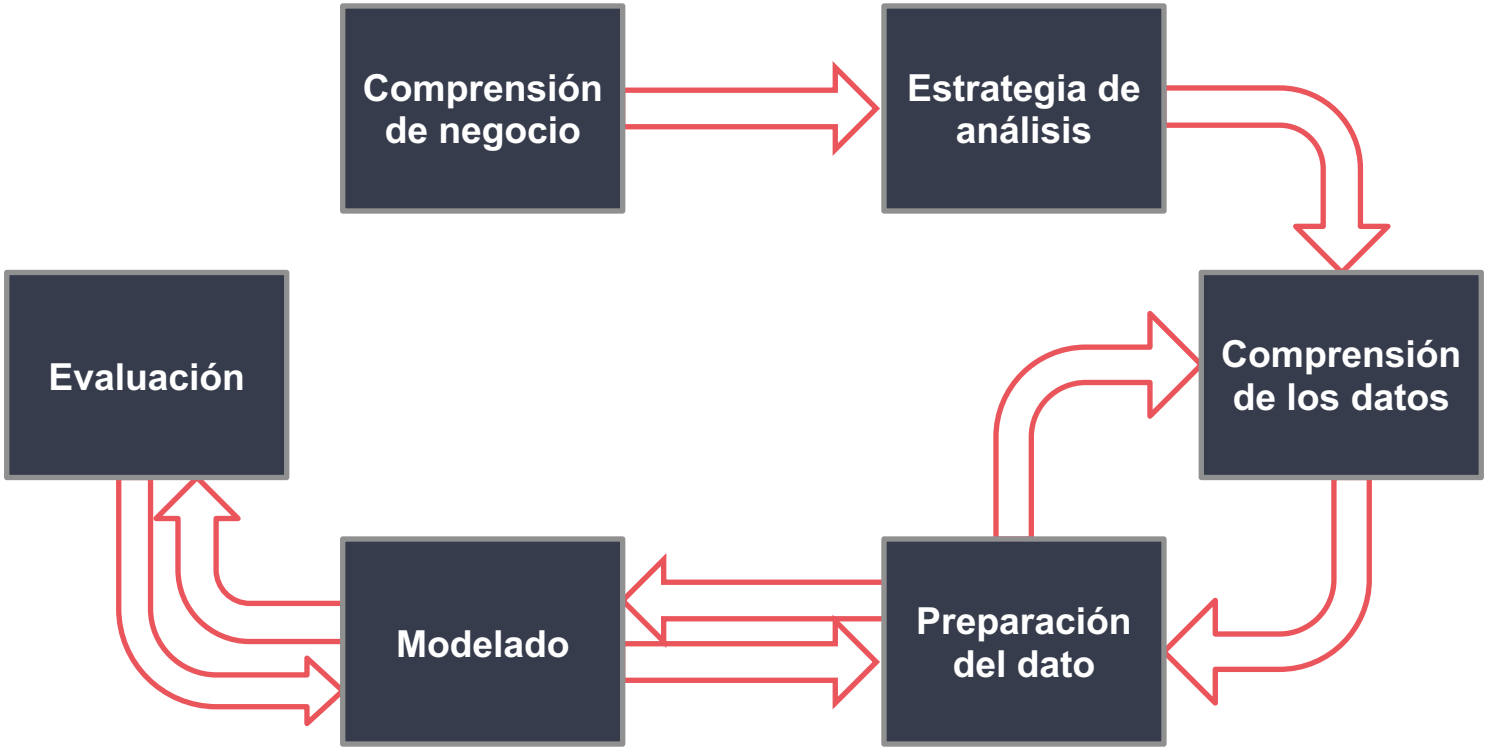
Modelo predictivo de detección de fraude en tarjetas bancarias

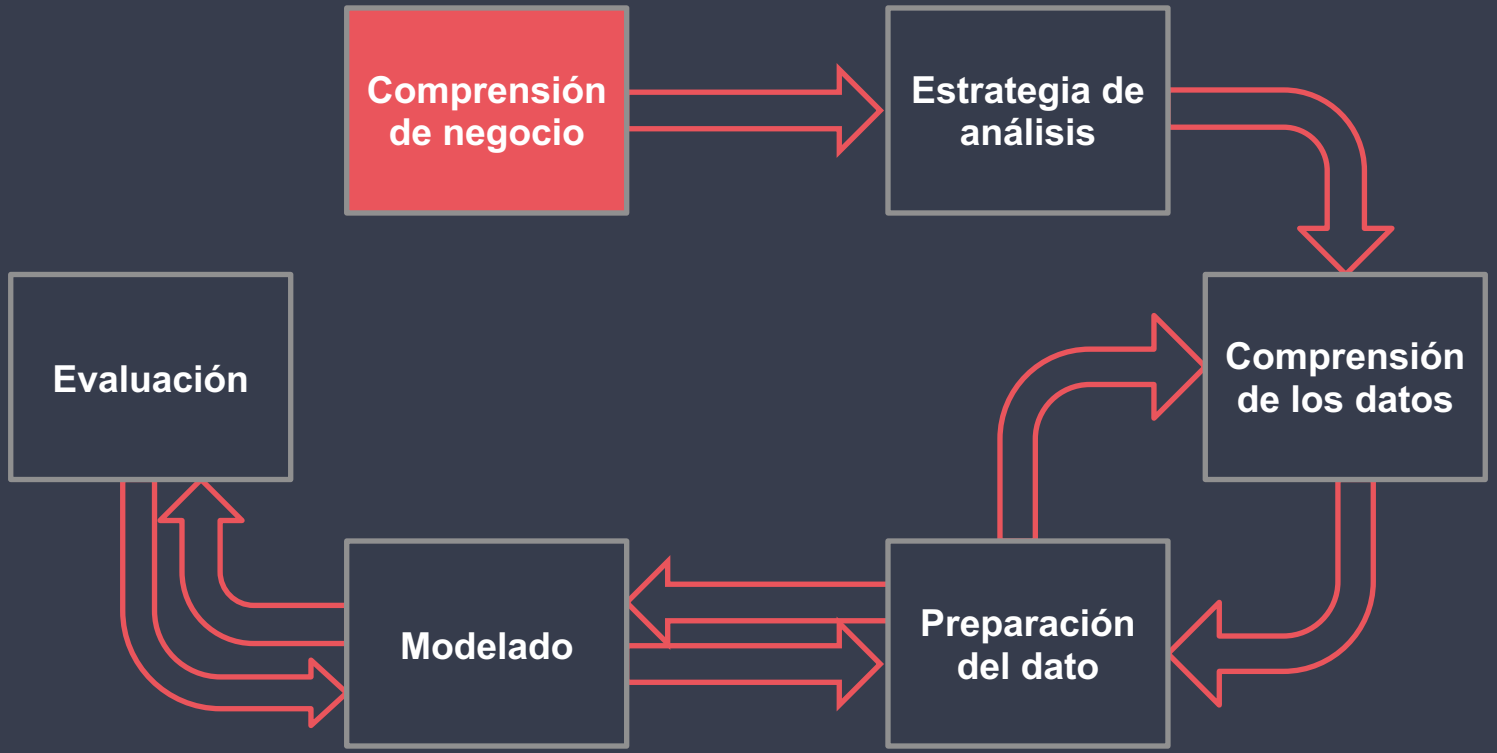
Sergio Mañanas López
Juan José Hernández Villlar
Diciembre 2020

