DATACOP

SISTEMA DE ANÁLISIS Y PREDICCIÓN DE CRIMINALIDAD

Eduardo Muñoz Lorenzo

Junio 2019

Índice

1. Introduction

El aprovechamiento de los datos públicos recogidos por la mayoría de las grandes ciudades internacionales, permite la generación de productos tecnológicos para el análisis y predicción de datos sociodemográficos en el tiempo. Desde este punto de vista, se pueden generar de manera similar, dashboards dinámicos con datos relacionados con crimen, emergencias, renta, etc que permitirán analizar en detalle los sucesos ocurridos y gestionar y manejar las infraestructuras y servicios públicos para la demanda futura en las diferentes áreas.

Los principales objetivos de este proyecto son:

- Análisis de criminalidad por departamentos municipales
- Predicción de eventos de criminalidad en los 12 meses posteriores a las adquisición de los datos.

La motivación de este proyecto es desarrollar un **producto** de principio a fin con el cual se puedan optimizar los recursos a corto medio plazo. De la misma manera, servirá como un analizador de éxito a posteriori en campañas de seguridad organizadas por el cuerpo de policía de la ciudad estudiada. Por último, destacar que la metodología utilizada para está solución, es aplicable a diferentes áreas (crimen, emergencias, paro, incendios, etc).

2. Data Description

El Dataset utilizado para este proyecto procede del portal abierto de datos de la ciudad de Philadelphia (www.opendataphilly.org). Se trata de datos de criminalidad desde 2006 hasta nuestros días. El dataset, hace una distinción entre tipo de crimen y distrito donde ha tenido lugar. También se han tratado los datos geoespaciales de la segmentación en distritos de la ciudad.

Nuestro base de datos de criminalidad (.csv) cuenta con los siguientes campos:

Field	**Description**	**Type**
DC_Dist	A two character field that names the District boundary.	Text
DC_Key	The unique identifier of the crime that consists of Year + District + Unique ID.	Text
Dispatch_Date_Time	The date and time that the officer was dispatched to the scene.	Date/Time
Hour	The generalized hour of the dispatched time.	Date/Time
Location_Block	The location of crime generalized by street block.	Text
PSA	A single character field that names the Police Service Area boundary.	Text
Text_General_Code	The generalized text for the crime code.	Text
UCR_General	The rounded crime code, i.e. 614 to 600.	Numeric
Point_x	Longitude	Numeric
Point_y	Latitude	Numeric

El archive geoespacial (.geojson) tiene los siguientes campos:

Field	**Description**	**Type**
PERIMETER	District perimeter	Numeric
*DISTRICT_*	Number of District	Numeric

PHONE	District phone extension	Text
AREA_SQMI	Square mile	Numeric
GEOMETRY	District polygon	Polygon

3. Methodology

3.1. Data Adquisition

Archivos asociados:

./data/01_Data_Adquisition_AWS_S3.ipynb

./src/data/dataAdquisitonAWS.py

Para la adquisición de los datos se realiza un request al portal de datos abiertos de la ciudad de Philadelphia. A continuación se genera un .csv que es alojado en un bucket en Amazon Web Service AWS S3 a través de la librería Boto3.

3.2. Data Preprocessing

Archivos asociados:

./notebooks/00_Districts.ipynb

./src/districts.py

Los datos del dataset son leídos desde AWS S3 a través de la librería Boto3. Estos datos, se correponden con sucesos o eventos criminales puntuales, por lo cual un registro corresponde a un crimen. Se ha decidido agregarlos por mes para su tratamiento como series temporales.

Antes de la agregación, se han eliminado los datos correspondientes al distrito 4, 23 y 92 por estar incompletos y no aparecer sus respectivos polígonos en el archivo geoespacial.

Se ha comprobado la existencia de nulos y se han agrupado posteriormente los datos de manera mensual y por distrito obteniendo un dataset de la siguiente forma:

	city	dist_1	dist_2	dist_18	dist_77	dist_3	dist_5	dist_6	dist_14
dispatch_date									
2006-01-31	19359	512	794	898	62	468	290	899	1134
2006-02-28	15894	385	712	676	52	444	283	723	830
2006-03-31	18627	561	817	771	61	533	257	777	1008
2006-04-30	18940	466	922	843	49	565	282	759	1025
2006-05-31	20041	458	1029	918	62	509	259	867	1100

Fig. 1 DataFrame preprocesado para su modelización como serie temporal.

