Universidad Ricardo Palma

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

PROGRAMA DE ESPECIALIZACIÓN DATA SCIENCE NIVEL II

R + Python

MÓDULO VI

Taller Aplicativo de Algoritmos de Machine Learning

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

PROGRAMA DE ESPECIALIZACIÓN EN DATA SCIENCE NIVEL II

« Un esfuerzo mas y lo que iba a ser un fracaso se convierte en un éxito; no existe el fracaso si nos esforzamos cada vez mas y mas»

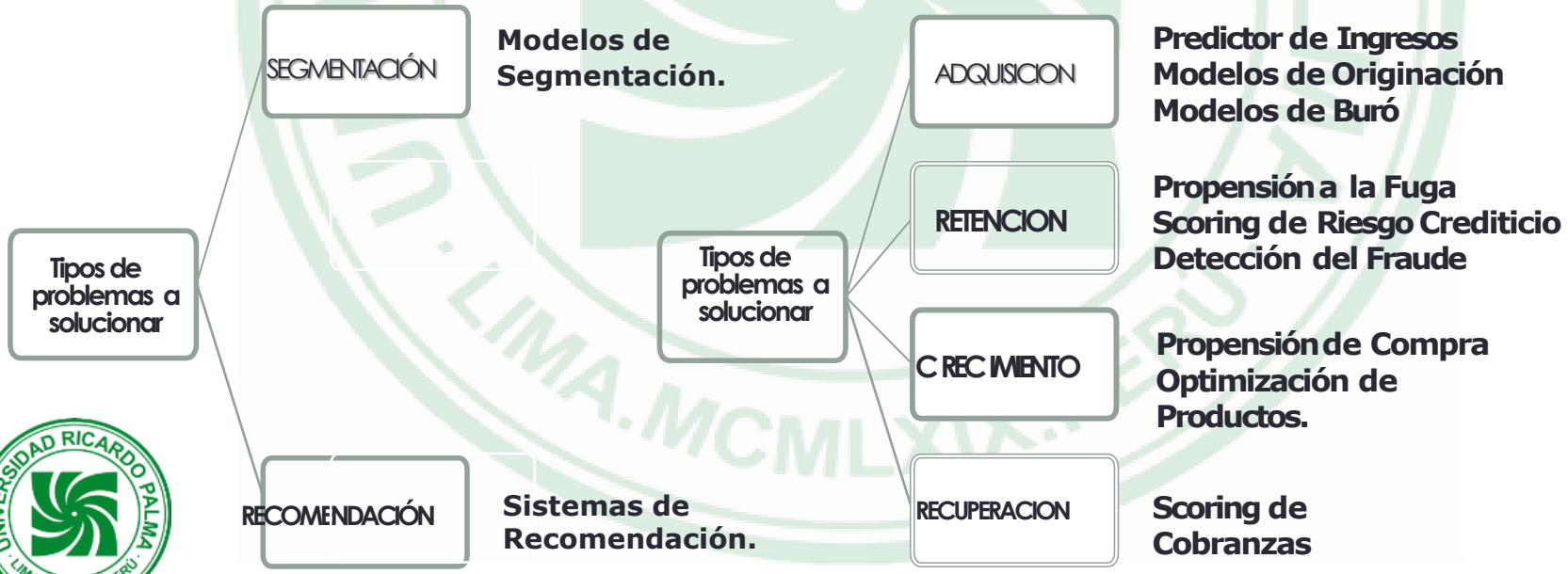
Marat

ENTENDIMIENTO DEL NEGOCIO :ENTENDIMIENTO DEL PROBLEMA

PROPÓSITO DEL ANÁLISIS

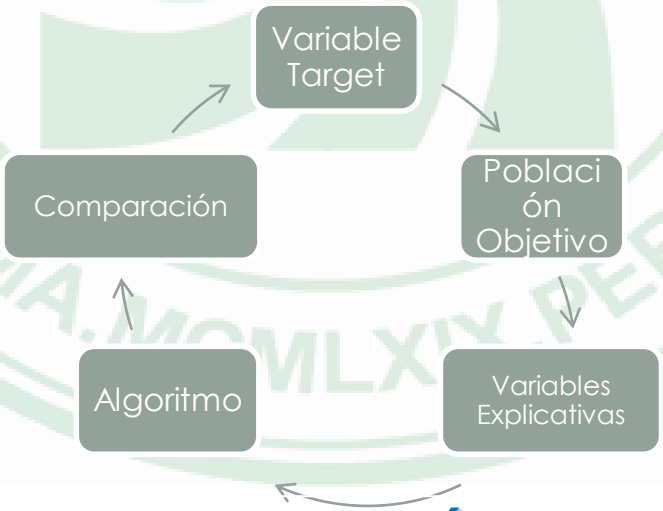
Descubrir eventos o resultados futuros en base al conocimiento previo de los datos, utilizando para ello métodos estadísticos, matemáticos, computacionales y de base de datos, así como de la aplicación de los algoritmos de machine learning. En cualquier negocio el éxito depende de:

- ✓ Capacidad de recopilar y limpiar la información para el análisis.
- ✓ Capacidad de Identificar los patrones y tendencias de los datos en relación a lo que se desea solucionar.
- ✓ Capacidad de crear el modelo que le de valor al negocio.



ENTENDIMIENTO DEL NEGOCIO :ENTENDIMIENTO DEL PROBLEMA

1.¿Que problema quiero solucionar?	Variable objetivo o de respuesta (Y)	Ejem: Estimar ingresos de personas no bancarizados
2. ¿Con qué población analizo el problema?	Población objetivo	Ejem: Dependientes e Independientes
3.¿Qué indicadores pueden explicar el problema?	Covariables (Xs)	Ejem: NSE (Reniec), Tipo de automovil (Sunarp)
4. ¿Qué técnica estadistica o informática se ajusta al análisis?	Métrica o algoritmo	Ejem: Arboles de decisión



Entendimiento del Negocio : Criterios de Inclusión y Exclusión de la Información

- ✓ Tienen poca información histórica
- ✓ Son riesgosos para el negocio
- ✓ Están fuera de las políticas o de las estrategias del negocio
- ✓ Son datos erróneos
- ✓ Son casos especiales que no se volverán a recopilar
- ✓ Ciclos económicos
- ✓ Temas regulatorios / legales

