ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Lab 01: Linear Regression Project

Môn học: Nhập Môn Học Máy

Sinh viên thực hiện:
Bùi Huy Giáp - 23127289
Lê Minh Đức - 23127351
Vũ Tiến Dũng - 23127354
Đinh Xuân Khương - 23127398
Nguyễn Đồng Thanh - 23127538

Giảng viên hướng dẫn:
Bùi Tiến Lên
Lê Nhựt Nam
Võ Nhật Tân

Ngày 16 tháng 11 năm 2025

Mục lục

1 Tổng quan về đồ án

1.1 Vấn đề hiện nay và động lực

1.1.1 Vấn đề hiện nay

Ngày nay ở các quốc gia phát triển, đặc biệt là Mỹ, chi phí y tế đang gia tăng theo thời gian, khiến các công ty bảo hiểm đối mặt với thách thức lớn trong việc xác định mức phí phù hợp cho từng khách hàng. Những yếu tố như tuổi tác, giới tính, chỉ số BMI, tình trạng hút thuốc, số lượng con cái và khu vực sinh sống đều có thể ảnh hưởng đáng kể đến mức chi trả bảo hiểm. Tuy nhiên, việc đánh giá thủ công các yếu tố này thường thiếu chính xác, dễ bị thiên lệch và không bắt kịp với quy mô dữ liệu lớn.
Bên cạnh đó, sự phức tạp trong mối quan hệ giữa các biến số khiến việc phân tích truyền thống trở nên kém hiệu quả. Do đó, các công ty bảo hiểm đang cần những phương pháp hiện đại, có khả năng tự động phân tích và dự đoán chi phí dựa trên dữ liệu thực tế.

1.1.2 Động lực

Trong dự án này, Bộ dữ liệu chi phí bảo hiểm y tế cung cấp cơ hội để áp dụng các mô hình học máy nhằm phân tích những yếu tố quyết định chi phí y tế và dự đoán mức phí cho khách hàng mới. Việc xây dựng mô hình dự đoán không chỉ hỗ trợ cải thiện độ chính xác trong định giá bảo hiểm mà còn giúp doanh nghiệp tối ưu hoá lợi nhuận, giảm rủi ro và minh bạch hơn trong quy trình đánh giá khách hàng.
Bên cạnh đó, bệnh nhân có thể xem được với tình trạng của mình, thì chi phí mà bảo hiểm có thể chi trả là bao nhiêu, từ đó có những góc nhìn mới hơn và trực quan hơn về việc mua bảo hiểm.

1.2 Mục tiêu

Mục tiêu chính của phân tích và mô hình hoá bộ dữ liệu chi phí bảo hiểm y tế là xây dựng một hệ thống dự đoán mức phí bảo hiểm dựa trên các yếu tố đầu vào của khách hàng. Cụ thể, báo cáo hướng đến các mục tiêu sau:

- Xác định mức độ ảnh hưởng của các yếu tố như tuổi, giới tính, BMI, tình trạng hút thuốc, số lượng con và khu vực sinh sống đối với chi phí y tế.
- Xây dựng và đánh giá các mô hình học máy có khả năng dự đoán chi phí bảo hiểm với mục tiêu là độ chính xác dự đoán cao.
- So sánh hiệu quả giữa các mô hình khác nhau nhằm lựa chọn mô hình tối ưu cho bài toán thực tế.
- Hỗ trợ doanh nghiệp bảo hiểm đưa ra quyết định định giá chính xác hơn, giảm thiểu rủi ro và tối ưu hóa lợi nhuận.
- Tạo nền tảng phân tích dữ liệu có thể mở rộng và áp dụng cho các bộ dữ liệu sức khỏe khác trong tương lai.

1.3 Input và output

1.3.1 Input

Dữ liệu đầu vào của mô hình bao gồm các thông tin mô tả đặc điểm nhân khẩu học và tình trạng sức khỏe của từng cá nhân. Các thuộc tính này được sử dụng để dự đoán chi phí bảo hiểm y tế. Bảng dưới đây minh họa chi tiết từng trường dữ liệu cùng với kiểu dữ liệu tương ứng.

Thuộc tính	**Ý nghĩa**	**Kiểu dữ liệu**
`age`	Tuổi của cá nhân.	`int`
`sex`	Giới tính của cá nhân (nam hoặc nữ).	`string`
`bmi`	Chỉ số BMI, thể hiện lượng mỡ cơ thể.	`float`
`children`	Số con mà người đó có.	`int`
`smoker`	Tình trạng hút thuốc (`"yes"` hoặc `"no"`).	`string`
`region`	Khu vực địa lý sinh sống (northeast, northwest, southeast, southwest).	`string`

1.3.2 Output

Đầu ra của mô hình là một giá trị số thực, biểu diễn chi phí bảo hiểm y tế dự đoán (`charges`) dựa trên các thông tin đầu vào của từng khách hàng. Kết quả này giúp công ty bảo hiểm ước tính chi phí dự kiến và xây dựng chính sách giá phù hợp.

1.4 Đóng góp của các thành viên

Thành viên	**Công việc**	**Mức độ hoàn thành**
Bùi Huy Giáp	Tiền xử lý dữ liệu, xây dựng mô hình 1 và 2, đánh giá các mô hình và viết báo cáo.	100%
Lê Minh Đức	Phân tích dữ liệu và khám phá (EDA), trực quan hóa các mối quan hệ và phân bố giữa các biến, xây dựng mô hình 3, 4 và viết báo cáo	100%
Vũ Tiến Dũng	Tìm dữ liệu phù hợp cho đồ án, xây dựng mô hình 5, 6, đánh giá các mô hình, và viết báo cáo.	100%
Đinh Xuân Khương	Thiết kế giao diện người dùng và tích hợp mô hình vào ứng dụng Streamlit, xây dựng mô hình 8, và viết báo cáo.	100%
Nguyễn Đồng Thanh	Quay video ứng dụng, xây dựng mô hình 7, 9, đánh giá các mô hình, và viết báo cáo.	100%

2 Mô tả dữ liệu (Data description)

Lưu ý: Các số liệu và đánh giá này được thực hiện trên toàn bộ tập dữ liệu

2.1 Nguồn dữ liệu và giấy phép

Nguồn: *Medical Insurance Cost Prediction*
Giấy phép sử dụng: *MIT*

2.2 Khám phá và tiền xử lý dữ liệu

Khi sử dụng hàm `info()`:

#	Column	Non-Null Count	Dtype
0	age	2772 non-null	int64
1	sex	2772 non-null	object
2	bmi	2772 non-null	float64
3	children	2772 non-null	int64
4	smoker	2772 non-null	object
5	region	2772 non-null	object
6	charges	2772 non-null	float64

Bảng 1: Thông tin cấu trúc DataFrame của bộ dữ liệu bảo hiểm y tế.

Theo như bảng thông số dữ liệu, ta thấy không có hàng hay cột nào bị thiếu dữ liệu. Vì thế, chúng ta không cần xử lý việc thiếu giá trị (missing value).
Ngoài ra, có một số cột có dữ liệu chưa được chuyển sang dạng số để phù hợp với việc tính toán, chúng ta có thể sử dụng **one-hot encoding** để xử lý.
Kiểm tra và xử lý trùng lặp
Sau khi kiểm tra trùng lặp, chúng tôi thấy có **1435** dòng bị trùng lặp. Vì thế chúng tôi đã xóa các dòng bị trùng lặp này, vì chúng tôi dữ liệu huấn luyện là duy nhất.
Khi sử dụng hàm `describe()` sau khi loại bỏ trùng lặp:

	age	bmi	children	charges
count	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487
std	14.044333	6.100468	1.205571	12110.359656
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.290000	0.000000	4746.344400
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	53.130000	5.000000	63770.428010

Nhận xét từ thống kê mô tả của bộ dữ liệu

Dựa trên bảng số liệu `describe()`, ta có một số đánh giá quan trọng như sau:

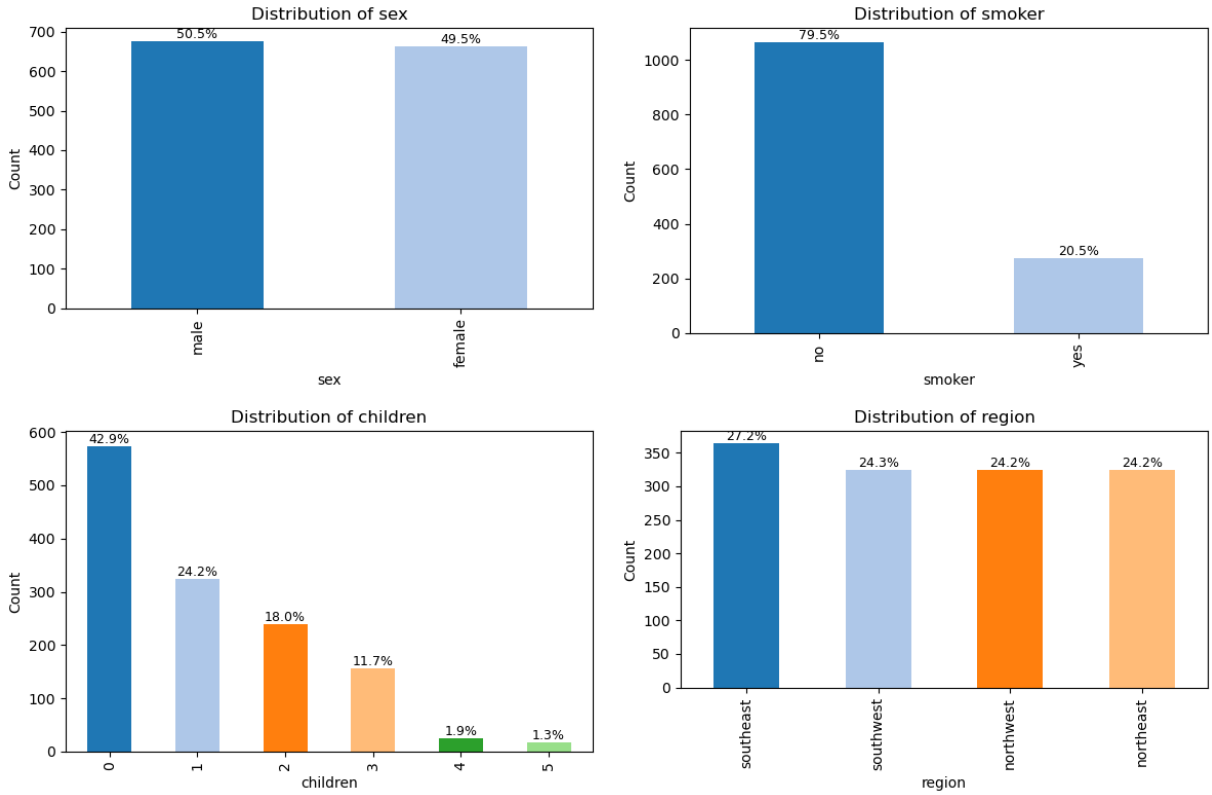
- **Độ tuổi trung bình của người được bảo hiểm là khoảng 39 tuổi**, với giá trị nhỏ nhất là 18 và lớn nhất là 64. Điều này cho thấy dữ liệu bao phủ nhiều nhóm tuổi, trong đó nhóm trung niên chiếm tỷ trọng lớn.
- **Chỉ số BMI trung bình vào khoảng 30.66**, thuộc mức thừa cân, từ đó có thể rút được rằng người có cân nặng thừa thường sử dụng các dịch vụ y tế nhiều hơn. Ngoài ra, giá trị BMI dao động từ 15.96 đến 53.13, thể hiện sự đa dạng đáng kể về thể trạng của người tham gia bảo hiểm.
- **Số con trung bình là 1.09**, với độ lệch chuẩn 1.20. Phần lớn các cá nhân có từ 0 đến 2 con, phù hợp với phân vị 25%, 50% và 75%.
- **Mức phí bảo hiểm trung bình là 13,279.12 USD**, với độ lệch chuẩn rất lớn (12,110.36 USD). Điều này cho thấy mức phí bị phân tán mạnh và có thể tồn tại các giá trị ngoại lai, đặc biệt khi mức phí tối đa lên đến hơn 63,000 USD.
- Sự chênh lệch lớn giữa trung vị (9,386 USD) và giá trị trung bình (13,279 USD) cho thấy **phân phối phí bảo hiểm lệch phải**, nghĩa là có một nhóm khách hàng phải trả mức phí rất cao, kéo trung bình lên.
- Khi kết hợp với đặc thù bảo hiểm y tế, có thể dự đoán rằng **các yếu tố như BMI cao, nhiều con, hoặc tuổi lớn** sẽ góp phần làm tăng phí bảo hiểm, do rủi ro sức khỏe cao hơn.

