Desafio Técnico

Ricardo Yamamoto Abe

1 de Junho de 2017

1 Introdução

Este relatório é o resultado da análise dos dados de faturamento e potencial de bairros do Rio de Janeiro, criação de modelos preditivos e suas aplicações para fins de classificação e regressão para bairros de São Paulo, além da segmentação por idade e classe social relacionados ao público alvo.

2 Pré-processamento

Antes de carregar o arquivo csv no sistema, o cabeçalho foi alterado para remover caracteres acentuados e cedilhas.

2.1 Missing values

Na primeira etapa de pré-processamento, foram detectados 2 problemas:

- Os bairros Eta Guaraú, Pico do Jaraguá e Reserva da Cantareira estavam com todos os totais de população por faixa etária e número de domicílios por classe zerados. Eles foram removidos da base.
- Seis bairros do RJ não possuiam renda média: Anil, Catumbi, Freguesia, Jacaré, Rio Comprido e Maracanã. Essa coluna foi preenchida via *MICE (Multiple Imputation by Chained Equations)*.

2.2 Normalização de faixas etárias e total de domicílios

Com o intuito de transformar as variáveis de faixa etária e domicílios em números reais no intervalo $[0, 1]$, as colunas `popAte9`, `popDe10a14`, `popDe15a19`, `popDe20a24`, `popDe25a34`, `popDe35a49`, `popDe50a59` e `popMaisDe60` foram divididas pela população do bairro, e as colunas `domiciliosA1`, `domiciliosA2`, `domiciliosB1`, `domiciliosB2`, `domiciliosC1`, `domiciliosC2`, `domiciliosD` e `domiciliosE` foram divididas pelo total de domicílios do bairro.

2.3 Análise de *outliers*

Foram encontrados 2 bairros do Rio de Janeiro com dados identificados como *outliers* e que foram removidos da base:

- Campo Grande – o *dataset* indica população de 667 mil habitantes, mais que o dobro do segundo bairro mais populoso. Em 2010, o total era de aproximadamente 328 mil pessoas.
- Lagoa – a renda média é superior a 63 mil Reais, praticamente o triplo do segundo bairro nesse quesito.

2.4 Normalização para renda média e populaçào

Para a normalização das colunas `rendaMedia` e `populacao`, foram testadas duas funções: logaritmo e *standardization* ($\frac{X-\mu}{\sigma}$, basicamente deixando a variável com média 0 e desvio padrão 1).

A normalização via logratimo obteve melhor resultado na classificação do potencial do bairro, enquanto que *standardization* foi superior para a regressão do faturamento.

2.5 Redução da dimensionalidade

Como ferramenta de redução de dimensionalidade, foi utilizada PCA (Análise de Componentes Principais) na classificação do potencial do bairro. Para regressão, optou-se por não utilizá-la.

Foram feitos alguns testes utilizando t-SNE (*t-Distributed Stochastic Neighbor Embedding*), sem sucesso.

3 Predição para São Paulo

A estimativa de faturamento e potencial para cada bairro de São Paulo encontram-se no arquivo `sp.csv`.

3.1 Classificação do potencial do bairro

O classificador utilizado para o potencial do bairro foi o XGBoost, utilizando PCA e normalização da renda média e população por logaritmo; sua acurácia na validação cruzada com 10 iterações foi de 86,61%.

3.2 Regressão para faturamento do bairro

Para regressão do faturamento, foi utilizada a métrica R2, que pode ser negativo para modelos arbitrariamente ruins, até 1.0, para um modelo perfeito. Novamente foi utilizado XGBoost com renda média e população normalizada via *standardization*. O *score* obtido via validação cruzada de 10 iterações foi 0,8837.

4 Segmentação

Assumindo que o total de domicílios por classe econômica e a população por faixa etária são variáveis independentes, podemos utilizar a proporção de cada uma delas para calcular a probabilidade conjunta via produto das marginais para cada um dos bairros. A segmentação foi feita sobre as classes e idades

do público alvo: 25 a 34 e 34 a 49 para idade; A1, A2, B1 e B2 para classes econômicas.

O algoritmo utilizado para encontrar os clusters foi o *KMeans*, utilizando $k = 9$, pois o processamento foi feito utilizando 2 proporções de idade e 4 de domicílio, totalizando 8 pares distintos; uma classe a mais foi gerada para agrupar os bairros menos interessantes do ponto de vista da pesquisa.

