Mariana Micaela Espejo Maldonado

Proyecto final

Informe de resultados

Análisis exploratorio

Para comenzar el análisis exploratorio de los datos, se utilizaron funciones como info(), describe(), dtypes, isnull(), para así obtener información detallada sobre los datos, ver con que tipo de datos estamos trabajando, observar si existen valores faltantes, obtener un resumen de estadísticas como el promedio, el mínimo y máximo de cada columna y demás. A continuación, se muestran las tablas obtenidas de esta primera parte.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3150 entries, 0 to 3149
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Call  Failure           3150 non-null   int64
 1   Complains               3150 non-null   int64
 2   Subscription  Length    3150 non-null   int64
 3   Charge  Amount          3150 non-null   int64
 4   Seconds of Use          3150 non-null   int64
 5   Frequency of use        3150 non-null   int64
 6   Frequency of SMS        3150 non-null   int64
 7   Distinct Called Numbers 3150 non-null   int64
 8   Age Group               3150 non-null   int64
 9   Tariff Plan             3150 non-null   int64
 10  Status                  3150 non-null   int64
 11  Age                     3150 non-null   int64
 12  Customer Value          3150 non-null   float64
 13  Churn                   3150 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 344.7 KB
```

***Tabla 1*.** Información y tipo de los datos.

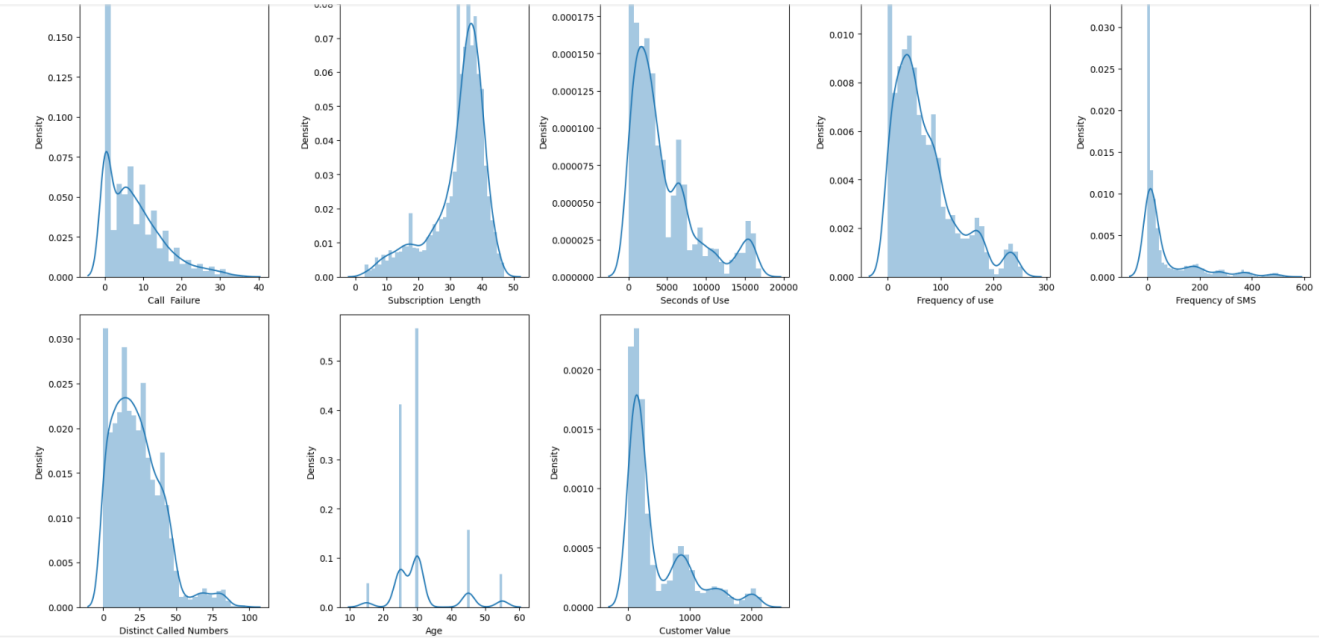
	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn
count	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	7.627937	0.076508	32.541905	0.942857	4472.459683	69.460635	73.174921	23.509841	2.826032	1.077778	1.248254	30.998413	470.972916	0.157143
std	7.263886	0.265851	8.573482	1.521072	4197.908687	57.413308	112.237560	17.217337	0.892555	0.267864	0.432069	8.831095	517.015433	0.363993
min	0.000000	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	15.000000	0.000000	0.000000
25%	1.000000	0.000000	30.000000	0.000000	1391.250000	27.000000	6.000000	10.000000	2.000000	1.000000	1.000000	25.000000	113.801250	0.000000
50%	6.000000	0.000000	35.000000	0.000000	2990.000000	54.000000	21.000000	21.000000	3.000000	1.000000	1.000000	30.000000	228.480000	0.000000
75%	12.000000	0.000000	38.000000	1.000000	6478.250000	95.000000	87.000000	34.000000	3.000000	1.000000	1.000000	30.000000	788.388750	0.000000
max	36.000000	1.000000	47.000000	10.000000	17090.000000	255.000000	522.000000	97.000000	5.000000	2.000000	2.000000	55.000000	2165.280000	1.000000

Tabla 2. Resumen estadístico de los datos.

