EXAMEN DE PREDICCIÓN

Máster en Data Science 2020-2021

Valentina Díaz Torres
22-01-2021
CUNEF

SEGUNDA PARTE

Introducción

El presente análisis trata sobre un tema que atrae a numerosos científicos habitualmente, este es, la detección de suicidios. Es decir, la ciencia busca encontrar datos adicionales que ayuden a predecir suicidios antes de que estos tengan lugar.

Google, recoge información sobre las búsquedas de las palabras "suicidio" y "depresión, las cuales podrían estar muy relacionados con el hecho de que ocurra un suicidio. Unos científicos demuestran que una tiene más relevancia que la otra, pero las opiniones no iguales. Es por eso, que se estudia la influencia de ambas búsquedas en el número de suicidios.

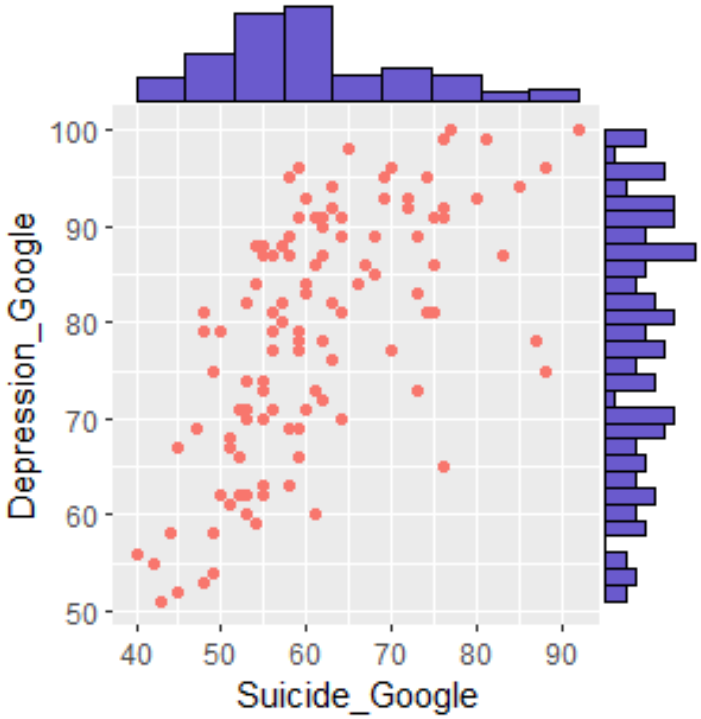
El objetivo de esta práctica será la predicción de una serie temporal, estudiando los años de los que se tiene información, con el fin de crear una predicción a corto plazo.

Información sobre los datos

Los datos recogen información mensual sobre el número de suicidios en Inglaterra, desde enero el año 2004 a diciembre de 2013, un total de 120 meses. La base de datos está compuesta por cuatro variables, en primer lugar, una columna de las fechas ya nombradas, otra que recoge el número de suicidios mensuales y dos columnas más, que recogen las búsquedas de la palabra suicidio y depresión respectivamente.

Se ha realizado el siguiente gráfico para observar cómo se distribuyen ambas búsquedas con el fin de tener una perspectiva general de los mismos.

SEGUNDA PARTE



Los datos de suicidios mensuales de todos los años son los siguientes:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2004	468	367	369	394	394	377	390	401	400	367	350	342
2005	376	323	421	386	377	342	367	350	387	352	365	351
2006	367	338	353	331	365	346	369	330	373	382	318	302
2007	345	345	326	383	381	404	359	338	339	320	317	330
2008	383	354	392	382	419	394	413	368	365	375	315	392
2009	409	315	329	371	373	357	360	353	363	372	369	325
2010	331	343	368	375	363	378	389	329	362	383	362	354
2011	386	354	391	390	377	379	363	369	333	324	355	413
2012	410	415	456	370	413	376	424	425	355	415	363	388
2013	442	345	387	411	379	385	424	390	366	366	345	378

