

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐIỆN – ĐIỆN TỬ**

**BỘ MÔN KỸ THUẬT MÁY TÍNH - VIỄN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**ĐỒ ÁN TỐT NGHIỆP  
 ĐỀ TÀI: DỰ ĐOÁN GIÁ ICO SỬ DỤNG**

**MACHINE LEARNING**

**NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH**

**GVHD: PGS.TS Võ Minh Huân**

Sinh viên: **Đào Minh Thuấn**

MSSV: 15119138

**Lê Thanh Tuấn**

MSSV: 15119152

TP.Hồ Chí Minh – tháng 01 năm 2020

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐIỆN – ĐIỆN TỬ**

**BỘ MÔN KỸ THUẬT MÁY TÍNH - VIỄN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**ĐỒ ÁN TỐT NGHIỆP  
 ĐỀ TÀI: DỰ ĐOÁN GIÁ ICO SỬ DỤNG**

**MACHINE LEARNING**

**NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH**

**GVHD: PGS.TS Võ Minh Huân**

Sinh viên: **Đào Minh Thuấn**

MSSV: 15119138

**Lê Thanh Tuấn**

MSSV: 15119152

TP.Hồ Chí Minh – tháng 01 năm 2020

**THÔNG TIN KHÓA LUẬN TỐT NGHIỆP**

1. **Thông tin sinh viên**

|  |  |
| --- | --- |
| Họ và tên sinh viên: Đào Minh Thuấn  Email: 15119138@student.hcmute.edu.vn | MSSV: 15119138  Điện thoại: 0394839208 |
| Họ và tên sinh viên: Lê Thanh Tuấn  Email: 15119152@student.hcmute.edu.vn | MSSV: 15119152  Điện thoại: 0903835225 |

1. **Thông tin đề tài**
   * Tên của đề tài: Dự đoán giá ICO sử dụng Machine Learning.
   * Đơn vị quản lý: Bộ môn Kỹ Thuật Máy Tính - Viễn Thông, Khoa Điện Điện Tử, Trường Đại Học Sư Phạm Kỹ Thuật Tp. Hồ Chí Minh.
   * Thời gian thực hiện: Từ ngày 09/09/2019 - 27/12/2019
   * Thời gian bảo vệ trước hội đồng: Ngày 9/1/2020
2. **Lời cam đoan của sinh viên**

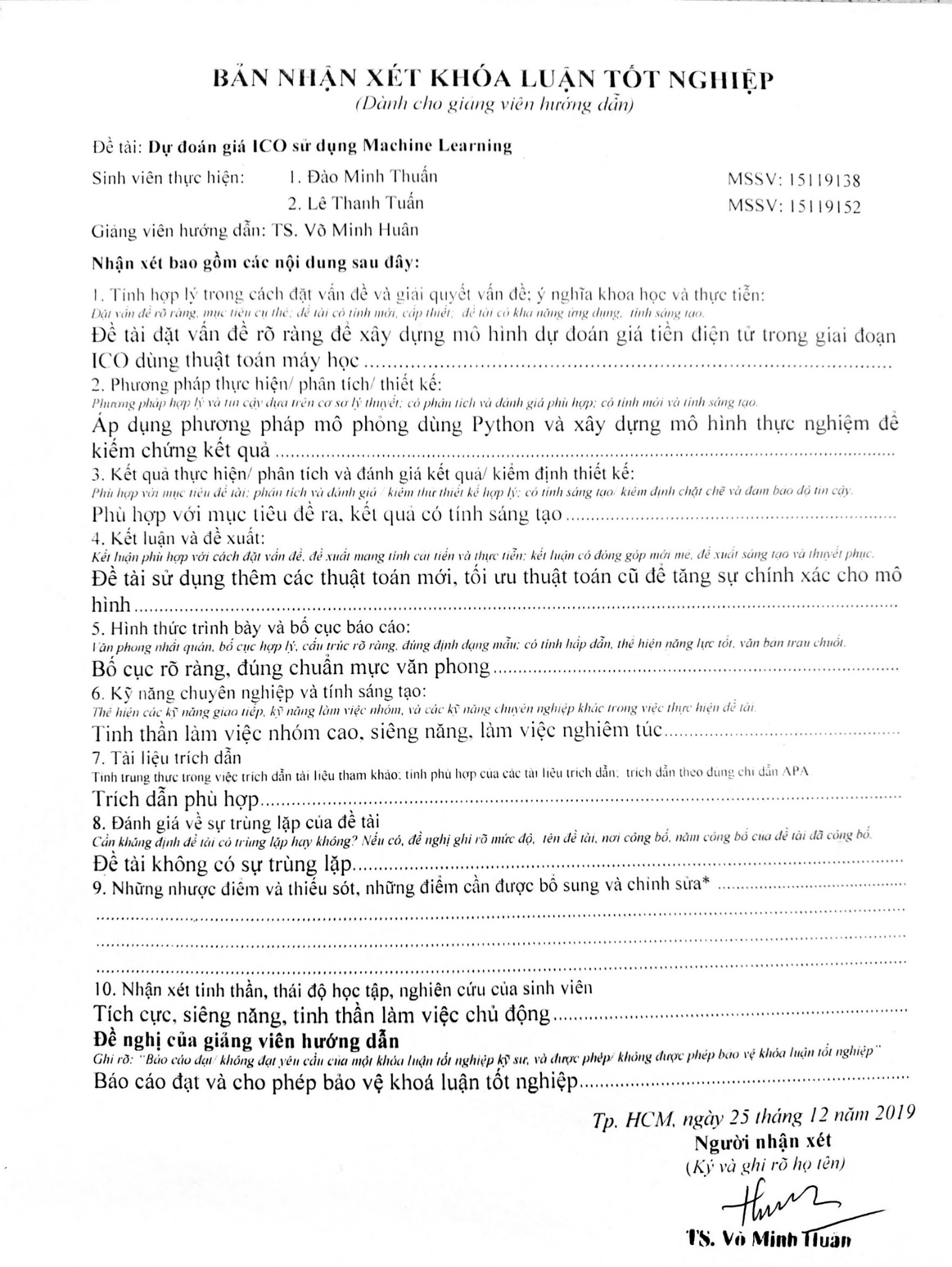
Chúng tôi – Đào Minh Thuấn, Lê Thanh Tuấn, cam đoan Khóa Luận Tốt Nghiệp là công trình nghiên cứu của chúng tôi dưới sự hướng dẫn của Phó giáo sư - Tiến sĩ Võ Minh Huân. Kết quả công bố trong Khóa Luận Tốt Nghiệp là trung thực và không sao chép từ bất kỳ công trình nghiên cứu nào khác.

|  |  |
| --- | --- |
|  | *Tp.HCM, ngày … tháng … năm 20…*  SV thực hiện đồ án  (*Ký và ghi rõ họ tên*) |

Giảng viên hướng dẫn xác nhận quyển báo cáo đã được chỉnh sửa theo đề nghị được ghi trong biên bản của Hội đồng đánh giá Khóa luận tốt nghiệp.

………………………………………………………………………………………………

|  |  |
| --- | --- |
| Xác nhận của Bộ Môn | *Tp.HCM, ngày … tháng … năm 20…*  Giáo viên hướng dẫn  *(Ký, ghi rõ họ tên và học hàm - học vị)* |

****

**BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP**

*(Dùng cho giảng viên phản biện)*

Đề tài: **Dự đoán giá ICO sử dụng Machine Learning.**

Sinh viên thực hiện: 1. Đào Minh Thuấn MSSV: 15119138

2. Lê Thanh Tuấn MSSV: 15119152

Giảng viên hướng dẫn: Ts. Võ Minh Huân

**Nhận xét bao gồm các nội dung sau đây:**

1. Tính hợp lý trong cách đặt vấn đề và giải quyết vấn đề; ý nghĩa khoa học và thực tiễn [15/100]:

*Đặt vấn đề rõ ràng, mục tiêu cụ thể [5]; đề tài có tính mới, cấp thiết [5]; đề tài có khả năng ứng dụng, tính sáng tạo [5].*

2. Phương pháp thực hiện/ phân tích/ thiết kế [25/100]:

*Phương pháp hợp lý và tin cậy dựa trên cơ sở lý thuyết [10]; có phân tích và đánh giá phù hợp [10]; có tính mới và tính sáng tạo [5].*

3. Kết quả thực hiện/ phân tích và đánh giá kết quả/ kiểm định thiết kế [25/100]:

*Phù hợp với mục tiêu [10]; phân tích và đánh giá / kiểm thử thiết kế hợp lý [10]; có tính sáng tạo/ kiểm định chặt chẽ và đảm bảo độ tin cậy [5].*

4. Kết luận và đề xuất [10/100]:

*Kết luận phù hợp với cách đặt vấn đề, đề xuất mang tính cải tiến và thực tiễn [5]; kết luận có đóng góp mới mẻ, đề xuất sáng tạo và thuyết phục[5].*

5. Hình thức trình bày, bố cục và chất lượng báo cáo [15/100]:

*Văn phong nhất quán, bố cục hợp lý, cấu trúc rõ ràng, đúng định dạng mẫu [5]; có tính hấp dẫn, thể hiện năng lực tốt, văn bản trau chuốt [15].*

6. Tài liệu trích dẫn [10/100]

Tính trung thực trong việc trích dẫn tài liệu tham khảo; tính phù hợp của các tài liệu trích dẫn; trích dẫn theo đúng chỉ dẫn APA.

7. Đánh giá về sự trùng lặp của đề tài

*Cần khẳng định đề tài có trùng lặp hay không? Nếu có, đề nghị ghi rõ mức độ, tên đề tài, nơi công bố, năm công bố của đề tài đã công bố.*

8.Những nhược điểm và thiếu sót, những điểm cần được bổ sung và chỉnh sửa\*

**Câu hỏi sinh viên phải trả lời trước hội đồng\*** *(ít nhất 02 câu)*

**Đánh giá chung**

* Điểm (Quy về thang điểm 10 không làm tròn): …………. /10.
* Xếp loại chung (Xuất sắc, Giỏi, Khá, Trung bình, Yếu, Kém): …………………………………….

**Đề nghị của giảng viên phản biện**

*Ghi rõ: “Báo cáo đạt/ không đạt yêu cầu của một khóa luận tốt nghiệp kỹ sư, và được phép/ không được phép bảo vệ khóa luận tốt nghiệp”*

*Tp. HCM, ngày … tháng …. năm 20…*

**Người nhận xét**

(*Ký và ghi rõ họ tên)*

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm thực hiện đề tài xin trân trọng cảm ơn GVHD PGS.TS Võ Minh Huân đã tận tình chỉ dẫn cho nhóm các bước thực hiện đồ án này đúng tiến độ và đạt được thành quả. Trong quá trình làm việc với thầy nhóm không những tiếp thu những kiến thức bổ ích, mà còn rèn luyện được thêm nhiều kĩ năng hay, thái độ nghiêm túc trong quá trình nghiên cứu khoa học, điều này là rất cần thiết trong quá trình học tập và đặc biệt hơn nó sẽ là hành trang lớn cho con đường sự nghiệp sau này.

Đồng thời nhóm xin chân thành cảm ơn các giảng viên khoa đã truyền dạy những kiến thức cần thiết cho ngành.

Cuối cùng, nhóm xin chúc Quý Thầy, Cô của trường Đại học Sư Phạm Kỹ Thuật TPHCM dồi dào sức khỏe và đạt được nhiều thành công trong công việc.

Nhóm xin chân thành cảm ơn!

*Tp. Hồ Chí Minh, tháng 01 năm 2020*

Sinh viên thực hiện

Đào Minh Thuấn

Lê Thanh Tuấn

**TÓM TẮT**

Từ khi tiền điện tử ra đời, nó đã tạo nên một sức hút lớn trong thị trường giao dịch trên thế giới. Với việc biết được giá trị và có thể dự đoán sự thay đổi của nó trong tương lai chính là một lợi thế cho những người sử dụng chúng. Có rất nhiều hệ thống dự đoán với các phương pháp khác nhau. Mỗi phương pháp đều có những ưu điểm và nhược điểm riêng. Việc ứng dụng phương pháp nào là tùy thuộc dữ liệu thu thập được, yêu cầu về cấu hình phần cứng cũng như mục đích của người xây dựng hệ thống.

Áp dụng thuật toán máy học để dự đoán giá trị chính là một bước đột phá lớn trong công nghệ. Trong quá trình thực hiện, người ta đã có thể dự đoán được sự biến động về giá của đồng tiền điện tử trong một thời gian ngắn.

Ứng dụng các thuật toán máy học là một trong những phương pháp mới, có nhiều ưu thế vượt trội trong việc giải quyết các vấn đề mà các phương pháp thông thường không giải quyết được. Các thuật toán máy học mang tính chất khách quan và đáp ứng được nhiều yêu cầu của thực tiễn. Với những ưu thế vượt trội ở trên như tính linh hoạt, giải quyết được các vấn đề khó khắn trong việc xử lý dữ liệu có tính biến động lớn. Đó là những lý do để nhóm chọn đề tài: Dự đoán giá ICO sử dụng Machine Learning.

Do thời gian nghiên cứu không nhiều và hiểu biết còn hạn chế dẫn đến việc trình bày các vấn đề không tránh khỏi thiếu sót. Rất mong nhận được những ý kiến đóng góp của quý Thầy, Cô cũng như các bạn để nhóm có thể hoàn thiện đề tài của mình.

# MỤC LỤC

[MỤC LỤC ix](#_Toc29344633)

[DANH MỤC HÌNH ẢNH xi](#_Toc29344634)

[DANH MỤC BẢNG xvi](#_Toc29344635)

[DANH MỤC CÁC TỪ VIẾT TẮT xvii](#_Toc29344636)

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc29344637)

[1.1. GIỚI THIỆU ĐỀ TÀI 1](#_Toc29344638)

[1.2. TÍNH CẤP THIẾT CỦA ĐỀ TÀI 1](#_Toc29344639)

[1.3. MỤC TIÊU CỦA ĐỀ TÀI 1](#_Toc29344640)

[1.4. ĐỐI TƯỢNG, PHẠM VI NGHIÊN CỨU 2](#_Toc29344641)

[1.4.1. Đối tượng nghiên cứu 2](#_Toc29344642)

[1.4.2. Phạm vi nghiên cứu 2](#_Toc29344643)

[1.5. PHƯƠNG PHÁP NGHIÊN CỨU 2](#_Toc29344644)

[1.6. BỐ CỤC CỦA ĐỀ TÀI 2](#_Toc29344645)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4](#_Toc29344646)

[2.1. KHÁI NIỆM VỀ INITIAL COIN OFFERING. 4](#_Toc29344647)

[2.2. KHÁI NIỆM VỀ TOKEN 5](#_Toc29344649)

[2.3. KHÁI NIỆM VỀ BLOCKCHAIN 6](#_Toc29344650)

[2.3.1. Ưu điểm blockchain 6](#_Toc29344651)

[2.3.2. Nhược điểm blockchain 7](#_Toc29344652)

[2.4. HỆ SỐ TƯƠNG QUAN 7](#_Toc29344653)

[2.5. HIỆN TƯỢNG ĐA CỘNG TUYẾN 8](#_Toc29344656)

[2.6. MÔ HÌNH THUẬT TOÁN HỒI QUY RIDGE 9](#_Toc29344657)

[2.6.1. Bài toán hồi quy đa biến 9](#_Toc29344658)

[2.6.2. Hiện tượng quá khớp (Overfitting) 10](#_Toc29344659)

[2.6.2.1 Early stopping 11](#_Toc29344660)

[2.6.2.2. Thêm số hạng vào hàm mất mát 12](#_Toc29344662)

[2.6.3. Hồi quy Ridge 13](#_Toc29344663)

[2.7. MÔ HÌNH MẠNG NƠRON 14](#_Toc29344664)

[2.7.1. Lan truyền tiến (Feedforward) 16](#_Toc29344668)

[2.7.2. Lan truyền ngược (Backpropagation) 17](#_Toc29344670)

[2.7.3. Activation Function 17](#_Toc29344672)

[CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ 21](#_Toc29344677)

[3.1. HÌNH THÀNH Ý TƯỞNG 21](#_Toc29344678)

[3.2. MÔ HÌNH HỆ THỐNG 22](#_Toc29344680)

[3.3. QUY TRÌNH THU THẬP DỮ LIỆU 22](#_Toc29344682)

[3.4. PHÂN TÍCH TẬP DỮ LIỆU 30](#_Toc29344694)

[3.4.1. Phân tích tương quan 30](#_Toc29344695)

[3.4.1.1. Phân tích tương quan giữa các input và output 30](#_Toc29344696)

[3.4.2. Đa cộng tuyến 37](#_Toc29344710)

[3.4.2.1. Các biến gần như độc lập với nhau 39](#_Toc29344717)

[3.4.2.2. Các biến có tương quan với nhau 46](#_Toc29344726)

[3.5.1. Hồi quy Ridge 51](#_Toc29344738)

[3.5.2. Mạng Nơron 54](#_Toc29344740)

[CHƯƠNG 4: KIỂM TRA THỰC NGHIỆM VÀ ĐÁNH GIÁ 56](#_Toc29344742)

[4.1. OVERFITTING 56](#_Toc29344743)

[4.2. THỰC NGHIỆM THUẬT TOÁN 58](#_Toc29344752)

[4.2.1. Mô hình hồi quy Ridge 59](#_Toc29344754)

[4.2.2. Mô hình mạng Nơron 65](#_Toc29344769)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 74](#_Toc29344786)

[5.1. KẾT LUẬN 74](#_Toc29344787)

[5.1.1. Những vấn đều nghiên cứu 74](#_Toc29344788)

[5.1.2. Những vấn đề hoàn thành 74](#_Toc29344789)

[5.2. HẠN CHẾ 74](#_Toc29344790)

[5.3. HƯỚNG PHÁT TRIỂN 74](#_Toc29344791)

[TÀI LIỆU THAM KHẢO 75](#_Toc29344792)

# DANH MỤC HÌNH ẢNH

[*Hình 2.1: Tổng số tiền mà hình thức ICO tạo ra.1* 5](#_Toc29535277)

[*Hình 2.2: Chiều tương quan.2* 7](#_Toc29535283)

[*Hình 2.3: Thể hiện độ tương quan.* 8](#_Toc29535284)

[*Hình 2.4: Early stopping.3* 12](#_Toc29535290)

[*Hình 2.5: Mạng nơron tự nhiên.4* 14](#_Toc29535294)

[*Hình 2.6: Mạng nơron nhân tạo.* 14](#_Toc29535295)

[*Hình 2.7: Mô hình mạng Nơron đa lớp 5* 15](#_Toc29535296)

[*Hình 2.8: Lan truyền tiến.* 16](#_Toc29535298)

[*Hình 2.9 Lan truyền ngược.* 17](#_Toc29535300)

[*Hình 2.10: Đồ thị kích hoạt tuyến tính.* 18](#_Toc29535302)

[*Hình 2.11: Đồ thị hàm Sigmoid.* 19](#_Toc29535303)

[*Hình 2.12: Đồ thị hàm tanh.* 19](#_Toc29535304)

[*Hình 2.13: Đồ thị hàm Relu.* 20](#_Toc29535305)

[*Hình 3.1: Sơ đồ tư duy hệ thống.* 21](#_Toc29535308)

[*Hình 3.2: Mô hình hệ thống.* 22](#_Toc29535310)

[*Hình 3.3: Quy trình thu thập dữ liệu.* 23](#_Toc29535312)

[*Hình 3.4: Rest API design.* 24](#_Toc29535313)

[*Hình 3.5: Ứng dụng thu thập dữ liệu.* 24](#_Toc29535314)

[*Hình 3.6: File docker-compose.yml* 25](#_Toc29535315)

[*Hình 3.7: Giao diện Command Prompt* 26](#_Toc29535316)

[*Hình 3.8: Thông báo khi chạy Docker.* 26](#_Toc29535317)

[*Hình 3.9: Giao diện cấu hình truy cập cơ sở dữ liệu.* 26](#_Toc29535318)

[*Hình 3.11: Cơ sở dữ liệu ICODb* 27](#_Toc29535319)

[*Hình 3.12: Dữ liệu không hợp lệ.* 28](#_Toc29535320)

[*Hình 3.13: Dữ liệu bị mất.* 28](#_Toc29535321)

[*Hình 3.14: Tập dữ liệu hoàn chỉnh.* 29](#_Toc29535322)

[*Hình 3.15: Tập dữ liệu sau khi được mã hóa One hot Encoding.* 29](#_Toc29535323)

[*Hình 3.16: Ví dụ về tập dữ liệu.* 30](#_Toc29535324)

[*Hình 3.17: Sơ đồ phán tán giữa giá usd và giá trị ngõ ra.* 31](#_Toc29535329)

[*Hình 3.18: Sơ đồ phân tán giữa giá bitcoin và giá trị ngõ ra* 31](#_Toc29535330)

[*Hình 3.19: Sơ đồ phân tán giữa biến total\_supply và giá trị ngõ ra* 32](#_Toc29535331)

[*Hình 3.20: Sơ đồ phân tán giữa biến market\_cap\_usd và giá trị ngõ ra* 32](#_Toc29535332)

[*Hình 3.21: Sơ đồ phân tán giữa biến available\_supply giá trị ngõ ra* 33](#_Toc29535333)

[*Hình 3.22: Sơ đồ phân tán giữa biến usd\_raised và giá trị ngõ ra* 33](#_Toc29535334)

[*Hình 3.23: Sơ đồ phân tán giữa biến eth\_price\_launch và giá trị ngõ ra.* 34](#_Toc29535335)

[*Hình 3.24: Sơ đồ phân tán giữa biến btc\_price\_launch và giá trị ngõ ra.* 34](#_Toc29535336)

[*Hình 3.25: Sơ đồ phân tán giữa biến month và giá trị ngõ ra.* 35](#_Toc29535337)

[*Hình 3.26: Sơ đồ phân tán giữa day và giá trị ngõ ra* 35](#_Toc29535338)

[*Hình 3.27: Sơ đồ phân tán giữa country và giá trị ngõ ra* 36](#_Toc29535339)

[*Hình 3.28: Sơ đồ phân tán giữa biến ico\_duration và giá trị ngõ ra.* 36](#_Toc29535340)

[*Hình 3.29: Biểu đồ tương quan giữa giá trị ngõ vào và giá trị ngõ ra* 37](#_Toc29535341)

[*Hình 3.30: Mối quan hệ giữa price\_usd và price\_btc.* 38](#_Toc29535343)

[*Hình 3.31: Mối quan hệ giữa total supply và available\_supply.* 38](#_Toc29535344)

[*Hình 3.32: Mỗi quan hệ giữa market\_cap\_usd và usd\_raised*. 38](#_Toc29535345)

[*Hình 3.33: Mối quan hệ giữa btc\_price và month.* 39](#_Toc29535346)

[*Hình 3.34: Mối quan hệ giữa usd\_raised và ico\_duaration.* 39](#_Toc29535347)

[*Hình 3.35: Mối quan hệ giữa market\_cap\_usd và ico\_duration.* 39](#_Toc29535348)

[*Hình 3.36: Biểu diễn Sum of Squared.* 41](#_Toc29535350)

[*Hình 3.37: Mối quan hệ giữa market\_cap\_usd và available\_supply.* 42](#_Toc29535351)

[*Hình 3.38: Phân tích tổng quan giữa biến market\_cap\_usd và ngõ ra.* 43](#_Toc29535353)

[*Hình 3.39: Phân tích tổng quan giữa biến available\_supply và ngõ ra.* 44](#_Toc29535354)

[*Hình 3.40: Phân tích tổng quan giữa 2 biến market\_cap\_usd và available\_supply với ngõ ra.* 45](#_Toc29535355)

[*Hình 3.41: Phân tích tổng quan giữa biến used\_raised và ngõ ra.* 46](#_Toc29535359)

[*Hình 3.42: Phân tích tổng quan giữa biến ico\_duration và ngõ ra.* 47](#_Toc29535360)

[*Hình 3.43: Phân tích tổng quan giữa 2 biến usd\_raised và ico\_duration với ngõ ra.* 48](#_Toc29535361)

[*Hình 3.44: Giá trị t và p trước khi ghép của usd\_raised.* 51](#_Toc29535366)

[*Hình 3.45: Giá trị t và p trước khi ghép của ico\_duration.* 51](#_Toc29535367)

[*Hình 3.46: Giá trị t và p sau khi ghép của usd\_raised và ico\_duration* 51](#_Toc29535368)

[*Hình 3.47: Lưu đồ thuật toán hồi quy Ridge.* 53](#_Toc29535371)

[*Hình 3.48: Lưu đồ thuật toán mạng Nơron.* 55](#_Toc29535373)

[*Hình 4.1: Chia dataset thành 2 tập dữ liệu cross và test holdout.* 56](#_Toc29535376)

[*Hình 4.2: Hàm trainingError dùng để tính toán training error.* 56](#_Toc29535377)

[*Hình 4.3: So sánh giá trị dự đoán và giá trị thực tế.* 57](#_Toc29535378)

[*Hình 4.4: Hàm testError dùng để tính toán test error.* 57](#_Toc29535379)

[*Hình 4.5: Giá trị dự đoán trả về từ tập dữ liệu test.* 57](#_Toc29535380)

[*Hình 4.6: So sánh giá trị dự đoán và giá trị thực tế.* 58](#_Toc29535381)

[*Hình 4.7: Sử dụng mô hình để dự đoán một mẫu ICO 0x.* 58](#_Toc29535383)

[*Hình 4.8: Thông số rMSE tối ưu nhất của mô hình.* 59](#_Toc29535387)

[*Hình 4.9: Kết quả dự đoán trên tập X\_test\_holdout với hình rMSE tối ưu.* 60](#_Toc29535388)

[*Hình 4.10: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu rMSE.* 60](#_Toc29535389)

[*Hình 4.11: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số rMSE tối ưu nhất.* 61](#_Toc29535390)

[*Hình 4.12: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có rMSE tối ưu nhất* 61](#_Toc29535391)

[*Hình 4.13: Thông số R2 của mô hình tối ưu R2.* 61](#_Toc29535392)

[*Hình 4.14: Kết quả dự đoán trên X\_test\_holdout với mô hình tối ưu R2.* 62](#_Toc29535393)

[*Hình 4.15: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu R2.* 62](#_Toc29535394)

[*Hình 4.16: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số R2 tốt nhất* 63](#_Toc29535395)

[*Hình 4.17: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có R2 tốt nhất* 63](#_Toc29535396)

[*Hình 4.19: Kết quả dự đoán trên tập dữ liệu X\_test\_holdout với mô hình tối ưu MAE.* 64](#_Toc29535397)

[*Hình 4.20: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu MAE.* 64](#_Toc29535398)

[*Hình 4.21: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số MAE tốt nhất* 65](#_Toc29535399)

