Criminalidade em Chicago

Antonio Gadelha, Rafael Albuquerque e Rodrigo Carneiro

December 11, 2017

Abstract

Esse documento tem por objetivo expor algumas análises feitas acerca dos índices de criminalidade do ano de 2016 da cidade de *Chicago, IL (EUA)* para tentar entender alguns aspectos desse ambiente. Foram feitas análises para entender melhor os crimes domésticos, o efeito da educação sobre a criminalidade e as características mais comuns em crimes que resultam em prisão para entender como melhorar essas taxas de criminalidade.

1 Introdução

A partir de conjuntos de dados sobre a criminalidade e educação da cidade de Chicago fizemos várias análises das características dos crimes na cidade, por exemplo em que dias e horários mais ocorrem e se o nível da educação ou a quantidade de escolas numa região é um fator que contribui para o aumento ou diminuição dos crimes. Analisamos também algumas dessas características nos crimes domésticos e por fim utilizamos uma árvore de decisão para predizer se, baseado em algumas características do crime, o infrator seria preso ou não.

2 Metodologia

2.1 Coleta de Dados

Os dados foram oletados do portal de dados abertos da cidade de Chicago(https://data.cityofchicago.org/), de onde foram retirados todos os datasets utilizados na análise. São eles:

- Crimes_16.csv - Dados de criminalidade da cidade de Chicago em 2016
- Schools_1617.csv - Dados sobre as escolas da cidade de Chicago no período 2016-2017
- CommAreas.csv - Lista das áreas comunitárias de Chicago
- shapeCH - ShapeFile da cidade de Chicago e suas áreas comunitárias

2.2 Pré-Processamento dos Dados

Os datasets continham muitas informações que seriam irrelevantes para as análises desejadas, então primeiramente selecionamos apenas as colunas com os dados que gostaríamos de analisar, fizemos também algumas conversões de dados e além disso fizemos algumas conversões de dados qualitativos para quantitativos para a fase de aprendizado. Por exemplo: Passamos a formatação dos horários dos crimes de 12h para 24h, atribuimos um número único para cada tipo de infração, dentre outras mudanças para facilitar as nossas análises.

2.3 Análise Exploratória

Iniciamos nossa análise classificando as variáveis do conjunto de dados:

```
schools_upd.dtypes

School_ID                int64
Zip                      int64
Student_Count_Total      int64
School_Latitude        float64
School_Longitude       float64
Overall_Rating          object
Area                     int64
dtype: object
```

(a)

```
crimes_upd.dtypes

Date                   object
Primary Type           object
Location Description   object
Arrest                   bool
Domestic                 bool
Community Area          int64
FBI Code               object
Location               object
dtype: object
```

(b)

Figure 1: (a) Tipos de dados das escolas (b) Tipos de dados dos crimes

onde int64 representa números inteiros, float64 representa números reais, bool representa dados que assumem valores Verdadeiro ou Falso e object está represetando outros tipos de dados, como texto por exemplo.

2.3.1 Análise geral dos crimes

As primeiras análises feitas foram análises mais gerais sobre o crime de Chicago, procurando econtrar padrões nas ocorrências de crimes. Com os crimes organizados em um dataframe, pudemos ver os horários com maior taxa de crimes.

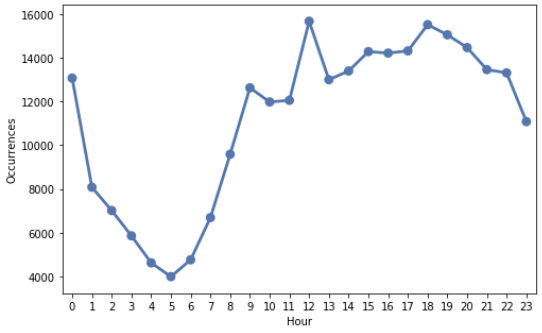

Figure 2: Ocorrência de crimes x hora

A maior quantidade de crimes se dá às 12 horas e às 18 horas, provavelmente por que essas são as horas em que muitas pessoas estão nas ruas para almoçarem e para irem dos seus trabalhos para casa, respectivamente. A grande queda que ocorre no horário da madrugada se deve a um menor número de pessoas circulando pela cidade.

Após isso fomos verificar se os horários de pico de crimes eram também os horários em que mais ocorriam prisões.

