

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

****

**BÀI TẬP THỰC HÀNH**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU KINH DOANH**

**LAB 04**

*Giảng viên hướng dẫn:*

***PGS.TS Nguyễn Đình Thuân***

*Sinh viên thực hiện:*

|  |  |
| --- | --- |
| ***Nguyễn Hoàng Long*** | ***- 19521788*** |
| ***Phan Phạm Quỳnh Hoa***  ***Huỳnh Minh Thư*** | ***- 19521520***  ***- 19522304*** |

*TP HCM, Ngày 20 tháng 04 năm 2022*

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM, đặc biệt là quý thầy cô khoa Hệ thống thông tin của trường đã giúp cho chúng em trang bị các kiến thức cơ bản, các kỹ năng thực tế và tạo điều kiện để chúng em có thể hoàn thành đồ án môn học của mình.

Đặc biệt chúng em xin chân thành cảm ơn thầy Nguyễn Đình Thuân đã nhiệt tình hướng dẫn hướng dẫn, quan tâm truyền đạt những kiến thức và kinh nghiệm, trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báu cho chúng em trong suốt thời gian học tập để chúng em có thể hoàn thành tốt môn học này.

Trong thời gian thực hiện, nhóm tác giả đã vận dụng những kiến thức nền tảng đã tích lũy. Từ đó, nhóm tác giả vận dụng tối đa những gì đã thu thập được để hoàn thành bài tập thực hành một cách tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía Thầy/Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn!

***Nhóm thực hiện***

**NHẬN XÉT CỦA GIẢNG VIÊN**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**MỤC LỤC**

[**I.** **Thực hiện bài tập** 5](#_Toc104112832)

[**4a. Các phương pháp đánh giá độ chính xác mô hình** 5](#_Toc104112833)

[**1.** **Mô hình hồi quy** 5](#_Toc104112834)

[**2.** **Mô hình phân loại** 10](#_Toc104112835)

[**4b. Hồi quy phi tuyến đa biến với dữ liệu data housing của Việt Nam.** 13](#_Toc104112836)

[**Phát biểu bài toán** 13](#_Toc104112837)

[**1.** **Sử dụng Excel** 13](#_Toc104112838)

[**2.** **Sử dụng R** 19](#_Toc104112839)

[**3.** **Sử dụng Python** 20](#_Toc104112840)

[**4c. Hồi quy Logistic với dữ liệu Bank Customer Data in VietNam.** 21](#_Toc104112841)

[**Phát biểu bài toán** 21](#_Toc104112842)

[**4.** **Sử dụng Excel** 21](#_Toc104112843)

[**5.** **Sử dụng R** 28](#_Toc104112844)

[**6.** **Sử dụng Python** 30](#_Toc104112845)

[**7.** **Dự báo ARIMA** 30](#_Toc104112846)

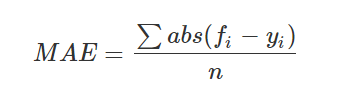
[**II.** **Bảng phân công công việc** 34](#_Toc104112847)

[**III.** **Tài liệu tham khảo** 35](#_Toc104112848)

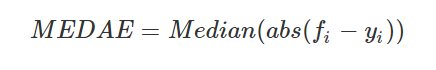
1. **Thực hiện bài tập**

## **4a. Các phương pháp đánh giá độ chính xác mô hình**

1. **Mô hình hồi quy**
   1. **Nhóm 1: Các tiêu chí dựa vào các giá trị sai biệt tuyệt đối**
      1. **MAE: trung bình của sai biệt tuyệt đối**



* + 1. **MEDAE: trung vị của sai biệt tuyệt đối**

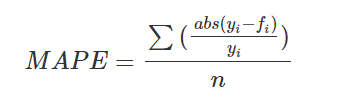
****

* + 1. **SAE: Tổng sai biệt tuyệt đối**

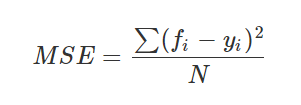
**A picture containing text, watch, clock

Description automatically generated**

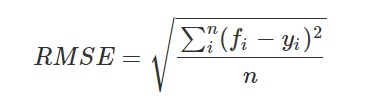
* + 1. **MAPE Mean absolute percentage error**

****

* 3 tiêu chí trên đo lường sai biệt giữa thực tế (yi) và tiên lượng của mô hình (fi), giá trị tuyệt đối được dùng để tránh sai lầm trong trường hợp mô hình đồng thời có nguy cơ đánh giá quá cao và quá thấp, dẫn đến việc sai số > 0 và <0 triệt tiêu lẫn nhau.
* Tiêu chí thứ 4: MAPE đo lường sai biệt theo tỉ lệ % , dùng cho những trường hợp mà biến kết quả có đơn vị quá thấp hoặc quá cao
  1. **Nhóm 2: Các tiêu chí dựa vào bình phương sai số**
     1. **MSE: trung bình bình phương sai số**



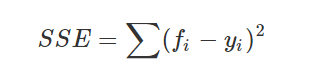
* + 1. **RMSE: căn bậc 2 của trung bình bình phương sai số**



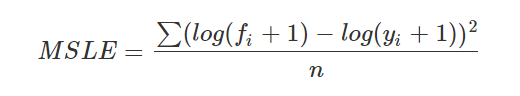
* + 1. **Trung vị bình phương sai số**

****

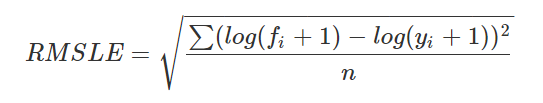
* + 1. **Tổng bình phương sai số**

****

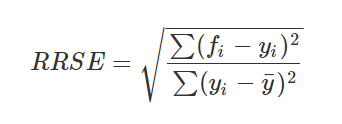
* + 1. **MSLE Mean squared logarithmic error**

