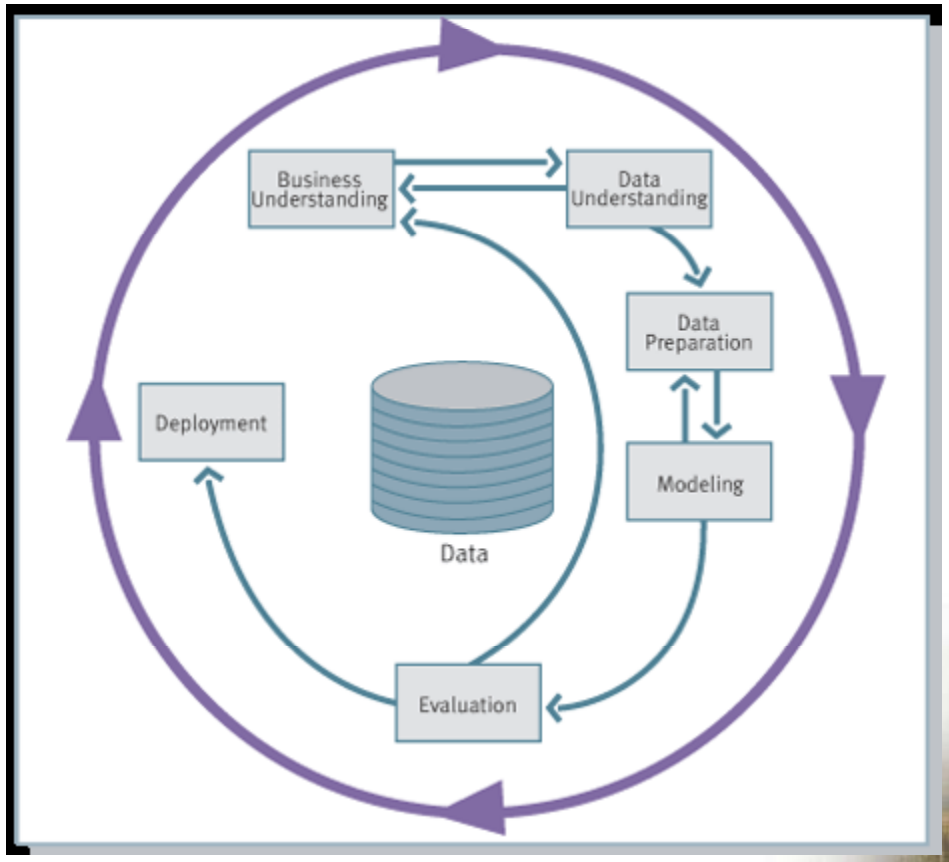
Aplicaciones de Minería de Datos

Ignacio Pérez Ph.D.

Generalidades

- Sectores: productos de consumo masivo, bancos.
- Intereses: Segmentación, pronósticos, scoring.
- Bases de datos: 2-10 GB, (1MM-10MM registros) x (19 – 130 atributos antes de derivados).
- Software: SAS, SPS, R, RapidMiner, Knime.

Metodología CRISP



Metodología CRISP

- Ventajas:
  - Estructurada.
  - Simplifica la relación con el cliente.
  - Facilita la continuidad del proceso.

Dificultades en el desarrollo de los proyectos

- Datos!!!
- Limpieza de las bases de datos (en ocasiones ha consumido el 75% del tiempo del proyecto). Es recomendable que sea el cliente quien “limpie”.
- Uniformidad en la comprensión del problema/datos.

Dificultades en el desarrollo de los proyectos

- Uso de los resultados: hay que involucrar a las áreas interesadas.
- Continuidad de los proyectos.

Caso 1

- Compañía productora de artículos de consumo masivo, distribución a nivel de tienda.
- Preocupación: ¿Quiénes son nuestros clientes?, ¿Quiénes responden a promociones?,