El archivo resultante es alojado en el siguiente path:

./data/CSV/city_districts.csv

3.3. Analysis

Archivos asociados:

./notebooks/10_DataAnalysis.ipynb

Los datos recogidos desde el portal abierto de la ciudad de Philadelphia tienen el siguiente comportamiento:

Fig. 2 Serie temporal de criminalidad en la ciudad de Philadelphia desde 2006 hasta la actualidad.

Se aprecia una clara estacionalidad, la cual, corresponde con un crecimiento de la criminalidad en las estaciones de primavera y verano, y una disminución de los incidentes en otoño e invierno. Estos datos, se pueden relacionar de forma directa con el comportamiento meteorológico. De manera global, se aprecia una disminución gradual del número de incidentes desde 2006 hasta la actualidad.

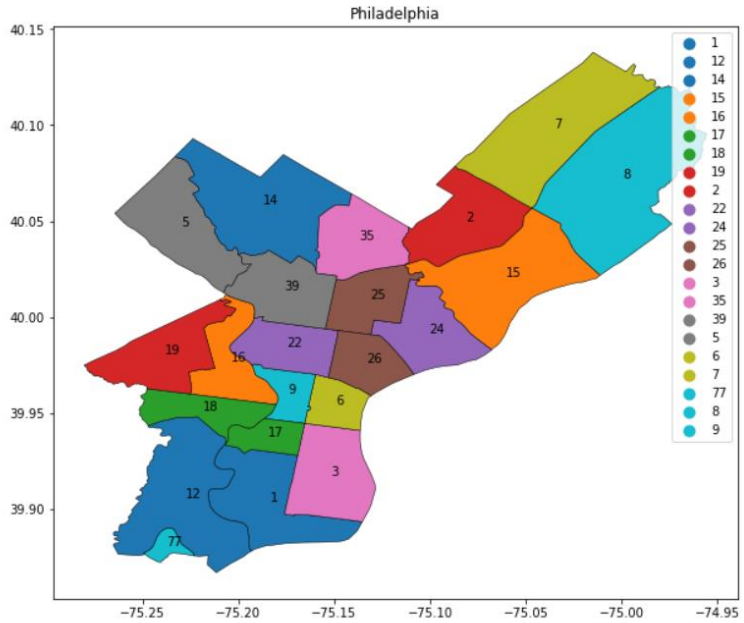

Fig. 3 Representación de los Distritos de la ciudad de Philadelphia.

En la figura anterior, se ha representado gráficamente los datos geoespaciales, donde se puede observar la disposición de los distritos en la ciudad de Philadelphia. Se ha utilizado la librería Geopandas para el tratamiento del archivo geoespacial.

A continuación, se ha realizado un análisis de los crímenes a nivel ciudad en la serie temporal:

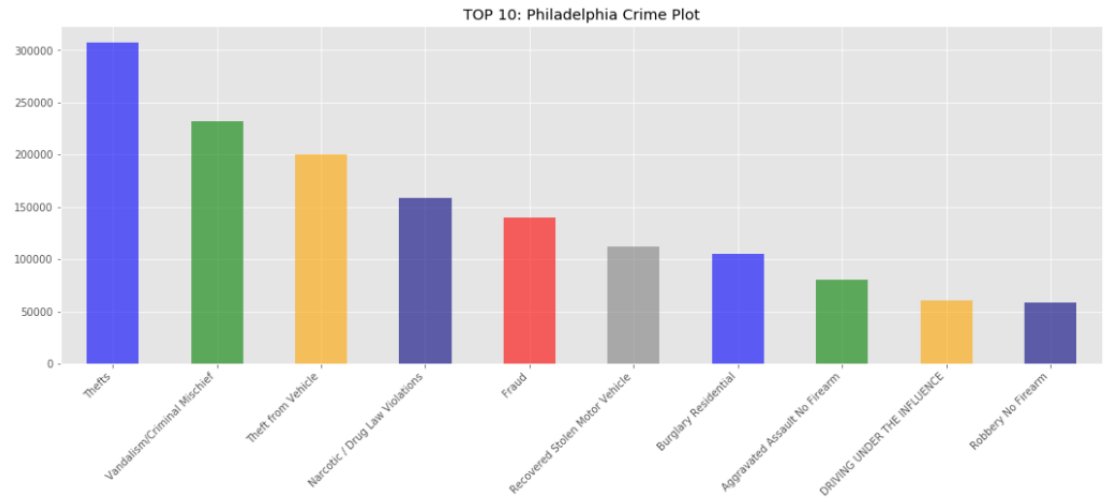

Fig. 4 Top 10 eventos de criminalidad en la ciudad de Philadelphia desde 2006.