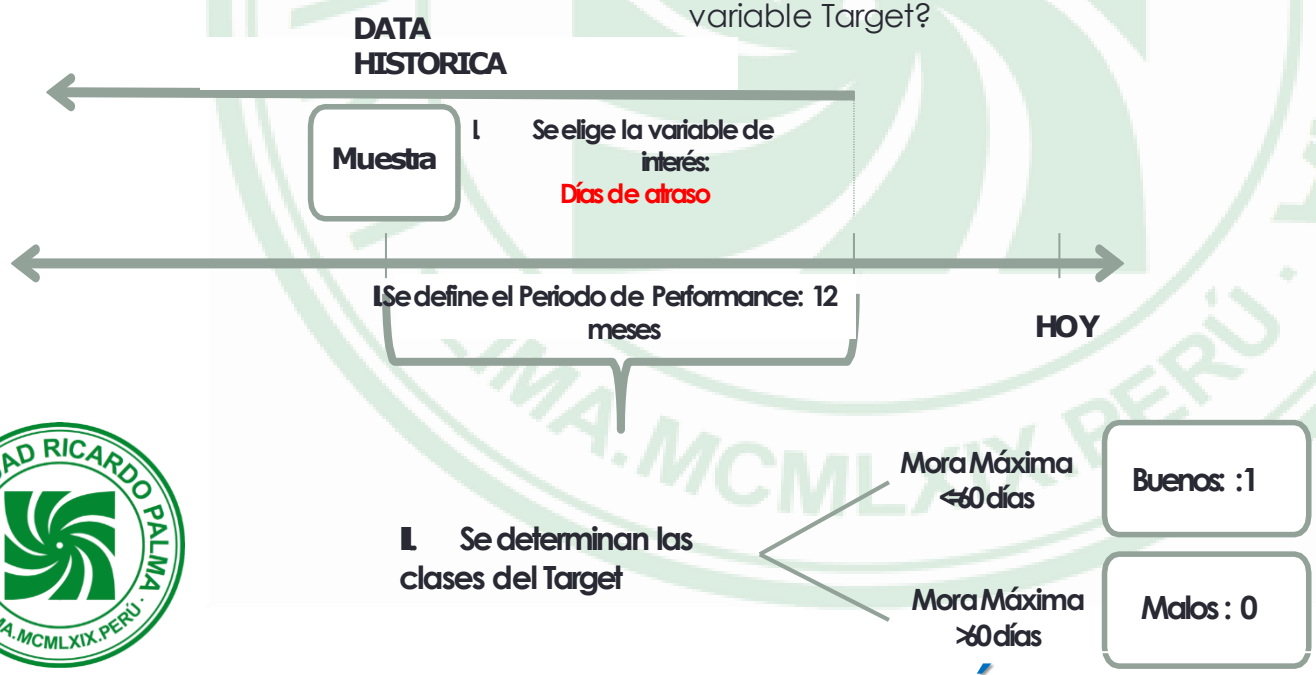
ENTENDIMIENTO DEL NEGOCIO : DEFINICIÓN DE LA VARIABLE TARGET

Los pasos para crear una variable target de clasificación son:

- **Primero**: Elegir la variable(es) de interés para crear el target.
- **Segundo**: Definir el horizonte temporal del periodo de performance o predicción.
- **Tercero**: Determinar las clases del indicador según los cortes de la variable(es) de interés.

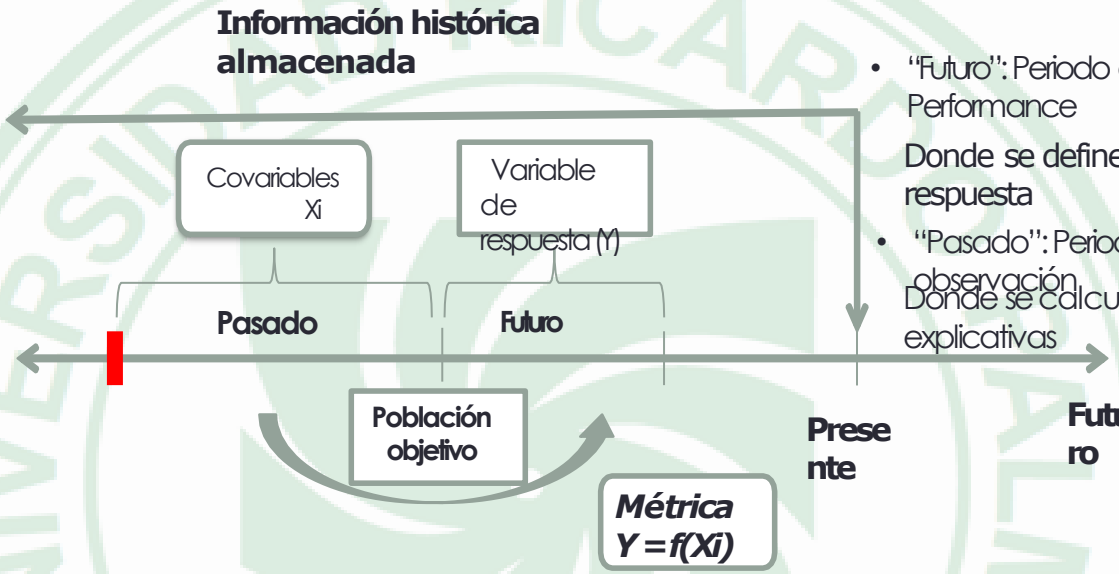
Por ejemplo para un modelo de buró:

Un modelo de buró es un modelo de riesgo crediticio que mide la incertidumbre de los clientes bancarizados. ¿Cómo podría ser su variable Target?



