Beatriz Moreira, 54514 Grupo 1 Rute Patuleia, 51780

REGRESSÃO E ANÁLISE DE VARIÂNCIA - 2024

FCUL DEIO

Regressão e Análise de Variância - Group 1

Problema 1

Beatriz Moreira, 54514 | Rute Patuleia, 51780

1. Resumo

A Análise de Variância (ANOVA) é uma ferramenta estatística crucial para comparar a variância entre diferentes grupos de dados. O presente estudo recorre à ANOVA e contrastes ortogonais para investigar a influência da qualidade da cozinha e da presença de ar condicionado central nos preços de venda de casas em Ames, Iowa. Os resultados indicam que a qualidade da cozinha tem um impacto significativo nos preços das casas, enquanto a presença de ar condicionado central não demonstra evidências significativas. Os contrastes ortogonais confirmam esses resultados, evidenciando as diferenças nas médias de preços entre os grupos de cada variável. Destaca-se, assim, a importância da qualidade da cozinha na determinação dos preços das casas, ao contrário da presença de ar condicionado central que pode não ter um efeito estatisticamente significativo. Desta forma, estudos futuros devem aumentar o tamanho da amostra e explorar outros fatores que possam influenciar os preços das casas em Ames, Iowa.

Keywords: Análise de Variância, Contrastes Ortogonais

2. Introdução

A Análise de Variância, ou ANOVA, compara a variância entre os diferentes grupos presentes nos dados com a variância dentro de cada grupo através da razão:

$$F_{ratio} = \frac{MS_{treat}}{MS_E} \textbf{ (1)}$$

Onde MS_{treat} e MS_E são as estimativas das variâncias entre grupos e em cada grupo, respetivamente. [1]

Se a variância entre os grupos for significativamente superior à variância dentro de cada grupo, podemos concluir que os grupos são diferentes; caso contrário, não podemos tirar conclusões. Depois de termos obtido o F_{ratio}, este valor é comparado a um valor correspondente a um determinado nível de confiança da distribuição de probabilidade de Fisher. [2]

Assim, o objetivo principal deste trabalho é utilizar a ANOVA para avaliar se alguma das duas variáveis explicativas e/ou a interação entre elas, é estatisticamente significativa na explicação da variação da variável em estudo.

3. Metodologia

3.1. Descrição dos Dados

O presente projeto foi desenvolvido utilizando um conjunto de dados referentes ao preço de casas em Ames, Iowa, nos Estados Unidos da América. Originalmente, estes dados continham 163 colunas com informações sobre a qualidade das casas em estudo. [3]

Esta análise é deveras importante no mercado imobiliário, pois permite que proprietários e profissionais do setor compreendam como certos atributos afetam o valor dos imóveis. O presente estudo focou-se apenas em duas variáveis e na sua relação com o preço final das casas (*SalePrice*): *CentralAir*, que indica se a casa tem ar condicionado ou não e *KitchenQual*, que avalia a qualidade da cozinha da residência em 5 níveis: excelente, boa, típica/mediana, razoável e pobre.

3.2. Pré-processamento

De modo a obter os dados necessários para a realização deste projeto foi implementado um pré-processamento dos mesmos.

Inicialmente, selecionou-se apenas as três variáveis de interesse (*SalePrice*, *CentralAir* e *KitchenQual*). Em seguida, dividiu-se os valores do preço de venda das casas por 10 mil, transformando assim os dados em unidades de dezenas de milhares de dólares (Figura 1).

Apesar da variável qualidade da cozinha (*KitchenQual*) originalmente apresentar 5 níveis, nomeadamente Excellent (Ex), Good (Gd), Typical/Average (TA), Fair (Fa) e Poor (Po), apenas foram considerados 3 níveis, “Gd”, “Fa” e “TA”. Para isso, converteu-se o nível “Ex” em “Gd”, conforme o código ilustrado na Figura 1. Além disso, uma vez que não existe a categoria “Po” nos dados, foi possível incluir os 3 níveis da variável *KitchenQual*.

De seguida, prosseguiu-se à criação do novo *dataset*, no qual selecionou-se aleatoriamente 6 observações por grupo, garantindo, assim, uma análise balanceada. O código encontra-se na Figura 2.

3. Resultados e Discussão

3. 1. Representações Gráficas

As representações gráficas revelam *insights* sobre a importância de cada fator e a sua interação.

Primeiramente, evidencia-se a influência da qualidade da cozinha na determinação dos preços de venda das casas. O boxplot ilustrado na Figura 3 destaca essa relação, indicando que casas com cozinha de qualidade "Gd" tendem a ter preços de venda mais elevados em comparação com aquelas com qualidade de cozinha "Fa" ou "TA".

De seguida, analisou-se o impacto/efeito da presença de ar condicionado central nos preços de venda das casas. O boxplot ilustrado na Figura 4, demonstra que, na verdade, casas sem ar condicionado tendem a ter preços de venda mais altos em comparação às casas com essa comodidade. No entanto, as diferenças significativas não parecem ser significativas.

Adicionalmente, pretendeu-se mostrar informações úteis sobre como os preços de venda variam de acordo com a qualidade da cozinha e a presença de ar condicionado central. O boxplot do preço de venda de casas em função da presença de ar condicionado central e da qualidade da cozinha, ilustrado na Figura 5, revela que casas com boa qualidade de cozinha (Gd) e ar condicionado central apresentam os preços de venda mais elevados, enquanto as casas com uma qualidade de cozinha mais baixa (Fa) e ausência de ar condicionado central tendem a ter os preços mais baixos.

Por fim, explorou-se a interação (*interaction effect*) entre as variáveis *KitchenQual* e *CentralAir* nos preços de venda das casas (*SalePrice*), através de um *interaction plot* como ilustrado na Figura 6. Observa-se uma clara interação entre os dois fatores, uma vez que as linhas se cruzam.

3. 2. Teste Global

Para averiguar se existiam diferenças significativas entre as médias dos grupos obtidos na amostra, foi efetuado um teste global, com duas hipóteses:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ **vs** $H_1: \mu_i \neq \mu_j$ para pelo menos um par (i,j).

Os resultados obtidos estão presentes na Tabela 1.

Tabela 1. Tabela ANOVA para o teste global.

Fontes de Variação	SS	DF	MS (Variância)	F-ratio	P-value
Entre grupos (Trat)	648,21	5	129,64	5,72	<.001***
Erro (E)	679,56	30	22,65		
Total	1327,77	35			

*** Estatisticamente significante a p < .001.

Para nível de significância de 5%, $p_{value} \leq \alpha$, logo rejeitamos H_0. Assim, podemos concluir que existem diferenças entre, pelo menos, dois grupos dos dados.

3. 3. Testes Simples

Com base nos resultados dos testes de significância, podemos inferir o impacto/influência das variáveis no preço final das casas.

Tabela 2. Tabela ANOVA para os testes simples.

	SS	DF	MS	F-ratio	P-value
KitchenQual	551,83	2	275,92	12,18	<.001***
CentralAir	84,78	1	84,78	3,74	0,063
Interação	11,59	2	5,79	0,26	0,776
Erro (E)	679,56	30	22,65		

Os *p-values* da Tabela 2 indicam que a variável *KitchenQual* demonstrou um efeito estatisticamente significativo no preço final das casas, com um p-value muito baixo, indicando forte evidência estatística. Por outro lado, a variável *CentralAir* não apresentou tal evidência, o que sugere que a presença de ar condicionado central pode não ter um impacto estatisticamente significativo no preço das casas.

Em relação à interação entre *KitchenQual* e *CentralAir*, o *p-value* associado a este teste foi consideravelmente alto, indicando falta de suporte estatístico para rejeitar a hipótese nula (não há interação entre as duas variáveis), ou seja, não há evidência estatística para suportar a presença de uma interação significativa entre essas duas variáveis. Tal significa que o efeito combinado da qualidade da cozinha e da presença de ar condicionado central no preço das casas não é estatisticamente diferente do esperado pela simples soma dos efeitos individuais de cada variável.

3.4. Contrastes Ortogonais

3.4.1. Desenho de Contrastes Ortogonais

Contrastes são basicamente comparações entre médias de grupos ou níveis de fatores num estudo. No presente estudo, pretende-se comparar os preços das casas com base na qualidade da cozinha, na presença de ar condicionado central e da sua interação.

Relativamente à *KitchenQual* (Qualidade da Cozinha) procedeu-se à comparação entre a média de "Fair" e a média de "Typical/Average" e "Good", e de seguida, comparou-se a média de "Typical/Average" com a média de "Good".

Para a variável *CentralAir* (Presença de Ar Condicionado Central), a comparação foi feita entre a média da existência ("Yes") ou não ("No") de ar condicionado.

