

Análisis de eficiencia de operadores - CallMeMaybe

PROYECTO SPRINT 14

ROCIOBOZOLAVIN@GMAIL.COM

TABLA DE CONTENIDOS

INTRODUCCION	1
Objetivos	3
Metodología.....	3
ANALISIS EXPLORATORIO DE DATOS.....	4
<i>Gráfico: Acumulación de valores Nan en operator_id a lo largo del tiempo.....</i>	<i>4</i>
<i>Gráfico: Histogramas de distribución por operadores.....</i>	<i>4</i>
<i>Gráfico: Duración promedio de llamadas por proporción de llamadas perdidas</i>	<i>5</i>
Clustering	6
<i>Gráfico: Método del codo</i>	<i>6</i>
<i>Gráfico: Scatter plot de clusters de operadores</i>	<i>6</i>
<i>Gráfico: Cantidad de llamadas por cluster</i>	<i>7</i>
<i>Gráfico: Promedio de llamadas perdidas por cluster</i>	<i>7</i>
<i>Gráfico: Duración promedio de llamadas por cluster</i>	<i>7</i>
<i>Gráfico: Duración promedio de una llamada</i>	<i>8</i>
<i>Gráfico: Mapa de calor de operadores por cluster y plan tarifario</i>	<i>8</i>
<i>Gráfico: Distribución del total de usuarios por plan</i>	<i>8</i>
PRUEBA DE HIPOTESIS ANOVA.....	9
<i>Gráfico: Mapa de calor proporción de llamadas perdidas vs duración promedio de una llamada</i>	<i>9</i>
<i>Gráfico: Scatter plot: Promedio llamadas perdidas por promedio de duración de llamadas según cluster</i>	<i>9</i>
CONCLUSIONES GENERALES	10
DASHBOARD	10

INTRODUCCION

El presente proyecto tiene como objetivo analizar el comportamiento de los usuarios de CallMeMaybe mediante técnicas de clustering, con el fin de identificar patrones de ineficiencia asociados a llamadas perdidas y duraciones atípicamente cortas.

Objetivos:

Mejorar el servicio de telefonía virtual CallMeMaybe a través de la identificación de operadores menos eficientes.

Metodología:

Datos utilizados:

- Datasets telecom_dataset_us.csv y telecom_clients_us.csv.

Procesamiento:

- Limpieza de datos, creación de métricas de eficiencia (duración promedio, proporción de llamadas perdidas).

Clusterización:

- Aplicación de K-Means Clustering para segmentar usuarios en grupos de comportamiento similar.

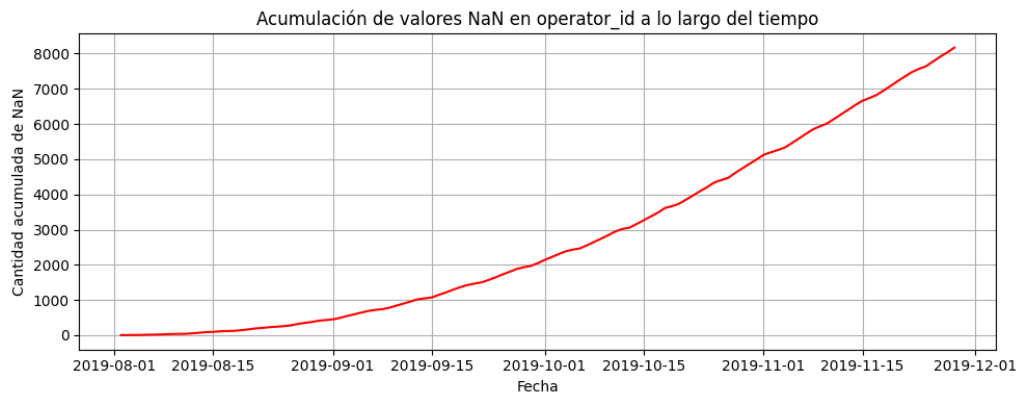
Prueba de hipótesis:

- Evaluación mediante ANOVA para comparar diferencias entre clusters

ANALISIS EXPLORATORIO DE DATOS

El proyecto comenzó con la limpieza de los datasets proporcionados (*dataset* y *clients*), en los cuales no se detectaron grandes discrepancias, excepto por valores ausentes en la columna *operator_id*. Estos valores, en más del 90% de los casos, correspondían a llamadas perdidas, como se evidencia en el siguiente gráfico, que muestra el aumento de valores ausentes a lo largo del tiempo.

