Minería de Datos Para Predicción de Riesgo de Compras en Retail

Preparación de los Datos

Equipo de Trabajo:
Juan Martínez
Cinthya Leonor Vergara Silva

Cátedra Introducción a la Minería de datos
Profesor Richard Weber/ Auxiliar Gastón L´Huiller

Tabla de Contenido

Resumen Ejecutivo

Data Mining es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos. Los algoritmos de Data Mining se enmarcan en el proceso completo de extracción de información conocido como KDD [Knowledge Discovery in Databases], que se encarga además de preparación de los datos y de la interpretación de los resultados obtenidos.

El desarrollo de este informe se enmarca en la etapa de selección, limpieza, preprocesamiento y transformación de datos. La preparación de los datos para su posterior uso no es un asunto trivial, ya que del resultado obtenido en esta etapa dependerá en gran medida la capacidad de predicción de modelo. Algunos de los temas desarrollados fueron el tratamiento de datos nulos y valores fuera de rango, recodificación de variables tanto numéricas como categóricas, selección de atributos, etc.

Por último, tras la limpieza, y la transformación de datos descritos en este resumen, se procedió a seleccionar nuevamente variables relevantes a través de utilización de alguna de las estrategias de selección de atributos. Para tales efectos, y con el fin de disminuir la varianza y el potencial ruido que se puede agregar al momento de desarrollar el modelo. En este caso, se estandarizaron las variables entre [0,1], para poder aplicar arboles de decisión o PCA.

Finalmente obtuvimos una base de datos con menor dimensionalidad, con un total de 32 atributos y 27218 registros y sin problemas de valores perdidos.

Introducción

El término Minería de Datos abarca una serie de técnicas utilizadas en una gran variedad de industrias con el objetivo de incrementar la competitividad, las utilidades, participación de mercado así como también mejorar la gestión de clientes, procesos, proveedores, entre otros (Parr Rud, 2001). El término minería de datos se utiliza para describir técnicas de exploración y análisis sobre grandes volúmenes de datos que permiten descubrir patrones significativos y reglas que no podrían ser descubiertas por la simple observación (Berry & Linoff, 2004).

La empresa de retail con ha integrado dentro de sus actividades la venta en línea de sus productos, frente a esta nueva estrategia de plaza se ha encontrado con una mayor cantidad de casos de clientes que no cumplen con sus obligaciones comerciales, y con el creciente aumento de las ventas mediante este medio, sobre todo en los meses de noviembre y diciembre de cada año, la empresa ha aumentado las pérdidas por no pago de productos. Actualmente, utiliza técnicas simples de validación de tarjetas de crédito y direcciones de clientes con un mecanismo de validación principalmente manual que ha generado graves errores y altos costos operacionales.

En este escenario y en vista del creciente volumen de información que generan las transacciones se presenta la necesidad de encontrar una nueva y efectiva forma de encontrar patrones de fraude que permitan determinar el puntaje de riesgo individual de acuerdo al pago de cada cliente. Como solución al problema, se plantea la opción de aplicar técnicas de minería de datos para poder generar un modelo logre mitigar este problema.

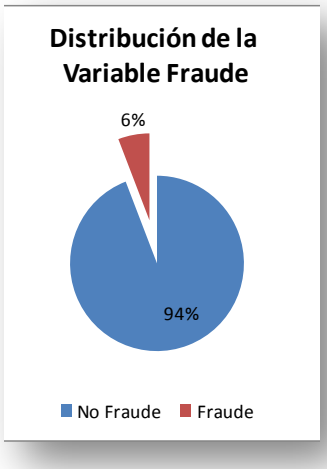
En relación a las técnicas de minería de datos, nos encontramos con una serie de etapas necesarias para poder llevarlas a cabo. El proceso completo para la obtención de conocimiento, en este caso la clasificación de riesgo de los clientes, es llamado KDD por sus siglas en inglés Knowledge Discovery from Database (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) donde se llevan a cabo las etapas de Selección, Preprocesamiento, Transformación, Minería de Datos, Interpretación de los resultados y obtención de conocimiento, etapas que serán llevadas a cabo en este análisis para lograr el objetivo planteado por la empresa.

Para ello, la empresa ha dispuesto una muestra de los pedidos en línea de 30.000 clientes a partir del mes de enero de 2005 que permita obtener una descripción o modelo de clasificación que prediga el riesgo de incumplimiento de pago dada una clasificación de los registros que indica sí cumplió o no el pago de la orden (TARGET_BETRUG).

Con ello, en esta etapa del análisis de clientes y su correspondiente clasificación de riesgo, nos enfocaremos en analizar el problema desde los datos evaluado tanto su estado como capacidad productiva, pasando principalmente por las etapas genéricas de selección, limpieza, preprocesamiento y transformación de manera de obtener una fuente de datos que permita aplicar distintas técnicas de minería de datos para solucionar el problema enfrentado por la empresa.

Descripción General de los Datos

Para comenzar con el estudio encomendado, contamos con una base de datos de clientes con 30.000 registros, cada uno de ellos categorizado de acuerdo a la variable objetivo TARGET_BETRUG que indica si el cliente pagó o no pagó la transacción. De acuerdo a ello, nos encontramos con un 6% de fraudes y un 94% de no fraudes, lo que nos lleva a un problema de clases desbalanceadas (Chawla, Japkowicz, & Ko lcz, 2004).



Para poder explorar los datos y entender de mejor el problema a solucionar se llevó a cabo cambios en el formato de los datos mediantes las herramientas MS Excel y SPSS de acuerdo a la complejidad del cambio. Los cambios realizados para poder realizar este primer análisis exploratorio del estado de los datos fueron los siguientes:

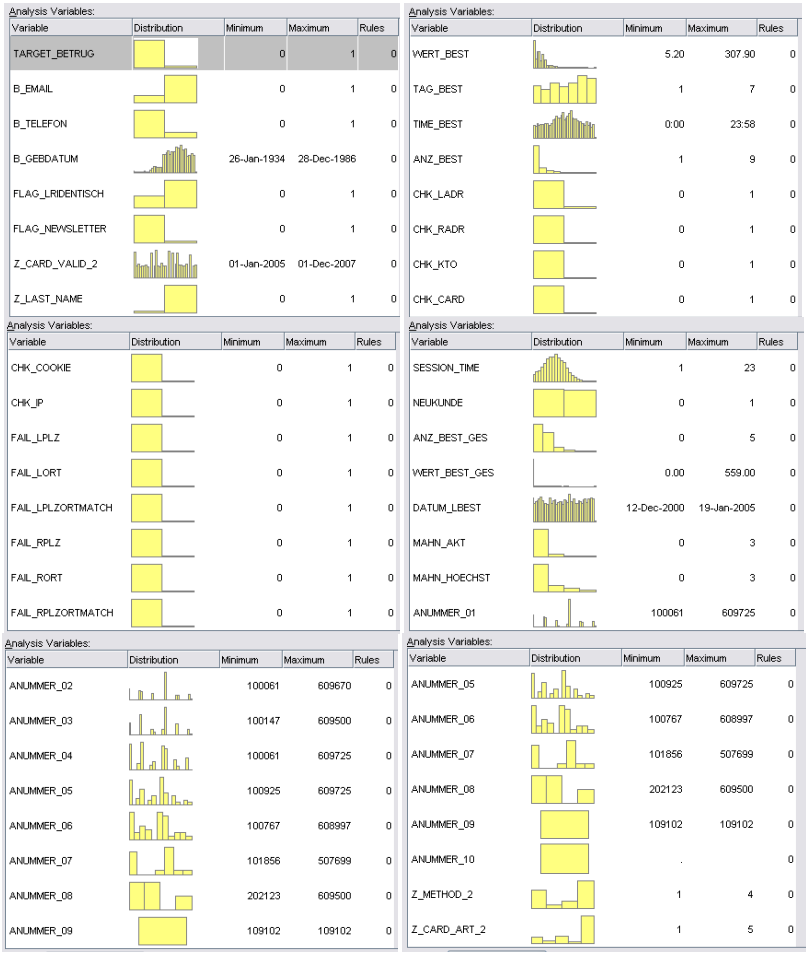
I. Cambio de Variables en Texto a Número: las variables modificadas fueron TARGET_BETRUG, B_EMAIL, B_TELEFON, FLAG_LRIDENTISCH, FLAG_NEWSLETTER, Z_LAST_NAME, TAG_BEST, CHK_LADR, CHK_RADR, CHK_KTO, CHK_CARD, CHK_COOKIE, CHK_IP, FAIL_LPLZ, FAIL_LORT, FAIL_LPLZORTMATCH, FAIL_RPLZ, FAIL_RORT, FAIL_RPLZORTMATCH, NEUKUNDE, Z_METHOD, Z_CARD_ART.

II. Estandarización del formato de fecha: en los campos B_GEBDATUM, DATUM_LBEST sólo se estandarizó el formato de fecha a “día-mes-año” y en la variable Z_CARD_VALID se generó una fecha válida ya que presentaba un formato “mesaño”(por ejemplo 52005) que fue modificado a “01-mes-año” para poder ser analizado.

De esta manera, los cambios fueron básicamente cambiar de texto a número, corregir el formato de las fechas y de las monedas. De acuerdo estas modificaciones iniciales los datos se pueden resumir como muestra la siguiente tabla:

Variable	Tipo	Tendencia Central	Frecuencias/Rangos	Missing	% Miss
TARGET_BETRUG	nominal	mode = no (28254), least = yes (1746)	no (28254), yes (1746)	0	0,0%
B_EMAIL	nominal	mode = 1.0 (23963), least = 0.0 (6037)	1.0 (23963), 0.0 (6037)	0	0,0%
B_TELEFON	nominal	mode = 0.0 (25493), least = 1.0 (4507)	0.0 (25493), 1.0 (4507)	0	0,0%
B_GEBDATUM	date	length = 19336 days	[20-01-1934 ; 29-12-1986]	2.942	9,8%
FLAG_LRIDENTISCH	nominal	mode = 1.0 (21157), least = 0.0 (8843)	1.0 (21157), 0.0 (8843)	0	0,0%
FLAG_NEWSLETTER	nominal	mode = 0.0 (28772), least = 1.0 (1228)	1.0 (1228), 0.0 (28772)	0	0,0%
Z_CARD_VALID_2	date	length = 1064 days	[31-12-2004 ; 30-11-2007]	0	0,0%
Z_LAST_NAME	nominal	mode = 1.0 (14437), least = 0.0 (755)	1.0 (14437), 0.0 (755)	14.808	49,4%
WERT_BEST	numeric	avg = 43.968 +/- 35.709	[5.200 ; 361.200]	0	0,0%
TAG_BEST	nominal	mode = Samstag (5958), least = Dienstag (2976)	Montag (3571), Dienstag (2976), Mittwoch (4296), Donnerstag (3618), Freitag (4224), Samstag (5958), Sonntag (5357)	0	0,0%
TIME_BEST	time	length = 3394272 hours	[08:03:00; 09:00:00]	20	0,1%
ANZ_BEST	numeric	avg = 1.442 +/- 0.921	[1.000 ; 9.000]	0	0,0%
CHK_LADR	nominal	mode = 0.0 (28865), least = 1.0 (1135)	0.0 (28865), 1.0 (1135)	0	0,0%
CHK_RADR	nominal	mode = 0.0 (29889), least = 1.0 (111)	0.0 (29889), 1.0 (111)	0	0,0%
CHK_KTO	nominal	mode = 0.0 (29912), least = 1.0 (88)	0.0 (29912), 1.0 (88)	0	0,0%
CHK_CARD	nominal	mode = 0.0 (29893), least = 1.0 (107)	0.0 (29893), 1.0 (107)	0	0,0%
CHK_COOKIE	nominal	mode = 0.0 (29905), least = 1.0 (95)	0.0 (29905), 1.0 (95)	0	0,0%
CHK_IP	nominal	mode = 0.0 (29879), least = 1.0 (121)	0.0 (29879), 1.0 (121)	0	0,0%
FAIL_LPLZ	nominal	mode = 0.0 (29835), least = 1.0 (165)	0.0 (29835), 1.0 (165)	0	0,0%
FAIL_LORT	nominal	mode = 0.0 (29849), least = 1.0 (151)	0.0 (29849), 1.0 (151)	0	0,0%
FAIL_LPLZORTMATCH	nominal	mode = 0.0 (29772), least = 1.0 (228)	0.0 (29772), 1.0 (228)	0	0,0%
FAIL_RPLZ	nominal	mode = 0.0 (29674), least = 1.0 (326)	0.0 (29674), 1.0 (326)	0	0,0%
FAIL_RORT	nominal	mode = 0.0 (29693), least = 1.0 (307)	0.0 (29693), 1.0 (307)	0	0,0%
FAIL_RPLZORTMATCH	nominal	mode = 0.0 (29637), least = 1.0 (363)	0.0 (29637), 1.0 (363)	0	0,0%
SESSION_TIME	numeric	avg = 8.578 +/- 3.863	[1.000 ; 24.000]	0	0,0%
NEUKUNDE	nominal	mode = 1.0 (15032), least = 0.0 (14968)	1.0 (15032), 0.0 (14968)	0	0,0%
ANZ_BEST_GES	numeric	avg = 0.607 +/- 0.766	[0.000 ; 6.000]	0	0,0%
WERT_BEST_GES	numeric	avg = 1.234 +/- 13.252	[0.000 ; 559.000]	13.849	46,2%
DATUM_LBEST	date	length = 1499 days	[11-12-2000 ; 18-01-2005]	15.856	52,9%
MAHN_AKT	nominal	mode = 0.0 (12703), least = 2.0 (126)	0.0 (12703), 1.0 (1182), 3.0 (133), 2.0 (126)	15.856	52,9%
MAHN_HOECHST	nominal	mode = 0.0 (10200), least = 3.0 (409)	0.0 (10200), 1.0 (2282), 2.0 (1253), 3.0 (409)	15.856	52,9%
ANUMMER_01	nominal	mode = 402845.0 (80), least = 408801.0 (33)	-	0	0,0%
ANUMMER_02	nominal	mode = 409662.0 (25), least = 200035.0 (4)	-	22.147	73,8%
ANUMMER_03	nominal	mode = 200578.0 (14), least = 503397.0 (1)	-	26.802	89,3%
ANUMMER_04	nominal	mode = 209725.0 (8), least = 201796.0 (1)	-	28.668	95,6%
ANUMMER_05	nominal	mode = 201373.0 (7), least = 308997.0 (1)	-	29.459	98,2%
ANUMMER_06	nominal	mode = 109102.0 (3), least = 509500.0 (1)	-	29.794	99,3%
ANUMMER_07	nominal	mode = 106520.0 (2), least = 405284.0 (1)	-	29.905	99,7%
ANUMMER_08	nominal	mode = 609500.0 (2), least = 404134.0 (1)	-	29.966	99,9%
ANUMMER_09	nominal	mode = 109102.0 (1), least = 109102.0 (1)	-	29.993	100,0%
ANUMMER_10	nominal	mode = unknown	-	30.000	100,0%
Z_METHOD_2	nominal	mode = Rechnung (14808), least = Kundenkarte (1550)	Kreditkarte (9796), Kundenkarte (1550), Lastschrift (3846), Rechnung (14808)	0	0,0%
Z_CARD_ART_2	nominal	mode = Eurocard (5096), least = Amex (773)	Visa (3927), Kundenkarte (1550), Eurocard (5096), Amex (773)	18.654	62,2%

Para cada variable además vemos el siguiente comportamiento reflejado en los histogramas.



Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rules
TARGET_BETRUG		0	1	0
B_EMAIL		0	1	0
B_TELEFON		0	1	0
B_GEBDATUM		26-Jan-1934	28-Dec-1986	0
FLAG_LRIDENTISCH		0	1	0
FLAG_NEWSLETTER		0	1	0
Z_CARD_VALID_2		01-Jan-2005	01-Dec-2007	0
Z_LAST_NAME		0	1	0

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rules
WERT_BEST		5.20	307.90	0
TAG_BEST		1	7	0
TIME_BEST		0:00	23:58	0
ANZ_BEST		1	9	0
CHK_LADR		0	1	0
CHK_RADR		0	1	0
CHK_KTO		0	1	0
CHK_CARD		0	1	0

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rules
CHK_COOKIE		0	1	0
CHK_IP		0	1	0
FAIL_LPLZ		0	1	0
FAIL_LORT		0	1	0
FAIL_LPLZORTMATCH		0	1	0
FAIL_RPLZ		0	1	0
FAIL_RORT		0	1	0
FAIL_RPLZORTMATCH		0	1	0

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rules
SESSION_TIME		1	23	0
NEUKUNDE		0	1	0
ANZ_BEST_GES		0	5	0
WERT_BEST_GES		0.00	559.00	0
DATUM_LBEST		12-Dec-2000	19-Jan-2005	0
MAHN_AKT		0	3	0
MAHN_HOECHST		0	3	0
ANUMMER_01		100061	609725	0

Analysis Variables:

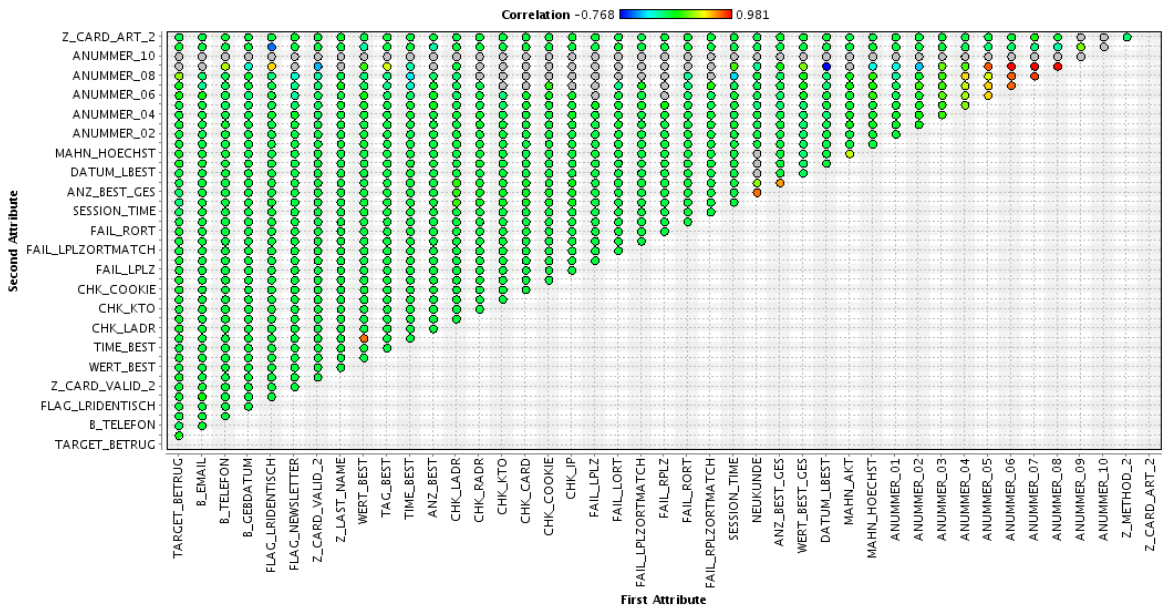
Variable	Distribution	Minimum	Maximum	Rules
ANUMMER_02		100061	609670	0
ANUMMER_03		100147	609500	0
ANUMMER_04		100061	609725	0
ANUMMER_05		100925	609725	0
ANUMMER_06		100767	608997	0
ANUMMER_07		101856	507699	0
ANUMMER_08		202123	609500	0
ANUMMER_09		109102	109102	0

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rules
ANUMMER_05		100925	609725	0
ANUMMER_06		100767	608997	0
ANUMMER_07		101856	507699	0
ANUMMER_08		202123	609500	0
ANUMMER_09		109102	109102	0
ANUMMER_10		.		0
Z_METHOD_2		1	4	0
Z_CARD_ART_2		1	5	0

Es importante recalcar que las en las variables ANUMMER_02, ANUMMER_03, ANUMMER_04, ANUMMER_05, ANUMMER_06, ANUMMER_07, ANUMMER_08, ANUMMER_09, ANUMMER_10 y Z_CARD_ART se identifican valores perdidos, pero en realidad es por la naturaleza de la variable que se encuentran vacíos, ya que se rellenan en caso que corresponda pero no constituye a un valor perdido propiamente tal.

Con estas modificaciones además revisamos la correlación de las variables para poder obtener información adicional acerca de la relación entre las variables.



De acuerdo a ello, podemos ver en este análisis exploratorio del estado de la data que la base de datos cuenta con 43 variables con una correlación similar entre cada una de ellas, de las cuales una corresponde a la variable objetivo que identifica si un cliente cometió o no fraude en su compra. En las variables restantes nos encontramos con 33 variables nominales, 3 de fecha, 1 de tiempo y sólo 5 variables continúas escalares.

Modificaciones Adicionales para iniciar el Análisis

En un segundo análisis y debido a la estructura de los datos, se llevó a cabo un análisis de los productos comprados dado que se detallaban en 10 columnas. Por ello se genero una tabla auxiliar que une a todos los productos en una sola columna identificados por el cliente y la clase. De acuerdo a ello la distribución de frecuencias indica que existen 560 productos distintos, el promedio de compra de cada producto es 77 unidades con una desviación estándar de 9,23. Si seleccionamos los productos más vendidos nos encontramos con que en la clase No Fraude 10 productos superan las 92 unidades mientras que, en la clase Fraude 9 productos superan las 10 unidades. Los productos más vendidos por clase son:

Rótulos de fila	Fraude	No Fraude	Total general
402062	11		11
401402	11		11
504077	11		11
406310	11		11
400061	12		12
403950	13		13
407703	13		13
409513	14		14
402845	17		17
406685		92	92
401241		93	93
405965		93	93
209725		93	93
107359		93	93
603639		94	94
206590		95	95
403533		96	96
201216		98	98
504792		101	101
Total general	113	948	1061

El atributo ANZ_BEST indica la cantidad de productos por orden de compra presentando una cantidad de máxima de 9 productos por orden. De acuerdo a ello se decidió mantener este atributo y trabajar sólo con los productos más vendidos por cada clase, ya que en su total no muestra mayor importancia respecto a la variable objetivo pero si posee cierta relación, las cual fue calcula por los test de Bartlett y Chi.-Cuadrado. El procedimiento realizado quitó las 10 columnas ANUMMER y agregó 10 columnas correspondiente a los productos más comprados (5 de la clase Fraude y 5 de la clase No Fraude) indicando con un 1 si está en la compra y con un 0 sino.

Correlation Matrix

		VAR00002	VAR00003
Correlation	VAR00002	1,000	,012
	VAR00003	,012	1,000
Sig. (1-tailed)	VAR00002		,008
	VAR00003	,008	

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,500
Bartlett's Test of Sphericity	Approx. Chi-Square	5,824
	df	1
	Sig.	,016

Test Statistics

	VAR00002	VAR00003
Chi-Square	616,017[a]	33480,414[b]
df	559	1
Asymp. Sig.	,047	,000

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 77,3.

b. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 21633,0.

Del mismo modo, para evitar el problema de los registros en blanco, se transformó la variable Z_CARD_ART a variables dummy generando 4 nuevas columnas (VISA, KUNDENKARTE, EUROCARD, AMEX).

A su vez y debido a que el formato de fecha no es fácil de manejar para evaluar tendencias, o relaciones por medio de software estadístico, una vez analizada la estructura inicial fue necesario transformar los campos de fecha para poder iniciar el análisis. Las transformaciones realizadas fueron las siguientes:

B_GEBDATUM: se realizó el cambio de fecha de nacimiento a edad considerando que el análisis fue en el año 2005.

Z_CARD_VALID: se separó en mes y año agregando dos campos más ZCV_MONAT que almacena el mes y ZVC_JAHR que almacena el año ambos en formato de número.

DATUM_LBEST: se separó en mes y año descartando el día ya que en la columna TAG_BEST se almacena el día de la semana de la compra, con lo que se agregaron las columnas DATUM_MONAT y DATUM_JAHR que almacenan el mes y el año respectivamente en formato de número.

Adicionalmente se creó un campo que calcula la diferencia de tiempo en años entre la última compra (DATUM_LBEST) y la fecha de vencimiento de la tarjeta (Z_CARD_VALID) llamado DIF_VKAUFEN.

Finalmente, los datos utilizados para el análisis quedaron de la siguiente manera:

Variable	Tipo	Tendencia Central	Frecuencias/Rangos	Missing	% Miss
BESTELLIDENT	integer	avg = 15000.500 +/- 8660.398	[1.000 ; 30000.000]	0	0,0%
Z_CARD_ART = VISA	numeric	avg = 0.131 +/- 0.337	[0.000 ; 1.000]	0	0,0%
Z_CARD_ART = EUROCARD	numeric	avg = 0.170 +/- 0.376	[0.000 ; 1.000]	0	0,0%
Z_CARD_ART = KUNDENKARTER	numeric	avg = 0.052 +/- 0.221	[0.000 ; 1.000]	0	0,0%
Z_CARD_ART = AMEX	numeric	avg = 0.026 +/- 0.158	[0.000 ; 1.000]	0	0,0%
TARGET_BETRUG	integer	avg = 0.058 +/- 0.234	[0.000 ; 1.000]	0	0,0%
B_EMAIL	integer	avg = 0.799 +/- 0.401	[0.000 ; 1.000]	0	0,0%
B_TELEFON	integer	avg = 0.150 +/- 0.357	[0.000 ; 1.000]	0	0,0%
ALTER	integer	avg = 33.444 +/- 9.643	[18.000 ; 70.000]	2.942	9,8%
FLAG_LRIDENTISCH	integer	avg = 0.705 +/- 0.456	[0.000 ; 1.000]	0	0,0%
FLAG_NEWSLETTER	integer	avg = 0.041 +/- 0.198	[0.000 ; 1.000]	0	0,0%
Z_METHODE	integer	avg = 2.789 +/- 1.344	[1.000 ; 4.000]	0	0,0%
ZCV_MONTAG	integer	avg = 6.480 +/- 3.468	[1.000 ; 12.000]	0	0,0%
ZCV_JAHR	integer	avg = 2005.999 +/- 0.818	[2005.000 ; 2007.000]	0	0,0%
DIF_VKAUFEN	real	avg = 3.459 +/- 1.473	[-0.040 ; 6.970]	15.856	52,9%
Z_LAST_NAME	integer	avg = 0.950 +/- 0.217	[0.000 ; 1.000]	14.808	49,4%
WERT_BEST	real	avg = 43.968 +/- 35.709	[5.200 ; 361.200]	0	0,0%
TAG_BEST	integer	avg = 4.375 +/- 2.003	[1.000 ; 7.000]	0	0,0%
TIME_BEST	real	avg = 12.335 +/- 6.258	[0.000 ; 23.900]	20	0,1%
ANZ_BEST	integer	avg = 1.442 +/- 0.921	[1.000 ; 9.000]	0	0,0%
CHK_LADR	integer	avg = 0.038 +/- 0.191	[0.000 ; 1.000]	0	0,0%
CHK_RADR	integer	avg = 0.004 +/- 0.061	[0.000 ; 1.000]	0	0,0%
CHK_KTO	integer	avg = 0.003 +/- 0.054	[0.000 ; 1.000]	0	0,0%
CHK_CARD	integer	avg = 0.004 +/- 0.060	[0.000 ; 1.000]	0	0,0%
CHK_COOKIE	integer	avg = 0.003 +/- 0.056	[0.000 ; 1.000]	0	0,0%
CHK_IP	integer	avg = 0.004 +/- 0.063	[0.000 ; 1.000]	0	0,0%
FAIL_LPLZ	integer	avg = 0.006 +/- 0.074	[0.000 ; 1.000]	0	0,0%
FAIL_LORT	integer	avg = 0.005 +/- 0.071	[0.000 ; 1.000]	0	0,0%
FAIL_LPLZORTMATCH	integer	avg = 0.008 +/- 0.087	[0.000 ; 1.000]	0	0,0%
FAIL_RPLZ	integer	avg = 0.011 +/- 0.104	[0.000 ; 1.000]	0	0,0%
FAIL_RORT	integer	avg = 0.010 +/- 0.101	[0.000 ; 1.000]	0	0,0%
FAIL_RPLZORTMATCH	integer	avg = 0.012 +/- 0.109	[0.000 ; 1.000]	0	0,0%
SESSION_TIME	integer	avg = 8.578 +/- 3.863	[1.000 ; 24.000]	0	0,0%
NEUKUNDE	integer	avg = 0.501 +/- 0.500	[0.000 ; 1.000]	0	0,0%
ANZ_BEST_GES	integer	avg = 0.607 +/- 0.766	[0.000 ; 6.000]	0	0,0%
WERT_BEST_GES	numeric	avg = 63.367 +/- 69.728	[5.200 ; 1047.800]	15.856	52,9%
DATUM_MONTAG	integer	avg = 6.475 +/- 3.496	[1.000 ; 12.000]	15.856	52,9%
DATUM_JAHR	integer	avg = 2002.501 +/- 1.179	[2000.000 ; 2005.000]	15.856	52,9%
MAHN_AKT	integer	avg = 0.130 +/- 0.432	[0.000 ; 3.000]	15.856	52,9%
MAHN_HOECHST	integer	avg = 0.425 +/- 0.771	[0.000 ; 3.000]	15.856	52,9%
PRODUCTO_402845	integer	avg = 0.004 +/- 0.059	[0.000 ; 1.000]	0	0,0%
PRODUCTO_409513	integer	avg = 0.002 +/- 0.048	[0.000 ; 1.000]	0	0,0%
PRODUCTO_407703	integer	avg = 0.003 +/- 0.054	[0.000 ; 1.000]	0	0,0%
PRODUCTO_403950	integer	avg = 0.002 +/- 0.046	[0.000 ; 1.000]	0	0,0%
PRODUCTO_400061	integer	avg = 0.003 +/- 0.054	[0.000 ; 1.000]	0	0,0%
PRODUCTO_504792	integer	avg = 0.003 +/- 0.059	[0.000 ; 1.000]	0	0,0%
PRODUCTO_201216	integer	avg = 0.003 +/- 0.058	[0.000 ; 1.000]	0	0,0%
PRODUCTO_403533	integer	avg = 0.003 +/- 0.058	[0.000 ; 1.000]	0	0,0%
PRODUCTO_206590	integer	avg = 0.003 +/- 0.057	[0.000 ; 1.000]	0	0,0%
PRODUCTO_603639	integer	avg = 0.003 +/- 0.057	[0.000 ; 1.000]	0	0,0%

Desarrollo del Análisis

Pregunta I: Datos *Nulos e Imputación de Datos*

Los valores nulos son un problema no menor en cualquier análisis de bases de datos, las razones por la cual aparecen son de distinta índole, desde errores en procedimientos manuales de ingreso de datos, errores en los equipos, mediciones incorrectas hasta censura de los datos (por ejemplo en una entrevista médica) (Allison, 2001)

La existencia de una cantidad considerable de valores nulos en una variable dificulta el análisis de los datos, ya que usualmente no permite la aplicación de las técnicas existentes que posibilitan la extracción de conocimiento. Existen tres tipos de problemas asociados con los valores nulos (Barnard, 1999):

- Pérdida de información y eficiencia en el análisis;
- Complicación en el uso de los datos, debido a inaplicabilidad de las herramientas o software estándar; y
- La existencia potencial de sesgo en el análisis atribuible a las diferencias sistemáticas entre los datos observados y los datos perdidos

En la práctica, la eficacia de las técnicas de tratamiento de valores nulos está directamente relacionada con la razón por la cual tuvo su origen el valor perdido. Si tenemos alguna información acerca de ella, es posible que encontremos una regla para completar estos valores, por el contrario, si no tenemos dicha información, es necesario aplicar técnicas de evaluación de los valores perdidos que encuentren algún patrón que permita ya sea completarlos o descartarlos (en el caso que no afecten el análisis), decisión que depende en gran medida del tipo del valor perdido y la importancia del registro en la base de datos (Allison, 2001).

