Desafío Swiss Medical

Alejandro Quiroga Alsina

12 de septiembre de 2022

DigitalHouse >
Coding School

Introducción

- Swiss Medical nació en 1989 con la construcción de la Clínica y Maternidad Suizo Argentina.
- La empresa ofrece servicios de medicina privada y seguros a empresas (a sus empleados) y a individuos.
- Su misión y filosofía es: “Cuidar la salud de nuestros clientes brindando un servicio integral de máxima calidad”.
- Ofrece más de 60.000 profesionales de todas las especialidades y más de 18.300 prestadores de diagnóstico y tratamiento, junto a las mejores clínicas de internación del país.
- Swiss Medical Group es unos de los principales grupos de Argentina dedicados a la protección de personas, compuesto por 8 clínicas, 12 centros médicos ambulatorios, ECCO Emergencia y Prevención, SMG Cells, Swiss Medical Seguros, Instituto de Salta, Fundación Swiss Medical y Blue Cross & Blue Shield en Uruguay.

Objetivos

La empresa tiene el objetivo de depurar su base de contactos, obtenida de diversas fuentes de información de empresas pertenecientes al grupo.

Esta base consta de una lista de correos electrónicos (emails) y números de teléfono ordenados por cliente y numerados (taggeados) por unicidad de adquisición.

La empresa necesita entender la calidad de sus datos con la finalidad de generar un mecanismo fiel de seguimiento y contacto con sus clientes.

El objetivo del trabajo es el siguiente:

1. Crear un scoring o puntuación sobre la similitud de los datos de contactos de una persona (email y teléfono) en los diferentes sistemas de la empresa.
2. Este **índice de contactabilidad** debe ser de fácil comprensión y uso automatizable.

Un primer desglose del objetivo

A partir del análisis de los datos (**500.000 registros**), se replantea el objetivo buscando una respuesta más efectiva:

1. Se necesita entender, de los emails, cuáles poseen **dominios activos válidos**. Este es un problema importante, dado que:
   a) Los dominios en Internet cambian frecuentemente dejando dominios inválidos o introduciendo nuevos dominios que reemplazan anteriores.
   b) Existen muchos dominios registrados (válidos) que son similares a dominios correctos. Esto busca aprovechar los errores de tipeo para capturar usuarios válidos, con el objetivo de construir bases de spam.
2. Se necesita entender, de los emails, cuáles poseen **usuarios confiables**. Esto se debe a que durante el ingreso de datos puede haber existido un error de tipeo y la verificación de la validez de un usuario es cara y problemática.
3. Lo mismo sucede con los teléfonos asociados a cada usuario.

Nota importante: *los datos de email y teléfono que se muestran son ficticios.*

Etapas de trabajo:

1. Emails: dominios
2. Emails: usuarios
3. Teléfonos

DigitalHouse >
Coding School

1. Emails: Dominios

DigitalHouse >
Coding School

Validación de los dominios en los emails

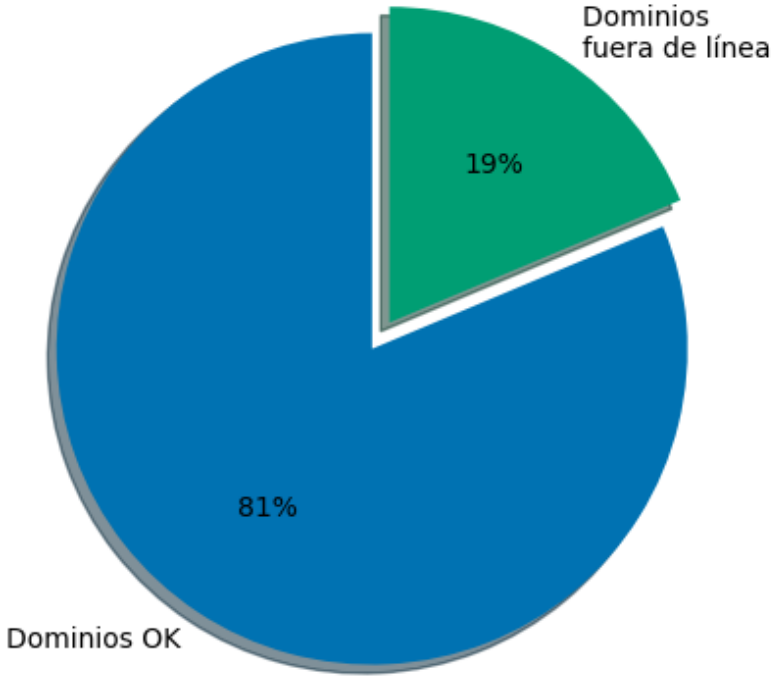
El dominio de un email es la parte de la dirección que está después del símbolo **@** (usuario@**dominio.com.ar**):