Tóm lại, bộ dữ liệu có sự phân tán mạnh về chi phí bảo hiểm và phản ánh rõ mối quan hệ giữa nhân khẩu học (tuổi, BMI, số con) và mức phí mà mỗi cá nhân phải chi trả.
Những thông tin trên là cơ sở để lựa chọn các đặc trưng cho việc dự đoán ở dưới.

2.3 Trực quan hóa dữ liệu

2.3.1 Phân bố của dữ liệu

Dữ liệu phân loại



Hình 1: Phân bố của dữ liệu phân loại

Nhận xét về phân phối các biến phân loại

- **Giới tính**: Tỷ lệ nam và nữ trong mẫu gần như tương đương, lần lượt khoảng 50.5% và 49.5%. Điều này cho thấy dữ liệu không bị lệch về giới tính, giúp các phân tích sau này mang tính khách quan hơn, không bị thiên vị cho giới tính nào.

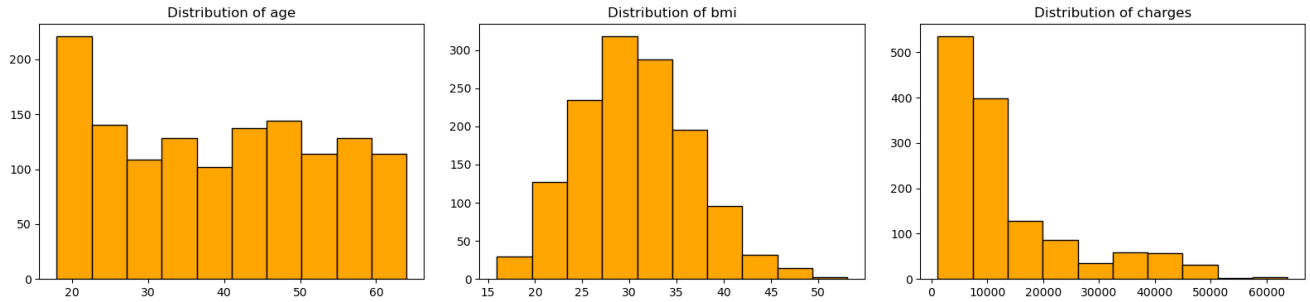
- **Tình trạng hút thuốc**: Phần lớn đối tượng trong mẫu không hút thuốc (khoảng 79.5%), trong khi chỉ khoảng 20.5% có hút thuốc. Sự chênh lệch rõ rệt này có thể dẫn đến sự khác biệt mạnh ở các biến đầu ra như chi phí y tế (*charges*), do hút thuốc là một yếu tố rủi ro quan trọng. Việc tỷ lệ người hút thuốc thấp cũng thể hiện rằng mẫu có cấu trúc thực tế tương đối hợp lý.

- **Số con**: Nhóm có 0 con chiếm tỷ lệ cao nhất (42.9%), tiếp theo là 1 con (24.2%), 2 con và giảm dần ở các mức 3–5 con. Phân phối này cho thấy phần lớn đối tượng trong dataset thuộc các hộ gia đình nhỏ hoặc chưa có con, phù hợp với cấu trúc dân số phổ biến.

- **Khu vực sinh sống**: Bốn vùng *southeast, southwest, northwest, northeast* có tỷ lệ phân phối tương đối đồng đều, dao động quanh mức 24–27%. Điều này giúp đảm bảo không có khu vực nào chiếm ưu thế quá lớn, hỗ trợ việc so sánh chi phí y tế theo vùng được công bằng hơn.

Tổng kết. Nhìn chung, phân phối các biến phân loại trong dữ liệu khá cân bằng, ngoại trừ biến *smoker* có sự mất cân đối đáng kể. Tuy nhiên, đây cũng là đặc điểm phản ánh thực tế xã hội. Cấu trúc dữ liệu như vậy giúp mô hình hóa và phân tích sau này có nền tảng tốt và ít chịu sai lệch do phân phối không đồng đều.

Dữ liệu số



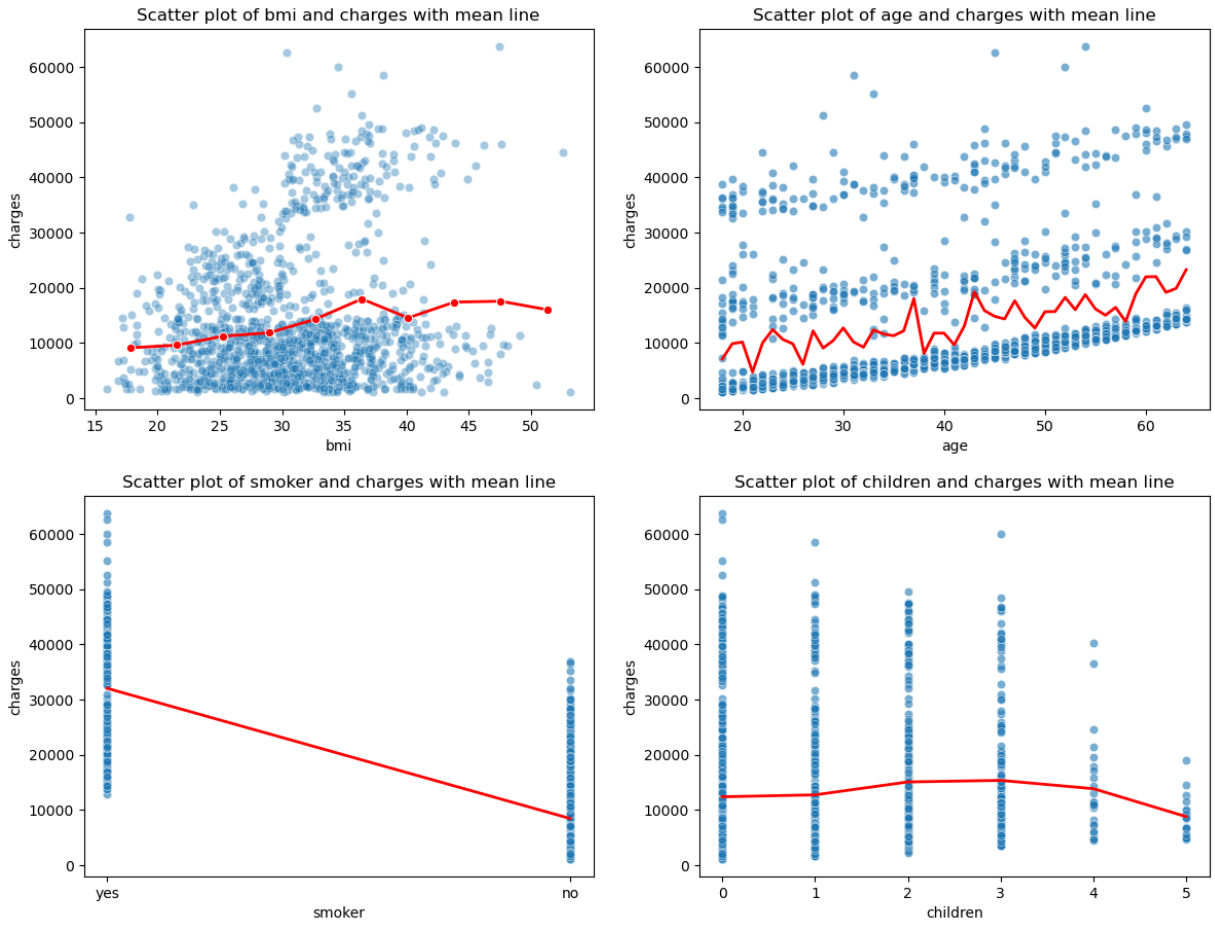
(a) (b) (c)

Hình 2: Phân bố của dữ liệu số

Nhận xét về phân phối của các biến mang giá trị số

- Tuổi tác có phân phối tương đối đồng đều trên toàn bộ dải giá trị (20-60), cho thấy tập dữ liệu có sự cân bằng về đại diện nhóm tuổi, yếu tố quan trọng trong việc dự báo chi phí y tế.
- BMI có phân phối gần chuẩn và tập trung mạnh mẽ xung quanh mức 30-35, điều này chỉ ra rằng phần lớn người tham gia bảo hiểm có BMI ở mức thừa cân/béo phì, một yếu tố nguy cơ cần được theo dõi sát trong mô hình dự báo.
- Chi phí có phân phối lệch phải mạnh, tập trung chủ yếu ở mức thấp và có nhiều giá trị ngoại lai cao. Điều này yêu cầu một phép biến đổi phù hợp.

