Phân tích các phương pháp dự báo bằng máy học trên dữ liệu thị trường ngoại hối

Le Tran Hoai An[1,2]
18520426@gm.uit.edu.vn

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Tóm tắt nội dung Tỷ giá hối đoái của mỗi cặp tiền có thể được dự đoán bằng cách sử dụng thuật toán học máy trong quá trình phân loại. Với sự trợ giúp của mô hình học máy được giám sát, xu hướng tăng hoặc giảm được dự đoán của tỷ giá FoRex có thể giúp các nhà giao dịch đưa ra quyết định đúng đắn về các giao dịch trên FoRex. Việc tiến hành phân tích kết quả mang lại một số thông tin nhất định cho việc sử dụng các phương pháp xử lý cũng như mô hình trên dữ liệu thị trường này.

Keywords: Technical analysis · Feature selection · Feature extraction · Machine-learning techniques · FoRex prediction

1 Giới thiệu

Thị trường ngoại hối là một thị trường phi tập trung toàn cầu cho việc trao đổi các loại tiền tệ. Với sự bùng nổ dữ liệu ngày nay, đặc biệt là dữ liệu lớn của doanh nghiệp, các kỹ thuật máy học có thể được ứng dụng để hỗ trợ doanh nghiệp, chủ kinh doanh xem xét và phân tích sâu hơn để có thể trích xuất những thông tin hữu ích cho mục đích của công ty cũng như đề xuất, đưa ra những quyết định, phương án kinh doanh phù hợp. Đặc biệt đáng nói tới trong đó là dự đoán giá cả cổ phiếu trên thị trường, tỷ giá ngoại hối.

Trong thị trường, các giao dịch trao đổi tiền có thể được dự đoán tăng hoặc giảm từ việc huấn luyện mô hình trên bộ dữ liệu về giá được thu thập trong quá khứ. Mục tiêu kết quả đầu ra của bài toán thường sẽ là các giá trị danh nghĩa (rời rạc) hoặc liên tục. Tuy nhiên, do tính chất phi cố định và biến động cao của thị trường FoRex(Foreign Exchange) nên hầu hết các thuật toán dự đoán với đầu ra là giá trị liên tục đều không đạt kết quả tốt khi đưa vào thực tế. Do đó nhóm không cố gắng dự đoán giá trị tỷ giá hối đoái thực tế giữa hai loại tiền tệ, mà xem vấn đề dự đoán như một nhiệm vụ phân loại xu hướng tiếp theo của thị trường sẽ tăng hoặc giảm

2 Công trình liên quan

Việc áp dụng các kỹ thuật máy học vào dự đoán, phân tích về lĩnh vực kinh tế thị trường mặc dù đã có từ lâu nhưng vì sự phức tạp, độ khó của bài toán và chủ yếu là những người dùng sử dụng với mục đích đem lại lợi nhuận về mặt kinh tế. Do đó những nội dung liên quan về đề tài dự đoán giá cả của thị trường được công bố ở các hội nghị khoa học lớn vẫn còn hiếm.

Trước khi thực hiện đề tài nhóm đã tham khảo và tìm hiểu về một số bài báo khoa học có liên quan, một trong số đó như FoRex Trend Classification using Machine Learning Techniques (Areej Abdullah Baasher et al.) [1] đề cập nghiên cứu phân tích kỹ thuật FoRex nhiều khung thời gian và các tính năng xử lý tín hiệu để dự đoán xu hướng tỷ giá cao hàng ngày. Dự đoán được đặt ra như một bài toán phân loại nhị phân mà hệ thống dự đoán liệu tỷ lệ giá cao sẽ tăng hay giảm. Các phương pháp lựa chọn tính năng dựa trên SVM và Bagging Trees cũng như năm kỹ thuật trích xuất đặc trưng đều được sử dụng để tìm các tập con đặc trưng tốt nhất cho bài toán phân loại. Các bộ phân loại học máy (RBF, MLP và SVM) đều được đào tạo bằng cách sử dụng nhiều tập hợp con tính năng khác nhau và kết quả được hiển thị và so sánh dựa trên hiệu suất phân loại phần trăm.