NEOLAND

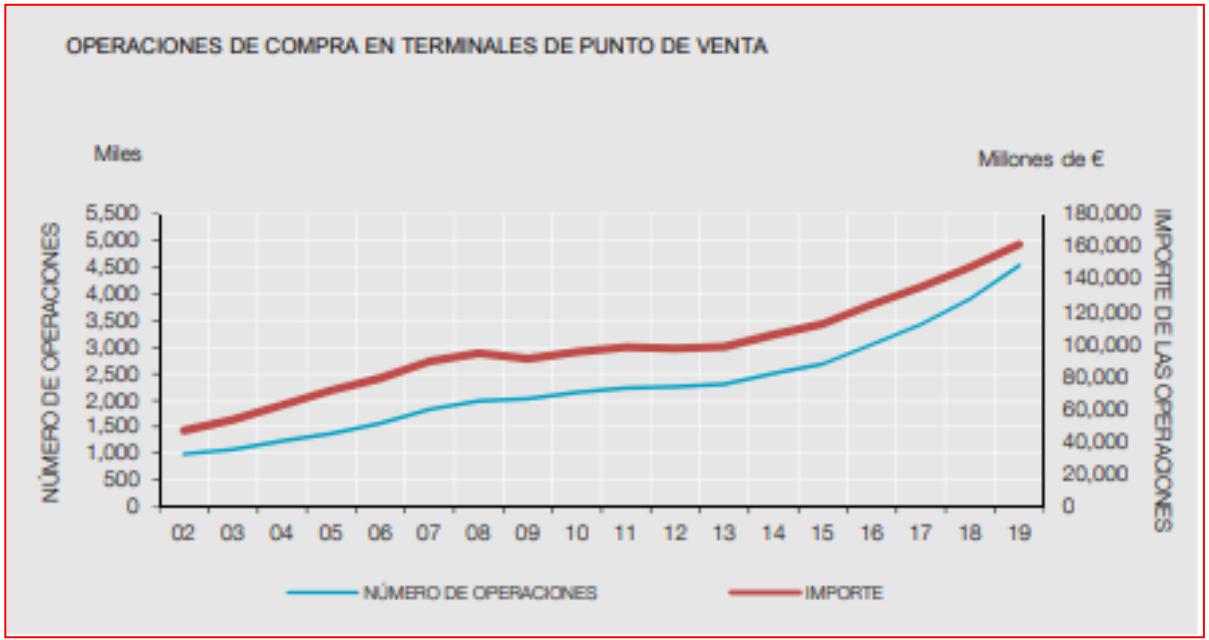
Metodología del proyecto



Basado en Metodología CRISP


Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

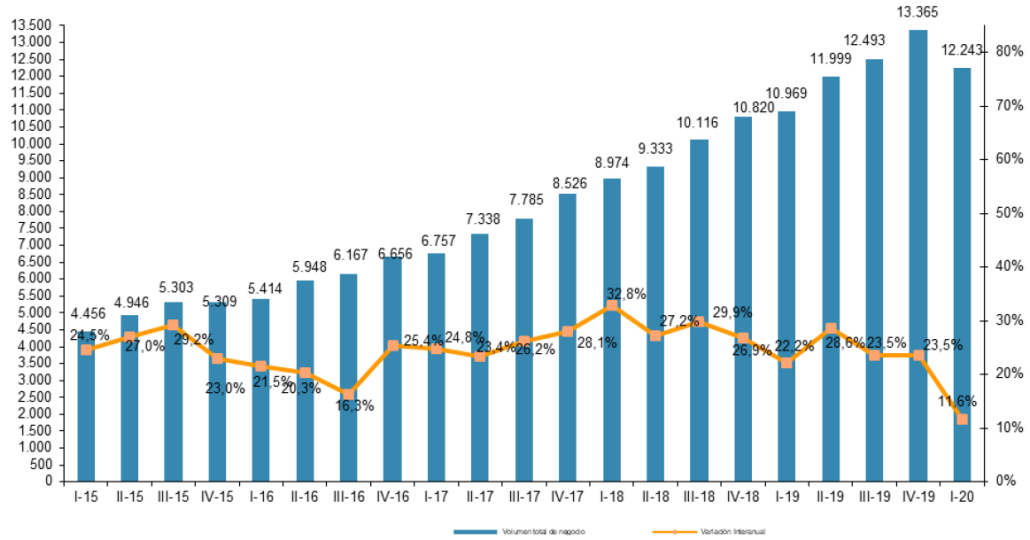
1. Comprensión de Negocio



- En 2019, se realizaron **4,5 billones de operaciones con tarjetas** en TPV's en España, con un importe de **160.000 millones de euros**, según el BDE.

