BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRẦN VĂN NGHIỆP

SỬ DỤNG HỒI QUY TUYẾN TÍNH TRONG DỰ ĐOÁN MỨC LƯƠNG CÔNG VIỆC TRÊN QUẢNG CÁO TUYỂN DỤNG

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 03 năm 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRẦN VĂN NGHIỆP

SỬ DỤNG HỒI QUY TUYẾN TÍNH TRONG DỰ ĐOÁN MỨC LƯƠNG CÔNG VIỆC TRÊN QUẢNG CÁO TUYỂN DỤNG

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS. TRẦN ĐỨC KHÁNH

TP. HỒ CHÍ MINH, tháng 03 năm 2016

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : TS. TRẦN ĐỨC KHÁNH
(Ghi rõ họ, tên, học hàm, học vị và chử ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 20 tháng 03 năm 2016.

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:
(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	**Họ và tên**	**Chức danh Hội đồng**
1	PGS.TSKH. Nguyễn Xuân Huy	Chủ tịch
2	PGS.TS. Vũ Đức Lung	Phản biện 1
3	TS. Hồ Đắc Nghĩa	Phản biện 2
4	TS. Cao Tùng Anh	Ủy viên
5	TS. Vũ Thanh Hiền	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận văn sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày 20 tháng 08 năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: TRẦN VĂN NGHIỆP — Giới tính: Nam

Ngày, tháng, năm sinh: 15/05/1987 — Nơi sinh: Cà Mau

Chuyên ngành: Công nghệ thông tin — MSHV: 1441860018

I- Tên đề tài:

"SỬ DỤNG HỒI QUY TUYẾN TÍNH TRONG DỰ ĐOÁN MỨC LƯƠNG CÔNG VIỆC TRÊN QUẢNG CÁO TUYỂN DỤNG"

II- Nhiệm vụ và nội dung:

- Tìm hiểu về học máy thống kê, quy trình khai thác dữ liệu, phân tích thống kê.
- Xây dựng mô hình dự đoán mức lương trên quảng cáo tuyển dụng ứng dụng phương pháp phân tích hồi quy.
- Đánh giá mô hình dự đoán ứng dụng các phương pháp đánh giá mô hình.

III- Ngày giao nhiệm vụ: *20/08/2015*

IV- Ngày hoàn thành nhiệm vụ: *15/01/2016*

V- Cán bộ hướng dẫn: *TS. Trần Đức Khánh*

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

KHOA QUẢN LÝ CHUYÊN NGÀNH
(Họ tên và chữ ký)

TS. TRẦN ĐỨC KHÁNH

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Thầy TS. Trần Đức Khánh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình. Trường Đại Học Công Nghệ TP.HCM không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện.

Học viên thực hiện luận văn

TRẦN VĂN NGHIỆP

LỜI CÁM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Trong suốt thời gian từ khi bắt đầu học tập tại trường đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý Thầy Cô, gia đình và bạn bè. Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý Thầy Cô ở Khoa Công Nghệ Thông Tin – Trường Đại Học Công Nghệ TP.HCM đã cùng với tri thức và tâm huyết của mình để truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Và đặc biệt, trong học kỳ này. Nếu không có những lời hướng dẫn, dạy bảo của các thầy cô thì em nghĩ bài luận văn này của em rất khó có thể hoàn thiện được. Bài luận văn thực hiện trong khoảng thời gian 6 tháng. Bước đầu của em còn rất hạn chế và còn nhiều bỡ ngỡ. Do vậy, em gặp rất nhiều khó khăn trong giai đoạn đầu làm luận văn. Nhưng với sự dìu dắt hướng dẫn tận tình của thầy TS. TRẦN ĐỨC KHÁNH em đã dần làm quen với việc nghiên cứu và hoàn thiện bài luận văn này.

Em xin gởi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của Trường Đại Học Công Nghệ TP.HCM, đặc biệt là các thầy cô Khoa Công Nghệ Thông Tin của trường đã tạo điều kiện cho em để em có thể hoàn thành tốt bài luận văn này. Và em cũng xin chân thành cám ơn các bạn học cùng khóa đã nhiệt tình đóng góp ý kiến để em hoàn thành tốt bài luận văn của em.

Trong quá trình làm bài luận văn, khó tránh khỏi những sai sót, rất mong quý Thầy, Cô bỏ qua. Đồng thời do trình độ lý luận cũng như kinh nghiệm thực tiễn còn hạn chế nên bài luận văn không thể tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp của Thầy, Cô để em học thêm được nhiều kinh nghiệm để tiếp tục hoàn thành tốt những nghiên cứu sắp tới.

Em xin chân thành cảm ơn!

TRẦN VĂN NGHIỆP

TÓM TẮT

Phân tích hồi quy là một phương pháp thống kê nhằm tìm ra mối liên hệ giữa một biến phụ thuộc (thường ký hiệu là Y) và một loạt các biến đổi khác (được biết đến như là các biến độc lập). Mối liên hệ này được mô tả trên hình thức của một phương trình đường thẳng (phương trình hồi quy) dựa trên các đặc trưng của dữ liệu cần phân tích. Phân tích hồi quy thường được sử dụng để xác định có bao nhiêu yếu tố cụ thể như giá của một mặt hàng, lãi suất, các ngành công nghiệp, ngành nghề đặc biệt ảnh hưởng đến sự biến động về lương của công việc trên quảng cáo tuyển dụng. Trong phạm vi đề tài này là ứng dụng phương pháp hồi quy để dự đoán mức lương của công việc trên các quảng cáo tuyển dụng. Nỗ lực tìm mối liên hệ giữa các đặc trưng ảnh hưởng đến mức lương công việc như: nhóm công việc, loại công việc, loại hợp đồng, địa điểm làm việc… Từ đó đưa ra mô hình dự đoán tối ưu nhất áp dụng các phương pháp phân tích hồi quy đơn giản, hồi quy đa biến, phân tích phương sai, phân tích thành phần cũng như phương pháp đánh giá mô hình dựa trên dữ liệu quảng cáo tuyển dụng được cung cấp bởi Kaggle (https://www.kaggle.com/c/job-salary-prediction/data).

ABSTRACT

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. Regression is often used to determine how much specific factors such as the price of a commodity, interest rates, particular industries or sectors influence the price movement of an asset.

On this thesis, I am using regression for predicting the salary of the job on job advertisement. Try to find the relationship between features that impacted to the job salary such as: Job Category, Contract Time, Contract Type, Location, and so on. Base on that points we generate a model that help employer or job seeker can forecast the rank salary of the job by applying simple linear regression, multiple regression, variables analysis, model evaluation on the Job Advertisement is provived by Kaggle (https://www.kaggle.com/c/job-salary-prediction/data).

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Cụm từ nguyên	Ý nghĩa
CRISP-DM	Cross-Industry Standard Process for Data Mining	Quy trình khai thác dữ liệu
KDD	Knowledge Discovery in Databases	
P	P-Value	Giá trị xác suất (p-value)
R		Công cụ phân tích R
SSR	Sum of Square Residual	Tổng bình phương hồi quy
SST	Total Sum of Squares	Tổng biến động của y
SSE	Sum of Square Error	Tổng sai số bình phương
MSE	Mean Square Error	Sai số trung bình bình phương
ANOVA	Analysis of Variables	Phân tích biến
ME	Mean Error	Sai số trung bình
MAE	Mean Absolute Error	Sai số trung bình tuyệt đối
MPE	Mean Percentage Error	Trung bình sai số theo phần trăm
MAPE	Mean Absolute Percentage Error	Sai số trung bình tuyệt đối theo phần trăm
MASE	Mean Absolute Scaled Error	Tỷ lệ sai số trung bình tuyệt đối

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH ẢNH

CHƯƠNG 1: GIỚI THIỆU

1.1. Lý do chọn đề tài

Trong lĩnh vực tuyển dụng ngày nay, khoảng một nửa số công ty họ không công khai mức lương tuyển dụng trên các quảng cáo tuyển dụng. Với vai trò là một người tìm kiếm công việc tác giả cảm thấy rất khó khăn để làm sao biết được mức lương công việc của quảng cáo tuyển dụng mà tác giả quan tâm, liệu rằng mức lương nào là phù hợp hoặc không phù hợp với từng loại công việc trên quảng cáo tuyển dụng đó. Và với vai trò là nhà tuyển dụng tác giả muốn biết được hoặc tham khảo để có thể đưa ra mức lương hợp lý trên các quảng cáo tuyển dụng của doanh nghiệp mình. Do đó rất cần một giải pháp để mang lại nhiều thông tin hơn trong lĩnh vực này. Từ đó có thể giúp người tìm kiếm việc làm và nhà tuyển dụng ước lượng được mức lương của một công việc hoặc nhóm công việc nào đó là phù hợp hoặc không phù hợp, họ sẽ có những điều chỉnh hoặc sự chuẩn bị tốt hơn trong công tác tuyển dụng hoặc tìm kiếm việc làm. Với những khó khăn và nhu cầu như trên nên tác giả nghiên cứu lựa chọn hướng đề tài xây dựng mô hình dự đoán mức lương công việc trên quảng cáo tuyển dụng với tên đề tài là: "**SỬ DỤNG HỒI QUY TUYẾN TÍNH TRONG DỰ ĐOÁN MỨC LƯƠNG CÔNG VIỆC TRÊN QUẢNG CÁO TUYỂN DỤNG**" để nghiên cứu xây dựng một công cụ dự báo cho mức lương của bất kỳ quảng cáo tuyển dụng nào. Nhằm giúp người tìm việc cũng như các nhà tuyển dụng có thể dự đoán được mức lương phù hợp cho các vị trí công việc khác nhau.

1.2. Mục tiêu nghiên cứu

Luận văn tập trung nghiên cứu về các nghiệp vụ về quảng cáo việc làm, nghiên cứu các nhân tố ảnh hưởng đến mức lương công việc trên quảng cáo tuyển dụng. Những nhân tố đó có thể là nhóm công việc, loại công việc, loại hợp đồng hoặc là địa điểm làm việc mà một quảng cáo tuyển dụng cần có. Từ đó xây dựng mô hình dự đoán dựa trên những nhân tố ảnh hưởng đó để đưa ra kết quả dự đoán với độ tin cậy và độ chính xác cao nhất.

Để giải quyết vấn đề đó luận văn sử dụng giải pháp học máy thống kê mà cụ thể là phân tích hồi quy ứng dụng những kỹ thuật phân tích liên quan theo quy trình khai phá dữ liệu chuẩn công nghiệp CRISP-DM để có thể xây dựng mô hình dự đoán đạt được kết quả tốt nhất.

1.3. Đối tượng nghiên cứu

Về nghiệp vụ, đối tượng nghiên cứu là những vấn đề liên quan đến quảng cáo tuyển dụng. Các nhân tố liên quan đến quảng cáo tuyển dụng như: nhóm công việc, loại công việc, loại hợp đồng, địa điểm làm việc, công ty tuyển dụng, nguồn quảng cáo tuyển dụng, vị trí tuyển dụng.v.v.

Về dữ liệu, đối tượng nghiên cứu liên quan đến quảng cáo việc làm được công bố tại Anh được cung cấp bởi Kaggle [5]: (https://www.kaggle.com/c/job-salary-prediction/data).

Về kỹ thuật, đối tượng nghiên cứu là các lý thuyết về học máy thống kê và khai thác dữ liệu, cụ thể là phân tích hồi quy dựa theo tài liệu tham khảo [1] trong phần tài liệu tham khảo của luận văn này.

Về quy trình, đối tượng nghiên cứu là quy trình khai thác dữ liệu chuẩn công nghiệp CRISP-DM dựa theo tài liệu tham khảo IBM SPSS Modeler CRISP-DM Guide [2] và CRISP-DM 1.0 [3] trong phần tài liệu tham khảo.

Về công cụ, đối tượng nghiên cứu là công cụ phân tích dữ liệu R. là một phần mềm mã nguồn mở phát triển và cung cấp miễn phí bởi CRAN (https://cran.r-project.org/).

1.4. Tổng quan nghiên cứu

Bài toán dự đoán mức lương trên quảng cáo tuyển dụng được tổ chức Kaggle đưa ra vào tháng 3 năm 2013. Dựa trên yêu cầu của Adzuna (một công ty về quảng cáo tuyển dụng ở Anh - https://www.adzuna.co.uk/) là muốn xây dựng một ứng dụng dự đoán mức lương của bất kỳ quảng cáo công việc ở Anh. Từ đó, họ có thể cải thiện rất lớn sự trải nghiệm của người dùng trong tìm kiếm việc làm, giúp nhà tuyển dụng

và người tìm việc tìm ra các giá trị thị trường của các vị trí việc làm khác nhau. Vì khoảng một nửa trong số các quảng cáo công việc họ không cho biết mức lương công khai. Vì thế Adzuna cần phát triển việc dự đoán mức lương để mang lại sự minh bạch cho thị trường quan trọng này.

Trước đó, có một vài công trình nghiên cứu liên quan đến thu nhập, mức lương. Nhưng chủ yếu các nghiên cứu chỉ tập trung nghiên cứu thu nhập và mức lương cho từng cá nhân cụ thể ví dụ như đề tài "**Dự báo và xác định thu nhập cố định**" [14] của Ramses H Abul Naga, University of Lausanne, 1997. Trong luận văn này là tập trung nghiên cứu phương pháp dự đoán mức lương cho từng công việc cụ thể được quảng cáo trên các quảng cáo tuyển dụng.

Trên cơ sở đó, nhiều cách tiếp cận được đưa ra mà cụ thể của một tác giả ẩn danh [15], họ đưa ra cách tiếp cận là kết hợp tuyến tính với phân loại rừng ngẫu nhiên (random forest) đưa ra kết quả khá ấn tượng là **4933**. Riêng với cách tiếp cận của luận văn này chỉ sử dụng mô hình tuyến tính và đi sâu khai thác, phân loại các biến "địa điểm làm việc", "tiêu đề công việc", và "mô tả công việc" từ đó xây dựng mô hình kết hợp tất cả các biến độc lập đó, tận dụng mọi dữ liệu có sẵn để đưa ra mô hình dự đoán tốt nhất.

1.5. Bố cục luận văn

Chương 1: Giới thiệu

Nội dung chương 1 giới thiệu bài toán dự đoán mức lương, lý do chọn đề tài, mục tiêu nghiên cứu và đối tượng nghiên cứu.

Chương 2: Cơ sở lý thuyết

Trình bày quy trình khai thác dữ liệu CRISP-DM. Các lý thuyết về phân tích hồi quy, kỹ thuật xây dựng mô hình, kỹ thuật đánh giá mô hình, kỹ thuật kiểm tra độ tin cậy của mô hình và công cụ phân tích R.

Chương 3: Ứng dụng hồi quy trong phân tích dự đoán mức lương

Trình bày việc xây dựng mô hình dự đoán mức lương trên quảng cáo tuyển dụng dựa trên quy trình khai thác dữ liệu CRISP-DM với các giai đoạn: tìm hiểu nghiệp vụ, tìm hiểu dữ liệu, chuẩn bị dữ liệu, mô hình hóa, và đánh giá mô hình.

Chương 4: Kết luận và hướng phát triển

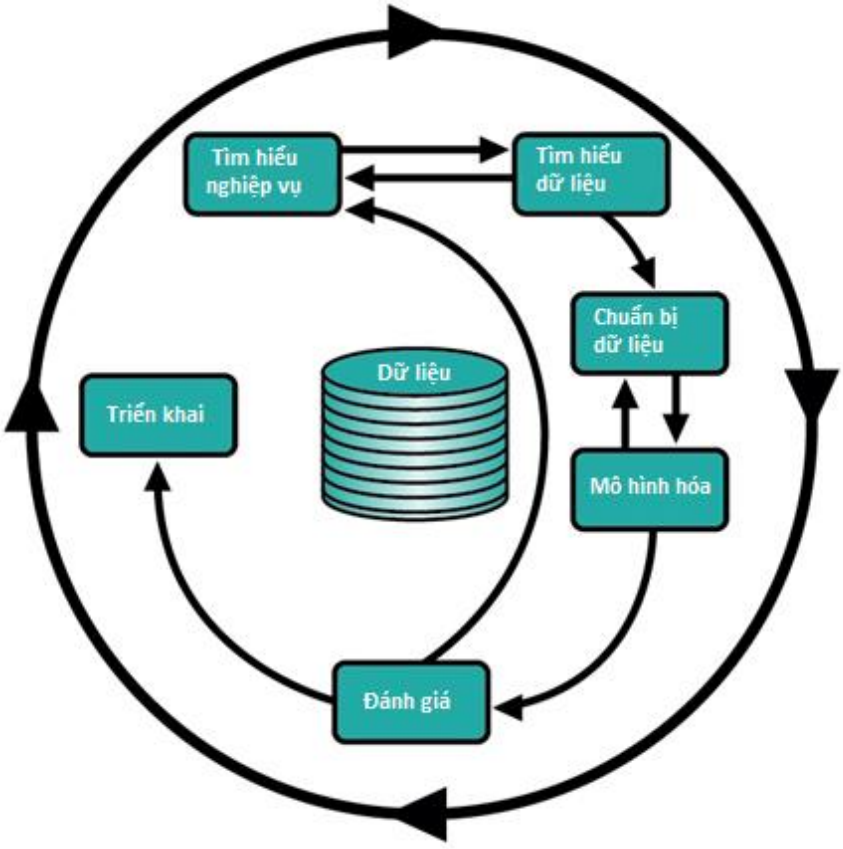
Tổng kết lại những nội dung chính của luận văn và trình bày hướng phát triển trong tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Mô hình khai thác dữ liệu CRISP-DM

Quá trình khai thác dữ liệu điển hình có thể trở nên phức tạp. Có rất nhiều thứ để theo dõi - những vấn đề phức tạp trong kinh doanh, với nhiều nguồn dữ liệu, thay đổi chất lượng dữ liệu qua các nguồn dữ liệu, một loạt các kỹ thuật khai thác dữ liệu, với nhiều cách khác nhau để việc khai thác dữ liệu được thành công.

Mô hình khai thác dữ liệu được đề nghị là mô hình CRISP - DM viết tắt của Cross - Industry Standard Process for Data Mining. Mô hình này được thiết kế như một mô hình chung có thể được áp dụng cho một loạt các ngành công nghiệp và các vấn đề kinh doanh.



Hình 1: Mô hình CRISP-DM

Mô hình quy trình của CRISP - DM bao gồm 6 giai đoạn giải quyết các vấn đề chính trong khai thác dữ liệu. Sáu giai đoạn kết hợp với nhau như một quá trình mang tính chu kỳ.

Bảng 1: Công việc từng giai đoạn trong CRISP-DM

GIAI ĐOẠN VÀ NHIỆM VỤ					
Tìm hiểu nghiệp vụ	**Tìm hiểu dữ liệu**	**Chuẩn bị dữ liệu**	**Mô hình hóa**	**Đánh giá mô hình**	**Triển khai ứng dụng**
Xác định mục tiêu: *Tổng quan nghiệp vụ* *Mục tiêu* *Tiêu chí để thành công* **Đánh giá tình huống:** *Inventory of resourse* *Tài liệu* *Giả định* *Những rang buộc* *Rủi ro và những điều không lường trước* *Thuật ngữ* *Chi phí và lợi nhuận* **Xác định mục đích khai thác dữ liệu:** *Mục tiêu khai phá dữ liệu* *Tiêu chí thành*	**Thu thập dữ liệu ban đầu:** *Báo cáo thu thập dữ liệu ban đầu* **Mô tả dữ liệu:** *Báo cáo mô tả dữ liệu* **Nghiên cứu dữ liệu:** *Báo cáo nghiên cứu dữ liệu* **Kiểm tra chất lượng dữ liệu:** *Báo cáo chất lượng dữ liệu*	*Tập dữ liệu* *Mô tả tập dữ liệu* **Lựa chọn dữ liệu:** *Phân tích nguyên nhân cho sự bao hàm hoặc loại trừ dữ liệu* **Làm sạch dữ liệu:** *Báo cáo làm sạch dữ liệu* **Xây dựng dữ liệu:** *Đưa ra các thuộc tính* *Tạo ra dòng dữ liệu* **Tích hợp dữ liệu:** *Hợp nhất dữ liệu* **Định dạng dữ liệu:** *Định dạng lại dữ liệu*	**Lựa chọn kỹ thuật mô hình hóa:** *Kỹ thuật mô hình hóa* *Giả định mô hình hóa* **Tạo mẫu kiểm tra:** *Mẫu kiểm tra* **Xây dựng mô hình:** *Thiết lập tham số cho mô hình* *Mô tả mô hình* **Kiểm tra mô hình:** *Đánh giá mô hình* *Duyệt lại các thiết lập tham số*	**Đánh giá kết quả:** *Đánh giá kết quả khai phá dữ liệu dựa trên các tiêu chí thành công của dự án* *Chấp thuận mô hình đã đưa ra* **Duyệt lại quy trình:** *Xem xét lại quy trình* **Đưa ra những bước tiếp theo:** *Đưa ra danh sách những hành động, quyết định tiếp theo*	**Lập kế hoạch triển khai:** *Kế hoạch triển khai* **Lập kế hoạch theo dõi và bảo trì:** *Kế hoạch theo dõi và bảo trì* **Đưa ra báo cáo hoàn tất:** *Báo cáo hoàn tất* *Trình bày hoàn tất* **Duyệt lại dự án:** *Tài liệu sử dụng*

công trong khai phá dữ liệu **Lập kế hoạch cho dự án:** *Kế hoạch* *Đánh giá ban đầu về công cụ và kỹ thuật*					

2.1.1. Tìm hiểu nghiệp vụ

Đây có thể là giai đoạn quan trọng nhất của việc khai thác dữ liệu. Tìm hiểu nghiệp vụ bao gồm việc xác định mục tiêu kinh doanh, đánh giá tình hình, xác định mục tiêu khai thác dữ liệu. Hoạt động trong giai đoạn này bao gồm:

- Xác định mục tiêu kinh doanh và tiêu chí thành công.
- Thực hiện đánh giá thực trạng (nguồn lực, khó khăn, giả định, rủi ro, chi phí và lợi ích).
- Xác định các mục tiêu khai thác dữ liệu.

2.1.2. Tìm hiểu dữ liệu

Nguồn dữ liệu cung cấp nguyên liệu cho việc khai thác dữ liệu. Sự cần thiết ở giai đoạn này phải hiểu biết các nguồn dữ liệu của một doanh nghiệp đang có và đặc điểm của dữ liệu. Bao gồm việc thu thập dữ liệu ban đầu, mô tả dữ liệu, khai thác dữ liệu và kiểm tra chất lượng dữ liệu.

2.1.3. Chuẩn bị dữ liệu

Sau khi chia ra từng loại dữ liệu, đến giai đoạn cần chuẩn bị dữ liệu để khai thác. Việc chuẩn bị bao gồm việc lựa chọn, làm sạch, xây dựng, tích hợp và định dạng dữ liệu. Những nhiệm vụ này sẽ được thực hiện nhiều lần và không có bất kỳ thứ tự quy định nào.

Những nhiệm vụ này có thể sẽ tốn nhiều thời gian nhưng là bước quan trọng cho sự thành công của việc khai thác dữ liệu. Chuẩn bị dữ liệu bao gồm:

- Giải nén dữ liệu
- Liên kết các bảng với nhau trong một cơ sở dữ liệu hoặc trong mô hình.
- Kết hợp các tập tin dữ liệu từ hệ thống khác nhau.

- Xác định giá trị bị mất, những dữ liệu không chính xác.
- Lựa chọn dữ liệu.
- Tái cấu trúc dữ liệu thành dạng phân tích yêu cầu.
- Chuyển đổi các lĩnh vực có liên quan.

2.1.4. Mô hình hóa

Giai đoạn này liên quan đến việc lựa chọn kỹ thuật tạo ra các thiết kế thử nghiệm, xây dựng và đánh giá mô hình. Mô hình hóa là một quá trình lặp đi lặp lại, như thế mới có được một mô hình thống kê chuẩn. Sử dụng nhiều mô hình để đưa ra các dự đoán.

2.1.5. Đánh giá

Một khi đã chọn được một mô hình chuẩn, chuẩn bị bước qua giai đoạn đánh giá kết quả khai thác dữ liệu có thể giúp đạt được mục tiêu kinh doanh. Trước khi viết báo cáo tổng kết và triển khai mô hình, đều quan trọng là đánh giá sâu hơn về mô hình và xem xét các bước thực hiện xây dựng các mô hình để chắc chắn nó đạt được mục tiêu kinh doanh.

2.1.6. Triển khai

Hoàn tất việc xây dựng mô hình không có nghĩa việc hoàn thành dự án. Cần phải thực hiện sử dụng các mô hình đã tạo ra. Đó là giai đoạn triển khai đơn giản là các báo cáo, phức tạp hơn là những ứng dụng dựa trên mô hình đã xây dựng được.

2.2. Hồi quy tuyến tính đơn

Phân tích hồi quy là nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc hay còn gọi là biến được giải thích) vào một biến hay nhiều biến khác (biến độc lập hay còn gọi là biến giải thích) với ý tưởng cơ bản là ước lượng (hay dự đoán) giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị đã biết của biến độc lập.

2.2.1. Phương trình hồi quy tuyến tính đơn

Đặt (x_1, y_1), (x_2, y_2),…, (x_n, y_n) là mẫu gồm n cặp quan sát trên đường hồi quy tổng thể:

$y = \alpha + \beta x_1 + \varepsilon_1$ [1]

Theo phương pháp bình phương bé nhất thì ước lượng các hệ số α và β là các giá trị a và b sao cho tổng bình phương sai số của phương trình sau đây là bé nhất:

SS = $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ [1]

Các hệ số a và b được tính như sau:

b = $\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - nx^{-2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_1 - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ [1]

Suy ra: a = $\bar{y}$ - b $\bar{x}$

Và phương trình hồi quy tuyến tính mẫu của y trên x là: y = a + b x

2.2.2. Khoảng tin cậy và kiểm định giả thuyết trong hồi quy đơn

Giả sử đường hồi quy tuyến tính có dạng: $y_i = \alpha + \beta x_1 + \varepsilon_1$

Và đặt σ_ε^2 là phương sai của sai số và được ước lượng từ công thức sau:

$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SEE}{n-2}$ [1]

Đặt b là ước lượng mẫu của β thì phương sai của b là

$\sigma_b^2 = \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma_e^2}{\sum x_i^2 - n\bar{x}^2}$ [1]

$\rightarrow$Ước lượng không chênh lệch của σ_ε^2 được xác định bởi:

$\sigma_b^2 = S_b^2 = \frac{S_e^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma_e^2}{\sum x_i^2 - n\bar{x}^2}$

Giả sử, sai số hồi quy (ε_1) có phân phối chuẩn thì ngẫu nhiên (t) dùng để kiểm định giả thuyết về β và ước lượng khoảng tin cậy của β được tính như sau:

t = $\frac{b - \beta}{S_b}$

Và khoảng tin cậy 100 (1 - α)% cho β là:

$b - t_{n-2,\frac{\alpha}{2}} S_b < \beta < b + t_{n-2,\frac{\alpha}{2}} S_b$

Trong đó, $t_{n-2,\frac{\alpha}{2}}$ là một số sao cho P($t_{n-2} > t_{n-2,\frac{\alpha}{2}}$) = $\frac{\alpha}{2}$

2.2.3. Kiểm định tham số hồi quy tổng thể (β)

Ở mức ý nghĩa α, giả thuyết H_0 có thể được kiểm định dưới các trường hợp:

Đặt giả thuyết: $\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta > \beta_0 \end{cases}$ (1) $\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta < \beta_0 \end{cases}$ (2) $\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \# \beta_0 \end{cases}$ (3)

Giá trị kiểm định: t = $\frac{b-\beta_0}{S_b}$

Quyết định bác bỏ giả thuyết H_0 khi: t < $t_{n-2,\alpha}$ t < $t_{n-2,\alpha}$ $\begin{cases} t > t_{n-2,\frac{\alpha}{2}} \\ t < -t_{n-2,\frac{\alpha}{2}} \end{cases}$

Giả thuyết H_0: β = 0

2.2.4. Phân tích phương sai hồi quy

**** Hệ số xác định***: R^2 là hệ số nhằm xác định mức độ quan hệ giữa X và Y có quan hệ hay không hoặc bao nhiêu phần trăm sự biến thiên của Y có thể giải thích bởi sự phụ thuộc tuyến tính của Y vào X.