```
Call  Failure              0
Complains                  0
Subscription  Length       0
Charge  Amount             0
Seconds of Use             0
Frequency of use           0
Frequency of SMS           0
Distinct Called Numbers    0
Age Group                  0
Tariff Plan                0
Status                     0
Age                        0
Customer Value             0
Churn                      0
dtype: int64
```

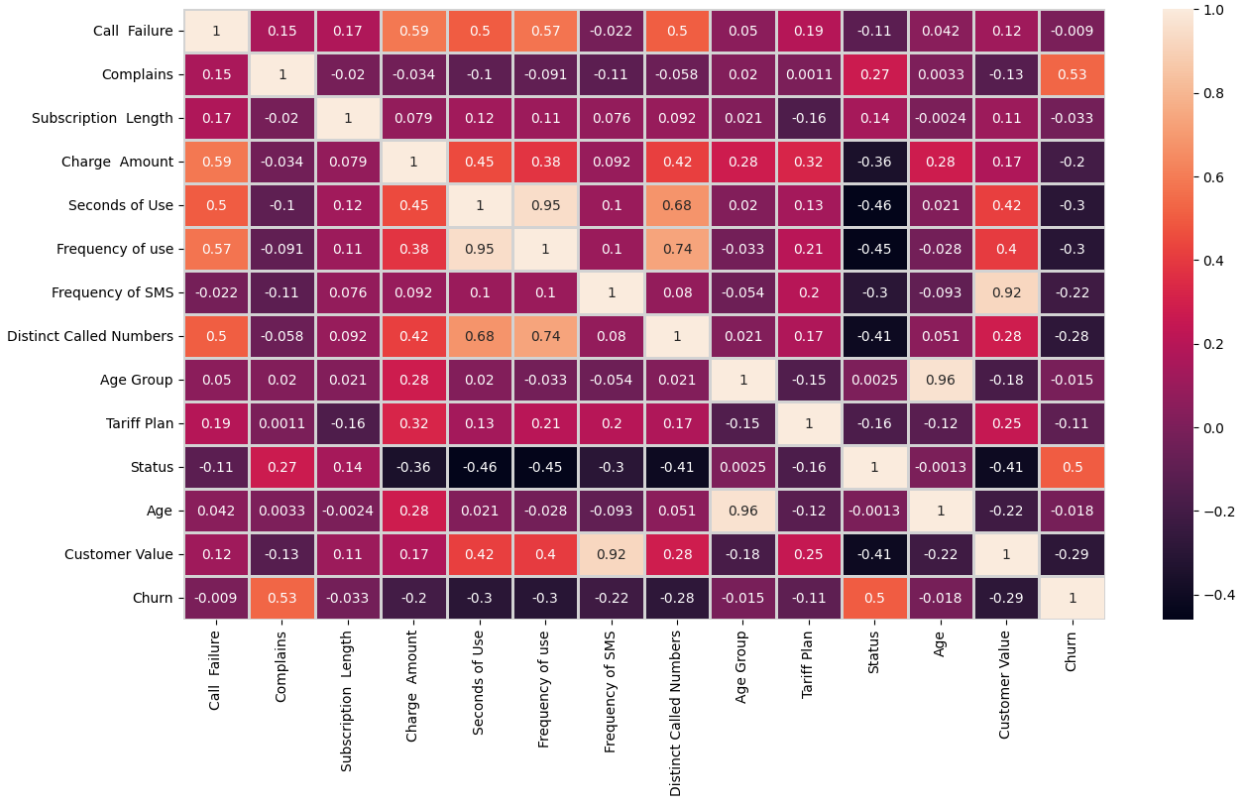
Tabla 3. Información sobre valores nulos. Observamos que no tenemos ningún valor faltante.

