



# Mentor Data Science @ DEVF

Santiago Fernando Utrilla Lack

02

# Planteamiento del problema

Contamos con dos sets de datos los cuales fueron captados por algunos sensores de tecnología de tracking de la posición de celulares mediante señales de wifi.

La diferencia entre los dos archivos es que uno contiene la etiqueta de si el registro fue de un visitante o no, este archivo nos permitirá con un modelo de clasificación inferir y etiquetar los registros del segundo archivo.



# Datos

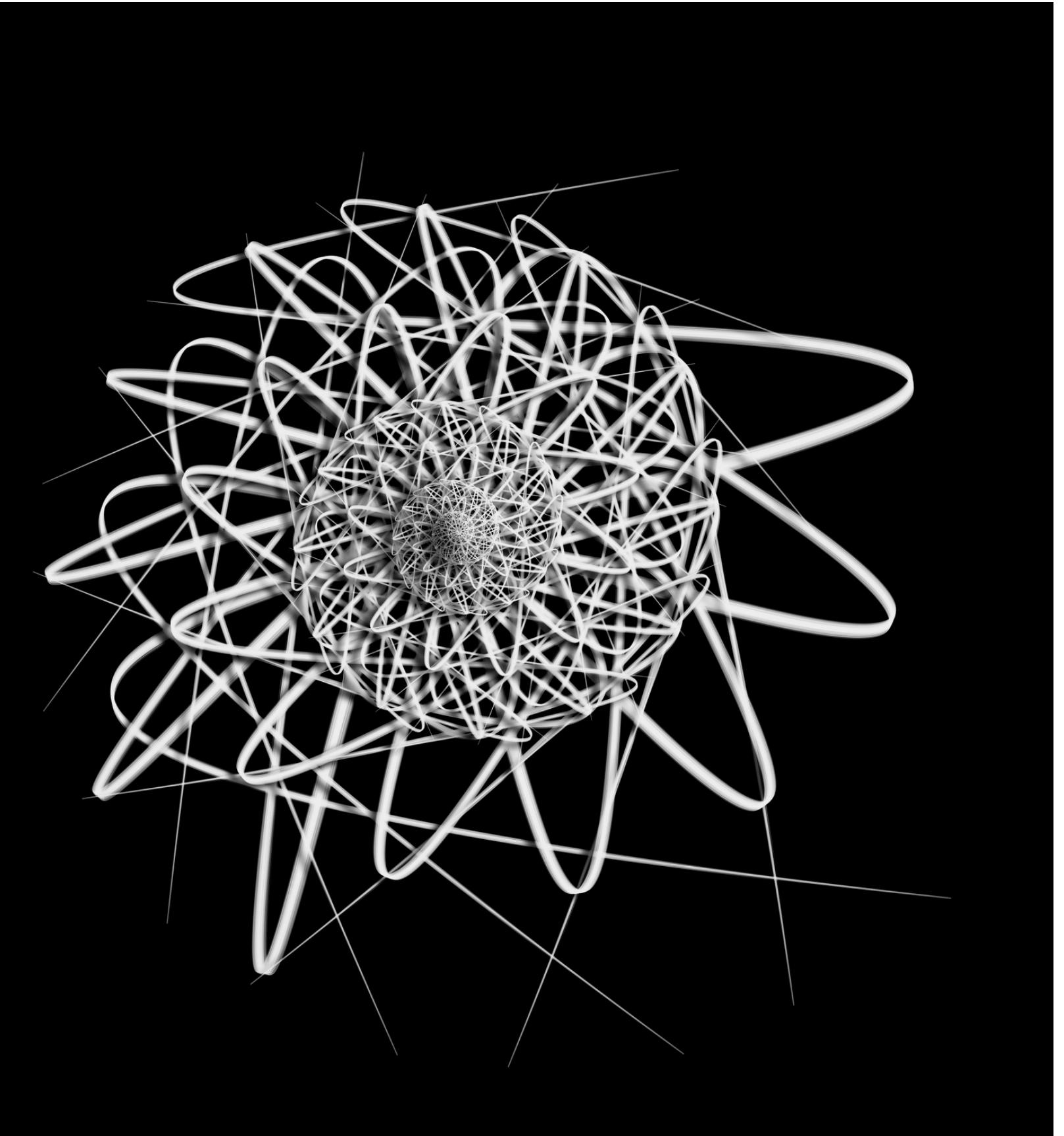
- Identificador único de un celular, cada célula asumimos es una persona (string).
- Sucursal del cliente (int), hay únicamente 3 sucursales.
- Variables que hacen referencia al tiempo como lo es el mes (Noviembre y Diciembre), días de la semana (Lunes a Domingo) y la hora del día (24h), haciendo referencia al momento de la visita.
- Duración de la sesión en segundos (int)
- Variable que nos indica si es visitante del lugar (bool)



## 04

El número de registros fue cercano a 25 mil datos, los cuales el 65.8% no eran visitantes y el 34.2% sí.

