SELECTION OF IMPORTANCE INDICATORS FOR MACHINE LEARNING MODELS IN FOREX TRADING AREA

Mai Van Hoan*, Dao Tran Chung, Vu Van Dien
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 22/4/2021
Revised: 21/5/2021
Published: 24/5/2021

KEYWORDS

Feature Selection
Machine Learning
Dimension reduction
Forex market
Random Uniform Forests

ABSTRACT

How to choose the best input variable for use in machine learning is the big question. In real life, the selection of indicators will help improve the results of forex market trend prediction, stock market based on machine learning models is always a topic of great interest to many scientists and investors. In this article, we focus on solving the problem of how to select the best indicators based on Random Uniform Forest. Our method consists of 3 steps: First, We collect data including indices commonly used in the forex sector; second, the data is standardized and labeled; finally, We use Random Uniform Forests to select indicators that are beneficial for prediction. Through the method done, In 17 common indicators in our interested domain, we found out 5 indicators (*vol, cci, adx, ar and chv*) are most important. We can explain why those indicators is beneficial for machine learning models, improving the model's performance, computation speed and reduced number of data dimensions.

LỰA CHỌN CÁC CHỈ SỐ QUAN TRỌNG CHO MÔ HÌNH HỌC MÁY ỨNG DỤNG TRONG GIAO DỊCH NGOẠI HỐI

Mai Văn Hoàn*, Đào Trần Chung, Vũ Văn Diện
Trường Đại học Công nghệ thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 22/4/2021
Ngày hoàn thiện: 21/5/2021
Ngày đăng: 24/5/2021

TỪ KHÓA

Lựa chọn thuộc tính
Học máy
Giảm chiều dữ liệu
Ngoại hối
Random Uniform Forests

TÓM TẮT

Lựa chọn các thuộc tính tốt cho các mô hình học máy cũng tương tự như việc lựa chọn các chỉ số tối ưu sẽ giúp ích cho việc nâng cao kết quả dự đoán xu hướng của thị trường ngoại hối, chứng khoán dựa trên các mô hình học máy luôn được các nhà đầu tư quan tâm. Trong bài báo này tập trung nghiên cứu giải quyết bài toán bằng cách nào có thể chọn lựa ra được những thuộc tính tốt nhất trong rất nhiều các thuộc tính ban đầu để phục vụ việc giảm chiều dữ liệu nâng cao tốc độ huấn luyện cho mô hình học máy. Phương pháp nhóm tác giả thực hiện gồm 3 bước chính: đầu tiên thu thập dữ liệu liên quan bao gồm các chỉ số được sử dụng phổ biến trong lĩnh vực ngoại hối; tiếp theo, dữ liệu được chuẩn hóa và gán nhãn; sau cùng, sử dụng thuật toán Random Uniform Forests với các thông tin về độ quan trọng của các thuộc tính để lựa chọn ra những chỉ số có lợi cho việc dự đoán. Kết quả nghiên cứu đã chỉ ra, trong 17 chỉ số thông dụng trong lĩnh vực nhóm đang quan tâm thì 05 chỉ số *(vol, cci, adx, ar và chv*) có ảnh hưởng nhất đến kết quả phân lớp dữ liệu có lợi cho mô hình học máy, cải thiện hiệu năng và tốc độ tính toán của các mô hình do số chiều dữ liệu được giảm xuống.