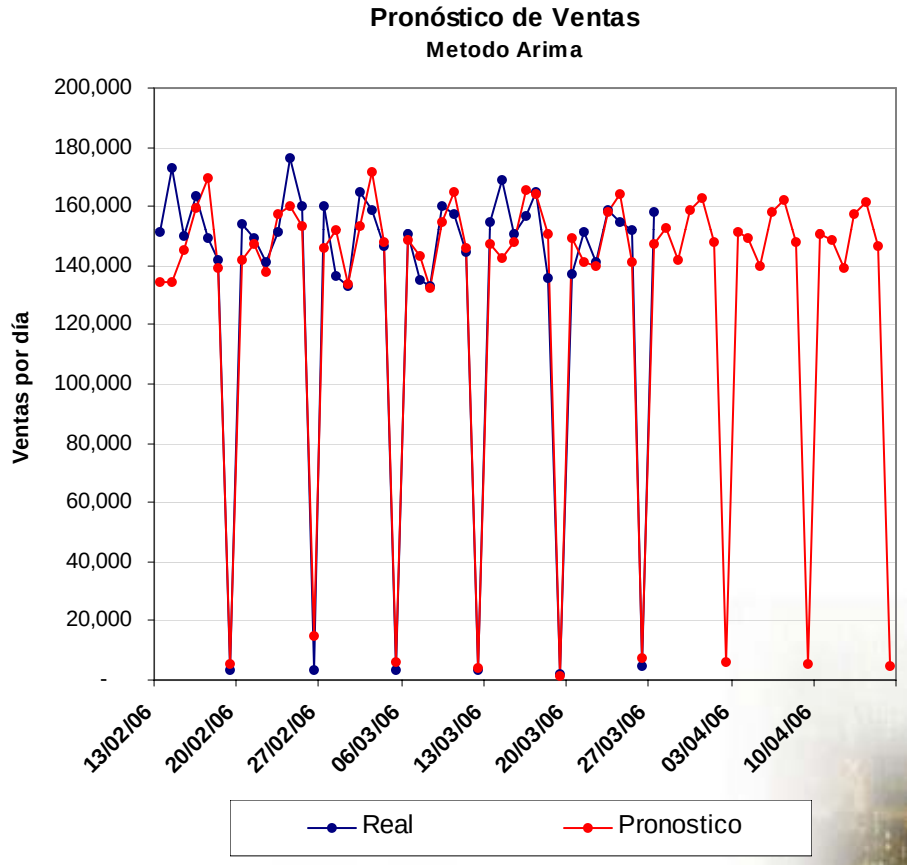
Técnicas

- Estadística Básica
- Analisis de Clusters (K-means).
- Sistemas de Pronósticos (Modelos ARIMA)
- Software: SAS, RapidMiner.

Resultados

- Segmentación de Clientes en tres categorías:
  - Platino
  - Oro
  - Plata.
- Reformulación de los esquemas de promoción.
- Formulación de pronósticos.

Esquema de Pronósticos



Caso 2

- Institución del sector financiero, requiere cumplir con la normativa de la Superfinanciera con relación a la prevención en el Lavado de Activos, Fraude y Terrorismo (SARLAFT).

Caso 2

- Problemática: Detectar las operaciones inusuales susceptibles de corresponder a operaciones de lavado.
- Requerimientos: Segmentación, scoring.

