Modelos de Regresión aplicados a Violencia Familiar

Martín Hernan Bellini, Joaquin Espona, Tamara Bertomeu

UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Introducción:

El presente trabajo esta basado en la información de los llamados atendidos por violencia familiar al #137 en la Ciudad Autónoma de Buenos Aires desde el inicio del 2017 al tercer trimestre del 2019 proporcionados por El Ministerio de Justicia y Derechos Humanos de la Nación y la Secretaria de Modernización de la Presidencia de la Nación.

Partiremos de una base de mas de 23.420 llamados realizados en el periodo de tiempo mencionado.

Tendremos como objetivo predecir la edad de la victima de violencia familiar mediante el uso de los modelos de regresión en aquellos casos donde dispongamos de toda la información a excepción de la edad de la victima u su rango etario.

Datos:

Los datos que se utilizaran como input a lo largo de los distintos modelos de regresión que verán son los siguientes:

- Descripción del llamante
- Género del llamante
- Vínculo del llamante con los niños presentes
- Tipo de violencia
- Edad de la víctima
- Rango etario de la víctima
- Género de la víctima
- Relación del agresor con la víctima
- Derivación del llamado
- Cantidad de víctima
- Cantidad de agresores
- Año del llamado
- Mes del llamado
- Día del llamado
- Día de la semana del llamado
- Hora del llamado
- Momento del día del llamado

Tras una exploración, análisis y limpieza de la información obtenida, la cantidad de llamados con la cual se trabajara a futuro será de 4256.

Situación pre-modelos:

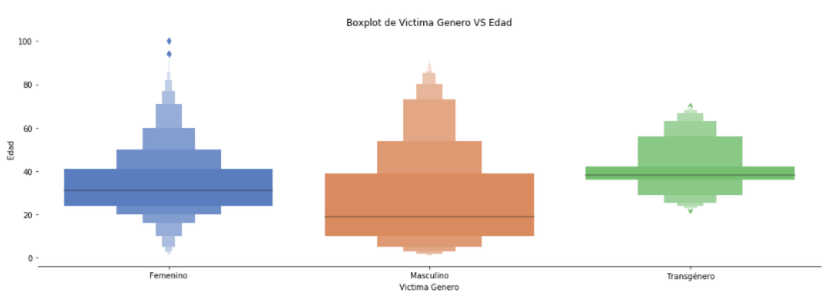
A continuación, observaremos como se encuentra la situación actual desde distintos aspectos.

Boxplot:

En la figura de la derecha, podrán observar como se distribuye la edad de la víctima según su género.

La recta de color negro representa la mediana de la edad de la víctima.

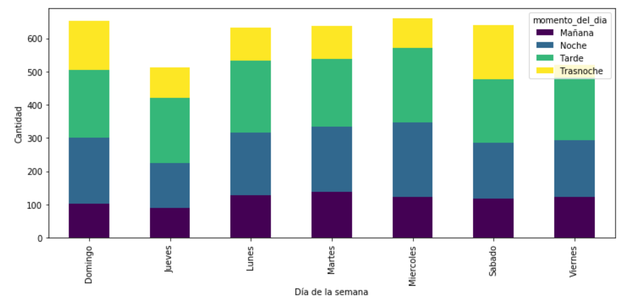
*Mediana: Valor que indica que el 50% de los datos es igual o menor a dicho valor.



Frecuencia:

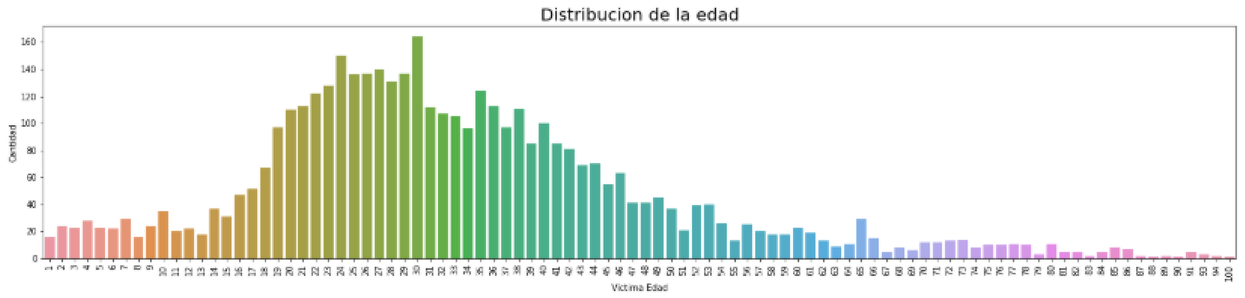
En los próximos 2 gráficos de la derecha, podrán observar que:

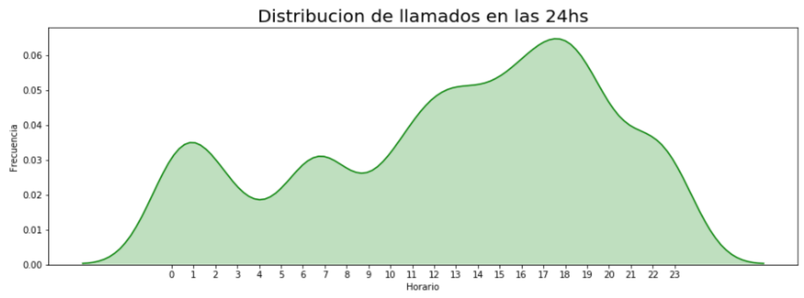
- ❑ La mayor cantidad de llamados se encuentra de 12:00 a 20:00hs. Esto comprende los momentos de la "Tarde" y "Noche".
- ❑ Por otro lado, el día de menor cantidad de llamados es el Jueves.
- ❑ Tanto el Sábado como el Domingo, muestran una notable cantidad de llamados durante la Trasnoche.



Resumen - Edad de la victima:

- Media: 33 Años
- Desvió Std: 15 Años
- Mínimo: 1 Año
- 1° Cuartil: 23 Años
- 2° Cuartil: 31 Años
- 3° Cuartil: 41 Años
- Máximo: 100 Años





Modelos de Regresión y sus Resultados:

Una vez explorado y ordenado la información, procedimos a entrenar distintos modelos de regresión utilizando técnicas de Feature Selection con el fin de mejorar su performance. A continuación detallaremos los modelos y técnicas de feature selection implementados:

Modelos de Regresión:

- ❑ KNN Regression [KNN]
- ❑ Logistic Regression [LR]
- ❑ Support Vector Machine – Regression [SVR]

Para este ultimo método, se utilizo la siguientes variantes:

- Lineal [L]
- Gaussiano [G]

Métodos de Feature Selection:

- ❑ Threshold [TH]
- ❑ Lasso [LA]

Adicionalmente, se entreno a los modelos con la base de datos entera, es decir, sin Feature Selection [SF]

Error:

Con el fin de evaluar la performance de los modelos de regresión y poder compararlos entre ellos, se midieron los siguientes errores:

- ❑ MAE: Error Absoluto Medio
- ❑ MSE: Error Cuadrático Medio
- ❑ RMSE: Raíz del Error Cuadrático Medio.

Los Resultados fueron los siguientes:

	Model	MAE	MSE	RMSE	
0	LR-SF	5.544114	96.565527	9.826776	Sin Feature Selection
1	KNN-SF	7.201988	116.512367	10.794089	
2	SVR-L-SF	4.938397	117.785760	10.852915	
3	SVR-G-SF	5.214431	84.497585	9.192257	
4	LR-TH	5.517641	99.126198	9.956214	Threshold
5	KNN-TH	6.950942	118.486057	10.885130	
6	SVR-L-TH	4.934553	121.197110	11.008956	
7	SVR-G-TH	4.484192	77.481729	8.802371	
8	LR-LA	5.543652	96.573017	9.827157	Lasso
9	KNN-LA	6.211749	96.955604	9.846604	
10	SVR-L-LA	4.939136	117.689975	10.848501	
11	SVR-G-LA	4.710262	78.795152	8.876663	

Modelos elegidos:

A partir de los errores y sus respectivas curvas de distribución e histogramas de los datos de testeo vs los valores predichos por los modelos, se eligió continuar con siguientes modelos para el proyecto:

- KNN Regression con Lasso de Feature Selection
- SVR Gaussiano con Threshold de Feature Selection

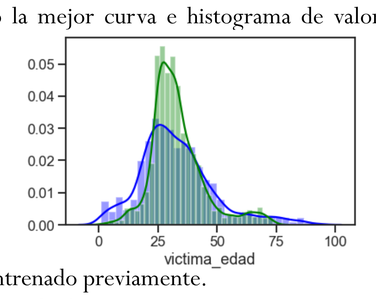
Modelo KNN con Lasso:

El KNN regression que utilizamos se caracteriza, no por tener el menor error, sino la mejor curva e histograma de valores predichos respecto a los valores reales de prueba como se puede ver a continuación.

- ❑ Verde: Edades predichas por el modelo.
- ❑ Azul: Edades reales de la base de datos de los llamados.