DOI: https://doi.org/10.34238/tnu-jst.4410

* Corresponding author. *Email: maihoan@ictu.edu.vn*

1. Giới thiệu

Trong vài năm trở lại đây, thị trường chứng khoán, ngoại hối nhận được sự quan tâm lớn của các nhà đầu tư. Đây là thị trường kỳ vọng mang lại lợi nhuận lớn cho các nhà đầu tư và cũng là thị trường có độ rủi ro rất cao và khó lường. Các nhà khoa học và các chuyên gia phân tích dữ liệu cũng không ngừng nghiên cứu nhằm áp dụng những kiến thức mới nổi về khai phá dữ liệu, xử lý dữ liệu lớn, các mô hình học máy, trí tuệ nhân tạo,... nhằm giải quyết các vấn đề đang được quan tâm của các nhà đầu tư. Một bài toán đang nhận được rất nhiều sự quan tâm là: Hàng ngày mỗi khi cần ra một quyết định mua, bán hay giữ một mã cổ phiếu, một cặp tiền tệ, các nhà đầu tư thường dựa vào một danh sách lớn các chỉ số thị trường và phân tích chúng. Câu hỏi đặt ra với họ là: trong các chỉ số này chỉ số nào thực sự có ý nghĩa và quan trọng trong việc ra quyết định của mình. Nhóm tác giả nhận thấy, việc lựa chọn các chỉ số có ý nghĩa trên cũng tương đồng với việc lựa chọn các thuộc tính đặc trưng của dữ liệu trong các mô hình máy học. Việc lựa chọn đúng các đặc trưng quan trọng sẽ giúp cho các mô hình học máy có kết quả chính xác hơn, giảm thời gian huấn luyện, giảm độ phức tạp của dữ liệu. Chính vì những lý do trên, nhóm nghiên cứu quyết định sử dụng việc lựa chọn các đặc trưng cho các mô hình máy học áp dụng cho việc chọn các chỉ số quan trọng nhằm giải quyết vấn đề chọn các chỉ số quan trọng trong lĩnh vực chứng khoán, ngoại hối. Trong phần này, nhóm tác giả giới thiệu sơ lược về thị trường ngoại hối, khai phá dữ liệu, lựa chọn thuộc tính và thuật toán Random Uniform Forests cùng các đặc trưng để giải quyết bài toán đặt ra.

Thị trường ngoại hối (Foreign Curency Exchange market – FOREX) là thị trường trao đổi tiền tệ với sự tham gia của trên 4.600 ngân hàng quốc tế và hàng triệu tổ chức, cá nhân nhỏ lẻ trên toàn thế giới [1]. Hàng ngày có đến trên 1,9 nghìn tỷ dollar giao dịch được ghi nhận vào năm 2016. Giao dịch cơ bản trong thị trường này là các giao dịch trao đổi ngoại tệ - mua một đồng này và bán một đồng khác. Lúc này, giá trị của một đồng tiền tệ được định giá thông qua việc so sánh với một đồng ngoại tệ khác thông qua tỉ giá. Một vài đồng ngoại tệ lớn trong thị trường gồm: EUR (euro), USD (United States dollar), JPY (Japanese yen), GBP (British pound),... Các đồng tiền tệ này sẽ được giao dịch thành các cặp như: EUR/USD, GBP/USD,... các nhà đầu tư sẽ quyết định mua hay bán một cặp tiền tệ nào đó với kỳ vọng sẽ mang về khoản lợi nhuận chênh lệch. Có hai trường phái cơ bản trong việc phân tích để đưa ra quyết định mua, bán hay giữ các cặp tiền tệ là phân tích kỹ thuật và phân tích cơ bản. Trong phần này, chúng tôi quan tâm đến phân tích kỹ thuật, đây là mô hình phân tích dựa trên các chỉ số và được xây dựng trên cơ sở những hiểu biết về giá nhằm giúp các nhà đầu tư đưa ra các quyết định. Các chỉ số được sử dụng phổ biến gồm: SAR, Bollinger Bands, MACD, Stochastic, RSI, MA, ADX,... Rất nhiều các chỉ số được giới thiệu, việc lựa chọn các chỉ số thích hợp cho việc phân tích của mình sẽ ảnh hưởng trực tiếp đến quyết định của các nhà đầu tư và lợi nhuận của họ. Chúng tôi nhận định rằng, việc lựa chọn các chỉ số này cũng tương tự như việc lựa chọn các thuộc tính đặc trưng quan trọng trong các mô hình học máy và kỳ vọng sẽ giúp giải quyết được vấn đề các nhà đầu tư đang gặp phải là làm sao chọn đúng những chỉ số quan trọng. Thông thường việc lựa chọn, kết hợp các chỉ số để đưa gia các quyết định giao dịch được thực hiện thủ công dựa trên kinh nghiệm của các nhà đầu tư. Những thông tin này thường là bí mật kinh doanh do vậy rất ít khi được công bố. Do vậy nhóm nghiên cứu cố gắng chỉ ra cách lựa chọn được những chỉ số có ảnh hưởng lớn đến kết quả đầu tư dựa trên đánh giá trực quan của mô hình máy học. Từ đó chứng minh việc chỉ ra các chỉ số tốt là dựa trên cơ sở khoa học không còn dựa trên cảm tính và kinh nghiệm cá nhân.