Seguidamente, se obtuvieron gráficas de distribución de las variables numéricas que tenemos, para así poder observar su comportamiento.



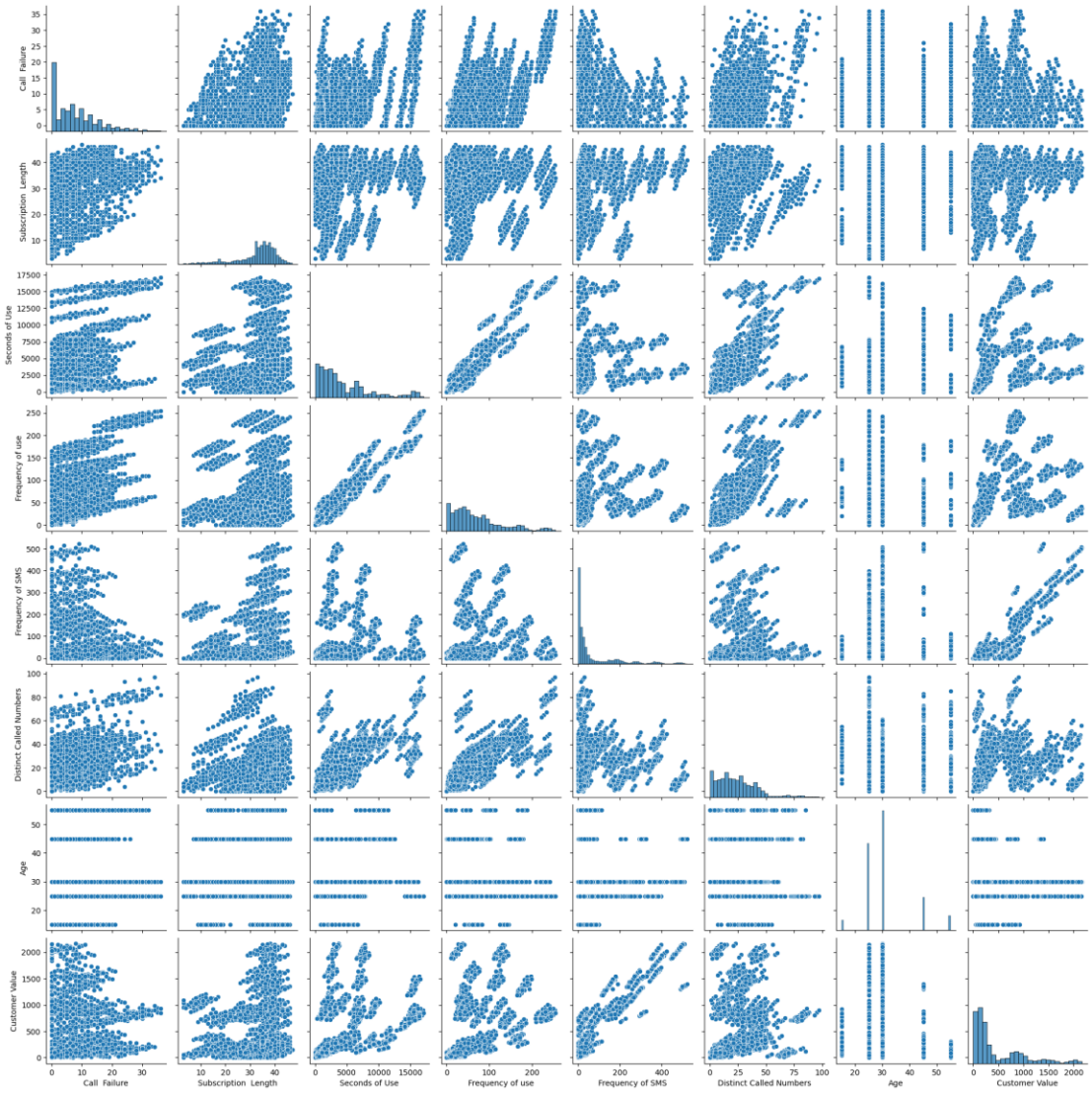
Gráfica 1. Distribución de las variables numéricas.

Luego, se realizó un análisis de correlación de las variables, mostrado a través de un mapa de calor. Acá podemos observar que las variables con mayor correlación a la variable objetivo (Churn) son *Complains, Charge Amount, Seconds of Use, Frequency of Use, Frequency of SMS, Distinct Call Numbers, Status* y *Customer Value.*



Gráfica 2. Mapa de calor de correlación de las variables.