1. Comprensión de Negocio

EVOLUCIÓN TRIMESTRAL DEL VOLUMEN DE NEGOCIO DEL COMERCIO ELECTRÓNICO Y VARIACIÓN INTERANUAL (millones de euros y porcentaje)



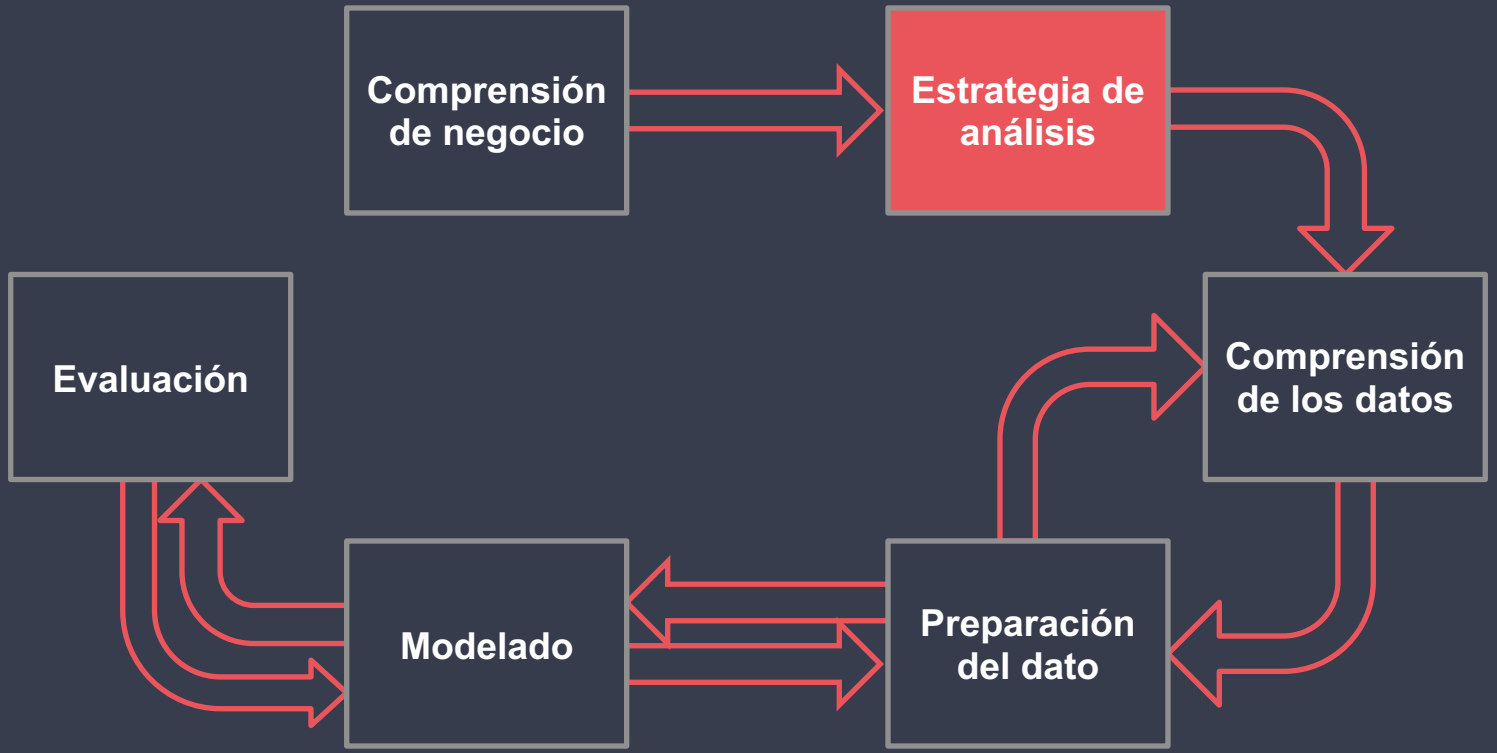
Fuente: CNMC

- En el primer trimestre de 2020, el **comercio electrónico** supero los **12.200 millones de euros** según la CNMC.

1. Comprensión de Negocio

- En 2018, se registraron más de **un millón de operaciones fraudulentas con tarjetas** por importe de **88 millones**, con el **64 % por operativa remota** y el 34% por TPV.

- El coste de este fraude **repercute tanto en el banco como en los clientes**, no solo por la cantidad defraudada, también los seguros contratados, departamentos de gestión, y el coste de reputación de la entidad financiera y de satisfacción de sus clientes.

