Modelo de Regressão Lasso para os custos pessoais em Seguro de Saúde nos Estados Unidos

Caroline Cogo Carneosso João Inácio Scrimini Joelmir Junior Renata Stone

Sumário

1 Introdução **1**

2 Análise exploratória **2**

3 Modelagem **8**

4 Conclusão **12**

1 Introdução

As áreas financeiras e de seguro são recorrentes na aplicação de modelos estatísticos para a tomada de decisão em negócios. O banco de dados utilizado em nosso estudo, seguindo esse contexto, o objetivo é fazer uma previsão de gastos com Seguro de Saúde privado nos Estados Unidos. O qual leva em consideração algumas variáveis, como, a idade do beneficiário, número de dependentes, o IMC que fornece uma compreensão do corpo, pesos relativamente altos ou baixos em relação à altura, índice objetivo de peso corporal (kg / m^2) usando a relação altura/peso, idealmente 18.5 a 24.9, dentre outras.

O banco de dados possui 1338 observações e 7 variáveis, que estão descritas na Tabela 1.

Tabela 1: Descrição da váriaveis

Variável	Descrição
Idade	Idade do beneficiário principal.
Genero	Gênero do contratante do seguro, feminino ou masculino.
IMC	Índice de massa corporal.
Filhos	Número de filhos cobertos pelo seguro saúde / Número de dependentes.
Fumante	Se fuma ou não.
Regiao	A área residencial do beneficiário nos EUA, nordeste, sudeste, sudoeste, noroeste.
Encargos	Custos médicos individuais faturados pelo seguro de saúde.

O presente trabalho tem como objetivo aplicar um modelo de Machine Learning, através da regressão Lasso no contexto de encargos de seguro saúde. A fim de obter um modelo que consiga identificar as características presente nos dados, e tenha um bom desempenho. Para isso, inicialmente, será realizado o tratamento dos dados e uma análise exploratória, a fim de entender seu comportamento. Na parte da modelagem o banco de dados é dividido em treino e teste, será realizada a validação cruzada e a identificação do hiperparâmetro.

2 Análise exploratória

Nesta seção, analisaremos as variáveis presentes no estudo, buscando pré identificar o comportamentos dos dados e associações relevantes. Na qual optamos por modificar a variável resposta Encargos dividido-a em mil, assim facilitando a apresentação dos resultados e visualização gráfica. As análises foram feitas por meio de tabelas descritivas, gráficos de colunas e boxplots. Na Tabela 2 é apresentado a análise descritiva para as variáveis quantitativas em estudo. Inicialmente, observamos que não é apresentado valores faltantes nos dados. A Idade mínima no banco de dados foi de 18 anos e a máxima de 64 anos, com média de aproximadamente 39 anos. No IMC, com base no índice ideal de 18.5 até 24.9, é notado que a média entre dos indivíduos presentes no banco de dados foi de 30.66, maior que o ideal. Assim, concluimos que a maioria das pessoas estão acima do IMC ideal, possivelmente apresentando algum nível de obesidade, chegando em alguns casos em IMC de 53.1, valor muito elevado, mais que o dobro do máximo ideal. Observamos também que existem alguns caso com valores de IMC inferiores a 18.5, e em alguns casos com um valor de 16. O máximo de filhos entre os beneficiários foi de 5, e a média de filhos é de 1 por pessoa/família. Para os Encargos médicos mínimo de gastos de 1122 e máximo de 63770.4, com média de 13270.42, pelo valor da média em relação ao máximo, já podemos esperar que existam *outliers*, devido à grande diferença.

Tabela 2: Análise descritiva das variáveis quantitativas em estudo.

variable	mean	median	sd	min	max	na_count
Idade	39.21	39.0	14.05	18	64.0	0
IMC	30.66	30.4	6.10	16	53.1	0
Filhos	1.09	1.0	1.21	0	5.0	0
Encargos	13270.42	9382.0	12110.01	1122	63770.4	0

Pelas Figuras 1 e 2 são apresentados os gráficos de boxplots para as variáveis quantitativas em estudo, mencionadas na Tabela 2, e separadas por gênero. Nesses gráficos a Idade e quantidade de Filhos exibiram pouca diferença entre feminino e masculino. Para o IMC, podemos visualizar que a média em relação ao gênero possui valores similares, porém temos *outliers*, mostrando que existem beneficiários com obesidade nos dois gêneros, entretanto o masculino apresentando valores do IMC mais elevados nesses *outliers*. Nos Encargos, como foi mencionado anteriormente sobre possíveis *outliers* devido a grande diferença entre a média e o máximo, conseguimos concluir isso agora, os dois gêneros apresentam *outliers*, visualmente o gênero feminino apresenta *outliers* em maiores quantidades, porém com gastos entre o primeiro e terceiro quantil em menor quantidade, em relação ao masculino.

