# UNIVERSITY OF ECONOMICS AND LAW

# FACULTY OF INFORMATION SYSTEMS

-- & --


# FINAL PROJECT

## Phân tích dữ liệu chuỗi thời gian và dự báo

**PERFORMED BY: Photo_Dump**

**CLASS: K20416C**

| Tên | MSSV | Tỉ lệ đóng góp |
|---|---|---|
| Lê Quang Chấn Phong | K204161998 | 100% |
| Lê Chí Hào | k204161988 | 100% |
| Nguyễn Thị Ái Linh | K204161991 | 100% |

## ACKNOWLEDGEMENTS

First of all, I'd like to thank the University of Economics and Law for allowing me to complete this report on the Time series data analysis and forecasting Project. The successful completion of any type of project necessitates the assistance of a number of people. For the course of this work, I also enlisted the assistance of several people. There is already a small effort to express my heartfelt gratitude to that helpful individual. I would like to express my heartfelt appreciation to Mr. Su, Academic Supervisor at the University of Economics and Law. This study would have been a little less successful without his kind direction and proper guidance. His supervision and guidance were instrumental in ensuring that this report was completed flawlessly at every stage of the project.

# MỤC LỤC

LIST OF FIGURES

LIST OF TABLES

# Chapter 1: Introduction

## 1.1. Economic forecasting and the evolution of stock price forecasting

Dự báo kinh tế và dự báo giá cổ phiếu có liên quan chặt chẽ với nhau vì cả hai đều liên quan đến việc dự đoán tình trạng tương lai của nền kinh tế và thị trường tài chính. Dự báo kinh tế là quá trình sử dụng các mô hình và kỹ thuật khác nhau để dự đoán tình trạng tương lai của nền kinh tế, trong khi dự báo giá cổ phiếu là quá trình sử dụng các mô hình và kỹ thuật khác nhau để dự đoán giá cổ phiếu trong tương lai trên thị trường tài chính.

Dự báo kinh tế liên quan đến việc phân tích các biến kinh tế vĩ mô như GDP, lạm phát, việc làm, lãi suất và các yếu tố khác để dự đoán hoạt động của một nền kinh tế. Thông tin này có thể được sử dụng để đưa ra các quyết định chiến lược về đầu tư, lập ngân sách và các hoạt động lập kế hoạch tài chính khác. Có nhiều mô hình và kỹ thuật khác nhau được sử dụng trong dự báo kinh tế, bao gồm mô hình chuỗi thời gian, mô hình kinh tế lượng và mô hình cấu trúc.

Mặt khác, dự báo giá cổ phiếu liên quan đến việc phân tích các yếu tố khác nhau có thể ảnh hưởng đến giá cổ phiếu trong tương lai của một công ty hoặc thị trường nhất định. Những yếu tố này có thể bao gồm dữ liệu giá trong quá khứ, dữ liệu tài chính của công ty, tin tức và sự kiện cũng như các yếu tố khác có thể ảnh hưởng đến tâm lý nhà đầu tư. Có nhiều mô hình và kỹ thuật khác nhau được sử dụng trong dự báo giá cổ phiếu, bao gồm phân tích kỹ thuật, phân tích cơ bản và thuật toán máy học.

Sự phát triển của dự báo giá cổ phiếu đã bị ảnh hưởng rất nhiều bởi những tiến bộ công nghệ và sự sẵn có của dữ liệu. Các phương pháp dự báo giá cổ phiếu truyền thống, chẳng hạn như phân tích cơ bản và kỹ thuật, đã tồn tại hàng thập kỷ. Tuy nhiên, những tiến bộ trong thuật toán học máy và sự sẵn có của các tập dữ liệu lớn đã cho phép sử dụng các mô hình và kỹ thuật phức tạp hơn trong dự báo giá cổ phiếu.

Ngày nay, dự báo giá cổ phiếu thường được thực hiện bằng cách sử dụng kết hợp các kỹ thuật truyền thống và hiện đại. Phân tích kỹ thuật và cơ bản vẫn được sử dụng rộng rãi, nhưng các thuật toán học máy và các mô hình thống kê nâng cao khác đang trở nên phổ biến hơn. Các mô hình này có thể tính đến nhiều loại điểm dữ liệu, cho phép dự báo giá cổ phiếu chính xác và chi tiết hơn.

Dự báo kinh tế và dự báo giá cổ phiếu là những lĩnh vực liên quan chặt chẽ với nhau liên quan đến việc dự đoán tình trạng tương lai của nền kinh tế và thị trường tài chính. Trong khi các mô hình và kỹ thuật truyền thống vẫn được sử dụng rộng rãi, những tiến bộ về công nghệ và tính sẵn có của dữ liệu đã cho phép các mô hình dự báo phức tạp và chính xác hơn được đưa ra.

## 1.2. Objectives of the research

Dưới đây là những mục tiêu của chúng em khi thực hiện bài nghiên cứu về cổ phiếu liên quan đến dầu khí:

- Để cung cấp cái nhìn sâu sắc về hiệu suất trong tương lai của các cổ phiếu dầu khí: Một trong những mục tiêu chính của nghiên cứu về dự báo trữ lượng dầu khí là cung cấp cái nhìn sâu sắc về hiệu suất trong tương lai của các cổ phiếu này. Điều này có thể giúp các nhà đầu tư và các bên liên quan khác đưa ra quyết định sáng suốt về việc mua, nắm giữ hoặc bán các cổ phiếu này.
- Để xác định các động lực chính đối với hoạt động của cổ phiếu dầu khí: Điều này có thể bao gồm các yếu tố như giá dầu khí, mức sản xuất, nhu cầu và các chỉ báo kinh tế vĩ mô khác.
- Để phát triển các mô hình dự báo trữ lượng dầu khí: Nghiên cứu về dự báo trữ lượng dầu khí cũng có thể nhằm mục đích phát triển các mô hình dự báo có thể được sử dụng để dự đoán giá cổ phiếu và hiệu suất trong tương lai. Các mô hình này có thể dựa trên nhiều nguồn dữ liệu khác nhau, chẳng hạn như dữ liệu giá lịch sử, dữ liệu tài chính của công ty, tin tức và sự kiện cũng như các yếu tố liên quan khác.

- Để đánh giá độ chính xác của các mô hình dự báo hiện có: Điều này có thể bao gồm so sánh hiệu suất của các mô hình khác nhau, sử dụng các bộ dữ liệu khác nhau và thử nghiệm các giả định và thông số khác nhau.
- Để cung cấp thông tin cho các quyết định chính sách: Điều này có thể bao gồm các khuyến nghị về chính sách, quy định của chính phủ và các biện pháp can thiệp khác có thể giúp ổn định giá cổ phiếu dầu khí và thúc đẩy tăng trưởng kinh tế bền vững.

Mặc dù,mục tiêu nghiên cứu dự báo trữ lượng dầu khí có thể rất đa dạng nhưng mục tiêu cuối cùng mà chúng em hướng đến là cung cấp cho các bên liên quan thông tin chính xác và đáng tin cậy về hiệu suất trong tương lai của các cổ phiếu này, có thể giúp hướng dẫn các quyết định đầu tư và thúc đẩy sự ổn định và tăng trưởng kinh tế.

## 1.3. Subject of the research

Đối tượng nghiên cứu về dự báo cổ phiếu dầu khí là **phân tích và dự đoán diễn biến tương lai của cổ phiếu dầu khí trên thị trường tài chính**. Bài làm tập trung vào việc phát triển các mô hình và kỹ thuật có thể được sử dụng để dự báo giá và lợi nhuận trong tương lai của các cổ phiếu này, cũng như để xác định các động lực chính và các yếu tố có thể ảnh hưởng đến hiệu suất của chúng.

Chủ đề của đồ án liên quan đến việc phân tích nhiều nguồn dữ liệu, chẳng hạn như dữ liệu giá lịch sử, dữ liệu tài chính của công ty và các chỉ số kinh tế vĩ mô khác, để xác định các mẫu và mối quan hệ có ý nghĩa có thể được sử dụng để dự đoán giá cổ phiếu và hiệu suất trong tương lai. Nghiên cứu này cũng liên quan đến việc thử nghiệm và tinh chỉnh các mô hình dự báo để đảm bảo độ chính xác và độ tin cậy của chúng, cũng như đánh giá hiệu suất của các mô hình này theo thời gian.

Ngoài ra, đối tượng nghiên cứu về dự báo trữ lượng dầu khí còn bao gồm việc nghiên cứu các yếu tố khác nhau có thể ảnh hưởng đến hoạt động của các trữ lượng này, chẳng hạn như thay đổi giá dầu khí, mức sản xuất, nhu cầu, chính sách và quy định của chính phủ, công nghệ. tiến bộ và điều kiện kinh tế toàn cầu. Nghiên cứu này nhằm

mục đích xác định các động lực và xu hướng chính có khả năng định hình hiệu quả hoạt động trong tương lai của những cổ phiếu này, cũng như cung cấp thông tin chuyên sâu về cách các tình huống và sự kiện khác nhau có thể ảnh hưởng đến hiệu quả hoạt động của chúng.

## 1.4. Introduce data

Description Data

Tập dữ liệu về các loại stock dầu khí Việt Nam có tổng cộng 14468 dòng và 7 cột, bao gồm 13 mã stock khác nhau được thu thập từ ngày 03-01-2017 đến 31-12-2021.

| Column1 | code | date | time | floor | type | close | nmVolume | snapshot_date | vietnameseName |
|---|---|---|---|---|---|---|---|---|---|
| 113 | TDG | 2021-12-03 | 15:12:06 | HOSE | STOCK | 12.0 | 1384100.0 | 2021-12-03 | Dầu Khí |
| 290 | PLX | 2021-12-01 | 15:12:03 | HOSE | STOCK | 54.3 | 1844900.0 | 2021-12-01 | Dầu Khí |
| 436 | TDG | 2021-12-01 | 15:12:03 | HOSE | STOCK | 11.2 | 727100.0 | 2021-12-01 | Dầu Khí |
| 452 | PVB | 2021-12-03 | 15:12:08 | HNX | STOCK | 16.2 | 84551.0 | 2021-12-03 | Dầu Khí |
| 526 | PTX | 2021-12-01 | 15:12:01 | UPCOM | STOCK | 0.3 | 0.0 | 2021-12-01 | Dầu Khí |
| 600 | PND | 2021-12-14 | 15:12:01 | UPCOM | STOCK | 14.9 | 0.0 | 2021-12-14 | Dầu Khí |
| 677 | PTV | 2021-12-15 | 15:12:01 | UPCOM | STOCK | 7.8 | 18200.0 | 2021-12-15 | Dầu Khí |
| 725 | PND | 2021-12-03 | 15:12:01 | UPCOM | STOCK | 14.3 | 200.0 | 2021-12-03 | Dầu Khí |
| 899 | PVO | 2021-12-15 | 15:12:01 | UPCOM | STOCK | 10.1 | 30800.0 | 2021-12-15 | Dầu Khí |
| 974 | TDG | 2021-12-16 | 15:12:03 | HOSE | STOCK | 12.1 | 706300.0 | 2021-12-16 | Dầu Khí |

*Table 1.1: 10 dòng đầu của bộ dữ liệu*

Dưới đây là các biến và giải thích ý nghĩa của từng biến:

| | Attribute | Type Description |
|---|---|---|
| 1 | code | nominal This variable represents the stock code or identifier for each stock. |
| 2 | date | datetime This variable represents the date for which the stock information was |

| | | |
|---|---|---|
| | | collected |
| 3 | time | datetime This variable represents the time for which the stock information was collected. |
| 4 | floor | nominal This variable represents the stock exchange where the stock is traded. |
| 5 | type | nominal This variable represents the type of stock, such as common stock or preferred stock. |
| 6 | close | numeric This variable represents the |
| 7 | nmVolume | numeric This variable represents the notional market volume of shares traded for a given day. |

*Table 1.2: Giới thiệu biến*

**Kiểm tra null data và type của từng loại trường dữ liệu.**

Sử dụng lệnh info() để tìm ra các biến bị null.

