**DỰ ĐOÁN NHỮNG GIAO DỊCH GIAN LẬN TRONG SÀN FOREX**

NGUYỄN TẤN HUY

Tháng 11, 2020

**1.Giới Thiệu**

**1.1 Lĩnh vực**  
Thị trường forex là thị trường mà người chơi sẽ đặt cược vào khả năng tăng hay giảm giá trị của một cặp tiền tệ được lưu hành trên thị trường để kiếm lời từ khả năng dự đoán của mình. Tại 1 thời điểm người chơi sẽ được đặt buy/sell một số lượng tiền tệ. Nếu người chơi buy, giá cặp tiền đó tăng thì người chơi có lời và ngược lại nếu sell và giá cặp tiền giảm thì người chơi có lời. Thị trường forex là một thì trường chơi khá lớn và có nhiều người chơi. Người chơi thông thường sẽ dựa vào thông tin thị trường, chính trị, kinh tế, hoặc sử dụng các loại phân tích tài chính để dự đoán giá tiền và từ đó kiếm lời từ từ dự đoán của chính mình.

Table

Description automatically generated

**Đây là hình ảnh một tài khoản có giao dịch thông thường**

**1.2 Vấn đề**Các sàn forex thường dùng các chương trình khuyến mãi để khuyến khích người chơi bằng cách tặng cho người chơi một khoảng tiền gọi là tiền thưởng (bonus). Người chơi được sử dụng tiền thưởng này để giao dịch và đặt lệnh nhưng không được rút ra. Nếu người chơi sử dụng tiền thưởng này để đặt lệnh và kiếm lời được từ lệnh này thì được phép rút ra bằng tiền mặt. Thông thường các sàn thường khuyến mãi theo phần trăm tiền nộp VD: Người chơi nộp tiền sẽ được tặng 100% tiền thưởng. Một số người chơi có thể lợi dụng tiền thưởng này để kiếm lời từ sàn giao dịch mà không cần phải phân tích và dự đoán. Khi 2 người chơi cùng mở 1 lệnh ở cùng một thời điểm nhưng ngược nhau BUY vs SELL thì một người sẽ lời và một người sẽ lỗ với số tiền tương ứng như nhau.  
A screenshot of a computer

Description automatically generated  
 **Hình ảnh một cặp giao dịch gian lận**  
  
Như hình ảnh cặp giao dịch gian lận ở trên thì nếu cả 2 tài khoảng cùng đóng lệnh cùng 1 lút. Thì 1 tài khoản sẽ lời 2.100.000 tiền thật và 1 tài khoản sẽ lỗ 700.000 tiền thật. Sàn giao dịch sẽ bị lỗ phần tiền thưởng. Hành vi này gây thiệt hại rất lớn cho sàn. Nên mọi sàn giao dịch forex đều nghiêm cấm hành vi này. Chúng ta có 2 loại giao dịch gian lận : 1. Gian lận trong cùng một sàn 2. Gian lận giữa 2 sàn khác nhau chúng ta sẽ không tìm được giao dịch đối xứng với giao dịch nghi ngờ. Việc xác định gian lận hay không dựa trên kinh nghiệm và cảm tính của người đánh giá. Sử dụng Machine Learning sẽ giúp giảm bớt khối lượng giao dịch phải đánh giá và giúp việc đánh giá chính xác hơn.

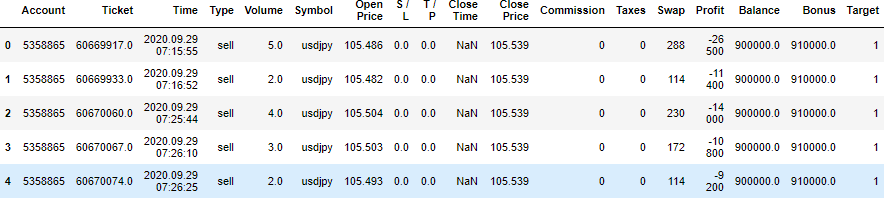
**1.3 Giải pháp**

1. **Giải pháp 1**: Modeling, dự đoán các giao dịch có khả năng gian lận dựa trên các thông tin có được.
2. **Giải pháp 2**: Xem xét các giao dịch giống nhau từ đó xác định các giao dịch gian lận.