La categoría de crimen más repetida en la ciudad de Philadelphia es el robo o hurto con más de 300000 registros. Seguidamente, encontramos la categoría de vandalismo, robos desde vehículo, narcóticos, fraude, etc.

Sería interesante poder desglosar el top 3 de categorías criminales por barrios.

Los distritos con más robos son el 6 y el 9, los cuales coinciden con el centro de la ciudad, lugar común de este tipo de sucesos.

Los actos de vandalismo son de manera notable más evidentes en el distrito 15. Lo mismo sucede con robos desde vehículos, el distrito número 15 encabeza la lista de este tipo de crimen.

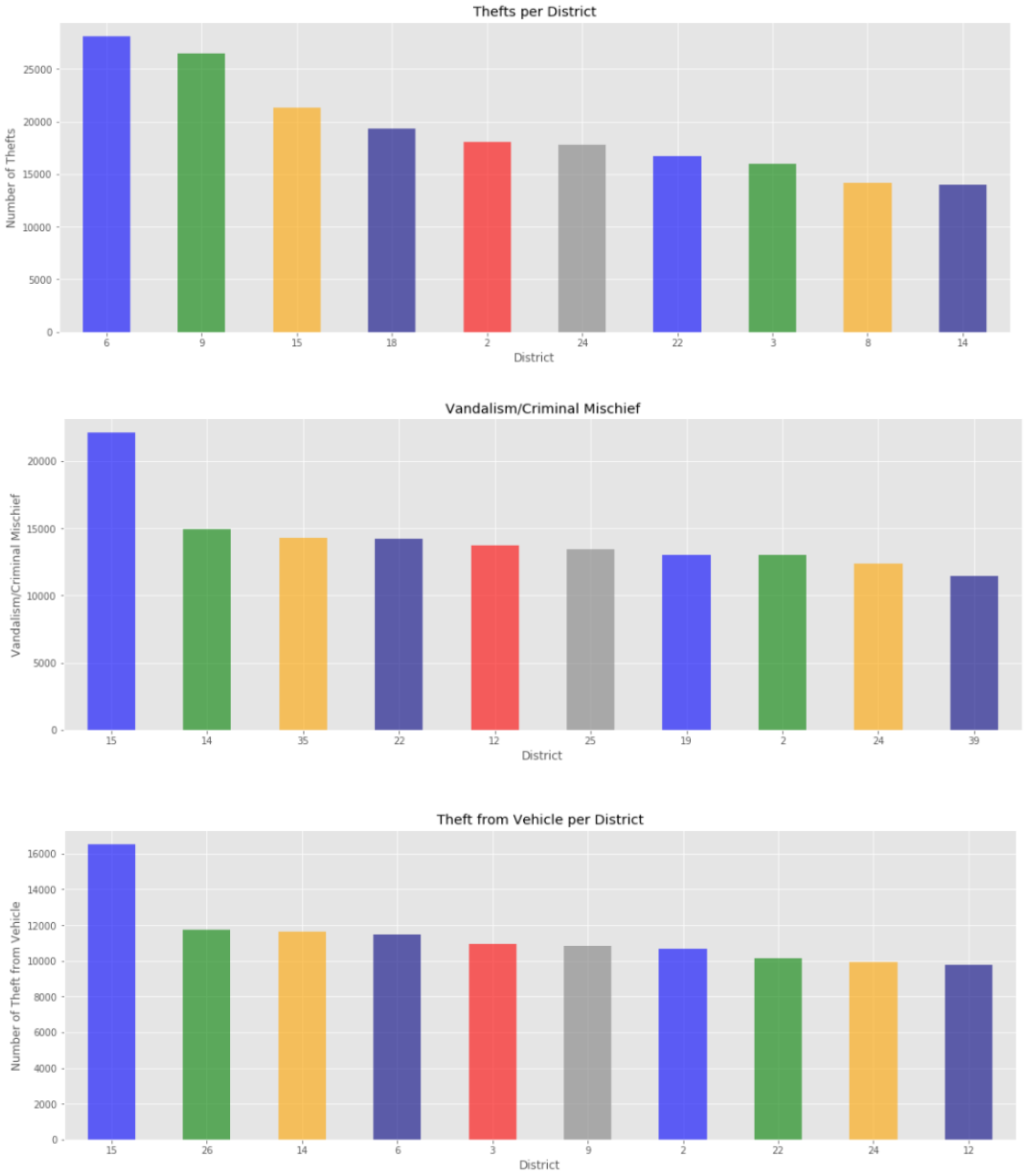

Fig. 5 Thefts/Vandalism/Theft from Vehicle Crimes per District

Por último, vamos a analizar en qué periodo del día se producen más crímenes:

Como se aprecia en la figura siguiente, la mayoría de crímenes tienen lugar en los periodos que comprenden la tarde (afternoon y evening) y la noche (night).

