Análise demográfica de bairros de São Paulo

Lucas Souza de Oliveira

1. PROBLEMA DE NEGÓCIO

A empresa X, situada no Rio de Janeiro, deseja abrir filiais na cidade de São Paulo. A empresa deseja estimar seu faturamento, classificar o potencial e segmentar os bairros de São Paulo de acordo com o perfil de renda e idade, para direcionamento de campanhas de marketing, utilizando como base os dados já conhecidos da cidade de Rio de Janeiro.

Os objetivos desse projeto são:

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

2. ESTRATÉGIA DE SOLUÇÃO

Para elucidar a estratégia adotada para alcançar os objetivos, esse relatório será organizado pelas seguintes etapas:

- Análise Exploratória dos Dados;
- Preparação dos Dados;
- Modelos de Machine Learning
- Tradução e Interpretação dos Erros
- Resultados do Modelo em Produção
- Próximos Passos

3. ANÁLISE EXPLORATÓRIA DE DADOS

A base de dados disponibilizada apresenta as colunas descritas abaixo na tabela 1.

Tabela 1 – Descrição das colunas da base de dados

Dicionário dos 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

3.1. Preenchimento de Valores Vazios

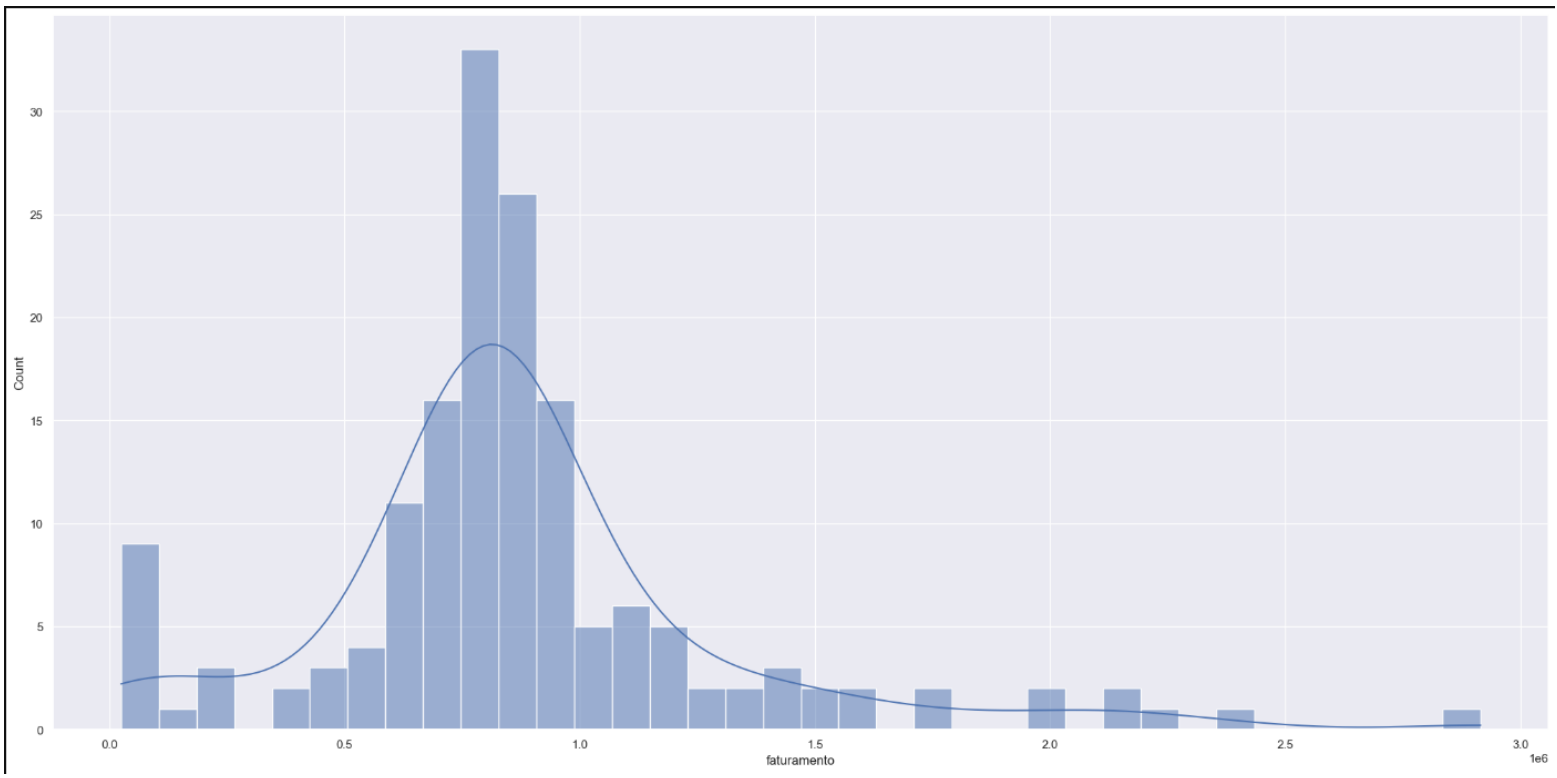
Algumas linhas da base de dados apresentavam informações faltantes. Para contornar esse problema, foi utilizado um algoritmo de preenchimento de valores vazios k-Nearest Neighbors. Este imputador utiliza o método k-Nearest Neighbors para substituir os valores ausentes nos conjuntos de dados com o valor médio dos vizinhos mais próximos encontrados no conjunto de treinamento. Por padrão, ele usa uma métrica de distância euclidiana para imputar os valores ausentes.

3.2. Análise Univariada

Nesse tópico será abordada as análises gráficas das variáveis presentes no modelo e suas implicações futuras.