Los datos no contenían valores nulos y venían en un formato limpio por lo que no se requirió hacer ningún tipo de limpieza profunda más que la manipulación de algunos datos referentes a fechas para poder observar comportamientos a través del tiempo.

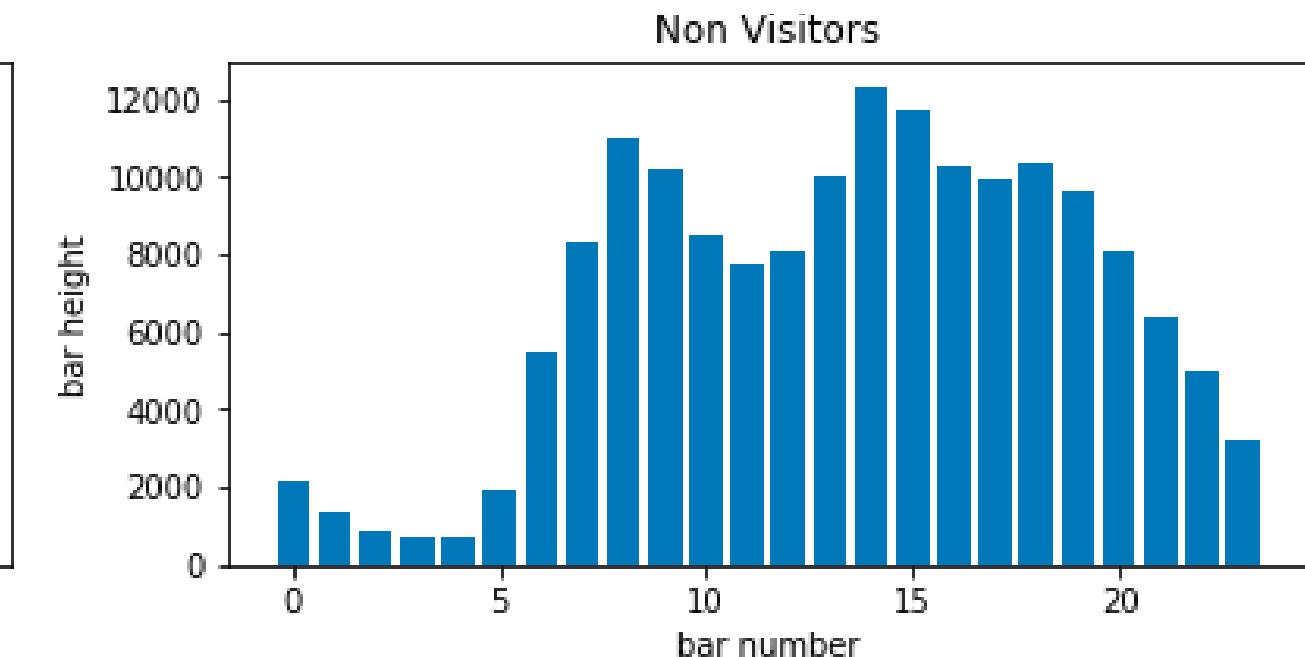
