Punto de Corte

Para seleccionar el punto de corte se seleccionó el modelo de regresión logística y se realizó un análisis de deciles para determinarlo, dicho análisis se puede encontrar al final del notebook y se obtuvo el siguiente resultado.

	decil	frauds	total	fraud_rate	cumulative_fraud	cumulative_fraud_rate	lift
0	9	19	19	1.000000	19	0.195876	1.958763
1	8	19	19	1.000000	38	0.391753	1.958763
2	7	19	19	1.000000	57	0.587629	1.958763
3	6	19	19	1.000000	76	0.783505	1.958763
4	5	12	19	0.631579	88	0.907216	1.814433
5	4	4	19	0.210526	92	0.948454	1.580756
6	3	2	19	0.105263	94	0.969072	1.384389
7	2	2	19	0.105263	96	0.989691	1.237113
8	1	0	19	0.000000	96	0.989691	1.099656
9	0	1	19	0.052632	97	1.000000	1.000000

Antes de iniciar, hay que analizar brevemente que significan algunas de las variables obtenidas. Primero, el “fraud rate” indica la proporción de fraudes de cada decil, es decir, en decil número 2 podemos ver que el 10.52% representa los 2 fraudes del total de 19 de datos. La variable de “cumulative_fraud_rate” indica la proporción acumulativa de fraudes hasta cada decil y nos da una idea del porcentaje de fraudes que se han detectado hasta ese punto. Finalmente, la variable de “lift” mide la efectividad del modelo en identificar fraudes comparado con una selección aleatoria, por lo que valores mayores a 1 indican que el modelo funciona mejor que un aleatorio.

Con esto en mente, el punto de corte final lo seleccionaría ya sea en el decil 5 o en el decil 6, probablemente en el **decil 6** sería la mejor opción ya que de este decil hacia arriba todas las observaciones son fraudes o tienen un fraud_rate de 1, además, hasta este decil se han acumulado un 78% de los fraudes totales. De igual manera, el lift en este decil es superior a 1 y en el siguiente decil la disminución en la tasa de fraudes disminuye en gran cantidad, lo cual haría al decil 5 una opción un poco menos óptima.