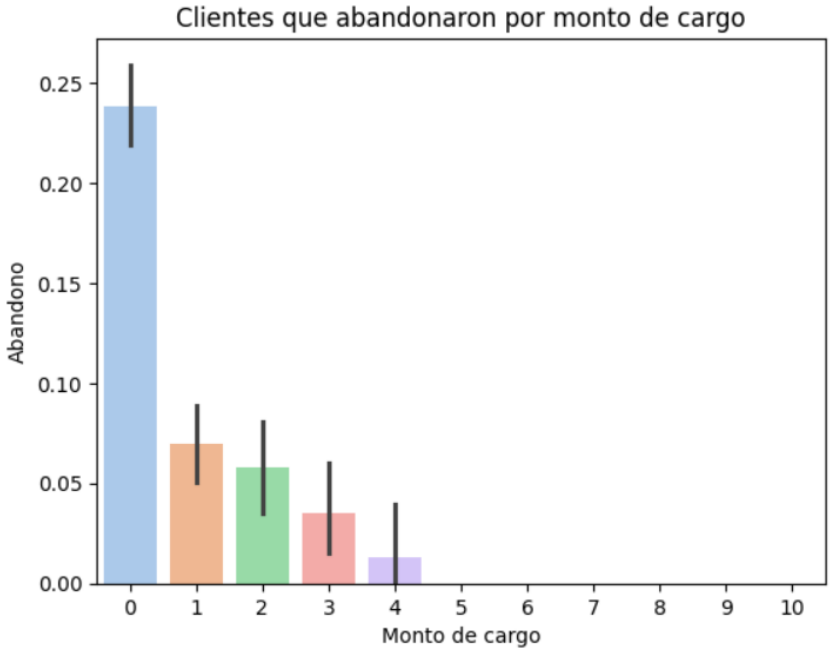
Se obtuvo también un gráfico de pares de cada una de las variables, mostrado a continuación.



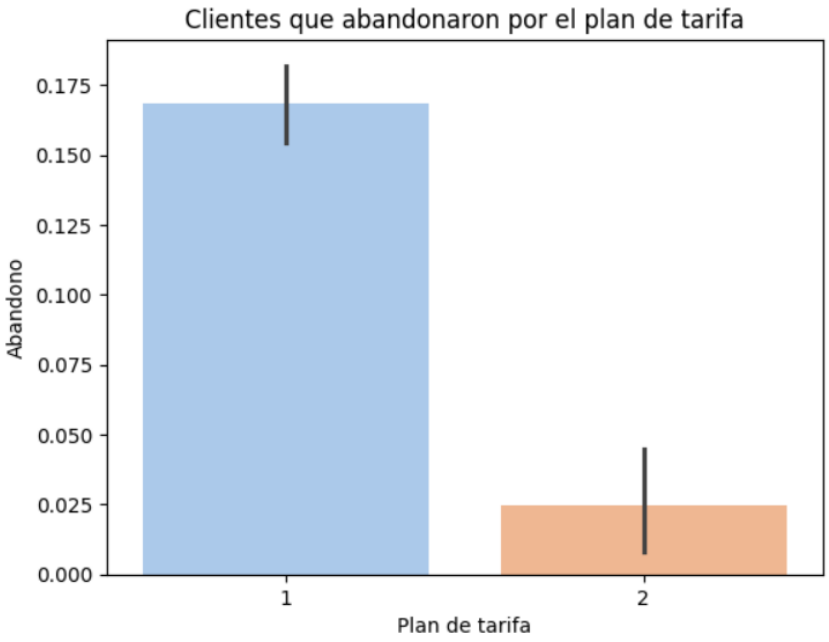
Gráfica 3. Gráfico de pares de todas las variables numéricas.

Seguidamente, se realizaron gráficos de barras de las variables más correlacionadas con la variable objetivo, para poder observar su comportamiento.