**GIẢI PHÁP 1**

**2. Thu thập dữ liệu và khám phá dữ liệu**

**2.1 Nguồn** : sử dụng người dữ liệu từ sàn forex và các giao dịch bị đánh giá là gian lận trong quá khứ.

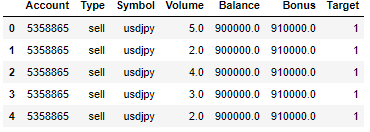


**Hình ảnh một phần của data**

|  |  |
| --- | --- |
| **Bảng mô tả sơ lược dataframe** | |
| Số lượng record | 88140 |
| Số lượng cột | 8 |
| Missing Values | 0 |
| Loại bỏ dữ liệu trùng lặp | YES |

**2.2 EDA**

|  |  |  |  |
| --- | --- | --- | --- |
| **Giới thiệu biến** | | | |
| Tên | Ý nghĩa | Loại biến | Giá trị |
| Type | Loại giao dịch | Biến phân loại | buy,sell |
| Time | Thời gian mở lệnh |  | Giá trị datetime |
| CloseTime | Thời gian đóng lệnh |  | Giá trị datetime |
| Symbol | Cặp tiền giao dịch | Biến phân loại | usdjpy,eurjpy,… |
| Volume | Khối lượng giao dịch | Biến liên tục | float |
| Profit | Lời lỗ hiện tại của gian dịch | Biến liên tục | Float |
| Balance | Tiền thật người chơi đã nộp | Biến liên tục | Integer |
| Bonus | Tiền thưởng người chơi được thưởng | Biến liên tục | Integer |
| Target | Biến Target | Biến phân loại | 0 -> bình thường 1 -> gian lận |
| **CloseTime, Profit** : rất quan trọng nhưng do việc đánh giá của chúng ta dựa trên hệ thống trực tuyến nên 2 biến này không thể lấy được vì lệnh vẫn chưa đóng. Chúng ta chỉ có thể đánh giá dựa trên các biến còn lại. Chúng ta sẽ loại bỏ 2 biến này và các biến khác không liên quan. | | | |



**Dataframe sau khi drop**

**2.3 Khám phá dữ liệu**

**2.3.1 Phân tích biến trên từng giao dịch**

**Volume**

**Shape, square

Description automatically generated  
Nhận xét: phần lớn giao dịch ở trong khoảng từ 0-1. Có một số giao dịch được có số volume rất lớn. Đây là một số giao dịch ở những cặp tiền hiếm gặp.**

Shape, rectangle

Description automatically generated

**Nhận xét: có rất nhiều outlier. Chúng ta không thể bỏ đi được bởi vì đây là những người chơi chuyên nghiệp và chuyên kiếm tiền từ sàn forex nên họ sẽ đặt cược với volume lớn. Các ngoại lệ này cũng chứa những giao dịch gian lận muốn kiếm tiền nhanh từ giao dịch gian lận nên thường đặt ở những volume lớn**

A picture containing chart

Description automatically generated

**Nhận xét: ở class gian lận thì volume thường đặt rất cao.Với 3.35+-4.5 trong khi đó các giao dịch bình thường chỉ giao động từ 0.04+-1.75**

Nhận xét biến volume: các giao dịch ở volume lớn có rất nhiều khả năng là giao dịch gian lận. Phần lớn người chơi chỉ giao dịch ở volume rất thấp. Có sự quan hệ rất lớn giữa Volume và giao dịch gian lận. Có khá nhiều ngoại lệ ở biến này nhưng chúng ta không thể loại bỏ chúng hay thay thế chúng được. Vì ngoại lệ này có liên quan rất nhiều đến khả năng gian lận của giao dịch. Các ngoại lệ của volume >= 0.61. Ta thấy có sự khác biệt rõ ràng giữa các class 0 và 1. Với mean,median của class 0 và class 1 khác nhau rất lớn.