Giá trị thực tế $y_i = a + bx_1 + e_1$

Giá trị dự đoán theo phương trình hồi quy: $y = a + bx_1$

$\Rightarrow y_1 = y_1 + e_1$

Vậy e_1 là sự khác biệt giữa giá trị thực tế với giá trị dự đoán của phương trình hồi quy tuyến tính. Như vậy e_1 thể hiện phần biến thiên của Y không thể giải thích bởi mối quan hệ tuyến tính giữ Y và X.

Ta có:

$\sum(y_i - \bar{y})^2 = \sum(y_i - \bar{y})^2 + \sum e_i^2$

Hay SST = SSR+ SSE

SSR càng lớn thì mô hình hồi quy tuyến tính càng có độ tin cậy cao trong việc giải thích sự biến động của Y

Hệ số xác định $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ là phần trăm biến động của Y được giải thích bởi mối quan hệ tuyến tính của Y vào X.

**** Phân tích phương sai***

Trong ước lượng các tham số của mô hình hồi quy tuyến tính đơn theo phương pháp bình quân nhỏ nhất, có thể chứng minh được rằng:

$\sum (y_i - y_{tb})^2 = \sum (y_i - \tilde{y}_i)^2 + \sum (\tilde{y}_i - y_{tb})^2$

Trong đó:

$\sum (y_i - y_{tb})^2$ = SST là tổng biến động của y.

$\sum (\tilde{y}_i - y_{tb})^2$ = SSR là tổng bình phương hồi quy, là đại lượng biến động của y được giải thích bởi đường hồi quy.

$\sum (y_i - \tilde{y}_i)^2$ = SSE là phần biến động còn lại hay còn gọi là dư số, là đại lượng biến động tổng gộp của nguồn biến động do các nhân tố khác gây ra mà không hiện diện trong mô hình hồi quy và phần biến động ngẫu nhiên.

- SSR càng lớn thì mô hình hồi quy càng có độ tin cậy cao trong việc giải thích biến động của y.
- Hệ số xác định: r^2 = SSR/ SST = 1 – (SSE/ SST) là phần trăm biến động của y được giải thích bởi mối quan hệ tuyến tính của y đối với x.
- Số thống kê F = SSR/ [SSE/ (n-2)] = MSR/MSE có phân phối F và thường được dùng để kiểm định mức ý nghĩa của mô hình hồi quy. F càng lớn mô hình càng có ý nghĩa.

Các nguồn biến động của hồi quy tuyến tính đơn được tóm tắt trong bảng phân tích phương sai hồi quy như sau:

Bảng 2: Biến động của hồi quy tuyến tính

Nguồn biến động	Độ tự do (d.f)	Tổng bình phương (SS)	Bình phương trung bình (MS)
Do hồi quy Dư số	1 (n-2)	SSR=$\sum (\tilde{y}_i - y_{tb})^2$ SSE=$\sum (y_i - \tilde{y}_i)^2$	SSE/(n-2)

Tổng cộng	(n-1)	SST= $\sum (y_i - y_{tb})^2$	SST/(n-1)

2.2.5. Dự báo trong phương pháp hồi quy tuyến tính đơn

Ước lượng khoảng giá trị thực của y_{n+1} với độ tin cậy (1 - α)

$$y \pm t_{n-2,\frac{\alpha}{2}} S_e \sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^{-2}}}$$

Ước lượng khoảng giá trị trung bình của y_{n+1} với độ tin cậy (1 - α)

$$y \pm t_{n-2,\frac{\alpha}{2}} S_e \sqrt{\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^{-2}}}$$

2.3. Hồi quy tuyến tính đa biến

2.3.1. Mô hình hồi quy

Giả sử Y phụ thuộc vào k biến độc lập X1…X_k. Nếu giá trị của k biến độc lập X_1...X_k mô hình hồi quy tuyến tính đa biến có dạng :

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$ [1]

Giải thích biến:

- Y (biến phụ thuộc): chỉ tiêu phân tích: mức lương công việc
- α (biến độc lập): hệ số chặn phản ánh mức độ ảnh hưởng của các nhân tố khác đến chỉ tiêu phân tích.
- β : hệ số ước lượng, các hệ số hồi quy này phản ánh mức độ ảnh hưởng của từng nhân tố đến biến giải thích.

Nếu β >0 thì ảnh hưởng thuận và ngược lại là ảnh hưởng nghịch. β càng lớn thì sự ảnh hưởng đến chỉ tiêu phân tích càng mạnh.

- X_i các yếu tố ảnh hưởng đến mức lương. Với i chạy từ 1 đến k.
- U là sai số

2.3.2. Phương trình hồi quy

Gọi các hệ số a, b_1…b_k ước lượng cho $\alpha, \beta_1 \ldots \beta_k$ được xác định bởi phương pháp bình phương bé nhất. Phương trình hồi quy có dạng:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k.$$

Các tham số a, $b_1, b_2, \ldots, b_n$ có thể được ước lượng dễ dàng nhờ các phần mềm có sẵn các biến độc lập $X_1, X_2, \ldots, X_k$.

2.3.3. Phân tích phương sai hồi quy

☞ **Hệ số xác định:**

Hệ số xác định R^2 là nói lên tính chặt chẽ giữa biến phụ thuộc Y và các biến độc lập X_i, tức là nó thể hiện phần trăm biến thiên của Y có thể được giải thích bởi sự biến thiên của tất cả các biến X_i.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad 0 \le R^2 \le 1 \text{ [1]}$$

Trong đó:

$SSE = \sum_{i=1}^{n} e_i^2$: là phần biến động còn lại hay còn gọi là số dư

$SSR = \sum_{i=1}^{n} (y_i - \bar{y})^2$: là tổng bình phương hồi quy, là đại lượng biến động của y được giải thích bằng đường hồi quy

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: là tổng biến động của y.

SSR càng lớn thì mô hình hồi quy càng có độ tin cậy cao trong việc giải thích biến động y

☞ **Hệ số tương quan bội R**

R nói lên tính chặt chẽ của mối quan hệ giữa biến phụ thuộc (y) và các biến độc lập (x_i).

$R = \sqrt{R^2}$ $\quad (-1 \le R \le 1)$

☞ **Phân tích ANOVA hồi quy:**

Kiểm định sự phù hợp của mô hình (ANOVA):

Giá trị được dùng để kiểm định là giá trị F. Việc kiểm định này nhằm đảm bảo cho việc phù hợp của mô hình hồi quy tuyến tính mẫu với các hệ số tìm được vẫn có giá trị khi suy diễn ra mô hình thực cho tổng thể.

Để kiểm định sự phù hợp của mô hình hồi quy tổng thể, ta sử dụng Sig.F để làm căn cứ cho việc chấp nhận hay bác bỏ giả thiết

Sig.F < α : mô hình có ý nghĩa.

Sig.F > α : mô hình không có ý nghĩa.

2.3.4. Ước lượng khoảng tin cậy và kiểm định giả thuyết trong hồi quy đa biến

Mô hình hồi quy đa biến cho tổng thể có dạng:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + U$$

Đặt a, b_1, b_2, … ,b_k là những tham số được ước lượng cho tổng thể ; S_a , S_{b_1} , S_{b_2} , …, S_{b_k} là những độ lệch chuẩn đã ước lượng, và U coi phân phối chuẩn thì biến ngẫu nhiên t được tính như sau:

$$t_\alpha = \frac{a-\alpha}{S_a};\ t_{b_1} = \frac{b_1-\beta_1}{S_{b_1}} \text{ có độ tự do (n –k -1)}$$

Vì vậy, khoảng tin cậy 100(1-α)% cho các hệ số hồi quy β_1 được tính như sau:

$$b_1 - t_{n-k-1,\frac{\alpha}{2}} S_{b_i} < \beta < b_1 + t_{n-k-1,\frac{\alpha}{2}} S_{b_i}$$

$t_{n-k-1,\frac{\alpha}{2}}$ là một số sao cho (P$\left(t_{n-k-1} > t_{n-k-1,\frac{\alpha}{2}} \right)$.

2.4. Phương pháp đánh giá độ chính xác của mô hình

Đánh giá độ chính xác của bộ phân lớp rất quan trọng, bởi vì nó cho phép dự đoán được độ chính xác của các kết quả phân lớp những dữ liệu tương lai. Độ chính xác còn giúp so sánh các mô hình phân lớp khác nhau.Một số phương pháp đánh giá phổ biến bao gồm:

2.4.1. Phương pháp chia ngẫu nhiên

Trong phương pháp holdout, dữ liệu dưa ra được phân chia ngẫu nhiên thành 2 phần là: tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Thông thường 2/3 dữ liệu cấp cho tập dữ liệu huấn luyện, phần còn lại cho tập dữ liệu kiểm tra.

- Toàn bộ tập ví dụ *D* được chia thành 2 tập con **không giao nhau**
- Tập huấn luyện *D_train* – để huấn luyện hệ thống
- Tập kiểm thử *D_test* – để đánh giá hiệu năng của hệ thống đã học

→ *D* = *D_train* ∪ *D_test*, và thường là |*D_train*| >> |*D_test*|

▪ Các yêu cầu:

- Bất kỳ ví dụ nào thuộc vào tập kiểm thử *D_test* đều không được sử dụng trong quá trình huấn luyện hệ thống.

- Bất kỳ ví dụ nào được sử dụng trong giai đoạn huấn luyện hệ thống (i.e., thuộc vào *D_train*) đều không được sử dụng trong giai đoạn đánh giá hệ thống.

- Các ví dụ kiểm thử trong *D_test* cho phép một đánh giá không thiên vị đối với hiệu năng của hệ thống.

▪ Các lựa chọn thường gặp: |*D_train*|=(2/3).|*D*|, |*D_test*|=(1/3).|*D*|

▪ Phù hợp khi ta có tập ví dụ *D* có kích thước lớn.

2.4.2. Kiểm tra chéo K-Fold

▪ Để tránh việc trùng lặp giữa các tập kiểm thử (một số ví dụ cùng xuất hiện trong các tập kiểm thử khác nhau).

▪ Kiểm tra chéo *k*-fold

- Tập toàn bộ các ví dụ *D* được chia ngẫu nhiên thành *k* tập con **không giao nhau** (gọi là "*fold"*) có kích thước xấp xỉ nhau.

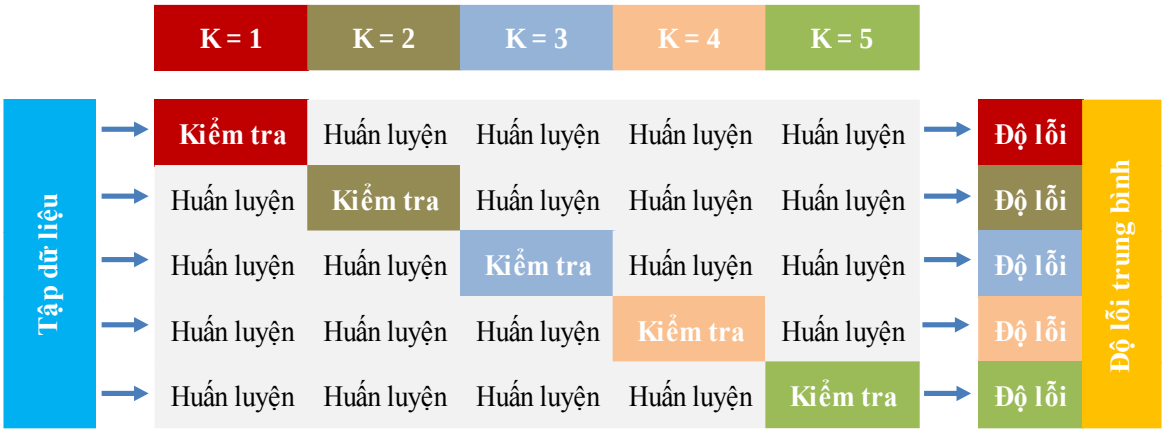
- Mỗi lần (trong số *k* lần) lặp, một tập con được sử dụng làm tập kiểm thử, và (*k*-1) tập con còn lại được dùng làm tập huấn luyện

- *k* giá trị lỗi (mỗi giá trị tương ứng với một *fold*) được tính trung bình cộng để thu được giá trị lỗi tổng thể.

▪ Các lựa chọn thông thường của k: 10, hoặc 5.

▪ Thông thường, mỗi tập con (fold) được lấy mẫu phân tầng (xấp xỉ phân bố lớp) trước khi áp dụng quá trình đánh giá kiểm tra chéo.

▪ Phù hợp khi ta có tập ví dụ D vừa và nhỏ.



Hình 2: Mô tả phương pháp thử nghiệm K-Fold với k=5

2.4.3. Kiểm tra chéo Leave-one-out

Có thể coi là thử nghiệm trên từng cá nhân, là việc tiến hành thử nghiệm với dữ liệu huấn luyện (training) và dữ liệu kiểm thử (test) trên cùng một người, tức là sử dụng dữ liệu thu được từ một người để huấn luyện, sau đó dùng dữ liệu cũng của người đó nhưng chưa được dùng trong huấn luyện để kiểm tra độ chính xác theo phương pháp kiểm tra chéo (cross-validation).

- Một trường hợp (kiểu) của phương pháp kiểm tra chéo
- Số lượng nhóm các (folds) bằng kích thước của tập dữ liệu ($k=|D|$)
- Mỗi nhóm (fold) chỉ bao gồm một ví dụ
- Khai thác tối đa (triệt để) tập ví dụ ban đầu
- Không hề có bước lấy mẫu ngẫu nhiên (no random subsampling)
- Áp dụng lấy mẫu phân tầng (stratification) không phù hợp

  → Vì ở mỗi bước lặp, tập thử nghiệm chỉ gồm có một ví dụ
- Chi phí tính toán (rất) cao
- Phù hợp khi ta có một tập ví dụ D (rất) nhỏ

2.5. Tổng quan công cụ R

2.5.1. Giới thiệu R

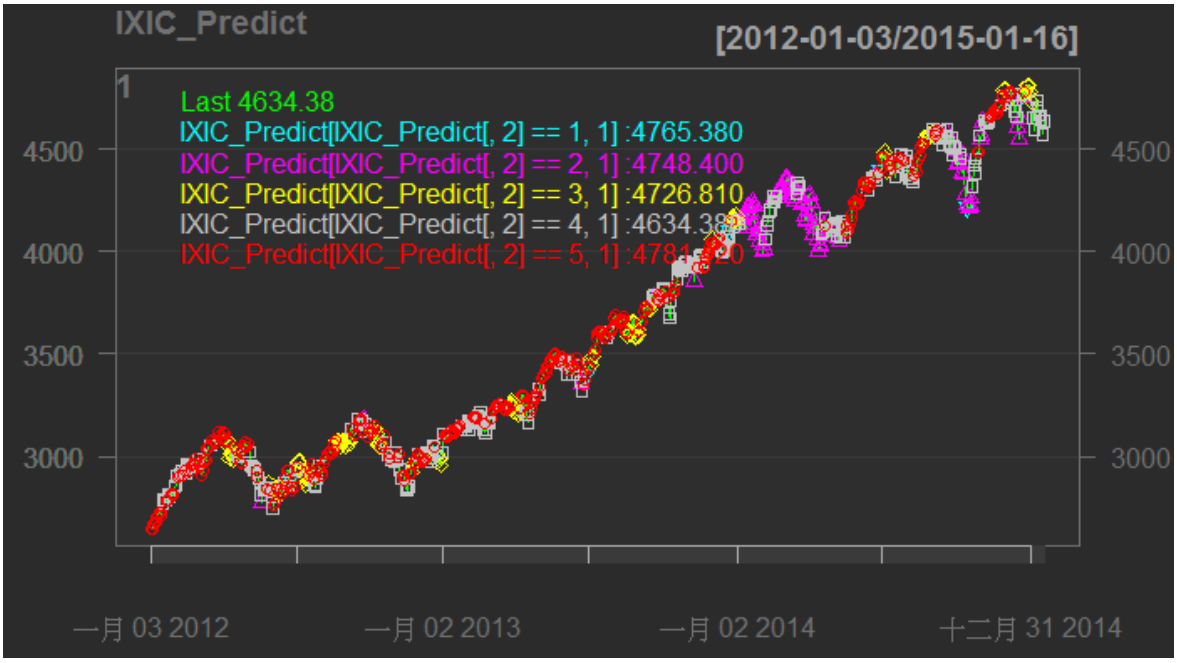
R là một ngôn ngữ lập trình và môi trường phần mềm dành cho tính toán và đồ họa thống kê. Đây là một bản hiện thực ngôn ngữ lập trình S với ngữ nghĩa khối từ vựng lấy cảm hứng từ Scheme. R do Ross Ihaka và Robert Gentleman tạo ra tại Đại học Auckland, New Zealand, đến nay do R Development Core Team chịu trách nhiệm

phát triển. Tên của ngôn ngữ một phần lấy từ chữ cái đầu của hai tác giả (Robert Gentleman và Ross Ihaka), một phần cũng là cách chơi chữ từ tên S.

Ngôn ngữ R đã trở thành một tiêu chuẩn trên thực tế giữa các nhà thống kê cho thấy sự phát triển của phần mềm thống kê, và được sử dụng rộng rãi để phát triển phần mềm thống kê và phân tích dữ liệu.

R là một bộ phận của dự án GNU. Mã nguồn của nó được công bố tự do theo giấy phép bản quyền công cộng GNU, và có các phiên bản dịch sẵn cho nhiều hệ điều hành khác nhau. R sử dụng giao diện dòng lệnh, tuy cũng có một vài giao diện đồ họa người dùng dành cho nó.

Sử dụng R để đơn giản hoá học máy. Tất cả những gì người dùng cần phải biết là làm thế nào mỗi thuật toán có thể giải quyết vấn đề của người dùng, và sau đó người dùng chỉ sử dụng một gói phần mềm được viết ra để nhanh chóng tạo ra mô hình dự đoán trên dữ liệu với một vài dòng lệnh. Ví dụ, người dùng có thể thực hiện Naïve Bayes cho lọc thư rác, sử dụng gom cụm k-means cho phân khúc khách hàng, sử dụng hồi quy tuyến tính để dự báo giá nhà, hoặc thực hiện một mô hình Markov để dự đoán thị trường chứng khoán, như thể hiện trong hình bên dưới:



Hình 3: Dự đoán chứng khoán sử dụng R

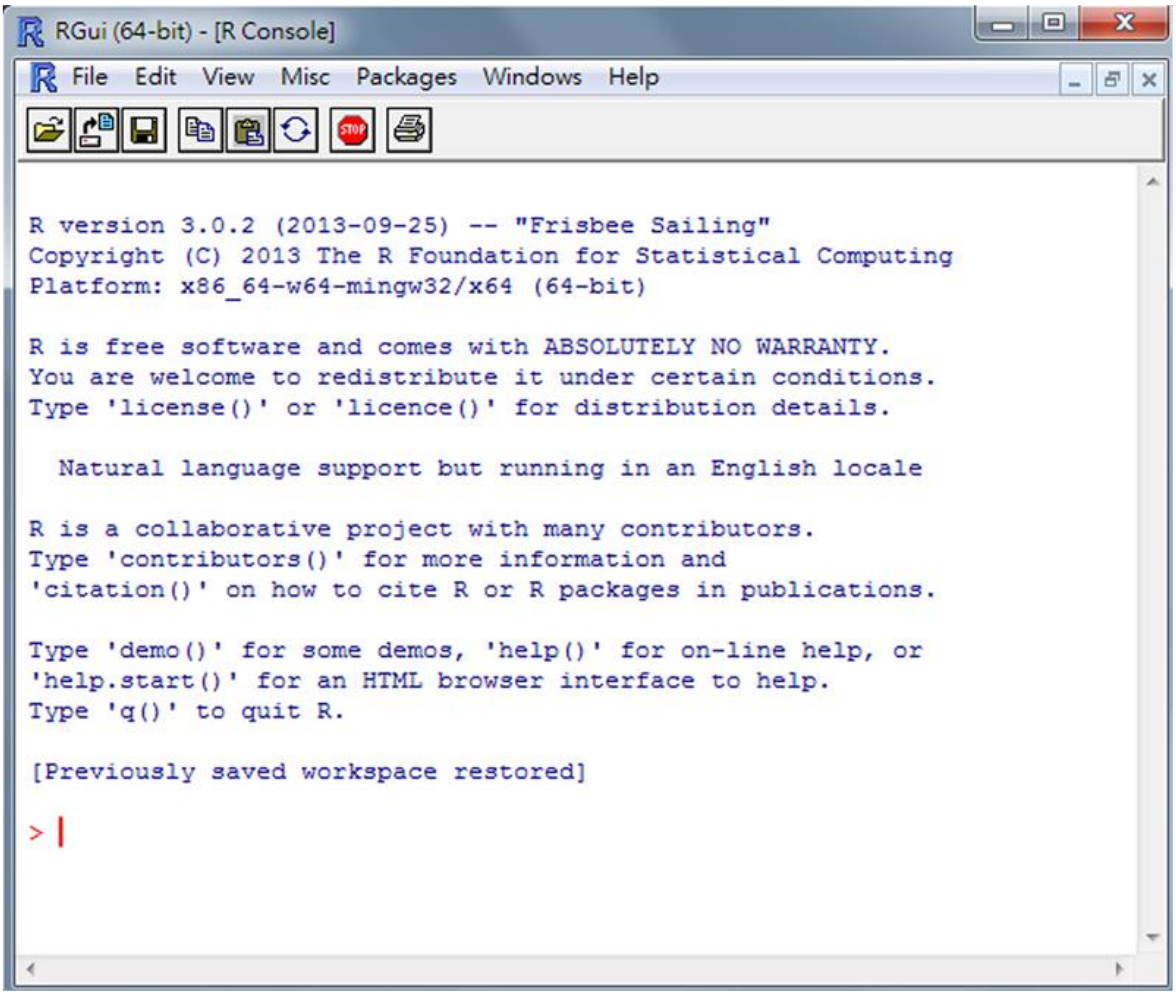
Hơn nữa, người dùng có thể thực hiện giảm chiều phi tuyến để tính toán các tính chất không giống nhau của dữ liệu hình ảnh, và thể hiện qua đồ thị gom cụm, như thể hiện trong hình bên dưới:

Hình 4: Biểu đồ gom cụm dữ liệu hình ảnh sử dụng R

2.5.2. Sử dụng R

Tất cả thông tin chi tiết liên quan đến ngôn ngữ R được giới thiệt chi tiết trên trang web chính thức của R (http://www.r-project.org/).

Để sử dụng R, trước tiên phải cài đặt nó trên máy tính của. Vào trang https://www.r-project.org/ bấm vào đường dẫn **Download R** (http://cran.r-project.org/mirrors.html) để tải và cài đặt R. Sau khi hoàn tất cài đặt, có thể mở công cụ R từ thanh công cụ khởi động hoặc biểu tượng công cụ R trên màn hình máy tính. Màn hình thao tác câu lệnh của công cụ R được mở lên.

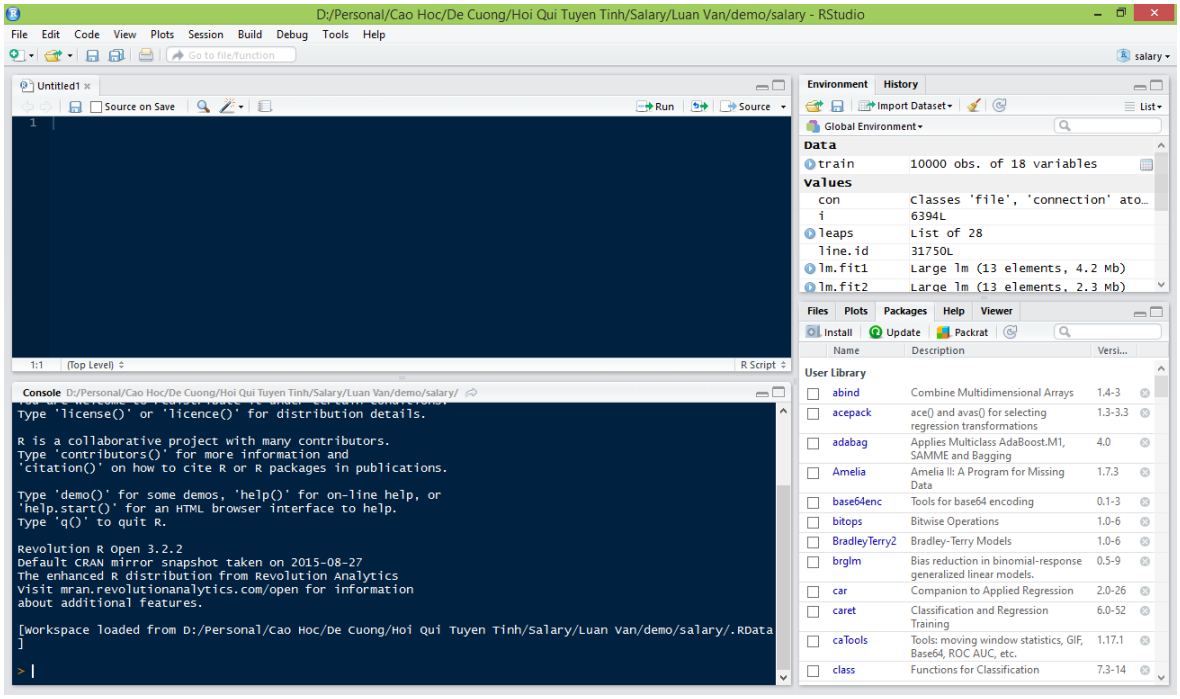


Hình 5: Màn hình thao tác câu lệnh của công cụ R

Hướng dẫn tải và cài đặt công cụ R cho môi trường Linux hoặc Mac OS X: tham khảo ở đườn dẫn sau: https://cran.rstudio.com/doc/manuals/r-release/R-admin.html.

2.5.3. Sử dụng RStudio

Để viết một kịch bản R, người dùng có thể sử dụng R Console, R Commander, hoặc bất kỳ trình soạn thảo văn bản (Emacs, VIM, hoặc sublime). Tuy nhiên, với sự hỗ trợ của RStudio, một môi trường phát triển tích hợp trên R, có thể giúp cho việc sử dụng R dễ dàng hơn rất nhiều. Mở trang RStudio theo đường dẫn sau: https://www.rstudio.com/products/rstudio/ Để tải và cài đặt công cụ RStudio. Sau khi cài đặt hoàn tất, mở công cụ RStudio để thao tác công việc phân tích thật thuận tiện và dễ dàng.



Hình 6: Màn hình làm việc của công cụ RStudio

2.5.4. Một số lệnh cơ bản trong R

a. Ghi chú trong R

Trong R để viết ghi chú hoặc đánh dấu dòng lệnh nào đó ta dùng ký tự (#) trước một ghi chú hoặc dòng lệnh cần ghi chú. R sẽ không thực thi những dòng ghi chú hoặc dòng lệnh có ký tự (#) ở trước. Ví dụ: `# Phần nội dung ghi chú`

b. Đọc và hiển thị dữ liệu trong R

`cars <- read.csv(file = "C:/…/cars.csv", stringsAsFactors = FALSE)` # Đọc dữ liệu. thay chuỗi "`C:/…/`" bằng đường dẫn chứa tập tin dữ liệu của bạn.