Tất cả các truong du lieu không có missing data. Dữ liệu gồm 3 biến định lượng và 7 biến định tính.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14468 entries, 0 to 14467
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      14468 non-null  int64
 1   code            14468 non-null  object
 2   date            14468 non-null  object
 3   time            14468 non-null  object
 4   floor           14468 non-null  object
 5   type            14468 non-null  object
 6   close           14468 non-null  float64
 7   nmVolume        14468 non-null  float64
 8   snapshot_date   14468 non-null  object
 9   vietnameseName  14468 non-null  object
dtypes: float64(2), int64(1), object(7)
memory usage: 1.1+ MB
```

*Figure 1.1: Kiểm tra null*

# Chapter 2: Literature Review

| Tên bài | Năm | Tác giả | Tổng kết | Thuật | Nguồn dữ liệu | Các biến sử dụng |
|---|---|---|---|---|---|---|

| **nghiên cứu** | | | | **toán** | | |
|---|---|---|---|---|---|---|
| Stock selection with random forest: An exploitation of excess return in the Chinese stock market | August 2019 | Zheng Tan * Ziqin Yan * Guangwei Zhu | bài báo chỉ ra rằng phương pháp họ sử dụng là đáng tin cậy và thuật toán học máy được sử dụng ở đây có khả năng tốt khái quát hóa khi tạo phân loại chứng khoán hợp lý về mặt kinh tế. Minh họa được mối quan hệ cấu trúc giữa lợi nhuận vượt trội trong dài hạn và không gian tính năng cơ bản/kỹ thuật có liên quan. Những quan sát trong bái báo là hữu ích cho các nhà giao dịch định lượng trong việc xây dựng các chiến lược có lợi nhuận. | random forest (RF) | Tất cả các feature được lấy từ cơ sở dữ liệu Wind, có mối tương quan chặt chẽ với cổ phiếu trung bình lợi nhuận trong thị trường chứng khoán trưởng thành. | EP, BP, ROE, Net profits yoy, Business income yoy, ROA, Market cap, SP. |
| Forecasting Model of High | May 16,2021 | Yi Tao,Yanyan Chen | Bài báo đã chỉ ra những điểm nổi bật như sau: Thứ nhất, các mô hình có khả | XGBoost, LightGB, CatBoost, Random | cổ phiếu ở Trung Quốc trong bảy năm qua với tổng số 367 | learning_rate, n_estimators, max_depth, num_leaves, |

| | | | | | | |
|---|---|---|---|---|---|---|
| Transfer Stock —— Based on Integrated Learning | | | năng dự đoán tốt hơn và khả năng khái quát hóa bằng sử dụng thuật toán học máy mới kết hợp với lý thuyết tài chính truyền thống; Thứ hai, do mất cân đối phân phối tập dữ liệu gây ra bởi tương đối ít truyền cao mẫu, bài báo này sử dụng Xếp chồng để tích hợp những người học giỏi, đã cải thiện khả năng nhận dạng của chuyển giao cao cổ phiếu. | Forest | chỉ số như cổ tức trên cổ phiếu và giá trị TM của thu nhập trên mỗi cổ phiếu. | max_bin, min_data_in_leaf, feature_fraction, bagging_fraction, lambda_l1, lambda_l2, min_split_gain |
| Demand Forecasting of a Multinational Retail | 2022 | Priyam Saha, Nitesh Gudheniya , Rony Mitra, Dyutim | Bài nghiên cứu đã chứng minh việc thực hiện tính ưu việt của dự báo doanh số bán hàng dựa trên LSTM và LGBM, sau đó tiếp tục so sánh rộng rãi trên một công ty bán | Machine Learning, Deep Learning, LSTM, LightGBM | Dữ liệu được sử dụng là doanh số bán lẻ của một công ty đa quốc gia có trụ sở tại Mỹ. | Sales vs Days |

| | | | | | | |
|---|---|---|---|---|---|---|
| Company using Deep Learning Frameworks | | oy Das, Sushmita Narayana, Manoj K. Tiwari | lẻ đa quốc gia của Mỹ. Cuối cùng, họ lập kế hoạch trước cho công ty ví dụ: dự trữ trước trong trường hợp dự kiến giá tăng đột biến bán hàng hoặc giảm lượng tiêu thụ trong thời gian bán hàng chậm, do đó giảm tổn thất. | | | |
| Wavelet-Aided Stock Forecasting Model based on Ensembled Machine Learnin | September 18–20, 2020 | Yuanyuan Qu, Zhongkai Zhang, Zhiliang Qin | Bài báo đã đưa ra dự đoán và hiển thị xu hướng giá đóng cửa trên từng cổ phiếu trong các ngành cụ thể. Ngoài ra bài báo còn sử dụng phép biến đổi wavelet để giảm nhiễu trình tự và trích xuất phần bổ sung các tính năng, và làm giảm bớt sự bất ổn cố hữu trong tài | LightGBM | - Sử dụng dữ liệu giá cổ phiếu từ tháng 1 năm 2015 đến tháng 1 năm 2019 để **đào tạo mô hình**,<br>**-đánh giá hiệu suất** dựa trên dữ liệu giá cổ phiếu từ tháng 1 năm 2019 đến tháng 9 năm 2019 | Real Estate ,Electric Power, Cement, Coal |

| | | | | | | |
|---|---|---|---|---|---|---|
| g | | | chính bộ dữ liệu, sử dụng giá cổ phiếu đa chiều các vectơ đặc trưng để tối ưu hóa mô hình tập hợp giúp cải thiện độ chính xác của dự đoán so với các phương pháp khác. | | | |
| Financial Trading Strategy System Based on Machine Learning | 28 July 2020 | Yanjun Chen, Kun Liu, Yuantao Xie , and Mingyu Hu | Bài báo này xem xét các rủi ro tài chính và lợi nhuận của cổ phiếu thị trường làm đối tượng nghiên cứu và sử dụng phương pháp học máy và khai thác dữ liệu để xây dựng giao dịch tài chính hệ thống chiến lược dựa trên.<br>Kết quả môt hình cho ra như sau:<br>1/ LightGBM đối với mô hình truyền thống: LightGBM | LightGBM,RandomForest , GLM, DNN, SVM | Sử dụng chứng khoán Trung Quốc từ 2008 đến 2018. | Close_Price,AP, N_CF_FR_Finan_A, CIP, Current Ratio, Intant Assets, T_Compr_Income, AR, Income_Tax, ROS, P/B Ratio, T_CA, Cash_C_Equiv, ROE, ROA, C_Paid_G_S, Surplus_Reser, Turnover_Value, N_CE_Beg_Bal, Price_Rate, |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | R-squared RMSE, GLM. -> Light GBM lỗi dự đoán nhỏ hơn suy ra thuật toán mang lại độ chính xác cao<br>2/So với phương pháp trọng số bình đẳng và phương pháp trọng số giá trị thị trường: danh mục đầu tư, được xây dựng bằng phương pháp trọng số phương sai tối thiểu của mô hình phương sai trung bình với ràng buộc CVaR, có độ ổn định và năng suất tốt nhất, tiếp theo là phương pháp trọng số giá trị thị trường và phương pháp cân bằng.<br>3/ ba yếu tố ảnh hưởng nhất là giá đóng cửa của cổ phiếu hiện tại, tài khoản phải trả, và tốc | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | độ tăng trưởng của giá cổ phiếu. | | | |
| MACHINE LEARNING BASED DEMAND FORECAST MODELS FOR E-COMMERCE INDUSTRY | 19-22 MAY 2022 | Bülent Bedir, Mert Erdoğan, Ersin Kanar, M. Fatih Akayand Sevtap Erdem | Theo các kết quả thu được, người ta đã quan sát thấy rằng những thay đổi trong các tùy chọn độ trễ thời gian được sử dụng trong các mô hình dựa trên LSTM và các tính năng được sử dụng trong các mô hình dựa trên Light-GBM có giá trị MAPE đáng kể,Các giá trị MAPE thay đổi trong khoảng từ 14,2% đến 27,2% đối với các mô hình dựa trên LSTM và từ 15,6% đến 22,4% đối với các mô hình dựa trên Light-GBM. Có thể kết luận rằng LSTM và Light-GBM | LSTM and Light-GBM | Sử dụng bộ dữ liệu 4 năm của một công ty thương mại điện tử. Tập dữ liệu bao gồm các phạm vi ngày 01.01.2017 - 31.12.2020. | sale, date, model (1-20) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | có thể được áp dụng để giải bài toán dự báo nhu cầu. | | | |
| Research on trend prediction of component stock in fuzzy time series based on deep forest | 24 September 2022 | Peng Li, Hengwen Gu, Lili Yin, Benling Li | Bài viết này kết hợp logic mờ và mô hình rừng tầng đa hạt để thu được xu hướng của chỉ số chứng khoán sau khi dự báo mờ. Theo kết quả thực nghiệm, các đặc trưng thu được từ việc xử lý mờ chỉ số chứng khoán có lợi cho việc dự báo các cổ phiếu cấu thành mô hình. Đây là trong rừng theo tầng đa hạt và các mô hình rừng cực đoan đã được xác thực. Kết quả của cuộc thử nghiệm cũng cho thấy chỉ số S&P 500 có tác dụng tích cực trong việc dự đoán xu hướng của các cổ phiếu cấu thành nó. | Decision tree, Random forest, gcForest | Nhận dữ liệu chứng khoán Hoa Kỳ thông qua giao diện Yahoo Finance, bao gồm cao, thấp, mở, đóng và khối lượng giao dịch của Facebook, AMD và Chỉ số S&P 500 từ ngày 3 tháng 1 năm 2017 đến ngày 6 tháng 6 năm 2020 | date, close, open, high, low, Adj close, Volume |

| | | | | | | |
|---|---|---|---|---|---|---|
| Stock Market Prices Prediction using Random Forest and Extra Tree Regression | September 2019 | Subba Rao Polamuri, K. Srinivas, A. Krishna Mohan | Bài nghiên cứu đưa ra kết luận rằng tất cả các thuật toán này đều tốt để đưa ra dự đoán tốt về giá cổ phiếu nhưng **Cây quyết định** và **hồi quy rừng ngẫu nhiên** là tốt nhất thuật toán hồi quy giữa chúng sau khi so sánh kết quả. Toàn bộ quá trình được thực hiện trong bối cảnh học máy. Các thuật toán và hệ thống được hệ thống truyền thống có thể không hiệu quả giải quyết vấn đề liên quan đến lượng dữ liệu khổng lồ này và có thể dẫn đến hệ thống chạy rất chậm và không thể mang lại kết quả tốt nhất và quả dự đoán chính xác | Linear Regression, Multi-Variant Linear Regression, Random Forest Regressor, Extra Tree Regressor | dữ liệu được lấy trong 5 năm qua giá cổ phiếu lịch sử được lấy làm thành viên tập dữ liệu cho tất cả các công ty hiện có trên chỉ số S&P 500. | date, close, open, high, low |

| | | | | | | |
|---|---|---|---|---|---|---|
| Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange | 2020 | Mehar Vijha, Deeksha Chandolab, Vinay Anand Tikkiwalb, Arun Kumar | Phân tích so sánh dựa trên các giá trị RMSE, MAPE và MBE chỉ ra rõ ràng rằng ANN đưa ra dự đoán tốt hơn về giá cổ phiếu so với RF. Kết quả bài nghiên cứu cho thấy các giá trị tốt nhất thu được từ mô hình **ANN** cho RMSE (0,42), MAPE (0,77) và MBE (0,013). Đối với công việc trong tương lai, các mô hình học sâu có thể được phát triển để xem xét các bài báo tài chính cùng với các thông số tài chính như giá đóng cửa, khối lượng giao dịch, báo cáo lãi lỗ, v.v., để có thể có kết quả tốt hơn. | Random Forest, Artificial Neural Network | Dữ liệu lịch sử của năm công ty đã được thu thập từ Yahoo Finance . Bộ dữ liệu bao gồm 10 dữ liệu năm từ 4/5/2009 đến 4/5/2019 của Nike, Goldman Sachs, Johnson and Johnson, Pfizer và JP Morgan Chase và Co. | High, Low, Open, Close, Adjacent close and Volume. |
| Estimating | July 2019 | Na Zenga | Trong nghiên cứu này, thuật toán RF được sử | random forest | Dữ liệu mẫu trường AGB đồng | AGB, EVI, NDVI, MAP, MAT, |

| | | | | | | |
|---|---|---|---|---|---|---|
| grassland aboveground biomass on the Tibetan Plateau using a random forest algorithm | | , Xiaoli Rena , Honglin Hea , Li Zhanga , Dan Zhao , Rong Gea , Pan Lie ,Zhongen Niua | dụng để kết hợp dữ liệu quan sát thực địa, chỉ số thảm thực vật viễn thám, dữ liệu khí tượng và dữ liệu địa hình để ước tính đồng cỏ AGB trên Cao nguyên Tây Tạng từ năm 2000 đến 2014. Thông qua so sánh thực nghiệm, 5 biến được chọn làm biến đầu vào của mô hình RF để ước tính AGB của đồng cỏ, bao gồm cả viễn thám VIs (NDVI và EVI), biến khí tượng (MAT và MAP) và biến địa hình (độ cao). Cuối cùng, nghiên cứu đã chứng minh rằng RF là một giải pháp hiệu quả phương pháp trong ước tính AGB đồng cỏ trên Cao nguyên Tây Tạng. | | cỏ được thu thập từ hai nguồn khác nhau: (1) tài liệu được ghi lại (Yang et al., 2010) và (2) đo đạc trường. Cuộc khảo sát thực địa được tiến hành tại khu vực đang phát triển mùa (tháng 5-tháng 9) trong giai đoạn 2005–2014. | Altitude, Slope, Aspect |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |

# Chapter 3: Methodology

## 3.1. Model description

### 3.1.1. Random Forest  Ensemble

**Random Forest** là một thuật toán học máy sử dụng kỹ thuật **Ensemble** (kết hợp) để giải quyết bài toán **phân loạ**i và **dự đoán**. Nó kết hợp nhiều cây quyết định (decision tree) để tạo thành một mô hình dự đoán mạnh mẽ hơn.

Thuật toán Random Forest hoạt động như sau:

- Lấy ngẫu nhiên một mẫu từ tập dữ liệu (có thể lấy được trùng lặp).
- Xây dựng một cây quyết định trên tập dữ liệu này. Khi xây dựng cây, thay vì sử dụng toàn bộ tập dữ liệu, chỉ sử dụng một phần để giảm thiểu việc overfitting.
- Lặp lại quá trình trên k lần (k là một tham số được định trước), để tạo ra k cây quyết định.
- Khi cần đưa ra dự đoán cho một mẫu mới, Random Forest sẽ trả về giá trị dự đoán bằng cách lấy trung bình của các giá trị dự đoán từ các cây quyết định.
- Nhờ vào kỹ thuật Ensemble, Random Forest giảm thiểu tình trạng overfitting và cải thiện khả năng dự đoán của mô hình. Ngoài ra, Random Forest có khả năng xử lý dữ liệu nhiễu và không cần quá nhiều xử lý dữ liệu trước khi huấn luyện mô hình.

Random Forest được sử dụng rộng rãi trong nhiều lĩnh vực như kinh tế, y học, tài chính, thương mại điện tử và các bài toán học máy phức tạp.

### 3.1.2. LightGBM

LightGBM là một khung tăng cường độ dốc phân tán (distributed gradient boosting framework), mã nguồn mở và miễn phí được phát triển bởi Microsoft. Nó được thiết kế để hoạt động nhanh, hiệu quả và có thể mở rộng, khiến nó trở nên lý tưởng cho các

vấn đề học máy quy mô lớn. LightGBM đã được chứng minh là rất hiệu quả đối với nhiều tác vụ khác nhau, bao gồm cả dự báo chuỗi thời gian.

Nguyên lý hoạt động của LightGBM bao gồm các bước sau:

- Khởi tạo cây quyết định với một lá.
- Tính toán gradient và hessian cho tất cả các điểm dữ liệu và lưu trữ chúng vào một ma trận.
- Chọn một tập con của các điểm dữ liệu để xây dựng cây con. LightGBM sử dụng phương pháp "leaf-wise" để chọn các lá có tỷ lệ lớn nhất của giá trị gradient và hessian. Điều này giúp LightGBM có thể xây dựng cây nhanh hơn bằng cách giảm số lượng lá cần tạo.
- Tính toán các gradient và hessian cho mỗi lá, và sử dụng chúng để tối ưu hóa các trọng số cho lá.
- Lặp lại các bước 3-4 cho đến khi đạt đến số lượng lá tối đa hoặc cây không còn cải thiện được nữa.
- Kết hợp các cây để tạo ra mô hình dự đoán cuối cùng.

LightGBM sử dụng nhiều kỹ thuật tối ưu hóa và sử dụng GPU để đạt được hiệu suất tốt hơn. Nó cũng hỗ trợ việc tăng tốc bằng cách sử dụng phiên bản song song trên nhiều CPU.

LightGBM là một thư viện rất mạnh mẽ để dự đoán chuỗi thời gian (time series forecasting) bởi vì nó cung cấp một cách tiếp cận tối ưu cho bài toán dự đoán chuỗi thời gian.

LightGBM là một thư viện rất mạnh mẽ để dự đoán chuỗi thời gian bởi vì nó cung cấp một cách tiếp cận tối ưu cho bài toán dự đoán chuỗi thời gian. Dựa vào những giá trị lịch sử, LightGBM có thể sử dụng để dự đoán giá cổ phiếu tương lai, doanh số bán hàng tương lai hoặc những số liệu trong tương lai mà con người có nhu cầu cần được dự đoán.

## 3.2. Các kỹ thuật trong time series

### 3.2.1. Sliding Window For Time Series Data

Số lượng quan sát được ghi lại trong một thời gian nhất định trong tập dữ liệu chuỗi thời gian có ý nghĩa quan trọng. Theo truyền thống, các tên khác nhau được sử dụng:

- Chuỗi thời gian đơn biến: Đây là những bộ dữ liệu chỉ quan sát một biến duy nhất tại mỗi thời điểm, chẳng hạn như nhiệt độ mỗi giờ. Ví dụ trong phần trước là tập dữ liệu chuỗi thời gian đơn biến.
- Chuỗi thời gian đa biến: Đây là những bộ dữ liệu có hai hoặc nhiều biến được quan sát tại mỗi thời điểm.

Hầu hết các phương pháp phân tích chuỗi thời gian và thậm chí cả sách về chủ đề này đều tập trung vào dữ liệu đơn biến. Điều này là do nó đơn giản nhất để hiểu và làm việc với. Dữ liệu đa biến thường khó xử lý hơn. Nó khó mô hình hơn và thường nhiều phương pháp cổ điển không hoạt động tốt. Trong bài nghiên cứu này, nhóm sẽ sử dụng cả phương pháp đơn biến và đa biến để xem xét.
Với kiểu dữ liệu chuỗi thời gian, bao gồm trường về thời gian và trường về giá đóng cửa là một ví dụ về chuỗi thời gian đơn biến.

| Date | PCG | PLX | PVB | PVC | PVO |
|---|---|---|---|---|---|
| 2017-01-03 00:00:00 | 8,6 | | 10,9 | 8 | 3,5 |
| 2017-01-04 00:00:00 | 8,6 | | 10,9 | 8,1 | 3,5 |
| 2017-01-05 00:00:00 | 8,6 | | 10,5 | 8,1 | 3,6 |
| 2017-01-06 00:00:00 | 8,6 | | 10,3 | 8 | 3,6 |
| 2017-01-07 00:00:00 | 8,6 | | 10,33333 | 8,033333 | 3,633333 |
| 2017-01-08 00:00:00 | 8,6 | | 10,36667 | 8,066667 | 3,666667 |
| 2017-01-09 00:00:00 | 8.6 | | 10.4 | 8.1 | 3.7 |

*Table 3.1: Ví dụ về chuỗi thời gian đơn biến*

Dữ liệu chuỗi thời gian có thể được xử lý như một bài toán học có giám sát vì chúng ta có thể chuyển đổi dữ liệu này thành dạng bài toán học có giám sát. Cụ thể, chúng ta

có thể sử dụng các giá trị quan sát trong quá khứ làm biến đầu vào và sử dụng giá trị quan sát kế tiếp làm biến đầu ra. Quá trình này được gọi là "sliding window" và cho phép tạo ra nhiều mẫu dữ liệu mới để sử dụng cho bài toán học có giám sát. Với việc áp dụng kỹ thuật này, chúng ta có thể sử dụng các mô hình học máy có giám sát thông thường để dự đoán giá trị của chuỗi thời gian trong tương lai.
Trong thống kê và phân tích chuỗi thời gian, điều này được gọi là phương pháp trễ (lag method). Số bước thời gian trước đó được gọi là chiều rộng cửa sổ hoặc kích thước trễ.
Cửa sổ trượt này là cơ sở cho cách chuyển đổi bất kỳ tập dữ liệu chuỗi thời gian nào thành một vấn đề học có giám sát:

- Chúng ta có thể chuyển đổi chuỗi thời gian thành một bài toán học có giám sát dựa trên giá trị chuỗi thời gian được xác định trước hoặc được nhãn.
- Chúng ta có thể thấy rằng khi tập dữ liệu chuỗi thời gian được chuẩn bị theo cách này thì bất kỳ thuật toán học máy tuyến tính và phi tuyến nào cũng có thể được áp dụng, miễn là thứ tự các hàng được bảo tồn.
- Chúng ta có thể thấy cách tăng kích thước cửa sổ trượt để bao gồm nhiều bước thời gian trước đó hơn.

Chúng ta có thể thấy cách tiếp cận của cửa sổ trượt có thể được sử dụng trên chuỗi thời gian có nhiều giá trị hơn một, hoặc gọi là chuỗi thời gian đa biến.

### 3.2.2. Walk-forward validation

Walk-Forward Validation là một kỹ thuật đánh giá mô hình trong machine learning cho các bài toán dự báo chuỗi thời gian (time series forecasting). Kỹ thuật này được sử dụng để đánh giá hiệu suất của mô hình trên các tập dữ liệu thời gian khác nhau.

Thay vì chia dữ liệu thời gian thành các tập train và test tĩnh và đánh giá hiệu suất của mô hình trên tập test một lần duy nhất, Walk-Forward Validation sử dụng một cách tiếp cận khác. Nó chia dữ liệu theo một cách trượt (sliding) dọc theo thời gian và thực

hiện các vòng lặp huấn luyện và đánh giá mô hình trên các tập dữ liệu con theo thời gian.

Ví dụ, trong một mô hình sử dụng Walk-Forward Validation với cửa sổ trượt (sliding window) là 1 năm, dữ liệu được chia thành các tập train và test tương ứng với từng năm, mô hình sẽ được huấn luyện trên dữ liệu của các năm trước đó, và được đánh giá trên dữ liệu của năm tiếp theo. Quá trình này được tiếp tục cho đến khi mô hình được đánh giá trên toàn bộ dữ liệu thời gian.

Walk-Forward Validation giúp đánh giá hiệu suất của mô hình trên các tập dữ liệu thời gian khác nhau, giúp mô hình có khả năng tổng quát hơn và đưa ra dự báo chính xác hơn.

Sẽ không hợp lệ nếu mô hình phù hợp với dữ liệu từ tương lai và để nó dự đoán quá khứ. Mô hình phải được đào tạo về quá khứ và dự đoán tương lai. Điều này có nghĩa là không thể sử dụng các phương pháp ngẫu nhiên hóa tập dữ liệu trong quá trình đánh giá, chẳng hạn như xác thực chéo k-fold. Thay vào đó, chúng ta phải sử dụng một kỹ thuật gọi là xác thực từ từ.

Trong xác thực walk-forward, trước tiên, tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra bằng cách chọn điểm giới hạn, ví dụ: tất cả dữ liệu ngoại trừ tháng 12 được sử dụng để đào tạo và dữ liệu tháng 12 được sử dụng để kiểm tra.

### 3.2.3. One-step prediction

One-step prediction là phương pháp dự đoán giá trị của một biến số tại thời điểm tiếp theo dựa trên giá trị của biến số tại thời điểm hiện tại. Ví dụ, nếu ta đang dự đoán giá trị của một chứng khoán vào ngày mai dựa trên giá trị của nó vào ngày hôm nay, thì đó là một one-step prediction. Nói cách khác, one-step prediction là phương pháp dự đoán giá trị của một chuỗi thời gian bằng cách sử dụng thông tin về các giá trị của

chuỗi thời gian tại các thời điểm trước đó để dự đoán giá trị của nó tại thời điểm tiếp theo.

Nếu chúng tôi quan tâm đến việc đưa ra dự báo một bước, ví dụ: chúng ta có thể đánh giá mô hình bằng cách huấn luyện trên tập dữ liệu huấn luyện và dự đoán bước đầu tiên trong tập dữ liệu thử nghiệm. Sau đó, chúng ta có thể thêm quan sát thực tế từ tập thử nghiệm vào tập dữ liệu huấn luyện, điều chỉnh lại mô hình, sau đó để mô hình dự đoán bước thứ hai trong tập dữ liệu thử nghiệm.

Việc lặp lại quy trình này cho toàn bộ tập dữ liệu thử nghiệm sẽ đưa ra dự đoán một bước cho toàn bộ tập dữ liệu thử nghiệm, từ đó có thể tính toán thước đo lỗi để đánh giá kỹ năng của mô hình.

### 3.2.4. Feature Selection

Feature selection là quá trình chọn lọc các thuộc tính (features) quan trọng nhất mà không giảm đi độ chính xác của mô hình.

Feature selection giúp cải thiện tốc độ huấn luyện và dự đoán, giảm độ phức tạp của mô hình và tránh overfitting.

Một số phương pháp thông dụng trong feature selection bao gồm:Feature Importance and RFE

#### 3.2.4.1. Feature Importance

là một phương pháp được sử dụng trong Machine Learning để đánh giá độ quan trọng của các features đóng vai trò trong việc xây dựng một mô hình.

Một số mô hình đánh giá feature importance bằng cách tính toán độ quan trọng của từng feature trong quá trình huấn luyện mô hình, dựa trên một số metric như entropy, gain, Gini index, hay coefficient trong linear models, ...

Feature Importance được sử dụng để chọn lọc các features quan trọng nhất khi có quá nhiều features, giúp cho mô hình trở nên đơn giản và tăng độ chính xác trong huấn luyện. Ngoài ra, Feature Importance còn giúp tìm hiểu sự ảnh hưởng của các features đến kết quả dự đoán, giúp tối ưu hóa các feature để đạt được độ chính xác tốt nhất.

Trong quá trình tính toán độ quan trọng của các features, một số mô hình sử dụng các phương pháp khác nhau để tính toán. Ví dụ:

Decision Trees: Đối với một decision tree, feature importance của một feature được tính bằng cách tính toán độ giảm impurity do sự chia nhánh của feature này gây ra. Trong đó, impurity được tính bằng metric như entropy hay Gini impurity. Những feature gây ra sự giảm impurity lớn hơn sẽ được coi là quan trọng hơn.

Random Forest: Đối với Random Forest, feature importance của một feature được tính toán theo đó nó làm giảm được độ sai khác trung bình (mean decrease impurity) của các cây con (subtree). Các thuật toán như ExtraTreesClassifier cũng có thể được sử dụng để tính toán feature importance.

Gradient Boosting: Đối với gradient boosting, feature importance của một feature được tính toán theo đó nó đóng góp bao nhiêu vào việc giảm lỗi trong một model được huấn luyện dựa trên gradient boosting (việc giảm lỗi được tính bằng mean squared error - MSE - hay mean absolute error - MAE).

Linear Regression/Logistic Regression: Đối với các mô hình linear regression hay logistic regression, feature importance của một feature thường được tính bằng giá trị tuyệt đối của hệ số (absolute coefficient value) tương ứng với feature đó trong phương trình mô hình. Các feature có giá trị hệ số lớn hơn sẽ được hiểu là quan trọng hơn.

Các phương pháp tính toán feature importance này đều hữu ích trong việc giúp tìm ra những feature quan trọng nhất để giảm số lượng features cần sử dụng để xây dựng mô hình và tăng hiệu quả của mô hình Huấn luyện.

#### 3.2.4.2. RFE

(Recursive Feature Elimination) là một thuật toán chọn tập features trong machine learning. Thuật toán này hoạt động bằng cách sử dụng một mô hình học máy (machine learning model) và đánh giá độ quan trọng của các features. Sau đó, nó loại bỏ các features không quan trọng và sử dụng lại các remaining features để huấn luyện một mô hình mới. Quá trình này lặp lại cho đến khi đạt được số lượng features cần chọn.

Cụ thể, quá trình RFE là như sau:

- Huấn luyện một mô hình học máy trên toàn bộ tập features.
- Đánh giá độ quan trọng của mỗi feature thông qua mô hình học được trên tập features hiện tại.
- Loại bỏ feature có độ quan trọng thấp nhất.
- Huấn luyện lại mô hình học máy trên các features còn lại.
- Lặp lại quá trình cho đến khi đạt đến số features đã định trước.

RFE thường được sử dụng trong các bài toán với dữ liệu có nhiều features và tránh overfitting. Thuật toán này có thể được sử dụng với nhiều loại mô hình học máy, nhưng thường được sử dụng với các mô hình linear regression và logistic

#### 3.2.4.3. RPECV

Recursive Partitioning and Evolutionary Selection (RPES) Cross-Validation, là một kỹ thuật chọn mô hình (model selection) cho các mô hình tuyến tính trong machine learning.

RPES Cross-Validation là một phương pháp tự động tối ưu hóa các tham số trong mô hình tuyến tính bằng cách sử dụng một thuật toán tối ưu hóa đa nhiệm để đồng thời tối ưu hóa cả số lượng biến độc lập và các tham số liên quan đến mô hình.

Trong quá trình chạy RPES Cross-Validation, thuật toán sẽ chia tập dữ liệu thành các tập con (folds) và tiến hành kiểm định chéo (cross-validation) trên từng tập con. Sau đó, thuật toán sẽ sử dụng các thông số tối ưu để đưa ra dự đoán trên tập kiểm tra và tính toán sai số dự đoán. Tiếp theo, thuật toán sẽ điều chỉnh lại các thông số và biến độc lập để tối ưu hóa sai số dự đoán. Quá trình này sẽ được lặp đi lặp lại cho đến khi sai số dự đoán được giảm đến mức thấp nhất có thể.

#### 3.2.4.4. SHAP

Shap (Shapley Additive exPlanations) là một phương pháp giải thích được sử dụng trong machine learning để hiểu rõ hơn về cách các đặc trưng của dữ liệu ảnh hưởng đến kết quả dự đoán của mô hình. Khi một mô hình machine learning thực hiện một dự đoán, Shap tính toán đóng góp của từng đặc trưng vào giá trị dự đoán.

Shap dựa trên lý thuyết giá trị Shapley từ lĩnh vực lý thuyết trò chơi. Lý thuyết này giải thích cách một phần tử trong một tập hợp các đặc trưng đóng góp vào giá trị tổng cộng của một hàm. Trong trường hợp của Shap, giá trị tổng cộng là kết quả dự đoán của mô hình và các đặc trưng đóng vai trò như các phần tử.

Để tính toán giá trị Shapley cho mỗi đặc trưng, Shap sử dụng một quy trình hợp lý và xấp xỉ. Nó chạy qua tất cả các tập con của các đặc trưng và tính toán đóng góp của từng tập con đối với giá trị dự đoán. Kết quả là một ước tính của giá trị Shapley cho mỗi đặc trưng.

Thông qua giá trị Shapley, người dùng có thể hiểu được độ quan trọng của từng đặc trưng đối với dự đoán của mô hình. Nếu một đặc trưng có giá trị Shapley lớn, nghĩa là

nó có ảnh hưởng lớn đến kết quả dự đoán. Ngoài ra, Shap cũng cung cấp thông tin về hướng tương quan của mỗi đặc trưng, liệu có ảnh hưởng tích cực hay tiêu cực đến giá trị dự đoán.

# Chapter 4: Experimental Results

## 4.1. Model Random Forest

### 4.1.1. Random Forest cho đơn biến

Chúng tôi sử dụng bộ dữ liệu chuỗi thời gian có cấu trúc **đơn biến** với **mục đích** sử dụng mô hình để đưa ra dự báo **1 tháng trong tương lai** (vượt ra khỏi thời gian tập dữ liệu thu thập).

| Date | PCG | PLX | PVB | PVC | PVO |
|---|---|---|---|---|---|
| 2017-01-03 00:00:00 | 8,6 | | 10,9 | 8 | 3,5 |
| 2017-01-04 00:00:00 | 8,6 | | 10,9 | 8,1 | 3,5 |
| 2017-01-05 00:00:00 | 8,6 | | 10,5 | 8,1 | 3,6 |
| 2017-01-06 00:00:00 | 8,6 | | 10,3 | 8 | 3,6 |
| 2017-01-07 00:00:00 | 8,6 | | 10,33333 | 8,033333 | 3,633333 |
| 2017-01-08 00:00:00 | 8,6 | | 10,36667 | 8,066667 | 3,666667 |
| 2017-01-09 00:00:00 | 8,6 | | 10,4 | 8,1 | 3,7 |
| 2017-01-10 00:00:00 | 8,6 | | 10,8 | 8 | 3,5 |
| 2017-01-11 00:00:00 | 8,6 | | 10,6 | 8 | 3,6 |
| 2017-01-12 00:00:00 | 8,6 | | 10,7 | 8 | 3,6 |
| 2017-01-13 00:00:00 | 7,8 | | 10,6 | 8 | 3,5 |
| 2017-01-14 00:00:00 | 7,566667 | | 10,33333 | 8 | 3,533333 |
| 2017-01-15 00:00:00 | 7,333333 | | 10,06667 | 8 | 3,566667 |

*Table 4.1: Dữ liệu chuỗi thời gian của 5 mã Stock*

Xét riêng với mã **Stock PCG**, chúng tôi thực hiện chia tập dữ liệu theo thời gian: Từ tháng 03 năm 2017 đến đầu tháng 11 năm 2021 sẽ là tập **train**; và 2 tháng còn lại sẽ là tập **test**. Theo kinh nghiệm từ những nghiên cứu trước, đối với thuật toán này chúng tôi cũng đã thực nghiệm thử và sai để lựa chọn cách **chia dữ liệu train và test theo thời gian**.

Như đã trình bày, đầu tiên chúng tôi cần phải thực hiện **chuyển đổi dữ liệu chuỗi thời gian** để phù hợp với bài toán học máy có giám sát. Sau khi thực hiện chuyển đổi, chúng tôi nhận lại một mảng numpy của các mẫu dữ liệu đầu vào và đầu ra có kích thước phù hợp cho mô hình học máy. Đồng thời, chúng tôi thực hiện **kết hợp kĩ thuật Sliding window**.

Quy trình chuyển đổi một tập dữ liệu chuỗi thời gian thành một tập dữ liệu được giám sát. Tập dữ liệu được giám sát có nghĩa là mỗi mẫu dữ liệu sẽ có một giá trị đầu ra được xác định trước đó (ví dụ: giá cổ phiếu của ngày hôm sau được dự đoán dựa trên giá cổ phiếu của n ngày trước đó). Cụ thể, chúng ta kiểm tra xem dữ liệu là một danh sách hay một mảng nhiều chiều. Nếu dữ liệu là một danh sách thì chúng ta sẽ tạo một đối tượng DataFrame với cột duy nhất, ngược lại, nếu data là một mảng nhiều chiều, chúng ta sẽ tạo DataFrame với số lượng cột bằng với số lượng biến của mảng đó. Ở trường hợp này là đơn biến.

Sau khi dữ liệu được chuyển đổi, chúng tôi nhận lại một mảng như bên dưới:

