Modelos de cancelación de suscripción y segmentación de clientes en telecomunicaciones

Jeniffer Abigail Martínez Baez

Resumen

A continuación se presentan técnicas de Machine Learning para responder a las siguientes preguntas: ¿Cómo saber si un cliente cancelará una suscripción? y ¿Cuáles son las razones por las que cancelan?. Para la primera pregunta se ocuparon técnicas de clasificación como Regresión Logísitica, Naive Bayes,K-Vecinos y Anaálisis Discriminante. En cuanto a las razones de cancelación se acudió a segmentar a los clientes que cancelaron con modelos no supervisados: K-Means y Gaussiano mixto.

1. Introducción

Actualmente, una de las problemáticas más interesantes y oportunas es prevenirse de escenarios que no sean convenientes para el ser humano. Para una empresa cuyo actividad radique en ofrecer servicios y productos, el riesgo básicamente depende de las decisiones del cliente. En este caso, se trabaja con un dataset de telecomunicaciones, donde cada registro reprensenta una línea telefónica. El dataset contiene 3333 registros y 20 variables, que en su mayoría son detalles de historial de llamadas. Algo muy importante es que tenemos una flag de cancelación, por lo tanto esta será nuestra variable objetivo. El objetivo es conocer qué clientes tienen mayor probabilidad de cancelar (modelo supervisado), al igual que sus principales razones para el uso de una línea telefónica, por lo tanto no solo basta con saber si cancelará o no su suscripción, sino también saber cuales son las características (modelo no supervisado) y así construir estrategias para reducir la tasa de cancelación.

2. Marco teórico

2.1. Modelo supervisado

En este ejercicio debido a la naturaleza de la variables objetivo, se ocuparon algoritmos clasícos de clasificación : **Regresión Logísitica**, **Clasificar Bayesiano**, **K vecinos** y **Análisis discriminante**. Esto debido a que el número de datos no es tan grande y los objetivos qué se tienen requieren de una explicación más aterrizada que difícilmente puede verse si acudimos a un Red Neuronal muy sofisticada o árboles de decisión.

2.1.1 Regresión Logística La regresión logísitca esta basada en el siguiente planteamiento matemático:

$$\frac{p}{p+q} \tag{1}$$

donde *p*:=Probabilidad de ocurrimiento de evento de interés y *q*:= Probabilidad de que no ocurra el evento de interés El cual se traslada a la distribución logísitica:

$$p(E) = \frac{1}{1 + exp(-\beta_0 - \sum_{n=1}^{m} \beta_n x_n} \tag{2}$$

donde *m* := Número de variables independendientes y $\beta_n := Pesodelavariable x_n$.

2.1.2 Naive Bayes Este clasificador esta basado en el Teorema de Bayes:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \tag{3}$$

, el cual supone independencia entre variables, por lo tanto no es recomendable utilizarlo basándose en su pobabilidad.

2.1.3 K-Vecinos K-Vecinos es un clasificador que no crea un modelo paramétrico internamente, sino que únicamente se fija en los k elementos "más cercanos", es decir recurre a una métrica de distancia, del punto estudiado para cada clase. Se dice que este modelo es muy "peresozo" porque realmente no está aprendiendo, "solo toma lo que se encuentra".

2.1.4 Análisis discrimante Este discriminador busca su propia función discriminadora que es similiar a una regresión lineal múltiple, en el que se busca encontrar el punto de corte para separar las distintas clases.

2.2. Modelo no supervisado

Dentro de la modelación supervisada utilizada fueron únicamente dos algoritmos: ***K-Mean*** y ***Gaussiano Mixto***. A *K-Means* se le conoce también como un algoritmo de optimización, ya que va segmentándose por distancias de centros de clústeres dados, aunque este muchas veces es afectado por datos atípicos, no es un modelo paramétrico, a comparación del modelo *Gaussiano Mixto* el cual asume que las clases se distribuyen con una Normal Gaussiana Multivariante, este caso puede ser de gran utilidad pero cuando se tiene una gran cantidad de datos.

2.3. Reducción de dimensiones

Para reducción de dimensiones en el caso del modelo supervisado fue utilizado el método **Weight of Evidence** con el cual puedes obtener el número de variables con variedad de las n clases através del *Information value*. También es una técnica para pasar variables discretas a continuas. En cuanto a modelo no supervisado es de gran utilidad un subconjunto de variables continuas y posteriormente ver cuánta varianza tiene mediante **Componente Principales**, el cual crea n (depende del desarrollador) variables ortogonales que son combinaciones lineales de las variables independientes originales.

3. Results

3.1. Análisis exploratorio y tratamiento de datos

La variable objetivo planteada es "CHURN", el cual tiene únicamente dos posibles valores:

True := Cancela suscripción
False := No cancela suscripción

La Tabla 1 muestra el porcentaje de valores nulos y tipo de dato, en el caso de variables continuas también considera mínimo, máximo, media y desviación estándar.