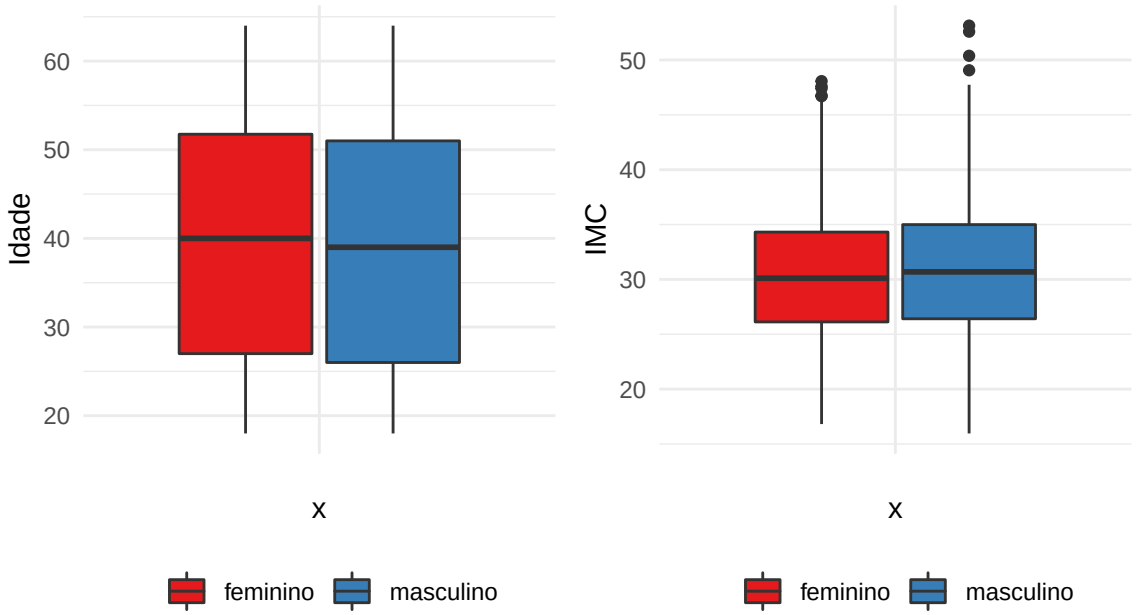

Figura 1: Boxplot para Idade e IMC.

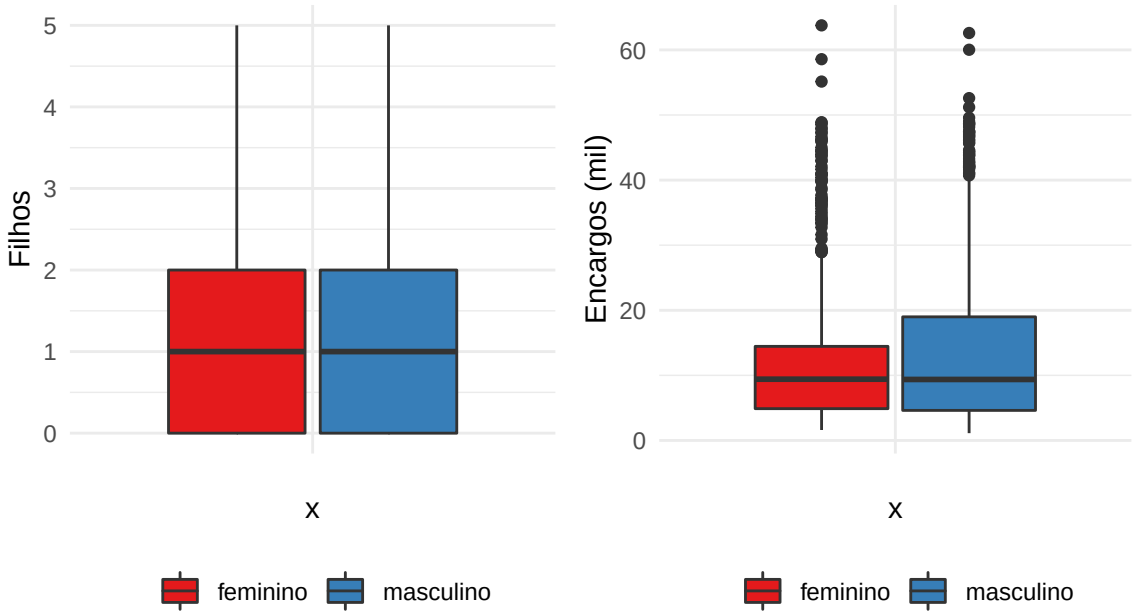

Figura 2: Boxplot para Filhos e Encargos (mil).

Para a sequência da análise, foi optado categorizar as variáveis Idade e IMC, com o intuito de facilitar algumas análises. Esta alteração será implementada somente para a parte exploratória em alguns casos, como análise primária, sendo desfeitas para a parte de modelagem. Assim, para a Idade temos os seguintes níveis (faixas etárias), 18 a 25 anos, 26 a 33 anos, 34 a 41 anos, 42 a 49 anos, 50 a 57 anos e 58 a 65 anos. E para IMC optamos por separar em 3

níveis, “abaixo”, que se refere as pessoas com IMC abaixo do ideal de 18.5, “acima, referindo-se aos que apresentam IMC acima do ideal de 24.9, e”ideal” para os que apresentam IMC dentro dos Índices ideais de 18.5 a 24.9.

Com isso, na Figura 3 temos o gráfico de boxplot para o IMC (não categorizado) vs as Faixas Etárias separados por gênero, percebe-se que as faixas de 18 a 25 anos, 34 a 41 anos e 42 a 49 anos, apresentam *outliers*, sendo com valores mais elevados nos pacientes de menor idade e masculinos, na faixa de 18 a 25 anos, com valores de IMC maiores que 50, sendo maior que o dobro do ideal de 24.9. Na faixa de 34 a 41 anos e do gênero masculino temos um único *outliers* abaixo dos limites, com valores menores que o ideal de 18.5 para o IMC. Outro fator interessante de observar neste gráfico é na faixa de 58 a 65 anos para o gênero masculino, como visualizado ele não apresenta *outliers*, porém é nítido como a média está acima dos demais, junto com seus limites máximo e mínimo. Já para o gênero feminino nesta mesma faixa isso não ocorre, apresentando limites mínimo e máximo menores comparado com o masculino da mesma faixa.