```
array([[ 7.17155172,  5.90089286,  8.08709677,  8.5       ,  7.82580645,
         6.97      ,  7.63225806],
       [ 5.90089286,  8.08709677,  8.5       ,  7.82580645,  6.97      ,
         7.63225806,  6.99032258],
       [ 8.08709677,  8.5       ,  7.82580645,  6.97      ,  7.63225806,
         6.99032258,  8.26777778],
       [ 8.5       ,  7.82580645,  6.97      ,  7.63225806,  6.99032258,
         8.26777778,  8.66989247],
       [ 7.82580645,  6.97      ,  7.63225806,  6.99032258,  8.26777778,
         8.66989247,  8.93      ],
       [ 6.97      ,  7.63225806,  6.99032258,  8.26777778,  8.66989247,
         8.93      ,  8.87419355],
```

*Figure 4.1: Hình dạng của Dataframe trả về*

Chúng tôi tiến hành khởi tạo và huấn luyện mô hình Random Forest Regression với 1000 cây quyết định. Sau đó, chúng tôi thực hiện chạy mô hình theo quy trình Walk-forward validation và thu được kết quả cụ thể như hình dưới đây:

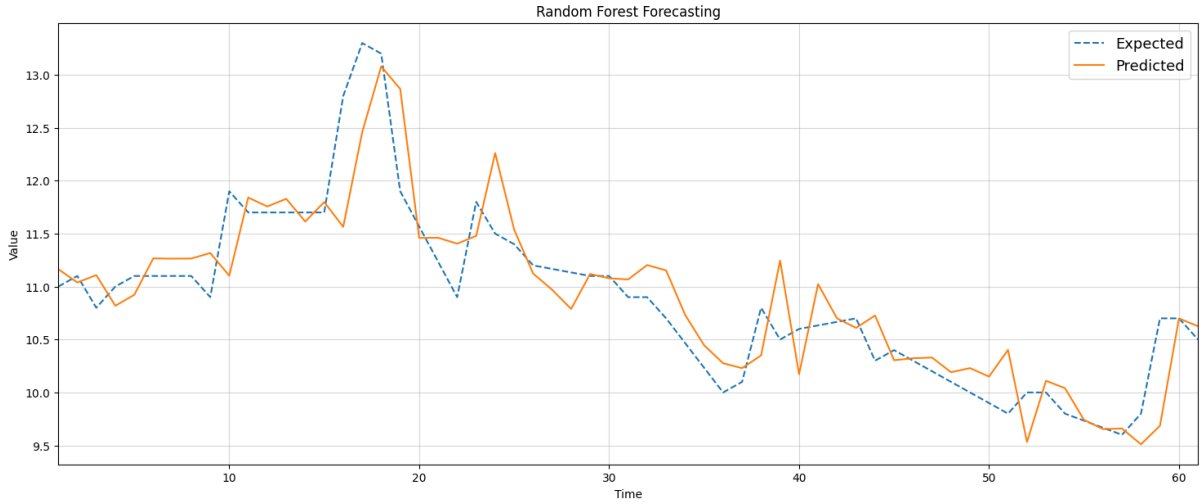
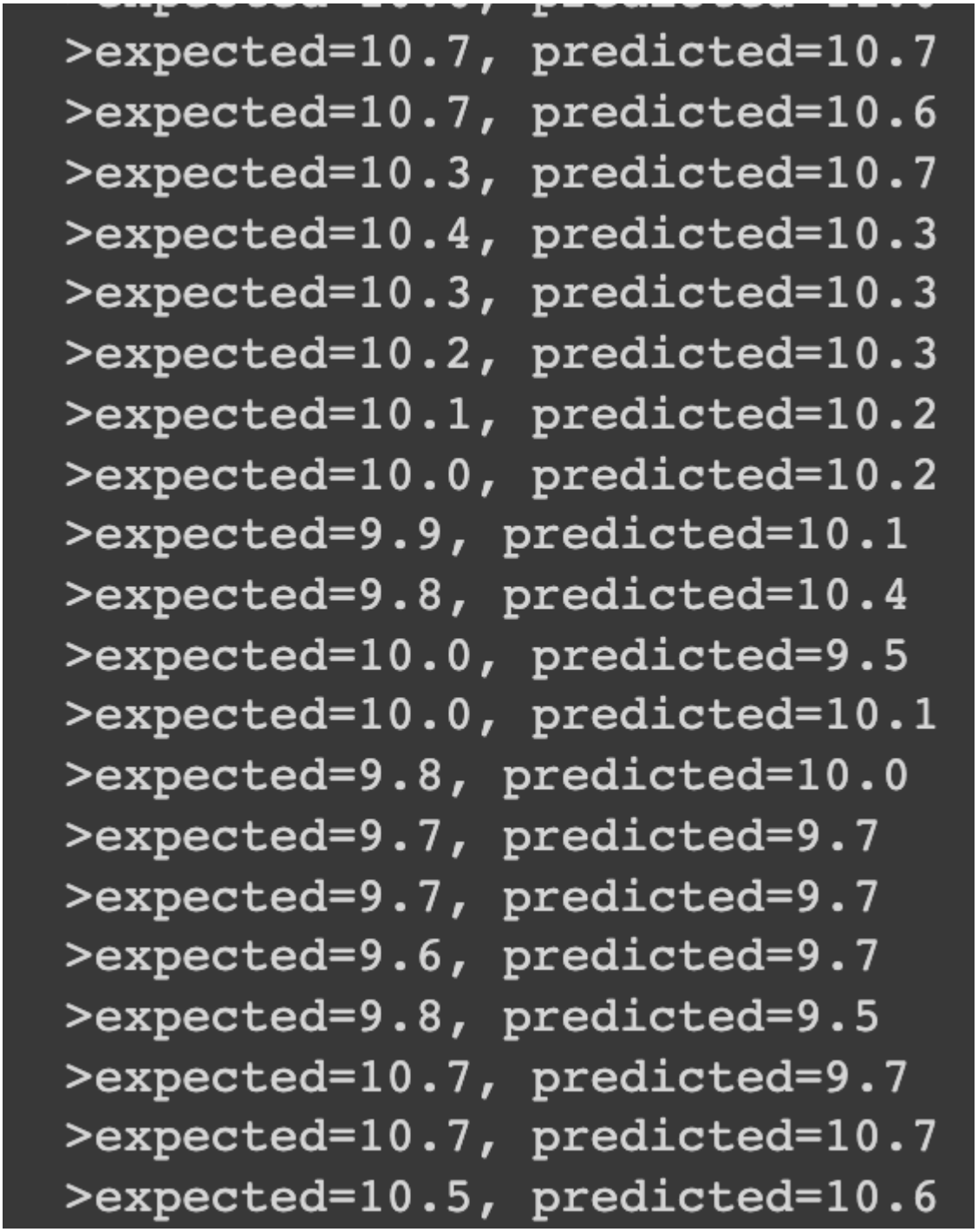



*Figure 4.2: Kết quả dự đoán giá đóng của trên tập test của mã PCG*

Chúng ta có thể thấy, kết quả thu được có khá tốt. Đường dự đoán có xu hướng khá tương đồng với dữ liệu thực tế.




Đánh giá kết quả của mô hình. Ta có thể thấy các chỉ số khá tốt. Điều này được đóng góp từ bởi cấu hình mô hình Rừng ngẫu nhiên cuối cùng được chọn. Chúng tôi đã thực hiện quy tắc thử và sai để tìm ra các đối số phù hợp với mô hình Random Forest.

```
MAE: 0.281
RMSE: 0.390
MAPE: 2.542%
95% confidence interval: (10.726274584885012, 11.099495642093583)
```

Như vậy, có thể thấy, mô hình hoạt động khá tốt trong việc dự đoán giá trong vòng 2 tháng đối với mã Stock này. Đặc biệt chỉ số MAPE khá nhỏ, thể hiện được khả năng dự đoán ổn định của mô hình khi so sánh với các mô hình khác.

Tiếp theo, chúng tôi thực hiện dự báo ngoài mẫu. Chúng tôi mong muốn có thể dự đoán giá 1 tháng đầu tiên sau khi kết thúc thu thập dữ liệu. Điều này được thực hiện giống hệt với việc đưa ra dự đoán trong quá trình đánh giá mô hình, vì chúng tôi luôn muốn đánh giá một mô hình bằng cách sử dụng cùng một quy trình mà chúng tôi dự kiến sẽ sử dụng khi mô hình được sử dụng để đưa ra dự đoán về dữ liệu mới.

Dễ hiểu hơn, chúng tôi thực hiện khớp mô hình Rừng ngẫu nhiên cuối cùng trên tất cả dữ liệu có sẵn và đưa ra dự đoán tuân tự theo quy trình đánh giá walk-forward cho toàn tập dữ liệu. Sau khi chạy mô hình, tương tự sử dụng quy trình walk-forward validation để dự đoán ngày tiếp theo (nằm ngoài tập dữ liệu). Chúng tôi thu được kết quả dự đoán cho 30 ngày tiếp theo, tức giá cho giá của tháng đầu tiên sau khi kết thúc tập dữ liệu như hình bên dưới.

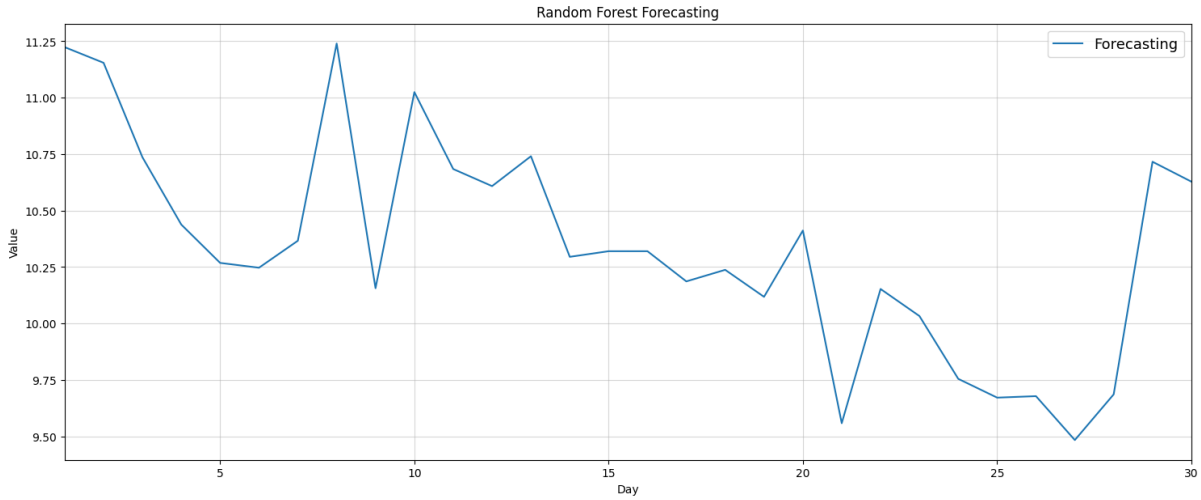



*Figure 4.3: Dự đoán giá đóng cửa 1 tháng tiếp theo của mã PCG*

Other Stocks

Chúng tôi thực hiện tương tự các bước chuẩn bị và dự đoán cho các mã stock còn lại:

- **PLX:**

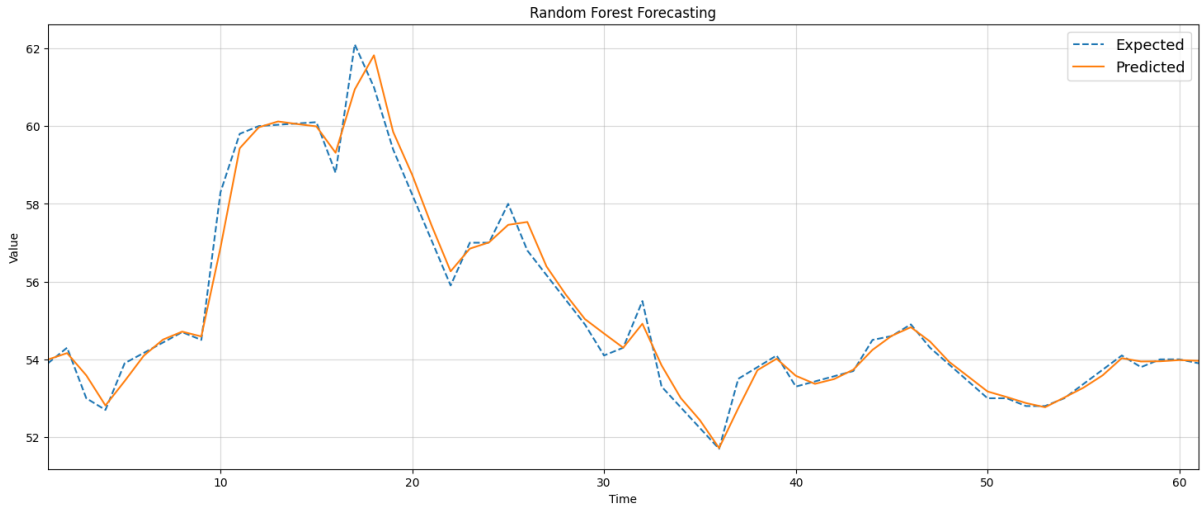



*Figure 4.4: Kết quả dự đoán giá đóng của trên tập test PLX*