- Es conveniente "**validarlos**", es decir confirmar que son dominios existentes y correctamente configurados para recibir emails.
- Esta tarea no es compleja y agrega mucha información sobre la **contactabilidad** de una dirección de email.
- Del total disponible de **500.000** emails de contacto se extrajeron **5300** dominios únicos, esto es, en posibles condiciones de recibir emails.
- De esa cantidad se evalúa cuáles son válidos y se les asigna un scoring o **índice de contactabilidad**.
- Esto no garantiza per se que la dirección de email sea correcta porque el usuario puede estar mal escrito o no existir en destino. El análisis de contactabilidad de los usuarios se hace más adelante en este trabajo.

Validación de dominios “on-line”

En la primera etapa del análisis se verificó únicamente la existencia de cada dominio.

- Sobre un total de **5300** dominios se encontraron **4200** correctamente configurados para recibir correos electrónicos (Registros MX válidos).
- Entre los restantes se encontraron un total de **150** dominios válidos de sitios web, puesto que en los sitios mal configurados los correos electrónicos poseen un mecanismo de recupero (fallback) que consiste en utilizar los **Registros A** válidos como destinatarios.



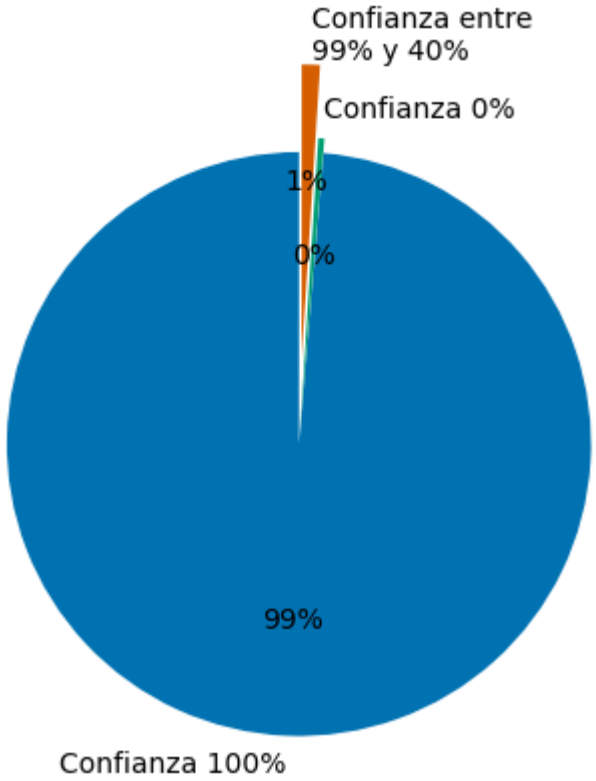
Entre los dominios “correctos”

Hay un total de **137.500** direcciones de email únicas.
Comparten sus dominios entre los llamados “**on-line**” o “**correctos**” y los llamados “**off-line**” (confianza de **0%**).