Bài báo FoRex Trading Using Supervised Machine Learning (Thuy Nguyen Thi Thu et al.) [2] đề cập sâu hơn về sử dụng Support Vector Machine cho bài toán này, hỗ trợ giúp dự báo xu hướng tăng hoặc giảm của FoRex. Các vấn đề FoRex hiện nay có thể được coi là các vấn đề phân loại, cụ thể là với các mô hình SVM. Thử nghiệm đã cho thấy sức mạnh của SVM trong việc thực hiện dữ liệu chuỗi thời gian khổng lồ, phức tạp để giúp đưa ra quyết định giao dịch phù hợp.

Bài báo Event-Driven LSTM For FoRex Price Prediction (Ling Qi et al.)[3] có đưa ra những nghiên cứu dựa trên dữ liệu, các tính năng rút trích thông tin từ dữ liệu và các chỉ số phân tích kỹ thuật dựa vào kỹ thuật AI tiên tiến.

3 Tiền xử lý dữ liệu

3.1 Dữ liệu thị trường FoRex

Dữ liệu thị trường ngoại hối chứa thông tin tỷ giá của một cặp tiền tệ trên thế giới, dữ liệu có thể được biểu diễn trên nhiều khung thời gian khác nhau, với mỗi điểm dữ liệu đại diện mỗi phiên giao dịch của khung thời gian đó. Mỗi điểm dữ liệu sẽ biểu diễn các thông tin: Thời điểm bắt đầu phiên giao dịch, tỷ giá lúc bắt đầu phiên giao dịch, tỷ giá ở thời điểm cuối của phiên giao dịch, tỷ giá cao nhất, thấp nhất trong cả phiên giao dịch đó và khối lượng giao dịch trong phiên.

Các kỹ thuật dự đoán chuỗi thời gian (time series) trong kinh tế được phân thành hai loại chính; cụ thể là các kỹ thuật cố gắng dự đoán giá trị thực tế của tỷ giá hoặc giá trị lợi nhuận thực tế và các kỹ thuật cố gắng dự đoán hướng xu hướng (xu hướng tăng, xu hướng giảm và đi ngang). Với kỹ thuật dự đoán xu hướng, nhóm tạo thêm một thuộc tính nhãn dự đoán mang các giá trị đại diện cho xu hướng về giá dựa trên sự chênh lệch về giá trị tỷ giá

Nhóm chọn ra các cặp tiền có mức thanh khoản tương đối trên thị trường như USD/JPY, GBP/NZD, EUR/CAD, AUD/CHF để tiến hành thu thập dữ liệu trên khung thời gian ngày và đưa vào bài toán dự đoán. Việc chọn khung thời gian sẽ tùy thuộc vào xu hướng muốn dự đoán là ở khoảng thời gian ngắn hạn trong tương lai hoặc trong dài hạn, tuy nhiên việc lựa chọn này ảnh hưởng lớn đến dữ liệu và hiệu suất hoạt động của mô hình dự đoán trong tương lai do tính chất biến động và gây nhiễu của thị trường. Sau nhiều lần thử nghiệm và tìm hiểu, nhóm chọn dữ liệu biểu diễn ở khung thời gian ngày với mục đích có được nhiều thông tin từ dữ liệu nhưng không bị nhiễu thông tin về dữ liệu quá nhiều.

3.2 Chỉ báo kỹ thuật

Hình 1: Chỉ báo kỹ thuật trong FoRex

Phân tích kỹ thuật (Technical analysis) được định nghĩa là việc sử dụng các chuỗi số được tạo ra bởi hoạt động thị trường, chẳng hạn như tỷ giá và khối lượng, để dự đoán các xu hướng trong tương lai trên thị trường đó. Các kỹ thuật này có thể được áp dụng cho bất kỳ thị trường nào có lịch sử giá cả toàn diện. Các tính năng phân tích kỹ thuật dựa trên việc kiểm tra thống kê thông tin về giá và khối lượng trong một khoảng thời gian nhất định (các khoảng thời gian khác nhau). Các chỉ báo phân tích kỹ thuật (TA) thường được sử dụng như các tính năng trong dự đoán ngoại hối và có nhiều bài báo có thể xác nhận hiệu quả của các chỉ số TA trong dự báo ngoại hối [4].

