ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ỨNG DỤNG THƯ VIỆN SKLEARN VÀO PHÂN TÍCH VÀ DỰ ĐOÁN MỨC LƯƠNG NHÓM NGÀNH KHOA HỌC DỮ LIỆU

Nhóm 19			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
64	Nguyễn Phú Tài	22521280	KHMT
65	Mai Văn Tân	22521301	KHMT
78	Nguyễn Công Trúc	22521550	KHMT
80	Trần Lê Nguyên Trung	22521568	KHMT

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Trong bối cảnh ngành Khoa học Dữ liệu đang ngày càng phát triển và đóng vai trò quan trọng trong nhiều lĩnh vực, việc hiểu rõ các yếu tố ảnh hưởng đến mức lương trong ngành này là một nhu cầu cấp thiết! Đề tài của chúng tôi tập trung vào việc phân tích và dự đoán mức lương của các vị trí trong ngành Khoa học Dữ liệu. Mục tiêu chính là xác định các yếu tố có tác động đến mức lương, từ đó xây dựng mô hình dự đoán hiệu quả dựa trên các thông tin thu thập được.

Để thực hiện nghiên cứu này, chúng tôi sử dụng bộ dữ liệu ***Data Science Jobs & Salaries 2024*** được công bố trên nền tảng Kaggle. Các công cụ như Python và các thư viện phân tích dữ liệu phổ biến (pandas, seaborn, matplotlib) được áp dụng để làm sạch dữ liệu và trực quan hóa. Đồng thời, các mô hình hồi quy tiên tiến như Random Forest, SVR, XGBoost và Linear Regression, với sự hỗ trợ của thư viện scikit-learn, đã được triển khai. Ngoài ra, các phương pháp tối ưu hóa tham số (RandomizedSearchCV) và đánh giá hiệu năng (K-Fold Cross Validation) được sử dụng để đảm bảo độ chính xác và tính khả dụng của mô hình. Kết quả nghiên cứu cho thấy mô hình XGBoost Regressor đạt hiệu năng cao nhất với R-squared = 0.829, MAE = 10650.671, và MSE = 268723315.569 trên tập kiểm tra. Chúng tôi cam kết minh bạch về đề tài và bộ dữ liệu, mã nguồn và các kết quả phân tích sẽ được công khai để đảm bảo tính toàn vẹn.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu này được trích xuất từ trang web Glassdoor, một nền tảng nổi tiếng cung cấp thông tin về đánh giá công ty và mức lương từ nhân viên. Dữ liệu này bao gồm các công việc liên quan đến lĩnh vực khoa học dữ liệu và các vị trí khác, cung cấp một cái nhìn tổng quan về cơ hội nghề nghiệp, mức lương, và yêu cầu công việc.

Bộ dữ liệu phân tích được tham khảo tại đường dẫn [1].

Bộ dữ liệu này nhằm mục đích cung cấp thông tin về xu hướng tuyển dụng trong lĩnh vực khoa học dữ liệu, bao gồm:

- Đánh giá mức lương ở nhiều công ty khác nhau.
- Phân tích nhu cầu tuyển dụng và yêu cầu về kỹ năng.
- So sánh đánh giá của các công ty dựa trên đánh giá của nhân viên.

- Khám phá các yếu tố ảnh hưởng đến mức lương như địa điểm, quy mô công ty, và loại hình sở hữu.

Thống kê ban đầu, bộ dữ liệu có:

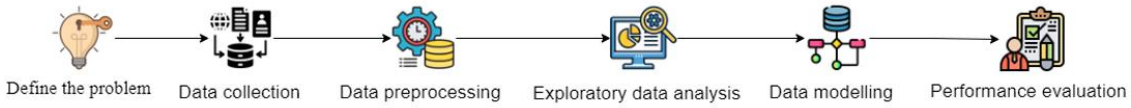
- 14 đặc trưng (cột) và 956 mẫu (dòng), là thông tin của công việc.
- Biến phân loại: “Job Title”, “Salary Estimate”, “Job Description”, “Company Name”, “Location”, “Headquarters”, “Size”, “Type of Ownership”, “Industry”, “Sector”, “Revenue”, “Competitors”.
- Biến số: “Rating”, “Founded”.

Bảng 2-1. Các thuộc tính của bộ dữ liệu

Tên thuộc tính	**Kiểu dữ liệu**	**Mô tả**	**Khoảng giá trị**
Job Title	object (string)	Tên công việc	“Data Scientist”, “Healthcare Data Scientist”, ...
Salary Estimate	object (string)	Khoảng lương, đi kèm là ghi chú về cách tính lương hoặc thông tin này do ai cung cấp	“$53K-$91K (Glassdoor est.)”, “$21-$34 Per Hour(Glassdoor est.)”, ...
Job Description	object (string)	Mô tả công việc, có chứa các kỹ năng yêu cầu (chuỗi rất dài)	“KnowBe4, Inc. is a high growth information sec...”, ...
Rating	float64	Điểm số đánh giá về công ty	[-1.0, 5.0]
Company Name	object (string)	Tên công ty, kèm với điểm đánh giá	“KnowBe4\n4.8”, “PNNL\n3.8”, ...
Location	object (string)	Địa điểm làm việc	“Linthicum, MD”, “Clearwater, FL”, ...

Headquarters	object (string)	Địa điểm trụ sở chính	"Baltimore, MD", "Clearwater, FL", ...
Size	object (string)	Quy mô công ty	"1 to 50 employees", "51 to 200 employees", ...
Founded	int64	Năm thành lập	[-1, 2019]
Type of ownership	object (string)	Loại hình sở hữu công ty	"Company - Private", "Government", ...
Industry	object (string)	Ngành công nghiệp	"Energy", "Security Services", ...
Sector	object (string)	Lĩnh vực	"Health Care", "Business Services", ...
Revenue	object (string)	Doanh thu công ty	"$2 to $5 billion (USD)", "$50 to $100 million (USD)", ...
Competitors	object (string)	Tên các công ty đối thủ	-1 hoặc "Novartis, Baxter, Pfizer", "Travelers, Allstate, State Farm", ...

3. PHƯƠNG PHÁP PHÂN TÍCH



Hình 3-1. Phương pháp phân tích

3.1. Define the problem (Định nghĩa vấn đề)

Bước đầu tiên, chúng tôi đã xác định rõ lĩnh vực và chủ đề mà mình quan tâm để từ đó chọn được đề tài phân tích. Đồng thời, các thành viên cũng đã tự đặt các câu hỏi về đầu ra của đồ án này cho chính mình: thông qua bài phân tích này, mỗi người có thể

làm được gì, học thêm gì kĩ năng nào, và có thể áp dụng được những kĩ năng đó vào hiện tại và mai sau như thế nào…

3.2. Data collection (Thu thập dữ liệu)

Ở giai đoạn này, mỗi thành viên cần tự tìm kiếm một bộ dữ liệu phù hợp với tiêu chí mà nhóm đã đề ra. Kết quả hầu hết các thành viên đều chọn các bộ dữ liệu trên Kaggle do tính phổ biến và dễ truy cập của trang này. Sau đó, chúng tôi đã thảo luận sôi nổi để chọn ra bộ dữ liệu phù hợp nhất và tiến hành bước tiếp theo.