`cars` # Hiển thị dữ liệu

`head(cars)` # Hiển thì một vài dòng dữ liệu đầu tiên trong dữ liệu cần phân tích

`names(cars)` # Hiển thị tên biến của dữ liệu cần phân tích

`cars$weight` # Xem dữ liệu cho một biến cụ thể (ví dụ: biến weight) trong bộ dữ liệu cần phân tích

c. Ma trận trong R

`mat <- matrix(0.0, nrow = 3, ncol = 2); mat` # Tạo một ma trận với 3 dòng 2 cột và tất cả dữ liệu bằng giá trị 0, xem dữ liệu ma trận *mat.*

`colnames(mat) <- c("Var 1", "Var 2")` # Đặt tên biến cho ma trận *mat.*

`colnames(mat)` # Hiển thị tên biến của ma trận *mat.*

d. Tập con và khai báo biến trong R

`cars.rsub <- cars[1:50,]` # Lấy tập con của dữ liệu cần phân tích theo dòng

`cars.csub <- cars[,1:3]` # Lấy tập con của dữ liệu cần phân tích theo cột

`cars.rcsub <- cars[c(1,3,5), c(2,4)]` # Lấy ra tập con của dữ liệu cần phân tích theo những dòng và cột cụ thể

`cars.vsub <- cars[which(cars$mpg > 30),]` # Lấy tập con của dữ liệu cần phân tích theo điều kiện

Khai báo biến mới và gán giá trị cho biến theo định dạng `<tên biến> "<-" hoặc "=" <giá trị của biến>` như ví dụ bên dưới:

```
firstletter <- "a"
weight <- cars$weight
```

e. Hiển thị số lượng biểu đồ ở một thời điểm

`par(mfrow=c(1,1))` # Kết quả biểu đồ phân tích hiển thị từng hình một, đây là tính năng mặc định của R.

`par(mfrow=c(2,3))` # Biểu đồ kết quả hiển thị theo 6 hình: 3 hình trên dòng đầu, 3 hình ở dòng dưới. Biểu đồ được hiển thị theo từng dòng của phần hiển thị biểu đồ.

f. Tải, cài đặt và mở gói thư viện trong R

Ví dụ muốn cài gói thư viện phân tích bằng biểu đồ `ggplot2` trong R, sử dụng câu lệnh sau:

`install.packages("ggplot2")` # R sẽ tự tìm trong máy chủ gần nhất chứa gói thư viện muốn cài đặt sau đó tải và cài đặt vào công cụ R.

Mở và sử dụng gói thư viện trong R (ví dụ: gói "`ggplot2`"):

```
library(ggplot2)
```

CHƯƠNG 3: ỨNG DỤNG PHÂN TÍCH HỒI QUY DỰ ĐOÁN MỨC LƯƠNG

3.1. Tìm hiểu dữ liệu

Ứng dụng phân tích hồi quy tuyến tính trên dữ liệu các quảng cáo việc làm được cung cấp bởi Kaggle [5]: (https://www.kaggle.com/c/job-salary-prediction/data). Dữ liệu được mô tả như sau:

Id – Một định danh duy nhất cho mỗi quảng cáo tuyển dụng.

Title – Tiêu đề của quảng cáo tuyển dụng. Đây là một bản tóm tắt mô tả công việc hoặc vị trí tuyển dụng.

FullDescription – Mô tả đầy đủ của quảng cáo tuyển dụng theo quy định của các nhà quảng cáo tuyển dụng. Không có thông tin lương xuất hiện trong các mô tả này.

LocationRaw – Nơi làm việc được cung cấp bởi các nhà quảng cáo tuyển dụng.

LocationNormalized – Nơi làm việc đã được chuẩn hóa dựa trên nơi làm việc được cung cấp bởi các nhà quảng cáo tuyển dụng. Chuẩn hóa này có thể không hoàn hảo.

ContractType – Mô tả công việc làm toàn thời gian hoặc công việc bán thời gian.

ContractTime – Mô tả hình thức làm việc lâu dài hoặc theo hợp đồng.

Company – Tên công ty được cung cấp bởi các nhà quảng cáo tuyển dụng.

Category – Nhóm công việc, phân ra 30 nhóm công việc chuẩn. Có rất nhiều dữ liệu hỗn tạp và sai trong trường dữ liệu này vì chúng đến từ nhiều nguồn quảng cáo khác nhau.

SalaryRaw – Mức lương được mô tả trong các quảng cáo việc làm từ các nhà quảng cáo tuyển dụng.

SalaryNormalised - Mức lương hàng năm dự trên mức lương trong các quảng cáo việc làm từ nhà tuyển dụng. Đây là giá trị để dự đoán.

Location Tree: Đây là một tập hợp dữ liệu bổ sung để mô tả mối quan hệ thứ bậc giữa các địa điểm khác nhau thể hiện trong các dữ liệu tuyển dụng. Nó cũng có thể là những mối quan hệ có ý nghĩa giữa tiền lương của công việc trong một khu vực

địa lý tương tự, ví dụ tiền lương trung bình ở London và Nam Trung Đông cao hơn trong phần còn lại của Vương quốc Anh.

Tổng quan 20 dòng dữ liệu đầu tiên của dữ liệu quảng cáo tuyển dụng và dữ liệu về địa điểm ở Anh được thể hiện như Hình 7 và Hình 8 bên dưới:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Id	Title	FullDescription	LocationRaw	LocationNormalized	ContractType	ContractTime	Company	Category	SalaryRaw	SalaryNormalized	SourceName
2	12612628	Engineeri	Engineering Sys	Dorking, Surrey	Dorking		permanent	Gregory Martin	Engineering Jobs	20000 - 30000/annum	25000	cv-library.co.uk
3	12612830	Stress Eng	Stress Engineer	Glasgow, Scotla	Glasgow		permanent	Gregory Martin	Engineering Jobs	25000 - 35000/annum	30000	cv-library.co.uk
4	12612844	Modelling	Mathematical M	Hampshire, Sou	Hampshire		permanent	Gregory Martin	Engineering Jobs	20000 - 40000/annum	30000	cv-library.co.uk
5	12613049	Engineeri	Engineering Sys	Surrey, South E	Surrey		permanent	Gregory Martin	Engineering Jobs	25000 - 30000/annum	27500	cv-library.co.uk
6	12613647	Pioneer, M	Pioneer, Miser	Surrey, South E	Surrey		permanent	Gregory Martin	Engineering Jobs	20000 - 30000/annum	25000	cv-library.co.uk
7	13179816	Engineeri	Engineering Sys	Dorking, Surrey	Dorking		permanent	Gregory Martin	Engineering Jobs	20000 - 30000/annum	25000	cv-library.co.uk
8	14131336	Senior Sul	A globally renov	Aberdeen, Bord	UK		permanent	Indigo 21 Ltd	Engineering Jobs	50000 - 100000/annun	75000	cv-library.co.uk
9	14663196	RECRUITM	THIS IS A LIVE VA	MANCHESTER, (	Manchester		permanent	Code Blue Recr	HR & Recruitmen	18000 - 26000/annum	22000	cv-library.co.uk
10	14663197	RECRUITM	This is an except	LEEDS, West Yo	Leeds		permanent	Code Blue Recr	HR & Recruitmen	18000 - 28000/annum	23000	cv-library.co.uk
11	15395797	Subsea Ca	A subsea engine	Aberdeen, UK	Aberdeen		permanent	Indigo 21 Ltd	Engineering Jobs	70000 - 100000/annun	85000	cv-library.co.uk
12	19047429	Trainee M	Are you a succes	East Midlands	East Midlands		permanent	Brite Recruitme	Accounting & Fin	17000 - 25000/annum	21000	cv-library.co.uk
13	20199757	PROJECT E	PROJECT ENGINE	Witney, Oxford	Witney		permanent	MatchBox Recru	Healthcare & Nur	35000 - 40000/annum	37500	cv-library.co.uk
14	20638787	Principal (	Aerospace Senic	Avon, South We	Avon		permanent	Gregory Martin	Other/General Jo	40000 - 50000/annum	45000	cv-library.co.uk
15	20638788	Senior Fat	Senior Fatigue S	Avon, South We	Avon		permanent	Gregory Martin	Other/General Jo	35000 - 45000/annum	40000	cv-library.co.uk
16	20797143	Chef de P	A well respecte	Derby Derbyshi	Derby			Chef Results	Hospitality & Cat	Upto 16,000 per annu	16000	caterer.com
17	22579462	Quality En	Our client are a	Gateshead, Tyn	Gateshead		permanent	Asset Appointm	Engineering Jobs	22000/annum	22000	cv-library.co.uk
18	22581547	Principal (	A leading Subse	Kent, South Eas	Kent		permanent	Indigo 21 Ltd	Engineering Jobs	70000 - 100000/annun	85000	cv-library.co.uk
19	22933091	Chef de P	A popular hotel	Norfolk East An	UK			Chef Results	Hospitality & Cat	18,000 per annum + T	18000	caterer.com
20	23528672	Senior Fat	Senior Fatigue S	Avon, South We	Avon		permanent	Gregory Martin	Engineering Jobs	34000 - 45000/annum	39500	cv-library.co.uk

Hình 7: Tổng quan 20 dòng dữ liệu đầu tiên trong dữ liệu quảng cáo tuyển dụng

location_tree.txt

```
"UK~London~East London~Mile End"
"UK~London~East London~Shadwell"
"UK~London~East London~Spitalfields"
"UK~London~East London~Stepney"
"UK~London~East London~Wapping"
"UK~London~East London~Whitechapel"
"UK~London~East London~Bethnal Green"
"UK~London~East London~Cambridge Heath"
"UK~London~East London~Haggerston"
"UK~London~East London~Shoreditch"
"UK~London~East London~Bow"
"UK~London~East London~Bromley-By-Bow"
"UK~London~East London~Old Ford"
"UK~London~East London~Chingford"
"UK~London~East London~Chingford Hatch"
"UK~London~East London~Friday Hill"
"UK~London~East London~Hale End"
"UK~London~East London~Highams Park"
"UK~London~East London~South Chingford"
"UK~London~East London~Clapton"
```

Hình 8: Dữ liệu địa điểm ở Anh

Các nhóm công việc chính và lương trung bình được đưa ra dưới đây:

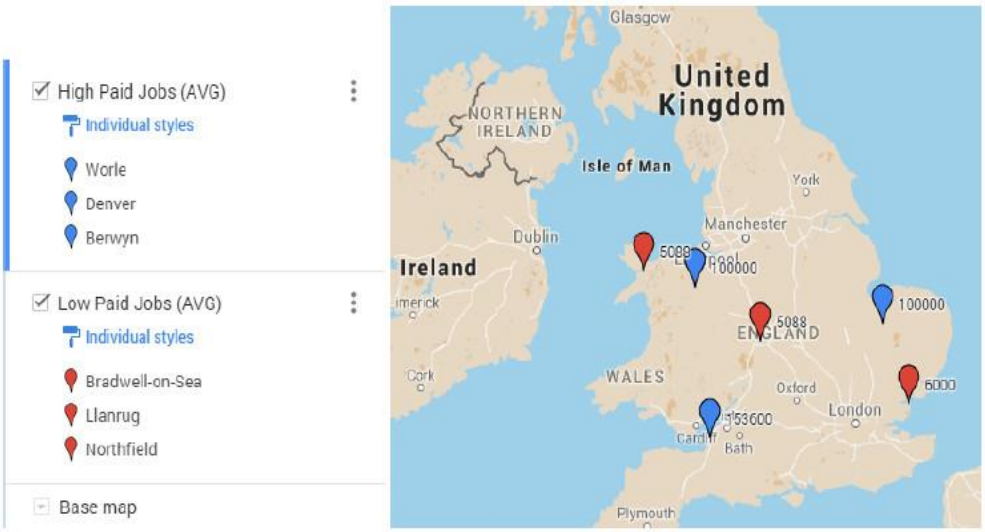
Bảng 3: Lương trung bình theo nhóm công việc

Nhóm công việc	**Mức lương trung bình**
Accounting & Finance Jobs	38622.456172
Admin Jobs	20916.362130
Charity & Voluntary Jobs	28200.204936
Consultancy Jobs	36374.208544
Creative & Design Jobs	32585.487409
Customer Services Jobs	19795.438648
Domestic help & Cleaning Jobs	17492.815789
Energy, Oil & Gas Jobs	**45384.110103** (lớn nhất)
Engineering Jobs	35608.058499
Graduate Jobs	28677.438703
HR & Recruitment Jobs	32386.298070
Healthcare & Nursing Jobs	32203.188169
Hospitality & Catering Jobs	23552.462798
IT Jobs	44081.780990
Legal Jobs	42350.550210
Logistics & Warehouse Jobs	25711.287983
Maintenance Jobs	17533.320433
Manufacturing Jobs	25653.276129
Other/General Jobs	34361.716029
PR, Advertising & Marketing Jobs	35521.313285
Part time Jobs	**10514.285714** (nhỏ nhất)
Property Jobs	32167.056647
Retail Jobs	32851.304998
Sales Jobs	30685.809207
Scientific & QA Jobs	33950.363683
Social work Jobs	32324.809030

Teaching Jobs	27240.637182
Trade & Construction Jobs	36050.166566
Travel Jobs	23913.173143

Mức lương trung bình ở Anh là 31,554.441 đơn vị. Mức lương trung bình cao nhất là ở thành phố "Worle" và "Denver" với giá trị là 153.600 và 100.000. Tuy nhiên, mức lương trung bình thấp nhất là ở thành phố "Northfield" và "Llanrug" chỉ là 5.088 đơn vị. Trong đó mức lương trung bình trong quần đảo Channel là khoảng 71.500 đơn vị trên mức trung bình. Phần còn lại của những nơi khác có khoảng 10.000 đơn vị trên hoặc dưới mức trung bình. Ngoài ra còn có một tập tin khác cung cấp thông tin về các địa điểm thành phố trong Vương quốc Anh với cấu trúc dạng cây. Nó giúp cho việc so sánh các quần đảo Channel với các thành phố khác.

Bộ dữ liệu này cũng cho biết mọi người làm việc theo hợp đồng hoặc là làm việc không thời hạn. Căn bản đầu tiên là những người làm việc theo hợp đồng sẽ được trả cao hơn vì họ không có bất kỳ bảo hiểm việc làm nào. Ví dụ tại Mỹ, dịch giả tự do được trả tiền cao hơn lao động thường xuyên. Nhưng tại Anh kông có nhiều sự khác biệt trong thu nhập giữa nhân viên làm việc theo hợp đồng và nhân viên làm việc không thời hạn. Những người làm việc theo hợp đồng kiếm được khoản 36,099.67 đơn vị và nhân viên làm việc không thời hạn có được khoản 35,327.47 đơn vị khác biệt khoản 700 đơn vị.



Hình 9: Mức lương trung bình cao nhất và thấp nhất theo thành phố ở Anh

Hình 9 cho thấy bản đồ của những thành phố có mức lương trung bình cao nhất và thấp nhất. Các ô màu xanh là mức lương bình quân cao nhất và màu đỏ là mức lương bình quân thấp nhất ở cấp thành phố. Bên cạnh đó mỗi ô là mức lương trung bình được đề cập. Không có xu hướng rõ ràng dựa trên vị trí địa lý cho số tiền lương kiếm được. Tuy nhiên, London và phía Đông Nam có thu nhập cao hơn mức trung bình.

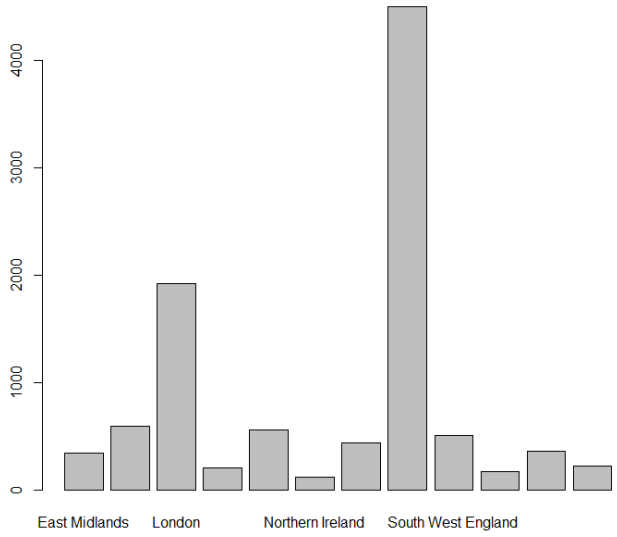
Tập dữ liệu này cũng tương tự như dữ liệu về điều tra dân số. Dữ liệu điều tra dân số có sẵn cho mỗi quốc gia. Và cũng có các cuộc thi khác diễn ra trên Kaggle (https://www.kaggle.com/c/us-census-challenge/leaderboard) cũng như những nơi khác như KDD (https://archive.ics.uci.edu/ml/datasets - điều tra về dân số và thu nhập). Tuy nhiên hầu hết bộ dữ liệu đòi hỏi để dự đoán xem thu nhập của một người là trên hoặc dưới một ngưỡng nào đó. Nó đơn giản hơn nhiều so với những vấn đề trong luận văn này là cố gắng để giải quyết, tức là dự đoán chính xác mức lương của công việc trên quảng cáo tuyển dụng cụ thể.

3.2. Chuẩn bị dữ liệu

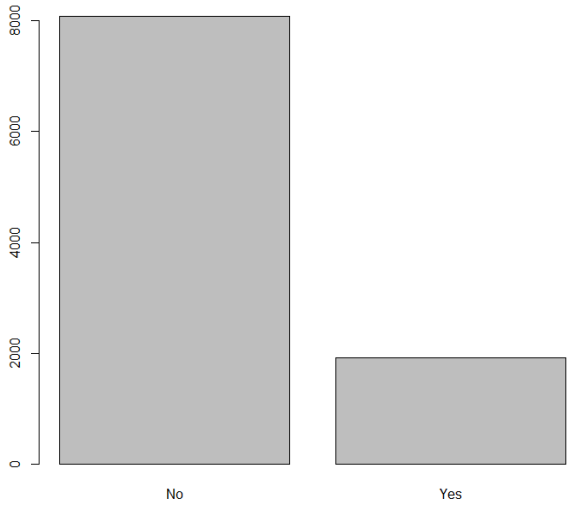
3.2.1. Phân loại dữ liệu Địa Điểm, Tiêu Đề, và Mô Tả công việc

a. Địa điểm làm việc

Địa điểm làm việc trong bộ dữ liệu này được phân bố khá đa dạng. Do đó trong luận văn này được phân thành 2 nhóm: Địa điểm cụ thể (ví dụ: Luân Đôn) và những địa điểm khác. Dựa trên tập tin dữ liệu về địa điểm thành phố khu vực ở Anh, tác giả dùng R để phân nhóm dữ liệu lại và được thể hiện như Hình bên dưới:



Hình 10: Dữ liệu địa điểm trước khi phân loại theo theo địa điểm là Luân Đôn

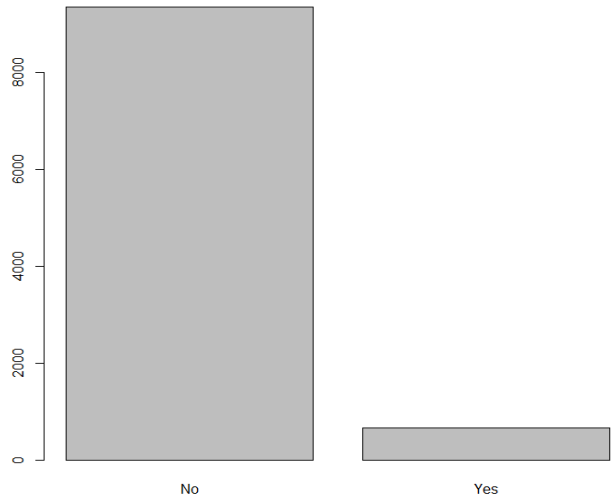


Hình 11: Dữ liệu được phân loại theo địa điểm làm việc ở Luân Đôn và Khác Luân Đôn

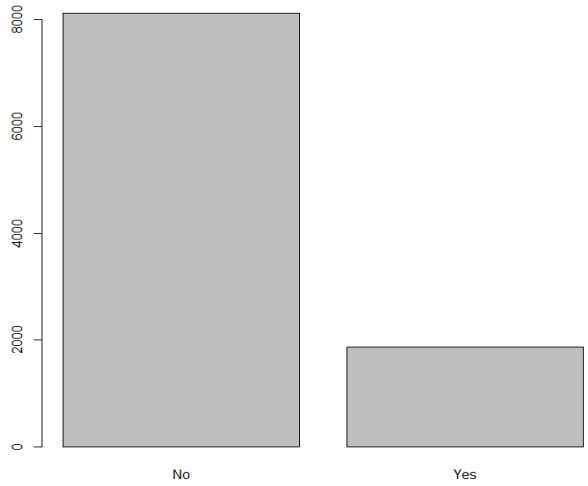
b. Tiêu đề công việc

Tiêu đề công việc mỗi quảng cáo tuyển dụng là khác nhau, nhưng ngược lại chúng có những điểm chung đó là tiêu đề công việc tập chung vào vị trí tuyển dụng công

việc. Ví dụ: tiêu đề việc làm cho vị trí có kinh nghiệm, vị trí quản lý, hoặc những vị trí tuyển dụng khác. Do đó ở đây trên cơ sở dữ liệu đang phân tích tác giả phân theo các nhóm tiêu đề công việc như sau: Tiêu đề cho vị trí ứng viên có kinh nghiệm, Tiêu đề cho vị trí quản lý. Dữ liệu phân loại lại được thể hiện như hình bên dưới:



Hình 12: Dữ liệu được phân loại theo tiêu đề công việc cho vị trí có kinh nghiệm

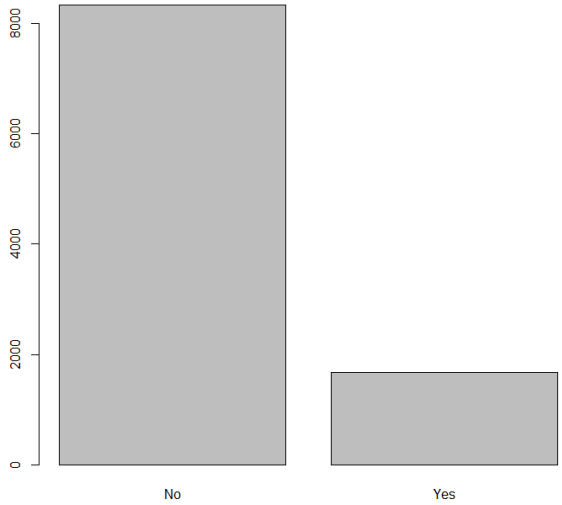


Hình 13: Dữ liệu được phân loại theo tiêu đề cho vị trí quản lý

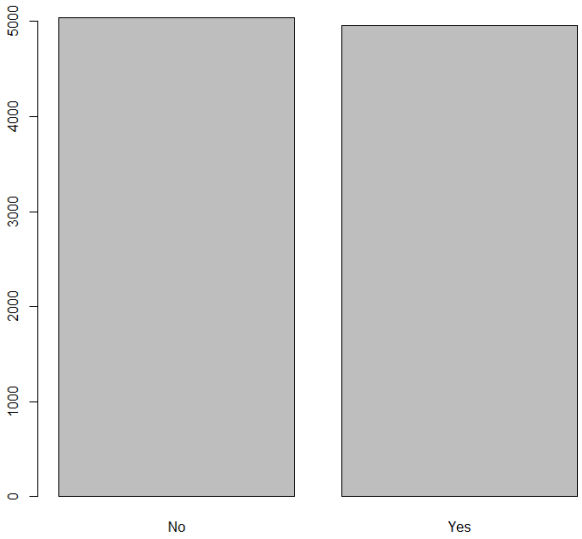
c. Mô tả công việc

Tương tự ở tiêu đề công việc, quảng cáo tuyển dụng được mô tả chi tiết với tiêu đề cho từng vị trí tương ứng. Do đó ở đây trên cơ sở dữ liệu đang phân tích tác giả

phân theo các nhóm mô tả công việc như sau: Mô tả công việc cho vị trí ứng viên có kinh nghiệm, Mô tả công việc cho vị trí quản lý. Dữ liệu phân loại lại được thể hiện như hình bên dưới:



Hình 14: Dữ liệu được phân loại theo mô tả công việc cho vị trí có kinh nghiệm

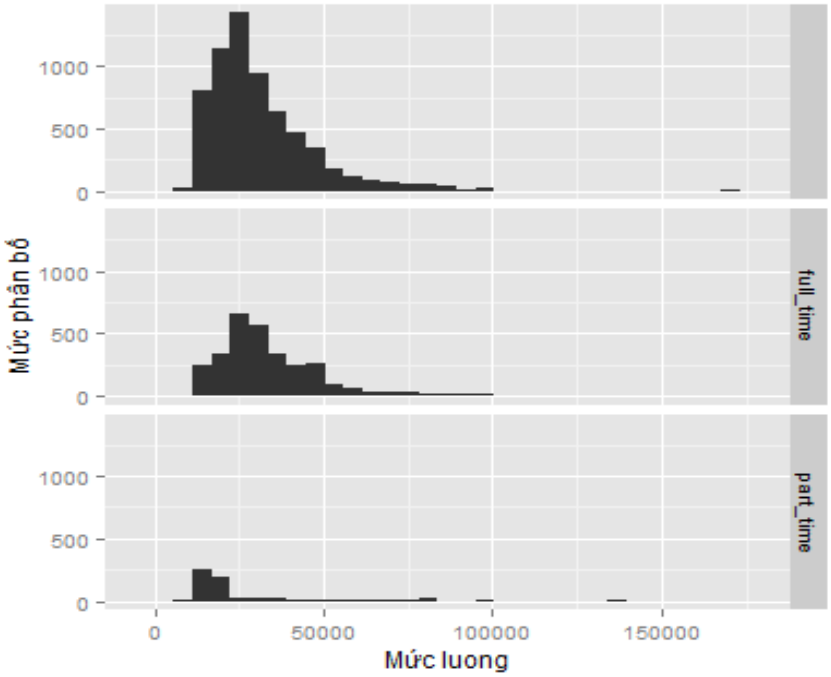


Hình 15: Dữ liệu được phân loại theo mô tả công việc cho vị trí quản lý

3.2.2. Phân bố dữ liệu việc làm theo loại công việc

Mức phân bố công việc dựa trên loại công việc ta thấy rằng dữ liệu cho loại công việc làm toàn thời gian phân bố với mức phân bố cao hơn so với loại công việc bán

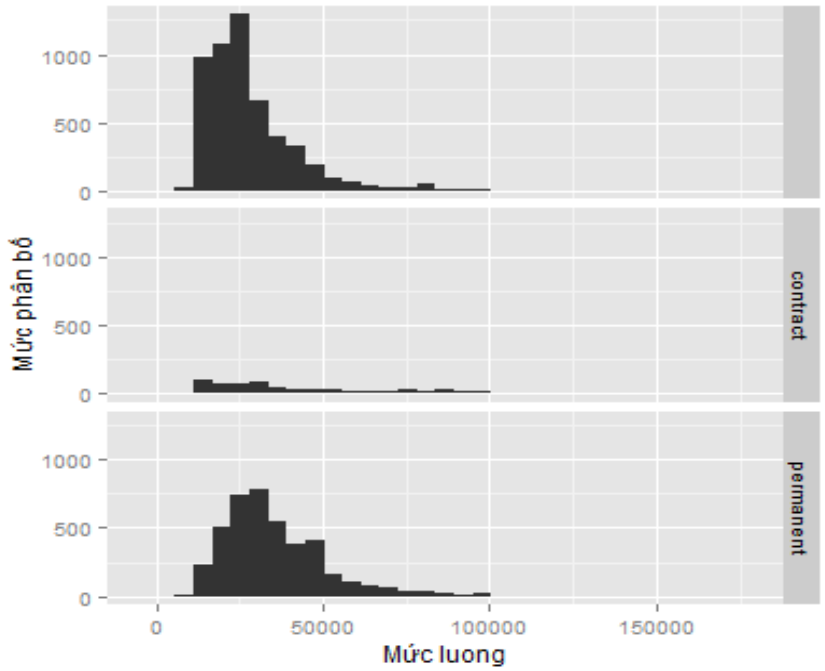
thời gian. Mức lương cho các loại công việc toàn thời gian và bán thời gian tập chung hầu hết ở mức lương từ 10,000 đơn vị đến 50,000 đơn vị. Mức phân bố dữ liệu loại công việc không được cung cấp trong quảng cáo việc làm cao hơn so với loại công việc toàn thời gian và loại công việc bán thời gian.