Acumulacion de valores Nan en operator_id a lo largo del tiempo



De la grafica y de los datos podemos deducir que las llamadas perdidas pueden no ser registradas por el usuario al no ser contestadas, por ende, debemos evitar tener la mayor cantidad de llamadas perdidas posibles.

Luego exploramos el comportamiento de los valores atípicos y es acá que descubrimos que tenemos valores atípicos que se escapan muchísimo de la media, y de los cuales debemos comenzar a tener en consideración en el resto del proyecto para que no ensucien nuestros datos.

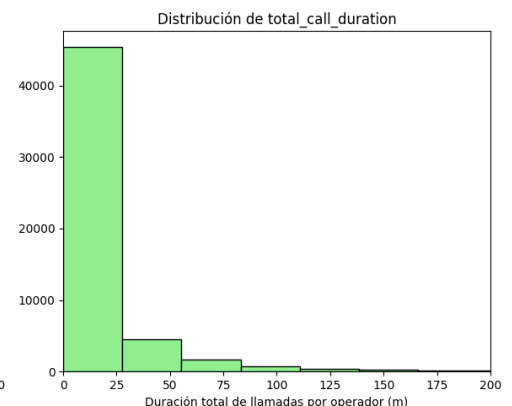
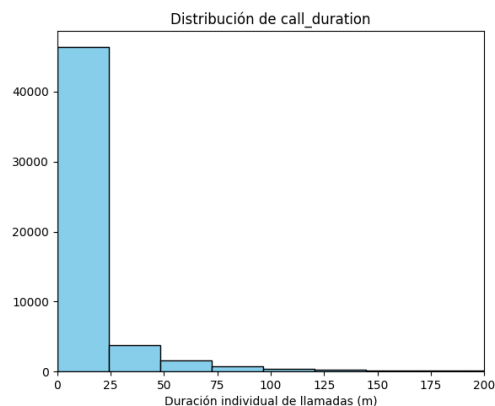
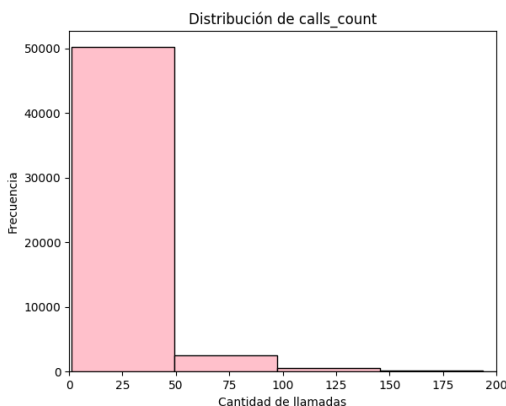
En la gráfica podemos ver las tres columnas mas importantes que nos ayudaran a definir a lo largo del proyecto a los operadores menos eficientes, acá vemos como los datos se escapan, aun cuando hubo una reducción del eje x en todos los gráficos.

Histogramas de distribución por operadores:

Cantidad de llamadas por operador

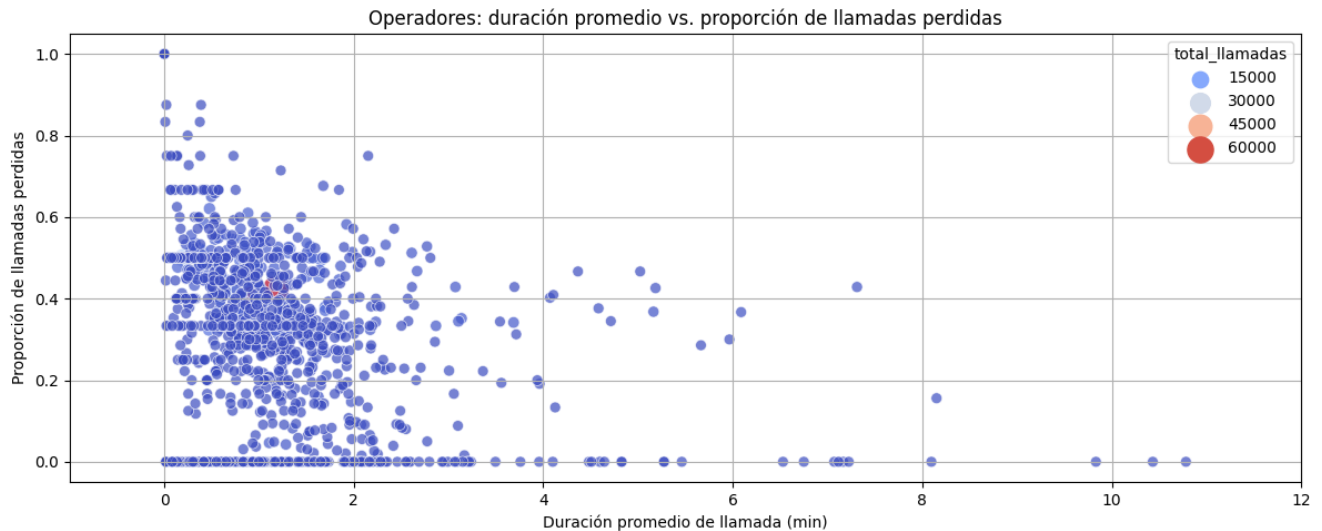
Duración de llamadas por operador duración