	VARIABLE	#NULL	%NULL	TYPE	MIN	MAX	MEAN	DESV
0	STATE	0	0.0	object	-	-	-	-
1	ACCT_LENGTH	0	0.0	int64	1	243	101.065	39.8221
2	AREA_CODE	0	0.0	object	-	-	-	-
3	PHONE_NUMBER	0	0.0	object	-	-	-	-
4	FLAG_INTL_PLAN	0	0.0	object	-	-	-	-
5	FLAG_VOICEMAIL_PLAN	0	0.0	object	-	-	-	-
6	NUM_EMAILMESSG	0	0.0	int64	0	51	8.09901	13.6884
7	TOTAL_DAY_MINUTES	0	0.0	float64	0	350.8	179.775	54.4674
8	TOTAL_DAY_CALLS	0	0.0	int64	0	165	100.436	20.0691
9	TOTAL_DAY_CHARGE	0	0.0	float64	0	59.64	30.5623	9.25943
10	TOTAL_EVE_MINUTES	0	0.0	float64	0	363.7	200.98	50.7138
11	TOTAL_EVE_CALLS	0	0.0	int64	0	170	100.114	19.9226
12	TOTAL_EVE_CHARGE	0	0.0	float64	0	30.91	17.0835	4.31067
13	TOTAL_NIGHT_MINUTES	0	0.0	float64	23.2	395	200.872	50.5738
14	TOTAL_NIGHT_CALLS	0	0.0	int64	33	175	100.108	19.5686
15	TOTAL_NIGHT_CHARGE	0	0.0	float64	1.04	17.77	9.03932	2.27587
16	TOTAL_INTL_MINUTES	0	0.0	float64	0	20	10.2373	2.79184
17	TOTAL_INTL_CALLS	0	0.0	int64	0	20	4.47945	2.46121
18	TOTAL_INTL_CHARGE	0	0.0	float64	0	5.4	2.76458	0.753773
19	NUM_CUSTSERV_CALLS	0	0.0	int64	0	9	1.56286	1.31549

Tabla 1. Fuente: Creación propia

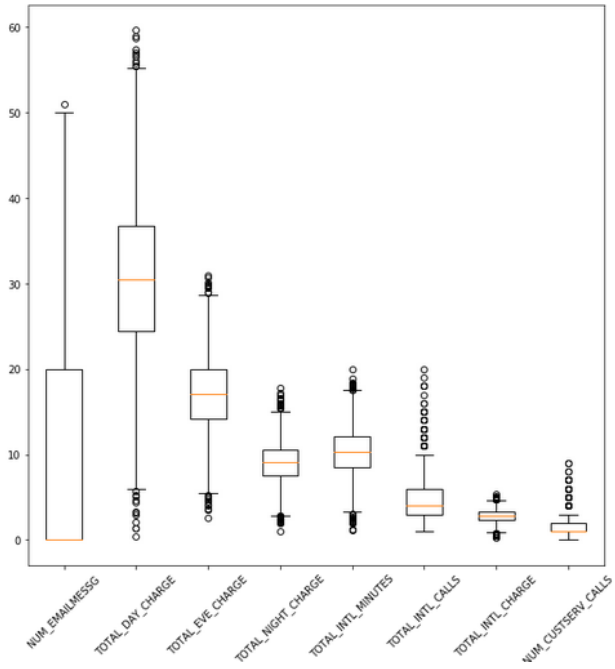


Figura 1. Boxplot de variables continuas. Fuente: Creación propia

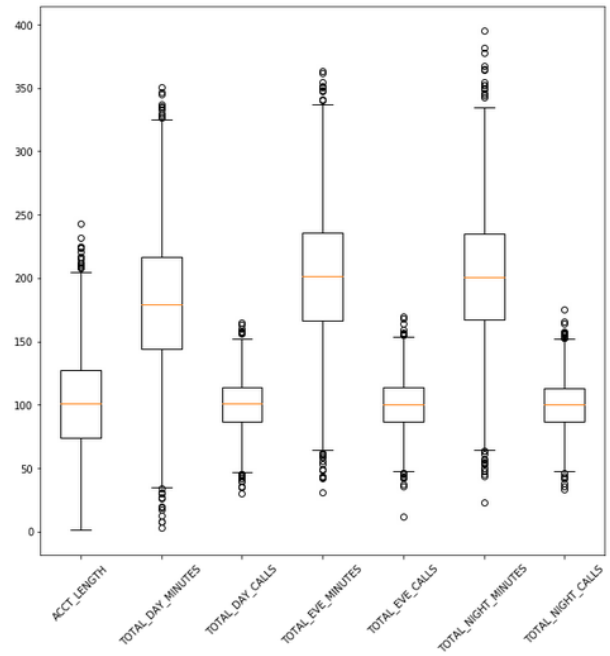


Figura 2. Boxplot de variables continuas. Fuente: Creación propia

Como puede apreciarse visualmente en la Figura 1 y 2, las variables continuas no presentan valores atípicos.