3.3. Data preprocessing (Tiền xử lý dữ liệu)

Trong bước này, chúng tôi đã tiến hành kiểm tra và làm sạch dữ liệu để đảm bảo chất lượng và tính nhất quán cho quá trình phân tích. Các công việc được thực hiện cụ thể như sau:

3.3.1. Kiểm tra dữ liệu

Chúng tôi đã thực hiện các bước kiểm tra nhằm đánh giá chất lượng dữ liệu ban đầu:

- Missing values: Kiểm tra các giá trị bị thiếu trong từng cột để xác định các vấn đề cần xử lý.
- Duplicates: Tìm và hiển thị các dòng bị trùng lặp trong DataFrame.
- Data Types: Kiểm tra kiểu dữ liệu của từng cột để đảm bảo dữ liệu được định dạng đúng.
- Unique Values: Đếm số lượng giá trị duy nhất trong từng cột để xác định tính đa dạng của dữ liệu.
- Statistics: Kiểm tra các thông số thống kê của dữ liệu số, bao gồm giá trị trung bình (mean), nhỏ nhất (min), lớn nhất (max), và độ lệch chuẩn (std).
- Categories: Xem xét các giá trị của các cột thuộc loại phân loại (category) để hiểu rõ phân phối dữ liệu.

3.3.2. Làm sạch dữ liệu

Dựa trên kết quả kiểm tra, chúng tôi đã thực hiện các bước làm sạch dữ liệu như sau:

- Loại bỏ các cột không cần thiết: Xóa cột 'Unnamed: 0', thường là cột index mặc định không cần thiết.
- Xử lý cột Job Title:
  + Tạo cột job_simplified bằng cách đơn giản hóa tên công việc. Ví dụ: "Data Scientist", "Healthcare Data Scientist" được chuyển thành "data scientist".
  + Tạo cột seniority để trích xuất thông tin về mức độ thâm niên từ tên công việc. Ví dụ: "Senior Data Scientist" → "Senior", "Data Analyst II" → "Associate", "Data Science Intern" → "Entry Level".
- Xử lý cột Salary Estimate:
  + Loại bỏ các dòng có giá trị '-1' (mức lương không xác định).
  + Tạo cột Hourly để đánh dấu các công việc có lương theo giờ.
  + Tạo cột employer_provided để đánh dấu các công việc có lương do nhà tuyển dụng cung cấp.
  + Tách cột Salary Estimate thành hai cột Min Salary và Max Salary bằng cách loại bỏ các ký tự không cần thiết.
  + Tính toán Average Salary từ giá trị trung bình của Min Salary và Max Salary.
- Xử lý cột Job Description: Trích xuất và lựa chọn các kỹ năng quan trọng từ mô tả công việc, bao gồm: 'python', 'spark', 'aws', 'excel', 'tableau', 'SQL', 'SAS', 'MS Access', 'Data Visualization Tools', 'Algorithmic Aptitude'.
- Xử lý cột Rating: Tạo cột Rating Category để phân loại công ty theo các ngưỡng đánh giá.
- Xử lý cột Size: Thay thế giá trị -1 bằng 'Unknown'.
- Xử lý cột Revenue: Thay thế giá trị -1 và ‘Unknown/Non/Applicable’ bằng ‘Unknown’.
- Xử lý cột Company Name: Loại bỏ các ký tự không cần thiết như "\n" và các số phía sau tên công ty.
- Xử lý cột Competitors: Xóa cột này vì không cần thiết cho phân tích.
- Xử lý cột Founded: Tạo cột Age để thể hiện tuổi của công ty bằng cách tính toán từ năm thành lập.

- Xử lý cột Type of Ownership: Thay thế giá trị -1 bằng 'Unknown'.

3.3.3. Lưu dữ liệu đã làm sạch

Sau khi hoàn thành các bước làm sạch, chúng tôi đã lưu bộ dữ liệu đã xử lý dưới dạng file CSV với tên data_EDA.csv để sử dụng cho các bước phân tích tiếp theo.

3.4. Exploratory data analysis (Phân tích thăm dò dữ liệu)

Nhằm hiểu rõ hơn về cấu trúc và mối quan hệ giữa các biến trong tập dữ liệu, chúng tôi đã thực hiện bước phân tích dữ liệu. Dưới đây là các bước phân tích chi tiết:

3.4.1. Phân tích đơn biến (Univariate Analysis)

- Phân phối lương: Sử dụng biểu đồ histogram và boxplot để phân tích phân phối của Average Salary, Min Salary, và Max Salary. Điều này giúp nhận diện các xu hướng và outliers trong dữ liệu lương.
- Phân phối đánh giá và tuổi: Vẽ histogram và boxplot cho Rating và Age để hiểu rõ hơn về sự phân phối của các đánh giá công ty và tuổi của các công ty.
- Phân phối của các biến phân loại nhị phân: Sử dụng biểu đồ cột chồng (stacked bar plot) để phân tích các biến nhị phân như same_state, SQL_yn, Tableau_yn, Excel_yn, AWS_yn, Spark, và Python_yn.
- Phân phối dữ liệu dạng phân loại: Phân tích các cột dạng object như job_simplified, Sector, và Type of ownership bằng cách sử dụng countplot để xem phân phối từng giá trị.

3.4.2. Phân tích hai biến (Bivariate Analysis)

Nhóm 1: Liên quan đến công việc

- Mối quan hệ giữa lương và các chức danh công việc: Sử dụng boxplot để phân tích sự phân phối lương theo job_simplified và seniority.
- Top ngành nghề theo lương trung bình: Tính toán lương trung bình theo Industry và Sector, sau đó vẽ barplot để xác định các ngành nghề có mức lương cao nhất.

Nhóm 2: Liên quan đến công ty

- Công ty có số lần tuyển dụng nhiều nhất: Đếm số lượng công việc theo Company Name và vẽ barplot cho top 10 công ty.
- Top công ty theo đánh giá: Tính trung bình Rating của các công ty và xác định top 10 công ty có đánh giá cao nhất.

- Công ty lâu đời nhất: Phân tích tuổi (Age) của công ty để xác định công ty lâu đời nhất.
- Lương trung bình của các công ty có nhiều công việc nhất: Phân tích lương trung bình cho top 10 công ty có số lượng công việc cao nhất.
- Mối quan hệ giữa Rating và các yếu tố công ty: Sử dụng boxplot để phân tích phân phối Rating theo Size, Type of ownership, và Revenue.

Nhóm 3: Liên quan đến kỹ năng

- Kỹ năng theo nhóm công việc: Lọc các công việc trọng tâm như Data Scientist, Data Analyst, và phân tích tần suất yêu cầu kỹ năng (Python, SQL, Tableau,...) bằng countplot.