2.3.2 Kiểm tra quan hệ tuyến tính của trung bình đặc trưng với Charges



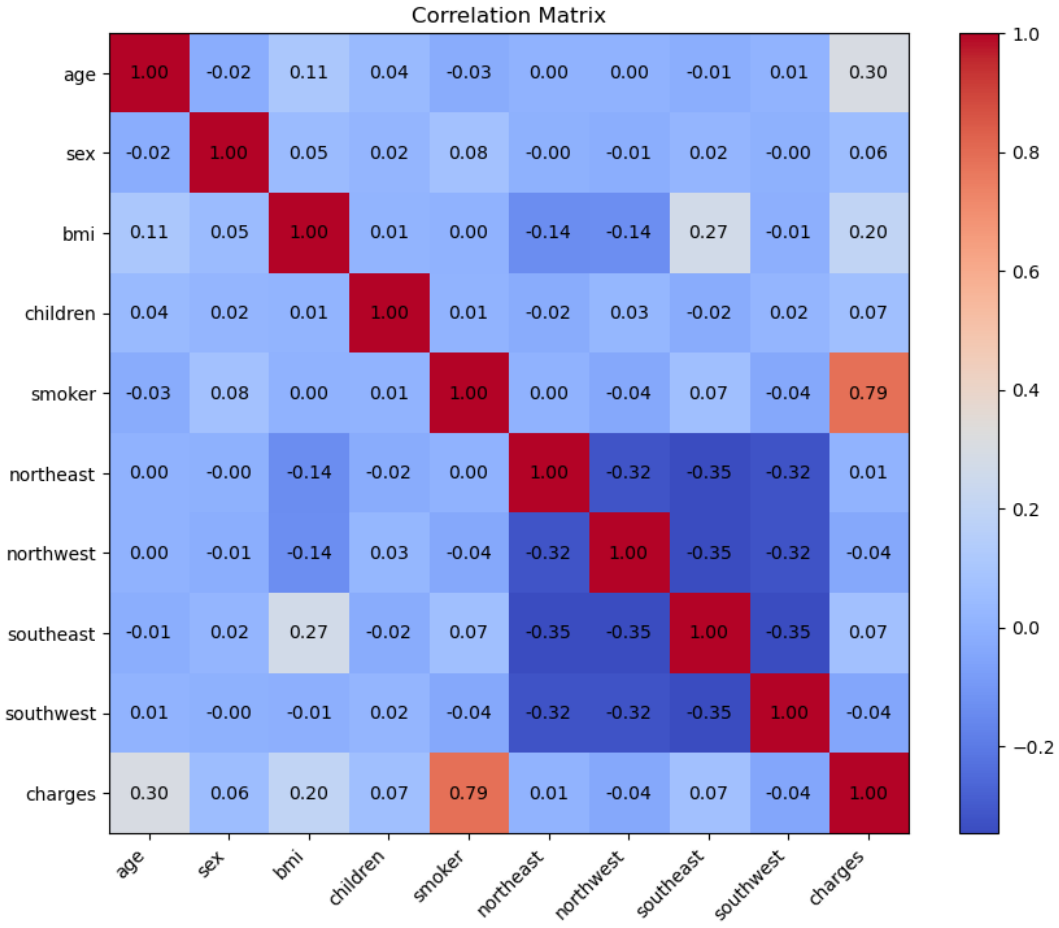
Hình 3: Phân bố của dữ liệu phân loại

Vì phân phối của **Region** và **Sex** tương đối đồng đều nhau, nên chúng ta không xét trung bình các đặc trưng này so với Charges nữa. Các quan đặc trưng mà có phân phối đồng đều thì thường không cho nhiều ý nghĩa trong mô hình tuyến tính (vì nếu đồng đều quá thì mô hình sẽ học như một sự ngẫu nhiên, giống như tung một đồng xu và xác suất 50 - 50, thì việc học chẳng có ý nghĩa).

Nhận xét về của các biểu đồ này

Dễ nhìn thấy: **BMI**, **Age**, và **Smoker** cho chúng ta được mối quan hệ tuyến tính so với **Charges**. Trong đó, **Smoker** cho chúng ta được mối quan hệ rõ ràng và lớn nhất. Các đặc trưng **BMI**, **Age** cho chúng ta thấy được mối quan hệ tăng tuyến tính, nhưng không quá nhiều.
Chúng ta có thể kiểm tra thêm về mỗi quan hệ giữa các biến bằng Ma trận hệ số tương quan.

2.3.3 Ma trận tương quan (correlation matrix)



Hình 4: Ma trận hệ số tương quan

Tương quan với Charges

- **Tương quan mạnh nhất:** Biến smoker có tương quan dương mạnh nhất, với hệ số $\boldsymbol{r = 0.79}$. Điều này cho thấy tình trạng hút thuốc là yếu tố dự báo hàng đầu và có tác động lớn nhất đến chi phí y tế.
- **Tương quan trung bình:** Biến age có tương quan dương ở mức $\boldsymbol{r = 0.30}$, khẳng định chi phí y tế tăng theo tuổi.
- **Tương quan yếu:** Biến bmi có tương quan dương yếu hơn là $\boldsymbol{r = 0.20}$. Các biến sex, children, và region có tương quan rất yếu ($\boldsymbol{\approx 0.00 - 0.07}$), cho thấy chúng không phải là các biến dự báo mạnh đối với chi phí.

Kết luận về Mô hình hóa

Mô hình dự báo chi phí y tế nên ưu tiên tập trung vào các biến smoker, age, và bmi, vì chúng giải thích phần lớn phương sai của biến mục tiêu charges.

3 Thiết kế mô hình và giải thích

3.1 Model 1 - Sử dụng tất cả đặc trưng

Chọn tất cả đặc trưng của mô hình: **age, sex, bmi, children, smoker, region.**
Lý do lựa chọn mô hình:

- Ở ma trận tương quan ở trên có thể thấy những biến đều có mối tương quan (có thể là không cao) với label (charges) nên có thể toàn bộ features và label đều quan hệ tuyến tính.
- Để tạo baseline so sánh.
- Dùng để xem ảnh hưởng từng feature (có thể diễn giải được).

Sau khi one-hot encoding thì biến region tách thành các biến: northeast, northwest, southeast, southwest.
Thêm hệ số bias $\boldsymbol{\theta_0}$ (intercept)

- Bias giúp tăng khả năng biểu diễn của mô hình, giảm training error.
- Cho phép dịch chuyển đường hồi quy để nó không bị ràng buộc đi qua gốc tọa độ.

Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}} + \theta_4 x_{\text{children}} \\ &+ \theta_5 x_{\text{smoker}} + \theta_6 x_{\text{northeast}} + \theta_7 x_{\text{northwest}} \\ &+ \theta_8 x_{\text{southeast}} + \theta_9 x_{\text{southwest}} \end{aligned}$$

3.2 Model 2 - Sử dụng đặc trưng smoker, age và bmi

Chọn tất cả đặc trưng của mô hình: **age, sex, bmi, children, smoker, region.**
Lý do lựa chọn mô hình:

- Theo ma trận tương quan có thể thấy ba biến có độ tương quan cao nhất với **target (charges)** là **smoker, age** và **bmi**.
- Giả thiết nếu model không chọn hết tất cả features mà chọn những feature có độ tương quan cao để xét mức ảnh hưởng đến output và so sánh giá trị hàm lỗi.

Sau khi one-hot encoding thì biến region tách thành các biến: northeast, northwest, southeast, southwest.
Công thức tổng quát của mô hình:

$$y_{\text{charges}} = \theta_0 + \theta_1 x_{\text{smoker}} + \theta_2 x_{\text{age}} + \theta_3 x_{\text{bmi}}$$

3.3 Model 3 - Sử dụng đặc trưng smoker, age, bmi, children and sex

Chọn các đặc trưng của mô hình: **smoker, age, bmi, children, sex**.
Lý do lựa chọn mô hình:

- Từ EDA và ma trận tương quan, các biến *smoker*, *age*, *bmi* có mối liên hệ rõ rệt với label (*charges*). Thêm *children* và *sex* để kiểm tra ảnh hưởng bổ sung — có thể nhỏ nhưng có ý nghĩa diễn giải.
- Mục tiêu: tạo một mô hình đơn giản, nhưng giàu đủ để nắm bắt các ảnh hưởng chính (smoking, tuổi, BMI, số con, giới tính).
- Giữ số lượng biến ở mức vừa phải (5 biến) giúp khi mở rộng bằng các tương tác hoặc tính đa thức (Polynomial) vẫn kiểm soát được số thông số.

Công thức tổng quát của mô hình:

$$y_{\text{charges}} = \theta_0 + \theta_1 x_{\text{smoker}} + \theta_2 x_{\text{age}} + \theta_3 x_{\text{bmi}} + \theta_3 x_{\text{children}} + \theta_3 x_{\text{sex}}$$

3.4 Model 4 - Ridge regression (sử dụng tất cả đặc trưng)

Chuẩn hóa (Áp dụng cho những Model 4, 5, 6, 7, 8, 9)

- Sử dụng kỹ thuật Standard Scaling: biến đổi dữ liệu về cùng một tỷ lệ để mô hình học hiệu quả hơn.
- Mục đích: một feature có giá trị lớn hơn nhiều so với các feature khác, nó sẽ chiếm ưu thế trong việc tính trọng số, dẫn đến mô hình không tối ưu.
- Chuẩn hóa mỗi feature sao cho:
  - Trung bình: $\mu = 0$
  - Độ lệch chuẩn: $\sigma = 1$
- Công thức chuẩn hóa các feature:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

Lý do lựa chọn mô hình:

- Các biến sau one-hot có thể gây đa cộng tuyến (ví dụ giữa các dummy của region). Ridge (L2 penalty) ổn định ước lượng hệ số trong bối cảnh đa cộng tuyến bằng cách co nhỏ các hệ số.
- Giảm variance, giúp tổng quát hoá tốt hơn so với LinearRegression thường trên tập feature khá nhiều mà không chuẩn hoá trước đây, có thể giảm thiểu oferfitting.
- Cho phép giữ tất cả đặc trưng, mọi biến đều có ảnh hưởng nhỏ nhưng cộng dồn.

Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}} + \theta_4 x_{\text{children}} \\ &+ \theta_5 x_{\text{smoker}} + \theta_6 x_{\text{northeast}} + \theta_7 x_{\text{northwest}} \\ &+ \theta_8 x_{\text{southeast}} + \theta_9 x_{\text{southwest}} \end{aligned}$$

Cơ sở lý thuyết

- Hàm loss theo Ridge Regression (L_2 norm) để tối ưu hóa tham số và hyperparater λ:
  - Hàm loss:
$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} \theta_j^2$$
  - Khi $\lambda = 0$, mô hình trở thành hồi quy tuyến tính bình thường.
  - Khi λ lớn, các trọng số bị phạt mạnh, giúp giảm phương sai nhưng tăng bias.
- Lựa chọn λ:
  - Dữ liệu huấn luyện và validation được sử dụng để tìm giá trị λ tối ưu.
  - Thực hiện Grid Search với k-fold (với k = 5) cross-validation trên tập giá trị:
$$\lambda \in [0.001, 0.01, 0.1, 1, 10, 100]$$
- Tiêu chí đánh giá: Negative Mean Squared Error = $-\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

3.5 Model 5 - Lasso regression (sử dụng tất cả đặc trưng)

Lý do lựa chọn mô hình:

- Các biến sau one-hot có thể gây đa cộng tuyến. Lasso (L1 penalty) giúp co một số hệ số về 0, vừa giảm phương sai, vừa giúp lựa chọn đặc trưng tự động.
- Giảm overfitting, đồng thời giữ các biến quan trọng (age, bmi, smoker) và loại bỏ các biến ít ảnh hưởng (ví dụ sex hay một số region dummy).
- Cho phép giữ các biến quan trọng và loại bỏ những biến có ảnh hưởng gần như bằng 0.

Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}} + \theta_4 x_{\text{children}} \\ &+ \theta_5 x_{\text{smoker}} + \theta_6 x_{\text{northeast}} + \theta_7 x_{\text{northwest}} \\ &+ \theta_8 x_{\text{southeast}} + \theta_9 x_{\text{southwest}} \end{aligned}$$

Cơ sở lý thuyết

- Hàm loss theo Lasso Regression (L_1 norm) để tối ưu hóa tham số và hyperparameter λ:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} |\theta_j|$$

- Lựa chọn λ cũng giống với Ridge regression: Thực hiện Grid Search với 5-fold cross-validation trên tập giá trị để tìm giá trị λ tối ưu.

$$\lambda \in [0.001, 0.01, 0.1, 1, 10, 100]$$

- Tiêu chí đánh giá: Negative Mean Squared Error $= -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

3.6 Model 6 - Polynomial regression (sử dụng tất cả đặc trưng)

Khi sử dụng Polynomial Regression với bậc 2, 3,... số lượng feature sau khi biến đổi tăng lên rất nhanh (có cả bình phương, tích chéo). Việc chuẩn hoá là bắt buộc, vì:

- Các feature bậc cao như x^2, x^3 có độ lớn khác nhau so với x
- Nếu không chuẩn hóa, các feature có giá trị lớn (đặc biệt là bậc cao) sẽ chi phối mô hình, làm các hệ số bị lệch.

Lý do lựa chọn mô hình:

- Dựa vào scatter plot có thể thấy mô hình nhìn có vẻ tuyến tính giữa target (charges) và các features, nhưng không thể chắc chắn là tuyến tính.
- Có thể thấy với chỉ số **bmi** thì scatter plots phân bố không đều và không theo một đường hay cung nhất định.
- Dữ liệu không đủ lớn khá phù hợp Polynomial Regression vì hoạt động tốt với dataset nhỏ.
- Giả thuyết mô hình không tuyến tính và thử mô hình hóa quan hệ phức tạp nhưng vẫn giữ tính "linear in parameters".

Có thể số feature bị nổ theo bậc đa thức, dễ bị overfitting nếu dữ liệu nhỏ nên chỉ dùng được với những mô hình có độ phức tạp vừa đủ (d = 2,3,4).

Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}} + \theta_4 x_{\text{children}} \\ &+ \theta_5 x_{\text{smoker}} + \theta_6 x_{\text{northeast}} + \theta_7 x_{\text{northwest}} \\ &+ \theta_8 x_{\text{southeast}} + \theta_9 x_{\text{southwest}} \end{aligned}$$

Cơ sở lý thuyết

- Giả sử bạn có m features đầu vào: $x = (x_1, x_2, x_3, ..., x_m)$, với degree = d (bậc giới hạn của feature).

$$\hat{y}(\mathbf{x}) = \sum_{\substack{a_1,\ldots,a_m \geq 0 \\ a_1+\cdots+a_m \leq d}} \theta_{a_1,\ldots,a_m} \prod_{i=1}^{m} x_i^{a_i}.$$

- Với mô hình ở đây, chọn tất cả features ($m = 9$), $d = 2$, công thức tổng quát có dạng:

$$\hat{y} = \theta_0 + \sum_{i=1}^{9} \theta_i x_i + \sum_{i=1}^{9} \theta_{ii} x_i^2 + \sum_{1 \leq i < j \leq 9} \theta_{ij} x_i x_j$$

$$\begin{aligned} &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + ... + \theta_8 x_8 + \theta_9 x_9 \\ &\quad + \theta_{11} x_1^2 + \theta_{22} x_2^2 + \cdots + \theta_{99} x_9^2 \\ &\quad + \theta_{12} x_1 x_2 + \theta_{13} x_1 x_3 + \cdots + \theta_{56} x_8 x_9. \end{aligned}$$

3.7 Model 7 - Bình phương một đặc trưng (bmi)

Lý do lựa chọn mô hình:

- Chỉ số **bmi** thì scatter plots phân bố không đều và không theo một đường hay cung nhất định nên giả thiết **target (charges)** quan hệ bình phương (Phi tuyến) với **bmi** hay $y_{charges} \sim \theta_3 x_{\text{bmi}}^2$.

Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}}^2 + \theta_4 x_{\text{children}} \\ &+ \theta_5 x_{\text{smoker}} + \theta_6 x_{\text{northeast}} + \theta_7 x_{\text{northwest}} \\ &+ \theta_8 x_{\text{southeast}} + \theta_9 x_{\text{southwest}} \end{aligned}$$

Mô hình được khởi tạo với λ tốt nhất được chọn: $\lambda_{\textbf{best}} = \mathbf{10}$. Sau khi train thì trả về vector hệ số sau:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \end{bmatrix} = \begin{bmatrix} -6866.4902 \\ 249.1932 \\ 77.9588 \\ 5.0085 \\ 587.3745 \\ 22753.8668 \\ 535.0765 \\ -49.9336 \\ -478.6347 \\ -6.5082 \end{bmatrix}$$

Giá trị hàm loss:

- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 36086658.28$
- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = 4144.86$

Nhận xét:

- Biến **bmi** có hệ số bình phương $\theta_3 = 5.0085$, cho thấy mối quan hệ phi tuyến (quadratic) giữa chỉ số BMI và chi phí y tế, phù hợp với giả thuyết $y_{charges} \sim x_{\text{bmi}}^2$.
- Biến **smoker** ($\theta_5 = 22753.87$) vẫn là yếu tố tác động mạnh nhất, chi phí tăng rõ rệt khi người dùng hút thuốc.
- Biến **age** ($\theta_1 = 249.19$) và **children** ($\theta_4 = 587.37$) tác động vừa phải, phản ánh ảnh hưởng tuyến tính đến chi phí.
- Một số biến vùng và sex có hệ số nhỏ hoặc âm ($\theta_2, \theta_7, \theta_8, \theta_9$), cho thấy ảnh hưởng yếu, mô hình đã tự động giảm tầm quan trọng của những biến này.
- Tổng thể, việc thêm x_{bmi}^2 là hợp lý để mô hình có thể giải thích các biến phi tuyến quan trọng mà không làm tăng quá mức độ phức tạp.

3.8 Model 8 - Polynomial regression sử dụng đặc trưng age, bmi, smoker

Lý do lựa chọn mô hình:

- Chọn đặc trưng tương quan lớn ($|\rho|$ cao) nghĩa là biến giải thích được nhiều biến thiên.
- Nếu không chuẩn hóa, các feature có giá trị lớn (đặc biệt là bậc cao) sẽ chi phối mô hình, làm các hệ số bị lệch.
- Giảm phương sai ước lượng và cải thiện tổng quát hóa.
- **smoker** có tương quan rất lớn và tác động thực tế, cạnh đó **bmi** và **age** có tương quan dương đáng kể.
- Hệ số dương của $r_{bmi,charges} \times r_{smoker,charges}$ nếu kết hợp có thể ảnh hưởng lớn đến target.

Có thể số feature bị nổ theo bậc đa thức, dễ bị overfitting nếu dữ liệu nhỏ nên chọn bậc nhỏ $d = 2$. Công thức tổng quát của mô hình:

$$\begin{aligned} y_{\text{charges}} = \theta_0 &+ \theta_1 x_{\text{age}} + \theta_2 x_{\text{sex}} + \theta_3 x_{\text{bmi}} + \theta_{12} x_{\text{age}} x_{\text{sex}} + \theta_{23} x_{\text{sex}} x_{\text{bmi}} + \theta_{13} x_{\text{age}} x_{\text{bmi}} \\ &+ \theta_{11} x_{\text{age}}^2 + \theta_{22} x_{\text{sex}}^2 + \theta_{33} x_{\text{bmi}}^2 \end{aligned}$$