O gráfico de distribuição da variável “faturamento” está na figura 1 abaixo:

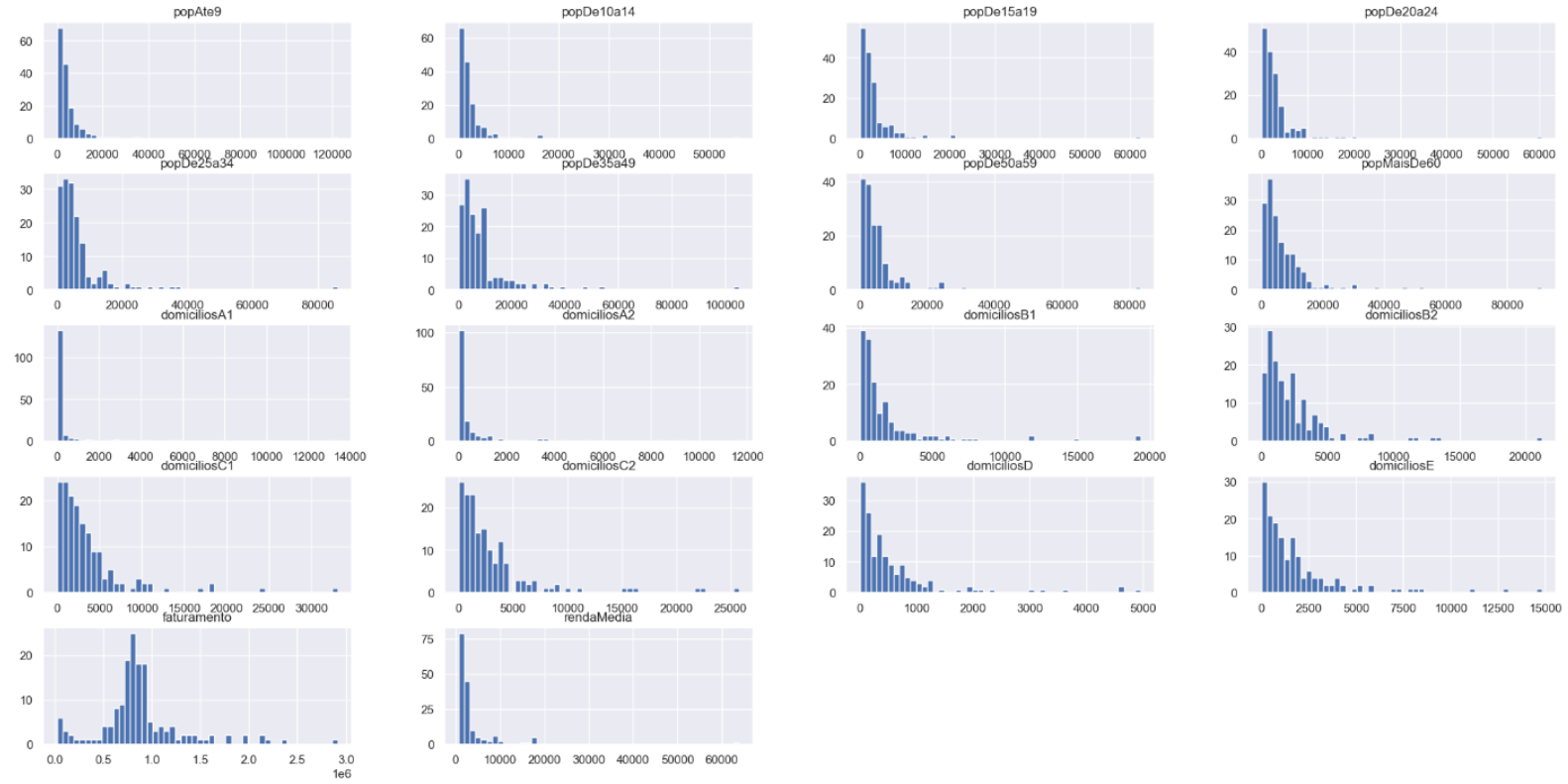
Figura 1 – Distribuição da variável “faturamento”



Nessa imagem, foi observada que a variável alvo “faturamento” possui uma distribuição gaussiana que colabora com o funcionamento de algoritmos de predição.

O gráfico de distribuição de demais variáveis numéricas está na figura 2 abaixo:

Figura 2 – Distribuição das variáveis numéricas auxiliares



Na figura 2, foi observada que as variáveis numéricas apresentam bastante outliers (valores distantes da concentração maior de dados). Há, também, uma diferença na escala dos valores que será tratada em tópicos futuros.

A análise estatística descritiva de parte das variáveis numéricas está apresentada na tabela 2 abaixo.

Tabela 2 – Descrição estatística das colunas da base de dados

	codigo	população	popAte9	popDe10a14	popDe15a19	popDe20a24	popDe25a34	popDe35a49	popDe50a59	popMaisDe60	domiciliosA1
count	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00
mean	3304557080.50	42654.35	5329.06	2484.44	3272.16	3349.38	6584.48	8826.86	5332.26	7475.71	388.48
std	46.33	64262.95	10816.20	4948.83	5752.16	5601.89	9049.90	11536.28	7986.26	10258.31	1403.61
min	3304557001.00	173.00	33.00	13.00	22.00	17.00	28.00	38.00	10.00	12.00	0.00
25%	3304557040.75	13362.75	1512.75	690.75	945.00	994.75	2131.25	2900.75	1598.75	2118.25	0.00
50%	3304557080.50	26076.00	3111.50	1434.00	2015.00	2142.50	4170.50	5657.00	3371.50	4575.50	0.00
75%	3304557120.25	46504.50	5622.25	2621.50	3448.75	3451.75	7145.25	9586.75	6333.25	9196.00	1.00
max	3304557160.00	667603.00	122578.00	55935.00	62342.00	60567.00	86116.00	105316.00	83341.00	91408.00	13408.00
skew	0.00	6.43	8.46	8.39	7.40	7.25	5.16	4.69	6.50	4.73	6.22
kurt	-1.20	56.99	88.00	86.77	71.04	69.29	38.47	32.10	57.99	30.97	49.00

Na tabela 2, pode-se observar que as variáveis possuem diferenças de escala de valores que serão tratados nos tópicos futuros.