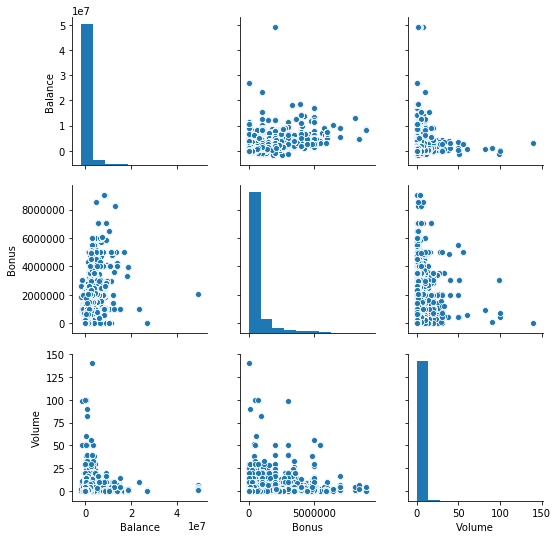
**Symbol**Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Nhận xét: các cặp tiền gian lận thì tập trung rất nhiều vào các cặp tiền chứa EUR, USD, GBP, JPY bởi vì đây là các đồng tiền thường có các biến động rất lớn. Dễ dàng kiếm lời từ các cặp tiền này.**

**Balance và Bonus**

****

**Nhận xét: không có sự tương quan trung bình giữa Balance và Bonus với độ tương quan là 0.58. Không có sự tương quan giữa Balance, Bonus với Volume. Điều này cho thấy không phải những người chơi có nhiều Balance và Bonus thì sẽ đặt số lượng volume lớn.**

**A picture containing chart

Description automatically generated**

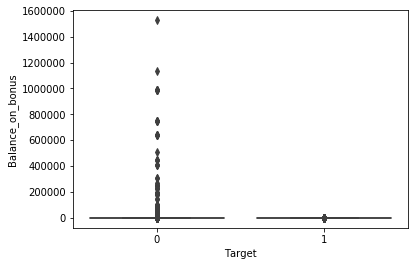
**A picture containing chart, box and whisker chart

Description automatically generated**

**Nhận xét: Đây là 2 biến quan trọng nhất trong mô hình của chúng ta. Sau khi phân tích ta thấy 2 biến này không có liên quan đến biến Volume. Dường như không có sự liên quan giữa số lượng tiền nộp và tiền thưởng trong số lượng Volume mà các account sẽ đặt. Quan sát 2 biến Bonus và Balance trên từng class thì ta thấy 2 biến trên đều nằm trong khoảng trung bình của 2 biến trên. Điều này không lạ, trong 1 sàn giao dịch forex thì số lượng đặt volume thường được giới hạn trong 1 khoảng nhất định theo từng loại tiền. VD: với EURJPY thì sàn giao dịch sẽ giới hạn chỉ được giao dịch với VOLUME 30 trong cùng 1 thời điểm. Vì vậy, các tất những giao dịch gian lận sẽ chỉ giao dịch trong 1 khoảng nhỏ vừa đủ để nhanh lấy được tiền lời từ giao dịch. Nếu các account gian lận nộp quá nhiều tiền và tiền thưởng thì sẽ rất lâu để có thể lấy được tiền lời. Theo kinh nghiệm, một giao dịch gian lận thường có 1 tỉ lệ nhất định giữa 2 biến balance và bonus. Thông thường 2 biến này có tỉ lệ là 1 hoặc 2. Tỉ lệ này tương ứng với tỉ lệ khuyến mãi của các sàn. Và thông thường các giao dịch gian lận sẽ xảy ra trong các lần khuyến mãi nên tỉ lệ này thường được không giao động quá lớn.**

**Tỉ lệ giữa Balance và Bonus**

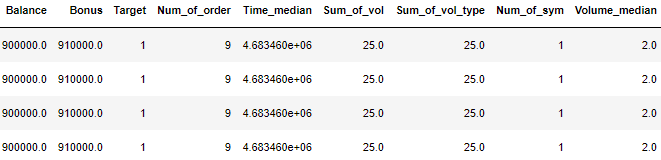
Theo quan sát thì tỉ lệ Bonus/Balance ở các giao dịch gian lận luôn là một hàng số cố định.



**Khuyến mãi :trên thực tế tỉ lệ này chính là tỉ lệ khuyến mãi của các sàn giao dịch trong 1 lần. Các giao dịch gian lận thông thường sẽ là tài khoản mới mở và nộp tiền trong đợt khuyến mãi để gian lận. Thông thường thì tỉ lệ này sẽ là 1. Tỉ lệ này ảnh hưởng rất nhiều đến model. Vì vậy, nếu tỉ lệ này được nâng lên là 2 hoặc 3 thì chúng ta phải thu thập thêm dữ liệu và cùng với đó là tạo ra 1 mô hình mới với tỉ lệ này. Thêm nữa nếu tỉ lệ này quá thấp thì số tiền kiếm được từ giao dịch gian lận sẽ thấp. Nên thông**

