

Case Geofusion

Cientista de Dados

| Versão | Data | Nome | Modificação |
|--------|------------|---------------|--|
| 1.0 | 22/04/2022 | Isac Carvalho | Estudo/ Análise Exploratória/Modelagem |

Visão Geral

O case para a vaga de Cientista de Dados da Geofusion consiste nos dados fictícios de uma empresa alimentícia situada no Rio de Janeiro, que deseja abrir filiais na cidade de São Paulo. Tendo como público alvo adultos de 25 a 50 anos e domicílios das classes A (rendas A1 e A2) e B (rendas B1 e B2). A empresa quer estimar o quanto ela poderia faturar em cada um dos bairros de São Paulo.

Com os dados em CSV fornecido pela empresa, temos 3 objetivos principais:

1. Estimar o faturamento que uma loja teria em cada um dos bairros de São Paulo.
2. Classificar o potencial de cada bairro de São Paulo como Alto, Médio ou Baixo.
3. Segmentar os bairros de São Paulo de acordo com a renda e a idade, e indicar aqueles com maior aderência ao público alvo.

Referente a natureza do problema apresentado no case, também foi solicitado para verificar se é possível usar fontes públicas ou fontes privadas, que poderiam agregar mais valor para análise realizada.

Etapa Análise Exploratória

Nesta etapa foi a fase de explorar a base de dados e entender o que precisa ser feito para responder os objetivos do case. Toda a análise exploratória será feita com o uso da Linguagem R.

Conhecendo a Base de Dados

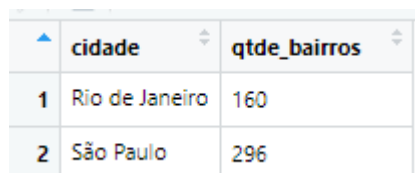
Os dados estão distribuídos em uma única tabela em formato csv. Essa base de dados possui 456 linhas e 24 colunas, onde a mesma possui dados das cidades de São Paulo e Rio de Janeiro. Cada estado é dividido por seus respectivos bairros e cada um desses bairros possui: a quantidade de população separada por classe de idade, a quantidade de domicílios separada por classe de renda, a renda média por domicílio, o faturamento total e a classificação de potencial de cada bairro.

Dicionário de Dados

| | |
|--------------|--------------------------------------|
| codigo | Código do bairro |
| nome | Nome do bairro |
| cidade | Cidade |
| estado | Estado |
| população | População total |
| popAte9 | População - até 9 anos |
| popDe10a14 | População - de 10 a 14 anos |
| popDe15a19 | População - de 15 a 19 anos |
| popDe20a24 | População - de 20 a 24 anos |
| popDe25a34 | População - de 25 a 34 anos |
| popDe35a49 | População - de 35 a 49 anos |
| popDe50a59 | População - de 50 a 59 anos |
| popMaisDe60 | População - 60 anos ou mais |
| domiciliosA1 | Quantidade de Domicílios de Renda A1 |
| domiciliosA2 | Quantidade de Domicílios de Renda A2 |
| domiciliosB1 | Quantidade de Domicílios de Renda B1 |
| domiciliosB2 | Quantidade de Domicílios de Renda B2 |
| domiciliosC1 | Quantidade de Domicílios de Renda C1 |
| domiciliosC2 | Quantidade de Domicílios de Renda C2 |
| domiciliosD | Quantidade de Domicílios de Renda D |
| domiciliosE | Quantidade de Domicílios de Renda E |
| rendaMedia | Renda Média por Domicílio |
| faturamento | Faturamento Total no Bairro |
| potencial | Potencial do Bairro |

Primeiras análises e conclusões

As primeiras análises foram feitas para entender como a nossa base de dados é distribuída e o primeiro passo foi verificar a quantidade de bairros em cada cidade.

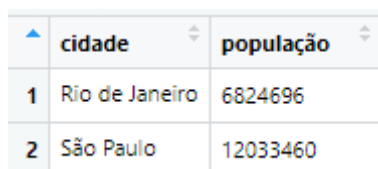


| | cidade | qtde_bairros |
|---|----------------|--------------|
| 1 | Rio de Janeiro | 160 |
| 2 | São Paulo | 296 |

Imagem R Studio, 2022.

