TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

Dự báo sự tăng hoặc giảm giá cổ phiếu lúc đóng phiên vào hôm sau của Amazon (2015-2022)

HỌ VÀ TÊN SINH VIÊN	LỚP HỌC PHẦN	ĐIỂM BẢO VỆ
Nguyễn Khoa Hoàng	19N10	
Lê Mạnh Duy	19N10	
Hoàng Minh Đức	19N10	

ĐÀ NẴNG, 06/2022

TÓM TẮT

Hiện nay, chúng ta theo dõi cổ phiếu một cách thường xuyên hơn là điều dễ hiểu trong thời đại công nghệ phát triển. Chúng ta quan tâm nhiều hơn đến các vấn đề liên quan đến tài chính và kinh tế.

Với chương trình Dự đoán sự tăng hoặc giảm của giá cổ phiếu lúc đóng phiên vào hôm sau của Amazon sẽ giúp chúng ta biết được có nên xuống tiền đầu tư vào lô cổ phiếu đó hay không.

Để xây dựng được chương trình đó chúng ta cần thu thập dữ liệu từ trang web Yahoo Finance và tiến hành xử lý dữ liệu đó để cho ra được kết quả mà mình mong muốn.

Cuối cùng xin chân thành cảm ơn thầy Ninh Khánh Duy đã tạo cơ hội để giúp chúng em có thể tiếp xúc với những kiến thức mới và hoàn thành bài tiểu luận này.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Kết quả
Hoàng Minh Đức	- Thu thập và mô tả dữ liệu	Hoàn thành
Lê Mạnh Duy	- Trích xuất đặc trưng	Hoàn thành
Nguyễn Khoa Hoàng	- Mô hình hoá dữ liệu	Hoàn thành

MỤC LỤC

MỤC LỤC HÌNH ẢNH

1. Giới thiệu

Một chương trình Dự đoán sự tăng hoặc giảm của giá cổ phiếu lúc đóng phiên vào hôm sau của Amazon sẽ giúp cho người xem có thể quyết định có nên đầu tư vào lô cổ phiếu đó hay là không.

Để xây dựng được chương trình đó chúng ta cần thu thập dữ liệu từ trang web Yahoo Finance và xuất ra các thống kê mô tả trực quan về các đặc trưng. Sau đó chúng ta sẽ tiến hành làm sạch dữ liệu và tạo ra các đặc trưng mới. Bước cuối cùng sẽ là lựa chọn ra các mô hình phù hợp để dự đoán giá cổ phiếu.

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

2.1.1. Nguồn thu thập dữ liệu

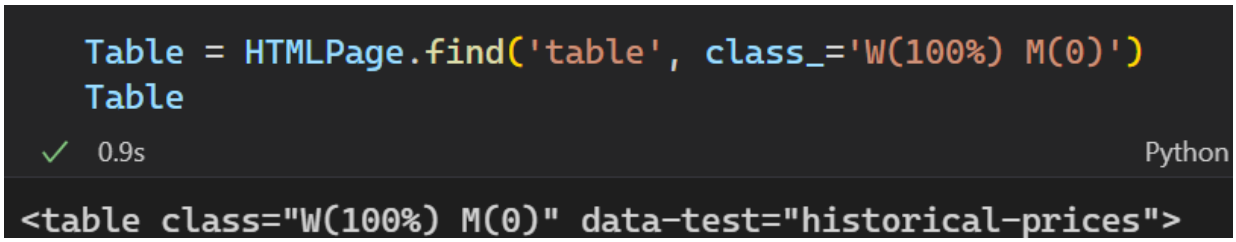
Nguồn thu thập dữ liệu được lấy từ trang web Yahoo Finance:
Link to data source

2.1.2. Công cụ thu thập

Công cụ thu thập được sử dụng ở đây là: Beautiful Soup và Selenium.