Na Figura 4 é apresentado o boxplot para os Fumantes vs Encargos separados por gênero. Neste gráfico podemos notar uma grande diferença entre fumar ou não nos valores de Encargos médicos, sendo muito influente a pessoa fumar, aumentando consideravelmente os Encargos médicos. Para os indíviduos que não fumam, nota-se que existem vários *outliers*, onde ocorrem gastos elevados devido a outros possíveis problemas médicos ocorridos. Para os que fumam o gênero masculino tem uma média mais elevada que para o gênero feminino, já os que não fumam apresentam-se bem similares a isso e também em relação aos *outliers*.

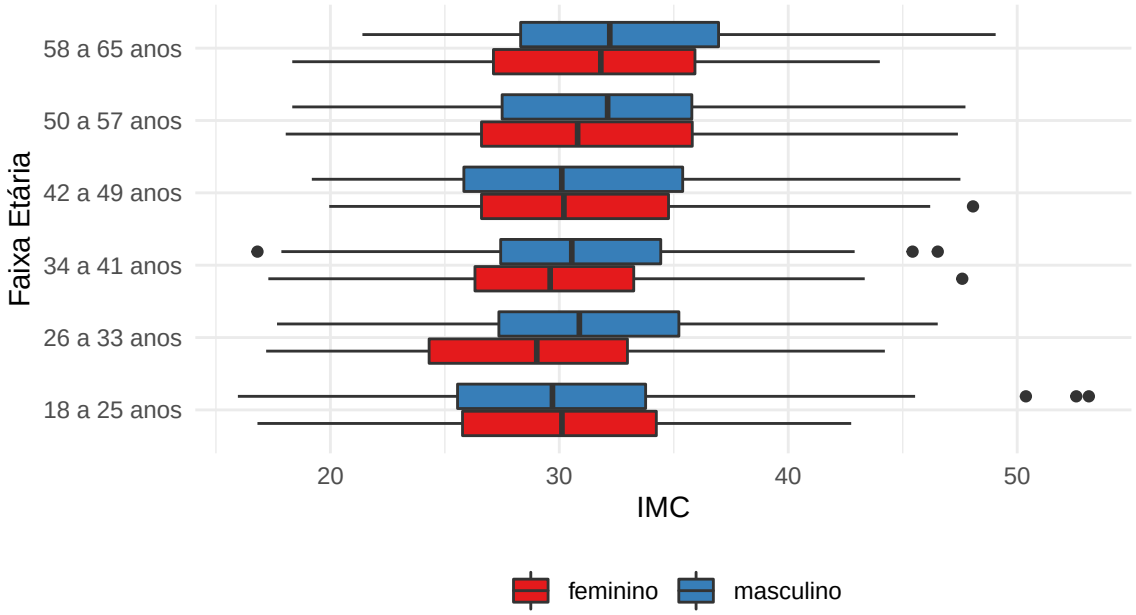

Figura 3: Box plot IMC vs Faixa etária.

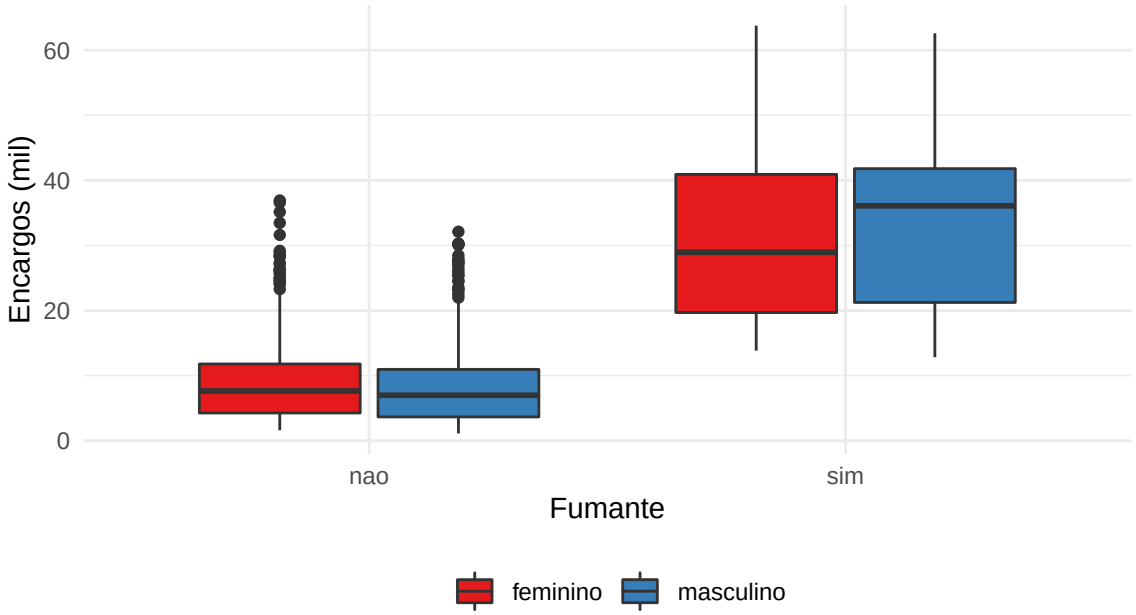

Figura 4: Box plot Fumante vs Encargos (mil).

Agora analisaremos por meio de gráficos de barras as seguintes variáveis, Faixa Etária (IDADE), Gênero, Fumante e IMC, em relação a variável resposta Encargos, buscando identificar alguma influência.

Primeiramente temos a Figura 5 que apresenta o gráfico de colunas para a Faixa Etária vs Encargos. Percebe-se que conforme o aumento da idade (faixa etária) maior se torna os Encargos dos beneficiários. Principalmente acima da idade de 42 anos, na qual já apresenta média de Encargos maiores que a média geral, 13.27 mil.

