2022

EVALUACIÓN Y CLASIFICACIÓN DE CLIENTES PARA UN CRÉDITO

Proyecto Final - Coderhouse

Carlos Arturo Velázquez Nolasco

ÍNDICE

CASO DE NEGOCIO

Problema abordado

Una entidad financiera quiere automatizar su proceso de aprobación de créditos para los posibles clientes, tomando como base 2 documentos:

- Solicitud de crédito del solicitante: Documento donde el solicitante llena información personal, entre los datos estan:
  - Salario anual total
  - Nivel de estudios
  - Si posee alguna propiedad o automóvil
  - Estado Civil, etc.

- Historial crédititicio del solicitante: Documento donde se listan por mes como ha sido el desempeño del solicitante con sus créditos (si tiene adeudos, etc).

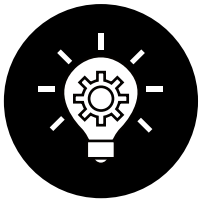
Con la información de esos documentos, la entidad necesita diferenciar entre los solicitantes a los que tengan buenos indicadores y mejor potencial para pagar sus créditos en tiempo y forma de los que tengan malos indicadores y sea muy probable que terminen siendo morosos

OBJETIVOS DEL PROYECTO

El objetivo final de este proyecto sera crear un modelo robusto de clasificación que haga la clasificación de los solicitantes buenos y malos y con ello, automatizar y mejorar la selección de solicitantes para la entidad financiera

N.º 01 — Analizar y tratar la información de los solicitantes

Uno de los principales objetivos del proyecto sera analizar la información disponible de los solicitantes, con base a los descubrimientos se haran un tratamiento de esta información para poder utilizarla



N.º 02 — Creación del modelo de clasificación

Una vez hecho el tratamiento de los datos, nuestro siguiente objetivo sera la creación del modelo. Para esto, se valoraran distintas propuestas de modelos y el que tenga mejor performance sera el seleccionado

DESCRIPCIÓN DE LOS DATOS

Como comentamos anteriormente, se ocupara la información de 2 archivos para la creación del modelo de clasificación. A continuación veremos la información detallada que provee cada archivo:

Historial créditicio de solicitante

Nombre de la variable	Descripción	Observaciones
ID	Identificador númerico del cliente	
MONTHS_BALANCE	El mes de la información recopilada	El mes de los datos extraídos es el punto de partida, al revés, 0 es el mes actual, -1 es el mes anterior, etc.
STATUS	El status del crédito del cliente	0: 1-29 días de atraso 1: 30-59 días de atraso 2: 60-89 días de atraso 3: 90-119 días de atraso 4: 120-149 días de atraso 5: Atrasos o deudas incobrables, cancelaciones por más de 150 días C: pagado ese mes X: No hay préstamo para el mes

Solicitud de crédito del solicitante

Nombre de la variable	Descripción	Observaciones
ID	Identificador númerico del cliente	
CODE_GENDER	Género del solicitante	
FLAG_OWN_CAR	¿El solicitante tiene un automovil?	
FLAG_OWN_REALTY	¿El solicitante tiene una propiedad?	
CNT_CHILDREN	Número de hijos	
AMT_INCOME_TOTAL	Ingreso anual en dolares	
NAME_INCOME_TYPE	Categoria del ingreso	
NAME_EDUCATION_TYPE	Nivel de estudios	
NAME_FAMILY_STATUS	Estado civil	
NAME_HOUSING_TYPE	Modalidad de vivienda	
DAYS_BIRTH	Dias de nacido	Cuenta hacia atrás desde el día actual (0), -1 significa ayer
DAYS_EMPLOYED	Dias empleado	Cuenta hacia atrás desde el día actual (0). Positiva si persona actualmente desempleada.
FLAG_MOBIL	¿El solicitante tiene teléfono móvil?	
FLAG_WORK_PHONE	¿El solicitante tiene telefono de trabajo?	
FLAG_PHONE	¿El solicitante tiene telefono fijo?	
FLAG_EMAIL	¿El solicitante tiene e-mail?	
OCCUPATION_TYPE	Tipo de ocupación	
CNT_FAM_MEMBERS	Tamaño de la familia	

Nuevas variables creadas

Nombre de la variable	Descripción	Observaciones
Target	Variable con los datos a predecir	Se creo en base a la variable STATUS
Edad	Edad del solicitante	Se creo en base a la variable DAYS_BIRTH
Empleo	¿El solicitante tiene empleo?	Se creo en base a la variable DAYS_EMPLOYED