Técnicas

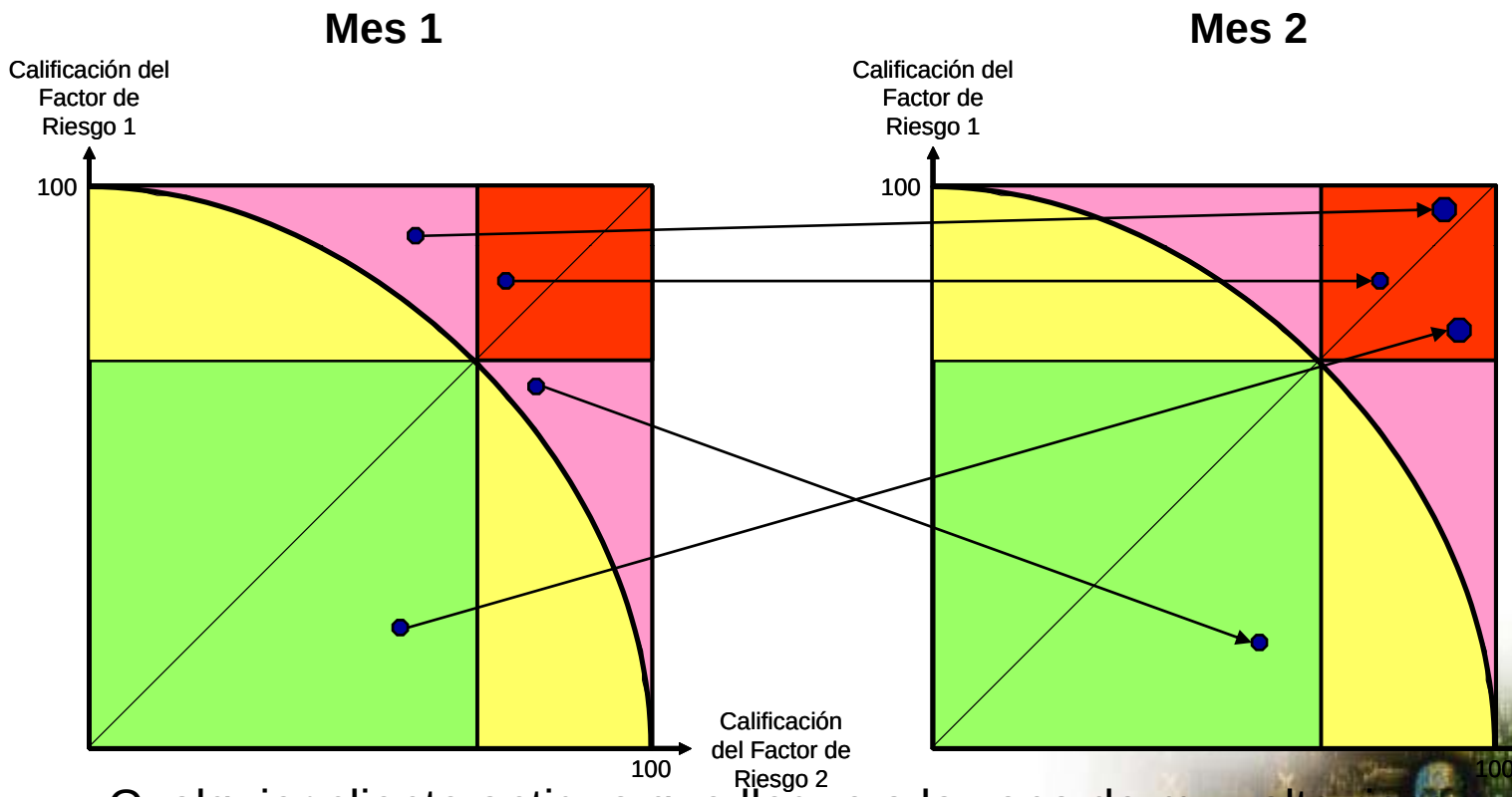
- Segmentación a priori complementada con segmentación a posteriori de los clientes.
- Scoring de clientes.
- Análisis de clusters (K-means)
- Componentes principales.

Resultados

- Construcción de un índice de Inusualidad.

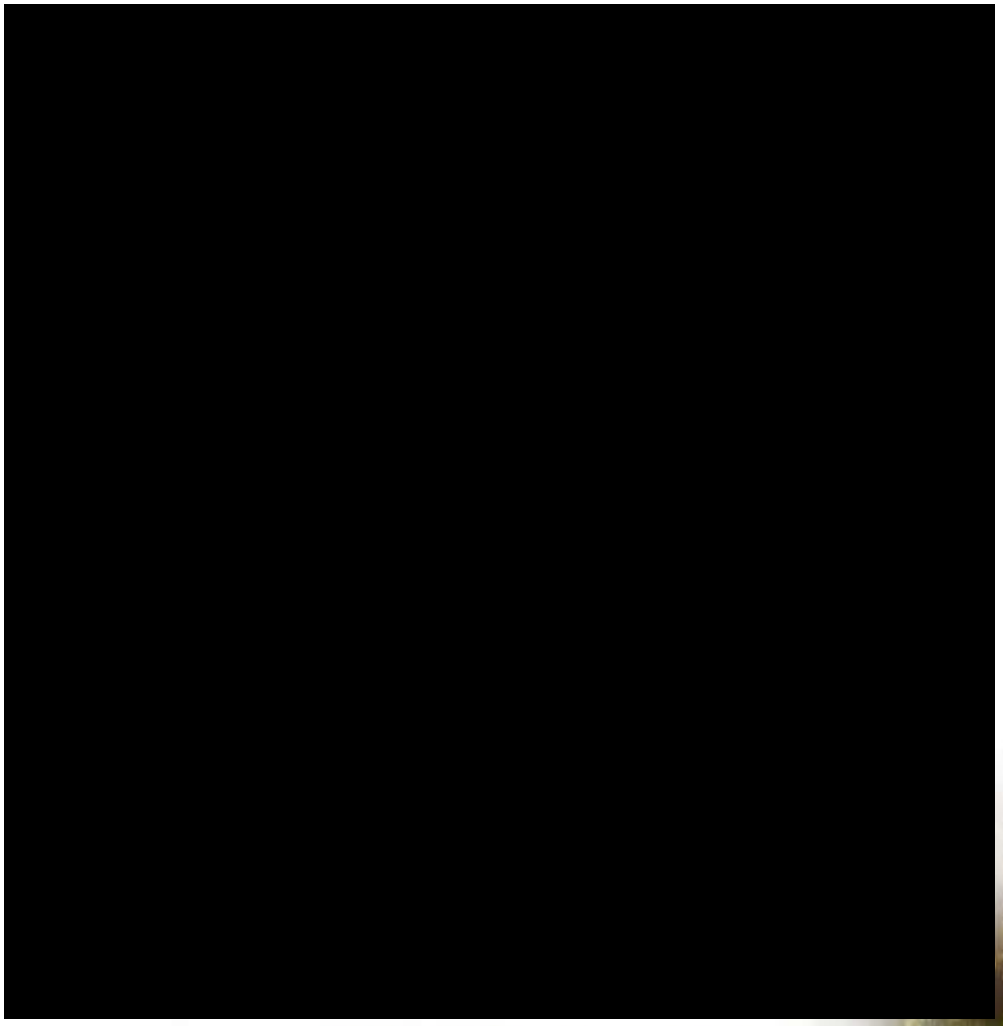
El cambio en el Índice de inusualidad es el que determina la sospecha en un cliente