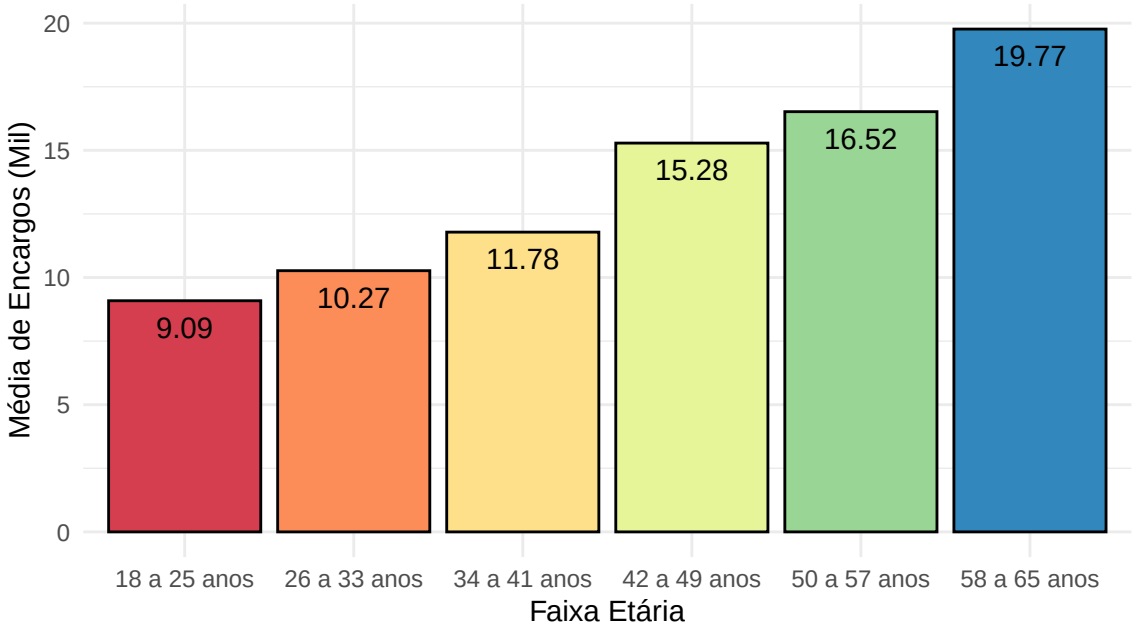

Figura 5: Média de Encargos (mil) por Faixa Etária.

As Figuras 6, 7, 8 e 9 apresentam gráficos de colunas para as variáveis Gênero, Fumante, IMC e Regiao, respectivamente, em relação aos Encargos médios (mil). Para o gênero não é observado diferença tão grande entre masculino e feminino, os dois apresentam médias relativamente próximas, mas considerando a unidade que os valores estão, sendo em mil, temos uma diferença de quase 1.5 mil, o que se torna considerável. Assim, o gênero masculino apresenta maior gasto com Encargos.

Já para os Fumantes é onde vemos a maior diferença, é muito discrepante a diferença entre fumar ou não em relação aos gastos com Encargos médicos.

Para o IMC é observado que os Encargos médicos é bem maior em situações que caracterizam algum nível de obesidade, valor de IMC acima do ideal. Uma observação relevante sobre as pessoas com IMC abaixo do ideal presente no banco em estudo é que poucos indíviduos foram registrados com essa situação, portanto temos poucas observações, dificultando a análise. Apesar disso, vemos que apresentam valores menores com Encargos médicos.

Pelas Regiões, a região sudeste apresentou os maiores gastos com Encargos médicos. As regiões noroeste e sudoeste, regiões mais voltadas ao oeste apresentaram menores gastos médicos, já as outras duas mais localizadas ao leste, apresentaram maiores gastos.

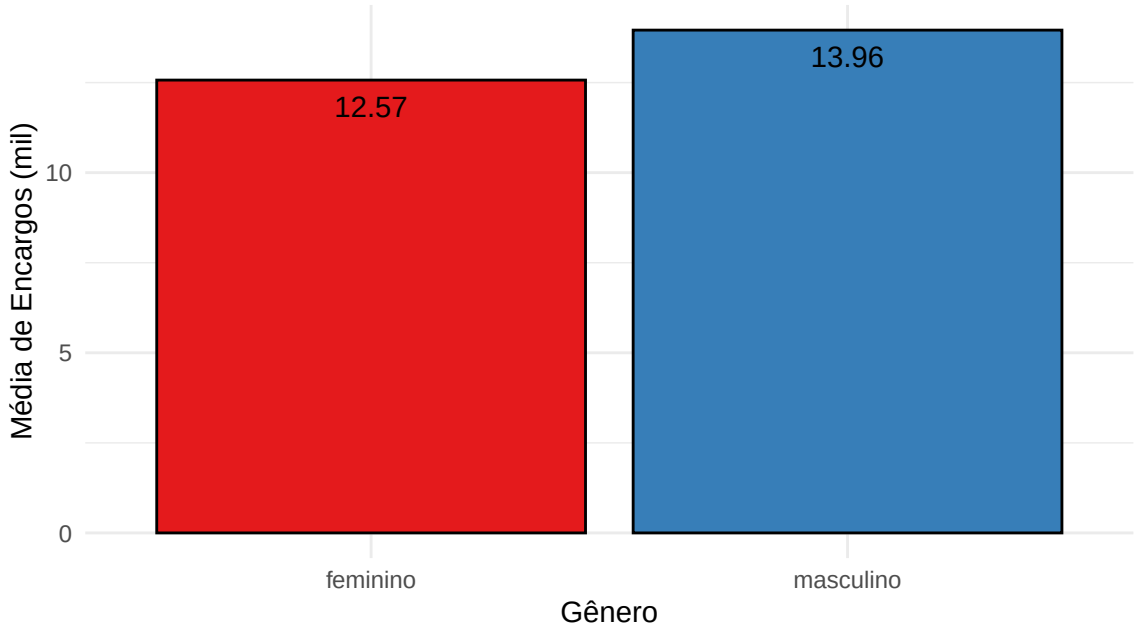

Figura 6: Média de Encargos (mil) por Gênero.

