DỰ ĐOÁN XU HƯỚNG GIÁ CỔ PHIẾU BẰNG BÀI TOÁN PHÂN LỚP

1st Nguyễn Triệu Kim Oanh
Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công Nghệ Thông Tin – ĐHQGTPHCM
Hồ Chí Minh, Việt Nam
20521729@gm.uit.edu.vn

2nd Nguyễn Phan Uyển Nhi
Khoa Khoa học và Kỹ thuật thông tin
Trường Đại học Công Nghệ Thông Tin - ĐHQGTPHCM
Hồ Chí Minh, Việt Nam
20521714@gm.uit.edu.vn

Bài toán của chúng tôi dự đoán giá lên hoặc xuống một cách an toàn cho người dùng lướt sóng. Tối đa hóa những dự đoán có giá tăng và nó thực sự tăng, giảm thiểu những dự đoán tăng nhưng giá thực tế lại giảm. Điều này sẽ giúp người dùng ít tổn thất. Nghiên cứu này so sánh 3 mô hình Random Forest, Logistic Regression, K-Nearest Neighbors (KNN) với hai cách xử lý dữ liệu đầu vào. Đánh giá này được thực hiện trên bộ dữ liệu chứng khoán trong 10 năm của 3 tập đoàn Microsoft, Apple và Netflix. Độ chính xác (Precision) của tất cả mô hình nằm từ 0.4 đến 0.8 cho toàn bộ tập dữ liệu. Kết quả thử nghiệm cho thấy rằng cách xử lý dữ liệu đầu tiên trong đó sử dụng phần trăm thay đổi trên giá đóng cửa hiệu chỉnh tạo ra nhiều đặc trưng gọi là "độ trễ" làm đầu vào cho mô hình, hồi quy logistic vượt trội so với 2 mô hình dự đoán khác về hiệu suất tổng thể. Kết quả nghiên cứu này dành cho những người mới chơi cổ phiếu và muốn an toàn. Trên bối cảnh dữ liệu sẽ như thế nào nếu không có sự tác động từ bên ngoài.

Từ khóa: dự đoán cổ phiếu, mô hình phân lớp, random forest, logistic regression, k-nearest neighbor, máy học, học có giám sát.

I. GIỚI THIỆU:

Dự đoán giá cổ phiếu là quá trình dự đoán giá trị tương lai của một cổ phiếu được giao dịch trên sàn giao dịch chứng khoán để thu về lợi nhuận. Với việc dự đoán giá cổ phiếu thì cần rất nhiều yếu tố, vì thế để dự đoán giá cổ phiếu với độ chính xác cao thì rất khó nên đây là lúc máy học đóng một vai trò quan trọng.

Tuy nhiên ứng dụng máy học trong việc dự đoán cổ phiếu cũng là một thách thức lớn bởi vì sự thay đổi của thị trường phụ thuộc vào nhiều thông số nhưng chúng ta chỉ có thể định lượng được một số thông số nhất định, những thông số nhất định như là khối lượng giao dịch, giá hiện tại, giá đóng, giá mở, giá cao nhất trong ngày, giá thấp nhất và giá đóng cửa hiệu chỉnh. Những thông số mơ hồ không thể định lượng được bao gồm các yếu tố cơ bản như giá trị nội tại của công ty, tài sản, hiệu suất hàng quý, các khoản đầu tư gần nhất và chiến lược đều ảnh hưởng đến lòng tin của các nhà giao dịch đối với công ty và do đó những thông số này cũng ảnh hưởng lớn đến giá cổ phiếu của công ty. Chỉ một vài trong số đó có thể kết hợp một cách hiệu quả vào một mô hình toán học. Đó là lý do khi nói sử dụng máy học trong việc dự đoán giá cổ phiếu trở nên khó khăn và không đáng tin cậy ở một mức độ nhất định.

Vì vậy, thay vì tập trung vào việc dự đoán ra một kết quả chính xác, ở bài toán này chúng ta chỉ tập trung vào việc đưa ra dự đoán giá đóng cửa lên hoặc xuống của ngày mai bằng cách sử dụng dữ liệu của quá khứ. Nếu mô hình nói rằng giá sẽ tăng, chúng ta sẽ mua cổ phiếu. Nếu mô hình nói rằng giá sẽ giảm, chúng ta sẽ không làm gì cả [1].