Para realizar os contrastes relativamente à interação dos dois fatores, primeiro foi feita a comparação entre a média dos grupos "Fair" e "Typical/Average" com o grupo "Good" em casas com ar condicionado central contra a média do grupo "Good" com "Fair" e "Typical/Average" em casas sem ar condicionado central. Depois, foi realizado o contraste entre as médias de "Fair" e "Typical/Average" com e sem ar condicionado central.

Todos os contrastes ortogonais utilizados e os resultados associados a estes estão presentes na Tabela 3.

Tabela 3. Tabela de Contrastes Ortogonais para *CentralAir* e *KitchenQual*.

		Contrastes Ortogonais							
		Central Air		**Estatísticas**					
	Kitchen Quality	**Yes**	**No**	**teta(j)^**	**Aux**	**SS(j)**	**f(j)_obs**	**p-value**	**Conclusão**
Contraste 1 (Kitchen Quality)	**Fair**	0,5	0,5	-5,57	0,13	248,36	10,96	<.001	Rejeita-se H0
	TA	-0,25	-0,25						
	Good	-0,25	-0,25						
Contraste 2 (Kitchen Quality) TA vs Good	**Fair**	0	0	-7,11	0,17	303,48	13,40	<.001	Rejeita-se H0
	TA	0,5	0,5						
	Good	-0,5	-0,5						
Contraste 3 (Central Air) Yes vs No	**Fair**	0,33	-0,33	3,07	0,11	84,78	3,74	0,06	Não se rejeita H0
	TA	0,33	-0,33						
	Good	0,33	-0,33						
Contraste 4 (Interação) Fair and TA vs Good (Yes) vs Good vs Fair and TA (No)	**Fair**	0,25	-0,25	1,19	0,13	11,28	0,49	0,61	Não se rejeita H0
	TA	0,25	-0,25						
	Good	-0,5	0,5						
Contraste 5 (Interação) Fair vs TA (Yes) vs Ta vs Fair (No)	**Fair**	0,25	-0,25	-0,11	0,04	0,32	0,01	0,99	Não se rejeita H0
	TA	-0,25	0,25						
	Good	0	0						

3.4.2. Comparação com F-value do Teste Global

Para contextualizar os resultados dos contrastes ortogonais, é importante compará-los com o F-value obtido no teste global realizado anteriormente. O teste global foi conduzido para investigar se existiam diferenças significativas entre as médias dos grupos obtidos na amostra, considerando todas as variáveis em conjunto.

Os resultados do teste global, apresentados na Tabela 1, revelaram uma estatística F de 5,72, com um p-value muito baixo, indicando uma forte evidência estatística. Isso levou à rejeição da hipótese nula, permitindo-nos concluir que há diferenças significativas entre pelo menos dois grupos dos dados.

Observou-se que a média dos valores F observados de cada contraste é 5,72, sendo este valor igual ao F observado do teste global, sugerindo consistência nos resultados entre os contrastes individuais e o teste global da ANOVA. Ou seja, os contrastes ortogonais capturaram efetivamente as diferenças significativas entre as médias dos grupos, que também são refletidas no teste global. Essa consistência reforça a confiança e fornece uma validação adicional nas conclusões tiradas a partir dos contrastes ortogonais e do teste global.

3.4.3. Comparação do SS parcial e SS_{Trat}

O valor da soma dos quadrados parciais (648,21) é igual à SS_{Trat}. Isso significa que toda a variabilidade dos dados está explicada pelos contrastes ortogonais escolhidos, o que era esperado, uma vez que o número de contrastes ortogonais seleccionados para cada uma das variáveis foram determinados pelos graus de liberdade correspondentes: dois para *KitchenQual,* um para *CentralAir*, e dois para a interação entre os fatores.

Essa equivalência entre os valores de SS Parcial e SS_{Trat} reforça a validade dos contrastes ortogonais e sugere que os contrastes ortogonais escolhidos foram adequados para capturar as diferenças significativas entre as médias dos grupos.

3.4.4. Resultados dos Contrastes

Os contrastes delineados são os seguintes:

- ***KitchenQual*: Fair vs TA e Good:** compara a média de preços das casas entre as categorias "Fair" e "Typical/Average" e a categoria "Good".

  Resultado: Rejeita-se a hipótese nula, logo há diferença significativa nas médias de preços entre as casas "Fair" e "Typical/Average" e "Good".
- ***KitchenQual*: TA vs Good:** compara a média de preços das casas entre as categorias "Typical/Average" e "Good".

  Resultado: Rejeita-se a hipótese nula, logo há diferença significativa nas médias de preços entre as casas "Typical/Average" e "Good".
- ***CentralAir*: Yes vs No:** compara a média de preços das casas entre casas com ar condicionado central ("Yes") e casas sem ar condicionado central ("No").

  Resultado: Não se rejeita a hipótese nula, logo não há diferença significativa nas médias de preços entre as casas com e sem ar condicionado.
- **Interação: (Fair e TA vs Good (Yes)) vs (Good vs Fair e TA (No)):** compara a média de preços das casas entre as categorias "Fair" e "Typical/Average" e "Good" em casas com ar condicionado e "Good" contra "Fair" e "Typical/Average", em casas sem ar condicionado central.

  Resultado: Não se rejeita a hipótese nula, logo não há diferença significativa nas médias de preços entre os dois grupos de casas.
- **Interação: (Fair vs TA (Yes)) vs (TA vs Fair (No)):** compara a média de preços das casas entre as categorias "Fair" e "Typical/Average" com ar condicionado central e "Fair" e "Typical/Average" sem ar condicionado.

  Resultado: Não se rejeita a hipótese nula, logo não há diferença significativa nas médias de preços entre os dois grupos de casas.

É importante salientar que os resultados dos contrastes ortogonais estão em consonância com os resultados obtidos nos testes simples, uma vez que houve um *main effect* do *KitchenQual* e realmente, ao fazer os contrastes ortogonais desta variável, existem entre os grupos (Fair vs TA e Good) e (TA vs Good). Assim, é possível perceber que a variável *KitchenQual* têm um efeito significativo na variável *SalePrice* e que ao fazer os contrastes entre cada combinação, observam-se diferenças significativas em cada contraste. Adicionalmente, nos testes simples não se obteve um *main effect* da *CentralAir* e ao fazer os contrastes ortogonais, como é de esperar, também não há diferenças significativas entre os grupos. E, por fim, nos testes simples não se obteve um *interaction effect* das variáveis *KitchenQual* e *CentralAir*, sendo que os contrastes ortogonais desta interação também não revelaram diferenças estatisticamente significativas.

3.5. Intervalo de Confiança (1 - α) x 100%

O cálculo do intervalo de confiança de 95% ($\alpha = 0.05$) para o preço médio de venda das casas com boa qualidade de cozinha e a presença de ar condicionado central foi feito utilizando apenas os três grupos com casas de diferentes qualidades de cozinha ("Fa", "Gd" e "TA") e com presença de ar condicionado. Os valores necessários para a obtenção do intervalo de confiança estão representados na Tabela 4.

Tabela 4. Valores utilizados para o cálculo do intervalo de confiança do grupo com boa qualidade de cozinha e a presença de ar condicionado central.

$\hat{\theta}$	**MSE**	$\sum_{i=1}^{3} \frac{c_i^2}{n_i}$	**SE($\hat{\theta}$)**	α	**qt**
19,48	22,65	0,17	1,94	5,00%	2,04

Assim, o intervalo de confiança de 95% obtido para este grupo foi de $[15{,}51; 23{,}44]$.

3.6. Conclusão

O estudo realizado visou investigar a influência da presença de ar condicionado central e da qualidade da cozinha no preço de venda das casas em Ames, Iowa, utilizando técnicas de Análise de Variância (ANOVA) e contrastes ortogonais.

Com base nos testes de significância efetuados, a qualidade da cozinha tem um impacto significativo no preço de venda das casas, enquanto a presença de ar condicionado central não. Embora tenha sido observada uma tendência em que casas com boa qualidade de cozinha e ar condicionado central apresentarem preços mais altos, a análise estatística não apontou uma evidência significativa para essa interação. Adicionalmente, os contrastes ortogonais realizados refletiram os resultados obtidos nos testes simples realizados.

Contudo, é importante ressaltar que este estudo se baseia numa amostra muito pequena e pode não capturar todas as nuances que condicionam o preço de casas em Ames, Iowa. Futuras pesquisas podem beneficiar de amostras maiores e de uma análise mais detalhada de outros fatores que influenciam os preços das casas.