Hình 16: Phân bố dữ liệu quảng cáo tuyển dụng theo loại công việc dựa trên mức lương

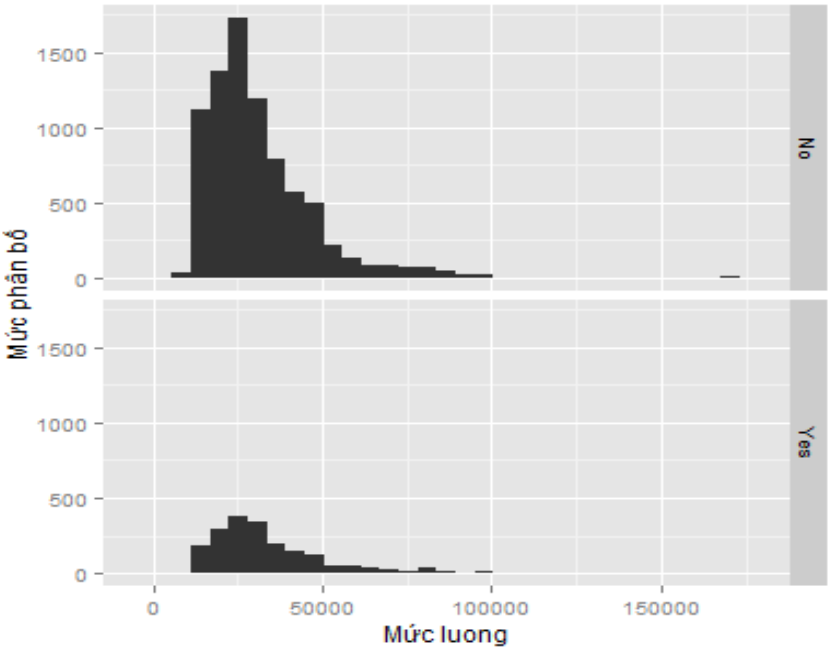
3.2.3. Phân bố dữ liệu việc làm theo loại hợp đồng công việc

Mức phân bố dữ liệu dựa trên loại hợp đồng là làm việc không xác định thời gian cao hơn so với làm việc theo hợp đồng.



Hình 17: Phân bố dữ liệu quảng cáo tuyển dụng theo loại hợp đồng dựa trên mức lương

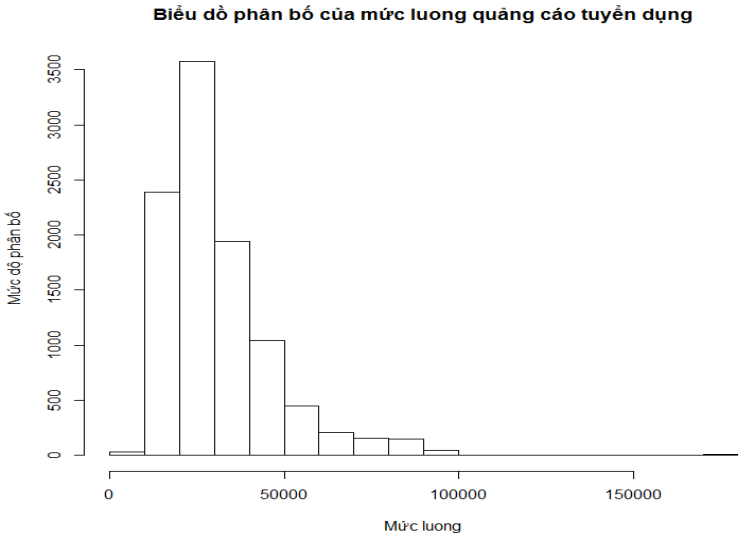
3.2.4. Phân bố dữ liệu việc làm theo địa điểm làm việc là Luân Đôn



Hình 18: Phân bố dữ liệu quảng cáo tuyển dụng theo địa điểm làm việc là Luân Đôn dựa trên mức lương

3.2.5. Phân bố dữ liệu việc làm theo mức lương

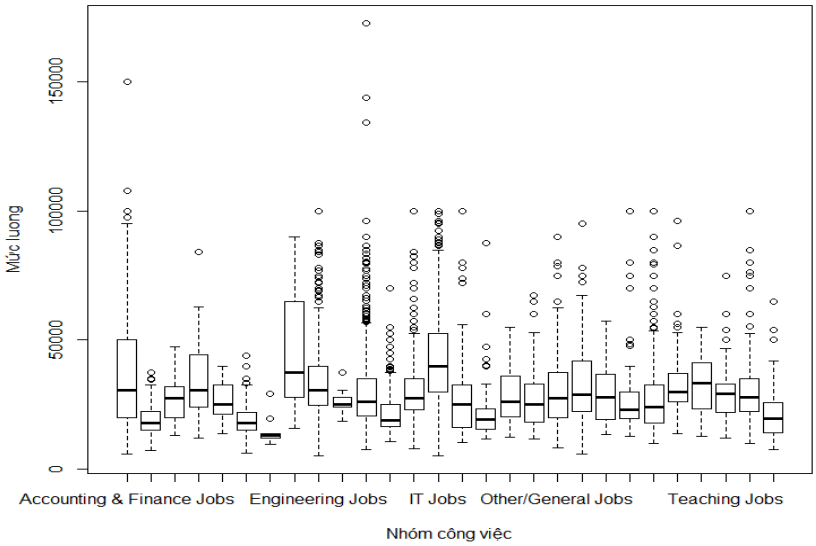
Dữ liệu phân bố theo mức lương tập trung nhiều vào khoản 10,000 đơn vị đến 50,000 đơn vị. Một số rất ít phân bố ở mức lương > 150,000 đơn vị.



Hình 19: Phân bố dữ liệu quảng cáo tuyển dụng dựa trên mức lương

3.2.6. Phân bố dữ liệu việc làm theo nhóm công việc dựa trên mức lương

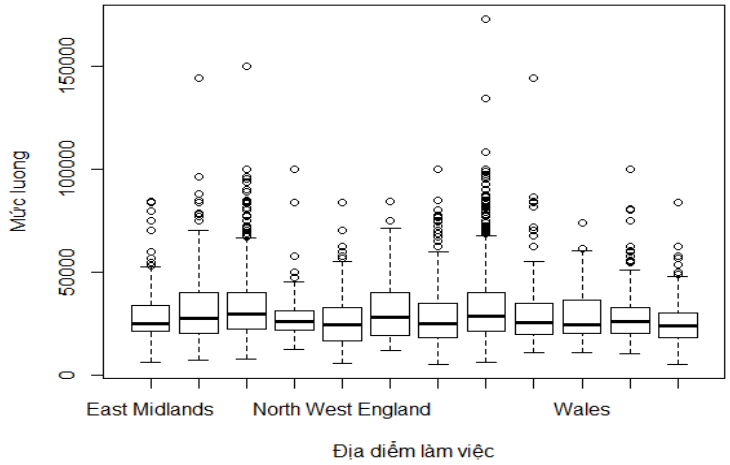
Mức phân bố dữ liệu theo nhóm công việc dựa trên mức lương tập trung khá đồng điều và tập trung nhiều ở nhóm công việc về Dầu khí năng lượng với mức lương trung bình vào khoản 45,000 đơn vị.



Hình 20: Phân bố dữ liệu quảng cáo tuyển dụng dựa trên nhóm công việc

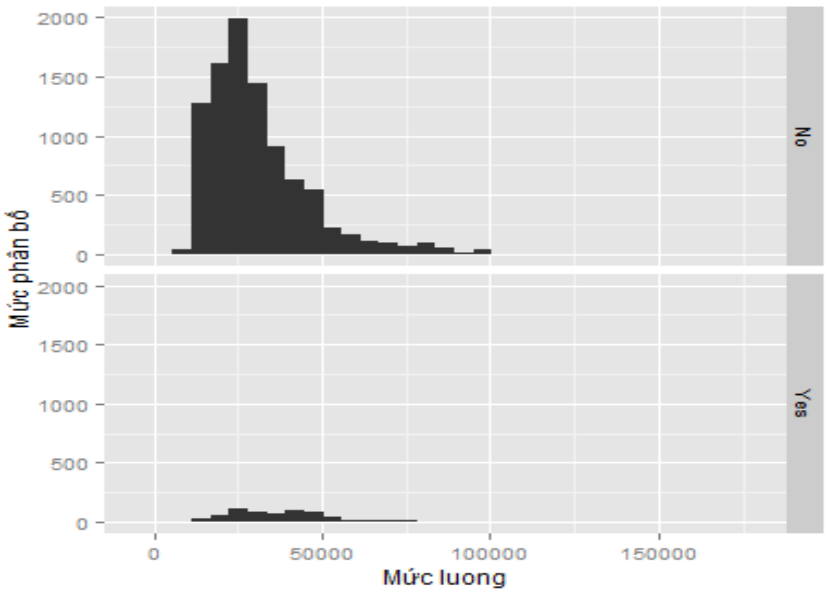
3.2.7. Phân bố dữ liệu việc làm theo địa điểm làm việc dựa trên mức lương

Mức phân bố dữ liệu theo địa điểm làm việc dựa trên mức lương tập trung khá đồng điều và tập trung nhiều ở địa điểm làm việc ở Denver có mức lương trung bình cao.



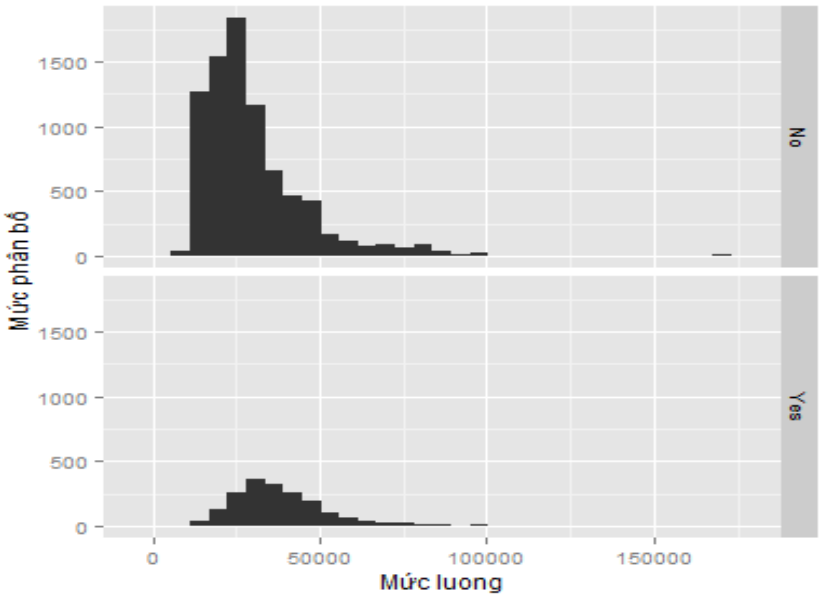
Hình 21: Phân bố dữ liệu quảng cáo tuyển dụng theo địa điểm làm việc dựa trên mức lương

3.2.8. Phân bố dữ liệu việc làm theo tiêu đề công việc là vị trí ứng viên có kinh nghiệm dựa trên mức lương



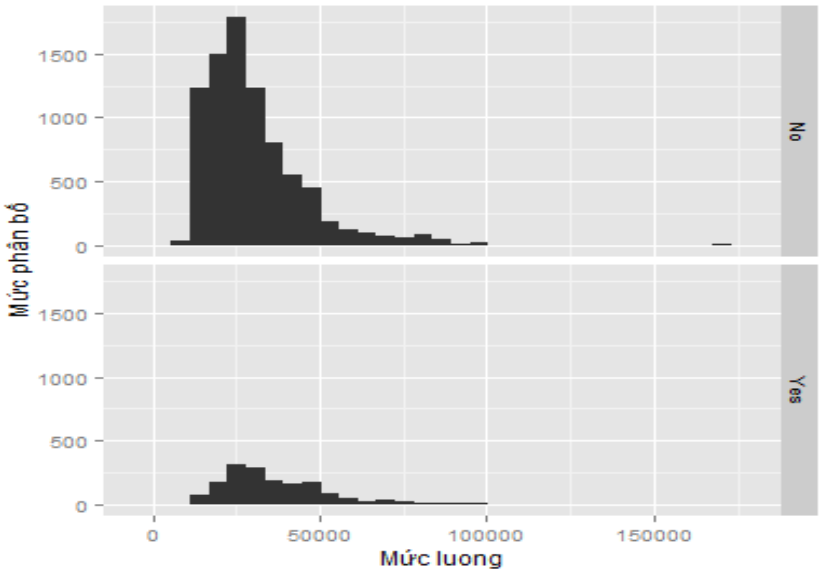
Hình 22: Phân bố dữ liệu quảng cáo tuyển dụng theo tiêu đề công việc là vị trí ứng viên có kinh nghiệm dựa trên mức lương

3.2.9. Phân bố dữ liệu việc làm theo tiêu đề công việc là vị trí quản lý dựa trên mức lương



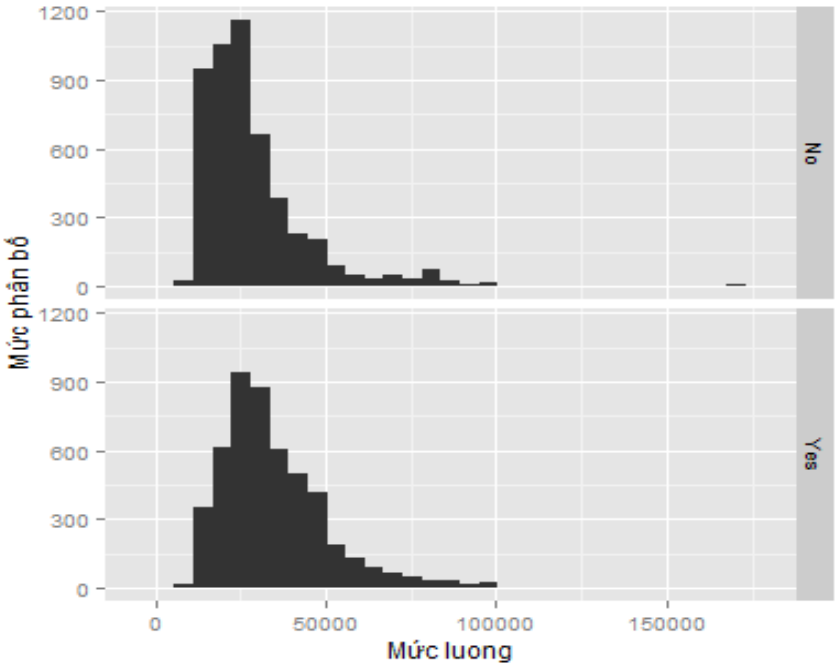
Hình 23: Phân bố dữ liệu quảng cáo tuyển dụng theo tiêu đề công việc là vị trí quản lý dựa trên mức lương

3.2.10. Phân bố dữ liệu việc làm theo mô tả công việc là vị trí ứng viên có kinh nghiệm dựa trên mức lương



Hình 24: Phân bố dữ liệu quảng cáo tuyển dụng theo mô tả công việc là ứng viên có kinh nghiệm dựa trên mức lương

3.2.11. Phân bố dữ liệu việc làm theo mô tả công việc là vị trí quản lý dựa trên mức lương



Hình 25: Phân bố dữ liệu quảng cáo theo mô tả công việc là vị trí quản lý dựa trên mức lương

3.3. Mô hình hóa

3.3.1. Biến độc lập và Biến phụ thuộc

Dựa trên mức độ tương quan và mức độ ảnh hưởng qua lại giữa các biến ta có biến phụ thuộc và các biến độc lập như bảng dưới đây:

Bảng 4: Biến độc lập và Biến phụ thuộc

Biến phụ thuộc	Biến độc lập
Mức lương công việc	Nhóm công việc
	Loại công việc
	Loại hợp đồng
	Địa điểm làm việc
	Tiêu đề công việc
	Mô tả công việc

3.3.2. Phân tích ảnh hưởng của nhóm công việc lên mức lương

Phương trình hồi quy một biến nhóm công việc ảnh hưởng lên mức lương quảng cáo tuyển dụng có dạng như sau:

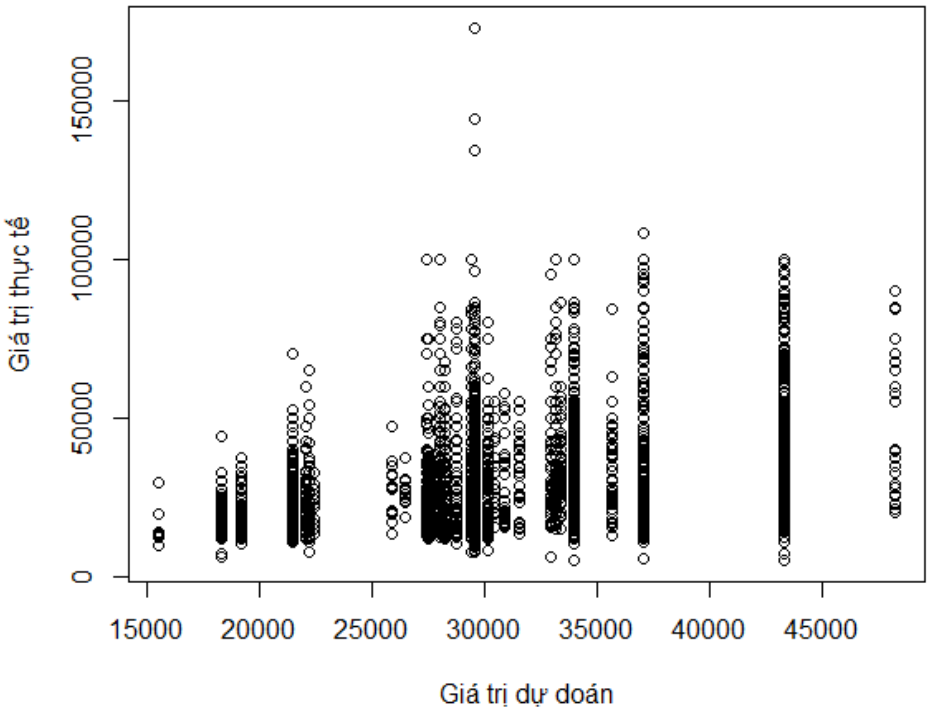
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

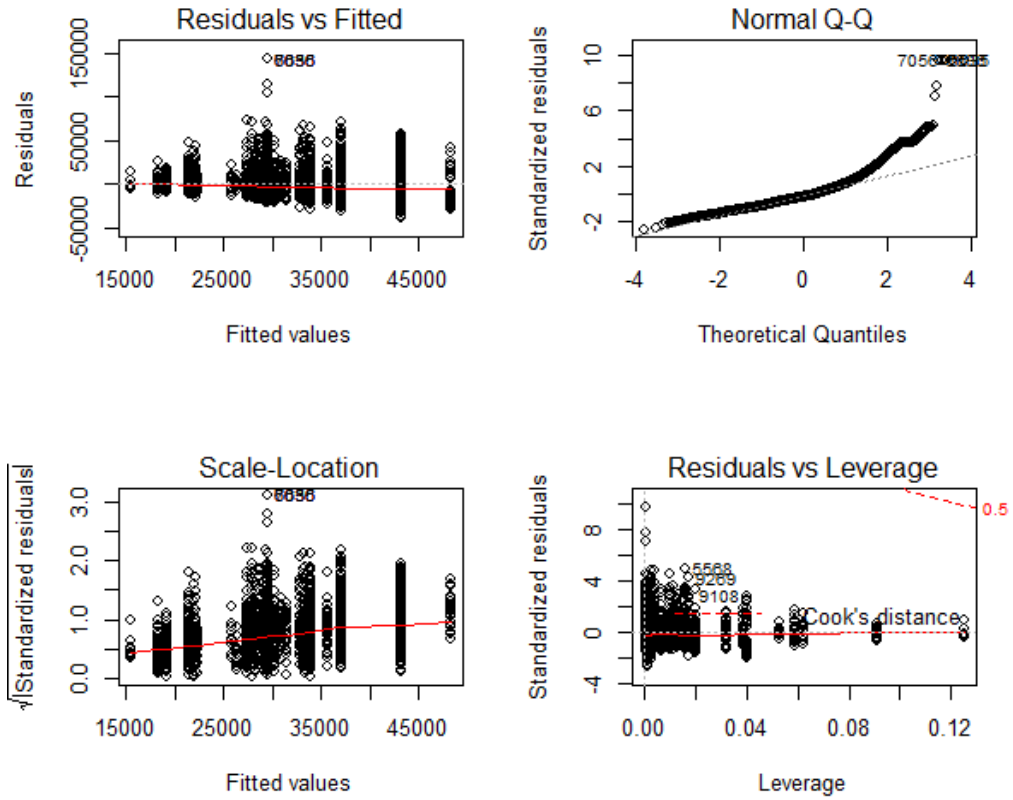
- Y: là biến Mức lương
- X_1: là biến Nhóm công việc

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit1 <- lm(SalaryNormalized ~ Category, data = tr)
summary(lm.fit1)
lm.pred1 <- predict(lm.fit1, newdata = val)
sqrt(mean((lm.pred1 - val$SalaryNormalized)^2))
```



Hình 26: Mối liên hệ giữa nhóm công việc và mức lương



Hình 27: Phân tích kiểm tra mối liên hệ giữa nhóm công việc và mức lương

3.3.3. Phân tích ảnh hưởng của loại công việc lên mức lương

Phương trình hồi quy một biến loại công việc ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

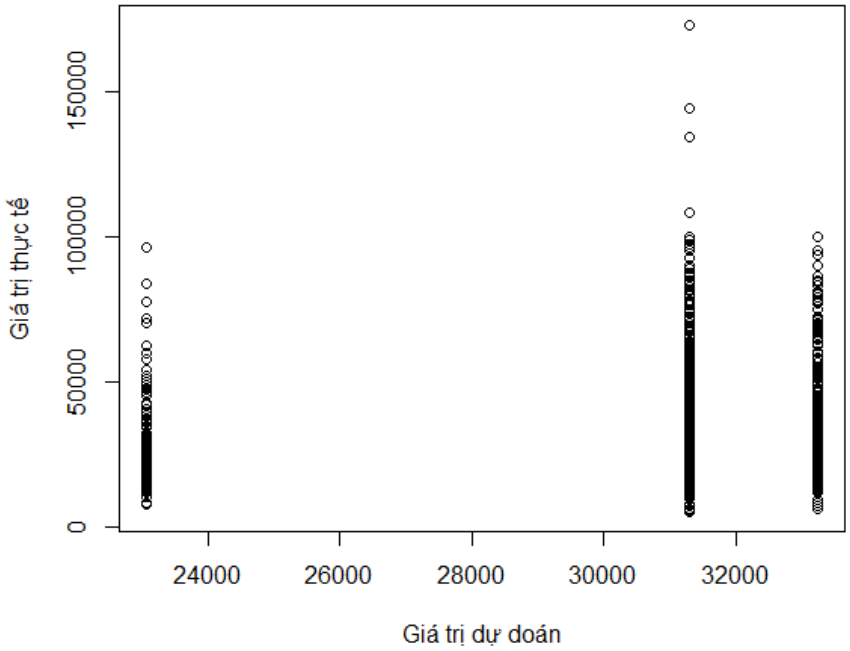
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

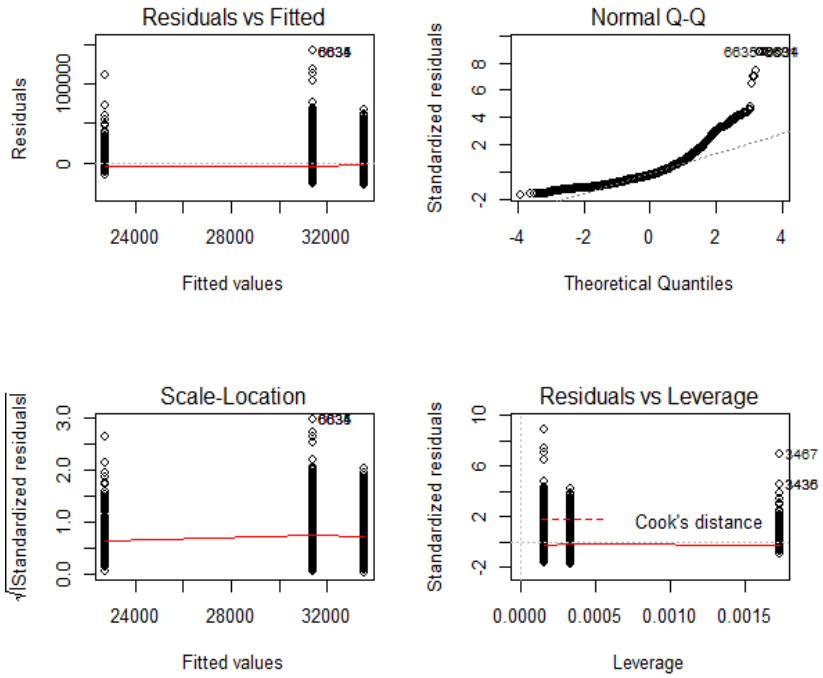
- Y: là biến Mức lương
- X_1: là biến Loại công việc

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit2 <- lm(SalaryNormalized ~ ContractType, data =
tr)
summary(lm.fit2)
lm.pred2 <- predict(lm.fit2, newdata = val)
sqrt(mean((lm.pred2 - val$SalaryNormalized)^2))
```



Hình 28: Liên hệ giữa loại công việc với mức lương



Hình 29: Phân tích kiểm tra mối liên hệ giữa loại công việc và mức lương

3.3.4. Phân tích ảnh hưởng của loại hợp đồng lên mức lương

Phương trình hồi quy một biến loại hợp đồng ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

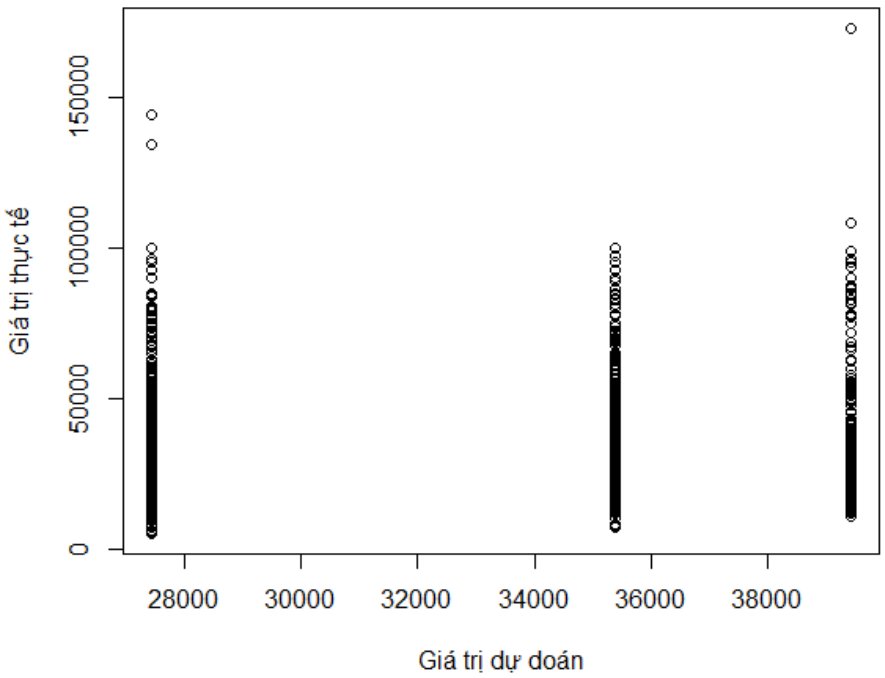
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

