DATA SCIENCE SALARIES PREDICTION

Dự đoán tiền lương ngành Khoa học Dữ liệu

Nguyễn Hồng Cát Thy

*Khoa Khoa học máy tính*

*Trường Đại học Công nghệ Thông tin*

*Thành phố Hồ Chí Minh, Việt Nam*

21522665@gm.uit.edu.vn

Trương Nguyên Hạo

*Khoa Khoa học máy tính*

*Trường Đại học Công nghệ Thông tin*

*Thành phố Hồ Chí Minh, Việt Nam*

21522051@gm.uit.edu.vn

*Tóm tắt* — Báo cáo đồ án dự đoán tiền lương ngành khoa học dữ liệu trình bày kết quả của một nghiên cứu về tiền lương trong lĩnh vực này. Nghiên cứu được thực hiện bằng cách thu thập thông tin về mức lương của các chuyên gia hàng đầu trong lĩnh vực này, và sử dụng các mô hình dự đoán là Multiple Linear Regression, Random Forest Regression và Gradient Boosting để đưa ra dự đoán về mức lương trong tương lai.

Kết quả của nghiên cứu cho thấy rằng, mức lương trung bình của các chuyên gia trong lĩnh vực khoa học dữ liệu là khá cao, và có xu hướng tăng dần theo thời gian. Các yếu tố ảnh hưởng đến mức lương bao gồm kinh nghiệm, trình độ học vấn, vị trí công việc và đặc biệt là quốc gia cư trú của nhân viên và công ty.

Nghiên cứu cũng đưa ra các khuyến nghị cho những ai muốn tìm kiếm công việc trong ngành khoa học dữ liệu, bao gồm tập trung vào việc rèn luyện kỹ năng và trau dồi kiến thức về các công nghệ mới nhất, và tìm kiếm các cơ hội để tham gia vào các dự án và hoạt động trong lĩnh vực này.

Từ khóa — Data Science, Khoa học dữ liệu, dự đoán, tiền lương

# giới thiệu

Hiện nay, ngành khoa học dữ liệu đang phát triển mạnh mẽ và trở thành một trong những ngành có tiềm năng trong tương lai.

Vì vậy, dự đoán tiền lương của các chuyên gia trong lĩnh vực này sẽ giúp các công ty, tổ chức, doanh nghiệp hoặc nhà tuyển dụng ở các nước phát triển dự đoán được mức lương phù hợp cho các vị trí việc làm, từ đó có thể giúp cho quản lý tài chính hiệu quả hơn.

Để giải quyết bài toán này, ta cần có một tập dữ liệu về các thông tin liên quan đến mức lương của các chuyên gia và nhà phân tích dữ liệu. Sau đó, ta sử dụng các thuật toán Machine Learning như Multiple Linear Regression, Random Forest Regression và Gradient Boosting để xây dựng mô hình dự đoán mức lương.

# dữ liệu

## Giới thiệu bộ dữ liệu

Bộ dữ liệu được nhóm thu thập từ trang web Kaggle [1] - một kho dữ liệu trực truyến lớn và cung cấp các công cụ và tài nguyên hỗ trợ cho việc phân tích dữ liệu và machine learning.

Có tổng cộng 607 mẫu với 11 trường dữ liệu bao gồm các thuộc tính là:

* Work year (năm làm việc): là thời gian lấy mẫu dữ liệu.
* Salary (lương): là lương được trả khi lấy mẫu (được tính theo mệnh giá riêng của từng quốc gia)
* Salary currency (mệnh giá lương): là mệnh giá tiền tệ riêng của từng quốc gia.
* Salary in usd (lương theo USD): là mức lương đã được chuyển sang mệnh giá USD (Mỹ).
* Job title (vị trí công việc): là vị trí làm việc ngành khoa học dữ liệu khi lấy mẫu.
* Experience level (kinh nghiệm): là kinh nghiệm làm việc khi lấy mẫu. Ví dụ: EN: Entry-level hoặc Junior (thực tập).
* Employment type (loại hình công việc): PT (Part-time), FT (Full-time), CT (Contract), FL (Freelance).
* Employee residence (nơi cư trú): là quốc gia cư trú chính của nhân viên trong thời gian làm việc, được ghi lại bằng mã quốc gia ISO 3166 [2].
* Company location (vị trí công ty): là vị trí công ty hoặc nơi đặt trụ sở chính của tập đoàn. Được ghi lại bằng mã quốc gia ISO 3166.
* Remote ratio (làm việc từ xa): là phần trăm thời gian nhân viên làm việc từ xa. 0 cho làm việc hoàn toàn trực tiếp hoặc ít hơn 20% từ xa; 50 cho khoảng từ 21% đến 79% từ xa; và 100 cho 80% hoặc hoàn toàn từ xa.
* Company size (quy mô công ty): là số người trung bình làm việc cho công ty trong năm: S cho công ty dưới 50 nhân viên (nhỏ); M: từ 50 đến 250 nhân viên (trung bình); L: trên 250 nhân viên (lớn)