```
MAE: 0.245
RMSE: 0.378
MAPE: 0.434%
95% confidence interval: (54.66091799165651, 55.85269764870213)
```

*Figure 4.5: Đánh giá kết quả dự đoán trên tập train PLX*

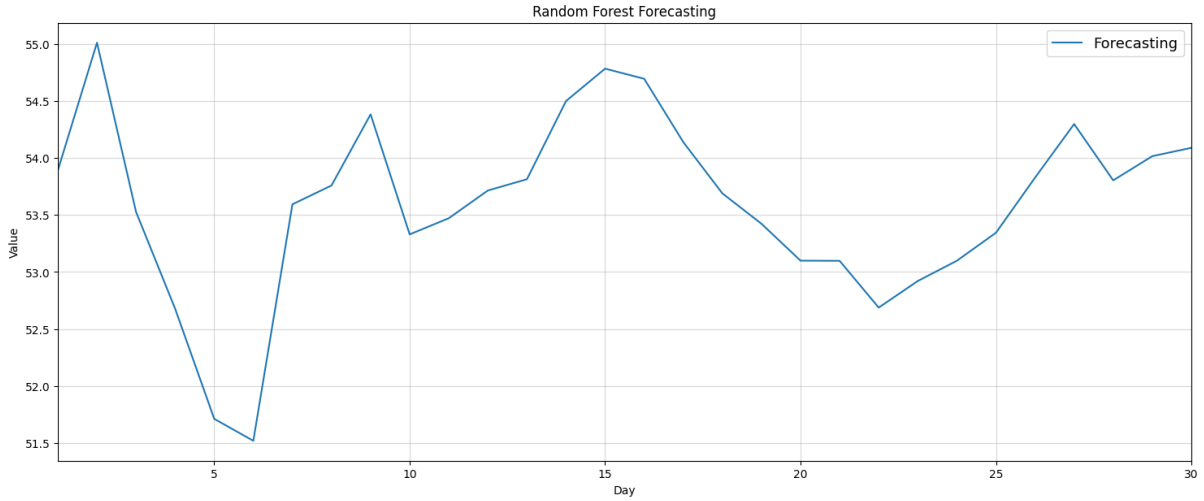


*Figure 4.6: Dự đoán giá đóng cửa 1 tháng tiếp theo của mã PLX*

- **PVB**

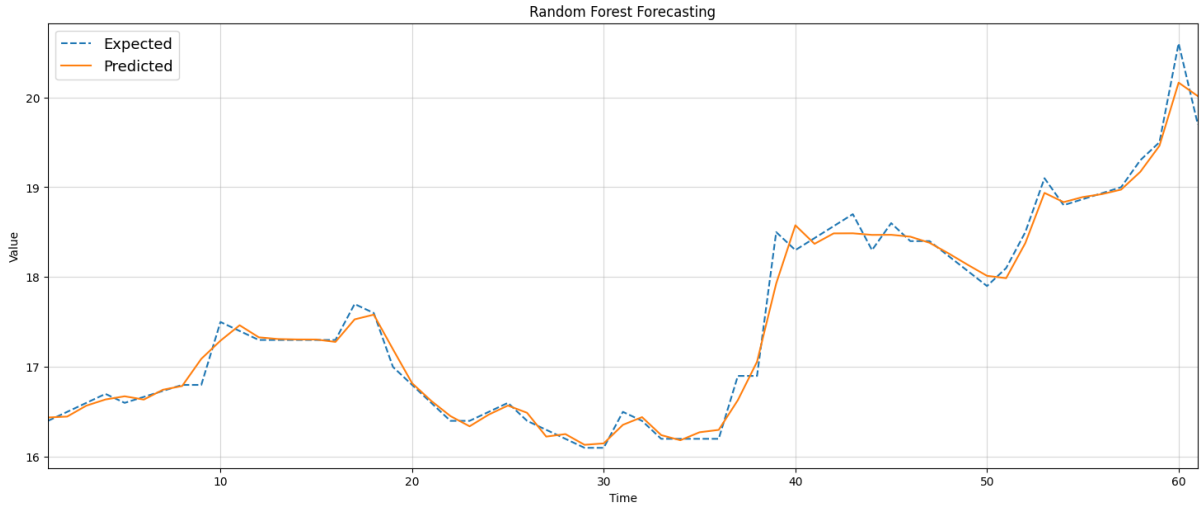



*Figure 4.7: Kết quả dự đoán giá đóng của trên tập test PVB*

```
MAE: 0.096
RMSE: 0.144
MAPE: 0.540%
95% confidence interval: (17.201910076844264, 17.72993560963109)
```

*Figure 4.8: Đánh giá kết quả dự đoán trên tập train PVB*

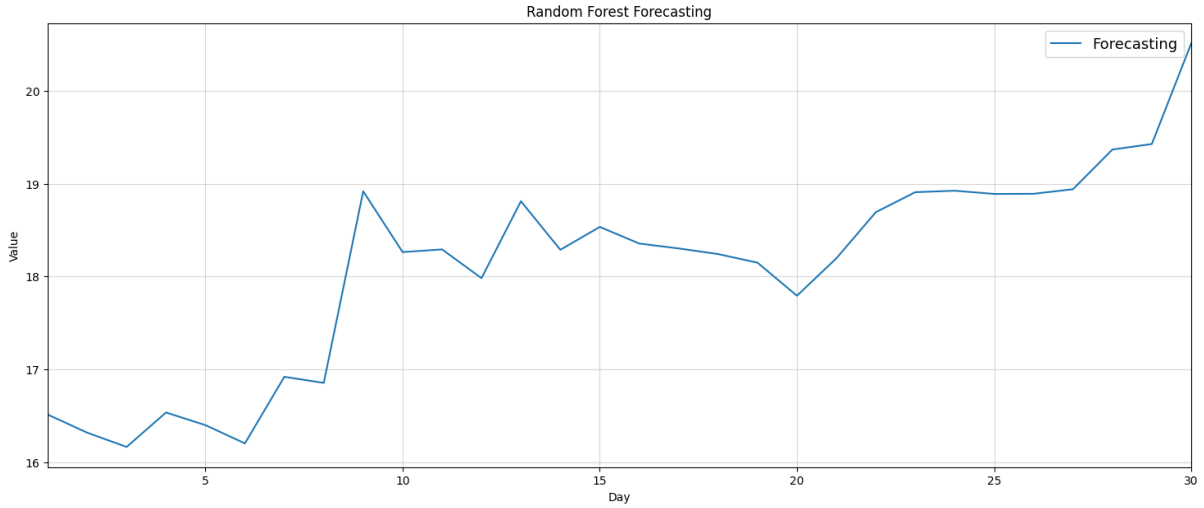


*Figure 4.9: Dự đoán giá đóng cửa 1 tháng tiếp theo của mã PVB*

- **PVO**

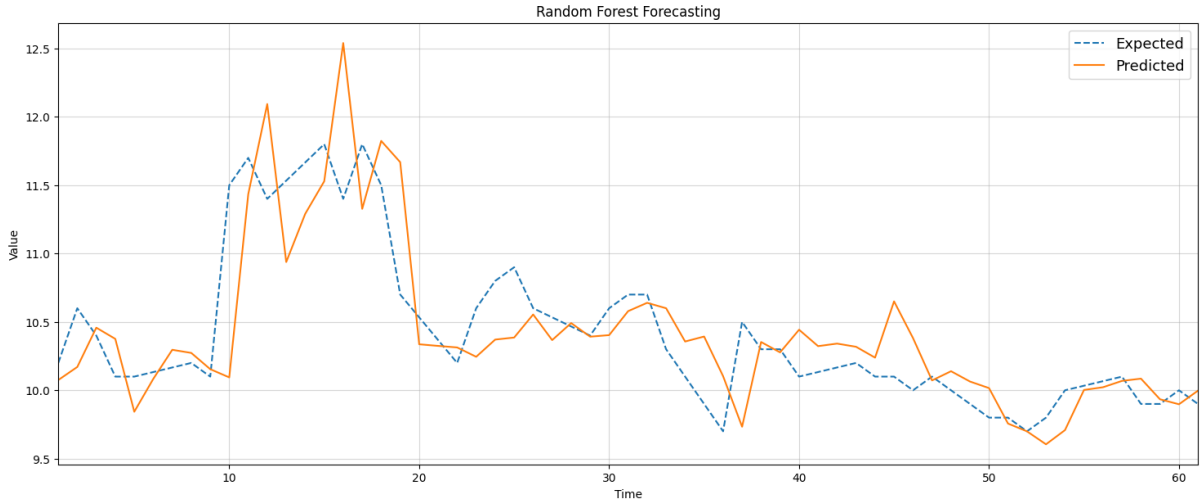



*Figure 4.10: Kết quả dự đoán giá đóng của trên tập test PVO*

```
MAE: 0.263
RMSE: 0.381
MAPE: 2.461%
95% confidence interval: (10.277056229508231, 10.571914637978168)
```

*Figure 4.11: Đánh giá kết quả dự đoán trên tập train PVO*

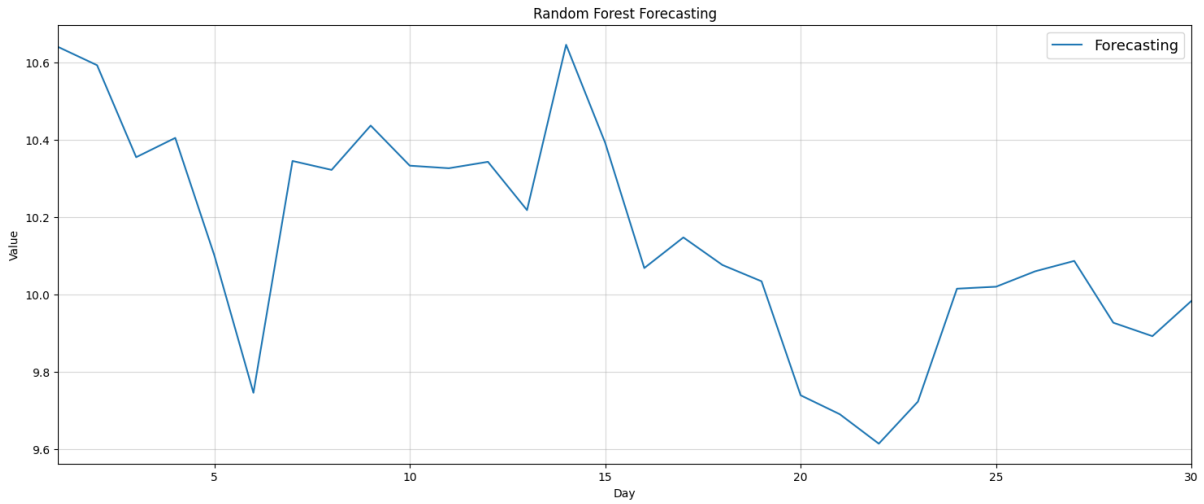


*Figure 4.12: Dự đoán giá đóng cửa 1 tháng tiếp theo của mã PVO*

- **PVC**

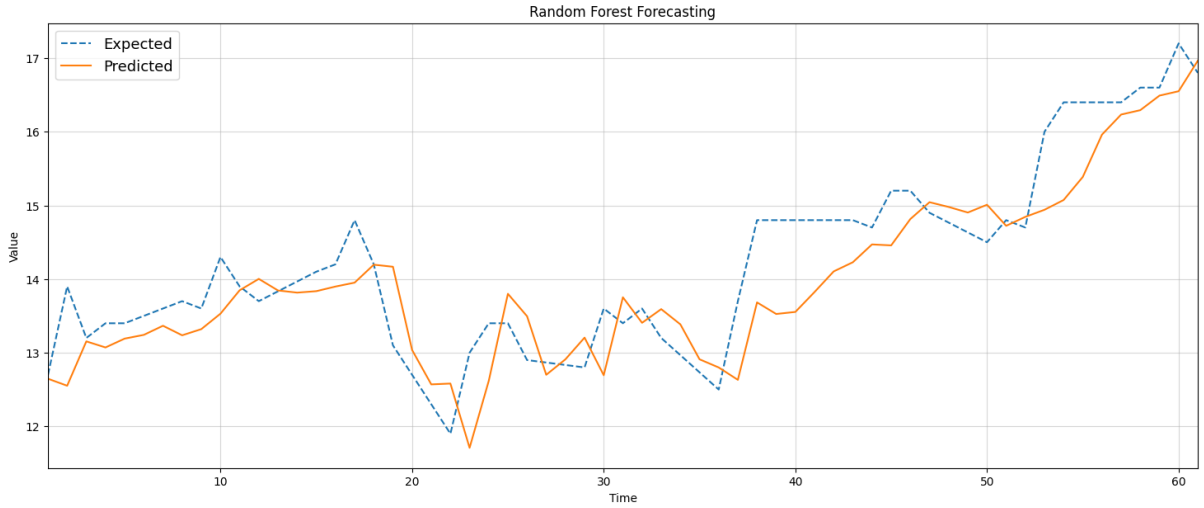



*Figure 4.13: Kết quả dự đoán giá đóng của trên tập test PVC*

