Detección de Fraude en Tarjetas de Crédito

Alumnos: Díaz Nathaly

Peña Jhonathan

CODER HOUSE

Junio 2022

Caso de Negocio

Siendo las tarjetas de crédito uno de los productos financieros más utilizados en la actualidad debido a la facilidad de acceso al compararlo con otras herramientas de financiación, brindando múltiples ventajas y beneficios al consumidor, también representa por otro lado un servicio que los bancos deben evaluar y determinar perfiles para el otorgamiento y aprobación dependiendo del tipo de cliente. De la mano de esta información se genera un importante perfil por definir: ¿Qué tipo de cliente tendrá dificultades para pagar la financiación con tarjeta de crédito?

Con el fin de minimizar los riesgos y ajustar montos o directamente negar el producto, se desarrolla un modelo de clasificación que tomando gran cantidad de variables y aprendiendo de estas, podrá identificar las principales características que son fuertes indicadores de incumplimiento de pago.

Tabla de Versionado

NOMBRE	CONTENIDO	VERSIÓN
PRIMERA ENTREGA	EDA (análisis univariado, bivariado, multivariado)	Versión 1
ALGORITMOS DE CLASIFICACIÓN	Aplicación algoritmo de clasificación	Versión 2
SEGUNDA ENTREGA	Aplicación varios algoritmos de clasificación	Versión 2
TERCERA ENTREGA	Transformaciones de la data	Versión 2

Objetivos del Modelo

Predecir el cliente que tendrá dificultades para pagar la financiación con tarjeta de crédito

Descripción de los Datos

Con la finalidad de alimentar el modelo se utilizaron dos dataset conformados por variables de tipo: numéricas, categóricas y binarias.

Previous_application.csv y Application_data.csv

Siendo un problema con fines financieros de las principales variables a trabajar fueron las contenedoras de monto de anualidad del préstamo que solicita, características sobre último cambio de teléfono, región poblacional a la que el cliente pertenece, edad del cliente en días, medición obtenida por el buró de crédito y otras variables que brindan información importante para definir el perfil que alimentará al modelo.

Hallazgos encontrados por el EDA

Siguiendo el objetivo de conocer el data set y las variables para realizar el análisis y posterior desarrollo del modelo, se aplicaron técnicas de exploratorias de los datos.

Análisis Univariado

En esta primera etapa de exploración de los datos se realizó un análisis estadístico generando las primeras métricas con el fin de conocer el tipo de datos con los que se va a trabajar, verificación de valores nulos y aplicación de técnicas para sustituirlos. De la mano de este procedimiento considerando cada variable por separado se generaron gráficos con el conteo de algunas de ellas, permitiendo una primera vista general de los datos.

Análisis Bivariado

En esta segunda instancia de exploración y ya teniendo un previo conocimiento de los valores se evalúa el comportamiento de los datos entre dos variables, comenzando a entrelazar la información para generar conclusiones como que los montos más altos son en retiros en efectivo, los días de más aplicaciones son lunes, martes y miércoles, la mayoría de los montos mayores a 30k no son aprobados y el tipo de cliente con mayor participación son clientes ya activos en el servicio.

Análisis Multivariado

Al analizar comportamientos de diversas variables dentro del dataset y ya unificando la información obtenida en los análisis anteriores resaltan las variables más importantes para el desarrollo del modelo, luego de detectar las que se van a emplear finalmente se realiza un conteo de la variable target que se utilizará y esta define el dataset como imbalanceado, dentro del cual:

1 = Clientes con dificultades para pagar

0 = Todos los demás casos

Las variables principales son:

VARIABLE	DESCRIPCIÓN
DAYS_BIRTH	Edad del cliente en días al momento de la solicitud
AMT_CREDIT	Monto del crédito del préstamo
REGION_POPULATION_RELATIVE	Población de la región donde vive el cliente
AMT_ANNUITY	Anualidad de la aplicación
EXT_SOURCE_2	Puntuación de fuente de datos externa
EXT_SOURCE_3	Puntuación de fuente de datos externa
DAYS_ID_PUBLISH	Cantidad de días antes de la solicitud que el cliente cambió de documento de identidad
DAYS_EMPLOYED	Cuántos días antes de la solicitud la persona comenzó el empleo actual
DAYS_REGISTRATION	Cuántos días antes de la solicitud el cliente cambió su registro
DAYS_LAST_PHONE_CHANGE	Cuántos días antes de la solicitud el cliente cambió su número telefónico

Algoritmo Elegido

El algoritmo elegido luego de probar varios modelos fue **Random Forest**, ya que al momento de evaluar otros entre ellos XGBOOST, SVM, REGRESIÓN LINEAL, ÁRBOL DE DECISIÓN, resaltaba que desde el principio las métricas de F1 score sin haber hecho ninguna optimización ya arrojaba una buena probabilidad al estimar la clase target, dando los primeros indicios de ser el que mejor podría funcionar

Métricas de Desempeño del Modelo

En primera instancia se evaluaron las métricas en todos los modelos para hacer un análisis comparativo

Train_Accuracy	Test_Accuracy	Precision	Recall	AUC	F1_Score	F2_Score	Roc_Auc_score	Matthews_corrcoef	name
0.919129	0.919429	0.426829	0.004716	0.730247	0.009330	0.494163	0.730247	0.037988	Logistic Regression
0.921053	0.921391	0.764890	0.032880	0.799723	0.063049	0.511663	0.799723	0.148241	XGBoost
0.919106	0.919440	0.422535	0.004043	0.732361	0.008009	0.493767	0.732361	0.034908	SVM
0.936174	0.935884	0.992157	0.204555	0.948694	0.339180	0.614661	0.948694	0.435327	Random Forest

Debido a que las métricas arrojadas en estas primeras iteraciones no eran buenas, el siguiente paso fue realizar optimizaciones hasta llegar al modelo de brindara las mejores métricas al momento de predecir la variable target

Iteraciones de Optimización

Se agregaron los parámetros stratify, técnicas de muestreo como oversampling, undersampling y oversampling y undersampling a la vez

Métricas Finales del Modelo Optimizado

Train_Accuracy	Test_Accuracy	Precision	Recall	AUC	F1_Score	F2_Score	Roc_Auc_score	Matthews_corrcoef	name
0.919129	0.919429	0.426829	0.004716	0.730247	0.009330	0.494163	0.730247	0.037988	Logistic Regression
0.921053	0.921391	0.764890	0.032880	0.799723	0.063049	0.511663	0.799723	0.148241	XGBoost
0.919111	0.919407	0.333333	0.001887	0.732460	0.003752	0.492466	0.732460	0.019844	SVM
0.936625	0.937423	0.994005	0.223420	0.948989	0.364837	0.625460	0.948989	0.455804	Random Forest
0.945178	0.946026	0.909233	0.955374	0.987866	0.931732	0.945981	0.987866	0.887891	Random Forest with undersampling and oversampling

Futuras Líneas

Finalmente luego de realizar un completo análisis exploratorio y evaluar las variables de mayor importancia para el modelo, los siguientes pasos para llevarlo al siguiente nivel y evaluarlo en otras instancias sería desarrollarlo a un entorno de testing en donde se podría explorar en tiempo real como se comporta, si las proyecciones se pueden llevar a cabo y luego trasladarlo a un entorno de producción usando técnicas de MLops (contenedores-Docker, monitoreo-Grafana, Cloud-Azure, AWS, GCP)

Se busca complementar el modelo al utilizar data de un periodo de tiempo determinado para hacer una comparativa de las pruebas de testeo y evaluar cómo se comporta con respecto a la data real, optimizando con cada aporte el servicio brindado y beneficiando financieramente los riesgos asumidos por el sector bancario al otorgar la aprobación del mismo