Các chỉ báo trong phân tích kỹ thuật được chia thành các chỉ báo mang những thuộc tính khác nhau như chỉ báo xu hướng, chỉ báo khối lượng, chỉ báo biến động, chỉ báo động lượng. Trong tài liệu, tổng số các thuộc tính phân tích kỹ thuật là 48, được tạo ra từ các tỷ giá và khối lượng giao dịch trong dữ liệu

3.3 Lựa chọn và trích xuất đặc trưng

3.3.1 NCA Neighborhood component analysis - Phân tích thành phần lân cận là một phương pháp học có giám sát phi tham số để phân loại dữ liệu đa biến thành các lớp khác nhau theo một thước đo khoảng cách nhất định trên dữ liệu. Phân tích các thành phần vùng lân cận nhằm mục đích "học" số liệu khoảng cách bằng cách tìm ra sự chuyển đổi tuyến tính của dữ liệu đầu vào sao cho hiệu suất phân loại leave-one-out (LOO) trung bình được tối đa hóa trong không gian đã biến đổi.

3.3.2 PCA Principal Component Analysis, (PCA hay Phân tích thành phần chính) là một thuật toán giảm chiều dữ liệu thuộc kỹ thuật trích xuất đặc trưng (Features extraction). PCA là phương pháp đi tìm một hệ tọa độ mới sao cho thông tin của dữ

Bảng 1: Thống kê số chỉ báo kỹ thuật sử dụng

Type	Indicator	Features	Total
Trend	EMA	2	10
	ADX	1	
	CCI	1	
	MACD	1	
	Ichimoku	4	
	PSAR	1	
Volume	OBV	1	9
	ADI	1	
	CMF	1	
	EMV	1	
	FI	1	
	MFI	1	
	NVI	1	
	VPT	1	
	VWAP	1	

Type	Indicator	Features	Total
Volatility	ATR	1	17
	Donchian	5	
	Keltner	5	
	ULI	1	
	Bolingerbands	5	
Momentum	RSI	1	12
	Oscillator	1	
	KAMA	1	
	PPO	1	
	PVO	1	
	StochRSI	3	
	TSI	1	
	Uoscillator	1	
	WilliamR	1	
	ROC	1	

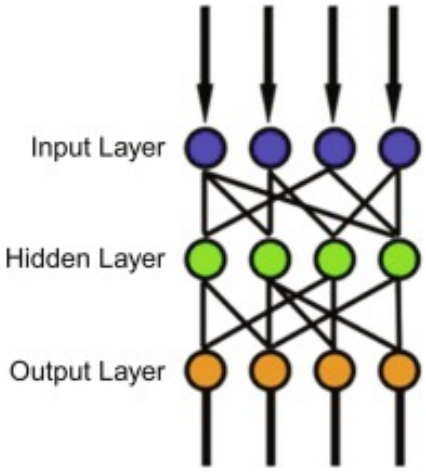
liệu chủ yếu tập trung ở một vài chiều và phần còn lại chỉ mang một lượng nhỏ thông tin. PCA có thể giúp đơn giản hóa dữ liệu và nhìn thấu được mối quan hệ giữa các biến độc lập, giữa các thuộc tính hay giữa các đặc trưng (các chiều)

3.3.3 BaggingTrees Rừng cây quyết định là một tập hợp các cây quyết định mà các dự đoán của chúng được kết hợp với nhau để đưa ra dự đoán tổng thể tạo thành một tập hợp bootstrap hoặc các cây đóng gói (Bagging Trees). Rừng cây quyết định tạo một số cây độc lập song song và chúng không tương tác với nhau cho đến khi tất cả chúng đã được xây dựng xong. Bagging Trees là công cụ tuyệt vời để lựa chọn tính năng. Đối với mỗi thuộc tính, sai số bình phương trung bình 'out-of-bag' khi loại bỏ thuộc tính đó được tính trung bình trên tất cả các cây. Điều này được lặp lại cho từng thuộc tính để đưa ra danh sách thuộc tính tốt nhất.