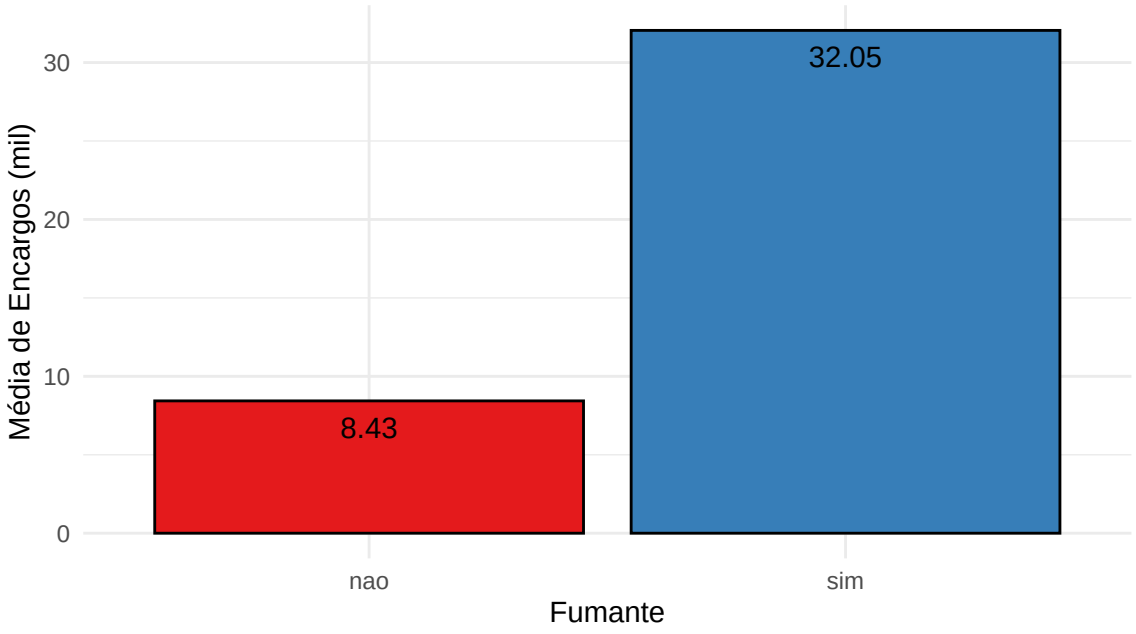

Figura 7: Média de Encargos (mil) pelo evento de ser Fumante ou não.

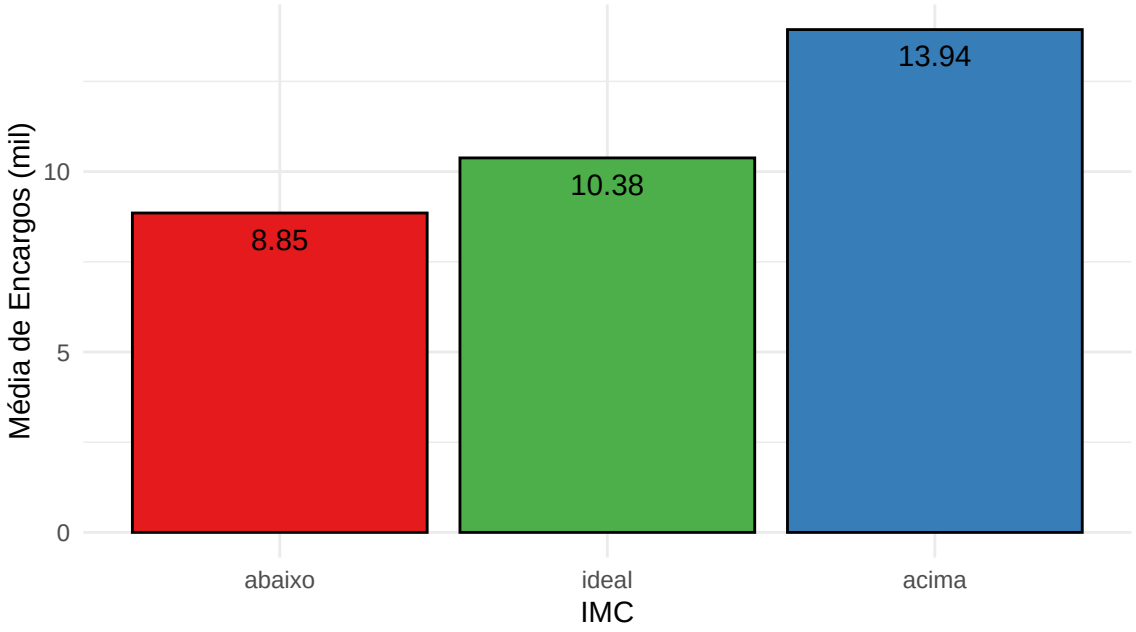

Figura 8: Média de Encargos (mil) pela categoria do IMC.