2.1.3. Cách thức sử dụng

- Selenium: Tạo ra một webdriver để truy cập vào trang web theo đường link
- Beautiful Soup: Phân tích cú pháp từ trang web lấy được
  + Lấy ra thành phần có thẻ table với class là 'W(100%) M(0)'



Hình 1: Mã chương trình và kết quả lấy ra thẻ table với class 'W(100%) M(0)'

+ Lấy ra các hàng có thẻ tr với class là 'BdT Bdc($seperatorColor) Ta(end) Fz(s) Whs(nw)'

```
Rows = Table.find_all('tr',
    class_='BdT Bdc($seperatorColor) Ta(end) Fz(s) Whs(nw)')
Rows
```

✓ 0.2s Python

```
Output exceeds the size limit. Open the full output data in a text editor
[<tr class="BdT Bdc($seperatorColor) Ta(end) Fz(s) Whs(nw)">
```

Hình 2: Mã chương trình và kết quả lấy ra thẻ tr với class 'BdT Bdc($seperatorColor) Ta(end) Fz(s) Whs(nw)'

+ Lấy ra các thẻ td trong hàng của Rows và lưu lại vào một dictionary sau đó chuyển nó thành một DataFrame

```
# Lấy ra các thẻ td chứa dữ liệu các cột của mỗi hàng
Values = Rows[i].find_all('td')

# Lấy ra 7 giá trị tương ứng với 7 cột lấy được lưu vào trong dictionary extracted_data
if len(Values) == 7:
 RowDict["Date"] = Values[0].find('span').text.replace(',', '')
 RowDict["Open"] = Values[1].find('span').text.replace(',', '')
 RowDict["High"] = Values[2].find('span').text.replace(',', '')
 RowDict["Low"] = Values[3].find('span').text.replace(',', '')
 RowDict["Close"] = Values[4].find('span').text.replace(',', '')
 RowDict["Adj Close"] = Values[5].find('span').text.replace(',', '')
 RowDict["Volume"] = Values[6].find('span').text.replace(',', '')
 extracted_data.append(RowDict)
```

Hình 3: Mã chương trình và kết quả lấy ra thẻ td và lưu vào một dictionary

```
extracted_data = pd.DataFrame(extracted_data)
extracted_data.to_csv('raw data/raw_data.csv',index=False)
extracted_data
```

✓ 0.8s

	Date	Open	High	Low	Close	Adj Close	Volume
0	May 20 2022	109.57	109.90	105.01	107.59	107.59	99500000
1	May 19 2022	106.28	110.03	106.19	107.32	107.32	88142000
2	May 18 2022	111.44	112.85	106.25	107.11	107.11	108380000
3	May 17 2022	113.28	115.80	111.28	115.37	115.37	76448000
4	May 16 2022	113.10	113.99	110.35	110.81	110.81	74566000
...	...	...	...	...	...	...	...
1759	May 28 2015	21.49	21.57	21.27	21.33	21.33	38248000
1760	May 27 2015	21.37	21.59	21.25	21.57	21.57	44622000
1761	May 26 2015	21.31	21.35	21.10	21.27	21.27	44884000
1762	May 22 2015	21.58	21.62	21.38	21.38	21.38	40412000
1763	May 21 2015	21.40	21.84	21.40	21.58	21.58	82428000

1764 rows × 7 columns

Hình 4: Dữ liệu sau khi thu thập được lưu vào file raw_data.csv

Đầu vào: Đường link dẫn đến trang web Yahoo Finance của Amazon (2015-2022).
Đầu ra: Một file có tên là raw_data.csv gồm 1764 mẫu x 7 cột chứa dữ liệu cổ phiếu được thu thập từ trang web.

