* No hay que explicarle que hacen los modelos, sino lo que fuimos experimentando y que sirvió o no.
* Si hay ideas de experimentos colectivos, aclarar que “de acuerdo a experimentos de colegas que hicieron previamente, utilicé …”.
* Si no hay score interno, lo de kaggle citarlo como el modelo puesto en producción (si es idea de los experimentos colectivos).
* Se puede armar estilo workflow del readme: primero feauture engineering, después bo de un lgbm (explicar meses usados), después semillerío (explicar por qué) y ensambles
* Storytelling:

1. Presentación máx de 30 segundos: comentar brevemente que voy a presentar la predicción de los clientes que se van a dar de baja en los siguientes 2 meses para nuestra campaña de retención proactiva, con un dataset de 155 feautures y dataos desde enero de 2019 hasta julio de 2021.
2. Presentación del modelo: Primero arranqué con un modelo sencillo de Árbol de Decisión, con fácil interpretabilidad, pero luego pasé al Modelo lgbm que la evidencia suele indicar que son mejores modelos para datos temporales (no explicar que es un lgbm, ya lo sabe Michelena) y en este dataset no fue la excepción.
3. Para la búsqueda de los hiperparámetros óptimos, empecé con una búsqueda grillada, pero rápidamente pasé a una optimización bayesiana de 66 iteraciones que, a pesar de llevar más tiempo, obtenía mejore ganancias internas. La BO era con un 5-fold cross validation, con junio de 2021 como mes de validación y julio como mes de testeo.
4. FE: Del tratamiento del dataset, primero usé un modelo base con sólo los últimas 6 meses, y feauture engineering de variables históricas con 6. Aquí, la optimización bayesiana con 5 semillas distintas me daba una ganancia máxima cercana a los $150m, y en los datos observados de septiembre cerca de $130m. Intenté corregir el data drifting de las variables monetarias ajustando por inflación, pero no conseguí mejoras significativas (no implica que en un futuro, con inflación creciente, sea un problema a tratar).
5. Pero gracias a experimentación propia y de colegas, aprendí que a mayor cantidad de datos/meses (expansión “a lo largo”) y de procesamiento/feautures (“a lo ancho”) logro mejores resultados. Por lo tanto, utilicé todos los meses desde enero de 2019 EXCEPTO los de los peores meses de la pandemia/cuarentena de entre marzo y julio de 2020, donde había mayor proporción de variables rotas o nulos o ceros o cambios de comportamientos distinto al patrón normal. Para las columnas, realicé los Lags 1, 3, 6; el Delta 1; la media móvil de 6 meses; tendencia lineal de 7 meses; agregué la columna 'total\_mcaja' como la suma de mcaja\_ahorro'+'mcuenta\_corriente' +'minversiones1'+minversiones2'+'mcaja\_ahorro\_adicional'+'mcaja\_ahorro\_dolares'+'mcuenta\_corriente\_adicional'+'mcuentas\_saldo'+'mcuenta\_debitos\_automaticos"; y saqué las columnas con más de 50% de nulos. Quedaron en total arriba de 1000 columnas.
6. Las salidas de las bayesianas de este nuevo modelo con una semilla ya me daba pico de casi $165m de ganancia, y en los datos observados de septiembre cerca de $120m. Los hiperparámetros de la mejor iteración (la 62) fueron:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | num\_iterations | | **1213** |   learning\_rate | feature\_fraction | num\_leaves | min\_data\_in\_leaf |
| **0.0292608373298842** | **0.136781998386116** | **342** | **43158** |

Este proceso, en una vm de 256gb de RAM demoró aproximadamente 6 días.

1. Luego, de acuerdo a experimentación de colegas, descubrieron un hallazgo interesantísimo: si se hacía un ensamble de un mismo modelo pero que sólo se variaban las semillas, entonces la solución terminaba siendo de mejor media y mucha menor varianza. Por lo que, del mejor bo del modelo con la primera semilla que experimenté, probé después con 100 semillas aleatorias, y a ellas les realicé el promedio del score/probabilidad de las 100 salidas. Ello, con 10.000 envíos (donde siempre me daba el pico en los datos observados), me terminó dando un score de $130m en los datos observados de septiembre, corroborando la hipótesis de las colegas. A considerar, armar este “semillerío” demoró otros 5 días.
2. En síntesis, creo haber llegado a un modelo muy trabajado en cantidad de datos, procesamientos y reducción de la varianza, y llegado a una ganancia interesante de terminar aplicándose. Como desventaja, pienso que el tiempo de demora en el procesamiento y en el semillerío se tiene que considerar si se pone en producción para arrancar rápidamente a principio de cada mes, pero sino se puede hacer un “undersampling” a costa de perder un poco de ganancia, pero no perder todo el trabajo hecho y ganar más tiempo. Muchas gracias.