Foi feita uma tentativa de clusterização via HDBSCAN, mas os resultados não foram bons.

Na figura 1, encontramos os seguintes boxplots:

(a) População de 25 a 34 anos, classe A1

(b) População de 25 a 34 anos, classe A2

(c) População de 25 a 34 anos, classe B1

(d) População de 25 a 34 anos, classe B2

(e) População de 34 a 49 anos, classe A1

(f) População de 34 a 49 anos, classe A2

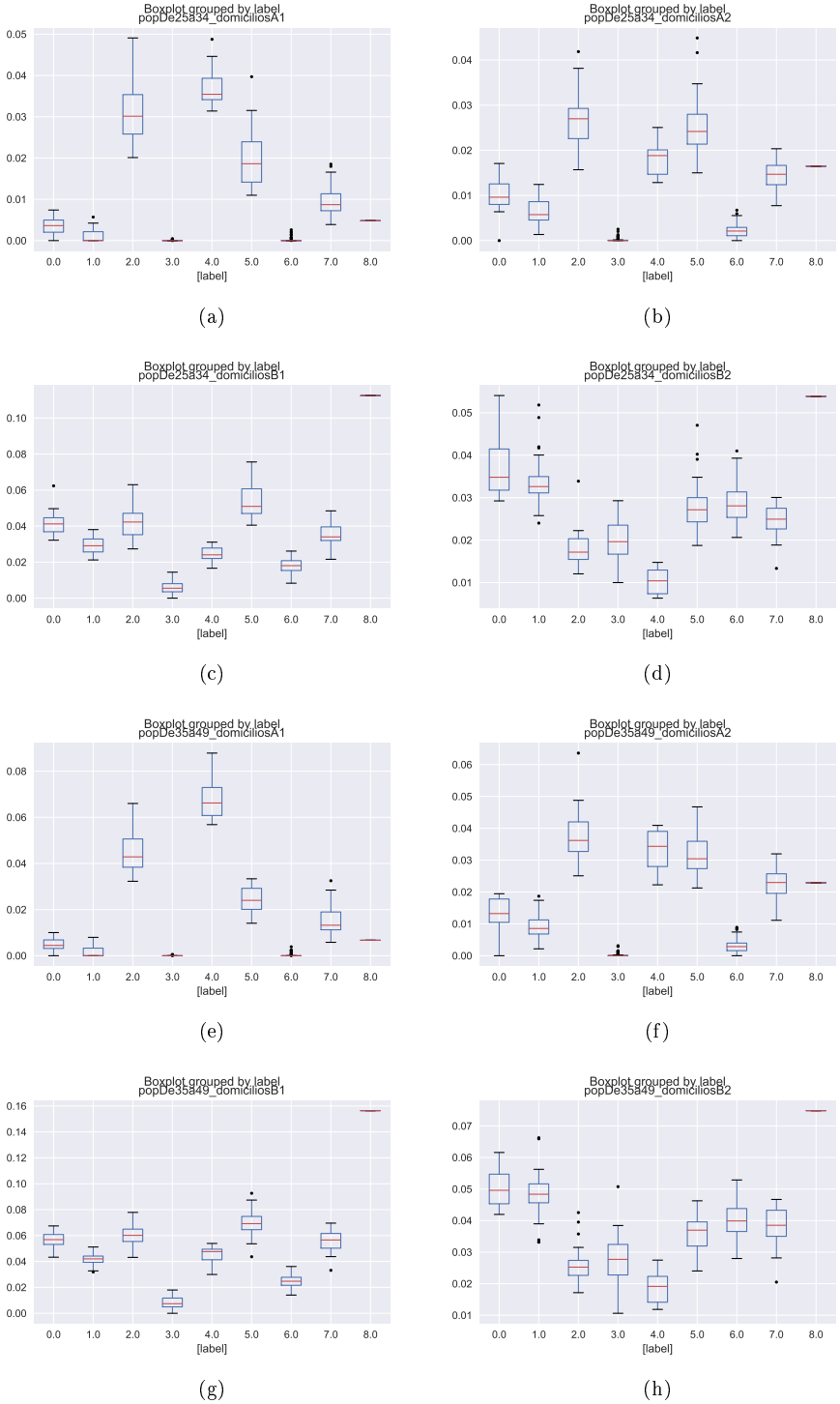
(g) População de 34 a 49 anos, classe B1