2.2. Mô tả dữ liệu

2.2.1 Mô tả các đặc trưng của tập dữ liệu

Tập dữ liệu có kích thước là: 1764 mẫu x 7 cột Tổng quan về tập dữ liệu:

Cột dữ liệu	Loại dữ liệu	Mô tả cột dữ liệu
Date	object	Ngày giao dịch
Open	float64	Giá lúc mở phiên
High	float64	Giá cao nhất trong ngày
Low	float64	Giá thấp nhất trong ngày
Close	float64	Giá lúc đóng phiên
Adj Close	float 64	Giá thay đổi sau khi đóng phiên
Volume	float64	Khối lượng giao dịch trong ngày

2.2.2 Các thống kê mô tả trực quan về các đặc trưng

Dưới đây sẽ là các thống kê mô tả trực quan về các đặc trưng:

- Tổng quan về cổ phiếu AMZ từ 5/2015 đến hiện tại

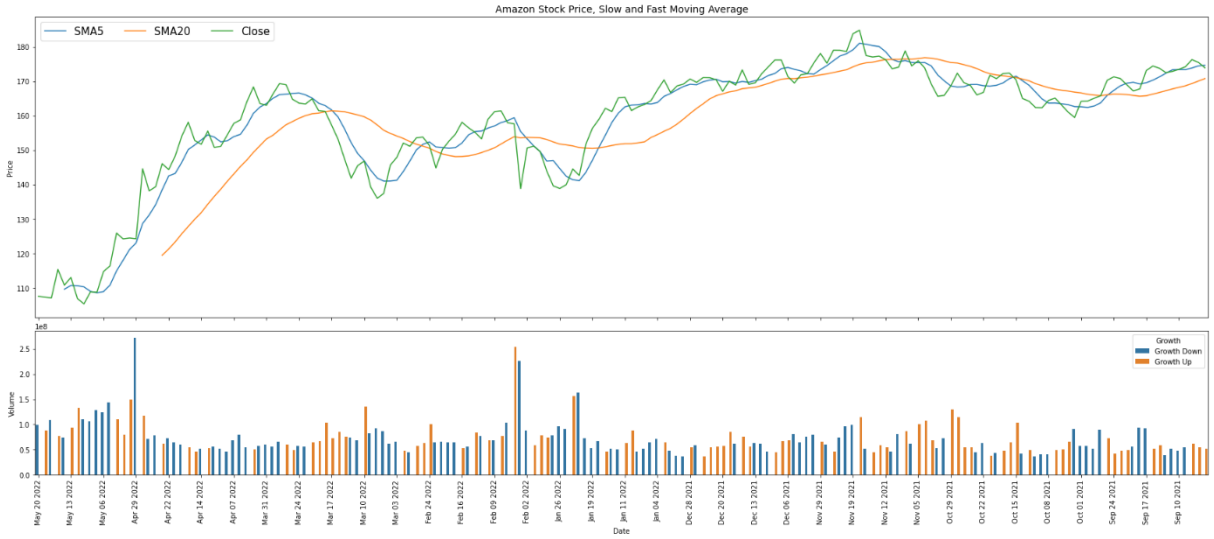
Hình 5: Tổng quan về cổ phiếu AMZ từ 5/2015 đến hiện tại

Nhận xét:

+ Giá cổ phiếu lẫn số giao dịch có sự biến động lớn, những lúc đỉnh điểm giá có thể tăng/giảm đến 15%.

+ Chỉ dựa vào những giá trị trên thì khó đoán được xu hướng cổ phiếu Amazon.

- Tổng hợp độ biến thiên các thành phần và biến động trong dataset dựa trên Volume, Price và SMA



Hình 6: Tổng hợp độ biến thiên các thành phần và biến động trong dataset dựa trên Volume, Price và SMA

Đồ thị trên là biểu diễn độ biến động về giá và khối lượng giao dịch của cổ phiếu facebook trong thời gian 6 tháng:

+ Đồ thị nằm ở phía trên là đồ thị biểu diễn biến động giá đóng sàn của cổ phiếu Amazon theo thời gian.

+ Đồ thị có chứa tham số SMA [4]: Với tham số SMA (Simple Moving Average) một dạng đồ thị đường phổ biến dùng để phân tích kỹ thuật, được tính toán bằng trung bình cộng giá đóng cửa của n phiên, đây là dạng đường dùng để dự đoán xu hướng của giá trị cổ phiếu gồm tăng, giảm và đi ngang.