Na cidade de Rio de Janeiro temos um total de 160 bairros e na cidade de São Paulo temos um total de 296 bairros.

Um outro fator importante é a distribuição da população nas duas cidades.

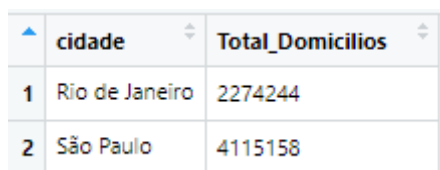


| | cidade | população |
|---|----------------|-----------|
| 1 | Rio de Janeiro | 6824696 |
| 2 | São Paulo | 12033460 |

Imagem R Studio, 2022.

Na cidade de Rio de Janeiro temos uma população de mais de 6,8 milhões de pessoas e na cidade de São Paulo temos uma população de mais de 12 milhões de pessoas.

Para verificar a distribuição em relação à quantidade de domicílios foi necessário acrescentar uma nova coluna como 'Total_Domicilios'.



| | cidade | Total_Domicilios |
|---|----------------|------------------|
| 1 | Rio de Janeiro | 2274244 |
| 2 | São Paulo | 4115158 |

Imagem R Studio, 2022.

Na cidade de Rio de Janeiro temos mais de 2,2 milhões de domicílios, na cidade de São Paulo temos mais de 4,1 milhões de domicílios.

Para verificarmos se seria um bom investimento para abrir filiais na cidade de São Paulo, podemos comparar as duas cidades com o público alvo. Para fazer essa comparação usamos a média da população de 25 a 50 anos e a média da quantidade de domicílios das classes A (A1 e A2) e B (B1 e B2).

| | cidade | Media_população25_34 | Media_população34_49 | Media_domiciliosA1 | Media_domiciliosA2 | Media_domiciliosB1 | Media_domiciliosB2 |
|---|----------------|----------------------|----------------------|--------------------|--------------------|--------------------|--------------------|
| 1 | Rio de Janeiro | 6584 | 8827 | 388 | 608 | 1872 | 2385 |
| 2 | São Paulo | 6441 | 9138 | 415 | 566 | 1911 | 2250 |

Imagem R Studio, 2022.

Vale ressaltar que a nossa base de dados tem os dados consolidados da população de 34 a 49 anos, a população de 50 anos está consolidada em um outro grupo (população de 50 a 59 anos). Com isso podemos considerar que o público alvo seria um pouco maior.

Como primeiras conclusões podemos perceber que dentro do público alvo a cidade de São Paulo apresenta Médias da população e de domicílios bem próximos e em alguns casos até superiores em relação a cidade do Rio de Janeiro. Esses indicadores nos mostram que investir em filiais na cidade de São Paulo possui um grande potencial de dar certo.

Preparando base de dados

Para sabermos onde investir na cidade de São Paulo, precisamos analisar bairro por bairro. Porém, para podermos olhar quais bairros seriam o melhor investimento, precisamos saber de duas informações cruciais: Faturamento e Potencial.

Obviamente, em nossa base de dados na categoria que traz a cidade de São Paulo não temos essas informações, afinal o desafio é esse. Por conta disso, vamos ter que criar um modelo de previsão que nos forneça esses dados.

Mas antes, precisamos preparar nossa base de dados. Isso é, verificar se cada variável está classificada corretamente, se temos dados vazios, se necessitamos de mais dados ou se esses dados estão na estrutura que precisamos.

O primeiro ponto já mencionado anteriormente, foi a necessidade de criar a coluna 'Total_Domicilios', que traz a somatória de todas as classes de domicílios.

Um segundo ponto foi que temos bairros com a população e a quantidade de domicílios com um total de zero. Isso ocorre com 3 bairros: "Reserva Da Cantareira", "Eta Guaraú" e "Pico Do Jaraguá", todos da cidade de São Paulo. Por falta de dados nesses bairros, os mesmos foram excluídos da nossa base de dados.

