ANÁLISE DE CIDADES PARA ABERTURA DE CLÍNICAS DE FISIOTERAPIA NO BRASIL

PROJETO EM GRUPO - PARTE 2

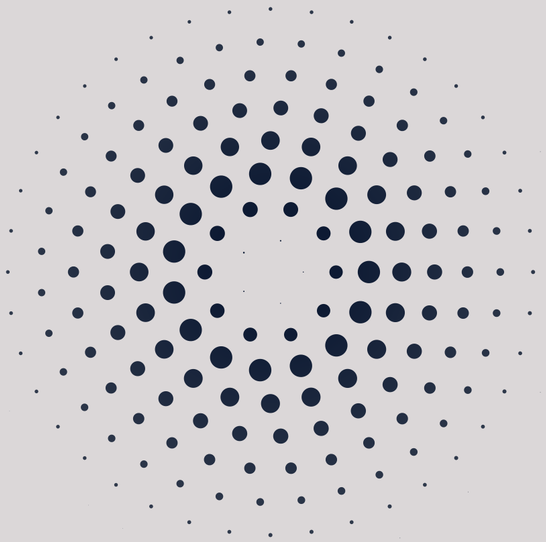
07/08/23

Brenda Farias
Fabricio Leal

ADM01007 - Introdução à Data Science

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Etapas



01	Pré-processamento e limpeza de dados
02	Análise exploratória
03	Selecionar modelos
04	Rodar previsões/inferência

Objetivos

- Analisar cidades no Brasil e identificar aquelas que indicam melhores oportunidades para a abertura de clínicas de fisioterapia.

- Por meio dessa análise, busca-se fornecer informações para orientar a abertura ou expansão de clínicas de fisioterapia em regiões que apresentem maior necessidade ou oportunidade desses serviços.

Contexto

- O contexto da análise engloba todos os municípios do Brasil, embora dados sobre todos os municípios não foram encontrados para todas as variáveis.
- A área escolha é saúde.

Pré-processamento e limpeza de dados

Adicionado a variável "Gasto público em saúde per capita"

	Codigo	Nome	Gasto_Saude_Muni/Cap
0	110001	Alta Floresta D'Oeste	463.51
1	110037	Alto Alegre dos Parecis	395.54
2	110040	Alto Paraíso	299.95
3	110034	Alvorada D'Oeste	320.78
4	110002	Ariquemes	363.66
...	...	...	...
5520	522190	Varjão	459.93
5521	522200	Vianópolis	409.71
5522	522205	Vicentinópolis	419.86
5523	522220	Vila Boa	525.31
5524	522230	Vila Propício	458.76

5525 rows × 3 columns

DESAFIOS

Muitos dados faltando. Quatro variáveis apresentam uma grande lacuna com muitos missings: PH_Articulacoes (96% faltando), PH_Ortopedicas (96% faltando), AcidTransito (73% faltando), Qtd empresas (53% faltando). Imputação não é uma opção razoável pois dado que o número de dados imputados seria muito maior que o número de dados existentes, os dados seriam distorcidos e não confiáveis.

DESAFIOS

Escolha da variável dependente. Considerando a nossa base de dados, escolhemos Total de beneficiários de plano de saúde para ser nossa variável dependente. Motivos:

- Relevância Direta
- Acessibilidade Financeira
- Dados Disponíveis
- Viabilidade de Negócios

DESAFIOS

	Missing Values	Percentage
Nome	0	0.000000
Codigo	0	0.000000
Estado	0	0.000000
20 a 29 anos	0	0.000000
30 a 39 anos	0	0.000000
40 a 49 anos	0	0.000000
50 a 59 anos	0	0.000000
60 a 69 anos	0	0.000000
70 a 79 anos	0	0.000000
80 anos e mais	0	0.000000
Menor que 1 a 9 anos	0	0.000000
10 a 19 anos	0	0.000000
Total_População	0	0.000000
Total_BeneficiariosPlanoSaude	12	0.215633
PH_Articulacoes	5366	96.424079
PH_Ortopedicas	5389	96.837376
AcidTransito	4090	73.495058
Qtd empresas	3183	57.196765
PIB	0	0.000000
PIB/capita	0	0.000000
VABServiços	0	0.000000
Gasto_Saude_Muni/Cap	40	0.718778

POSSIBILIDADES

Possibilidade 1: Analisar municípios com dados completos

Possibilidade 2: Excluir variáveis com muitos missings.