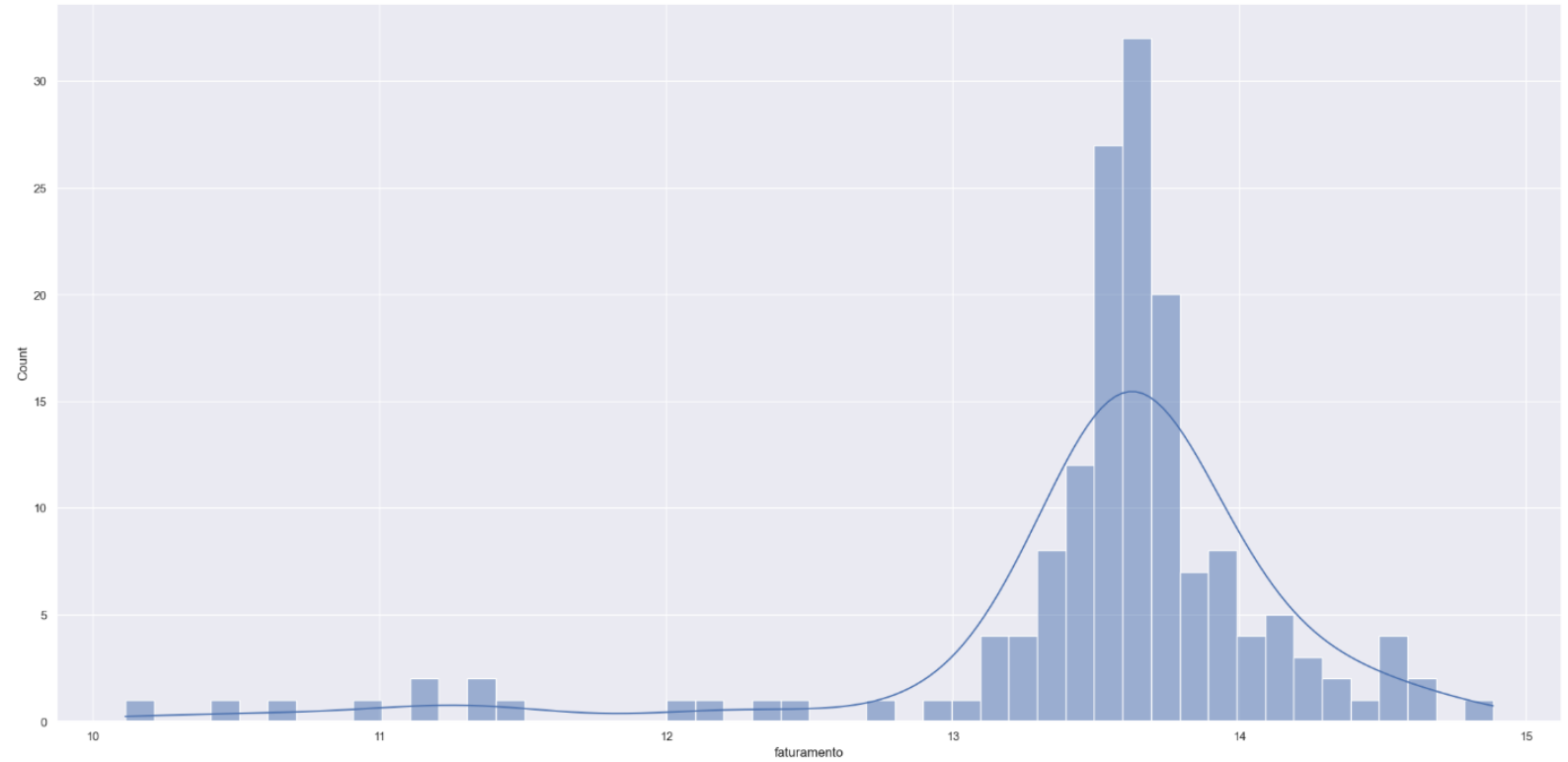
Algumas colunas foram tiradas da análise. As colunas “código”, “nome”, “cidade” e “estado” foram retiradas por não apresentarem informações relevantes para a análise e a coluna “população” foi retirada por apresentar redundância, já que ela é o valor da soma das colunas de faixa etária.

4. PREPARAÇÃO DOS DADOS

Nesse tópico, será mostrado algumas transformações estatísticas necessárias para prosseguir com a estratégia de solução.

Como mostrado na figura 1, a variável resposta “faturamento” possui uma distribuição próxima de uma normal. Foi realizada apenas uma transformação logarítmica para deslocamento de posição no eixo x, como mostrado na figura 3 abaixo.

Figura 3 – Distribuição da variável “faturamento” com transformação logarítmica



Para contornar os problemas das demais variáveis numéricas descritas no tópico 3, foram utilizadas as transformações e reescalas:

- RobustScaler;
- Ordinal Encoding.

RobustScaler é uma reescala utilizada com o objetivo de padronizar a escala das variáveis analisadas e diminuir a influência negativa de outliers.

Ordinal Encoding é uma transformação utilizada em variáveis categóricas. Nesse caso, foi utilizada para transformar as informações na coluna “potencial” para variáveis numéricas. Atribuiu-se “0” para potenciais “Baixo”, “1” para potenciais “Médio” e “2” para potenciais “Alto”.

5. MODELOS DE MACHINE LEARNING

Nesse tópico, será mostrado o resultado de diferentes algoritmos de Machine Learning utilizados para solucionar o problema de negócio. Esse tópico será subdividido em:

- “Modelos de Regressão”, utilizados na tarefa de predição de faturamento;
- “Modelos de Classificação”, utilizados na tarefa de classificação de potenciais;
- “Modelos de Clusterização”, utilizados na tarefa de segmentação demográfica;

Para análise de resultados, a base de dados de São Paulo foi dividida de 5 maneiras diferentes entre “banco de dados de teste” e “banco de dados de treino”. Base de “treino” são utilizadas para treinar os modelos de Machine Learning e gerar um resultado resposta, já a base de “teste” é utilizada para comparar a resposta com o resultado real. Todos os bairros foram, em algum momento durante as 5 divisões da base de dados, utilizadas como “treino” e “teste”.

5.1. Modelos de Regressão

Os modelos de Regressão testados foram:

- Lasso;
- Random Forest Regressor;
- XGBoost Regressor.

Para avaliar suas performances foram utilizadas as métricas:

- MAE (Mean Absolute Error): Erro absoluto médio do faturamento predito;
- MAPE (Mean Aboslute Percentage Error): Erro percentual médio do faturamento predito;
- RMSE (Root Mean Square Error): Erro quadrático médio. Métrica para avaliar outliers.

Os resultados observados estão na tabela 3 abaixo.

Tabela 3 – Resultados dos modelos de regressão

Model Name	MAE CV	MAPE CV	RMSE CV
XGBoost Regressor	66898.68 +/- 15593.44	0.11 +/- 0.01	134477.21 +/- 51754.76
Random Forest	82359.89 +/- 21250.11	0.14 +/- 0.04	167223.38 +/- 58764.73
Linear Regression - Lasso	154200.01 +/- 72037.36	0.45 +/- 0.57	417868.76 +/- 288248.72

Na tabela 3, pode-se observar que o algoritmo com melhor desempenho foi o XGBoost, com um erro percentual médio de 11% ± 1. Esse foi o algoritmo escolhido para a produção das soluções requisitadas.