O terceiro ponto é que temos dados vazios na coluna 'rendaMedia'. Isso ocorre com 6 bairros: "Catumbi", "Rio Comprido", "Maracanã", "Anil", "Freguesia (Jacarepaguá)" e "Jacaré", todos da cidade do Rio de Janeiro. Apesar da ausência desses dados ser apenas em 6 observações, são dados que podemos estimar. Pensando em um cenário real, esse problema poderia ocorrer novamente se o mesmo estudo fosse aplicado para outras cidades. Criar uma solução para esse caso evitaria a perda de dados e faríamos ganhar tempo a longo prazo.

Com todas essas mudanças, agora temos uma base de dados de 453 linhas e 25 colunas. Onde a cidade do Rio de Janeiro se mantém com 160 bairros e a cidade de São Paulo passou a ter 293 bairros.

Etapa Modelagem de Dados

Nesta etapa vamos continuar usando a linguagem R para fazer a modelagem e fazer as estimativas que precisamos.

Primeiro vamos estimar a Renda Média dos bairros que estão faltando na base do Rio de Janeiro.

Com a renda média dos bairros que faltam, vamos poder estimar o Faturamento dos bairros de São Paulo.

Enfim, com os valores de Faturamento estimados, poderemos estimar o Potencial de cada bairro da cidade de São Paulo

Modelo Renda Média

Para estimar a Renda Média foi usado o modelo de regressão Random Forest.

Para a base de dados de treino e teste do modelo foi usada uma proporção 70/30 e as variáveis usadas no modelo foram 18 (relacionadas à população, à quantidade de domicílios, faturamento e a Renda Média que foi a variável preditora).

Com o modelo de regressão Random Forest chegamos ao Coeficiente de Determinação R^2 de $\sim 0,8$ na base de treino e $\sim 0,94$ na base de teste. Esse coeficiente é importante porque nos mostra o quanto o modelo se adapta a nossa base de dados. O coeficiente R^2 varia de 0 a 1, sendo assim o coeficiente da base de teste indica um ótimo resultado.

Com o modelo de regressão Random Forest também tivemos resultados interessantes com um Erro Médio Absoluto(MAE) de ~ 551 . Isso quer dizer que temos uma margem de erro de 551 para mais ou para menos nos valores previstos para a Renda Média.

Em comparação com os valores (mínimo, 1º quartil, mediana, média, 3º quartil e máximo) da Renda Média da base de dados original, o Erro Médio Absoluto é relativamente mais baixo.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|-------|
| 654 | 1486 | 1916 | 3608 | 2954 | 63887 |

Imagem R Studio, 2022.

Modelo Faturamento

Para estimar o Faturamento foi usado o modelo de regressão redes neurais H2O.

Para a base de dados de treino e teste do modelo foi usada uma proporção 70/30 e as variáveis usadas no modelo foram 18 (relacionadas à população, à quantidade de domicílios, Renda Média e faturamento que foi a variável preditora).

Com o modelo de regressão redes neurais H2O chegamos ao Coeficiente de Determinação R^2 de $\sim 0,997$ na base de treino e ~ 0.994 na base de teste, o que nos indica um ótimo resultado.

Com o modelo de regressão redes neurais H2O também tivemos bons resultados com um Erro Médio Absoluto(MAE) de ~24387 em Faturamento.

Em comparação com os valores (mínimo, 1º quartil, mediana, média, 3º quartil e máximo) do Faturamento da base de dados original, o Erro Médio Absoluto é relativamente mais baixo.

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
24651 711279 824229 876160 934321 2915612
```

Imagem R Studio, 2022.

Modelo Potencial

Para estimar o Potencial foi usado o modelo de classificação redes neurais H2O.

Para a base de dados de treino e teste do modelo foi usada uma proporção 80/20 e no modelo foram usadas 9 variáveis. Onde 6 destas variáveis estão relacionadas ao público-alvo (pop25a34, pop35a49, domiciliosA1, domiciliosA2, domiciliosB1 e domiciliosB2) e as demais são Renda Média, Faturamento e Potencial que é a variável preditora).

Com o modelo de classificação redes neurais H2O chegamos em uma acurácia de ~96.8%. Isso quer dizer que a cada 100 previsões feitas pelo modelo, aproximadamente 97 delas estarão certas. Para observar podemos ver a matriz de confusão onde tivemos apenas um erro.

