Pauta Publicitaria

Ciudad de Buenos Aires

Ades Lucas
Paez Facundo Nahuel
Vinacur Damian

Abstract

Este artículo tiene el objetivo de poder analizar y obtener información sobre los diferentes tipos de publicidades que ha estado invirtiendo Ciudad de Buenos Aires desde 2016 a 2019. Para poder lograr este objetivo se realizó un preprocesamiento de los datos, una exploración de los mismos y se aplicaron 3 modelos de Machine Learning.

Keywords

Publicidad, TV, Graficas, Medios Vecinales, Radio

I. INTRODUCCION

Si bien hoy en día sabemos que existe publicidad por parte del Gobierno de la Ciudad de Buenos Aires por distintos medios, muy pocos saben exactamente en qué medios se invierten, ni cuáles son los montos estimados de los mismos.

De parte del Gobierno de la Ciudad, creemos que es importante poder reflejar con claridad y transparencia todos estos datos a toda la población que realiza aportes en esta jurisdicción, y ver en que forma se gasta la recaudación.

Por lo tanto, el objetivo que buscamos con el siguiente Paper, es poder conseguir información que nos enseñe el importe que conllevan las distintas publicidades en Ciudad de Buenos Aires, y poder entender, o tener una noción más clara, de hacia donde va el importe de los contribuyentes en lo que respecta a publicidad.

II. DATA SET

Para el siguiente informe, hemos tomado 16 data sets. El gran numero, es debido a que hemos tomado el dataset de cada tipo de publicidad en los años 2016, 2017, 2018 y 2019. Cada dataset está separado por tipo de publicidad, los cuales son: “TV”, “Radio”, “Grafica” y “Medios Vecinales”.

A continuación, el link de acceso a los data sets:

https://data.buenosaires.gob.ar/dataset/pauta-publicitaria

En lo que respecta al formato de los datasets, todos conservan el mismo. Este consiste en 4 features, las cuales son: “Fecha de publicación”, “Tipo”, “Medio” e “Importe”.

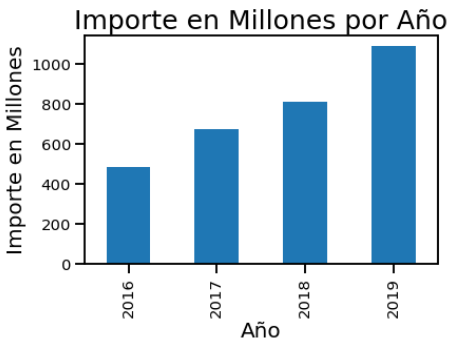
Antes de comenzar con el análisis exploratorio de datos, se tuvo que realizar un proceso de concatenado de los data set de tal manera que conseguimos 4 datasets consolidados de cada tipo de publicidad. Por último, concatenamos estas cuatro bases de datos para obtener un último datasets que aglomere la totalidad de la publicidad. Obteniendo para este último un total de 29274 muestras.

III. ANALISIS EXPLORATORIO DE DATOS

La mejor manera de poder ver e interpretar información y orientarnos un poco sobre que tipo de data set manejamos, es obtener primero gráficos simples y luego proceder a gráficos más complejos.

Análisis anual

El primer grafico del que partimos es el siguiente:



Aquí podemos ver un incremento del importe en millones de publicidad total por año. Sin embargo, nos interesa ver si la inversión en publicidad ha aumentado en términos reales, ya que debemos tener en cuenta la inflación del país a lo largo de esos cuatro años de estudio.

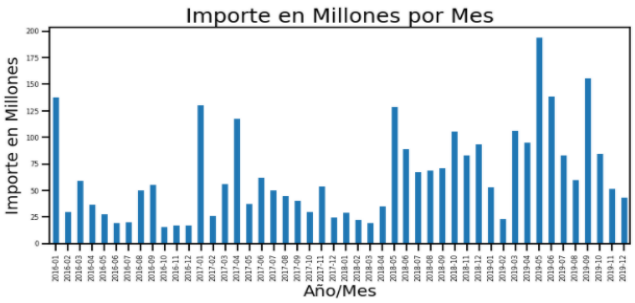
Año	Incremento de Importe	Inflacion*
2017	38%	24,80%
2018	20%	47,65%
2019	34%	53,83%

*Fuente: Indec

Podemos apreciar que a pesar de que la publicidad aumento su inversión año tras año, en términos reales, hubo un importante aumento en 2017 de 14 puntos de diferencia con respecto a la inflación, pero luego tanto en el año 2018 como 2019 este decreció.