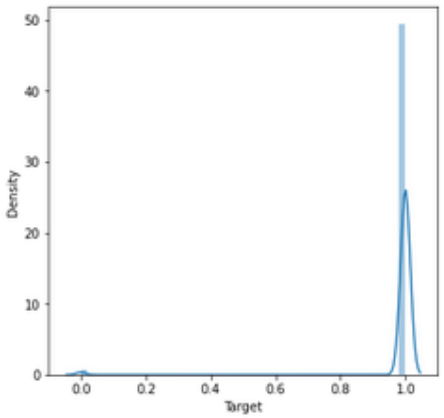
Variables seleccionadas para el modelo

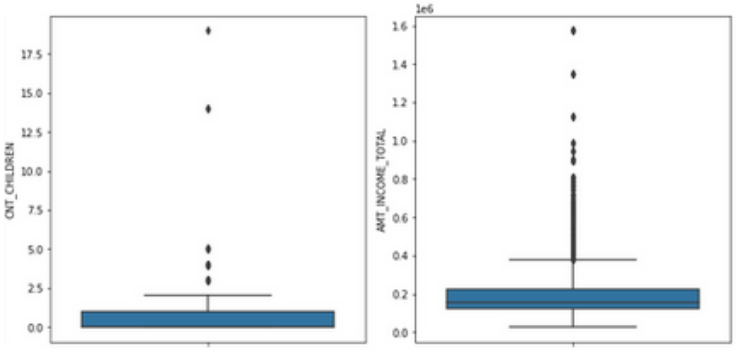
Nombre de la variable	Descripción	Observaciones
FLAG_OWN_CAR	¿El solicitante tiene un automovil?	
FLAG_OWN_REALTY	¿El solicitante tiene una propiedad?	
CNT_CHILDREN	Número de hijos	
AMT_INCOME_TOTAL	Ingreso anual en dolares	
CNT_FAM_MEMBERS	Tamaño de la familia	
Edad	Edad del solicitante	
Empleo	¿El solicitante tiene empleo?	
Target	Variable con los datos a predecir	

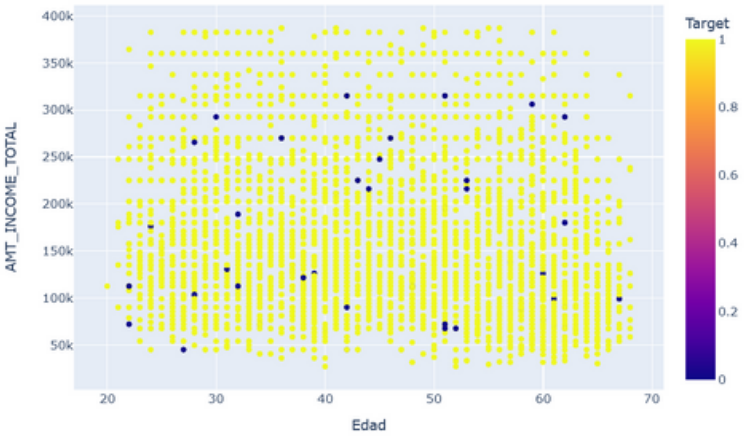
Hallazgos encontrados en el EDA

Durante el análisis y tratamiento de la información proporcionada, se realizaron eliminación y transformación de variables que no aportaban para la creación de nuestro modelo, de las variables seleccionadas para el modelo se encontraron los siguientes hallazgos:

- **Variable Target:** Los datos no contenian una columna target (columna con la información a la que queremos llegar) por lo que se tuvo que construir con la variable STATUS.
- **Valores nulos:** La información de los datasets no presento valores nulos a excepción de la variable OCCUPATION_TYPE, la cual tenia 34,203 datos nulos que representa el 30.59% de los datos. Esta fue una de las columnas que se retiro por lo que dataset final estuvo libre de datos nulos
- **Desbalanceo en los datos:** Una vez que tuvimos nuestra variable Target, se descubrio que la proporción entre buenos y malos clientes estaba desbalanceada (teniendo solo 1.33% de clientes malos vs un 98.77% de clientes buenos)
- **Valores outliers:** 3 de nuestras variables (CNT_CHILDREN, AMT_INCOME_TOTAL y CNT_FAM_MEMBERS) presentaban en menos del 5% de sus datos totales valores outliers. Se decidio eliminar estos datos para que no afectaran de ninguna forma al modelo
- **Correlación:** La variable Target tuvo nula correlación con todas las demás variables, esto sin embago, significa que no hay relación lineal más no que carece de algun tipo de relación.
- **Análisis Multivariado con base en la variable Target:** Al hacer un analisis de los datos con las variables Target, Edad y AMT_INCOME_TOTAL se descubrio que no hay una edad ni un salario que parezca abarcar los clientes malos de nuestro target, esta bastante bien distribuido.