****

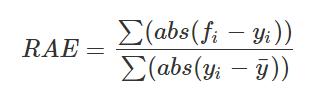
* + 1. **RMSLE: Root mean squared logarithmic errors**

****

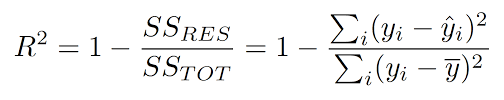
* 1. **Nhóm 3: Khảo sát sai biệt tương đối của mô hình**
     1. **RRSE Root relative squared error**

****

* + 1. **RAE : Relative absolute error**

****

* 2 tiêu chí này không đo lường sai biệt của mô hình một cách độc lập và tuyệt đối, nhưng đánh giá tương đối tỉ lệ giữa sai biệt của mô hình và sai biệt nội tại của biến kết quả bên trong quần thể
  1. **Nhóm 4: Dựa vào hệ số xác định**
     1. **R-square**

****

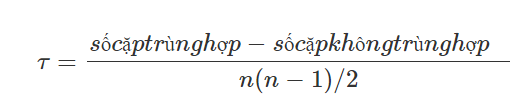
* SSR: Tổng bình phương sai số
* SST: Tổng bình phương tất cả sai lệch giữa yi và giá trị trung bình

Bản thân hệ số này không cung cấp thông tin về độ chính xác của mô hình

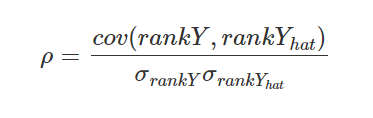
nhưng cho biết mô hình giải thích được bao nhiêu phần phương sai của biến kết quả trong mẫu.

càng cao (càng gần 1) thì mô hình càng giải thích được biến động của biến phụ thuộc (VD: = 0.85 cho thấy biến độc lập giải thích được 85% sự thay đổi của biến phụ thuộc, còn 15% còn lại là do các yếu tố ngẫu nhiên gây ra

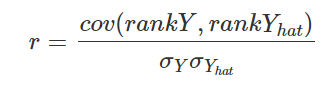
* 1. **Nhóm 5 : Các tiêu chí đánh giá mối tương quan tuyến tính giữa giá trị thực và tiên lượng:**
     1. **Kendall’s Tau (thứ hạng)**

****

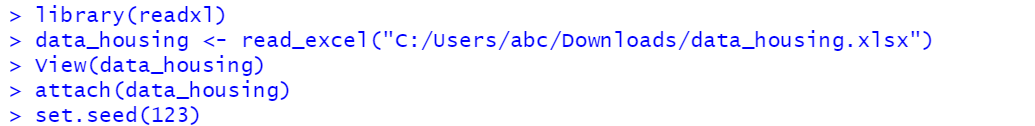
* + 1. **Rho của Spearman (phi tham số)**

****

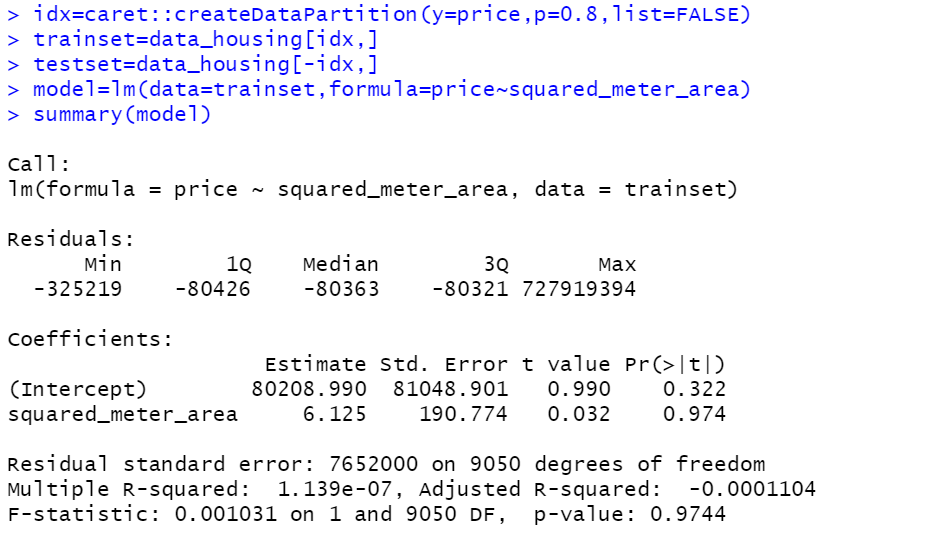
* + 1. **Rho của Spearman (phi tham số)**