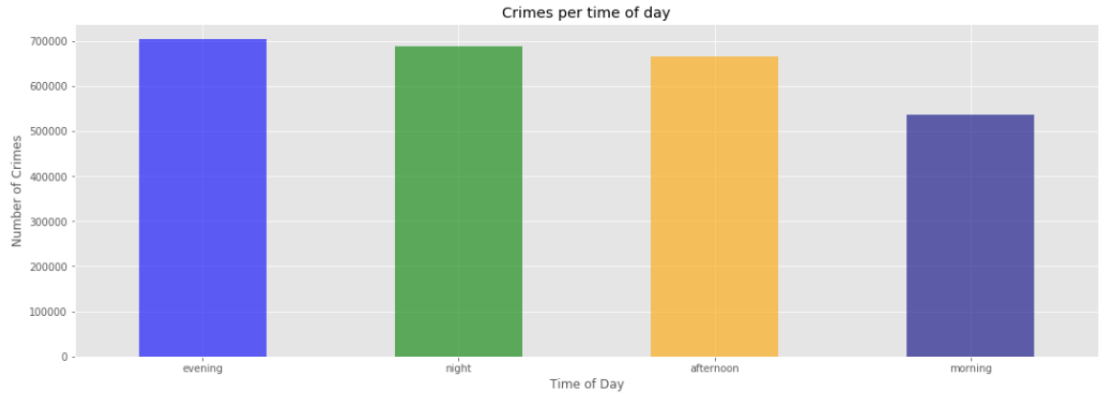

Fig. 6 Crimes per time of day

3.4. Time Series Models

Archivos asociados:

./notebooks/01_SARIMA.ipynb **Testeo de SARIMA para la ciudad completa**
./notebooks/02_Prophet.ipynb **Testeo de Prophet para la ciudad completa**
./notebooks/03_Keras.ipynb **Testeo de Keras LSTM para la ciudad completa**
./notebooks/06_SARIMA_all_Districts_Tableau.ipynb **SARIMA para ciudad y distritos**
./notebooks/07_Prophet_all_Districts_Tableau.ipynb **Prophet para ciudad y distritos**
./notebooks/08_Join_Results.ipynb Uniendo **Resultados para representación en Dashboard**

Para la predicción de criminalidad en la ciudad de Philadelphia se han utilizado diferentes modelos de series temporales.

- SARIMA
- Facebook Prophet
- LSTM Keras

SARIMA (Seasonal Autoregressive Integrated Moving Average Model) es un modelo ARIMA que tiene en cuenta la estacionalidad. Como se puede apreciar en la figura 2, los datos muestran una estacionalidad anual.

Por otro lado, Prophet es una librería de Facebook para el tratemiento de series temporales. El input que recibe Prophet siempre es de la misma forma, un dataframe con dos columnas:

- ds: Fecha
- y: Datos numéricos

LSTM (Long Short Term Memory) aplica una mejora a las redes neuronales recurrentes (RNN). Ya que soluciona la pérdida de memoria sobre los primeros inputs de la red y la perdida de información que se produce en cada paso.

Como primer paso, se han aplicado estos tres modelos al conjunto total de criminalidad (Nivel Ciudad) para testear sus resultados. Para ello, se ha utilizado un conjunto de 12 meses para el conjunto de validación (test), quedando 148 meses para el conjuento de entrenamiento (train).

La métrica utilizada para testear la precisión del modelo ha sido RMSE (root mean squared error)

3.4.1. SARIMA

- SARIMA Parametros usando Pyramid ARIMA SARIMA(1,1,0)(2,0,1,12)
- RMSE: 362,76
- Media en el conjunto de validación: 13552,91

