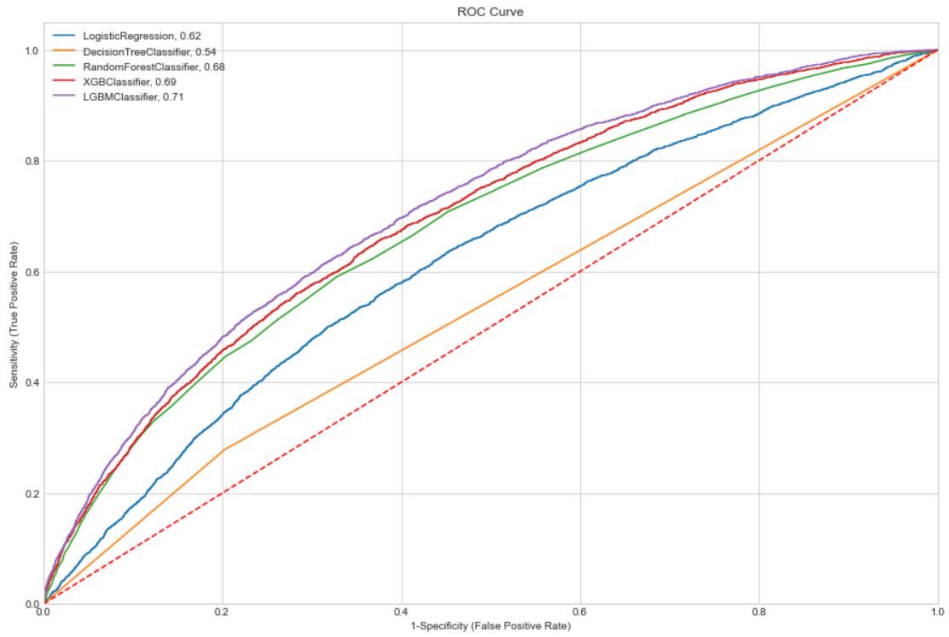
CASO PRÁCTICO

Pregunta 1

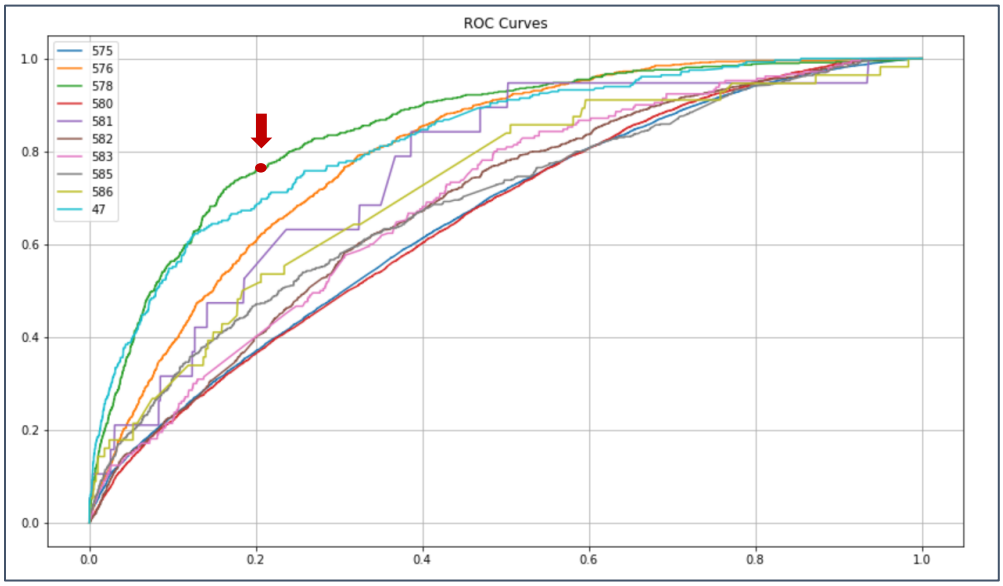
En el siguiente gráfico se muestra el resultado de diferentes modelos de Machine Learning. Por favor indique, cual recomendaría utilizar y el motivo de su selección.