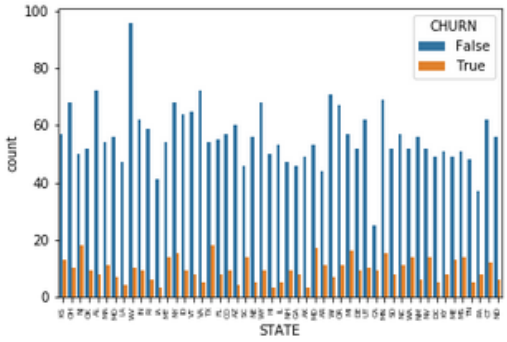


Figura 3. Conteo de clientes por estado por CHURN. Fuente: Creación propia

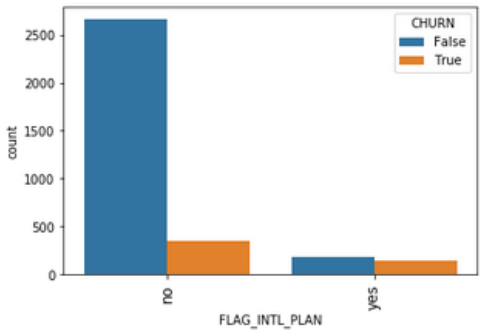


Figura 4. Conteo de clientes que cuentan o no cuentan con plan internacional en cada CHURN. Fuente: Creación propia

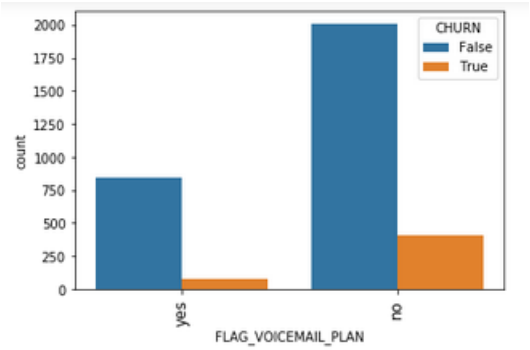


Figura 5. Conteo de clientes que cuentan o no cuentan con plan de buzón de voz en cada CHURN. Fuente: Creación propia

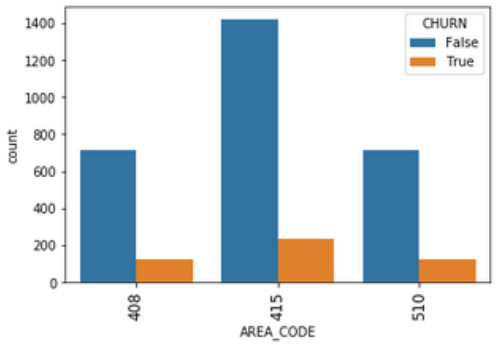


Figura 6. Conteo de clientes que hay por cada código de área en cada CHURN. Fuente: Creación propia

Al igual que para las variables discretas, no hay variables que contengan un solo tipo de CHURN en alguno de sus valores. Después de crear, através de ratios, 210 variables nuevas. Primeramente debe realizarse otra limpieza de datos donde veamos si hay valores atípicos o nulos. Hay algunas variables, como se muestra en la Tabla 2, que cuentan con más del 5% de datos nulos, por lo tanto estos campos serán removidos.

	0	%NULL
NUM_CUSTSERV_CALLS/NUM_EMAILMESSG	504	0.151215
NUM_EMAILMESSG/NUM_CUSTSERV_CALLS	504	0.151215

Tabla 2.Fuente: Creación propia

Y para las demás variables a lo más había un 0.005% de registros nulos, los cuales fueron borrados obteniendo un total de 3312 registros.

Instanciamos la clase StandardScaler (variables continuas) para tener una escala más tratable, normalizando los valores con una distribución gaussiana de media 0 y varianza 1.

Posteriormente, hay que reducir variables con el IV (Information Value) que el WoE nos arroje. Tomamos IV entre 0.02 y 0.5 sin valores nulos o atípicos, dando como resultado 36 variables seleccionadas.

Debido a que para los modelos se necesitan variables continuas, ocupemos WoE para darle un valor numérico a las variables discretas y posteriormente asignarle una clase numérica al resto de variables continuas.