Có nhiều nghiên cứu sử dụng công nghệ học sâu như LSTM, GAN, RNN trong việc dự đoán giá cổ phiếu. Ở đây, chúng tôi muốn xem xét những mô hình học máy bình thường và cụ thể là những mô hình phân lớp để phân loại giá cổ phiếu lên hoặc xuống, thay vì những mô hình có độ khó cao với kết quả chính xác giá cổ phiếu. Kết quả nghiên cứu cho ra một mô hình tham khảo dành cho những người mới chơi cổ phiếu hoặc những người chơi lướt sóng. Chúng tôi đề xuất ba phương pháp phân loại: Random Forest, Logistic Regeression và K-Nearest Neighbors (KNN) trong dự đoán giá cổ phiếu. Kết quả thực nghiệm cho thấy mô hình đề xuất đạt được kết quả khả quan về độ đo chính xác (Precision) trên bộ dữ liệu đã xử lý được sử dụng để kiểm tra và huấn luyện.

II. DỮ LIỆU:

A. *Giới thiệu tập dữ liệu*

Chúng tôi đã chọn 3 cổ phiếu của các tập đoàn có khối lượng giao dịch lớn là cổ phiếu của tập đoàn MSFT (Microsoft), NFLX (Netflix) và AAPL (Apple) với khoảng thời gian giao dịch là 10 năm từ năm 2012 đến 2022, với khoảng 7551 mẫu. Với mỗi tập dữ liệu, có 2517 dòng, 6 cột bao gồm những thuộc tính như: giá cao nhất trong một phiên giao dịch (High), giá thấp nhất trong một phiên giao dịch (Low), giá mở cửa (Open), giá đóng cửa (Close), giá đóng cửa hiệu chỉnh (Adj close).

B. *Phân tích sơ bộ tập dữ liệu*

Chúng tôi đánh giá, xây dựng mô hình trên dữ liệu cổ phiếu MSFT, NFLX và AAPL. Tất cả dữ liệu cổ phiếu được lấy từ trang web https://finance.yahoo.com/. Năm dòng đầu của từng tập dữ liệu được chỉ ra trong Hình 1, 2, 3.

Cổ phiếu của tập đoàn MSFT, số lượng 2517 mẫu.

Date	Open	High	Low	Close	Adj Close	Volume
2012-11-01	28.840000	29.559999	28.820000	29.520000	24.142986	72047900
2012-11-02	29.590000	29.770000	29.330000	29.500000	24.126631	57131600
2012-11-05	29.620001	29.740000	29.330000	29.629999	24.232956	38070800
2012-11-06	29.820000	30.200001	29.610001	29.860001	24.421057	43401500
2012-11-07	29.530001	29.830000	29.049999	29.080000	23.783125	57871800

Hình 1. Cổ phiếu của tập đoàn MSFT

Cổ phiếu của tập đoàn NFLX, số lượng 2517 mẫu.

Date	Open	High	Low	Close	Adj Close	Volume
2012-11-01	11.121429	11.384286	10.664286	11.098571	11.098571	62669600
2012-11-02	11.035714	11.407143	10.852857	10.985714	10.985714	29650600
2012-11-05	10.652857	11.277143	10.642857	11.177143	11.177143	29016400
2012-11-06	11.008571	11.421429	10.815714	10.910000	10.910000	39102700
2012-11-07	10.918571	11.321429	10.918571	11.097143	11.097143	31099600

Hình 2. Cổ phiếu của tập đoàn NFLX

Cổ phiếu của tập đoàn AAPL, số lượng 2517 mẫu.

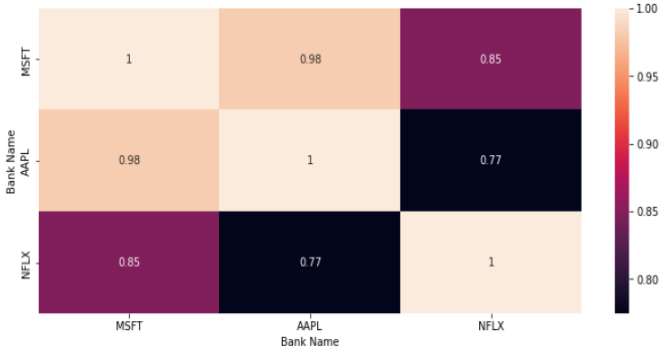
Date	Open	High	Low	Close	Adj Close	Volume
2012-11-01	21.365000	21.535713	21.220358	21.305000	18.238737	361298000
2012-11-02	21.281786	21.319643	20.526787	20.600000	17.635201	599373600
2012-11-05	20.840000	20.991785	20.628571	20.879286	17.874287	529135600
2012-11-06	21.079643	21.097857	20.717501	20.816071	17.820166	374917200
2012-11-07	20.494286	20.519285	19.848213	19.928572	17.138329	793648800