```
MAE: 0.486
RMSE: 0.619
MAPE: 3.408%
95% confidence interval: (13.681055327868856, 14.25206056693992)
```

*Figure 4.14: Đánh giá kết quả dự đoán trên tập train PVC*

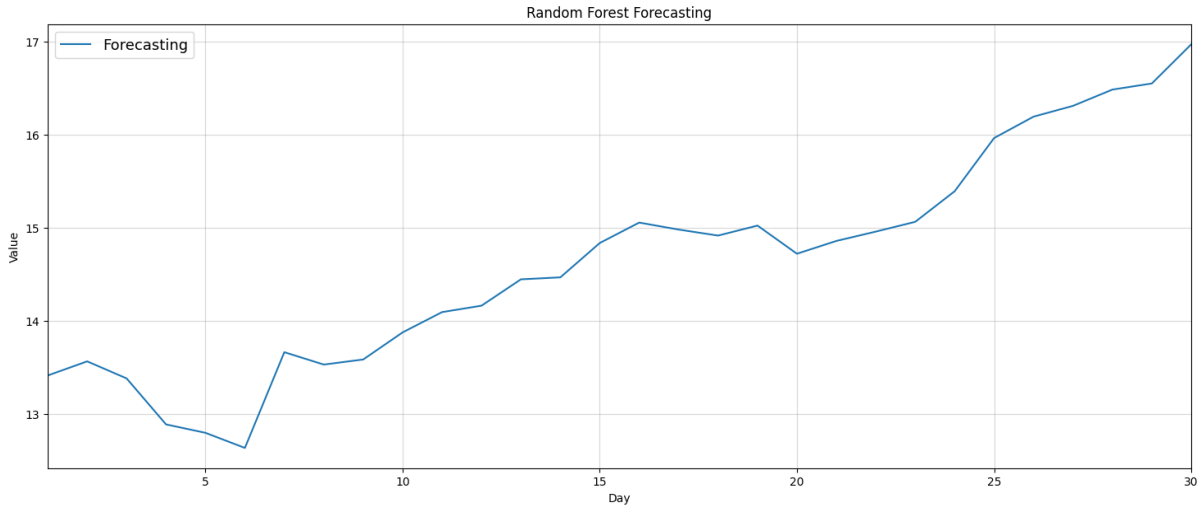


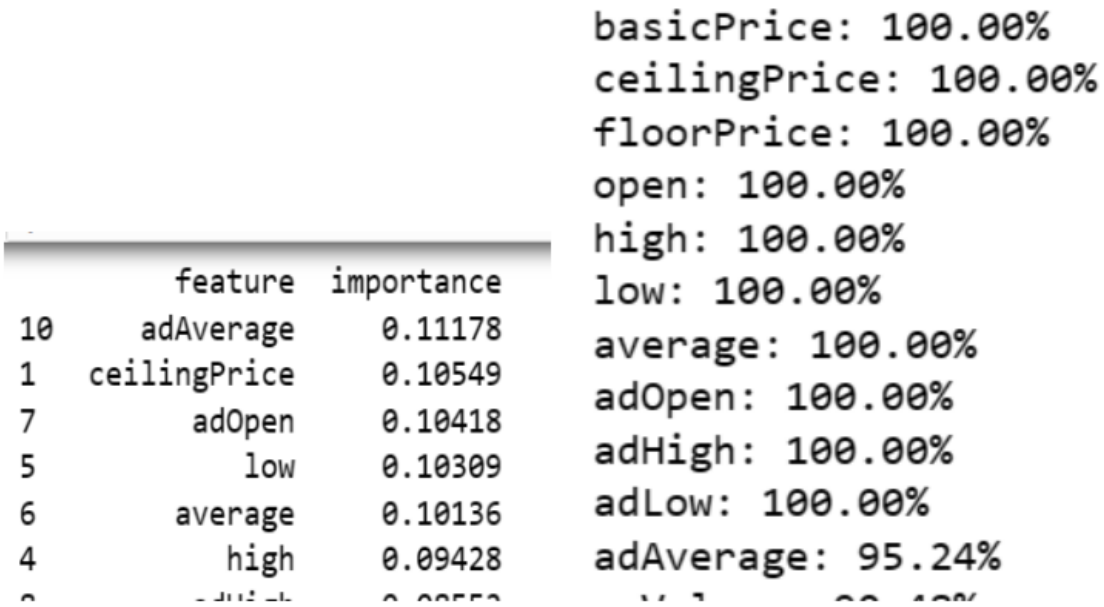
*Figure 4.15: Dự đoán giá đóng cửa 1 tháng tiếp theo của mã PVC*

Có thể thấy, các kết quả trên cho thấy mô hình cũng thể hiện được khả năng dự đoán khá tốt. Cụ thể, chỉ số MAPE tốt nhất ở mã PLX với 0.434% ở tập test và cũng khá ổn định thông qua các mã còn lại.

### 4.1.2. Random Forest cho đa biến

Sự khác biệt rõ ràng nhất khi sử dụng model cho đơn biến và đa biến của nhóm là với model đơn biến, dữ liệu sẽ chỉ bao gồm giá trị date và close. Tại mô hình đa biến, sẽ lấy toàn bộ các dữ liệu đã có trong dataset để thêm vào mô hình. Tuy nhiên, việc thêm quá nhiều trường dữ liệu như vậy sẽ ảnh hưởng tới mô hình. Vậy nên cần phải sử dụng lượng biến số phù hợp. Trong các mô hình đa biến sau, nhóm sẽ sử dụng thêm Feature Selection để lựa chọn các biến số phù hợp.

Nhóm đã thực nghiệm đầu tiên là sử dụng Feature Importance và RPE để lựa chọn các biến đầu vào để phù hợp mô hình. Về cơ bản, ý tưởng của nhóm là sẽ sử dụng Feature Importance để xem mô hình tương ứng với mã code nào thì cần các biến đầu vào nào để phù hợp. Vì với mã code sẽ cần một biến đầu vào khác nhau. Sau đó sẽ sử dụng thêm RPE để kiểm tra lại một lần nữa. Tuy nhiên cách này khá mất thời gian cũng như sẽ có một vài sai xót. Ví dụ như hai ảnh dưới, tại  Feature Importamce sẽ có các biến đầu được xem là quan trọng ( nhóm đặt các biến quan trọng là >10%), trong đó có adVerage và biến này được coi là biến quan trọng nhất, tuy nhiên tại RPE, adAvegare lại được coi là biến kém quan trọng hơn so với những biến khác.



| | feature | importance |
|---|---|---|
| 10 | adAverage | 0.11178 |
| 1 | ceilingPrice | 0.10549 |
| 7 | adOpen | 0.10418 |
| 5 | low | 0.10309 |
| 6 | average | 0.10136 |
| 4 | high | 0.09428 |

basicPrice: 100.00%
ceilingPrice: 100.00%
floorPrice: 100.00%
open: 100.00%
high: 100.00%
low: 100.00%
average: 100.00%
adOpen: 100.00%
adHigh: 100.00%
adLow: 100.00%
adAverage: 95.24%

*Figure 4.16: Feature Importance (left) and RPE (right)*

Xác định được ý tưởng so sánh hai kết quả Feature Selection trên không thực sự tốt, nhóm quyết định sử dụng RPECV để tự động tối ưu hóa các tham số trong mô hình tuyến tính bằng cách sử dụng một thuật toán tối ưu hóa đa nhiệm để đồng thời tối ưu hóa cả số lượng biến độc lập và các tham số liên quan đến mô hình.

Tùy vào mô hình mà chọn tham số RPECV cho phù hợp. Trong mô hình Random Forest, nhóm sử dụng RandomForestRegressor để estomator. KFold sẽ bằng 5.

Nhóm chạy thử nghiệm và xác định các biến quan trọng cho mô hình Random Forest đa biến mã PCG. Sau đó Lấy các biến đó tiếp tục chạy để dự đoán tệp test ( là 2 tháng sau - từ tháng 11/2021 đến tháng 12 năm 2021 ). Kết quả như hình bên dưới

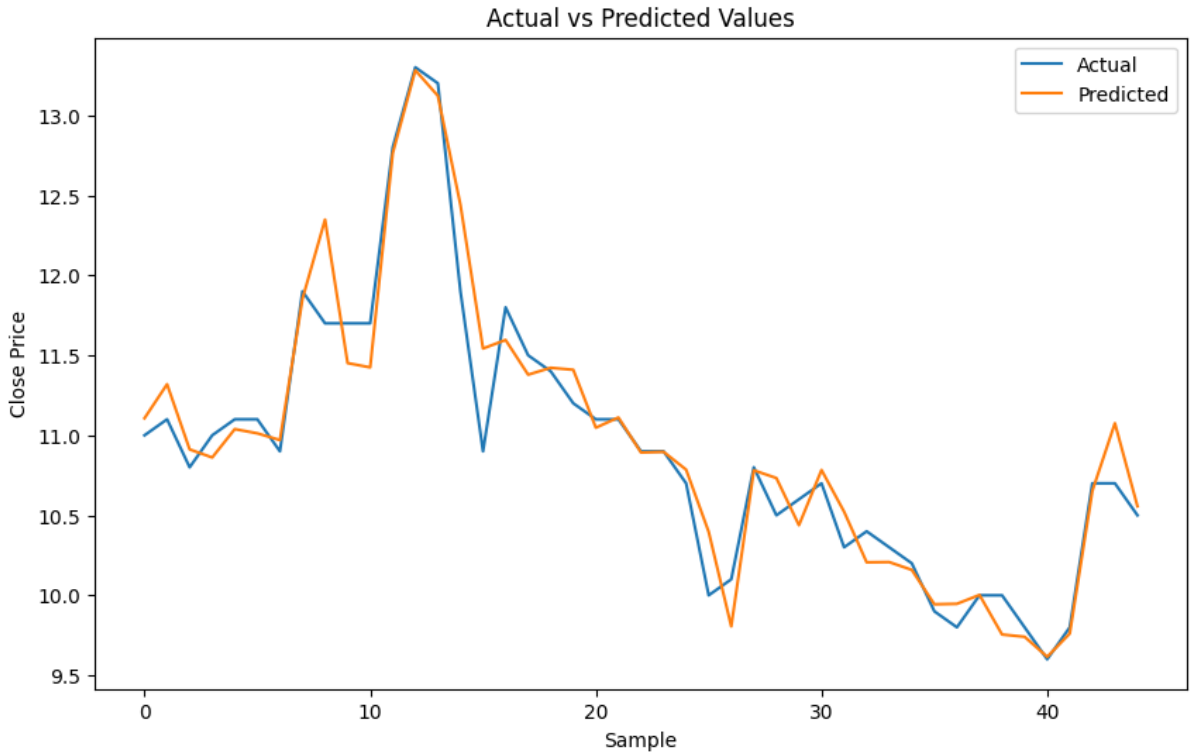



*Figure 4.17: So sánh test và predict RF*

```
MAE (selected features): 0.15146101851096888
RMSE (selected features): 0.21882477482416782
R^2 (selected features): 0.9320940157110126
```

*Figure 4.18: Kết quả so sánh model trên tập test RF*

Nhóm sử dụng Learning Curve Score để đánh giá xem model có bị overfitting hay không. Hình dưới đây cho thấy đường cong huấn luyện và đường cong kiểm tra đều tăng dần và hội tụ về một giá trị cao. Sự tương đồng giữa hai đường cong, cùng với sự tăng dần của độ chính xác trên cả hai tập huấn luyện và kiểm tra, cho thấy mô hình có

khả năng học tốt từ dữ liệu huấn luyện và tổng quát hóa tốt trên dữ liệu mới.

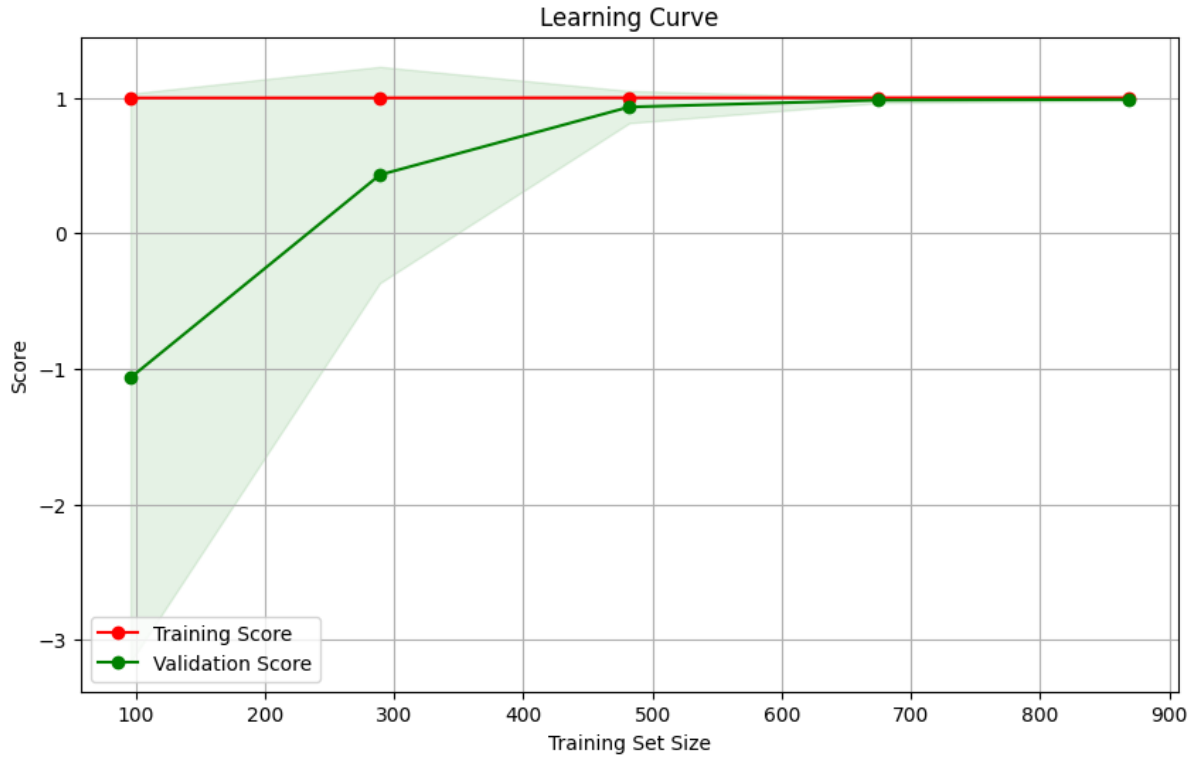



*Figure 4.19 : Learning Curve RF*

*Ngoài ra nhóm cũng đã tìm các siêu tham số bằng phương pháp Grid-Search để phù hợp với dữ liệu. Kết quả cho thấy tốt hơn so với tham số mặc định của mô hình*

```
MAE (selected features): 0.14873333333333205
RMSE (selected features): 0.21464549585045706
R^2 (selected features): 0.9346630845833863
```

*Figure 4.20 : Kết quả sau khi chạy mô hình tham số tối ưu RF*

Other Stocks

PLX

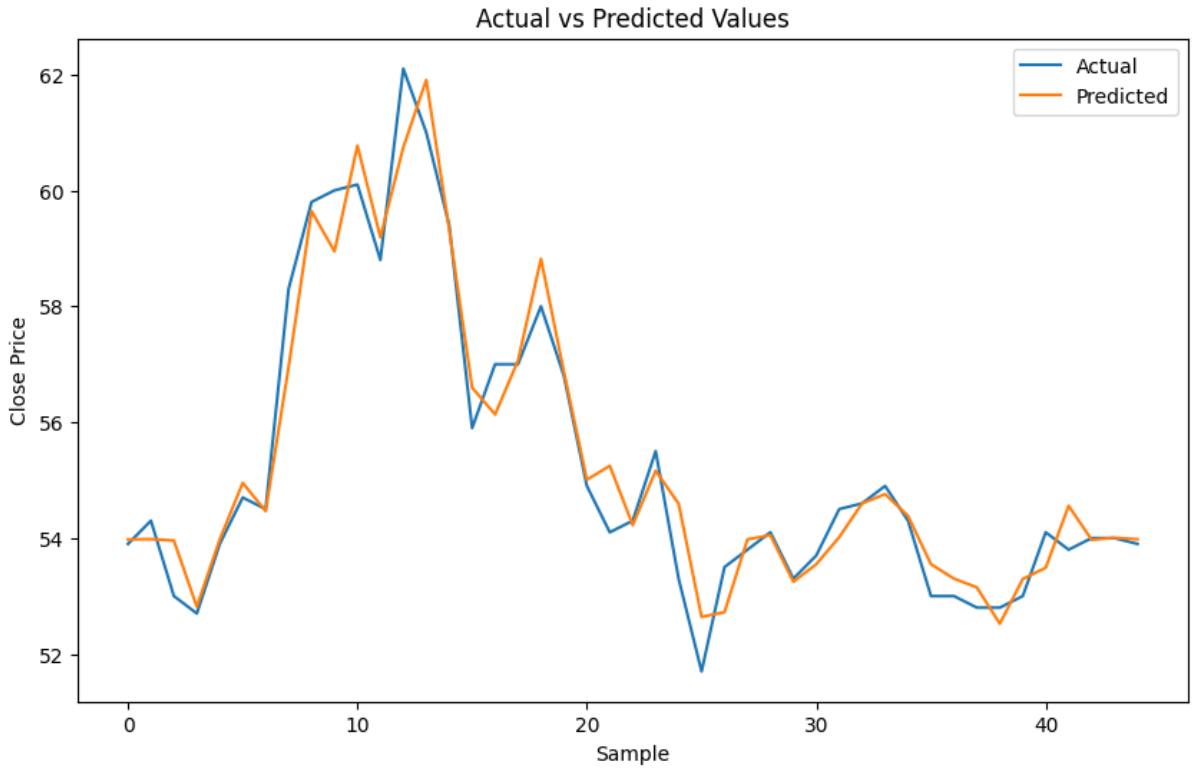



*Figure 4.21: So sánh test và predict PLX*

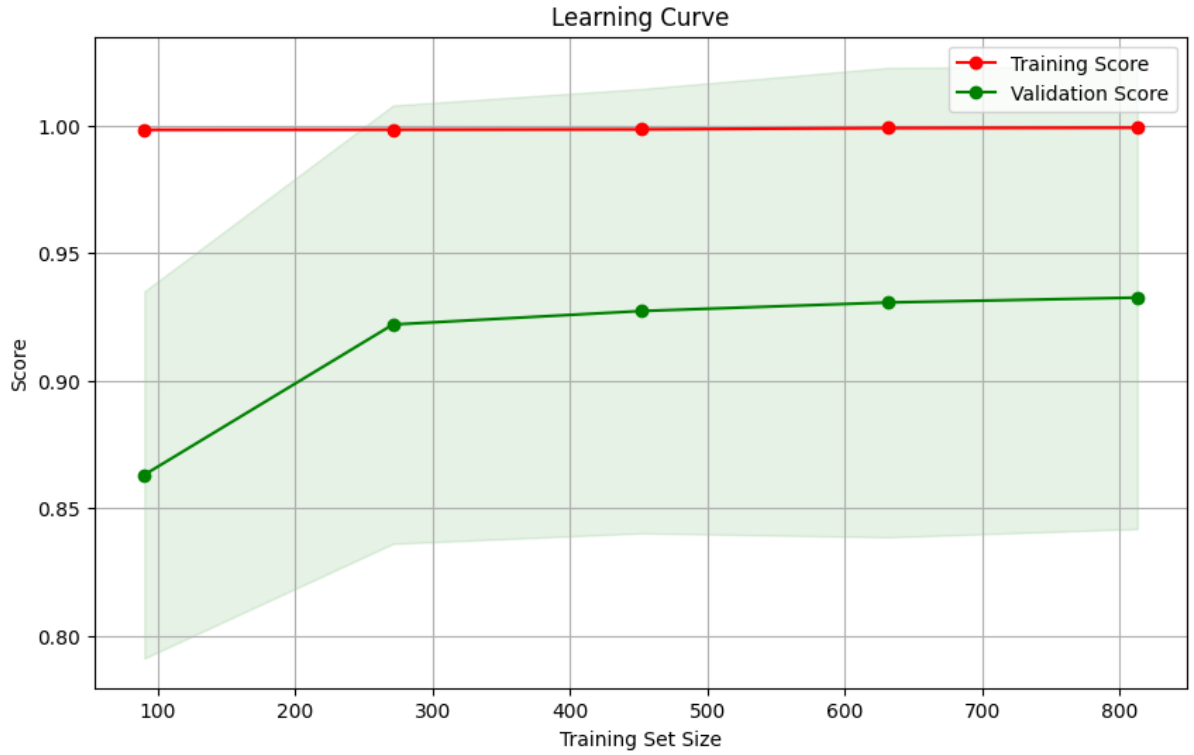


*Figure 4.22 : Learning Curve PLX*

