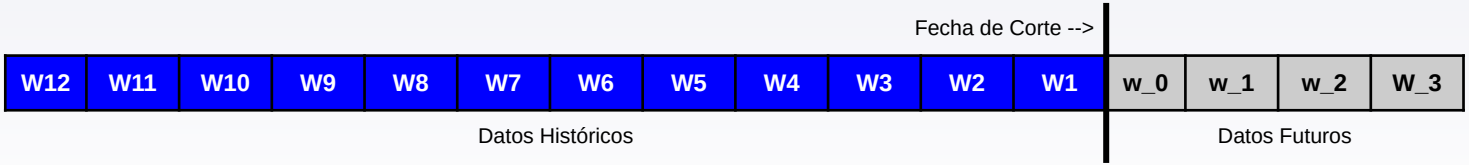
Modelo de Churn en Empresa de Telefonia Prepaga

Situación y Objetivo

Predecir para una empresa de telefonía prepaga,dentro del conjunto de sus clientes, cuál de ellos dejará de realizar recargas a su línea telefónica en las próximas 4 semanas.

Datos



Archivo de Datos:

- 12 semanas anteriores
- Entrenamiento y la predicción del modelo

Archivo de Target

- 4 semanas futuras
- Construir los targets del modelo
- Entrenamiento y la evaluación del modelo.

Datos

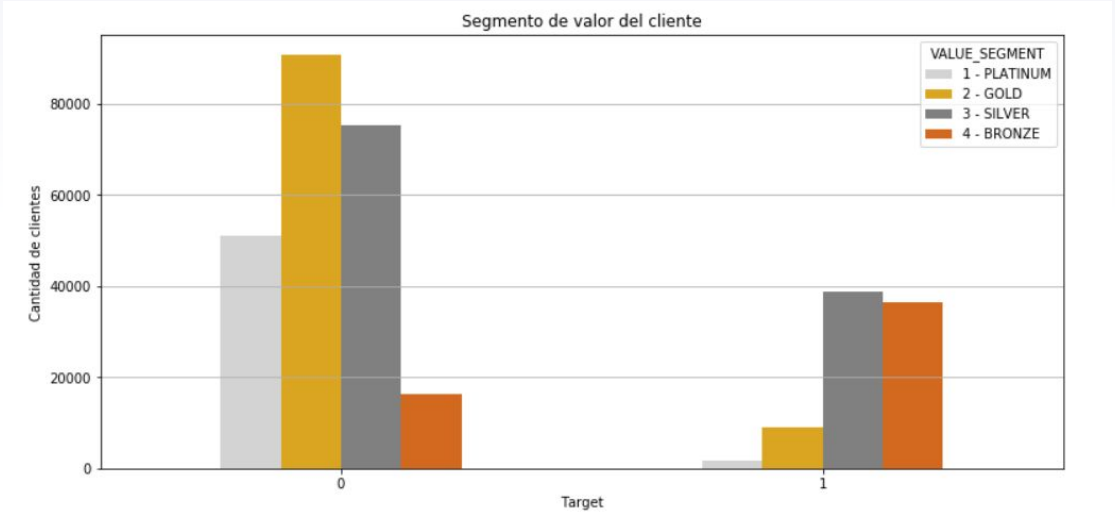
- Datos del cliente
- Montos de paquetes mensuales, semanas y recargas.
- Cantidades de paquetes mensuales, semanales y recargas
- Tráfico de datos, voz y SMS
- Otros

