Predizendo Prognóstico de Mortalidade com Dados Sintéticos

Mortality Prognostic Prediction with Sintetic Data

Apresentação

O presente projeto foi originado no contexto das atividades da disciplina de pós-graduação, Ciência e Visualização de Dados em Saúde, oferecida no primeiro semestre de 2022, na Unicamp, e foi desenvolvido por Mariângela Lima Rodrigues, RA 183863, aluna de mestrado em Estatística.

Contextualização da Proposta

O objetivo deste estudo foi predizer o prognóstico de um paciente morrer em um período de até 1 ano após ter dado entrada na emergência devido a um infarto no miocárdio ou uma parada cardíaca.

Ferramentas

- Software R Studio versão 4.1.2

Metodologia

Dado o objetivo de predizer a morte de um paciente que chega à emergência devido a um infarto ou uma parada cardíaca em um período de até um ano (365 dias) após seu encontro na emergência, observa-se que o modelo ajustado deverá ser capaz de classificar um paciente, dado um conjunto de fatores observados, de acordo com uma resposta binária, morre ou não em até um ano. Deste modo, o problema apresentado trata da proposição de um modelo de classificação com respeito ao prognóstico de morte de um paciente. E, consideradas estas informações, propõe-se a construção de um Modelo de Regressão Logística Binária.

A Regressão Logística trata de uma metodologia estatística empregada com propósitos como previsão de risco, classificação e determinação de características, por exemplo. No presente trabalho será empregada a vertente classificatória, na qual é ajustado um modelo de tal modo que, a partir da busca por identificação de padrões de associação entre as observações, seja possível classificá-las em suas respectivas categorias.

Modelo de Regressão Logística Binária

A Regressão Logística é um modelo que permite a estimação da probabilidade de ocorrência de um evento a partir de observações aleatórias e, a partir desta informação, é possível classificar observações em suas respectivas categorias de acordo com a probabilidade de pertencer a cada uma destas.

No cenário de Regressão Logística Binária a variável resposta, geralmente, é identificada como

$$Y = \begin{cases} 1, & \textit{se ocorre o evento de interesse}, \\ 0, & \textit{se não ocorre o evento de interesse} \end{cases} \tag{1}$$

onde o evento de interesse é determinado de acordo com os objetivos do estudo.

Dito isto, o modelo de classificação é dado por

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}, \tag{2}$$

onde $g(x) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ trata de uma função linear das p variáveis independentes a serem utilizadas no ajuste da variável resposta.

Para que o ajuste do modelo seja realizado é necessária a estimação dos parâmetros da função linear $g(x)$, os coeficientes β_i, $i = 1, 2, ..., p$. O processo de estimação ocorre a partir do Método de Máxima Verossimilhança, que busca pelos parâmetros estimados $\hat{\beta}_i$ tais que tenham maior probabilidade de se comportarem como os parâmetros reais observados na população.

Mais detalhes com respeito ao modelo e todo o processo de construção e análise do mesmo podem ser observadas em Gonzalez (2018).

Além disso, neste estudo utilizou-se a técnica de seleção de variáveis, Stepwise (Seber e Lee, 2012), técnica esta que permite, no contexto de análise de regressão logística, avaliar quais variáveis são estatisticamente significativas para a classificação de uma observação e, com isto, otimiza o modelo ajustado de acordo com as covariáveis disponíveis para a modelagem.

Bases adotadas para o Estudo

- scenario01
- scenario02

Resultados Obtidos

Foram construídos dois modelos de regressão logística que serão apresentados a seguir. Para a construção destes modelos foram utilizadas as seguintes variáveis descritivas:

- Doença (DESCRIPTION - CODE): infarto ou parada cardíaca;
- Gênero do paciente (GENDER): masculino ou feminino;
- Etnia do paciente (ETHNICITY): hispânico ou não hispânico;
- Alergias (ALLERGIES): indica se o paciente possui algum tipo de alergia ou não;
- Número de encontros (N_ENCOUNTERS): quantas vezes o paciente foi até a emergência;
- Valor total de despesas com saúde ao longo da vida (HEALTHCARE_EXPENSES) e
- Valor total de despesas cobertas pelo seguro de saúde ao longo da vida (HEALTHCARE_COVERAGE).

Como mencionado na seção de metodologia, o modelo ajustado foi um modelo de regressão logística. O primeiro modelo foi treinado e validado para os dados do scenario01 e testado nos dados do scenario02, este caso será referido como caso 1. Já o segundo modelo, foi treinado e validado no scenario02 e testado nos dados do scenario01, caso 2.