Và không có dữ liệu null.

Các kiểu dữ liểu của các trường bao gồm:

* Kiểu int: work\_year, salary, salary\_in\_usd, remote\_ratio.
* Kiểu object: experience\_level, employment\_type, job\_title, salary\_currency, employee\_residence, company\_location, company\_size.

Bảng 1. Trực quan hóa dữ liệu đích

(salary in usd)

A picture containing text, screenshot, diagram, line

Description automatically generated

Bảng 2. Mức lương (usd) trung bình theo Employment level

A picture containing screenshot, text, line, diagram

Description automatically generated

Bảng 3. Trực quan hóa mức lương (usd) theo Remote ratio

A picture containing text, screenshot, diagram, line

Description automatically generated

Bảng 4. Trực quan hóa mức lương (usd) theo Employment type

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

Bảng 5. Mức lương (USD) theo vị trí công ty

A picture containing text, line, screenshot, diagram

Description automatically generated

Bảng 6. Trung bình lương (USD) theo quốc gia cư trú của nhân viên

A picture containing text, screenshot, diagram, line

Description automatically generated

Bảng 7. Trung bình lương (usd) theo mối liên hệ giữa nơi cư trú của nhân viên và vị trí công ty

A screen shot of a graph

Description automatically generated with medium confidence

Bảng 8. Mức lương (USD) trung bình của từng vị trícông việc qua các năm

A picture containing text, screenshot, font, design

Description automatically generated

# phương pháp máy học

## Multiple Linear Regression

Multiple Linear Regression (hồi quy tuyến tính đa biến) là một kỹ thuật thống kê được sử dụng để kiểm tra mối quan hệ giữa một biến phụ thuộc và nhiều biến độc lập. Nó là một phần mở rộng của hồi quy tuyến tính đơn biến (Simple Linear Regression) – thuật toán chỉ dùng đến một biến độc lập.

Trong hồi quy tuyến tính đa biến, biến phụ thuộc được dự đoán dưới dạng một hàm của hai hoặc nhiều biến độc lập như sau [3]:

Trong đó:

* : là biến phụ thuộc, là biến chịu tác động của biến khác.
* : là các biến độc lập, là biến tác động lên biến khác.
* : là hằng số hồi quy, hay còn được gọi là hệ số chặn. Đây là chỉ số nói lên giá trị của sẽ là bao nhiêu nếu tất cả đều bằng 0. Khi biểu diễn trên đồ thị Oxy, là điểm trên trục Oy mà đường hồi quy cắt qua.
* : là các hệ số hồi quy, hay còn được gọi là các hệ số góc. Chỉ số này cho chúng ta biết về mức thay đổi của gây ra bởi tương ứng. Nói cách khác, chỉ số này nói lên có bao nhiêu đơn vị sẽ thay đổi nếu tăng hoặc giảm một đơn vị.
* : là sai số, hay còn được gọi là phần dư. Chỉ số này càng lớn càng khiến cho khả năng dự đoán của hồi quy trở nên kém chính xác hơn hoặc sai lệch nhiều hơn so với thực tế.

Nếu mô hình không đạt được độ chính xác mong muốn, ta có thể tinh chỉnh lại mô hình bằng cách thay đổi các thông số hoặc biến đầu vào để cải thiện hiệu suất của mô hình.

## Random Forest Regression

Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Như một khu rừng gồm nhiều cây. Random Forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu.

Random Forests có thể được mô tả với một mã giả sau [4]:

• Bước 1: Chọn ngẫu nhiên k đặc trưng từ tập n đặc trưng (k ≤ n).