Hình 3. Cổ phiếu của tập đoàn AAPL

Dưới đây là biểu đồ giá đóng cửa của ba công ty, có thể thấy rằng giá đóng cửa của NFLX có sự biến động trong giai đoạn từ năm 2018 đến năm 2023, và từ giữa năm 2021 giá đang có xu hướng giảm mạnh. Giá đóng cửa của MSFT cho thấy rằng bắt đầu từ 2019 giá đã tăng lên, tăng lên cao nhất vào năm 2022 sau đó có xu hướng giảm. Cuối cùng, giá đóng cửa của AAPL rất ổn định, tăng dần qua các năm.

Hình 4. Biểu đồ giá đóng cửa của ba công ty

Từ biểu đồ thể hiện sự tương đồng của tập dữ liệu, cho chúng ta thấy rằng mức độ tương đồng giữa các tập dữ liệu rất cao. AAPL và MSFT là 0.98, NFLX và MSFT là 0.85, AAPL và NFLX là 0.77. Nói một cách khác thì hai tập dữ liệu có độ tương đồng cao khi xu hướng tăng giảm của nó gần giống nhau trên thị trường chứng khoán.



Hình 5. Biểu đồ thể hiện sự tương đồng của tập dữ liệu

C. Xử lý dữ liệu:

Chúng tôi có hai phương pháp xử lý dữ liệu: Phương pháp đầu tiên, chúng tôi tạo một đặc trưng "Today" bằng cách lấy phần trăm thay đổi giữa giá hiện tại và giá trong quá khứ của giá đóng cửa hiệu chỉnh bằng hàm pct. Sau đó tạo thêm các đặc trưng Lag (Độ trễ) dựa vào đặc trưng "Today". Độ trễ chính là Today nhưng được dịch chuyển "về phía trước" tương ứng với đánh số của Độ trễ. Thêm đặc trưng Mục tiêu (Target) thể hiện sự tăng giảm của đặc trưng "Today", 1 là tăng và 0 là giảm [2].

	Date	Today	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Volume	Target
8	2012-11-14	-0.922881	-3.215418	-2.115861	0.069430	-0.928471	-2.612213	0.776259	0.440657	0.131689	0
9	2012-11-15	-0.670620	-0.922881	-3.215418	-2.115861	0.069430	-0.928471	-2.612213	0.776259	0.076086	0
10	2012-11-16	-0.525132	-0.670620	-0.922881	-3.215418	-2.115861	0.069430	-0.928471	-2.612213	0.050956	0
11	2012-11-19	0.791857	-0.525132	-0.670620	-0.922881	-3.215418	-2.115861	0.069430	-0.928471	0.064083	1
12	2012-11-20	-0.074820	0.791857	-0.525132	-0.670620	-0.922881	-3.215418	-2.115861	0.069430	0.057179	0
...	...	...	...	...	...	...	...	...	...	...	...
2512	2022-10-26	-7.715629	1.379166	2.118791	2.528052	-0.139546	-0.846962	0.408368	3.924572	0.034776	0
2513	2022-10-27	-1.975619	-7.715629	1.379166	2.118791	2.528052	-0.139546	-0.846962	0.408368	0.082543	0
2514	2022-10-28	4.022046	-1.975619	-7.715629	1.379166	2.118791	2.528052	-0.139546	-0.846962	0.040425	1
2515	2022-10-31	-1.585614	4.022046	-1.975619	-7.715629	1.379166	2.118791	2.528052	-0.139546	0.040648	0
2516	2022-11-01	-1.705939	-1.585614	4.022046	-1.975619	-7.715629	1.379166	2.118791	2.528052	0.028357	0