Análisis Mensual

En cuanto al análisis de los 48 meses que conforman el tiempo de estudio, conseguimos la siguiente información:

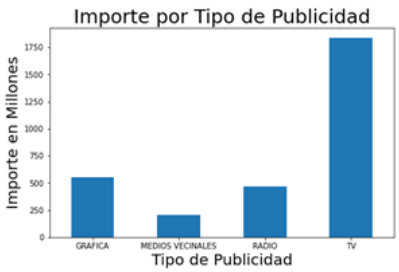


Aquí, podemos visualizar algunos picos de importe, siendo el mayor, de casi 200 millones de pesos, en Mayo de 2019, mes previo a las PASO que se realizaron el 12 de Agosto. En Junio, el importe bajó a casi 125 millones de pesos.

Además, vemos otro pico importante en Septiembre, previo a Octubre, mes en el que se voto presidente en Argentina. Este fue un importe de mas de 150 millones de pesos.

Análisis por tipo de Publicidad

Una vez analizado el data set en dimensiones de tiempo, nos interesa ahora ver qué tipo de publicidad es la que mas importe conllevo a lo largo de los 4 años.



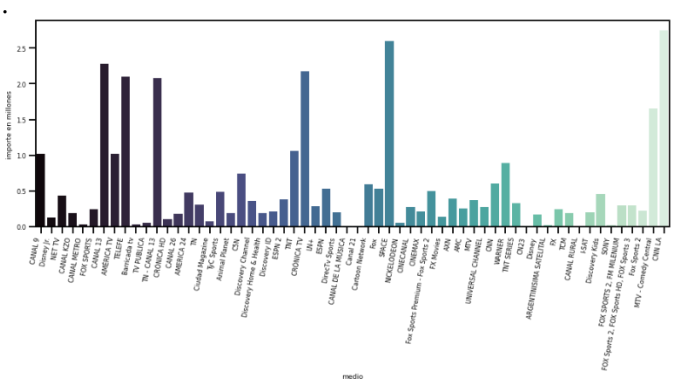
Intuitivamente, se puede decir que TV debería ser el tipo de publicidad con mas importe invertido, pero gracias al análisis de los datos, corroboramos esta información.

Que TV sea el tipo de publicidad en el que mas importe se invierta, y por amplia diferencia con respecto a las otras, es debido a varios factores:

- Mayor audiencia se encuentra en este rubro, así como a su vez, el mayor lapso de atención.
- Por medio de la TV, es la mejor manera en la que un político puede transmitir sus ideas/emociones, dado que gran parte de la comunicación es corporal, y mediante TV, se pueden reflejar estos comportamientos.

Debido a esto, decidimos incursionar un poco mas en TV y ver que medios son los que absorbieron la publicidad más cara en el año 2019.

.



Para nuestra sorpresa, vemos que el mayor importe en publicidad fue destinado a CNN con más de 2.5 millones de pesos. Luego, le sigue muy de cerca Space, y detrás Canal 13.

También podemos mencionar a Telefe, TN y Crónica TV, con importes aproximados de 2.1 y 2.2 millones.

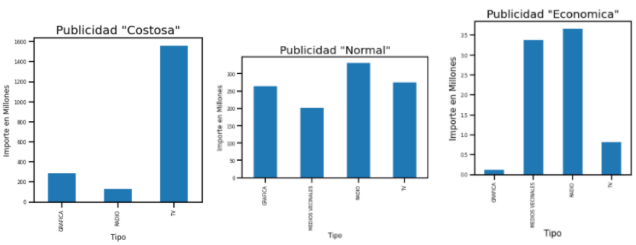
Análisis Contemplando Outliers

A la hora de analizar datos, es importante adoptar un criterio en el cual se eliminen outliers, es decir, puntos que quizá no guardan mucha correlación con el resto de los datos, sino que se encuentran por encima de un cuantil elevado, o por debajo de un cuantil muy inferior.

Es por eso, que adoptamos como publicidad excesivamente cara a la que tiene un importe por encima del cuantil 0.95, y publicidad económica a la que pertenece por debajo del cuantil 0.05.

Por lo tanto, procedemos a analizar el importe por tipo de publicidad, eliminando publicidades excesivamente caras y baratas.

Obtuvimos lo siguiente:



Encontramos algo muy interesante, y es que cuando hablamos de publicidad costosa, la mayor cantidad de importe es destinada a TV.