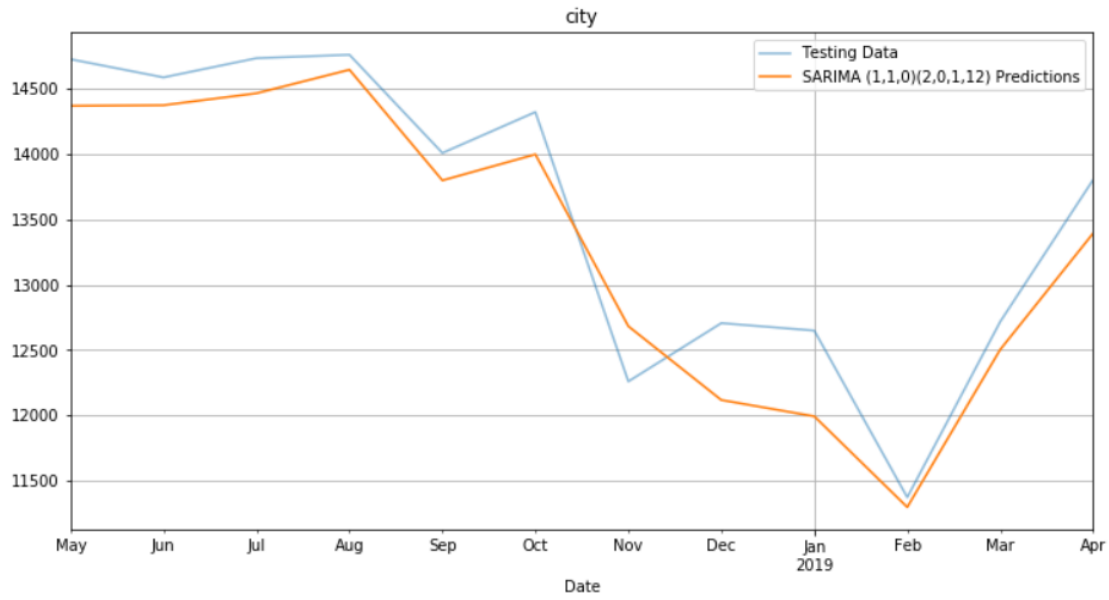

Fig. 7 SARIMA: Predicción conjunto de validación vs Valores Reales.

Finalmente, se ha entrenado un nuevo modelo con todo el dataset y se ha calculado un forecast de 12 meses.

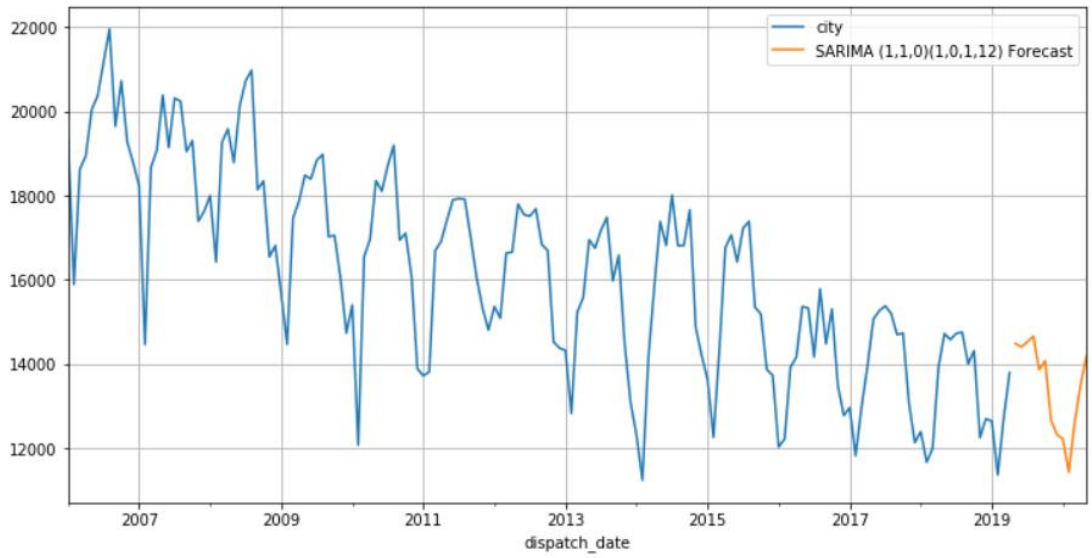

Fig. 8 Predicciones a un año vista.

3.4.2. Prophet

- Prophet modelo multiplicativo
- RMSE: 481.42
- Media en el conjunto de validación: 13552,91