[*Hình 4.22: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có MAE tốt nhất* 65](#_Toc29535400)

[*Hình 4.23: Mô hình mạng Noron cho bài toán dự đoán giá ICO* 66](#_Toc29535402)

[*Hình 4.24: Thông số rMSE tối ưu nhất của mô hình mạng Nơron* 66](#_Toc29535403)

[*Hình 4.25: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số rMSE tối ưu nhất.* 67](#_Toc29535404)

[*Hình 4.26: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu rMSE.* 67](#_Toc29535405)

[*Hình 4.27: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số rMSE tối ưu nhất.* 68](#_Toc29535406)

[*Hình 4.28: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có rMSE tối ưu nhất* 68](#_Toc29535407)

[*Hình 4.29: Thông số R2 tối ưu của mô hình mạng Nơron.* 68](#_Toc29535408)

[*Hình 4.30: Kết quả dự đoán trên tập X\_test\_holdout.* 69](#_Toc29535409)

[*Hình 4.31 Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu R2.* 69](#_Toc29535410)

[*Hình 4.32: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số R2 tối ưu nhất.* 70](#_Toc29535411)

[*Hình 4.33: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có R2tối ưu nhất.* 70](#_Toc29535412)

[*Hình 4.34: Thông số MAE tối ưu nhất của mô hình mạng Nơron.* 70](#_Toc29535413)

[*Hình 4.35: Kết quả dự đoán trên tập X\_test\_holdout với mô hình tối ưu MAE.* 71](#_Toc29535414)

[*Hình 4.37: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số MAEtối ưu nhất.* 72](#_Toc29535415)

[*Hình 4.38: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có MAEtối ưu nhất* 72](#_Toc29535416)

# DANH MỤC BẢNG

[*Bảng 3.1: Xem xét sự thay đổi các giá trị của các biến market\_cap\_usd và available\_supply.* 45](#_Toc27596688)

[*Bảng 3.2: Xem xét sự thay đổi các giá trị của các biến ico\_duration và usd\_raised.* 48](#_Toc27596693)

[*Bảng 3.3: So sánh hệ số hồi quy trước của usd\_raised và ico\_duration sau khi ghép.* 48](#_Toc27596695)

[*Bảng 3.4: So sánh sai số chuẩn hệ số hồi quy của usd\_raised và ico\_duration trước và sau khi ghép.* 49](#_Toc27596696)

[*Bảng 3.5: So sánh RSS của 2 biến usd\_raised và ico\_duration trước và sau khi ghép.* 49](#_Toc27596697)

[*Bảng 4.1: Bảng phân tích overfitting* 57](#_Toc27596707)

[*Bảng 4.2: Phân tích dữ liệu thành 3 phần.* 58](#_Toc27596708)

[*Bảng 4.3: So sánh độ chính xác của 2 thuật toán.* 71](#_Toc27596711)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Nghĩa đầy đủ** |
| ICO | **Initial Coin Offering** |
| API | Application programming interface |
| MSE | Mean Square Error |
| RMSE | Root Mean Square Error |
| R2 | R squared |
| RSS | Residual Sum of Squared |
| API | Application Programming Interface |
| REST | Representational State Transfer |
| Df | Degree of Freedom |
| MAE | Mean Absolute Error |
| SVM | Support Vector Machine |
| SSE | Sum of Squared Error |
| SST | Sum of Squared Total |
| SSR | Sum of Squared Regression |

# CHƯƠNG 1: GIỚI THIỆU

## GIỚI THIỆU ĐỀ TÀI

Hiện nay, các công ty start-up mới thành lập sử dụng một hình thức huy động vốn liên quan đến tiền điện tử, gọi là hình thức Initial Coin Offering (ICO). Trong giai đoạn ICO, một số lượng tiền điện tử sẽ được mở bán cho nhà đầu tư dưới dạng token. Sau khi kết thúc giai đoạn này, đồng tiền đó sẽ được niêm yết lên các sàn giao dịch để nhà đầu tư giao dịch với nhau. Việc dự đoán giá sau khi kết thúc giai đoạn này sẽ mang lại lợi ích cho các nhà đầu tư. Trong đề tài này nhóm đã nghiên cứu hai mô hình thuật toán máy học để dự đoán giá trị của token sau giai đoạn ICO. Dự đoán giá bằng máy học đang trở thành xu thế mới trên thế giới, vì tính chính xác và hiệu quả của nó trong các ứng dụng đời sống. Do đó, có rất nhiều thuật toán dùng để dự đoán như: Multiple Regression (hồi quy đa biến), Ridge Regression (hồi quy Ridge), Neural Networks (Mạng Nơron). Trong phạm vi đề tài này sẽ dùng hai thuật toán chính là hồi quy Ridge và mạng Nơron. Bên cạnh đó, đề tài cũng sẽ đánh giá hiệu quả của các phương pháp dự đoán.

## TÍNH CẤP THIẾT CỦA ĐỀ TÀI

Trong lĩnh vực tài chính, kinh doanh, để đạt được lợi nhuận cao nhất thì việc dự đoán giá trị của một mặt hàng. Đối với lĩnh vực tiền điện tử, trong giai đoạn ICO, việc đầu tư vào một loại token và mong muốn bán chúng với giá cao hơn để thu được lợi nhuận chính là mục đích lớn nhất của các nhà đầu tư. Nhận thấy tầm quan trọng của giá trị token sau giao đoạn ICO, nhóm đã quyết định tìm hiểu và xây dựng mô hình dự đoán giá trị token dùng thuật toán máy học.

## MỤC TIÊU CỦA ĐỀ TÀI

Trong luận văn này, nhóm sẽ tìm hiểu khái niệm ICO, Blockchain, ưu, nhược điểm của chúng. Quy trình thu thập dữ liệu từ các trang web tiền điện tử. Tập dữ liệu được sử dụng trong mô hình máy học này sẽ gặp phải một số vấn đề như đa cộng tuyến, overfitting cũng sẽ được nhóm trình bày. Ngoài ra nhóm cũng sẽ tìm hiểu về thuật toán máy học là hồi quy Ridge và mạng Nơron, áp dụng chúng vào việc dự đoán giá token sau 6 tháng.

## ĐỐI TƯỢNG, PHẠM VI NGHIÊN CỨU

Với mục tiêu đề tài ở trên, nhóm cần phải xác định đối tượng nghiên cứu là gì, và phạm vi nghiên cứu đến đâu để từ đó có thể thực hiện nghiên cứu đề tài một cách có hệ thống và dễ dàng trình bày một cách hiệu quả nhất.

### Đối tượng nghiên cứu

* Tiền điện tử và hình thức huy động vốn ICO.
* Vấn đề đa cộng tuyến, overfitting của tập dữ liệu.
* Thuật toán máy học: hồi quy Ridge, mạng Nơron

### Phạm vi nghiên cứu

* Nghiên cứu quy trình thu thập dữ liệu.
* Nghiên cứu thuật toán hồi quy Ridge.
* Nghiên cứu thuật toán mạng Nơron.

## PHƯƠNG PHÁP NGHIÊN CỨU

Để hoàn thành đề tài, việc thực hiện đề tài theo đúng phương pháp sẽ rất thuận lợi cho việc định hướng đề tài của mình cần làm gì và nó sẽ có những vấn đề gì cần tập trung nghiên cứu, chính vì những vấn đề như vậy nhóm phải áp dụng những phương pháp nào cho phù hợp? Dưới đây là một số phương pháp nhóm sẽ thực hiện xuyên suốt trong quá trình thực hiện đề tài.

* Nghiên cứu tài liệu.
* Phân tích, tổng hợp để trình bày các vấn đề.
* Liệt kê, sưu tầm tài liệu.

## BỐ CỤC CỦA ĐỀ TÀI

Có tổng cộng 5 chương trong cuốn báo cáo để làm rõ đề tài này:

* *Chương 1: Giới thiệu* bao gồm các nội dung giới thiệu đề tài, tính cấp thiết đề tài, hình thành ý tưởng, mục tiêu đề tài, phương pháp nghiên cứu, đối tượng, phạm vi nghiên cứu.
* *Chương 2: Cơ sở lý thuyết* bao gồm khái niệm cơ bản về ico, blockchain, quy trình thu thập dữ liệu, hệ số tương quan, hiện tượng đa cộng tuyến, overfitting, các thuật toán.
* *Chương 3: Phân tích và thiết kế* bao gồm các nội dung hình thành ý tưởng, phân tích quy trình thu thập dữ liêu, phân tích đa cộng tuyến và ảnh hưởng của nó, xây dựng thuật toán.
* *Chương 4: Kết quả thực nghiệm và đánh giá* bao gồm thực nghiệm và đánh giá mô hình.
* *Chương 5: Kết luận và hướng phát triển*.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

1. **KHÁI NIỆM VỀ INITIAL COIN OFFERING.**

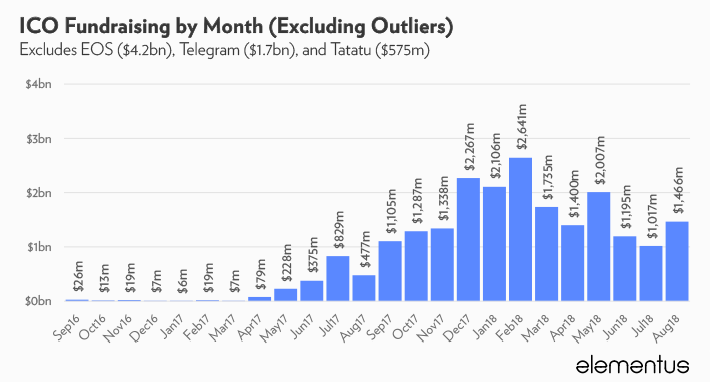
3

Initial Coin Offering (**ICO)** là một hình thức huy động vốn đầu tư của các công ty startup khá phổ biến trong các dự án tiền điện tử. Khi một công ty hay đội ngũ muốn phát triển một dự án, họ thường tạo ra số lượng token nhất định và bán các token này cho những nhà đầu tư trong các đợt mở bán để kêu gọi vốn. Công ty hay đội ngũ sẽ sử dụng số vốn này để phát triển sản phẩm. Khi sản phẩm được thương mại hoá, họ có thể bán token để thu lại lợi nhuận, hoặc dùng token để trải nghiệm sản phẩm đó.

ICO đại diện cho hình thức kêu gọi vốn đầu tư thế hệ mới, dựa trên niềm tin của cộng đồng cũng như đội ngũ phát triển và tiềm năng của nó. ICO ra đời nhằm giải quyết các vấn đề là ngày càng nhiều start-up có ý tưởng hay, công nghệ đột phá nhưng chưa đủ vốn để xây dựng thì lúc này họ sẽ tìm đến ICO.

Tuy nhiên có những công ty tạo ra dự án ma, sau đó dùng hình thức ICO này để kêu gọi vốn. Các dự án này không thể phát triển, dẫn đến công ty phá sản và các nhà đầu tư sẽ mất trắng số tiền họ đã đầu tư cho công ty. Nhưng nếu nhà đầu tư chọn lọc kỹ càng, đầu tư cho những dự án có tiềm năng, khi sản phẩm ra đời, token được lên sàn, thì giá token sẽ tăng so với giá khi nhà đầu tư mua vào lúc phát hành. Lúc này nhà đầu tư sẽ bán các token để kiếm lợi nhuận.

* Các yếu tố ảnh hưởng đến giá của token
* Chất lượng công ty phát hành: Các công ty hoạt động với nhiều năm kinh nghiệm thì chất lượng sản phẩm cũng sẽ tốt hơn.
* Thông tin về ICO như: ngày bắt đầu và kết thúc bán token, giao dịch bằng cách nào, giá, tổng nguồn cung, vốn hóa thị trường.
* Ý tưởng sản phẩm: Công ty cần phải trình bày ý tưởng bằng video thuyết trình, công nghệ, nền tảng, dịch vụ sử dụng là gì. Các cột mốc mà công ty muốn đạt được khi tung ra sản phẩm.
* Phương tiện truyền thông, mạng xã hội: Một dự án nổi bật là khi nó được đề cập nhiều ở các mạng xã hội như Facebook, Twitter.
* Ý kiến của các chuyên gia trong lĩnh vực về tiền điện tử.



*Hình 2.1: Tổng số tiền mà hình thức ICO tạo ra.1*

Theo hình 1, tổng số tiền mà các công ty có được thông qua hình thức ICO từ cuối năm 2017 tăng mạnh và giảm nhẹ vào cuối năm 2018. Nguyên nhân là do thời kì bùng nổ của hình thức gây quỹ này là vào năm 2017, năm 2018 thì thị trường tiền điện tử có xu hướng bão hoà vì đã xuất hiện một vài công ty lừa đảo.

1. **KHÁI NIỆM VỀ TOKEN**

Khi một công ty phát hành một số lượng nhất định token và bán nó cho các nhà đầu tư để kiếm vốn xây dựng dự án.

Token là đồng tiền mã hóa được phát hành trong các đợt ICO. Token được phát hành riêng và dựa trên một nền tảng đồng tiền điện tử nào đó. Đa số token được phát hành dựa trên nền tảng Ethereum. Một số token khác dựa trên NEO, WAVES, STELLAR và Bitcoin.

Có thể phân loại Token thành 2 loại dựa theo tính năng:

* Ultility Token: Là token tiện ích được sinh ra để phục vụ một dự án với mục tiêu cụ thể.
* Security Token: Là token chứng khoán, một dạng cổ phiếu điện tử phát hành dưới dạng token. Nhà đầu tư sẽ được hưởng cổ tức dựa trên số cổ

1 Mike Kalomeni, “The ICO Market is not collapsing. It’s maturing”.Available: <https://elementus.io/blog/ico-market-august-2018/>

phần họ sở hữu trong dự án đó. Ngoài ra, dạng token này còn cho phép những người sở hữu nó có quyền tham gia quyết định một số công việc của dự án.

Ví dụ: BNB Token của Binance có tính năng giảm giá phí giao dịch…

1. **KHÁI NIỆM VỀ BLOCKCHAIN**

Blockchain(hay cuốn sổ cái) là hệ thống cơ sở dữ liệu cho phép lưu trữ và truyền tải các thông tin. Chúng được liên kết với nhau nhờ mã hóa. Các khối thông tin này hoạt động độc lập và mở rộng theo thời gian. Chúng được quản lý bởi những người tham gia hệ thống không thông qua đơn vị trung gian. Khi một khối thông tin được ghi vào hệ thống Blockchain thì không có cách nào thay đổi được, chỉ có thể bổ sung khi đạt được sự thỏa thuận của tất cả mọi người. Khối thông tin là những cuộc trao đổi, giao dịch trong thực tế.

1. **Ưu điểm blockchain**

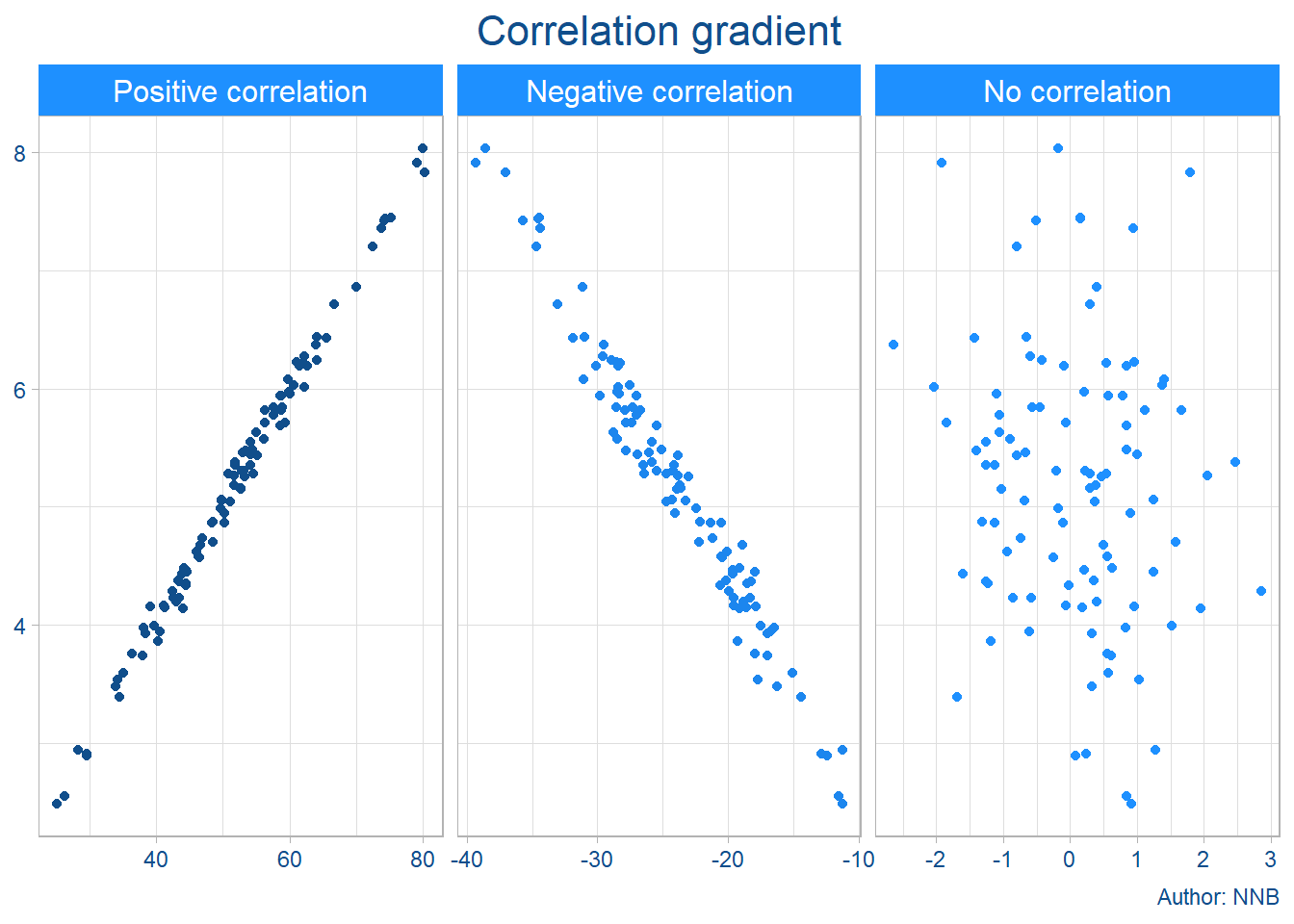
* Tính minh bạch: Các hệ thống xây dựng dựa trên nền tảng Blockchain cung cấp nhiều bước tiến về tính minh bạch cụ thể là về cách lưu trữ hồ sơ hiện hành. Những thay đổi này cho phép mọi người trong mạng lưới có thể xem xét sổ cái và ngay khi nhập vào Blockchain, các giao dịch không thể bị sửa đổi hay xóa bỏ.
* Loại bỏ đơn vị trung gian: Một lợi thế của công nghệ Blockchain so với các hệ thống hiện hành là khả năng xóa bỏ các đơn vị trung gian bằng cách cho phép mọi người giao dịch trực tiếp với nhau thay vì qua một bên thứ ba nào đó.
* Niềm tin: Các phương thức giao dịch hiện hành giữa mọi người cần tới lòng tin vào đơn vị trung gian để quá trình diễn ra thuận lợi. Blockchain cho phép loại bỏ các đơn vị trung gian nhưng vẫn duy trì sự tin tưởng giữa mọi người tham gia giao dịch. Niềm tin được đặt vào mạng lưới Blockchain thay vì vào bên thứ ba.
* Độ bảo mật: Dữ liệu một khi đã được đưa vào Blockchain không thể thay đổi hoặc chỉnh sửa.
* Tiết kiệm chi phí: Công nghệ Blockchain có thể giúp giảm bớt rất nhiều chi phí nhờ loại bỏ được các đơn vị trung gian.
* Tăng tốc độ giao dịch:Các hệ thống sử dụng công nghệ Blockchain không chỉ giúp giảm chi phí giao dịch mà còn thực sự tăng tốc độ giao dịch. Nhờ loại bỏ được các đơn vị trung gian. Blockchain có thể xử lý giao dịch gần như ngay lập tức.

1. **Nhược điểm blockchain**

* Dễ bị hacker nhòm ngó:Dù được bảo vệ bởi các thuật toán mạnh mẽ nhưng các ứng dụng dựa trên công nghệ Blockchain vẫn còn là “con mồi” của rất nhiều cuộc tấn công mạng.
* Việc sửa dữ liệu rất khó khăn:Một khi dữ liệu đưa vào Blockchain thì sẽ rất khó thay đổi. Độ bảo mật cao là một lợi thế nhưng nó cũng là nhược điểm.

1. **HỆ SỐ TƯƠNG QUAN**

Hệ số tương quan thể hiện sự tương quan giữa hai hay nhiều biến không phân biệt biến này phụ thuộc vào biến khác. Ký hiệu là *r*.

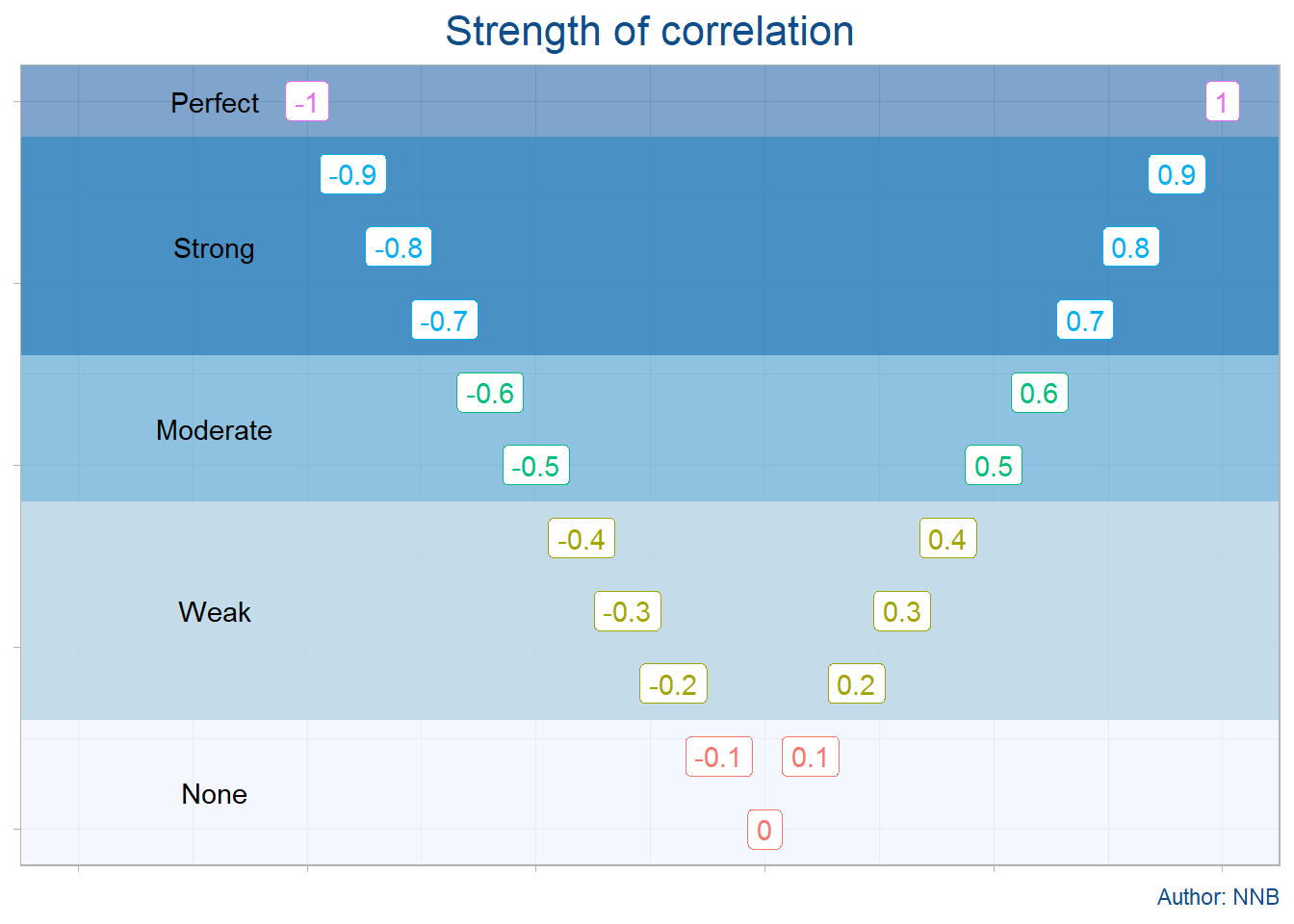


*Hình 2.2: Chiều tương quan.2*

**Đặc điểm:**

* Hệ số tương quan không có đơn vị.
* *r* > 0: Tương quan dương.
* *r* < 0: Tương quan âm.
* *r* = 0: Không tương quan với nhau.
* Nằm trong khoảng [-1, 1].

2 Nguyễn Ngọc Bình, “Phân tích hồi quy và tương quan”, Available: <https://rpubs.com/nguyenngocbinhneu/432206>.



*Hình 2.3: Thể hiện độ tương quan.*

Đánh giá mức độ tương quan giữa 2 biến tiên lượng. Cho 2 biến số x, y và *n* mẫu, hệ số tương quan được tính toán bằng công thức [1]:

(2.1)

Trong đó:

* x, y là 2 biến số cần xem xét độ tương quan.
* i: Chỉ số chạy từ 1 đến n.
* xmean, ymean: Lần lượt là trung bình của biến x và y.
* n: Kích thước mẫu.
  1. **HIỆN TƯỢNG ĐA CỘNG TUYẾN**

Hiện tượng đa cộng tuyến là hiện tượng có 2 hoặc nhiều biến dự đoán được xác định trong mô hình hồi quy tuyến tính có mối tương quan cao. Sự hiện diện của đa cộng tuyến tác động tiêu cực đến toàn bộ phân tích và hạn chế nghiêm trọng các kết luận của nghiên cứu, thường xảy ra trong các nghiên cứu mà chúng ta không thể tác động vào quá trình thu thập dữ liệu. Có 2 dạng đa cộng tuyến:

* Đa cộng tuyến cấu trúc (Structural multicollinearity): Là 1 dạng “nhân tạo”, tạo ra biến mới từ biến nào đó.