**2.3.2 Phân tích trên từng tài khoản**

Khi dự đoán, chúng ta sẽ dự đoán dựa trên tất cả giao dịch hiện tại trên từng tài khoản. Dựa trên kinh nghiệm đánh giá, chúng ta sẽ biến đổi các biến sau đây :



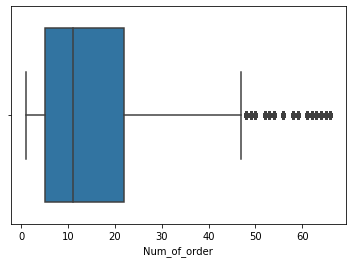
|  |  |  |  |
| --- | --- | --- | --- |
| **Giới thiệu biến** | | | |
| Tên | Ý nghĩa | Loại biến | Giá trị |
| Num\_of\_order | Số lượng giao dịch hiện có của 1 tài khoản | Biến liên tục | Giá trị integer |
| Sum\_of\_vol | Tổng số volume hiện đang giao dịch trong 1 tài khoản | Biến liên tục | Giá trị integer |
| Num\_of\_sym | tổng số cặp tiền được giao dịch hiện đang giao dịch trong 1 tài khoản | Biến liên tục | Giá trị integer |
| Volume\_median | Giá trị trung vị tất cả các giao dịch hiện tại trong 1 tài khoản | Biến liên tục | Giá trị float |

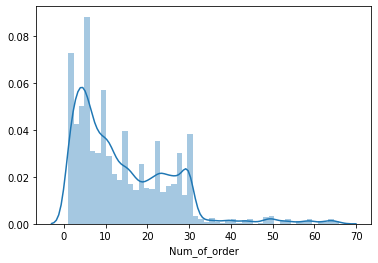
**Num\_of\_order**

Shape, square

Description automatically generated

Chúng ta có một số ngoại lệ ở đây với một số tài khoản mở hơn 66 (ngoại lệ trên) giao dịch 1 lần. Các tài khoảng có quá nhiều một lần không thể có khả năng gian lận được. Chúng ta sẽ loại bỏ các tài khoản này.





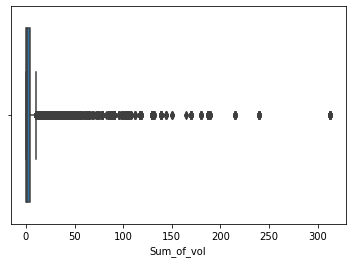
Ở đây ta thấy được các tài khoản gian lận có số lượng giao dịch mở khá ít so với các tài khoản bình thường.

Chart, box and whisker chart

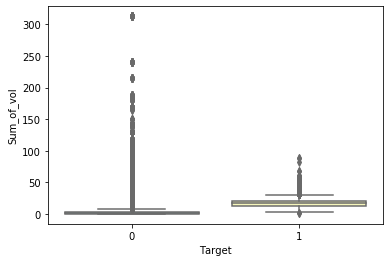
Description automatically generated

Khối lượng mở sẽ lớn hơn để đảm bảo chắc chắn có lời. Chúng ta sẽ loại bỏ tất cả các ngoại lệ ở biến này. Những người chơi với số lượng rất nhiều giao dịch trong 1 lần gần như không có khả năng gian lận và giao dịch ở VOLUME rất nhỏ có thể gây ảnh hưởng đến mô hình.

**Sum\_of\_vol**

****

**Rất nhiều ngoại lệ ở biến này. Tổng số volume trong 1 tài khoản đa phần nằm trong khoảng 1. Ta không thể loại bỏ ngoại lệ của biến này bởi vì các ngoại lệ này chính là các giao dịch gian lận, và một số người chơi chuyên nghiệp.**



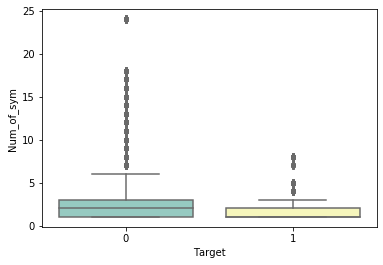
Khác biệt khá rõ ràng giữa class gian lận và class không gian lận. class không gian lận có tổng số volume giao dịch giao động trong khoảng 0.84+-15. Trong khi đó class gian lận giao động trong khoảng 18+-8.