Như chúng ta đã biết các mô hình học máy là một phần của việc khai phá dữ liệu và phát hiện tri thức từ các tập dữ liệu lớn [2]. Việc khai phá tri thức từ dữ liệu lớn luôn gặp phải những khó khăn như: dữ liệu lớn thường được biết tới với hai vấn đề, đó là lớn về mặt số lượng dữ liệu và lớn về mặt số chiều dữ liệu [3]. Trong đó, số chiều dữ liệu được quan tâm hơn cả vì nó ảnh hưởng trực tiếp đến tốc độ huấn luyện và kết quả của các mô hình học máy. Do đó, việc giảm chiều dữ liệu thực sự nhận được rất nhiều sự quan tâm của cộng đồng nghiên cứu khoa học. Giảm

chiều dữ liệu có thể được thực hiện thông qua việc lựa chọn các thuộc tính hữu ích [4]. Việc lựa chọn thuộc tính được thực hiện bằng cách chọn ra một tập con thuộc tính từ các thuộc tính ban đầu sao cho kết quả của mô hình học máy xấp xỉ với tập thuộc tính ban đầu nhưng có tốc độ học nhanh hơn, tốn ít chi phí hơn. Thuật toán đánh giá vai trò của các thuộc tính được sử dụng phổ biến là Random Uniform Forests.

Random Uniform Forests (RUF) là thuật toán học có giám sát, có thể được sử dụng cho cả phân lớp và hồi quy [5]. RUF tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng.

Trong bài toán phân lớp nhị phân, giả sử chúng ta có tập huấn luyện $D_n = \{(X_i, Y_i), 1 \le i \le n\}$ với $Y \in \{0,1\}$, ta có cây quyết định dữ liệu thuộc về một nhãn lớp nào đó được viết như sau:

$$g_p(x, A, D_n) = g_p(x) = \begin{cases} 1, nếu \sum_{i=1}^{n} I_{\{X_i \in, Y_i = 1\}} > \sum_{i=1}^{n} I_{\{X_i \in A, Y_i = 0\}}, x \in A \\ 0, \quad trong\ các\ trường\ hợp\ khác \end{cases}$$

Ngoài ra, RUF cũng cung cấp cho chúng ta các thông tin liên quan đến tầm quan trọng của các thuộc tính trong tập dữ liệu huấn luyện, các thông tin bao gồm:

- ***Biến quan trọng toàn cục:*** Mức độ quan trọng của biến được đo lường trên tất cả các nút và tất cả các cây, dẫn đến tất cả các biến đều có một giá trị, vì các điểm cắt là ngẫu nhiên. Do đó, mỗi biến có cơ hội được chọn như nhau nhưng nó sẽ chỉ có tầm quan trọng nếu nó là biến làm giảm nhiều nhất entropy tại mỗi nút. Mức độ quan trọng của biến toàn cục tạo ra các biến làm giảm sai số dự đoán nhiều nhất nhưng nó không cho chúng ta biết gì về cách một biến quan trọng ảnh hưởng đến các phản hồi. Do vậy, RUF cung cấp cho chúng ta thông tin về độ quan trọng toàn cục của các biến, nó chỉ ra những biến quan trọng ảnh hưởng trực tiếp đến lỗi của mô hình dự đoán.
- ***Biến quan trọng cục bộ***: Thể hiện mức độ quan trọng của mỗi biến tới kết quả dự đoán tại mỗi nhánh.
- ***Biến quan trọng từng phần***: Thể hiện độ quan trọng của các biến tới từng lớp dữ liệu.