Para obter os modelos que serão apresentados a seguir, foram separados os conjuntos de dados para treinamento (70%) e validação (30%) de maneira aleatória, conforme apresentado no código abaixo. Foram fixadas as sementes, 001 e 002, a fim de que a partir dos dados seja possível obter a mesma amostra utilizada para treino/validação do modelo.

```r
# CASO 1: Modelo treinado no scenario01 e testado no scenario02
set.seed(001)

# Amostra de treino e validação no scenario01
n_scenario01 <- round(nrow(base_scenario01)*0.7)
sample_train_scenario01 <- sample(nrow(base_scenario01),
                                  size = n_scenario01, replace = FALSE)

train_scenario01 <- base_scenario01[sample_train_scenario01,] %>%
  select(DESCRIPTION, ETHNICITY, GENDER, HEALTHCARE_EXPENSES,
         HEALTHCARE_COVERAGE, ALLERGIES, N_ENCOUNTERS, DEATH)

test_scenario01 <- base_scenario01[-sample_train_scenario01,] %>%
  select(DESCRIPTION, ETHNICITY, GENDER, HEALTHCARE_EXPENSES,
         HEALTHCARE_COVERAGE, ALLERGIES, N_ENCOUNTERS, DEATH)

# CASO 2: Modelo treinado no scenario02 e testado no scenario01
set.seed(002)

# Amostra de treino e validação no scenario01
n_scenario02 <- round(nrow(base_scenario02)*0.7)
sample_train_scenario02 <- sample(nrow(base_scenario02),
                                  size = n_scenario02, replace = FALSE)

train_scenario02 <- base_scenario02[sample_train_scenario02,] %>%
  select(DESCRIPTION, ETHNICITY, GENDER, HEALTHCARE_EXPENSES,
         HEALTHCARE_COVERAGE, N_ENCOUNTERS, DEATH)

test_scenario02 <- base_scenario02[-sample_train_scenario02,] %>%
  select(DESCRIPTION, ETHNICITY, GENDER, HEALTHCARE_EXPENSES,
         HEALTHCARE_COVERAGE, N_ENCOUNTERS, DEATH)
```

Dadas as amostras conforme apresentado, foi ajustado um modelo de regressão logística em cada um dos cenários, com o auxílio da função `glm()`, e em seguida a este ajuste inicial, contendo todas as covariáveis supracitadas, foi aplicada a técnica de seleção de variáveis setpwise e obteve-se o modelo que melhor ajusta os dados e possui melhores características preditivas. A Figura 1 apresenta as curvas ROC, da performance do modelo no conjunto de dados de validação do modelo, para os casos 1 e 2, respectivamente.

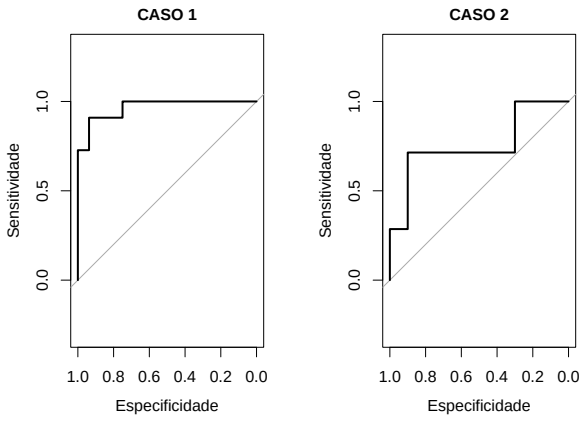

Figura 1: Curva ROC (Receiver Operating Characteristic) para os modelos dos casos 1 e 2 aplicados aos dados de validação.

Respectivamente, os modelos no contexto de validação apresentaram AUC de 0.9659 e 0.7571, valores bastante satisfatórios. Logo em seguida, dado que os modelos propostos apresentaram boa performance, os mesmos foram testados nos conjuntos de dados de interesse, para o caso 1, conforme aponta a Figura 2, a performance do modelo nos dados de teste foi também satisfatória, com AUC igual a 0.8528.

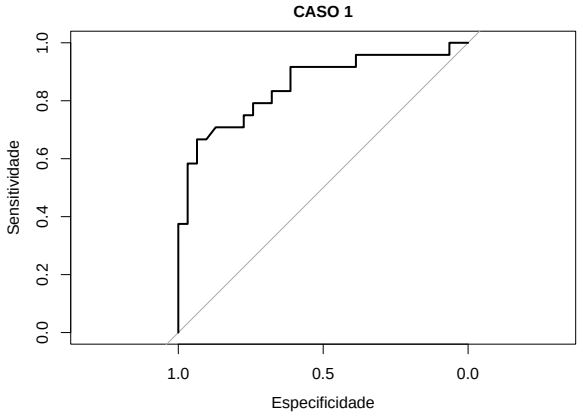

Figura 2: Curva ROC (Receiver Operating Characteristic) para o modelo do caso 1 aplicado aos dados do scenario02.