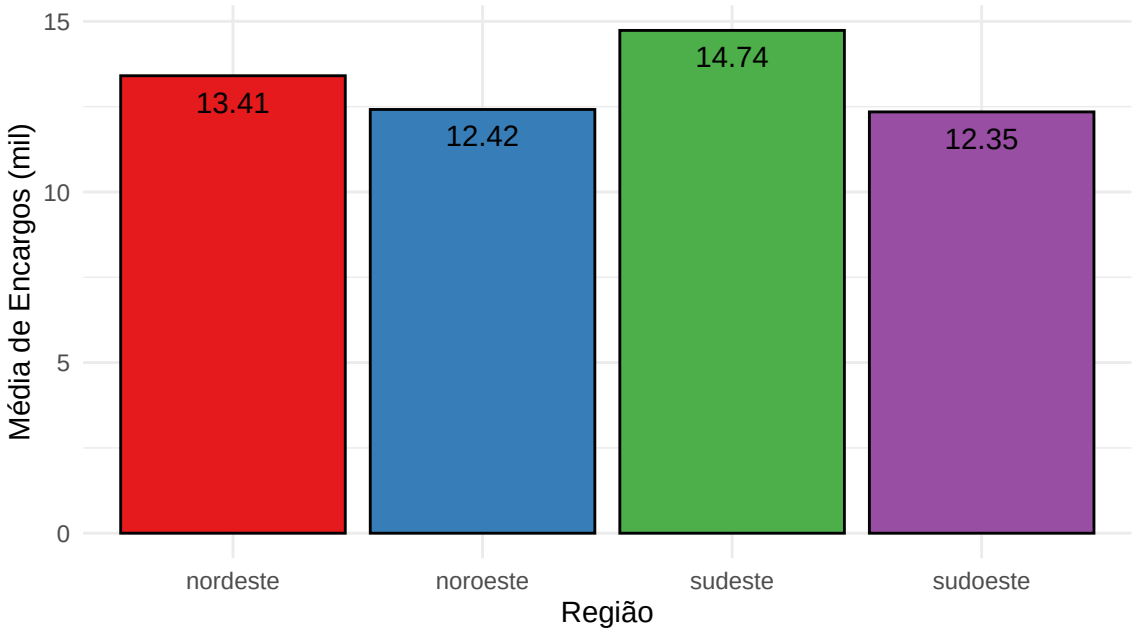

Figura 9: Média de Encargos (mil) pela Região.

3 Modelagem

Depois de toda análise exploratória dos dados realizada anteriormente , identificamos várias possíveis influências para a variável resposta Encargos. Assim, nesta seção ajustaremos um modelo de regressão Lasso, buscando minimizar o erro do modelo, assim como, não obter overfitting. Ou seja, o objetivo é o modelo conseguir identificar as características presente nos dados, para que em outros dados ele consiga ter um desempenho similar.

Para iniciar, vamos separar o banco de dados em duas partes, uma para treino e outra para teste. Essa divisão ficou aproximadamente 75% para treino, equivalente a 1003 observações, e 25% para teste, equivalente a 335 observações. Desta forma, conseguiremos trabalhar na base de treino, ajustando o modelo, e ao fim conseguir testar a eficiência do modelo ajustado sobre a base de treino. Abaixo está apresentado essa divisão, onde foi setado uma semente "set.seed(12345)" para que conseguissemos manter a mesma divisão e resultados em toda análise:

```
# Separando as bases de treino e teste -----------------------------------
set.seed(12345)
base_quebra <- rsample::initial_split(dados)
treino <- training(base_quebra)
teste <- testing(base_quebra)
```

Nesta etapa, iremos especificar o modelo que iremos trabalhar, definindo a regressão LASSO, que é identificada pelo valor 1 em "mixture", onde 0 corresponderia a regressão RIDGE. Logo, após é realizado a validação cruzada, que consiste na divisão que irá ser realizada na base de treino para assim fazer varias verificações em proporções diferentes da base de treino, uma por vez. Assim, para amostras de tamanho maiores que 1000, de forma mais usual é utilizado 5 ou 10 dobras, neste caso usaremos 5 dobras, o que significa que iremos quebrar em 5 partes iguais, assim sendo realizados 5 validações onde em cada vez 1/5 da base de treino será utilizada para teste. Desta forma, conseguimos minimizar o erro e não ter overfitting, já que conseguiremos de forma mais precisa, descrever o comportamento dos dados. Abaixo é mostrado essas especificações:

```r
# Especificando o modelo
modelo <- linear_reg(
  penalty = tune(),
  mixture = 1 # LASSO
)|>
  set_engine("glmnet")|>
  set_mode("regression")

# Cross-validation
set.seed(12345)
bases_Cross<- vfold_cv(treino, v = 5)
```

Com o modelo selecionado e realizada a separação do modelo em 5 dobras, é realizado a tunagem do modelo, identificando qual o valor do hiperparâmetro que melhor minimiza o erro sem ocorrer overfitting. Dessa forma, especificamos, o modelo selecionado, as variáveis selecionadas para o modelo, a base de validação cruzada, o tamanho do grid, que seria a quantidade de regularização, neste caso definimos em 300. Utilizando 2 métricas, rmse (Raiz do Erro Quadrático Médio), rsq (R^2). Podemos analisar pelos códigos executados abaixo e pela Figura 10 a tunagem do modelo, que apresentou R^2 aproximadamente de 74% e rmse de 6116.

```r
# Escolhendo o valor do hiperparâmetro
set.seed(12345)
tunagem <- tune_grid(
  modelo,
  Encargos ~ .,
  bases_Cross,
  grid=300,
  metrics= metric_set(rmse, rsq),
  control = control_grid(verbose = TRUE, allow_par = FALSE))
```