```
MAE (selected features): 0.438355185185193 06
RMSE (selected features): 0.6058027939619939
R^2 (selected features): 0.9427505403202144
```

*Figure 4.23 : Kết quả sau khi chạy mô hình tham số tối ưu PLX*

*PVB*

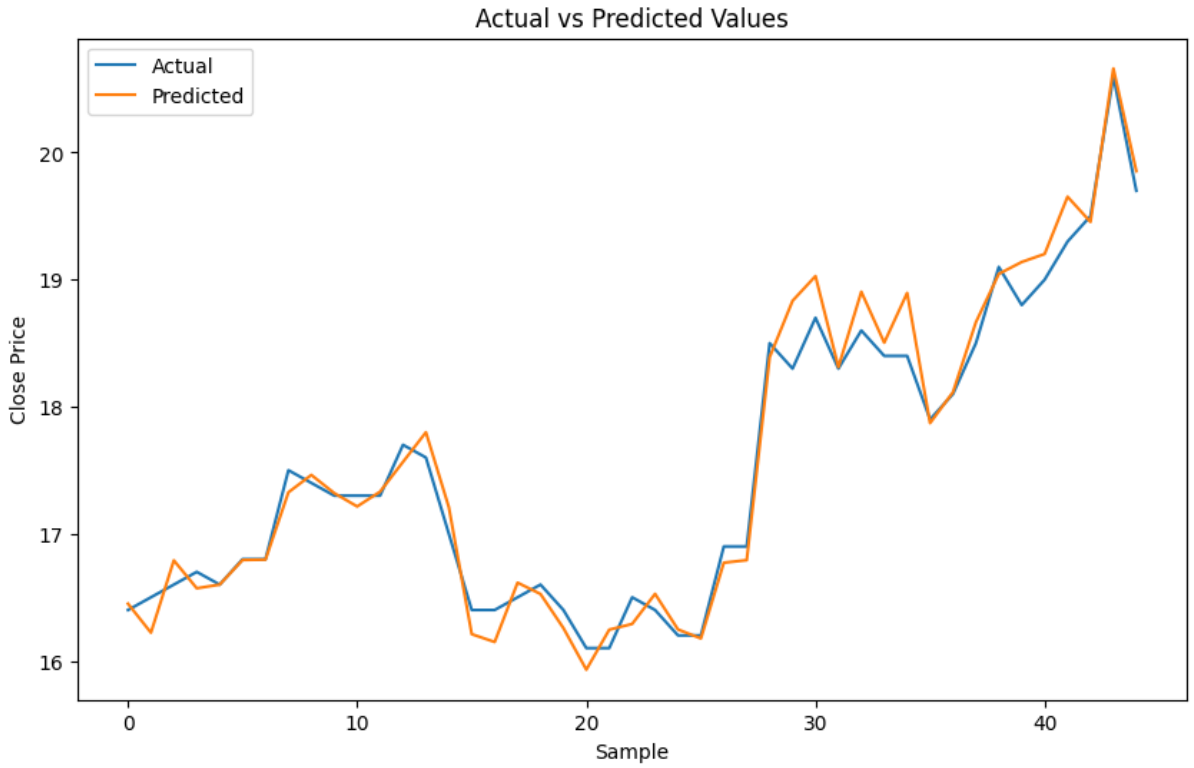


*Figure 4.24: So sánh test và predict PVB*

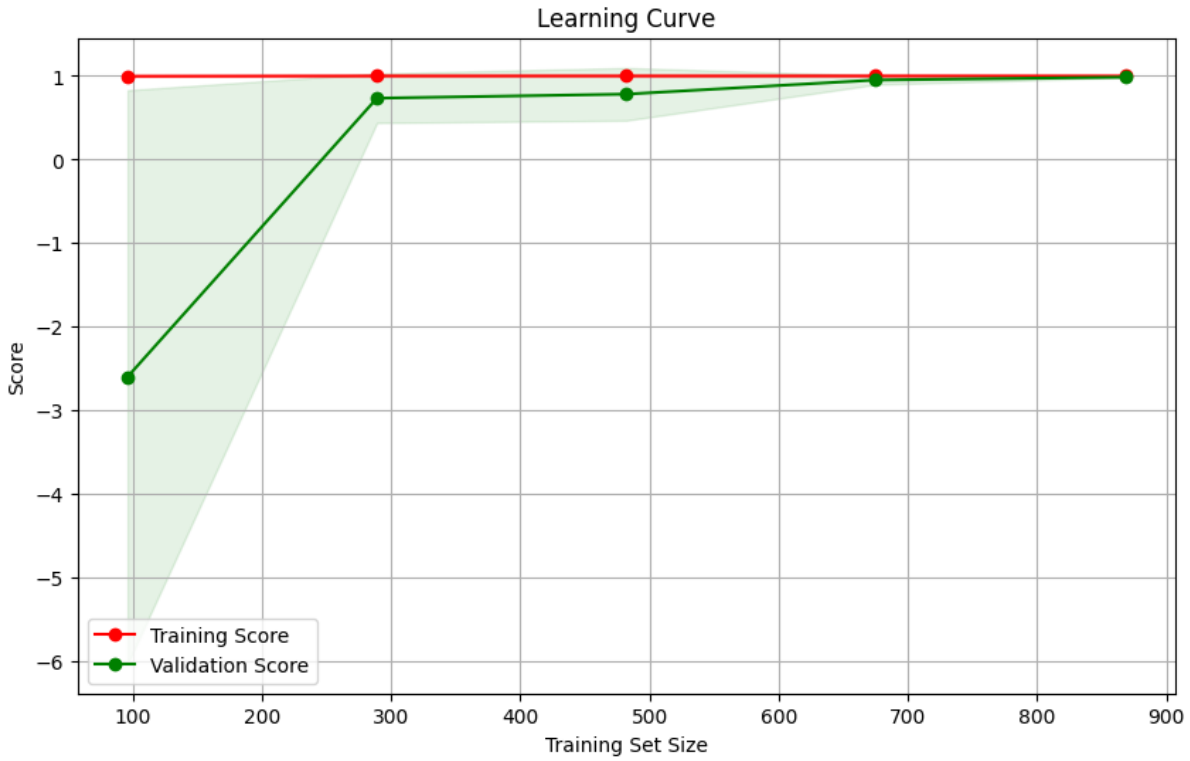



*Figure 4.25 : Learning Curve PVB*

```
MAE (selected features): 0.12669190870551736
RMSE (selected features): 0.160309698551981 85
R^2 (selected features): 0.9799824295758454
```

*Figure 4.26 : Kết quả sau khi chạy mô hình tham số tối ưu PVB*

*PVO*

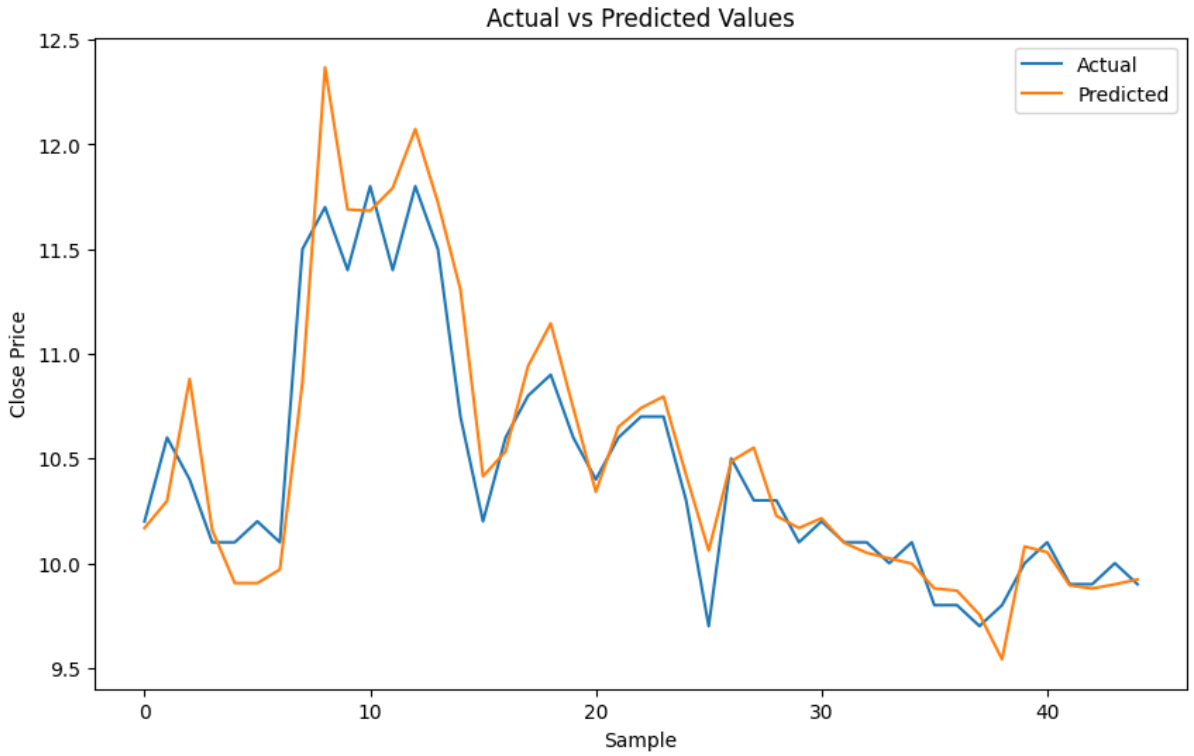



*Figure 4.27: So sánh test và predict PVO*

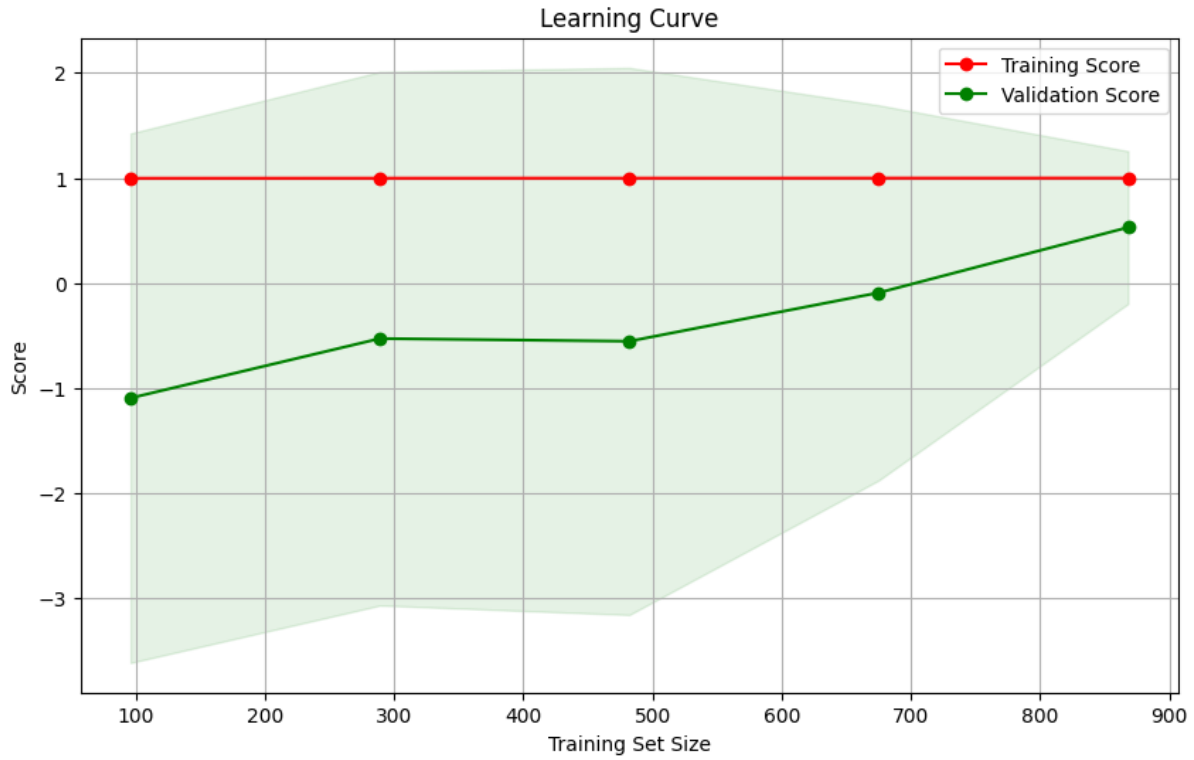


*Figure 4.28 : Learning Curve PVO*

```
MAE (selected features): 0.14673333333333274
RMSE (selected features): 0.196687851966284444
R^2 (selected features): 0.8841202073841787
```

*Figure 4.29 : Kết quả sau khi chạy mô hình tham số tối ưu PVO*

*PVC*

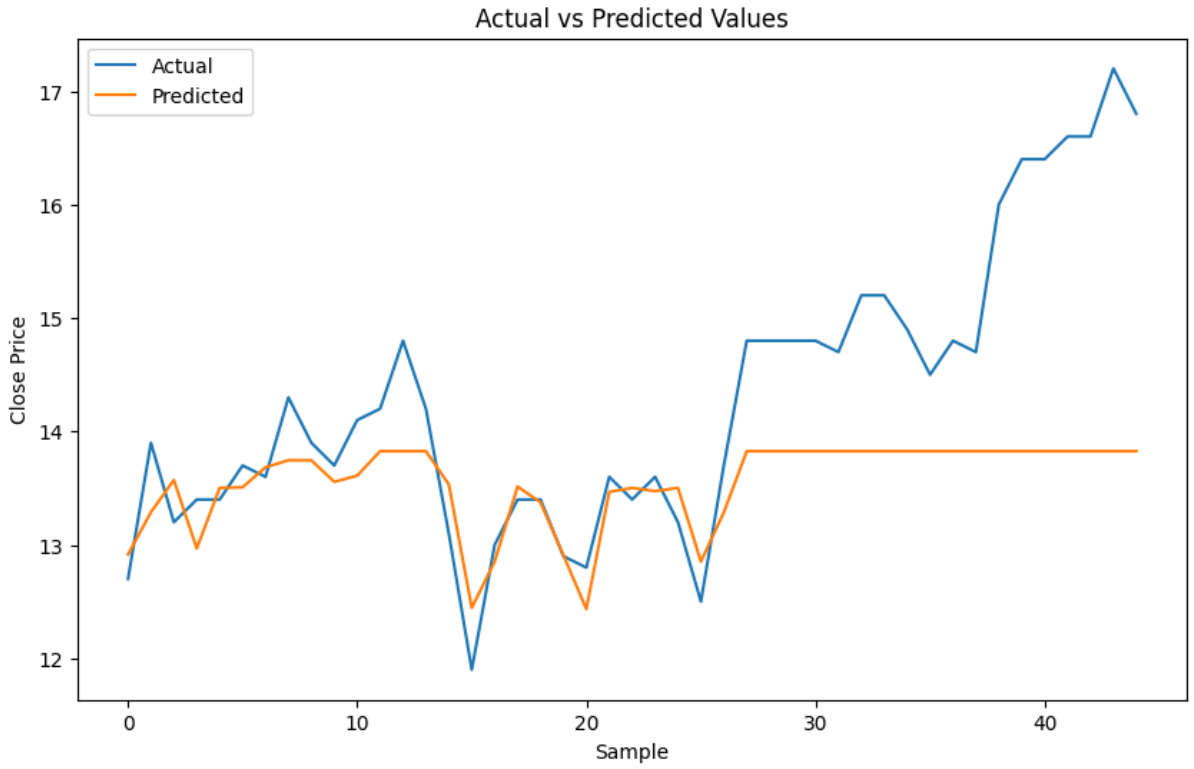



*Figure 4.30: So sánh test và predict PVC*

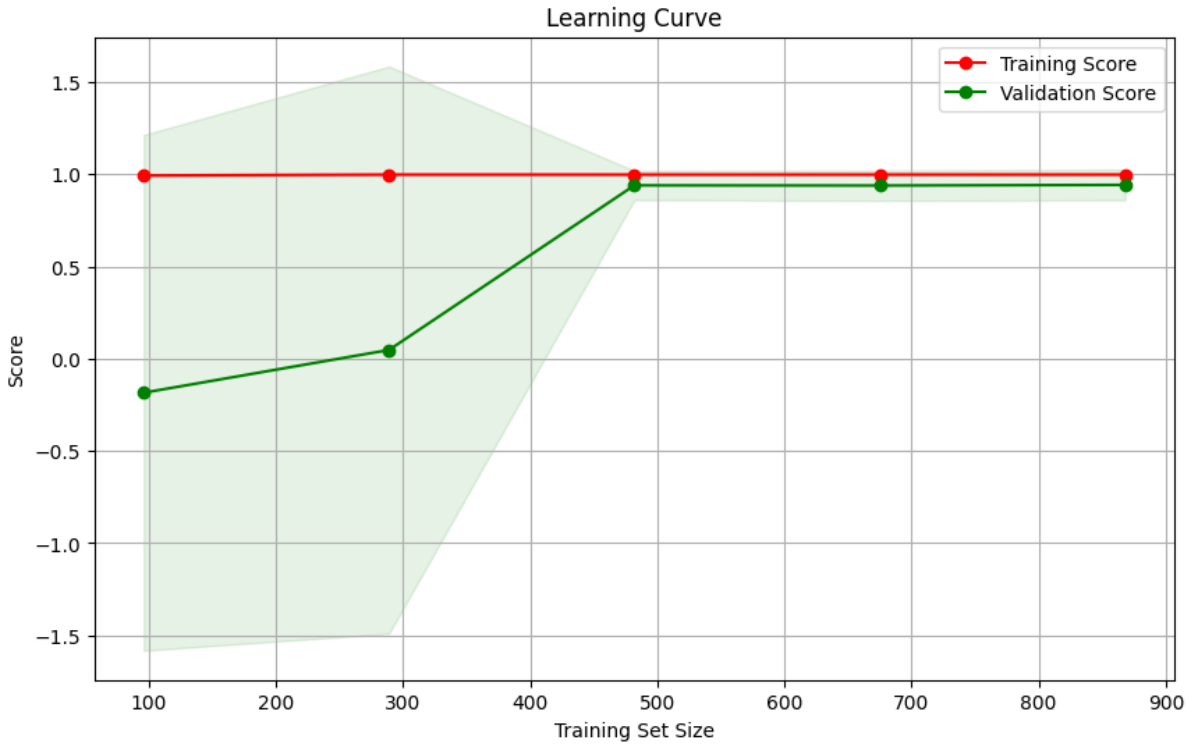



*Figure 4.31: Learning Curve PVC*

```
MAE (selected features): 0.8591407407407565
RMSE (selected features): 1.2406419127030972
R^2 (selected features): -0.0026715006305455535
```

*Figure 4.32 : Kết quả sau khi chạy mô hình tham số tối ưu PVC*

## 4.2. Model LightGBM

### 4.2.1. LightGBM cho đơn biến

Tại mô hình LightGBM đơn biến, tương tự như Random Forest đơn biến, với mã PCG, đây là kết quả sau khi chạy mô hình:

```
RMSE: 3.4220864795331254
MAE: 2.528177557297475
R-squared: -0.8267991843338485
```

*Figure 4.33 : Kết quả sau khi chạy mô hình  (LightGBM đơn biến)*

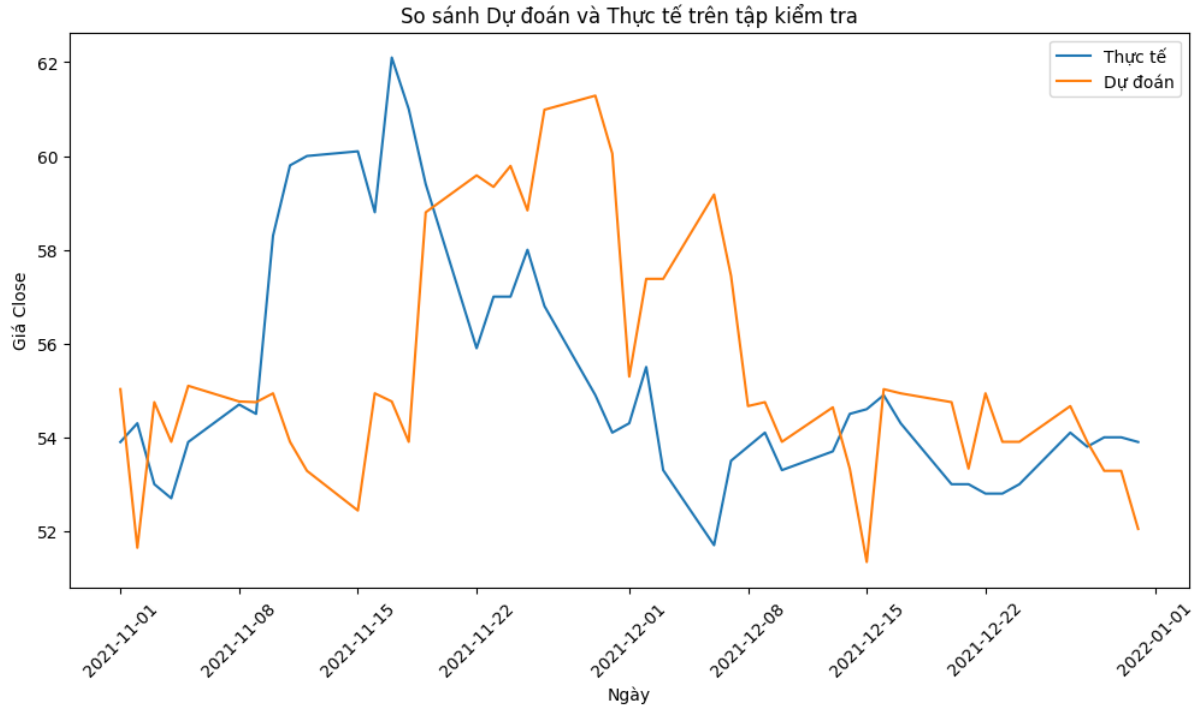



*Figure 4.34: So sánh test và predict (LightGBM đơn biến)*

Ta có thể thấy được các đường dự đoán được thể hiện xu hướng khá là tốt so với thực tế

Tương tự với các mã khác, ta có kết quả sau:

PLX