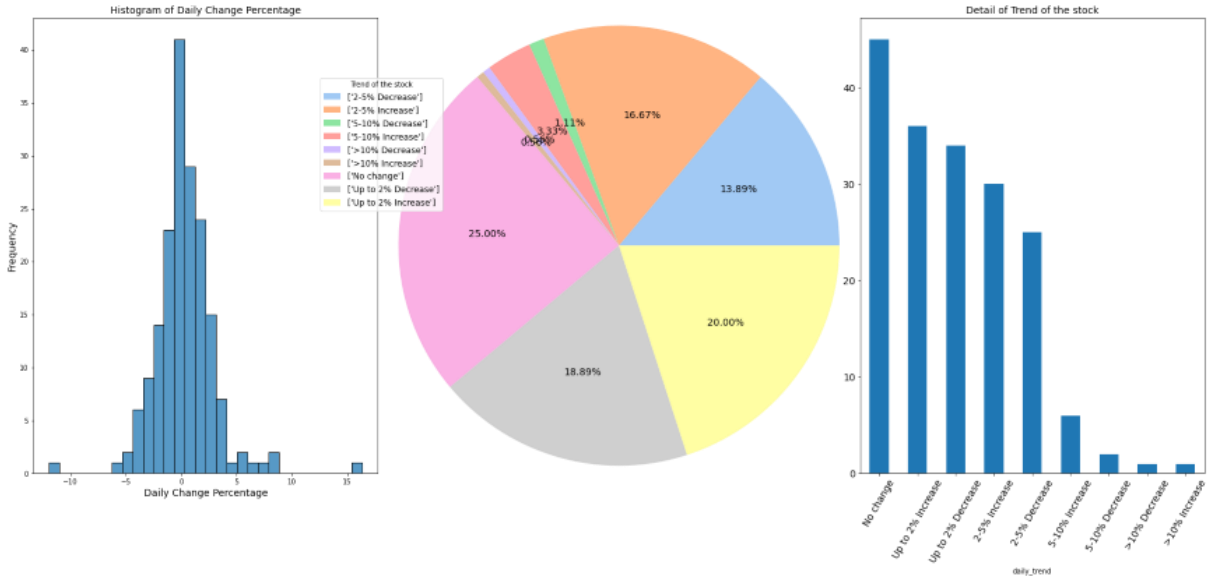
+ Đồ thị nằm ở phía dưới là một đồ thị cột thể hiện khối lượng giao dịch cổ phiếu với hai loại đó là màu xanh là tăng trưởng âm và màu cam tức là tăng trưởng dương.

Nhận xét:

+ Ta có thể thay rằng cổ phiếu Amazon là một cổ phiếu biến đổi theo xu hướng ngắn hạn bằng cách ta có thể nhìn vào đường SMA5 bám rất sát so với đường Close điều đó báo hiệu rằng việc biến đổi của các phiên giao dịch của chứng khoán có một xu thế biến đổi nhanh với khoảng thời gian 5 - 7 ngày.

+ Quan sát được đường giá đóng sàn ta có thể nhận thấy rằng là trong suốt thời điểm mà giá cả tăng thì không có sự gia tăng đột biến của khối lượng giao dịch và sự gia tăng chỉ xảy ra khi có sự biến thiên đột ngột của thị trường sau khi giá trị đóng của sàn bị biến đổi tức là khi dấu hiệu chạm đáy xảy ra từ đó ta thấy ta có thể nhìn thấy khối lượng giao dịch tăng cao là một xu thế có thể mua vào và bán ra để thu được lợi nhuận.

- Xu hướng của cổ phiếu Amazon trong 6 tháng đầu năm 2022



Hình 7: Xu hướng cổ phiếu Amazon trong 6 tháng đầu năm 2022

Nhận xét:

+ Tập đồ thị bao gồm 3 đồ thị trước hết ta đi tới đồ thị phần tay trái.