• Bước 2: Từ tập k đặc trưng đó, chọn ra node tốt nhất cho node phân loại.

• Bước 3: Chia các node con theo node tốt nhất tìm được.

• Bước 4: Lặp lại bước 1 đến 3 cho đến khi đạt k nodes.

Nếu mô hình không đạt được độ chính xác mong muốn, ta có thể thay đổi các giá trị như số thuộc tính tối đa, số nhánh cây,… đồng thời ta cũng có thể thay đổi số lượng thuộc tính k đầu vào để tinh chỉnh mô hình.

## Gradient Boosting

Gradient Boosting (tăng cường độ dốc) là một kỹ thuật học máy được sử dụng trong các nhiệm vụ hồi quy và phân loại, được phát triển bởi GS Jerome H. Friedman năm 1999. Với ý tưởng kết hợp các mô hình dự đoán yếu (thường là các decision tree) để cho ra một mô hình dự đoán mạnh duy nhất, Gradient Boosting được ra đời. Khi cây quyết định là cây học yếu, thuật toán kết quả được gọi là cây tăng cường độ dốc (Gradient Boosting Trees); nó thường hoạt động tốt hơn Random Forest.

Mô hình cây được tăng cường độ dốc được xây dựng theo kiểu kết hợp bằng cách thêm tuần tự các mô hình cơ sở sao cho mô hình sau sửa lỗi cho mô hình liền trước như trong các phương pháp tăng cường khác. Nhưng khác biệt ở chỗ nó tổng quát hóa các phương pháp khác bằng cách cho phép tối ưu hóa hàm mất mát khả vi tùy ý.

Đầu vào: tập huấn luyện , hàm mất mát khả vi tùy ý , số lần lặp .

Thuật toán Gradient Boosting [5]:

1. Khởi tạo mô hình với một giá trị không đổi:
2. Vòng lặp với đến :
3. Tính phần dư:

Với

1. Điều chỉnh mô hình cơ sở (mô hình yếu) gần với tỉ lệ để giảm mất mát, huấn luyện nó với tập huấn luyện .
2. Tính toán số nhân bằng cách giải hàm tối ưu một chiều sau:
3. Nâng cấp mô hình:

.

Đầu ra: .

## Độ đo sử dụng để đánh giá

Vì đây là một bài toán dạng hồi quy nên nhóm sẽ dùng kết hợp các độ đo chính xác và đánh giá sai số.

### Accuracy

### Mean Absolute Error (MAE)

MAE tính giá trị trung bình của sự khác biệt tuyệt đối giữa các giá trị dự đoán và các giá trị quan sát thực tế. Nó cho biết mức độ sai lệch trung bình giữa các giá trị dự đoán và giá trị thực tế. Công thức tính MAE như sau [6]:

.

### Mean Squared Error (MSE)

MSE được tính bằng cách lấy trung bình của tổng bình phương sai số giữa giá trị dự báo và giá trị thực tế trên toàn bộ các điểm dữ liệu [7]:

### Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) hoặc Root Mean Square Deviation (RMSD) là căn bậc hai của mức trung bình của các sai số bình phương. RMSE là độ lệch chuẩn của các phần dư (sai số dự đoán):

# quy trình xây dựng mô hình học máy dự đoán

## Tiền xử lý dữ liệu

Bảng 9. Bộ dữ liệu trước khi được xử lý

A picture containing screenshot, text, line, parallel

Description automatically generated

Đầu tiên, 2 trường dữ liệu là salary\_currency và salary sẽ bị loại bỏ để quy về cùng một đơn vị tiền tệ là USD (salary\_in\_usd). Từ đó dữ liệu sẽ dễ dàng so sánh hơn.

Hầu hết thuật toán máy học chỉ làm việc tốt với thuộc tính dạng số, do đó phải chuyển kiểu object thành số. Do dữ liệu thuộc kiểu Object trong bộ dữ liệu khá nhiều (6 trường) nên nhóm quyết định sử dụng LabelEncoder() có trong thư viện sklearn để gán nhãn dữ liệu.