En este caso de clasificación binaria, el gráfico muestra que el modelo LightGBM (LGBMClassifier) entrega los mejores resultados, ya que su curva ROC es la que más se acerca al punto de predicción perfecta (esquina superior izquierda, cuando el TPR es 1 y el FPR es 0). Este comportamiento se mantiene con todos los valores de Tresholds (la curva se mantiene encima de los demás en todo momento). Sin embargo, es conocido que los modelos LGBM pueden sufrir de overfitting (memorizado de los datos de entrenamiento) por su naturaleza de crecimiento por hojas. Dependiendo del tamaño del dataset, el hardware disponible y la naturaleza del caso de uso, quizás el modelo XGB sea el más adecuado. De todas formas, con los datos disponibles, **escogería el modelo LGBMClassifier**.

Pregunta 2

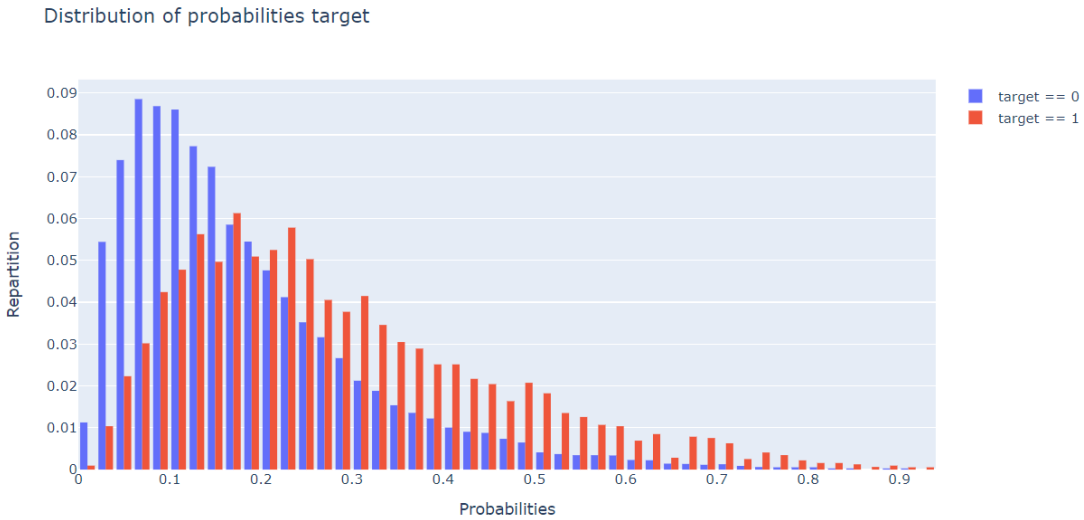
A continuación, se muestra el resultado de un modelo de propensión a la compra para múltiples productos. Por favor interprete el punto señalado sobre el gráfico.



Significa que para el producto 578, cerca del 0.8 (TPR) de las personas recomendadas compraran el producto. Mientras que cerca del 20% (0.2 FPR) no lo hará. El valor de treshold donde se encuentra el punto valora muy bien el balance entre TPR y FPR, que en el caso de un modelo de propensión a la compra representa la cantidad de personas recomendadas que compraran el producto (TPR) de un porcentaje de toda la población disponible (FPR).

