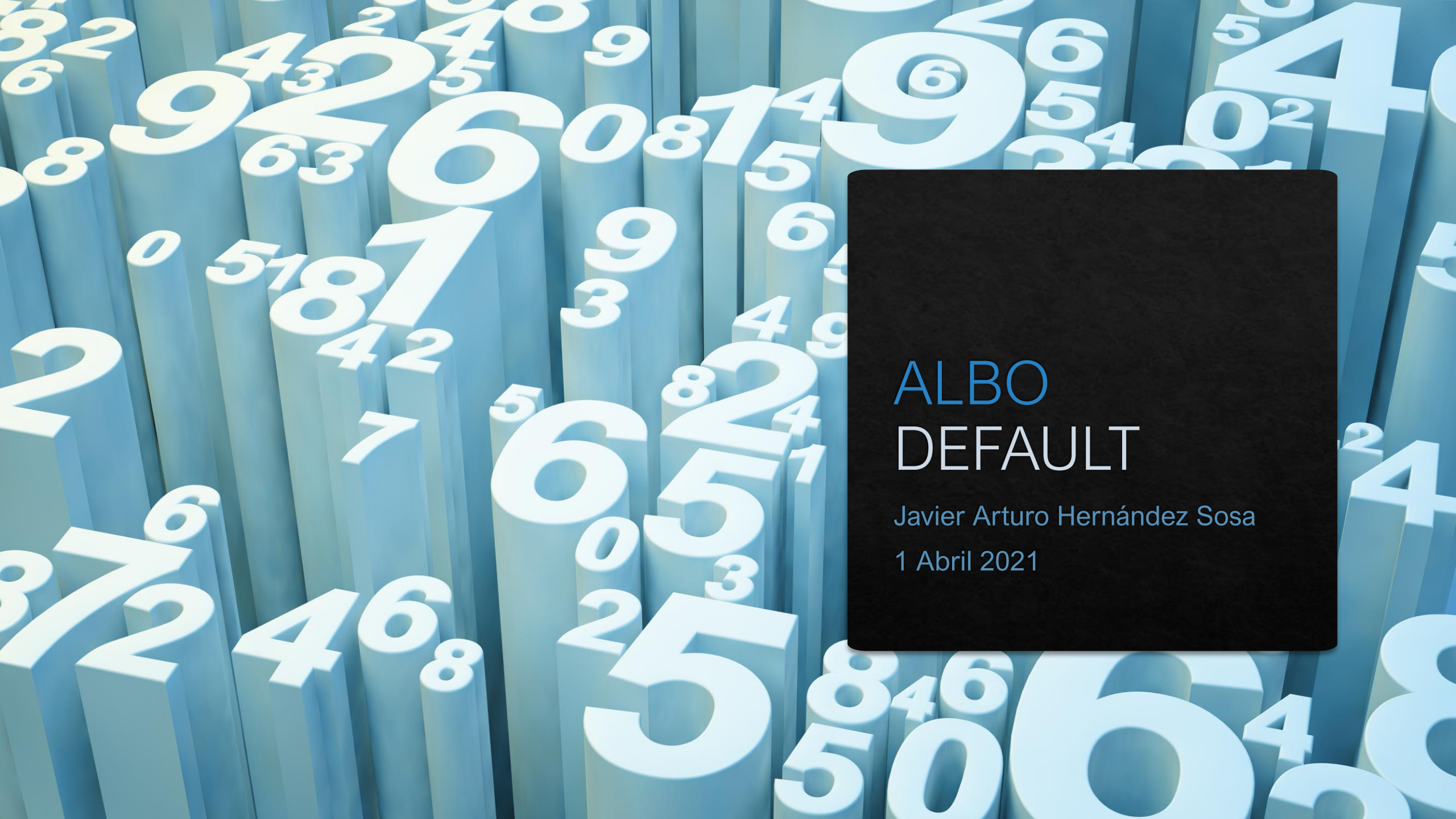

ALBO DEFAULT

Javier Arturo Hernández Sosa

1 Abril 2021

Procedimiento

Análisis de Datos

Información encontrada en los datos.

Desarollo del Modelo

Procedimiento para el Desarrollo del Modelo

Respuestas

Respuestas a preguntas y opinión

Análisis de Datos

Información	Datos
La población de mujeres es mayor.	El 60% son mujeres.
Los hombres tienen mayor probabilidad de default.	24% hombres – 20% mujeres
Hay más solteros que casados en el conjunto de datos.	53% solteros – 45% casados
Los casados tienen mayor probabilidad de default.	23% Casado – 20% soltero
Hay más clientes con estudios universitarios.	47% Uni – 35% estudios posgrado En Total 82% con estudios superiores
Entre menos educación mayor probabilidad de default.	19% GS - 23% U - 25% HS
Créditos menores o iguales a $90000 con mayor probabilidad de default.	30% <= $90000 17% > $90000
Estatus de pago promedio mayor o igual a 1, mayor probabilidad de default.	63% >= 1 – 17% < 1
Entre más porcentaje pagues de tu factura mensual tiendes a tener menor probabilidad de default.	25% < rate(.5) – 15% > rate(.5)

Desarrollo del Modelo

- Se dividió el dataset en 3 partes desde el inicio, 1 para validar con 1000 registros y otro con 29000 para dividir en entrenamiento y prueba.
- No se encontraron valores duplicados o ausentes.
- Para minimizar la presencia de outliers de colocó un límite superior en el percentil 99 en las variables continuas.
- Para minimizar la presencia de outliers se agruparon categorías en las variables categóricas.