* 1. **Thực hiện trên R**
* Cài đặt Package caret
* Import dữ liệu



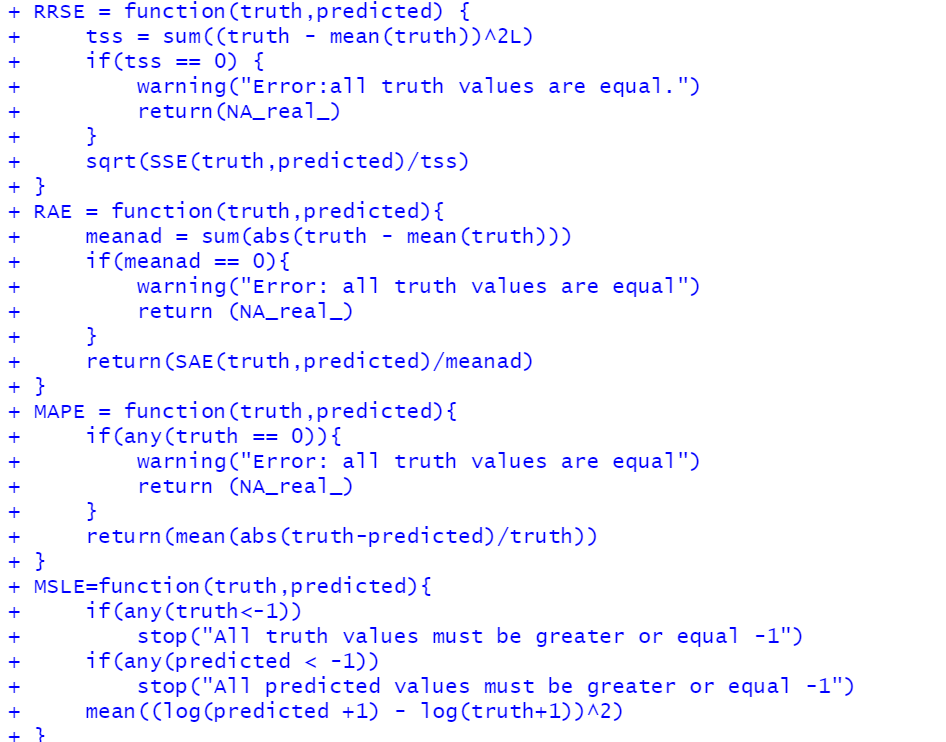
* Chia dữ liệu thành 2 tập test và train
* Dựng model

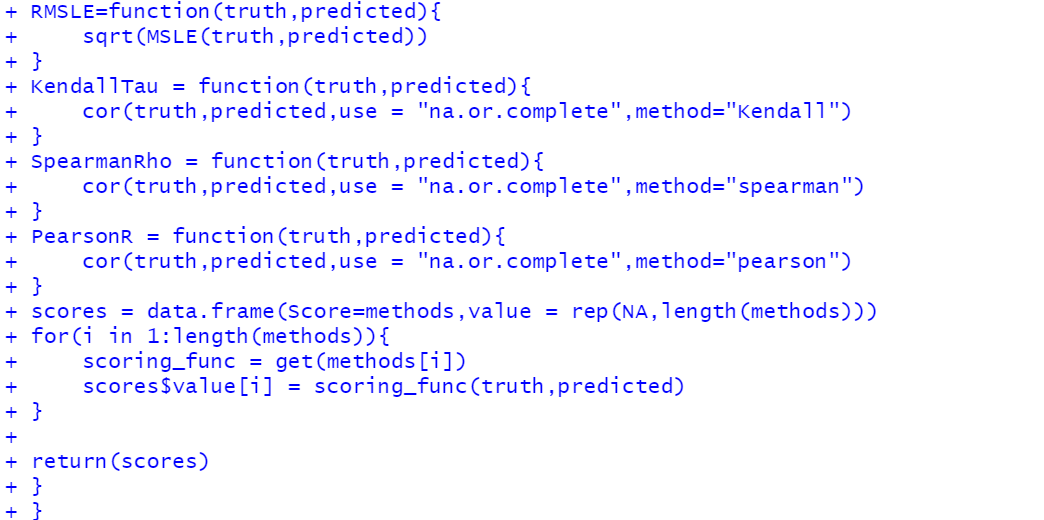


* Tính các tiêu chí đánh giá model

Text

Description automatically generated





1. **Mô hình phân loại**
   1. **Accuracy**

Accuracy là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu, nhằm tìm ra tỷ lệ dự đoán đúng của mô hình

Ảnh có chứa văn bản

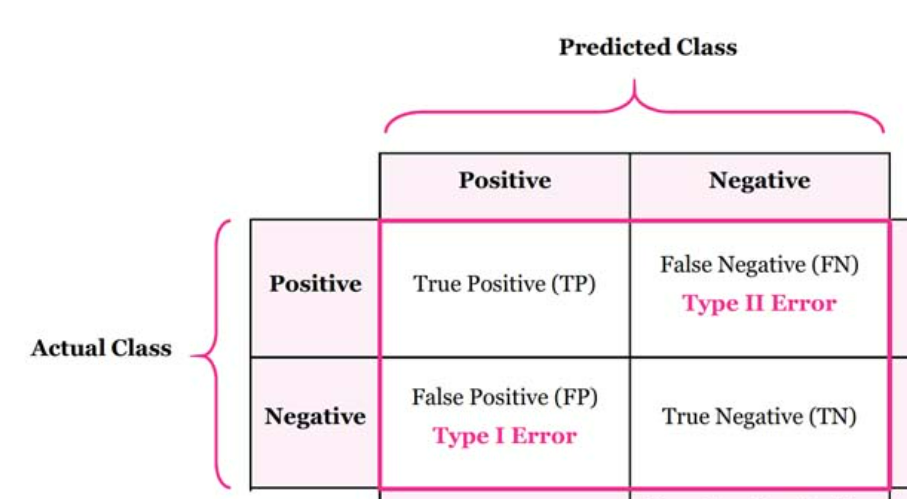
Mô tả được tạo tự động

* 1. **Confusion matrix**

Cách tính sử dụng accuracy chỉ cho chúng ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác

Confusion matrix là một kỹ thuật đánh giá hiệu năng của mô hình cho các bài toán phân lớp. Confusion matrix là một ma trận thể hiện số lượng điểm dữ liệu thuộc vào một class và được dự đoán thuộc vào class