5.2. Modelos de Classificação

Os modelos de Classificação testados foram:

- KNN;
- Support Vector Machines;
- Naive Bayes Classifier;
- Random Forest Classifier.

Para avaliar suas performances foi utilizada a métrica de acurácia. Os resultados estão na tabela 4 abaixo.

Tabela 4 – Resultados dos modelos de classificação

Model Name	Accuracy CV
Random Forest Classifier	87.5 +/- 3.42
k-Nearest Neighbors	81.25 +/- 6.56
Support Vector Machine	70.0 +/- 9.4
Gaussian Naive Bayes	61.88 +/- 7.5

Na tabela 4, pode-se observar que o algoritmo com melhor desempenho foi o Random Forest Classifier, com uma acurácia de 87,5 % ± 3,42. Esse foi o algoritmo escolhido para a produção das soluções requisitadas.

5.3. Modelos de Clusterização

A tarefa de segmentação foi separada em dois. Uma segmentação etária e outra social dos bairros da base de dados.

Para a segmentação, foi utilizada os seguintes algoritmos:

- UMAP (Uniform Manifold Approximation and Projection): Esse algoritmo utiliza técnicas de redimensionamento do espaço de variáveis. Muito utilizado quando há mais de 3 variáveis sendo analisadas e impossível de se observar graficamente.
- GMM (Gaussian Mixture Model): Algoritmo de clusterização que trabalha com distribuição probabilística de segmentação.

Para avaliar a performance dos modelos escolhidos, será necessária uma análise gráfica que será abordada no tópico seguinte.

6. TRADUÇÃO E INTERPRETAÇÃO DOS ERROS

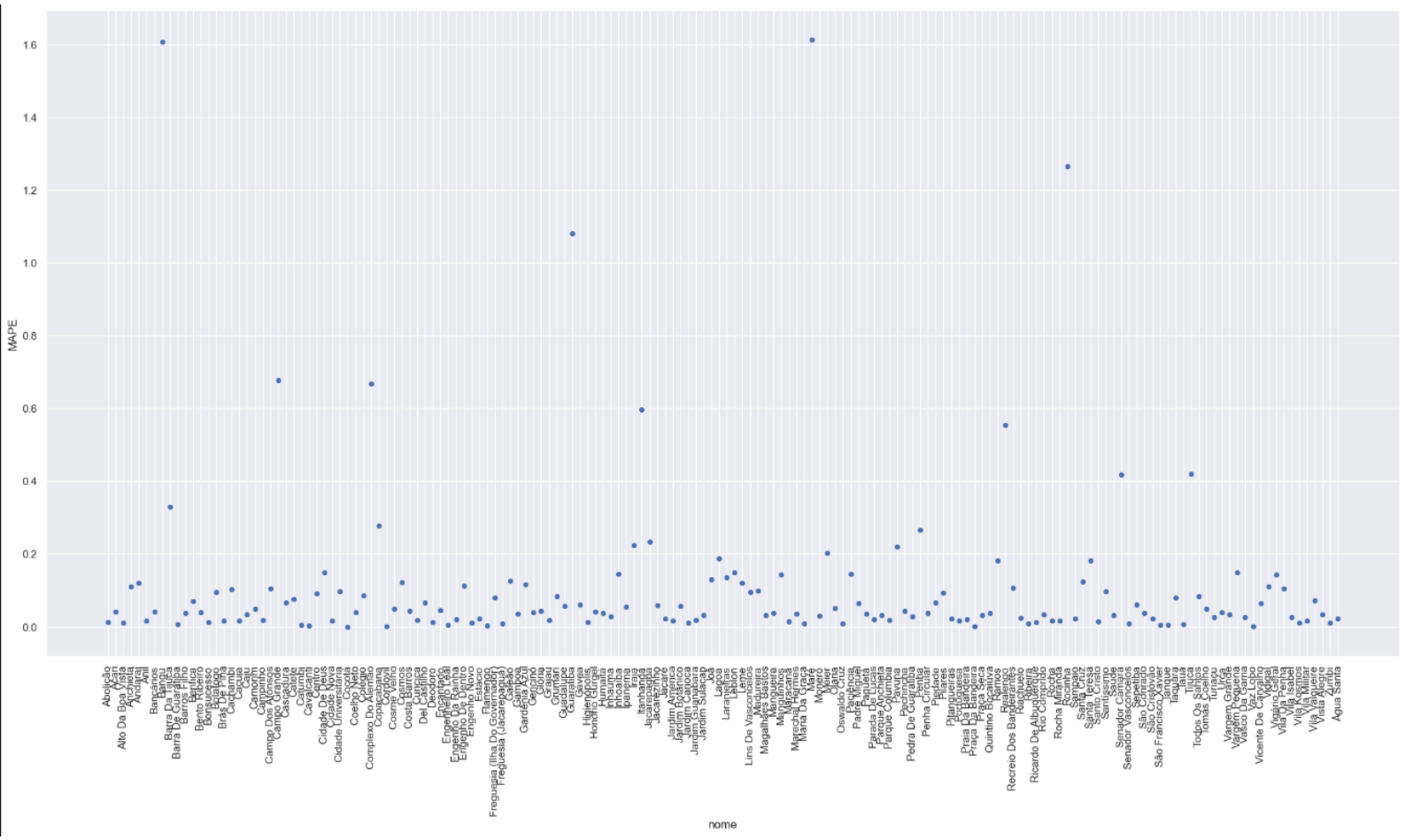
Esse tópico conterá informações visuais que buscam explicar a performance de cada solução proposta. Será subdividido em:

- Resultados de Regressão: Resultados obtidos na predição de faturamento.
- Resultados de Classificação: Resultados obtidos na classificação de potencial.
- Resultados de Clusterização: Resultados obtidos na segmentação dos bairros.

6.1. Resultados de Regressão

Na figura 4 abaixo está o gráfico de pontos de valor de MAPE do modelo treinado para cada bairro de São Paulo.