3.7. Referências

[1] https://www.investopedia.com/terms/a/anova.asp
[2] https://statsandr.com/blog/anova-in-r/#aim-and-hypotheses-of-anova
[3] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Regressão e Análise de Variância - Group 1

Problema 2

Beatriz Moreira, 54514 | Rute Patuleia, 51780

1. Resumo

O presente estudo focou-se em técnicas de regressão linear para analisar a relação entre os atributos das casas em Ames, Iowa, e os seus preços de venda e tentar prever esses mesmos valores. Inicialmente, foram selecionadas 61 variáveis de interesse e realizadas análises para verificar a multicolinearidade e a influência dessas variáveis nos preços de venda. Os resultados indicaram que algumas variáveis não contribuem significativamente para a explicação dos preços das casas, levando à exclusão das mesmas do modelo. Em seguida, foram utilizadas técnicas para encontrar um modelo parcimonioso, eficiente na explicação da variabilidade nos dados. O modelo final foi testado em dados não utilizados anteriormente, dados de teste, mostrando uma capacidade razoável de explicar a variabilidade nos preços das casas. Os intervalos de previsão e o erro médio absoluto indicaram uma precisão razoável nas previsões do modelo. Em conclusão, este estudo forneceu *insights* importantes sobre os fatores que influenciam os preços das casas em Ames, Iowa, demonstrando a utilidade da regressão linear no mercado imobiliário.

Keywords: Regressão Linear, Previsão

2. Introdução

A regressão linear é um modelo estatístico que estima a relação linear entre a variável dependente e uma ou mais variáveis independentes ajustando uma linha reta que minimiza as discrepâncias entre os valores observados e reais.[1][2] A equação do modelo fornece coeficientes que elucidam o impacto de cada variável independente na variável dependente. A análise deste modelo é, assim, usada para prever o valor de uma variável com base no valor de outra variável.[1]

Existem dois tipos de regressão linear, nomeadamente, a simples e a múltipla, no qual a primeira envolve apenas uma variável independente e a segunda mais do que uma variável independente.[1] A equação para a regressão linear simples e múltipla são, respetivamente:

$$Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X + \cdots + \beta_n X$$

Assim, o objetivo principal deste modelo é encontrar a melhor equação da fit line que possa prever os valores com base nas variáveis independentes.

Existem várias suposições para ambos os tipos de modelos, no caso da regressão linear múltipla, que é o foco deste trabalho, deve-se aplicar as quatro suposições da regressão linear simples, nomeadamente, a linearidade, independência, homocedasticidade e normalidade e também verificar se não há multicolinearidade, aditividade, seleção das variáveis independentes e *overfitting*.[1]

Relativamente à multicolinearidade, este é um fenómeno estatístico que ocorre quando duas ou mais variáveis independentes estão altamente correlacionadas, dificultando a avaliação dos efeitos individuais de cada variável sobre a variável dependente. É possível detetar a multicolinearidade através de duas técnicas, nomeadamente uma matriz de correlação, no qual correlação altas (próximas de 1 ou -1) indicam potencial multicolinearidade e VIF (fator de inflação da variância), sendo esta uma medida que quantifica quanto a variância de um coeficiente de regressão estimado aumenta se os seus preditores estiverem correlacionados. Um VIF alto (normalmente acima de 10) sugere multicolinearidade.[1]

Por fim, é necessário aplicar medidas de avaliação para determinar a força do modelo de regressão linear. Estas métricas de avaliação dão frequentemente uma indicação de quão bem o modelo está a produzir os resultados observados. As medidas mais comuns são o erro quadrático médio (MSE), erro médio absoluto (MAE), raiz do erro quadrático médio (RMSE), o coeficiente de determinação (R^2) e o erro do R^2 ajustado.[1]

3. Metodologia

3.1. Descrição dos Dados

O presente projeto foi desenvolvido utilizando um conjunto de dados referentes ao preço de casas em Ames, Iowa, nos Estados Unidos da América. Originalmente, estes dados continham 163 colunas com informações sobre a qualidade das casas em estudo.[3]

Esta análise é deveras importante no mercado imobiliário, pois permite que proprietários e profissionais do setor compreendam como certos atributos afetam o valor dos imóveis. O presente estudo focou-se na relação entre as 61 variáveis escolhidas e o preço final das casas (*SalePrice*), tendo sido estudado o impacto individual de cada variável e as relações entre elas.

3.2. Pré-processamento

De modo a obter os dados necessários para a realização deste projeto foi implementado um pré-processamento dos mesmos. De forma a obter dados conformizados, este pré-processamento foi efetuado nos três datasets (train, test e sample submission).

Inicialmente, selecionou-se apenas as 62 variáveis de interesse, contendo 29 variáveis contínuas, 32 variáveis categóricas e a variável *target*, *SalePrice* . Depois, procedeu-se à remoção de *missing values*, sendo eliminadas todas as linhas com pelo menos um *missing value*. Em seguida, dividiu-se os valores do preço de venda das casas por 10 mil, transformando assim os dados em unidades de dezenas de milhares de dólares. Por fim, todas as variáveis categóricas foram convertidas em fatores, para facilitar a análise desses mesmos dados.

4. Resultados e Discussão

4. 1. Multicolinearidade para Variáveis Contínuas

Para analisar a presença de multicolinearidade entre as variáveis contínuas, criou-se um modelo de regressão linear com apenas as variáveis contínuas e *SalePrice*. A partir deste modelo, testou-se a multicolinearidade através do fator de inflação da variância (função vif no R), e concluiu-se que não existia problema de multicolinearidade, visto que nenhum valor se encontrava acima de 10.

Contudo, o teste realizado com o modelo de regressão linear com todas as variáveis contínuas também leva em conta a interação das variáveis entre si e não só o seu impacto na variável *target*. Desta forma, de modo a testar o impacto de cada variável contínua em *SalePrice*, realizaram-se regressões lineares simples para cada variável contínua e reteve-se os *p-values* e os valores de beta correspondentes a cada variável. Os resultados obtidos revelaram que existiam 7 variáveis que não contribuiam significativamente para a explicação de *SalePrice*: *BsmtFinSF2* (Rating of basement finished area) *LowQualFinSF* (Low quality finished square feet) *BsmtHalfBath* (Basement half bathrooms) *X3SsnPorch* (Three season porch area in square feet), *MiscVal* (Value of miscellaneous feature), *MoSold* (Month Sold), and *YrSold* (Year Sold). Assim, estas variáveis foram removidas do modelo.

4. 2. Impacto na Variável *Target*

As representações gráficas revelam *insights* sobre a importância de cada fator e a sua interação. Para averiguar quais das variáveis que tinham um impacto negativo ou positivo na variável *SalePrice*, foram realizados gráficos de cada uma das variáveis significativas para a explicação da variável target *SalePrice,* que estão representados em anexo nas Figuras 1, 2 e 3.

É possível observar que algumas variáveis como, por exemplo, *YearBuilt*, *YearRemodAdd*, *FullBath* e *TotRmsAbsGrd* têm um impacto positivo na variável *target* e que *KitchenAbvGr* parece ter um impacto negativo. Tal é confirmado pelos respectivos betas. Contudo, é difícil perceber qual o impacto das restantes variáveis utilizando apenas os gráficos, sendo assim necessário levar em consideração outros fatores, como os coeficientes, e testes, como o t-value, para avaliar a importância destas variáveis para a explicação de *SalePrice*.

4. 3. Multicolinearidade para Variáveis Categóricas

Para analisar a presença de multicolinearidade entre as variáveis categóricas, realizou-se duas análises, nomeadamente, o valor de Cramer e a análise de variância (ANOVA). Inicialmente, todas as variáveis categóricas foram convertidas em fatores. Os valores de V de Cramér foram calculados para cada par de variáveis categóricas. O V de Cramér é uma medida de associação entre duas variáveis categóricas e varia de 0 (nenhuma associação) a 1 (associação perfeita). Na presente análise, identificou-se alguns pares de variáveis com V de Cramér superior a 0.3, indicando uma correlação significativa.

De seguida, para determinar a significância de cada variável categórica em relação ao *SalePrice* procedeu-se a análise de variância (ANOVA) com o objetivo de verificar se existiam diferenças significativas nas médias de *SalePrice* para diferentes níveis

das variáveis categóricas. A partir dos resultados da ANOVA, foram obtidos os valores de F e os p-values correspondentes para cada variável categórica, sendo que variáveis com p-values superiores a 0.05 foram consideradas como não tendo um impacto significativo na *SalePrice*, nomeadamente a *Street*, *Utilities* and *LandSlope*.

4.4. Modelo com todas as Variáveis

De modo a avaliar a utilidade do modelo de regressão que inclui as variáveis selecionadas construiu-se dois modelos, um com todas as variáveis disponíveis (modelo completo) e outro com as variáveis reduzidas selecionadas a partir das análises anteriores com maior significância estatística e menor multicolinearidade (modelo reduzido). Em seguida, ambos os modelos foram comparados recorrendo à ANOVA.

O teste F foi usado para avaliar se a redução do modelo resultou numa perda significativa de capacidade explicativa. No entanto, obteve-se uma estatística F significativa ($F = 6.782$, $p < 0.001$), indicando que a redução do modelo não resultou numa perda significativa de informação e que os modelos são essencialmente idênticos em termos de ajuste aos dados. Desta forma, a presente análise demonstra que o modelo reduzido é altamente eficaz na previsão de *SalePrice*, explicando 93% da variabilidade na variável dependente. Portanto, a redução das variáveis não comprometeu a capacidade explicativa do modelo, tornando-o mais parcimonioso e interpretável.