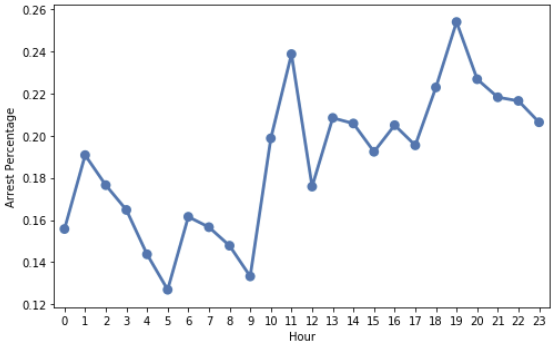

Figure 3: Ocorrência de prisões x hora

E percebemos que as horas que mais ocorrem prisões não são as horas em que os crimes são mais

altos. Por conta disso fomos verificar o que aconteciam nessas horas para que as prisões fossem tão altas em relação as outras horas e verificamos que furtos e agressões(Theft e battery) são os crimes que mais ocorrem durante todo o dia, mas às 11 horas é onde ocorre um número maior de crimes relacionados a narcóticos, o que pode justificar a quantidade de prisões nessa hora. E o grande número às 19 horas pode ser uma consequência do número de crimes ocorrido às 18 horas.

	Hour	1st Crime	1st Number	2nd Crime	2nd Number	3rd Crime	3rd Number
0	0	BATTERY	2486	THEFT	2341	CRIMINAL DAMAGE	1764
1	1	BATTERY	2264	THEFT	1445	CRIMINAL DAMAGE	1190
2	2	BATTERY	2087	THEFT	1226	CRIMINAL DAMAGE	1072
3	3	BATTERY	1696	THEFT	957	CRIMINAL DAMAGE	924
4	4	BATTERY	1300	THEFT	778	CRIMINAL DAMAGE	778
5	5	BATTERY	923	THEFT	739	CRIMINAL DAMAGE	618
6	6	THEFT	989	BATTERY	876	CRIMINAL DAMAGE	711
7	7	THEFT	1442	BATTERY	1125	CRIMINAL DAMAGE	827
8	8	THEFT	2154	BATTERY	1538	CRIMINAL DAMAGE	1050
9	9	THEFT	2633	DECEPTIVE PRACTICE	2336	BATTERY	1773
10	10	THEFT	2696	BATTERY	1845	DECEPTIVE PRACTICE	1248
11	11	THEFT	2857	BATTERY	2092	NARCOTICS	1013
12	12	THEFT	4071	DECEPTIVE PRACTICE	2307	BATTERY	2267
13	13	THEFT	3602	BATTERY	2172	DECEPTIVE PRACTICE	1078
14	14	THEFT	3901	BATTERY	2221	CRIMINAL DAMAGE	1149
15	15	THEFT	3907	BATTERY	2517	ASSAULT	1295
16	16	THEFT	3753	BATTERY	2601	CRIMINAL DAMAGE	1425
17	17	THEFT	3951	BATTERY	2457	CRIMINAL DAMAGE	1600
18	18	THEFT	4036	BATTERY	2552	CRIMINAL DAMAGE	1842
19	19	THEFT	3519	BATTERY	2546	CRIMINAL DAMAGE	1748
20	20	THEFT	3144	BATTERY	2720	CRIMINAL DAMAGE	1915
21	21	BATTERY	2712	THEFT	2695	CRIMINAL DAMAGE	1925
22	22	BATTERY	2849	THEFT	2613	CRIMINAL DAMAGE	2030
23	23	BATTERY	2656	THEFT	2100	CRIMINAL DAMAGE	1614

Figure 4: Crimes mais comuns por hora

Após algumas análises a mais, vimos que era mais interessante explorar mais a fundo os crimes domésticos, pois em uma de nossas análises, após desconsiderar os crimes domésticos, percebemos uma mudança na tabela de crimes mais comuns por horário. Agressão deixava de figurar como o crime mais comum durante a madrugada.