Nhóm 4: Liên quan đến vị trí

- Lương trung bình theo bang: Tính toán lương trung bình theo job_state và vẽ biểu đồ barh.
- Phân tích lương trung bình tại các thành phố có số lượng công việc cao nhất: Xác định các thành phố có nhiều công việc nhất và phân tích lương trung bình tại các thành phố đó.

3.4.3. Phân tích sâu hơn (Bivariate Analysis với các nhóm cụ thể)

- Mối quan hệ giữa lương và doanh thu: Sử dụng boxplot để phân tích phân phối lương trung bình (Average Salary) theo nhóm Revenue.
- Mối quan hệ giữa độ tuổi và quy mô công ty: Phân tích độ tuổi (Age) theo Size bằng boxplot.
- Mối quan hệ giữa quy mô và loại hình sở hữu: Phân tích số lượng công ty (Count) theo Type of ownership và Size.

3.4.4. Trích xuất/ chọn ra các biến quan trọng cho huấn luyện mô hình

- Phân tích ma trận tương quan của các đặc trưng số: Sử dụng ma trận tương quan để xác định mối quan hệ giữa các đặc trưng số và lựa chọn các biến quan trọng.
- Tìm biến phân loại có khả năng ảnh hưởng đến mức lương: Sử dụng phân tích ANOVA để xác định các biến phân loại có ảnh hưởng đáng kể đến mức lương.

3.5. Model training (Huấn luyện mô hình)

Huấn luyện mô hình là một bước quan trọng trong việc xây dựng hệ thống dự đoán. Để hiện thực công việc này, chúng tôi đã thực hiện các bước như sau:

- Đọc và kiểm tra dữ liệu.
- Sử dụng hàm train_test_split từ thư viện scikit-learn để chia dữ liệu thành hai phần: 80% cho tập huấn luyện và 20% cho tập kiểm tra và đặt random_state=42 để đảm bảo tính tái lập của việc chia dữ liệu.
- Xây dựng đường ống dữ liệu (build model pipeline) bằng cách sử dụng StandardScaler để chuẩn hóa các biến số, OneHotEncoder để mã hóa các biến phân loại.
- Lựa chọn mô hình:
    + Xây dựng một số pipeline cho các mô hình hồi quy khác nhau (SVR, XGBoost, KNeighbors, DecisionTree, RandomForest, LinearRegression, HistogramGradientBoosting).
    + Chạy XGBoost với tham số mặc định. Tinh chỉnh siêu tham số cho XGBoost (tuning).
    + Khởi tạo XGBRegressor với tham số mặc định. Tinh chỉnh siêu tham số cho XGBRegressor.
    + Sử dụng RandomizedSearchCV với n_iter=50 để tìm kiếm bộ tham số tốt nhất dựa trên tập huấn luyện.

3.6. Performance evaluation (Đánh giá hiệu suất)

Để mô hình dự đoán hoạt động tốt trên dữ liệu chưa từng thấy, chúng tôi cần đánh giá hiệu suất của mô hình. Dưới đây là các bước đã được thực hiện:

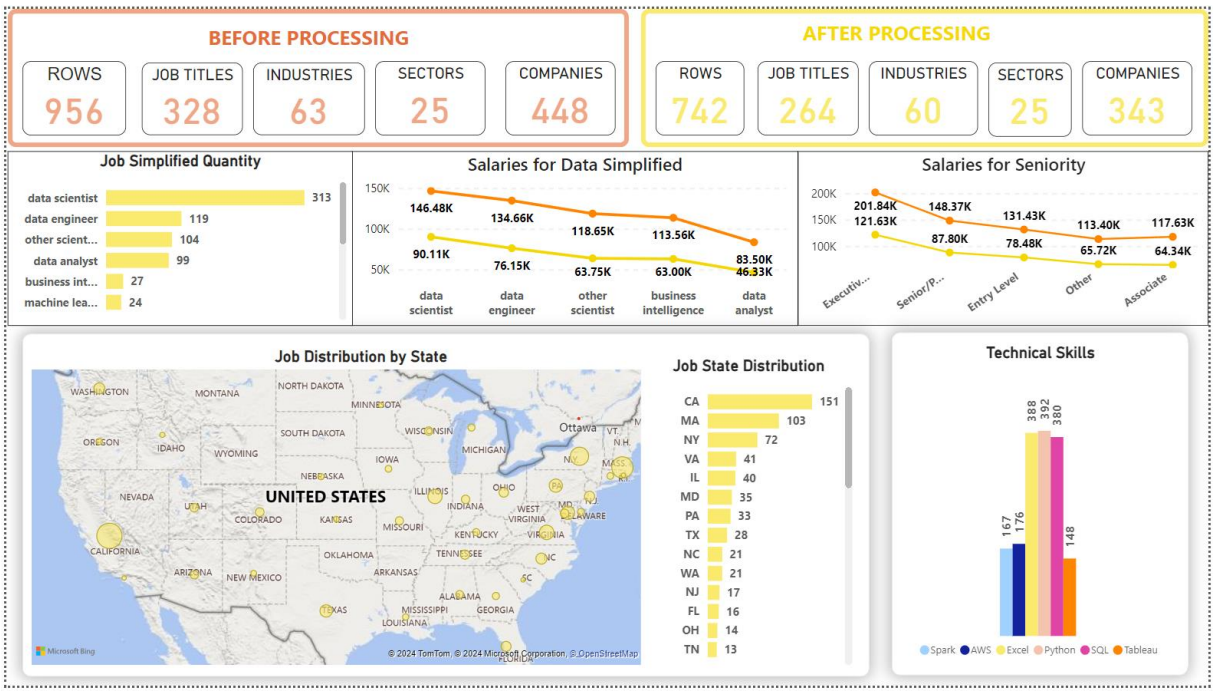
- Huấn luyện và đánh giá từng mô hình bằng các độ đo MSE, MAE và R^2. Kết quả được lưu vào một DataFrame và sắp xếp theo R^2 giảm dần để dễ dàng so sánh hiệu suất của chúng.
- Đánh giá mô hình XGBoost với tham số tốt nhất.
- Kiểm tra chéo (Cross Validation - K-Fold): Sử dụng KFold với 10 folds để đánh giá mô hình một cách tổng quát hơn, tính toán MSE, MAE, R^2 cho mỗi fold và tính trung bình của các độ đo này.

- Trực quan hóa (Visualization): Vẽ biểu đồ cột hiển thị MSE, MAE và R^2 cho mỗi fold của K-Fold cross-validation. Cũng vẽ thêm một biểu đồ tổng hợp thể hiện cả ba độ đo trên từng fold.

4. PHÂN TÍCH THĂM DÒ/SƠ BỘ

4.1. Kết quả

Để hỗ trợ quá trình phân tích và trình bày kết quả một cách trực quan, chúng tôi đã sử dụng Power BI để trực quan hóa dữ liệu trước và sau khi xử lý. Kết quả thu được như ảnh sau:



Hình 4-1. Dữ liệu trước và sau xử lý: Sự thay đổi về số lượng hàng, chức danh công việc, ngành nghề, và công ty, cùng với thông tin về mức lương và kỹ năng kỹ thuật.