Desarrollo del estudio

SEGUNDA PARTE

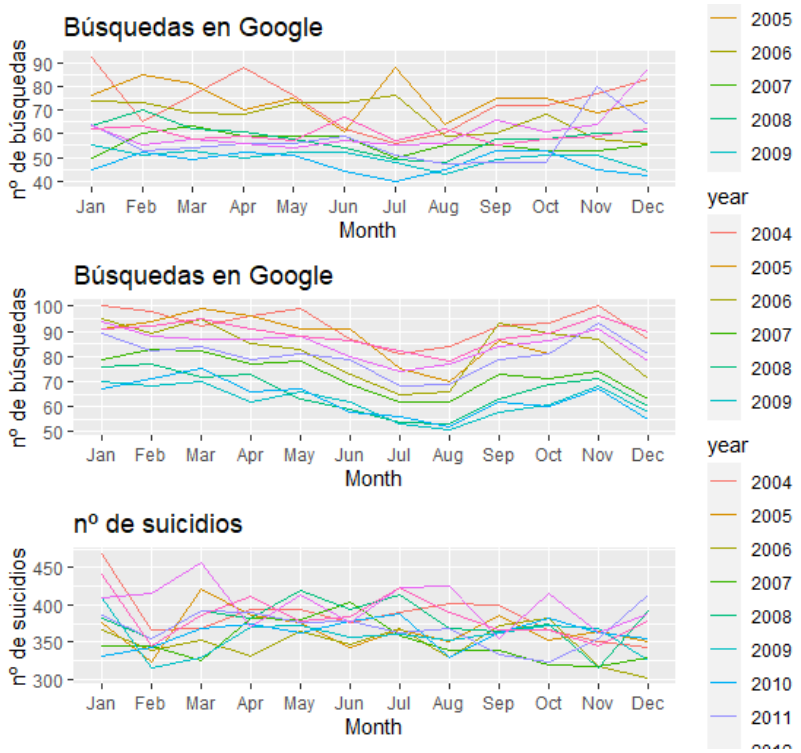
El objetivo principal es crear dos modelos, por un lado, el modelo base, compuesto por la variable que recoge el total de suicidios mensual y por otro, el modelo que aporta las búsquedas de Google. Esto es fundamental para poder comprobar si efectivamente esta información ayuda a predecir suicidios y mejora el modelo, haciéndolo más preciso o si, por el contrario, estos datos no aportan nada al estudio de los suicidios.

En primer lugar, ha sido preciso pasar cada una de las variables con las que se va a trabajar a series temporales, con el fin de poder formar modelos a lo largo del tiempo. La representación de la serie temporal de los suicidios y las búsquedas es la siguiente:

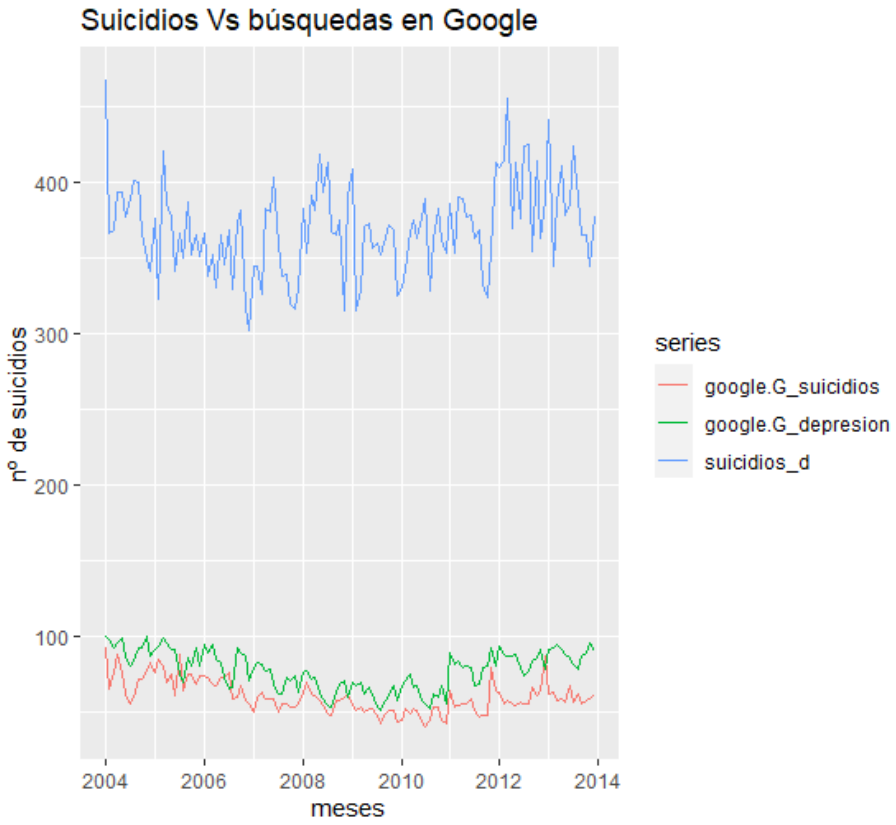
En estas gráficas se ve, que ambas series de búsquedas, parecen no tener estacionalidad, lo cual se comprueba más tarde con un análisis de estacionalidad. Además, se observa como el número de suicidios es muy superior con respecto al número de búsquedas.

En cuanto a la estacionalidad, como se ha comentado, tras realizar un test, ambas distribuciones de las búsquedas, no tenían una media estacionaria, por lo que fue necesario hacerla estacionaria, para poder realizar los modelos. Esto se realiza mediante la prueba de Dickey Fuller, donde aparece que ambas variables de búsquedas no son ruido blanco, que su media estacionaria no es cero, la varianza no es constante y el valor de p valores es mayor a 0.5.