4 Phương pháp tiếp cận

4.1 Xây dựng mô hình máy học

4.1.1 Multi-layer Perceptron Perceptron nhiều lớp (MLP) là một phần bổ sung của mạng nơ-ron truyền thẳng. Nó bao gồm ba lớp — lớp đầu vào, lớp đầu ra và lớp ẩn



Hình 2: Ba lớp mạng nơ-ron truyền thẳng

Perceptron nhiều lớp (MLP) là một phần bổ sung của mạng nơ-ron truyền thẳng. Nó bao gồm ba lớp — lớp đầu vào, lớp đầu ra và lớp ẩn. Lớp đầu vào nhận tín hiệu đầu vào để xử lý. Tác vụ cần thiết như dự đoán và phân loại được thực hiện bởi lớp đầu ra. Một số lượng tùy ý các lớp ẩn được đặt giữa lớp đầu vào và đầu ra là công cụ tính toán thực sự của MLP. Tương tự như mạng nơ-ron truyền thẳng, trong MLP, dữ liệu truyền theo hướng thuận từ lớp đầu vào đến đầu ra. Các nơ-ron trong MLP được huấn luyện với thuật toán học truyền ngược. MLP được thiết kế để tính gần đúng bất kỳ hàm liên tục nào và có thể giải quyết các vấn đề không thể phân tách tuyến tính. Các trường hợp sử dụng chính của MLP là phân loại, nhận dạng, dự đoán và xấp xỉ mẫu.

4.1.2 Gaussian Processes Classifier Gaussian Processes Classifier là một thuật toán máy học phân loại.

Quy trình Gaussian là sự tổng quát hóa của phân phối xác suất Gauss và có thể được sử dụng làm cơ sở cho các thuật toán học máy phi tham số phức tạp để phân loại và hồi quy.

Chúng là một loại mô hình hạt nhân, giống như SVM nhưng chúng có khả năng dự đoán xác suất thành viên lớp được hiệu chỉnh cao, mặc dù sự lựa chọn và cấu hình của hạt nhân được sử dụng ở trung tâm của phương pháp có thể khó khăn.

4.1.3 Catboost Catboost xây dựng dựa trên cây quyết định được tăng cường gradient bao gồm tập dữ liệu đào tạo, với độ chính xác được xác định trên tập dữ liệu xác thực. Trong quá trình huấn luyện, những cây quyết định đó được xây dựng liên tiếp với mỗi cây được giảm bớt sự mất mát. Dựa trên các tham số bắt đầu của Catboost, lượng tử hóa được sử dụng cho các thuộc tính giá trị số khi xác định các cách tốt nhất để chia dữ liệu thành các nhóm.

Catboost có một số lợi ích hữu ích, với việc thực hiện dễ dàng. Một số tính năng chính của thư viện cạnh tranh này là ngay cả khi không điều chỉnh tham số, các tham

số mặc định vẫn mang lại kết quả tuyệt vời, các tính năng phân loại không cần xử lý trước, tính toán nhanh, tăng độ chính xác với ít trang bị. Lợi ích chính của thuật toán này là nó xử lý việc chuyển đổi tính năng phân loại theo cách tốt nhất khi so sánh với các thuật toán máy học khác.