POSSIBILIDADE 1

INFORMAÇÕES

Ao analisar apenas os municípios que possuem dados completos para todas as variáveis, focamos em um subconjunto menor de dados, mas que são mais confiáveis. Nesse novo dataframe, municipios_completos, temos 54 linhas e 22 colunas. Os municípios que possuem dados completos podem ser conferidos logo abaixo do dataframe.

POSSIBILIDADE 1

MODELOS

1. Regressão linear
2. Regressão polinomial
3. Regressão ridge
4. LASSO
5. Random forest
6. XGBoost

POSSIBILIDADE 1

CONCLUSÃO

Verificou-se diversos problemas: multicolinearidade, resíduos não estão normalmente distribuidos, presença de heteroscedasticidade, overfitting, etc. Os problemas persistem mesmo removendo algumas variáveis independentes.

R-squared e Adj. R-squared: O valor de R-squared é muito alto (0.995), um sinal de overfitting.

POSSIBILIDADE 2

INFORMAÇÕES

Considerando que algumas variáveis possuem muitos valores faltando, tentaremos modelar sem essas variáveis.

O novo dataframe, df_menor possui 18 colunas, sendo que excluimos PH_Articulacoes, PH_Ortopedicas, Qtd empresas e AcidTransito.

POSSIBILIDADE 2

MODELOS

1. Regressão linear
2. Ridge
3. Lasso
4. ElasticNet
5. Random Forest Regressor
6. Gradient Boosting Regressor
7. XGBoost

POSSIBILIDADE 1

RESULTADOS E PROBLEMAS DOS MODELOS

Todos os modelos de regressão (Linear, Ridge, Lasso, ElasticNet) mostraram um desempenho semelhante, com o R^2 variando em torno de 0.87, o que indica que eles são capazes de explicar cerca de 87% da variância no conjunto de dados. No entanto, tanto o modelo Lasso quanto o ElasticNet apresentaram problemas de convergência, indicando que podem não ter sido totalmente otimizados.

POSSIBILIDADE 1

MELHOR MODELO

Entre os modelos de Machine Learning, o XGBoost teve o melhor desempenho com o R^2 mais alto (0.923), seguido pelo Gradient Boosting (0.920) e pelo Random Forest (0.919). Considerando tanto o desempenho quanto os recursos de cada modelo, o XGBoost poderia ser considerado o melhor modelo para este conjunto de dados.

POSSIBILIDADE 1

VALIDAÇÃO DAS HIPOTESES

Com base no resultado do OLS, as variáveis '30 a 39 anos', '50 a 59 anos', '60 a 69 anos', '70 a 79 anos', '80 anos e mais', 'Menor que 1 a 9 anos', '10 a 19 anos', 'Total_População', 'PIB', e 'VABServiços' são estatisticamente significativas em um nível de significância de 0.05. Portanto, a hipótese nula era de que essas variáveis não tinham relação com a variável dependente, foi rejeitada.

POSSIBILIDADE 1

IMPORTÂNCIA DOS RECURSOS

No que diz respeito à importância dos recursos, os modelos baseados em árvore (Random Forest, Gradient Boosting e XGBoost) deram informações úteis. No entanto, a importância dos recursos variou bastante entre os modelos, o que sugere que a contribuição relativa de cada recurso para o resultado previsto pode depender do modelo específico utilizado.

CONCLUSÃO

O modelo XGBoost é mais adequado para prever o número de beneficiários de plano de saúde em cidades do Brasil em comparação com a regressão linear. Sua melhor capacidade de generalização e identificação de características relevantes torna-o recomendado para decisões sobre clínicas de fisioterapia, proporcionando informações valiosas para a tomada de decisão.

Links úteis

Colab

https://colab.research.google.com/drive/1PfkxdeyRwC3OKCHQRGuqK8vDaDTr3zWI?usp=sharing

Github

https://githubusercontent.com/bsf94/trabgrupoDS/main