+ Đây là đồ thị thể hiện tần số của tỷ lệ phần trăm thay đổi theo tỷ giá đóng sàn của ngày hiện tại so với ngày trước đó, thông qua đồ thị histogram ta có thể thấy được thể hiện dưới dạng gồm 2 khối, một là khối các giá trị ngoại lại xuất hiện nằm tách biệt so với đồ thị làm cho histogram có dạng phân phối đảo nhỏ , hai là khối thể hiện được một phân phối lệch chuẩn với xu hướng tập trung các giá trị ở chính giữa và giảm dần về hai bên, điều, thế nhưng phần tỷ lệ dương lớn hơn tỷ lệ âm và ta có thể thấy một điều là tỷ giá của cổ phiếu Amazon có xu hướng tập trung ở vùng trung tâm và ít có xu hướng tăng đột ngột trừ những giá trị ngoại lại xuất hiện ngẫu nhiên.

+ Để thấy rõ phân bố các xu hướng ta tiếp tục đi vào đồ thị số 2 và số 3 thì đây lần lượt hai đồ thị biểu thị tần suất của loại xu hướng của cổ phiếu Amazon.

+ Với độ thị hai đây là loại đồ thị tròn dùng để thể hiện các trending dưới dạng tỷ lệ phần trăm và độ thị thứ ba là một đồ thị chi tiết hơn để minh họa số liệu phần trăm cho đồ thị hai.

3. Trích xuất đặc trưng

3.1 Làm sạch dữ liệu và tạo đặc trưng mới

Ở bước này, chúng ta tiến hành làm sạch dữ liệu bằng cách loại bỏ cột dữ liệu không mong muốn. Cột 'Adj Close' sẽ được loại bỏ.

Sau đó chúng ta sẽ tiếp tục tạo thêm các đặc trưng mới bao gồm [3]:

- 'Have_Increase' - Giá trị tăng trưởng của cổ phiếu
    + Nếu giá mở cửa trừ giá đóng cửa < 0 -> giá trị sẽ bằng 0
    + Nếu giá mở cửa trừ giá đóng cửa > 0 -> giá trị sẽ bằng 1
- 'Target' - So sánh giá trị đóng phiên trước và phiên sau
    + Nếu giá trị đóng ngày sau > giá trị đóng ngày trước -> giá trị sẽ bằng 1
    + Nếu giá trị đóng ngày sau < giá trị đóng ngày trước -> giá trị sẽ bằng 0

- Tỉ số giữa weekly_mean, quarterly_mean, annual_mean
    + Nếu giá trị weekly_mean (quarterly_mean, annual_mean) / giá trị đóng vào ngày hôm đó > 1 -> cổ phiếu có xu hướng tăng và ngược lại
- Tỉ số giữa giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa của phiên trong ngày
    + Nếu giá trị giá mở cửa / giá trị đóng cửa > 1 -> cổ phiếu có xu hướng giảm vào cuối ngày.

Tổng quan về tập dữ liệu hiện tại:

Column	**Loại dữ liệu**
Date	object
Open	float64
High	float64
Low	float64
Close	float64
Volume	float64
Have_Increase	float64
Target	float64
weekly_mean	float64
quaterly_mean	float64
annual_mean	float64
open_close_ratio	float64
high_close_ratio	float64
low_close_ratio	float64

Việc tạo ra các đặc trưng mới như weekly_mean, quaterly_mean và annual_mean sẽ gây trống dữ liệu như hình vẽ dưới đây:

Hình 8: Đồ thị của các giá trị trung bình hàng tuần, hàng quý và hàng năm khi chưa xử lý dữ liệu trống

Nhận xét trước khi xử lý dữ liệu trống:

- Dữ liệu weekly_mean, quaterly_mean và annual_mean bị mất dữ liệu lần lượt là 7 ngày, 3 tháng và 1 năm.
- Có nhiều kĩ thuật xử lý dữ liệu trống như: thay thế bằng Mean/Median/Mode, thay thế bằng giá trị đuôi phân bố, thay thế bằng giá trị ngẫu nhiên,... Tuy nhiên thay thế bằng giá trị ngẫu nhiên sử dụng trong trường hợp này sẽ mang lại kết quả tốt nhất.