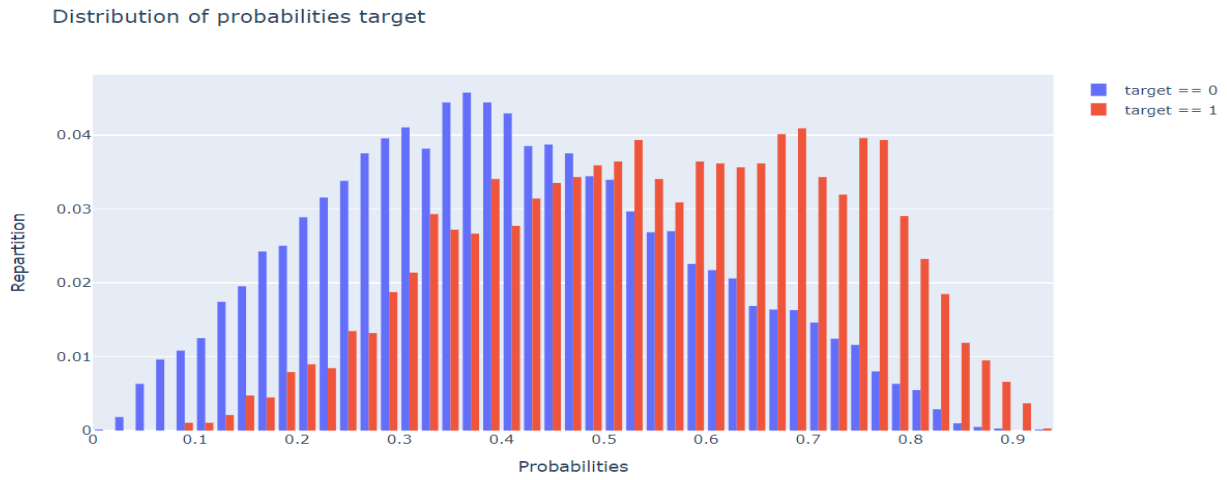
Pregunta 3

Interprete el siguiente gráfico de probabilidad. ¿Considera que el modelo discrimina fuerte el evento 0 del evento 1?



Dado que las medias de ambas poblaciones de los targets se encuentran a la izquierda de la probabilidad del 0.5, el modelo fallará en identificar a la gran mayoría de los targets 1, ya que su media se encuentra en la probabilidad 0.2 aproximadamente. Es decir, el modelo tendrá muy pocos falsos positivos pero tendrá muchos casos de falsos negativos (error tipo 2). En síntesis, **el modelo no discrimina fuerte el evento 0 del 1.**

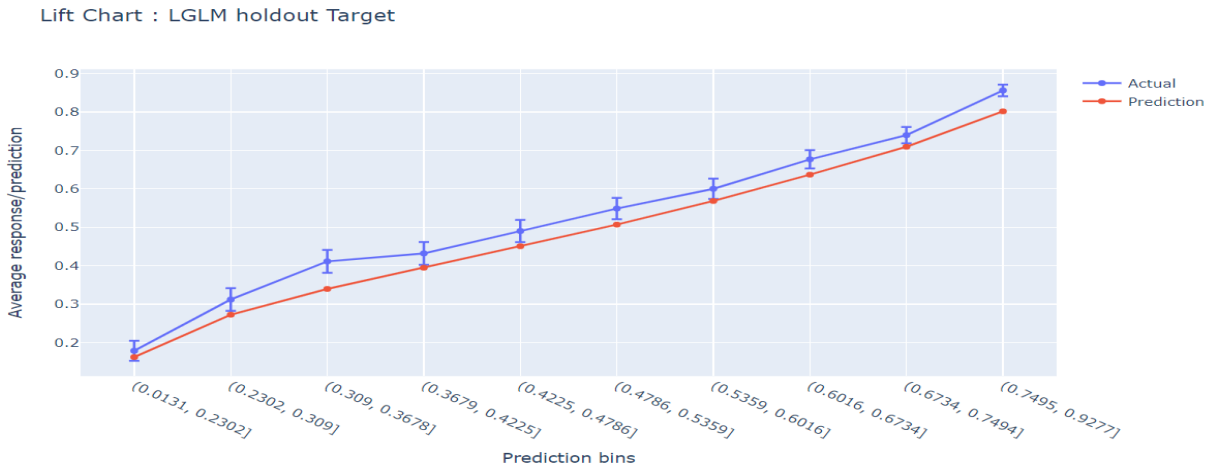
¿Qué parámetros se han podido modificar en el algoritmo XGBOOST para que se tenga el siguiente resultado? Tome en consideración que se trata el mismo caso presentado en el gráfico anterior.



El treshold (0.5) discrimina muy bien entre ambos tipos de targets. Se puede inferir que el learning rate del modelo fue aumentado ligeramente, logrando así posicionar la barrera de decisión (treshold) en un lugar muy adecuado. Dado que las distribuciones resultantes de los targets son similares, se puede señalar que no se cambió el número de estimadores (nodos del modelo) drásticamente.