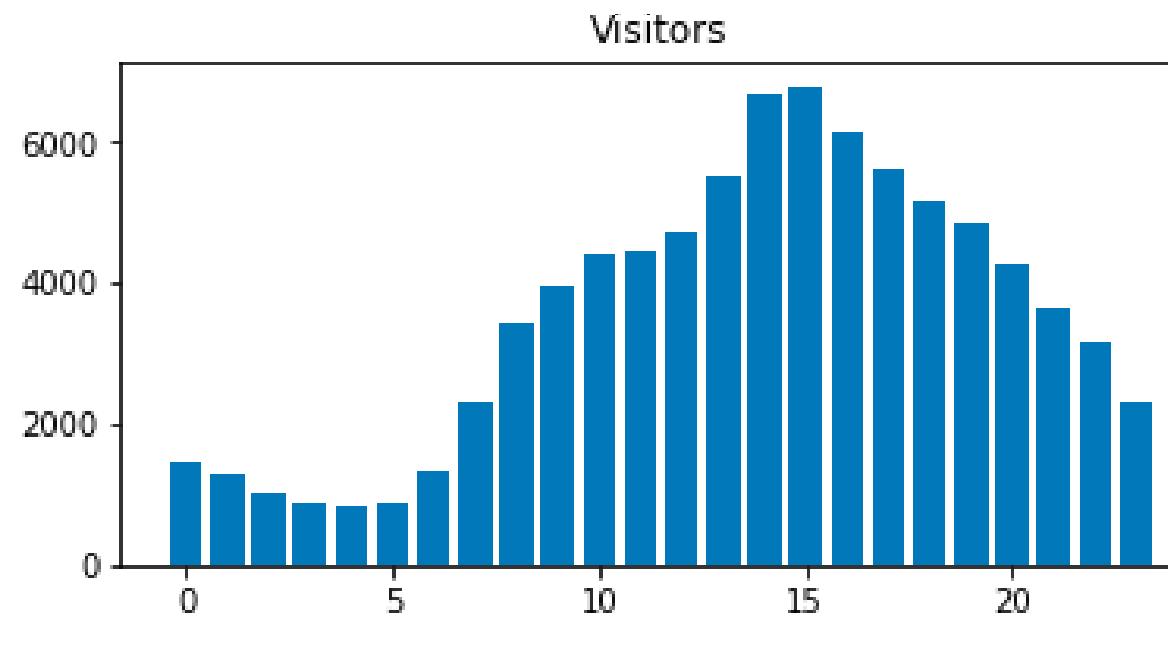
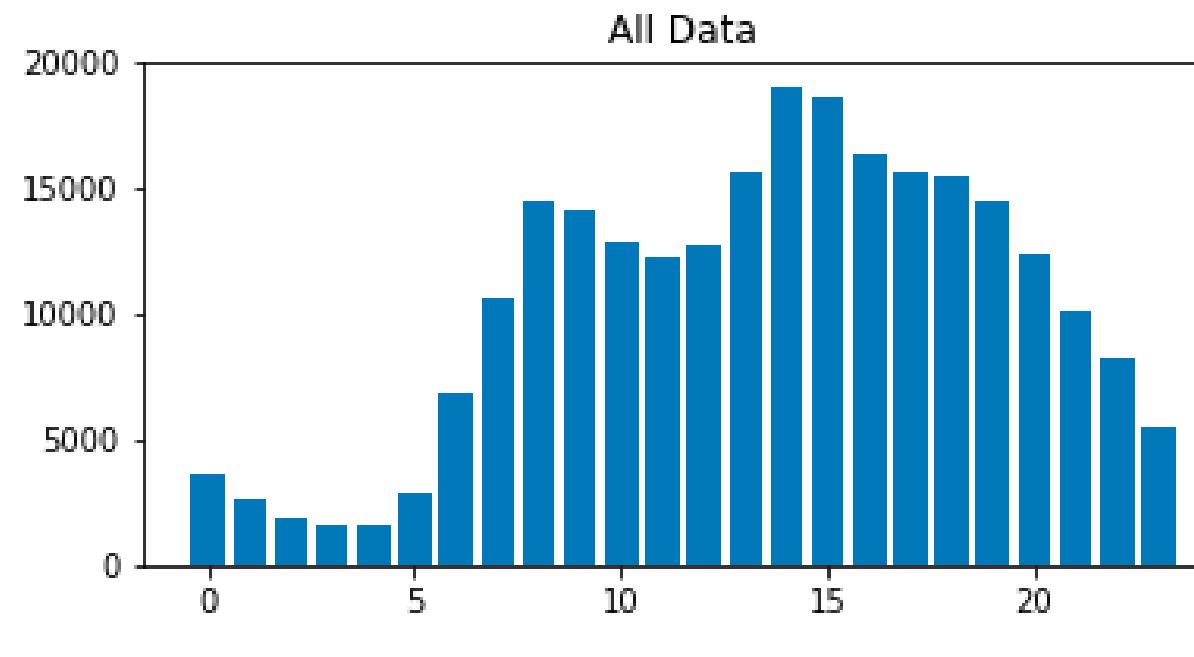


05

# Exploración

Aunque la proporción de visitantes es pequeña comparada con la de los no visitantes, mantienen una tendencia en el tráfico muy parecida excepto entre las ocho y nueve de la mañana donde podemos ver un pico de los no visitantes.