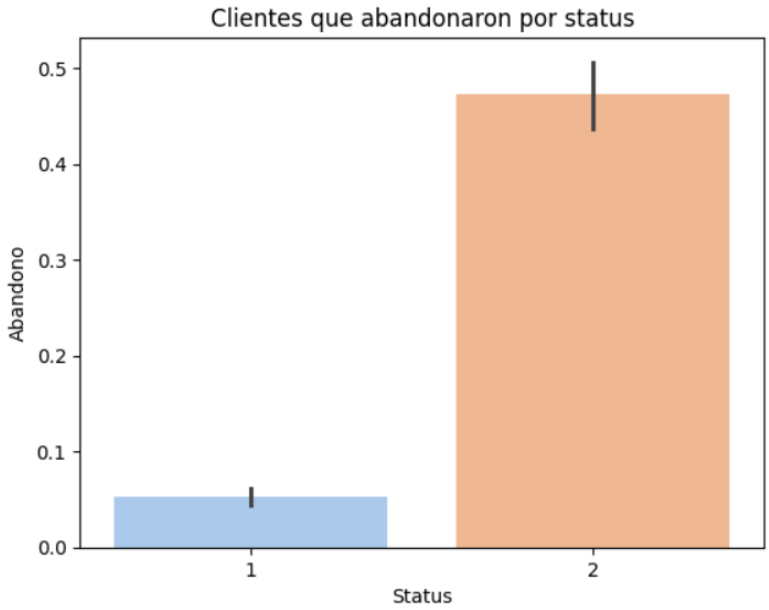
Gráfica 4. Gráfico de barras de Quejas vs Abandono. Observamos que mientras más quejas existan, mayor probabilidad de abandono existe.



Gráfica 5. Gráfico de barras de Monto de cargo vs Abandono. Contrariamente a lo que podría pensarse, observamos que mientras menor sea el monto de cargo al cliente, mayor probabilidad de abandono existe.

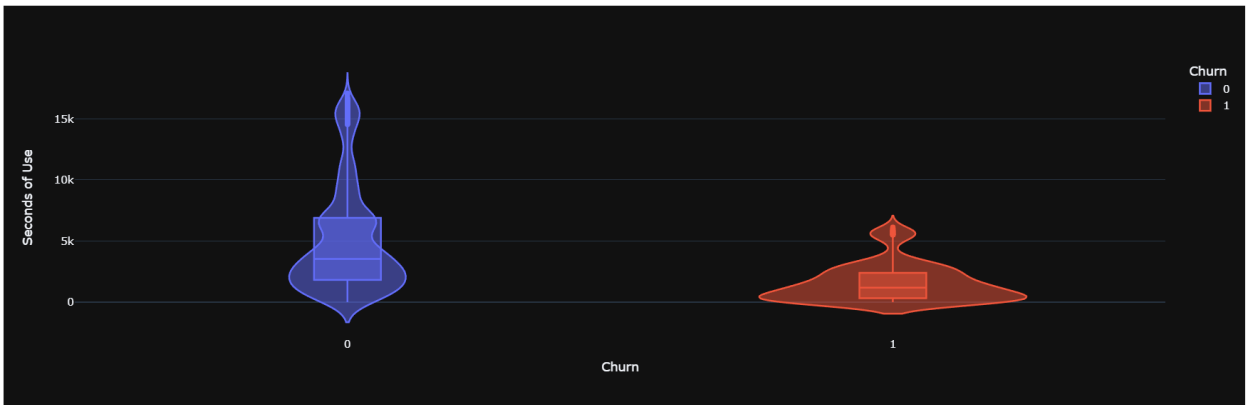


Gráfica 6. Gráfico de barras de Plan de tarifa vs Abandono. Los clientes con un plan de tarifa de pago por uso tienen más posibilidad de abandonar la empresa, a comparación de los clientes con plan contractual.

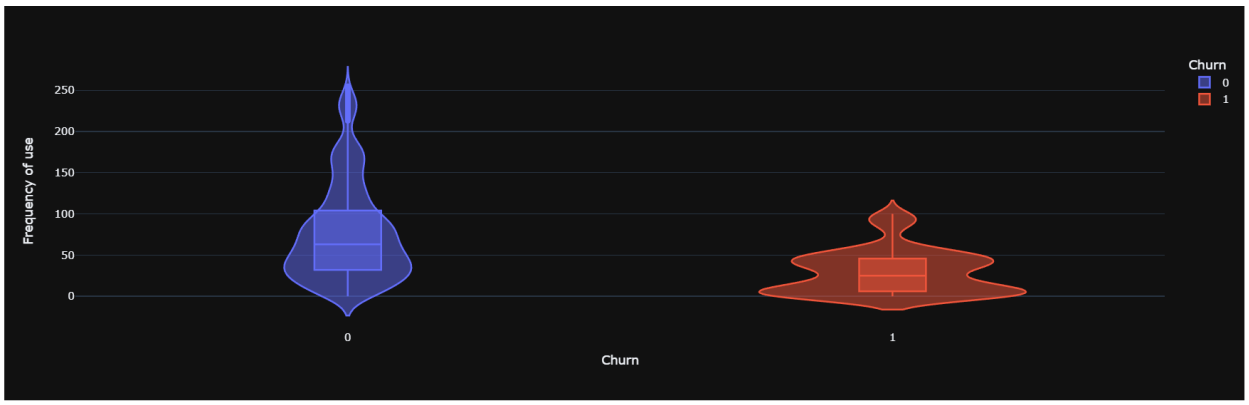


Gráfica 7. Gráfico de barras de Status vs Abandono. Los clientes con status inactivo tienen muchas chances de abandonar la compañía.