	Hour	1st Crime	1st Number	2nd Crime	2nd Number	3rd Crime	3rd Number
0	0	THEFT	2235	DECEPTIVE PRACTICE	1676	CRIMINAL DAMAGE	1592
1	1	THEFT	1361	BATTERY	1130	CRIMINAL DAMAGE	1021
2	2	THEFT	1154	BATTERY	1023	CRIMINAL DAMAGE	927
3	3	THEFT	887	BATTERY	835	CRIMINAL DAMAGE	801
4	4	THEFT	720	CRIMINAL DAMAGE	664	BATTERY	602
5	5	THEFT	695	CRIMINAL DAMAGE	529	BATTERY	384
6	6	THEFT	931	CRIMINAL DAMAGE	625	BURGLARY	424
7	7	THEFT	1372	CRIMINAL DAMAGE	744	BURGLARY	698
8	8	THEFT	2063	CRIMINAL DAMAGE	922	DECEPTIVE PRACTICE	920
9	9	THEFT	2500	DECEPTIVE PRACTICE	2313	CRIMINAL DAMAGE	1002
10	10	THEFT	2582	DECEPTIVE PRACTICE	1237	CRIMINAL DAMAGE	954
11	11	THEFT	2762	BATTERY	1055	NARCOTICS	1013
12	12	THEFT	3946	DECEPTIVE PRACTICE	2278	BATTERY	1235
13	13	THEFT	3508	BATTERY	1232	DECEPTIVE PRACTICE	1074
14	14	THEFT	3803	BATTERY	1253	DECEPTIVE PRACTICE	1034
15	15	THEFT	3797	BATTERY	1609	CRIMINAL DAMAGE	1152
16	16	THEFT	3653	BATTERY	1525	CRIMINAL DAMAGE	1269
17	17	THEFT	3861	CRIMINAL DAMAGE	1460	BATTERY	1397
18	18	THEFT	3931	CRIMINAL DAMAGE	1695	BATTERY	1371
19	19	THEFT	3412	CRIMINAL DAMAGE	1602	BATTERY	1385
20	20	THEFT	3039	CRIMINAL DAMAGE	1770	BATTERY	1457
21	21	THEFT	2584	CRIMINAL DAMAGE	1748	BATTERY	1394
22	22	THEFT	2497	CRIMINAL DAMAGE	1842	BATTERY	1369
23	23	THEFT	2009	CRIMINAL DAMAGE	1430	BATTERY	1213

Figure 5: Crimes não domésticos mais comuns por hora

2.3.2 Análise da influência da educação sobre os crimes

Antes de explorar mais a fundo os crimes domésticos, procuramos ver a importância da educação nas taxas de criminalidade. A nossa hipótese inicial era de que regiões mais bem educadas teriam menos crimes. Ao fazer uma regressão entre o número de escolas da região e a taxa de crimes, percebeu-se o contrário. As regiões com mais escolas, também tinham mais crimes.

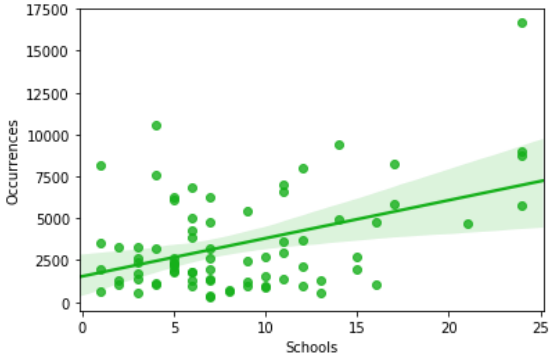

Figure 6: Ocorrência de crimes x número de escolas

Mas apesar de contradizer nosso pensamento inicial, essa conclusão faz sentido, pois lugares com mais escolas, supõe-se que têm mais moradores e um maior fluxo de pessoas, o que explica uma maior taxa de crime em comparação com lugares menos povoados ou com menor fluxo de pessoas.

Por outro lado, tínhamos disponível também dados de qualidade das escolas. Fizemos uma análise considerando a influência da quantidade de escolas de cada ranking para as taxas de crimes da região. E essa análise corroborou a primeira. Independente do ranking das escolas, a regressão levando em conta apenas a quantidade continuava dizendo que quanto mais escolas, mais crimes ocorrem na região.

Uma terceira abordagem deu um resultado diferente. Ao levar em consideração a proporção de escolas com nota 1+(Maior nota do ranking) em relação ao número de escolas totais, podemos entender quais os bairros tem a melhor educação. Não basta pensar apenas na quantidade pura de escolas nível 1+, mas deve-se levar em consideração o quanto do total de escolas da região

elas representam. Ao ajustar um modelo e fazer uma regressão levando em conta essa proporção notou-se que regiões com maior proporção de escolas boas possuem menos crimes.