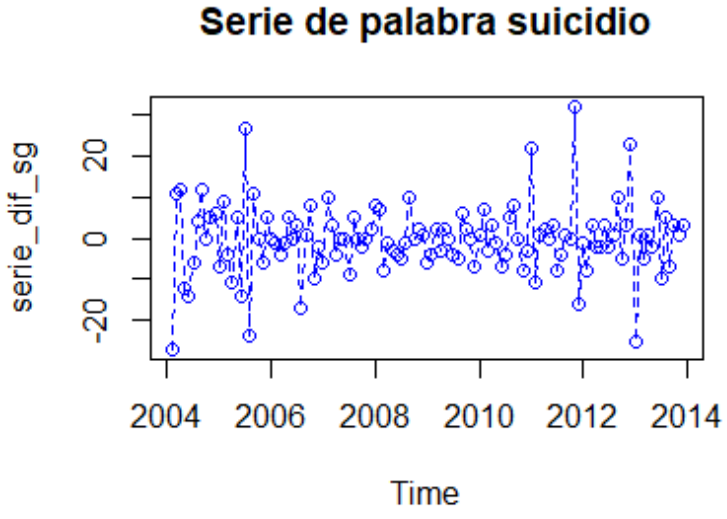
Para solventar este problema, en primer lugar, se probó a realizar el logaritmo de ambas, no obstante, los resultados obtenidos no fueron los correctos, ya que seguían no siendo estacionarias. Por eso, hizo falta la realización de una diferencia de cada una de las variables.

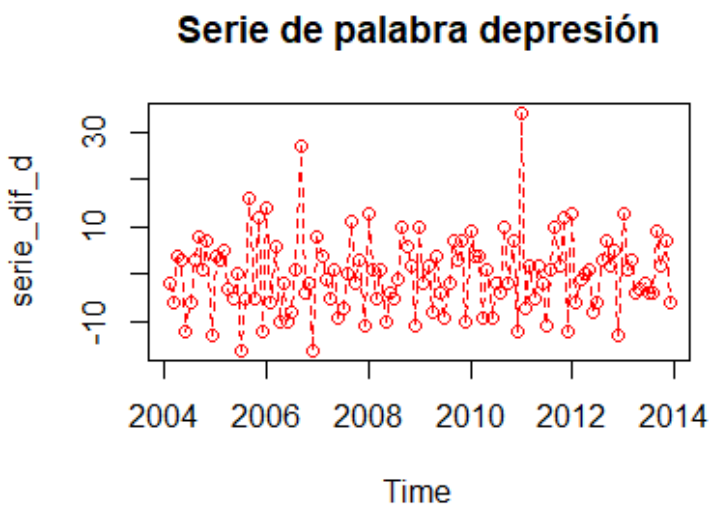


SEGUNDA PARTE



Cuando se realizó de nuevo el test de Dickey Fuller, estas ya lo eran, por lo que se podía proceder a la creación de los dos modelos de los que se ha hablado al principio. También se realizan pruebas de autocorrelación y autocorrelación parcial de cada variable, para conocer la media móvil de cada variable y sus autoregresivos. En las siguientes gráficas se muestra cómo ya existe estacionalidad:





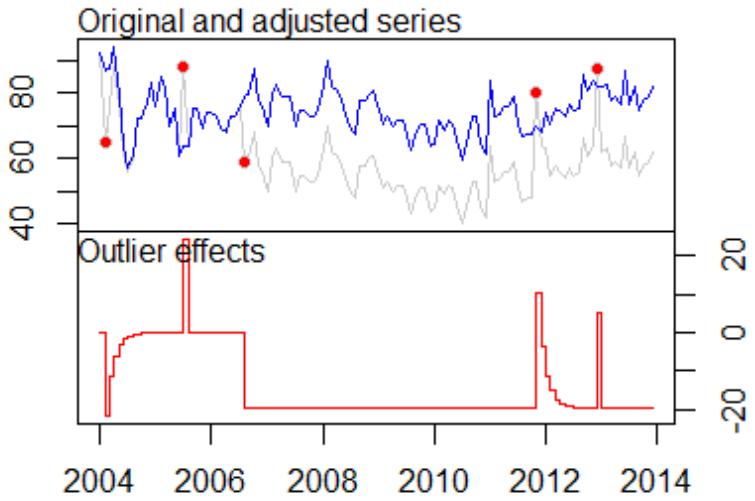
Tras ello se lleva a cabo la realización de los dos modelos, ambos modelos ARIMA, razón por la cual se necesita la estacionalidad de las series. En primer lugar, se crea el modelo ARIMA de suicidios, que compone el modelo base, mediante la función auto.arima. El resultado obtenido es (0,1,1) (1,0,0). El error porcentual absoluto medio (MAPE) obtenido es de 6.222639. En el caso del segundo modelo este es formado mediante la unión de las dos variables de búsquedas. Cuando esto se lleva a cabo, los resultados obtenidos son un ARIMA (1,1,1) (1,0,0) y un valor de MAPE de 5.985509. También el error RMSE baja de 28.74747, en el caso del modelo base a 27.97308.