- Tổng quan dữ liệu, sau khi xử lý:
  - + Dữ liệu giảm từ 956 dòng xuống còn 742.
  - + Số lượng chức danh công việc: từ 328 còn 264.
  - + Số lượng công ty: từ 448 giảm còn 343.
- Phân phối và mức lương của công việc:
  - + Các chức danh công việc phổ biến trong khối ngành dữ liệu: data scientist (313), data engineer (119), data analyst (99), business intelligence (27).

+ Data Scientist là công việc phổ biến nhất, với mức lương cao ($146.48K).
+ Ngoài ra, mức lương các vị trí công việc Data Engineer cũng khá cao $134.66K và mức lương của Data Analyst là thấp nhất $83.5K.

- Mức lương theo cấp bậc: Việc này khá rõ ràng khi mức lương trung bình giảm dần theo cấp bậc (Executive/Director: cao nhất với $201.84K), (Entry Level: $78.48K), (Associate: $64.34K).
- Phân phối địa lý: California (151 công việc) dẫn đầu, tiếp theo là Massachusetts (103) và New York (72).
- Kỹ năng hàng đầu: Python, SQL, và Excel là những kỹ năng được yêu cầu nhiều nhất (hơn 380 lần mỗi kỹ năng).

4.2. Giải thích, bình luận

- Phân tích chức danh giúp nhóm các công việc tương tự vào cùng một nhóm như *Data Scientist*, *Data Engineer*, *Data Analyst*,... Điều này giúp chúng ta hiểu về xu hướng nghề nghiệp và nhu cầu thị trường trong lĩnh vực dữ liệu.
- Dữ liệu về địa lý cho thấy các bang như **California** và **Massachusetts** có cơ hội việc làm cao, phản ánh sự tập trung công việc tại các trung tâm công nghệ và khoa học như Silicon Valley, Boston.
- Từ cột kỹ năng, chúng ta nhận thấy những kỹ năng như **Python**, **SQL**, và **Excel** là bắt buộc, cho thấy nhu cầu về phân tích dữ liệu và lập trình là cốt lõi.

5. KẾT QUẢ PHÂN TÍCH

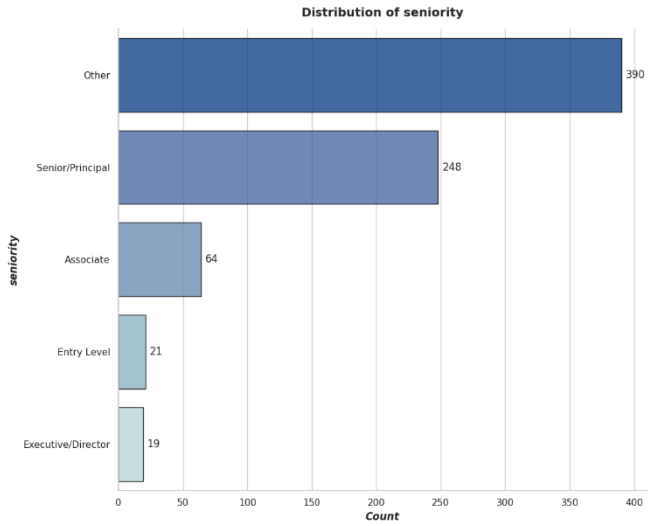
5.1. Kết quả của Phân tích thăm dò dữ liệu

Dựa vào dữ liệu từ những bài đăng tuyển nhân sự trong khối ngành khoa học dữ liệu được thu thập và phân tích, chúng tôi nhận thấy thị trường làm việc trong lĩnh vực này hiện nay đang rất được quan tâm.

5.1.1. Thông tin về công việc

- Thị trường việc làm trong lĩnh vực khoa học dữ liệu rất đa dạng, chủ yếu tập trung vào các chức danh như Data Scientist, Data Engineer, Data Analyst, và Business Intelligence.
- Thị trường chủ yếu tuyển dụng nhân lực có kinh nghiệm, với các vị trí Senior/Principal chiếm đa số.

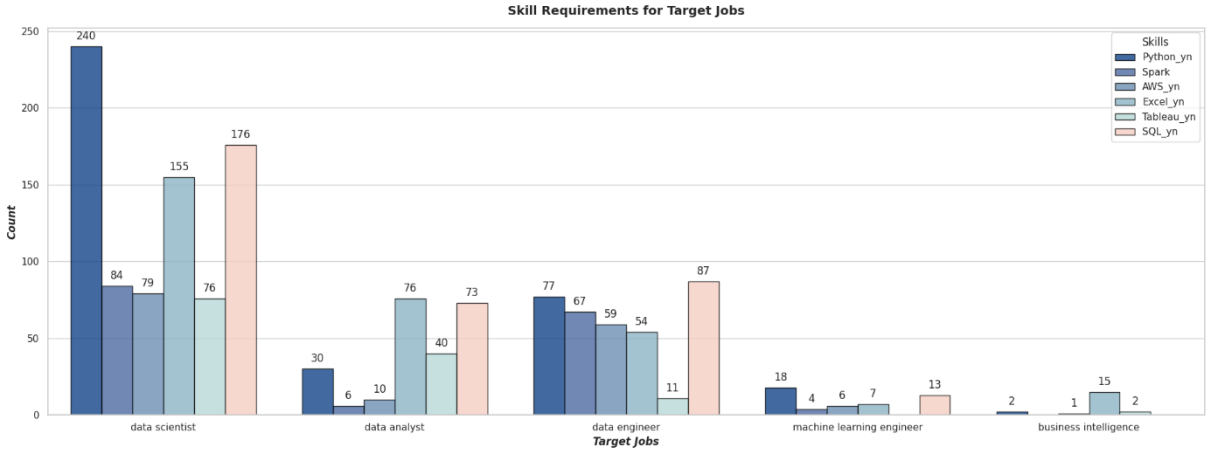
- Trong các bài tuyển dụng, nhóm "other" (không rõ thâm niên) chiếm 390 vị trí, cho thấy một số cơ hội cho ứng viên mới vào nghề. Tuy nhiên vẫn tập trung chủ yếu vào nhân lực có kinh nghiệm.