Hình 6. Kết quả sau khi xử lý dữ liệu của phương pháp 1

Phương pháp thứ hai, chúng tôi sử dụng phương pháp dịch chuyển DataFrame để di chuyển tất cả các hàng "về phía trước" trong một ngày giao dịch. Điều này nhằm đảm bảo rằng chúng ta đang dự đoán giá trong tương lai bằng cách sử dụng dữ liệu trong quá khứ. Sau đó chúng tôi lấy bảng có các giá trị đóng, mở, cao, thấp, khối lượng giao dịch đã được dịch chuyển này nối với bảng giá đóng cửa hiệu chỉnh ban đầu chưa được dịch chuyển. Chúng ta sẽ được bảng dữ liệu thể hiện được giá đóng cửa hiệu chỉnh của ngày hôm trước. Và cũng thêm đặc trưng Mục tiêu (Target) thể hiện sự tăng giảm của giá đóng cửa, 1 là tăng và 0 là giảm [1].

	Date	Adj Close	Target	Close	Volume	Open	High	Low
0	2012-11-02	24.126627	0.0	29.520000	72047900.0	28.840000	29.559999	28.820000
1	2012-11-05	24.232943	1.0	29.500000	57131600.0	29.590000	29.770000	29.330000
2	2012-11-06	24.421057	1.0	29.629999	38070800.0	29.620001	29.740000	29.330000
3	2012-11-07	23.783127	0.0	29.860001	43401500.0	29.820000	30.200001	29.610001
4	2012-11-08	23.562311	0.0	29.080000	57871800.0	29.530001	29.830000	29.049999
...	...	...	...	...	...	...	...	...
2511	2022-10-25	249.955582	1.0	247.250000	24911200.0	243.759995	247.839996	241.300003
2512	2022-10-26	230.669937	0.0	250.660004	34775500.0	247.259995	251.039993	245.830002
2513	2022-10-27	226.112778	0.0	231.320007	82543200.0	231.169998	238.300003	230.059998
2514	2022-10-28	235.207138	1.0	226.750000	40424600.0	231.039993	233.690002	225.779999
2515	2022-10-31	231.477661	0.0	235.869995	40647700.0	226.240005	236.600006	226.050003

Hình 7. Kết quả sau khi xử lý dữ liệu của phương pháp 2

III. PHƯƠNG PHÁP MÁY HỌC:

A. Mô hình Random Forest

Random Forest là một phương pháp học máy có giám sát. Mô hình này bao gồm nhiều cây quyết định, tuy nhiên mỗi cây quyết định sẽ khác nhau, vì vậy mà nó tránh được hiện tượng thường gặp của mô hình cây quyết định là Overfitting. Đầu ra của Random Forest được tổng hợp từ các cây quyết định [3]. Random Forest vừa có thể sử dụng được trong bài toán hồi quy vừa sử dụng được trong bài toán phân lớp. Ở bài toán này, chúng tôi sử dụng Random Forest Classifier [4].
Mô hình này chúng tôi sử dụng phương pháp xử lý dữ liệu thứ hai. Chúng tôi lấy các giá trị là giá mở, giá đóng, giá cao, giá thấp, khối lượng giao dịch làm đầu vào của mô hình và giá trị Mục tiêu (Target) là đầu ra của mô hình. Sau đó, chia tập train sử dụng dữ liệu trước năm 2022 và tập test sử dụng dữ liệu năm sau 2022 để kiểm tra mô hình dự đoán gì cho năm 2022, so sánh với giá trị thực tế ở năm 2022.

Mô hình ban đầu của chúng tôi các giá trị tham số cơ bản như sau: Số cây trong rừng (n_estimators) là 100, số lượng mẫu tối thiểu tại mỗi leaf node để có thể tiếp tục mở rộng tree (min_sample_split) là 200, chúng tôi thiết lập min_samples_split cao để tránh tình trạng Overfitting và random_state = 1, tham số này giống nhau thì các tập dữ liệu con sinh ra sẽ giống nhau mỗi khi gọi lại mô hình.

B. Mô hình Logistic Regression

Trong mô hình Logistic Regression, chúng tôi sử dụng Statsmodels.api để dự đoán giá cổ phiếu. Statsmodels là một mô-đun Python cung cấp các chức năng khác nhau để ước tính các mô hình thống kê khác nhau và thực hiện các bài kiểm tra thống kê [5].