Việc sử dụng các điểm cắt ngẫu nhiên giúp cho RUF giảm tình trạng Overfitting, tạo được độ lệch thấp, phương sai thấp. Việc sử dụng thuật toán RUF với các thông tin về biến quan trọng toàn cục, quan trọng cục bộ, quan trọng từng phần sẽ giúp chúng ta thuận tiện trong việc chỉ ra những thuộc tính nào được cho là quan trọng và ảnh hưởng trực tiếp đến kết quả của các mô hình dự đoán.

Chọn lựa thuộc tính là một quá trình để tìm ra một tập con tốt nhất các thuộc tính theo một số tiêu chí nào đó. Việc lựa chọn các thuộc tính quan trọng dựa trên RUF giúp chúng ta có thể giải quyết được bài toán đặt ra ban đầu nhanh chóng và hiệu quả.

2. Phương pháp nghiên cứu

Để thực hiện tìm ra những chỉ số quan trọng ảnh hưởng đến chất lượng của một số mô hình học máy, cụ thể trong trường hợp cần nghiên cứu là phân loại dữ liệu. Phương pháp của nhóm thực hiện qua một số bước sau:

2.1. Thu thập dữ liệu và chuẩn hóa dữ liệu

Việc thu thập dữ liệu được thực hiện thông qua phần mềm MT4, dữ liệu thu thập gồm các thông tin về giá của các cặp ngoại tệ phổ biến, trên các khung thời gian khác nhau trong một khoảng thời gian cụ thể đủ lớn. Dữ liệu thô thu thập gồm các thông tin như giá mở cửa (Open), đóng cửa (Close), giá cao nhất trong phiên (High), giá thấp nhất trong phiên (Close).

2.2. Chuyển đổi dữ liệu

Dữ liệu sau khi thu thập sẽ được tiến hành làm sạch, loại bỏ các điểm thiếu dữ liệu. Từ dữ liệu thô ban đầu sẽ được sử dụng để xây dựng các chỉ số thường được các nhà đầu tư sử dụng để phân tích nhằm đưa ra các quyết định như: ADX, ATR, CCI, MACD, RSI,...

2.3. Đánh nhãn dữ liệu

Việc chúng ta dựa vào các chỉ số để thực hiện việc phân tích và đưa ra quyết định mua hoặc bán một cặp tiền tệ cũng tương tự việc phân loại thị trường thành hai loại: 1 tương ứng với quyết định mua và kỳ vọng thị trường đi lên; -1 với quyết định bán và kỳ vọng thị trường đi xuống. Như vậy, mô hình học máy của chúng ta định sử dụng là phân lớp với 2 nhãn lớp. Việc đánh nhãn lớp có thể thực hiện theo nhiều cách khác nhau. Cách đơn giản nhất chúng ta có thể sử dụng kỹ thuật One Step a Head, tức là dữ liệu phần nhãn của ngày hôm trước sẽ được đánh nhãn bằng kết quả thực tế của ngày hôm sau. Ví dụ, nhãn lớp của ngày hôm qua sẽ được gán nhãn là 1 nếu giá đóng cửa của phiên giao dịch ngày hôm nay cao hơn giá mở cửa và nhãn là -1 nếu giá đóng cửa của ngày hôm nay thấp hơn giá mở cửa của ngày hôm nay. Hoặc chúng ta sẽ sử dụng chỉ số Zigzag để đánh nhãn cho các lớp tương ứng, kỹ thuật này sẽ giúp ta đánh giá được nhãn lớp theo tính chu kỳ của dữ liệu. Tóm lại, chúng ta sẽ có 2 nhãn dữ liệu là -1 và 1 để phục vụ bài toán lựa chọn những thuộc tính quan trọng đối với bài toán phân lớp dữ liệu.

2.4. Đánh giá và đưa ra các chỉ số quan trọng

Dựa trên các các thông tin thu được trong quá trình học mô hình phân lớp dựa trên Random Uniform Forests, đặc biệt các thông tin liên quan đến thuộc tính quan trọng toàn cục, thuộc tính quan trọng cục bộ và thuộc tính quan trọng thành phần sẽ giúp chúng ta lựa chọn ra các chỉ số quan trọng ảnh hưởng đến mô hình phân lớp, những thuộc tính nào ảnh hưởng đến các lớp như thế nào.