Categorías de datos perdidos

Los datos faltantes o perdidos pueden ser categorizados en tres tipos (Meng XL, 1999):

- **Missing Completely at random (MCAR):** corresponden a variables con datos perdidos que no poseen relación alguna a los valores de otros registros dentro de la misma variable ni a los valores de otras variables. Cuando ocurre esto, las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas. A modo de ejemplo, asumamos la siguiente tabla con datos faltantes.
- **Missing at Random (MAR):** corresponde a variables con datos perdidos que tiene alguna relación con otras variables, es decir, que si tenemos una variable Y con valores faltantes, y otra variable X, se dice que los datos son MAR si $P(Y = null|Y, X) = P(Y = null|X)$. Dicho de otra forma, el dato perdido puede predecirse a partir de otros datos (existentes) para el registro.
- **Not Missing at random (NMAR):** corresponde a variables con datos perdidos donde el mismo dato perdido determina en si mismo por qué es desconocido.

Método para hacer frente a los datos perdidos

Existe una serie de métodos para enfrentar el problema de los datos perdidos, entre ellos nos encontramos con (Farhangfara, Kurganb, & Dy, 2008):

- Descartar los registros con datos faltantes: Este método es práctico sólo cuando los datos contienen una relativamente baja cantidad de registros con datos perdidos y cuando el análisis de todos los datos no produce un sesgo importante por no utilizar los registros con datos perdidos.;
- Reemplazar los datos faltantes con otro valor: Este método tiende a producir serios problemas de inferencia. No se entrará en detalle en este método; y
- Imputar los datos faltantes: Este método es aplicable cuando la cantidad de atributos con datos faltantes es relativamente pequeño en relación al número de registros que presentan dicha condición.

Métodos de Descarte o Eliminación

En estos métodos existen dos grandes grupos, Listwise Deletion y Pairwise Deletion, en el primero se conservan sólo aquellos registros que están completos (es decir se elimina cualquier registro que posea a lo menos una variable perdida); en el segundo grupo, se conservan aquellos registros completos y aquellos que sólo tienen datos perdidos que no son variables necesarias para el análisis.

Estos métodos son razonables debido a su simplicidad, sin embargo no se recomienda su aplicación en los siguientes casos:

i. Si el análisis no puede tolerar datos faltantes (por ejemplo en el caso de Pairwise Deletion, las variables con datos nulos no permiten aplicar técnicas como regresión);
ii. Si la fracción de registros a eliminar (en Listwise o PairwiseDeletion) es excesivamente grande respecto al total de datos; y
iii. Si los datos nulos no son ignorables (por ejemplo si aportan información al análisis requerido).

Métodos de Imputación

Existen dos grandes tipos de técnicas que pueden ser agrupadas en dos grupos, Single Imputation y Multiple Imputation.

Single Imputation

Single Imputation [(Farhangfara, Kurganb, & Dy, 2008) ; (Pearson, Mining Imperfect Data: Dealing with Contamination and Incomplete Records, 2005)] es quizás el enfoque más utilizado en la práctica. En este método se estima el dato faltante usando otros datos relacionados que estén disponibles, algunos de los métodos más utilizados son:

Reemplazar los datos faltantes a través de la imputación por el promedio (en el cual se reemplaza el valor faltante de acuerdo al valor promedio de un grupo apropiadamente definido de valores disponibles). Single Imputation a través de la imputación por el promedio posee tres limitaciones potenciales:

i. Uso de la imputación por el promedio generalmente disminuye la variabilidad inherente al conjunto original de datos, particularmente en el caso que el mismo valor promedio sea utilizado para reemplazar varios datos faltantes;
ii. Puede ser implementado de diferentes formas a través de obtener el promedio de diferentes conjuntos de datos existentes candidatos, y lo razonable del resultado depende de esta elección, es decir, es dependiente de las elecciones de grupos de datos realizada;
iii. Si existen outliers entre los conjuntos de datos candidatos para obtener el promedio que se empleará para realizar la imputación, este valor puede generar un importante sesgo en el resultado.

Otra forma de reemplazo se conoce por Hot Deck Imputation, en este caso cada registro que contiene datos perdidos se busca el registro más parecido que no tenga datos perdidos y de esta forma, el dato perdido se imputa con el valor del dato existente en el registro más parecido.

También se utiliza la Imputación por Regresión, que utiliza modelos de regresión que utilizan datos de otras variables para predecir las observaciones faltantes.

Los mecanismos anteriormente mencionados caen dentro del grupo de los determinísticos, sin embargo existe un grupo de análisis estocástico dentro del cual se encuentra el método de Naive-Bayes, que es una técnica de clasificación que asume que los valores de las variables son condicionalmente independientes dentro de una clase, es decir: $P(v_1, v_2, \dots, v_d|C) = \prod_{i=1}^{d} P(v_i, C)$ donde v_i es la i-ésima variable o atributo, C representa la clase y *d* es el número de variables.

Naive-Bayes a menudo funciona bien, sin embargo se debe tener en consideración lo siguiente:

i. Opera sobre atributos categóricos
ii. No es recomendable aplicarlo si el número de instancias para un atributo dado (o valor de este) es muy pequeño, puesto que esto conlleva a que muchas de las probabilidades calculadas sean cero.

Multiple Imputation

En Multiple Imputation [(Farhangfara, Kurganb, & Dy, 2008) ; (Pearson, Mining Imperfect Data: Dealing with Contamination and Incomplete Records, 2005); (Wayman, 2003)], los valores perdidos de cualquier variable son estimados usando los valores existentes en otras variable. Los valores estimados (o imputados) sustituyen a los valores faltantes con lo cual se obtiene un conjunto de datos completo denominado "conjunto de datos imputados". Este proceso es realizado varias veces, produciendo varios conjuntos de datos imputados (de aquí el nombre de Multiple Imputation). Se realizan análisis estadísticos sobre cada uno de los conjuntos de datos imputados, obteniéndose múltiples resultados. Estos resultados posteriormente son combinados para producir un análisis final.

Multiple Imputation permite restaurar la variabilidad natural de los datos perdidos (a través de la imputación de valores que se basan en variables correlacionadas con los datos perdidos); e incorporar la incertidumbre atribuida a la estimación de dichos datos (esto producto de la creación de diferentes versiones de datos perdidos y observados). Es importante mencionar que los valores imputados deben mantener la variabilidad de la población y preservar las relaciones con otras variables, es decir, con la imputación múltiple se desea preservar importantes características del conjunto de datos como un todo (por ejemplo, medias, varianzas, parámetros de regresión, entre otros). La imputación múltiple es una solución atractiva para hacer frente al problema de los datos perdidos debido a que representa un buen balance entre calidad del resultado y facilidad de uso.

El proceso de Imputación Múltiple consta de las siguientes etapas:

i. Creación de los conjuntos de datos imputados, algunas herramientas que pueden ser utilizadas son: Regresión Lineal, Regresión Spline, Regresión Logística, etc.;
ii. Análisis de los conjuntos de datos imputados, las en esta etapa comúnmente se utiliza el test ANOVA;
iii. Combinación de los resultados.

De este modo, podemos ver que existen varias técnicas para tratar los valores faltantes dependiendo del tipo valor al que corresponda, el nivel de conocimiento que tengamos de él y de su relación con las otras variables. En general podemos resumir las técnicas de acuerdo a la siguiente tabla con sus ventajas y desventajas (Wagstaff & Laidler, 2005):

Técnica	Ventajas	Deventajas
Marginalización de Variables: Omite Variables con Valores Perdidos	Simple	Se pierde información de todos los objetos
Marginalización de Objetos: Omite Objetos con Valores Perdidos	Simple	Se pierden Objetos
Imputación por la Media: Remplaza Cada Valor Perdido con la Media del Set de Datos	Simple	Tiende a ser impreciso; el valor promedio realmente nunca ocurre
Imputación Probabilística: Reemplaza con un valor aleatorio de acuerdo a la distribución de los valores en el set de datos	Valores inferidos son Reales (observaciones actuales)	Los valores inferidos tienden pueden no tener relación con los objetos
Imputación Por el Vecino Más Cercano: Reemplaza con el valor del vecino más cercano	Los valores inferidos son "los mejores posibles estimados"	Los valores inferidos pueden aún ser inapropiados (inobservables)

Por si parte, Scheffer (Scheffer, 2002) da las siguientes sugerencias para tratar casos perdidos:

- No usar imputación por el promedio a menos que el dato sea MCAR.
- Eliminar cuidadosamente verificando antes que el dato es MCAR.
- Single Imputation trabaja bien para datos faltantes MAR, siempre que menos del 10% de ellos sean datos nulos.
- Si se debe usar Single Imputation, use EM o Regresión.
- Si las estructuras de varianza en los datos son importantes, no use el método de Eliminación o Single Imputation si más del 5% de los datos están perdidos.
- Multiple Imputation opera correctamente para casos sobre el 25% de los datos perdidos.
- Para NMAR, sólo se podría usar Multiple Imputation, y preferiblemente con niveles de datos perdidos menores a 25%.
- Siempre que sea factible usar Multiple Imputation debido a sus características.

Consideraciones para Eliminación de una Variable

La eliminación de datos perdidos es una medida poco recomendable debido a la pérdida de información que genera, sin embargo puede ser una buena opción en caso que los datos perdidos sean de naturaleza MCAR y no puedan ser imputados fehacientemente (Scheffer, 2002). Si separamos la eliminación de datos perdidos en dos tipos: (a) Eliminación de Columnas y (b) Eliminación de registros, las consideraciones para eliminar los datos cambian. A su vez, podemos encontrar distintas categorías de datos fuera de rango dependiendo de la razón que los origina.

En primer lugar tenemos aquellas observaciones que provienen de un error de procedimiento (como podría ser una codificación, error de entrada de datos, etc.), son observaciones que provienen de un error de procedimiento, por ejemplo un error de entrada de datos, un error de codificación, etc. Si es que estos datos no son detectados mediante filtrado, se deben eliminar o recodificar como datos vacíos. En segundo lugar, también existen aquellas observaciones que ocurren a partir de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Los datos de esta categoría normalmente se retienen en la muestra, salvo que su significancia no sea relevante. En tercer lugar tenemos observaciones atípicas, las cuales corresponde a las observaciones extraordinarias donde no existe explicación de su origen, en este caso estas observaciones normalmente son eliminadas del análisis. Y finalmente, en cuarto lugar, nos encontramos con observaciones fuera de rango propiamente tal, las cuales corresponden a los casos atípicos, los cuales suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población.

En el caso que se esté pensando en eliminar una columna se debe tener en consideración la cantidad de datos perdidos que posee en total y si existe o no alguna relación con otra variable. Sería recomendable eliminar la columna en el caso que la cantidad de valores perdidos supere un umbral mínimo que permita análisis (por ejemplo que posea más de la mitad de datos perdidos probablemente no genere información fidedigna), o cuando la variable sea MCAR y no pueda ser deducida de ninguna forma con los datos existentes, por lo que imputarlos generaría un mayor costo de error que por pérdida de información.

Por su parte, al eliminar registros en primer lugar hay que ser cuidadoso con la proporción de las clases que están evaluando. Si la el problema a resolver constituye uno de clases desbalanceadas, la eliminación de un registro del cual hay pocas filas puede ser una gran pérdida de información inclusive su posee gran parte de sus atributos con datos perdidos o vacios. Por ello se debe tener cuidado con qué tipo de registro se está eliminando y tomar medidas de acuerdo a la información relativa que se pierde con su eliminación. Ahora, se recomienda eliminar registros siempre y cuando posean una gran cantidad de valores perdidos o blancos y correspondan a una pequeña cantidad dentro del total de los datos.

De esta manera, las medidas generalmente utilizadas para tratar valores fuera de rango son:

- Ignorar y mantener dentro del análisis
- Eliminar la columna
- Eliminar la fila
- Reemplazar el valor por: nulo, máximo o mínimo;
- Discretizar y hacer que los anómalos sean “muy alto” o “muy bajo”

Datos Nulos y Variable Objetivo

En un primer análisis de la data original sin modificaciones podemos ver que las variables que no tienen todos su valores completos son: B_GEBDATUM, Z_LAST_NAME, TIME_BEST, WERT_BEST_GES, DATUM_LBEST, MAHN_AKT, MAHN_HOECHST, ANUMMER_02, ANUMMER_03, ANUMMER_04, ANUMMER_05, ANUMMER_06, ANUMMER_07, ANUMMER_08, ANUMMER_09, ANUMMER_10, Z_CARD_ART. De las cuales podemos no todas son precisamente datos perdidos.

Variable Checks

Categorical	Cases Missing > 50	MAHN_AKT MAHN_HOECHST ANUMMER_02 ANUMMER_03 ANUMMER_04 ANUMMER_05 ANUMMER_06 ANUMMER_07 ANUMMER_08 ANUMMER_09 ANUMMER_10
	Cases Constant > 95	FLAG_NEWSLETTER Z_LAST_NAME CHK_LADR CHK_RADR CHK_KTO CHK_CARD CHK_COOKIE CHK_IP FAIL_LPLZ FAIL_LORT FAIL_LPLZORTMATCH FAIL_RPLZ FAIL_RORT FAIL_RPLZORTMATCH
	Categories Containing One Case > 90	ANUMMER_07 ANUMMER_08 ANUMMER_09
Scale	Cases Missing > 50	DATUM_LBEST

Each variable is reported with every check it fails.