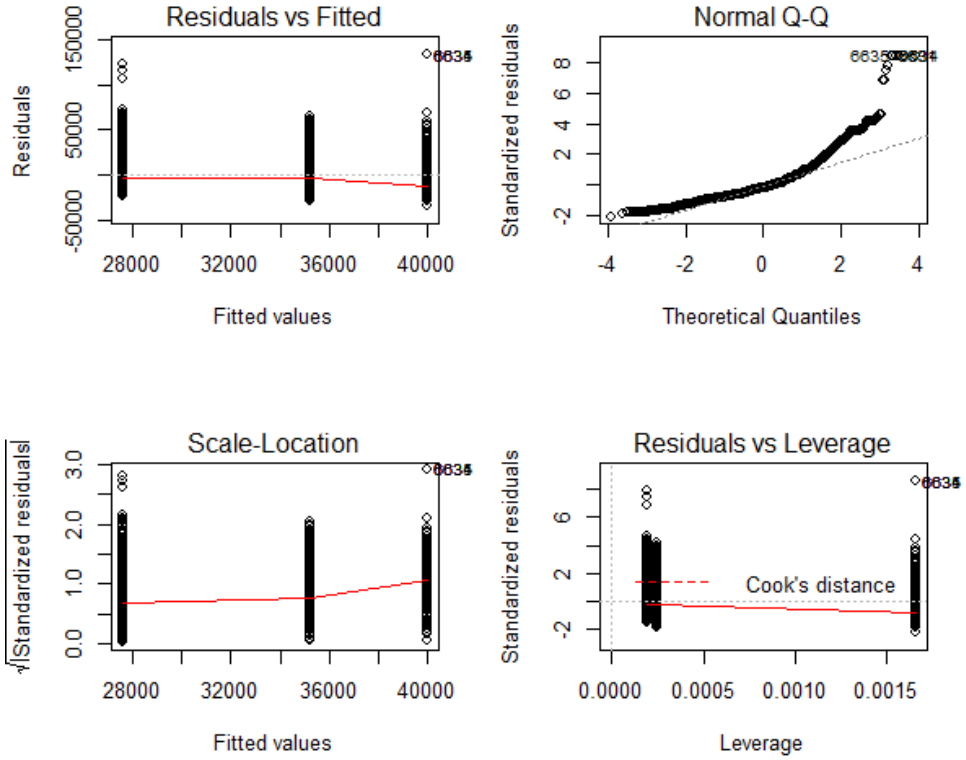
- Y: là biến Mức lương
- X_1: là biến Loại hợp đồng

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit3 <- lm(SalaryNormalized ~ ContractTime, data =
tr)
summary(lm.fit3)
lm.pred3 <- predict(lm.fit3, newdata = val)
sqrt(mean((lm.pred3 - val$SalaryNormalized)^2))
```



Hình 30: Liên hệ giữa loại hợp đồng và mức lương



Hình 31: Phân tích kiểm tra mối liên hệ giữa loại hợp đồng và mức lương

3.3.5. Phân tích ảnh hưởng của địa điểm làm việc lên mức lương

Phương trình hồi quy một biến địa điểm làm việc ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

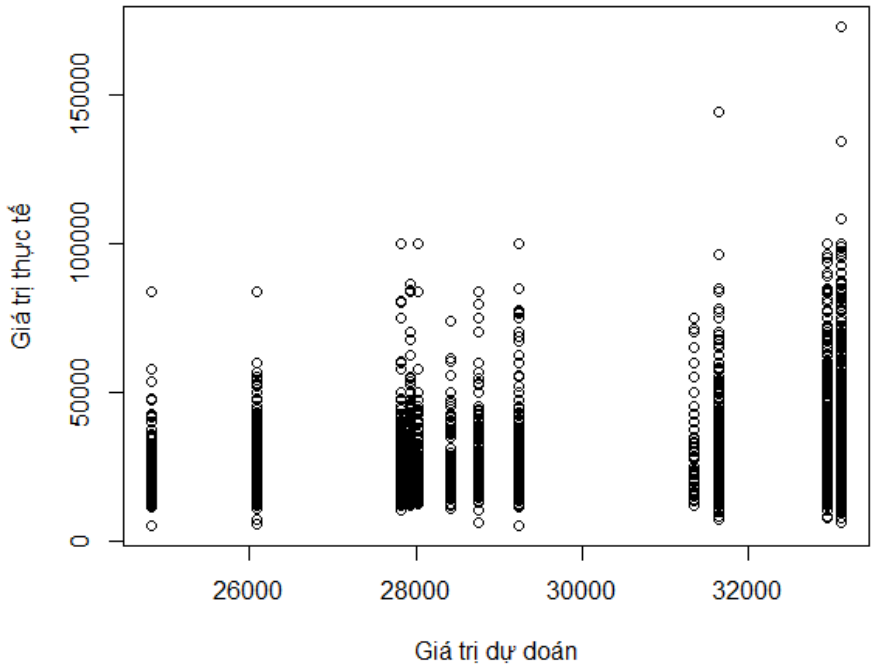
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

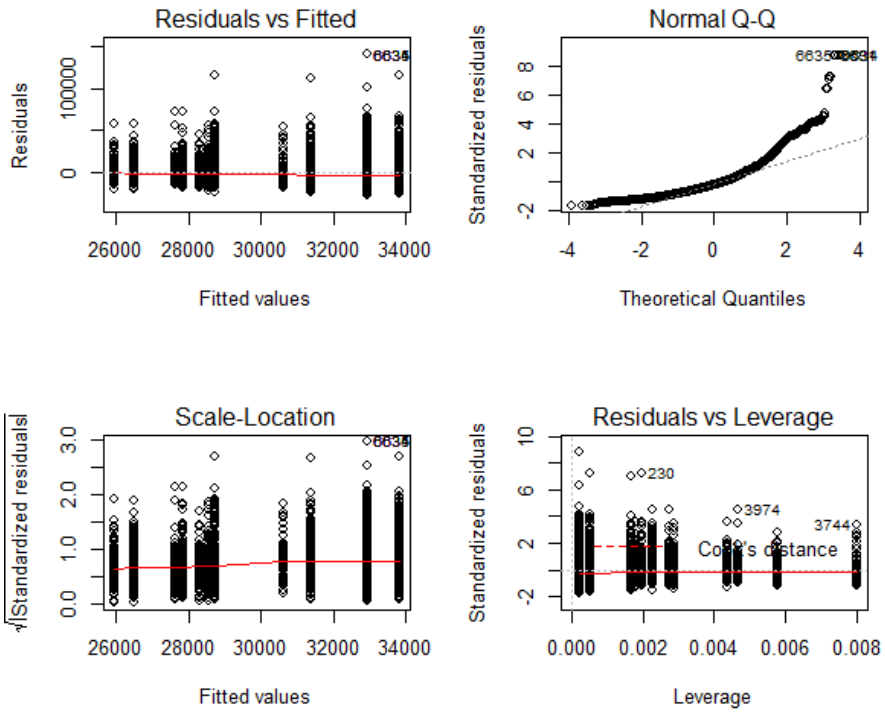
- Y: là biến Mức lương
- X_1: là biến Địa điểm làm việc

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit4 <- lm(SalaryNormalized ~ Location, data = tr)
summary(lm.fit4)
lm.pred4 <- predict(lm.fit4, newdata = val)
sqrt(mean((lm.pred4 - val$SalaryNormalized)^2))
```



Hình 32: Liên hệ giữa địa điểm làm việc và mức lương



Hình 33: Phân tích kiểm tra mối liên hệ giữa địa điểm làm việc và mức lương

3.3.6. Phân tích ảnh hưởng của địa điểm làm việc là Luân Đôn lên mức lương

Phương trình hồi quy một biến địa điểm làm việc là Luân Đôn ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

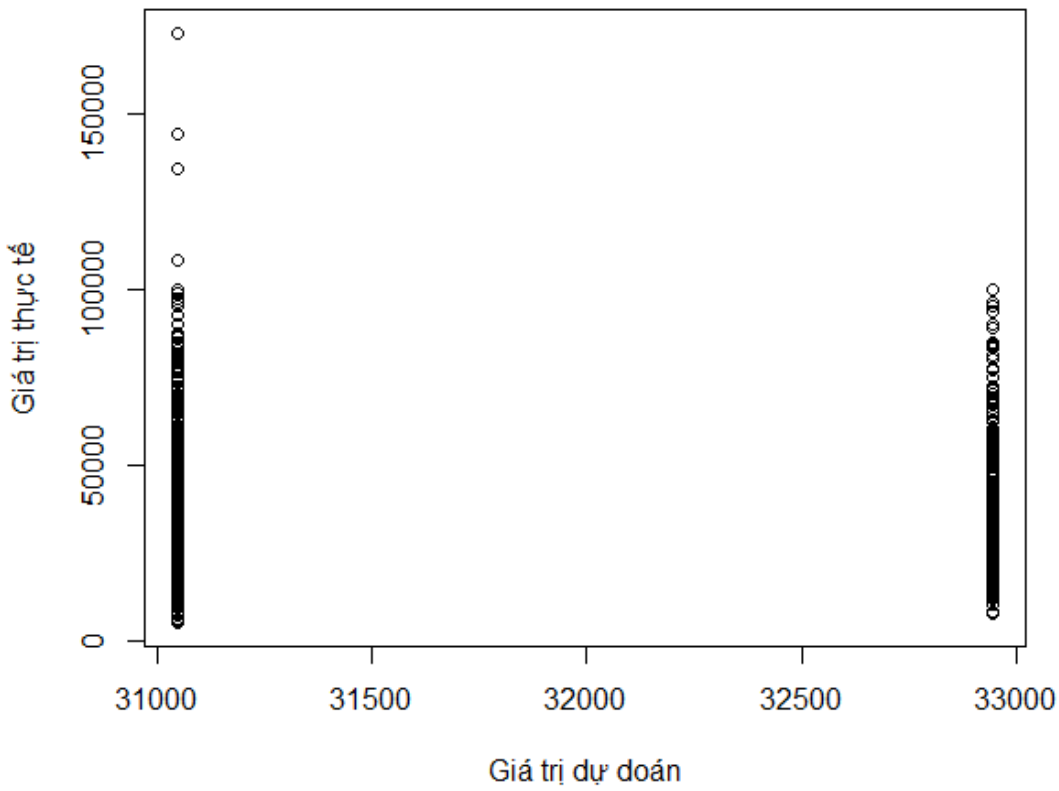
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

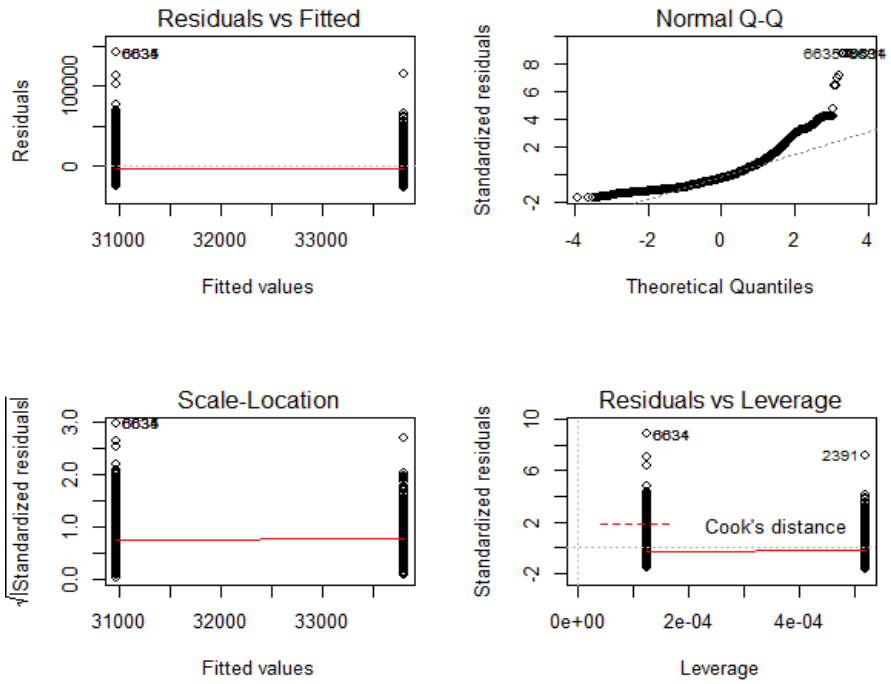
- Y: là mức lương
- X_1: là biến địa điểm làm việc được phân loại là Luân Đôn

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit5 <- lm(SalaryNormalized ~ London, data = tr)
summary(lm.fit5)
lm.pred5 <- predict(lm.fit5, newdata = val)
sqrt(mean((lm.pred5 - val$SalaryNormalized)^2))
```



Hình 34: Liên hệ giữa địa điểm làm việc là Luân Đôn và mức lương



Hình 35: Phân tích kiểm tra mối liên hệ giữa địa điểm làm việc là Luân Đôn và mức lương

3.3.7. Phân tích ảnh hưởng của tiêu đề công việc cho vị trí ứng viên có kinh nghiệm lên mức lương

Phương trình hồi quy một biến tiêu đề công việc cho vị trí ứng viên có kinh nghiệm ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

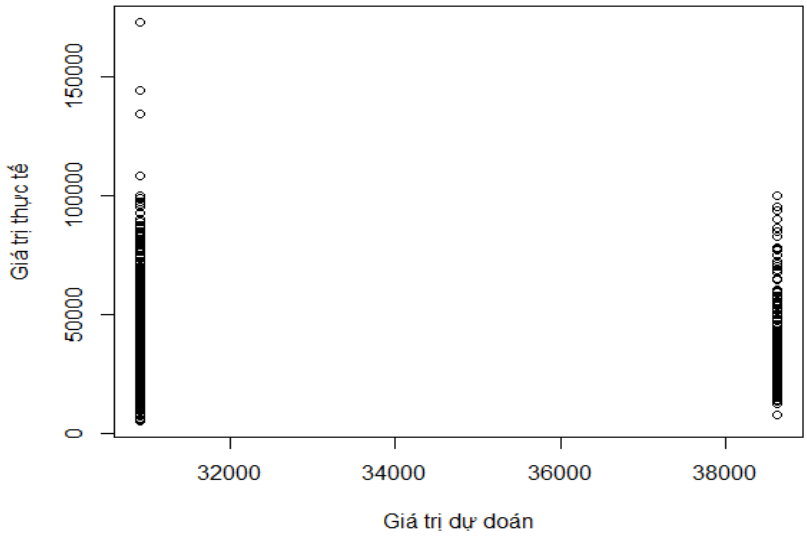
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

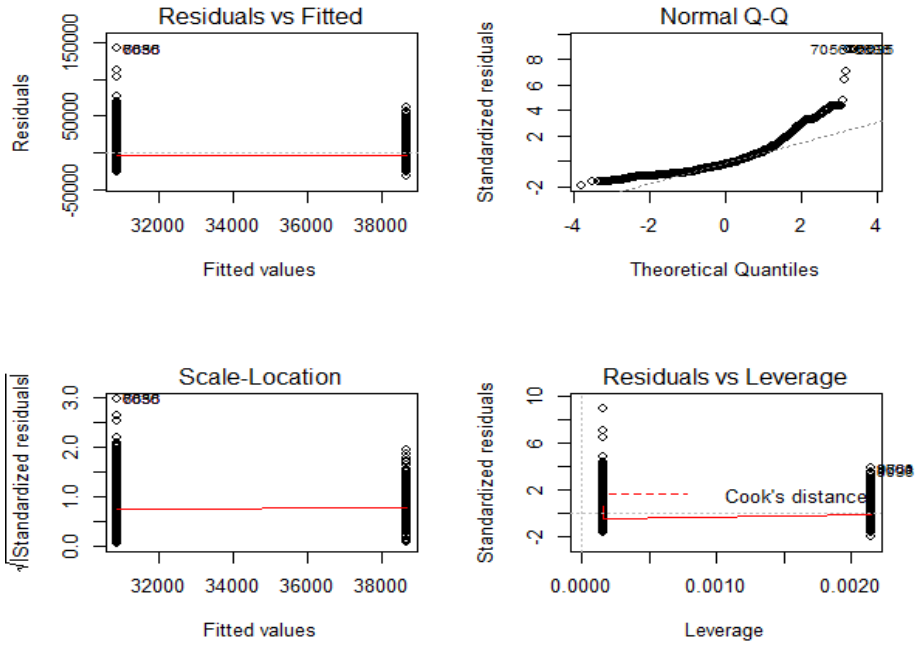
- Y: là biến Mức lương
- X_1: là biến Tiêu đề công việc cho vị trí ứng viên có kinh nghiệm

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit6 <- lm(SalaryNormalized ~ TitleSenior, data = tr)
summary(lm.fit6)
lm.pred6 <- predict(lm.fit6, newdata = val)
sqrt(mean((lm.pred6 - val$SalaryNormalized)^2))
```



Hình 36: Liên hệ giữa tiêu đề công việc cho vị trí ứng viên có kinh nghiệm và mức lương



Hình 37: Phân tích kiểm tra mối liên hệ giữa tiêu đề công việc cho vị trí ứng viên có kinh nghiệm và mức lương

3.3.8. Phân tích ảnh hưởng của tiêu đề công việc cho vị quản lý lên mức lương

Phương trình hồi quy một biến tiêu đề công việc cho vị trí quản lý ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

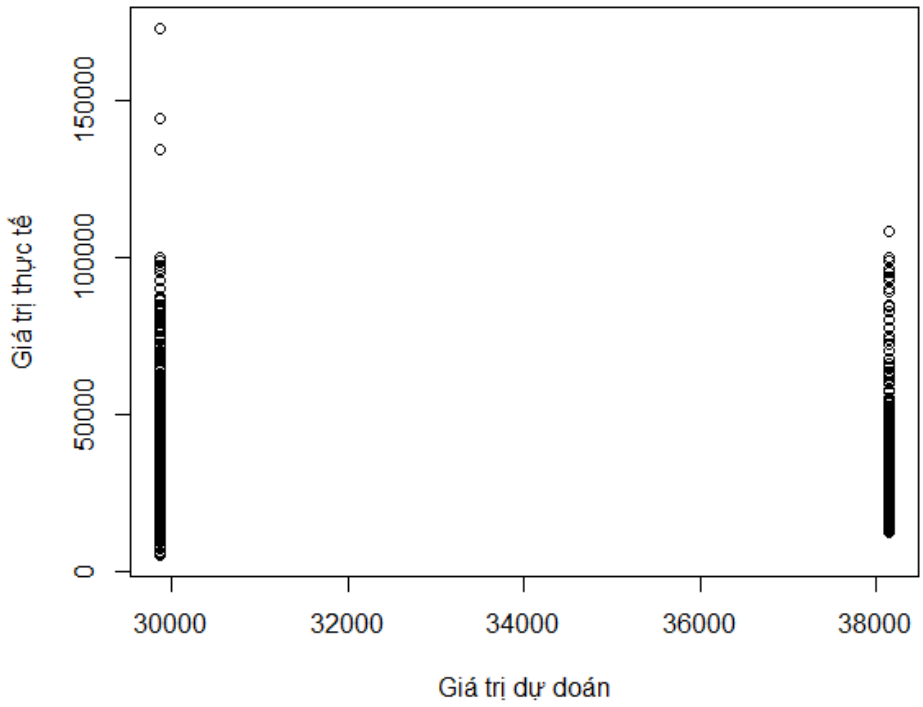
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

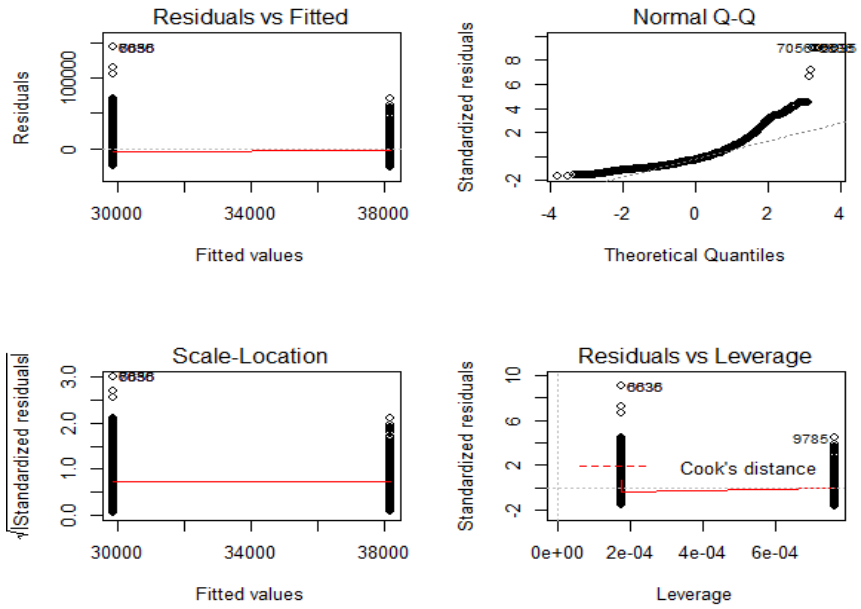
- Y: là biến Mức lương
- X_1: là biến Tiêu đề công việc cho vị trí quản lý

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit7 <- lm(SalaryNormalized ~ TitleManage, data = tr)
summary(lm.fit7)
lm.pred7 <- predict(lm.fit7, newdata = val)
sqrt(mean((lm.pred7 - val$SalaryNormalized)^2))
```



Hình 38: Liên hệ giữa tiêu đề công việc cho vị trí quản lý và mức lương



Hình 39: Phân tích kiểm tra mối liên hệ giữa tiêu đề công việc cho vị trí quản lý và mức lương

3.3.9. Phân tích ảnh hưởng của mô tả công việc cho vị trí ứng viên có kinh nghiệm lên mức lương

Phương trình hồi quy một biến mô tả công việc cho vị trí ứng viên có kinh nghiệm ảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

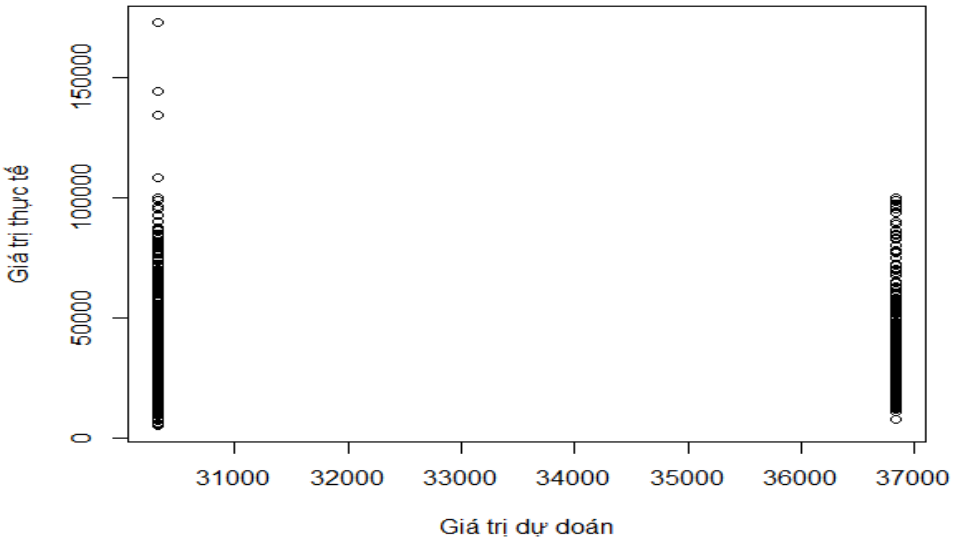
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

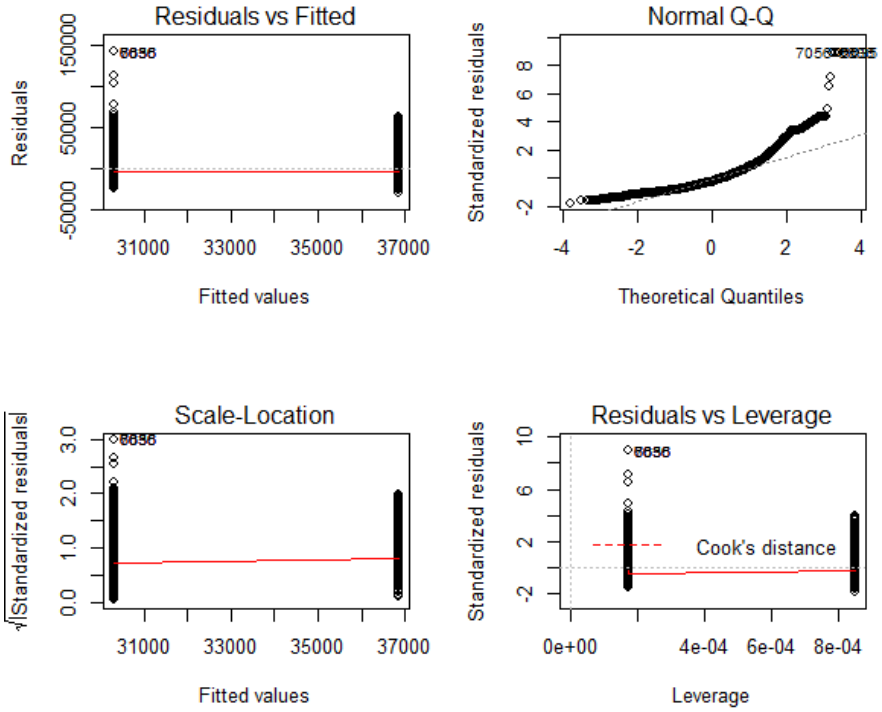
- Y: là biến Mức lương
- X_1: là biến Mô tả công việc cho vị trí ứng viên có kinh nghiệm

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit8 <- lm(SalaryNormalized ~ DescripSenior, data =
tr)
summary(lm.fit8)
lm.pred8 <- predict(lm.fit8, newdata = val)
sqrt(mean((lm.pred8 - val$SalaryNormalized)^2))
```



Hình 40: Liên hệ giữa mô tả công việc cho vị trí ứng viên có kinh nghiệm và mức lương



Hình 41: Phân tích kiểm tra mối liên hệ giữa mô tả công việc cho vị trí ứng viên có kinh nghiệm và mức lương

3.3.10.Phân tích ảnh hưởng của mô tả công việc cho vị trí quản lý lên mức lương

Phương trình hồi quy một biến mô tả công việc cho vị trí quản lýảnh hưởng lên mức lương quảng cáo tuyển dụng như sau:

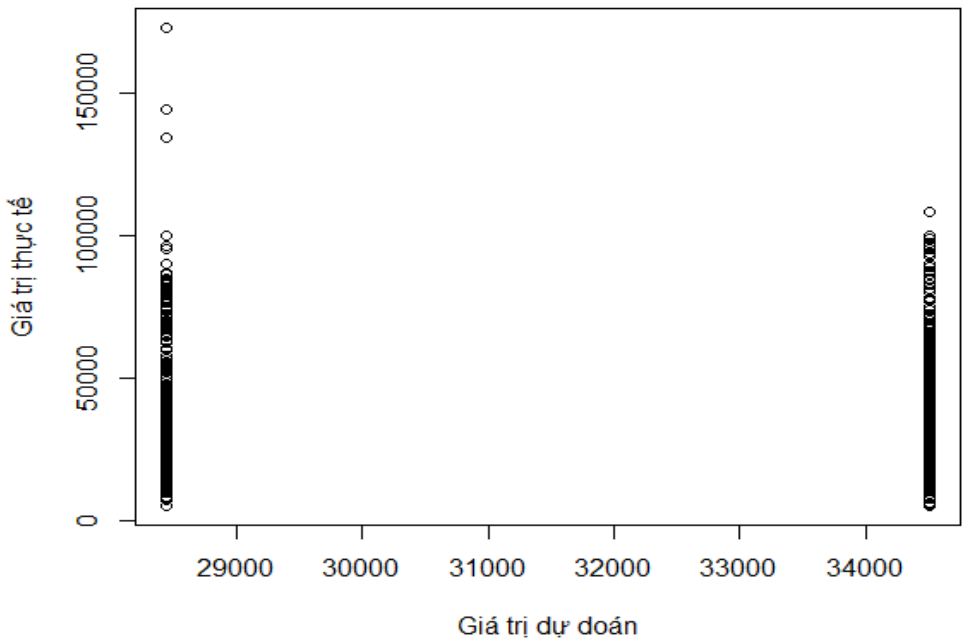
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Trong đó:

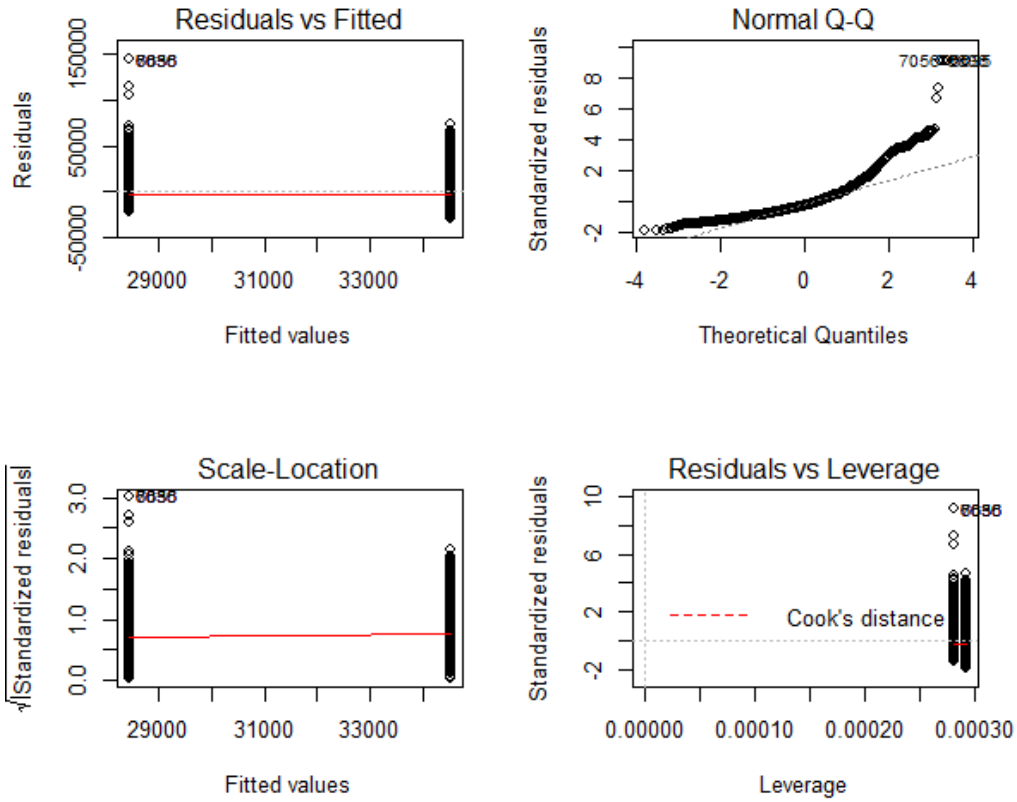
- Y: là biến Mức lương
- X_1: là biến Mô tả công việc cho vị trí quản lý

Phương trình được xây dựng trong công cụ R như sau:

```
lm.fit9 <- lm(SalaryNormalized ~ DescripManage, data =
tr)
summary(lm.fit9)
lm.pred9 <- predict(lm.fit9, newdata = val)
sqrt(mean((lm.pred9 - val$SalaryNormalized)^2))
```



Hình 42: Liên hệ giữa mô tả công việc cho vị trí quản lý và mức lương



Hình 43: Phân tích kiểm tra mối liên hệ giữa mô tả công việc cho vị trí quản lý và mức lương

3.3.11. Mô hình 0

Mô hình được xây dựng trên cơ sở chia tập dữ liệu làm 2 phần để phân tích: Tập huấn luyện và Tập kiểm tra:

Mô hình được xây dựng trong R như sau:

```
set.seed(100)
train.index <- sample(10000, 7000, replace = FALSE)
tr <- train[train.index, ]
val <- train[-train.index, ]
```

Kiểm tra giá trị sai số trung bình mô hình này:

```
sqrt(mean((mean(tr$SalaryNormalized) - 
val$SalaryNormalized)^2))
```

Sai số trung bình của mô hình này là **16200.71** tác giả đi xây dựng những mô hình hồi quy kế tiếp để so sánh với mô hình này.

3.3.12. Mô hình 1

Xây dựng mô hình dự đoán mức lương sử dụng 3 biến độc lập: nhóm công việc, loại hợp đồng, và loại công việc như sau:

Mô hình:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Trong đó:

Y: là mức lương công việc

X_1: loại công việc

X_2: loại hợp đồng

X_3: nhóm công việc

Mô hình được xây dựng trong R như sau:

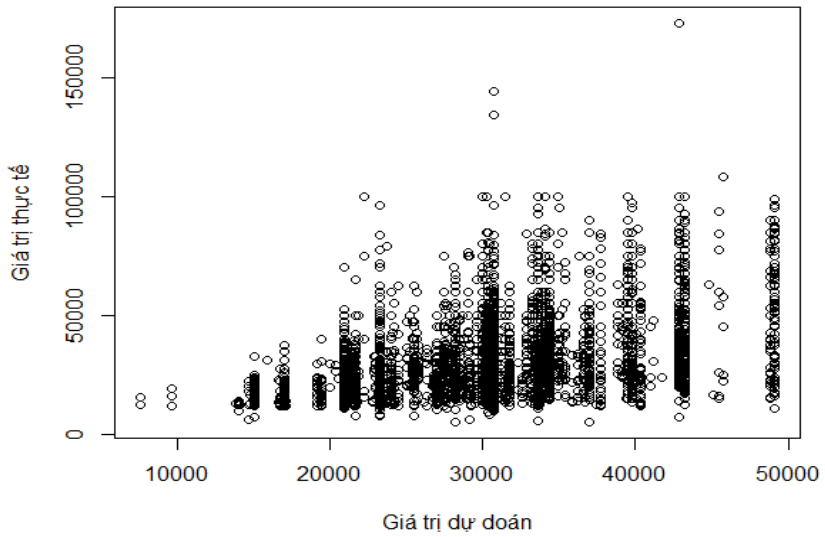
```
lm.fit10 <- lm(SalaryNormalized ~ ContractType +
ContractTime + Category, data =tr)
summary(lm.fit10)
lm.pred10 <- predict(lm.fit10, newdata = val)
```

Kiểm tra giá trị sai số trung bình của mô hình 1:

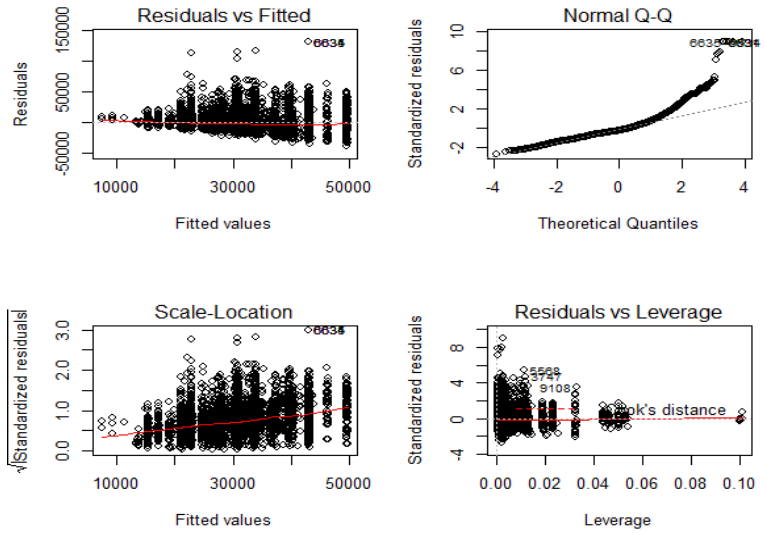
```
sqrt(mean((lm.pred10 - val$SalaryNormalized)^2))
```

Kết quả sai số trung bình mô hình 1 là **14701.27**

Với kết quả này so sánh với mô hình cơ bản ban đầu là **16200.71** ta thấy cách tiếp cần với mô hình hồi quy tuyến tính cải thiện được sai số trung bình khoản **1499.44**. Tiếp tục với cách tiếp cận này luận văn xây dựng tiếp các mô hình với nhiều biến độc lập hơn để phân tích mức độ ảnh hưởng và mức cải thiện sai số qua từng mô hình. Từ đó tìm ra mô hình tốt nhất trong bài toán dự đoán mức lương theo phương pháp hồi quy tuyến tính.



Hình 44: Liên hệ giữa Nhóm công việc, Loại công việc, và Loại hợp đồng ảnh hưởng lên Mức lương



Hình 45: Phân tích kiểm tra mối liện hệ giữa nhóm công việc, loại công việc và loại hợp đồng ảnh hưởng lên mức lương

3.3.13. Mô hình 2

Xây dựng mô hình hồi quy dự đoán mức lương sử dụng 4 biến độc lập: nhóm công việc, loại hợp đồng, loại công việc và địa điểm làm việc như sau:

Mô hình:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Trong đó:

Y: là mức lương công việc

X_1: loại công việc

X_2: loại hợp đồng

X_3: nhóm công việc

X_4: địa điểm làm việc

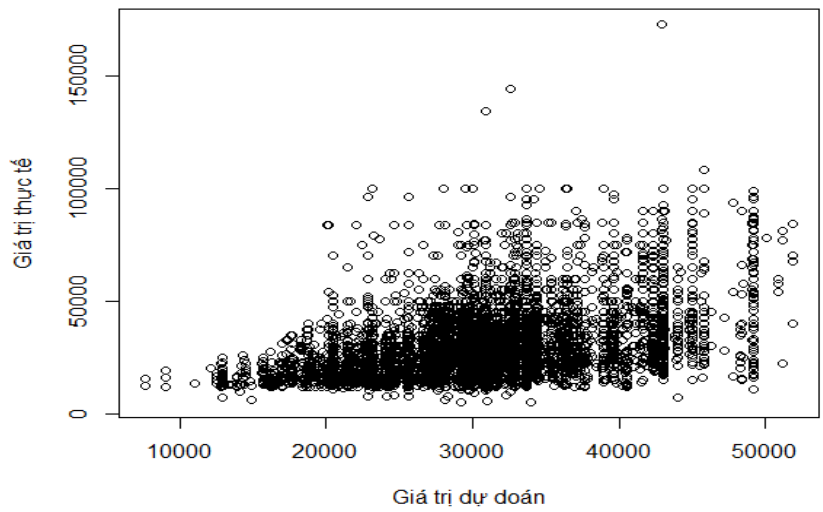
Mô hình được xây dựng trong R như sau:

```
lm.fit11 <- lm(SalaryNormalized ~ ContractType +
ContractTime + Category + Location, data = tr)
summary(lm.fit11)
lm.pred11 <- predict(lm.fit11, newdata = val)
```

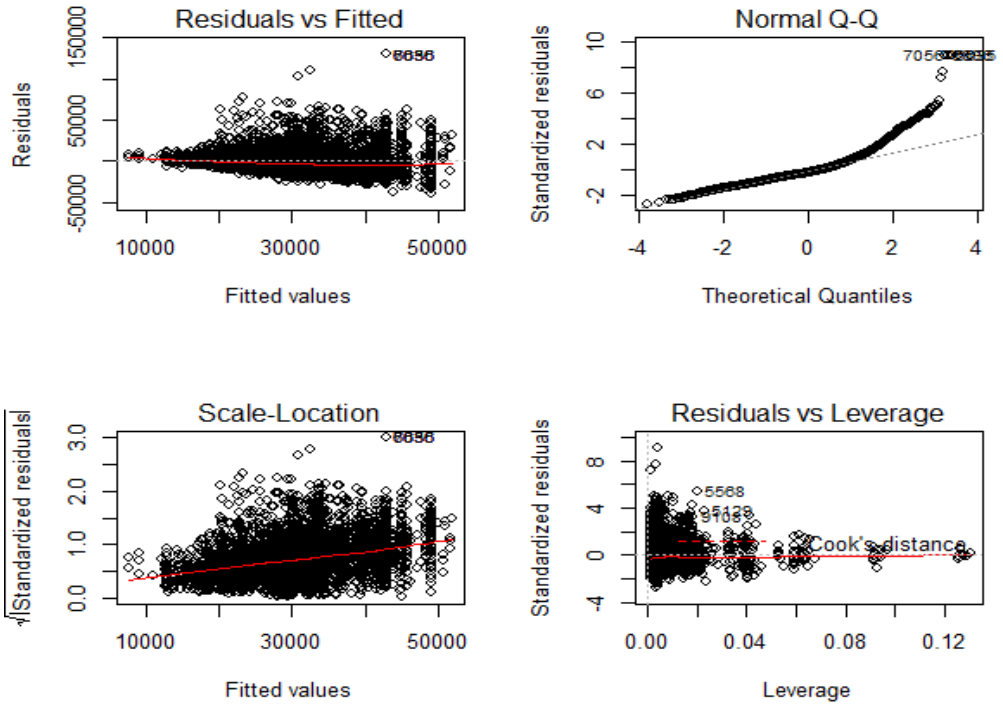
Kiểm tra giá trị sai số trung bình mô hình 2:

```
sqrt(mean((lm.pred11 - val$SalaryNormalized)^2))
```

Với việc thêm biến số "địa điểm làm việc" mô hình 3 cho kết quả cải thiện sai số là **14618.94** so với mô hình 2. Sai số trung bình giảm **82.33** sai số giảm tuy không cao nhưng cho thấy với việc thêm biến địa điểm đã cho kết quả khả quan.



Hình 46: Liên hệ giữa Nhóm công việc, loại công việc, loại hợp đồng và địa điểm làm việc ảnh hưởng lên mức lương



Hình 47: Phân tích kiểm tra mối liên hệ nhóm công việc, loại công việc, loại hợp đồng và địa điểm làm việc ảnh hưởng lên mức lương

3.3.14. Mô hình 3

Xây dựng mô hình hồi quy dự đoán mức lương dựa trên các biến độc lập sau: nhóm công việc, loại hợp đồng, loại công việc và địa điểm làm việc là Luân Đôn:

Mô hình:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Trong đó:

Y: là mức lương công việc

X_1: loại công việc

X_2: loại hợp đồng

X_3: nhóm công việc

X_4: địa điểm làm việc là Luân Đôn

Mô hình được xây dựng trong R như sau:

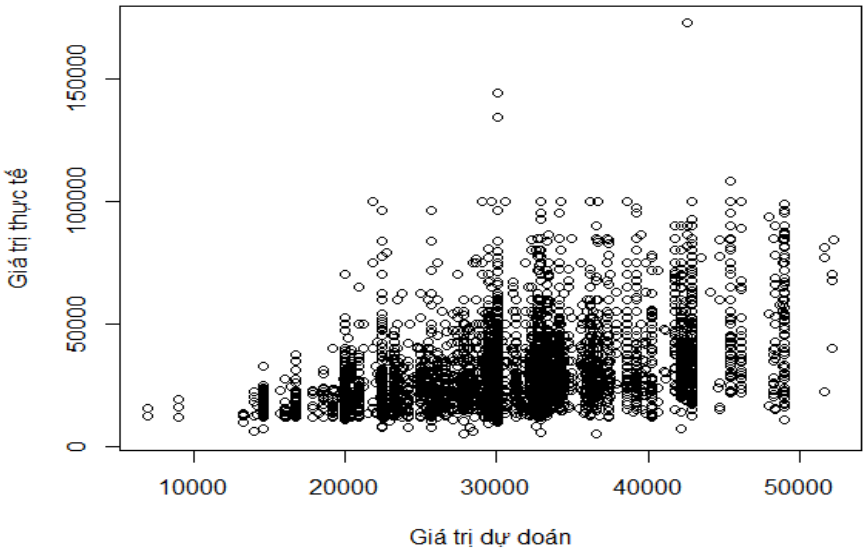
```
lm.fit12 <- lm(SalaryNormalized ~ ContractType +
ContractTime + Category + London, data = tr)
```

```
summary(lm.fit12)
lm.pred12 <- predict(lm.fit12, newdata = val)
```

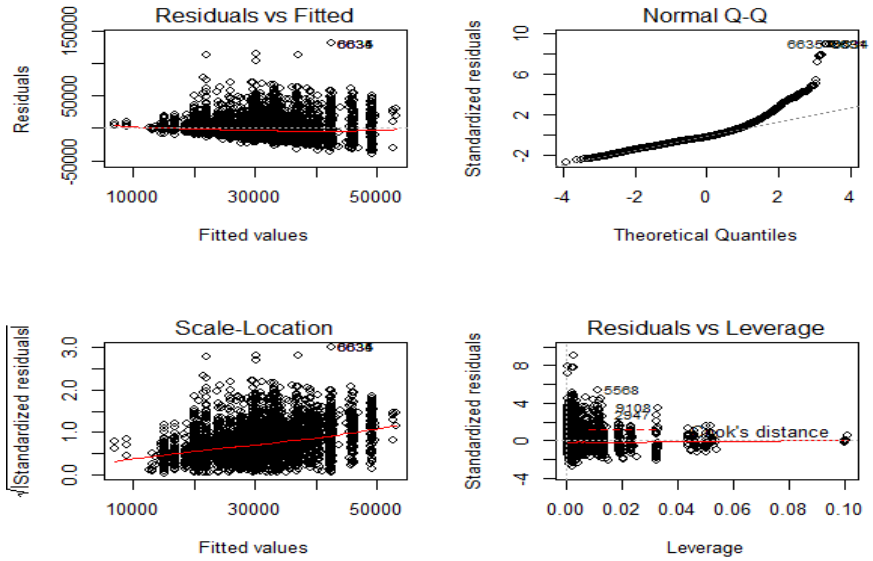
Kiểm tra sai số trung bình mô hình 3:

```
sqrt(mean((lm.pred12 - val$SalaryNormalized)^2))
```

Với mô hình này việc phân loại biến địa điểm làm việc theo địa điểm cụ thể (trong phạm vi luận văn này là Luân Đôn) cho thầy kết quả sai số của mô hình này là **14590.61**. Điều này cho thấy việc phân loại địa điểm thành địa điểm cụ thể cho kết quả sai số giảm **110.66** đơn vị. Với việc này chúng ta có thể dự đoán rằng việc phân loại tiêu đề công việc và mô tả công việc cũng sẽ cho kết quả rất tốt trong mô hình 4 tiếp theo.



Hình 48: Liên hệ giữa Nhóm công việc, loại công việc, loại hợp đồng và địa điểm là Luân Đôn ảnh hưởng lên mức lương



Hình 49: Phân tích kiểm tra mối liên hệ giữa nhóm công việc, loại công việc, loại hợp đồng và địa điểm là Luân Đôn ảnh hưởng lên mức lương

3.3.15. Mô hình 4

Xây dựng mô hình hồi quy dự đoán mức lương dựa trên các biến độc lập: nhóm công việc, loại hợp đồng, loại công việc, địa điểm làm việc là Luân Đôn, tiêu đề và mô tả công việc theo vị trí ứng viên có kinh nghiệm hoặc là vị trí quản lý được phân loại như mục 3.2.1 phần xử lý dữ liệu như sau:

Mô hình:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

Trong đó:

Y: là mức lương công việc

X_1: loại công việc

X_2: loại hợp đồng

X_3: nhóm công việc

X_4: địa điểm làm việc là Luân Đôn

X_5: tiêu đề công việc cho vị trí ứng viên có kinh nghiệm

X_6: tiêu đề công việc cho vị trí ứng viên quản lý

X_7: mô tả công việc cho vị trí ứng viên có kinh nghiệm

X_8: mô tả công việc cho vị trí ứng viên quản lý

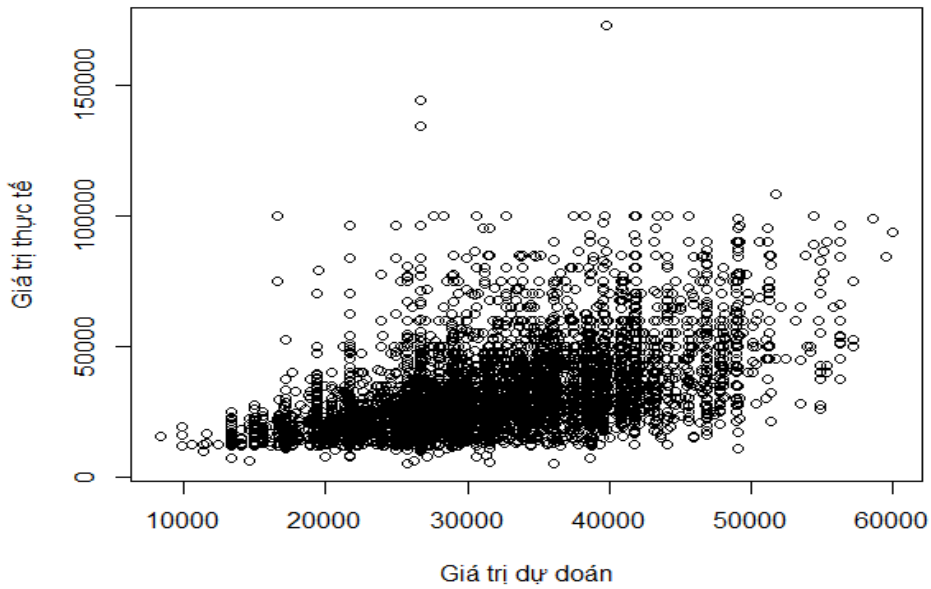
Mô hình được xây dựng trong R như sau:

```
lm.fit13 <- lm(SalaryNormalized ~ ContractType +
ContractTime + Category + London + TitleSenior +
TitleManage + DescripSenior + DescripManage, data = tr)
summary(lm.fit13)
lm.pred13 <- predict(lm.fit13, newdata = val)
```

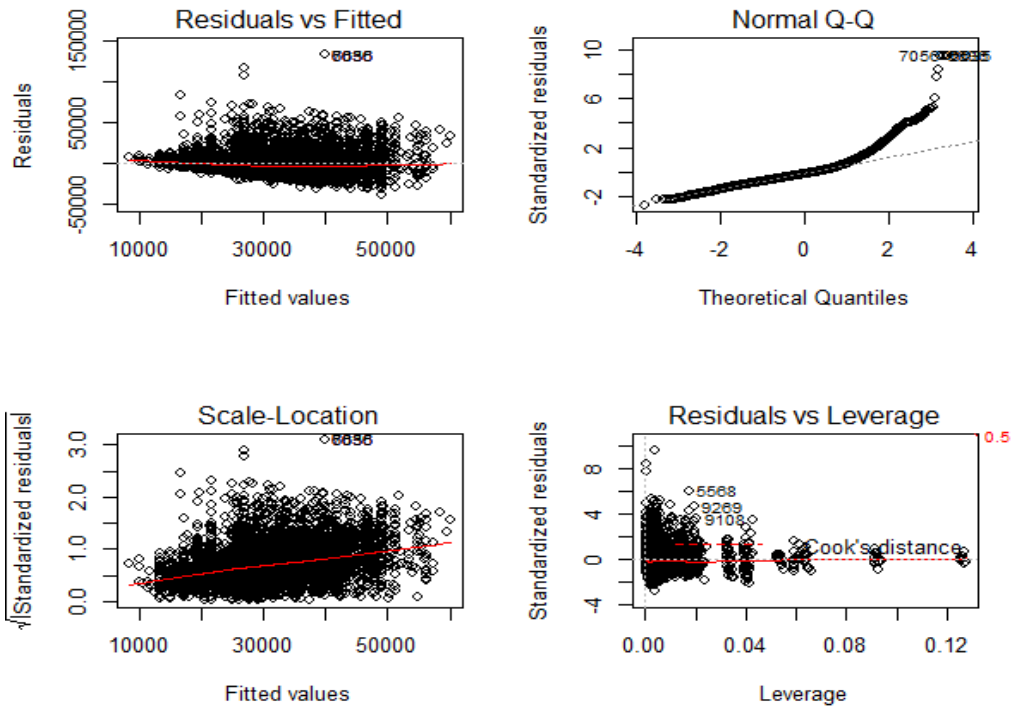
Kiểm tra giá trị sai số trung bình:

```
sqrt(mean((lm.pred13 - val$SalaryNormalized)^2))
```

Giá trị sai số trung bình mô hình này là **14088.45**



Hình 50: Liên hệ giữa Nhóm công việc, loại công việc, loại hợp đồng, địa điểm, tiêu đề và mô tả công việc ảnh hưởng lên mức lương



Hình 51: Phân tích kiểm tra mối liên hệ giữa nhóm công việc, loại công việc, loại hợp đồng, địa điểm, tiêu đề và mô tả công việc ảnh hưởng lên mức lương

Tác giả thấy việc thay đổi các biến trong phân tích hồi quy so với mô hình ban đầu có những biến đổi đáng kể. Ở mô hình 4 có giá trị sai số trung bình **14088.45** so với mô hình ban đầu là **16200.71**. Việc phân loại dữ liệu bằng cách thêm các đặc trưng địa điểm làm việc là vùng hoặc thành phố cụ thể (ví dụ tác giả dùng ở đây là Luân Đôn), Tiêu đề công việc và Mô tả công việc được phân loại thành các nhóm tiêu đề công việc hoặc mô tả công việc cho vị trí ứng viên có kinh nghiệm và vị trí quản lý đã cải thiện sai số trung bình rất đáng kể cụ thể là. Tác giả sẽ so sánh với một số mô hình khác trong phần đánh giá mô hình với những gì đạt được ở **mô hình 4**.