Thông qua 5 bước trên sẽ giúp chúng ta giải quyết được bài toán đã đặt ra ban đầu là bằng cách nào có thể chọn lựa ra các thuộc tính có ảnh hưởng thực sự đến quyết định của các nhà đầu tư và giúp giảm bớt chiều dữ liệu từ đó giúp cho chúng ta có thể tăng tốc độ học và chất lượng của các mô hình học máy.

3. Kết quả

3.1. Dữ liệu gốc

Bảng 1. *Thông tin dữ liệu*

	Min,	1st Qu,	Median	Mean	3rd Qu	Max,	NA's
DX	*0,007527*	*10,18*	*21,64*	*24,73*	*36,96*	*80,91*	*16*
ADX	*7,207*	*16,22*	*22,45*	*24,65*	*31,24*	*56,44*	*31*
oscDX	*-70,37*	*-4,556*	*12,14*	*10,09*	*27,49*	*65,62*	*16*
ar	*-100*	*-50*	*0*	*1,048*	*56,25*	*100*	*16*
tr	*0*	*0,00016*	*0,00035*	*0,00047*	*0,00064*	*0,0048*	*1*
atr	*0,000141*	*0,000348*	*0,000446*	*0,00047*	*0,000569*	*0,00126*	*16*
cci	*-417,9*	*-84,85*	*6,332*	*2,624*	*88,79*	*408,8*	*15*
chv	*-0,6307*	*-0,3113*	*-0,01492*	*0,1006*	*0,4292*	*2,732*	*31*
cmo	*-98,39*	*-28,06*	*1,669*	*2,378*	*31,28*	*96,28*	*16*
sign	*-0,156*	*-0,03687*	*0,002712*	*0,000936*	*0,03982*	*0,1935*	*33*
vsig	*-0,2014*	*-0,02604*	*0,000525*	*0,000114*	*0,02722*	*0,1792*	*34*
rsi	*11,55*	*40,4*	*50,53*	*50,94*	*61,57*	*87,28*	*16*
slowD	*0,03721*	*0,2782*	*0,5225*	*0,5186*	*0,7609*	*0,9736*	*19*
oscK	*-0,4095*	*-0,04513*	*0,00233*	*3,50E-07*	*0,04829*	*0,36*	*17*
SMI	*-75,55*	*-25,96*	*2,755*	*3,482*	*34,31*	*80,05*	*25*
signal	*-72,27*	*-23,34*	*2,718*	*3,412*	*31,09*	*76,99*	*33*
vol	*0,002066*	*0,004582*	*0,006125*	*0,006313*	*0,007671*	*0,01432*	*16*

Chúng tôi sử dụng dữ liệu gồm 5.000 bản ghi (gồm giá mở cửa (Open), giá cao nhất (High), giá thấp nhất (Low) và giá đóng cửa (Close)) của cặp ngoại tệ EUR/USD tại khung thời gian M30 để phục vụ quá trình nghiên cứu.

Thông tin tổng hợp về dữ liệu các chỉ số được thể hiện ở bảng 1, bao gồm 17 chỉ số thông dụng trong lĩnh vực ngoại hối.

Dữ liệu phân loại gồm 2.438 có nhãn -1 (tương ứng với trạng thái thị trường đi xuống) và 2.526 nhãn 1 (tương ứng với thị trường đi lên).

3.2. Công cụ sử dụng

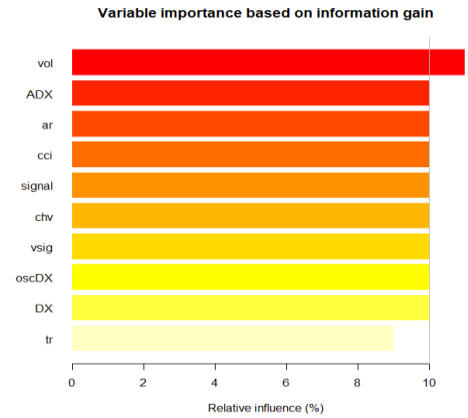
Việc thu thập dữ liệu được thực hiện trên phần mềm MT4. Xử lý và chuẩn hóa dữ liệu thực hiện trên R và thư viện *randomUniformForest* để thực thi mô hình phân lớp dữ liệu sử dụng thuật toán Random Uniform Forests.