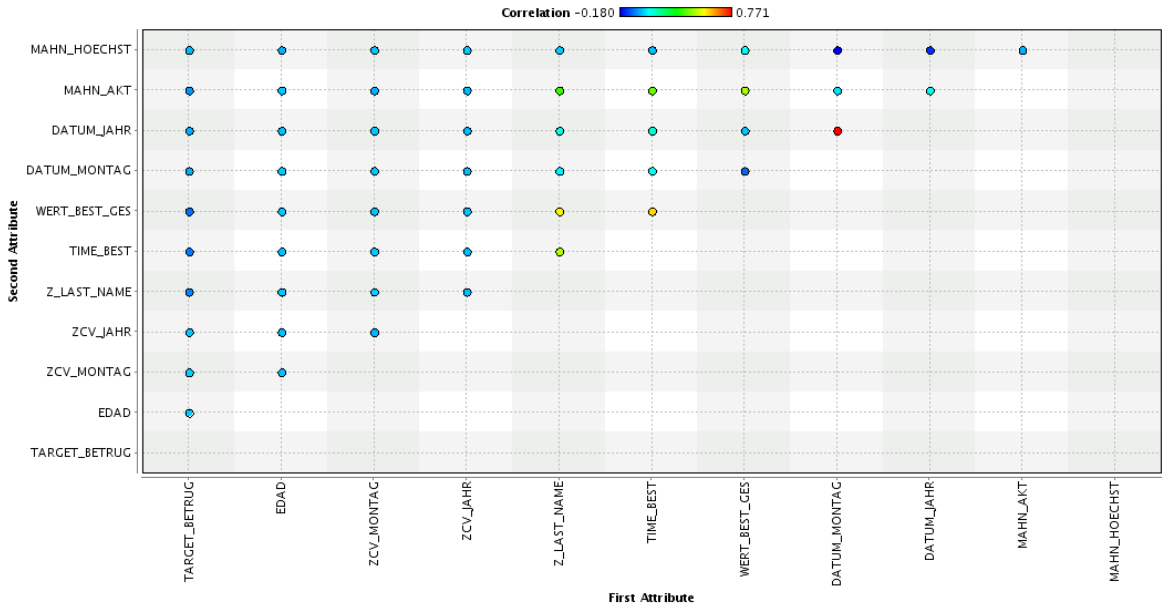
Las variables NUMMER_02, ANUMMER_03, ANUMMER_04,ANUMMER_05, ANUMMER_06, ANUMMER_07 ,ANUMMER_08,ANUMMER_09, ANUMMER_10, Z_CARD_ART por su naturaleza

dejan casos en blanco que no son vacíos. En las variables ANUMBER_XX se rellena en forma horizontal el detalle de los productos en caso que exista dentro de una compra, lo que genera valores en blanco, mientras que Z_CARD_ART sólo tiene datos cuando las compras son con Kundercarter o Kreditcarter en la variable Z_METHOD,generando también valores blancos. Luego, se revisó si para las filas de Kunercarter y Kreditcarter existía un valor asociado, y se encontró que estas variables no tienen valores perdidos. Las variables con valores perdidos propiamente tal y sus porcentajes sobre el total son:

B_GEBDATUM	10%
Z_LAST_NAME	49%
TIME_BEST	0%
WERT_BEST_GES	46%
DATUM_LBEST	53%
MAHN_AKT	53%
MAHN_HOECHST	53%

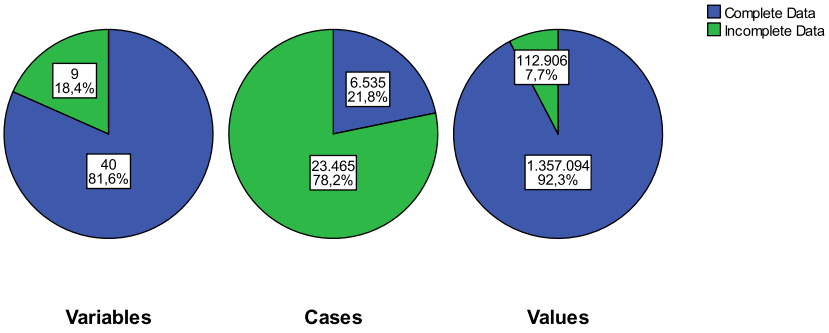
En el caso de la variable WERT_BEST_GES hay que tener cuidado porque no es claro si son datos perdidos o datos en blanco, ya que es el valor de la compra anterior en caso que corresponda, con lo que podían existir tanto datos perdidos como datos en blanco.

Para poder manejar los datos y realizar el análisis de los datos perdidos, se trabajó con la base de datos original modificada, con la cual se obtuvieron las siguientes correlaciones con la variable objetivo:



Claramente vemos que no existe mayor correlación entre la variable objetivo y las variables que poseen valores nulos. Ahora, realizando un análisis de los valores perdidos y su comportamiento en la base de datos llevamos a cabo un "Análisis de Patrones" en las opciones de "Imputación de Datos Perdidos" del editor SPSS, encontramos los siguientes resultados:

Overall Summary of Missing Values



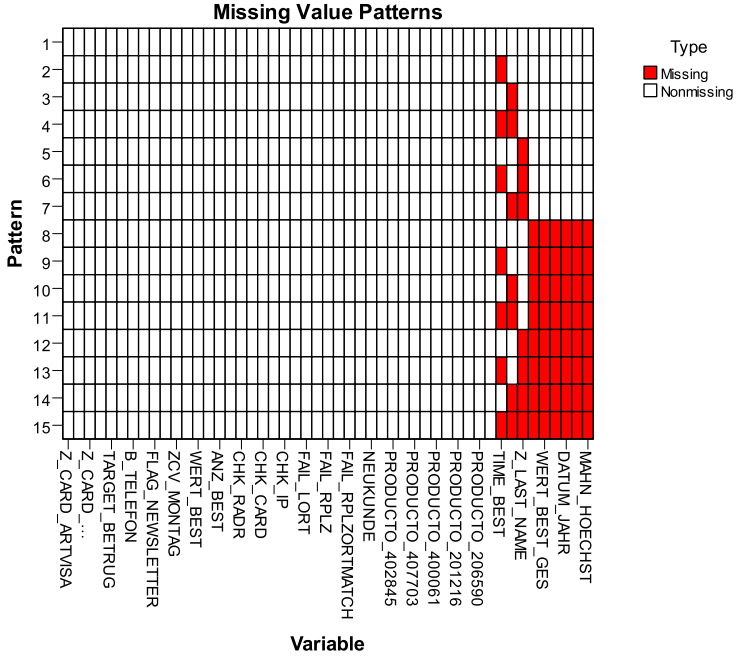
Variable Summary[a,b]

	Missing				
	N	Percent	Valid N	Mean	Std. Deviation
MAHN_HOECHST	15856	52,9%	14144		
MAHN_AKT	15856	52,9%	14144		
DATUM_JAHR	15856	52,9%	14144		
DATUM_MONTAG	15856	52,9%	14144		
WERT_BEST_GES	15856	52,9%	14144	63,3675	69,72758
DIF_VKAUFEN	15856	52,9%	14144	3,4589	1,47320
Z_LAST_NAME	14808	49,4%	15192		

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 10,0%

Missing Value Patterns



Lo que nos muestra que dentro del total de las variables existe un bajo porcentaje de (18%) de valores perdidos, pero que estos a través de los registros generan una gran cantidad de filas incompletos (al 78% le falta algún dato). Sin embargo, el 92% de los valores totales se encuentra completo.

El algoritmo EM es utilizado en el mundo estadístico para encontrar los estimadores de máxima verosimilitud de los parámetros en modelos probabilísticos. Este algoritmo varía entre desarrollar una fase de esperanza donde se calcula la esperanza de la similitud incluyendo las variables disponibles, y una fase de maximización donde se calcula el estimador de máxima verosimilitud de los parámetros maximizando la similitud esperada encontrada en la fase E. Los parámetros encontrados en la etapa M son usados para comenzar otra etapa E. Esto se utiliza para realizar reemplazo de variables por valores estimados que se calculan iterativamente hasta llegar a un punto de convergencia. La fase E calcula la esperanza condicional de los datos faltantes dados los datos observados y la estimación de los parámetros, luego estas esperanzas sustituyen a los datos faltantes, mientras que, la fase M, realiza la estimación máxima-verosimilitud del parámetro de interés como si no existieran datos faltantes (Dempster, Laird, & Rubin, 1977).
Para poder utilizar la imputación mediante EM es necesario que las variables no sean MCAR, por lo que, se realizó el análisis de EM para verificar que las variables no fuesen MCAR obteniendo los siguientes resultados:

EM Correlations[a]

	ALTER	DIF_VKAUFEN	WERT_BEST	TIME_BEST	SESSION_TIME	WERT_BEST_GES
ALTER	1					
DIF_VKAUFEN	-,001	1				
WERT_BEST	-,001	,009	1			
TIME_BEST	,000	,012	,006	1		
SESSION_TIME	,007	,018	,017	-,001	1	
WERT_BEST_GES	,003	-,022	-,020	,020	,017	1

a. Little's MCAR test: Chi-Square = 201,902, DF = 26, Sig. = ,000

EM Means[a]

ALTER	DIF_VKAUFEN	WERT_BEST	TIME_BEST	SESSION_TIME	WERT_BEST_GES
33,42	3,4451	45,8583	12,3352	8,59	63,6637

a. Little's MCAR test: Chi-Square = 201,902, DF = 26, Sig. = ,000

Summary of Estimated Standard Deviations

	ALTER	DIF_VKAUFEN	WERT_BEST	TIME_BEST	SESSION_TIME	WERT_BEST_GES
All Values	9,643	1,47320	35,70943	6,25770	3,863	69,72758
EM	9,646	1,47200	38,20312	6,25949	3,872	70,50550

Summary of Estimated Means

	ALTER	DIF_VKAUFEN	WERT_BEST	TIME_BEST	SESSION_TIME	WERT_BEST_GES
All Values	33,44	3,4589	43,9681	12,3352	8,58	63,3675
EM	33,42	3,4451	45,8583	12,3352	8,59	63,6637

Vemos entonces que por el test de Little se rechaza la Hipótesis nula de que las variables son MCAR dado que el pvalue es muy cercano a cero. De acuerdo a este test podemos tomar medidas de imputación en cada una de las variables con valores perdidos.

Consideraciones y Aplicación de Tratamiento de datos nulos en la base de datos Retail

Complementando el análisis realizado en el punto (a), Farhangfara et al. (Farhangfara, Kurganb, & Dy, 2008), sugiere que es factible eliminar registros con datos perdidos siempre y cuando:

- La cantidad de registros a eliminar sea comparativamente baja en relación a la totalidad de los registros.

- Cuando no se produce un sesgo importante en el análisis producto de la no utilización de los registros con datos perdidos; es decir, no se pierde otra información valiosa para el análisis debido a la eliminación del registro.

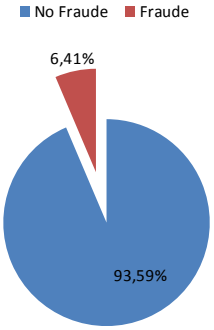
En este caso, nos encontramos con un problema donde las clases son desbalanceadas no sólo en la variable objetivo, sino además en la mayoría de las variables. Además el 90% de las variables son categóricas, los que hace bastante difícil decidir la eliminación de los registros a simple vista.

En la base de datos, se decidió eliminar los registros que poseían en blanco TIME_BEST, ya que correspondían a sólo 20 registros y sólo 1 era de la clase fraude, lo cual en comparación a los 30.000 registros originales genera una pérdida insignificante de información.

A su vez, se revisó qué registros poseían una cantidad alta de datos vacíos a través de las variables y nos encontramos con 719 (2% de la data) registros que poseen las variables ALTER, DIF_VKAUFEN, Z_LAST_NAME WERT_BEST_GES,DATUM_MONTAG, DATUM_JAHR, MAHN_AKT, MAHN_HOECHST de los cuales 62 (4% dentro de la clase) son de la clase FRAUDE y 657 (2% dentro de la clase) de la clase No FRAUDE . Como existe mayor información sobre la clase No Fraude, consideramos que no es una gran pérdida de información eliminar los 657 registros, por lo que sólo eliminamos los registros de esta clase bajo esta regla.

Por su parte, fue evaluada también la variable B_GETDATUM que posee un 10% de valores nulos, pero en su caso la cantidad de registros, si bien no es de gran magnitud, si puede afectar el análisis. En general se acepta trabajar con el 95% de los casos, por lo que eliminar estos registros es arriesgado en un primer análisis. En este caso existen 177 registros “Fraude” (10% del total de Fraudes) con valores perdidos en la variable B_GETDATUM, lo cual es una gran pérdida de información en vista que la clase Fraude tiene muy pocos registros. En este sentido, se decidió eliminar los registros 2762 (10% del total de “No Fraudes”) que poseen valores perdidos en la variable B_GETDATUM debido a que en esta clase hay suficiente información y su eliminación no debería generar mayor pérdida, casos que luego de la eliminación anterior (registros con varios blancos) corresponden a 2105 registros.

Con ello trabajaríamos con el 91% (27.218) de los datos originales sin generar pérdida de información en la clase objetivo (“Fraude”). Lo cual se considera aceptable para nuestro análisis ya que además no cambia la proporción entre las clases y, además, con estos cambios se evidencia una mejora en la calidad de los datos.

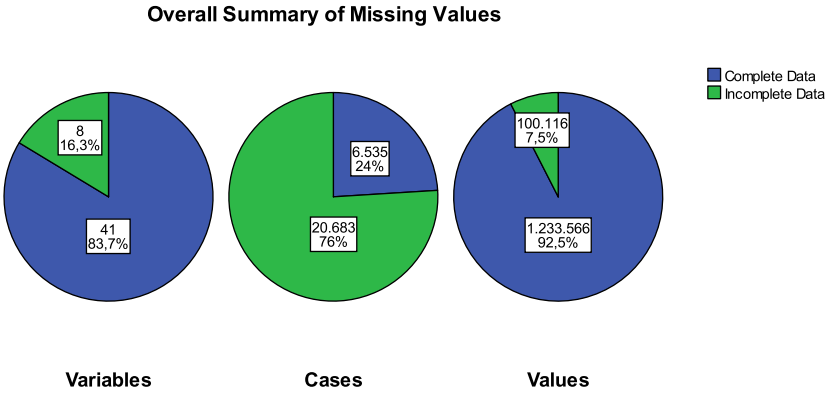


Variable Summary[a,b]

	Missing				
	N	Percent	Valid N	Mean	Std. Deviation
MAHN_HOECHST	14412	53,0%	12806		
MAHN_AKT	14412	53,0%	12806		
DATUM_JAHR	14412	53,0%	12806		
DATUM_MONAT	14412	53,0%	12806		
WERT_BEST_GES	14412	53,0%	12806	62,763470	68,7675442
DIF_VKAUFEN	14412	53,0%	12806	2,971107	1,5003139
Z_LAST_NAME	13467	49,5%	13751		

a. Maximum number of variables shown: 25

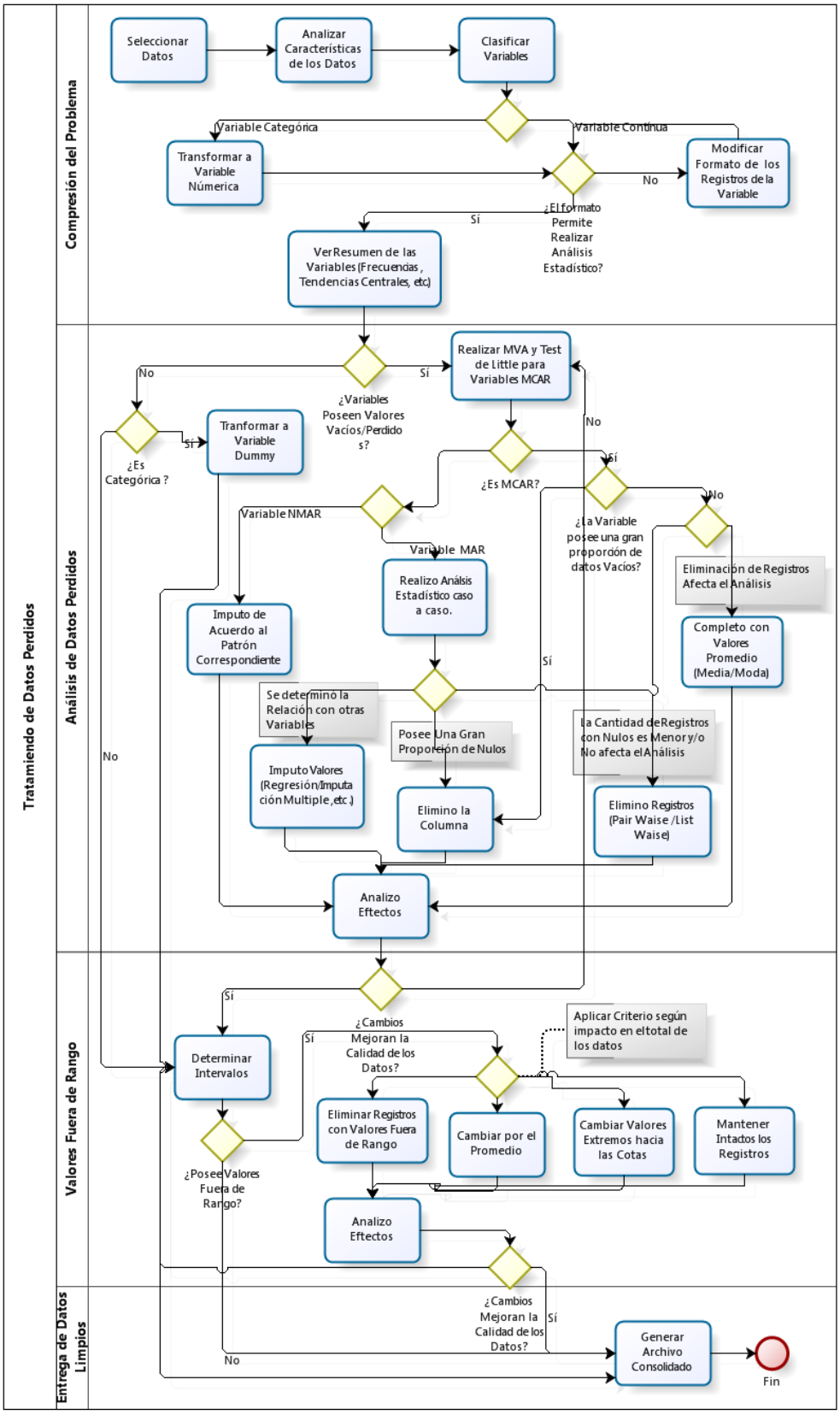
b. Minimum percentage of missing values for variable to be included: 10,0%



Metodología

A grandes Rasgos la metodología sigue las siguientes fases mostradas en la siguiente figura, las cuales pueden ir iterando a medida que se va entendiendo mejor el problema o si se mejora el análisis.