Đối với mô hình này chúng tôi dùng phương pháp xử lý dữ liệu thứ nhất. Đầu tiên, chúng tôi xác định biến phụ thuộc (y) và biến độc lập (X). Trong đó y là "Target" và X là "const, Lag 1, Lag 2, Lag 3, Lag 4, Lag 5, Lag 6, Lag 7, Volume". Sau đó, chia tập train sử dụng dữ liệu trước năm 2022 và tập test sử dụng dữ liệu năm 2022 để kiểm tra mô hình dự đoán gì cho năm 2022, so sánh với giá trị thực tế ở năm 2022. Dùng hàm Logit() để thực hiện hồi quy Logistic. Hàm Logit () chấp nhận y và X làm tham số và trả về đối tượng Logit [5]. Kết quả đầu ra của hàm là:

Đối với MSFT số lần lặp lại để tối ưu hóa mô hình là 5 lần và giá trị function hiện tại là 0.68557.

```
Optimization terminated successfully.
         Current function value: 0.685570
         Iterations 5
```

Hình 8. Kết quả đầu ra của MSFT

NFLX số lần lặp lại để tối ưu hóa mô hình là 4 lần và giá trị function hiện tại là 0.689321.

```
Optimization terminated successfully.
         Current function value: 0.689321
         Iterations 4
```

Hình 9. Kết quả đầu ra của NFLX

AAPL số lần lặp lại để tối ưu hóa mô hình là 5 lần và giá trị function hiện tại là 0.687797.

```
Optimization terminated successfully.
         Current function value: 0.687797
         Iterations 4
```

Hình 10. Kết quả đầu ra của AAPL

Cuối cùng, chúng tôi dùng confusion_matrix để hình dung hiệu suất của thuật toán và từ đó tính precision.

C. Mô hình K-Nearest Neighbors

Mô hình KNN là một thuật toán học máy có giám sát dạng lazy learning (thuộc dạng lười học): Khi có dữ liệu mới thì thuật toán mới đi thực hiện tính toán để ra kết quả dự đoán. Nhãn của 1 dữ liệu mới được dự đoán dựa vào k "láng giềng" (neighbor) gần nhất của nó (k = 1...n). Điểm cốt lõi của KNN là viêc tính khoảng cách (distance) của dữ liệu với các điểm lân cận của nó [3]. Chúng tôi sử dụng phương pháp K – Neighbors Classifier để phân loại giá lên hoặc giá xuống [6].

Tương tự với mô hình hồi quy Logistic, đầu vào là "const, Lag 1, Lag 2, Lag 3, Lag 4, Lag 5, Lag 6, Lag 7, Volume" và đầu ra là "Target" với các tham số K – Neighbors Classifier là k "láng giềng" (n_neighbors) = 5, metric = 'minkowski' và p = 2 theo mặc định của mô hình.

D. Độ đo sử dụng để đánh giá

Chúng tôi sử dụng một độ đo duy nhất để đánh giá các mô hình là độ đo chính xác (Precision) bởi vì mục tiêu bài toán của chúng tôi là đề cao sự an toàn, có nghĩa là tối đa hóa các dự đoán True Positive và giảm thiểu tối đa các dự đoán False Negative. Vì vậy, Precision càng cao thì các dự đoán sẽ càng an toàn cho người chơi chứng khoán.

Dưới đây là bảng độ đo của 3 mô hình trên từng tập dữ liệu khi Random Forest sử dụng tập dữ liệu của phương pháp xử lý thứ hai, Logistic Regression sử dụng tập dữ liệu của phương pháp xử lý thứ nhất, K- Nearest Neighbors sử dụng tập dữ liệu của phương pháp xử lý thứ nhất.

Mô hình	Precision		
	MSFT	NFLX	AAPL
Random Forest	0.46	0.45	0.48
Logistic Regression	0.81	0.72	0.7
K- Nearest Neighbors	0.43	0.42	0.48

Bảng 1. Bảng độ đo ban đầu của 3 mô hình

IV. CÁC THỬ NGHIỆM TINH CHỈNH MÔ HÌNH:

Chúng tôi quyết định xử lý dữ liệu bằng hai cách bởi vì mô hình Random Forest có độ chính xác Precision thấp khi thử nghiệm trên tập dữ liệu thứ nhất. Để tinh chỉnh mô hình này, chúng tôi muốn tạo thêm các đặc trưng tiêu biểu khi làm việc với dữ liệu thời gian. Vì vậy, với mô hình Random Forest chúng tôi thử tạo một phương pháp xử lý dữ liệu mới cùng với đầu vào sẽ khác với dữ liệu ban đầu. Đầu vào mới cho Random Forest là giá mở, giá đóng, giá cao, giá thấp và khối lượng giao dịch cùng với các đặc trưng trung bình thời gian. Và nó sử dụng 1000 dòng dữ liệu ban đầu và dự đoán 750 dòng tiếp theo.