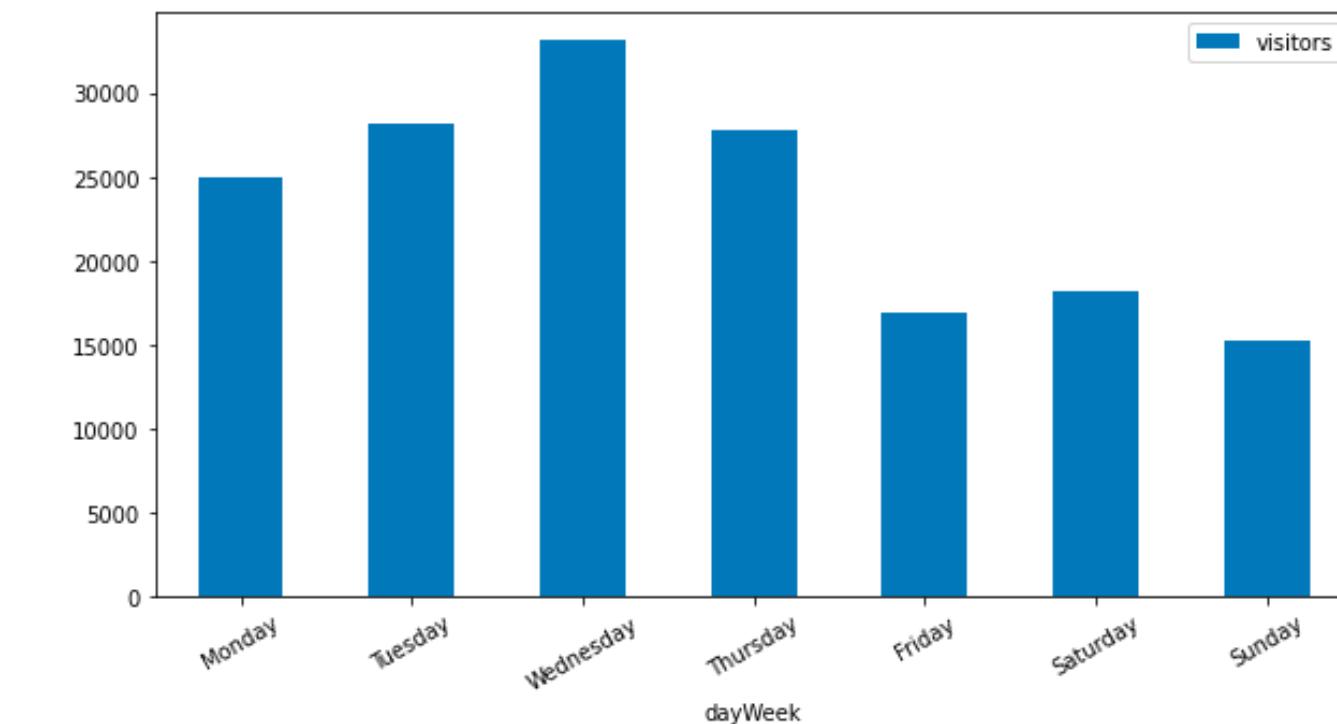
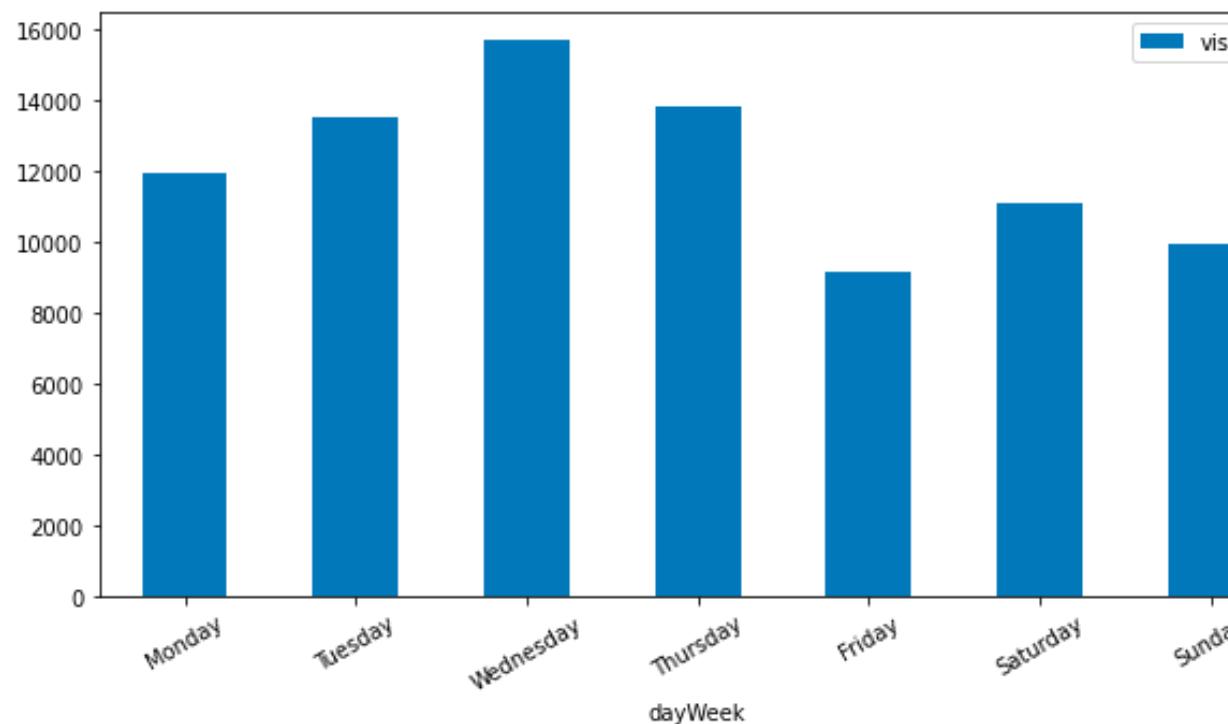
Fuera de esto no hay ningún otro factor visual que nos pueda hacer pensar que hay relación entre la hora y la veracidad de la visita.