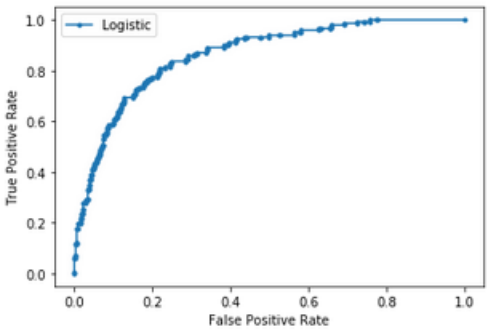
4. Resultados

Para calificar la eficacia de los modelos supervisado ocupamos una métrica originada por la curva ROC, Receiver Operating Characteristic Accuracy, la cual representa una curva de Verdaderos Positivos respecto Falsos positivos. Hablo del área bajo la curva ROC, el ROC Accuracy Score.

Regresión Logística

ROC= 0.8533594657375144 (train)

ROC= 0.8464327146171693 (train)

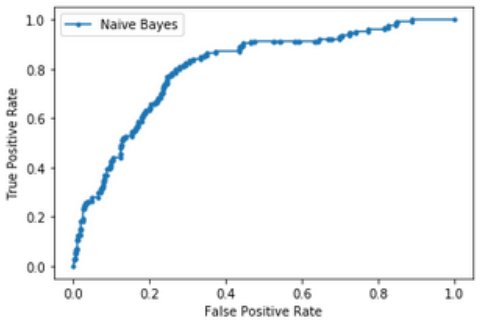


Gráfica 7. Curva ROC modelo de Regresión Logística. Fuente: Creación propia

Naive Bayes

ROC= 0.7927385774210247 (train)

ROC= 0.7631166134799493 (train)



Gráfica 7. Curva ROC modelo Naive Bayes. Fuente: Creación propia

K-Vecinos

ROC= 0.9027413561338945 (train)

ROC= 0.892414550843632 (train)

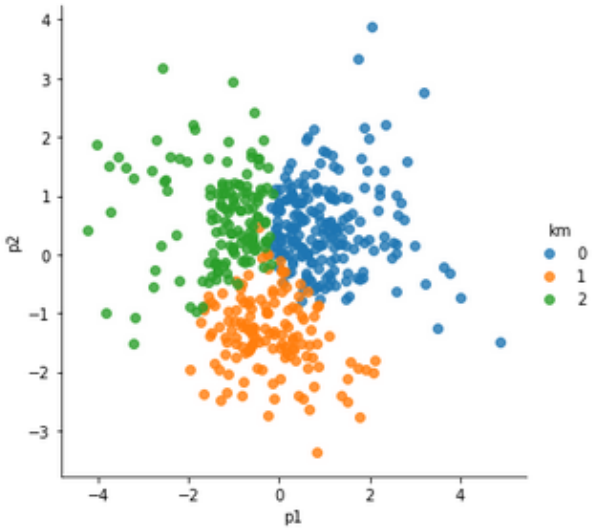
Análisis discrementa

ROC= 0.84000822700661845 (train)

ROC= 0.8708531713923123 (train)

Se elige el modelo de Regresión Logística debido a que la ROC accuracy es más alta y no hay diferencia tan grande entre la métrica Train y Test. Además de que posteriormente puede realizarse Credit Scoring para el monitoreo del modelo y ver si está discriminando a la variable objetivo con las variables seleccionadas para el modelo.

Para perfilar los tipos de clientes que se tienen en la clase que cancela su suscripción tenemos:



Gráfica 9. Clustering K-Means. Fuente: Creación propia

Cluster 0: Clientes sin solución a sus quejas Cargos promedio al día por llamadas es del 0.26, tiene un promedio de 4.4 llamadas a servicio al cliente, la duración promedio de sus llamadas internacionales es de casi 3 minutos y tiene un poco más de 4 llamadas internacionales en promedio, sus cargos en llamadas internacionales son de 0.75 en promedio.

Cluster 1: Clientes con altos cargos a llamadas internacionales Promedio de cargos por llamadas es casi de 0.34, tiene casi 2 llamadas en promedio hacia servicio al cliente, sus llamadas internacionales duran en promedio 6.5 minutos y los cargos promedio por llamadas rebasan al 1.75.

Cluster 3: Clientes que no están dispuestos a pagar cargos Los cargos promedio por llamada son de casi 0.5 por llamadas, tiene únicamente una llamada en promedio hacia servicio al cliente, la duración de llamadas internacionales son de 2.2 minutos en promedio y tiene casi 6 llamadas internacionales con cargos promedio de 0.6.

5. Conclusiones

El modelo de predicción es un paso adelante con el que se puede saber a qué clientes se deben enfocar más, en especial al momento en que estos empiecen a dar alertas de que tienen problemáticas con su línea telefónica, como una llamada de servicio al cliente o altos cargos a su línea. **Como reflexión me queda decir que es asombroso cómo con únicamente 20 variables puedes hacer toda una estrategia de negocio.**

References

- Scikitlearn https://scikit-learn.org/stable/index.html
- Understanding Data Science Classification Metrics in Scikit-Learn in Python https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3bc336865019