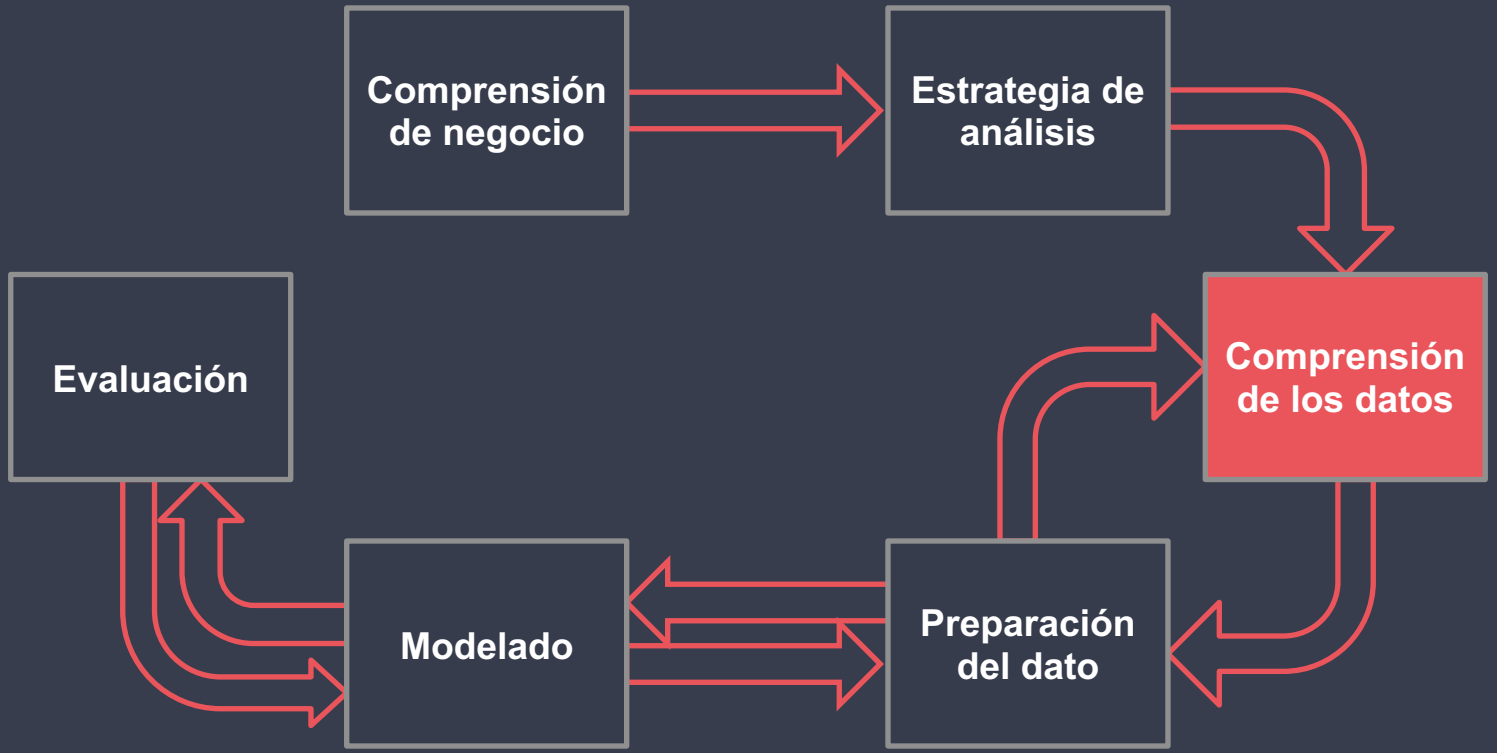

Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

2. Estrategia de análisis

El problema principal en la detección de fraude **es el desbalance de clases.**

En nuestro dataset, representa **solo un 3.5%** de todos los registros.

1. Para resolver el modelo, se elegirá entre varios **clasificadores de boosting.**
2. La métrica a utilizar es el **área bajo la curva ROC (AUC).**
3. Analizaremos los resultados a través de **Uplift modelling** para segmentar las transacciones y prescribir que cuantiles detener para maximizar la detección de fraude reduciendo los falsos positivos.


Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

3. Comprensión de los datos

La principal dificultad para el análisis de las variables es **la anonimización** de los datos reales. Las columnas poseen contenido con significado enmascarado.

Transaction columns	C_i columns	D_i columns	M_i columns	V_i columns
Categóricas y numéricas	Numéricas	Numéricas	Categóricas	Numéricas
…	…	…	…	…
…	…	…	…	…

1 transacción por fila

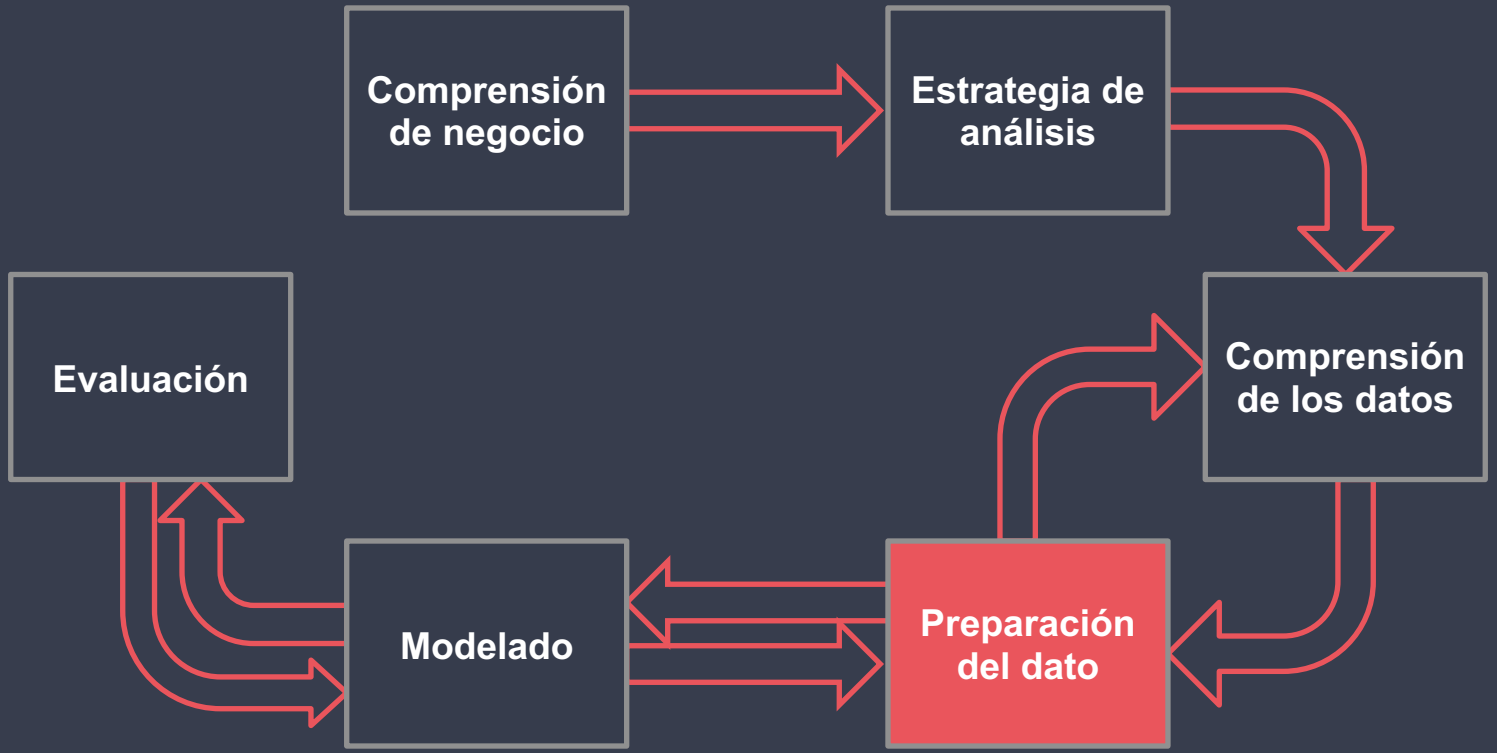
Tamaño de **Train**: 590540 filas x 394 columnas
Tamaño de **Test**: 506691 filas x 393 columnas

3. Comprensión de los datos

1. **Análisis exploratorio** de las variables.