En particular, tomando la etapa de Limpieza de Datos y Preprocesamiento de las variables. La metodología es la siguiente:



Luego, y de acuerdo a esta metodología fue aplicada en la base de datos RETAIL. En primer lugar, y como está documentado en la primera sección de este informe, realizamos un análisis de los datos verificando si era posible realizar análisis estadístico mediante software especializado, en este caso SPSS y RapidMiner que se utilizaron de manera complementaria. Con ello obtuvimos los estadísticos de tendencia central (media, modas, desviaciones) y las frecuencias, identificando además gráficamente los valores fuera de rango y la proporción de valores en blanco. Se analizó además los tipos de variables y fueron clasificadas entre categóricas y continuas, modificando las categóricas de manera que pudiesen ser analizadas de mejor manera convirtiéndolas en dummys.

Luego, se llevo a cabo la revisión de los valores en vacíos y los valores perdidos, identificando cuáles eran efectivamente nulos y cuáles eran errores o problemas de almacenamiento (ver punto b pregunta I). Con ellos se eliminaron las filas con baja importancia relativa y se corrigió la estructura de las variables con valores blancos. Luego, se llevó a cabo un análisis de correlaciones entre datos, en particular se utilizó el test de PCA para identificar grupos de variable, ver el nivel de significancia (nivel de ajuste con todas las variables) y de acuerdo a la tabla de Comunalidades se evalúo la significancia de las variables con valores perdidos mayor al 40%.

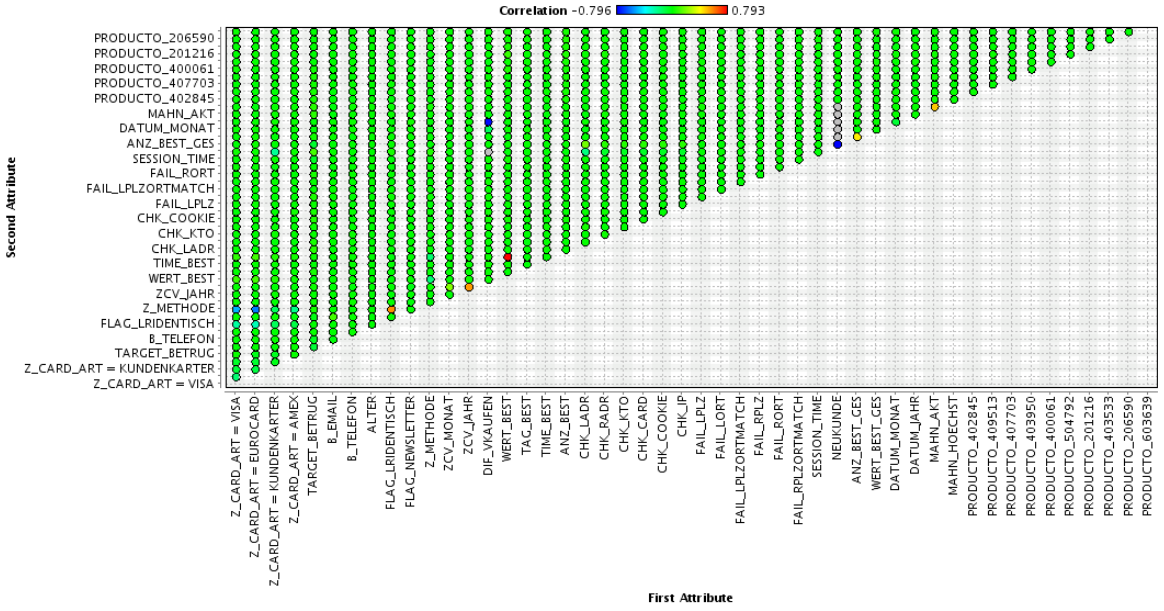
Variable	Inicial	Extracción
WERT_BEST_GES	1,000	,588
DIF_VKAUFEN	1,000	,941
DATUM_MONAT	1,000	,639
DATUM_JAHR	1,000	,887
MAHN_AKT	1,000	,740
MAHN_HOECHST	1,000	,739
Z_LAST_NAME	1,000	,476

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,259
Bartlett's Test of Sphericity	Approx. Chi-Square	42573,624
	df	1081
	Sig.	,000

Con ello se decidió eliminar la variable Z_LAST_NAME por la gran cantidad de valores perdidos y la poca relevancia en la explicación del modelo (extracción<0,5).

Luego, realizando un análisis de correlación entre las variables se determinó que variables podían explicar mejor las variables con valores perdidos, para luego ocupar esas variables como base para una Imputación Múltiple. De lo cual se obtuvo el siguiente resultado:



En este análisis, si bien hay correlación entre las variables, no existe un patrón entre ellas que permita distinguir cuáles son las mejores variables para imputar los datos. Es por ello que se decidió rellenar con la moda las variables categóricas dependiendo de la variable objetivo, lo que corresponde a:

MAHN_AKT→ Moda Fraude = 0 ; Moda No Fraude = 0
MAHN_HOECHST → Moda Fraude = 0 ; Moda No Fraude = 0

Luego las variables DATUM_JAHR, DATUM_MONAT, WERT_BEST_GES y DIF_VKAUFEN fueron completadas mediante imputación múltiple de la siguiente manera:

Imputation Constraints

	Role in Imputation		Imputed Values	
	Dependent	Predictor	Minimum	Maximum
TARGET_BETRUG	No	Yes		
Z_METHODE	No	Yes		
WERT_BEST	No	Yes		
NEUKUNDE	Yes	Yes		
ANZ_BEST	No	Yes		
WERT_BEST_GES	Yes	No	(none)	(none)
DIF_VKAUFEN	Yes	No	(none)	(none)
DATUM_MONAT	Yes	No		
DATUM_JAHR	Yes	No		

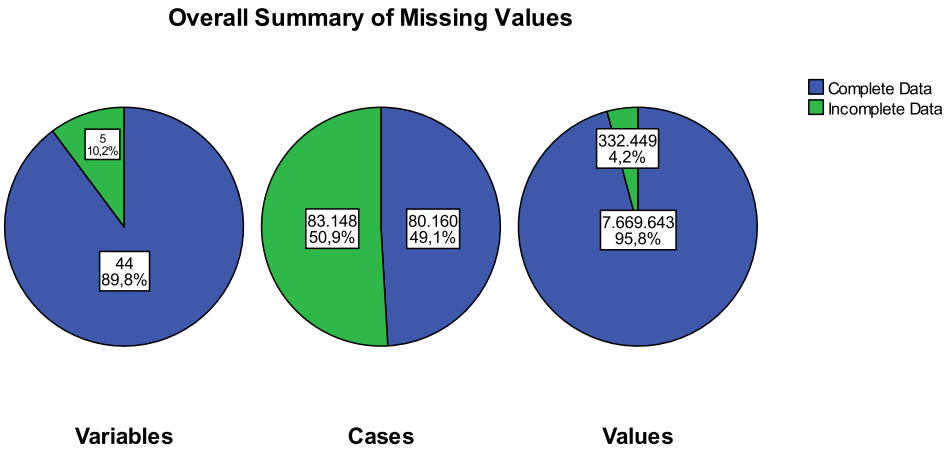
Imputation Results

Imputation Method		Monotone
Fully Conditional Specification Method Iterations		n/a
Dependent Variables	Imputed	WERT_BEST_GES,DIF_VKAUFEN, DATUM_MONAT,DATUM_JAHR
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	TARGET_BETRUG,Z_METHODE, WERT_BEST,NEUKUNDE,ANZ_BEST
Imputation Sequence		TARGET_BETRUG,Z_METHODE, WERT_BEST,NEUKUNDE,ANZ_BEST, WERT_BEST_GES,DIF_VKAUFEN, DATUM_MONAT,DATUM_JAHR

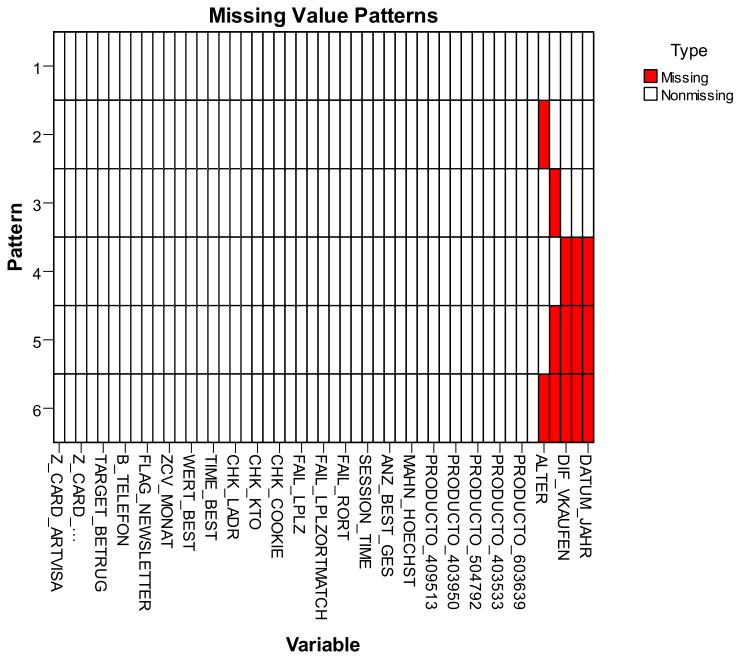
Imputation Models

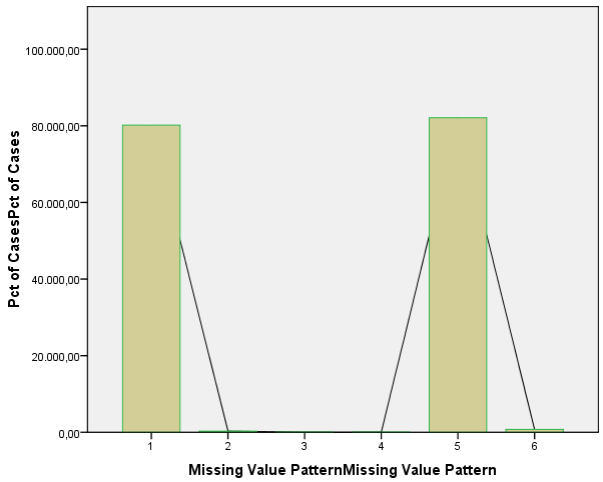
	Model		Missing Values	Imputed Values
	Type	Effects		
WERT_BEST_GES	Linear Regression	TARGET_BETRUG, Z_METHODE, NEUKUNDE, ANZ_BEST, WERT_BEST	14412	3626
DIF_VKAUFEN	Linear Regression	TARGET_BETRUG, Z_METHODE, NEUKUNDE, ANZ_BEST, WERT_BEST	14412	3625
DATUM_MONAT	Logistic Regression	TARGET_BETRUG, Z_METHODE, NEUKUNDE, ANZ_BEST, WERT_BEST	14412	3625
DATUM_JAHR	Logistic Regression	TARGET_BETRUG, Z_METHODE, NEUKUNDE, ANZ_BEST, WERT_BEST	14412	3625

Las variables de predicción fueron seleccionadas de acuerdo a las comunalidades de extracción y de acuerdo a la importancia extraída del análisis. Como vemos en el resultado, no todos los valores perdidos fueron completados, por lo que se ocupó un segundo criterio para completar los valores restantes. Revisando los cambios generados, vemos que la imputación genera una mejora en las proporciones de datos perdidos:



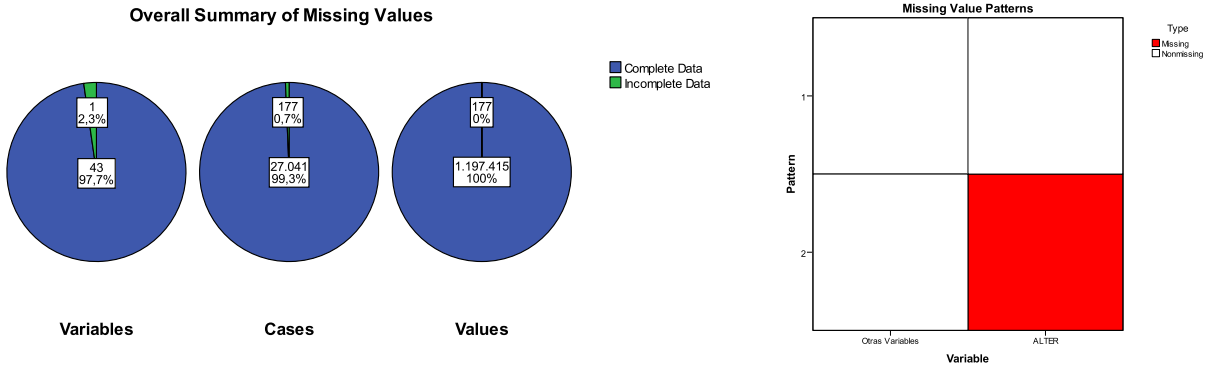
Se encontraron 6 patrones, donde el primero muestra los casos sin valores perdidos, el segundo con valores perdidos en ALTER, el tercero en WERT_BEST_GES, el cuarto en DIF_VKAUFEN, DATUM_MONAT y DATUM_JAHR, el quinto en WERT_BEST_GES , DIF_VKAUFEN, DATUM_MONAT y DATUM_JAHR y finalmente el sexto en ALTER, WERT_BEST_GES , DIF_VKAUFEN, DATUM_MONAT y DATUM_JAHR. Luego, la mayoría de los casos posee el patrón 1 y el 5.