Sau khi qua các bước Encoding dữ liệu có hình dạng như sau:

Bảng 10. Bộ dữ liệu sau khi được xử lý

A picture containing text, screenshot, number, parallel

Description automatically generated

Cùng với lượng lớn trường dữ liệu được hình thành từ quá trình Encoding. Việc ta vẽ biểu đồ correlation sẽ giúp loại bỏ đi các trường ít ảnh hướng đến dữ liệu mà ta cần dự đoán.

Bảng 11. Hệ số tương quan giữa các thuộc tính còn lại với thuộc tính Salary in USD

|  |  |
| --- | --- |
| **Thuộc tính** | **Hệ số tương quan** |
| Company size | -0.151205 |
| Employment type | -0.123545 |
| Job title | 0.131016 |
| Remote ratio | 0.132122 |
| Work year | 0.170493 |
| Experience level | 0.315312 |
| Company location | 0.428994 |
| Employee residence | 0.452501 |

Ta thấy được giá trị tuyệt đối của các trường dữ liệu tương đối ổn. Nên nhóm quyết định giữ lại tất cả các trường dữ liệu này để xây dựng mô hình dự đoán.

Tiếp theo ta chia bộ dữ liệu thành 2 tập là Training và Testing với tỷ lệ lần lượt là 0.8 và 0.2. Bởi vì tập dữ liệu không cân bằng nên nhóm sẽ sử dụng phương pháp K-Fold Cross-validation (Tập kiểm tra chéo K-Fold): Sử dụng phương pháp này để đánh giá hiệu suất mô hình trên các tập dữ liệu khác nhau. K-Fold Cross-validation chia tập dữ liệu thành K phần bằng nhau (gọi là fold), sau đó lần lượt sử dụng K-1 fold để huấn luyện và fold còn lại để kiểm tra. Quá trình này được lặp lại K lần và kết quả được tính trung bình. Ở đây nhóm sử dụng K = 10.

## Huấn luyện mô hình

Sau khi tiền xử lý dữ liệu, nhóm quyết định chọn ra 5 thuật toán để huấn luyện mô hình: Multiple Linear Regression, Random Forest, Gradient Boosting. Nhóm sử dụng các lớp có sẵn trong scikit-learn lần lượt là LinearRegression, GradientBoostingRegressor, RandomForestRegressorr để huấn luyện mô hình.

Với mỗi mô hình máy học có những siêu tham số riêng. Nhóm đã sử dụng GridSearchCV() của scikit-learn để đi tìm siêu tham số tốt nhất cho mô hình. Với phương pháp grid search, chương trình sẽ tự động thử tất cả các khả năng kết hợp, đánh giá theo cross validation.

Bảng 12. Kết quả tinh chỉnh mô hình sau khi dùng grid search trong scikit-learn

|  |  |
| --- | --- |
| Thuật toán | Siêu tham số |
| Multiple Linear Regression | copy\_X = True,  fit\_intercept = False,  positive = True |
| Random Forest | max\_features = 'log2',  n\_estimators = 50 |
| Gradient Boosting | max\_depth = 3  n\_estimators = 50 |

# kết quả

Kết quả của các mô hình với các tập huấn luyện phù hợp nhất với từng mô hình như sau [8]:

Bảng 13. Kết quả đánh giá mô hình

A picture containing text, screenshot, font, line

Description automatically generated

Theo bảng, mô hình Random Forest Regression và Gradient Boosting có điểm số cao hơn so với Multiple Linear Regression, với R2-Score là 47.11% và 44.11% tương ứng, so với 29.97% của Multiple Linear Regression. Điều này cho thấy rằng Random Forest Regression và Gradient Boosting có khả năng dự đoán tốt hơn các biến độc lập dựa trên các thông số đã được cung cấp.

Bên cạnh đó, khi so sánh MAE trong ba mô hình, Random Forest Regression có MAE thấp nhất (33055.15), tiếp đến là Gradient Boosting (33758.85) và cuối cùng là Multiple Linear Regression (41062.21).

Tổng kết lại, bảng cho thấy rằng Random Forest Regression và Gradient Boosting có hiệu quả dự đoán tốt hơn so với Multiple Linear Regression. Và trong hai mô hình, các độ đo sai số cho thấy mô hình Random Forest Regression có tiềm năng hơn.