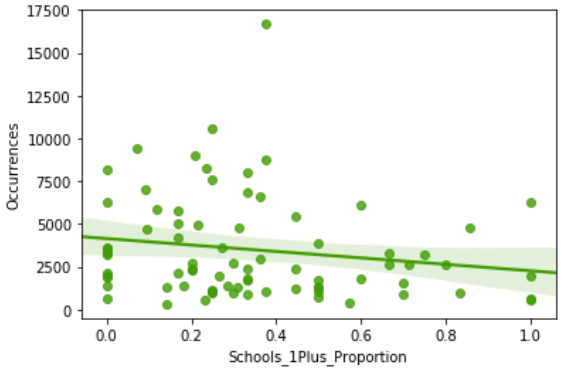

Figure 7: Ocorrência de crimes x número de escolas de nível 1+

2.3.3 Análise dos crimes domésticos

Começamos a análise dos crimes domésticos observando que a quantidade de crimes desse tipo tem um aumento ao final do dia.

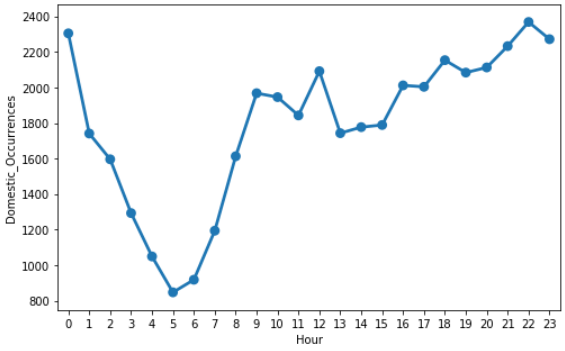

Figure 8: Crimes domésticos x hora

O que pode ser justificado por essa ser a hora em que as pessoas saem de seus trabalhos e vão normalmente para o ambiente familiar. Também observamos que esse tipo de crime é constante durante a semana e vai crescendo com o decorrer do fim de semana. Isso se deve ao fato de que no fim de semana a maioria das pessoas não trabalham, sendo que domingo é o dia que mais ocorre crimes domésticos pois é o dia em que menos pessoas passam o dia no trabalho, ja que há pessoas que trabalham aos sábados.

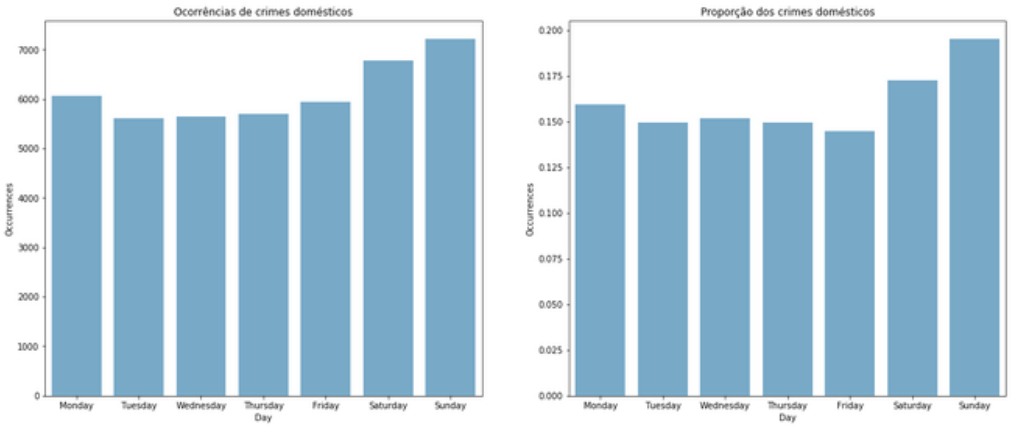

Figure 9: Crimes domésticos x dia da semana

Podemos assim deduzir que quanto mais tempo as pessoas passam perto de sua família, maior a chance de algum crime ocorrer nesse âmbito.

Ao realizar outra análise sobre os crimes mais comuns, vemos que em qualquer dia, a agressão é o crime que mais ocorre domesticamente.

	Day	1st Crime	1st Number	2nd Crime	2nd Number	3rd Crime	3rd Number
0	Friday	BATTERY	3298	OTHER OFFENSE	772	ASSAULT	618
1	Monday	BATTERY	3348	OTHER OFFENSE	837	ASSAULT	677
2	Saturday	BATTERY	4010	OTHER OFFENSE	756	ASSAULT	694
3	Sunday	BATTERY	4358	OTHER OFFENSE	769	ASSAULT	737
4	Thursday	BATTERY	3189	OTHER OFFENSE	792	ASSAULT	637
5	Tuesday	BATTERY	3110	OTHER OFFENSE	801	ASSAULT	658
6	Wednesday	BATTERY	3101	OTHER OFFENSE	775	ASSAULT	679

Figure 10: Crimes domésticos mais comuns durante a semana

Uma ultima análise que fizemos foi levando em consideração se o número de escolas influenciava no número de crimes domésticos, observamos tanto quantidade em geral como quantidade de escolas por nível. Na quantidade em geral pudemos perceber que o aumento no número de escolas em uma região afetava os crimes domésticos, podendo novamente se dar ao fato de que quanto mais escolas há em uma região, mais pessoas há nela também.