Para além disso, quando observada a acurácia do modelo do caso 1, verificou-se, conforme a matriz de confusão apresentada abaixo (Tabela 1), acurácia igual à 0.8, com intervalo de confiança de 95% dado por (0.6703, 0.8957), que reafirma a qualidade da predição. E ainda mais, quando analisados os resultados do modelo também podemos observar a sensibilidade do modelo, definida pela razão entre o número de predições verdadeiras positivas sobre a quantidade total de observações positivas, que matematicamente equivale a

$$Sensibilidade = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo\ +\ Falso\ Negativo},$$

uma métrica que permite mensurar o quanto o modelo detecta corretamente os resultados classificados como os eventos de interesse (neste caso, predizer corretamente a morte do paciente), observou-se sensibilidade de 67%. Apesar de não ser tão alta, a sensibilidade ainda aponta para um ajuste razoável dos dados, indicando a qualidade do modelo de predição proposto.

Tabela 1: Matriz de Confusão - Modelo do caso 1 aplicado aos dados do scenario02

		Real	
		Óbito	Não óbito
Predito	Óbito	16	8
	Não óbito	3	16

Já no caso 2, conforme aponta a Tabela 2 e a Figura 3 juntamente com o valor do AUC (0.8716) do modelo testado no scenario02, o segundo modelo também apresentou boa performance. Entretanto, é necessário ressaltar que neste segundo cenário foi preciso remover a covariável que indica a presença de alergias, uma vez que no scenario02 nenhum paciente da amostra possuía algum tipo de alergia.

Tabela 2: Matriz de Confusão - Modelo do caso 2 aplicado aos dados do scenario01

		Real	
		Óbito	Não óbito
Predito	Óbito	23	5
	Não óbito	6	57

Exceto este detalhe, considerando também as métricas do modelo, observou-se acurácia de 0.8791, com intervalo de confiança de 95% igual a (0.7940, 0.9381), e além disso, sensibilidade de 82%. Um desempenho melhor do que o modelo apresentado para o caso 1.

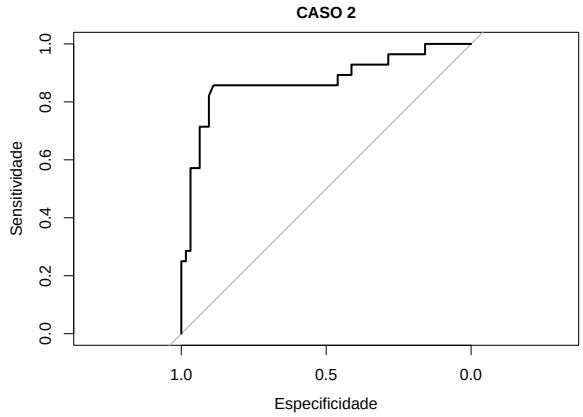

Figura 3: Curva ROC (Receiver Operating Characteristic) para o modelo do caso 2 aplicado aos dados do scenario01.

Discussão

Conforme apresentado, os modelos, tanto no caso 1 quanto no caso 2, apresentaram performance razoáveis no cenário de predição do prognóstico dos pacientes dos dados do scenario01 e scenario02. Apesar disto, observou-se que para o caso 2 não foi possível considerar a variável com respeito a alergias, uma vez que todos os pacientes da amostra não possuíam nenhum tipo de alergia e, portanto, o modelo não seria capaz de utilizar a informação.

Conclusão

Com o intuito de predizer o prognóstico de morte, em até um ano, de um paciente que chega a emergência devido a um infarto ou uma parada cardíaca, foram construídos modelos de regressão logística em diferentes cenários, usando diferentes conjuntos de dados. Dito isto, treinados e validados os modelos, observou-se um bom desempenho de ambas as propostas, entretanto, ressalta-se que o número de observações para o ajuste do modelo foi razoavelmente pequeno, menos de 100 observações, e este fato pode comprometer os resultados obtidos.

Como considerações finais, ficam as seguintes ressalvas a respeito do trabalho desenvolvido:

1. O ajuste poderia ser melhorado caso aumentássemos o tamanho amostral e
2. A amostra não possui uma distribuição equilibrada com respeito a todas as covariáveis, o que implicou na necessidade de remoção de uma delas para que o segundo modelo pudesse ser ajustado.

Com respeito às dificuldades enfrentadas, ressalto mais uma vez a dificuldade com a linguagem de programação e acredito que se tivesse mais tempo poderia ter me aprofundado mais nas técnicas de machine learning e assim teria produzido um trabalho melhor estruturado e com um nível de absorção de conteúdo muito maior do que o que de fato consegui absorver. Além disso, também senti muita dificuldade para conseguir estruturar o projeto no formato solicitado, pois confesso que não tenho familiaridade com o Github.

Referências Bibliográficas

GONZALEZ, Leandro de Azevedo. Regressão Logística e suas Aplicações, 2018.

SEBER, George AF; LEE, Alan J. Análise de regressão linear . John Wiley & Filhos, 2012.