4 Đánh giá mô hình và kết luận

4.1 Mô hình linear regression bậc nhất (mô hình 1, 2, 3)

4.1.1 Model 1 - Sử dụng tất cả đặc trưng

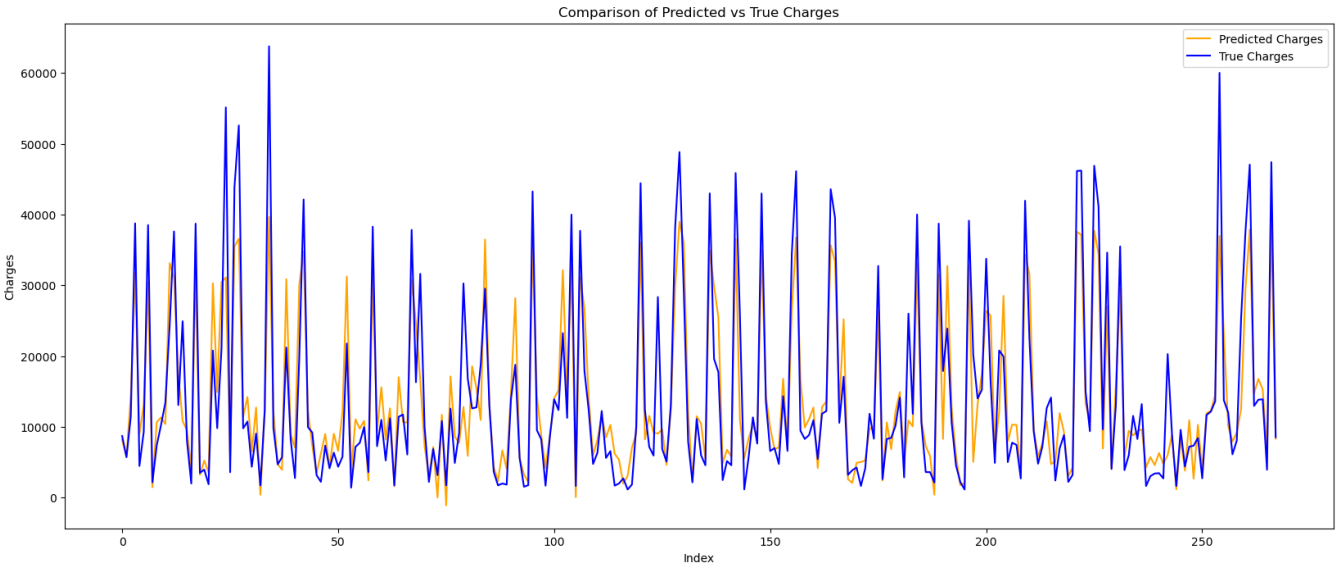
Kết quả thực nghiệm
Sau khi train thì trả về vector hệ số sau:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \end{bmatrix} = \begin{bmatrix} -11565.1075 \\ 248.2107 \\ -101.5421 \\ 318.7014 \\ 533.0100 \\ 23077.7646 \\ 472.4552 \\ 80.6938 \\ -366.4644 \\ -186.6845 \end{bmatrix}$$

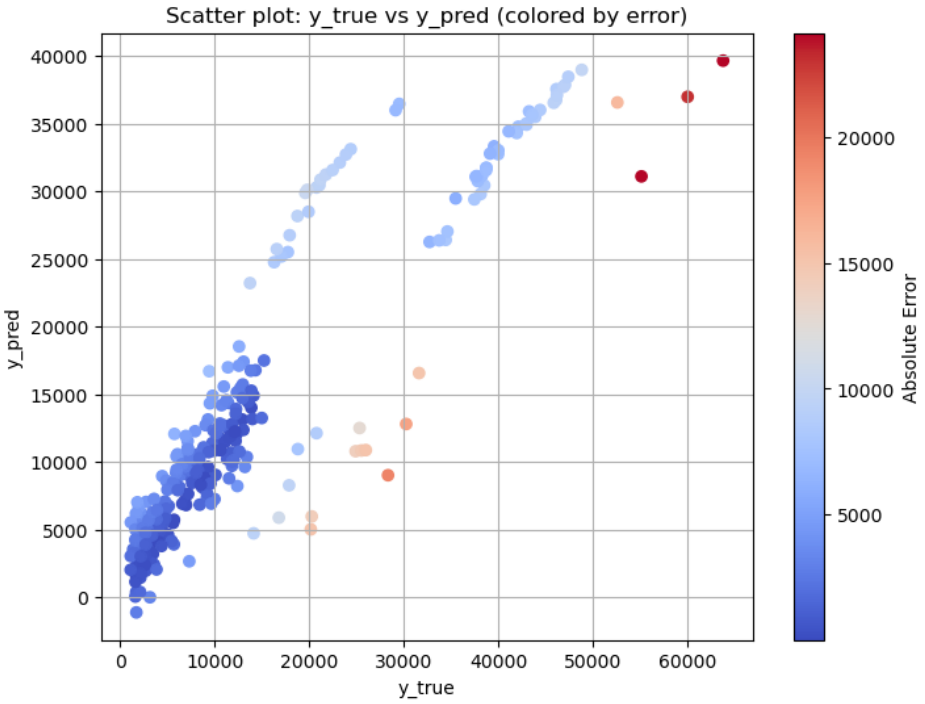
Giá trị hàm loss:

- MSE = 35478020.67
- MAE = 4177.04

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test:



Hình 5: So sánh giữa kết quả dự đoán và kết quả thực tế của model 2 trên tập test



Hình 6: Biểu đồ Scatter giữa kết quả dự đoán và kết quả thực tế

Nhận xét:

- Kết quả dự đoán của mô hình có xu hướng chưa bám sát tốt với giá trị thực tế trên tập kiểm tra, thể hiện qua đường dự đoán (màu cam) gần trùng với đường giá trị thật (màu xanh) ở hầu hết các mẫu.
- Sai số trung bình bình phương (MAE) đạt 4177.04, cho thấy mức độ chênh lệch giữa dự đoán và thực tế là tương đối cao.
- Hệ số $\boldsymbol{x}_{\text{bmi}}$, $\boldsymbol{x}_{\text{age}}$, $\boldsymbol{x}_{\text{children}}$ cũng khá ảnh hưởng đến chi phí, tức là với những người dân thừa cân hoặc cao tuổi hay đông con thì chi phí cũng tăng với 1 chỉ số bmi tương ứng với 318,7, tăng 1 tuổi tương ứng với tăng 248,2 và 533 cho mỗi con.
- Hệ số $\boldsymbol{x}_{\text{smoker}}$ rất cao chứng tỏ có ảnh hưởng rất nhiều đến charges, khi smoker là 1 thì chi phí trả tăng xấp xỉ 23077. Các hệ số còn lại có ảnh hưởng khá ít.
- Với data chưa được chuẩn hóa (scale) → nếu dùng regularization sau này sẽ bị ảnh hưởng
- Về tính đa cộng tuyến giữa one-hot region và sex/smoker có thể làm hệ số nhiễu, nhưng đối với ma trận tương quan giữa các biến thì không có cặp đặc trưng nào có hệ số tương quan quá cao, đa phần: $\boldsymbol{r_{x_1,x_2} \leq 0.3}$
- Model không có phần kiểm soát overfitting.

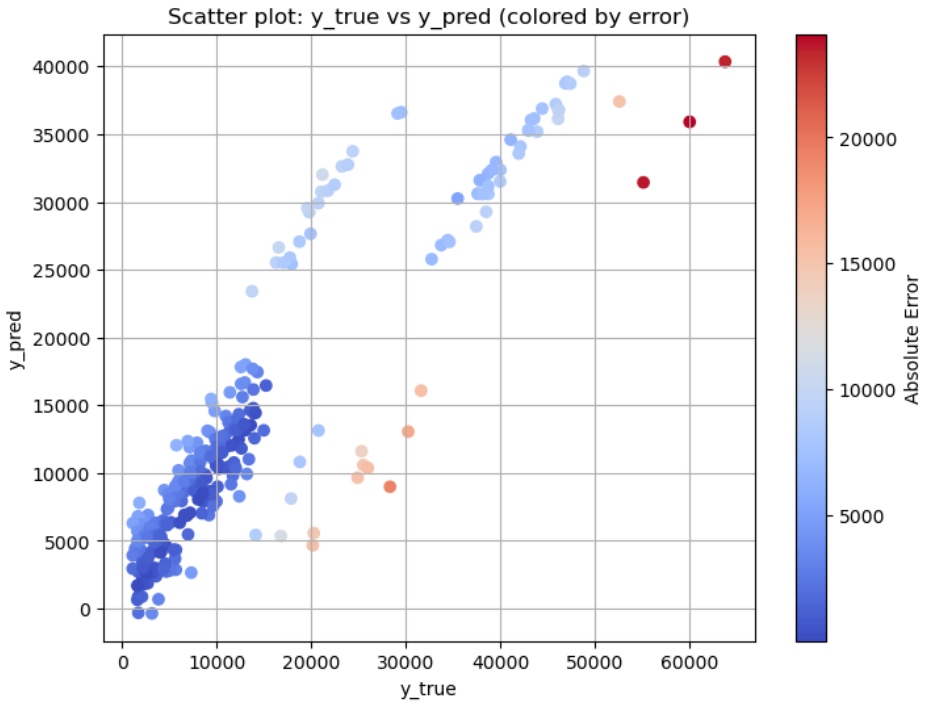
4.1.2 Model 2 - Sử dụng đặc trưng smoker, age và bmi

Kết quả thực nghiệm

Sau khi train thì trả về vector hệ số sau:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} -10770.8878 \\ 23074.2617 \\ 251.8762 \\ 304.8026 \end{bmatrix}$$

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test:



Hình 7: Biểu đồ Scatter giữa kết quả dự đoán và kết quả thực tế

Giá trị hàm loss:

- MSE = 35841574.81
- MAE = 4191.7

Nhận xét:

- Kết quả dự đoán của mô hình có xu hướng chưa bám sát tốt với giá trị thực tế.
- Sai số trung bình tuyệt đối (**MAE**) đạt 4191.7 và Sai số trung bình bình phương (**MSE**), cho thấy mức độ chênh lệch giữa dự đoán và thực tế là khá cao.
- Hệ số của $\boldsymbol{x}_{\text{smoker}}$, $\boldsymbol{x}_{\text{bmi}}$, $\boldsymbol{x}_{\text{age}}$, $\boldsymbol{x}_{\text{children}}$ cũng rất cao nhưng thấp hơn (không đáng kể) so với model 1 (Tất cả đặc trưng).

- Hệ số bias âm lớn, cho thấy mô hình extrapolate kém khi age/bmi rất nhỏ → Có thể dự đoán ra âm nếu xét những trường hợp nhỏ tuổi và chưa hút thuốc ($\boldsymbol{x_{smoker} = 0}$).
- Với model 2, data cũng chưa được chuẩn hóa và cũng không kiểm soát overfitting.