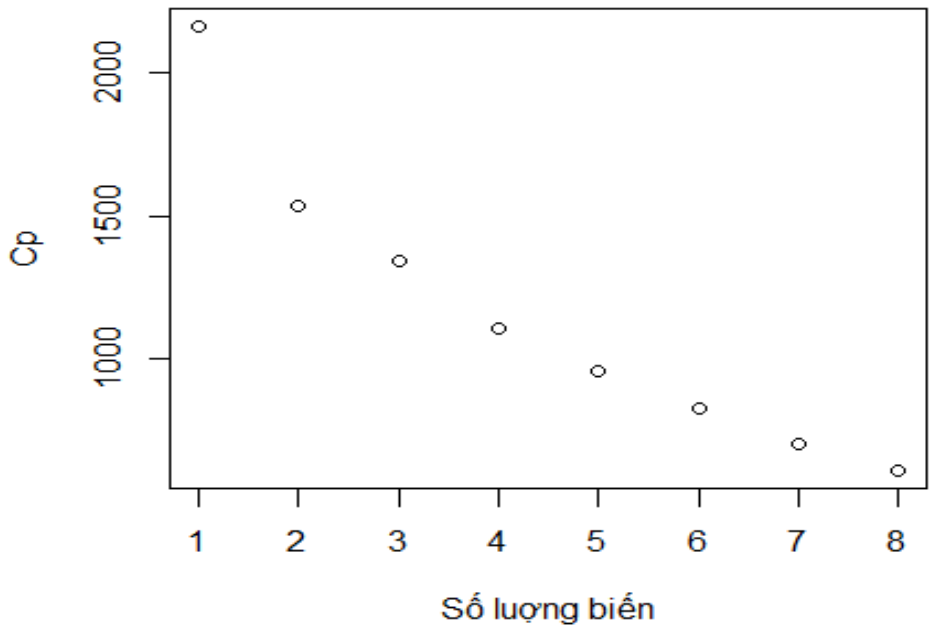
3.4. Đánh giá mô hình

3.4.1. Phương pháp lựa chọn từng bước

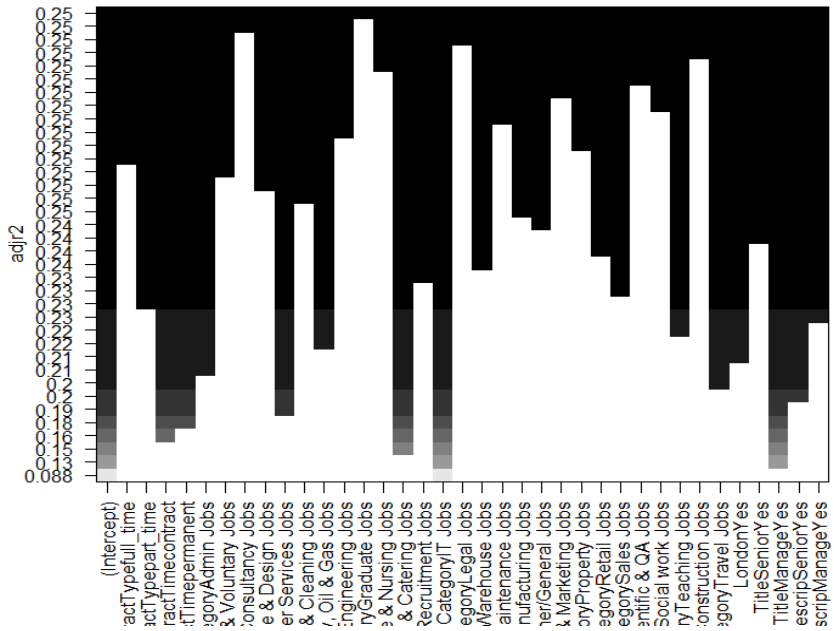
Lựa chọn theo từng bước về phía trước bắt đầu với một mô hình không có đặc trưng nào, sau đó đưa thêm đặc trưng vào cùng một lúc cho đến khi tất cả các đặc trưng được thêm vào. Một đặc trưng được lựa chọn được thêm vào trong tạo ra một mô hình với RSS thấp nhất. Vì vậy, về mặt lý thuyết, đặc trưng đầu tiên được lựa chọn

nên là nó giải thích các biến đáp ứng tốt hơn so với bất kỳ những đặc trưng khác, và tương tự.

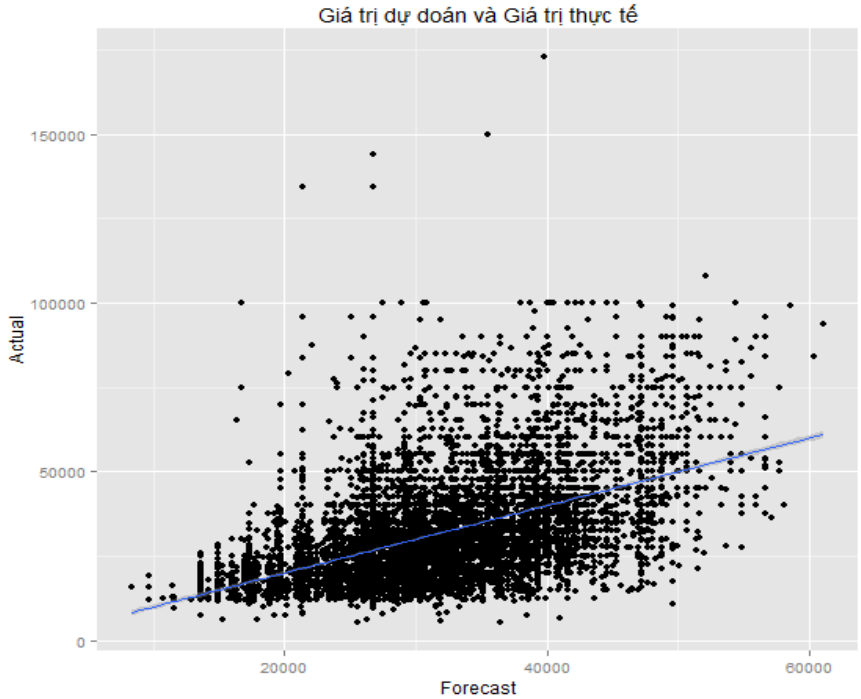
Dựa trên những sai số trung bình của các mô hình trên, so sánh để chọn ra mô hình tối ưu nhất trong phân tính dự đoán mức lương quảng cáo tuyển dụng.



Hình 52: Số lượng biến của mô hình và điểm Cp tương ứng



Hình 53: Chỉ số điều chỉnh giá trị trung bình nhỏ nhất

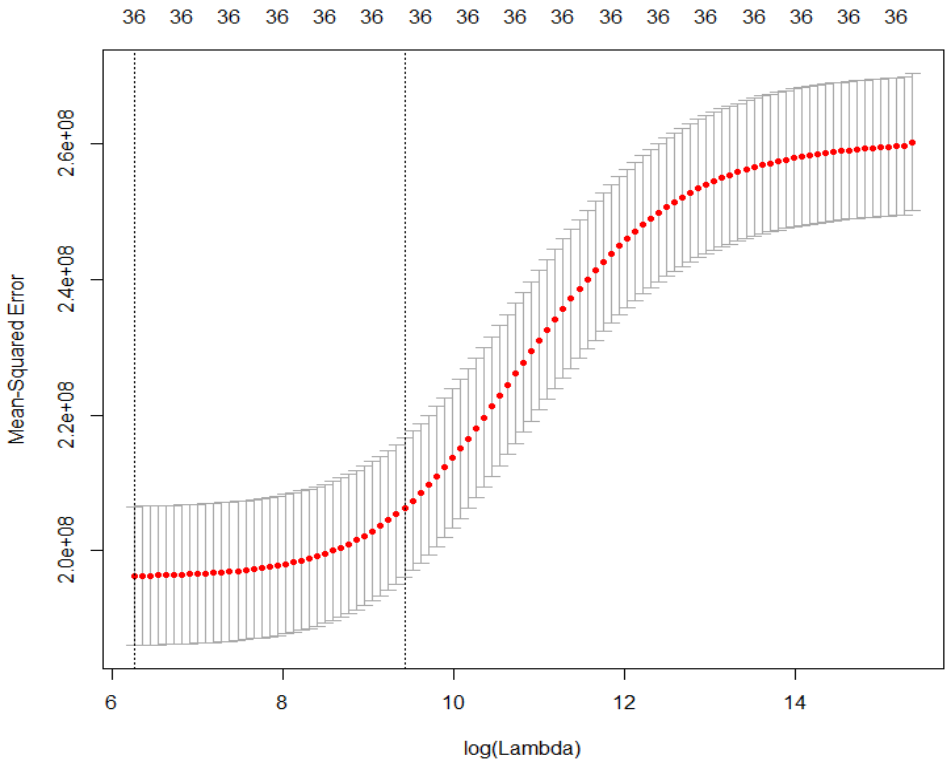


Hình 54: Thể hiện giá trị dự đoán so với giá trị thực tế của mô hình dự đoán dựa trên nhóm công việc, loại công việc, loại hợp đồng, địa điểm làm việc là Luân Đôn, tiêu đề và mô tả công việc

Sau khi lặp kiểm tra lựa trọn từng bươc trên Mô hình 4, kết quả giá trị sai số trung bình là **14088.45**. Tác giả thấy phương pháp lựa chọn từng bước cho kết quả tương tự như Mô hình 4.

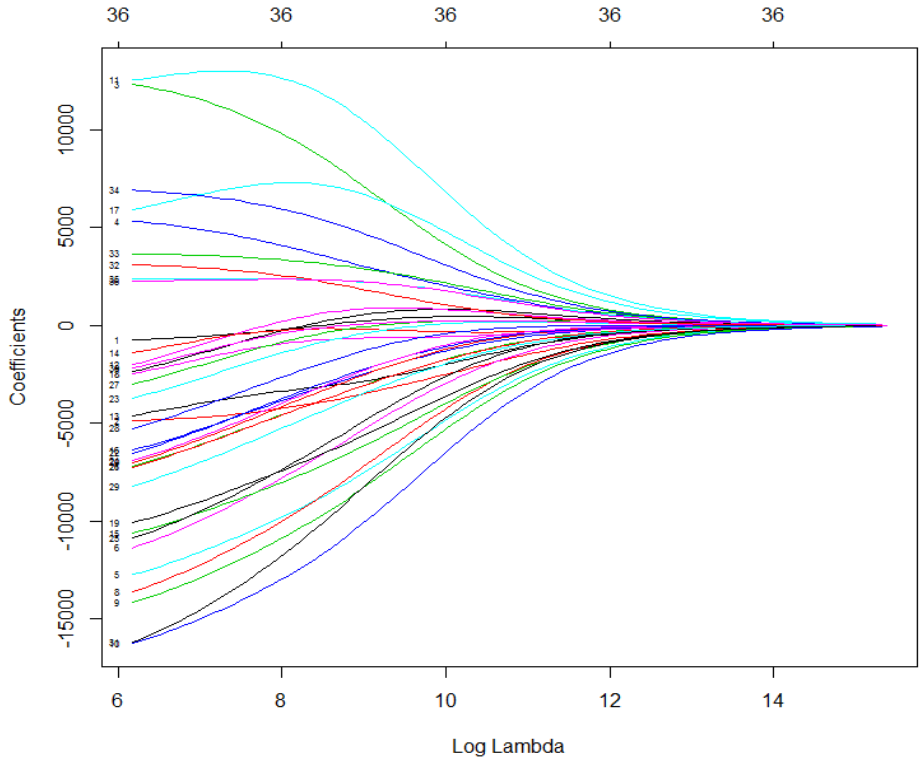
3.4.2. Mô hình hồi quy Ridge

Thử trên mô hình hồi quy Ridge sử dụng các đặc trưng là các biến độc lập trong mô hình 4 để xây dựng ma trận cho tập huấn luyện và tập kiểm tra và thu được giá trị Lambda tốt nhất.



Hình 55: Hệ số Lambda và sai số trung bình

Xây dựng mô hình dự đoán với giá trị Lambda tốt nhất:

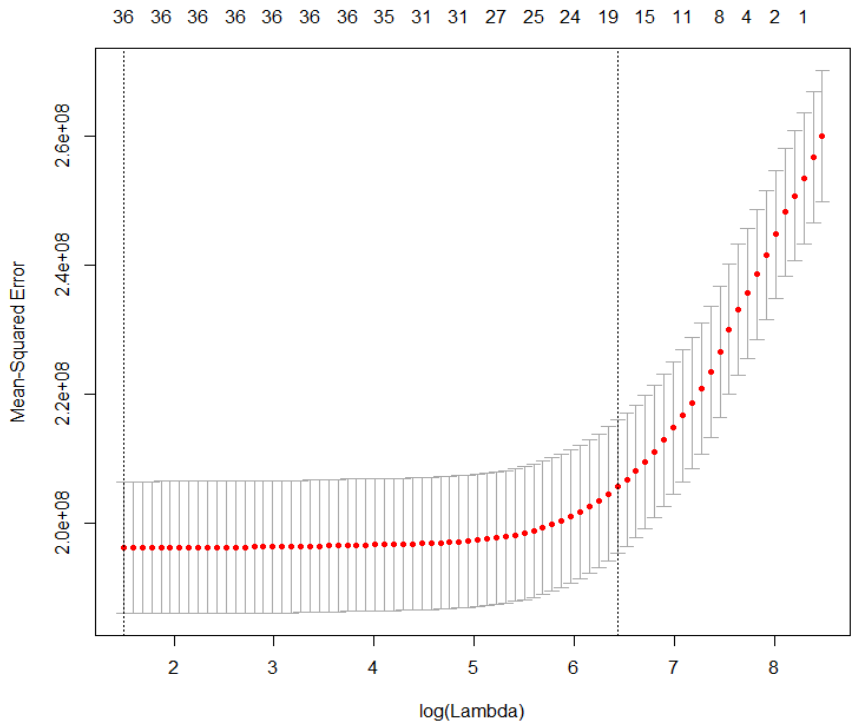


Hình 56: Hệ số tương quan và hệ số Lambda

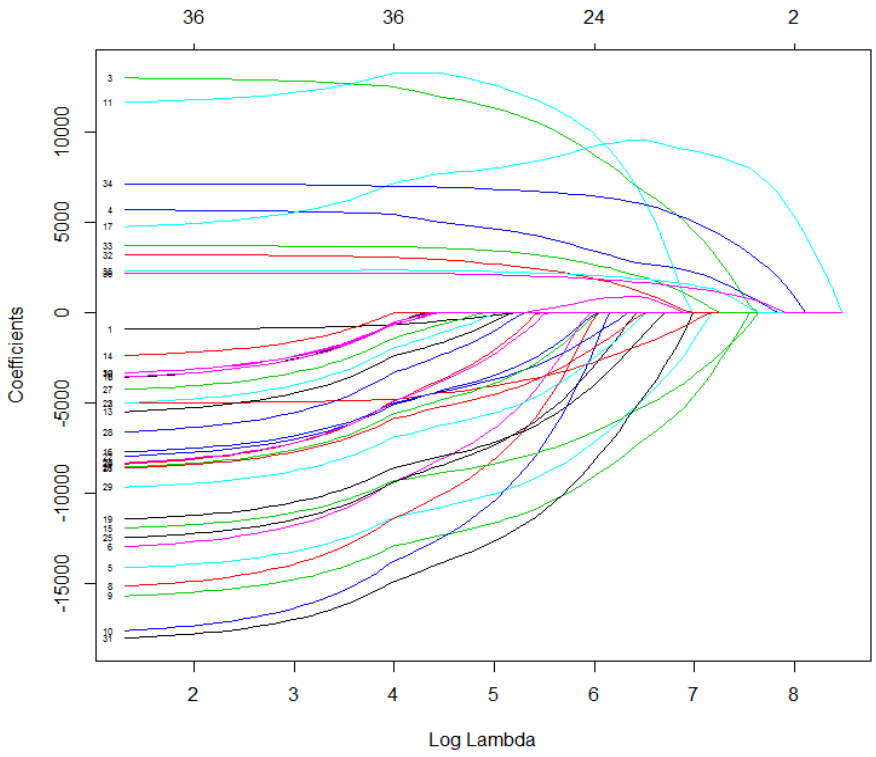
Kết quả sai số trung bình mô hình hồi quy Ridge là **14093.25**, ta thấy mô hình hồi quy Ridge cũng cho kết quả sai số gần giống như **mô hình 4**.

3.4.3. Mô hình Lasso

Mô hình hồi quy Ridge cũng cho kết quả rất tốt. Câu hỏi đặt ra là liệu mô hình Lasso có thể làm tốt hơn nữa hay không. Tiếp tục phân tích với mô hình Lasso bằng việc cũng tìm ra giá trị Lambda tốt nhất sau đó xây dựng mô hình dự đoán trên giá trị Lamda đó.



Hình 57: Giá trị Lambda trong mô hình Lasso



Hình 58: Biểu đồ hệ số tương quan và giá trị Lambda

Kết quả sai số trung bình của mô hình Lasso là **14088.92**. Một lần nữa mô hình Lasso cũng cho kết quả tương tự như **Mô hình 4**.

3.4.4. Kiểm tra với bộ dữ liệu giả định

Bộ dữ liệu tình huống có 5000 dòng dữ liệu so với tập dữ liệu huân luyện. Tác giả kiểm tra mô hình 4 có làm tốt trên bộ dữ liệu này hay không.

Kiểm tra giá trị sai số trung bình của toàn bộ dữ liệu huấn luyện trên bộ dữ liệu tình huống này trong R:

```
sqrt(mean((mean(train$SalaryNormalized) -
solution$SalaryNormalized)^2))
```

Giá trị sai số trung bình là **18760.7**. Huấn luyện mô hình hồi quy (mô hình 4) trên toàn bộ tập dữ liệu huấn luyện là 10000 dòng dữ liệu. Và kiểm tra lại với bộ dữ liệu tình huống trong công cụ R như sau:

```
lm.fit5 <- lm(SalaryNormalized ~ ContractType +
ContractTime + Category + London + TitleSenior +
TitleManage + DescripSenior + DescripManage, data =
train)
lm.pred5 <- predict(lm.fit5, newdata = solution)
```

Kiểm tra giá trị sai số trung bình:

```
sqrt(mean((lm.pred5 - solution$SalaryNormalized)^2))
```

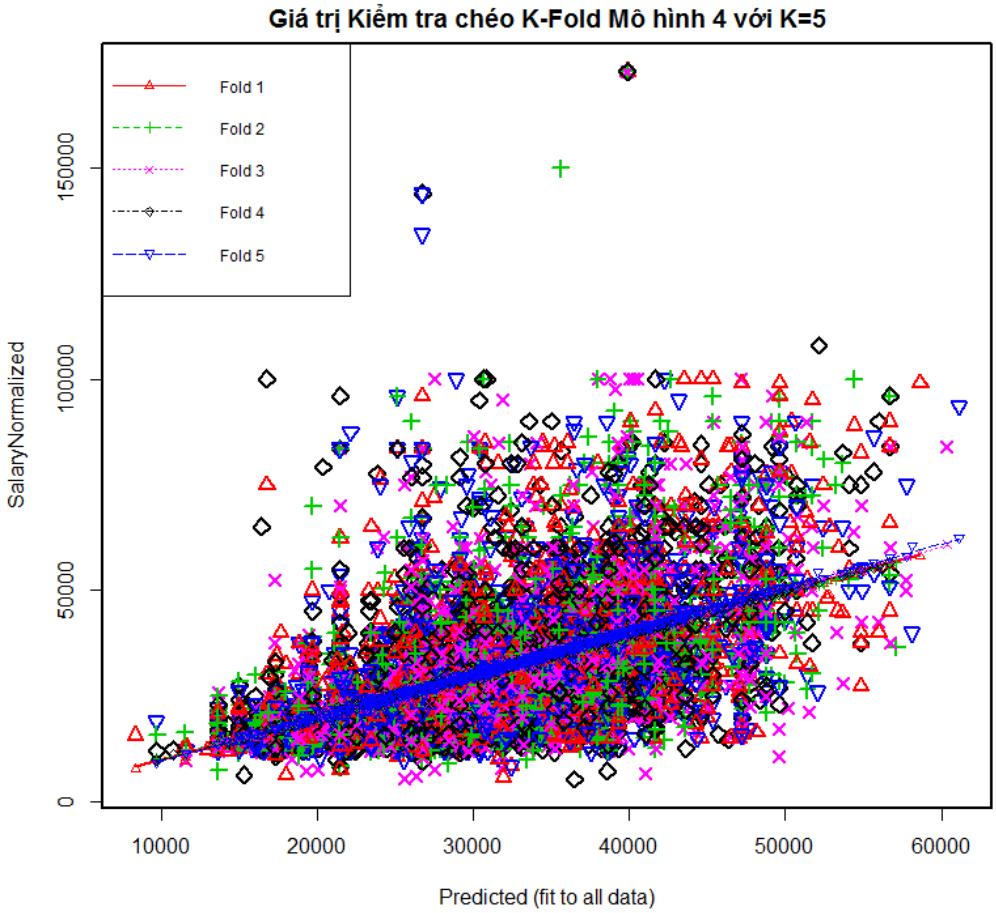
Kết quả giá trị sai số trung bình là **16469.98** giảm so với **18760.7** khi thực hiện kiểm tra với bộ dữ liệu tình huống.

Sự giảm sai số trung bình so với mô hình ban đầu tương tự với những gì chúng ta đã thấy Mô hình 4 làm trên tập kiểm tra, và ở đây Mô hình 4 cũng đã làm tốt trên tập dữ liệu tình huống. Điều này cho thấy việc việc thêm một vài đặc trưng trong phân tích từ Tiêu đề công việc và Mô tả đầy đủ công viêc. Với cách làm đó độ tin cậy của mô hình dự đoán cải thiện rất đáng kể để làm tốt công việc dự đoán.

3.5. Kiểm tra chéo với K-Fold

Sau khi đánh giá trên mô hình 4 qua các phương pháp lựa chọn từng bước, mô hình hồi quy Ridge, Lasso và bộ dữ liệu giả định được cung cấp kèm theo bởi Kaggle. Kết quả của việc đánh giá đó cho thấy mô hình 4 cho kết quả phân tích dự đoán với sai số trung bình nhỏ nhất điều đó có nghĩa là giá trị dự đoán trên mô hình 4 rất đáng tin cậy. Và sau đây tác giả tiếp tục thực hiện kiểm tra chéo bằng phương pháp K-Fold với mô hình 4 với một số k tập con như sau:

- Kiểm tra chéo K-Fold trên Mô hình 4 với k=5:



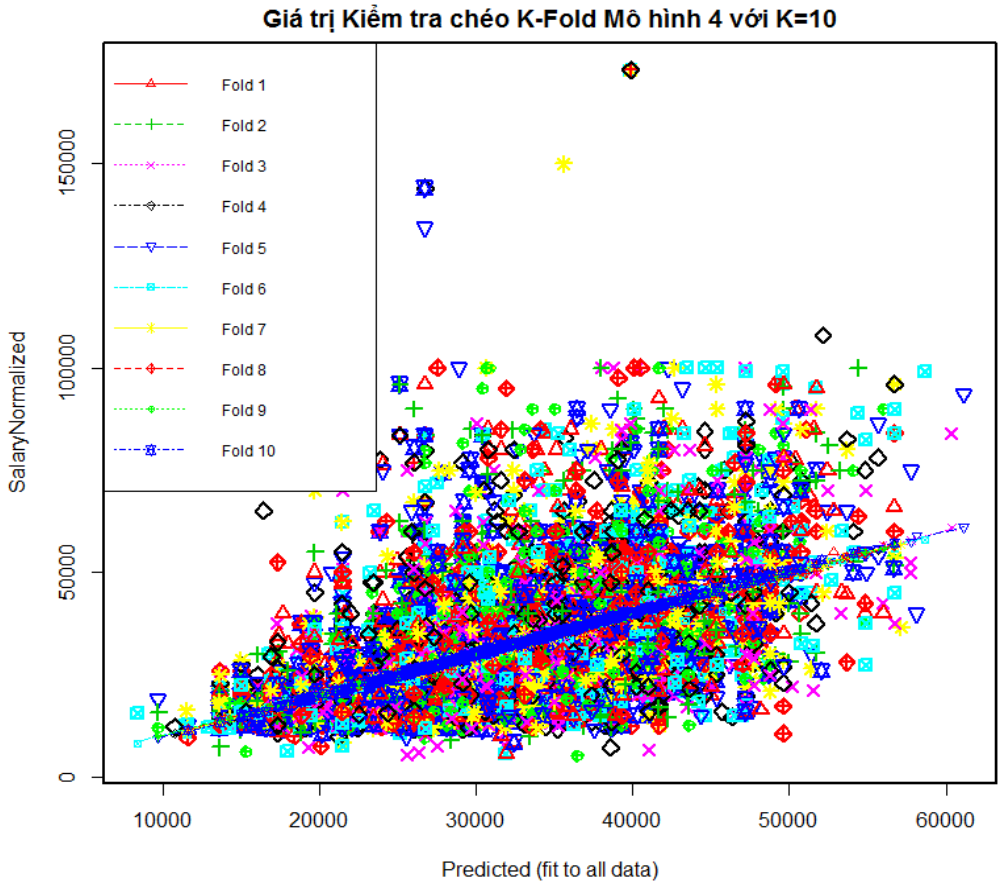
Hình 59: Kết quả kiểm tra chéo mô hình 4 với k = 5

Bảng 5: Kết quả sai số trung bình với k=5

Tập K	Sai số trung bình
k=1	13820.27
k=2	13928.39
k=3	13784.05
k=4	14798.65
k=5	13711.31 (nhỏ nhất)

Giá trị sai số trung bình với k=5 là: **14008.53**

- Kiểm tra chéo K-Fold trên Mô hình 4 với k=10:



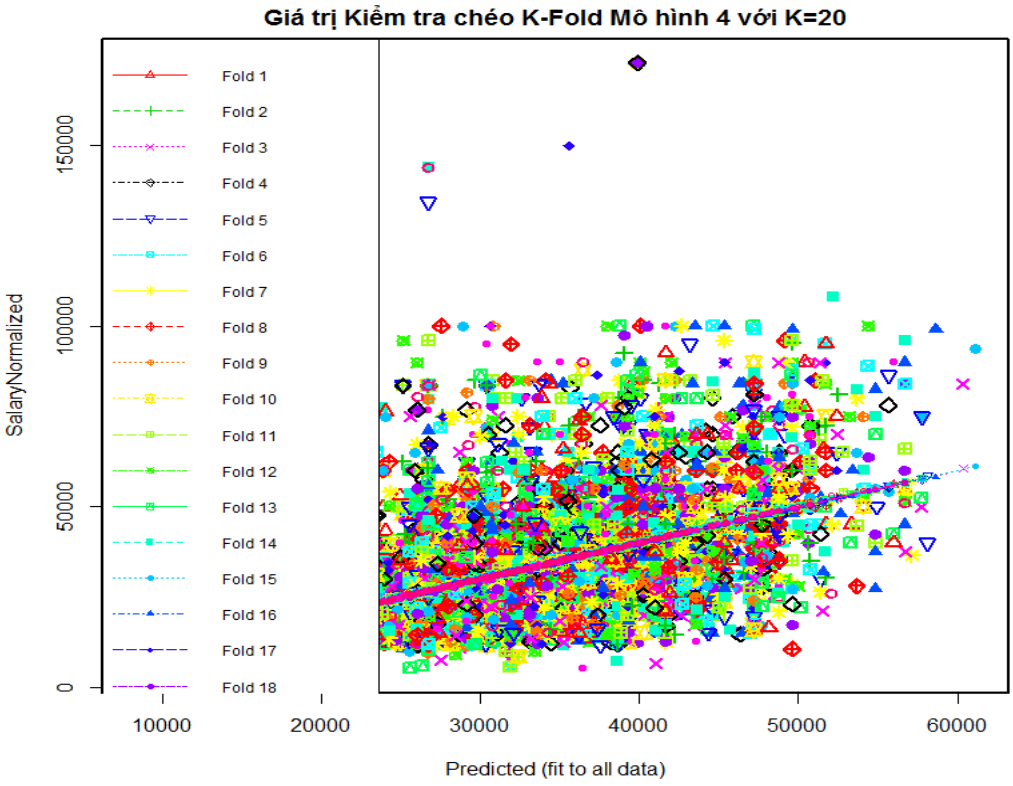
Hình 60: Kết quả kiểm tra chéo trên mô hình 4 với k=10

Bảng 6: Kết quả sai số trung bình với k=10

Tập K	Sai số trung bình
k=1	13564.66
k=2	13964.24
k=3	13416.41 (nhỏ nhất)
k=4	15779.73
k=5	13564.66
k=6	14142.14
k=7	13928.39
k=8	14142.14
k=9	13784.05
k=10	13820.27

Giá trị sai số trung bình với k =10 là: **14010.67**

- Kiểm tra chéo K-Fold trên Mô hình 4 với K=20:



Hình 61: Kết quả kiểm tra chéo trên mô hình 4 với k=20

Bảng 7: Kết quả sai số trung bình với k=20

Tập K	Sai số trung bình
k=1	12041.59 (nhỏ nhất)
k=2	13304.13
k=3	13076.70
k=4	16431.68
k=5	13928.39
k=6	13152.95
k=7	14035.67
k=8	14525.84
k=9	14071.25
k=10	13038.40
k=11	14899.66
k=12	14594.52
k=13	13711.31
k=14	15099.67
k=15	13152.95
k=16	15000.00
k=17	13820.27
k=18	13711.31
k=19	13490.74
k=20	14491.38

Giá trị sai số trung bình với k=20 là: **13978.92**

So sánh kết quả sau khi huấn luyện dữ liệu với mô hình 4 với k = 5, 10 và 20 ta thấy với k = 20 mô hình dự báo mô hình 4 có giá trị sai số trung bình nhỏ nhất **13978.92**. Do đó tác giả chọn huấn luyện tập dữ liệu với tập k = 20.