(h) População de 34 a 49 anos, classe B2

Analisando os boxplots, encontramos a seguinte segmentação:

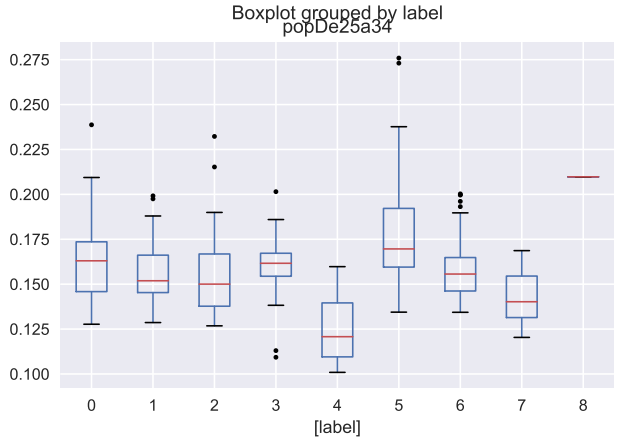
- Cluster 0 – representatividade na classe B2, para ambas as faixas de idade.
- Cluster 1 – representatividade na classe B2 (menor que no cluster 0), para ambas as faixas de idade.
- Cluster 2 – representatividade nas classes A1 e A2, para ambas as faixas de idade.
- Cluster 3 – é o cluster onde os bairros menos interessantes para investimento.
- Cluster 4 – é o cluster com maior proporção de domicílios na classe A1, para ambas as faixas de idade, mas pela análise da marginal popDe25a34 (figura 2), abaixo de todas os outros clusters, parece ser mais aderente à faixa de idade 35-49.
- Cluster 5 – representatividade nas classes B1 e A2, para ambas as faixas de idade. Também é relevante para a classe A1 em ambas as faixas de idade.
- Cluster 6 – representatividade em B2, para ambas as faixas de idade.
- Cluster 7 – representação equilibrada entre todas as classes econômicas e faixas de idade.
- Cluster 8 – o bairro do Parque Anhembi, tem uma proporção acima do esperado nas classes B1 e B2, para ambas as faixas de idade.

Os arquivos com os bairros separdos por segmentos são: `segment_00.csv`, `segment_01.csv`, ..., `segment_08.csv`.

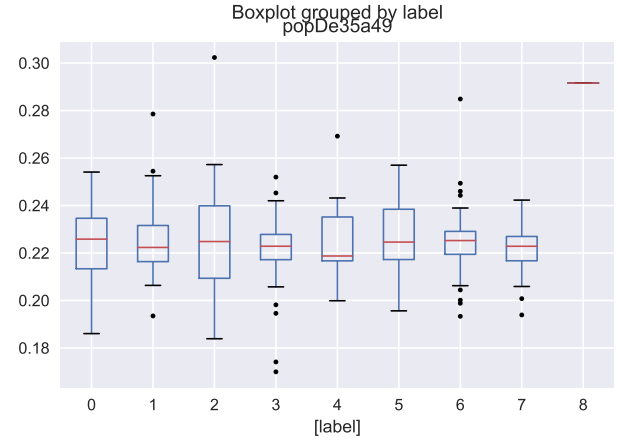


(a) (b) (c) (d) (e) (f) (g) (h)

Figura 1: Boxplots agrupados por label de cada cluster



(a)



(b)

Figura 2: Boxplots das faixas etárias 25 a 34 e 35 a 49 para clusters de São Paulo

5 Outras bases de dados

Para este projeto, imagino que os dados de pesquisa origem-destino do metrô de SP (feito em 2017, dado público) e do plano diretor de transporte urbano do RJ (2002/2003, dado público, mas aparentemente faltam as planilhas com mais detalhes dos dados) serviriam para indicar não só onde as pessoas moram, mas também quais seus outros pontos de interesse: trabalho, estudo e lazer. Assim sendo, uma pessoa do público alvo poderia ser observada não apenas em sua moradia, mas em locais prováveis para onde ela vai.

A API do Google Places (dado privado) e a base da wikimapia (dado público) poderiam ser utilizados para indicar estabelecimentos comerciais já existentes para analisar prováveis concorrentes em um determinado bairro.