2. **Feature Selection** a partir de correlacion de Pearson , número de NaNs y **Feature Importance.**

3. **Análisis** de la **identificación de cliente** / tarjeta utilizando **Adversarial Validation**, con Catboost y **LightGBM.**


Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

4. Preparación del dato

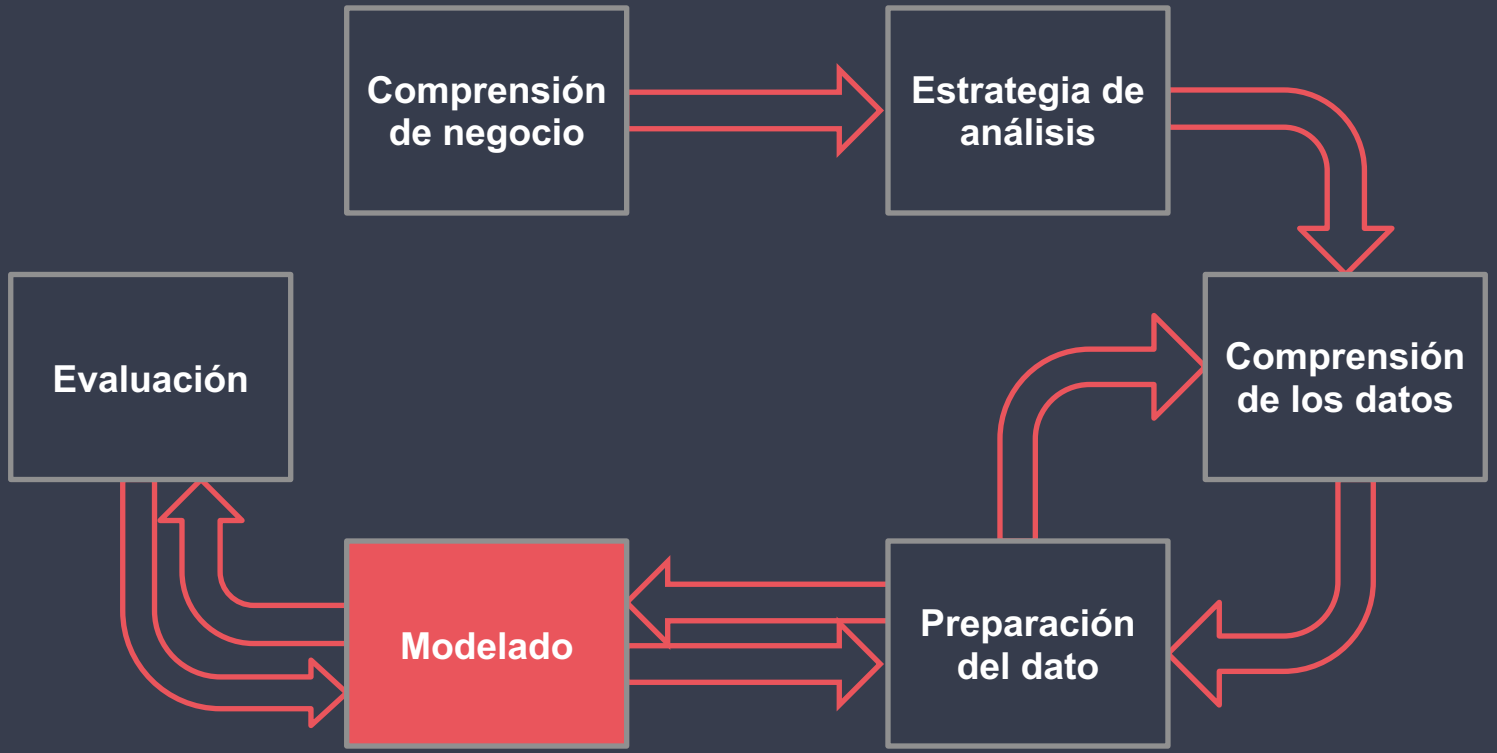
Preprocesamiento adaptado a LightGBM y XGBoosting, dependiendo del resultado obtenido en cada uno de los modelos:

LightGBM XGBoosting

- **Conversión de valores NaNs** de campos de tarjeta ($card_i$), direcciones ($addr_i$), columnas de concordancias (M_i) y dominios correo (emaildomain).
- Conversion de columnas de strings a numéricas a través de **Label Encoding.**

XGBoosting

- **Conversión del resto de NaNs** a un valor definido: -999, -1.
- **Feature Engineering**: Frequency encoding a las mejores columnas según Feature Selection.