Total de llamadas por operador



Luego seguimos con el análisis de los operadores con respecto al promedio de llamadas perdidas y duración de las llamadas, a través de un scatter plot podemos ver que existen grupos de operadores que poseen demasiadas llamadas perdidas, o llamadas demasiado largas en comparación al resto de los operadores

Duración promedio de llamadas por proporción de llamadas perdidas



Es acá cuando entendemos que debemos enfocarnos en ese grupo de usuarios de llamadas demasiado cortas y porcentaje de llamadas perdidas demasiado alto. Estos hallazgos fundamentaron la necesidad de segmentar los usuarios mediante clustering.

Clustering

Previo a definir nuestros clusters, seleccionamos las variables con las que trabajaremos, las cuales nos permiten medir el comportamiento operativo de los operadores de forma cuantitativa y facilitan la división de los operadores basada en la eficiencia, estas variables representan la eficiencia tanto en términos de rendimiento como de carga de trabajo.

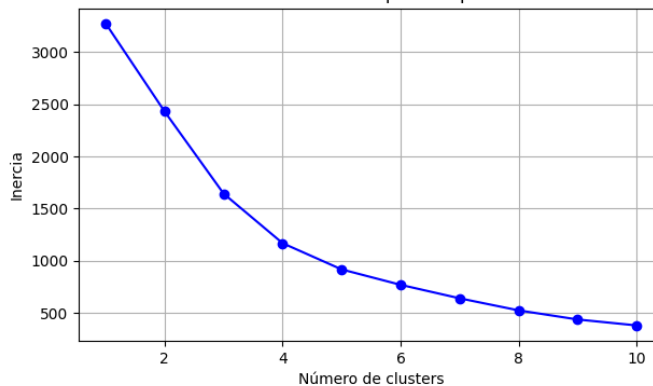
en base a esto se seleccionaron las siguientes tres variables:

- *Proporcion_perdidas*: proporción de llamadas no contestadas (ineficiencia).
- *Duracion_promedio_min*: duración promedio de llamadas (calidad/tiempo).
- *Total_llamadas*: volumen total de llamadas realizadas por operador (actividad).

Al siguiente paso lo nombramos selección de cluster, es aquí donde la curva nos muestra la cantidad de cluster con los que se sugiere trabajar, es donde la curva se estabiliza que definimos cuantos codos deben ser, en este caso serian cuatro codos.