4.5. Modelos Parcimonioso

Para tentar encontrar um modelo parcimonioso em relação ao modelo completo criado na alínea anterior, foram utilizados modelos criados a partir do critério de informação de Akaike (AIC). Este critério foi desenhado de maneira a encontrar o modelo que explica a maior quantidade de variabilidade nos dados e penalizar os modelos que usam parâmetros desnecessários.[4] Quanto menor for o AIC, melhor é o modelo. A função *stepAIC* no R retira ou acrescenta (ou ambos) variáveis do modelo previsto iterativamente até encontrar o modelo com o menor valor de AIC.

Para encontrar o melhor modelo possível, criaram-se três modelos a partir dos métodos possíveis: *backward*, que começa com o modelo completo e vai retirando variáveis; *forward*, que parte do modelo nulo (criado apenas com a variável *target* e sem qualquer outra variável) e acrescenta variáveis; e *both*, que é uma mistura dos dois métodos anteriores. Os modelos obtidos pelos métodos *backwards* e *both* foram idênticos, contendo 35 variáveis cada, enquanto que o modelo obtido com *forward* continha 38 variáveis. Como o modelo obtido por *backward/both* obteve um valor inferior de AIC em comparação ao modelo *forward*, este foi escolhido para prosseguir com a análise dos dados.

Para confirmar a escolha de modelo feita, comparámos os valores de AIC do modelo *both* e do modelo reduzido. O modelo reduzido obteve um valor de 6721.818 e o modelo *both* obteve 6649.481, valor inferior ao modelo reduzido, validando assim a escolha feita.

4.6. Análise de Resíduos

Para avaliar as suposições do modelo de regressão linear, é essencial analisar os resíduos gerados pelo modelo. A análise de resíduos fornece *insights* sobre a adequação do modelo, indicando se as suposições de homocedasticidade, normalidade dos erros e independência dos resíduos são atendidas.

O gráfico Scale-Location (raiz quadrada dos resíduos normalizados versus valores ajustados), apresentado na Figura 4, é utilizado para verificar a homocedasticidade (erros do modelo (resíduos) são constantes). A linha vermelha não é completamente horizontal, indicando uma ligeira tendência ascendente, o que sugere a presença de heterocedasticidade. No entanto, a variabilidade dos resíduos parece ser razoavelmente constante em torno da linha do zero para a maioria dos valores ajustados, exceto para alguns pontos de valores ajustados mais altos, onde a variância aumenta.

O gráfico de Residuals versus Leverage é utilizado para identificar pontos de alavancagem (*leverage points*) alta e potenciais outliers que podem distorcer os resultados do modelo. O *leverage* mede o potencial de uma observação influenciar a estimativa dos coeficientes de regressão. No gráfico da Figura 4, os pontos numerados (826, 589, 524, 804, 870) representam observações com alta *leverage*, indicando que esses pontos têm um impacto significativo na estimativa dos coeficientes do modelo.

No entanto, a maioria dos resíduos está concentrada perto de zero. Desta forma, apesar da presença de variabilidade não constante dos resíduos nos valores ajustados mais altos indicar heterocedasticidade e de se ter identificado alguns pontos de alta *leverage* que podem influenciar de maneira desproporcional os resultados do modelo, o modelo atual é robusto e explica uma parte significativa da variabilidade em *SalePrice*.

4.7. Teste do Modelo

Para testar a qualidade do modelo escolhido, procedeu-se a uma série de testes em dados que não tinham sido utilizados para a sua criação. Assim, utilizou-se o ficheiro 'test.csv', que continha as mesmas variáveis que o dataset original menos a variável *target*. Além disso, utilizou-se o ficheiro 'sample_submission.csv', que só continha os valores reais da variável *SalePrice*. Os passos de pré-processamento são os mesmos que os detalhados na seção 3.2, que incluiu a conversão de variáveis categóricas para fatores e a remoção de linhas com *missing values* e, para garantir consistência e que o tamanho de ambos os datasets fosse igual, as mesmas linhas foram removidas do dataset *sample_submission*.

O modelo selecionado (modelo *both*) foi então aplicado aos dados do dataset test. De seguida, as previsões geradas pelo modelo foram comparadas aos valores reais de *SalePrice* presentes no dataset *sample_submission*. Essa comparação permitiu avaliar a capacidade preditiva do modelo *both* em dados previamente não observados.

4.7.1. R^2

O coeficiente de determinação, R^2, foi calculado para avaliar o desempenho do modelo. Sendo que o R^2 indica a proporção da variabilidade total em *SalePrice* que pode ser explicada pelo modelo, ou seja, quantifica o quão bem o modelo explica a variabilidade dos dados. É um valor entre 0 e 1, onde valores mais próximos de 1 indicam um melhor ajuste do modelo aos dados.

O cálculo do R^2 foi calculado como a razão entre SSR e SST:

$$R^2 = \frac{SSR}{SSR + SSE} = \frac{SSR}{SST}$$

Tal que:

SSR (Sum of Squares Regression): soma dos quadrados das diferenças entre as previsões do modelo e a média dos valores reais de *SalePrice*.

SSE (Sum of Squares Error): soma dos quadrados das diferenças entre as previsões do modelo e os valores reais de *SalePrice*.

SST (Total Sum of Squares): soma dos quadrados das diferenças entre os valores reais de *SalePrice* e a média desses valores, que é a soma de SSR e SSE.

O valor obtido de R^2 foi de aproximadamente 0.51, o que indica que o modelo é capaz de explicar cerca de 51% da variabilidade nos dados de teste, sugerindo que, embora o modelo tenha capacidade preditiva razoável, é possível melhorar a precisão preditiva do modelo. No entanto, o modelo ainda é útil para entender os fatores que influenciam os preços das casas em Ames, Iowa.

4.7.2. Intervalos de Previsão

Os intervalos de previsão fornecem um intervalo, com determinado nível de confiança, neste caso 95%, no qual se espera que os valores reais estejam contidos. Para cada observação no dataset test foi gerado um intervalo de previsão com um limite inferior e superior.

De seguida, os intervalos de previsão foram utilizados para verificar quantas previsões do modelo continham o valor verdadeiro do *SalePrice* dentro do intervalo. Os resultados mostraram que cerca de 50,35% das previsões do modelo continham o valor verdadeiro de *SalePrice* dentro do intervalo de previsão. Tal sugere que o modelo possui uma precisão razoável nas suas precisões, capturando a incerteza inerente aos dados de teste.

4.7.3. Erro Médio Absoluto

Por fim, calculou-se o erro médio absoluto (MAE) para avaliar o desempenho do modelo em termos de precisão das previsões. O MAE mede a média das diferenças absolutas entre as previsões do modelo *both* e os valores reais presentes no dataset *sample_submission*.

O valor obtido foi de aproximadamente 5,77, indicando que cada valor previsto do modelo *both* no test set tinha, em média, um desvio de 5,77 mil dólares ao valor real de *SalePrice*. Quanto menor o valor do MAE mais precisa é a previsão do modelo, uma vez que indica um desvio menor entre as previsões e os valores reais.

4.8. Conclusão

O presente estudo foca-se na utilização da regressão linear na previsão dos preços das casas em Ames, Iowa. A análise minuciosa das variáveis e modelos permitiu uma compreensão abrangente da relação entre os atributos das casas e os valores dos preços das casas. Embora o modelo tenha demonstrado uma capacidade preditiva razoável, há margem para melhorias na precisão

das previsões. A identificação de multicolinearidade e variáveis estatisticamente não significativas facilitou a simplificação do modelo sem comprometer a sua capacidade explicativa. A análise de resíduos validou até certo ponto as suposições do modelo. Adicionalmente, os testes aplicados validaram a capacidade preditiva do modelo, embora o erro médio absoluto tenha indicado que é possível fazer melhorias.

Em suma, este estudo fornece informações valiosas para compreender os determinantes dos preços das casas em Ames, Iowa, destacando a importância da análise de regressão linear em modelar previsões no mercado imobiliário.

4.9. Referências

[1] https://www.geeksforgeeks.org/ml-linear-regression/
[2] https://www.ibm.com/topics/linear-regression
[3] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data
[4] https://www.statology.org/stepaic-r/

Problema 3

Beatriz Moreira, 54514 | Rute Patuleia, 51780

1. Resumo

A regressão logística é utilizada para classificação binária, transformando a regressão linear através da função sigmoide. Esta função mapeia valores reais para um intervalo de 0 a 1, representando probabilidades de uma instância pertencer a uma classe específica.