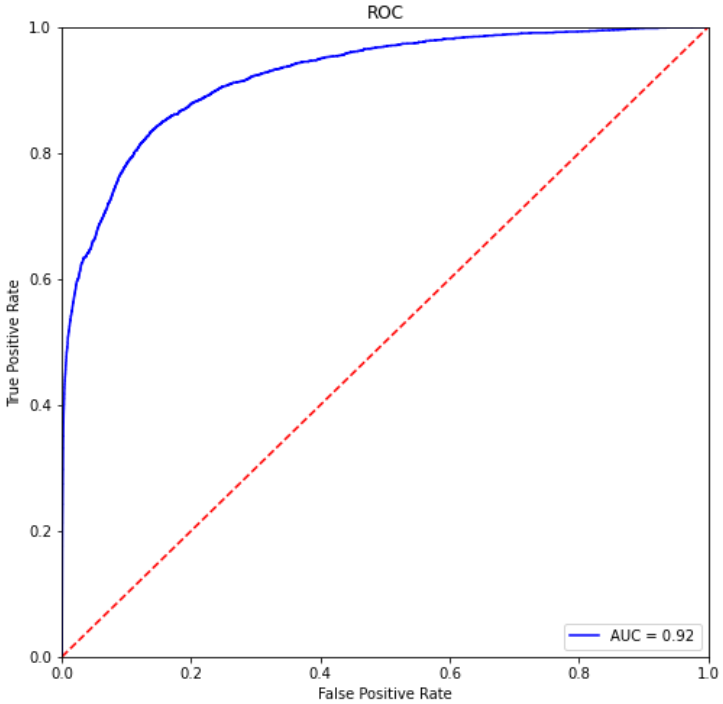

Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

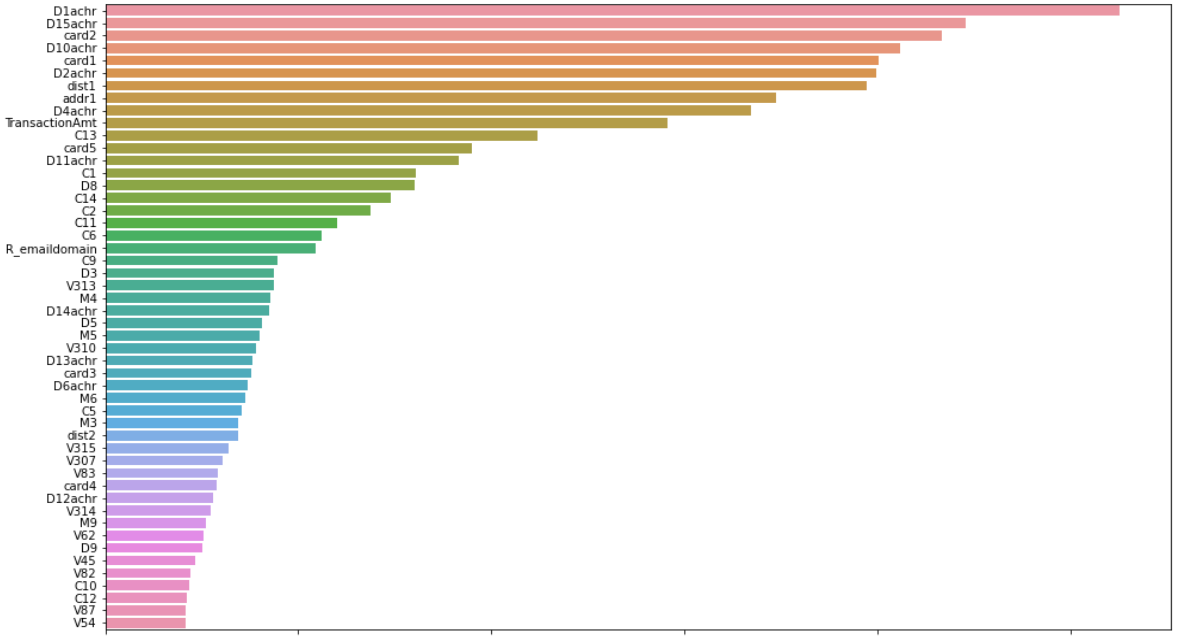
5. Modelado

#	learning_rate	num_leaves	max_depth	colsample_bytree	subsample	AUC
1	0.03215518473989215	294	14	0.4	0.8	0.9194457912401148
2	0.04	278	13	0.4	0.9	0.9164233793906151
3	0.027832622205538907	268	15	0.3740990054742914	0.589171882228726	0.9188929389573968
4	0.029827755272285032	262	15	0.7823490596970961	0.8631490082348747	0.92719808846762

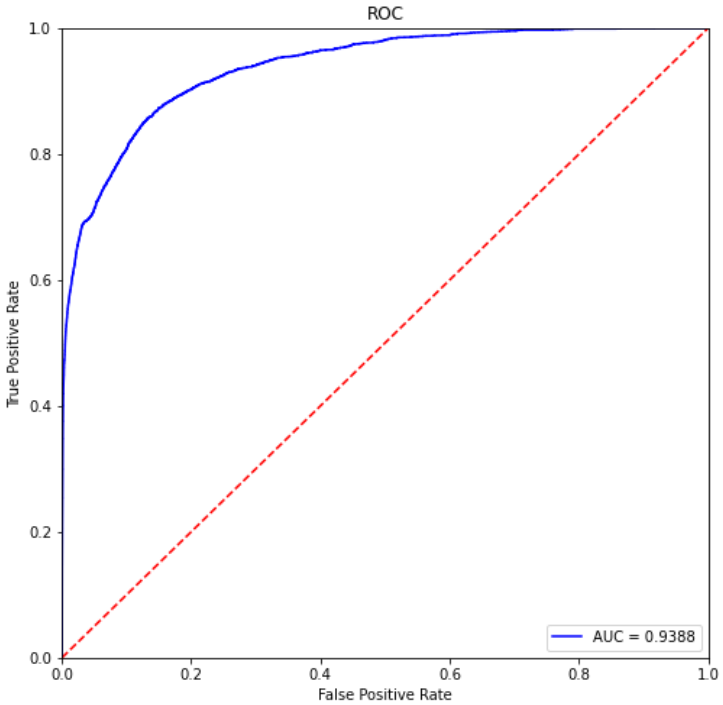
Optimización de parámetros a través de Bayesian Optimization