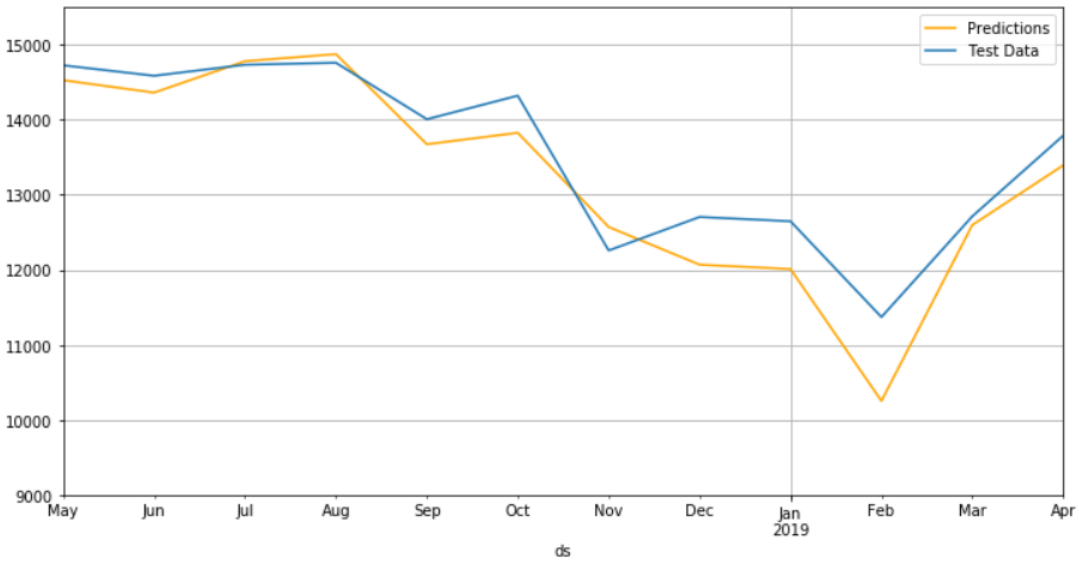

Fig. 9 Prophet: Predicción conjunto de validación vs Valores Reales

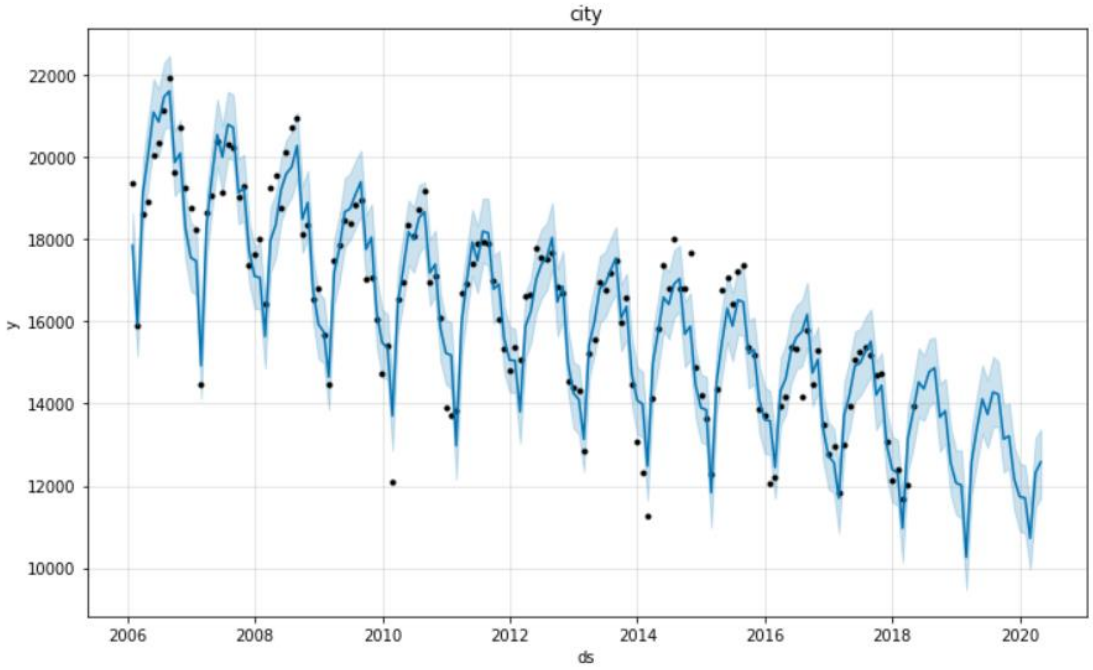

Fig. 10. Predicciones a un año vista.