212 columnas

Análisis de los datos

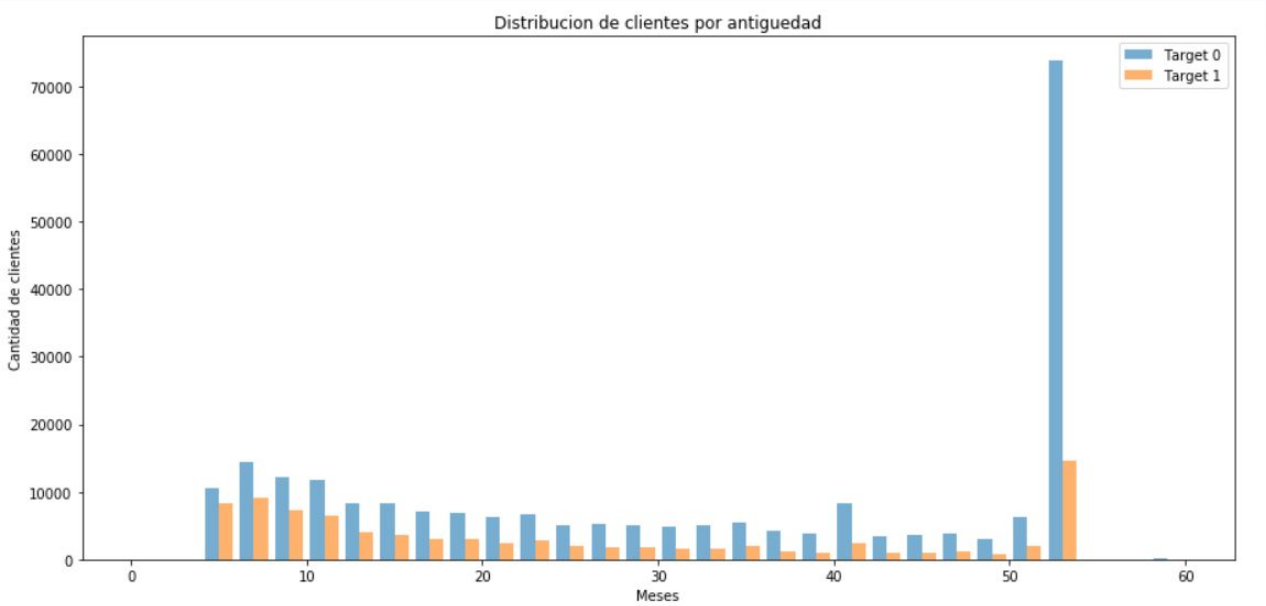
Segmento de valor del cliente

Se puede observar de esta columna a que clasificación de valor de la empresa corresponde nuestro target. Se determina que la gran mayoría entran en categoría Silver y Bronze.



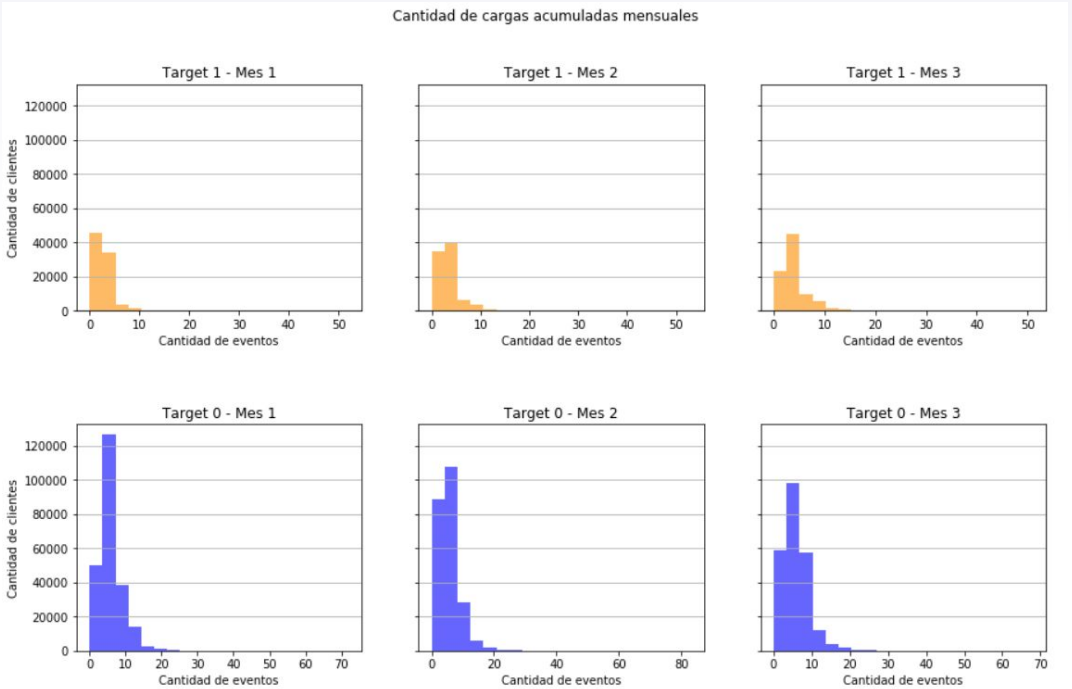
Antigüedad

Se observa que a mayor antigüedad se incrementa la diferencia entre los clientes que se quedan y los que se van.