**Ví dụ:**

* Biến x tạo ra biến x2 và x2 = x + 10, trong trường hợp này x và x2 có liên quan với nhau.
* Hồi quy đa thức (polynomial regression): có x và tạo x2 thì lúc này x và x2 chắc chắn có liên quan đến nhau.
* Đa cộng tuyến dựa trên dữ liệu (Data-based multicollinearity)**:** Là hệ quả của các quá trình nghiên cứu thiết kế hoặc thu thập dữ liệu kém chất lượng.
  1. **MÔ HÌNH THUẬT TOÁN HỒI QUY RIDGE** 
     1. **Bài toán hồi quy đa biến**

Đối với hồi quy đa biến thì mối quan hệ giữa đầu ra và đầu vào có thể được thông qua bằng một hàm tuyến tính [2]:

(2.2)

Trong đó:

* y là giá trị thực tế.
* ŷ là giá trị dự đoán.
* là một vector chứa các tham số mô hình và cũng là ma trận chuyển vị của . Đây cũng chính là tham số mô hình bài toán.

Sau khi xây dựng xong mô hình dự đoán đầu ra, cần tìm một phép đánh giá phù hợp với bài toán. Với bài toán hồi quy nói chung, mong muốn sai số *e* (*error)* giữa giá trị thực *y* và giá trị dự đoán *ŷ* là nhỏ nhất. Nói cách khác, giá trị sau đây càng nhỏ càng tốt [2]:

(2.3)

Lấy bình phương vì có thể là số âm.Việc sai số nhỏ nhất cũng có thể mô ta bằng cách lấy trị tuyệt đối , tuy nhiên cách này ít được sử dụng vì không thuận tiện cho việc tối ưu. Hệ số sẽ bị triệt tiêu sau khi lấy đạo hàm của *e* theo tham số mô hình .

Hàm *J()* là hàm mất mát của hồi quy tuyến tính với tham số mô hình là .

(2.4)

Trong đó:

* là ký hiệu cho các input của mẫu huấn luyện thứ i.
* N là cỡ mẫu.
* là giá trị thực tế tương ứng với i.

Trước khi xây dựng nghiệm bài toán tối ưu hàm mất mát, hàm số này có thể được viết dưới dạng vector như sau [3]:

(2.5)

Với là vector chứa các giá trị thực tế , là vector chứa các input của mẫu huấn luyện, là vector chứa các tham số mô hình.

Nghiệm bài toán hồi quy đa biến [3]:

(2.6)

Việc tìm trung bình sai số nhỏ nhất tương đương với việc tìm để hàm đạt giá trị nhỏ nhất. Giá trị của có thể tính thông qua việc giải phương trình đạo hàm *J(*theo

Hạn chế của hồi quy đa biến:

* Rất nhạy cảm với nhiễu. Khi gặp nhiễu thì điều này sẽ ảnh hưởng đến kết quả.
* Không biểu diễn được các mô hình có độ phức tạp lớn. Khi mô hình quá phức tạp có thể dẫn đến overfitting.
  + 1. **Hiện tượng quá khớp (Overfitting)**

Trong các bài toán học có giám sát, thường phải tìm một mô hình ánh xạ đặc trưng thành các kết quả tương ứng trong tập huấn luyện. Tức là đi tìm hàm số *f* sao cho *yi f(xi),* i = 1,2,…,N. Nếu mô hình quá khớp với dữ liệu thì nó sẽ phản tác dụng. Hiện tượng này trong máy học được gọi là quá khớp*.* Nó xảy ra khi mô hình quá khớp với tập huấn luyện nhưng với tập chưa nhìn thấy thì kết quả sẽ sai lệch. Một mô hình chỉ mô tả tốt tập huấn luyện là mô hình không có tính tổng quát.

Dưới đây là 2 đại lượng đơn giản đánh giá chất lượng của mô hình trên tập huấn luyện và tập chưa nhìn thấy.

**Training error**:Đại lượng này là sự sai khác giữa giá trị thực và giá trị dự đoán của mô hình, thường là giá trị của hàm mất mát lên mô hình. Hàm mất mát này cần có 1 thừa số là để tính giá trị trung bình, tức mất mát trung bình trên mỗi điểm dữ liệu. Với mô hình hồi quy, bài toán được xác định là mean squared error [2].

*Training error =*  (2.7)

**Test error**:Tương tự với đại lượng training error, nhưng mô hình tìm được, được áp dụng vào tập dữ liệu chưa nhìn thấy [2].

*Test error =*  (2.8)

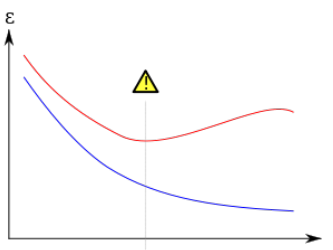
Việc lấy trung bình là rất quan trọng vì lượng dữ liệu trong tập huấn luyện có thể chênh lệch rất nhiều.

Một mô hình được gọi là tốt nếu training error và test error đều thấp. Nếu training error thấp và test error cao thì đó là hiện tượng quá khớp. Nếu training error cao và test error cao thì đó là hiện tượng chưa khớp (underfitting). Xác suất để xảy ra việc training error cao nhưng test error thấp là rất nhỏ. Kỹ thuật phổ biến giúp tránh được hiện tượng quá khớp là regularization*.*

Trong nhiều bài toán, lượng tham số cần xác định thường lớn. Điều này dẫn đến việc huấn luyện mô hình cũng khó khăn. Kỹ thuật regularization là chấp nhận hy sinh độ chính xác trong tập huấn luyện, nhưng sẽ giảm đi độ phức tạp mô hình, giúp tránh đượchiện tượng quá khớp, nhưng vẫn giữ được tính tổng quát của mô hình. Có 2 kỹ thuật regularization đó là Early stopping và Thêm số hạng vào hàm mất mát.

* + - 1. **Early stopping**

Rất nhiều mô hình sử dụng thuật toán lặp cho đến khi hàm mất mát hội tụ để tìm ra nghiệm. Giá trị của hàm mất mát giảm dần khi số vòng lặp tăng lên. Một giải pháp đưa ra đó là dừng thuật toán trước khi nó hội tụ. Giải pháp này có tên là early stopping.



*Hình 2.4: Early stopping.3*

Một kỹ thuật thường dùng là tách từ tập huấn luyện ra một tập xác nhận (validation set). Trong khi huấn luyện, tính toán cả training error và validation error, nếu training error vẫn có xu hướng giảm nhưng validation error có xu hướng tăng lên thì dừng thuật toán. Hình 2.4 mô tả cách tìm điểm *stopping.* Đường màu xanh là training error, màu đỏ là validation error, trục hoành thể hiện số vòng lặp còn trục tung là giá trị error.

* + - 1. **Thêm số hạng vào hàm mất mát**

Kỹ thuật regularization phổ biến hơn là thêm vào hàm mất mát một số hạng. Hàm mất mát mới này được gọi là *regularized loss function***,** thường được định nghĩa như sau [2]:

(2.9)

Trong đó:

* là tham số mô hình.
* : là hàm mất mát phụ thuộc vào training set và .
* : Là số hạng regularization chỉ phụ thuộc vào .
* : Thường là một số thực dương và được gọi là tham số regularization (paramater regularization).

3 Vũ Hữu Tiệp, “Machine Learning Cơ Bản,” [Trực tuyến]. Available: <http://machinelearningcoban.com/ebook/>.

* + 1. **Hồi quy Ridge**

Hồi quy Ridge là biến thể của hồi quy tuyến tính nhằm khắc phục hiện tượng đa cộng tuyến và quá khớp.

Cũng giống như hồi quy đa biến thì hồi quy Ridge cũng tìm cách khớp với số liệu bằng cách tối thiểu hóa sai số. Phương trình của hồi quy Ridge có dạng [1]:

(2.10)

Trong đó:

* là ký hiệu cho các input của mẫu huấn luyện thứ i.
* N là cỡ mẫu.
* là giá trị thực tế tương ứng với i.
* λ: Tham số điều chỉnh độ phức tạp mô hình.

Từ công thức 2.10, có thể được viết lại [1]:

(2.11)

Với là vector chứa các giá trị thực tế , là vector chứa

các input của mẫu huấn luyện, là vector chứa các tham số mô hình.

Tương tự như bài toán hồi quy đa biến là đi tìm giá trị các tham số mô hình sao cho *J()* cực tiểu.

Nghiệm bài toán hồi quy Ridge [1]:

(2.12)

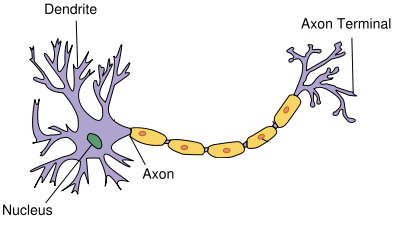
Trong đó:

* X là ma trận chứa các mẫu huấn luyện và là ma trận chuyển vị của X.
* I là ma trận đơn vị với kích thước phù hợp.
* y là ma trận chứa các giá trị thực tế.

Khác với hồi quy đa biến thì nghiệm hồi quy có thêm một lượng *λI.* Điều này sẽ giúp giảm đi độ phức tạp mô hình bằng cách giảm các tham số mô hình xuống.

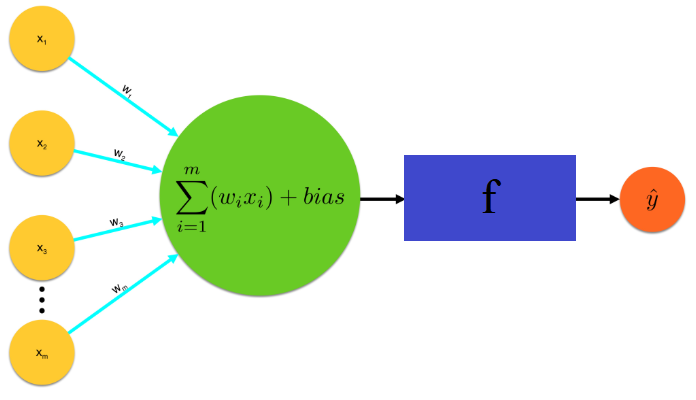
* 1. **MÔ HÌNH MẠNG NƠRON**

Mạng nơron là một mô hình toán học mô phỏng nơron trong hệ thống thần kinh của con người. Mô hình này biểu hiện cho một số chức năng của nơron thần kinh con người.



# *Hình 2.5: Mạng nơron tự nhiên.4*

Với hệ thống thần kinh con người, nơron nhận các tín hiệu từ sợi nhánh (dendrite), nhân tế bào (nucleus) sẽ xử lý thông tin, ngõ ra của 1 nơron sẽ được truyền sang nơron khác thông qua sợi trục axon.



# *Hình 2.6: Mạng nơron nhân tạo.*

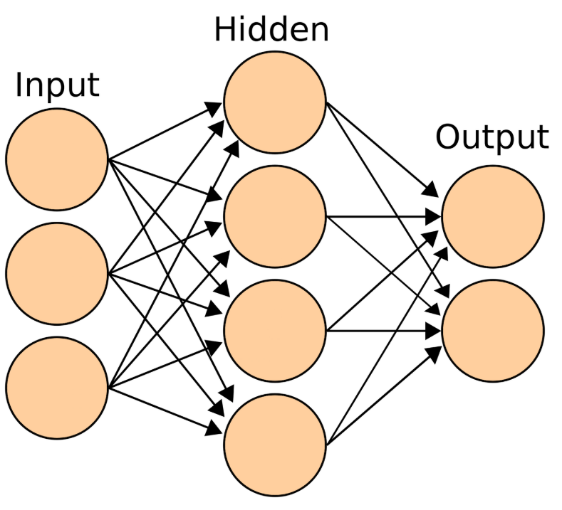
4 Scott Robinson,”Introduction to Neural Networks with Scikit-Learn, available: <https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/>

Mạng nơron của mô hình toán học được mô phỏng tương tự dựa trên công thức sau:

*= f* (2.13) [3]

Trong đó:

* X: Dữ liệu ngõ vào.
* Trọng số w (weight): Khi mô hình nhận dữ liệu ngõ vào, việc xử lý từng dữ liệu sẽ được gắn với 1 trọng số w. Nếu dữ liệu ít quan trọng sẽ có trọng số w thấp, ngược lại dữ liệu quan trọng sẽ có trọng số w cao.
* Hàm f: Được gọi là hàm kích hoạt (Activation Function), có nhiệm vụ chuẩn hoá dữ liệu ngõ ra.
* Bias: Tham số được thêm vào để điều chỉnh ngõ ra.



# *Hình 2.7: Mô hình mạng Nơron đa lớp 5*

Một mạng Nơron đa lớp sẽ gồm:

* Lớp ngõ vào (*input layer*): Là lớp bên trái của mạng thể hiện cho các đầu vào của mạng.
* Lớp ẩn (*hidden layer*): Là lớp nằm giữa lớp ngõ vào và lớp ngõ ra, thể hiện cho việc suy luận logic của mạng.
* Lớp ngõ ra (*output layer*): Là lớp bên phải của mạng thể hiện cho các đầu ra của mạng.

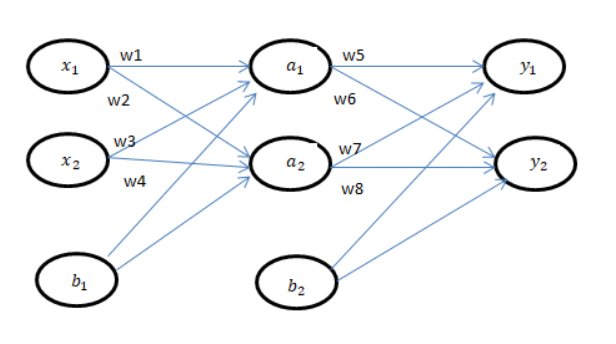
5 Scott Robinson,”Introduction to Neural Networks with Scikit-Learn, available: <https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/>

* + 1. **Lan truyền tiến (Feedforward)**

Dữ liệu từ lớp ngõ vào chỉ được truyền thẳng. Việc xử lý dữ liệu có thể xảy

ra trên nhiều lớp nhưng không có liên kết ngược, nghĩa là không có sự liên kết từ

lớp ngõ ra đến lớp ngõ vào. Sai số được tính bằng cách so sánh giá trị thực của mạng với giá trị mong muốn, trong đó sai số của quá trình huấn luyện thường được lấy bằng tổng bình phương tất cả sai số thành phần.



# *Hình 2.8: Lan truyền tiến.*

Trong đó:

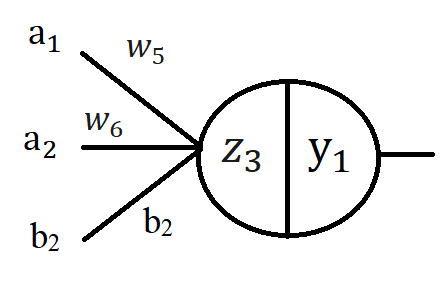
* x1, x2 là các dữ liệu ngõ vào.
* y1, y2 là các dữ liệu ngõ ra.
* b1, b2 là các hệ số bias.
* w1, w2…w8 là các trọng số.

Giả sử trong mạng Nơron ở hình 7 sử dụng hàm kích hoạt Sigmoid.

* + 1. **Lan truyền ngược (Backpropagation)**

Lan truyền ngược là một kĩ thuật dùng để cập nhật trọng số w trong mạng nơron, nhằm giảm sự sai số.

Kĩ thuật này sẽ lấy đạo hàm của hàm sai số tại ngõ ra theo trọng số đã có, sau đó sẽ cập nhật lại trọng số nhờ sử dụng kĩ thuật Gradient Descent nhằm mục địch giảm tối đa sai số. Quy trình cập nhật trọng số sẽ bắt đầu từ trọng số ở lớp cuối cùng, khi cập nhật xong sẽ bắt đầu cập nhật trọng số ở các lớp trước đó.



# *Hình 2.9 Lan truyền ngược.*

Giả sử đạo hàm của hàm sai số theo trọng số áp dụng quy tắc dây chuyền (chain rule) được tính theo công thức

(2.14)

Sau khi tính được giá trị , trọng số mới sẽ được tính bằng công thức

(mới) = – η.(2.15)

Trong đó: η là tốc độ học (learning rate)

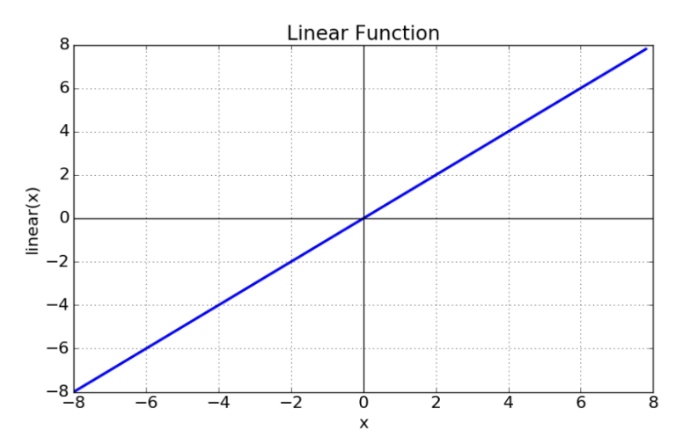
* + 1. **Activation Function**

Activation Function hay còn gọi là hàm kích hoạt có chức năng mô phỏng tỷ lệ truyền xung qua axon của một nơron thần kinh. Giống như trong mô hình sinh học, các xung thần kinh được truyền qua axon với tỷ lệ nào đó. Ở mô hình học máy mô phỏng xây dựng, các hàm kích hoạt sẽ điều chỉnh tỷ lệ truyền này.

Có 2 loại hàm kích hoạt:

* Hàm kích hoạt tuyến tính (Identify function): Là hàm có dạng

*f(x) = x.* (2.16)



# *Hình 2.10: Đồ thị kích hoạt tuyến tính.*

Ngõ ra của hàm kích hoạt tuyến tính không bị giới hạn, vì vậy nó không thể được sử dụng cho mạng nơron phức tạp.

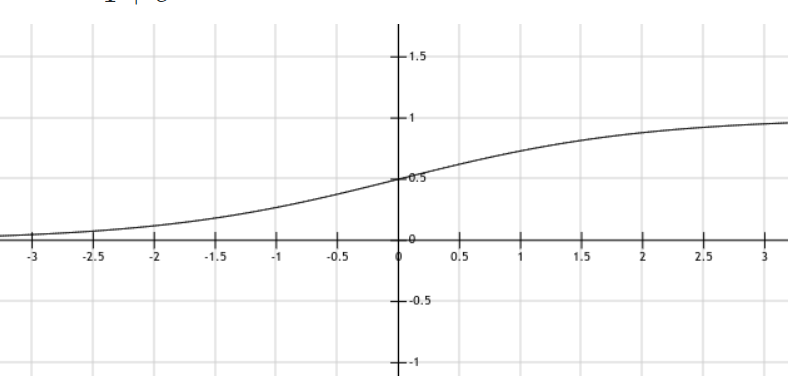
* Hàm kích hoạt phi tuyến tính: Hàm kích hoạt phi tuyến tính là một phép biến đổi phi tuyến tính được áp dụng với tín hiệu ngõ vào. Nếu không có hàm kích hoạt phi tuyến tính thì mạng nơron sẽ biến đổi tuyến tính. Xử lý một mạng nơron tuyến tính sẽ đơn giản hơn nhưng nó khó có thể mô hình hoá và giải quyết những vấn đề phức tạp.

Một số hàm kích hoạt phổ biến:

**Sigmoid**

Công thức:

(2.17)

****

*Hình 2.11: Đồ thị hàm Sigmoid.*

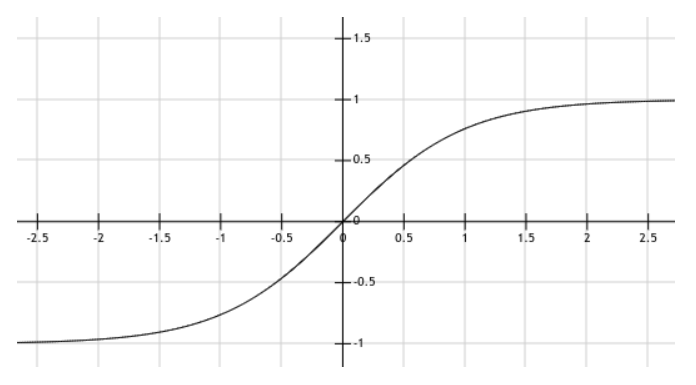
Hàm Sigmoid sẽ nhận dữ liệu ngõ vào là một số thực và chuyển thành một giá trị output trong khoảng (0;1). Ngõ vào là một số thực âm rất nhỏ sẽ cho ngõ ra tiệm cận với 0, ngược lại nếu ngõ vào là một số thực dương lớn sẽ cho ngõ ra là một số tiệm cận 1. Hàm Sigmoid phù hợp với mô hình dự đoán xác suất trong khoảng (0;1).

Nhược điểm của hàm Sigmoid: Khi ngõ vào có trị tuyệt đối lớn (rất âm hoặc rất dương), gradient của hàm số này sẽ rất gần với 0. Điều này đồng nghĩa với việc các hệ số tương ứng với node đang xét sẽ gần như không được cập nhật, hay còn gọi là hiện tượng Vanishing Gradients.

**Tanh**

Công thức:

(2.18)



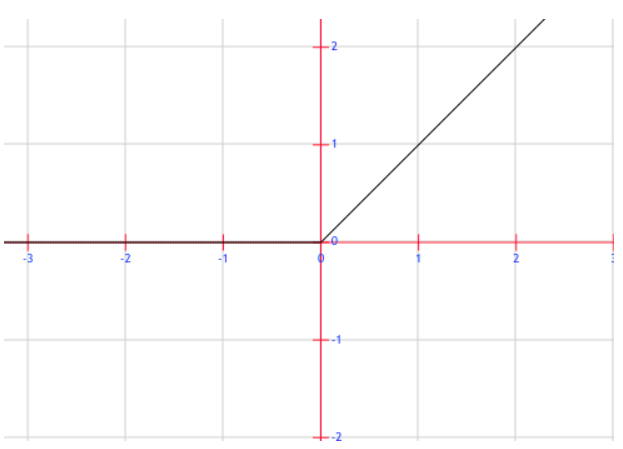
# *Hình 2.12: Đồ thị hàm tanh.*

Hàm tanh nhận ngõ vào là một số thực và ngõ ra là một giá trị trong khoảng (-1;1). Giống như Sigmoid, hàm Tanh bị bão hoà ở hai đầu. Tuy nhiên hàm Tanh đối xứng qua 0, vì vậy nó giúp cho quá trình hội tụ sẽ nhanh hơn, nhưng nó vẫn chưa khắc phục được hiện tượng Vanishing Gradients.

**ReLu**

Công thức:

(2.19)



# *Hình 2.13: Đồ thị hàm Relu.*

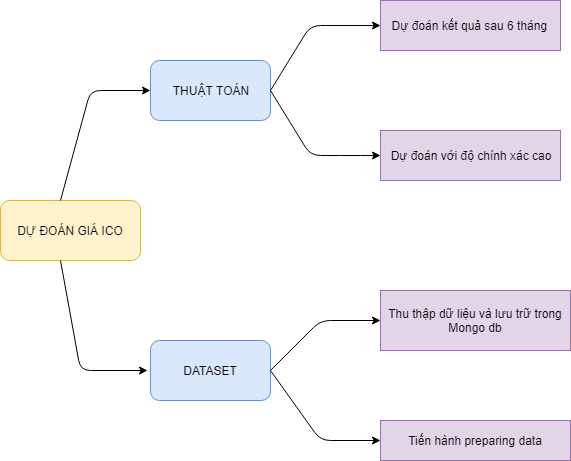
Hàm ReLu sẽ lọc các giá trị nhỏ hơn 0. Hàm ReLu sẽ có tốc độ hội tụ nhanh hơn do không bị bão hoã ở hai đầu. Công thức của hàm ReLu đơn giản hơn công thức của hàm Tanh và Sigmoid, vì vậy sẽ tốn ít chi phí khi tính toán.

Tuy nhiên hàm ReLu cũng có những nhược điểm sau: Với các giá trị nhỏ hơn 0, hàm ReLu sẽ loại bỏ chúng, hiện tượng này gọi là Dying ReLu. Nếu các node bị chuyển thành 0 thì sẽ không có ý nghĩa ở lớp tiếp theo và các trọng số tương ứng sẽ không được cập nhật với Gradient Descent.

# CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ

1. **HÌNH THÀNH Ý TƯỞNG**

Trong đề tài nhóm đã hình thành ý tưởng để có thể dự đoán giá của token một cách chính xác nhất bằng cách áp dụng thuật toán sử dụng tập dữ liệu được thu thập từ các trang web và huấn luyện nó để nó có thể căn cứ vào đó dự đoán được giá token trong sau 6 tháng kể từ khi một ICO được phát hành. Thuật toán nhóm sử dụng là hồi quy Ridge và mạng Nơron.

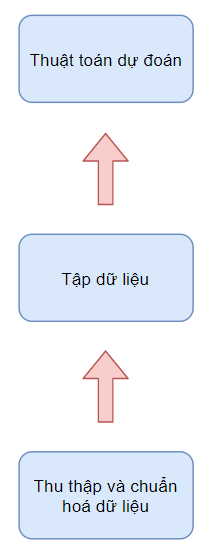


*Hình 3.1: Sơ đồ tư duy hệ thống.*

Dựa theo sơ đồ tư duy hệ thống bên thì hệ thống này sẽ được chia làm 2 phần đó là: Phần thuật toán và phần tập dữ liệu.

* **Phần tập dữ liệu (dataset)**: Dữ liệu thu thập được gồm rất nhiều thông tin liên quan đến các ICO. Dữ liệu sẽ được chọn lọc để loại bỏ đi những dữ liệu sai. Sau đó dữ liệu sẽ được kết xuất ra file csv để làm đầu vào cho thuật toán dự đoán.
* **Phần thuật toán**: Với dữ liệu thu thập và chọn lọc được, tiến hành đưa vào 2 thuật toán là hồi quy Ridge và mạng Nơron. Kết quả trả về sẽ là giá của một ICO sau 6 tháng. Tiến hành so sánh hiệu năng giữa 2 thuật toán để xem thuật toán nào tốt hơn.