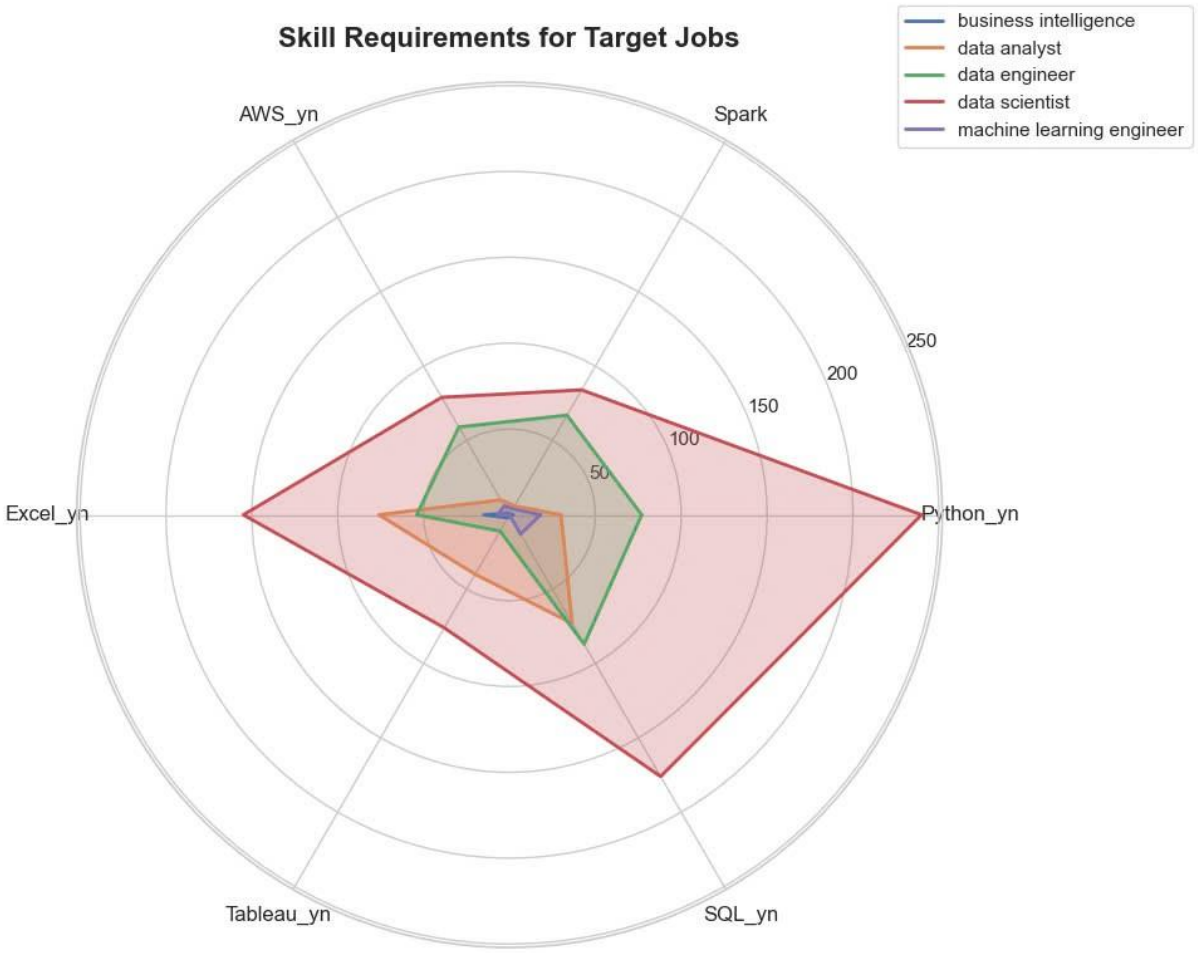
Hình 5-1. Biểu đồ phân phối về thâm niên làm việc

5.1.2. Kỹ năng cần thiết

- Python và SQL là hai kỹ năng quan trọng nhất, đặc biệt cho các vị trí như Data Scientist và Data Engineer.
- Excel và Tableau thường được yêu cầu nhiều cho Data Analyst và Business Intelligence, nhưng ít quan trọng hơn đối với Data Engineer và Machine Learning Engineer.
- AWS và Spark được yêu cầu nhiều hơn cho các vị trí kỹ thuật như Data Engineer và Data Scientist.



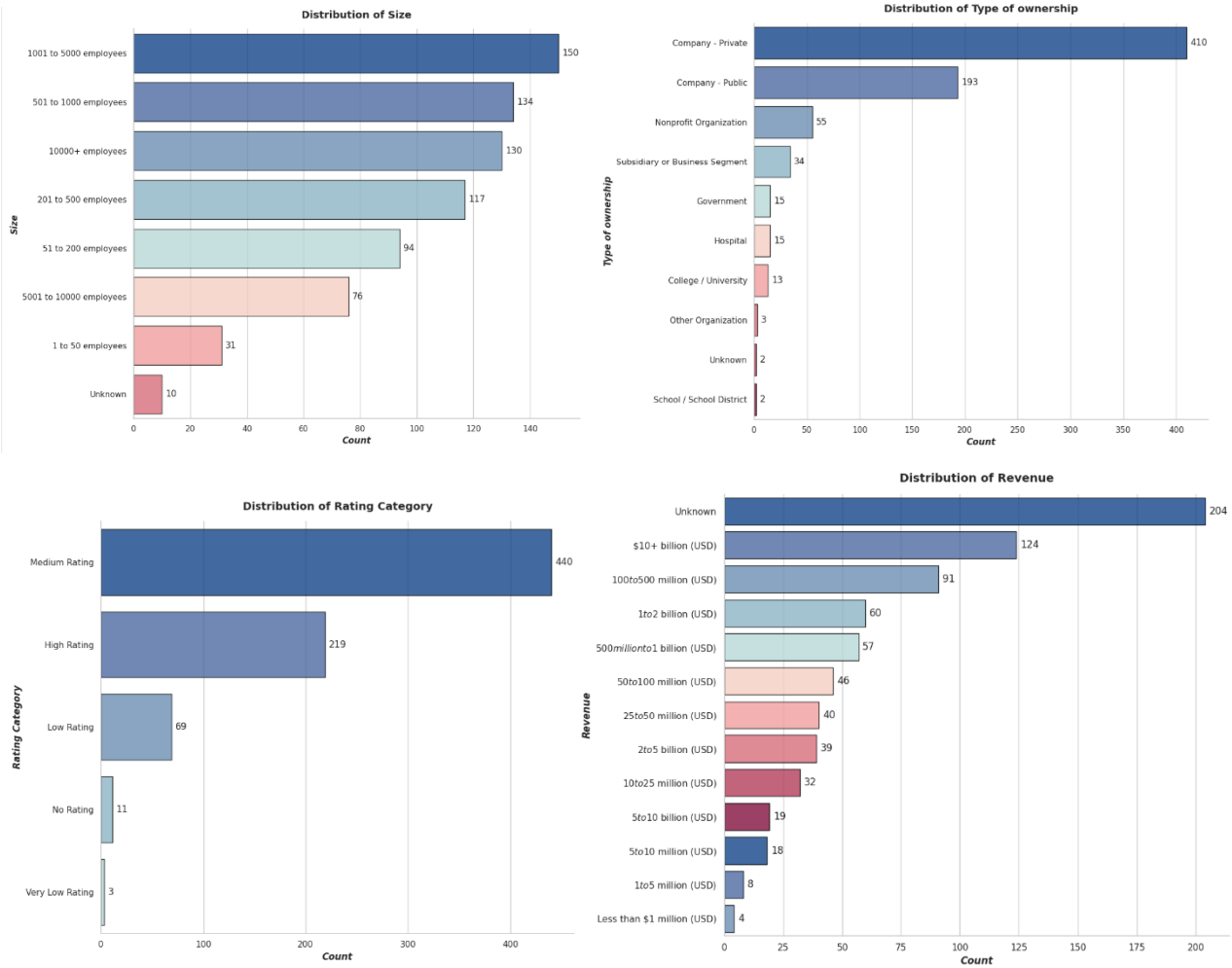
Hình 5-2. Countplot Kỹ năng yêu cầu cho từng vị trí công việc