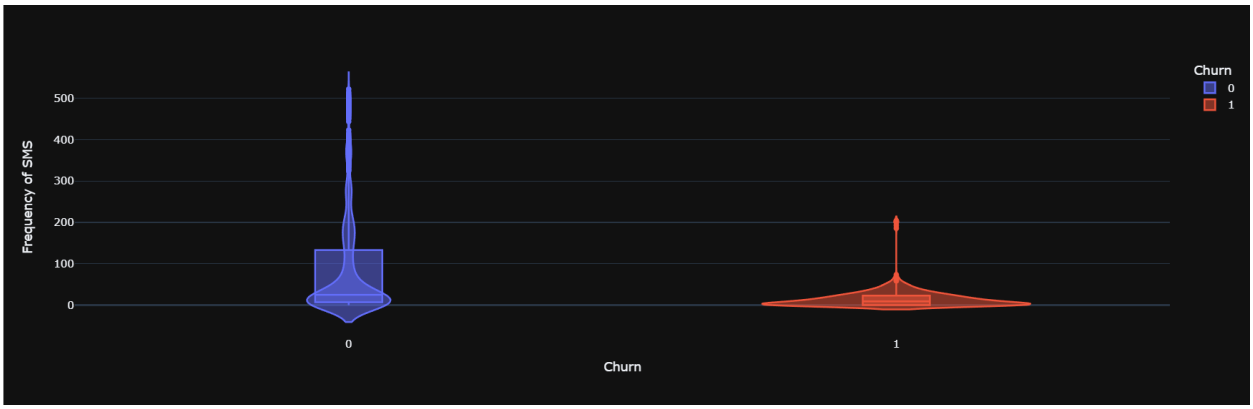
Por otro lado, se realizaron gráficas de violín de las variables numéricas más correlacionadas con el abandono.



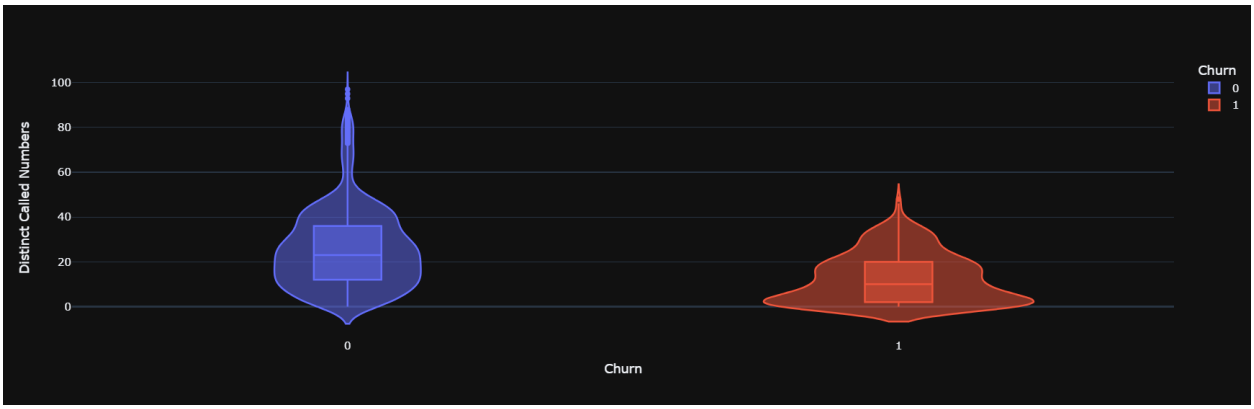
Gráfica 8. Gráfica de violín de Abandono vs total de segundos de llamadas. Los clientes que no abandonaron la compañía tienen un mayor número del total de segundos en llamadas.