3.3. Lựa chọn các thuộc tính quan trọng

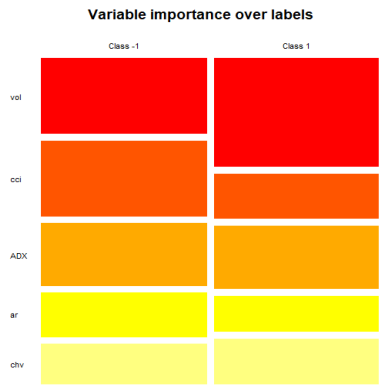
Để lựa chọn ra những thuộc tính quan trọng, có ảnh hưởng đến chất lượng phân lớp của mô hình phân lớp dữ liệu bằng cách: đầu tiên, sử dụng thuật toán Random Uniform Forests và lấy ra các thông tin liên quan đến các biến quan trọng toàn cục, biến quan trọng cục bộ và quan trọng thành phần. Tiếp theo, chúng tôi tìm hiểu cách các biến ảnh hưởng đến chỉ số lỗi của mô hình, xem xét mối quan hệ giữa các chỉ số. Cuối cùng, chúng tôi có được khi nào và làm thế nào biến có thể/ quan trọng bằng cách xem xét thông qua các tập dữ liệu training và test.

3.3.1. Các chỉ số quan trọng toàn cục

Chỉ số quan trọng toàn cục cho phép ta giảm tối đa lỗi dự đoán. Cụ thể nó cho chúng ta biết chỉ số nào có ảnh hưởng mạnh hơn đến việc phân lớp dữ liệu. Việc này được thực hiện bởi thuật toán Random Uniform Forests trong trình cắt tỉa ngẫu nhiên cây quyết định giúp tạo ra cơ hội ảnh hưởng của các chỉ số là như nhau lên kết quả phân loại. Hình 1 cho chúng ta thấy danh sách 10 chỉ số có ảnh hưởng toàn cục.



Hình 1. *Danh sách 10 chỉ số có ảnh hưởng toàn cục*



Hình 2. *Độ quan trọng của mỗi chỉ số trong mỗi lớp*

3.3.2. Tầm quan trọng của các chỉ số với các lớp

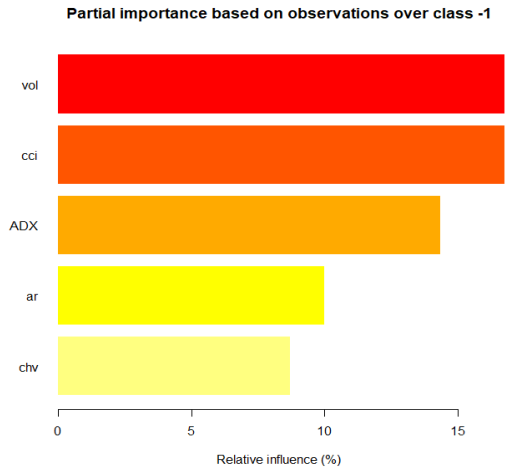
Các biến quan trọng cục bộ có vai trò chỉ ra ảnh hưởng của mình trong mỗi lớp. Hình 2 cho ta thấy được độ quan trọng của các chỉ số trong mỗi lớp.