ENTENDIMIENTO DEL NEGOCIO : DEFINICIÓN DE LA VARIABLE TARGET

Enfoque temporal:



- "Futuro": Periodo de Predicción o Performance

Donde se define a la variable de respuesta

- "Pasado": Periodo de observación

Donde se calculan a las variables explicativas

Linea de tiempo

Enfoque matricial:

ID	Segment_Target	Var_Target	Var_X1	Var_X2	Var_X3	Var_X4	Var_X5	Var_X6
1	Segment 1	1	-0.243257655	216	952.4800	1	4	3
2	Segment 2	1	1.696358794	191	633.4949	0	7	4
3	Segment 3	1	0.561226988	192	637.5107	0	6	3
4	Segment 1	1	-1.673888687	205	927.2513	0	8	3
5	Segment 2	0	-0.315746538	200	988.0877	0	2	3
6	Segment 3	0	0.402197729	201	927.5218	1	6	2
7	Segment 1	1	0.668736379	202	582.0028	0	6	2
8	Segment 2	1	1.489475004	197	701.1748	0	6	2
9	Segment 3	0	0.308647509	201	526.3747	0	8	4
10	Segment 1	1	0.090616380	189	989.2571	0	7	4
11	Segment 2	1	0.081223506	200	789.0298	0	8	2
12	Segment 3	1	-0.443663814	207	937.3809	0	2	2
13	Segment 1	1	-1.416088194	220	819.6118	0	9	1
14	Segment 2	1	-0.316298576	187	995.7736	1	2	5

Población objetivo

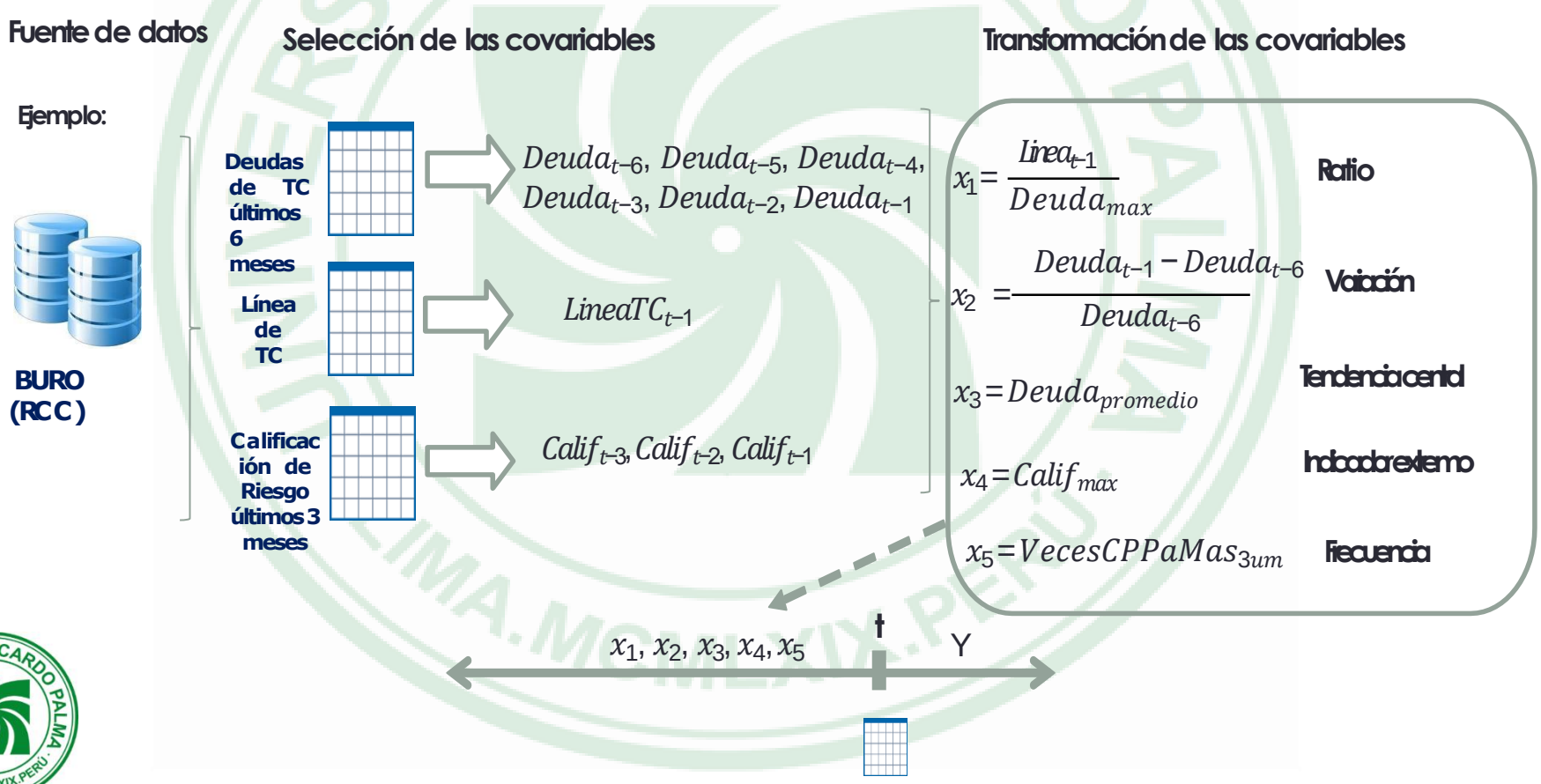
Variable de respuesta (Y)