4.1.3 Model 3 - Sử dụng đặc trưng smoker, age, bmi, children and sex

Kết quả thực nghiệm

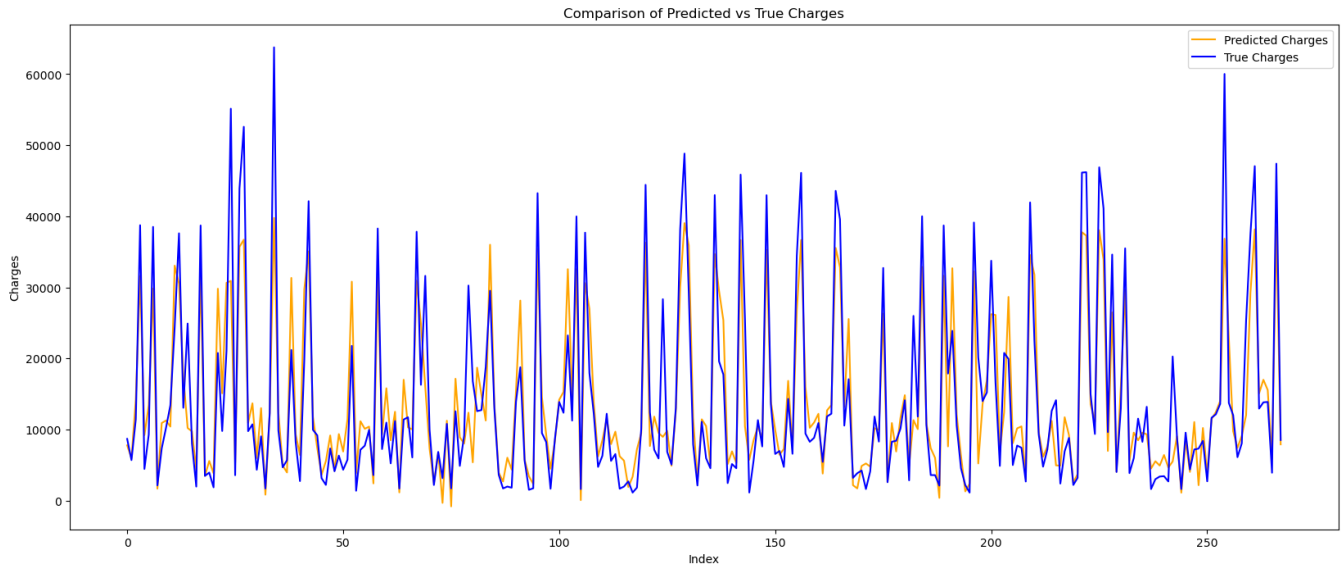
Sau khi train thì trả về vector hệ số sau:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{bmatrix} = \begin{bmatrix} -11221.0517 \\ 23051.3681 \\ 249.0952 \\ 305.5965 \\ 537.9634 \\ -85.027 \end{bmatrix}$$

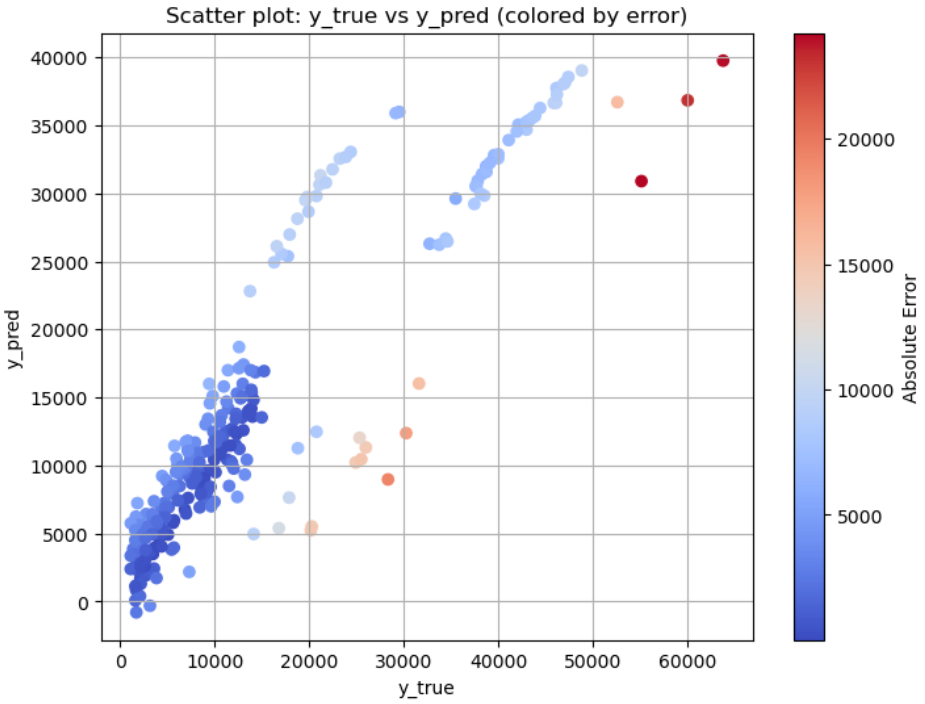
Giá trị hàm loss:

- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 35901914.11$
- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = 4198.11$

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test:



Hình 8: So sánh giữa kết quả dự đoán và kết quả thực tế của model 3 trên tập test



Hình 9: Biểu đồ Scatter giữa kết quả dự đoán và kết quả thực tế

Nhận xét:

- Kết quả dự đoán của mô hình chưa thật sự fit với giá trị thực tế, khá giống với 2 model trên.
- MSE đạt 35901914.11 và MAE đạt 4198.11, cho thấy chênh lệch vẫn còn cao, thông số lỗi vẫn xấp xỉ model 1, 2.
- Kỳ vọng hệ số của $\boldsymbol{x}_{\text{smoker}}$ sẽ là dương và lớn (smoker thường làm tăng mạnh chi phí) và những hệ số còn lại (age, bmi).
- Hệ số của $\boldsymbol{x}_{\text{children}}$ khá nhỏ, nếu nhiều con thì tăng tầm 538 cho mỗi đứa con và giới tính ($\boldsymbol{x}_{\text{sex}}$ = -85.027) dường như không ảnh hưởng đến output.
- Dữ liệu model 3 vẫn chưa được chuẩn hóa, cũng như 2 model trên.

4.2 Mô hình regression với Regularization (mô hình 4, 5)

4.2.1 Model 4 - Ridge regression (sử dụng tất cả đặc trưng)

Kết quả thực nghiệm

Mô hình được khởi tạo với λ tốt nhất được chọn: $\lambda_{\textbf{best}} = \textbf{10}$. Sau khi train thì trả về vector hệ số

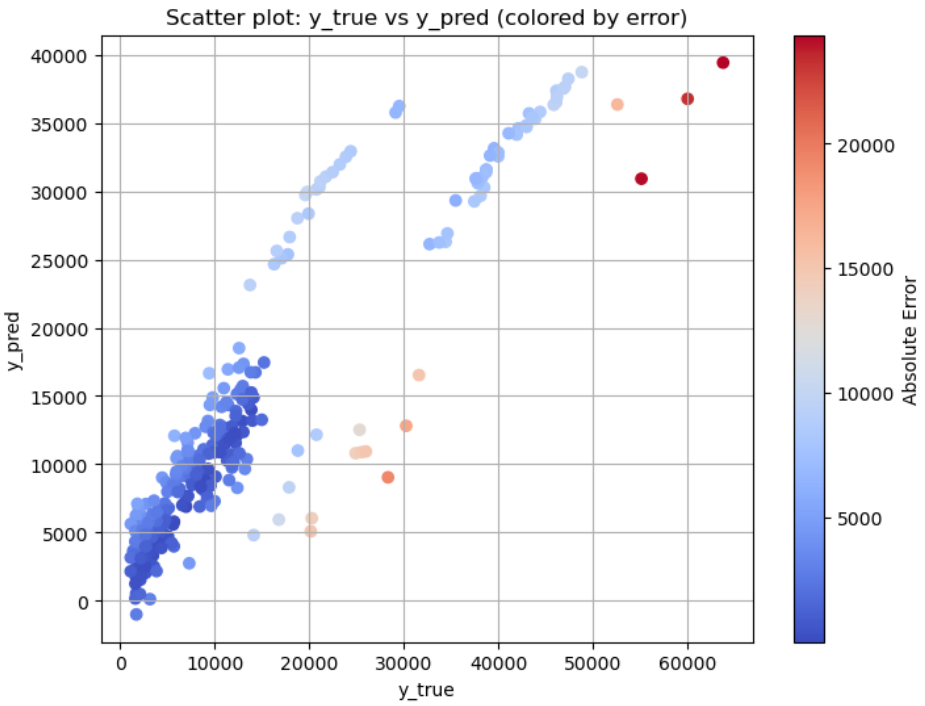
sau:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \end{bmatrix} = \begin{bmatrix} 13048.1342 \\ 3441.2215 \\ -42.9564 \\ 1913.2063 \\ 634.3586 \\ 9170.1726 \\ 201.6211 \\ 34.2218 \\ -148.8671 \\ -77.2448 \end{bmatrix}$$

Giá trị hàm loss:

- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 35826718.50$
- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = 4202.59$

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test:



Hình 10: Biểu đồ Scatter giữa kết quả dự đoán và kết quả thực tế

Nhận xét:

- Kết quả dự đoán của mô hình chưa thật sự khớp với giá trị thực tế, **MSE** đạt 35826718.5 và **MAE** đạt 4202.59, thậm chí cao hơn cả những model linear bậc nhất.

- $\theta_5 = 9170.17$ (smoker) tác động mạnh, chi phí tăng rõ rệt khi người dùng hút thuốc. $\theta_1 = 3441.22$ (age) phí tăng 3441 với mỗi đơn vị tuổi và $\theta_3 = 1913.21$ (bmi) phí tăng 1913 với mỗi đơn vị BMI cao hơn so với những model trên. $\theta_4 = 634.36$ (children) giá trị dương nhưng nhỏ hơn so với age và bmi, thể hiện mức độ ảnh hưởng vừa phải.
- Dữ liệu đã được chuẩn hóa các feature trước khi train, ridge Regression đã giúp ổn định các hệ số khi dữ liệu có đa cộng tuyến, nhờ L2 penalty co nhỏ các trọng số. Ridge Regression giảm phương sai (variance) nhưng có thể làm tăng bias một chút, là sự đánh đổi trong regularization.
- Mô hình phù hợp khi số lượng feature nhiều, đặc biệt khi có các feature liên quan tuyến tính với nhau, nhưng với bộ dữ liệu này thì tuyến tính giữa các features không cao nên chỉ giảm thiểu được phần nhỏ và hạn chế overfitting không đáng kể.