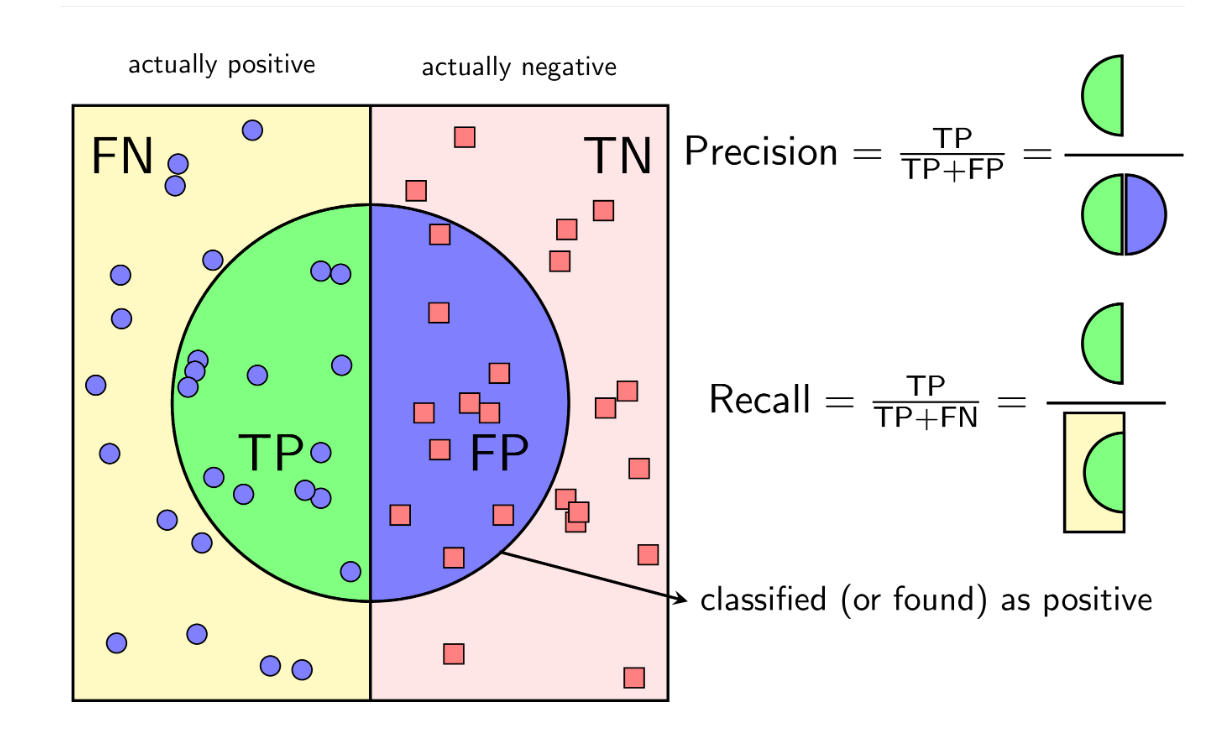
Confusion matrix cung cấp thêm thông tin về tỉ lệ phân lớp đúng giữa các lớp, hay giúp phát hiện các lớp có tỉ lệ phân lớp nhầm cao nhờ vào các khái niệm True (False) Positive (Negative)



* True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)
* True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)
* False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error
* False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error
  1. **Precision và Recall**

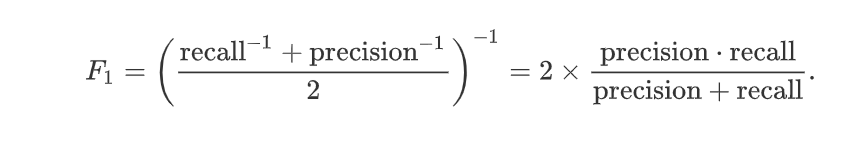
Precision: Trong tất cả các dự đoán Positive được đưa ra, bao nhiêu dự đoán là chính xác?

Recall: Trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác?



* 1. **F1-Score**

Cả precision và recall đều quan trọng trong từng trường hợp khác nhau, F1-Score ra đời như 1 thông số duy nhất để điều chỉnh hợp lí precision và recall.



Một trong 2 chỉ số recall hoặc precision thấp đều sẽ kéo điểm F1-score xuống. Trường hợp xấu nhất khi một trong hai chỉ số Precison và Recall bằng 0 sẽ kéo điểm F1-score về 0. Trường hợp tốt nhất khi cả điểm chỉ số đều đạt giá trị bằng 1, khi đó điểm F1-score sẽ là 1

## **4b. Hồi quy phi tuyến đa biến với dữ liệu data housing của Việt Nam.**

### **Phát biểu bài toán**

* Nguồn dữ liệu: https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi
* Ý nghĩa nguồn dữ liệu: Một số thông tin về giá nhà
* Phát biểu bài toán: Với độ tin cậy 95% có thể tìm ra mối liên hệ giữa price với num\_bed\_rooms, Squared\_meter\_area hay không
* Đặt giả thuyết
  + H0: Phương trình hồi qui không phù hợp.
  + H1: Phương trình hồi qui phù hợp
* Confident level = 95% tức α = 0.05

1. **Sử dụng Excel**

Tập dữ liệu data\_housing

**Table

Description automatically generated**

* Giả sử phương trình hồi phi tuyến tính đa biến có dạng:

price = b0 + b1 \*log(num\_bed\_rooms) + b2\* log(squared\_meter\_area)

Hay y = b0 + b1\*log(x1) + b2\*log(x2)

* Với các ma trận:

* Sử dụng hàm LN(number) để tính giá trị log(num\_bed\_rooms) và log(squared\_meter\_area)

Graphical user interface, table

Description automatically generated

Ta có các ma trận tương ứng với tập dữ liệu

* Ma trận F bao gồm 3 cột, 11313 hàng

Table

Description automatically generated

Ta có hệ phương trình trước được viết lại:

M=FT.F

Ma trận hệ số A=(FT.F)-1 .FT . Y

Chuyển vị của ma trận F, FT bao gồm 11313 cột, 3 hàng

Graphical user interface, text, application, table

Description automatically generated

* Trong excel sử dụng hàm MMULT(array1, array2) để nhân hai ma trận

Nhân hai ma trận FT, F thu được kết quả ma trận gồm 3 hàng, 3 cột Table

Description automatically generated

* Trong excel sử dụng hàm MINVERSE(array) để nghịch đảo một ma trận

Kết quả của phép tính ( FT.F)-1 là một ma trận gồm 3 hàng, 3 cột

Table

Description automatically generated

* Kết quả của phép tính ( FT.F)-1 FT là một ma trận gồm 11313 hàng, 3 cột

Table

Description automatically generated

* Kết quả của phép tính A=( FT.F)-1 FT.Y là một ma trận gồm 1 cột, 3 hàng

Table

Description automatically generated

* Vậy ta có được phương trình hồi quy phi tuyến:

**Y=** -300483 - 797629\***log(X1)** + 370679**\*log(X2)**

Hay

**Price=** - 300483 - 797629\***log(num\_bed\_rooms)** + 370679**\*log(squared\_meter\_area)**

* Phép kiểm định Fisher
* Giả thuyết:

H0: Phương trình hồi qui không phù hợp.

H1: Phương trình hồi qui phù hợp.

* Thực hiện tính toán trên Excel tương tự phần 4d.1.

Table

Description automatically generated

Table, Excel

Description automatically generated

* Ta thấy Significance F < α = 0.05
* Vậy ta bác bỏ H0.
* Phương trình hồi quy phù hợp.
* Sử dụng công cụ Data Analysis trên Excel kiểm tra kết quả
* Data -> Data Analyst -> chọn Regression

Graphical user interface, application

Description automatically generated

* Hộp thoại Regression xuất hiện

Graphical user interface, application

Description automatically generated

Input Y Range: chọn vùng dữ liệu cột price.

Input Y Range: chọn vùng dữ liệu cột log(num\_bed\_rooms), log(squared\_meter\_area).

Level: chọn độ tin cậy là 95%.

* Thu được kết quả như hình:

Application, table, Excel

Description automatically generated

* Đánh giá độ chính xác của phương trình

R square = 0.00163631 -> Có 0.163% dữ liệu phù hợp với mô hình

R square Adjusted = 0.00145976 R square < 50% nghiên cứu không được đánh giá tốt.

Vì giá trị Significance F < α = 0.05 Từ chối H0. Nên mô hình hồi qui phù hợp với tổng thể.

1. **Sử dụng R**

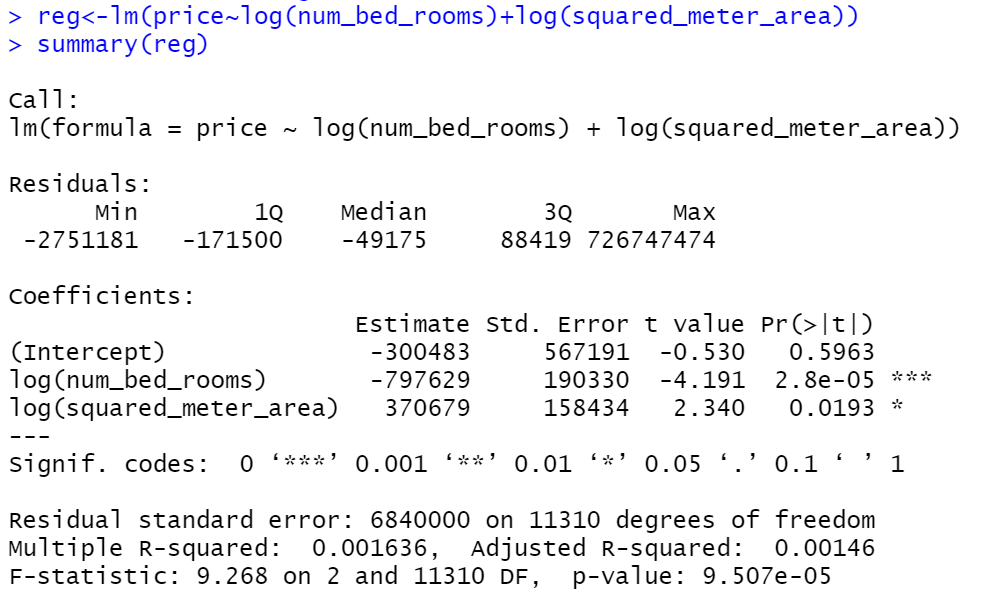
Dùng ngôn ngữ R thực hiện các phép tính Hồi quy với tập tin dữ liệu về Việt Nam.

* Import dataset

Table

Description automatically generated

* Sử dụng hàm **lm(y~log(x))** để biểu diễn mô hình hồi quy phi tuyến tính non-linear model.



* Nhìn vào Coefficients ta thấy Pr(>|t|) cột num\_bed\_rooms và cột squared\_meter\_area đều < 0.05.
* Nhận sét về độ tương quan R-squared, Adjusted R-squared ta thấy mô hình sau là phù hợp:

**Price** = -300483 – 797629 \* **num\_bed\_rooms** + 370679 \* **squared\_meter\_area**

* + **Kết luận:** Ta thấy p-value < α -> bác bỏ H0

1. **Sử dụng Python**

## **4c. Hồi quy Logistic với dữ liệu Bank Customer Data in VietNam.**

### **Phát biểu bài toán**

* Nguồn dữ liệu: https://www.kaggle.com/datasets/tomculihiddleston/bank-customer-data-in-vietnam
* Ý nghĩa nguồn dữ liệu: Thông tin của các khách hàng tại một ngân hàng Việt Nam
* Phát biểu bài toán: Kiểm tra mối liên hệ giữa output label **loan** với feature input **balance**

1. **Sử dụng Excel**

Tập dữ liệu Bank Customer Data in VietNam

Graphical user interface, application, table, Excel

Description automatically generated

Lấy cột dữ liệu *loan* tương ứng với biến Y, cột dữ liệu *balance* tương ứng với biến X.