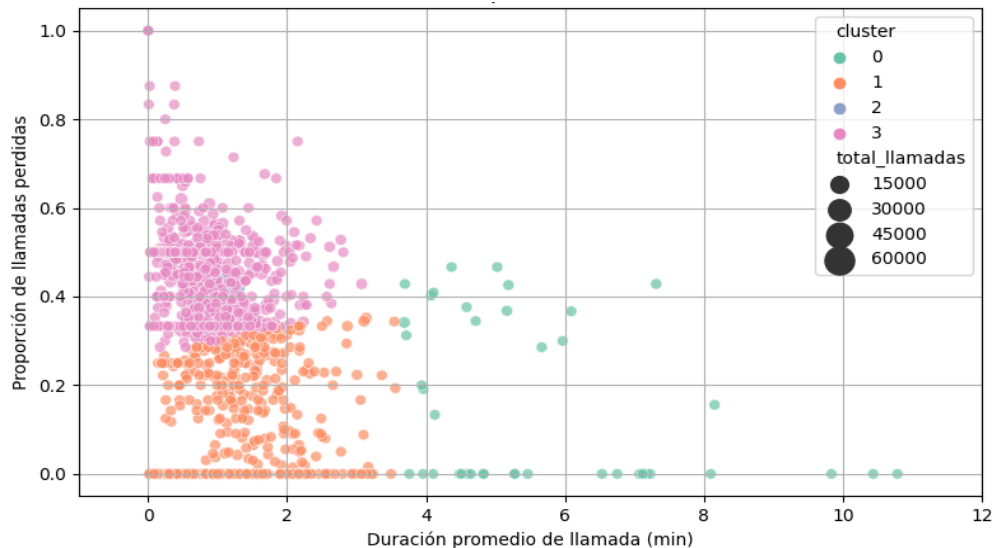
Método del codo

Método del codo para K óptimo



Luego de que se definen 4 clusters los cuales son los más adecuados a estudiar, los representamos en un scatter plot nuevamente.

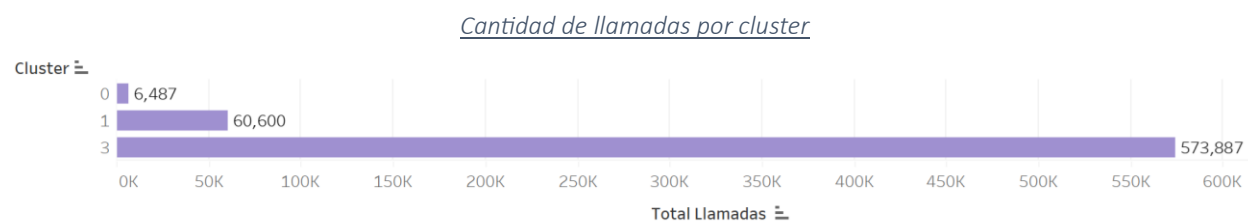
Scatter plot de clusters de operadores



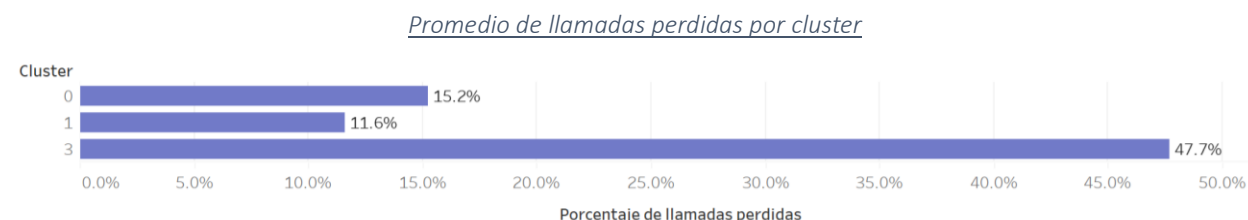
Durante el análisis de clusters, se identificó que el cluster 2 contenía únicamente dos operadores, ambos asociados a un mismo cliente, con volúmenes de más de 60.000 llamadas cada uno. Esto sugiere actividad de bots, motivo por el cual se decidió excluir este cluster del análisis posterior.

También descubrimos que el cluster 3 posee un comportamiento bastante malo con una muy alta tasa de llamadas perdidas, y una duración de llamadas muy pequeña. Lo que nos va dando indicios de donde podemos encontrar los operadores menos eficaces.

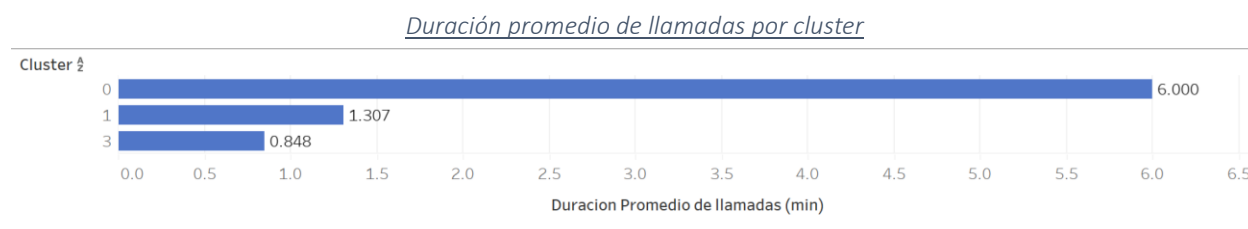
Luego seguimos indagando con respecto a los clusters y graficamos su comportamiento en los siguientes aspectos:



Podemos apreciar que el clúster 3 presenta un volumen de llamadas considerablemente mayor que el de los otros clusters. Es importante considerar que debido a esto probablemente posee mayor cantidad de llamadas perdidas

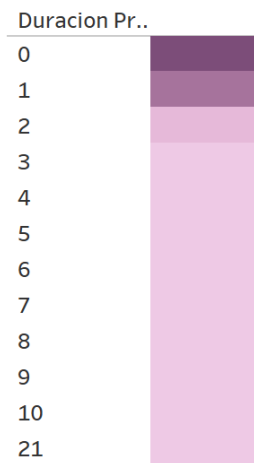


El análisis de la duración promedio de llamadas resulta crucial para este estudio ya que se espera como comportamiento del clúster 3 que, debido a la mayor cantidad de llamadas, sea mayor su promedio de duración.



En esta grafica es donde finalmente podemos apreciar el bajísimo promedio por llamadas que tiene el clúster 3 a pesar de la gran cantidad de llamadas que posee, lo que es un gran indicador de donde se encuentran los operadores menos eficientes También podemos decir que el análisis de la duración promedio reveló que el cluster 0 posee una duración considerablemente más alta en comparación con los otros clusters, aunque no necesariamente se traduce en mayor eficiencia.

Duración promedio de una llamada



Podemos ver este grafico en general que nos indica la duración de una llamada en promedio es de menos de un minuto, lo que podría ser fuertemente perjudicial para nuestro sistema, eso dependerá de cuanto queremos que nuestra llamada dure en realidad, pero podríamos decir que en un minuto representa más lo que buscamos en un operador.

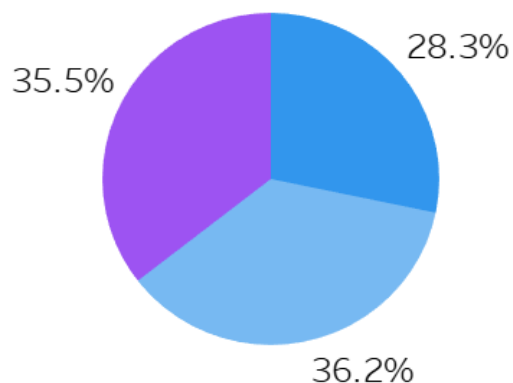
En paralelo también hicimos un breve análisis de los planes tarifarios de las llamadas y llegamos a la conclusión de no existe correlación entre el tipo de plan con la eficacia de los operadores. Ya que cada plan posee operadores de diferentes clústeres, sin excepción.

Mapa de calor de operadores por cluster y plan tarifario

Distribución de operadores por Cluster y Plan Tarifario

Cluster	A	B	C	Total
0	1	19	22	42
1	97	159	198	454
3	210	217	167	594
Total	308	395	387	1090

Distribución del total de usuarios por plan



En este grafico también podemos apreciar que la distribución de clientes por plan es si bien no exactamente igual no posee mayor diferencia entre ellas, lo que representa aun mas que no existe correlación que nos ayude a definir el grupo de operadores menos eficientes, ya que cada plan posee operadores de todo tipo.

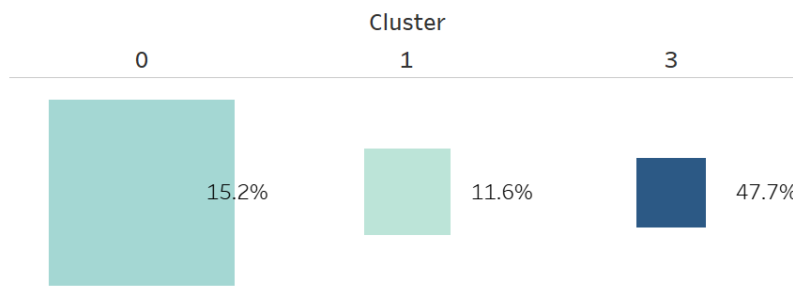
PRUEBAS DE HIPOTESIS ANOVA

En la prueba de hipótesis comparamos los 3 clústeres de los cuales obtuvimos resultados esperados, como que la media de los clústeres es altamente diferente entre si con respecto al promedio de llamadas perdidas y al promedio de duración de llamadas, lo que nos indica que independiente de la cantidad de operadores en cada clúster su comportamiento aun así no es proporcional a su cantidad, es decir a pesar de la densidad del clúster 3, no se justifica la gran perdida de llamadas perdidas o lo corto de duración de sus llamadas promedio.