Covariables Xi

Métrica
Var_Target= f(Var_X1,Var_X2, Var_X3, Var_X4,Var_X5,Var_X6)

ENTENDIMIENTO DEL NEGOCIO : DEFINICIÓN Y CREACIÓN DE DRIVERS

Las variables a seleccionar para la solución del problema propuesto deben tener **sentido para el negocio**. En otras palabras al seleccionarlas se espera que estén correlacionadas con la variable de respuesta del modelo. La transformación tiene como propósito optimizar el aporte de las Xi en el modelo.



Sesgo y Varianza en un Modelo Predictivo o Algoritmo de Machine Learning

Sesgo: Representa el bajo nivel de precisión del modelo como consecuencia de que no se ajusta lo suficiente a los datos.

$$\text{Bias}\left[\hat{f}(x)\right] = \text{E}\left[\hat{f}(x) - f(x)\right]$$

Varianza: Representa la volatibilidad del predictor debido a que está excesivamente ajustada a una data particular (data con la que se construyó). Así el modelo pierde su propiedad de generalización y se dice que existe sobreajuste (over-fitting)

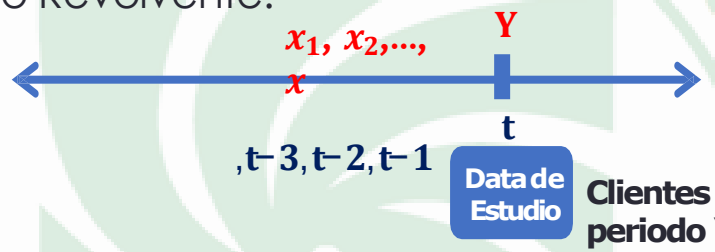
$$\text{Var}\left[\hat{f}(x)\right] = \text{E}[\hat{f}(x)^2] - \text{E}[\hat{f}(x)]^2$$

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

CASO : PREDICTOR DE INGRESOS

El negocio necesita contar con un **predictor de ingresos** para ser usado en las campañas masivas. La idea es conocer prospectivamente el ingreso de los clientes potenciales del mercado para poder ofrecerles algún crédito de consumo: Revolvente y No Revolvente.

Esquema temporal:



Mapping del Modelo

Nivel del modelo (Unidad Muestral)	A nivel de Cliente (ID Cliente)
Variable de Respuesta	Ingreso Real de los clientes en el corte de tiempo "t"
Data de estudio	Clientes bancarizados y con buen riesgo crediticio para el negocio
Tipos de Covariables	Créditos vigentes del cliente Flujo de Saldos del cliente Signos de Riqueza del cliente
Algoritmo	SVM Modelo Lineal Bagging Boosting Tree

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

CASO: PREDICTOR DE INGRESOS

Estructura de la Variables

Por ejemplo, consideremos la siguiente información recopilada de 3 meses para el cliente ID=1:

Enero:

ID	Entidad	Producto	Linea	Deuda	Departamento	Provincia	Distrito	TieneVivienda	TipoVivienda	TieneAuto	MarcaAuto
1	Falabella	Tarjeta de Crédito	6,000	2,500	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	Interbank	Tarjeta de Crédito	4,000	500	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	BCP	Hipotecario		84,000	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi

Febrero:

ID	Entidad	Producto	Linea	Deuda	Departamento	Provincia	Distrito	TieneVivienda	Propiedades 1	TieneAuto	MarcaAuto
1	Falabella	Tarjeta de Crédito	6,000	1,000	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	Interbank	Tarjeta de Crédito	4,000	1,000	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	BCP	Hipotecario		82,000	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi

Marzo:

ID	Entidad	Producto	Linea	Deuda	Departamento	Provincia	Distrito	TieneVivienda	Propiedades 1	TieneAuto	MarcaAuto
1	Falabella	Tarjeta de Crédito	6,000	800	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	Interbank	Tarjeta de Crédito	4,000	200	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi
1	BCP	Hipotecario		80,000	Lima	Lima	Miraflores	Si	Dpto Duplex	Si	Audi

Fuente: RCC **Fuente: RENIEC** **Fuente: SUNARP**

Data estructurada con variables propuestas para el Modelamiento:

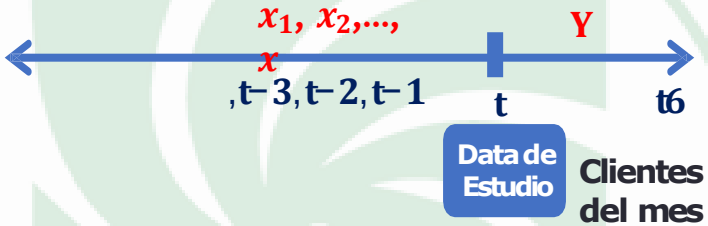
ID	Linea Promedio u1m	Utilizacion de TC u1m	Nro Entidades u1m	Variacion TC u3m	Variacion Otros u3m	CreditoHipo	Residencia Cat	MarcaAuto Cat	NroPropiedades
1	5,000	10%	3	-1,000	-2,000	Si	Top	Alta Gama	2

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

PROPENSION A LA FUGA DE CLIENTES

El negocio necesita contar con un **modelo de Propensión de Fuga de clientes** para su estrategia de retención de clientes del portafolio de TC. Si el modelo logra detectar a los potenciales clientes fuga se puede crear ofertas especiales para ellos y evitar que se desvinculen del negocio.

Esquema temporal:



Clientes de TC al corte del mes "t"

Mapping del Modelo

Nivel del modelo (Unidad Muestral)	A nivel de Producto (ID)
Variable de Respuesta (Y)	El cliente se retiró dentro de los 6 meses de seguimiento: Fuga El cliente está con el negocio más de 6 meses: No Fuga
Data de estudio	Clientes existentes del portafolio TC
	RFM de las Transacciones
Tipos de Covariables	Línea de crédito y/o Utilización Score de Riesgo
Algoritmo	Tree Logit Bagging Boosting SVM Naive Bayes KNN

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

CASO: PROPENSION DE FUGA DE CLIENTES

Estructura de la Variables

Por ejemplo, consideremos la siguiente información recopilada de 6meses de seguimiento para el cliente ID=1:

Enero:

ID	Linea	Saldo	Monto Trxs	Nro Trxs	Score
1	2,000	200	100	5	600

Febrero:

ID	Linea	Deuda	Monto Trxs	Nro Trxs	Score
1	2,000	100	80	2	600

Marzo:

ID	Linea	Deuda	Monto Trxs	Nro Trxs	Score
1	2,000	100	25	1	550

Abril:

ID	Linea	Saldo	Monto Trxs	Nro Trxs	Score
1	2,000	10	0	0	500

Mayo:

ID	Linea	Deuda	Monto Trxs	Nro Trxs	Score
1	2,000	0	0	0	500

Junio:

ID	Linea	Deuda	Monto Trxs	Nro Trxs	Score
1	2,000	0	0	0	350

Linea y Saldo de TC | Transacciones | Score de Riesgo

Data estructurada con variables propuestas para el Modelamiento:

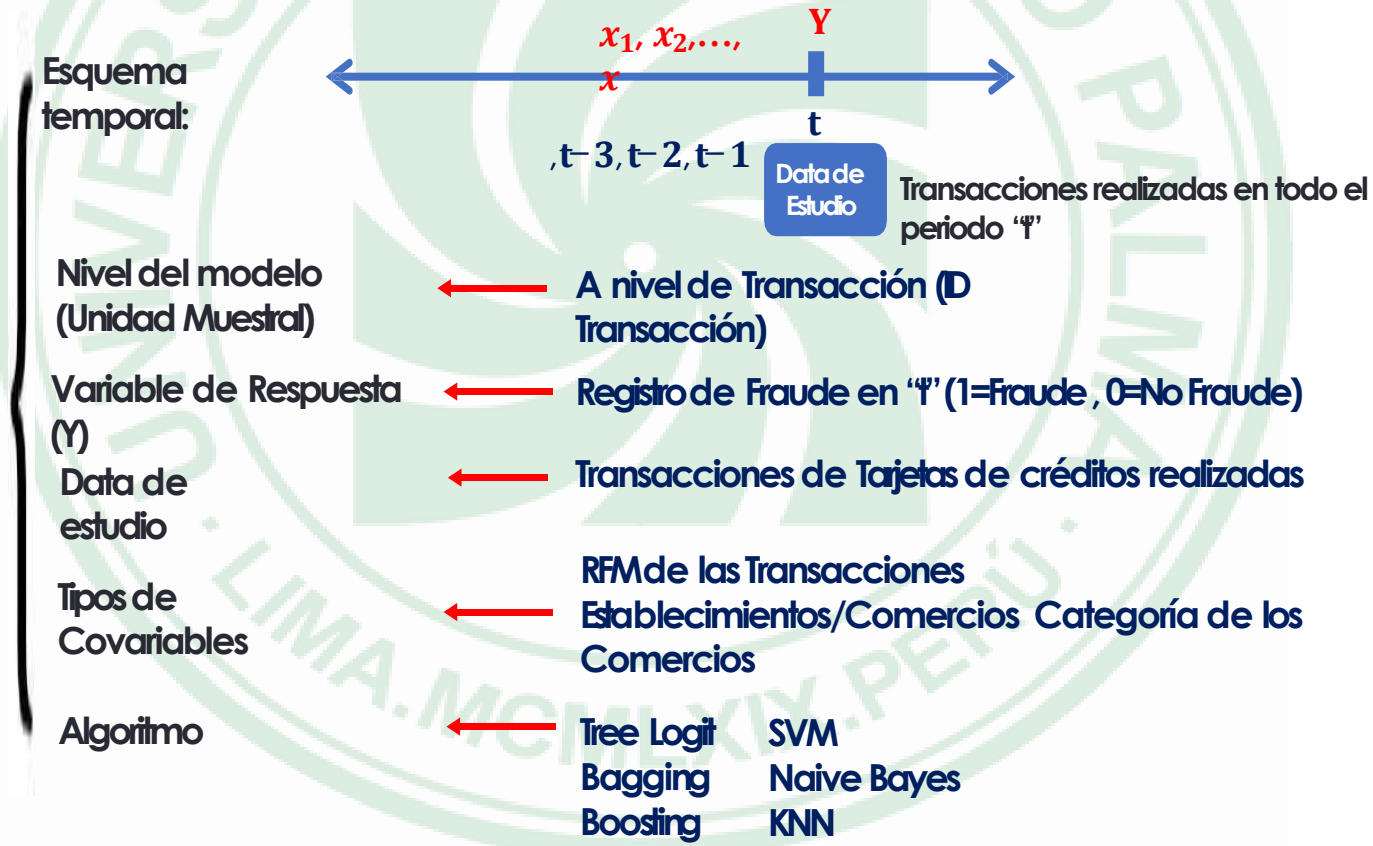
ID	Variación Score	Recencia (Mensual)	Frecuencia (Mensual)	Monto (Mensual)	Utilización 6m	Utilización Actual
1	-41.6%	2	2.6	68.3	10.0%	0.0%

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

CASO: FRAUDE TRANSACCIONAL

El negocio necesita contar con un **modelo de Fraude Transaccional** para fortalecer su estrategia de prevención y detección del fraude por transacciones de las tarjetas de crédito.