**Num\_of\_sym**

A picture containing histogram

Description automatically generated

Chúng ta sẽ loại bỏ ngoại lệ ở biến này sau khi ở cả 2 class. Phần lớn là những giao dịch không có gian lận, ở tài khoản gian lận thì rất khó phân biệt được gian lận nếu không được xem xét qua một nhân viên chuyên nghiệp.



Theo kinh nghiệm, hầu hết các giao dịch gian lận chỉ giao dịch dưới 4 cặp tiền. Bởi vì giao dịch quá nhiều cặp tiền rất khó kiểm soát được lời lỗ khi một loại tiền lời và các loại tiền khác lỗ.

Biến Target

A picture containing shape

Description automatically generated

**2.3.3 Kết luận**

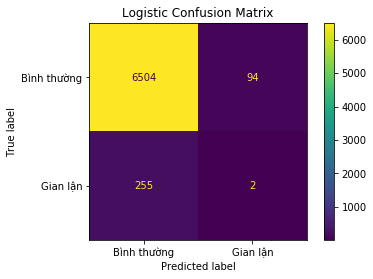
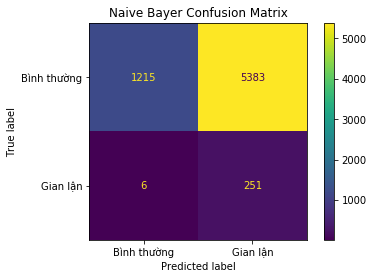
Các giao dịch gian lận thì sẽ thường mở ít giao dịch trong cùng 1 lần, với volume trung bình rất lớn trong khoảng từ 1-4, với balance và bonus nhỏ, và các cặp tiền thường được giao dịch có chứa EUR,GPB,USD,JPY là những cặp tiền có biến động tỉ giá trên profit lớn.

**2.4 Feature Engineering**

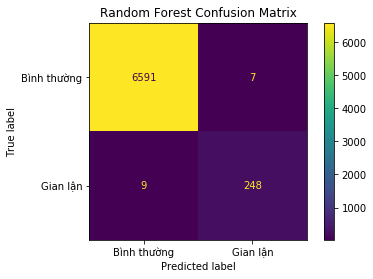
Biến đổi onehot encoder cho biến Symbol.

**3. Modeling**

Chúng ta sử dụng classification để dư đoán giao dịch gian lận. Ở đây ta có trường hợp unbalance dataset. Chúng ta sẽ sử dụng với tất cả các thuật toán bao gồm naïve bayer,logistic, decision tree và radom forest classification.

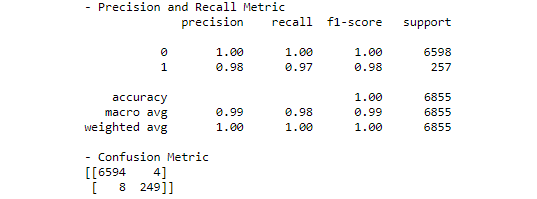






Thuật toán Random Forest cho ta kết quả tốt nhất. Xem xét các giao dịch thường bị lẫn lộn trong class 1 thì ta thấy được các giao dịch gian lận có số lượng number of order khá là cao với volume khá thấp chỉ bằng 1, 0.5. Rất dễ làm cho mô hình bị lẫn lộn với giao dịch thường. Các giao dịch bình thường bị lẫn lộn là gian lận thực chất là rất giống với giao dịch gian lận. Nên việc mô hình bị lẫn lộn là khá bình thường. Vì việc đánh giá gian lận hay không không chỉ phụ thuộc vào 1 giao dịch mà còn lịch sử các giao dịch trong tài khoản đó nữa. Và việc này rất cảm tính nên khó có thể chính xác 100%.

**3.1 Cải tiến mô hình dựa trên thuật toán RANDOM FOREST**

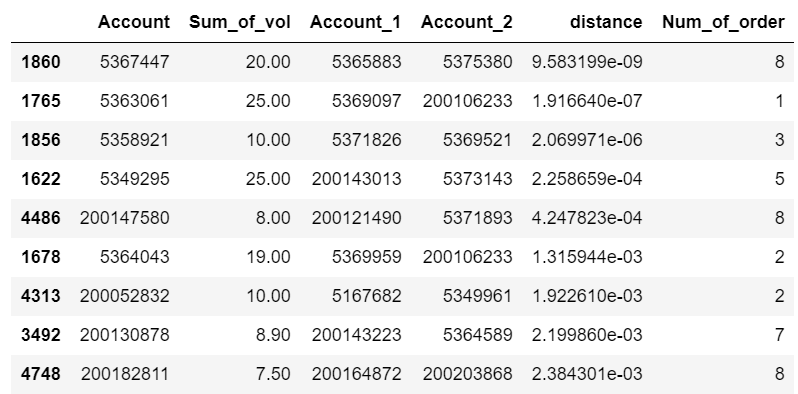
****

**Confusion matrix với adaboost**

Adaboost có giúp ta tăng Precision, Recall của class 1 lên 1%. Ta sẽ sử dụng Adaboost với Random forest.