3.4.3. LSTM

- LSTM 150 neuronas y activación relu
- RMSE: 794.93
- Media en el conjunto de validación: 13552,91

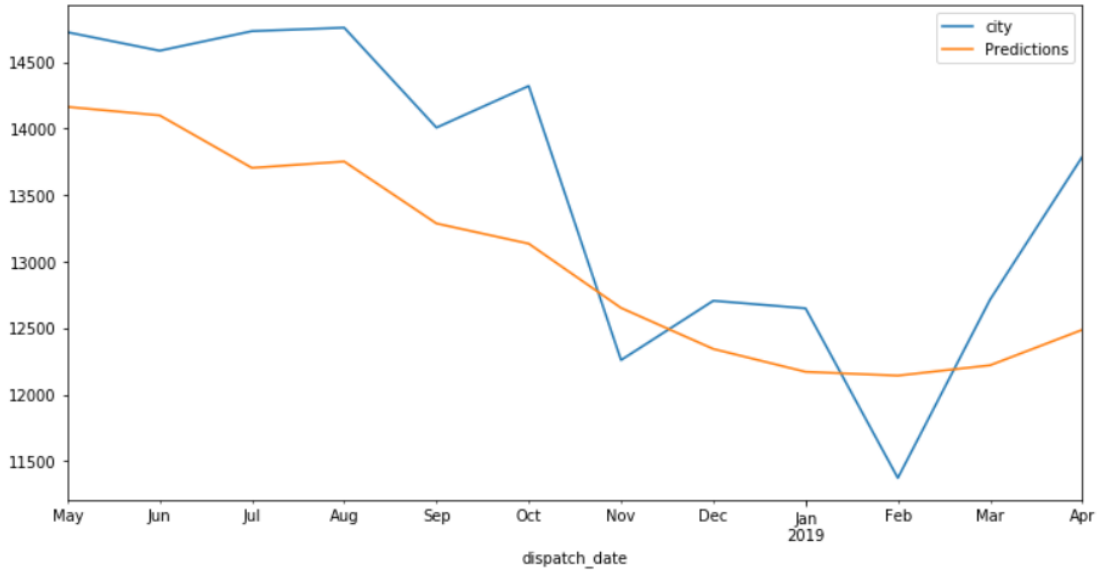

Fig. 11 LSTM: Predicción conjunto de validación vs Valores Reales

Observando los resultados, se aprecia que el mejor modelo es SARIMA, seguido de Prophet y en último lugar las redes neuronales LSTM, las cuales no han dado un buen resultado y se tendría que estudiar su optimización más a fondo para un mejor forecast.

A continuación, en los notebooks 6, 7 y 8 se ha ejecutado el modelo tanto para el nivel ciudad como para todos los distritos. Los resultados se han unido para su posterior análisis en el dashboard realizado con Tableau.

3.5. Dashboard

Archivos asociados:

./dashboard/DataCopTFM_Final.twb

Enlace al Dashboard público: [Tableau Dashboard DataCop]

Se ha realizado un dashboard utilizando la herramienta Tableau para la representación y análisis interactivo de los resultados.

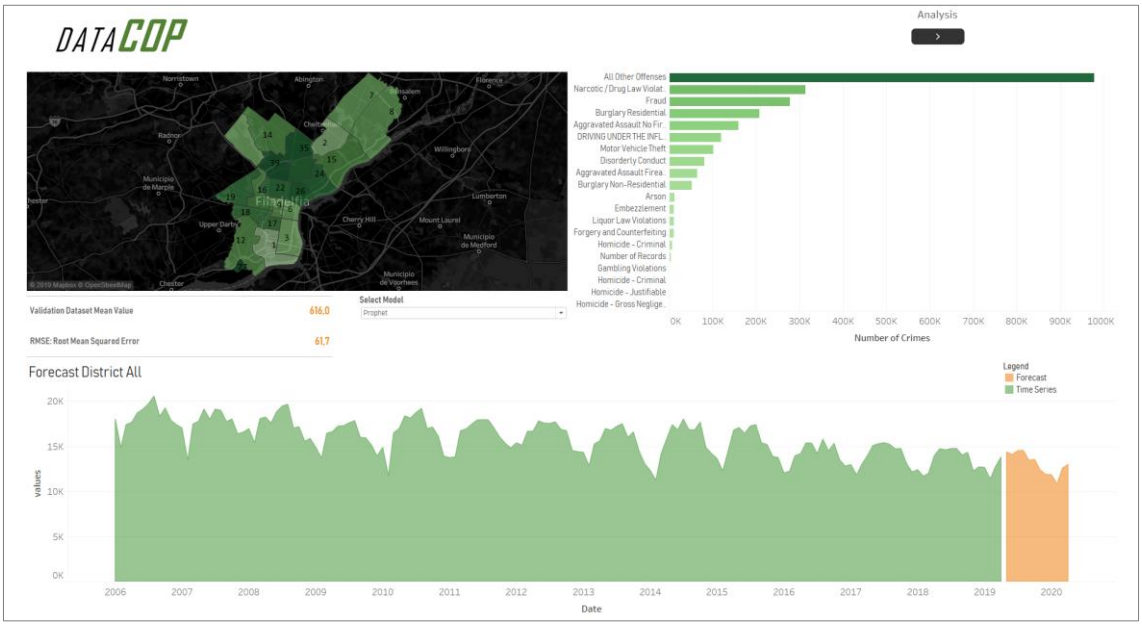

Fig. 12 Dashboard con desglose de tipo de crimen, distrito y predicciones en los próximos 12 meses.