3.3.3. Các chỉ số quan trọng của mỗi lớp

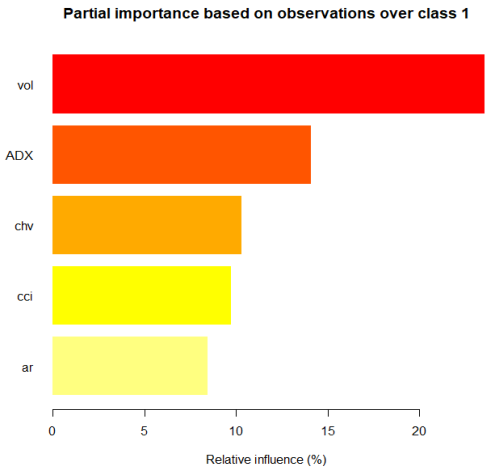
Việc tìm kiếm các chỉ số quan trọng toàn cục, quan trọng cục bộ và quan trọng từng phần trong Random Uniform Forests sẽ giúp chúng ta xác định được những chỉ số quan trọng đối với mỗi lớp dữ liệu, từ đó sẽ giúp chúng ta lựa chọn các chỉ số thực sự hiệu quả cho các mô hình máy học.

Như vậy với các thông tin thu được sau khi sử dụng thuật toán Random Uniform Forests, chúng ta đã tìm ra những chỉ số nào là quan trọng và tìm hiểu được sự ảnh hưởng của chúng lên mỗi lớp thông qua thuộc tính quan trọng toàn cục. Cùng đó là xem xét điều gì khiến nó có ảnh hưởng như

vậy và sự tương tác giữa các biến lên từng phần của cây quyết định thông qua các biến quan trọng cục bộ. Cùng với các thông tin liên quan đến độ quan trọng từng phần của các chỉ số thì chúng có thể thấy được những chỉ số ảnh hưởng lớn đến kết quả phân lớp dữ liệu cho lớp -1 thông qua hình 3 và hình 4 thể hiện những chỉ số quan trọng đối với việc phân lớp ở lớp 1.



Hình 3. *Các chỉ số quan trọng với nhãn lớp -1*



Hình 4. *Các chỉ số quan trọng với nhãn lớp 1*

4. Kết luận

Thông qua việc thực nghiệm trên 5.000 bản ghi dữ liệu bằng phương pháp kết hợp thuật toán Random Uniform Forests và các thông tin liên quan đến các biến quan trọng toàn cục, quan trọng cục bộ và quan trọng từng phần. Nghiên cứu đã chỉ ra được 05 chỉ số có ảnh hưởng nhất đến kết quả phân lớp dữ liệu, điều này giúp giải quyết được câu hỏi lớn đặt ra trong phần đầu của các nhà đầu tư. Hơn nữa, việc này đã giúp giảm chiều dữ liệu (từ 17 chiều xuống 05 chiều, tương ứng với giảm bớt 70% số chiều dữ liệu); từ đó giúp nâng cao tốc độ huấn luyện và chất lượng của mô hình học máy, cụ thể là mô hình phân lớp dữ liệu và hồi quy. Kết quả mà nhóm thực hiện được giúp mở ra cách thức lựa chọn các thuộc tính có ảnh hưởng thật sự tới tập dữ liệu bất kỳ, đều rất quan trọng đối với sự thành công của phát hiện tri thức và khai phá dữ liệu và đảm bảo rằng dữ liệu được cung cấp tốt về chất lượng và phù hợp về số lượng nhằm đưa ra các kết quả tin cậy và đầy đủ như người sử dụng mong muốn.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] M. Nassimi, Y. S. Asfaranjan, A. Keshvarsima, and F. Baradari, “Trading in the Foreign Exchange Market (Forex): A Study on Purchase Intention,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 4, no. 3, pp. 1-10, 2014.

[2] M. N. O. Sadiku, A. E. Shadare, and S. M. Musa, "Data mining: a brief introdcution," *European Scientific Journal*, vol. 11, no. 2, pp. 509-513, 2015.

[3] S. Velliangiria, S. Alagumuthukrishnanb, and S. I. T. Joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," *International Conference on Recent Trends in Advanced Computing (ICRTAC)*, 2019, pp. 104-111.

[4] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Advances in Bioinformatics*, vol. 2015, pp. 1-13, 2015.

[5] A. Pretorius, S. Bierman, and S. J. Steel, "A meta-analysis of research in random forests for classification," Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016, pp. 1-6.