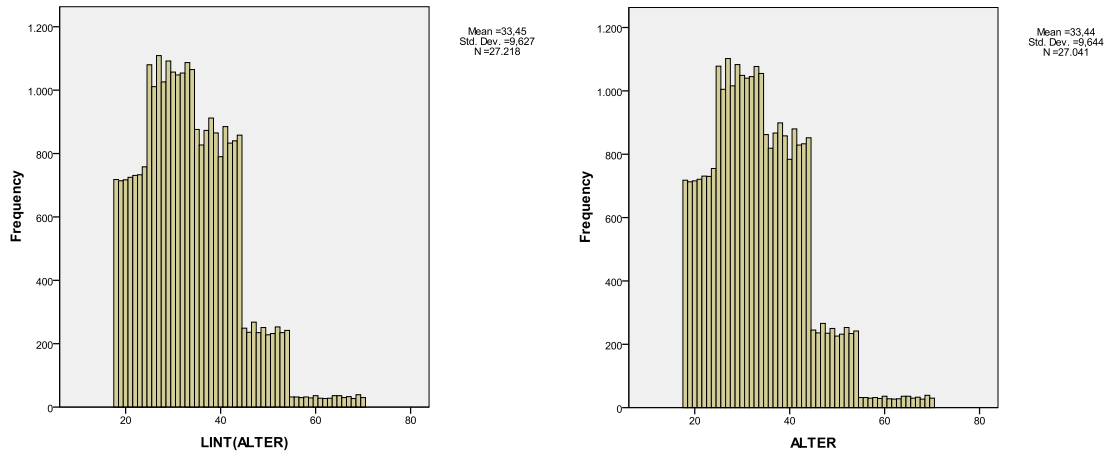


De acuerdo a ello y recordando las transformaciones realizadas, se decidió eliminar las variables WERT_BEST_GES, DIF_VKAUFEN, DATUM_MONAT y DATUM_JAHR ya que en sí DIF_VKAUFEN, DATUM_MONAT y DATUM_JAHR son obtenidas desde DATUM_LBEST y DIF_VKAUFEN también fue creada a partir de DATUM_LBEST, por lo que en su origen sería equivalente a eliminar sólo una columna con una gran cantidad de valores perdidos. Por su parte WERT_BEST_GES es el monto en de las compras anteriores si es que se conocía el dato, por lo que en su totalidad no era una variable totalmente confiable.

Luego, los valores perdidos quedan de la siguiente manera:



Luego se completo la variable ALTER con interpolación lineal, luego de ver los efectos de la distribución con imputación con la media y con la media de los 10 valores cercanos, de acuerdo a ello, la interpolación lineal mantiene la distribución original.



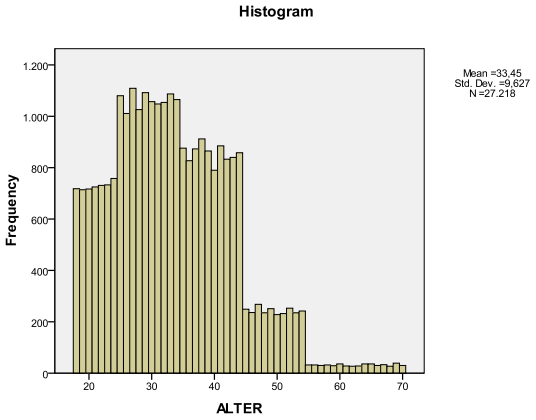
La base queda sin valores perdidos y con ello, se comenzó el análisis de valores extremos o extraños (outliers). Para ello, en primer lugar se realiza análisis visual mediante histogramas y bloxspots (ver salida SPSS de outliers).

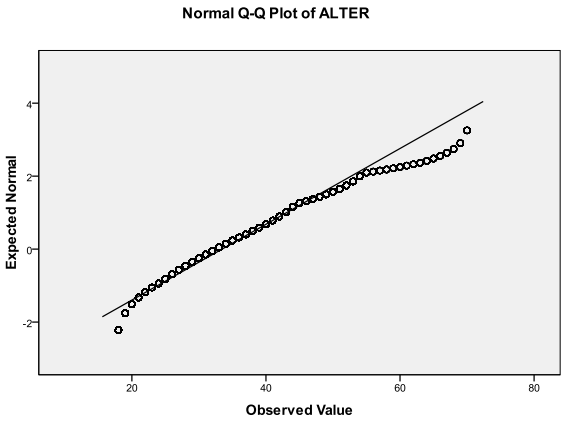
La corrección de los valores será del tipo univariado. En las variables categóricas se decidió no realizar acciones. Para las variables continuas se llevó un análisis más en profundidad obteniendo los siguientes resultados:

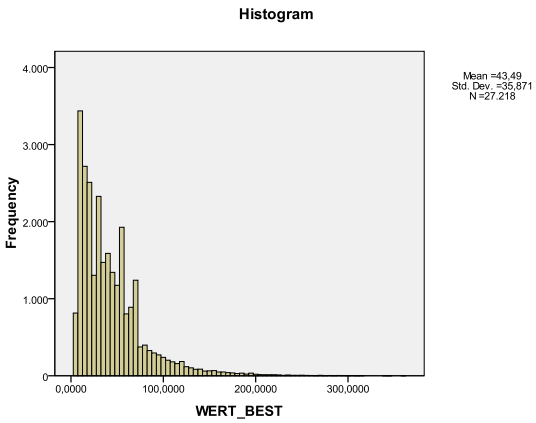
Case Processing Summary

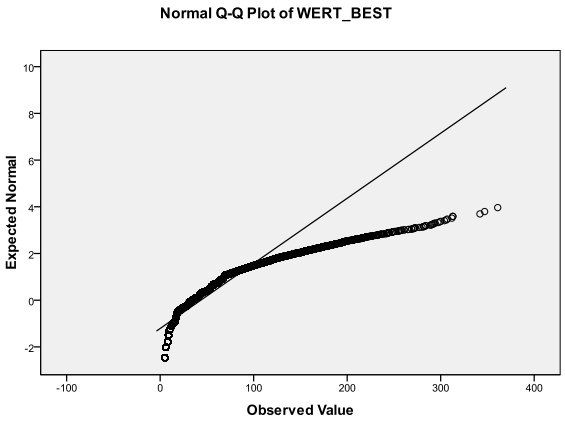
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ALTER	27218	100,0%	0	,0%	27218	100,0%
WERT_BEST	27218	100,0%	0	,0%	27218	100,0%
TIME_BEST	27218	100,0%	0	,0%	27218	100,0%
ANZ_BEST	27218	100,0%	0	,0%	27218	100,0%
SESSION_TIME	27218	100,0%	0	,0%	27218	100,0%

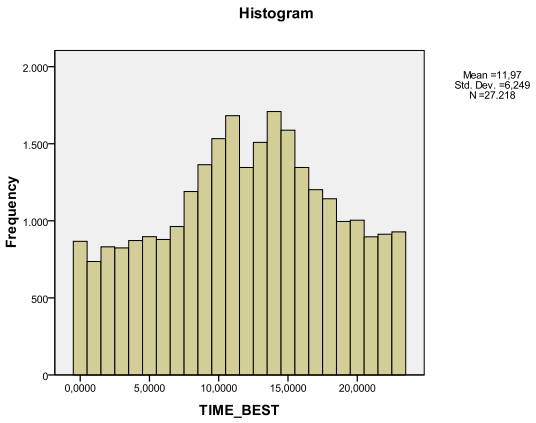
Histogramas

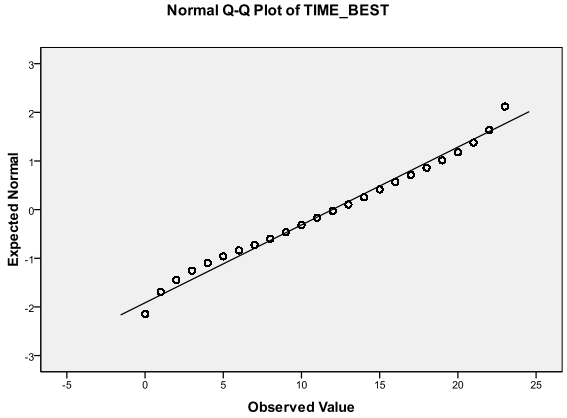


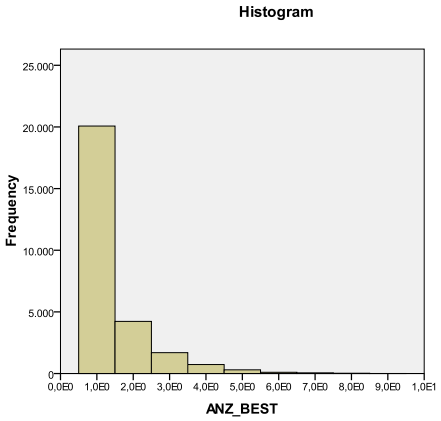


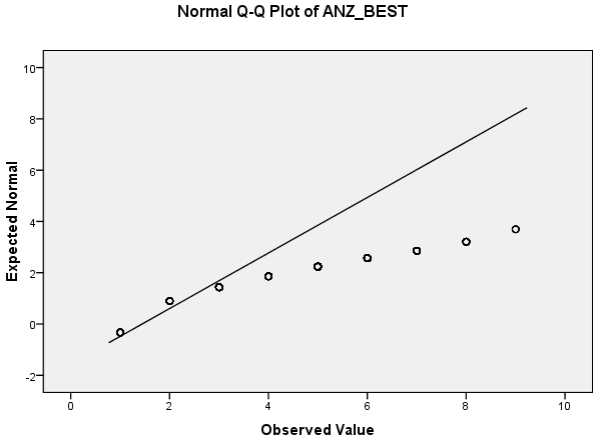


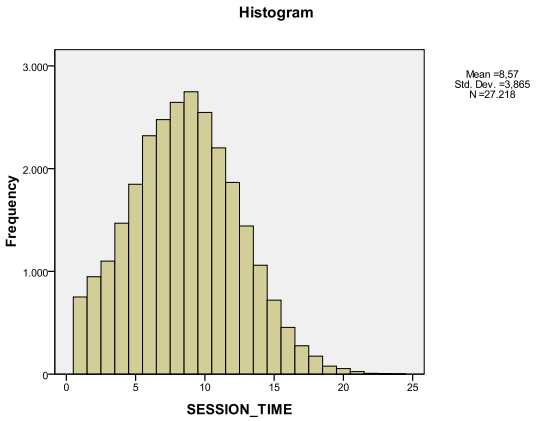


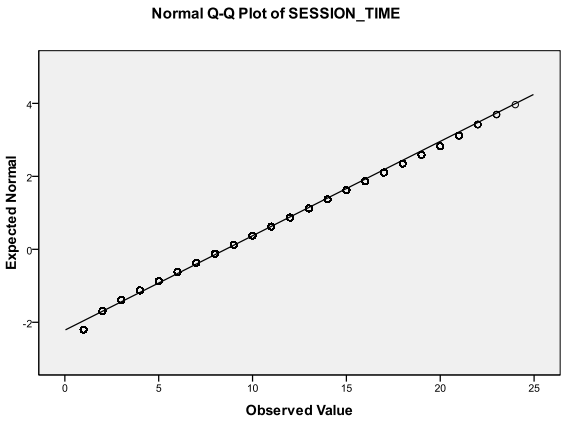






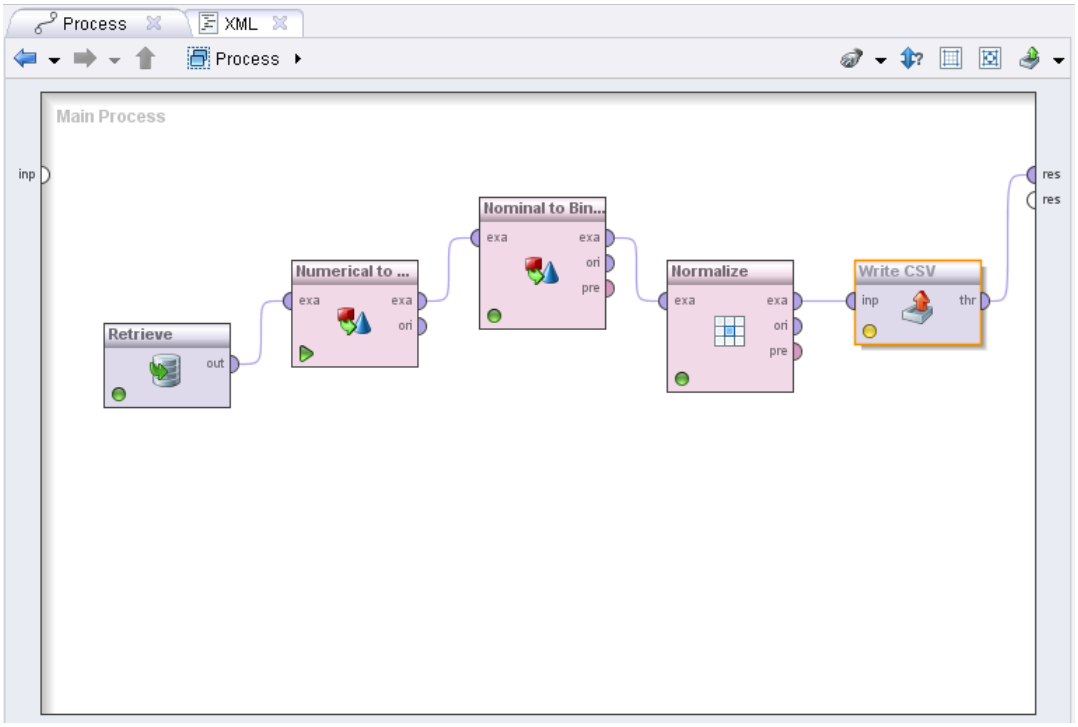






En vista de estos resultados y la dificultad computacional de realizar el cálculo de los outaliers multivariados, se decidió normalizar los atributos continuos y los categóricos convertir en variables dummy. Con ello la base de datos quedó sin valores vacíos y con 52 atributos con valores entre 0 y 1.

El procedimiento efectuado fue realizado en RapidMiner. Siguiendo los pasos de leer base de datos desde un archivo .cvs, pasar valores numéricos a nominales luego nominales a binominales y finalmente normalizar todos los atributos y con ello se generó el archivo para realizar el análisis de selección de variables.



Pregunta 2 : Valores fuera de Rango (Outliers)

Los Outliers, también conocidos como anormalidades en los datos, son observaciones dentro de una secuencia de datos que es anómala respecto a la conducta presente en la mayoría de las observaciones dentro del mismo conjunto de datos. (Pearson, Mining Imperfect Data, 2005). La existencia de outliers es una complicación común en los análisis de datos y es necesario tomar ciertas medidas sobre ellos para evitar generar sesgos en el análisis.

Podemos clasificar a los outlayers en univariados (dentro de la misma variable), bivariados y multivariados (varias anomalías en un solo registro). En esta ocasión, haremos referencia a la clase de outliers univariados debido a que son los más simples de detectar y su corrección es necesaria para mejorar el análisis.