1. **MÔ HÌNH HỆ THỐNG**

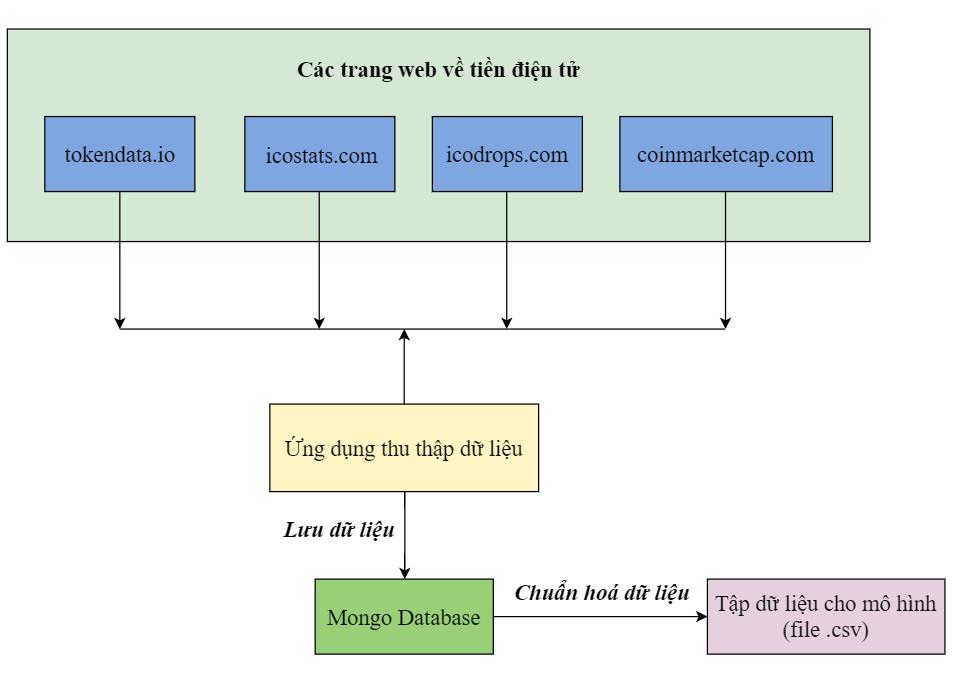


*Hình 3.2: Mô hình hệ thống.*

* Một API được viết bằng ngôn ngữ C# chịu trách nhiệm lấy dữ liệu từ các trang web, lưu trữ trong cơ sở dữ liệu và sau đó sẽ tiến hành chuẩn hoá dữ liệu (preparing data).
* Dữ liệu sẽ được kết xuất dưới định dạng csv, dữ liệu lúc này đã hoàn chỉnh.
* Thuật toán dữ đoán sẽ lấy dữ liệu từ tập dữ liệu, tính toán và dữ đoán.
  1. **QUY TRÌNH THU THẬP DỮ LIỆU**

API được viết bằng C# có chức năng là sẽ lấy dữ liệu từ các trang web uy tín như tokendata.io, icostats.com, icodrops.com, prediction.vc, coinmartketcap.com. Đây là các trang web uy tín về tiền điện tử, dữ liệu thu thập được từ các trang web này khá chính xác. Sau khi tổng hợp thông tin từ các trang web này, nhóm đã chọn ra các yếu tố ảnh hưởng đến giá Token, đó là:

* Price USD (giá đô la)
* Price BTC (giá bitcoin).
* Total supply (Tổng cung ứng).
* Market Cap (Vốn hóa thị trường).
* Available Supply (Nguồn cung có sẵn).
* USD raised (Tiền nhận được).
* Ethereum price at lunch (Giá Ethereum tại thời điểm mở bán).
* Bitcoin price at lunch (Giá bitcoin tại thời điểm mở bán).
* Month ICO was lunched (Tháng bắt đầu mở bán ICO).
* Date ICO was lunched (Ngày ICO mở bán).
* County an ICO was lunched from (Quốc gia mở bán ICO).
* ICO Durations day (Thời lượng phát hành ICO).

****

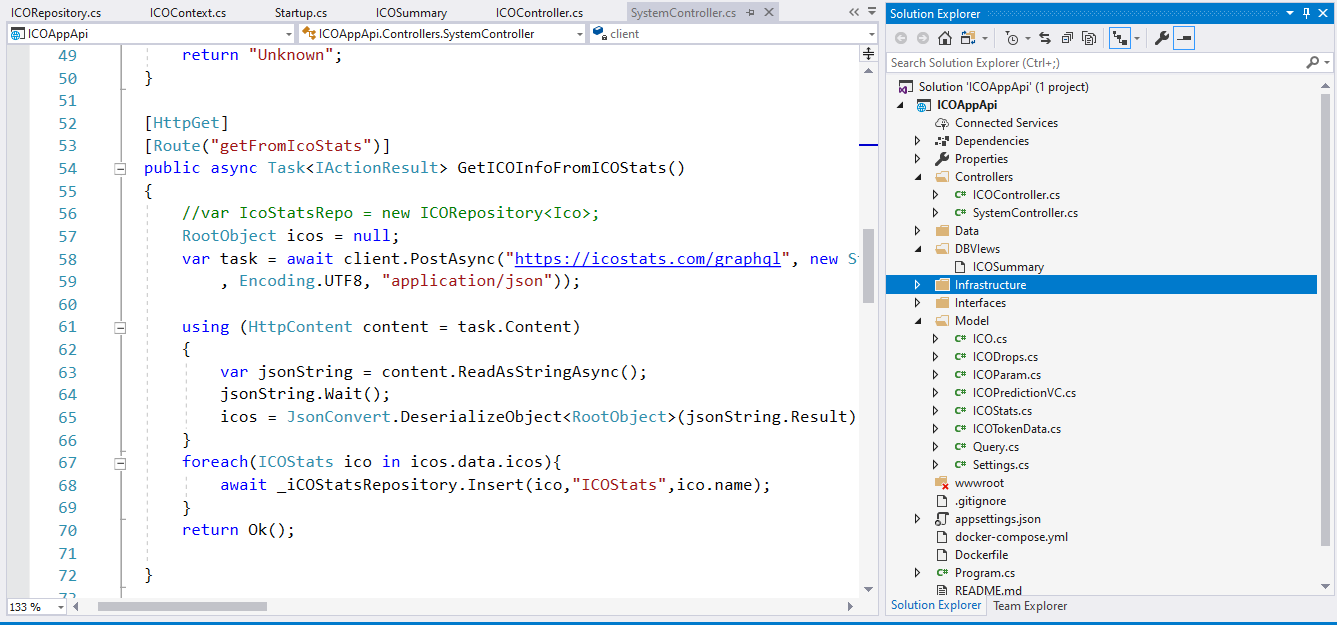
*Hình 3.3: Quy trình thu thập dữ liệu.*

REST là một kiểu kiến trúc để viết API. Nó sử dụng phương thức HTTP để tạo giao tiếp giữa các máy. Chức năng quan trọng nhất của REST là quy định cách sử dụng các HTTP method (như GET, POST, PUT, DELETE…).



# *Hình 3.4: Rest API design.*

**RESTful API** là một tiêu chuẩn dùng trong việc thiết kế các API cho các ứng dụng web để quản lý các resource. RESTful là một trong những kiểu thiết kế API được sử dụng phổ biến ngày nay để cho các ứng dụng (web, mobile…) khác nhau giao tiếp với nhau. API này chỉ lấy dữ liệu từ các trang web nên chỉ cần sử dụng HTTP method GET. Dữ liệu sau đó được lưu vào cơ sở dữ liệu *mongodb*trong đó sẽ bao gồm thông tin các ICO thu thập được từ các trang web.



# *Hình 3.5: Ứng dụng thu thập dữ liệu.*

Hình 19 là tổng quan về API dùng để lấy dữ liệu. Một vấn đề xảy ra là khi lấy dữ liệu thì cần cài đặt chương trình Visual Studio. Khi chuyển qua máy tính khác thì vẫn yêu cầu cài đặt Visual Studio, điều này sẽ làm mất rất nhiều thời gian. Để tránh mất thời gian cài đặt thì nhóm sẽ sử dụng nền tảng Docker.

**Docker** là một nền tảng để cung cấp cách để xây dựng, phát triển và thực thi ứng dụng dễ dàng hơn bằng cách sử dụng các hộp chứa (containers). Lập trình viên đóng gói một ứng dụng với tất cả các phần cần thiết, chẳng hạn như thư viện và các phụ thuộc khác, và gói tất cả ra dưới dạng một gói, được gọi là containers. Nhờ vào containers, ứng dụng sẽ chạy trên các máy tính khác nhau, bất kể mọi cài đặt tùy chỉnh mà máy có thể có khác với máy được sử dụng để viết code. Tuy nhiên việc quản lý Docker trên một hệ thống lớn với rất nhiều các containers khác nhau sẽ rất phức tạp và khả năng sinh ra nhiều lỗi. Để khắc phục vấn đề này, Docker đã đưa ra một công cụ rất hữu ích đấy là Docker Compose.

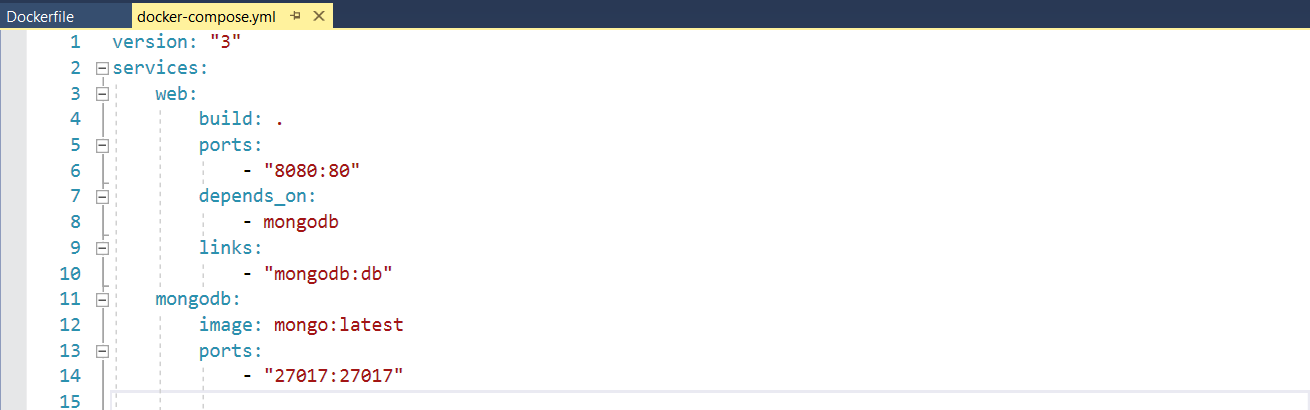
Docker Compose giúp cho việc tạo và thiết lập các ứng dụng cần sự liên kết giữa nhiều Docker Container với nhau.

* **Cách cài đặt Docker Compose:**

Cài đặt Compose có thể sử dụng link từ trang chủ của Docker. Để sử dụng Compose thông thường có ba bước sau:

* Tạo Dockerfile cho mỗi môi trường container của từng service mình muốn. Dockerfile là bắt buộc để khởi tạo container.
* Tạo file docker-compose.yml để định nghĩa mối liên kết giữa các containers với nhau.
* Chạy lệnh docker-compose up để khởi động Compose và chạy toàn bộ ứng dụng

Ứng dụng sau khi thu thập sẽ lưu trữ dữ liệu bằng database MongoDB. Vì vậy trong file docker-compose.yml sẽ khai báo dịch vụ database MongoDB và port được sử dụng.



*Hình 3.6: File docker-compose.yml*

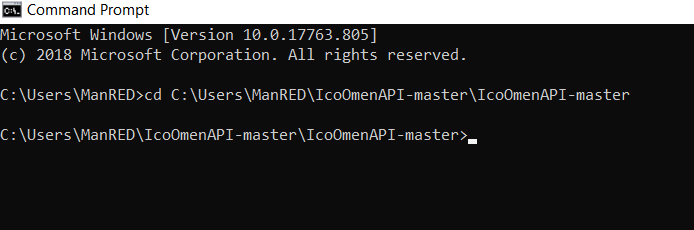
* **Cách cài đặt ứng dụng MongoDB Compass Community**

MongoDB Compass là ứng dụng dùng để chạy các truy vấn, tìm kiếm và quản lý cơ sở dữ liệu MongoDB. Ứng dụng được tải miễn phí tại trang chủ.

* **Cách sử dụng Docker Compose để chạy ứng dụng**

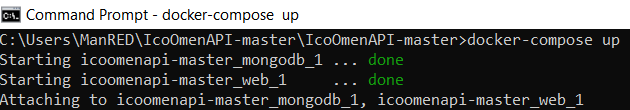
Bước 1: Khởi động ứng dụng Docker Desktop.

Bước 2: Mở Command Prompt và di chuyển tới đường dẫn của thư mục ứng dụng.



*Hình 3.7: Giao diện Command Prompt*

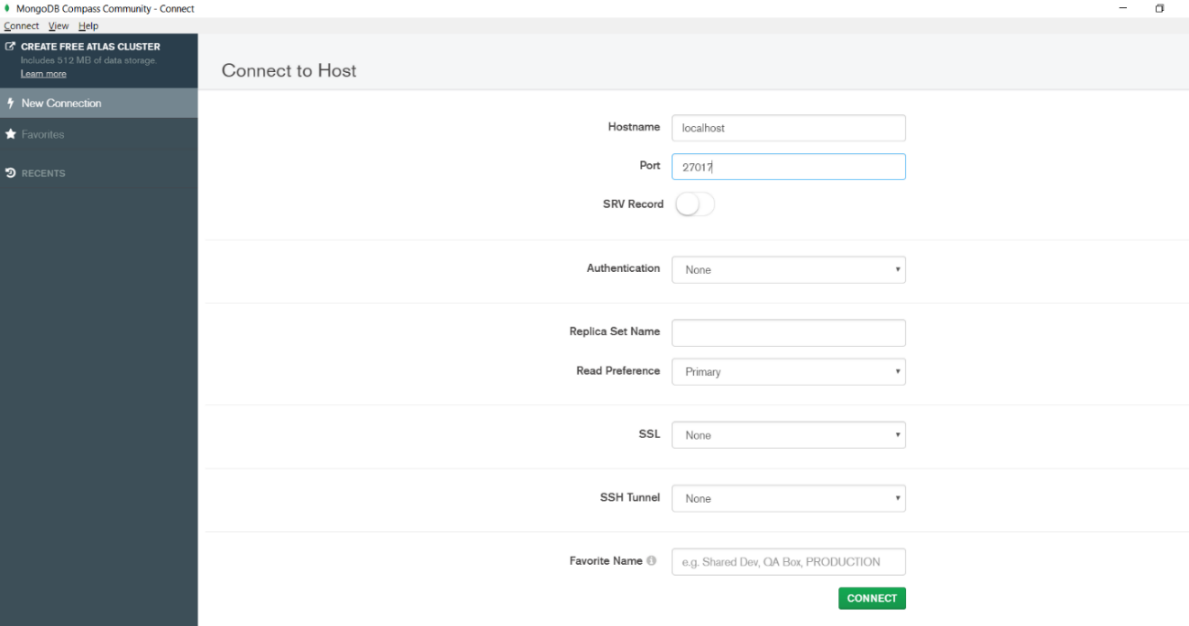
Bước 3: Chạy lệnh docker-compose up



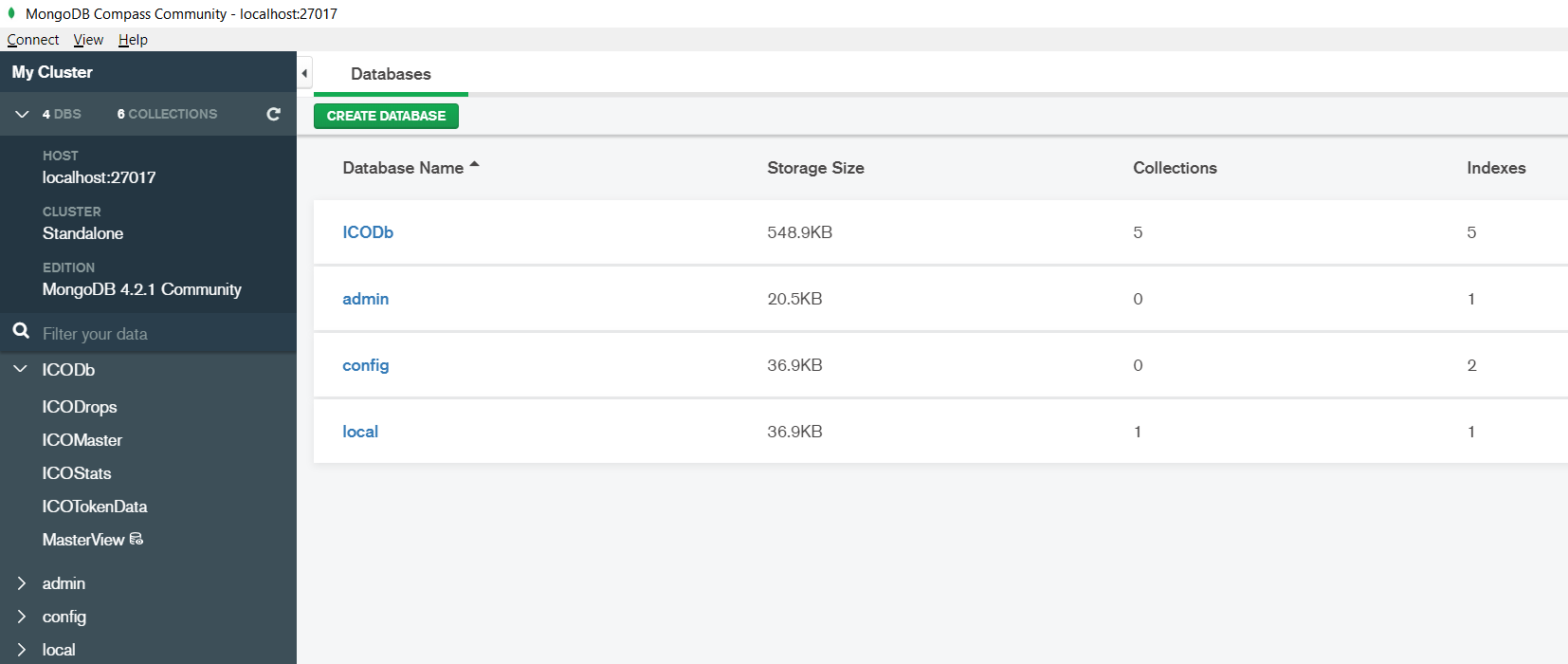
*Hình 3.8: Thông báo khi chạy Docker.*

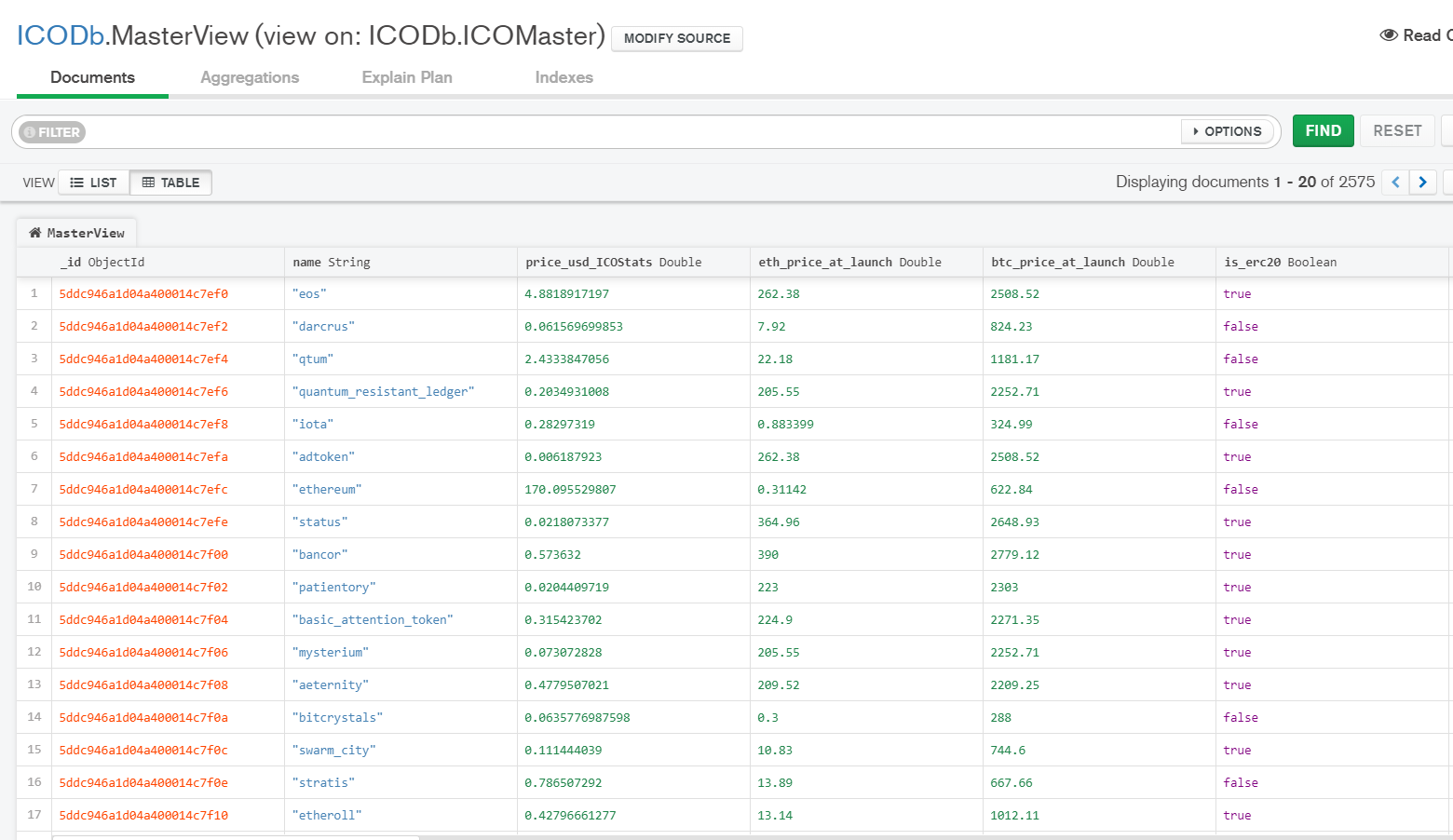
Bước 4: Khởi động ứng dụng MongoDB Compass Community

Nhập Hostname: localhost và Port: 27017 để truy cập cơ sở dữ liệu.



*Hình 3.9: Giao diện cấu hình truy cập cơ sở dữ liệu.*

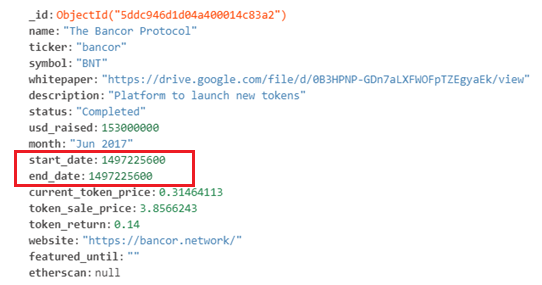
*Hình 3.10: Giao diện Mongodb compass.*



*Hình 3.11: Cơ sở dữ liệu ICODb*

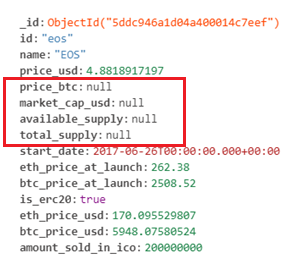
Dữ liệu thu được là thông tin của hơn 2500 ICO. Mặc dù dữ liệu thu thập về lớn nhưng dữ liệu rất thưa thớt do đó cần phải chuẩn hoá dữ liệu. Preparing data hay còn gọi là giai đoạn chuẩn hóa dữ liệu, là quá trình làm sạch và biến đổi dữ liệu thô trước khi xử lý và phân tích chúng. Đây là một bước quan trọng trước khi xử lý và thường liên quan đến việc định dạng lại dữ liệu, chỉnh sửa dữ liệu và kết hợp các bộ dữ liệu để để có được tập dữ liệu hoàn chỉnh. Trong quá trình chuẩn hoá dữ liệu sẽ gặp các hiện tượng như: Dữ liệu không hợp lệ (Unclean data), dữ liệu bị mất (Missing data), dữ liệu bị xung đột (Conflict data).

Dữ liệu không hợp lệ: Các trường dữ liệu start\_date và end\_date có giá trị không hợp lệ.



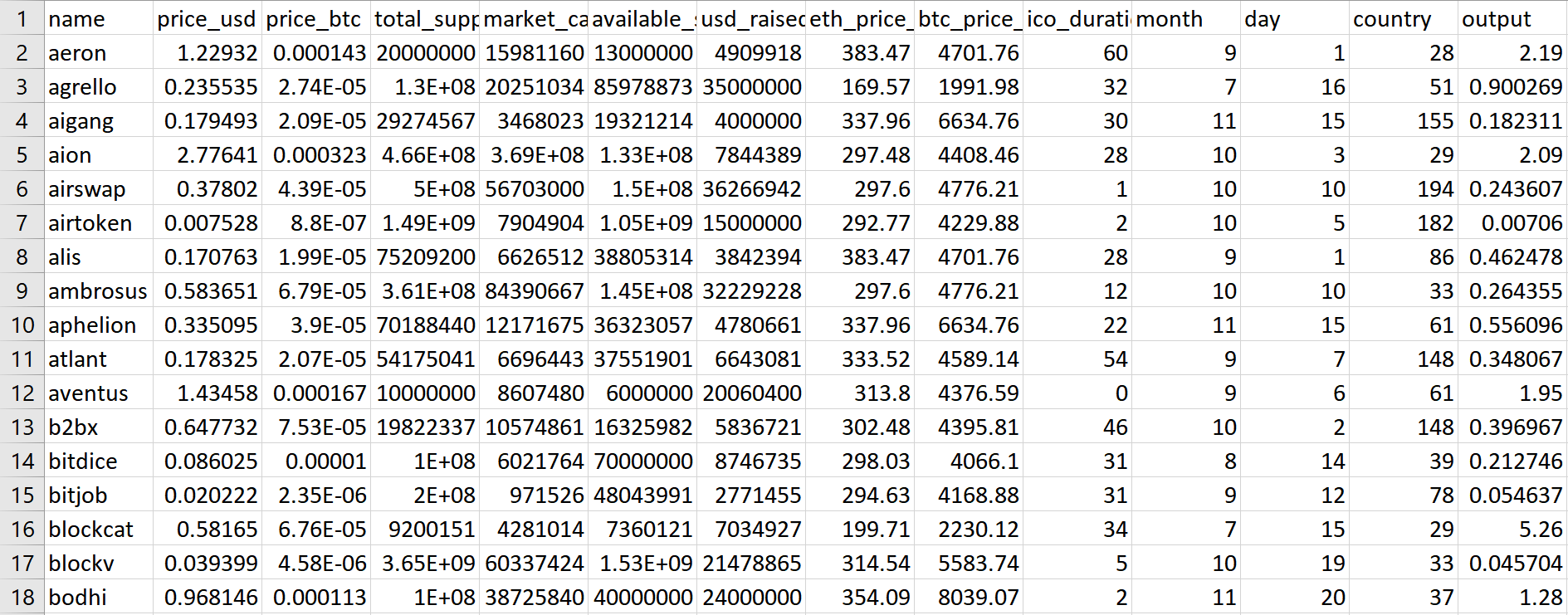
# *Hình 3.12: Dữ liệu không hợp lệ.*

Dữ liệu bị mất: Các trường dữ liệu price\_btc, market\_cap\_usd, available\_supply, total\_supply bị thiếu giá trị.