Mô hình KNN không có Precision khả quan khi sử dụng tập dữ liệu của phương pháp thứ hai vì vậy mà chúng tôi quyết định sử dụng mô hình này trên tập dữ liệu thứ nhất. KNN và Random Forest là hai mô hình cần tinh chỉnh vì độ chính xác ở dưới trung bình.

A. Mô hình Random Forest

Để tối ưu hóa mô hình Random Forest khi làm việc trên dữ liệu thời gian, chúng tôi thêm các đặc trưng dự đoán mô hình là các giá trị trung bình của từng tuần, từng quý và từng năm, tỷ lệ giữa các trung bình khác nhau (Điều này giúp thuật toán hiểu xu hướng hàng tuần hay hàng quý có liên quan gì đến xu hướng hàng năm), tỷ lệ giữa giá mở cửa, thấp, cao với giá đóng cửa (Điều này giúp thuật toán hiểu xu hướng giá trong ngày vừa rồi, nếu mức cao lớn hơn nhiều so với giá đóng cửa, điều đó có thể có nghĩa là cổ phiếu đang có xu hướng giảm vào cuối ngày). Khi thêm các đặc trưng trên, độ chính xác Precision lạc quan hơn rất nhiều, từ 0.5 đến 0.6 khi kết hợp kỹ thuật GridSearchCV.

Mô hình tinh chỉnh	Precision		
	MSFT	NFLX	AAPL
Random Forest ban đầu	0.46	0.45	0.48
Random Forest sử dụng GridSearchCV	0.46	0.46	0.5
Random Forest thêm đặc trưng kết hợp GridSearchCV	0.54	0.58	0.64

Bảng 2. Bảng độ đo tinh chỉnh của Random Forest

B. Mô hình KNN

Để tinh chỉnh mô hình KNN [7], tương tự với Random Forest, chúng tôi sử dụng GridSearchCV để tìm ra k "láng giềng" phù hợp nhất. Lúc này, độ chính xác vẫn còn khá thấp nên chúng tôi thêm trọng số trung bình lân cận dựa trên khoảng cách. Tìm trọng số phù hợp nhất cho mô hình nhờ vào kỹ thuật GridSearchCV. Sau khi có k "láng giềng" và trọng số phù hợp thì độ chính xác Precision của mô hình KNN đã tăng lên đáng kể so với ban đầu nhưng trung bình vẫn dưới 0.5.

Mô hình tinh chỉnh	Precision		
	MSFT	NFLX	AAPL
KNN ban đầu	0.43	0.42	0.48
KNN sử dụng GridSearchCV	0.45	0.48	0.5
KNN thêm trọng số kết hợp GridSearchCV	0.47	0.48	0.5

Bảng 3. Bảng độ đo tinh chỉnh của KNN

V. PHÂN TÍCH, HƯỚNG PHÁT TRIỂN:

A. Phân tích lỗi

Bài toán của chúng tôi đã trở nên phức tạp hơn khi có hai phương án xử lý dữ liệu. Các mô hình làm việc trên tập dữ liệu mà có cách xử lý khác nhau sẽ dẫn đến việc các mô hình không thể so sánh với nhau mặc dù có cùng tập dữ liệu. Ngoài ra, các mô hình sau khi tinh chỉnh cũng không đạt được kết quả mong muốn, Random Forest có độ đo nằm từ 0.5 cho đến 0.6 trên toàn bộ tập dữ liệu, KNN có độ đo nằm từ 0.4 cho đến 0.5 trên toàn bộ tập dữ liệu. Độ chính xác thể hiện rằng hai mô hình này chưa đủ an toàn cho người dùng thử nghiệm.