Existe una serie de procedimientos en la detección de outliers univariados, pero se presentan en este trabajo los tres métodos más recurrentes:

- La regla 3σ: $x_0 = \overline{x},\ \zeta = \hat{\sigma}$;
- El identificador de Hampel: $x_0 = Me,\ \zeta = S$;
- La regla del Boxplot: $x_0 = Me;\ \zeta = Q$

En general, para tratar outlayers se considera una muestra de datos (de alguna variable) $\{x_k\}$, sobre la cual se determinan un valor referencial x_0 y una medida de variación ζ. Luego, se elige un umbral t que permite determinar los valores outliers de acuerdo a la siguiente regla genérica:

$$Si\ \ |x_k - x_0| > t \cdot \zeta \Longrightarrow\ x_k\ es\ un\ outlier$$

Intuitivamente en la ecuación anterior, si una observación se encuentra muy alejada de x_0 será considerada un outlier. La medida de variación ζ caracteriza la "dispersión natural" de los valores de la variable, lo que permite calibrar las distancias desde x_0; y finalmente, el parámetro t determina que tan tolerantes seremos en la detección de outliers, en particular, si $t = 0$, entonces todos los valores distintos de x_0 serán considerados outliers.

Para convertir esta regla en un procedimiento práctico de detección de outliers, es necesario responder las siguientes preguntas: distensión

- ¿Cómo determino el valor referencial x_0?
- ¿Cómo determino la variación natural de la escala ζ?
- ¿Cómo elijo el umbral o tolerancia t?

Es importante siempre hacer la separación entre variables categóricas y variables continuas. En el caso de las variables categóricas, una vez declarados los valores válidos, el recuento de frecuencias es un buen indicador para identificar registros extraños. Mientras que, en el caso de las variables continuas, las estadísticas descriptivas, en particular el rango y un histograma, pueden ayudar a identificar los valores extremos y fuera de rango.

Procedimientos clásicos en la detección de outliers

Bajo el supuesto de distribución simétrica respecto del valor x_0 de los datos, existen dos alternativas obvias para la elección del valor referencial x_0 que son el promedio ($\bar{x}$) y la mediana (Me). De manera similar, existe una amplia variedad de posibles elecciones para la medida de variación ζ, pero las siguientes tres son las más utilizadas: la desviación estándar ($\hat{\sigma}$), el estimador MAD (S) y el primer cuartil (Q), estos parámetros se pueden desprender a partir de los métodos que describiremos a continuación.

La Regla 3σ (Wright, 1884)

Dada una secuencia de datos $\{x_k\}$ cuya distribución se aproxima a la secuencia de variables aleatorias i.i.d. Gaussiana con media μ y varianza σ^2, entonces la probabilidad que una observación se encuentre más allá de 3σ de distancia respecto de la media es aproximadamente 0,3 %. Para esta regla, los tres parámetros requeridos son:

a. Elegir el promedio como el valor referencial x_0
b. Elegir la desviación estándar estimada $\hat{\sigma}$ como la medida de variación ζ
c. Elegir el umbral de tolerancia $t = 3$, según lo indicado anteriormente

Desafortunadamente, más allá de su importancia histórica y simplicidad, este procedimiento de detección de outliers tiende a no ser efectivo en la práctica. La razón de ellos es que la presencia de outliers tiende a generar errores en dos de los parámetros requeridos por la regla, la media y la desviación estándar.

El Identificador de Hampel

A partir de la regla anterior se puede deducir que es necesario identificar parámetros para el procedimiento de detección de outliers que sean resistentes a la presencia de outliers en la secuencia de datos $\{x_k\}$

La alternativa obvia al promedio es la mediana Me, y una atractiva alternativa a la desviación estándar es la estimación de escala MAD (Median Absolute Deviation) S definida por:

$$S = \frac{1}{0{,}6745} \times Mediana\{|x_k - Mediana(\{x_k\})|\}$$

Es decir, el estimador de escala MAD es definido como una versión escalada de la mediana de las deviaciones absolutas de la mediana del conjunto de datos; es decir, la secuencia $\{|x_k - Mediana(\{x_k\})|\}$ mide la distancia de cada elemento de la secuencia de datos respecto del valor referencial de la mediana de los datos. El factor $\frac{1}{0{,}6745} = 1{,}4826$ permite que el estimador S sea un estimador insesgado de la desviación estándar σ cuando la secuencia de datos $\{x_k\}$ tiene distribución Gaussiana.

Ahora, debido a que la mediana Me y el estimador de escala MAD S tienen (ambos) muy poca sensibilidad a la existencia de outliers, en general el identificador de Hampel es mucho más efectivo que la regla anterior. Es importante mencionar que el estimador de escala MAD tiene un problema numérico importante, si más del 50% de los datos en la secuencia de datos son idénticos, entonces $S = 0$, lo que en práctica hace que todos los valores distintos a la Mediana sean considerados outliers.

La detección basada en Quartiles y Boxplots (La regla del Boxplot)

Otro posible estimador de escala resistente a la presencia de outliers es el estimador IQD Q, o amplitud inter-cuantílica $Q = Q_3 - Q_1 = P_{75} - P_{25}$

De esta manera el detector queda determinado por:

$$|x_k - Me| > t \cdot Q \implies x_k \; es \; un \; outlier$$

Ninguno de los tres procedimientos presentados es superior a los otros dos, en particular, el desempeño relativo de estos procedimientos depende fuertemente de dos cosas, el nivel de contaminación de los datos y el carácter de los outliers presentes en los datos.

Finalmente, complementando los tres métodos presentados, cabe destacar que las últimas investigaciones al respecto han considerado como un factor relevante la explosiva cantidad de datos que se generan hoy en día. En este sentido, los métodos Bloxplot y Hampel resultan ineficientes desde el punto de vista computacional debido a lo costoso que puede resultar la determinación de la mediana en un conjunto importante de datos, por lo cual la investigación se ha orientado a buscar parámetros para la detección de outliers que sean resistentes a la presencia de éstos y además resulten menos costosos de determinar, como por ejemplo una versión simplificada de LTS. (Rousseeuw & Driessen, 2002)

Pregunta 3 : Procesamiento y re-codificación de variables y Estrategias para Disminuir la Dimensionalidad

El preprocesamiento y recodificación de las variables son las etapas donde nos cercioramos que los datos estén las mejores condiciones para poder ser analizados. Dentro de las actividades que se incluyen en estas fases se encuentran el análisis de valores perdidos, identificación de atributos clave, corrección de ruido e identificación de valores anómalos o fuera de rango e identificación de inconsistencias (Pei, Kamber, & Han, 2005). De este modo, las tareas que se originan en el preprocesamiento de datos son

- **Data Cleaning:** etapa donde se lleva a cabo la limpieza de la base de datos limpiado los valores perdidos, suavizando el ruido, identificar o remover outliers, resolver inconsistencias, etc. Este ítem ha sido tratado en las preguntas anteriores de este informe.
- **Data Integration:** los datos disponibles para el análisis generalmente provienen de diversas fuentes de datos, lo que hace necesario llevar la estructura de los datos a un estándar dónde todas las variables tengan el mismo significado y la misma estructura. En el caso de la tarea los datos venían ya integrados, por lo cual no fue necesario llevar cabo esta etapa.
- **Data Transformation:** consiste en cambiar o, como su nombre lo dice, transformar las variables con operaciones como normalización o agregación de los datos con el objetivo de poder mejorar la calidad de la información y permitir el análisis mediante distintos modelos, que muchas veces requieren que los datos se encuentren con estructuras determinadas. En la base de datos de retail, las transformaciones fueron realizadas luego de encontrar los valores perdidos.
- **Data Reduction:** esta fase se lleva a cabo cuando el volumen de datos genera dificultades para realizar el análisis. Algunas técnicas para reducir los datos son Data aggregation (resumir las variables agrupadas por algún criterio), Attribute subset selection (remover atributos de poca relevancia analítica), dimensionality reduction (usar esquemas de codificación para representar las variables en un número menor de atributos) y Numerosity Reduction (disminuir la cantidad de registros, ocupar grupos, etc). También se pueden ocupar técnicas de Data Generalization, donde mediante estructuras jerárquicas se lleva a cabo un resumen de acuerdo a niveles más altos y resumidos).
- **Data Discretization y Automatic Generation of Concept Hierarchies:** esta fase puede ser vista como una transformación de variables donde se generan rangos de datos y/o de variables. En el caso de los valores numéricos se utilizan técnicas como binning, análisis de histogramas, discretización basada en la entropía, análisis $\chi 2$, análisis de clúster y discretización por particiones intuitivas. En el caso de los valores categóricos, se aplican métodos que se generan jerarquías conceptuales basadas en los distintos valores que definen la jerarquía en cuestión.

En esta las secciones anteriores vimos en detalle la limpieza de datos. Ahora sección veremos en particular la Transformación y Reducción de datos.

La transformación de variables, como su nombre lo indica, es la fase donde se transforman y consolidan los datos de acuerdo a una estructura apropiada para realizar minería de datos. Las estrategias utilizadas para realizar esta tarea son (Han et al, 2005):

1. **Smoothing:** su objetivo es encontrar ruido en los datos y suavizarlos mediantes técnicas como binning, regresión y clustering.
2. **Aggregation:** consiste en realizar resumen de las variables mediante el uso de operaciones de sumarización, agregación o generalización.
3. **Normalization/Standardization:** consiste en llevar a los atributos a alguna distribución conocida como la normal, donde los valores que posee son escalados dentro de un rango específico, como por ejemplo [-1, 1] o [0,1].
4. **Attribute Construction (o feature construction):** consiste en generar nuevos atributos a partir de los atributos existentes.

Por su parte, la Reducción de Datos, consiste en disminuir tanto la cantidad de datos como la cantidad de variables, de manera tal que el análisis mejore y/o pueda ser llevado a cabo. Muchas veces la cantidad de variables o de datos existentes dificulta e incluso imposibilita aplicar técnicas de minería de datos debido tanto a la dificultad de análisis como a la capacidad de los modelos y aplicaciones para manejar datos. Las estrategias existentes para reducir la cantidad de datos son (Pei, Kamber, & Han, 2005):

1. **Data Cube Aggregation:** consiste en aplicar operaciones que generan estruturas de datos a modo de Cubos Olap donde se realiza una agrupación de los valores bajo alguna variable que permita agrupar para luego resumir otras variables por medio de agregación de datos (sumas, cuentas, promedios, etc.).
2. **Attribute Subset Selection:** consiste en seleccionar sólo los atributos que permiten obtener el conocimiento buscado con el análisis, donde las variables que generan ruido, son redundantes o tienen poca relevancia son eliminadas del análisis en cuestión.
3. **Dimensionality reduction:** consiste en reducir la dimensionalidad mediante mecanismos de codificación (encoding) que disminuyen la cantidad de variables como las técnicas de análisis discrete wavelet transform y principal component.
4. **Numerosity reduction:** consiste en reemplazar o estimar la base de datos existente por una representación menor de los datos obtenida mediantes métodos paramétricos, que solo necesitan almacenar los parámetros del modelo en vez de los datos disponibles, así como también métodos no paramétricos como sampling, clustering o histogramas.
5. **Discretization y Concept Hierarchy Generation:** consiste en tomar los datos actuales y llevarlos a nuevos rangos de valores o a niveles más altos en una visión jerárquica. Esta es una manera de reducir en gran medida la cantidad de datos en el repositorio. Las técnicas de Discretización puede ser utilizadas para reducir el número de valores contínuos mediante la división del atributo en una escala de rangos o intervalos, lo cual genera mejoras en la comprensión y simplificación de la data original. Éstas técnicas pueden ser categorizadas en supervisadas, si existe una clase de referencia y no supervisadas, en los otros casos. Por su parte, las jerarquías pueden reducir datos mediante la creación de estructuras conceptuales jerárquicas que vayan recolectando y reemplazando de acuerdo a Niveles Altos y Bajos.

Luego, y en particular para las técnicas de selección de atributos veremos disntos métodos en detalle en la sección 4.

Alternativas que se pueden considerar para el procesamiento de variables cuantitativas continuas y variables cuantitativas discretas.

En general, existe un tratamiento distinto para las variables cualitativas ordinales, ya que éstas presentan una relación de orden que es necesario preservar cada vez que se aplica algún tipo de transformación en las variables. Generalmente, es necesario asignar un valor numérico a variables del tipo categórico el cual es elegido en forma arbitraria sin ningún método particular, lo cual en las variables categóricas ordinales es un riesgo ya que puede destruir la estructura interna de los datos ya que poseen un orden preestablecido que debe ser tomado en cuenta.

Como hemos visto en los puntos anteriores, existen varios métodos para el procesamiento de variables, para el caso particular de las variables cualitativas podemos tomar las siguientes acciones:

I. Cualitativas Discretas:

- No realizar ningún cambio
- Re-codificar los valores que toma la variable. Por ejemplo, aquellas variables de la base de datos de entrenamiento que toman dos valores como en el caso de la base retail de si posee la misma dirección de entrega se reemplazan los valores por 0 y 1.
- Si los posibles valores que puede tomar una variable son reducidos, se podrían crear variables dummies, como se realizaco con las variables que representaban los tipos de tarjetas en la base de datos de Retail.
- Se puede considerar el disminuir la cantidad de valores que puede tomar la variable. Esto se lograría agrupándolos según algún criterio como el de generación de jerarquías.

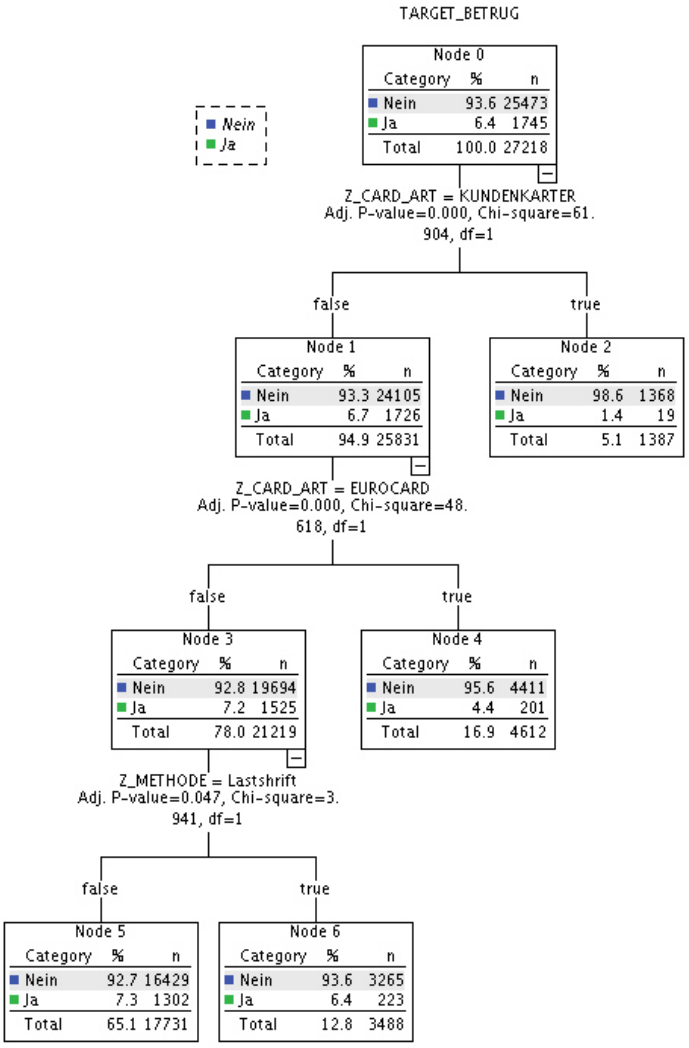
II. Cualitativas Contínuas u Ordinales:

- Se puede realizar recodificación disminuyendo los rangos posibles y agrupando rangos contiguos. Por ejemplo, separar la escala de una variable dejándola en sólo valores como Bajo, Medio y Alto.
- Crear variables dummies.