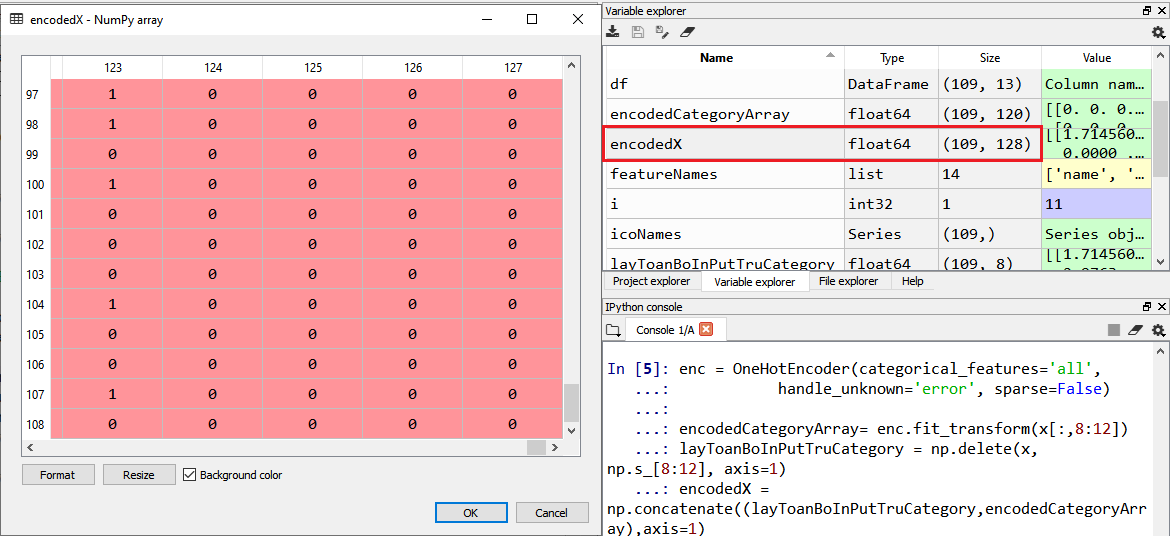
****

# *Hình 3.13: Dữ liệu bị mất.*

Trong quá trình tra cứu thì nhóm đã gặp hiện tượng dữ liệu xung đột. Đó là do các trang web có dữ liệu không đồng nhất. Vì vậy để xử lý ba hiện tượng trên, nhóm sẽ tra cứu dữ liệu từ nhiều nguồn khác nhau, tìm ra dữ liệu chung giữa các trang web và bổ sung chúng vào tập dữ liệu. Sau khi thực hiện xử lí dữ liệu thì dữ liệu sẽ còn 109 ICO.



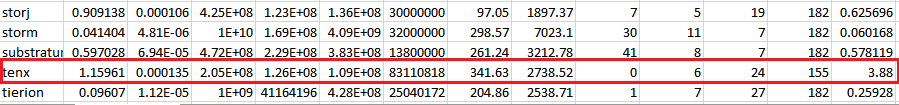
*Hình 3.14: Tập dữ liệu hoàn chỉnh.*



*Hình 3.15: Tập dữ liệu sau khi được mã hóa One hot Encoding.*

Áp dụng kĩ thuật One hot Encoding để phân loại cho các trường dữ liệu như ICO duration, Day, Month và County. Tập dữ liệu có 109 hàng nhưng số cột sẽ tăng từ 12 lên cột. Hình 3.15 với biến encodeX chính là tập dữ liệu sau khi được mã hóa Onehot.

Bốn trường dữ liệu cuối trong tập dữ liệu gồm ICO duration, Day, Month, country là 4 trường dữ liệu phân loại. Việc sử dụng kĩ thuật One hot Encoding để phân loại là để mô hình dự đoán tốt hơn.



*Hình 3.16: Ví dụ về tập dữ liệu.*

Tất cả các mô hình học máy đều thực hiện dựa trên một phương trình toán học, đối với tập dữ liệu của nhóm thì 4 trường dữ liệu cuối là dạng chữ, yêu cầu phải chuyển từ dạng chữ sang dạng số. Hình 3.16, các giá trị của 4 biến phân loại của ICO tenx đó là ico\_duration là 0, month là 6, day là 24, country là 155 (Singapore) và ICO tierion đó là ico\_duration là 1, month là 7, day là 27, country là 182 (United State). Các giá trị của của các biến phân loại thuộc ICO tierion lớn hơn các giá trị của biến phân loại thuộc ICO tenx thì phương trình trong mô hình sẽ nghĩ rằng giá trị ngõ ra của tierion sẽ lớn hơn của tenx và điều này là sai. Từ đó phải sử dụng mã hoá One hot Encoding để tránh hiện tượng này.

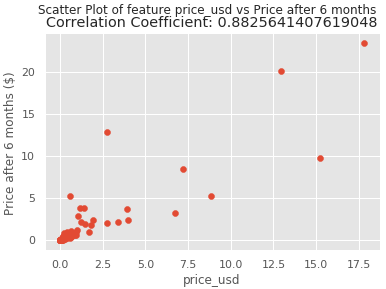
* 1. **PHÂN TÍCH TẬP DỮ LIỆU**
     1. **Phân tích tương quan**

1. **Phân tích tương quan giữa các input và output**

Dựa theo danh sách các yếu tố ảnh hưởng đến output, có thể phân tích độ tương quan mạnh yếu giữa các input với ouput thông qua các biểu đồ sự phân tán và tính toán hệ số tương quan tương ứng giữa các input với ouput.

* **Giá đô la (Price USD)**

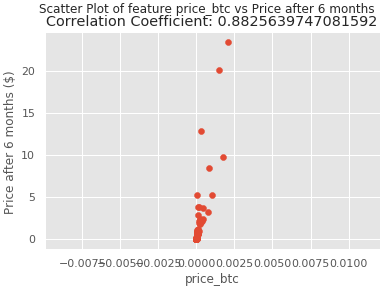
Đây là giá của 1 token được tính dựa trên giá đô la. Hệ số tương quan giữa 2 biến giá đô la và giá trị ngõ ra là 0.882. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan mạnh và là sự tương quan dương, nghĩa là giá đô la tăng dẫn đến giá trị ngõ ra tăng.

****

*Hình 3.17: Sơ đồ phán tán giữa giá usd và giá trị ngõ ra.*

* **Giá bitcoin (Price BTC)**

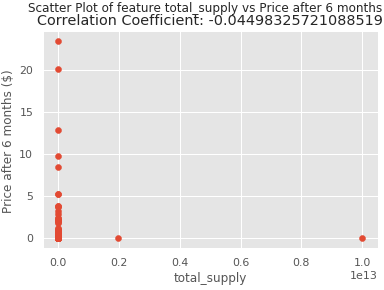
Đây là giá của của token được tính theo bitcoin. Hệ số tương quan giữa 2 biến giá bitcoin và giá trị ngõ ra là 0.882. Dựa theo hế số tương quan thì 2 biến này có sự tương quan mạnh.

****

*Hình 3.18: Sơ đồ phân tán giữa giá bitcoin và giá trị ngõ ra*

* **Total Supply**

Total Supply hay Max Supply, là tổng số token có thể cung ứng cho thị trường. Hệ số tương quan giữa 2 biến Total Supply và giá trị ngõ ra là -0.0452. Dựa theo hệ số tương quan thì 2 biến này gần như không có sự tương quan.

****

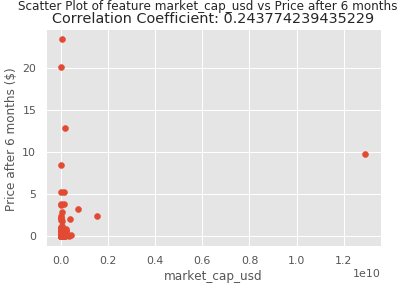
*Hình 3.19: Sơ đồ phân tán giữa biến total\_supply và giá trị ngõ ra*

* **Market Cap**

[Market Cap](https://vnsending.com/market-cap-la-gi-volume-24h-la-gi-cac-khai-niem-khi-dau-tu-coin.html) là tổng vốn hóa thị trường được tính dựa theo công thức:

*Market cap = Giá token tại thời điểm bán \* Số lượng token đã giao dịch*

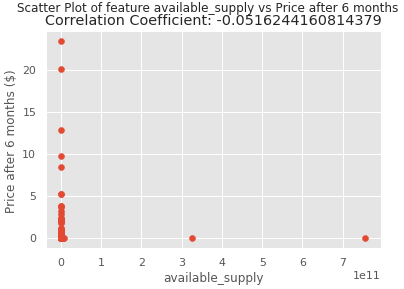
Hệ số tương quan giữa 2 biến market\_cap\_usd và giá trị ngõ ra là 0.242. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.20: Sơ đồ phân tán giữa biến market\_cap\_usd và giá trị ngõ ra*

* **Available Supply**

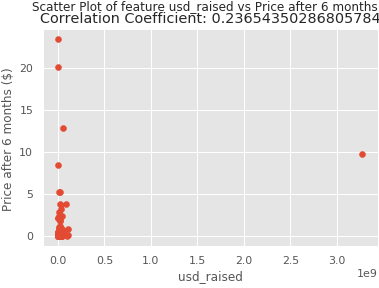
[Available supply](https://cafebitcoin.info/trading/cam-nang-su-dung-coinmarketcap/) là lượng token đã khai thác được và đang lưu thông trên thị trường. Hệ số tương quan giữa 2 biến là -0.0519. Dựa theo hệ số tương quan thì 2 biến này gần như không có sự tương quan với nhau.

****

*Hình 3.21: Sơ đồ phân tán giữa biến available\_supply giá trị ngõ ra*

* **USD raised**

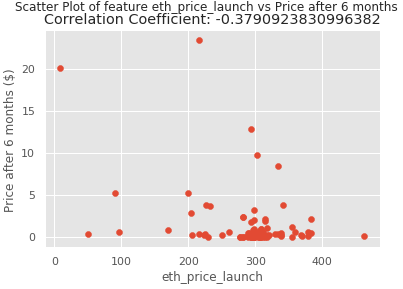
USD raised là số tiền (tính bằng đô la) mà công ty phát hành đồng tiền đó nhận được kể từ lúc phát hành đến lúc kết thúc mở bán. Hệ số tương quan giữa 2 biến usd\_raised và giá trị ngụ ra là 0.236. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.22: Sơ đồ phân tán giữa biến usd\_raised và giá trị ngõ ra*

* **Ethereum price at launch**

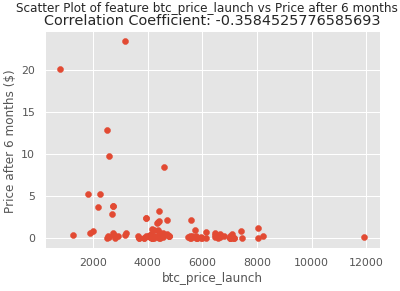
Đây là giá của Ethereum tại thời điểm token được phát hành. Hệ số tương quan giữa 2 biến eth\_price\_launch và giá trị ngõ ra là -0.3790. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.23: Sơ đồ phân tán giữa biến eth\_price\_launch và giá trị ngõ ra.*

* **Bitcoin was at launch**

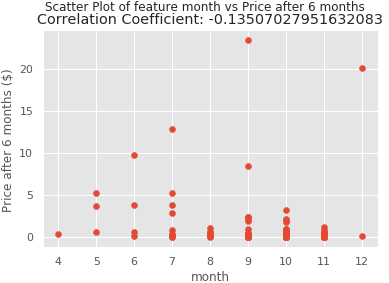
Đây là giá của Bitcoin tại thời điểm token được phát hành. Hệ số tương quan giữa 2 biến btc\_price\_launch và giá trị ngõ ra là -0.3584. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.24: Sơ đồ phân tán giữa biến btc\_price\_launch và giá trị ngõ ra.*

* **Month ICO was launched**

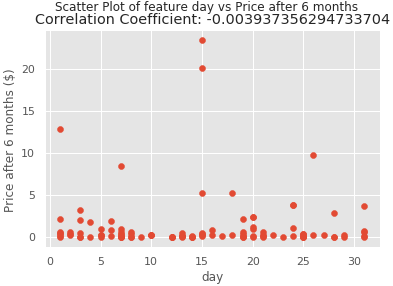
Month ICO was launched là tháng phát hành ICO. Hệ số tương quan giữa 2 biến month và giá trị ngõ ra là -0.135. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.25: Sơ đồ phân tán giữa biến month và giá trị ngõ ra.*

* **Date ICO was launched**

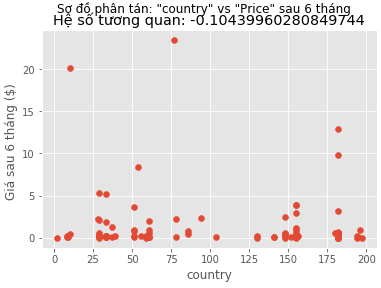
Date ICO was launched là ngày phát hành ICO. Hệ số tương quan giữa 2 biến day và giá trị ngõ ra là -0.0039. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.

****

*Hình 3.26: Sơ đồ phân tán giữa day và giá trị ngõ ra*

* **Country an ICO was launched from**

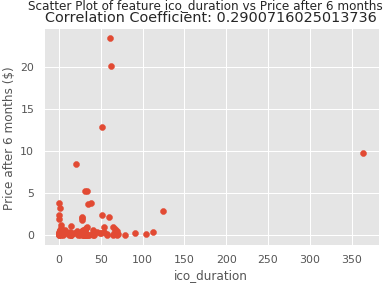
Country an ICO was launched from là quốc giá phát hành ICO. Hệ số tương quan giữa 2 biến country và giá trị ngõ ra là -0.104. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu.



*Hình 3.27: Sơ đồ phân tán giữa country và giá trị ngõ ra*

* **ICO Duration in day**

ICO Duration in day là tổng thời gian phát hành ICO tính theo ngày. Hệ số tương quan giữa 2 biến ico\_duration và giá trị ngõ ra là 0.29. Dựa theo hệ số tương quan thì 2 biến này có sự tương quan yếu

****

*Hình 3.28: Sơ đồ phân tán giữa biến ico\_duration và giá trị ngõ ra.*

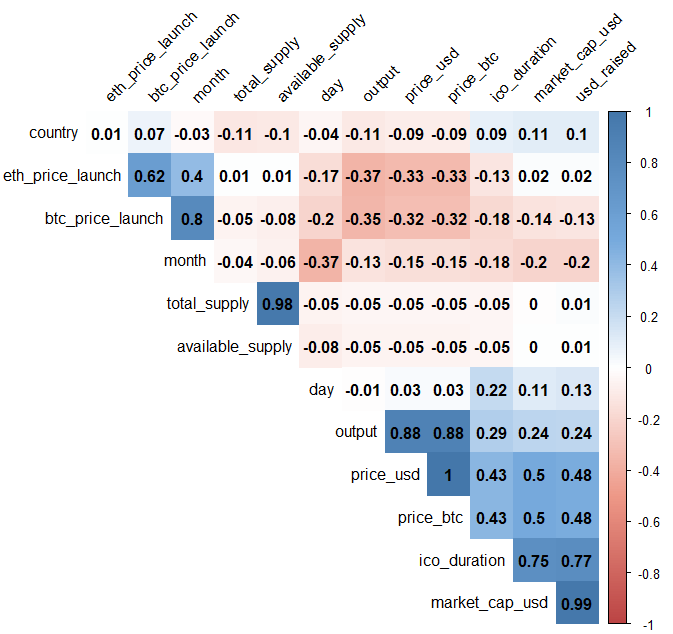
**Nhận xét:**

Sau khi phân tích sự tương quan giữa các giá trị ngõ vào với các giá trị ngõ ra, có thể nhận thấy có 2 giá trị ngõ vào là giá USD (Price USD) và giá Bitcoin (Price BTC) có sự tương quan mạnh với ngõ ra, trong khi các giá trị ngõ vào còn lại thì có sự tương quan yếu hoặc gần như là không có sự tương quan.

Nếu hệ số tương quan giữa một giá trị ngõ vào với giá trị ngõ ra cao thì có thể xây dựng một mô hình hồi quy đa biến thông thường để dự đoán và đạt được kết quả tốt. Các dữ liệu ngõ vào trong tập dữ liệu mà nhóm thu thập được thì có nhiều dữ liệu ngõ vào có mối liên hệ với ngõ ra, tuy là mức độ tương quan yếu nhưng vẫn có sự tương quan với nhau. Ví dụ có 2 dữ liệu ngõ vào có hệ số tương quan cao đó là price\_usd và price\_btc nhưng không thể chỉ dự đoán ngõ ra thông qua 2 dữ liệu này và xem các dữ liệu còn lại vô nghĩa

Tổng quát:

Hình 3.27 diễn tả tổng quát sự tương quan giữa các dữ liệu ngõ vào với ngõ ra cũng như giữa các dữ liệu ngõ vào với nhau. Những cặp có hệ số tương quan lớn thì màu sẽ tiến về màu xanh dương đậm hoặc màu đỏ đậm (tương ứng 1 và -1). Những cặp có hệ số tương quan yếu thì màu sẽ nhạt dần còn những cặp có hệ số tương quan gần như bằng 0 thì màu sẽ là trắng.

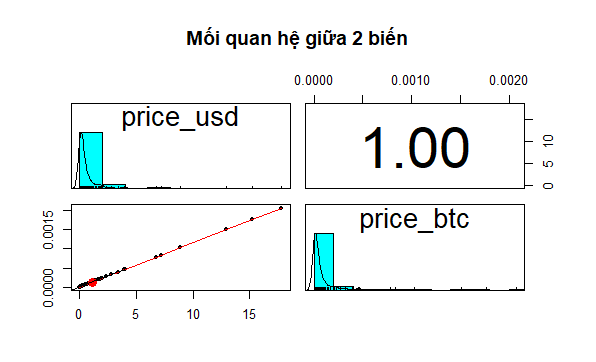


*Hình 3.29: Biểu đồ tương quan giữa giá trị ngõ vào và giá trị ngõ ra*

* + 1. **Đa cộng tuyến**

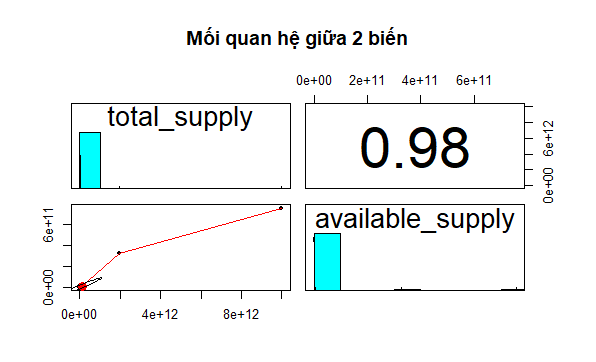
Dựa vào hình 43 có thể xác định các biến ngõ vào nào có mối liên hệ với nhau, đó là:

* Hai biến price\_usd & price\_btc



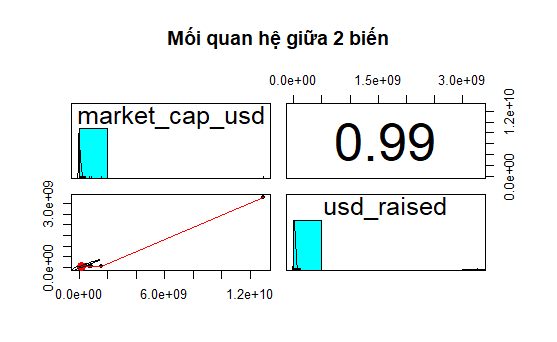
*Hình 3.30: Mối quan hệ giữa price\_usd và price\_btc.*

* Hai biến total\_supply & available\_supply

****

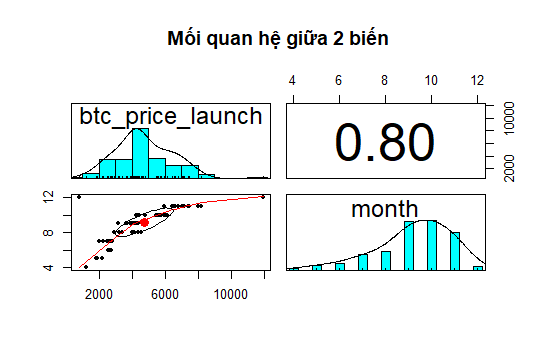
*Hình 3.31: Mối quan hệ giữa total supply và available\_supply.*

* Hai biến market\_cap\_usd & usd\_raised



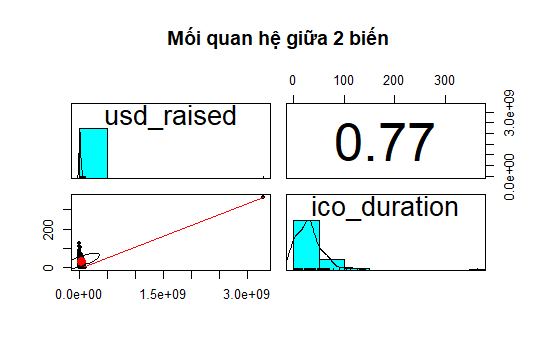
*Hình 3.32: Mỗi quan hệ giữa market\_cap\_usd và usd\_raised*.

* Hai biến btc\_price\_launch & month



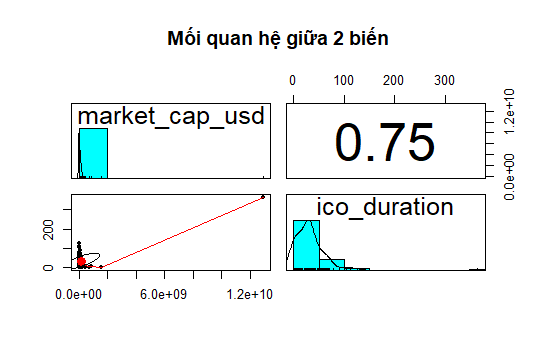
*Hình 3.33: Mối quan hệ giữa btc\_price và month.*

* Hai biến usd\_raised & ico\_duration



*Hình 3.34: Mối quan hệ giữa usd\_raised và ico\_duaration.*

* Hai biến market\_cap\_usd & ico\_duration.



*Hình 3.35: Mối quan hệ giữa market\_cap\_usd và ico\_duration.*

Nhóm sẽ phân tích rõ 2 trường hợp là các biến gần như độc lập và các biến có hệ số tương quan cao.

1. **Các biến gần như độc lập với nhau**

Ý nghĩa các thông số:

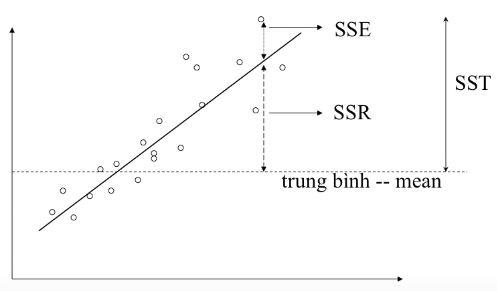
* Residual**:**  Là phần khác biệt giữa giá trị quan sát thực tế với giá trị được tiên lượng.
* Phần dư dao động trong khoảng Min và Max. Hai đại lượng 1Q và 3Q là hai chỉ số chất lượng (*qualities*)**.**
* 1Q:Có nghĩa là 25% đối tượng nghiên cứu có phần khác biệt giữa giá trị mẫu quan sát thực tế với giá trị tiên lượng nhỏ hơn hoặc bằng -1.1596.
* 3Q: Có nghĩa là 75% đối tượng nghiên cứu có phần khác biệt giữa giá trị mẫu quan sát thực tế với giá trị tiên lượng nhỏ hơn hoặc bằng -0.4531.
* Median: Có nghĩa là 50% đối tượng nghiên cứu có phần khác biệt giữa giá trị mẫu quan sát thực tế với giá trị tiên lượng nhỏ hơn hoặc bằng -0.9838.
* Hệ số (Cofficients):Là phần tiếp theo gồm có 4 phần nhỏ là ước tính (estimated), sai số chuẩn (std. Error)**,** giá trị t(t-value) và giá trị P (P-value)**.**
* Ước tính (estimated):Là giá trị tham số mô hình được ước tính, gồm giá trị hệ số chặn (intercept) và giá trị hệ số market\_cap\_usd.
* Sai số chuẩn (Std. Error): Là hệ số dùng để đo độ chính xác của hệ số hồi quy ước tính. Giá trị này cảng nhỏ thì hệ số hồi quy ước tính càng chính xác.
* t- value: Là giá trị kiểm định. Giá trị kiểm định này nhỏ hơn 2 thì có ý nghĩa thống kê.

Công thức để tính giá trị này

*t-value =* (3.1)

Trong đó: là hệ số hồi quy ước tính và *Std. Error* là sai số hệ số hồi quy.

* p-value: Đây cũng là giả trị kiểm định thống kê. Nếu giá trị p tính toán nhỏ hơn alpha bằng 0.05 thì input có ý nghĩa thống kê.
* **Phân tích phương sai**
* Df:Là bậc tự do. Trong ví dụ trên chúng ta có mẫu n = 109, thì df = 109 – 1 = 108. Vì là biến liên tục nên số bậc tự do của market\_cap\_usd là 1, số bậc tự do của residual là 108– 1 (df của market\_cap\_usd) = 107.
* Sum of Squared:



*Hình 3.36: Biểu diễn Sum of Squared.*

Sum of squared total (SST) là tổng của SSE (Sum of squared error) và SSR (Sum of squared regression):

*SST = SSE + SSR* (3.2)

Trong đó: SSE là tổng sai số bình phương giữa giá trị thực tế với giá trị dự đoán.

SSR: Là khoảng cách từ điểm dự đoán đến đường trung bình (đường đứt đoạn).

* Mean Sq: là trung bình bình phương sai số. Là khoảng cách trung bình từ 1 mẫu quan sát thực tế đến đường dự đoán.

(3.3)

Trong đó:

* SSE là tổng sai số bình phương tức giữa giá trị thực tế với giá trị dự đoán.
* N: Kích thước mẫu.
* y: giá trị dự đoán.
* : giá trị thực tế.
* **Hệ số R2**: Được xác định theo công thức:

*R2 = 1 -* (3.4)

Trong đó:

* SSE là tổng sai số bình phương tức giữa giá trị thực tế với giá trị dự đoán.
* SST: Tổng của SSE và SSR

Ý nghĩa: Phần trăm sự biến động của giá trị ngõ ra được giải thích bởi các giá trị ngõ vào, nghĩa là mô hình phù hợp với bao nhiều phần trăm dữ liêu. Thông số có giá trị càng gần 1 thì mô hình càng tốt.