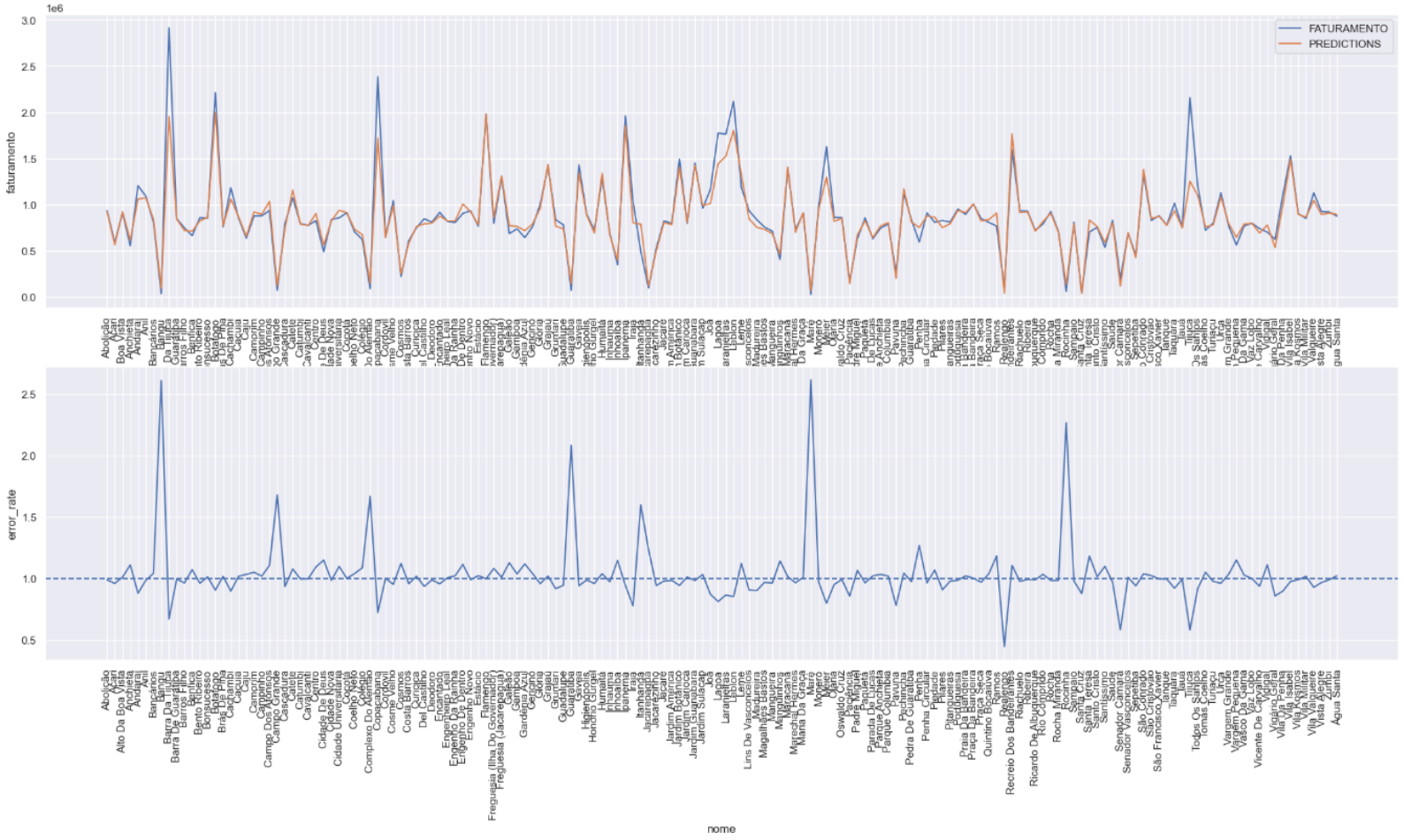
Figura 4 – Scatterplot do MAPE para cada bairro de SP



Nesse gráfico, pode-se observar que os MAPE dos bairros estão concentrados num valor menor que 0.2. Isso significa que o modelo apresentou um acerto de predição de faturamento de mais de 80 % na maioria dos bairros. Há de se observar, também, quatro bairros com um resultado abaixo de 50 %.

Na figura 5 abaixo, está o gráfico em linha relacionando o valor de faturamento real com o valor de faturamento predito para cada bairro.