06

# Exploración

En la gráfica del lado izquierdo podemos observar qué parece que el volumen de visitantes aumenta ligeramente el fin de semana pero no considero que se podría decir que a simple vista podamos encontrar una diferencia significativa entre el día de la semana de la visita.



07

# Inferencia / ML

Para poder predecir a partir de los datos, si es o no el registro un visitante se decidió utilizar un modelo de Machine Learning de Clasificación ya que resulta ideal para este tipo de problemas donde a partir de ciertas variables, se necesita decidir si el registro es verdadero o falso.

Primero se hizo la prueba con únicamente un arbol y posteriormente al ver los resultados positivos se hizo el entrenamiento de un conjunto de arboles con el algoritmo de Random Forest.





08

# Rendimiento del Algoritmo

El 96% de las veces el algoritmo determinó de manera correcta el resultado.

También tuvo una precisión del 96% para determinar los registros que efectivamente eran visitantes.

Accuracy = 96%

Precision = 96%

Recal = 92%

En general los resultados del algoritmo fueron muy buenos para haber sido el primer entrenamiento. Posterior a esto se podrían tomar diferentes acciones para mejorarlo como por ejemplo entrenar el modelo con la misma proporción de visitantes y no visitantes, identificar cuáles son las variables que tienen un mayor y menor impacto, conservarlas y descartarlas respectivamente.

09

# Conclusiones

La precisión del algoritmo es muy buena en proporción, pero le es imposible con el entrenamiento que se realizó identificar a un visitante en caso de que la duración de su sesión sea cero, por lo que se recomendaría continuar en identificar como resolver este problema entrenando otros algoritmos o agregando/quitando variables o registros.



DEV.FL

Santiago Utrilla



# 10 Agradecimientos

De todo corazón les agradezco mucho. Como en cada proyecto se aprenden cosas nuevas y este no fue la excepción. Disfrute mucho el desarrollo del código y la presentación, y por eso les doy las gracias.

Me encantaría poder colaborar con ustedes porque desde esta parte del proceso ya se siente la manera tan disruptiva de hacer las cosas y estoy seguro haríamos un gran equipo en el que podríamos lograr el mayor beneficio para ambas partes.

¡Muchas gracias!

Santiago Fernando Utrilla Lack

**DEV.FE**