* **Hệ số Adjust R2:** Là hệ số xác định điều chỉnh, có công thức:

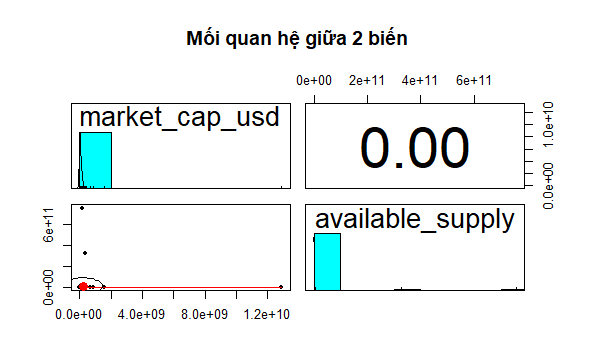
*R2adj =*  (3.5)

Trong đó:

* R2 là hệ số xác định bên trên.
* N là kích thước mẫu.
* p là số lượng các biến dự đoán.

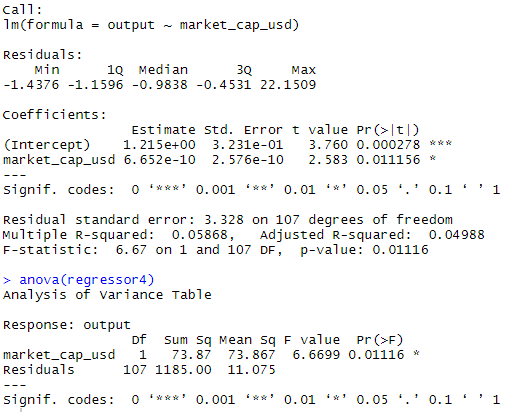
Ý nghĩa: Liên quan đến việc thêm biến vào mô hình. Giá trị này so sánh với giá trị R2, xem xét sự chênh lệch giữa chúng rồi từ đó kết luận việc thêm biến vào mô hình có nên hay không.

Dựa theo hình mô tả sự tương quan giữa các biến, chọn 2 biến market\_cap\_usd và available\_supplyvì nó có hệ số tương quan là 0.00006.

****

*Hình 3.37: Mối quan hệ giữa market\_cap\_usd và available\_supply.*

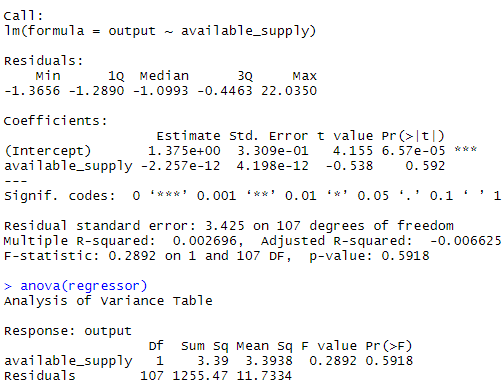
* Giả sử mô hình chỉ có một biến là market\_cap\_usd và ngõ ra.



*Hình 3.38: Phân tích tổng quan giữa biến market\_cap\_usd và ngõ ra.*

Tổng quan:

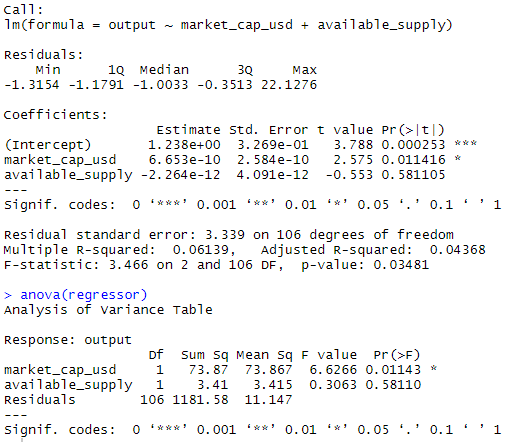
* Hệ số hồi quy của market\_cap\_usd: 6.652e – 10.
* Sai số chuẩn của hệ số hồi quy: 3.231e – 01.
* Giá trị t: 2.583.
* Giá trị p: 0.011156.
* R2: 0.05868.
* Adjusted R2: 0.04988.
* Sum of Squared: 73.87.
* Mean Squared: 73.867.
* Giá sử mô hình chỉ có 1 biến là available\_supplyvàngõ ra.

****

*Hình 3.39: Phân tích tổng quan giữa biến available\_supply và ngõ ra.*

Tổng quan:

* Hệ số hồi quy của available\_supply: -2.257e - 12.
* Sai số chuẩn của hệ số hồi quy: 4.198e – 12.
* Giá trị t: -0.538.
* Giá trị p: 0.592.
* R2: 0.02696.
* Adjusted R2: -0.006625.
* Sum of Squared: 3.39.
* Mean Squared: 3.3938.
* Giả sử mô hình có 2 biến là market\_cap\_usd, available\_supplyvà ngõ ra.



*Hình 3.40: Phân tích tổng quan giữa 2 biến market\_cap\_usd và available\_supply với ngõ ra.*

Tổng quan:

* Hệ số hồi quy của available\_supply: -2.265e - 12.
* Hệ số hồi quy của market\_cap\_usd: 6.653e – 10.
* Sai số chuẩn của hệ số hồi quy của available\_supply: 4.092e – 12.
* Sai số chuẩn của hệ số hồi quy market\_cap\_usd: 2.584e – 10.
* Giá trị t của available\_supply: -0.553.
* Giá trị t của market\_cap\_usd: 2.575.
* Giá trị p của available\_supply: 0.581137.
* Giá trị p của market\_cap\_usd: 0.011416.
* R2: 0.06139.
* Adjusted R2: 0.04368.
* Sum of Squared của available\_supply: 3.41.
* Sum of Squared của market\_cap\_usd: 73.87.
* Mean Squared của available\_supply: 3.414.
* Mean Squared của market\_cap\_usd: 73.867.

Nhận xét:

Giả sử hệ số hồi quy của market\_cap\_usd là X1, hệ số hồi quy của available\_supply là X2, Se(X1) là sai số chuẩn hệ số hồi quy của X1, Se(X2) là sai số chuẩn hệ số hồi quy của X2.

Dựa theo kết quả bảng 4.1 có thể kết luận với các ngõ vào độc lập với nhau thì các giá trị hệ số hồi quy, sai số chuẩn hệ số hồi quy và giá trị Sum Sq gần như không thay đổi.

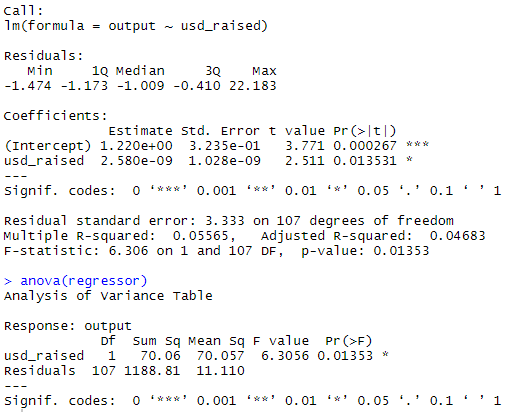
Bảng 3.1: Xem xét sự thay đổi các giá trị của các biến market\_cap\_usd và available\_supply.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mô hình** | **X1** | **Se(X1)** | **X2** | **Se(X2)** | **Sum Sq** |
| market\_cap\_usd | 6.652e-10 | 2.57e-10 | -- | -- | 73.87 |
| available\_supply | -- | -- | -2.25e-12 | 4.198e-12 | 3.39 |
| Mô hình gồm 2 biến trên | 6.653e-10 | 2.584e-10 | -2.26e-12 | 4.092e-12 | 73.87/3.41 |

1. **Các biến có tương quan với nhau**

Dựa theo hình mô tả sự tương quan giữa các biến, chọn 2 biến usd\_raised và ico\_durationvì nó này có hệ số tương quan gần bằng 0.77.

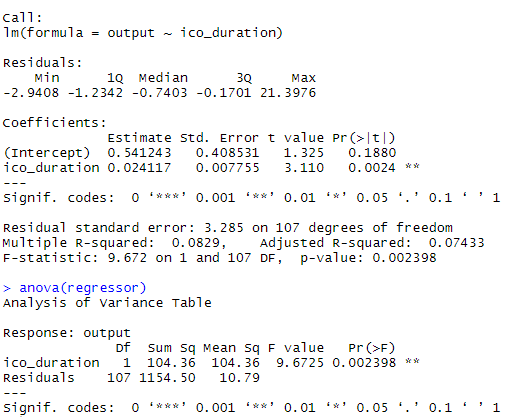
* Giả sử mô hình chỉ có biến usd\_raised và ngõ ra.



*Hình 3.41: Phân tích tổng quan giữa biến used\_raised và ngõ ra.*

Tổng quan:

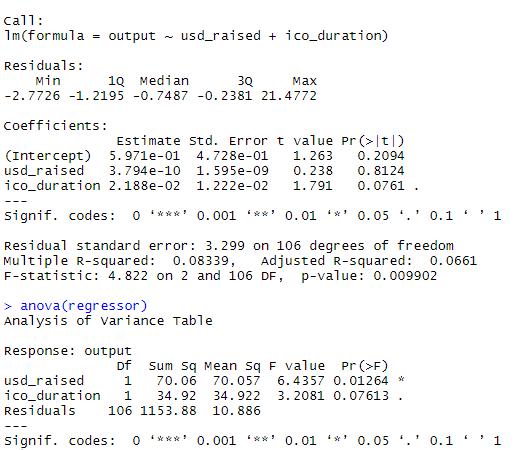
* Hệ số hồi quy của market\_cap\_usd: 2.508e – 09.
* Sai số chuẩn của hệ số hồi quy: 1.028e – 09.
* Giá trị t: 2.511.
* Giá trị p: 0.013531.
* R2: 0.05565.
* Adjusted R2: 0.04683.
* Sum of Squared: 70.06.
* Mean Squared: 70.057.
* Giả sử mô hình chỉ có biến ico\_duration và ngõ ra.



*Hình 3.42: Phân tích tổng quan giữa biến ico\_duration và ngõ ra.*

Tổng quan:

* Hệ số hồi quy của market\_cap\_usd: 2.508e – 09.
* Sai số chuẩn của hệ số hồi quy: 1.028e – 09.
* Giá trị t: 2.511.
* Giá trị p: 0.013531.
* R2: 0.05565.
* Adjusted R2: 0.04683.
* Sum of Squared: 70.06.
* Mean Squared: 70.057.
* Giả sử mô hình có 2 biến usd\_raised, ico\_duration và ngõ ra.



*Hình 3.43: Phân tích tổng quan giữa 2 biến usd\_raised và ico\_duration với ngõ ra.*

Tổng quan:

* Hệ số hồi quy của *usd\_raised*: 3.794e - 10.
* Hệ số hồi quy của *ico\_duration*: 2.188e – 02.
* Sai số chuẩn của hệ số hồi quy của *usd\_raised*: 1.595e – 09.
* Sai số chuẩn của hệ số hồi quy *ico\_duration*: 1.222e – 02.
* Giá trị t của *usd\_raised*: 0.283.
* Giá trị t của *ico\_duration*: 1.791.
* Giá trị p của *usd\_raised*: 0.8124.
* Giá trị p của *ico\_duration*: 0.0761.
* R2: 0.08339.
* Adjusted R2: 0.0661.
* Sum of Squared của *usd\_raised*: 70.06.
* Sum of Squared của *ico\_duration*: 34.92.
* Mean Squared của *usd\_raised*: 70.057.
* Mean Squared của *ico\_duration*: 34.922.

Nhận xét:

Giả sử hệ số hồi quy của usd\_raised là X1, hệ số hồi quy của ico\_duration là X2, Se(X1) là sai số chuẩn hệ số hồi quy của X1, Se(X2) là sai số chuẩn hệ số hồi quy của X2.

Bảng 3.2: Xem xét sự thay đổi các giá trị của các biến ico\_duration và usd\_raised.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mô hình** | **X1** | **Se(X1)** | **X2** | **Se(X2)** | **Sum Sq** |
| usd\_raised | 2.580e-09 | 1.028e-09 | -- | -- | 70.06 |
| ico\_duration | -- | -- | 0.024117 | 0.007755 | 104.36 |
| Mô hình gồm 2 biến trên | 3.794e-10 | 1.595e-09 | 2.188e-02 | 1.222e-02 | 70.06/34.92 |

Khi sử dụng mô hình có 2 biến có sự tương quan mạnh thì các giá trị hệ số hồi quy, sai số chuẩn hệ số hồi quy và giá trị Sum Sq thay đổi.

* + - 1. **Ảnh hưởng đa cộng tuyến**
* Tham số mô hình bị thay đổi

Bảng 3.3: So sánh hệ số hồi quy của usd\_raised và ico\_duration trước và sau khi kết hợp.

|  |  |  |
| --- | --- | --- |
| **Các biến trong mô hình** | **Hệ số hồi quy usd\_raised** | **Hệ số hồi quy ico\_duration** |
| usd\_raised | 2.580e-09 | -- |
| ico\_duration | -- | 0.024117 |
| Mô hình có cả 2 biến | 3.794e-10 | 2.188e-02 |

Khác với 2 biến gần như độc lập thì 2 biến có mối tương quan cao khi kết hợp thì hệ số hồi quy khác biệt rất nhiều so với khi sử dụng mô hình hồi quy từng biến độc lập. Hệ số hồi quy của usd\_raised và ico\_duration đã giảm đi khi đưa vào mô hình.

* Độ chính xác của các tham số mô hình (hệ số hồi quy) sẽ giảm.

Bảng 3.4: So sánh sai số chuẩn hệ số hồi quy của usd\_raised và ico\_duration trước và sau khi kết hợp.

|  |  |  |
| --- | --- | --- |
| **Các biến trong mô hình** | **Se(usd\_raised)** | **Se(ico\_duration)** |
| usd\_raised | 1.028e-09 | -- |
| ico\_duration | -- | 0.007755 |
| Mô hình có cả 2 biến | 1.595e-09 | 1.222e-02 |

Sự thay đổi sai số chuẩn của hệ số hồi quy có thể ảnh hưởng đến độ chính xác của hệ số. Khi sử dụng mô hình hồi quy đơn biến nếu chỉ sử dụng ico\_duration thì sai số (Standard Error) là 0.007755, nhưng khi đưa vào mô hình hồi quy đa biến thì lại tăng lên thành 0.0122. Tương tự cho tham số usd\_raised thì sai số cũng tăng nhưng không đáng kể.

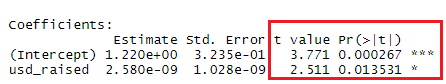
* Sum of Squared Regression (SSR) thay đổi.

Bảng 3.5**:** So sánh SSR của 2 biến usd\_raised và ico\_duration trước và sau khi kết hợp.

|  |  |  |
| --- | --- | --- |
| **Các biến trong mô hình** | **SSR của usd\_raised** | **SSR của ico\_duration** |
| usd\_raised | 70.06 | -- |
| ico\_duration | -- | 104.36 |
| Mô hình có cả 2 biến | 70.06 | 34.92 |

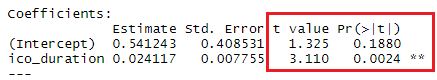
Mô hình có 2 biến kết hợp lại với nhau thì giá trị SSR của ico\_duration giảm so với mô hình chỉ có biến ico\_duration. Điều này ảnh hưởng đến mức độ phù hợp dữ liệu của mô hình.

* Vì độ chính xác giảm nên có thể dẫn đến kết luận rất khác nhau có khi không có ý nghĩa thống kê.
* Mô hình chỉ có usd\_raised:



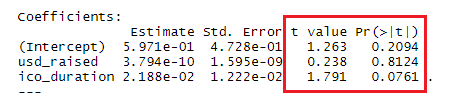
*Hình 3.44: Giá trị t và p trước khi ghép của usd\_raised.*

* Mô hình chỉ có ico\_duration:



*Hình 3.45: Giá trị t và p trước khi ghép của ico\_duration.*

* Mô hình có cả 2 biến usd\_raised và ico\_duration:



*Hình 3.46: Giá trị t và p sau khi ghép của usd\_raised và ico\_duration*

Đối với mô hình hồi quy đơn biến áp dụng cho từng biến usd\_raised và ico\_duration thì có thể thấy 2 biến này có các trị số t và p thỏa điều kiện là t > 2 và p < 0.05 nên có ý nghĩa thống kê. Khi kết hợp vào hồi quy đa biến thì cả 2 biến không còn ý nghĩa thống kê.

* 1. **XÂY DỰNG THUẬT TOÁN**

Sau khi phân tích dữ liệu trong tập dữ liệu, nhóm tiến hành thiết kế và xây dựng thuật toán hồi quy Ridge và mạng Nơron để dự đoán.

* + 1. **Hồi quy Ridge**

Với hồi quy đa biến thì nghiệm phương trình mất mát J() là :

(3.6)

Phương trình trên có nghiệm khi và chỉ khi tích của ma trận X với chuyển vị của nó khả nghịch, nghĩa là *det(* > 0 nhưng đối với tập dữ liệu thì số lượng ngõ vào sau khi mã hóa là 128, nhưng số mẫu là 109. Điều này sẽ tạo ra một ma trận có có kích thước 109x128 với số hàng nhỏ hơn số cột. Nếu một ma trận có số hàng nhỏ hơn số cột thì *det(* = 0 dẫn đến () không khả nghịch.

Hướng giải quyết:

* Xóa một vài biến thừa.

Không thể sử dụng cách này đối với tập dữ liệu vì các ngõ vào có sự tương quan với ngõ ra, ví dụ 2 ngõ vào price\_usd và price\_btc đều có sự tương quan rất lớn đến ngõ ra, được thể hiện qua hệ số tương quan. Nếu bỏ đi sẽ gấy ảnh hưởng đến kết quả.

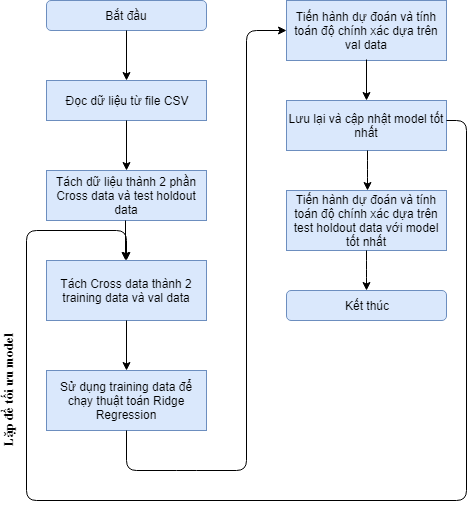
* Sử dụng kỹ thuật Regularization.

Ý tưởng của kỹ thuật Regularization: Nếu giá trị của các tham số nhỏ sẽ dẫn đến.

* Hàm giả thuyết (dự đoán) sẽ đơn giản hơn.
* Tránh được hiện tượng quá khớp.

Nhận xét:

Sau khi mã hóa onehot, tập dữ liệu có nhiều ngõ vào, rất khó chọn ra ngõ vào nào ít liên quan và cũng rất khó để biết tham số nào cần được giảm, thực hiện cách lấy *cost function* dựa trên công thức (2.11). Hiệu chỉnh cost function bằng cách cộng thêm một cụm ở sau. Số hạng này sẽ giảm tất cả các tham số từ .



*Hình 3.47: Lưu đồ thuật toán hồi quy Ridge.*

**Giải thích:**

* Đọc dữ liệu từ file csv. Trong đó là tất cả 109 ICO nhưng sẽ tách ra 3 ICO để làm 3 ICO mẫu dự đoán, còn lại là 106 ICO dùng để xây dựng mô hình.
* Tách dữ liệu từ datatset ra làm 2 phần: Tập cross (X\_cross, y\_cross) và tập test holdout (X\_test\_holdout, y\_test\_holdout) với tỷ lệ 80% cho tập cross và 20% cho tập test holdout.
* Từ tập dữ liệu Cross tiến hành tách ra 2 tập con: Tập training (X\_train, y\_train) và tập test (X\_val, y\_val) với tỷ lệ 80% cho tập training và 20% cho tập val.
* Sử dụng tập training để chạy thuật toán hồi quy Ridge, mô hình sẽ sử dụng X\_train và y\_train để tiến hành huấn luyện.
* Mô hình sau khi được huấn luyện sẽ dự đoán bằng cách sử dụng X\_val. Sau đó sẽ kiểm nghiệm độ chính xác giũa giá trị dự đoán với giá trị thực tế là y\_val bằng cách sử dụng đến các phương pháp đánh giá mô hình như rMSE, MAE và R2.
* Lưu mô hình dưới định dạng file sav. Tạo một vòng lặp để thực hiện lại công việc giống như trên hình. Điều này giúp tối ưu mô hình, ý tưởng là sẽ lưu mô hình nào có hiệu năng tốt nhất tức là sẽ có chỉ số rMSE, MAE hoặc R2 tối ưu nhất. Mỗi lần lặp sẽ là mỗi lần lấy mẫu khác nhau nên sẽ sau khi tính toán các giá trị rMSE, MAE và R2 của mô hình sẽ thay đổi. Sau khi kết thúc vòng lặp sẽ thu được 2 mô hình tối ưu nhất với mô hình có chỉ số rMSE, MAE và mô hình có chỉ số R2 tốt nhất.
* Sử dụng tập X\_test\_holdout đưa vào 2 mô hình tốt nhất, sau đó tính toán hiệu năng thông qua 2 chỉ số rMSE, MAE và R2 giữa giá trị dự đoán với giá trị thực tế y\_test\_holdout.

Việc lựa chọn tham số *λ* rất quan trọng vì tham số này ảnh hưởng đến 2 mục tiêu:

* Áp dụng tốt cho tập huấn luyện.
* Giữ các tham số nhỏ giúp hàm giả thuyết đơn giản hơn tránh được hiện tượng quá khớp.

Nếu *λ* quá lớn thì các tham số sẽ giảm rất mạnh dẫn đến gần bằng 0, hàm giả thuyết sẽ là . Lúc này sẽ xuất hiện hiện tượng *chưa khớp*.

Nếu *λ* quá nhỏ thì sẽ quay lại bài toán hồi quy đa biến thông thường và sẽ dẫn đến hiện tượng quá khớp.

Việc chọn *λ* được nhóm chọn thông qua việc chạy thực nghiệm.

### Mạng Nơron

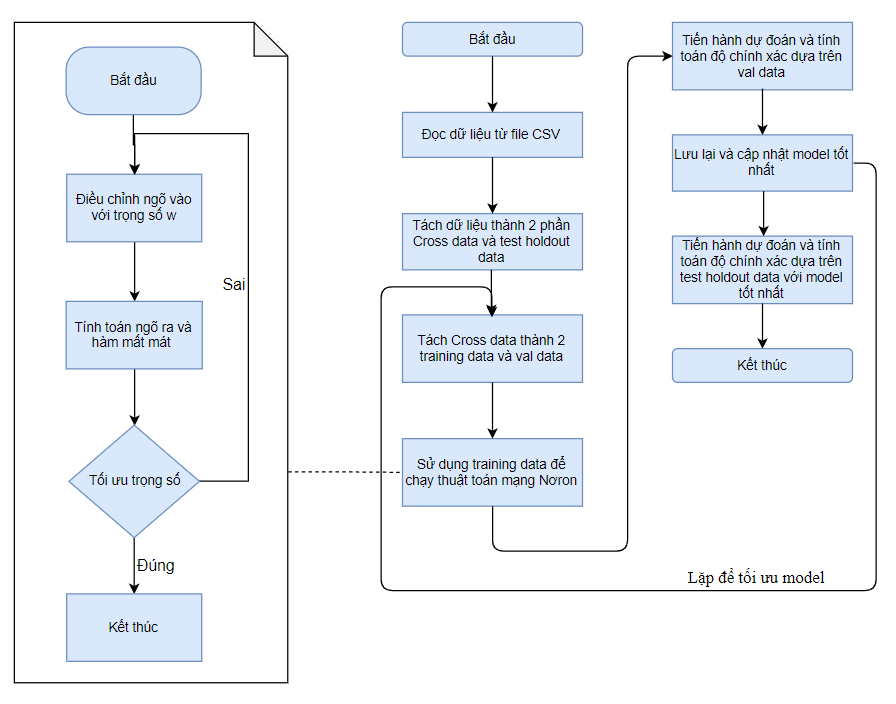
Trong mô hình mạng Nơron, để có thể xác định số lượng lớp ẩn, số nút trong lớp ẩn thì nhóm đã chạy mô hình nhiều lần và tìm ra các thông số phù hợp, cho ra kết quả chính xác nhất.

Với mô hình dự đoán giá ICO, nhóm đã chọn các thông số như sau:

* Lớp ngõ vào: 128 trường dữ liệu.
* Số lớp ẩn: 3.
* Số lượng nút mạng trong lớp ẩn: 100.
* Lớp ngõ ra: 1.
* Hàm kích hoạt được áp dụng trong lớp ẩn: hàm Tanh.

Quy trình thực hiện của thuật toán mạng Nơron:

* Giai đoạn lan truyền tiến: Mô hình sẽ tính toán ngõ vào với trọng số w, sau đó truyền sang lớp ẩn, kết hợp với hàm Tanh để tính toán giá trị ở lớp ẩn, lớp ngõ ra, hàm mất mát.
* Giai đoạn lan truyền ngược: Cập nhật lại trọng số đồ thị sử dụng kĩ thuật Gradient Descent để hàm mất mát giảm. Quá trình cập nhật các trọng số sẽ bắt đầu từ trọng số ở lớp cuối cùng. Cập nhật xong trọng số này, mô hình sẽ quay ngược lên lớp trước đó để cập nhật tiếp. Khi hàm mất mát bão hoà thì trọng số hiện tại sẽ là trọng số thích hợp cho mô hình. Khi đó thuật toán sẽ kết thúc.

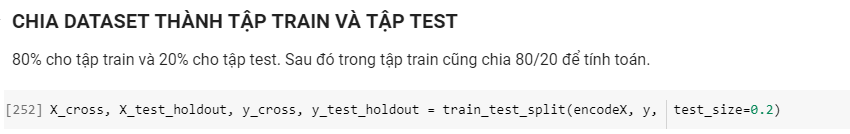


*Hình 3.48: Lưu đồ thuật toán mạng Nơron.*

**CHƯƠNG 4: KIỂM TRA THỰC NGHIỆM VÀ ĐÁNH GIÁ**

1. **OVERFITTING**

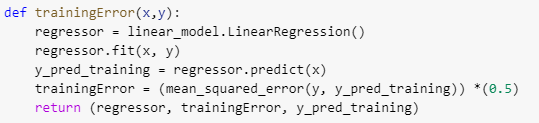
Hiện tượng quá khớp (*overfitting*), là hiện tượng dự đoán trên tập huấn luyện rất tốt nhưng dự đoán bộ dữ liệu chưa nhìn thấy thì lại cho kết quả rất tệ. Nhóm sẽ chia bộ dữ liệu ngẫu nhiên theo hình 58.