Dạng tuyến tính của phương trình hồi quy logistic:

Y = b0 + b1X + e hay loan = b0 + b1\*balance + e

* Giả sử phương trình hồi quy logistic có dạng:

log() = b0 + b1X + ℇ

* Bước 1: tạo thêm cột biến chặn trước cột giá trị balance

Table

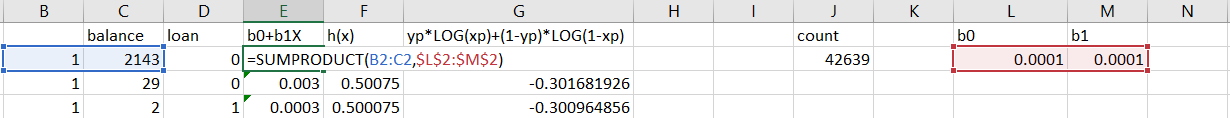
Description automatically generated with medium confidence

* Bước 2: gán cho b0, b1 hai giá trị bất kỳ, với bài này nhóm lấy giá trị 0.0001

Table

Description automatically generated

* Bước 3: Sử dụng hàm SUMPRODUCT(array1, array2) để tính giá trị cột b0+b1X



Ta được kết quả

Table

Description automatically generated with medium confidence

* Bước 4: áp dụng công thức

A picture containing text

Description automatically generated

Ta thu được kết quả cột h(x)

Text

Description automatically generated with low confidence

Table

Description automatically generated with medium confidence

* Bước 5: Tính log của hàm hợp lý

A picture containing diagram

Description automatically generated

Áp dụng công thức trên Excel

Timeline

Description automatically generated with low confidence

Ta được kết quả

A picture containing table

Description automatically generated

Sum: A picture containing text

Description automatically generated

* Bước 6: sử dụng công cụ Solver trên excel để tối ưu log(L) đạt giá trị lớn nhất

Chọn Data -> Solver -> xuất hiện hộp thoại Solver Parameters

Graphical user interface, application

Description automatically generated

Set Objecttive: giá trị cần tối ưu.

To: chọn Max, tối ưu giá trị đạt cực đại.

By Changing Variable Cells: chọn những giá trị cần thay đổi.

Bỏ chọn phần Make Unconstrained Variables Non-Negative

Graphical user interface, text, application

Description automatically generated

Nhấn chọn OK

* Ta thu được kết quả như hình

Table

Description automatically generated with medium confidence

* Từ kết quả trên ta có được phương trình hồi quy như sau:

Log() = -1.44827 - 0.001613 \* balance + ℇ

* Ta thu được kết quả sau:

() =

* Ta đặt hệ số p/(1-p) là Odd
* Đặt Odd0 và balance = 0 thì Odd0 =
* Đặt Odd1 và balance = 1 thì

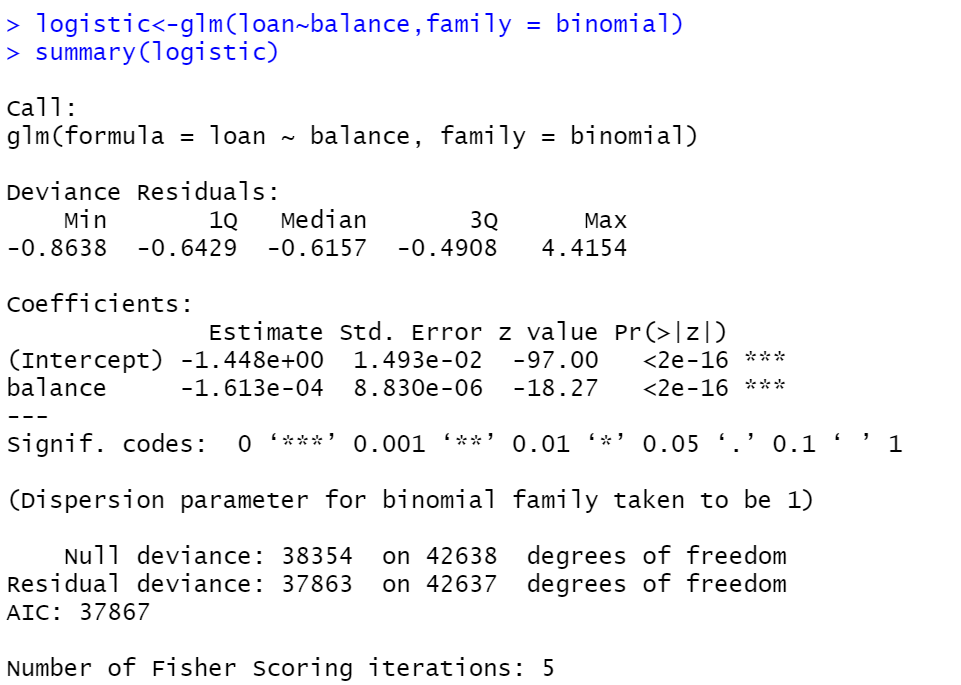
Odd1 =

* Tỉ số
* Như vậy giá trị *balance* lên 1 đơn vị thì khả năng có khoản nợ *loan* tăng lên 1.001614 lần.

1. **Sử dụng R**

Dùng ngôn ngữ R thực hiện Hồi quy Logistic với dữ liệu thực tế tùy chọn về/của Việt Nam.