Mô hình trước khi huấn luyện với k-fold, k=20:

Residual standard error:	13960 on 6963 degrees of freedom
Multiple R-squared:	0.2546
Adjusted R-squared:	0.2508
F-statistic:	66.07 on 36 and 6963 DF
p-value:	< 2.2e-16

Mô hình sau khi huấn luyện với k-fold, k =20:

Residual standard error:	14000 on 6963 degrees of freedom
Multiple R-squared:	0.255
Adjusted R-squared:	0.251
F-statistic:	66.1 on 36 and 6963 DF
p-value:	<2e-16

Sau khi huấn luyện với k-fold, k =20. Kết quả mô hình thể hiện khoảng 26% độ dao động của mức lương liên quan đến các nhân tố việc làm trong quảng cáo tuyển dụng trong mô hình 4. Điều đó không có nghĩa mô hình dự đoán không tốt. Và trong mô hình này để đánh giá độ tin cậy của mô hình, tác giả dựa vào giá trị sai số trung bình và sai số trung bình tuyệt đối, với mô hình 4 sau khi thực hiện với k-fold, k-20 ta có giá trị sai số trung bình là **13978.92** giảm **2221.79** so với mô hình dự đoán ban đầu là **16200.71** với phần trăm sai số như sau:

Bảng 8: Giá trị sai số dùng để đo lường độ chính xác của mô hình

	Sai số trung bình (ME)	Tổng sai số trung bình (RMSE)	Sai số tuyệt đối (MAE)	% Sai số trung bình (MPE)	% Sai số tuyệt đối trung bình (MAPE)	Tỷ lệ sai số trung bình tuyệt đối (MASE)
Mô hình 4	-4.85E-14	13927.94	4939	-15	33.1	0.802

Mô hình 4 với sai số trung bình tuyệt đối khoản 4939 và phần trăm sai số là **33.1%** điều đó cho thấy mức độ chính xác của mô hình dự đoán khoảng **66.9%** giá trị dự đoán về mức lương. So sánh với mô hình kết hợp của tác giả ẩn danh [15] sai số trung bình tuyệt của mô hình 4 gần giống nhau, chênh lệch khoản 6 đơn vị. Nhưng ngược lại mô hình 4 được xây dựng với cách tiếp cận mô hình hồi quy tuyến tính một các đơn giản nhất, tận dụng mọi dữ liệu có sẵn, phân loại và kết hợp lại tạo nên mô hình dự đoán mức lương cho kết quả khá tốt.

CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1. Kết luận

Phương pháp hồi quy có là một trong những phương pháp phân tích số liệu thông dụng nhất trong thống kê học, đặt biệt là phân tích dự đoán. Đã có nhiều nghiên cứu phân tích thống kê ứng dụng hồi quy như dự đoán giá bất động sản, dự đoán chứng khoáng .v.v. Ở mỗi dữ liệu phân tích điều có những đặc trưng khác nhau, tùy vào dự liệu phân tích mà có những mô hình dự đoán riêng. Và trong khuôn khổ luận văn này cũng không ngoại lệ.

Trong phạm vi luận văn này đã thực hiện xây dựng mô hình dự đoán trên dữ liệu quảng cáo tuyển dụng ở Anh được công bố trong một cuộc thi được cung cấp bởi Kaggle. Trên cơ sở đó tác giả đi xây dựng mô hình dự đoán mức lương dựa trên các đặc trưng mà dữ liệu cung cấp: Nhóm công việc, loại công việc, loại hợp đồng, địa điểm làm việc, tiêu đề công việc và mô tả công việc. Bằng việc thêm vào một số đặc trưng như thay vì dùng biến đặc trưng địa điểm để phân tích, tác giả đi vào việc sử dụng một địa điểm làm việc cụ thể (ví dụ: Luân Đôn) hoặc phân loại dữ liệu của các đặc trưng về tiêu đề công việc và mô tả công việc theo vị trí ứng viên có kinh nghiệm hoặc vị trí quản lý. Sau đó phân tích mức độ ảnh hưởng của các đặc trưng lên mức lương tuyển dụng để đưa ra một mô hình dự đoán tối ưu nhất (**Mô hình 4**).

Luận văn cũng đã thực hiện theo mô hình khai thác dữ liệu CRISP-DM, thực hiện các bước từ tìm hiểu nghiệp vụ, tìm hiểu dữ liệu, chuẩn bị dữ liệu cho đến mô hình hóa và đánh giá mô hình dựa trên các kiến thức về thống kê học mà tác giả nghiên cứu được.

4.2. Hướng phát triển

Phạm vi nghiên cứu hiện tại chỉ dừng lại ở mức phân tích thống kê trên dữ liệu quảng cáo tuyển dụng: tìm hiểu nghiệp vụ, tìm hiểu dữ liệu, chuẩn bị dữ liệu, mô hình hóa dữ liệu bằng phương pháp hồi quy, và đánh giá mô hình. Vì lý do thời gian có hạn do đó giai đoạn triển khai chưa được thực hiện. Tác giả sẽ thực hiện giai đoạn triển khai ứng dụng ở những nghiên cứu sâu hơn về sau.

TÀI LIỆU THAM KHẢO

[1] An Introduction to Statistical Learning with Applications in R (Fourth Printing), G. James, D. Witten, T. Hastie and R. Tibshirani, Springer-Verlag, 2014

[2] IBM SPSS Modeler CRISP-DM Guide, IBM Corporation, 1994-2011

[3] CRISP-DM 1.0, Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), 1999-2000

[4] The Elements of Statistical Learning (Second Edition), T. Hastie, R. Tibshirani and J. Friedman, Springer-Verlag, 2009

[5] Introduction to the Practice of Statistics (Sixth Edition), S. Moore, P. Mccabe, A. Craig, 2007

[6] Job Advertisement Dataset: https://www.kaggle.com/c/job-salary-prediction/data

[7] Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), Ian H. Witten, Eibe Frank and Mark A. Hall, 2011

[8] https://the-modeling-agency.com/crisp-dm.pdf

[9] Machine learning with R Cookbook, Yu-Wei, Chiu (David Chiu), Published by Packt Publishing Ltd., ISBN 978-1-78398-204-2, 2015

[10] Data Mining and Predictive Analytics, Daniel T.Larose and Chantal D.Larose, Published by John Wiley & Son, Inc., 2015

[11] Learning Predictive Analytics with R, Eric Mayor, Published by Packt Publishing Ltd., ISBN 978-1-78216-935-2, 2015

[12] Mastering Machine Learning With R, Cory Lesmeister, Published by Packt Publishing Ltd., ISBN 978-1-78398-452-7, 2015

[13] Phân tích dữ liệu với R, Nguyễn Văn Tuấn, NXB. Tổng Hợp TP.HCM, 2014

[14] Prediction And Determination Of Household Permanent Income, Ramses H Abul Naga, University of Lausanne, 1997

[15] http://www.cs.ubc.ca/~nando/540-2013/projects/p58.pdf

PHỤ LỤC

A. Code R sử dụng trong luận văn

A.1. Thêm tính năng phân loại dữ liệu quảng cáo tuyển dụng theo địa điểm

```
con <- file("data/location_tree.txt", "r")
tree <- readLines(con)
close(con)
for (i in 1:nrow(train)) {
    # lấy tên thành phố
    loc <- train$LocationNormalized[i]
    # Tìm dòng thứ nhất trong cây mà có hiển thị tên
thành phố
    line.id <- which(grepl(loc, tree))[1]
    # Dùng regular expressions để lấy ra tên thành phố
    r <- regexpr("~.+?~", tree[line.id])
    match <- regmatches(tree[line.id], r)
    # Lưu lại tên thành phố vào biến Location
    train$Location[i] <- gsub("~", "", match)
}
```

A.2. Thêm tính năng nhóm dữ liệu theo địa điểm làm việc là Luân Đôn

```
train$London <- as.factor(ifelse(train$Location ==
"London", "Yes", "No"))
table(train$London)
```

A.3. Thêm tính năng nhóm dữ liệu theo tiêu đề và mô tả công việc

```
train$TitleSenior <- as.factor(ifelse(grepl("[Ss]enior",
train$Title), "Yes", "No"))
train$TitleManage <- as.factor(ifelse(grepl("[Mm]anage",
train$Title), "Yes", "No"))
```

```
train$DescripSenior <-
as.factor(ifelse(grepl("[Ss]enior",
train$FullDescription), "Yes", "No"))
train$DescripManage <-
as.factor(ifelse(grepl("[Mm]anage",train$FullDescription
), "Yes", "No"))
table(train$TitleSenior)
table(train$TitleManage)
table(train$DescripSenior)
table(train$DescripManage)
```

A.4. Mô hình hồi quy Ridge:

```
library(glmnet)
x.tr <- model.matrix(SalaryNormalized ~ ContractType +
  ContractTime + Category + London + TitleSenior +
  TitleManage + DescripSenior + DescripManage, data =
  tr)[,-1]
y.tr <- tr$SalaryNormalized
x.val <- model.matrix(SalaryNormalized ~ ContractType +
  ContractTime + Category + London + TitleSenior +
  TitleManage + DescripSenior + DescripManage, data =
  val)[,-1]
y.val <- val$SalaryNormalized
set.seed(10)
rr.cv <- cv.glmnet(x.tr, y.tr, alpha = 0)
plot(rr.cv)
rr.bestlam <- rr.cv$lambda.min
rr.goodlam <- rr.cv$lambda.1se
rr.fit <- glmnet(x.tr, y.tr, alpha = 0)
plot(rr.fit, xvar = "lambda", label = TRUE)
```

```
rr.pred <- predict(rr.fit, s = rr.bestlam, newx = x.val)
sqrt(mean((rr.pred - y.val)^2))
```

A.5. Mô hình Lasso:

```
set.seed(10)
las.cv <- cv.glmnet(x.tr, y.tr, alpha = 1)
plot(las.cv)
las.bestlam <- las.cv$lambda.min
las.goodlam <- las.cv$lambda.1se
las.fit <- glmnet(x.tr, y.tr, alpha = 1)
plot(las.fit, xvar = "lambda", label = TRUE)
las.pred <- predict(las.fit, s = las.bestlam, newx =
  x.val)
sqrt(mean((las.pred - y.val)^2))
```

A.6. Phương pháp lựa chọn từng bước:

```
library(leaps)
fwd.fit <- regsubsets(SalaryNormalized ~ ContractType +
  ContractTime + Category + London + TitleSenior +
  TitleManage + DescripSenior + DescripManage, data =
  tr, nvmax = 36, method = "forward")
plot(fwd.fit, scale = "adjr2")
fwd.errors <- rep(NA, 36)
val.mat <- model.matrix(SalaryNormalized ~ ContractType
  + ContractTime + Category + London + TitleSenior +
  TitleManage + DescripSenior + DescripManage, data =
  val)
for (i in 1:36) {
    coefi <- coef(fwd.fit, id = i)
    pred <- val.mat[, names(coefi)] %*% coefi
```

```
    fwd.errors[i] <- sqrt(mean((y.val - pred)^2))
}
fwd.errors
```

A.7. Kiểm tra chéo K-Fold:

```
printOut <-
capture.output(mse_CVlm=CVlm(train,lm.fit4,10,main = "Giá
trị Kiểm tra chéo Mô hình 4 với K=10"))
GetValues <- function(itemName,printOut){
  line <- printOut[grep(itemName,printOut)]
  items <- unlist(strsplit(line,"[=]|  +"))
  itemsMat <- matrix(items,ncol=2,byrow=TRUE)
  vectVals <-
as.numeric(itemsMat[grep(itemName,itemsMat[,1]),2])
  return(vectVals)
}
MS <- GetValues("Mean square",printOut)
```

B. Kết quả chạy thực nghiệm

B.1. Kết quả mô hình 1:

```
Call:
lm(formula = SalaryNormalized ~ ContractType + ContractTime +
    Category, data = tr)

Residuals:
   Min     1Q Median     3Q    Max
-38567  -8189  -2896   5623 129929

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        33567.2      764.7  43.896  < 2e-16 ***
ContractTypefull_time               -311.6      419.7  -0.742 0.457813
ContractTypepart_time              -7430.8      804.0  -9.243  < 2e-16 ***
ContractTimecontract               12181.8      837.4  14.547  < 2e-16 ***
```

```
ContractTimepermanent                       6243.3      546.9  11.415  < 2e-16 ***
CategoryAdmin Jobs                        -16522.8     1559.9 -10.592  < 2e-16 ***
CategoryCharity & Voluntary Jobs          -14437.9     3698.5  -3.904 9.56e-05 ***
CategoryConsultancy Jobs                    -664.9     2169.2  -0.307 0.759232
CategoryCreative & Design Jobs            -16643.9     4428.9  -3.758 0.000173 ***
CategoryCustomer Services Jobs            -18558.9     1283.3 -14.461  < 2e-16 ***
CategoryDomestic help & Cleaning Jobs     -19524.5     5175.2  -3.773 0.000163 ***
CategoryEnergy, Oil & Gas Jobs              9324.5     2989.0   3.120 0.001819 **
CategoryEngineering Jobs                   -5413.3      892.4  -6.066 1.38e-09 ***
CategoryGraduate Jobs                      -7506.1     3418.7  -2.196 0.028153 *
CategoryHealthcare & Nursing Jobs          -2878.3      829.0  -3.472 0.000520 ***
CategoryHospitality & Catering Jobs       -12670.8     1057.6 -11.981  < 2e-16 ***
CategoryHR & Recruitment Jobs              -8040.6     1011.8  -7.947 2.22e-15 ***
CategoryIT Jobs                             3378.2      867.2   3.896 9.89e-05 ***
CategoryLegal Jobs                         -6588.7     1951.1  -3.377 0.000737 ***
CategoryLogistics & Warehouse Jobs        -14141.3     1772.3  -7.979 1.71e-15 ***
CategoryMaintenance Jobs                   -8281.6     3596.4  -2.303 0.021320 *
CategoryManufacturing Jobs                 -9866.8     1778.2  -5.549 2.99e-08 ***
CategoryOther/General Jobs                 -9080.4     1389.1  -6.537 6.72e-11 ***
CategoryPR, Advertising & Marketing Jobs   -4472.0     2045.4  -2.186 0.028824 *
CategoryProperty Jobs                      -6547.9     2932.8  -2.233 0.025603 *
CategoryRetail Jobs                       -11373.5     1977.5  -5.751 9.22e-09 ***
CategorySales Jobs                         -9561.9     1115.2  -8.574  < 2e-16 ***
CategoryScientific & QA Jobs               -5563.4     1692.7  -3.287 0.001018 **
CategorySocial work Jobs                   -6011.2     2703.2  -2.224 0.026196 *
CategoryTeaching Jobs                     -12028.5     1165.0 -10.325  < 2e-16 ***
CategoryTrade & Construction Jobs          -4568.4     1596.1  -2.862 0.004218 **
CategoryTravel Jobs                       -18149.7     1840.2  -9.863  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14490 on 6968 degrees of freedom
Multiple R-squared:  0.1964,     Adjusted R-squared:  0.1928
F-statistic: 54.93 on 31 and 6968 DF,  p-value: < 2.2e-16
```

B.2. Kết quả mô hình 2:

```
Call:
lm(formula = SalaryNormalized ~ ContractType + ContractTime +
    Category + Location, data = tr)

Residuals:
   Min     1Q Median     3Q    Max
```

```
-38574  -8376  -2816   5425 129864

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          32886.2     1205.8  27.274  < 2e-16 ***
ContractTypefull_time                 -767.7      422.9  -1.816 0.069489 .
ContractTypepart_time                 -7995.3      806.5  -9.913  < 2e-16 ***
ContractTimecontract                  12055.5      855.7  14.088  < 2e-16 ***
ContractTimepermanent                  5955.3      567.6  10.493  < 2e-16 ***
CategoryAdmin Jobs                   -16712.6     1561.8 -10.701  < 2e-16 ***
CategoryCharity & Voluntary Jobs     -14842.4     3678.8  -4.035 5.53e-05 ***
CategoryConsultancy Jobs               -773.4     2158.4  -0.358 0.720115
CategoryCreative & Design Jobs       -16678.9     4406.3  -3.785 0.000155 ***
CategoryCustomer Services Jobs       -18071.4     1280.1 -14.118  < 2e-16 ***
CategoryDomestic help & Cleaning Jobs -19579.0     5153.2  -3.799 0.000146 ***
CategoryEnergy, Oil & Gas Jobs         9411.1     2973.8   3.165 0.001560 **
CategoryEngineering Jobs              -5266.2      893.8  -5.892 3.99e-09 ***
CategoryGraduate Jobs                 -7287.4     3409.6  -2.137 0.032609 *
CategoryHealthcare & Nursing Jobs     -2827.3      832.4  -3.397 0.000686 ***
CategoryHospitality & Catering Jobs  -13289.0     1057.3 -12.568  < 2e-16 ***
CategoryHR & Recruitment Jobs         -8124.3     1018.9  -7.973 1.79e-15 ***
CategoryIT Jobs                        3370.8      866.3   3.891 0.000101 ***
CategoryLegal Jobs                    -6639.4     1942.5  -3.418 0.000634 ***
CategoryLogistics & Warehouse Jobs   -14097.0     1769.5  -7.967 1.89e-15 ***
CategoryMaintenance Jobs              -9208.0     3581.9  -2.571 0.010170 *
CategoryManufacturing Jobs            -9598.5     1775.7  -5.405 6.68e-08 ***
CategoryOther/General Jobs            -9404.3     1382.7  -6.801 1.12e-11 ***
CategoryPR, Advertising & Marketing Jobs -4736.4     2036.7  -2.326 0.020073 *
CategoryProperty Jobs                 -7064.8     2918.2  -2.421 0.015504 *
CategoryRetail Jobs                  -11256.7     1969.3  -5.716 1.14e-08 ***
CategorySales Jobs                    -9714.1     1112.4  -8.732  < 2e-16 ***
CategoryScientific & QA Jobs          -5522.0     1686.5  -3.274 0.001065 **
CategorySocial work Jobs              -5905.9     2693.7  -2.193 0.028374 *
CategoryTeaching Jobs                -12272.7     1168.0 -10.507  < 2e-16 ***
CategoryTrade & Construction Jobs     -4323.4     1595.6  -2.710 0.006754 **
CategoryTravel Jobs                  -18255.2     1831.1  -9.969  < 2e-16 ***
LocationEastern England                2556.6     1179.6   2.167 0.030238 *
LocationLondon                         3574.5     1023.3   3.493 0.000480 ***
LocationNorth East England             -432.7     1513.1  -0.286 0.774926
LocationNorth West England            -1920.7     1184.4  -1.622 0.104904
LocationNorthern Ireland                118.4     1904.6   0.062 0.950418
LocationScotland                       1542.2     1260.4   1.224 0.221139
LocationSouth East England              821.9      993.7   0.827 0.408196
```

```
LocationSouth West England                -1993.8     1211.5  -1.646 0.099850 .
LocationWales                              -493.5     1636.9  -0.301 0.763066
LocationWest Midlands                      -327.7     1310.2  -0.250 0.802491
LocationYorkshire And The Humber          -2307.2     1490.4  -1.548 0.121668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14410 on 6957 degrees of freedom
Multiple R-squared:  0.207,      Adjusted R-squared:  0.2022
F-statistic: 43.23 on 42 and 6957 DF,  p-value: < 2.2e-16
```

B.3. Kết quả mô hình 3:

```
Call:
lm(formula = SalaryNormalized ~ ContractType + ContractTime +
    Category + London, data = tr)

Residuals:
   Min     1Q Median     3Q    Max
-38479  -8319  -2935   5404 130243

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          32955.9      766.5  42.995  < 2e-16 ***
ContractTypefull_time                 -663.7      420.9  -1.577 0.114911
ContractTypepart_time                -7635.3      801.5  -9.527  < 2e-16 ***
ContractTimecontract                 12458.6      835.2  14.918  < 2e-16 ***
ContractTimepermanent                 6274.0      544.9  11.513  < 2e-16 ***
CategoryAdmin Jobs                  -16253.3     1554.6 -10.455  < 2e-16 ***
CategoryCharity & Voluntary Jobs    -14694.9     3685.0  -3.988 6.74e-05 ***
CategoryConsultancy Jobs              -641.9     2161.2  -0.297 0.766482
CategoryCreative & Design Jobs      -16529.2     4412.5  -3.746 0.000181 ***
CategoryCustomer Services Jobs      -18327.5     1279.0 -14.330  < 2e-16 ***
CategoryDomestic help & Cleaning Jobs -19643.5   5156.1  -3.810 0.000140 ***
CategoryEnergy, Oil & Gas Jobs        9674.5     2978.4   3.248 0.001167 **
CategoryEngineering Jobs             -5134.3      889.9  -5.770 8.28e-09 ***
CategoryGraduate Jobs                -6753.7     3407.6  -1.982 0.047525 *
CategoryHealthcare & Nursing Jobs    -2857.1      825.9  -3.459 0.000545 ***
CategoryHospitality & Catering Jobs  -12951.4     1054.4 -12.284  < 2e-16 ***
CategoryHR & Recruitment Jobs         -7982.7     1008.1  -7.919 2.78e-15 ***
CategoryIT Jobs                       3624.2      864.6   4.192 2.81e-05 ***
CategoryLegal Jobs                   -6369.4     1944.1  -3.276 0.001057 **
CategoryLogistics & Warehouse Jobs   -13755.9     1766.6  -7.787 7.87e-15 ***
```

```
CategoryMaintenance Jobs                   -9474.9     3586.8  -2.642 0.008270 **
CategoryManufacturing Jobs                 -9671.5     1771.9  -5.458 4.97e-08 ***
CategoryOther/General Jobs                 -9149.2     1384.0  -6.611 4.10e-11 ***
CategoryPR, Advertising & Marketing Jobs   -4469.1     2037.9  -2.193 0.028337 *
CategoryProperty Jobs                      -6815.1     2922.2  -2.332 0.019718 *
CategoryRetail Jobs                       -11086.8     1970.6  -5.626 1.91e-08 ***
CategorySales Jobs                         -9516.6     1111.1  -8.565  < 2e-16 ***
CategoryScientific & QA Jobs               -5854.8     1686.9  -3.471 0.000522 ***
CategorySocial work Jobs                   -5554.6     2693.9  -2.062 0.039250 *
CategoryTeaching Jobs                     -12500.8     1162.5 -10.753  < 2e-16 ***
CategoryTrade & Construction Jobs          -4312.9     1590.5  -2.712 0.006712 **
CategoryTravel Jobs                       -18265.1     1833.5  -9.962  < 2e-16 ***
LondonYes                                   3242.9      445.9   7.272 3.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14440 on 6967 degrees of freedom
Multiple R-squared:  0.2025,     Adjusted R-squared:  0.1988
F-statistic: 55.27 on 32 and 6967 DF,  p-value: < 2.2e-16
```

B.4. Kết quả mô hình 4:

```
Call:
lm(formula = SalaryNormalized ~ ContractType + ContractTime +
    Category + London + TitleSenior + TitleManage + DescripSenior +
    DescripManage, data = tr)

Residuals:
   Min     1Q Median     3Q    Max
-38496  -8005  -2400   4293 133040

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          29344.7      776.1  37.813  < 2e-16 ***
ContractTypefull_time                 -932.7      407.7  -2.288 0.022195 *
ContractTypepart_time                 -5017.2      784.9  -6.393 1.74e-10 ***
ContractTimecontract                  13020.3      808.5  16.104  < 2e-16 ***
ContractTimepermanent                  5704.0      527.8  10.807  < 2e-16 ***
CategoryAdmin Jobs                   -14391.9     1506.7  -9.552  < 2e-16 ***
CategoryCharity & Voluntary Jobs     -13263.9     3565.8  -3.720 0.000201 ***
CategoryConsultancy Jobs              -3850.6     2102.1  -1.832 0.067023 .
CategoryCreative & Design Jobs       -15475.3     4268.1  -3.626 0.000290 ***
CategoryCustomer Services Jobs       -15938.2     1243.7 -12.816  < 2e-16 ***
```

```
CategoryDomestic help & Cleaning Jobs    -17944.4     4986.8  -3.598 0.000322 ***
CategoryEnergy, Oil & Gas Jobs            11460.8     2881.8   3.977 7.05e-05 ***
CategoryEngineering Jobs                  -3579.0      865.5  -4.135 3.59e-05 ***
CategoryGraduate Jobs                     -5769.4     3300.9  -1.748 0.080540 .
CategoryHealthcare & Nursing Jobs         -2605.4      806.5  -3.230 0.001242 **
CategoryHospitality & Catering Jobs      -12170.9     1024.9 -11.875  < 2e-16 ***
CategoryHR & Recruitment Jobs             -7961.4      977.3  -8.146 4.41e-16 ***
CategoryIT Jobs                            4524.2      837.9   5.399 6.91e-08 ***
CategoryLegal Jobs                        -3820.1     1884.6  -2.027 0.042694 *
CategoryLogistics & Warehouse Jobs       -11697.3     1715.9  -6.817 1.01e-11 ***
CategoryMaintenance Jobs                  -8682.1     3469.9  -2.502 0.012368 *
CategoryManufacturing Jobs                -8818.7     1715.2  -5.141 2.80e-07 ***
CategoryOther/General Jobs                -8200.6     1340.0  -6.120 9.87e-10 ***
CategoryPR, Advertising & Marketing Jobs  -5282.1     1975.7  -2.674 0.007523 **
CategoryProperty Jobs                     -8609.9     2828.8  -3.044 0.002346 **
CategoryRetail Jobs                      -12731.3     1912.0  -6.659 2.98e-11 ***
CategorySales Jobs                        -8857.2     1077.1  -8.223 2.34e-16 ***
CategoryScientific & QA Jobs              -4525.0     1634.6  -2.768 0.005651 **
CategorySocial work Jobs                  -6901.6     2611.2  -2.643 0.008233 **
CategoryTeaching Jobs                     -9912.1     1131.5  -8.760  < 2e-16 ***
CategoryTrade & Construction Jobs         -3582.8     1543.1  -2.322 0.020270 *
CategoryTravel Jobs                      -18292.1     1775.9 -10.300  < 2e-16 ***
LondonYes                                  3225.9      431.2   7.480 8.32e-14 ***
TitleSeniorYes                             3693.4      787.7   4.689 2.80e-06 ***
TitleManageYes                             7156.8      503.3  14.220  < 2e-16 ***
DescripSeniorYes                           2323.7      532.0   4.368 1.27e-05 ***
DescripManageYes                           2166.8      393.7   5.504 3.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13960 on 6963 degrees of freedom
Multiple R-squared:  0.2546,     Adjusted R-squared:  0.2508
F-statistic: 66.07 on 36 and 6963 DF,  p-value: < 2.2e-16
```