Cantidad de cargas acumuladas mensuales

Se alcanza a determinar que el target presenta menor cantidad de operaciones realizadas. A su vez, se observa que el comportamiento entre ambos es muy similar. No se considera información crítica.



Limpieza y Transformación de datos

- Nulos
- Outliers
- Categóricos a numéricos
- Período

Nulos

Los valores nulos de las columnas se pueden asignar a un valor específico, ó, en su defecto, a la moda de la columna.

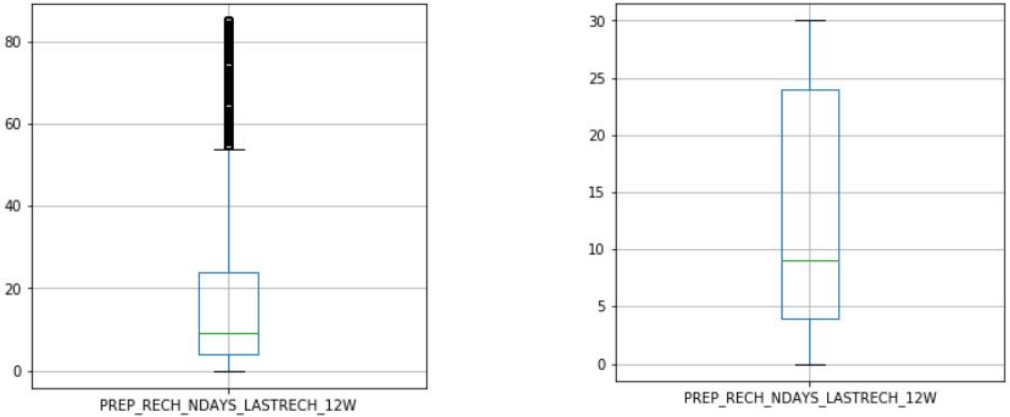
Ejemplo:

NETWORK_TECH posee demasiados nulos, al sólo tener 3 posibles valores (2G,3G,LTE) se decide asignar todos estos nulos a la moda, en este caso LTE.

Outliers

Los Outliers se detectan mediante boxplots y la limpieza de los mismos se realiza mediante el rango intercuartil (IQR).

Ejemplo: Columna 'PREP_RECH_NDAYS_LASTRECH_12W'



Variables Categóricas a numéricas.

Las columnas con datos o variables categóricas son reemplazadas, de ser posibles, en números que identifiquen cada categoría.

Ejemplo:

Los valores de la columna 'VALUE_SEGMENT' que describe las 4 posibles categorías de un cliente (Platinum, Gold, Silver y Bronze) son intercambiadas a números enteros (1,2,3,4).

Períodos

Respecto de los periodos se detecta una irregularidad en la colección de datos mediante análisis de tablas. Previo a 2015 se recolectan pocos datos, y en 2015 se observan 2 meses donde hay muchos más valores. Estos datos se ignoran por la irregularidad de los datos.

	size											
Month	1	2	3	4	5	6	7	8	9	10	11	12
Year												
2011	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	2.0	NaN
2012	NaN	NaN	NaN	5.0	3.0	NaN	NaN	NaN	7.0	NaN	NaN	NaN
2013	NaN	NaN	NaN	NaN	2.0	NaN	24.0	NaN	NaN	NaN	NaN	NaN
2014	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	49.0	73.0
2015	14.0	NaN	14.0	NaN	2322.0	34682.0	55817.0	3208.0	1926.0	2068.0	1968.0	2829.0
2016	2284.0	2160.0	2231.0	2261.0	6674.0	3943.0	2310.0	2569.0	2621.0	2846.0	2626.0	4515.0
2017	3481.0	3245.0	3591.0	2967.0	3178.0	3743.0	3231.0	3511.0	3273.0	3790.0	3426.0	5582.0
2018	4730.0	3945.0	5102.0	4927.0	4517.0	5194.0	5848.0	6149.0	5959.0	6395.0	6614.0	10802.0
2019	9615.0	9536.0	11358.0	11440.0	13163.0	8855.0	NaN	NaN	NaN	NaN	NaN	NaN

Eliminación de columnas innecesarias

- Valores únicos. Ejemplos: Columnas 'FECHA_CORTE' y 'SOURCE' .
- Conocimiento de dominio.
- Que no aportan al estudio del dataset.
- Correlación alta.