Error:

- ❑ MAE: 6.211
- ❑ MSE: 96.955
- ❑ RMSE: 9.846

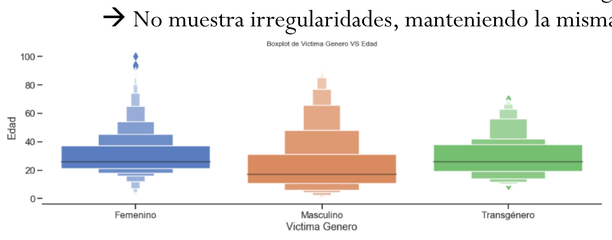


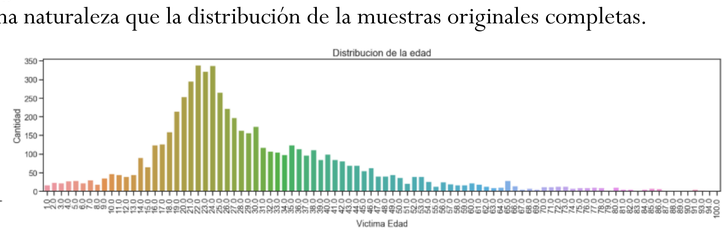
A partir de las muestras, cuya edad estaba sin completar, se aplico el KNN con Lasso entrenado previamente.

Con las edades predichas + las muestras que ya poseían edad, obtuvimos los siguientes resultados:

→ Una curva de distribución mas centrada en el rango de edad de 20 a 30 años.

→ No muestra irregularidades, manteniendo la misma naturaleza que la distribución de la muestras originales completas.





Modelo SVR Gaussiano con Threshold:

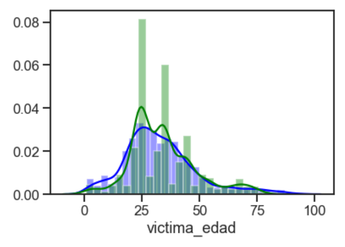
Por otro lado, el Support Vector Regresion de tipo Gaussiano que utilizamos tiene como principal característica que es el modelo de regresión que menor error MAE, MSE y RMSE dio.

Pero por el contrario, peor curva de distribución e histograma dio.

- ❑ Verde: Edades predichas por el modelo.
- ❑ Azul: Edades reales de la base de datos de los llamados.

Error:

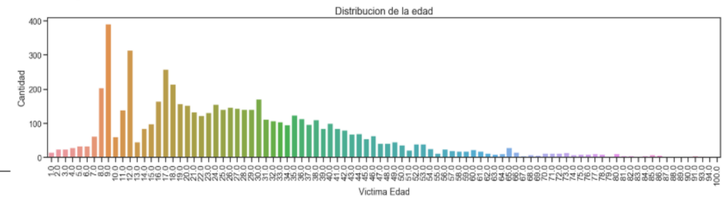
- ❑ MAE: 4.484
- ❑ MSE: 77.481
- ❑ RMSE: 8.802



De la misma manera que hicimos con el KNN con Lasso, procedimos a aplicar el SVR Gaussiano y unir las muestras con las edades predichas con aquellas que ya poseían dicho campo, los resultados fueron los siguientes:

→ La distribución de las edades sufrió grandes cambios producto de las bajas edades predichas

→ La mayoría de las edades predichas se encuentran en el rango de 8 a 18 años.



Conclusión:

El modelo de 'menor error' era el SVR Gausiano con Threshold, pero aun así nos resultados arrojados en la retroalimentación muestran irregularidades en su distribución de las edades. En cambio el modelo KNN con Lasso, que si bien no tenia la mejor performance en base a los errores, presenta una curva de distribución de las edades sin irregularidades y con una clara media. Al mismo tiempo, es la que mas se asemeja a la curva pre-modelos de regresión.

Por lo que guiarse únicamente por valores numérico como los errores, media o desvió no es lo mas optimo, y será necesario complementarlo con herramientas visuales como curvas de distribución u histogramas para notar rápidamente si hay alguna irregularidad.

Hallazgos:

→ Hay una fuerte correlación lineal negativa entre el género del agresor y el de la víctima.

→ Predomina fuertemente la violencia hacia las mujeres, donde en el 90% de los llamados la víctima es una mujer.

→ Por otro lado, en el 88% de los llamados el agresor es un hombre.

→ El principal tipo de violencia es la violencia física con un 70%, la psicológica ocupa un 25% de los llamados y el 5% cae en la categoría "otros".

→ En el aproximadamente 70% de los casos, la víctima pidió orientación u no acepto la intervención de equipos móviles.

→ El agresor es en el 74% de los casos la actual o ex pareja de la victima.

→ Los llamados proviene en 55% de los casos desde la comisaria, un 25% provienen de la víctima y otro 10% por parte de un familiar de la víctima.

Resumen de los datos:

→ La cantidad total de llamados registrados fueron de 23.420.

→ La cantidad de muestras utilizado para el entrenamiento de modelos fueron 4.256.

→ La cantidad de llamados a los cuales se le predijo la edad fueron 1.818.

→ La cantidad de llamados sin edad eran 9.434.

→ En total, se pudo re-utilizar un 20% de los llamados sin edad.

→ Se paso de utilizar el 19% de todos los llamados al 27%. Pasando de 4.256 a 6.083 llamados, esto ultimo representa un aumento del 42%.