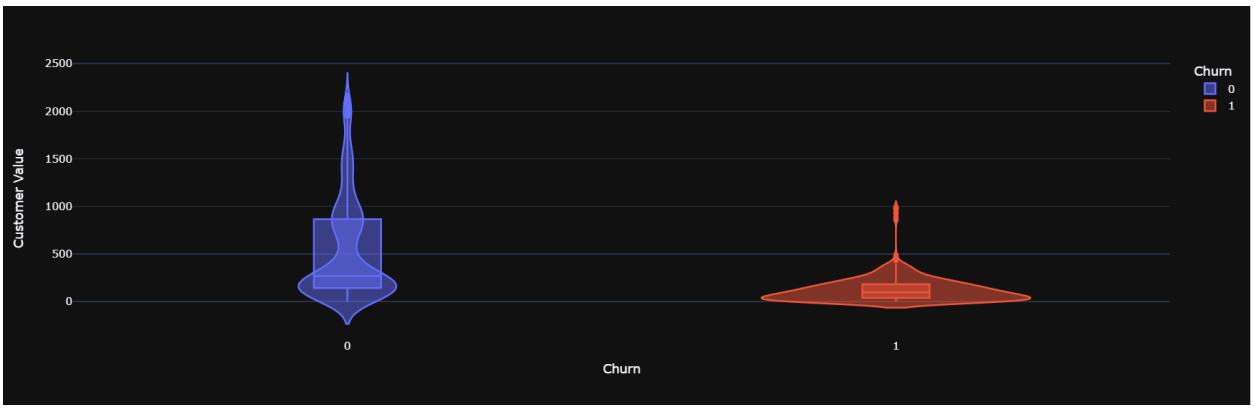
Gráfica 9. Gráfica de violín de Abandono vs Frecuencia de uso. Los clientes que tienen una mayor frecuencia de uso tienen menos probabilidades de abandonar.



Gráfica 10. Gráfica de violín de Abandono vs Número total de SMS. Los clientes que tienen un mayor número de SMS tienen menos probabilidades de abandonar.



Gráfica 11. Gráfica de violín de Abandono vs Total de números distintos a los que se ha llamado. Los clientes que han llamado a más números distintos tienen más probabilidades de seguir en la compañía.



Gráfica 12. Gráfica de violín de Abandono vs Valor del cliente. Los clientes con un mayor valor estimado tienen mayores chaces de quedarse con la compañía.

Comparación de modelos

Seguidamente, se realizaron tres modelos diferentes de machine learning para predecir el abandono. Para esto se dividieron los datos en entrenamiento y prueba, se entrenó al modelo con los datos de entrenamiento, se realizaron predicciones con los datos de prueba y se obtuvieron las métricas de cada uno, realizando una comparación entre ellos para el elegir el modelo más adecuado.

a) **Modelo de Regresión Logística:** Con este modelo obtuvimos una exactitud de 89%.

	precision	recall	f1-score	support
0	0.90	0.97	0.93	796
1	0.75	0.42	0.53	149
accuracy			0.89	945
macro avg	0.82	0.69	0.73	945
weighted avg	0.88	0.89	0.87	945

Tabla 4. Reporte de clasificación del modelo de Regresión Logística.

b) **Modelo de Random Forest:** Con este modelo obtuvimos una exactitud de 95%.

	precision	recall	f1-score	support
0	0.96	0.98	0.97	796
1	0.87	0.77	0.82	149
accuracy			0.95	945
macro avg	0.91	0.88	0.89	945
weighted avg	0.94	0.95	0.94	945

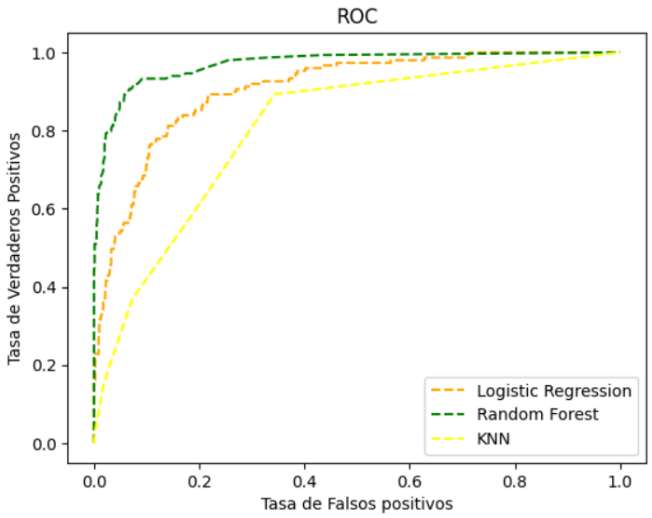
Tabla 5. Reporte de clasificación del modelo de Random Forest.