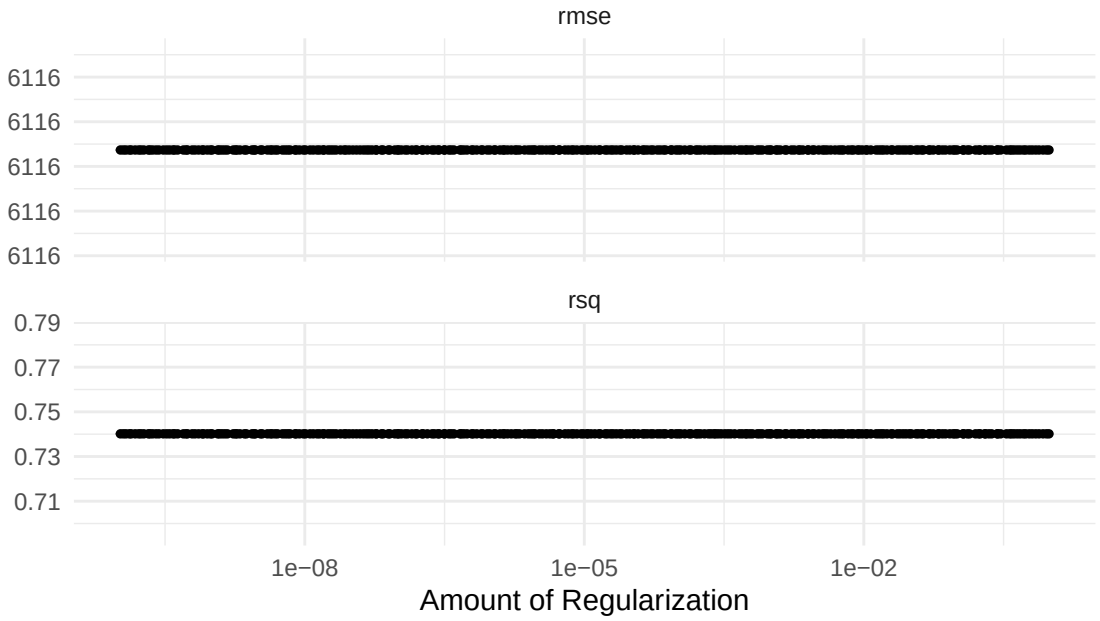

Figura 10: Tunagem do modelo.

Realizada a tunagem do modelo, iremos selecionar o valor do hiperparâmetro, como podemos ver anteriormente na Figura 10, conseguimos notar que independente do valor do hiperparâmetro o erro mantem-se igual, desta forma o menor valor do hiperparâmetro está apresentado na Tabela 3. Assim, o valor do hiperparâmetro selecionado foi de $1.05e^{-10}$, o qual é um valor muito próximo de 0, e como vimos o valor do hiperparâmetro não está interferindo no valor do erro, desta forma utilizaremos o valor de 0 como penalidade no modelo.

```
show_best(tunagem,"rmse", n=1)
```

Tabela 3: Resultado da tunagem com a escolha do melhor hiperparâmetro.

penalty	.metric	.estimator	mean	n	std_err
$1.05e^{-10}$	rmse	standard	6116	5	145

Abaixo é apresentado a nova especificação para o modelo, onde vamos introduzir o valor de penalidade para o modelo, o hiperparâmetro definido acima. E logo após, ajustando o novo modelo para a base de treino.

Com o modelo ajustado, podemos ver na Tabela 4 e na Figura 11 a importância das variáveis selecionadas pelo modelo em relação a variável resposta Encargos. Podemos notar que, a variável de maior influência no modelo é se o paciente é Fumante, aumentando em 23760 o valor de gasto com Encargos médicos quando fumante. Então, as covariáveis Fumante, filhos, IMC e Idade inteferem de forma positiva no modelo, ou seja, maior o valor sob a média de Encargos médicos. Já para as covariáveis, Região sudeste, sudoeste e noroeste, e também o genêro for masculino, influênciam negativamente no modelo, diminuindo o valor em gastos com Encargos médicos. Desta forma, o modelo final ajustado é apresentado abaixo:

$$Y = -11718 + 23760x_1 + 252x_2 - 350x_3 + 336x_4 + 338x_5 - 128x_6 - 804x_7 - 499x_8,$$

onde, Y = Encargos, x_1 = Fumante, x_2 = Idade, x_3 = Gênero Masculino, x_4 = IMC, x_5 = Filhos, x_6 = Região Noroeste, x_7 = Região Sudoeste, e x_8 = Região Sudoeste.

Tabela 4: Importância das variáveis do modelo ajustado em relação a variável resposta Encargos.

Variable	Importance	Sign
Fumantesim	23760	POS
Regiaosudeste	804	NEG
Regiaosudoeste	499	NEG
Filhos	388	POS
Generomasculino	350	NEG
IMC	336	POS
Idade	252	POS
Regiaonoroeste	128	NEG