Figura 5 – Gráfico de linha relacionando Faturamento com Predição

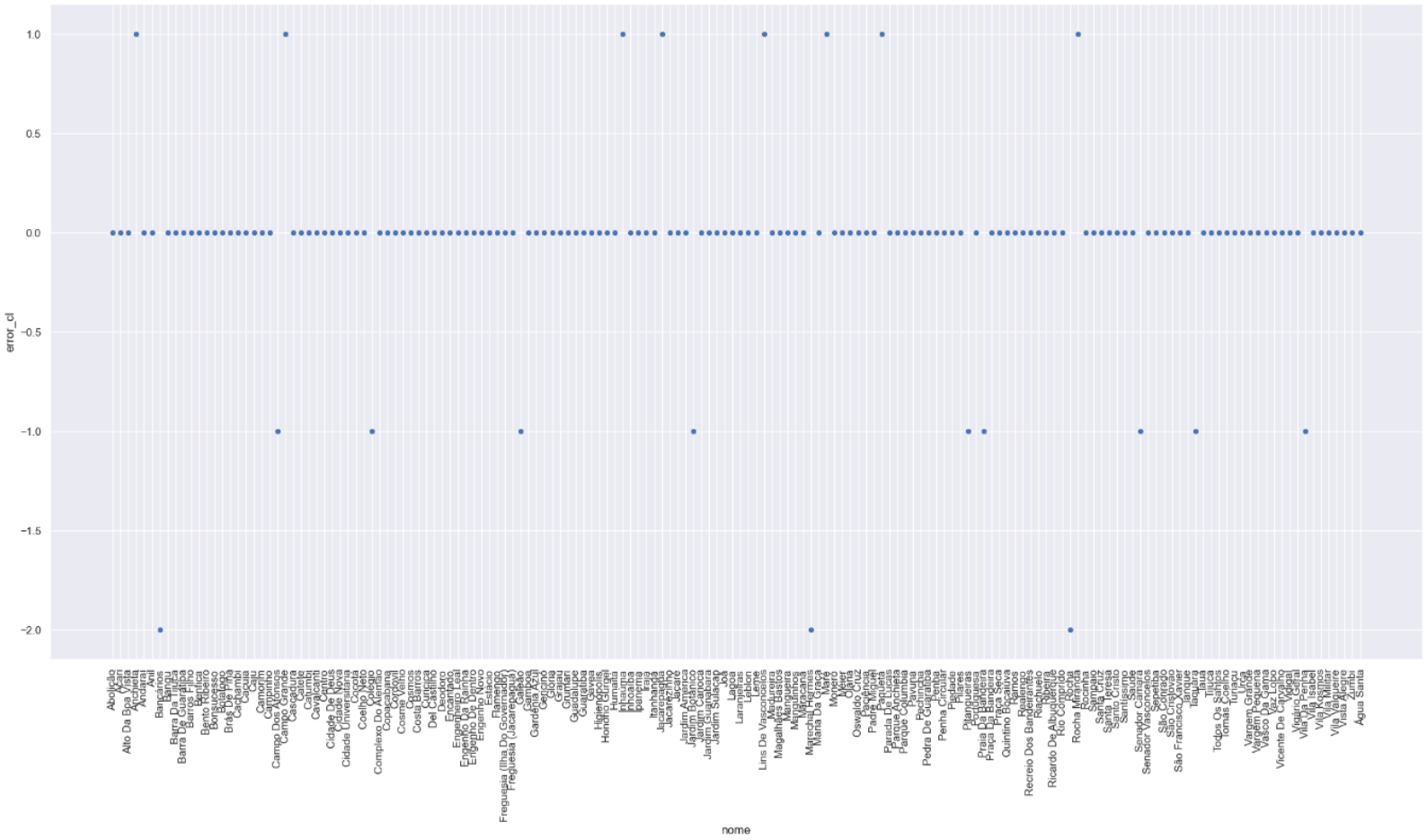


Na figura 5, pode-se observar que os valores das previsões e do faturamento seguem uma relação próxima, mostrando a efetividade do modelo utilizado. No gráfico de baixo da mesma figura, quando mais próximo de 1.0, melhor será o modelo. Novamente foi observado quatro bairros que apresentaram resultados anômalos.

6.2. Resultados de Classificação

Na figura 6 abaixo está o gráfico de erros de classificação de potencial para cada bairro de SP.

Figura **6 – Resultado gráfico do modelo de classificação**