B. Hướng phát triển

Trong tương lai, chúng tôi muốn kết hợp hai phương pháp xử lý dữ liệu này theo hướng tối ưu như xem xét đặc trưng nào giữ lại, đặc trưng nào không cần thiết và đánh giá mô hình lại một lần nữa. Và chúng tôi muốn thử thêm nhiều phương pháp tinh chỉnh mô hình và thử nghiệm thêm nhiều mô hình phân lớp khác hoặc những phương pháp học sâu để tìm ra những mô hình phù hợp trong việc dự đoán giá cổ phiếu. Ngoài ra, chúng tôi cũng sẽ đánh giá lại mô hình trên những tập dữ liệu chứng khoán có độ tương đồng thấp với nhau để thu được kết quả tổng quát hơn. Cuối cùng, chúng tôi sẽ phát triển thành ứng dụng hoặc web để thân thiện với người dùng.

VI. KẾT LUẬN:

Dự đoán giá cổ phiếu là một bài toán đầy thách thức do giá trị cổ phiếu thay đổi liên tục phụ thuộc vào nhiều yếu tố. Bộ dữ liệu có sẵn bao gồm một số tính năng như giá cao, giá thấp, giá mở, giá đóng, giá đóng cửa hiệu chỉnh, và khối lượng cổ phiếu được giao dịch trong ngày. Để có được độ chính xác cao thì bài toán này trở thành một bài toán phức tạp, bởi nhiều thông số mơ hồ không thể định lượng và những sự kiện ngẫu nhiên từ môi trường bên ngoài ảnh hưởng trực tiếp đến giá cổ phiếu ví dụ như Đại dịch Covid. Vì thế, chúng tôi chỉ dự đoán giá lên hoặc xuống, cho ra một kết quả tham khảo dành cho những người mới chơi hoặc người dùng lướt sóng thích sự an toàn. Sau khi ứng dụng ba mô hình Random Forest, Logistic Regression, KNN trên ba tập dữ liệu Microsoft, Apple, Netflix. Chúng tôi nhận thấy Logistic Regression sử dụng phân tích thống kê phù hợp nhất trong việc dự đoán chứng khoán vì nó có độ an toàn cao, phù hợp với mục tiêu ban đầu mà chúng tôi đặt ra, mô hình Random Forest và KNN sau khi tinh chỉnh mặc dù tốt hơn so với ban đầu nhưng vẫn chưa đủ để người dùng có thể tham khảo kết quả này. Trong tương lai, chúng tôi muốn kết hợp hai phương pháp xử lý dữ liệu của chúng tôi và đánh giá các mô hình lại lần nữa, và chúng tôi cũng sẽ xem xét thêm nhiều phương pháp để tối ưu mô hình.

Để tiếp cận người dùng dễ dàng hơn, chúng tôi muốn phát triển bài toán này thành ứng dụng hoặc web.

TÀI LIỆU THAM KHẢO

[1] V. Paruchuri, "DATAQUEST," Dataquest, 16 12 2021. [Trực tuyến]. Available: https://www.dataquest.io/blog/portfolio-project-predicting-stock-prices-using-pandas-and-scikit-learn/. [Đã truy cập 16 11 2022].

[2] G. James, D. Witten, T. Hastie và R. Tibshirani, An Introduction to Statistical Learning, Springer Science & Business Media, 2013.

[3] N. Tuấn, "TabML - Machine Learning cho dữ liệu bảng," Jupyter Book, [Trực tuyến]. Available: https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html. [Đã truy cập 15 11 2022].

[4] J. d. Boisberranger, J. V. d. Bossche, L. Estève, ... , M. Zain, "Scikit-learn," scikit-learn developers (BSD License), [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Đã truy cập 20 10 2022].

[5] Cosine1509, "GeeksforGeeks," GeeksforGeeks, 26 11 2022. [Trực tuyến]. Available: https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/?fbclid=IwAR2AaTf6euVNx_lgMFSvY96QRmD1nAxq1Cz3QMAzyZCs9jNEexP-YLoZiVY. [Đã truy cập 5 12 2022].

[6] J. d. Boisberranger, J. V. d. Bossche, L. Estève, ... , M. Zain, "Scikit-learn," scikit-learn developers (BSD License), [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Đã truy cập 30 11 2022].

[7] J. Korstanje, "Real Python," Real Python, [Trực tuyến]. Available: https://realpython.com/knn-python/. [Đã truy cập 5 12 2022].

[8] J. Patel, S. Shah, P. Thakkar và KKotecha, "Expert Systems with Applications," *Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques*, tập 42, số 1, pp. 259-268, 1 2015.