* Sử dụng hàm **glm(y~x)** để phân tích hồi quy logistic với tham số **family = binomial**



* Từ kết quả mô hình ta có được phương trình hồi quy như sau:

Log() = -1.448 -0.001613 \* balance + ℇ

* Ta thu được kết quả sau:

() =

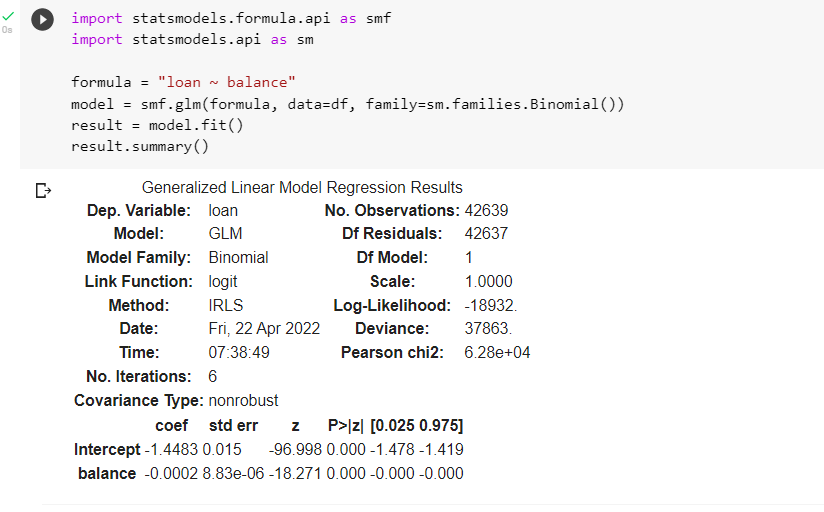
Đặt hệ số p/(1-p) là odd

* Đặt và balance = 0 thì =
* Đặt và balance = 1 thì =

Tỉ số = = 0.998

Lúc này ta có thể đánh giá, balance tăng 1 đơn vị thì khả năng nợ (loan) tăng 1 lần, balance tăng 0.1 đơn vị thì khả năng nợ tăng 0.1 lần,

1. **Sử dụng Python**



* Từ kết quả mô hình ta có được phương trình hồi quy như sau:

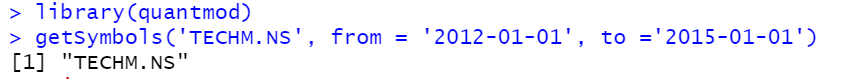
Log() = -1.448 -0.001613 \* balance + ℇ

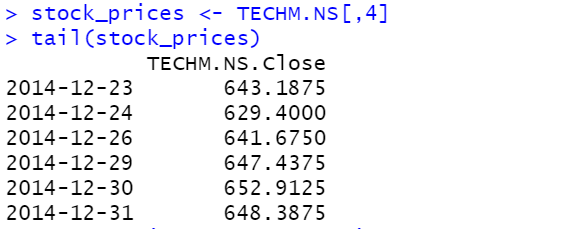
* Ta thu được kết quả sau:

() =

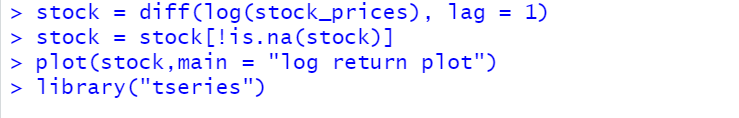
1. **Dự báo ARIMA**

* Sử dụng dữ liệu giá chứng khoán từ Yahoo finance trên R, thêm thư viện Library(quantmod)

****

****

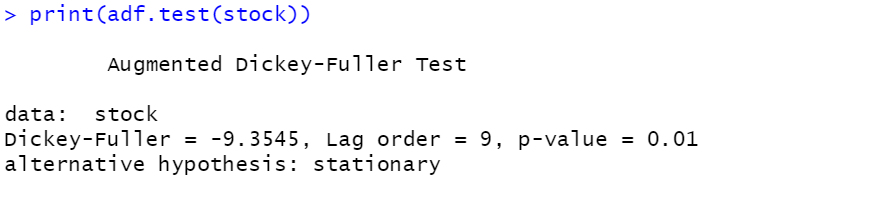
* Thay vì tính ra lợi nhuận của chứng khoán và hồi qui ARIMA trực tiếp trên giá trị này chúng ta sẽ tính sai phân của logarit và sử dụng như đầu vào của model ARIMA.

****

Chart

Description automatically generated

* Để kiểm định ADF, thêm thư viện library(tseries)

****

* p-value = 0.01 < 0.05 nên chuỗi có tính dừng.