```
previsoes
  2  3  4
2  9  1  0
3  0 12  0
4  0  0 10
```

Matriz de Confusão R Studio, 2022.

O que nos indica que o modelo obteve um bom resultado.

Pontuações

Podemos perceber que com os resultados dos modelos podemos confiar nos dados estimados de Renda Média, Faturamento e Potencial.

O Modelo de Renda Média pode ser aplicado com dados de outras cidades se houver ausência desses dados. No Modelo de Faturamento foi usado todas as variáveis numéricas, porém no Modelo de Potencial usando todas as variáveis numéricas não obteve um bom resultado. Quando foi treinado somente com as variáveis relacionadas ao público-alvo o modelo trouxe um resultado melhor.

Os modelos ainda possuem margem de melhora. Se tivermos dados mais refinados, mais detalhados e usando a base real de São Paulo em conjunto com a do Rio de Janeiro, iremos ter uma base maior e mais dados para os modelos treinarem, o que poderá nos trazer resultados ainda melhores.

Etapa Segmentando Público-Alvo

Nesta etapa vamos segmentar os bairros de São Paulo de acordo com a renda e a idade, e indicar aqueles com maior aderência ao público alvo. Assim como nas outras etapas, vamos usar a linguagem R para criar as análises e chegar no resultado.

Para segmentar por público-alvo é uma tarefa relativamente simples. Primeiro precisamos saber qual a fatia(percentual) que a população e a classe de domicílios correspondentes ao público-alvo tem em relação ao total.

Ou seja devemos fazer:

$$[\text{Índice Público Alvo População}] = [\text{População de 25 a 50 anos}] / [\text{Total População}]$$

$$[\text{Índice Público Alvo Domicílios}] = [\text{Domicílios A1, A2, B1 e B2}] / [\text{Total Domicílios}]$$

Com isso temos o segmento da população e classe de domicílios em relação ao público-alvo. Agora para obter o índice do segmento do público-alvo precisamos relacionar os dois índices:

$$[\text{Índice Segmento Público Alvo Geral}] =$$

$$([\text{Índice Público Alvo População}] + [\text{Índice Público Alvo Domicílios}]) / 2$$

Pronto. Já temos o nosso índice de segmento do público-alvo. Esse índice vai variar de 0 a 1 e quanto mais próximo este índice estiver de 1, mais aderente ao público o bairro é.

| | codigo | cidade | nome | indice_publico_alvo |
|----|-----------|-----------|------------------------|---------------------|
| 1 | 35503096 | Sao Paulo | Parque Anhembi | 0.70 |
| 2 | 35503086 | Sao Paulo | Pompeia | 0.63 |
| 3 | 355030149 | Sao Paulo | Vila Leopoldina | 0.63 |
| 4 | 35503046 | Sao Paulo | Chacara Klabin | 0.61 |
| 5 | 355030133 | Sao Paulo | Vila Olimpia | 0.61 |
| 6 | 355030136 | Sao Paulo | Berrini - Vila Funchal | 0.61 |
| 7 | 35503024 | Sao Paulo | Masp | 0.60 |
| 8 | 35503044 | Sao Paulo | Vila Clementino | 0.60 |
| 9 | 35503049 | Sao Paulo | Paraiso | 0.60 |
| 10 | 35503081 | Sao Paulo | Perdizes | 0.60 |

OS 10 mais aderentes - R Studio, 2022.

Com este Índice podemos ver, por exemplo, os 10 bairros de São Paulo mais aderentes ao público-alvo.

Conclusões

Com esse estudo podemos ver que São Paulo tem um grande potencial de sucesso e que a cidade tem bairros promissores para abrir novas filiais. Com esse estudo agora temos uma estimativa de quanto essas filiais iriam faturar e o potencial de cada uma. Sem falar que agora sabemos quais desses bairros têm mais aderências com o público.

Apesar disso, esse estudo ainda possui muita margem para melhorar, principalmente em relação em ter mais detalhes na base de dados. Como por exemplo a população total de 50 anos, a quantidade média de pessoas e a renda média por classes de domicílios.

O que podemos trazer de novas etapas é testar os índices de aderência do público que foi calculado nos modelos para verificar se a performance deles podem melhorar e também pesquisar novas fontes de dados que podem agregar para as análises.