5 Kết quả thực nghiệm

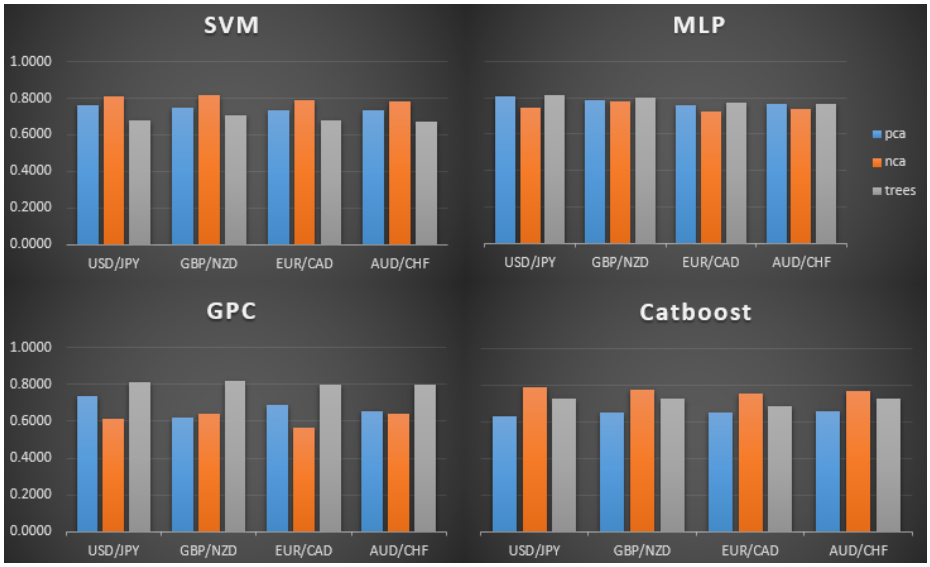
Bảng 2: Kết quả tối ưu sơ bộ

Features technique	Models	USD/JPY	GBP/NZD	EUR/CAD	AUD/CHF
PCA	MLP	**0.8059**	0.7880	0.7621	0.7670
	Catboost	0.7439	**0.7767**	0.7233	0.7362
	GPC	**0.8113**	0.7977	0.7751	0.7686
	svm	0.6334	0.6537	0.6505	**0.6570**
NCA	MLP	**0.7898**	0.7767	0.7573	0.7670
	Catboost	0.7251	**0.7265**	0.6861	0.7249
	GPC	**0.7358**	0.6197	0.6893	0.6570
	svm	0.6119	0.6408	0.5631	**0.6408**
Bagging Trees Selection	MLP	0.8167	**0.8220**	0.8026	0.8010
	Catboost	**0.7601**	0.7492	0.7330	0.7362
	GPC	0.8113	**0.8204**	0.7929	0.7799
	svm	0.6766	**0.7071**	0.6796	0.6699

Sau khi tiến hành chạy thực nghiệm ba phương pháp xử lý đặc trưng kết hợp bốn mô hình máy học với nhiều bộ tham số mô hình cũng như thay đổi số lượng đặc trưng được chọn, rút trích nhóm thu được nhiều kết quả độ chính xác đem lại một số thông tin nhất định về hiệu suất các mô hình và phương pháp trên bài toán. Kết quả sơ bộ với tham số cho những kết quả tốt nhất đối với mô hình và phương pháp xử lý được hiển thị ở bảng 2.

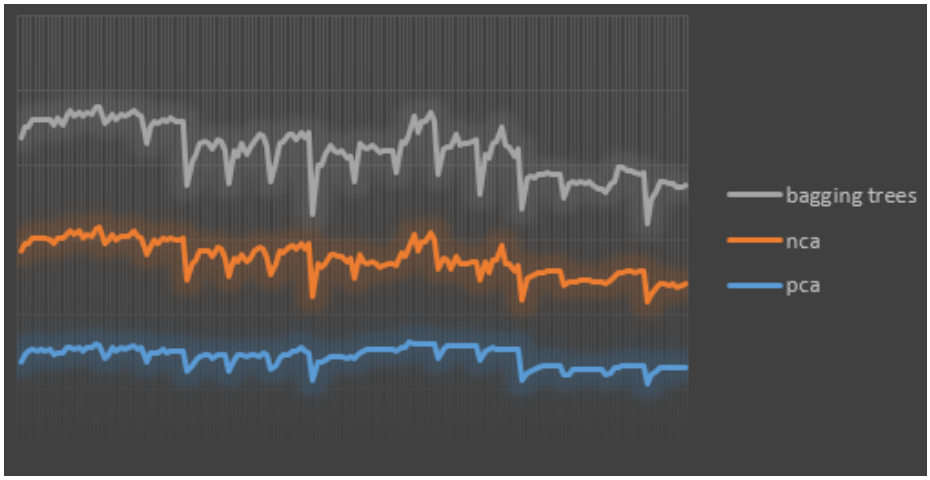
5.1 Kết quả trên phương pháp xử lý

Với kết quả hiệu suất trên các phương pháp thu được ở hình 3 ta có thể nhận thấy kết quả đạt được tốt hơn khi kết hợp phương pháp giảm chiều PCA, Bagging Trees Selection cho mô hình GPC và MLP thay vì NCA. Nhưng ngược lại với mô hình svm và Catboost thì phương pháp NCA lại cho kết quả tốt hơn