Conocimiento de dominio.

Eliminamos de nuestro Dataset las columnas que no aportan valor al modelo, ya sea por especificación o por conocimiento del campo.

Ejemplo:

Dato 'SPNDG_VOI_ONNET_ARPU_M1' y equivalentes, especificados por el cliente que no se deben tener en cuenta.

Correlación.

Se realizaron distintos análisis de correlación para determinar qué columnas aportan la misma información para el modelo (correlación = 1).

Ejemplo:

PREP_RECH_AMT_X por aportar la misma información que PREP_RECH_Q_EVT_X

-Por la misma lógica, se descarta la columna PREP_RECH_Q_EVT_Wx

-Relacionadas con cada categoria de trafico de datos (Streaming, Redes Sociales, etc). Se descartan por su alta correlación con la columna TRD_Mx.

Features para el modelo:

22 columnas

	Target
Target	1.000000
PREP_RECH_NDAYS_LASTRECH_12W	0.538946
VALUE_SEGMENT	0.481524
SEGMENTATION	0.421330
SUSCRIBER_KEY	-0.003540
TENURE_CUSTOMER	-0.187627
TRD_M3	-0.191822
TRD_M2	-0.232534
TRV_ONNET_DUR	-0.244262
PACK_DATA_Q_X3	-0.251169
TRV_OFFNET_DUR	-0.287269
PREP_RECH_Q_EVT_X3	-0.293794
TRD_M1	-0.325109
PACK_DATA_Q_X2	-0.338626
PACK_DATA_Q	-0.384348
PACK_DATA_EXP	-0.390268
PREP_RECH_Q_EVT_X2	-0.429638
PACK_DATA_Q_X1	-0.441303
PREP_RECH_Q_EVT_X1	-0.571420

Churn Prediction

Para la predicción utilizamos el modelo **XGBoost** con las variables que se muestran en la imagen conformando un total de 22 columnas de las más de 200 que teníamos inicialmente.

Para evaluar nuestro modelo analizamos las métricas:

- Matriz de confusión
- Exactitud
- Precision, Recall y f1-score

	Target
Target	1.000000
PREP_RECH_NDAYS_LASTRECH_12W	0.538946
VALUE_SEGMENT	0.481524
SEGMENTATION	0.421330
SUSCRIBER_KEY	-0.003540
TENURE_CUSTOMER	-0.187627
TRD_M3	-0.191822
TRD_M2	-0.232534
TRV_ONNET_DUR	-0.244262
PACK_DATA_Q_X3	-0.251169
TRV_OFFNET_DUR	-0.287269
PREP_RECH_Q_EVT_X3	-0.293794
TRD_M1	-0.325109
PACK_DATA_Q_X2	-0.338626
PACK_DATA_Q	-0.384348
PACK_DATA_EXP	-0.390268
PREP_RECH_Q_EVT_X2	-0.429638
PACK_DATA_Q_X1	-0.441303
PREP_RECH_Q_EVT_X1	-0.571420

Modelo XGBoost

Matriz de confusión

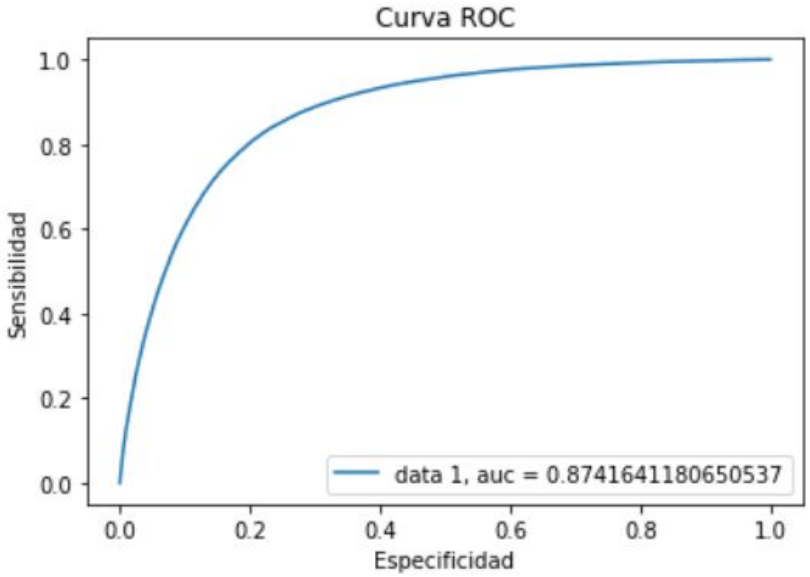
```
[[101704  15104]
 [ 16554  36293]]
```