*Hình 4.1: Chia dataset thành 2 tập dữ liệu cross và test holdout.*

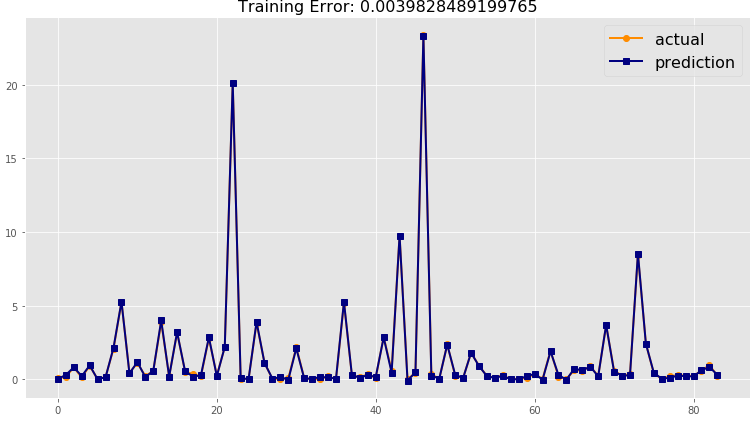
Chia tập dữ liệu thành 80% cho tập dữ liệu Cross và 20% cho tập dữ liệu test holdout. Đưa dữ liệu vào mô hình hồi quy đa biến thông thường.

* **Training Error**



*Hình 4.2: Hàm trainingError dùng để tính toán training error.*

Truyền 2 thông số X\_cross và y\_cross vào hàm trainingError. Tiến hành huấn luyện mô hình dựa trên 2 tập dữ liệu vừa truyền vào. Sau đó dự đoán dựa trên tập cross và tính toán *training error* dựa theo công thức 2.7, giá trị sau đó trả về gồm có mô hình đã được huấn luyện, giá trị *training error* và giá trị dự đoán dựa trên tập cross truyền vào.

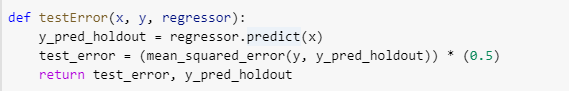


*Hình 4.3: So sánh giá trị dự đoán và giá trị thực tế.*

Theo hình 4.3, đường dự đoán trên tập *cross* với đường dữ liệu thực tế gần như khớp với nhau, với traning error là 0.00398 thì giá trị này có thể xem như dự đoán trên tập train rất tốt.

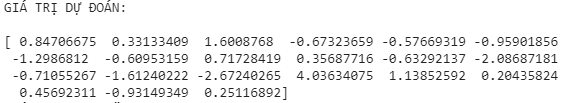
* **Test error**

Lấy mô hình được trả về từ hàm trainingError cùng với tập dữ liệu test holdout (X\_test\_holdout, y\_test\_holdout) truyền vào hàm testError. Giá trị trả về là một mảng giá trị chứa các giá trị dự đoán *y\_pred\_holdout*, và giá trị *test error.*

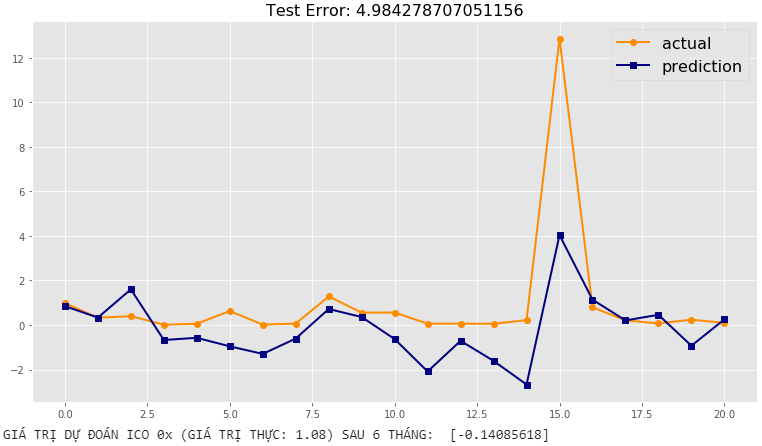


*Hình 4.4: Hàm testError dùng để tính toán test error.*

Giá trị trả về có 11 giá trị âm:



*Hình 4.5: Giá trị dự đoán trả về từ tập dữ liệu test.*



*Hình 4.6: So sánh giá trị dự đoán và giá trị thực tế.*

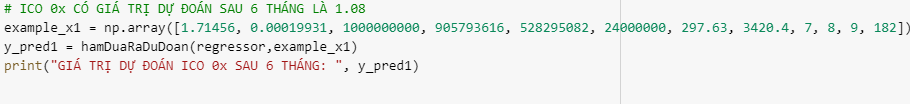
Bảng 4.1: Bảng phân tích overfitting

|  |  |  |
| --- | --- | --- |
|  | Training Error | Test Error |
| Mô hình | 0.00398 | 4.98427 |

Kết luận:

Dựa theo 2 thông số training error và test error hoàn toàn có thể kết luận hiện tượng quá khớp (overfitting) đang tồn tại khiến mô hình dự đoán trở nên sai lệch.

Lấy một mẫu dự đoán kết quả cụ thể là lấy giá trị của ICO 0x để xem giá trị dự đoán.



*Hình 4.7: Sử dụng mô hình để dự đoán một mẫu ICO 0x.*

Giá trị dự đoán trả về là -0.1408.

1. **THỰC NGHIỆM THUẬT TOÁN**

Tập dữ liệu gồm 109 ICO, nhóm sẽ tách ra 106 ICO dùng để huấn luyện mô hình, và 3 ICO dùng để test và đánh giá độ chính xác của thuật toán. Trong 106 ICO dùng để huấn luyện, nhóm sẽ chia thành Training data và Validation data theo tỉ lệ 80-20. Training data dùng dể chạy thuật toán, validation data dùng để theo dõi độ chính xác của từng thông số và điều chỉnh nó.

# Bảng 4.2: Phân tích dữ liệu thành 3 phần.

|  |  |  |  |
| --- | --- | --- | --- |
| Cross data | | Test holdout data | Test data |
| Train data | Val data |
| 68 Token | 17 Token | 21 Token | 3 Token |

Hiệu năng của hai thuật toán sẽ được tính thông qua công thức RMSE (Root Mean Square Error) và R2 (R Squared) và MAE (Mean Absolute Error)

(4.1)

(4.2)

(4.3)

1. **Mô hình hồi quy Ridge**

Chạy thực nghiệm với bộ mẫu ngẫu nhiên ta được 3 mô hình. Một mô hình có chỉ số rMSE tốt nhất, một mô hình có chỉ số R2 tốt nhất và một mô hình có chỉ số MAE tốt nhất.

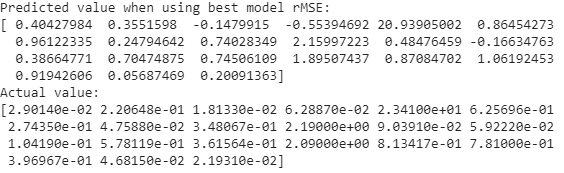
* **Kết quả rMSE:**

Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả rMSE tốt nhất.



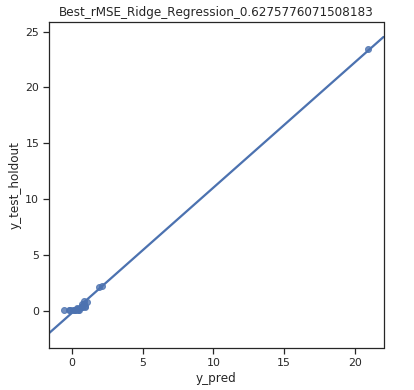
*Hình 4.8: Thông số rMSE tối ưu nhất của mô hình.*

Kết quả dự đoán trả về khi sử dụng mô hình có thông số rMSE tốt nhất dự đoán. Giá trị rMSE như trên hình 4.8 là giá trị tối ưu khi chạy vòng lặp 10000 lần trong đó mô hình nào có giá tri rMSE tốt nhất sẽ được giữ lại và giá trị này là giá trị được tính toán dựa trên giá trị dự đoán và giá trị thực tế (y\_val).



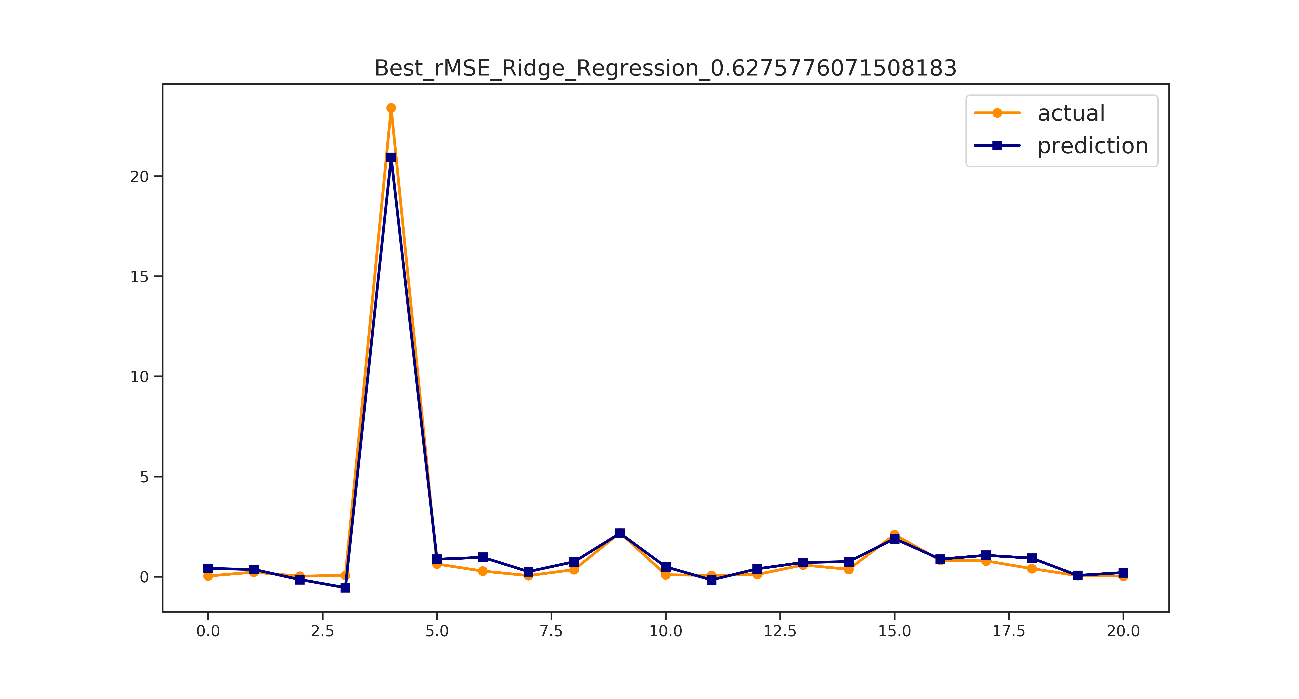
*Hình 4.9: Kết quả dự đoán trên tập X\_test\_holdout với hình rMSE tối ưu.*

So với mô hình chạy hồi quy đa biến bị overfitting thì có thể thấy các giá trị âm đã giảm đi nhiều. Mô hình hồi quy đa biến ở hình 4.4 có đến 11 giá trị âm nhưng với mô hình tối ưu giá trị rMSE của hồi quy Ridge thì số giá trị âm đã giảm còn có 3 giá trị.



*Hình 4.10: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu rMSE.*

Với hình 4.10 thì kết quả trả về giá trị rMSE là 0.627. Giá trị này có được khi lấy tập dự liệu test holdout (X\_test\_holdout) cùng với mô hình có giá trị rMSE tối ưu nhất (0.4265) đi tính toán. Dự đoán bằng cách đưa dữ liệu X\_test\_holdout vào mô hình sau khi kết quả trả về, lấy kết quả dự đoán đó cùng với dự liệu trong y\_test\_holdout để tính toán hiệu năng.

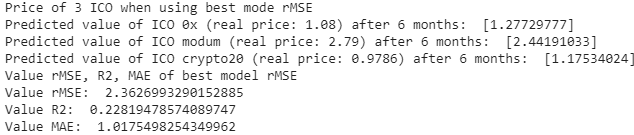


*Hình 4.11: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số rMSE tối ưu nhất.*

Ý nghĩa: Thông số rMSe thể hiện khoảng cách trung bình giữa giá trị dự đoán và giá trị thực tế là 0.6275.

Cuối cùng, nhóm sẽ sử dụng mô hình đã huấn luyện, đi tính toán kết quả của 3 ICO 0x, Modum, Crypto20.

* Giá trị thực tế của ICO1 sau 6 tháng là 1.08
* Giá trị thực tế của ICO2 sau 6 tháng là 2.79
* Giá trị thực tế của ICO3 sau 6 tháng là 0.9786



# *Hình 4.12: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có rMSE tối ưu nhất*

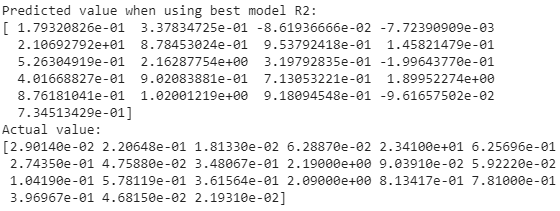
* **Kết quả R2**

Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả R2 tốt nhất.



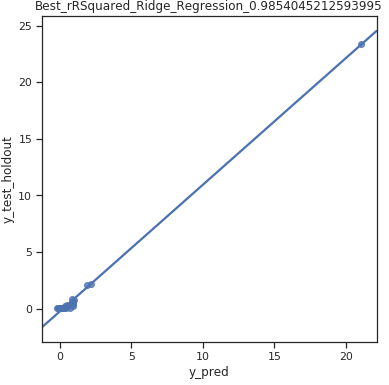
*Hình 4.13: Thông số R2**của mô hình tối ưu R2.*

Kết quả dự đoán trả về khi sử dụng mô hình có thông số R2 tốt nhất dự đoán. Giá trị R2 như trên hình 4.13 là giá trị tối ưu khi chạy vòng lặp 10000 lần trong đó mô hình nào có giá tri R2 tối ưu nhất sẽ được lưu lại và giá trị này là giá trị được tính toán dựa trên giá trị dự đoán và giá trị thực tế (y\_val).



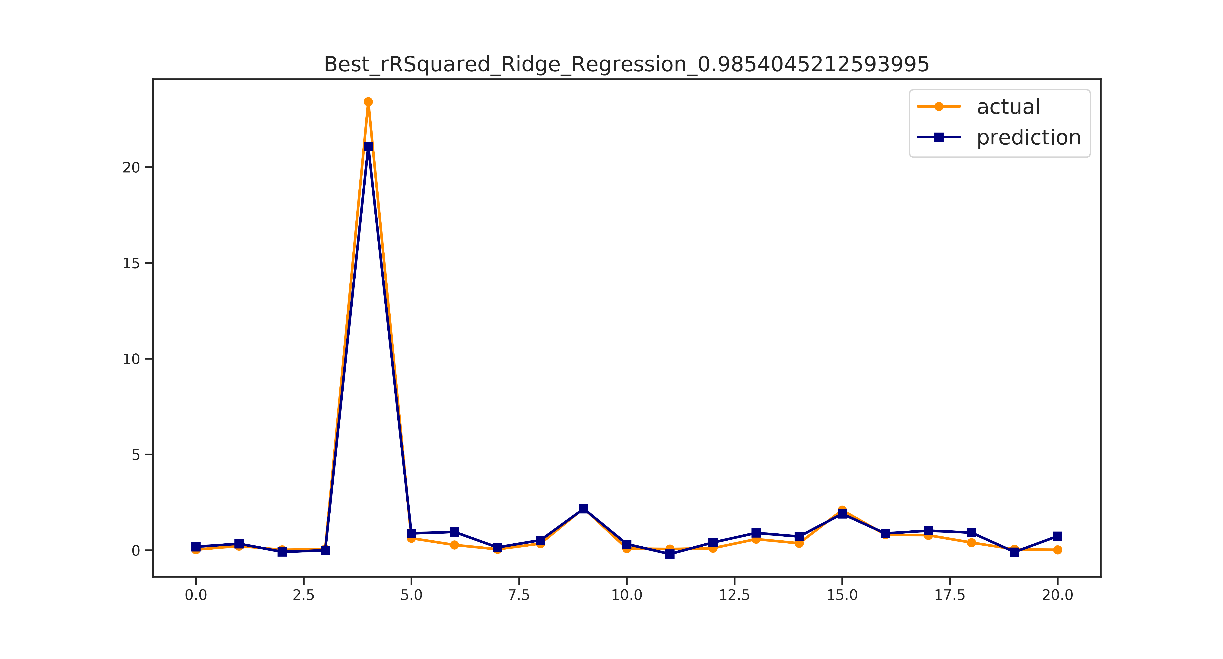
# *Hình 4.14: Kết quả dự đoán trên X\_test\_holdout với mô hình tối ưu R2.*

So với mô hình chạy hồi quy đa biến bị overfitting thì có thể thấy các giá trị âm đã giảm đi nhiều. Mô hình hồi quy đa biến ở hình 4.4 có đến 11 giá trị âm nhưng với mô hình tối ưu R2 sử dụng thuật toán Ridge thì số giá trị âm đã giảm còn có 2 giá trị.



*Hình 4.15: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu R2.*

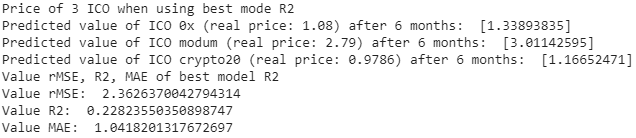
Với hình 4.15 thì kết quả trả về giá trị R2 là 0.985. Giá trị này có được khi lấy tập dự liệu test holdout (X\_test\_holdout) cùng với mô hình có giá trị R2 tối ưu nhất (0.9094) đi tính toán. Dự đoán bằng cách đưa dữ liệu X\_test\_holdout vào mô hình sau khi kết quả trả về, lấy kết quả dự đoán đó cùng với dự liệu trong y\_test\_holdout để tính toán hiệu năng.



# *Hình 4.16: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số R2 tốt nhất*

Ý nghĩa: Mô hình phù hợp với 98.5% dữ liệu từ tập X\_test\_holdout.

Cuối cùng, nhóm sẽ sử dụng mô hình đã huấn luyện, đi tính toán kết quả của 3 ICO 0x, Modum, Crypto20.



# *Hình 4.17: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có R2 tốt nhất*

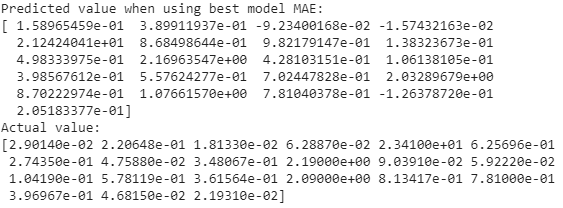
* **Kết quả MAE**

Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả MAE tốt nhất.



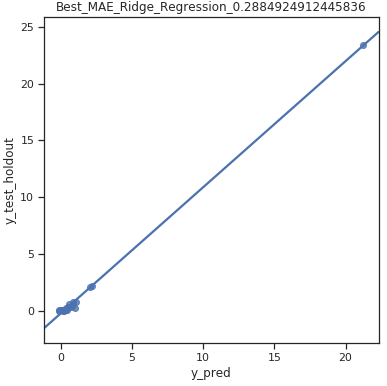
*Hình 4.18: Thông số MAE của mô hình tối ưu MAE.*

Kết quả dự đoán trả về khi sử dụng mô hình có thông số MAE tốt nhất dự đoán. Giá trị MAE như trên hình 4.18 là giá trị tối ưu khi chạy vòng lặp 10000 lần trong đó mô hình nào có giá tri MAE tối ưu nhất sẽ được giữ lại và giá trị này là giá trị được tính toán dựa trên giá trị dự đoán và giá trị thực tế (y\_val).



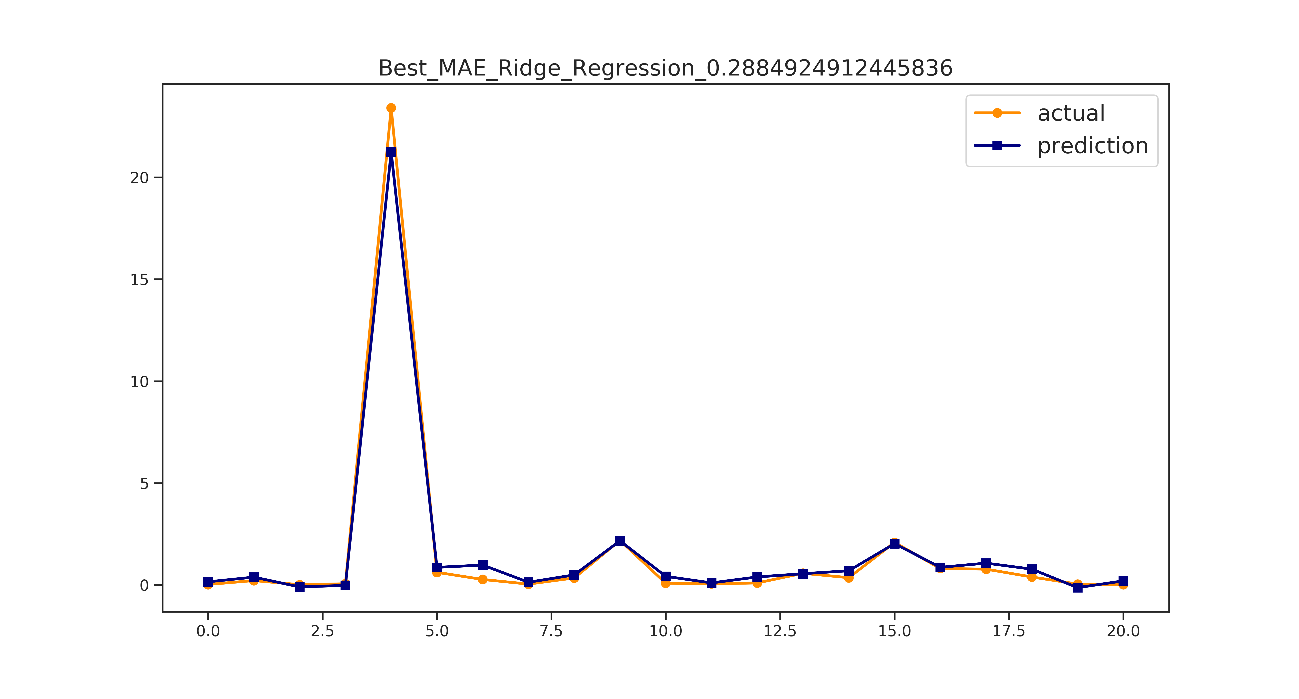
# *Hình 4.19: Kết quả dự đoán trên tập dữ liệu X\_test\_holdout với mô hình tối ưu MAE.*

So với mô hình chạy hồi quy đa biến bị overfitting thì có thể thấy các giá trị âm đã giảm đi nhiều. Mô hình hồi quy đa biến ở hình 4.4 có đến 11 giá trị âm nhưng với mô hình tối ưu MAE sử dụng thuật toán Ridge thì số giá trị âm đã giảm còn có 2 giá trị.



*Hình 4.20: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu MAE.*

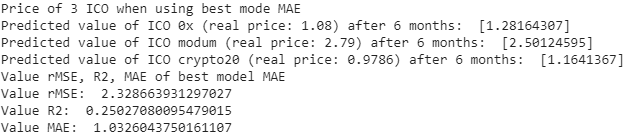
Với hình 4.20 thì kết quả trả về giá trị MAE là 0.288. Giá trị này có được khi lấy tập dự liệu test holdout cùng với mô hình có giá trị MAE tối ưu nhất (0.2894) đi tính toán. Dự đoán bằng cách đưa dữ liệu X\_test\_holdout vào mô hình sau khi kết quả trả về, lấy kết quả dự đoán đó cùng với dự liệu trong y\_test\_holdout để tính toán hiệu năng.



# *Hình 4.21: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình hồi quy Ridge có thông số MAE tốt nhất*

Ý nghĩa: Thông số MAE thể hiện sự khác biệt tuyệt đối giữa các giá trị được mô hình dự đoán và giá trị thực là 0.288.

Cuối cùng, nhóm sẽ sử dụng mô hình đã huấn luyện, đi tính toán kết quả của 3 ICO 0x, Modum, Crypto20.

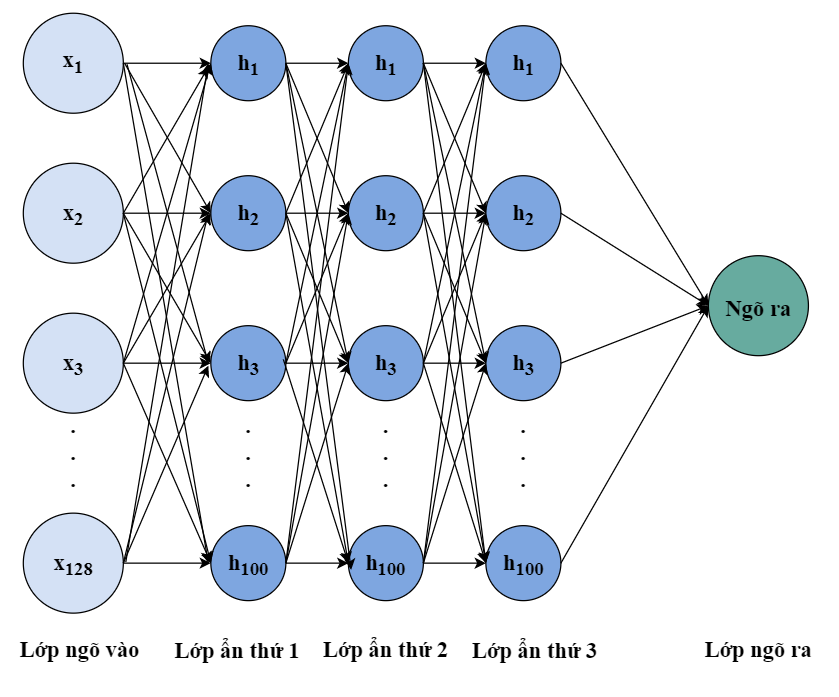


# *Hình 4.22: Kết quả dự đoán của 3 ICO dùng mô hình hồi quy Ridge có MAE tốt nhất*

1. **Mô hình mạng Nơron**

Với thuật toán mạng Nơron, nhóm thiết kế mô hình có 3 lớp ẩn với mỗi lần lớp có 100 nút và áp dụng hàm Tanh cho các lớp ẩn.