3.2 Xử lý dữ liệu trống

Chúng ta sẽ sử dụng kĩ thuật xử lý dữ liệu trống đó là: Thay thế bằng giá trị ngẫu nhiên để có thể vừa giữ được độ chính xác của dữ liệu và cũng tránh làm mất đi những đặc trưng.

Và đây là kết quả sau khi xử lý dữ liệu trống:

Hình 9: Đồ thị của các giá trị trung bình hàng tuần, hàng quý và hàng năm khi đã xử lý dữ liệu trống

Nhận xét sau khi xử lý dữ liệu trống:

- Dữ liệu lấy ngẫu nhiên được giới hạn tùy thuộc vào đặc trưng mà ta đã chọn.

  Ví dụ: Đặc trưng weekly_mean sẽ bị trống 7 ngày, vì vậy ta sẽ lựa chọn dữ liệu ngẫu nhiên của 7 ngày tiếp theo để điền vào phần dữ liệu trống.

- Dữ liệu sau khi được xử lý ít bị nhiễu và nằm ở mức chấp nhận được.

4. Mô hình hóa dữ liệu

4.1 Lựa chọn mô hình

Hai mô hình được lựa chọn cho bài toán này đó chính là: Logistic Regression và Random Forest.

Đầu tiên là về mô hình Logistic Regression [1], đây là 1 thuật toán phân loại được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Thuật toán trên dùng hàm sigmoid logistic để đưa ra đánh giá theo xác suất. Trong mô hình này có các bộ tham số như sau:

- penalty: Xác định các tiêu chuẩn penalty, một số tham số solver chỉ hỗ trợ với các tham số penalty cụ thể:
    + 'l1': penalty được hỗ trợ bởi liblinear và saga solvers
    + 'l2': penalty được hỗ trợ bởi cg, sag, saga, lbfgs solvers
    + 'elasticnet': penalty chỉ được hỗ trợ bởi saga solvers
    + 'none': Quy định về penalty sẽ không được áp dụng. Không hoạt động với liblinear solver.
- solver: Thuật toán được sử dụng trong tối ưu hóa
    + Đối với các tập dữ liệu nhỏ, 'liblinear' là một lựa chọn tốt, trong khi 'sag' và 'saga' nhanh hơn cho các tập lớn;
    + Đối với các bài toán đa thức, chỉ có 'newton-cg', 'sag', 'saga' và 'lbfgs' xử lý mất đa thức;
    + 'Liblinear' được giới hạn trong các lược đồ một so với phần còn lại.
- C: Biểu thị độ mạnh của regularization và nhận một giá trị float dương. C và cường độ regularization có tương quan nghịch (C càng nhỏ thì regularization càng mạnh).

Tiếp theo chúng ta sẽ đi đến mô hình thứ hai đó là Random Forest [2], đây là 1 thuật toán sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định. Trong mô hình này có các bộ tham số như sau :

- n_estimators: Số lượng cây trong rừng [10, 100, 1000]
- max_features: Số lượng các features cần xem xét khi tìm kiếm sự phân chia tốt nhất:
    + 'sqrt': max_features=sqrt(n_features)
    + 'log2': max_features=log2(n_features)

4.2 Phân chia dữ liệu thành các tập Train/Test và bộ tham số sử dụng

Chúng ta sẽ phân tập dữ liệu thành tập Train và Test với tỉ lệ là 80 - 20 % với random_state = 5.