Na figura 6, pode-se observar os erros e acertos do modelo de classificação em classificar os potenciais dos bairros. Há uma grande concentração de bairros com o erro em 0, mostrando a efetividade do modelo usado. Em valores positivos, o modelo previu um potencial maior do que o real. Em valores negativos, um potencial menor que o real.

6.3. Resultados de Clusteriação

Na figura 7 e 8 abaixo estão as distribuições multidimensionais geradas através do UMAP.

Figura 7 – Espaço gerado através do UMAP para faixa etária

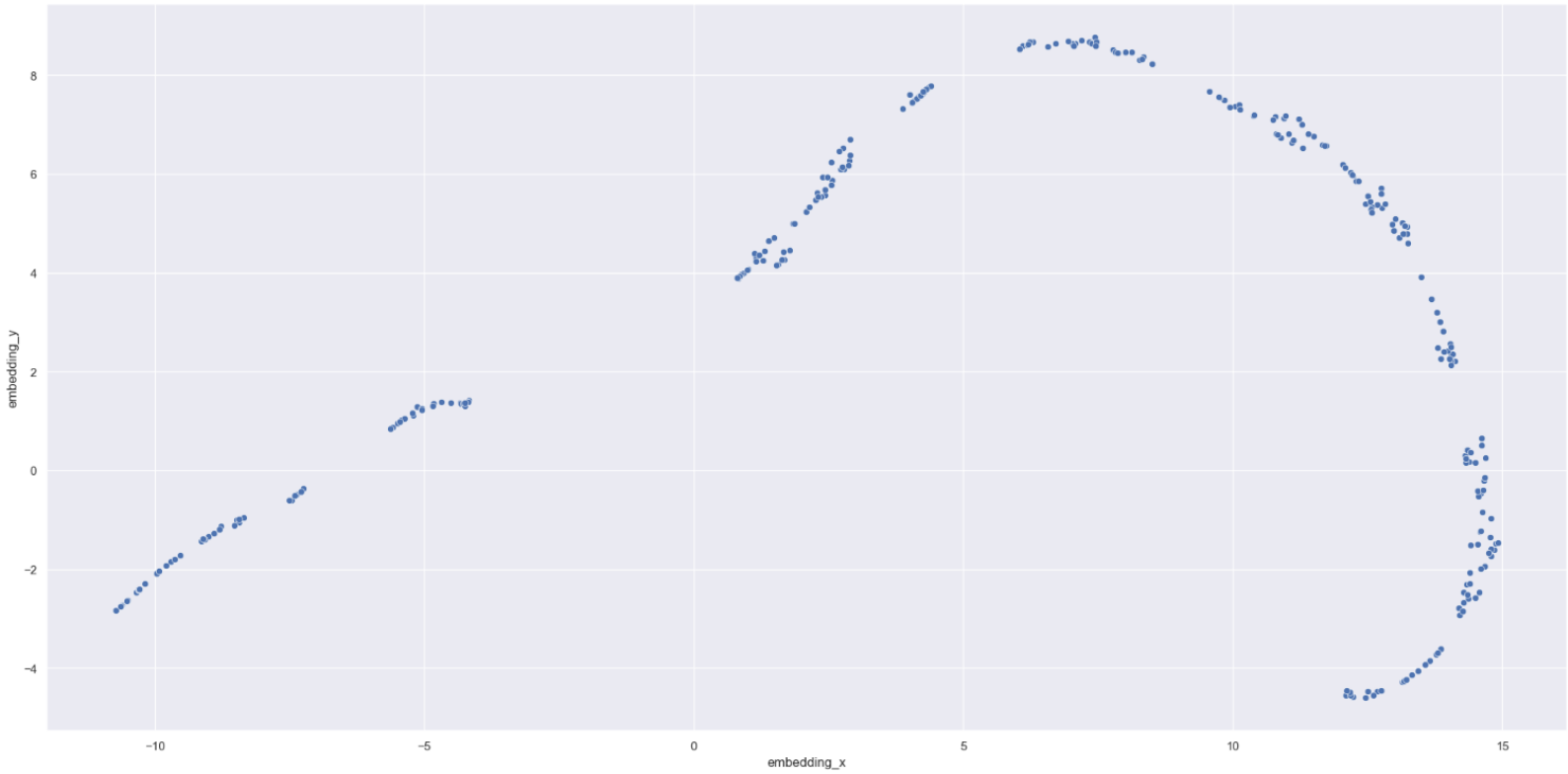
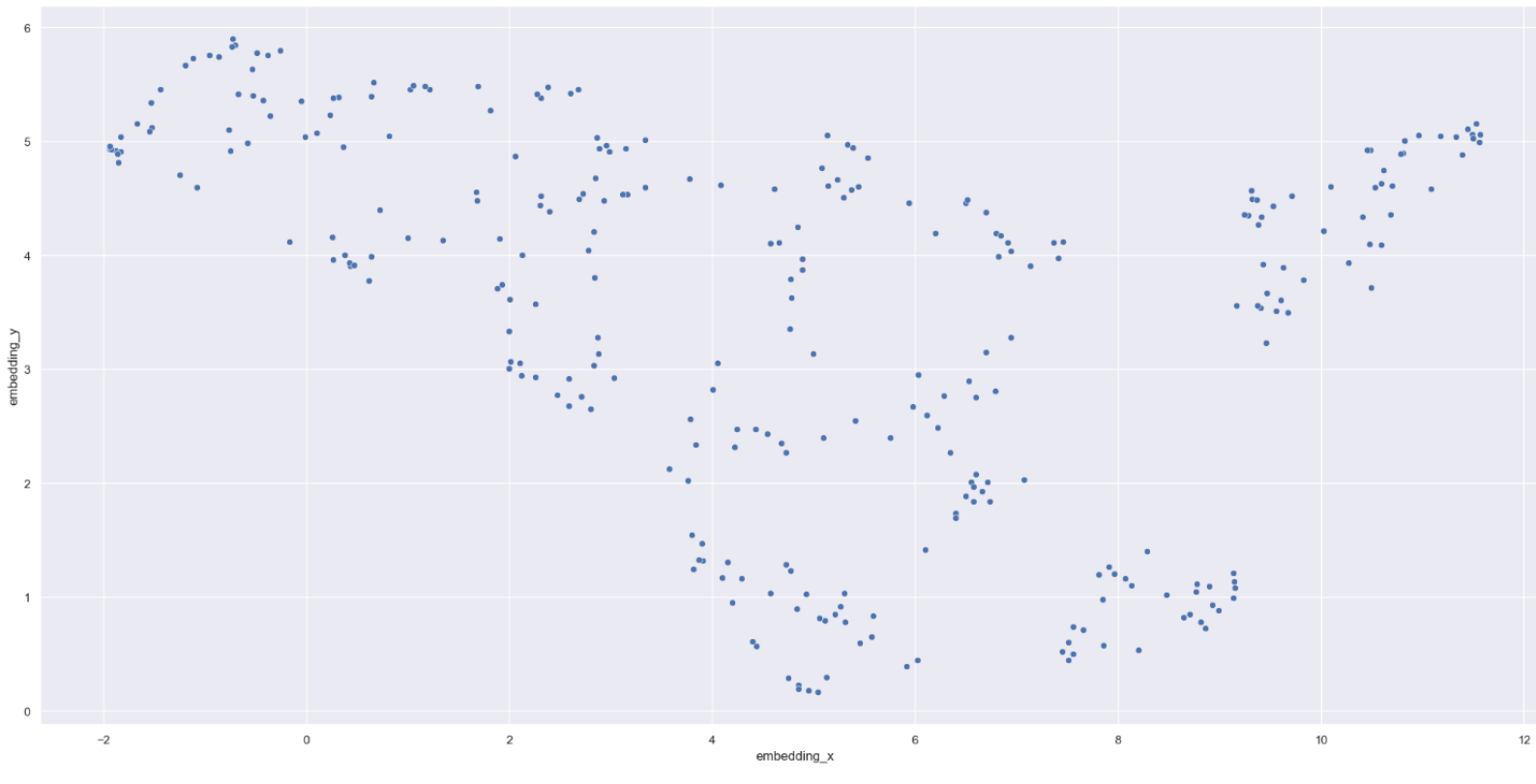