Chia tập dữ liệu thành tập train và tập test

****

****

* Tính chỉ số ACF và PACF

****

Chart, histogram

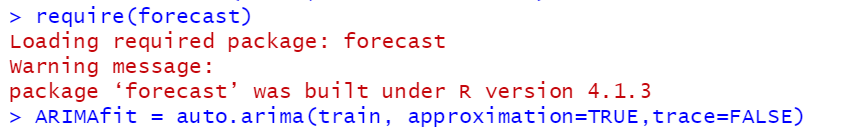
Description automatically generated

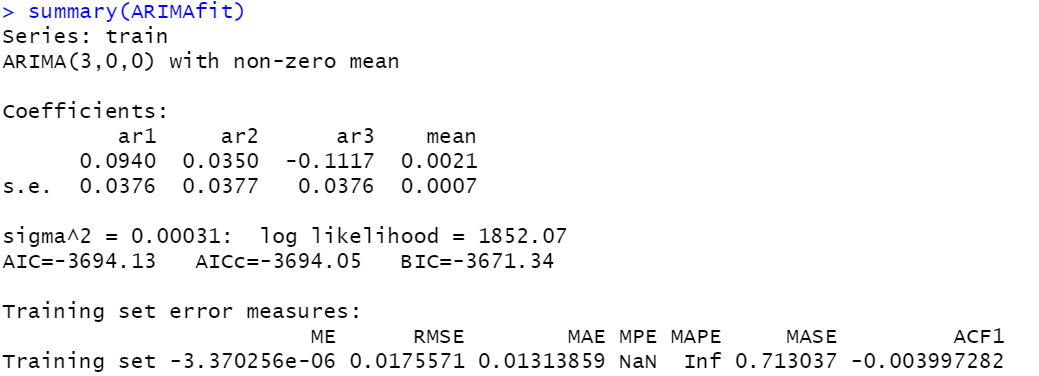
****

Chart

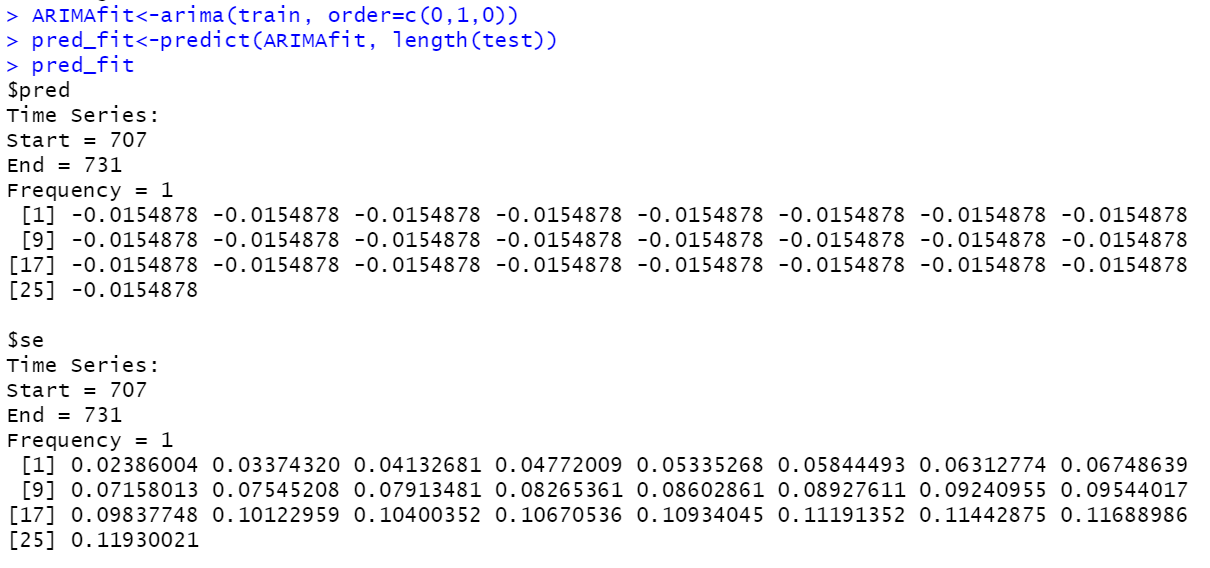
Description automatically generated

* Require package (forecast), xây dựng mô hình ARIMA

****

****

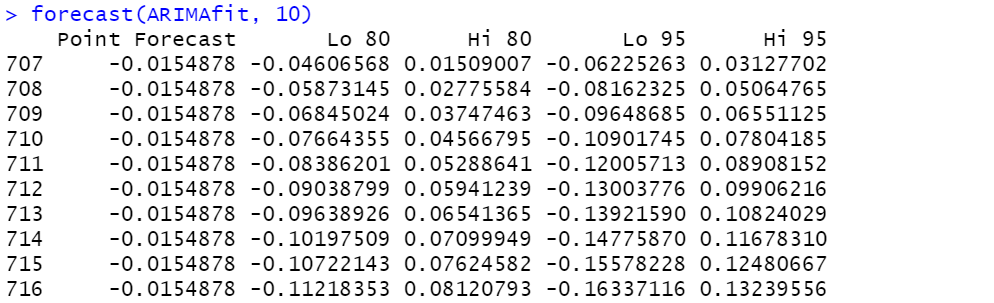
* Dự báo mô hình, trong đó pred là dự báo sai phân bậc nhất

****

* Require library (Metrics), dùng tập test kiểm tra mô hình.

****

****

****

****

Chart

Description automatically generated

1. **Bảng phân công công việc**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nguyễn Hoàng Long | Phan Phạm Quỳnh Hoa | Huỳnh Minh Thư |
| Xây dựng Template báo cáo | X |  |  |
| Tìm hiểu và sử dụng Excel |  |  | X |
| Tìm hiểu và sử dụng R |  | X |  |
| Tìm hiểu và sử dụng Python | X |  |  |
| Làm bài tập 4a,4b, 4c (Sử dụng Excel) |  |  | X |
| Làm bài tập 4a, 4b, 4c (Sử dụng R) |  | X |  |
| Làm bài tập 4a, 4b, 4c (Sử dụng Python) | X |  |  |

1. **Tài liệu tham khảo**

[1] <https://nguyenvantien0405.files.wordpress.com/2018/03/hd_excel_xstk_161223.pdf>

[2] <https://corporatefinanceinstitute.com/resources/excel/functions/>

[3] <https://nghiencuugiaoduc.com.vn/hoi-quy-logistic-nhi-thuc-binomial-logistic-regression/>

[4] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html>

[5] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html>

[6] <https://rpubs.com/lengockhanhi/445130>

[7] <http://rstudio-pubs-static.s3.amazonaws.com/271055_b970cf16eda047259a1a3d4b46a856e3.html>