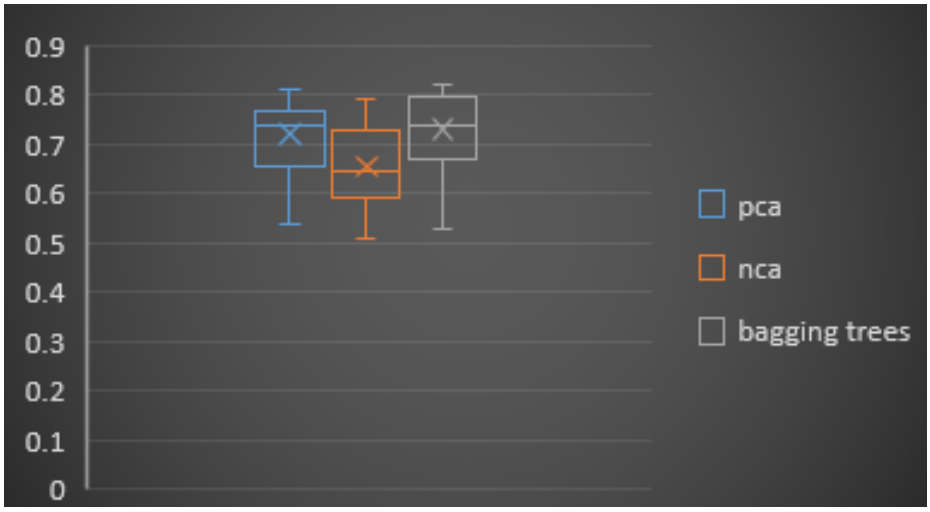
Hình 3: Kết quả hiệu suất phương pháp xử lý trên các mô hình

Nhìn chung trong các mô hình thì mô hình MLP cho thấy sự đồng đều và ổn định về hiệu suất dự đoán ở cả ba phương pháp xử lý và dữ liệu khác nhau, trong đó thì PCA và Bagging Trees Selection có hiệu suất nhỉnh hơn so với NCA.



Hình 4: Kết quả ổn định hiệu suất theo phương pháp xử lý

Hình 4 cho thấy mức độ dao động về hiệu suất của mô hình theo ba phương pháp xử lý khi thay đổi mô hình kết hợp, số lượng thuộc tính và các tham số đầu vào khác. Phương pháp giảm chiều dữ liệu PCA cho thấy được sự ổn định về hiệu suất tốt hơn so với hai phương pháp còn lại