Finalmente tenemos dos graficas claves que nos ayudan a cerrar nuestro proyecto,

Se desarrolló un mapa de calor donde el color representa la proporción de pérdidas y el tamaño de los círculos la duración promedio. El Cluster 3 destaca negativamente en ambas métricas.

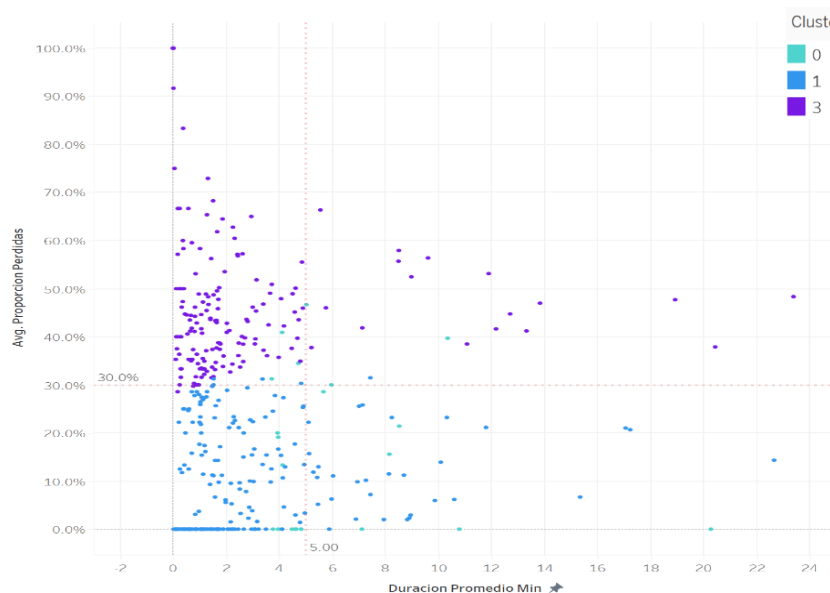
Mapa de calor proporción de llamadas perdidas vs duración promedio de una llamada



Mientras mas oscuro mas llamadas perdidas, mientras más pequeño meno promedio de duración de llamada

Y finalmente se realizó un gráfico de dispersión incluyendo líneas de referencia que marcan umbrales aceptables de duración y pérdida. El Cluster 3 se ubica fuera de las zonas óptimas, confirmando su bajo desempeño.

Scatter plot: Promedio llamadas perdidas por promedio de duración de llamadas según cluster



Conclusiones generales

El análisis revela que el **Clúster 3** presenta la mayor ineficiencia entre los grupos evaluados. Este clúster exhibe una **alta proporción de llamadas perdidas (48%)** junto con una **baja duración promedio de llamadas (0,85 minutos)**, lo que sugiere una gestión deficiente de las comunicaciones.

Adicionalmente, aunque el **Clúster 0** muestra una **duración promedio elevada (6 minutos)**, sus niveles de pérdidas no son alarmantes; sin embargo, llamadas excesivamente largas podrían implicar costos operativos innecesarios si no corresponden a un servicio de mayor calidad.

En conjunto, el **Clúster 3** destaca negativamente tanto por su alta tasa de pérdida como por la baja duración de las llamadas, evidenciando **un comportamiento claramente ineficiente**. Se recomienda enfocar esfuerzos de mejora y optimización en este segmento para incrementar la eficiencia general del servicio.

Acceso al Dashboard Interactivo

Para complementar el análisis presentado, se encuentra disponible un dashboard dinámico desarrollado en Tableau Public. Puede accederse a través del siguiente enlace:

[Dashboard de Eficiencia de Clusters - CallMeMaybe](#)

El presente análisis y el dashboard desarrollado buscan ofrecer herramientas prácticas para mejorar la eficiencia operativa de CallMeMaybe, enfocando esfuerzos en los segmentos de usuarios identificados como críticos.