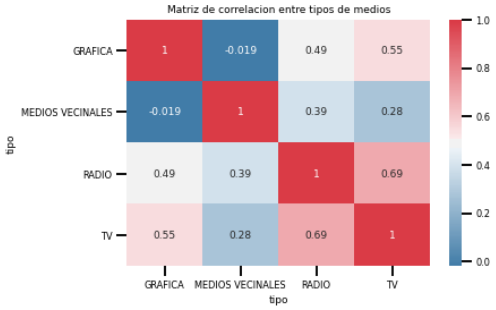
En el otro extremo, cuando analizamos publicidad económica, el mayor importe es destinado a Medios Vecinales y Radio.

Sin embargo, cuando tomamos publicidades de precio 'Normal', se empareja la destinación de los importes.

Esto nos demuestra que la mayor parte de las publicidades de gran importe, son de TV, mientras que las de menos importe, pertenecen a Radio y Medios Vecinales, que es lo que se podría suponer.

Relación entre Tipos de Publicidad

Antes de proceder a la aplicación de Machine Learning, corroboraremos la relación (o no correlación) de los tipos de publicidad.



A través del Heat Map, podemos ver que no existe gran correlación entre los tipos de publicidad, por sobre todo Medios Vecinales, la cual posee un R muy bajo con los otros tres tipos de publicidad.

La mayor correlación se da entre TV y Radio, con una R de 0,69. Podríamos decir que al aumentar el importe de publicidad en TV, hay más posibilidades de que aumente el importe en publicidad por Radio, aunque la correlación no la consideramos fuerte.

IV APLICACION DE MACHINE LEARNING

El objetivo que buscamos al realizar el modelo es poder predecir cual será el importe total de cada mes en relación con los datos de TV, dado que este, por gran diferencia, es el tipo de publicidad que incurre en el mayor importe. Por lo que se podrá determinar un aproximado de cuanto se invertirá, permitiendo determinar un presupuesto y a su vez, analizar posibles variaciones de la realidad con respecto a la predicción.

Para lograr esto decidimos analizarlo utilizando los data sets del 2016 al 2019 de tv para luego llevar a la práctica tres modelos distintos para su análisis.

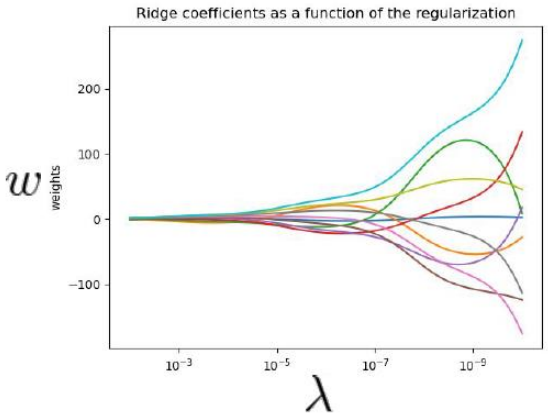
Linear Regression [1]:

En este método, se utilizan como inputs las meses (X) y los parámetros "w", y a través de entrenamiento de los datos utilizando los importes, nos dará como salida una función lineal, la cual va a tratar de asemejarse lo máximo posible a los valores reales, con el objetivo de minimizar el error (disminución de la función de costo). A simple vista, podríamos deducir que la línea de tendencia tendera a una recta de pendiente postitiva, ya que al avance de los años, el importe es mayor.

$$\hat{y} = f(x, w)$$

Ridge Regression [2]:

La diferencia de este modelo con el de regresión lineal mencionado anteriormente, es que en este modelo se utiliza un híper parámetro (λ) llamado L2, el cual es elegido por el usuario para penalizar los parámetros "w". En el siguiente gráfico muestra como ante un mayor L2 elegido por el usuario, el parámetro "w" va a tender a 0



Support Vector Regression (SVR) [3]:

El modelo de Support Vector Regression es el tercer modelo que utilizamos. En este caso, se busca el hiperplano que maximice el margen, en el cual se tiene una función de costo (c) en la que se penalizan las muestras que caen fuera del margen.

$$C \sum_{n=1}^{N} \xi_n + 1/2 \|w\|^2$$

V RESULTADOS

Luego de aplicar los 3 modelos lineales, observamos los resultados que obtuvo cada modelo con sus errores. Estos nos permitieron averiguar cual es el modelo más conveniente para este data set.