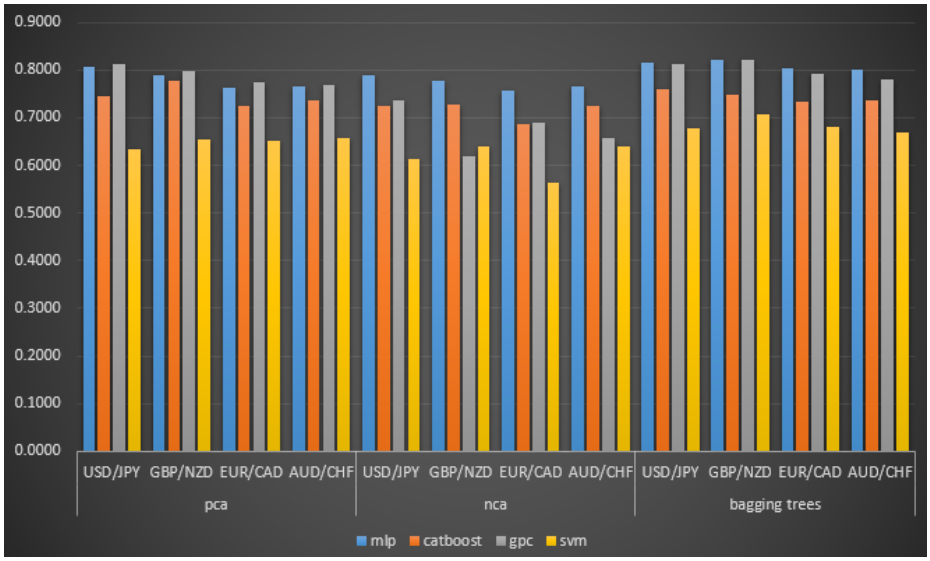


Hình 5: Kết quả boxplot theo phương pháp xử lý

Mặc dù có đạt được hiệu suất tối ưu khá cao nhưng Bagging Trees cho thấy mức độ phân bố hiệu suất không đồng đều trên các phương pháp (hình 5, hình 4), nên phương pháp lựa chọn đặc trưng này chỉ nên được sử dụng khi có được các thông tin để lựa chọn các đặc trưng phù hợp cũng như các tham số mô hình, số lượng đặc trưng. Trong hai phương pháp trích xuất đặc trưng thì PCA là phương pháp có hiệu suất phân bố tập trung ở mức cao hơn so với NCA.

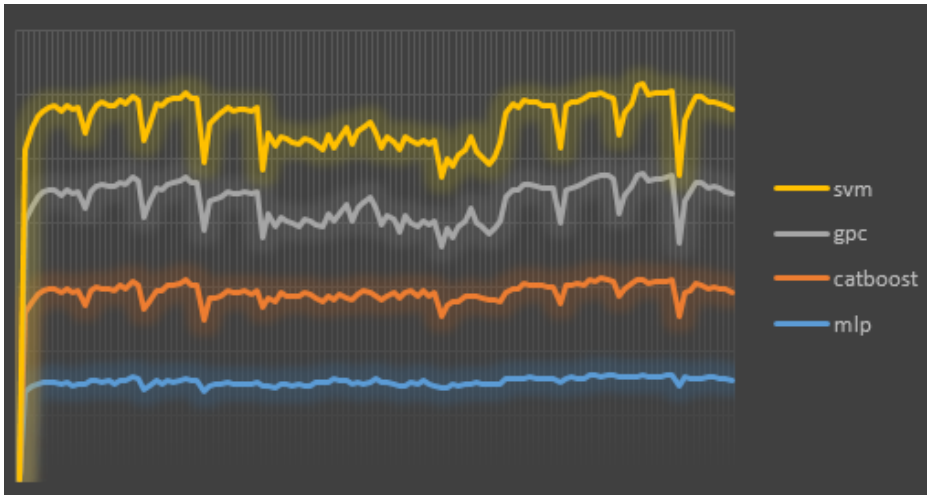
5.2 Kết quả trên mô hình dự đoán

Hình 6 cho thấy MLP và GPC thường có hiệu suất mô hình tối ưu tương đối trội hơn so với svm và catboos. Hiệu suất tối ưu của mỗi mô hình ở phương pháp lựa chọn đặc trưng thường cao hơn so với hai phương pháp trích xuất đặc trưng.



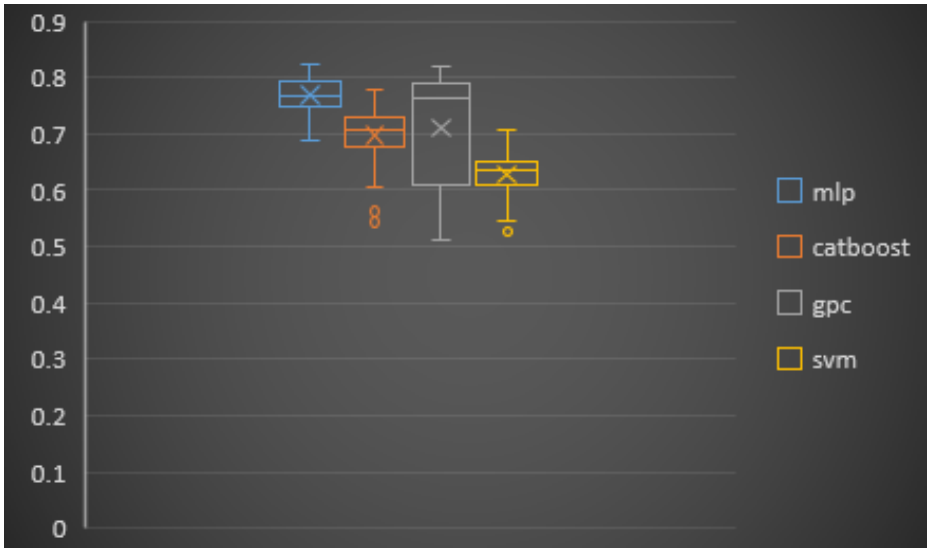
Hình 6: Kết quả hiệu suất mô hình theo phương pháp xử lý