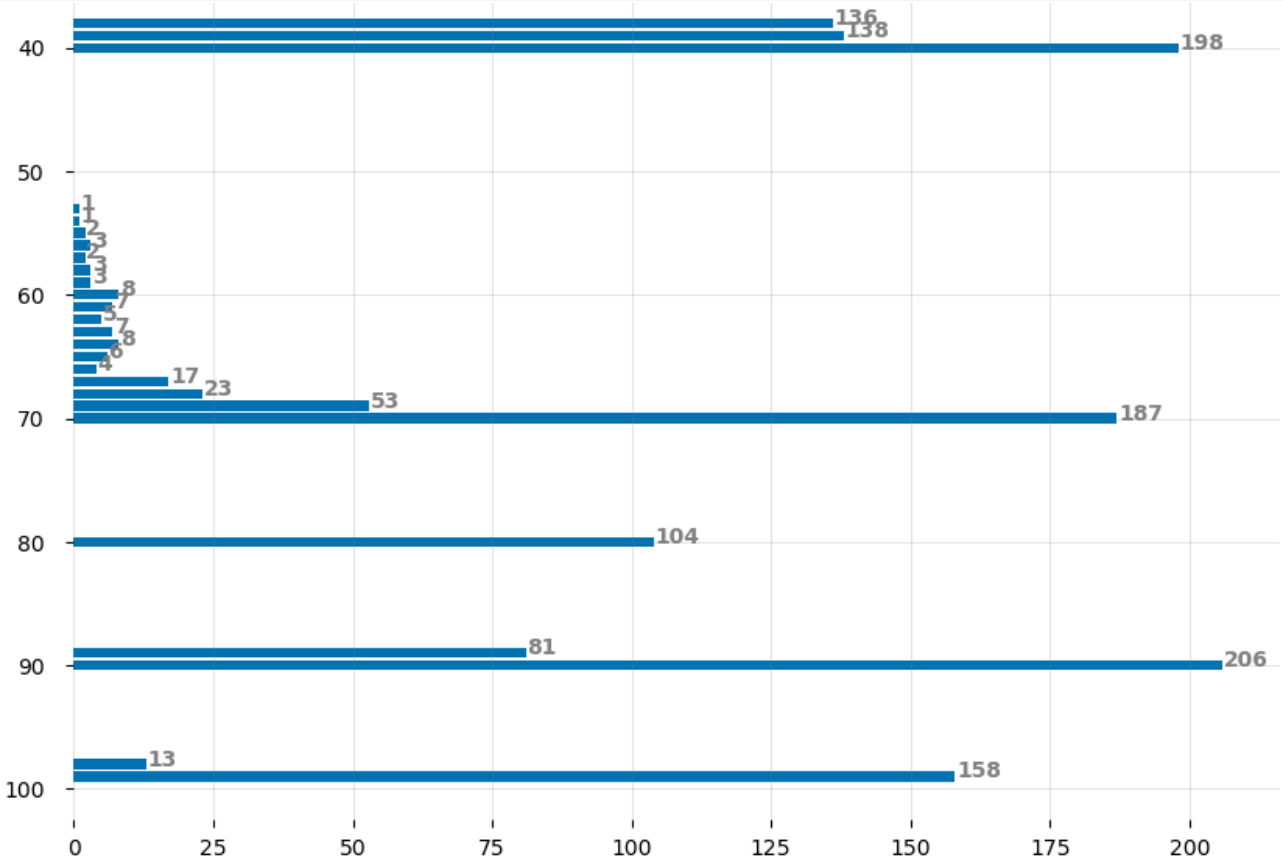
Los dominios **off-line** fueron reasignados con diferentes criterios logrando una cantidad de direcciones no válidas muy baja sobre el total: **500**).

Entre las direcciones con dominios “**correctos**” hay en realidad distintos niveles de confianza...

- Hay un total de **135.600** direcciones de email únicas con dominios con una confianza de **100%**.
- Entre las restantes **1400** direcciones se asignaron distintos niveles de confianza, tal como se muestra a continuación...



Dominios corregidos (se muestran subtotales)



- **Confianza media**, **40%**, fueron asignados utilizando distancia de Levenshtein máxima de 3.
- **Confianza media+**, entre **70%** y **50%**, Asignación de confianza por dominio válido dentro del **mismo cliente**. *Una menor confianza significa una mayor distancia de Levenshtein.*
- **80%**, cambios de dominio y fuentes de spam, como ".com.com"
- **90%** fuentes de spam corregidas, como "hotmal.com" vs "hotmail.com".
- **99%** y **98%** son dominios válidos, pero sin registro MX configurado.