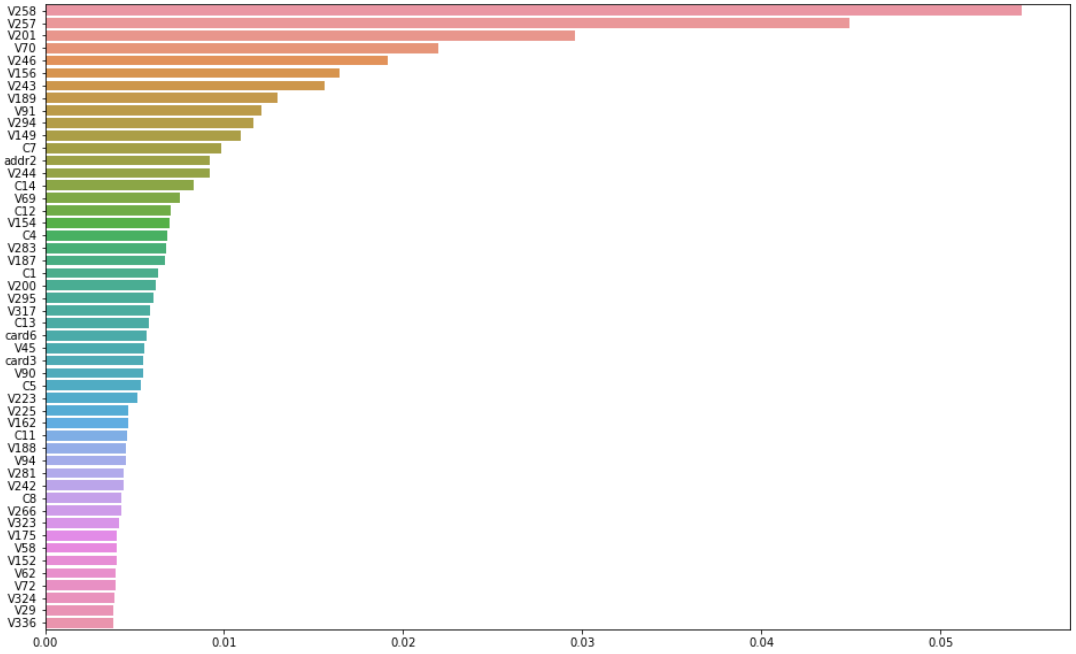
5. Modelado: LightGBM

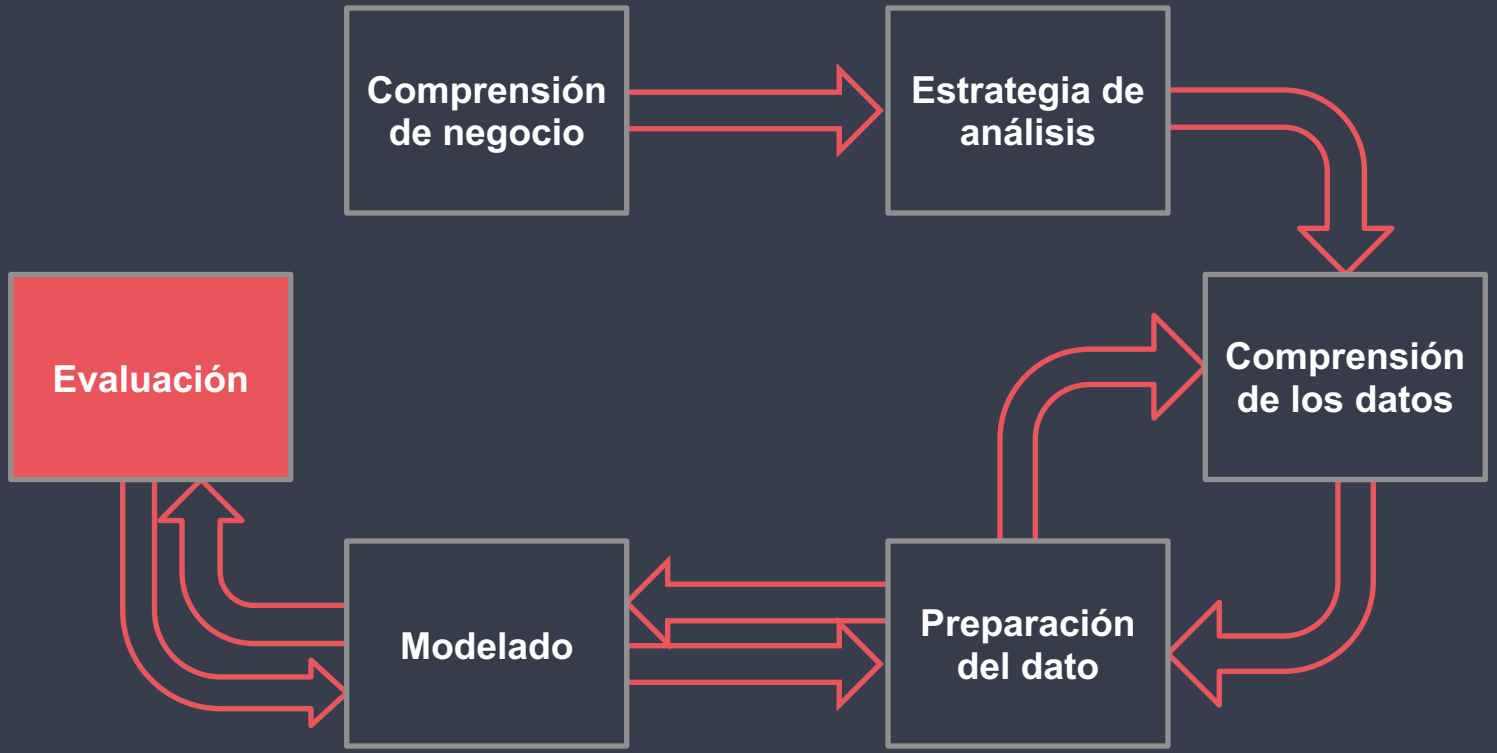




5. Modelado: XGBoost

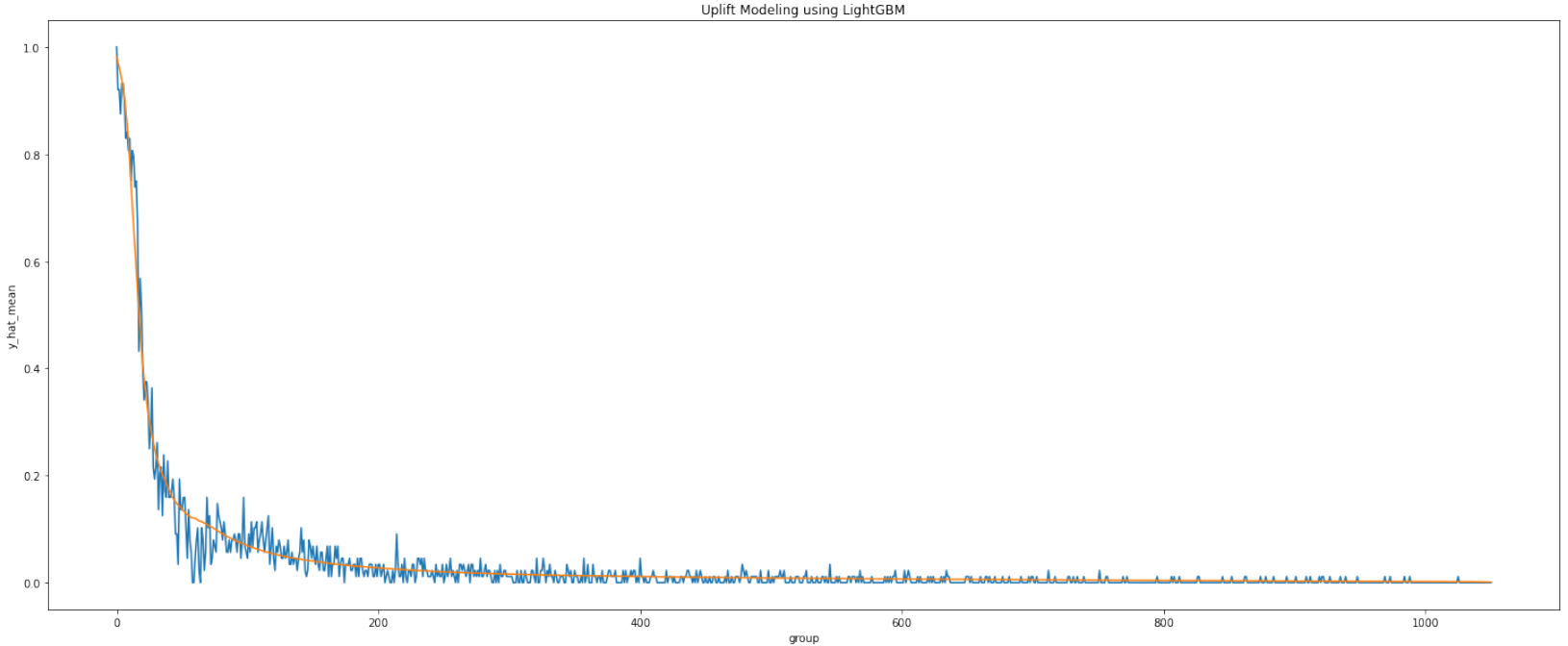





Comprensión de negocio
Estrategia de análisis
Comprensión de los datos
Preparación del dato
Modelado
Evaluación

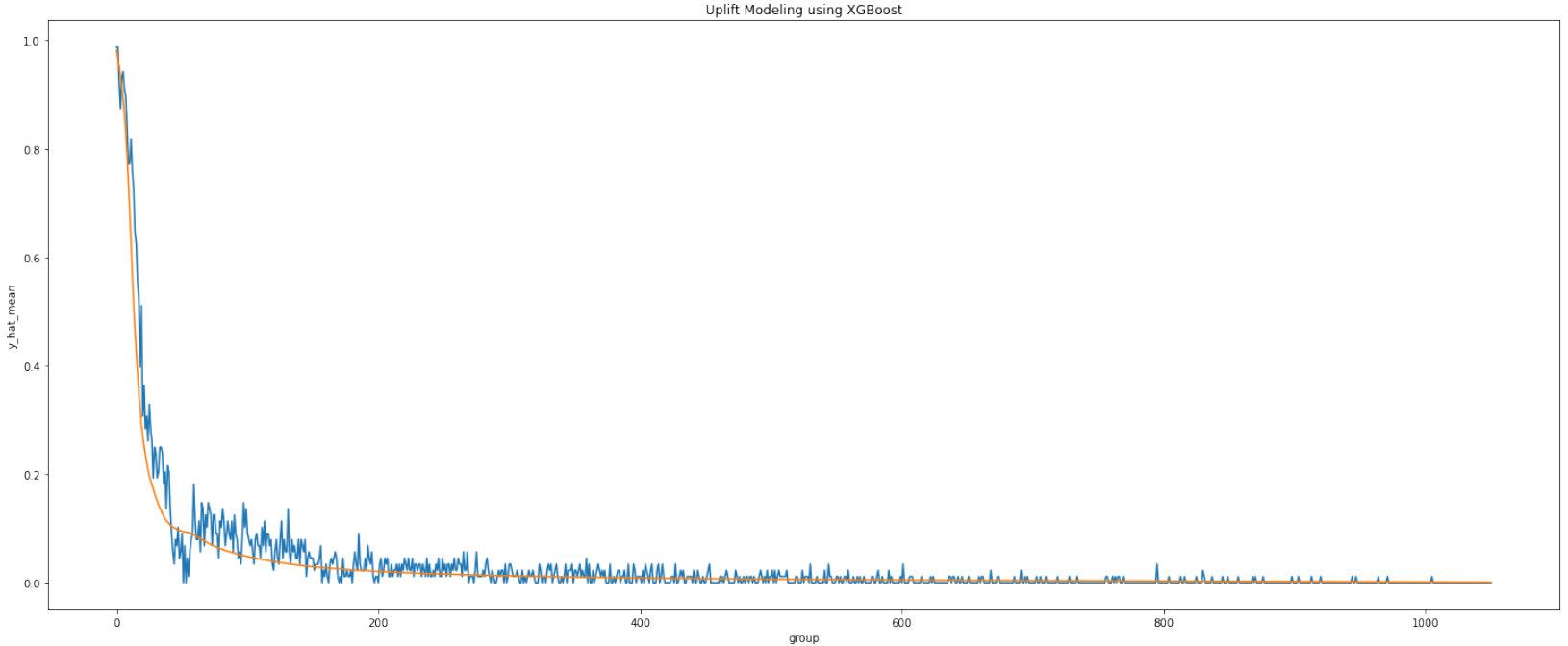
6. Evaluación: LightGBM

Model	Local Score	Private Score	Public Score	Overfitting with Local
LightGBM	0.923519	0.905291	0.927122	1.97%
XGBoost	0.938792	0.905261	0.929851	3.42%



6. Evaluación: XGBoost

Model	Local Score	Private Score	Public Score	Overfitting with Local
LightGBM	0.923519	0.905291	0.927122	1.97%
XGBoost	0.938792	0.905261	0.929851	3.42%



7. Trabajo a futuro

1. Evaluar **nuevas variables generadas por Feature Engineering**, controlando AUC y overfitting, hasta obtener el máximo área debajo de la curva ROC posible y poder reducir con ello los falsos positivos al mínimo.

2. Comparar los clasificadores de boosting con **utilizar redes neuronales**.

3. Aplicar **SMOTE** para intentar balancear las clases.

4. Aplicar **optimizadores de hiperparámetros en XGBoost**.

¡GRACIAS!

Pasamos a la presentación del código...

NEOLAND