La matriz de confusión nos da valores absolutos de nuestro resultado. Estos pueden ser útiles pero conviene analizar precisión, recall y f1-score.

	precision	recall	f1-score	support
0.0	0.86	0.87	0.87	116808
1.0	0.71	0.69	0.70	52847
accuracy			0.81	169655
macro avg	0.78	0.78	0.78	169655
weighted avg	0.81	0.81	0.81	169655

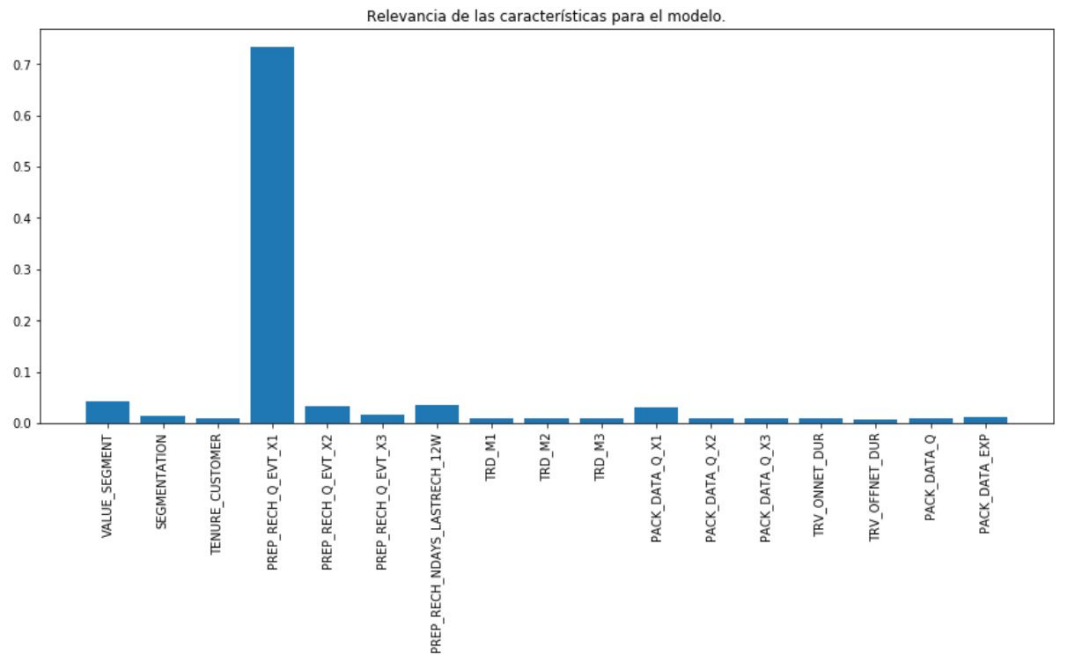
Respecto de nuestro target obtenemos un buen valor de f1-score (que tiene en cuenta precision y recall), es decir podemos predecir bien aquellos clientes que se van, aunque todavía hay varios que no podemos detectar y nos baja el recall.

Curva ROC

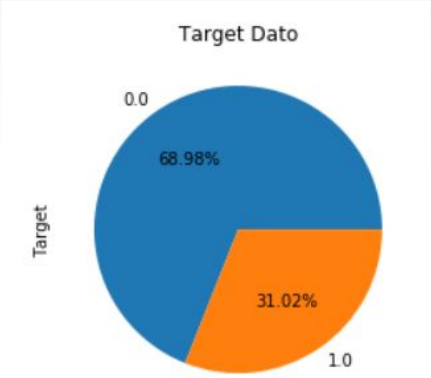


La curva ROC nos da un indicio de que tan bien separa nuestro modelo entre Target 1 y 0. Nuestra curva ROC tiene un valor de 0.85 (siendo 1 el valor ideal). Por lo que podemos concluir que nuestro modelo clasifica de manera satisfactoria entre target 0 y 1

Relevancia de las columnas



Predicción y Conclusión



Como conclusión nuestro sistema permite predecir de forma más que aceptable aquellos clientes que están por abandonar la empresa. Aún así, se considera factible modificar el modelo para perfeccionar algunos parámetros como el f1-score y el recall.