Predecir el abandono de clientes

El ratio de abandono (Churn rate) es un término de negocio que describe el ratio en el cuál los clientes abandonan o dejan de pagar por un producto o servicio. Es un punto crítico en muchos negocios, en los que es habitual el caso en el que adquirir nuevos clientes es mucho más "costoso" que retener los existentes (en algunos casos entre 5 y 20 veces más costoso). Comprender por lo tanto que es lo que mantiene a los clientes enganchados, es de mucho valor, puesto que es la base para poder desarrollar estrategias de retención y desplegar los procesos operacionales encaminados a evitar que el cliente se vaya.

Existe por lo tanto un interés creciente entre las empresas, en desarrollar mejores técnicas de detección de abandonos utilizando técnicas de Machine Learning. Este concepto es especialmente importante en aquellos negocios que están basados en suscripciones, tales como gimnasios, empresas de telefonía, TV por Internet, etc.. Desde un punto de vista de negocio, este tipo de empresas tienen equipos especializados en detección de abandono, que definen las estrategias de como acercarse a un cliente que tiene un determinado riesgo de abandono, con técnicas de retención.

Nuestro escenario: El mundo Telco

En este ejercicio vamos a trabajar con un conjunto de datos del mundo de Telefonía. El conjunto de datos ya está bastante preparado para servir como punto de partida a un análisis de abandono. Este conjunto de datos tiene 3.333 muestras de clientes con 21 atributos de cada uno de ellos, entre los que se encuentra el tiempo que ha sido cliente, duraciones de diferentes tipos de llamadas en diferentes franjas horarias, si tiene planes internacionales y el campo Churn, que indica si el cliente se ha ido o no.

	Account_Length	Vmail_Message	Day_Mins	Eve_Mins	Night_Mins	Intl_Mins	CustServ_Calls	Churn	Intl_Plan	Vmail_Plan	Day_Calls	Day_Charge	Eve_Calls	Eve_Charge	Night_Calls	Night_Charge	Intl_Cal
0	128	25	265.1	197.4	244.7	10.0	1	no	no	yes	110	45.07	99	16.78	91	11.01	
1	107	26	161.6	195.5	254.4	13.7	1	no	no	yes	123	27.47	103	16.62	103	11.45	
2	137	0	243.4	121.2	162.6	12.2	0	no	no	no	114	41.38	110	10.30	104	7.32	
3	84	0	299.4	61.9	196.9	6.6	2	no	yes	no	71	50.90	88	5.26	89	8.86	
4	75	0	166.7	148.3	186.9	10.1	3	no	yes	no	113	28.34	122	12.61	121	8.41	

El primer paso, la pregunta de negocio

Como primer paso deberíamos de intentar acotar la pregunta de negocio a la que queremos dar respuesta. Ejemplos en este caso:

- Quiere saber cuales son las características de los clientes que me abandonan (patrones de abandono)
- Quiere saber que clientes están en riesgo de abandono en los próximos 30 días
- Quiero saber cual es la probabilidad de que me abandone un determinado cliente en los próximos 60 días.

Aunque puedan parecer preguntas similares, desde un punto de vista de proyecto de Machine Learning son preguntas diferentes, que probablemente requieran de técnicas distintas a todos

los niveles, desde el conjunto de datos que tengo que utilizar para entrenamiento, hasta el algoritmo a utilizar, pasando por las características a extraer.

Y esto puede complicarse todavía más, si incluimos las variables "económicas", es decir, ¿Cuánto dinero voy a dejar de ingresar el mes que viene por abandonos de clientes? ¿Cuál es la ponderación entre el % de probabilidad de abandono y el importe que eso supone?

Veamos que podemos hacer

Tarea 1: Preparando los datos

Recuerda los conceptos que hemos visto durante el curso:

- Nulos y valores erróneos
- Correlación de características
- Encoding de variables categóricas
- Escalado de características
- Balanceo de etiquetas

¿Qué técnicas aplicarías en este caso?

Tarea 2: Selección del Algoritmo

Vamos a empezar por intentar resolver el problema de la clasificación binaria, para intentar detectar que clientes me abandonarán, sin bajar más en detalle en la pregunta de negocio a resolver. En este caso, comencemos comparando utilizando Accuracy y Matriz de Confusión, esos tres algoritmos.

Debemos de fijarnos en los siguientes escenarios para tomar la decisión de que algoritmo utilizar:

- Cuando un cliente abandona, ¿Con qué frecuencia mi clasificador lo predice correctamente?
- Cuando un clasificador predice que un cliente va a abandonar, ¿con qué frecuencia ese cliente realmente se va?

Tarea 3: La probabilidades

Queremos enriquecer nuestro modelo proporcionando al cliente final no solo un listado de posibles abandonos, sino también la probabilidad de dicho abandono. Para ello recuerda que existen algoritmos que nos proporcionan directamente esa información, a través del método `predict_proba()`.