Figura 8 – Espaço gerado através do UMAP para classe socia



Em ambas figuras acima, pode-se observar a separação dos pontos em grupos próximos. A seguir, na figura 9 e 10, estão os resultados de segmentação com o algoritmo GMM aplicado.

Figura 9 – Resultado do modelo de clusterização para faixa etária

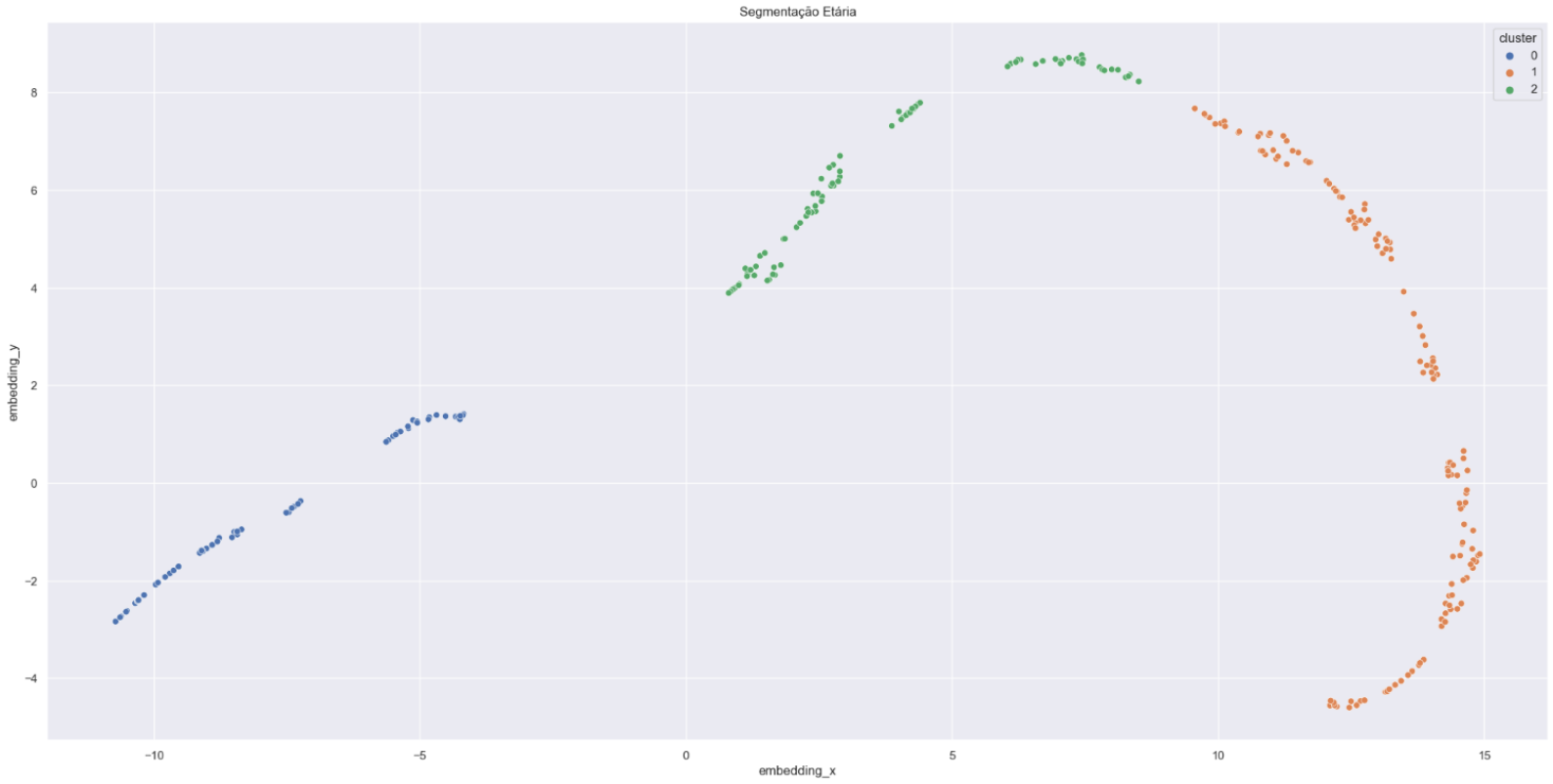
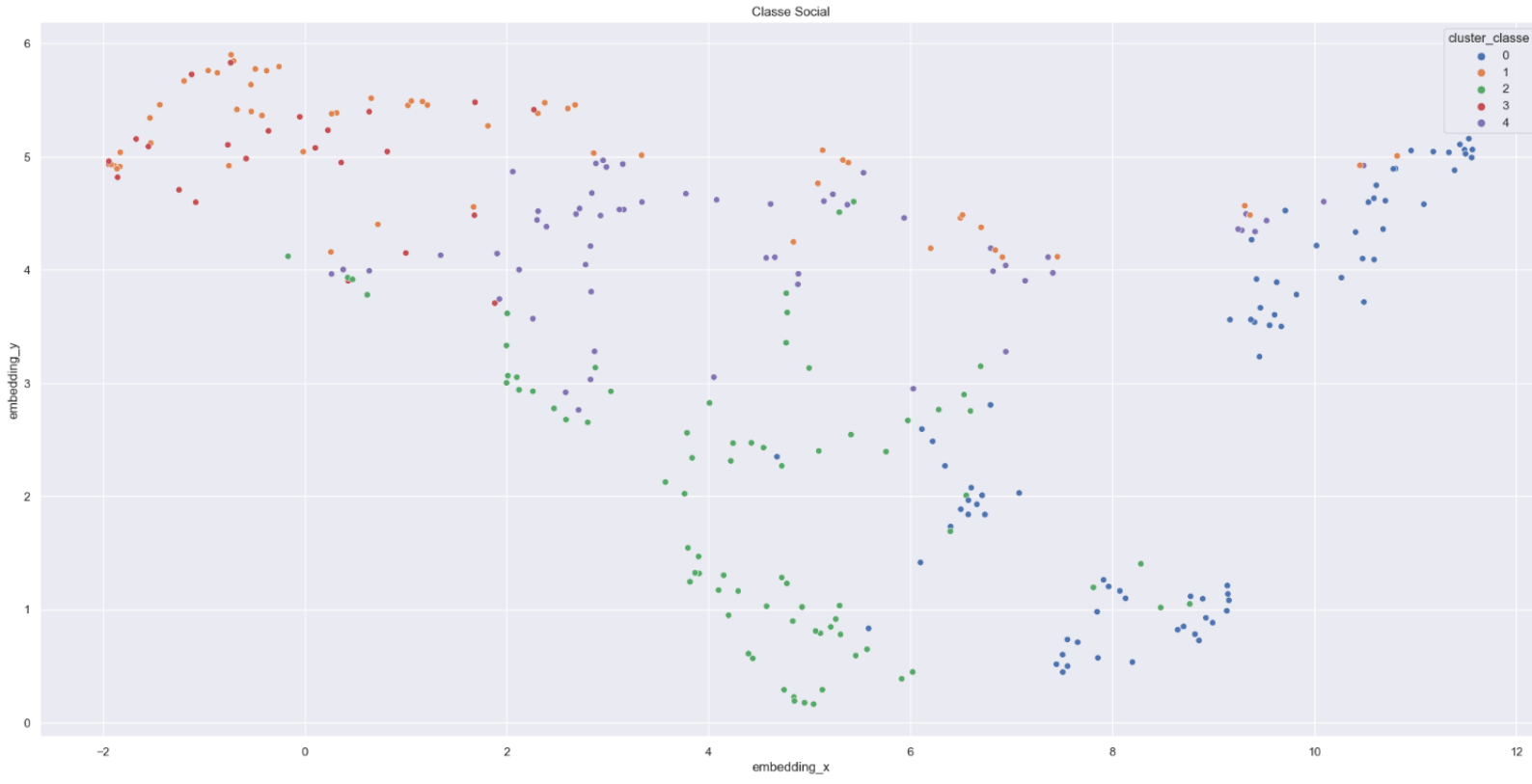


Figura 10 – Resultado do modelo de clusterização para classe social



Nas figuras acima, pode-se observar que o modelo para segmentação etária apresentou um ótimo resultado, com clusters (grupos) bem separados entre si. Para a segmentação social, o modelo não distinguiu com clareza o pertencimento de cluster para cada bairro. Numa eventual melhoria, será necessário repensar as variáveis utilizadas.

7. RESULTADOS DO MODELO EM PRODUÇÃO

O modelo apresentou resultados muito bons. Para predição de faturamento, possui um erro percentual médio por bairro (MAPE) de apenas 11 % ± 1. Para classificação de potenciais, uma acurácia de 87,5 % ± 3,42. E para segmentação, uma divisão satisfatória de faixa etária.

- Faturamento: Erro de 11 % ± 1.
- Potencial: Acurácia de 87,5 % ± 3,42.

8. PRÓXIMOS PASSOS

Para as próximas versões do modelo, recomenda-se uma análise aprofundada nos bairros que apresentaram resultados anômalos de predição de faturamento e classificação de potencial. A adição de novas variáveis no dataset podem melhorar sua performance. Algumas variáveis pensadas foram:

- Competidores próximos;
- Valor do m² no bairro;
- Fluxo de pessoas;