**GIẢI PHÁP 2**

Chúng ta sẽ dựa trên độ sự giống nhau của các giao dịch hiện tại trong một tài khoản và thời gian trung bình của các giao dịch trong tài khoản đó. Thuật toán sử dụng NearestNeighbors của thư viện sklearn.



**Dataframe sau khi biến đổi.**

***Kết luận***: Giải pháp này giúp dò tìm được những cặp gian lận khó xác định và thường có thể bị bỏ qua bởi model. Nhưng có một rất nhiều account chơi rất nhỏ nên nên distance ở đây có thể cũng sẽ rất nhỏ hoặc bằng = 0 nhưng thực chất không phải là gian lận. Vì vậy người đánh giá cần lọc lại theo một số tiêu chí nhất định. Một số tiêu chí quan trọng là Sum\_of\_vol ta là tổng số volume giao dịch. Lấy theo các giao dịch gian lận thì min[Sum\_of\_vol]=2. Num\_of\_order <= 15. Sum\_of\_sym <=4. Ngoài ra, giải pháp này còn giúp hoàn thiện giải pháp 1 khi giúp xác định được cặp giao dịch đối xứng khi ta biết được Account ID. Giúp chúng ta đánh giá một cách chính xác giao dịch là gian lận trong cùng sàn hay giao dịch gian lận với sàn khác vốn là một công việc tốn khá nhiều thời gian.

**ĐÁNH GIÁ CẢ 2 GIẢI PHÁP VÀ KHẢ NĂNG ÁP DỤNG VÀO THỰC TẾ DOANH NGHIỆP**

|  |  |  |
| --- | --- | --- |
|  | ***Giải pháp 1 ( Modeling )*** | ***Giải pháp 2 ( Nearest Find )*** |
| **Ưu điểm** | 1. Chi phí tính toán thấp, có thể áp dụng cho hệ thống trực tuyến. 2. Có thể tìm ra các giao dịch gian lận bên ngoài sàn giao dịch. | 1. Có thể tìm được các giao dịch gian lận bất thường. Có thể nói là ngoại lệ mà giải pháp 1 không thể tìm được. |
| **Nhược điểm** | 1. Có thể không thể tìm ra các giao dịch gian lận bất thường. Ví dụ như những giao dịch gian lận với volume quá nhỏ.. thì mô hinh không thể tìm ra. 2. Mô hình phụ thuộc khá nhiều vào tỉ lệ giữa bonus và balance. Thông thường tỉ lệ là 1, ngược lại nếu tỉ lệ này là 2 thì mô hình không thể tìm được. | 1. Người sử dụng giải pháp 2 phải có kiến thức, kinh nghiệm về việc đánh giá gian lận để lọc các giao dịch theo điều kiện chính xác. 2. Qua khá nhiều bước để biến đổi và tính toán distance đến các tài khoản đối xứng. Không thể áp dụng cho hệ thống trực tuyến. 3. Phải có người thực hiện và điều chỉnh TOOL của giải pháp này. |
| **Khắc phục** | 1. Đối với tỉ lệ balance/bonus khác 1. Chúng ta có thể sử dụng giải pháp 2 trong 1 thời gian đầu để thu thập đủ dữ liệu cho tỉ lệ này và tạo ra 1 mô hình phù hợp với tỉ lệ này. | 1. Sử dụng spark để tính toán nhanh hơn. 2. Có thể cân nhắc việc tạo ra một mô hình dựa trên khoảng độ khác biệt này để có thể áp dụng cho mô hình trực tuyến. |
| **Khả năng áp dụng vào thực tế** | Cả 2 giải pháp trên đều áp dụng rất tốt trong hệ thống hiện tại. Có thể cân nhắc sử dụng spark và một số hệ thống khác để cải thiện tốc độ của hệ thống. | |