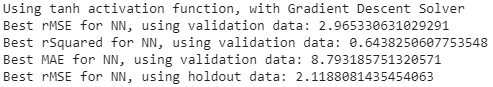
Vì ngõ ra là một giá trị liên tục nên ở lớp ngõ ra, nhóm sẽ sử dụng hàm Identity, có công thức là *f(x) = x* để giữ nguyên toàn bộ giá trị.



# *Hình 4.23: Mô hình mạng Noron cho bài toán dự đoán giá ICO*

* **Kết quả rMSE**

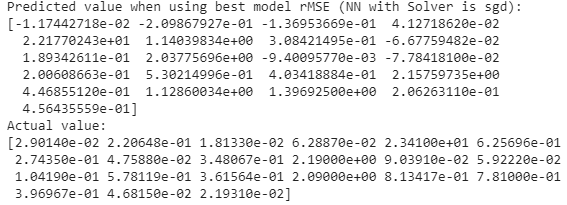
Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả rMSE tốt nhất.

****

# *Hình 4.24: Thông số rMSE tối ưu nhất của mô hình mạng Nơron*

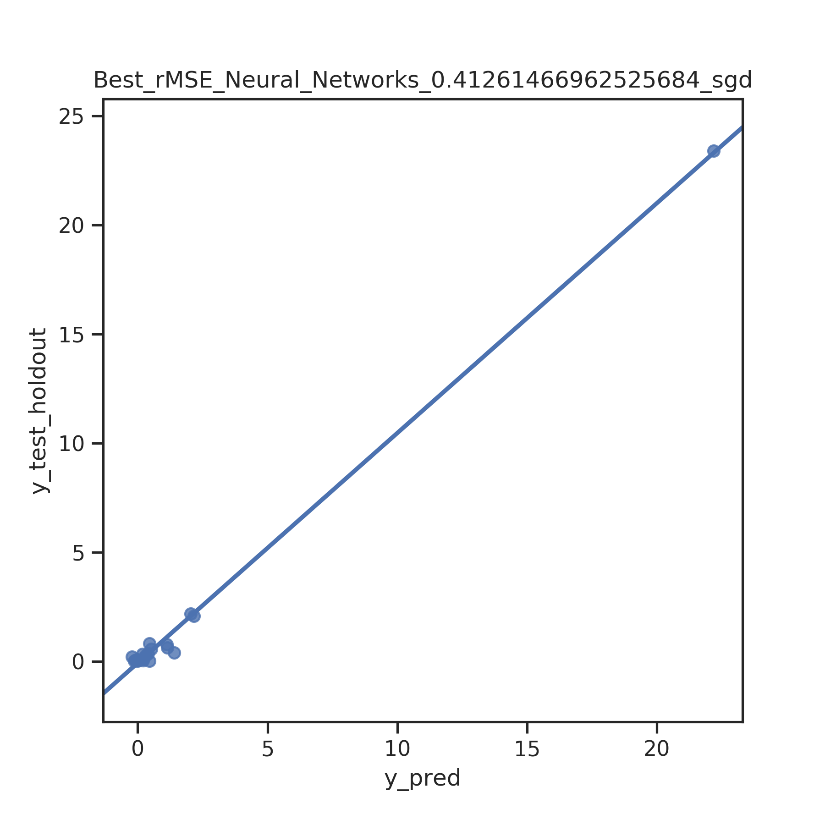
Với hình 4.24 thì kết quả trả về giá trị rMSE là 2.1188. Giá trị này có được khi lấy tập dự liệu test holdout cùng với mô hình có giá trị rMSE tối ưu nhất (2.9653) đi tính toán. Dự đoán bằng cách đưa dữ liệu X\_test\_holdout vào mô hình sau khi kết quả trả về, lấy kết quả dự đoán đó cùng với dự liệu trong y\_test\_holdout để tính toán hiệu năng.

Ý nghĩa: Thông số rMSe là khoảng cách trung bình giữa giá trị thực tế đến giá trị dự đoán là 2.1188

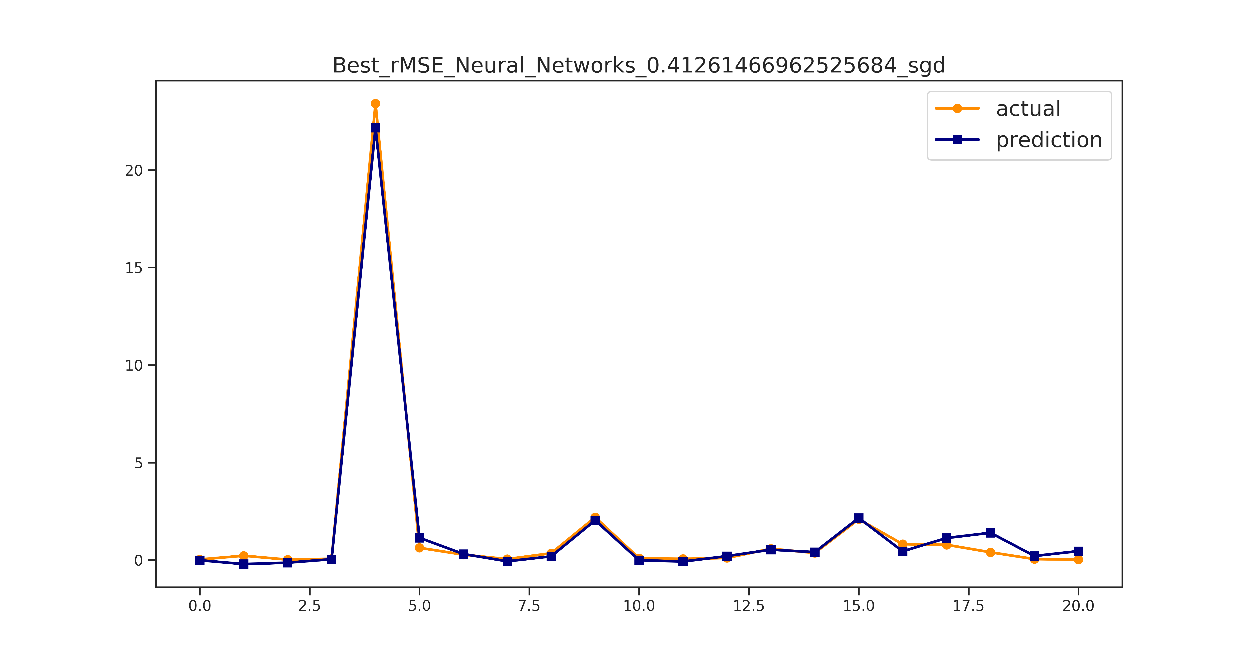


*Hình 4.25: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số rMSE tối ưu nhất.*

So với mô hình hồi quy đa biến bị quá khớp thì có thể thấy các giá trị âm đã giảm đi. Mô hình hồi quy đa biến ở hình 4.4 có đến 11 giá trị âm nhưng với mô hình tối ưu giá trị rMSE của mạng Nơron thì số giá trị âm đã giảm đi 5 giá trị.



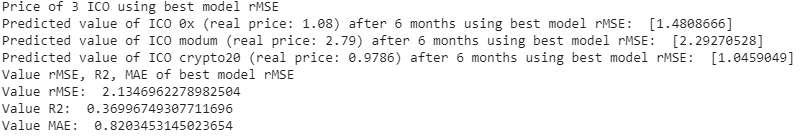
*Hình 4.26: Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu rMSE.*



*Hình 4.27: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số rMSE tối ưu nhất.*

Lưu ý:

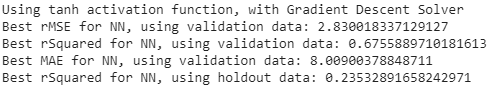
Ở đây có thể thấy giá trị rMSE sau khi chạy vòng lặp là 2.1188 ứng với mô hình tối ưu đó nhưng kết quả giá trị trên hình 4.26 và 4.27 là 0.412 vì mô hình đã lưu trước đó tối ưu hơn mô hình mới nên vẫn giữ mô hình cũ và sử dụng mô hình cũ để đi tính toán giá trị 3 ICO.



# *Hình 4.28: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có rMSE tối ưu nhất*

* **Kết quả**

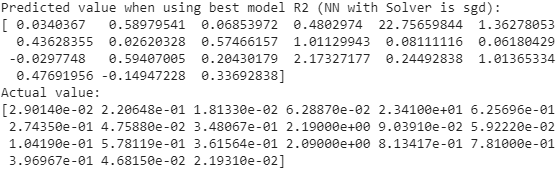
Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả R2 tốt nhất.



*Hình 4.29: Thông số R2 tối ưu của mô hình mạng Nơron.*

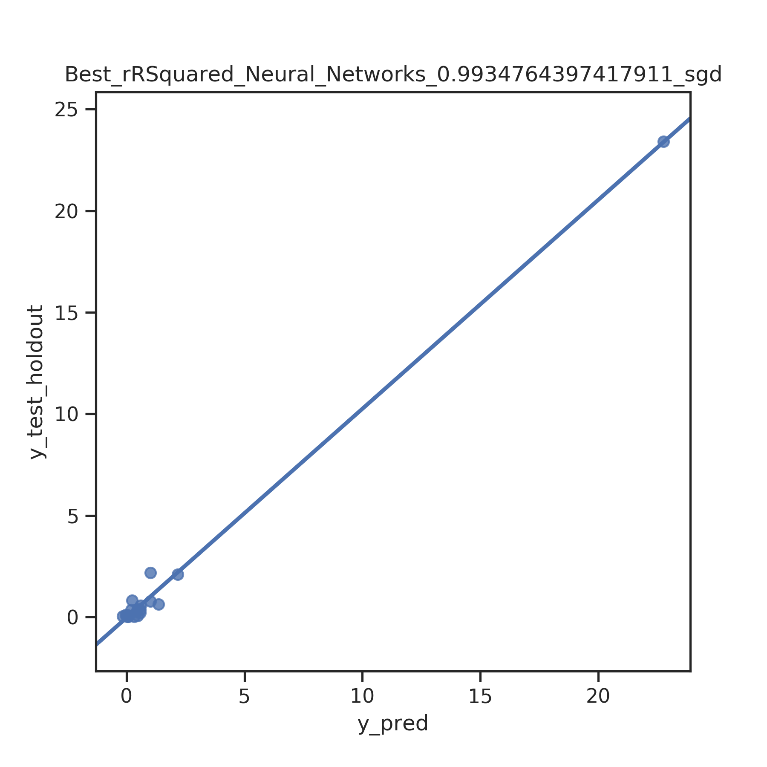
Với hình 4.29 thì kết quả trả về giá trị R2 là 0.2352. Giá trị này có được khi lấy tập dự liệu test holdout cùng với mô hình có giá trị R2 tối ưu nhất (0.6755) đi tính toán.

Ý nghĩa: Mức độ phù hợp dữ liệu của mô hình đối với tập X\_test\_holdout là 23.52%.

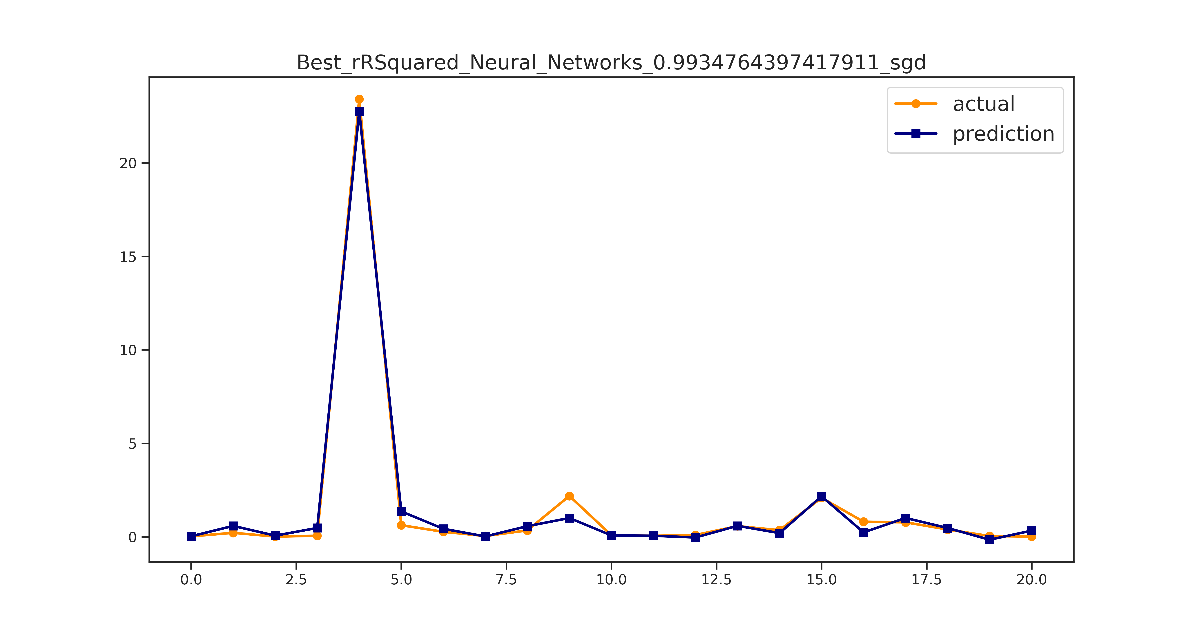


# *Hình 4.30: Kết quả dự đoán trên tập X\_test\_holdout.*

So với mô hình hồi quy đa biến bị quá khớp thì các giá trị âm đã giảm. Mô hình hồi quy đa biến ở hình 4.4 có 11 giá trị âm nhưng mô hình tối ưu giá trị R2 của mạng Nơron thì số giá trị âm đã giảm chỉ còn có 3 giá trị âm.



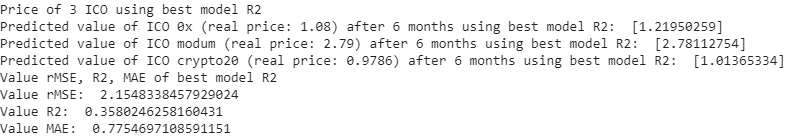
*Hình 4.31 Biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán của mô hình tối ưu R2.*



# *Hình 4.32: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số R2 tối ưu nhất.*

Lưu ý:

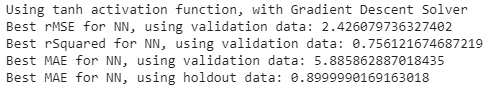
Ở đây có thể thấy giá trị R2 sau khi chạy vòng lặp là 0.2352 ứng với mô hình tối ưu đó nhưng kết quả giá trị trên hình 4.31 và 4.32 lại là 0.993 vì mô hình đã lưu trước đó tối ưu hơn mô hình mới nên vẫn giữ mô hình cũ. Sử dụng mô hình cũ để đi dự đoán giá trị 3 ICO.



# *Hình 4.33: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có R2tối ưu nhất.*

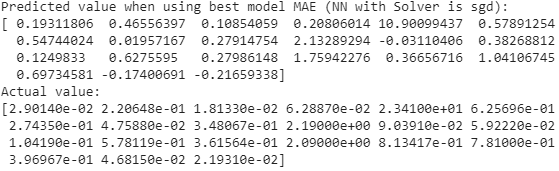
* **Kết quả MAE**

Sau khi chạy vòng lặp thì kết quả trả về là mô hình với các thông số đã tối ưu cũng như kết quả MAE tốt nhất.



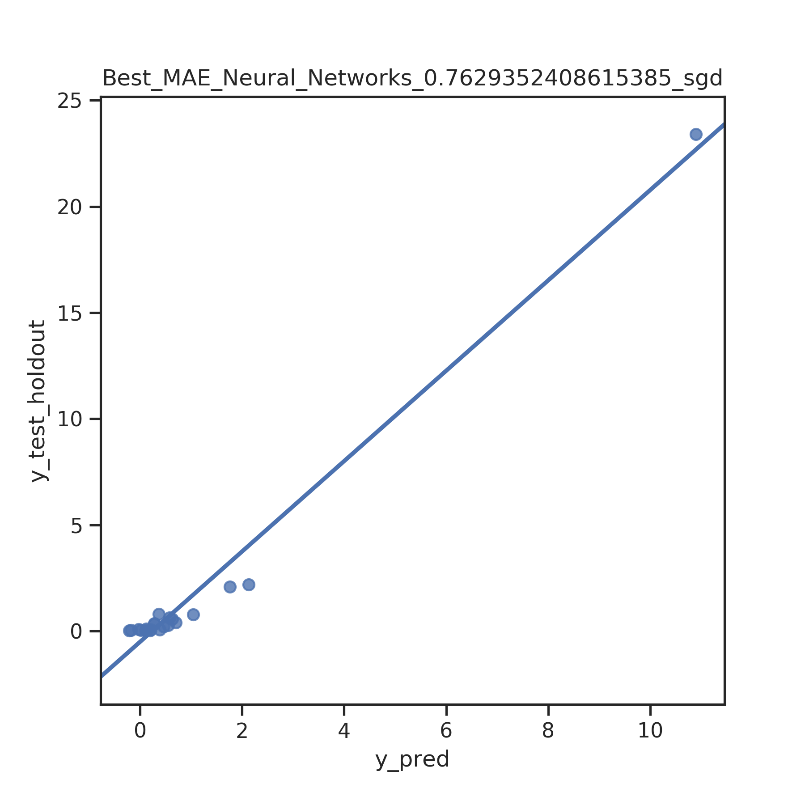
*Hình 4.34: Thông số MAE tối ưu nhất của mô hình mạng Nơron.*

Với hình 4.34 thì kết quả trả về giá trị MAE là 0.8999. Giá trị này có được khi lấy tập dự liệu test holdout cùng với mô hình có giá trị MAE tối ưu nhất (5.885) đi tính toán. Thông số MAE thể hiện sự khác biệt giữa các giá trị được mô hình dự đoán và giá trị thực tế là 0.899.

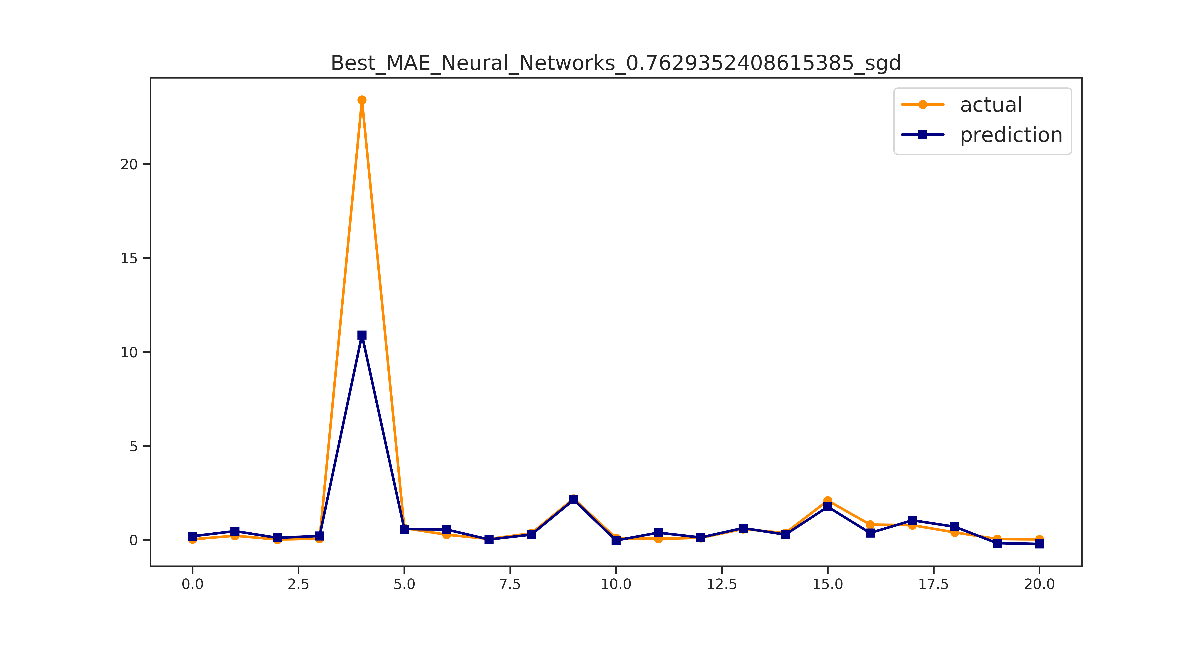


# *Hình 4.35: Kết quả dự đoán trên tập X\_test\_holdout với mô hình tối ưu MAE.*

So với mô hình hồi quy đa biến bị quá khớp thì các giá trị âm đã giảm. Mô hình hồi quy đa biến ở hình 4.4 có 11 giá trị âm nhưng mô hình tối ưu giá trị MAE của mạng Nơron thì số giá trị âm đã giảm chỉ còn có 1 giá trị âm.



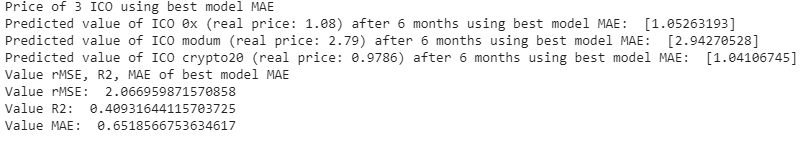
*Hình 4.36: Biểu đồ phân tán giữa giá trị dự đoán và giá trị thực tế của mô hình tôi ưu MAE.*



# *Hình 4.37: So sánh kết quả giữa y\_test\_holdout và y\_pred\_holdout trong mô hình mạng Nơron có thông số MAEtối ưu nhất.*

Lưu ý:

Ở đây có thể thấy giá trị MAE sau khi chạy vòng lặp là 0.8999 ứng với mô hình tối ưu đó nhưng kết quả giá trị trên hình 4.36 và 4.37 lại là 0.762 vì mô hình đã lưu trước đó tối ưu hơn mô hình mới nên vẫn giữ mô hình cũ.



# *Hình 4.38: Kết quả dự đoán 3 ICO dùng mô hình mạng Nơron có MAEtối ưu nhất*

# Bảng 4.3: So sánh độ chính xác của 2 thuật toán.

|  |  |  |  |
| --- | --- | --- | --- |
| Hồi quy Ridge | Đồng 0x | Đồng modum | Đồng crypto20 |
| rMSE tối ưu | 82.41% | 87.45% | 79.9% |
| tối ưu | 76.85% | 92.11% | 80.8% |
| MAE tối ưu | 81.48% | 89.64% | 81.04% |

|  |  |  |  |
| --- | --- | --- | --- |
| Mạng Nơron | Đồng 0x | Đồng modum | Đồng crypto20 |
| rMSE tối ưu | 63% | 82.07% | 93.12% |
| tối ưu | 87.96% | 99.64% | 96.42% |
| MAE tối ưu | 97.2% | 94.6% | 93.62% |

Dựa theo bảng 4.3 thì:

* Với mô hình có thông số rMSE tối ưu nhất thì thuật toán hồi quy Ridge có kết quả dự đoán tốt hơn thuật toán mạng Nơron.
* Với mô hình có thông số và MAE tối ưu nhất thì thuật toán mạng Nơron có kết quả dự đoán tốt hơn thuật toán hồi quy Ridge.

**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

1. **KẾT LUẬN**
2. **Những vấn đều nghiên cứu**

* Nghiên cứu về tiền điện tử, giai đoạn ICO.
* Nghiên cứu về quy trình thu thập dữ liệu.
* Nghiên cứu về thuật toán máy học là hồi quy Ridge và mạng Nơron.

1. **Những vấn đề hoàn thành**

* Sử dụng ứng dụng để thu thập dữ liệu.
* Sử dụng thành công thuật toán máy học trong việc dự đoán giá ICO sau 6 tháng.

1. **HẠN CHẾ**

* Tập dữ liệu còn ít.
* Quá trình thu thập dữ liệu chưa hoàn toàn bằng ứng dụng, sau khi thu thập với các trường hợp không có dữ liệu, nhóm phải tự thu thập bằng cách tham khảo các trang web.
* Thuật toán máy học dự đoán giá ICO đạt độ chính xác chưa cao. Cần tối ưu thuật toán cũng như phần cứng để có thể đạt độ chính xác tốt hơn.

1. **HƯỚNG PHÁT TRIỂN**

Từ những vấn đề nhóm đã hoàn thành và các hạn chế của đề tài, nhóm đề xuất ra hướng phát triển như sau:

* Sử dụng các thuật toán máy học khác nhằm tìm ra thuật toán tối ưu nhất cho bài toán dự đoán giá ICO.
* Bổ sung tập dữ liệu.
* Bài toán về dự đoán giá ICO cũng như các bài toán về vấn đề tài chính còn chịu sự ảnh hưởng của yếu tố con người. Việc tìm hiểu yếu tố con người và áp dụng vào mô hình sử dụng thuật toán máy học sẽ giúp mô hình đạt độ chính xác và độ tin cậy cao hơn.

**TÀI LIỆU THAM KHẢO**

[1] Nguyễn Ngọc Bình, “Phân tích hồi quy và tương quan” [Trực tuyến]. Available: <https://rpubs.com/nguyenngocbinhneu/432206> [Đã truy cập 4.10.2019]

[2] Vũ Hữu Tiệp, “Machine Learning Cơ Bản,” [Trực tuyến]. Available: <http://machinelearningcoban.com/ebook/>. [Đã truy cập 14.09.2019].

[3] Andrew Ng, “Machine Learning by Prof. Andrew Ng” [Trực tuyến]. Available: <https://github.com/vkosuri/CourseraMachineLearning> [Đã truy cập 10.09.2019]

[4] Nguyễn Thanh Tuấn, “Deep Learning Cơ Bản,” [Trực Tuyến]. Available:   
<https://drive.google.com/file/d/1UQQ5L2KqCgy5P7Z2hmiuzeCQg4Oqqexu/view>. [Đã truy cập 16.10092019].

[5] Deanna N. Schreiber - Gregory, “Ridge Regression and Multicollinearity: An In-Depth Review,” [Trực tuyến]. Available:  
<https://pdfs.semanticscholar.org/7b6c/1a15adcb62fca49e0fa28df4d50c3610a82f.pdf?fbclid=IwAR13FsElmnuFXE2KmMl13Y-I9cAP-FnngKrl_DBI79Mo_ODdhQM56wNpjAk>. [Đã truy cập 28.9.2019]

[6] Scott Robinson, “Introduction to Neural Networks with Scikit-Learn” [Trực tuyến]. Available:

<https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/> [Đã truy cập 2.10.2019]

[7] Avinash Sharma V, “Understanding Activation Functions in Neural Networks”, [Trực tuyến]. Available:

<https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0> [Đã truy cập 10.10.2019]

[8] Martin T. Hagan và cộng sự, “*Neural Network Design*”, United States, Martin T. Hagan, 2014.

[9] Matt Mazur, “A Step by step Backpropagation Example” [Trực tuyến]. Available: <https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/> [Đã truy cập 3.10.2019]