Para calcular el error medimos el R2, el error cuadrático medio (MSE) y la media del error (MAE). El R2 explica cuál es la proporción de varianza de las etiquetas que explica el

modelo de regresión. Y analizando el MSE y el MAE, la principal diferencia en los cálculos de estos errores es que el MSE al ser cuadrático permite sumar en términos absolutos tanto diferencias positivas como negativas.

Calculo de MAE:

$$\boldsymbol{MAE} = \frac{|\sum (\widehat{y}_t - y_t)|}{n}$$

Cálculo de MSE:

$$\boldsymbol{MSE} = \frac{\sum (\widehat{y}_t - y_t)^2}{n}$$

Cálculo de Coeficiente de Determinación (R2):

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}$$

Los resultados obtenidos con respecto a estos errores fueron los siguientes:

Model	Features	R2	MSE	MAE
Linear	Lineal	0.086	848.999	23.757
SVR	Linear	0.183	759.165	21.520
Ridge	Lineal	0.085	850.431	23.623

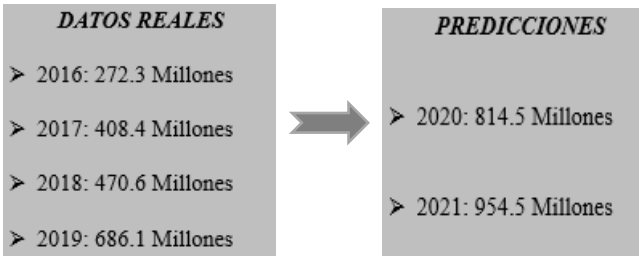
Haciendo un primer análisis de estos datos podemos notar que no se hallaron valores significativos ya que todos los valores de R2 muestran que nuestra variable independiente no logra explicar con efectividad a la variable dependiente. Al analizar los errores podemos darnos cuenta que son de muy alta magnitud, en especial el MSE, ya que al tener en cuenta los errores absolutos sumará positivamente tanto errores en los que predecimos valores superiores a los reales como inferiores.

A pesar de conseguir estos valores que se encuentran lejos del valor deseado, si tuvieramos que elegir uno de estos modelos, con seguridad elegiriamos el modelo de Linear Regression, incluso por encima de SVR. El SVR nos dio un error cuadratico medio menor, sin embargo, apelando al sentido comun, el cual siempre debe estar presente, verificamos y llegamos a la conclusion que no esta prediciendo bien.

VII PREDICCIONES

A la hora de realizar las predicciones, elegimos como nuevos input todos los meses de 2020 y todos los meses de 2021. De esta forma, el modelo nos lanzara las predicciones de cada mes, y de esta forma, sumando los valores, tendremos las predicciones para 2020 y para 2021.

Nos queda de la siguiente manera:



Son numeros que a simple vista son razonables, y nos da una leve nocion de que el modelo sigue un criterio coherente a la hora de predecir.

VIII CONCLUSIONES

A la hora de analizar las publicidades actuales ya existentes, podemos notar que mientras transcurren ciertos fenomenos en la sociedad se puede observar un aumento o disminución del importe en la publicidad. Algunos de estos ejemplos son epocas de elecciones tanto a nivel nacional como municipal u eventos importantes que hacen que haya muchos televidentes observando un mismo (o distintos) canal(es) de televisión.

Por otro lado, estamos seguros que este modelo no estaria cerca de predecir la publicidad invertida en el año 2020. Este ha sido un año atipico, y esta nueva normalidad, nunca antes habia sido plasmada en datos, siendo la cuarentena y la pandemia, un nuevo escenario global, impredescible por ningun modelo de Machine Learning, por mas buenas predicciones que este hubiera lanzado.

Por otro lado, teniendo en cuenta las predicciones realizadas con los modelos lineales podemos llegar a la conclusión de que la alta variabilidad de los datos mes a mes, generan un gran error al utilizar modelos sub-ajustados. Si bien aun no hemos probado modelos sobre ajusados, creemos que si se implementara, podríamos llegar a tener un bajo error en el entrenamiento, pero al momento de testear nos daría un alto error ya que es díficil encontrar una función que explique de forma lo más correcta posible la variabilidad de los datos analizados.

VII REFERENCIAS

[1] Laguna, C (2014). Correlacion y regresion lineal. *Instituto Aragonés de Ciencia de la Salud*, 1-18

[2] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[3] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). *Support vector regression machines. In Advances in neural information processing systems* (pp. 155-161).