add more functions which we could not do currently.

- Plot more figures
Programmers can define other functions of plotting figures in base class and then inherit it in sub class to run it.
- Training date
It should be noted that when users call predict function and enter the period of training data, the end date should not be the last day of the original dataset, since the test data could be zero and return nothing to users.
- Getting data
Actually it might be an inevitable problem. That is, when we get data from yahoo, sometimes we cannot connect the source and fail to download data, but we succeed in the end. We don't know the reason but it should not be a technical problem.