Neste estudo, foram analisados dados de 299 pacientes com insuficiência cardíaca, considerando 13 variáveis. O trabalho efetuado focou-se na variável *Time*, que se mostrou a mais relevante para prever a sobrevivência dos pacientes.

A regressão logística mostrou que o aumento no tempo de acompanhamento reduz a probabilidade de morte. O modelo apresentou bom ajuste e capacidade de discriminação, validado pela AUC de 0.84. A análise de custo identificou um *cutoff* de 0.51 como o mais eficiente.

Keywords: Regressão Logística, Previsão, Custo

2. Introdução

A regressão logística é um algoritmo de aprendizagem de máquina supervisionado utilizado para classificação binária, apesar de ter a palavra regressão no seu nome.[1] Essa nomenclatura está relacionada com o facto da regressão logística ser construída a partir da aplicação de uma transformação/função (denominada função logística ou sigmoide) sobre a regressão linear.[1] E, em vez de ajustar uma linha de regressão, ajusta-se uma função sigmoide ou logística numa curva em forma de "S", que prevê dois valores máximos (0 ou 1).[2] Ou seja, a função sigmoide é uma função matemática usada para mapear os valores previstos em probabilidades.[2]

A função sigmoide é matematicamente definida por:

$$y = b_0 + b_1 \times x$$

$$p = \frac{1}{1 + e^{-y}}$$

Tal que:

p: representa a probabilidade de uma dada instância pertencer à classe analisada.

y: número real dado pela combinação linear dos atributos utilizados na predição, derivado da regressão linear.

Desta forma, a função sigmoide recebe um número real (combinação linear de variáveis) como *input* e devolve um valor de *output* entre o intervalo de 0 e 1 que corresponde a uma probabilidade de pertencer a uma classe ou outra.[1] Por exemplo, temos duas classes, nomeadamente, Classe 0 e Classe 1.[2] Para realizar uma classificação a partir de uma probabilidade define-se um limiar de decisão (*threshold*), no qual os *inputs* cuja probabilidade ultrapassar esse limiar são classificados como pertencendo à Classe 1, caso contrário serão considerados da Classe 0.[1] Sendo assim, a regressão logística não só é capaz de classificar instâncias, mas também de informar a certeza/incerteza associada com a classificação, através do valor de probabilidade calculada.[1]

Para entender como chegamos a esse resultado, realiza-se os cálculos de forma reversa sobre a função logística. Sabendo que:

$$y = f(x) = b_0 + b_1 x_1 + \cdots + b_n x_n$$

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_n x_n)}}$$

Podemos reformular e reescrever as equações:

$$e^{b_0 + b_1 x_1 + \cdots + b_n x_n} = \frac{p}{1-p}$$

$$b_0 + b_1 x_1 + \cdots + b_n x_n = \ln\left(\frac{p}{1-p}\right)$$

Esta última equação é a função inversa da função logística (log-odds, também conhecido como função logit), representando, assim, o logaritmo natural das probabilidades.[1]

Note-se que *odds* (chance) é a razão entre algo que ocorre e algo que não ocorre, enquanto probabilidade é a razão entre algo que ocorre e tudo o que poderia ocorrer. Portanto, é a partir da definição de *odds* que deriva a capacidade da regressão logística em calcular a probabilidade e incerteza associada a uma instância classificada, por meio da função logit.[2]

O treino de um modelo de regressão logística consiste em encontrar a função sigmoide que melhor se ajusta aos dados de treino. Isto é, encontrar a combinação dos coeficientes (b0, b1, ..., bn) que minimiza os erros de predição e resulta no melhor desempenho possível. Para isso, utilizamos uma função erro/custo chamada Entropia Cruzada Binária.[1] Outra forma é recorrendo à estimativa de máxima verossimilhança. Este método é utilizado para estimar os coeficientes do modelo de regressão logística, que maximiza a probabilidade de observação dos dados fornecidos pelo modelo.[2]

Também é importante mencionar que existem várias suposições para este tipo de modelo. Nomeadamente, cada observação é independente da outra; a variável dependente deve ser binária; a relação entre as variáveis independentes e o logaritmo de probabilidades da variável dependente deve ser linear; não deve haver valores discrepantes no *dataset* e o tamanho da amostra deve ser suficientemente grande.[2]

Por fim, para avaliar o modelo de regressão logística pode-se usar as seguintes métricas:

- **Exatidão (Accuracy)**: A precisão fornece a proporção de instâncias classificadas corretamente.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total}$$

- **Precisão (Precision)**: A precisão concentra-se na precisão das previsões positivas.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- **Recall (Sensitivity ou True Positive Rate)**: A recall mede a proporção de instâncias positivas previstas corretamente entre todas as instâncias positivas reais.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **Pontuação F1 (F1 Score)**: A F1 score é a média harmónica de precisão e recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: A curva ROC representa graficamente a taxa de verdadeiros positivos em relação à taxa de falsos positivos em vários limites. A AUC-ROC mede a área sob esta curva, fornecendo uma medida agregada do desempenho de um modelo em diferentes limites de classificação.

- **Area Under the Precision-Recall Curve (AUC-PR)**: Semelhante ao AUC-ROC, a AUC-PR mede a área sob a curva de recall-precisão, fornecendo um resumo do desempenho de um modelo em diferentes compensações de recall-precisão.[2]

3. Metodologia

3.1. Descrição dos Dados

O presente projeto foi desenvolvido utilizando um conjunto de dados referentes aos registos médicos de 299 pacientes com insuficiência cardíaca recolhidos no Instituto de Cardiologia de Faisalabad e no Hospital dos Aliados em Faisalabad (Punjab, Paquistão), no período entre abril e dezembro de 2015. Os dados abrangiam pacientes entre os 40 e 95 anos de idade, com disfunção sistólica no ventrículo esquerdo e histórico de insuficiências cardíacas anteriores.[3]

No total, foram analisadas 13 colunas com informação sobre o estado clínico, características corporais e estilo de vida de cada paciente, incluindo 7 variáveis contínuas, 5 variáveis categóricas e a variável *target*, *Death_Event* (na qual um valor de 0 significa que o paciente sobreviveu e um valor de 1 que faleceu).[3]

Esta análise é deveras importante, uma vez que permite estudar quais as variáveis mais importantes para explicar a possibilidade de um paciente falecer depois de sofrer uma insuficiência cardíaca. O presente estudo focou-se na relação entre as 13 variáveis escolhidas e a sobrevivência do paciente, tentando perceber qual a variável mais importante para a explicação da variável *target*.

4. Resultados e Discussão

4. 1. Impacto das Variáveis Contínuas na Variável *Target*

Para analisar o impacto de cada variável contínua na variável *target*, realizaram-se regressões lineares simples para cada variável contínua e reteve-se os *p-values*, valores de beta e valores de R^2 correspondentes a cada variável. As variáveis contínuas com *p-values* significativos foram *Age*, *Ejection_Fraction*, *Serum_Creatinine*, *Serum_Sodium* e *Time*.

Quanto ao valor de R^2 obtido para cada variável, sabemos que o maior valor de R^2 corresponde à variável que tem o maior impacto na variável *target*. Assim, de acordo com os resultados obtidos, a variável *Time* é a que melhor descreve *Death_Event*, uma vez que tem o maior valor de R^2 (0.277).

4. 2. Representação Gráfica

De modo a visualizar a importância da variável *Time* na explicação da variável *Death_Event* criou-se um *boxplot* de *Time* em função de *Death_Event*, apresentado na Figura 1, em anexo.

Nesta figura podemos observar o impacto da variável *Time* na variável *target*, sendo que tempos de acompanhamento menores correspondem a uma quantidade maior de mortes.

4. 3. Categorização da Variável Contínua

Para categorizar a variável contínua escolhida, primeiro calculámos a percentagem de valores em cada quartil e dividimos a variável *Time* em *bins* de acordo com os valores obtidos. De seguida, calculou-se a média de *Death_Event* em cada intervalo de valores e fez-se o gráfico dos intervalos de valores ordenados pela proporção de valores em cada intervalo, apresentado em anexo na Figura 2.

Este gráfico permite-nos perceber que a maioria dos pacientes tende a morrer nos primeiros dias de acompanhamento, e que a proporção de mortes diminui com o aumento dos dias de acompanhamento. Tal implica que a taxa de sobrevivência aumenta com o passar do tempo.

4.4. Estimativas dos Coeficientes da Máxima Verossimilhança para o Modelo Logístico

Para estimar os coeficientes da máxima verossimilhança para o modelo logístico, inicialmente foi formulado o modelo de regressão logística, onde a variável dependente é a ocorrência de óbito (*Death_Event*) e a variável independente é o tempo de acompanhamento (*Time*). Utilizando a função de regressão logística generalizada (glm) com distribuição binomial e a função de ligação logística (logit), foi possível obter os parâmetros do modelo.