2. Emails: Usuarios

DigitalHouse >
Coding School

Análisis de los usuarios en los emails

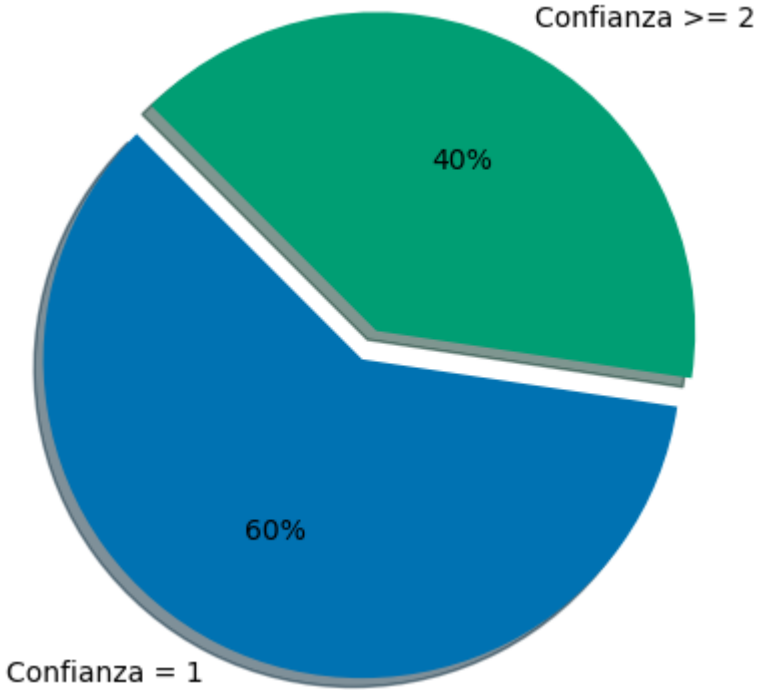
El usuario de un email es la parte de la dirección que está antes del símbolo **@** (**usuario**@dominio.com.ar):

- Es conveniente "**validarlos**", es decir confirmar que son usuarios existentes y correctamente configurados para recibir emails.
- Esta tarea **sí es compleja** y es generalmente lenta y onerosa (cuesta dinero), por lo que hay que realizarla para la mínima cantidad posible de usuarios.
  (ver en https://www.emailhippo.com/products/more#PricingCalculator)
- Del total de **500.000** emails de contacto se extrajeron **137.500** emails únicos.
- Luego se evalúa cuáles de sus usuarios son posiblemente válidos y se les asigna un scoring o **índice de contactabilidad** contando cuántas veces se repite en los datos.
- Aún así, la dirección de email puede no existir porque el usuario puede estar mal escrito. Pero para aquellos usuarios en los que el índice de contactabilidad es igual o mayor a 2, la probabilidad de que sea correcto aumenta.

Índice de Contactabilidad para usuarios

Se buscaron usuarios repetidos **siempre dentro del mismo cliente**.

- Sobre un total de **137.500** emails únicos se encontraron **82.000** usuarios que aparecen por única vez dentro de un cliente dado.
- Los usuarios que sí están repetidos dentro de un cliente permiten "dar menor valor relativo" a aquel usuario que posee un índice de contactabilidad de 1, y es muy parecido a otro de mayor índice (*siempre en el mismo cliente*).
  Por ejemplo, en el cliente 9203xxx:
  - **carinadie@hotmail.com** (índice = 2)
  - **cannadie@hotmail.com** (índice = 1)
- Igualmente no se descarta ningún usuario.



3. Teléfonos

DigitalHouse >
Coding School

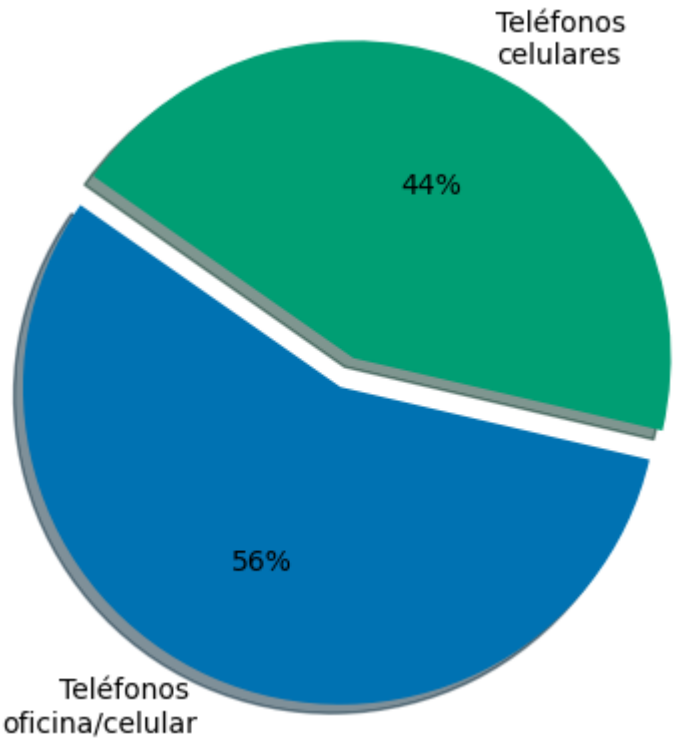
Validación de los teléfonos

Para cada cliente pueden existir varios teléfonos adquiridos desde los distintos sistemas de la empresa:

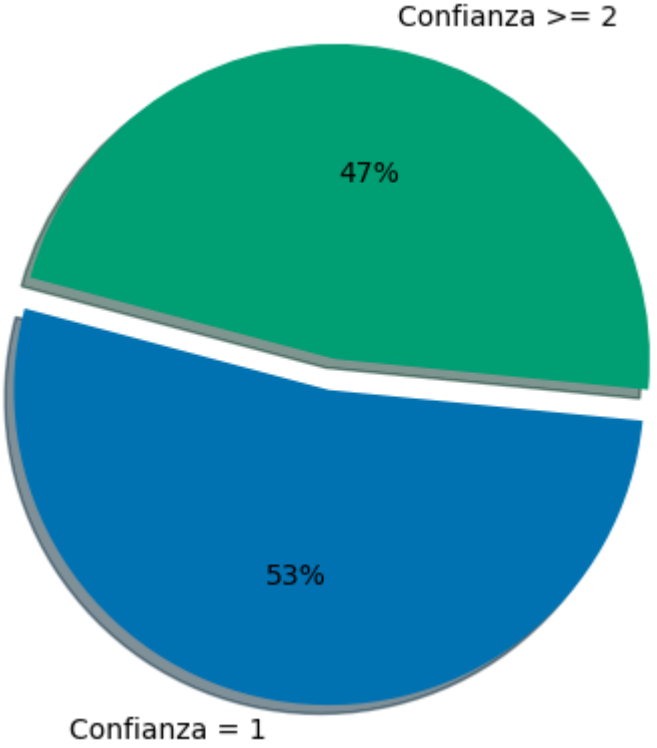
- Los teléfonos están asociados no sólo al cliente, sino también al email de una persona dentro de ese cliente.
- Los teléfonos pueden ser **celulares** o **fijos**. Los primeros son identificables porque llevan un “15” luego de la característica de área. Aquellos teléfonos que no tienen el “15” pueden a su vez ser fijos (oficina) o también celulares, dado que hoy en día no es necesario incluir el “15” al hacer la llamada a un celular.
- Del total disponible de **500.000** contactos se identificaron **164.000** teléfonos únicos, de los cuales **72.000** son celulares y el resto pueden ser de oficina o celular.
- A los teléfonos se les asigna un scoring o **índice de contactabilidad** basado en cuántas veces se repite dentro de los datos disponibles.
- No se descarta ningún teléfono, **salvo los que por su longitud son inválidos** (< 10 dígitos).

Algunas estadísticas sobre los teléfonos

- Entre los teléfonos únicos encontrados



- Índice de confianza de los teléfonos



Conclusiones

DigitalHouse >
Coding School

Conclusiones (1/2)

Se obtuvieron dos índices de contactabilidad:

- Para los emails su formato es el siguiente: "**indice de usuario @ índice de dominio**".

  Por ejemplo, un caso de distintos índices de usuario:
  - `llorenadie@hotmail.com`, **1 @ 100**
  - `lorenadie@hotmail.com`, **3 @ 100**

  Un ejemplo con índice de dominio distinto de 100:
  - `sofianadie@hotmail.com`, **2 @ 70**
    (*el usuario aparece dos veces, email original:* `sofianadie@jotmail.com`)

- Para los teléfonos su formato es únicamente un número que indica cuántas veces ese teléfono se encuentra repetido en los datos originales.

  - `1155533893`, **1**
  - `3415553893`, **4**

Nota importante: *los datos de email y teléfono que se muestran son ficticios.*

Conclusiones (2/2)

Se reconstruyó la base de datos original:

- En la nueva base de datos se listan las siguientes columnas:
  - Número de cliente.
  - Email, número de código único, fuente de adquisición, contactabilidad.
  - Teléfono, número de código único, fuente de adquisición, ¿celular?, contactabilidad.
- Los datos se entregan en el siguiente archivo: **`df_joins_final.csv`**
- Adicionalmente se entregan las siguientes bases de datos:
  - **`df_mails_final.csv`** (contiene los emails originales y los dominios evaluados)
  - **`df_phones_final.csv`** (contiene los teléfonos originales y su evaluación)
- También se entregan las herramientas con las que se trabajó, en formato de Jupyter Notebook (Python), separadas y numeradas en lotes de proceso.

Muchas gracias.

DigitalHouse >
Coding School