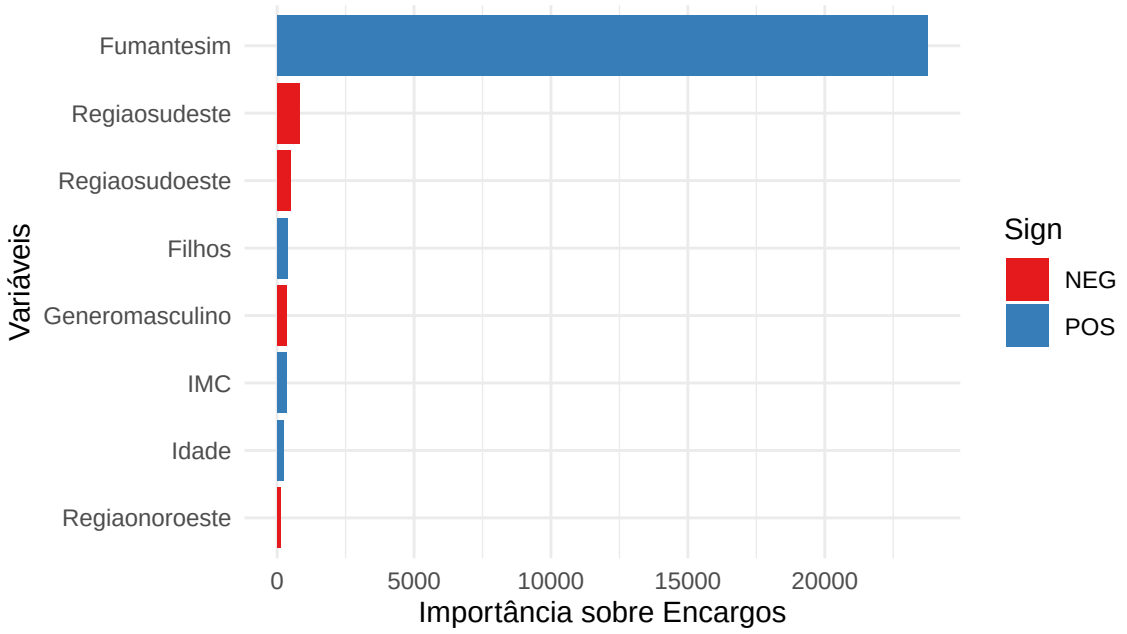

Figura 11: Importância das variáveis do modelo ajustado em relação a variável resposta Encargos.

Com o modelo ajustado e analisado, utilizamos o banco de treino para verificar a proximidade do modelo em relação ao banco de teste, fazendo uma previsão com os dados selecionados neste banco e conferindo com os reais. Desta forma, na Figura 12 apresentamos essa análise, onde conseguimos notar que o modelo se saiu muito bem em relação ao banco de teste, conseguiu seguir o comportamento dos dados, notamos apenas aqueles *outliers* onde não foi aproximado, mas levando em conta que o modelo deve seguir o comportamento e não tanto as irregularidades, *outliers*, o modelo se saiu bem, descrevendo com bastante proximidade a maioria dos dados e seguindo seu comportamento.

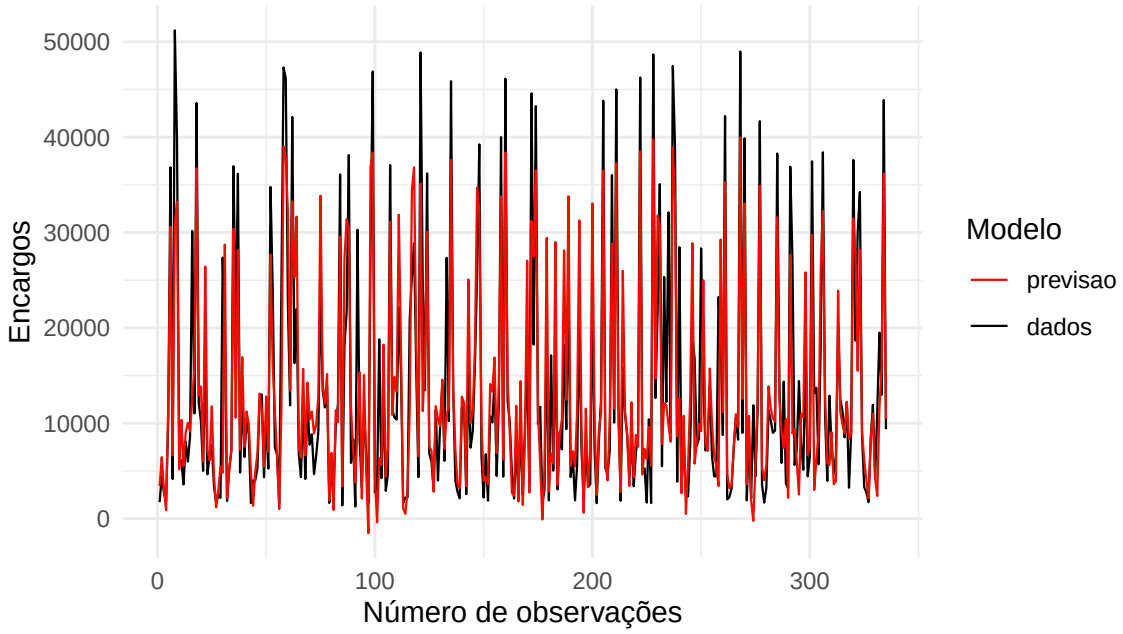

Figura 12: Previsão do modelo ajustado em relação ao banco de teste.

4 Conclusão

O presente trabalho teve a finalidade de apresentar os estudos desenvolvidos na disciplina de Machine Learning através da análise de um banco de dados com algumas informações pessoais para que pudéssemos fazer a previsão de gastos com Seguro Saúde nos Estados Unidos. Para isto, iniciamos com uma análise exploratória onde podemos ter um entendimento prévio porém básico dos dados em questão bem como das relações existentes entre as váriaveis, como por exemplo a relação entre a Idade e Encargos que indicam que beneficiários com mais de 42 anos de idade já pagam valores acima da média. Além disso, fumantes representam de forma muito discrepante os maires preços dos encargos do seguro, mostrando o quanto fumar é prejudicial a saúde.

Com o conhecimento prévio, adquirido na análise exploratória dos dados, partimos então para a estruturação do modelo, o qual foi ajustado através da regressão Lasso, com o objetivo de minimizar os erros de previsão sem que ocorra "overfitting" em nosso ajuste. A partir disto, podemos quantificar a influência de cada variável em relação aos encargos, sendo a variável "Fumante" a que mais interfere na média da nossa variável resposta. Por fim, foi realizado uma previsão do modelo utilizando a base de teste, para observar a proximidade com os valores reais, onde conclui-se que o modelo teve um bom comportamento.