- Logistic Regression với bộ tham số:
    - solver
    - penalty
    - C
- Random Forest với bộ tham số:
    - n_estimators
    - max_features

4.3 Đánh giá mô hình

Với việc phân chia như trên chúng ta sẽ thu được kết quả dưới đây:

- Logistic Regression

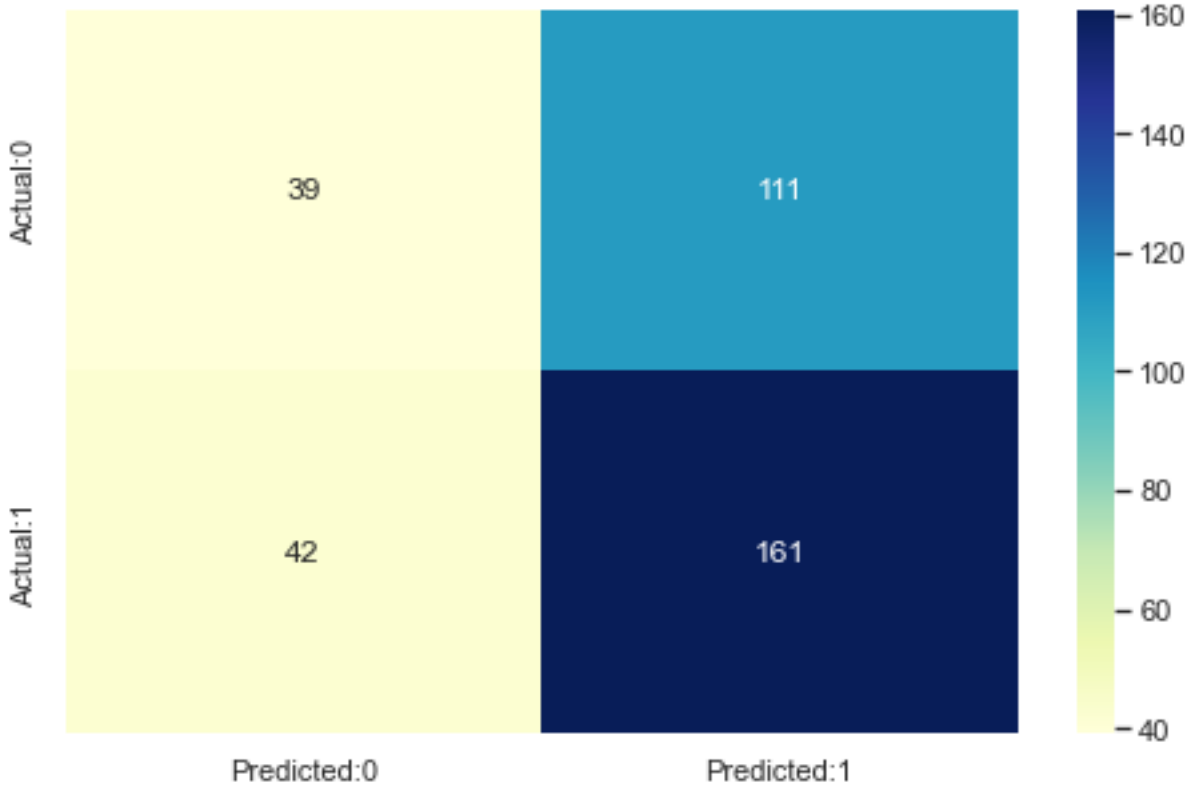
  Độ chính xác của thuật toán:

Siêu tham số và đặc trưng mặc định	Siêu tham số và đặc trưng mới
55.24 %	56.66%

  Các đặc trưng mới làm tăng độ chính xác của việc dự đoán.

  Độ chênh lệch giữa có đặc trưng mới và mặc định: 1.42%

  Ma trận nhầm lẫn trường hợp cấu hình siêu tham số và đặc trưng mới:



Hình 10: Ma trận nhầm lẫn trường hợp cấu hình siêu tham số và đặc trưng mới

Với mô hình Logistic Regression thì kết quả có xu hướng dự đoán ra 1 nhiều hơn so với kết quả thật.