O modelo logístico resultou numa equação representada por:

$$logit(P(Death)) = \beta_0 + \beta_1 \times Time$$

Onde:

- β_0 é a interceptação do modelo, representando o log das probabilidades quando todas as variáveis independentes são iguais a zero.
- β_1 é o coeficiente estimado para a variável independente *Time*, indicando como a probabilidade de óbito varia com o tempo de acompanhamento.

A interpretação desses coeficientes permite-nos entender como a variável *Time* influencia a probabilidade de ocorrência do evento de morte ao longo do tempo. Neste caso, a análise do modelo logístico ajustado revela que ambos os interceptos e coeficientes para a variável *Time* são estatisticamente significativos, conforme indicado pelos *p-values*.

O intercepto β_0 tem um valor estimado de 1.452, com um desvio padrão de aproximadamente 0.282 e um valor de z de 5.149. Isso significa que quando todas as outras variáveis independentes são iguais a zero, o log das probabilidades de ocorrer o evento (*Death_Event*) é significativamente diferente de zero.

O coeficiente para *Time* (β_1) tem um valor estimado de -0.020, com um desvio padrão de aproximadamente 0.003 e um valor de z de -7.804. Isso indica que, para cada unidade de aumento no tempo de acompanhamento (*Time*), o log das probabilidades de ocorrer o evento (*Death_Event*) diminui em média 0.020 unidades, mantendo todas as outras variáveis constantes.

Além disso, o modelo logístico apresenta um bom ajuste aos dados, como evidenciado pelo baixo valor do desvio residual e pelo teste de deviance. O teste de deviance compara o modelo ajustado com um modelo nulo, onde um valor menor de desvio-padrão dos resíduos (deviance residual) indica um melhor ajuste do modelo aos dados observados. Nesse caso, o desvio-padrão dos resíduos do modelo ajustado é de 279.07, em comparação com 375.35 para o modelo nulo, indicando uma melhoria significativa no ajuste do modelo.

O AIC (Critério de Informação de Akaike) fornece uma medida da qualidade relativa do modelo estatístico, levando em consideração a complexidade do modelo. Quanto menor o valor do AIC, melhor o ajuste do modelo. Neste caso, o AIC é de 283.07, o que sugere que o modelo logístico ajustado é uma escolha razoável para explicar os dados.

Para visualizar essa relação, traçou-se a função logística que mostra a probabilidade de morte em função do tempo, conforme apresentado na Figura 3, em anexo.

4.5. Avaliação do Modelo

4.5.1. Avaliação do Modelo Logístico usando o Teste de Razão de Verossimilhança

Para avaliar o modelo logístico obtido, comparamos este modelo ao modelo *null* (o modelo que apenas contém a variável *target* e nenhuma das outras variáveis) através do teste de razão de verossimilhança. Este teste compara o desvio-padrão dos resíduos do modelo ajustado com o desvio-padrão dos resíduos do modelo nulo e fornece uma estatística de teste que segue uma distribuição qui-quadrado.

No teste de razão de verossimilhança, as hipótese são as seguintes:

- H_0: O modelo nulo é igualmente eficaz em explicar a variabilidade nos dados em comparação com o modelo logístico.
- H_1: O modelo logístico é significativamente melhor em explicar a variabilidade nos dados do que o modelo nulo.

A tabela mostra que o modelo 2 (modelo ajustado com a inclusão da variável *Time*) tem um desvio-padrão dos resíduos de 279.07 com 297 graus de liberdade residuais. Enquanto isso, o modelo 1 (modelo nulo) tem um desvio-padrão dos resíduos de 375.35 com 298 graus de liberdade residuais. A diferença no desvio-padrão dos resíduos entre os dois modelos é de 96.275, com um *p-value* extremamente baixo (< 0.001), indicando que o modelo ajustado é significativamente melhor do que o modelo nulo em explicar a variabilidade nos dados.

Portanto, com base no teste de razão de verossimilhança, podemos rejeitar a hipótese nula e concluir que o modelo logístico ajustado, que inclui a variável *Time*, é estatisticamente significativo e fornece um melhor ajuste aos dados do que o modelo nulo.

Em suma, o teste foi realizado através de uma ANOVA, utilizando o teste do χ^2, e o *p-value* obtido foi de < 0.001. Como o *p-value* obtido é significativo, rejeita-se a hipótese nula e podemos concluir que o modelo logístico é melhor do que o modelo nulo, neste caso.

4.5.2. Curva ROC e Estimação da AUC

Para avaliar a eficácia do modelo logístico na classificação dos eventos de óbito, utilizamos a curva ROC (Receiver Operating Characteristic) e estimamos a Área Sob a Curva (AUC). Para obter a curva ROC (Receiver Operating Characteristic), criámos o modelo de previsões através do modelo logístico estudado nas etapas anteriores. A curva ROC é então o gráfico da taxa de verdadeiros positivos (TPR) em função da taxa de falsos positivos (FPR) para diferentes pontos de corte na probabilidade prevista de óbito, estando representada na Figura 4, em anexo.

Como podemos observar na figura, o modelo obtido tem uma boa performance, uma vez que a curva obtida não é uma linha reta. A partir de 0.6 de taxa de verdadeiros positivos o modelo tem boa discriminação, e a partir de 0.8, temos uma excelente discriminação.

A estimativa da AUC (Area Under the Curve) é uma indicação da qualidade do modelo obtido. Um modelo com AUC igual a 1 consegue classificar perfeitamente as observações em classes, enquanto que um modelo com AUC de 0.5 tem uma performance igual a um modelo de classificação *random*.[4]

O resultado mostrou que a AUC do modelo foi de aproximadamente 0.84, ou 84%. Assim, o modelo proposto tem um bom resultado de AUC, logo é capaz de prever razoavelmente bem em que classes se insere cada observação.

A curva ROC também mostra um bom afastamento da linha diagonal, o que confirma a capacidade discriminativa do modelo.

De modo a ter uma visão mais detalhada da performance do modelo, foi também traçado o gráfico da *accuracy* em função do *cutoff* (Figura 4). Deste gráfico, podemos concluir que o maior valor de precisão é atingido utilizando um *cutoff* de aproximadamente 0.5.

4.5.3. Análise de Custo

Para obter o *cutoff* que minimizava o custo do modelo obtido, tendo em conta que falsos positivos tinham um custo de 1000 euros e falsos negativos de 1500 euros, criámos uma lista com possíveis *thresholds* de 0 a 1, com um passo de 0.01 entre cada valor testado. De seguida, foi criado um ciclo em que cada valor previsto era comparado com o *threshold* escolhido, e se era inferior ao *threshold*, era classificado como 0, caso contrário, como 1. A classificação obtida era então comparada com a classificação real, e os casos de falsos positivos e falsos negativos eram contabilizados, de modo a calcular o custo total do *threshold* utilizado. No fim do ciclo, escolhemos o *threshold* com o custo total mínimo, que neste caso, foi de 0.51 (com um custo associado de 65,000 euros). Este *threshold* permite simultaneamente obter a melhor performance do modelo em classificar os pacientes em casos de morte ou não e o menor custo possível para o caso proposto.

Para visualizar a evolução do custo total contra o *cutoff* escolhido, foi feito o gráfico destas duas variáveis, apresentado na Figura 5. A representação gráfica confirma os resultados obtidos anteriormente e mostra que há uma tendência decrescente no custo total inicialmente, mas que depois de atingido o valor de *threshold* com o custo total mínimo, este tende a voltar a aumentar.

4.6. Conclusão

O presente estudo foca-se na utilização de um modelo de regressão logística para prever se um paciente irá falecer ou não após uma insuficiência cardíaca, destacando a variável *Time* como crucial para prever a sobrevivência dos pacientes.

O modelo logístico ajustado mostrou bom desempenho estatístico e capacidade discriminativa, avaliado através de medidas como a curva ROC e a estimativa de AUC. Além disso, a análise de custo aprimorou a eficiência do modelo na classificação de eventos de óbito, identificando um *cutoff* ideal tendo em conta as circunstâncias descritas.

Estes resultados indicam que a regressão logística é uma ferramenta valiosa para estudos médicos e outras aplicações de classificação binária, proporcionando tanto previsões precisas quanto *insights* sobre a incerteza associada a cada classificação.