Los resultados obtenidos muestran que los datos sobre las búsquedas de Google de la palabra “depresión” y la palabra “suicidio”, mejoran el modelo para predecir los suicidios, entre los casos estudiados. Se ha podido comprobar, según el test de error, como el modelo base, parte de un error, el cual, cuando se le añaden a un segundo modelo las variables de búsquedas de palabra depresión y suicidio, baja. Esto querría decir, por tanto, que tener en cuenta los datos de búsquedas recogidos por Google, ayudaría a predecir de una forma más precisa y con menor error los suicidios.

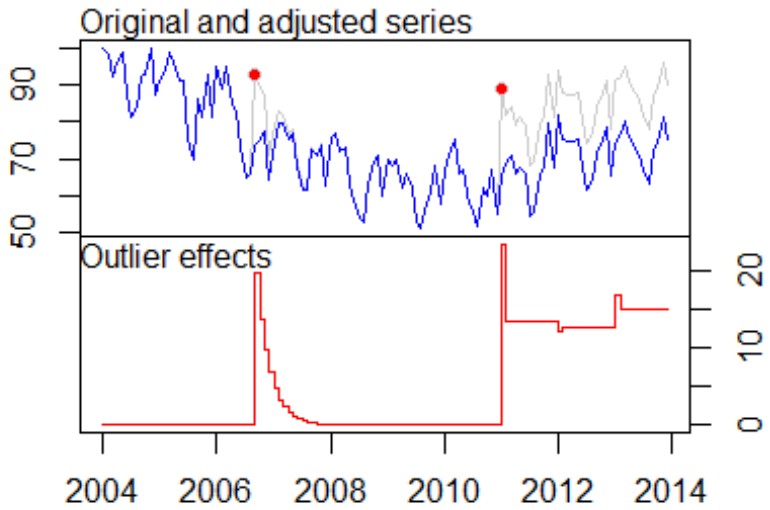
Estos resultados, con un MAPE cercano a 6, son muy similares con los obtenidos por los científicos. No obstante, se ha decidido indagar más allá. Debido a que se trata de una serie temporal, sería oportuno realizar una búsqueda de los valores atípicos u outliers. Una vez que estos se analizan, mediante un test de outliers, se ha detectado que tanto en las búsquedas de la palabra suicidio, como en la de depresión, existen algunos de ellos. En la palabra suicidio existen dos, que se establecen en los meses de septiembre de 2006 y enero de 2011, en ambos casos suponen picos más altos en la distribución. También, en el caso de la palabra depresión,

se han encontrado algunos, en este caso cuatro, en las fechas de febrero de 2004, julio de 2005, agosto de 2006, noviembre de 2011 y diciembre de 2012, el primero y el tercero un pico negativo, es decir, se detectó un valor atípico de menos incidencia y en el resto de casos mayor.

Estos son los valores atípicos en el caso de la palabra suicidio:



Estos son los outliers de la búsqueda de la palabra depresión en Google:



Estos podrían estar ocasionados porque son simplemente palabras que se buscan en internet, que pueden no estar relacionadas con un suicidio, también puede ser que se hayan buscado más veces una palabra por alguna tendencia, como la el suicidio de un famoso o una depresión.

Por lo tanto, estos fueron eliminado, construyéndose un nuevo modelo, esta vez sin valores atípicos en ambos casos, además de la variable de suicidios, que conformaba el modelo base y en la cual no se habían detectado outliers. La sorpresa de los valores ofrecidos por este nuevo modelo es que, en este caso, sin estos valores atípicos, no mejoraba el modelo base, es más, se obtenían valores muy cercanos o incluso mayores de los del modelo base. De este modo, se ha decidido que deshacerse de estos valores atípicos perjudicaría al análisis, por lo que se ha decidido conservar el primer modelo que se realizó.

Conclusión

En conclusión, se ha comprobado que efectivamente las búsquedas de Google de palabras relacionadas con suicidios, como en este caso son suicidio y depresión, tiene un impacto positivo en un modelo de serie temporal para predecir el número de suicidios. Si bien la diferencia del error no es demasiado grande, esta es suficiente como para suponer que ambos aportarían más información al modelo base, que solo tiene en cuenta el número de suicidios.

Los suicidios cada vez son más frecuentes, especialmente entre los jóvenes, es por eso, que hacer uso de los datos de internet, donde más están los jóvenes precisamente, podría ayudar bastante a predecirlos y poder usar este modelo para prevenir los mismosl