```
RMSE: 3.4220864795331254
MAE: 2.528177557297475
R-squared: -0.8267991843338485
```

*Figure 4.35 : Kết quả sau khi chạy mô hình PLX*

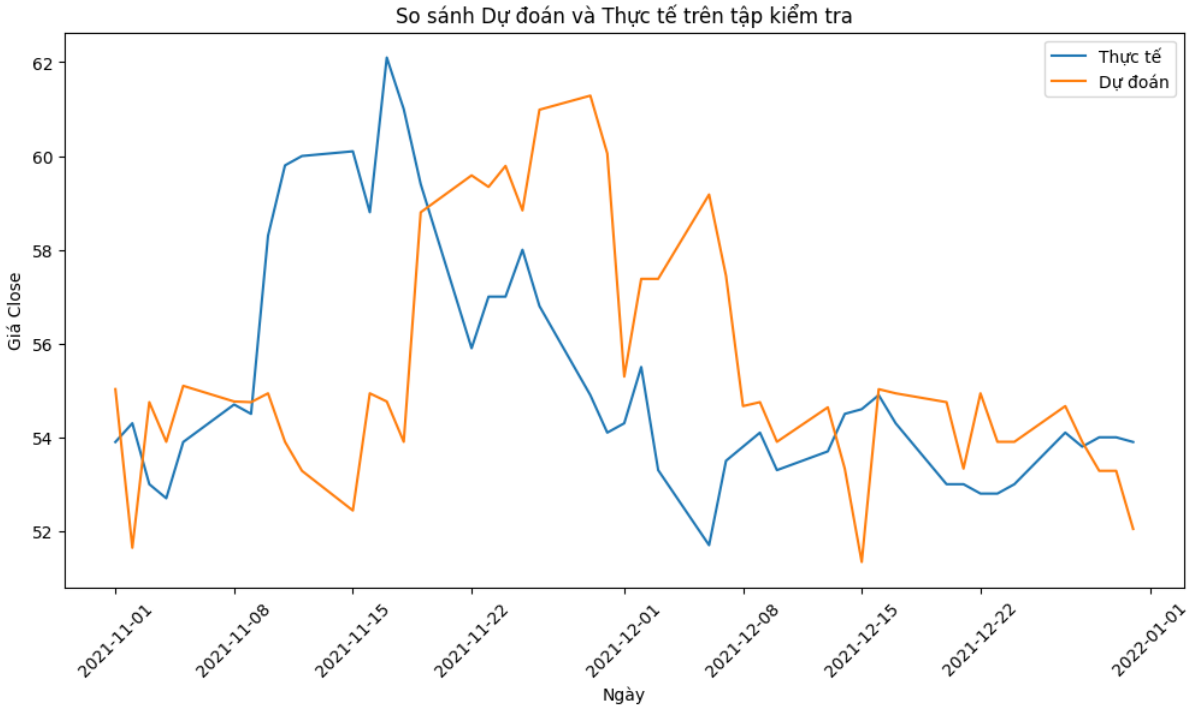



*Figure 4.36: So sánh test và predict PLX*

PVB

```
RMSE: 0.9990298540726475
MAE: 0.8312595066484203
R-squared: 0.22259254127687145
```

*Figure 4.37 : Kết quả sau khi chạy mô hình PVB*

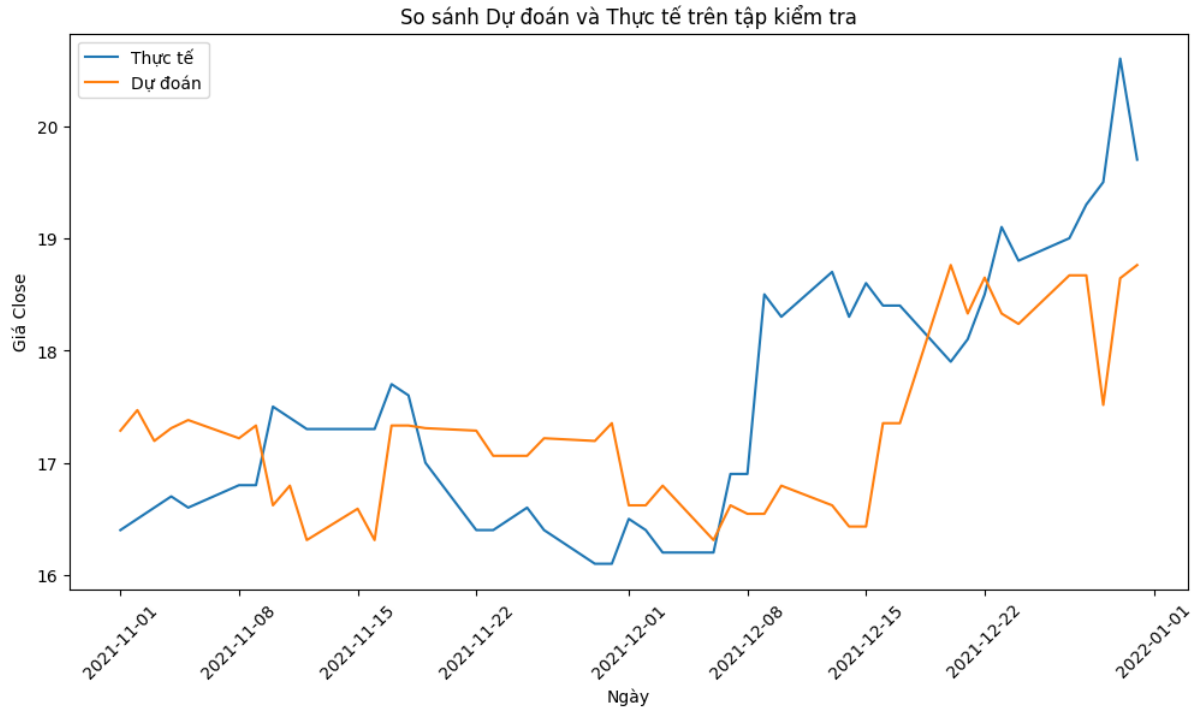



*Figure 4.38: So sánh test và predict PVB*

### 4.2.2. LightGBM cho đa biến

Tại mô hình LightGBM đa biến, nhóm sử dụng SHAP để tìm ra mức độ quan trọng của các feature. ( Phương pháp SHAP yêu cầu tính toán giá trị Shapley cho từng đặc trưng bằng cách xem xét tất cả các tập con của các đặc trưng trong mô hình. Khi áp dụng Shap trực tiếp cho Random Forests, việc tính toán các giá trị Shapley cho từng cây quyết định trong rừng và kết hợp chúng lại để có được một giá trị Shapley tổng quát cho mô hình là rất phức tạp và tốn kém về mặt tính toán vậy nên Shap sẽ chỉ được áp dụng cho mô hình LighGBM ). Nhóm đã tính trung bình giá trị tuyệt đối của các giá trị SHAP cho mỗi feature. Điều này giúp đánh giá mức độ quan trọng trung bình của các feature. Từ đó lấy các feature đó để training model

Sau khi tìm ra những đặc trưng quan trọng và đem vào mô hình, kết quả đem lại như hình dưới đây:

```
Best R^2 score (selected features): 0.9429295971002212
MAE (selected features): 0.15703503024236293
RMSE (selected features): 0.20060785543585502
R^2 (selected features): 0.9429295971002212
```

*Figure 4.39: Kết quả sau khi chạy model với tham số mặc định (LightGBM đa biến)*

Ta có thể thấy kết quả mô hình được coi như khá ổn định. So sánh giữa giá trị thực tế và giá trị dự đoán, nhóm đã làm rõ hơn bằng chart dưới đây. Ngoài ra để kiểm tra xem có bị overfitting, nhóm cũng vẽ thêm đồ thị Learning curve

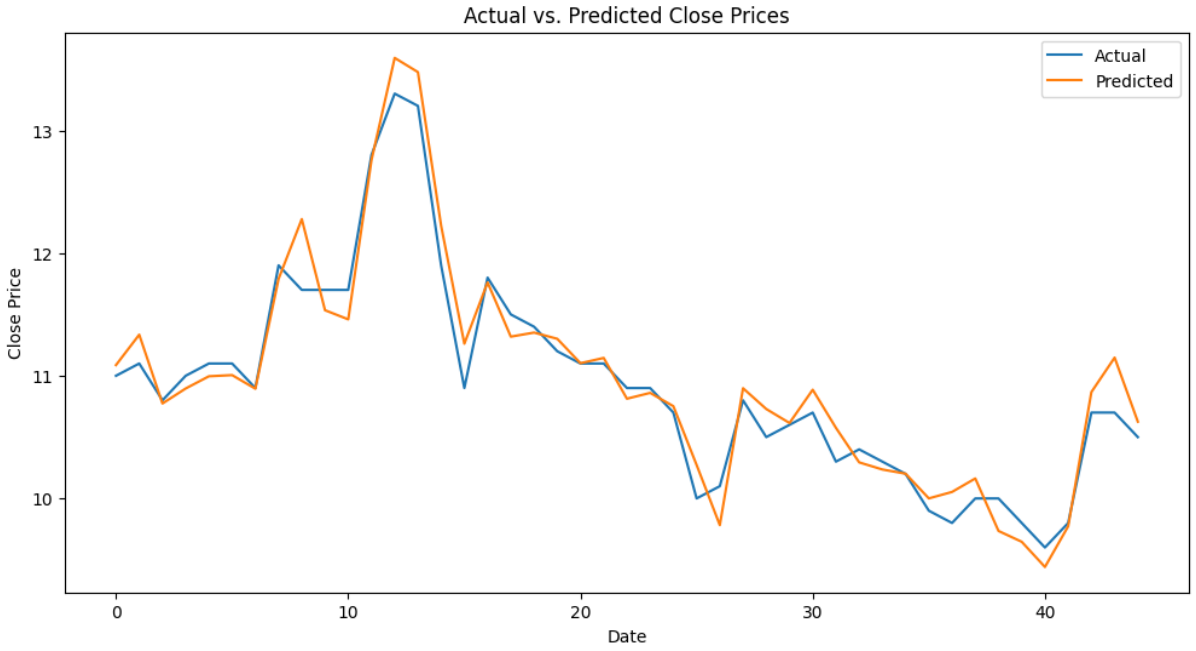



*Figure 4.40: Kết quả dự đoán giá đóng của trên tập test với tham số mặc định (LightGBM đa biến)*

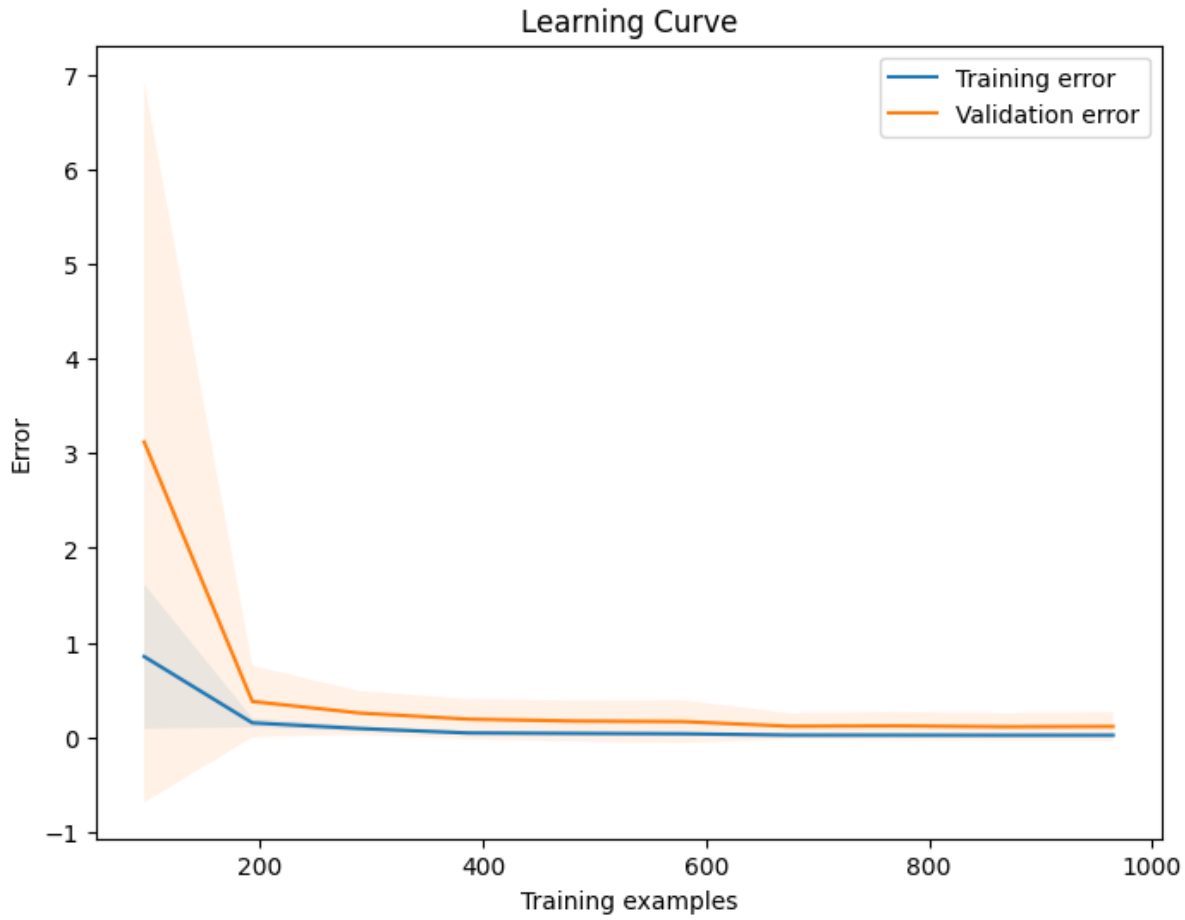



*Figure 4.41: Learning Curve  (LightGBM đa biến)*

Ta có thể thấy đồ thị hiển thị xu hướng cũng như bám khá sát với những dữ liệu thực tế. Đồ thị learning curve sẽ hiển thị sai số huấn luyện và sai số xác thực trên trục y và kích thước tập huấn luyện trên trục x. Ta có thể xem xét liệu sự sai lệch giữa hai đường cong này có nghiêng biểu thị overfitting hay không. Với kết quả, cả hai đường cong tiến gần đến nhau và hội tụ ở một giá trị thấp, điều đó cho thấy mô hình của chúng ta không bị overfitting và có khả năng tổng quát hóa tốt trên dữ liệu mới.

Để model có thể tốt hơn, nhóm sẽ tinh chỉnh các siêu tham số bằng phương pháp grid search.

```
Best parameters: {'colsample_bytree': 0.6, 'learning_rate': 0.05, 'n_estimators': 150, 'num_leaves': 30, 'subsample': 0.6}
MAE: 0.1494254654031749
RMSE: 0.1995990821295241
R^2: 0.9435021205277737
```

*Figure 4.42: Kết quả sau khi chạy model với tham số đã được tối ưu (LightGBM đa biến)*


*Figure 4.43: Kết quả sau khi chạy model với tham số đã được tối ưu (LightGBM đa biến)*

Kết quả cho thấy sau khi tuning, kết quả mô hình tốt hơn khi các giá trị R^2 cao hơn và RMSE thấp hơn so với sử dụng mô hình LightGBM với tham số mặc định

Other Stocks

Chúng tôi thực hiện tương tự các bước chuẩn bị và dự đoán cho các mã stock còn lại: PLX


*Figure 4.44: Kết quả dự đoán giá đóng của trên tập test với tham số đã tối ưu  PLX*

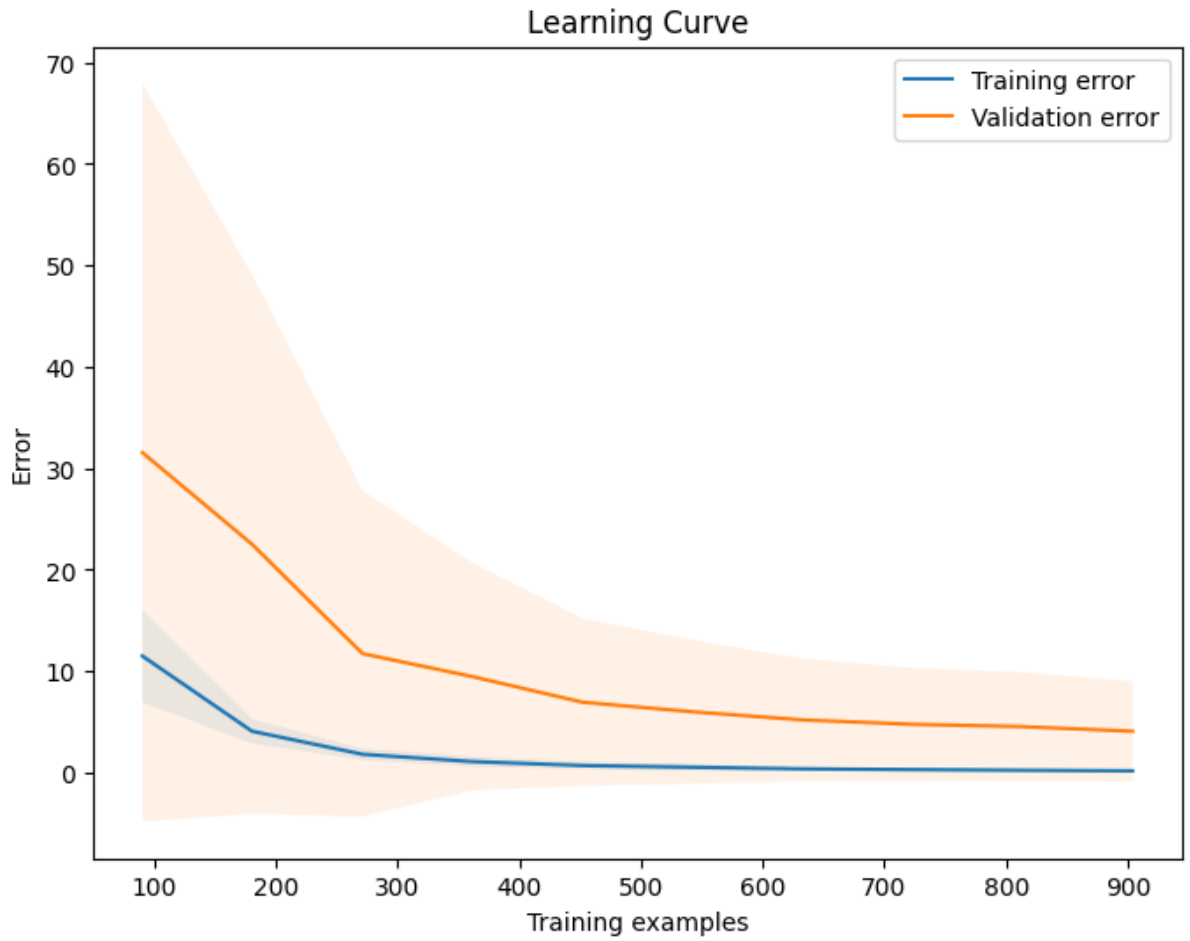



*Figure 4.45: Learning Curve PLX*