Hình 5-3. Radar plot Kỹ năng yêu cầu cho từng vị trí công việc

5.1.3. Thông tin về công ty

- Trang tuyển dụng tiếp cận hơn 400 công ty trên thị trường việc làm tại Mỹ, phân phối trên 25 lĩnh vực khác nhau.
- Các công ty trên trang tuyển dụng chủ yếu là công ty tư nhân, với quy mô từ trung bình đến lớn.
- Doanh thu của các công ty rất đa dạng, với một tỷ lệ lớn không rõ thông tin cụ thể. Tuy nhiên, phần lớn các công ty có mức xếp hạng trung bình hoặc cao, cho thấy mức uy tín và chất lượng của các công ty này.
- Trang tuyển dụng này là một nơi lý tưởng cho các ứng viên muốn tham gia vào các công ty lớn, nơi có cơ hội về lương bổng, điều kiện làm việc, và chất lượng công việc tốt.



Hình 5-4. Phân bố về kích thước, loại hình, đánh giá và doanh thu của các công ty

5.1.4. Thông tin về địa điểm và khu vực:

- Phân bố công việc: Các công việc trong lĩnh vực khoa học dữ liệu được phân bố ở hầu hết các tiểu bang trên nước Mỹ, đặc biệt tập trung tại các bang lớn như California, Massachusetts, và New York. Đây là những khu vực phát triển mạnh về kinh tế và công nghệ, thu hút nhiều công ty lớn đến đầu tư và hoạt động.
- Cơ hội và thách thức:
  + Những thành phố lớn này, giống như TP Hồ Chí Minh ở Việt Nam, mang đến nhiều cơ hội việc làm và sự lựa chọn phong phú cho các ứng viên. Thị trường lao động tại đây rất dồi dào, cho phép ứng viên tìm kiếm các vị trí phù hợp với kỹ năng và kinh nghiệm của mình.
  + Tuy nhiên, đi kèm với cơ hội là những thách thức, đặc biệt là chi phí sinh hoạt cao. Ứng viên từ các thành phố nhỏ hơn có thể gặp khó khăn trong

việc thích nghi với mức sống đắt đỏ tại các thành phố lớn như San Francisco, Boston, hay New York.

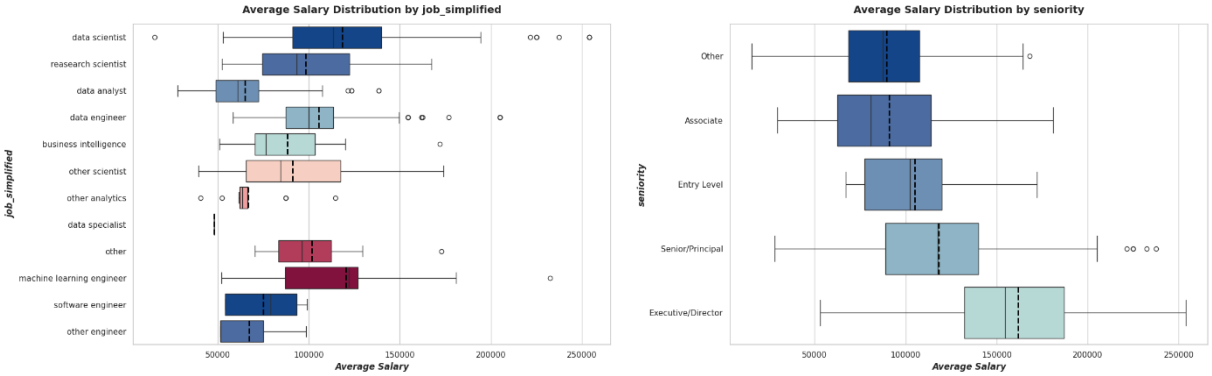
5.1.5. *Thông tin về mức lương trên thị trường*

Đây là thông tin được coi như quan trọng nhất trong bài báo cáo của nhóm. Lương là yếu tố mà cả nhà tuyển dụng và cả ứng viên trên trang tuyển dụng đều quan tâm. Khi có sự quan tâm từ 2 phía thì bắt buộc nhà phân tích phải đưa ra những lý giải hợp lý nhất đối với thông tin này.

Mức lương trung bình trên thị trường dao động từ $74,700 đến $128,200.

Ảnh hưởng của kinh nghiệm làm việc

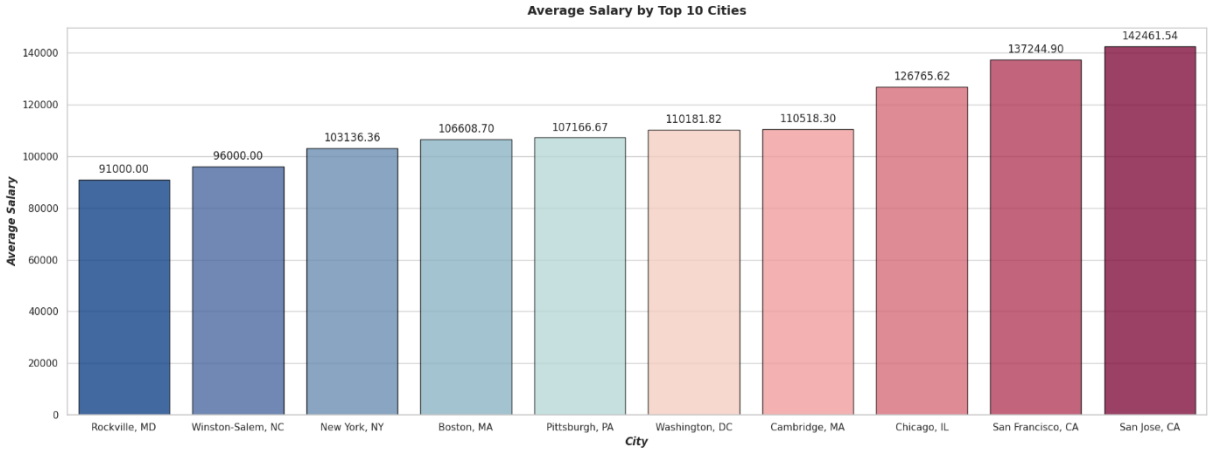
Mức lương khác nhau tùy thuộc vào từng chức danh, nhưng rõ ràng Data Scientist luôn có mức lương cao nhất trên thị trường. Bên cạnh đó, kinh nghiệm làm việc cũng ảnh hưởng đáng kể đến mức lương. Nhân viên ở cấp độ sơ cấp (Entry-level) thường nhận lương thấp hơn so với những người có kinh nghiệm lâu năm (Senior) hoặc giữ vị trí quản lý (Manager/Director).