- INGENIERÍA DE CARACTERÍSTICAS
- Para la ingeniería de características se crearon nuevas variables por mes y con estas nuevas variables se intentó generar interacción entre ellas para poder darle información al modelo.
- Las variables también se agruparon en meses 2,3,4,5 y 6 y se utilizaron para crear nuevas características utilizando métricas como “promedio", "mínimo", "máximo", "mediana", "suma", “desviación estándar“.

Variable	Fórmula
RATE_REMINDER_LIMIT-BILL_LIMITn	(LIMIT_BAL-BILL_AMTn)/LIMIT_BALL
RATE_REMINDER_LIMIT-PAY_LIMITn	(LIMIT_BAL-PAY_AMTn)/LIMIT_BALL
RATE_BILL_LIMITn	BILL_AMTn/LIMIT_BAL
RATE_PAYAMT_LIMITn	PAY_AMTn/LIMIT_BAL
RATE_PAY_BILL_AMTn	PAY_AMTn/BILL_AMTn
DIF_BILL_PAYn	BILL_AMTn-PAY_AMTn

- SELECCIÓN DE VARIABLES
- Para la selección de variables de intent usar “K Best”, pero es un modelo que considero es de fuerza bruta por lo que se tiene que iterar en el valor K hasta encontrar uno óptimo por lo que tarda demasiado, por el hecho de que temenos cerca de 300 características después de la ingenieria de variables.
- Paso el mismo caso con el modelo de “Eliminación Recursiva de atributos”.
- También se probo eliminar variables por su correlación, en el caso que fuera mayor a .8 y se obtuvieron 59 variables pero no se llego a la puntuación de las variables generadas por “IV”.
- Para la selección de características se optó usar una selección basa en el “Valor de la Información (IV)” lo que nos dejo con 214.
- Para el escalamiento no se usó “WOE” debido a que no ibamos a crear un credit scoring por lo que se usó, el modelo Robust Scaler para así tener en cuenta algún valor atípico que pudieras haber dejando en el camino.

ALGUNAS VARIABLES DE IMPORTACIA (mejores 10 según Método IV)

Variables	Descripción	Relación
PAY_3, PAY_4, PAY_5, PAY_6	Histórico del estatus de pago.	Es el comportamiento del cliente en sus pagos por lo que hace sentido tenerlas en cuenta.
RATE_PAY_BILL_AMT_min (1,2),(1,3),(1,4)	La razón minima entre el PAY_AMT y BILL_AMT en los períodos indicados.	Es la razón minima que a cubierto el cliente para pagar la factura del mes, esto puede indicar que si la razón es más baja tiene mayor probabilidad de default.
PAY_AMT_min (1,2),(1,3)	Pago mínimo en los períodos indicados.	Esto podría dar más información al modelo para complementar la variable anterior.
PAY_AMT_sum (1,2)	Suma del los pagos en el período de tiempo indicado.	Esta variable podría indicar que el monto que pague en los dos períodos anteriores puede ayudar a indicar cuanto pagará en el siguiente.

- MODELOS Y SELECCIÓN
- Se compararon 9 modelos de clasificación:
  LogisticRegression, XGBClassifier, MLPClassifier, DecisionTreeClassifier, ExtraTreesClassifier, KNeighborsClassifier, AdaBoostClassifier, GradientBoostingClassifier y RandomForestClassifier.
- Se escogió XGBClassifier por sus métricas de estabilidad:
  Promedio: 0.74 Desviación Estándar: 0.007
- Se utilizó Random Search para reducir el sobreajuste y tratar de mejorar el modelo, pero se mantuvo en sus métricas.
- La puntación del modelo con la evaluación fue de .74.
- Se utilizó curva ROC para evaluar el modelo.

Respuestas

- PROPUESTAS
- Una de las recomendaciónes que se haría es en base a los clientes, que pagan más del monto facturado, se podrían buscar promociones para intentar que ese dinero que paga de más, lo gaste y tenga alguna oportunidad de retrasarse en el pago, esto si se quiere generar dinero a través del interés por mora.
- Buscar campañas enfocadas a personas con estudios universitarios o superiores además de solteros.

- LIMITACIONES DEL MÉTODO Y MEJORA

- Una de las limitaciones es la puntación, se puede predecir el estado de default de 2 de cada 3 clientes y 8 de cada 10 que son buenos clientes.
- Se puede mejorar buscando más variables como la transaccionalidad de sus pagos, en donde compra, cuánto, la frecuencia, etc.
- Además se pueden usar modelos con más capacidad para encontrar relaciones entre las variables como Redes Neuronales Profundas.

GRACIAS