Clientes Antiguos



Cualquier cliente antiguo que llegue a la zona de muy alto riesgo, viniendo de otra zona, debe ser revisado

Se analizan todas las variables para determinar si es posible detectar comportamientos inusuales



Ejemplo

La calificación de cada cliente se obtiene sumando su calificación en cada factor de riesgo, ponderada por el peso de cada factor

Peso de cada por Factor	
Factor de riesgo 1	28%
Factor de riesgo 2	32%
Factor de riesgo 3	17%
Factor de riesgo 4	23%

	Calificación				
nit	Factor Riesgo 1	Factor Riesgo 2	Factor Riesgo 3	Factor Riesgo 4	Calificación Total
51779706	99	94	93	46	84.12
900013866	61	91	79	39	68.51
1086102088	70	34	94	84	65.80
5825443	80	82	21	50	63.62
13013073	79	65	33	58	61.78
74355054	78	47	61	61	61.33
5825443	53	34	70	95	59.52
890901130	33	55	32	100	55.28
52413064	79	17	12	74	46.55
805017025	36	49	28	64	45.26
20886939	30	35	29	73	41.39
16686658	58	11	44	5	28.32
860025519	11	14	12	27	15.74

A su vez, para calificar cada factor se ponderan los valores de cada indicador o componente principal, por su peso en el respectivo factor

Peso de cada indicador	
Indicador 1	15%
Indicador 2	24%
Indicador 3	21%
Comp Principal 1	17%
Comp Principal 2	23%

	Calificación Factor de Riesgo 4					
nit	Indicador 1	Indicador 2	Indicador 3	Comp Principal 1	Comp Principal 2	Calificación Total
51779706	40	70	60	20	30	46
900013866	50	90	10	30	10	39
1086102088	90	90	70	80	90	84
5825443	30	50	0	100	70	50
13013073	90	20	40	60	90	58
74355054	80	40	90	0	90	61
5825443	100	80	100	100	100	95
890901130	100	100	100	100	100	100
52413064	90	80	40	70	90	74
805017025	50	80	80	40	60	64
20886939	40	100	70	60	80	73
16686658	20	0	0	10	0	5
860025519	20	20	30	20	40	27

Gran parte del esfuerzo del proyecto se concentró en limpieza y análisis de las bases de datos

Pasos desarrollados en el proyecto:

- Obtención de las bases de datos de clientes y transaccionales;
- Limpieza de las bases de datos, quitando campos no requeridos, campos sin información, y casos no relevantes;
- Definición inicial de variables que podrían ser de interés;
- Segmentación a priori de los clientes;
- Construcción de las bases de datos de análisis para calcular las variables definidas como de interés;
- Análisis y calificación de las variables de interés que permiten identificar comportamientos inusuales de LA/FT;
- Definición de las variables que se utilizarán en los modelos;
- Asignar las variables utilizadas a los factores de riesgo;
- Calificar cada factor de riesgo;
- Ponderar los factores para obtener el índice de inusualidad;
- Caracterizar a posteriori el índice de inusualidad para determinar si hay diferencias significativas por factor de riesgo.

Conclusiones

- Amplias posibilidades para el desarrollo de la minería de datos en el país.
- Involucrarse con el cliente.
- Comprender el negocio.
- Continuidad.