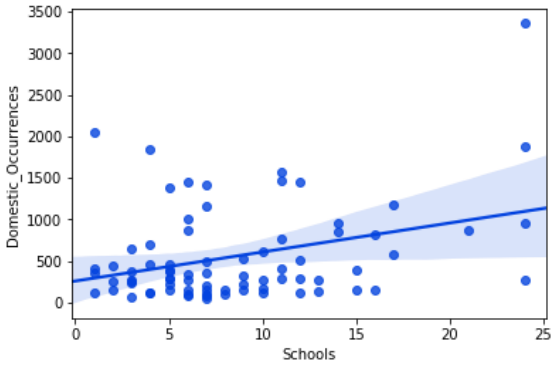

Figure 11: Crimes domésticos x número de escolas

Porém ao analisar tais crimes pela proporção de escolas nível 1+ percebemos que os seus números diminuiam. Deduzindo-se assim que ao elevar o nível de educação, as pessoas são ligeiramente menos propensas a infrigir a lei no âmbito familiar.

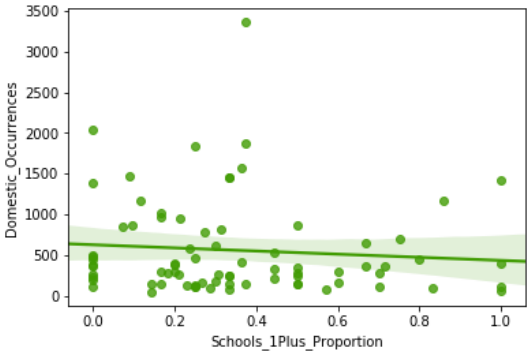

Figure 12: Crimes domésticos x número de escolas de nível 1+

2.4 Aprendizagem de Máquina

Utilizamos aprendizagem de máquina para predizer se o infrator vai ser preso ou não com base em algumas características do crime realizado, como a hora, o dia, o local, o tipo do crime e se

o crime foi doméstico ou não. Inicialmente fizemos uma regressão logística utilizando o pacote statsmodels de Python para ver a relevância dos atributos na predição do problema. Pudemos observar que a hora em que o crime ocorre é o que mais impacta nesse problema, enquanto que a definição do crime sendo doméstico ou não é o que tem menos impacto.

Após isso utilizamos a árvore de decisão da biblioteca do scikit-learn para predição do problema. Fizemos uso de k-fold estratificado, sendo 10 o número de folds, para separar o conjunto entre treino e teste, e a cada iteração do k-fold nós somamos o score do conjunto de teste, que é a acurácia em que o algoritmo prediz corretamente a classe em que o dado pertence (infrator preso ou não), e tiramos a média do score. Também foi feita a variação no número máximo de folhas, em que foi analisado que quanto maior o número de folhas maior a acurácia do algoritmo, porém a partir do número 50, o aumento na acurácia é muito baixo, havendo casos em que ela oscilava entre piora e melhora. Como um número muito grande de folhas pode causar overfitting, decidimos definir o número máximo de folhas em 50. Por fim nosso algoritmo conseguiu predizer corretamente em média 87,4% do conjunto de teste.

3 Conclusão

A partir do uso de diversas técnicas, conseguimos entender melhor sobre o crime da cidade de Chicago. Claro que não chegamos a uma solução final para o problema, pois várias outros fatores da cidade não estavam no escopo do projeto, mas deu para ter uma boa visão do comportamento da cidade. Para nós foi interessante confirmar algumas hipóteses inicias, como também foi muito construtivo falhar em uma de nossas hipóteses(de que quanto mais escolas na região, menos crimes existiriam) e perceber que a situação funcionava de outra maneira(ao levar em conta o nível das escolas, pudemos perceber a influência da educação na região).

Para o futuro, podemos aprofundar as análises, levando mais fatores em conta, para entender a criminalidade de Chicago com maior profundidade. Podemos também levar essa análise para outras grandes cidades e ver se o padrão se repete, ou se os resultados são peculiares da cidade de Chicago.