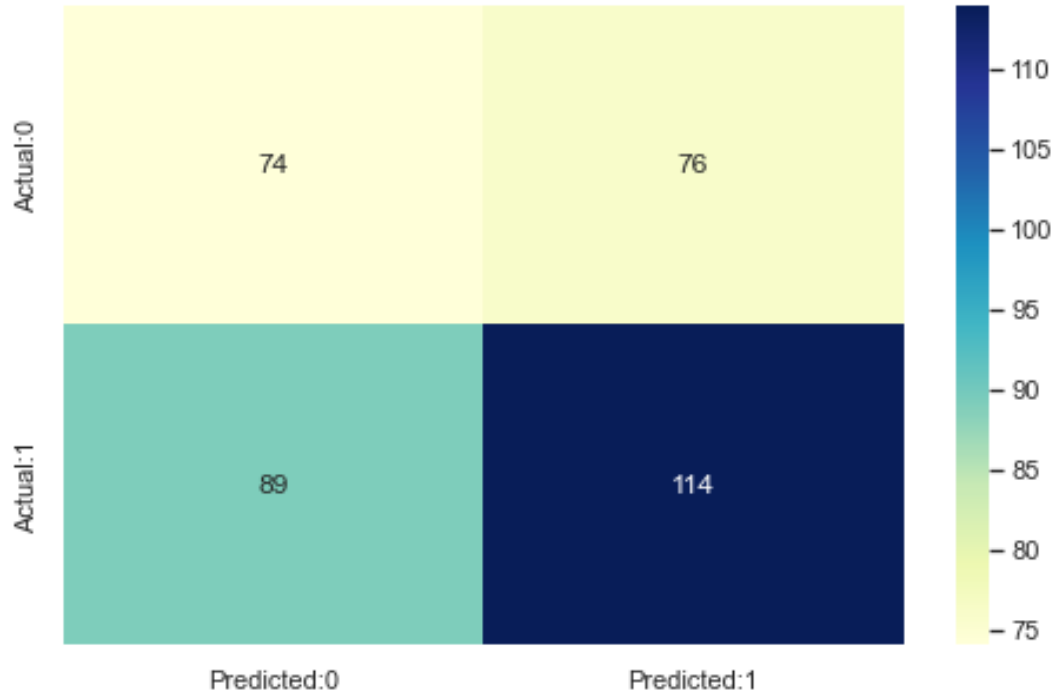
- Random Forest

Random Forest với đặc trưng mới	Logistic Regression với đặc trưng mới
53.26 %	56.65 %

Thuật toán Logistic cho độ chính xác đúng hơn so với thuật toán Random Forest.

Độ chênh lệch giữa 2 mô hình Logistic Regression và Random Forest: 3.4 %.

Ma trận nhầm lẫn thuật toán Random Forest:



Hình 11: Ma trận nhầm lẫn thuật toán Random Forest

Với mô hình Random Forest thì mô hình dự đoán ra kết quả 0 chính xác nhiều hơn nhưng dự đoán kết quả ra 1 lại không chính xác bằng Logistic Regression.

5. Kết luận

5.1 Kết quả đạt được:

- Các đặc trưng mới giúp cải thiện độ chính xác của mô hình 1 cách đáng kể.
- Mô hình Logistic có kết quả dự đoán tốt hơn so với mô hình Random Forest đối với bài toán này.
- Việc dự đoán có thể giúp ta dự đoán giá của cổ phiếu trong thời gian tương lai nhưng chỉ ở mức độ từ 51 - 56%.
- Việc dự đoán sự tăng giảm cổ phiếu là không khả thi nếu làm với các hình thức thông thường, xác suất luôn là 50% - 50% vô cùng rủi ro.

5.2 Hướng phát triển:

- Có thể áp dụng các kĩ thuật chỉ hướng để làm tăng độ chính xác của mô hình.

6. Tài liệu tham khảo

[1] Khái niệm về Logistic Regression,

Logistic Regression - Bài toán cơ bản trong Machine Learning (viblo.asia)

[2] Khái niệm về Random Forest,

Random Forest algorithm — Machine Learning cho dữ liệu dang bảng (machinelearningcoban.com)

[3] Dự đoán giá cổ phiếu sử dụng Pandas và Scikit-learn,

Portfolio Project: Predicting Stock Prices Using Pandas and Scikit-learn – Dataquest

[4] Khái niệm SMA,

Chỉ báo SMA là gì? Tác dụng của SMA trong giao dịch (topforexvn.com)