4.7. Referências

[1] https://medium.com/@msremigio/regress%C3%A3o-log%C3%ADstica-logistic-regression-997c6259ff9a
[2] https://www.geeksforgeeks.org/understanding-logistic-regression/
[3] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*(1), 16. https://doi.org/10.1186/s12911-020-1023-5
[4] https://www.statology.org/what-is-a-good-auc-score/

Anexo do Projeto 1

```r
houses<-read.csv('train.csv')
houses<-data.frame(houses)
head(houses)
central_air<-houses$CentralAir
kitchen_quali<-houses$KitchenQual
sale_price<-houses$SalePrice
sale_price<-sale_price/10000
kitchen_quali[kitchen_quali=='Ex']<-'Gd'
fixed_data<-cbind(sale_price,central_air,kitchen_quali)
write.table(fixed_data,"dataset.txt")
```

Figura 1. Código em R para a conversão do preço de venda para unidades de dezenas de milhares de dólares e transformação do nível "Ex" em "Gd".

```r
data<-read.table("dataset.txt")
head(data)
library(dplyr)
sampled_data <- data %>%
  group_by(central_air, kitchen_quali) %>%
  sample_n(size = 6, replace = FALSE)


# Write the sampled data to a .txt file
write.table(sampled_data,"Group1.txt", sep = "\t", row.names = FALSE)
```

Figura 2. Código em R para a obtenção da amostra aleatória.

Figura 3. Boxplot do Preço de Venda pela Qualidade da Cozinha. Fa corresponde a *Fair*, Gd corresponde a *Good* e TA corresponde a *Typical/Average*.

Figura 4. Boxplot do Preço de Venda pela Presença de Ar Condicionado Central. N e Y corresponde à presença e ausência de ar condicionado, respetivamente.

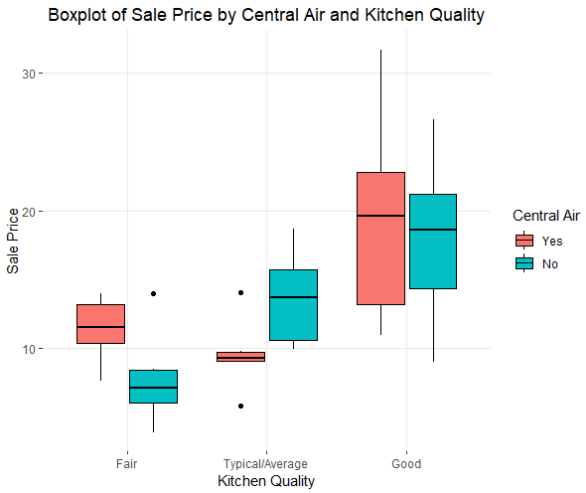

Figura 5. Boxplot do Preço de Venda de Casas em função da Presença de Ar Condicionado Central e da Qualidade da Cozinha.

Figura 6. Gráfico de Interação da Qualidade da Cozinha pela presença do Ar Condicionado.

Anexo do Projeto 2

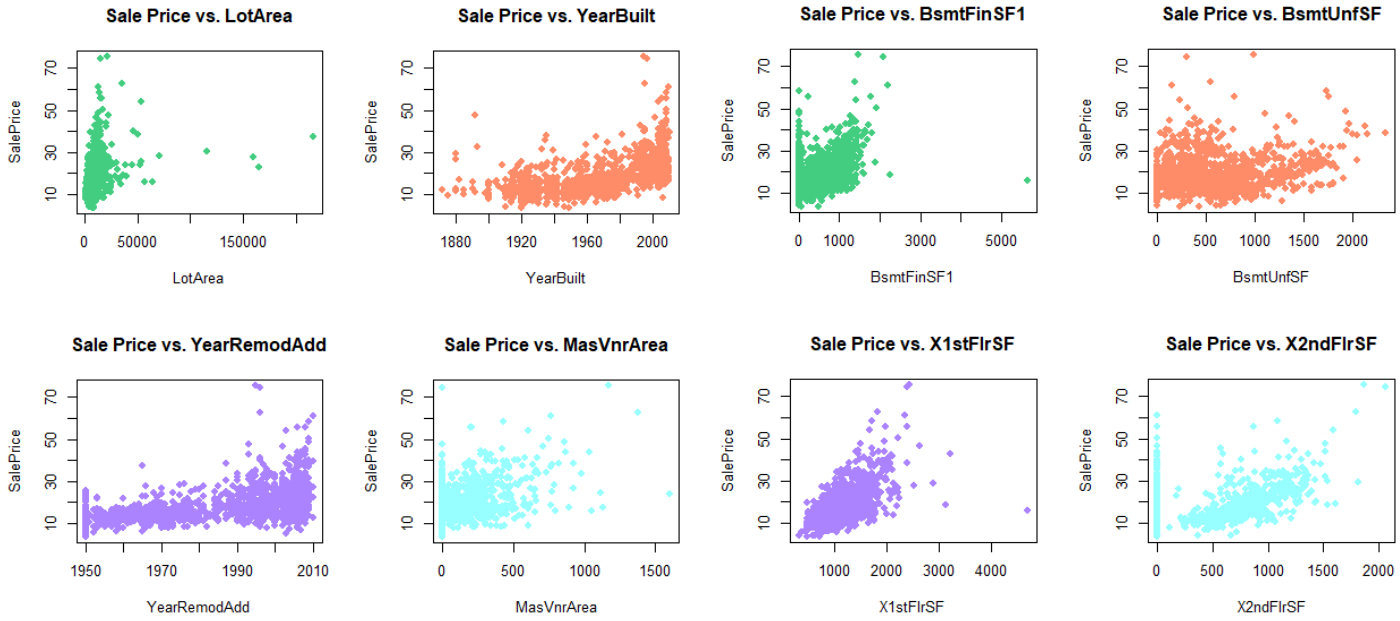

Figura 1. Representação das regressões lineares entre *LotArea*, *YearBuilt*, *BsmtFinSF1*, *BsmtUnfSF*, *YearRemodAdd*, *MasVnrArea*, *X1stFlrSF* , *X2stFlrSF* e *SalePrice*.

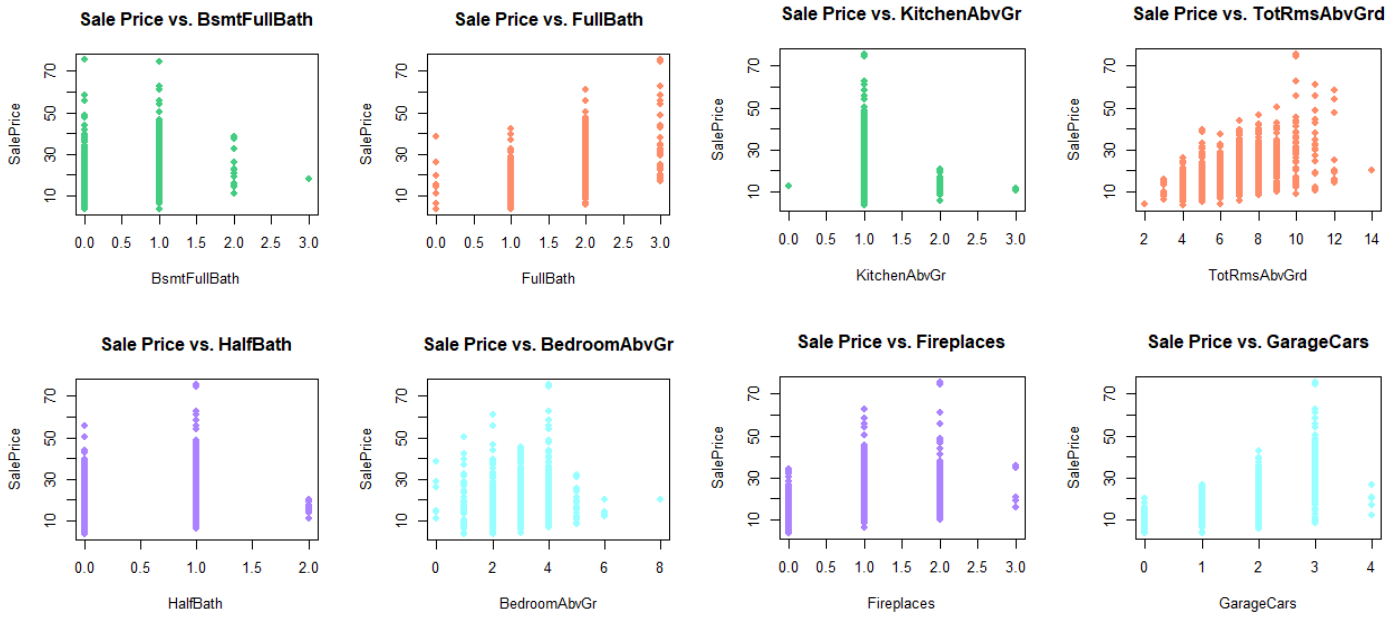

Figura 2. Representação das regressões lineares entre *BsmtFullBath*, *FullBath*, *KitchenAbvGr*, *TotRmsAbsGrd*, *HalfBath*, *BedroomAbvGr*, *Fireplaces* , *GarageCars* e *SalePrice*.

Figura 3. Representação das regressões lineares entre *GarageArea*, *WoodDeckSF*, *ScreenPorch*, *PoolArea*, *OpenPorchSF*, *EnclosedPorch* e *SalePrice*.