4.2.2 Model 5 - Lasso regression (sử dụng tất cả đặc trưng)

Kết quả thực nghiệm

Mô hình được khởi tạo với λ tốt nhất được chọn: $\lambda_{\text{best}} = 100$. Sau khi train thì trả về vector hệ số:

$$\boldsymbol{\theta} = \begin{bmatrix} 13048.1342 \\ 3441.2215 \\ -42.9564 \\ 1913.2063 \\ 634.3586 \\ 9170.1726 \\ 201.6211 \\ 34.2218 \\ -148.8671 \\ -77.2448 \end{bmatrix}$$

Giá trị hàm loss:

- MSE = 36258685.49
- MAE = 4199.58

Nhận xét:

- **MSE** đạt 36258685.49 và **MAE** đạt 4199.58, mô hình Lasso regression khớp khá giống, thông số hàm loss không thua kém bao nhiêu với Ridge Resression.
- Các biến quan trọng với chi phí y tế như smoker ($\theta_5 = 9141.57$), age ($\theta_1 = 3395.42$) và bmi ($\theta_3 = 1787.91$) vẫn giữ trọng số lớn. Một số hệ số biến ít quan trọng về 0 (sex: $\theta_2 = 0$, northwest: $\theta_7 = 0$, southwest: $\theta_9 = 0$), thể hiện khả năng lựa chọn đặc trưng tự động của Lasso. Hệ số children ($\theta_4 = 548.68$) và một số vùng còn lại có giá trị nhỏ, phản ánh ảnh hưởng vừa phải đến chi phí.

- Dữ liệu đã chuẩn hóa, giúp Lasso hoạt động hiệu quả và các trọng số có thể so sánh được, giảm phương sai và chọn lọc biến.

4.3 Mô hình linear regression bậc cao - Polynomial Regression (mô hình 7, 8)

4.3.1 Model 7 - Bình phương một đặc trưng (bmi)

Kết quả thực nghiệm

Sau khi train thì trả về vector hệ số - biểu diễn dạng bảng:

Đặc trưng	Hệ số
x_1	-0
x_2	3411.2622
x_3	-247.7251
x_4	2041.6764
x_5	1089.3506
x_6	2916.3025
x_7	81.5829
x_8	-18.1436
x_9	-23.2473

Bảng 2: Hệ số bậc 1

Đặc trưng	Hệ số
x_1^2	-37.5984
x_2^2	930.6313
x_3^2	47.0217
x_4^2	67.5703
x_5^2	3600.3155
x_6^2	121.2817
x_7^2	149.3303
x_8^2	-257.3271
x_9^2	-1.4046

Bảng 3: Hệ số bậc 2

Đặc trưng	Hệ số
x_1x_2	-1.4046
x_1x_3	-143.9279
x_1x_4	-120.3971
x_1x_5	132.3754
...	...
x_8x_9	-43.7771

Bảng 4: Hệ số tổ hợp giữa 2 đặc trưng

Giá trị hàm loss:

- MSE = 21954573.13
- MAE = 2910.43

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test:

Hình 11: Biểu đồ scatter so sánh giữa giá trị dự đoán của mô hình 7 và giá trị thực tế

Nhận xét:

- Mô hình 7 đạt MSE = 21.95M và MAE = 2910.43, thấp hơn đáng kể so với mô hình linear cơ bản (khoảng 35M). Điều này cho thấy việc thêm thành phần phi tuyến x_{bmi}^2 giúp mô hình nắm bắt tốt hơn mối quan hệ giữa BMI và chi phí y tế.
- Các biến quan trọng với chi phí y tế như **smoker** ($\theta_5 = 1089.35$), **age** ($\theta_1 = -0$), và **bmi** ($\theta_3 = -247.73$) vẫn giữ trọng số đáng kể trong mô hình, phản ánh ảnh hưởng rõ rệt đến

target. Hệ số bậc 2 lớn, ví dụ $x_2^2 = 930.63$, $x_5^2 = 3600.32$, cho thấy các mối quan hệ phi tuyến giữa features và target có ảnh hưởng lớn.

- Một số biến tuyến tính hoặc tương tác có hệ số gần bằng 0, ví dụ **sex** (θ_2) và một số hệ số interaction, cho thấy mô hình tự động giảm ảnh hưởng của những biến ít quan trọng, giúp tránh overfitting.
- Mô hình đã chuẩn hóa các feature trước khi train, giúp các hệ số có thể so sánh trực tiếp và giảm phương sai, đồng thời giữ được hiệu quả dự đoán trên dataset nhỏ.
- Nhìn chung, mô hình phù hợp để nắm bắt các quan hệ phi tuyến trong dữ liệu.

4.3.2 Model 8 - Polynomial regression sử dụng đặc trưng age, bmi, smoker

Kết quả thực nghiệm
Sau khi train thì trả về vector hệ số:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_{12} \\ \theta_{23} \\ \theta_{13} \\ \theta_{11} \\ \theta_{22} \\ \theta_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 3594.7615 \\ 1881.7831 \\ 2882.1517 \\ 525.7191 \\ 220.2195 \\ -47.2635 \\ -347.4131 \\ 3527.6282 \\ 4319.0171 \end{bmatrix}$$

Giá trị hàm loss:

- MSE = 21400305.65
- MAE = 2814.39

Biểu đồ so sánh nhãn dự đoán và nhãn dự đoán trên tập test

Hình 12: Biểu đồ scatter so sánh giữa giá trị dự đoán và giá trị thực tế

Nhận xét:

- Model 8 cho kết quả tốt hơn Model 7 với **MSE = 21.40M** và **MAE = 2814.39**, tiếp tục giảm so với mô hình trước. Việc đưa vào đầy đủ các thành phần bậc hai và tương tác giữa **age**, **bmi** và **smoker** giúp mô hình diễn tả được các mối quan hệ phi tuyến mạnh hơn, đặc biệt là tương tác giữa hút thuốc và BMI cao.
- Các biến quan trọng nhất vẫn là **smoker** ($\theta_3 = 2882.15$), **age** ($\theta_1 = 3594.76$) và **bmi** ($\theta_{33} = 4319.02$), phản ánh tác động mạnh mẽ đến chi phí y tế. Hệ số tương tác dương giữa các biến, ví dụ $x_{\text{age}}x_{\text{sex}}$ ($\theta_{12} = 525.72$) và $x_{\text{sex}}x_{\text{bmi}}$ ($\theta_{23} = 220.22$), cho thấy khi kết hợp các feature này, chi phí có xu hướng tăng thêm.
- Có một số hệ số âm hoặc nhỏ như $x_{\text{age}}x_{\text{bmi}}$ ($\theta_{13} = -47.26$) hay x_{age}^2 ($\theta_{11} = -347.41$), cho thấy mối quan hệ phi tuyến nhưng có ảnh hưởng nhỏ. Hệ số bậc 2 lớn, đặc biệt $x_{\text{bmi}}^2 = 4319.02$, chứng tỏ mối quan hệ phi tuyến giữa BMI và chi phí y tế là đáng kể.
- Tổng thể, mô hình capture tốt mối quan hệ phi tuyến quan trọng giữa BMI, age và smoker với chi phí y tế, đồng thời giữ số feature và bậc polynomial hợp lý để kiểm soát độ phức tạp.

4.4 Đánh giá mô hình

4.4.1 So sánh 8 mô hình

Dựa trên các kết quả thực nghiệm của các mô hình từ Linear Regression cơ bản đến Ridge, Lasso và Polynomial Regression, ta tổng hợp được bảng so sánh như sau:

Model	Đặc trưng sử dụng	Dạng mô hình	MSE	MAE
1	Tất cả features	Linear (bậc 1)	35.48M	4177
2	age, bmi, smoker	Linear (bậc 1)	35.84M	4191
3	age, bmi, smoker, children, sex	Linear (bậc 1)	35.90M	4198
4	Tất cả features	Ridge	35.83M	4202
5	Tất cả features	Lasso	36.26M	4199
7	Bậc 2, tất cả feature	Polynomial	21.95M	2910
8	**Bậc 2, age–bmi–smoker**	**Polynomial**	**21.40M**	**2814**

Bảng 5: So sánh hiệu năng các mô hình hồi quy

4.4.2 Nhận xét tổng quan

Dựa trên các thí nghiệm, ta rút ra một số nhận định quan trọng như sau:

Các mô hình Linear Regression (1, 2, 3) cho kết quả tương đối giống nhau: Các mô hình tuyến tính có **MSE** ổn định quanh **35M–36M** và **MAE** khoảng **4200**, cho thấy quan hệ giữa đầu vào và đầu ra không hoàn toàn tuyến tính. Việc thêm hoặc bớt một số đặc trưng chỉ tạo ra thay đổi rất nhỏ. Những đặc trưng quan trọng nhất liên tục lặp lại ở mọi mô hình gồm *smoker*, *bmi*, và *age*.

Các mô hình regularization (Ridge, Lasso) ổn định hệ số nhưng không giảm sai số: Ridge giúp giảm độ lớn của hệ số và giảm variance, trong khi Lasso thực hiện chọn lọc đặc trưng. Tuy vậy, cả hai mô hình đều không giúp giảm **MSE** hay **MAE**. Nguyên nhân là quan hệ trong dữ liệu không phải do đa cộng tuyến, mà mô hình tuyến tính thiếu khả năng biểu diễn.

Polynomial Regression cải thiện hiệu năng vượt trội: Các mô hình bậc hai giúp giảm MSE từ khoảng **35M** xuống còn **21M**, và giảm **MAE** từ khoảng **4200** xuống còn gần **2800**. Điều này chứng tỏ quan hệ phi tuyến giữa **age**, **bmi**, **smoker** và chi phí y tế là rất rõ ràng. Đặc biệt, **Model 8** sử dụng ít đặc trưng nhưng vẫn đạt hiệu năng tốt nhất, cho thấy việc chọn đúng đặc trưng quan trọng hiệu quả hơn việc thêm nhiều đặc trưng nhiễu.

4.4.3 Kết luận cuối cùng.

- Các mô hình tuyến tính bậc một không đủ mô tả dữ liệu do bản chất phi tuyến mạnh.
- Ridge và Lasso cải thiện sự ổn định mô hình nhưng không cải thiện sai số.