Hình 5-5. Phân bố lương trung bình theo vị trí và thâm niên làm việc

Ảnh hưởng của vị trí địa lý

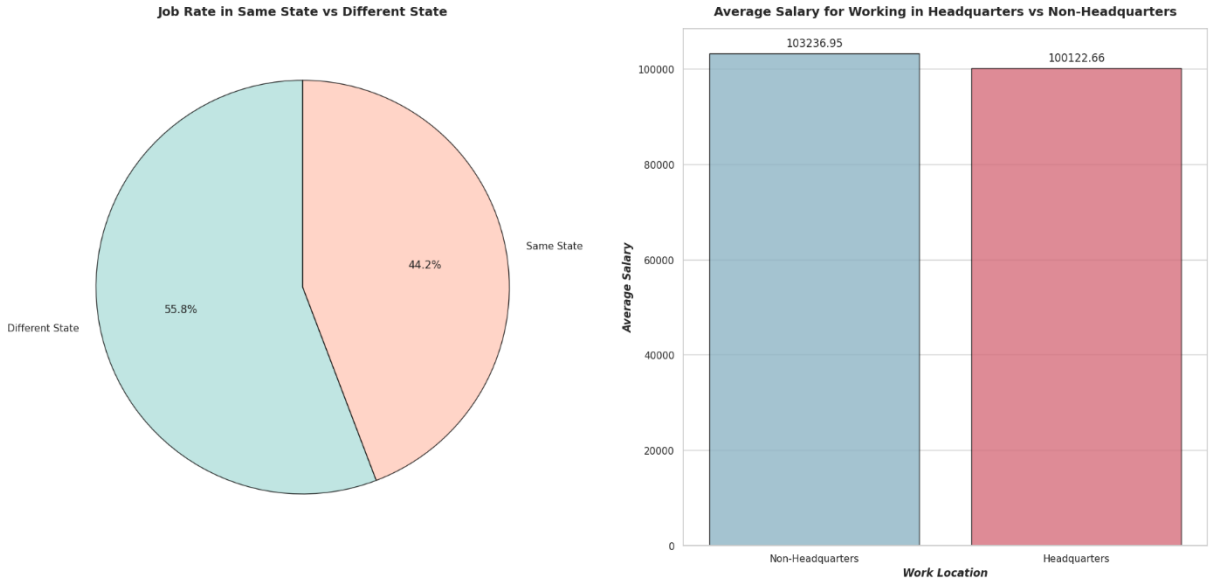
Việc giới hạn tìm kiếm công việc trong phạm vi một bang có thể là một trở ngại lớn đối với các ứng viên. city là một feature lý tưởng để thể hiện rõ những nơi đang có phân phối lương cao nhất ở Mỹ.



Hình 5-6. Phân phối lương trung bình theo 10 thành phố hàng đầu

Xu hướng và linh hoạt địa điểm làm việc:

- Tỷ lệ công việc ở bang khác cao hơn cho thấy xu hướng linh hoạt trong địa điểm làm việc.
- Lương ngoài trụ sở cao hơn có thể do yếu tố cạnh tranh hoặc điều chỉnh theo khu vực.
- Xu hướng làm việc từ xa và mở rộng địa lý ảnh hưởng đến cách các công ty tổ chức công việc và trả lương.



Hình 5-7. Tỉ lệ việc làm ở cùng bang, khác bang và Lương trung bình khi làm việc ở trụ sở chính và không làm việc ở trụ sở chính

Kết luận

Mức lương cao minh chứng cho một thị trường yêu cầu cao, dẫn đến xu hướng con người đổ vào học ngành nghề và lĩnh vực đó nhiều hơn.

Lĩnh vực khoa học dữ liệu mang lại nhiều cơ hội việc làm với mức lương hấp dẫn. Để tối đa hóa tiềm năng thu nhập, cần phát triển kỹ năng chuyên môn, tích lũy kinh nghiệm, và theo dõi xu hướng công nghệ mới nhất.

Đối với cá nhân, cần đầu tư vào học tập và rèn luyện kỹ năng chuyên môn, xây dựng portfolio, và tham gia các dự án thực tế để tích lũy kinh nghiệm. Đối với doanh nghiệp, cần xây dựng chiến lược tuyển dụng và giữ chân nhân tài hiệu quả, bao gồm chính sách lương thưởng cạnh tranh và môi trường làm việc hấp dẫn.

5.2. Kết quả của Mô hình

Chúng tôi đã thực hiện thử nghiệm với nhiều mô hình hồi quy khác nhau. Kết quả cho thấy XGBoost là mô hình tốt nhất cho bài toán dự đoán lương. Dưới đây là các kết quả chi tiết:

Bảng 5-1. Hiệu suất của các mô hình

STT	Mô hình	MSE	MAE	R^2
0	XGBoost	2.687e+08	1.065e+04	8.286e-01
1	RandomForest	2.936e+08	1.127e+04	8.127e-01
2	HistogramGradientBoosting	4.009e+08	1.456e+04	7.443e-01
3	DecisionTree	4.605e+08	1.024e+04	7.062e-01
4	KNeighbors	8.300e+08	2.173e+04	4.706e-01
5	SVR	1.616e+09	3.117e+04	-3.078e-02
6	LinearRegression	2.511e+33	1.946e+16	-1.602e+24

Hiệu suất của XGBoost: là mô hình có MSE thấp (2.687e+08) và MAE thấp (1.065e+04), cho thấy nó dự đoán chính xác hơn so với các mô hình khác. R^2 = 8.286e-01, thể hiện khả năng giải thích biến phụ thuộc rất tốt.

Vì XGBoost đang là mô hình tốt nhất và XGBoost có nhiều tham số ảnh hưởng đến hiệu xuất của nó nên chúng ta cần tinh chỉnh để tối ưu hóa siêu tham số bằng RandomizedSearchCV.

Kết quả tuning cho XGBoost như sau:

- Điểm số tốt nhất: 0.727
- Các tham số tốt nhất:
  + reg_lambda: 1.000e-05
  + reg_alpha: 1.000e+00
  + n_estimators: 4.500e+02
  + max_depth: 1.800e+01
  + learning_rate: 1.000e-01
  + gamma: 4.000e-01
  + colsample_bytree: 4.000e-01

Chúng tôi đã sử dụng các tham số tốt nhất (reg_lambda, reg_alpha, n_estimators, max_depth, learning_rate, gamma, colsample_bytree) để huấn luyện lại mô hình XGBoost và cho thấy kết quả được cải thiện đáng kể:

- MSE (248,466,344.827) và MAE (7,420.090) đều giảm, cho thấy độ chính xác của dự đoán tăng lên.
- $R^2 = 0.842$, mô hình hiện đã giải thích được 84.2% phương sai trong dữ liệu.

