PORTO SEGURO'S CLAIM PREDICTION

Las compañías de seguros habitualmente se ven enfrentadas a la problemática de tener clientes insatisfechos con la resolución de sus aseguradoras. Dichas situaciones suponen un importante gasto económico y en algunos casos, demandas judiciales. En este proyecto se va a estudiar el caso de una empresa brasileña de seguros automovilísticos, Porto Seguro.

El objetivo de este proyecto es el de predecir que clientes serán los que el año que viene presentarán una reclamación al seguro, basándonos en el análisis de los datos que la empresa ha registrado de sus clientes durante los últimos años.

Cuando eres conocedor de los factores que influyen, puedes anteponerte a los acontecimientos y actuar en consecuencia. Como por ejemplo adaptando las tarifas de los seguros en función del tipo de cliente o lanzando campañas que beneficien a los conductores más seguros y ayuden a modificar el comportamiento de los más irresponsables.

Conocer mejor a tus clientes es una valiosa información para realizar operaciones en consecuencia y hacer que la cobertura del seguro de automóviles sea más accesible para más conductores.

DATOS

La empresa ha proporcionado un dataset con casi 600.000 registros de clientes diferentes con 57 variables para cada registro. Una gran dificultad para sustraer información de estos datos ha sido que las variables están encriptadas y no se ha proporcionado información acerca de qué significa cada una de ellas realmente. Tan solo informan de que hay 4 grupos de variables: relacionadas con el individuo, con el coche, con la región y otras que han sido calculadas previamente.

PROCEDIMIENTO

Se han identificado algunos registros que contenían missings en algunas de las variables y se ha procedido a imputarlos con la mediana en las variables numéricas y añadiendo la categoría "missing" en las categóricas.

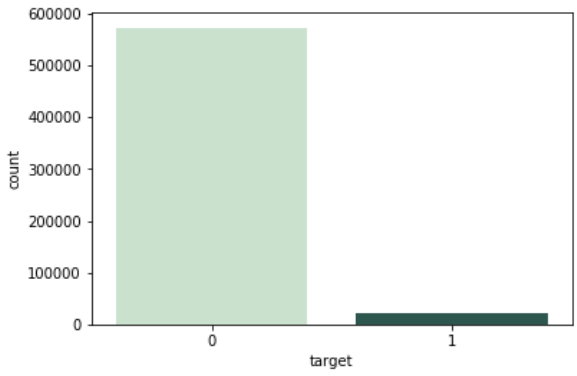

Figura 1: Countplot del target. Se aprecia claramente el desbalanceo de los registros.

Otra dificultad para obtener resultados sólidos con este dataset ha sido que el target, representado por el valor numérico 1 en caso de que un cliente presente una reclamación y un 0 en caso de que no lo haga, está muy desbalanceado, como se muestra en la *Figura 1*. Los reclamadores son tan solo el 3,6% del total de registros.

Para hacerse una idea de cuáles son las variables más relevantes en este problema se ha ajustado un modelo random forest para hacer uso de su capacidad discriminatoria de la importancia de las variables. De este modo se han obtenido las variables por orden de importancia.

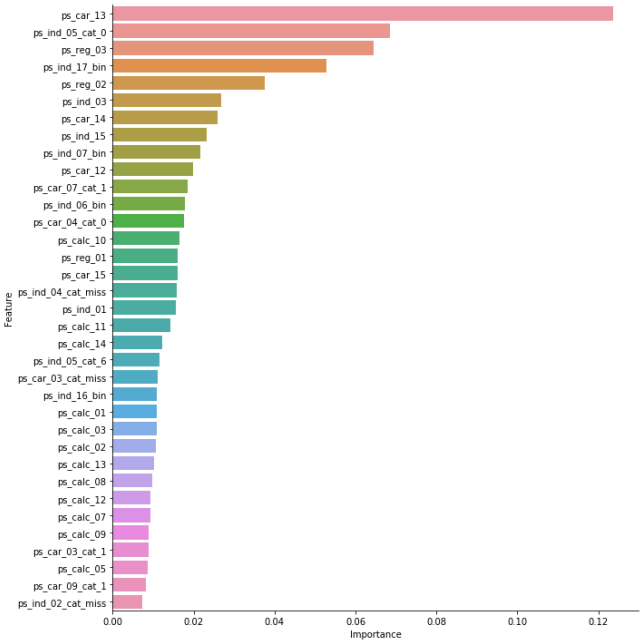

Figura 2: Las 35 variables más relevantes ordenadas por orden de importancia.