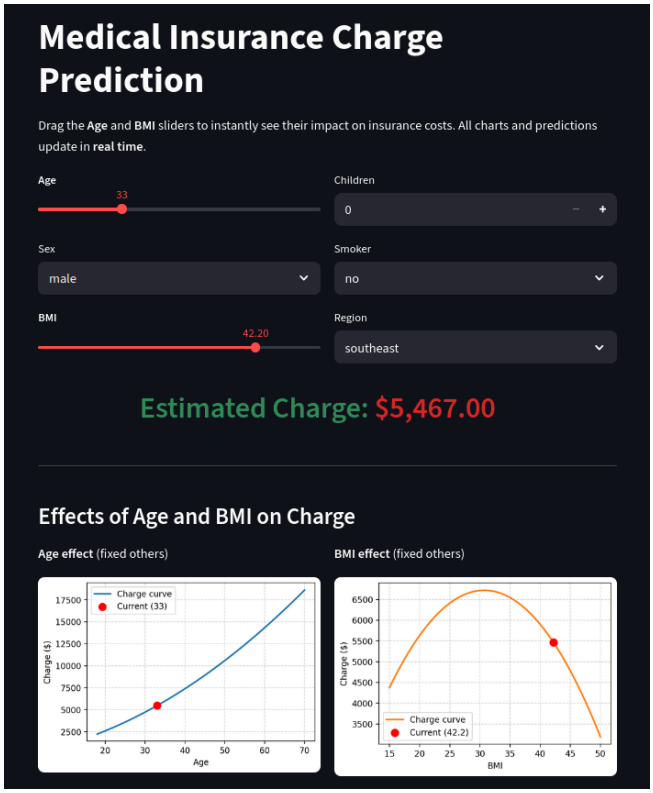
- Polynomial Regression cho kết quả vượt trội, đặc biệt với các đặc trưng quan trọng nhất.

Mô hình tốt nhất: Model 8 – Polynomial Regression (**age**, **bmi**, **smoker**, bậc 2), với **MSE = 21.40M** và **MAE = 2814**, đạt hiệu năng cao nhất và tổng quát hoá tốt nhất.

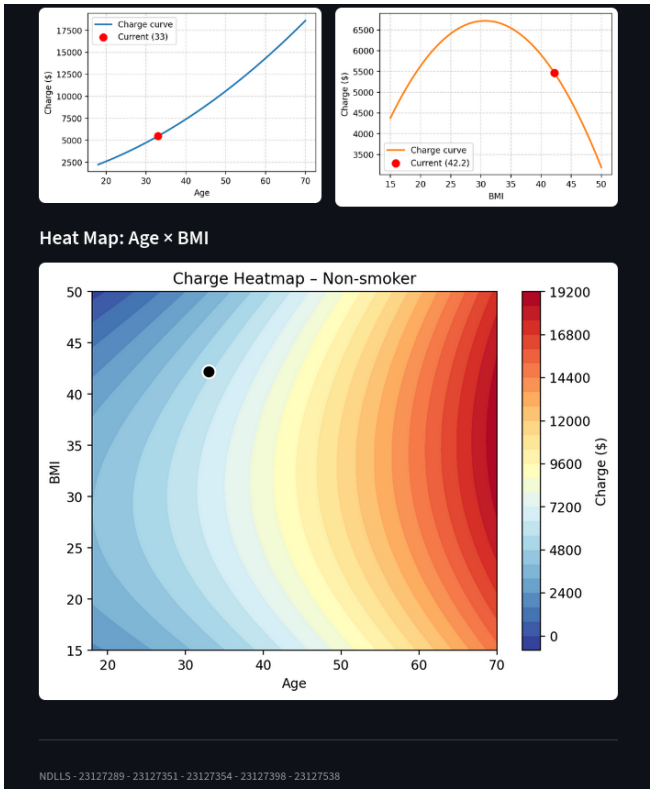
5 Mô tả ứng dụng

- Ứng dụng web Dự đoán Chi phí Bảo hiểm Y tế được phát triển nhằm mục đích trực quan hóa và dự báo tức thời chi phí bảo hiểm cho một cá nhân dựa trên các thông số đầu vào.
- Mục tiêu chính là cung cấp một công cụ trực quan và dễ sử dụng để người dùng (ví dụ: nhân viên công ty bảo hiểm, nhà phân tích) có thể đánh giá nhanh mức độ ảnh hưởng của từng yếu tố (như tuổi tác, tình trạng hút thuốc, ...) lên chi phí cuối cùng.

5.1 Tính năng Chính và Giao diện Người dùng



Hình 13: Giao diện Nhập liệu và Kết quả Dự báo



Hình 14: Trực quan hóa Tác động của Biến số

- **Giao diện nhập liệu (Hình 13):** Giao diện được thiết kế bao gồm các thành phần:
  1. Thanh trượt (Slider) cho các biến liên tục như Age và BMI, cho phép người dùng thay đổi giá trị và xem kết quả cập nhật theo thời gian thực.

2. Hộp chọn (Dropdown) và Thanh đếm (Counter) cho các biến phân loại/rời rạc như Sex, Smoker, Children, và Region.

- **Hiển thị kết quả (Hình 13):** Kết quả dự báo (Estimated Charge) được hiển thị rõ ràng và nhấn mạnh ngay bên dưới khu vực nhập liệu.

- **Trực quan hóa tác động (Hình 14):** Ứng dụng tích hợp các biểu đồ phân tích để giải thích kết quả. ***Khi chúng ta thay đổi thanh trượt thì những biểu đồ này cũng thay đổi theo***:

1. Effects of Age/BMI on Charge: Biểu đồ thể hiện tác động của Age và BMI lên chi phí khi các yếu tố khác được giữ cố định.
2. Heat Map: Age x BMI: Bản đồ nhiệt trực quan hóa mối quan hệ kết hợp giữa Age và BMI đối với Charge, giúp người dùng hiểu rõ hơn về sự tương tác của hai biến này.

5.2 Chức năng nhập/xuất

- **Đầu vào (Input):** Ứng dụng chấp nhận **7** biến đầu vào từ người dùng: Age, Sex, BMI, Children, Smoker, và Region. Các biến này được chuẩn hóa/mã hóa tương ứng với yêu cầu của mô hình đã huấn luyện.

- **Đầu ra (Output):** Đầu ra chính là giá trị Estimated Charge (Chi phí Dự kiến) dưới dạng số tiền ($), được làm tròn và hiển thị ngay lập tức sau khi mô hình xử lý.

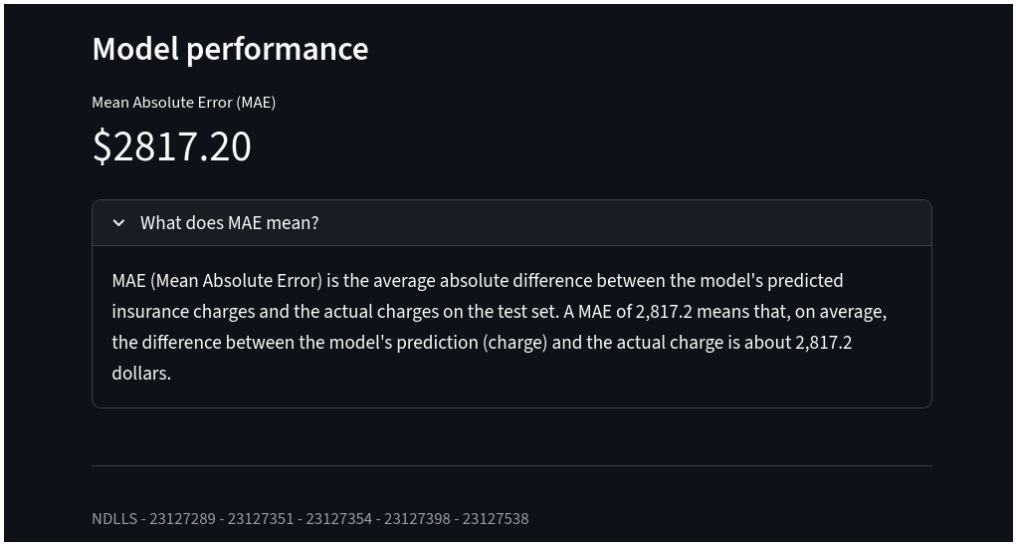
5.3 Tích hợp mô hình huấn luyện

- **Mô hình nền tảng:** Ứng dụng sử dụng **mô hình 8** đã được huấn luyện và nêu ở trên.

- **Quy trình tích hợp:**

1. Các giá trị đầu vào của người dùng được thu thập từ giao diện người dùng.
2. Dữ liệu được tiền xử lý theo đúng cách đã áp dụng trong quá trình huấn luyện mô hình (theo như pipepline).
3. Mô hình này nhận input mà người dùng nhập và kéo ở trên giao diện và trả về kết quả dự báo.
4. Sau khi nhận kết quả trả về, kết quả được hiển thị trên giao diện.
5. Bên cạnh việc dựa vào input của người dùng để dự đoán số tiền, mô hình còn dự đoán thêm các kết quả trên các biến số liên tục, và vẽ hình trực quan trong thời gian thực.

5.4 Hiệu suất mô hình



Hình 15: Hiệu suất của mô hình trên tập kiểm tra

Ứng dụng của chúng tôi được thêm phần hiệu suất của mô hình ở phần giao diện, từ đó người dùng có thể có những góc nhìn tốt hơn, hiểu rõ hơn về mô hình và tự quyết định rằng có nên tin tưởng mô hình hay không.

6 Tài liệu tham khảo

1. Standard Scaler, scikitlearn, url
2. scikitlearn, Linear Regression, url
3. scikitlearn, Ridge Regression, url
4. scikitlearn, Lasso Regression, url
5. scikitlearn, Polynomial Regression, url
6. Machine Learning with Scikit Learn, Keras and TensorFlow book
7. A Modern Approach to Regression with R book - Springer
8. Grok, Create UI using streamlit, link chat
9. Ken Jee, Titanic Project Example, url, [07/11/2025]
10. veronika kachmar, Medical Insurance Cost Prediction and EDA, url, [07/11/2025]
11. aly ashoush, Medical Insurance Cost Prediction, url, [07/11/2025]
12. Geeksforgeeks, Data Pre-Processing with Sklearn using Standard and Minmax scaler, url, [07/11/2025]
13. NeuralNine, Matplotlib Full Python Course - Data Science Fundamentals, url,[07/11/2025]

7 Lời cảm ơn

Em xin cảm ơn thầy *Bùi Tiến Lên*, thầy *Lê Nhựt Nam*, và thầy *Võ Nhật Tân* đã giúp đỡ, dời deadline, và giải đáp thắc mắc của chúng em trong quá trình học cũng như trong quá trình làm đồ án này.
Đồ án này có sự giúp đỡ chatGPT, Grok, Gemini trong việc viết Latex, tìm kiếm thông tin các hàm để trực quan hóa dữ liệu.