```
MAE: 0.24691353195063498
RMSE: 0.32343740253662223
R^2: 0.9836811583058888
```

*Figure 4.46: Kết quả sau khi chạy model với tham số đã được tối ưu PLX*

PVB


*Figure 4.47: Kết quả dự đoán giá đóng của trên tập test với tham số đã tối ưu  PVB*

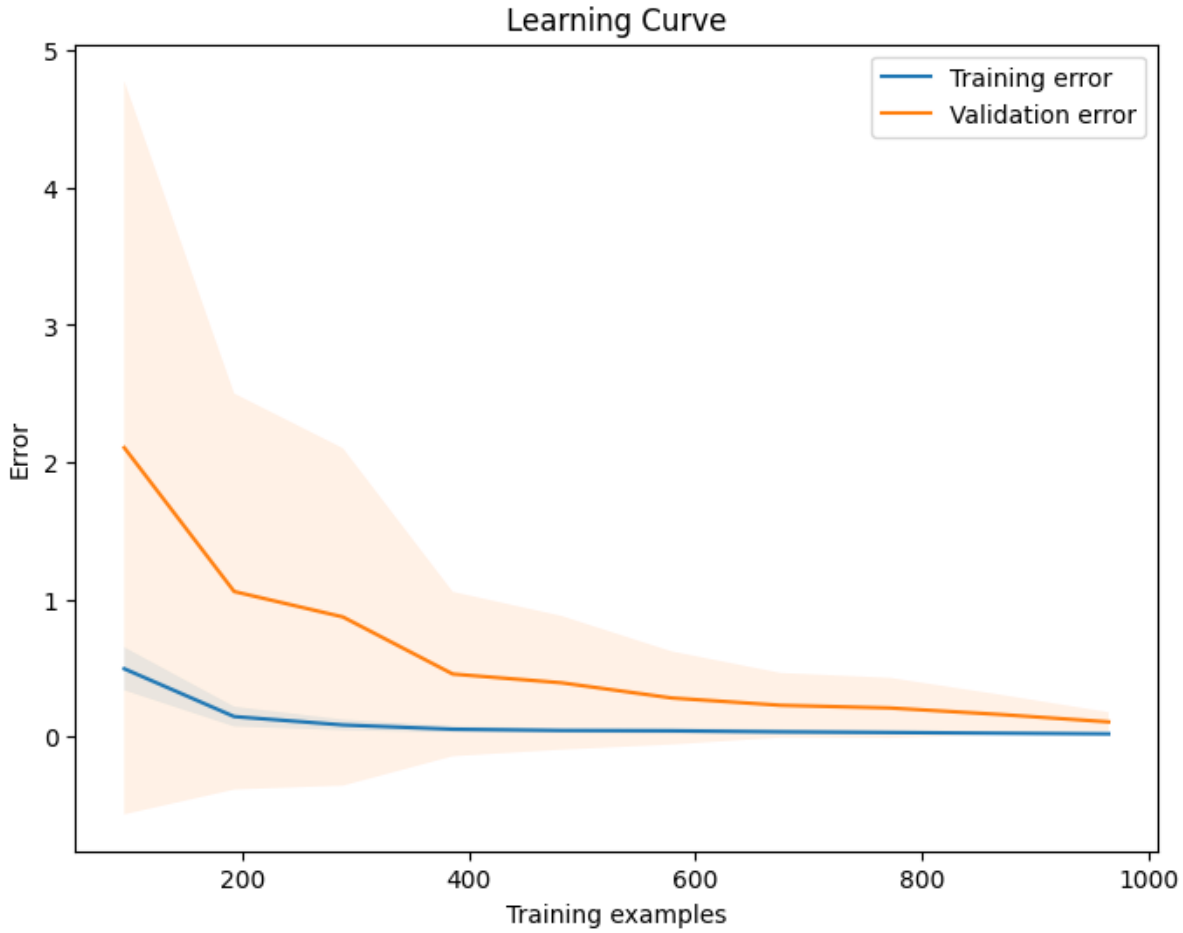


*Figure 4.48: Learning Curve PVB*

```
MAE: 0.0960146030867856
RMSE: 0.1236318925128145
R^2: 0.9880943583970221
```

*Figure 4.49: Kết quả sau khi chạy model với tham số đã được tối ưu PVB*

PVO

*Figure 4.50: Kết quả dự đoán giá đóng của trên tập test với tham số đã tối ưu PVO*

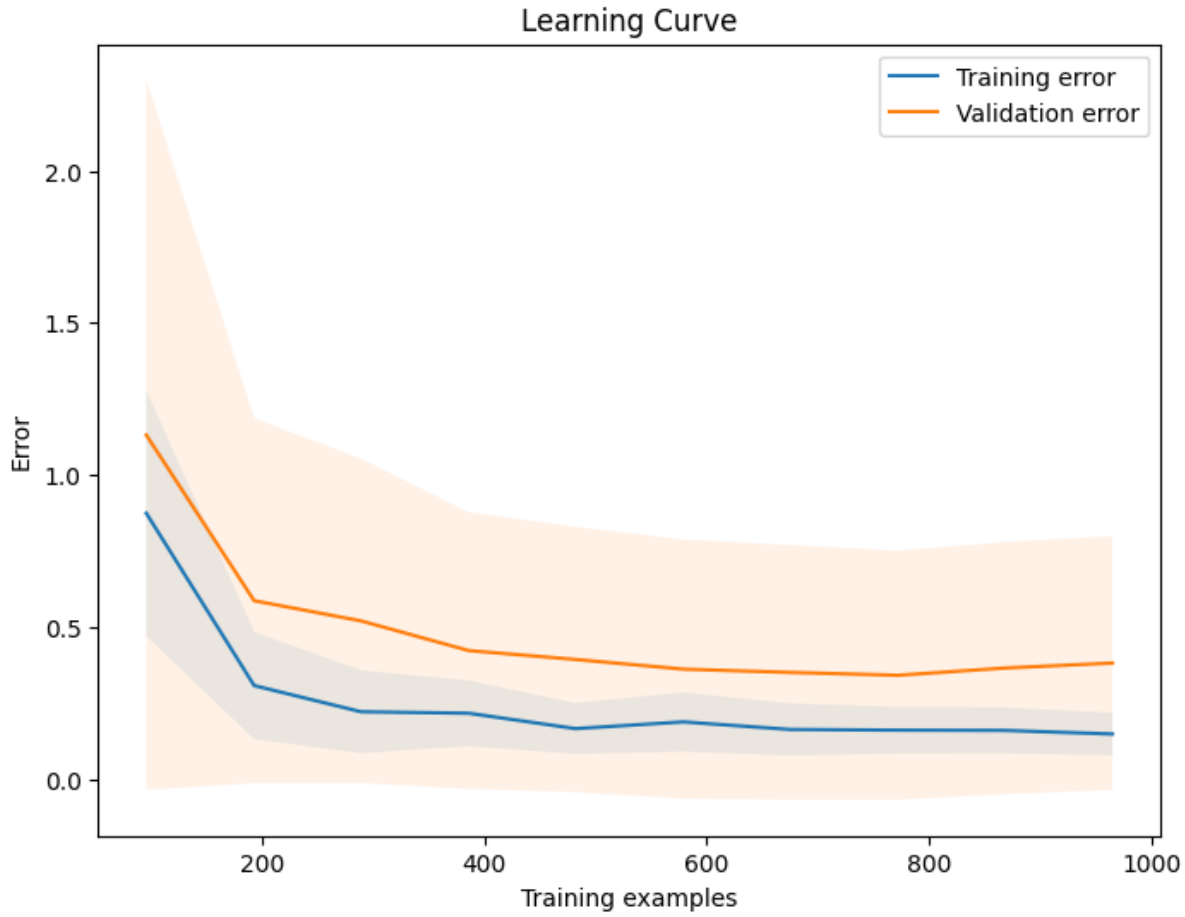



*Figure 4.51: Learning Curve PVO*

```
MAE: 1.5318587661941858
RMSE: 1.6350012287456 24
R^2: -7.00734980393082
```

*Figure 4.52: Kết quả sau khi chạy model với tham số đã được tối ưu PVO*

PVC

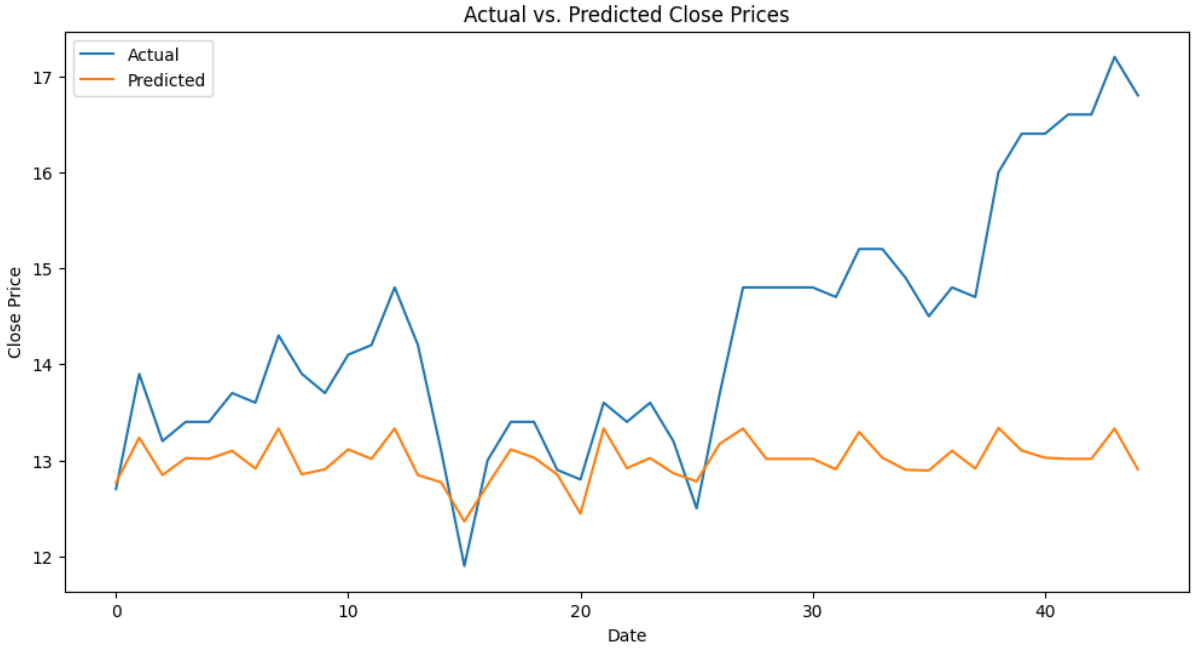



*Figure 4.53: Kết quả dự đoán giá đóng của trên tập test với tham số đã tối ưu PVC*

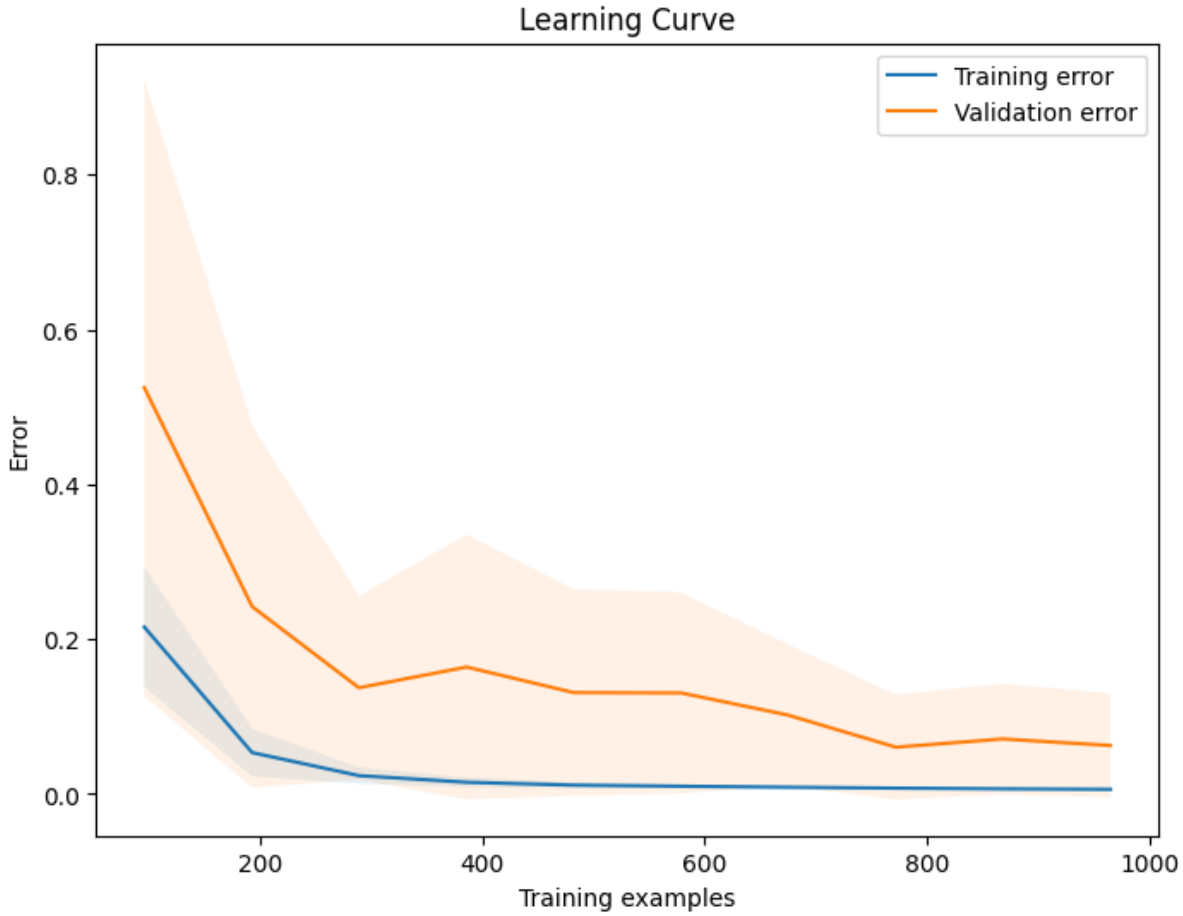


*Figure 4.54: Learning Curve PCV*

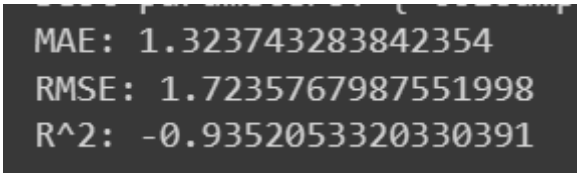



*Figure 4.55: Kết quả sau khi chạy model với tham số đã được tối ưu PVC*

## 4.3. Comparison

Nhóm lấy đại diện một mã là PCG để so sánh các model này với nhau. Với từng mã khác nhau sẽ có model phù hợp khác nhau. Dựa vào kết quả mô hình chạy được và xem xét các kết quả đầu ra như điểm RMSE, MAE, R2 và các chart như learning curve để xem xét model có kết quả bao nhiêu và xem xét có dấu hiệu overfitting hay không.

| Code | Model | Dấu hiệu overfitting | RMSE | MAE | R2 |
|---|---|---|---|---|---|
| **PCG** | Random Forest cho đơn biến | Không | 0.39 | 0.28 | 0.6 |
| **PCG** | Random Forest cho đa biến | Không | 0.21 | 0.14 | 0.93 |
| **PCG** | LightGBM cho đơn biến | Không | 03.42 | 2.52 | -0.82 |

| | | | | | |
|---|---|---|---|---|---|
| **PCG** | LightGBM cho đa biến | Không | 0.2 | 0.14 | 0.94 |

Kết luận lại dựa vào các chỉ số cũng như hình, với mã PCG phù hợp với LightGBM đa biến ( đã điều chỉnh siêu tham số cho phù hợp ). Với những mã kế tiếp, có thể mô hình LightGBM hoặc Random Forest sẽ phù hợp hơn.

# Chapter 5: Conclusion

Machine learning có thể được ứng dụng trong dự báo chứng khoán để phân tích các xu hướng, dự đoán giá cổ phiếu và đưa ra quyết định đầu tư. Nghiên cứu của nhóm đã hoàn thành việc cung cấp cho các bên liên quan thông tin chính xác và đáng tin cậy về hiệu suất trong tương lai của các cổ phiếu này, có thể giúp hướng dẫn các quyết định đầu tư và thúc đẩy sự ổn định và tăng trưởng kinh tế. Việc ra quyết định dựa trên những dữ liệu trước đây cùng với các mô hình máy học sẽ giúp nhà đầu tư cũng như các bên liên quan sẽ xem xét và cân nhắc đầu tư cho phù hợp.

# Reference

**[1]** Bülent, B. Mert, E., Ersin, K. M. Fatih, A & Sevtap, E. (2020). *MACHINE LEARNING BASED DEMAND FORECAST MODELS FOR E-COMMERCE INDUSTRY*. 5TH INTERNATIONAL CONFERENCE ON LIFE AND ENGINEERING SCIENCES, ALANYA, TURKEY ICOLES 2022.

**[2]** Mehar, V. Deeksha, C. Vinay, A. T. & Arun Kumar. (2020). *Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange* (Vol. Volume 167, 2020, Pages 599-606). Procedia Computer Science.

**[3]** Na, Z. Xiaoli, R. Honglin, H. Li, Z. Dan, Z. Rong, G. Pan, L. & Zhongen, N. (2019). *Estimating grassland aboveground biomass on the Tibetan Plateau using a random forest algorithm* (Vol. Volume 102, July 2019, Pages 479-487). Ecological Indicators.

**[4]** Peng, L. Hengwen, G. Lili, Y & Benling, L. (2022). *Research on trend prediction of component stock in fuzzy time series based on deep forest.* CAAI Transactions on Intelligence Technology.

**[5]** Priyam, S. Nitesh, G. Rony, M. Dyutimoy, D. Sushmita, N. Manoj, K & Tiwari. (2022). *Demand Forecasting of a Multinational Retail Company using Deep Learning Frameworks.* (Vol. 395–399). . IFAC PapersOnLine 55-10 (2022).

**[6]** Subba Rao Polamuri, K. Srinivas & A. Krishna Mohan. (2019). *Stock Market Prices Prediction using Random Forest and Extra Tree Regression.* (Vol. 1224 - 1228). International Journal of Recent Technology and Engineering 8(3).

**[7]** Yanjun, C. Kun, L. Yuantao, X. & Mingyu, H. (2020). *Financial Trading Strategy System Based on Machine Learning* (Vol. Volume 2020 | Article ID 3589198). Research Article | Open Access.

**[8]** Yi, T & Yanyan, C. (2021). *Forecasting Model of High Transfer Stock —Based on Integrated Learning* (Vol. pages 234-238). IEEE Xplore.

**[9]** Yuanyuan, Q. Zhongkai, Z. & Zhiliang, Q. (2020). *Wavelet-Aided Stock Forecasting Model based on Ensembled Machine Learning* . (Vol. 3 pages.). In 2020 The3rd International Conference on Machine Learning and Machine Intelligence(MLMI ’20), September 18–20, 2020, Hangzhou, China. ACM, New York, NY,USA,.

**[10]** Zheng, T. Ziqin, Y & Guangwei, Z. (2019). *Stock selection with random forest: An exploitation of excess return in the Chinese stock market* (Vol. Volume 5, Issue 8, August 2019, e02310). Heliyon.