ALGORITMO ELEGIDO Y METRICAS DE DESEMPEÑO

Antes de pasar con la creación de modelo, hay que recordar que estaba la situación de la información desbalanceada (con una proporción de 1.33% clientes malos contra 98.67% de clientes buenos en nuestra variable Target). Para contrarrestar esta situación se decidio implementar un sobre-muestreo de nuestra clase minoritaria (clientes malos), creando información sintentica basada en esos datos para asi tener una proporción balanceada en ambas clases

Una vez balanceados los datos, se implementaron los 3 algoritmos pre-seleccionados dando estos resultados:

- ***Decision Tree: 64% de exactitud***
- ***Random Forest: 94% de exactitud***
- ***Logistic regression: 49% de exactitud***
- ***XgBoost: 91% de exactitud***

Tomando en cuenta los resultados obtenidos, se tomo la decisión de elegir al algoritmo **Random Forest** como el algoritmo para nuestro modelo.

Algoritmos pre-seleccionados

- Decision Tree
- Random Forest
- Logistic regression
- XgBoost (algoritmo de boosting)

OPTIMIZACIÓN Y METRICAS FINALES DEL MODELO

Una vez seleccionado nuestro modelo, se decidio implementar 2 métodos de optimización para mejorar el rendimiento de nuestro modelo

El método de optimización ocupado fue:

- RandomSearchCV: Ayudara al algoritmo a encontrar el mejor conjunto de hyperparametros para nuestro modelo.

Método de Optimización	Accuracy
RandomSearchCV	95% de exactitud

Como podemos ver, la exactitud lograda por nuestro modelo optimizado es la misma que consiguio en un inicio, esto nos confirma que nuestro modelo en el apartado de hyperparametros esta lo más optimizado posible.

Con este optimización, nuestro modelo esta al 100% de su capacidad, por lo que sus métricas finales son:

Algoritmo y métricas finales

Random Forest

Métricas de desempeño:

Accuracy: 95%
Sensibilidad: 94%
Precisión: 96%
F1 Score: 95%

FUTURAS INICIATIVAS

Una vez terminado el modelo, consideramos necesario presentar estas posibles futuras iniciativas que puedan complementar el proyecto o ayudar a conseguir un mejor resultado en el futuro

N.º 01 — Complementar la información

Al no tener una variable Target, tuvimos que ocupar una de las variables para poder conseguirla. Esto hizo que esa variable ya no fuera utilizable y pudo haber ayudado a mejorar las métricas del modelo. Por lo que en futuros proyecto, se recomienda contar con esta columna para la fase de creación del modelo.

N.º 02 — Desbalanceo de los datos

Una de las situaciones que se tuvieron que abordar en el proyecto fue el desbalanceo de los datos entre clientes buenos y malos, a pesar de haber propuesto una solución la cual ayudo mucho a eliminar este problema, se podrian explorar más opciones o incluso, seleccionar para la información de entrenamiento un conjunto más balanceado de datos.

N.º 03 — Implementación de más algoritmos

Con 4 algoritmos pre-seleccionados pudimos obtener dos de ellos con un score bastante alto, sin embargo si se desea una exactitud cercana al 99%, se podrian explorar más algoritmos con distantas combinaciones de hyperparametros y aun más optimizaciones para lograr este objetivo

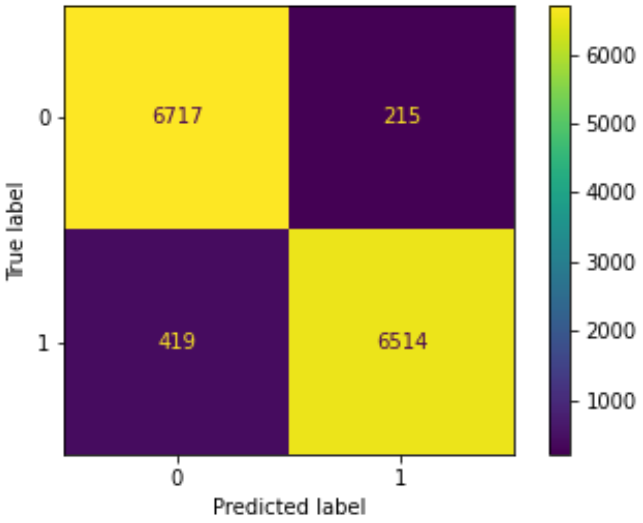
CONCLUSIONES

Como se pudo notar en la comparación nuestro algortimo de **Random Forest** fue el que mejor performance consiguio para clasificar a los solicitantes de todos nuestros algoritmos pre-seleccionado

Con este algoritmo, nuestro modelo es capaz de clasificar correctamente a 13,271 solicitantes (95% de los solicitantes de nuestro conjunto de test)

Del otro 5% restante:

- El 3% fueron falsos negativos (clientes buenos clasificados como malos)
- El 2% fueron falsos postivos (clientes malos clasificados como buenos)



Con estos resultados, nuestro proyecto puede automatizar el proceso de selección de los solicitantes de crédito con un porcentaje de predicción bastante bueno y con un error que no sobresale del 5%.

Hemos finalizado todas las etapas de nuestro proyecto, cumpliendo con los objetivos planteados y nuestro modelo esta listo para ser puesto en producción.