# phân tích lỗi, hướng phát triển

Phân tích lỗi của mô hình:

* Bộ dữ liệu chưa được cân bằng lắm, ở thuộc tính employee\_residence thì US chiếm tới 55% trong tổng số 233 đối tượng. Đối với thuộc tính company\_location, US chiếm 58% trong 207 đối tượng. Cả 2 thuộc tính này đều tham gia vào việc huấn luyện mô hình nên đều gây ảnh hưởng.
* Bộ dữ liệu khiêm tốn, chưa tới 700 mẫu và không đại diện cho toàn bộ ngành khoa học dữ liệu.
* Accuracy của mô hình ở mức 46 - 47%, đang ở mức thấp. RMSE trong khoảng (45000 - 46000) đánh giá không quá cao nhưng có thể thấy được dự đoán còn nhiều sai lệch.

Hướng phát triển mô hình:

* Cân bằng lại bộ dữ liệu bằng cách tìm thêm hoặc chỉnh sửa lại.
* Sử dụng thêm các thuật toán Optimizers.
* Thử thêm các thuật toán khác.

# kết luận

Nhóm đã thu thập dữ liệu biến động tiền lương ngành khoa học dữ liệu từ năm 2020 đến năm 2022 và huấn luyện được mô hình để dự đoán tiền lương dựa trên các thông số về vị trí việc làm, năm làm việc, kinh nghiệm, v.v… Mô hình mà nhóm đã xây dựng đã có tính chính xác tương đối ổn và độ đo sai số có sự lệch không quá lớn khi sử dụng thuật toán Gradient Boosting và Random Forest. Từ đó, nhóm có triển vọng để tiếp tục khắc phục các lỗi đã nêu ra, tiếp tục xây dựng hoàn thiện ứng dụng mà nhóm muốn hướng đến trong tương lai.

##### tài liệu tham khảo

[1] “Data Science Job Salaries dataset”, 2022. [Trực tuyến]. Địa chỉ: [Data Science Job Salaries | Kaggle](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries?datasetId=2268489&searchQuery=pre) [Truy cập 15/06/2023].

[2] “Mã ISO các quốc gia trên thế giới”, 2021. [Trực tuyến]. Địa chỉ: [Mã ISO của các nước trên thế giới: 2 chữ, 3 chữ và mã số | Các Nước (cacnuoc.vn)](https://cacnuoc.vn/ma-iso-cac-nuoc-tren-the-gioi/) [Truy cập 15/06/2023].

[3] “Multiple Linear Regression”, 2020. [Trực tuyến]. Địa chỉ: [Multiple Linear Regression - Overview, Formula, How It Works (corporatefinanceinstitute.com)](https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/#:~:text=Multiple%20linear%20regression%20refers%20to%20a%20statistical%20technique%20that%20uses,variable%20in%20the%20total%20variance.) [Truy cập 15/06/2023].

[4] “Random forest, thế nào là một rừng ngẫu nhiên”, 2018. [Trực tuyến]. Địa chỉ: [Random Forest, thế nào là một rừng ngẫu nhiên. – Finding Data (wordpress.com)](https://couhpcode.wordpress.com/2018/01/24/random-forest-the-nao-la-mot-rung-ngau-nhien/) [Truy cập 15/06/2023].

[5] “Gradient boosting”, 2021. [Trực tuyến]. Địa chỉ: [Gradient boosting - Wikipedia](https://en.wikipedia.org/wiki/Gradient_boosting#:~:text=Gradient%20boosting%20is%20a%20machine,which%20are%20typically%20decision%20trees.) [Truy cập 16/06/2023].

[6] “Mean absolute error”, 2021. [Trực tuyến]. Địa chỉ: [Mean absolute error - Wikipedia](https://en.wikipedia.org/wiki/Mean_absolute_error) [Truy cập 16/06/2023].

[7] “Sai số toàn phương trung bình”, 2021. [Trực tuyến]. Địa chỉ: [Sai số toàn phương trung bình – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Sai_s%E1%BB%91_to%C3%A0n_ph%C6%B0%C6%A1ng_trung_b%C3%ACnh) [Truy cập 16/06/2023].

[8] Code, 2023. [Trực tuyến]. Địa chỉ:

https://drive.google.com/file/d/1KzMwX1P7QlrJIqEQUk82uiITT49Hqqjd/view?usp=sharing