Sử dụng Cross Validation (K-Fold) để kiểm tra độ ổn định và tránh overfitting. Kết quả thu được như sau:

Bảng 5-2. Kết quả sau khi sử dụng Cross Validation (K-Fold)

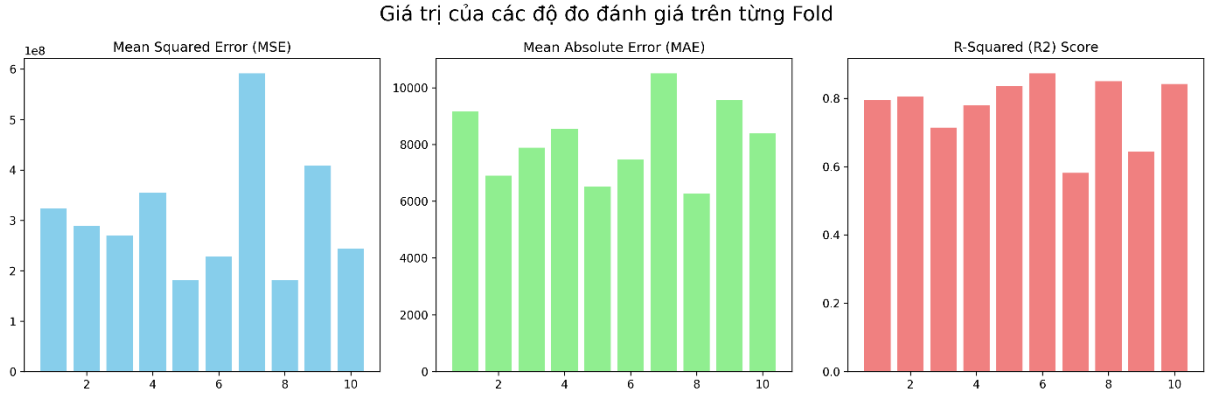
Fold	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R2) Score
Fold 1	323,509,244.889	9,170.349	0.795
Fold 2	288,311,970.653	6,899.549	0.805
Fold 3	270,256,278.143	7,887.944	0.714
Fold 4	355,090,927.602	8,544.482	0.780
Fold 5	181,863,423.341	6,508.935	0.836

Fold 6	228,430,120.322	7,472.681	0.873
Fold 7	591,616,426.198	10,506.991	0.581
Fold 8	180,901,631.391	6,261.799	0.850
Fold 9	407,922,399.500	9,569.530	0.644
Fold 10	243,493,054.242	8,405.791	0.842
Mean	307,139,547.628	8,122.805	0.772

Kết quả cho thấy Cross-Validation khá tốt, nhưng độ dao động giữa các fold cho thấy mô hình có thể nhạy cảm với dữ liệu và cần thêm các cải thiện:

- Mean MSE (307,139,547.628) và Mean MAE (8,122.805) cho thấy mô hình hoạt động tốt trong hầu hết các fold. Tuy nhiên, có sự dao động đáng kể giữa chúng, chẳng hạn MSE dao động từ 181,863,423.341 (fold 5) đến 591,616,426.198 (fold 7).
- R^2 dao động từ 0.581 (Fold 7) đến 0.873 (Fold 6).
- Fold 7 có hiệu suất kém nhất (MSE cao, R^2 thấp), cho thấy dữ liệu trong fold này có thể chứa nhiều nhiễu hoặc khác biệt đáng kể so với các fold khác.

Các biểu đồ dưới đây trực quan hóa giá trị của các độ đo đánh giá trên từng Fold:



Hình 5-8. Giá trị của các độ đo đánh giá trên từng fold

Khả năng tránh overfitting: với Mean R^2 = 0.772, mô hình giải thích được khoảng 77.2% phương sai trong dữ liệu trung bình qua các fold. Đây là mức độ tương đối ổn định và không có dấu hiệu rõ ràng của overfitting.

6. CHỈNH SỬA SAU BÁO CÁO

Sau khi nhận được phản hồi từ buổi báo cáo, chúng tôi đã tiến hành chỉnh sửa về hình thức trình bày để đảm bảo báo cáo rõ ràng và dễ theo dõi hơn. Phần giới thiệu đã được tách thành hai đoạn riêng biệt theo yêu cầu. Đồng thời, cách xưng hô được chuyển sang "chúng tôi" nhằm tạo sự chuyên nghiệp và nhất quán trong toàn bộ tài liệu. Ngoài ra, các kết quả được minh họa trong báo cáo không còn lấy trực tiếp từ ảnh chụp màn hình của mã nguồn mà đã được viết lại dưới dạng bảng hoặc đồ họa phù hợp.

Chúng tôi cũng đã bổ sung các câu dẫn ngắn gọn trước mỗi hình minh họa để giải thích ý nghĩa và bối cảnh sử dụng, giúp người đọc hiểu rõ hơn mục đích của từng hình. Đặc biệt, chúng tôi đã bổ sung radar plot phần Kỹ năng yêu cầu cho từng vị trí công việc, đảm bảo báo cáo có tính trực quan cao hơn.

7. KẾT LUẬN

Qua quá trình nghiên cứu và phân tích, chúng tôi đã đạt được những kết quả đáng kể trong việc dự đoán mức lương ngành Khoa học Dữ liệu. Bằng việc sử dụng thư viện scikit-learn và XGBoost, chúng tôi đã xây dựng mô hình dự đoán với hiệu suất cao nhất, đạt $R^2 = 0.842$ sau khi tối ưu hóa siêu tham số. Điều này cho thấy mô hình có khả năng giải thích 84.2% phương sai trong dữ liệu, minh chứng cho độ chính xác và sự tin cậy của nó.

Cross-validation với K-Fold cho thấy độ ổn định của mô hình, mặc dù có sự dao động giữa các fold, đặc biệt là fold 7. Tuy nhiên, Mean $R^2 = 0.772$ cho thấy mô hình không có dấu hiệu overfitting rõ ràng. Kết quả này khẳng định khả năng ứng dụng của mô hình trong nhiều tình huống khác nhau.

Điểm mạnh của nghiên cứu nằm ở việc sử dụng các kỹ thuật tiên tiến để tối ưu hóa mô hình và phân tích dữ liệu một cách toàn diện. Công thức và phương pháp của chúng tôi có thể áp dụng rộng rãi trong các bài toán dự đoán khác, mở ra nhiều cơ hội nghiên cứu và cải tiến trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Data Science Jobs & Salaries 2024. Link: https://www.kaggle.com/datasets/fahadrehman07/data-science-jobs-and-salary-glassdoor (Ngày truy cập: 11/11/2024)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Phú Tài	− Phân tích thăm dò dữ liệu: vẽ và thiết kế đồ thị − Đánh giá mô hình − Làm slide
2	Mai Văn Tân	− Phân tích thăm dò dữ liệu: mô tả các đồ thị − Lựa chọn và huấn luyện mô hình − Viết báo cáo
3	Nguyễn Công Trúc	− Tiền xử lý bộ dữ liệu − Phân tích thăm dò dữ liệu − Đánh giá mô hình − Viết báo cáo
4	Trần Lê Nguyên Trung	− Phân tích thăm dò dữ liệu − Lựa chọn và huấn luyện mô hình − Viết báo cáo