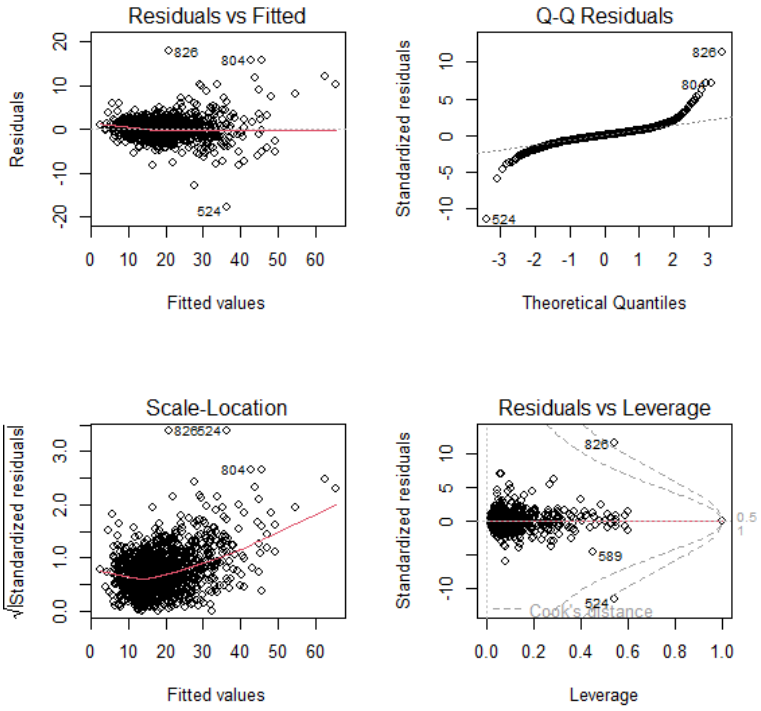

Figura 4. Representação dos resíduos.

Anexo do Projeto 3

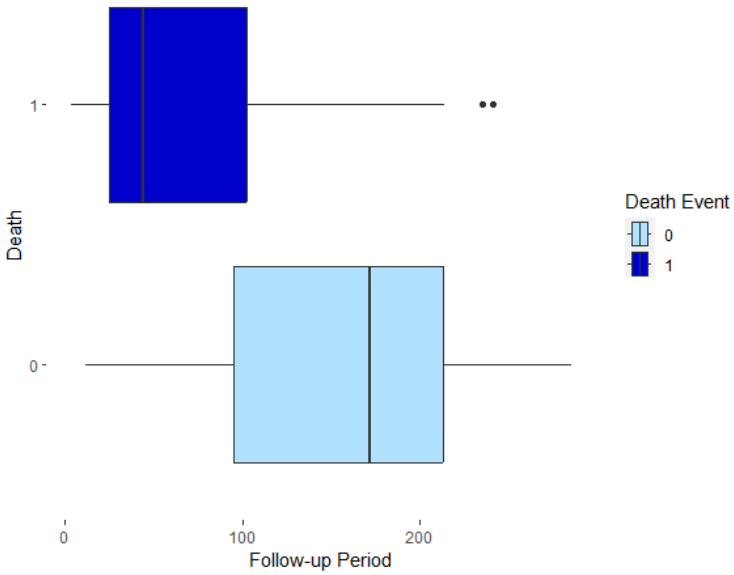

Figura 1. *Boxplot* da variável *Time* em função do *Death_Event*.

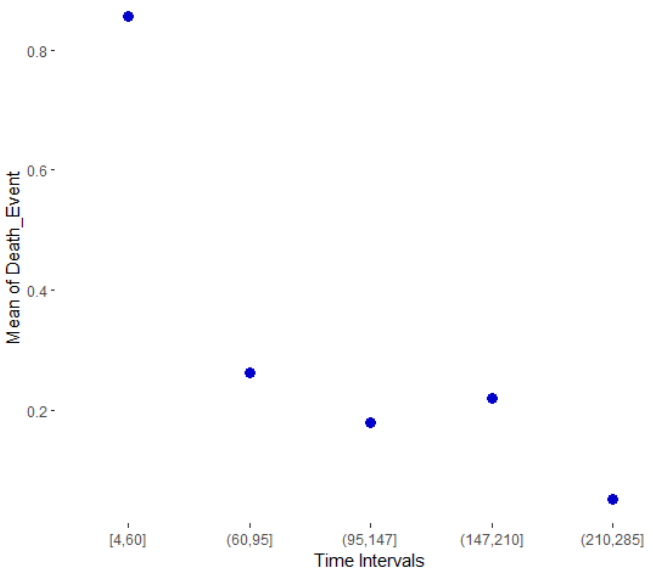

Figura 2. Gráfico da média de *Death_Event* por intervalo de tempo.

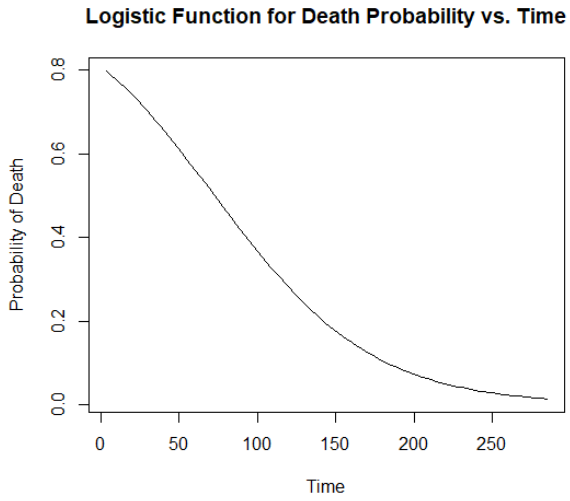

Figura 3. Gráfico da Função Logística da Probabilidade de Mortes vs Tempo.

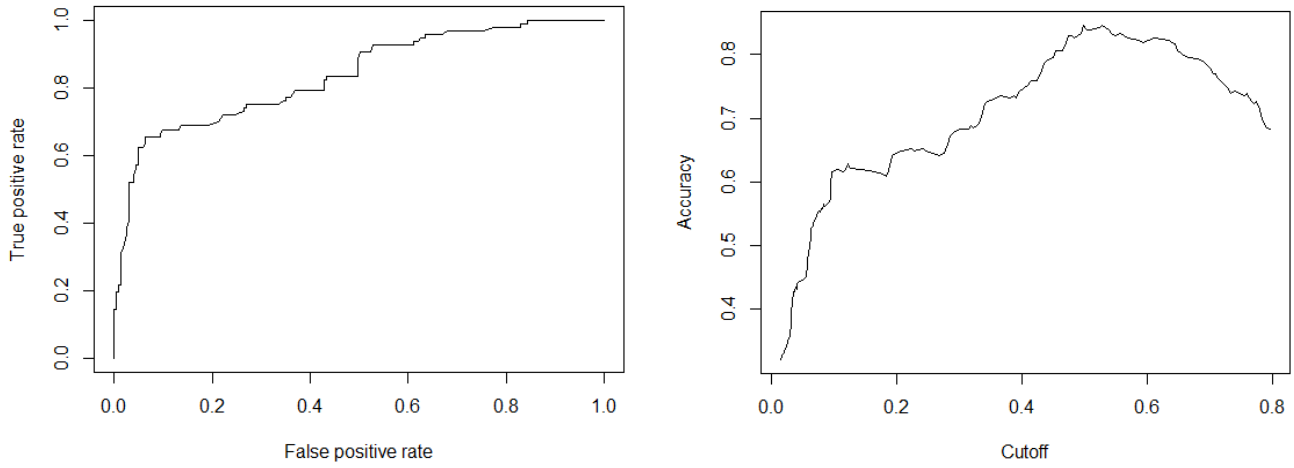

Figura 4. Gráfico da curva ROC (à esquerda) e da precisão (à direita) do modelo de previsões.

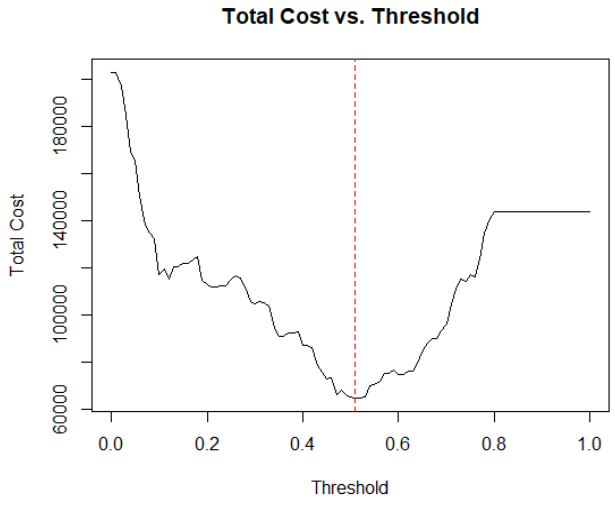

Figura 5. Gráfico do custo total em função do *threshold* escolhido.