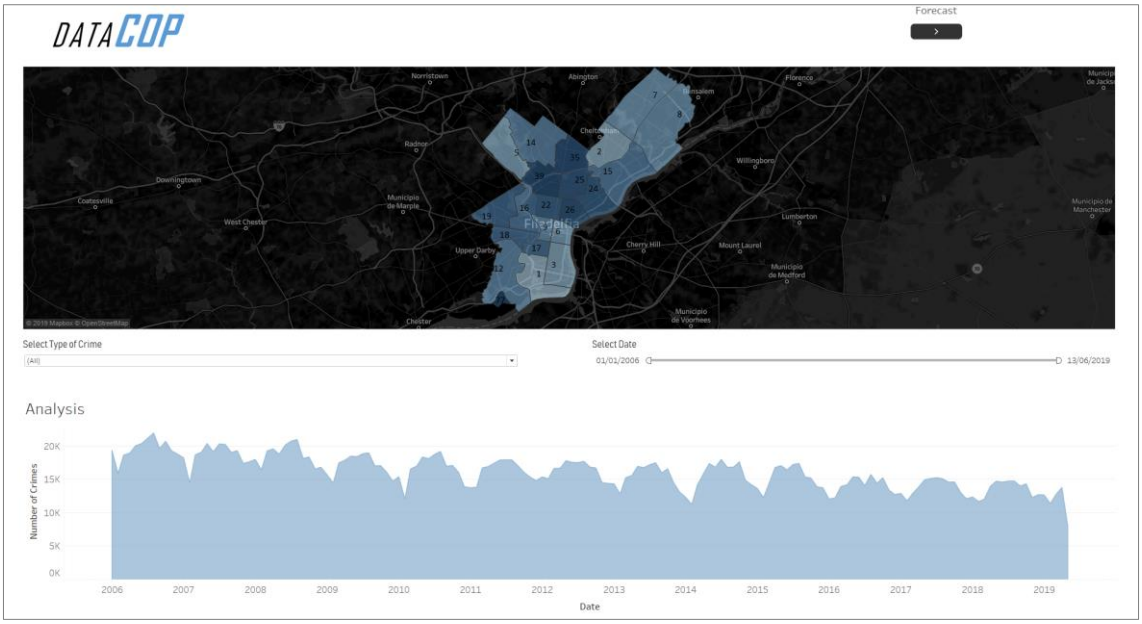

Fig. 13 Dashboard de análisis de criminalidad según distrito, tipo de crimen y fecha

4. Results

La serie temporal escogida en un primer momento engloba desde principios del 2006 hasta la actualidad. Los modelos fueron calculados tomando como último punto de valores reales el mes de abril de 2019. Se ha tomado el mes de mayo como testeo para evaluar la efectividad de las predicciones y estos son los resultados.

Distrito	Valor Real Mayo 2019	SARIMA Mayo 2019	Prophet Mayo 2019
City	14973	**14535**	14366
1	206	240	**221**
2	593	660	**632**
3	693	601	**653**
5	189	178	**186**
6	735	**723**	640
7	213	**239**	252
8	363	393	**356**
9	586	604	**574**
12	659	732	**721**
14	889	788	**810**
15	1025	**1048**	1164
16	706	**608**	540
17	411	**387**	300
18	818	749	**771**
19	1029	**1084**	1190
22	1016	**1028**	1052
24	1175	1154	**1189**
25	836	**833**	883
26	514	559	**533**
35	1126	**966**	841
39	930	**915**	799
77	61	47	**59**

5. Conclusions and Future Lines

Como futuros trabajos para la comercialización y mejora del producto se proponen los siguientes pasos:

- Introducción de variables exógenas SARIMAX
- Implementación de más modelos de predicción: Random Forest, XGBoost, Google Casual Impact.
- Automatización de procesos con periodicidad mensual
    - Ingesta
    - Entrenamiento
    - Visualización
- Gestión interdistrito de recursos humanos y materiales de los cuerpos de seguridad a partir de las predicciones a futuro.