Mapping del Modelo



APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

CASO: FRAUDE TRANSACCIONAL

Estructura de la Variables

1. Por ejemplo, consideremos la siguiente información recopilada de 1 semana de transacciones de una tarjeta de crédito "XXXXXXXXXXXXXXXX" de un cliente. ¿La transacción ID_trx=9 es propensa al fraude?

Tiempo	ID_trx	Comercio	Grupo	Monto
Lunes 9 am	1	METRO	Supermercados	60
Lunes 11 am	2	INKAFARMA	Farmacias	40
Martes 3 pm	3	INTERNACIONAL	Clinicas	80
Miercoles 9 am	4	METRO	Supermercados	100
Miercoles 12 pm	5	METRO	Supermercados	30
Miercoles 8 pm	7	WONG	Supermercados	70
Jueves 8 am	8	INTERNACIONAL	Clinicas	120
Sabado 7 pm	9	TOTTUS	Supermercados	1000

2. Reconocimiento de los patrones previos de consumo en 3 niveles:

A nivel de Comercio:

Comercio	Grupo	Recencia (dias)	Frecuencia (dias)	Monto (dias)	Tiempo medio entre compras (dias)
METRO	Supermercados	3.29	1.5	95	1.06
WONG	Supermercados	2.96	1	70	0
INKAFARMA	Farmacias	5.33	1	40	0
INTERNACIONAL	Clinicas	2.45	1	100	1.7

A nivel de Grupo:

Grupo	Recencia (dias)	Frecuencia (dias)	Monto (dias)	Tiempo medio entre compras (dias)
Supermercados	2.96	2	130	0.79
Farmacias	4.17	1	40	0
Clinicas	2.45	1	100	1.7

A nivel de Cliente:

Recencia (dias)	Frecuencia (dias)	Monto (dias)	Tiempo medio entre compras (dias)
2.45	1.75	125	0.49

3. Así tenemos un registro para la data de entrenamiento:

Flag Primera Compra | Recencia | Frecuencia | Monto | Tiempo medio entre compras

ID Trx	Monto Trx	Comercio	Grupo	Cliente	Comercio	Grupo	Cliente	Comercio	Grupo	Cliente	Comercio	Grupo	Cliente	Comercio	Grupo	Cliente
9	1000	Si	No	No	7	2.96	2.45	0	2	1.75	0	130	125		0.79	0.48

APLICACIONES DE ALGORITMOS DE MACHINE LEARNING

Métricas de Validación de los Predictores

Modelos de Clasificación:

ACCURACY

Métrica: *Área bajo la curva COR*

Curva generada por las distribuciones acumuladas de los eventos de éxito y fracaso del modelo.
Cuanto más alejado este la curva de la recta diagonal, el desempeño del modelo es mejor, por ello es de interés conocer el **área bajo la curva (AUC)**.

Umbrales:

- **AUC<60% -> Malo**
- **60%<AUC<70% -> Medio**
- **70%<AUC<80% -> Bueno**
- **80%<AUC<90% -> Muy bueno**
- **AUC>90% -> Sospechoso**

VARIANZA

Métrica: Índice de estabilidad poblacional (PSI)

El objetivo es confirmar que la predicción es estable en el tiempo. Consiste en comparar 2 distribuciones entre si:

- Distribución de la muestra de desarrollo:
  Benchmark (Lo esperado)
- Distribución en el periodo actual.

$$PSI = \sum(DistActual - DistEsperado) * Ln(DistActual / DistEsperado)$$

Umbrales:

- **<10%**: Buena estabilidad,
- **10% - 25%**: Pequeño cambio
- **>25%**: Cambio significante

¡Gracias!

PROGRAMA DE ESPECIALIZACIÓN EN DATA SCIENCE NIVEL II