c) **Modelo de K-Nearest Neighbour:** Con este modelo obtuvimos una exactitud de 84%.

	precision	recall	f1-score	support
0	0.89	0.93	0.91	796
1	0.49	0.37	0.42	149
accuracy			0.84	945
macro avg	0.69	0.65	0.66	945
weighted avg	0.82	0.84	0.83	945

Tabla 6. Reporte de clasificación del modelo de K-Nearest Neighbour.

Luego, se obtuvo la gráfica de la Curva de ROC para poder comparar todos los modelos. Como podemos observar a continuación, el modelo más adecuado es claramente el modelo de Random Forest.



Gráfica 13. Curva de ROC de los tres modelos.

Por estos resultados obtenidos, elegimos al modelo de Random Forest como el ganador, pues tiene una exactitud bastante alta y es el más adecuado de los tres para predecir el abandono del cliente.

Implicaciones del negocio

Con estos hallazgos importantes que se obtuvieron con el análisis de datos y el desarrollo del modelo de Machine Learning, es crucial implementar estrategias en la compañía basándose en los resultados. De esta manera, podemos asignar de manera más eficiente recursos y esfuerzos para retener a los clientes más propensos a abandonar.

Para esto, es importante tomar en cuenta las variables que más implicaciones tienen en el abandono del cliente que se vieron tras el análisis de datos. Entre estas, por ejemplo, el número de quejas de clientes. Como se puede observar en la G*ráfica 4*, mientras más quejas existan por parte del cliente, mayor será su probabilidad de abandono. Para esto, sería importante realizar encuestas de satisfacción hacia los clientes y obtener retroalimentación por parte de ellos para poder brindarles un mejor servicio.

De igual manera, como se observa en la *Gráfica 5*, un monto de cargo menor hacia el cliente representa una mayor probabilidad de abandono, algo que podría pensarse que fuera de manera contraria. Sin embargo, esto puede deberse a que los clientes más comprometidos con la empresa están más dispuestos a pagar montos altos. Para poder disminuir el abandono en este caso, es esencial implementar estrategias que aumenten la satisfacción, promuevan la lealtad y agreguen valor a las experiencias con el servicio.

Por otro lado, se pudo observar que los clientes que se encuentran con un status inactivo, tienen mayores probabilidades de abandonar la compañía, como se ve en la *Gráfica 7*. Para evitar este problema, es crucial implementar estrategias específicas para reactivar y retener a estos clientes, como, por ejemplo, ofrecer ofertas personalizadas para clientes con estado inactivo que incluyan descuentos especiales y promociones exclusivas para motivarlos a volver a utilizar los servicios.

De igual manera, se observó que los clientes con un mayor total de segundos en llamadas, mayor número de llamadas, mayor número de mensajes, mayores números distintos a los que se llamó y mayor sea su valor estimado tienen mayores probabilidades de permanecer en la empresa de telefonía. Para promover esto, se pueden implementar estrategias para implementar la interacción continua, por ejemplo, destacando características adicionales, promoviendo ofertas especiales a través de llamadas o mensajes, o proporcionando incentivos para un mayor compromiso. Podrían diseñarse también programas de lealtad que recompensen la actividad continua.

Finalmente, es importante implementar el modelo de predicción que se obtuvo, para así desarrollar estrategias más personalizadas, adaptadas a las necesidades de cada cliente, mejorando así la efectividad de las acciones preventivas e incrementar el valor de cada cliente a lo largo del tiempo. Además, al realizar esto, la empresa podrá posicionarse en lo más alto del mercado, pues al adoptar

estos enfoques avanzados en la ciencia de datos, la empresa se diferenciará competitivamente del resto al ofrecer soluciones innovadoras.

En conclusión, con los resultados obtenidos del análisis de datos de la empresa de telefonía, es importante comenzar a implementar las estrategias propuestas para promover la retención de los clientes y continuar fortaleciendo la lealtad de ellos. Esto promoverá el crecimiento de la empresa y el posicionamiento en el mercado.