Pregunta 4

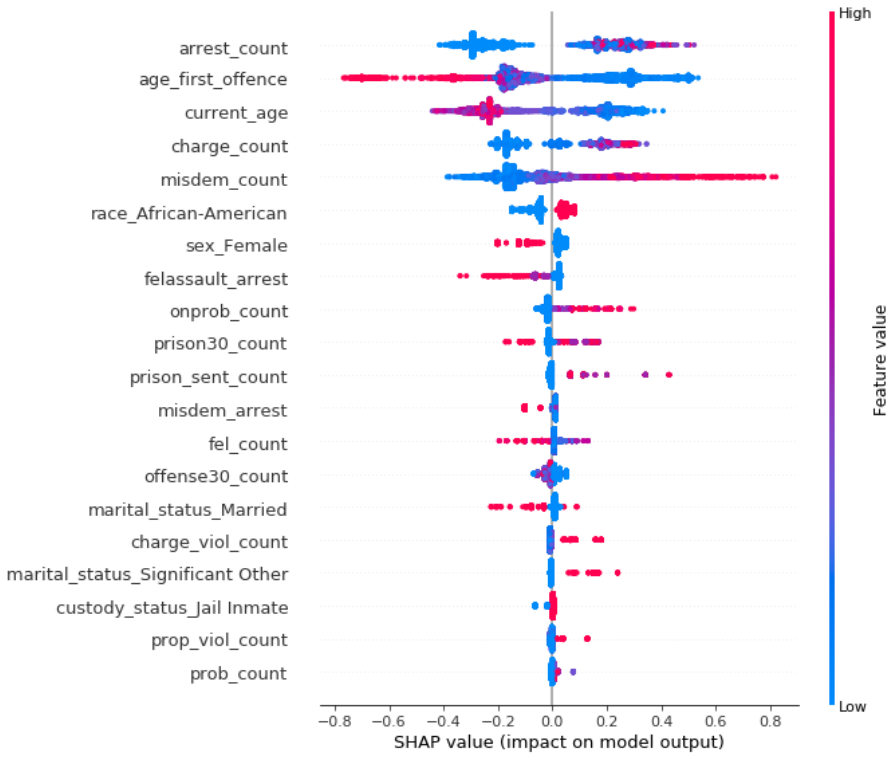
Tomando en consideración que se trata del mismo caso anterior, por favor interprete el siguiente gráfico. ¿El modelo consigue asignar una mayor probabilidad a aquellos individuos que tienen una mayor tendencia a que ocurra el evento 1?



Si, dado que las predicciones siguen el comportamiento creciente de las predicciones. Es decir, mientras el individuo tenga más probabilidad de ser del tipo 1, el modelo también entregará un valor alto, cercano al de la posibilidad real.

Pregunta 5

Interprete el siguiente gráfico. Nota: No es necesario que tenga el detalle del significado de cada variable, sólo se necesita validar el impacto de las variables.



Los puntos más cercanos al color rojo son los que generan un mayor efecto en la salida del modelo, si se encuentran en la parte derecha, contribuyen a que la salida tenga mayor magnitud, caso contrario (se encuentran en la derecha) contribuyen negativamente a la salida (menor magnitud). Las 2 variables que generan más varianza son “age_first_offence” y “misdem_count”. Esto quiere decir que el modelo es muy sensible a los valores de estas variables.

Práctica

En la empresa “Seguros CDF” se comercializa sus productos a través de distintos canales. Estos productos son adquiridos de manera voluntaria por los clientes y pueden ser cancelados en cualquier momento. El equipo comercial ha detectado que muchas de las ventas se cancelan dentro de los primeros 4 meses y necesita definir una estrategia proactiva para retener a sus clientes. Es por ello que se está apoyando en el equipo de Data Analytics de Seguros CDF para que mediante un modelo predictivo pueda ayudar a definir la estrategia más adecuada.

Para el presente caso se pide al candidato lo siguiente:

1. Elaborar un modelo predictivo que permita determinar la probabilidad de cancelación dentro de los primeros cuatro meses de haber comprado el seguro. Para ello se le entrega un conjunto de datos.

2. Use su creatividad para elaborar una estrategia de retención proactiva.
3. Elabore una presentación que sustente sus decisiones e impacto al negocio.
4. Fecha de Inicio de Vigencia: Fecha en la que se compró el seguro.
5. Fecha de cancelación: Fecha en la que el cliente solicitó la cancelación del seguro.
6. El candidato debe utilizar Python.