Kết quả thu được cho thấy svm có kết quả tối ưu không được tốt như các mô hình khác và mức độ dao động về hiệu suất tương đối phức tạp khi thay đổi phương pháp, tham số. Ngược lại MLP là mô hình có hiệu suất tối ưu tương đối hơn so với ba mô hình còn lại(hình 7) và hiệu suất cũng dao động không nhiều trên nhiều kết quả thử nghiệm khác nhau



Hình 7: Kết quả ổn định hiệu suất mô hình

Kết quả hiệu suất tốt của MLP là tín hiệu đáng mừng, nhưng vẫn cần phải thử nghiệm và xác định lại độ tin cậy của mô hình này đối với bài toán. Bên cạnh đó Catboost và GPC cũng cho thấy các chỉ số hiệu suất ở mức tương đối nhưng mức độ phân tán còn phức tạp chưa được ổn định như MLP. Mặc dù svm cho thấy hiệu suất tập trung đều hơn trên các phương pháp nhưng kết quả hiệu suất chưa cao và vẫn thiếu ổn định



Hình 8: Kết quả boxplot hiệu suất

6 Tổng kết và hướng phát triển

Sau khi thực nghiệm kết hợp các phương pháp xử lý đặc trưng và mô hình với nhiều bộ tham số khác nhau nhóm thu được kết quả tổng quan cho thấy hiệu suất của các mô hình, kỹ thuật xử lý trên các bộ dữ liệu. Việc tiến hành phân tích kết quả mang lại một số thông tin nhất định cho việc sử dụng các phương pháp xử lý cũng như mô hình trên dữ liệu thị trường FoRex.

Kết quả độ chính xác nói chung đạt được nằm ở mức tương đối trong bài toán phân loại xu hướng thị trường FoRex. Song, vẫn còn nhiều thách thức về việc xác định rõ ràng được phương thức để giúp tăng hiệu quả mô hình ở bài toán nói trên.

Có nhiều hướng đi hay phương pháp mà nhóm có thể phát triển thêm/từ đề tài này như thay đổi cách phân loại nhãn huấn luyện theo một công thức khác để tạo thêm mức độ hiệu qua cho giao dịch. Thay đổi hoặc thêm nhiều khung thời gian dữ liệu hơn là một cách có thể giúp ích khai thác được nhiều đặc tính hơn của dữ liệu. Phát triển thuộc tính theo phân tích cơ bản, cập nhật thông tin từ các bản tin kinh tế hoặc thông tin đưa ra từ các cá nhân, tổ chức có ảnh hưởng để tạo các thuộc tính phân tích cơ bản cho dữ liệu. Phát triển và chỉnh sửa hệ thống để có thể áp dụng cả những thị trường khác có dữ liệu giống hoặc gần giống như dữ liệu thị trường cổ phiếu, tiền kỹ thuật số.

Tài liệu

1. Areej Abdullah Baasher, Mohamed Waleed Fakhr.: FoRex Trend Classification using Machine Learning Technique (2011).
2. Thuy Nguyen Thi Thuy, Vuong Dang Xuan.:FoRex TradingUsing Supervised Machine Learning (2018).
3. Ling Qi, Matloob Khushi, Josiah Poon.: Event-Driven LSTM For FoRex Price Prediction (2020).
4. Dong-xiao Niu, Bing-en Kou, Yun-yun Zhang.: Mid-long Term Load Forecasting Using Hidden Markov Model (2009).
5. Yiqi Zhao, Matloob Khushi.: Wavelet Denoised-ResNet CNN and LightGBM Method to Predict FoRex Rate of Change (2020).
6. Magdalena Daniela Nemes, Alexandru Butoi.: Data Mining on Romanian Stock Market Using Neural Networks for Price Prediction