Con esta orientación como guía frente a la incertidumbre de nuestras variables, se ha procedido a ajustar el primer modelo sencillo con las 5 primeras variables más relevantes para evaluar las métricas y establecer un Benchmark como punto de referencia. El modelo ha sido una regresión logística y la métrica elegida ha sido el Gini, calculada a partir del AUC.

En un problema tan desbalanceado, la accuracy del modelo es muy elevada, debido a que el modelo tiende a predecir a todos los registros con el valor de la categoría dominante, en este caso, no reclamadores. Por lo tanto, se han elegido las métricas mencionadas anteriormente para evaluar nuestros modelos.

El tratamiento de variables en este problema era sumamente complicado por los motivos mencionados y hubiera requerido de mucho tiempo de investigación profunda en la empresa y de ensayo y error a la hora de crear combinaciones de variables que pudieran ofrecer información clave a los modelos. Tras sondear un poco el tratamiento de variables y medir las variaciones en la métrica con la regresión logística, se optó por ajustar otro tipo de modelos más sofisticados utilizando todas las variables sin más tratamiento que la imputación de los missings.

En concreto se han ajustado 3 modelos: random forest, lightgbm y red neuronal. Había bastante fe en que la red neuronal, por su naturaleza, pudiera encontrar algún tipo de combinación o patrón oculto en las variables que, por el hecho de ser anónimas, la interpretación humana no

pueda encontrar fácilmente. Sin embargo, lejos de ser así, el modelo que ha proporcionado la mejor métrica ha sido lightgbm. De todos modos, no descarto que una red neuronal pueda funcionar bien en este tipo de problema en particular, pero requiere invertir gran cantidad de tiempo en ajustar sus hyperparámetros y experimentar con diferentes diseños de estructura.

En un último intento por mejorar los resultados, se construyó un ensamble con los 3 modelos empleados anteriormente, ajustando una regresión logística a sus resultados. Una vez más, el vencedor seguía siendo lightgbm.

RESULTADOS

Lightgbm consigue un valor del Gini = 0.27214 y una accuracy = 0.963. Como se comentaba anteriormente, a pesar de haber obtenido una accuracy muy elevada, el modelo tan solo es capaz de predecir con certeza a 1 cliente como futuro reclamador.

Ajustando el threshold, el parámetro que establece la frontera que clasifica las probabilidades de la predicción de los modelos, se puede sacrificar parte de la accuracy, aumentando así el número de errores en la predicción, a cambio de aumentar el número de aciertos en los targets.

Con un valor del threshold = 0.348, obtenemos las siguientes predicciones representadas en una matriz de confusión.

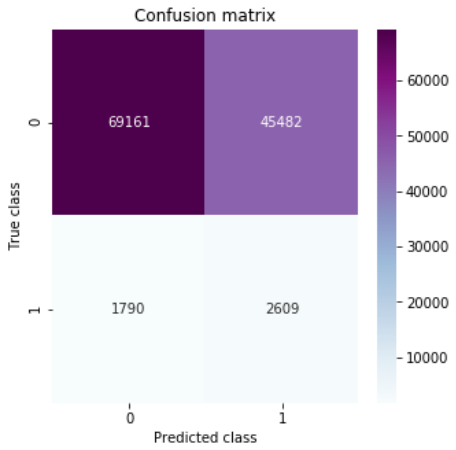

Figura 3: matriz de confusión del modelo lightgbm con un threshold = 0.348

De este modo se ha aumentado la tasa de aciertos del target en un 60%, aumentando también en un 40% la tasa de errores en la predicción de los no reclamadores.

CONCLUSIONES

Los resultados obtenidos darían información orientativa a la aseguradora para actuar en función de su política, sin embargo, para obtener unos resultados más precisos en la predicción, sería indispensable que la empresa proporcionara más información acerca de las variables empleadas. Sin conocer el significado de las variables se hace prácticamente imposible sustraer información de éstas y es aquí donde reside la magia del machine learning.

Por otro lado, al tener tan pocos registros del target, se complica la labor de aprendizaje para los modelos, ya que no tienen referencias suficientes para basarse en sus predicciones. Por lo tanto, para lograr unas predicciones sólidas, ayudaría muchísimo trabajar con datos de mayor calidad y con más registros en el target.

Sería interesante para este problema, dadas las circunstancias de que no se conoce la naturaleza de las variables, ajustar otro tipo de modelos como el support vector machine, que por como transforma las variables y discrimina los registros, probablemente pueda encontrar patrones entre las variables que ayuden a mejorar la predicción del target.