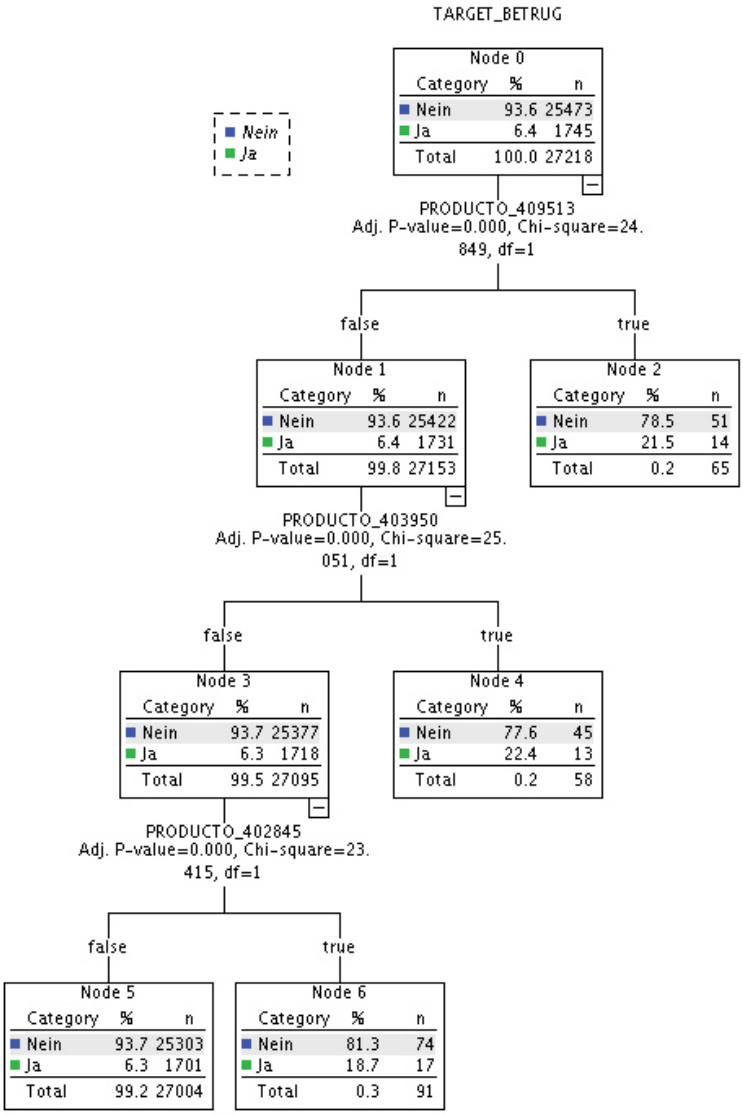
Aplicación en la base de datos Retail

De acuerdo al las técnicas ya mencionadas y su aplicación a la base de datos de Retail, nos encontramos con un problema de alta cantidad de datos y con una gran cantidad de columnas (52) con las transformaciones ya realizadas (ver sección 2.d).

Por ello en primer lugar hicimos un análisis de las variables categóricas y decidimos que las variables categóricas ordinales con muchas categóricas serían tratadas como variables continuas normalizando sus valores entre 0 y 1. Las variables que cumplen esta regla son: ZCV_MONAT y ZCV_JAHR. Luego nos encontramos con que las variables Z_METHODE y Z_CART_ART están relacionadas de manera jerárquica donde la segunda corresponde a la clasificación de las tarjetas de crédito, ya sea Kundenkarte o Kreditkarte. Debido a ello decidimos eliminar alguna de las categóricas con dependencia, para lo cual evaluamos la importancia de la dependencia mediante un árbol de decisión CHAID lanzando por SPSS considerando las 8 categorías ligadas al tipo de pago. De acuerdo a ello, las categorías KUNDENKARTER, LASTSHRIFT y EUROCARD entregan información valiosa para encontrar a la variable objetivo, por lo que la decisión tomada fue eliminar las columnas KREDITKARTER y KUNDENKARTER de manera que ellas se encuentra intrínsecamente y en más detalle en las categorías dadas por Z_CART_ART y no se pierde la información de las otras dos categorías de Z_METHOD.



A su vez analizamos las variables categóricas creadas al comienzo en relación a los productos, ya que la transformación se realizó en primera instancia para poder corregir los datos y ahora es necesario ver si aquella transformación es importante mantenerla para el análisis. Con ello obtuvimos que de los 10 productos incluidos se seleccionan 3 para seguir con el análisis PRODUCT_409513, PRODUCT_403950 y PRODUCT_402845, los demás fueron eliminados.



El resto de las variables categóricas fue transformada en dummys como ya vimos en la sección 1.d mientras que las variables continuas fueron normalizadas mediante la técnica Z-transformation implementada en la herramienta Rapidminer.

Con ello la base de datos quedó con 42 columnas y 27218 registros. A continuación seguiremos el proceso de Reducción de Variables de acuerdo a las técnicas estudiadas y descritas con mayor detalle.

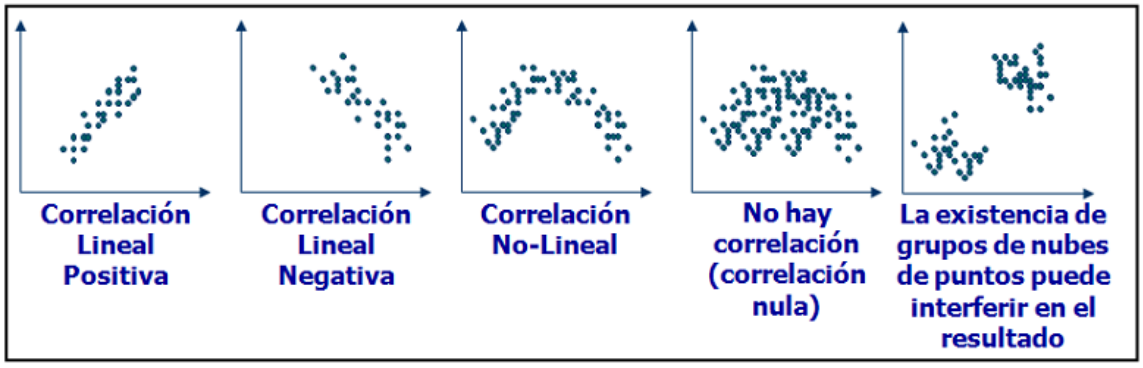
Pregunta 4 : Selección de Atributos y Extracción de Atributos

A continuación se detallaran estrategias de selección de atributos utilizando cada una de las siguientes técnicas:

- Análisis de correlación
- Tablas de Contingencia (CrossTabs)
- Test χ^2
- Test ANOVA
- Information Gain
- Gini Index
- Árboles de Decisión
- Fordward, Backward Feature Selection
- V de Cramer
- Criterio de Información de Akaike
- Coeficiente de Contingencia

Análisis de correlación

En este caso se realiza un test de correlación entre las variables independientes y la variable objetivo, una agrupación adecuada de éstas permite observar patrones de semejanza entre las variables independiente como respecto a la variable objetivo.



Esto permite, eventualmente, reducir la dimensión al eliminar variables independientes que presenten alta correlación con otras. Empíricamente, es recomendable eliminar variables independientes que tengan correlación sobre 0,75 entre ellas y en el caso de la correlación entre alguna variable independiente y la variable objetivo, se recomienda eliminar aquellas que tengan poca correlación con la variable objetivo.

Tablas de Contingencia

En el caso de variables categóricas, nominales u ordinales, es posible aplicar esta estrategia cuando estas variables toman dos valores posibles, o sólo uno, frente a la variable objetivo del problema que se trata como dummy. Para el análisis se toma una muestra aleatoria de observaciones de las variables independientes (de los tipos señalados) para una cantidad representativa de ambas realizaciones de la variable objetivo, digamos -1 y 1 (ya que se consideran dummys), junto al valor tomado por la variable objetivo y realizar tablas de contingencia.

El objetivo de estas tablas es que al observar los totales, ya sea cantidades de casos o frecuencias marginales, estadísticamente se puede comprobar si existe o no asociación entre las variables mediante un test Chi-Cuadrado de Pearson.

La tabla de contingencia organiza los datos de siguiente manera:

		VAR B					
		B_1	B_2	B_3	…	B_c	Total
VAR A	A_1	O_{11}	O_{12}	O_{13}		O_{1c}	R_1
	A_2	O_{21}	O_{22}	O_{23}		O_{2c}	R_2
	A_3	O_{31}	O_{32}	O_{33}		O_{3c}	R_3
	…					…	
	A_r	O_{r1}	Or_2	O_{r3}	…	O_{rc}	R_r
	Total	C_1	C_2	C_3	…	C_c	n

Donde:

- Ai y Bj : representan las categorías de la variable A y B respectivamente.
- Oij: es el número de casos que tienen las características Ai y Bj a la vez.
- Ri (i = 1,…,r): es la suma de la i-ésima fila de la tabla característica Bi.
- Cj (j = 1,…,c): es la suma de la j-ésima columna de la característica Ai.
- n: representa el total de observaciones tomadas.

Luego, A y B serán independientes si cada entrada de la tabla es igual al producto de los totales marginales dividido entre el número de datos. Esta estrategia es útil para los casos en que una variable independiente toma dos valores posibles, o sólo uno. Si una variable categórica presenta una cantidad numerosa de posibles valores, el análisis se vuelve más complejo.

Test Chi-Cuadrado

El test Chi-Cuadrado puede ser empleado para medir la asociación entre variables según un nivel de significancia estadística. Este test es aplicable sólo sobre variables categóricas (un test similar aplicable a variables numéricas es el test de Kolmogorov-Smmirnov). En el test se generan dos hipótesis de trabajo, H_0 = Las variables son independientes y H_1 = No se cumple H_0; normalmente la hipótesis nula se rechaza de acuerdo al nivel de significancia escogido del 5% o el 1% dependiendo del caso. Si se rechaza H_0 en la práctica implica que se puede elegir descartar ciertas variables independientes que tienen alta asociación con otras variables independientes con el fin de evitar sobre ajuste, o bien para descartar variables independientes que no estén asociadas a la variable objetivo.

Test ANOVA

El test ANOVA (ANalysis Of VAriance) permite analizar la relación entre la variable objetivo y variables independientes del tipo 0-1. Para poder aplicar de manera correcta el test es necesario que se den dos condiciones: que los datos tengan distribución Normal y que las varianzas de ellos sean significativamente diferentes. La dificultad de su aplicación radica en la comprobación, mediante tests estadísticos, del cumplimiento de las condiciones a fin de evitar resultados erróneos.

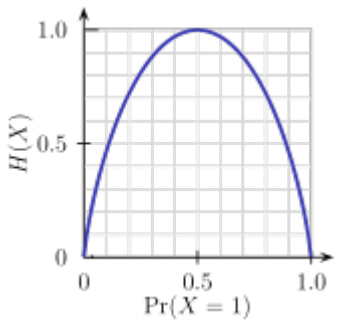
Information Gain (Andrew, 2003)

Para comprender el concepto de Information Gain, es necesario introducir el concepto de Entropía. La Entropía es la medida de "desorden" atribuible a una variable X. Supongamos que nuestra variable X presenta la siguiente distribución:

$\Pr(X = v_1) = p_1$	$\Pr(X = v_2) = p_2$	...	$\Pr(X = v_n) = p_n$

Entonces, se define la Entropía promedio como: $H(X) = -\sum_{j=1}^{m} p_j \times log_2(p_j)$

Donde, "Alta Entropía" significa que la variable X tiene una distribución sin mucho desorden; si viéramos la distribución de la variable en un histograma se vería un gráfico plano, y "Baja Entropía" quiere decir que la variable toma muy variados resultados; el cual en un histograma veríamos tal vez como muchos valores bajos para la variable y uno o dos muy altos. En el recuadro siguiente, se observa la distribución de la entropía para una variable X que se distribuye Bernoulli.



En ella podemos observar que en el caso de una variable de Bernoulli, la entropía máxima ocurre cuando la probabilidad es $\frac{1}{2}$, es decir, todos los resultados son igualmente posibles o desde el punto de vista de la información, la variable toma valores con muchos valores diferentes (o está muy desordenada)

Otro concepto importante es el de Entropía Condicional, para explicarlo usaremos la siguiente tabla de datos, donde la variable X es la materia preferida y la variable Y es su preferencia por la comida China.

X	Y
Matemáticas	Si
Historia	No
Ciencias	Si
Matemáticas	No
Matemáticas	No
Ciencias	Si
Historia	No
Matemáticas	Si

En la tabla anterior podemos ver que las Entropías para las variables X e Y son: $H(X) = 1{,}5$ y $H(Y) = 1{,}0$. Dicho antecedente sólo nos indica que hay más "desorden" en la variable X que en Y.

Ahora, la Entropía Condicional Específica nos permite medir el nivel de entropía de una variable Y, para aquellos registros donde otra variable X toma un valor particular x, es decir, $H(Y|X = x)$. Según los datos de la tabla anterior:

$$H(Y|X = Matematicas) = 1$$

$$H(Y|X = Historia) = 0$$

$$H(Y|X = Ciencias) = 0$$

Entropía Condicional Promedio

Se define la Entropía Condicional Promedio de la variable Y con respecto a la variable X por:

$$X(Y|X) \stackrel{\text{def}}{=} \sum_{x \in X} \Pr(X = x) \cdot H(Y|X = x)$$

$$= -\sum_{x \in X} \Pr(X = x) \cdot \sum_{y \in Y} \Pr(\mathrm{Y = y|X = x}) \cdot log_2(\Pr(\mathrm{Y = y|X = x}))$$

$$= -\sum_{x \in X} \sum_{y \in Y} \Pr(\mathrm{X = x; Y = y}) \cdot log_2(\Pr(\mathrm{Y = y|X = x}))$$

$$= -\sum_{x \epsilon X; y \epsilon Y} \Pr\,(X = x; Y = y) \cdot log_2(\Pr(Y = y|X = x))$$

$$= -\sum_{x \in X; y \in Y} \Pr(X = x; Y = y) \cdot log_2\,(\frac{\Pr(X=x;Y=y)}{\Pr(X=x)})$$

En nuestro ejemplo podemos calcular la entropía condicional promedio de la variable Y con respecto a la variable X, usando la siguiente tabla:

x_j	$\Pr(X = x_j)$	$H(Y\|X = x_j)$
Matemáticas	0,5	1
Historia	0,25	0
Ciencias	0,25	0

Luego $H(Y|X) = 0{,}5\times1 + 0{,}25\times0 + 0{,}25\times0 = 0{,}5$

Ganancia de Información

La ganancia de información se determina por $I = H(X) - H(X|Y)$.

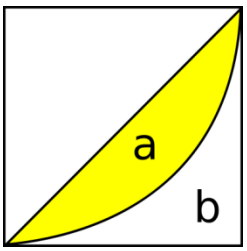
Desde el punto de vista de la selección de atributos lo que se hace es calcular la Ganancia de Información para toda variable (dada la variable objetivo), y ordenarlos en forma decreciente, de esta forma se seleccionan los K primero atributos.

Gini Index

Se utiliza para evaluar la distribución de una variable independiente de tipo numérica respecto a la muestra que se tiene de la variable dependiente y de éste modo aportar antecedentes sobre el equilibrio de la distribución de las variables independientes examinadas respecto a la variable objetivo, lo cual es útil para determinar si las variables independientes examinadas tienen más o menos relación con el valor 1 o -1 que puede tomar nuestra variable objetivo. El índice de Gini es el valor del coeficiente de Gini expresado en porcentaje, que toma calores entre 0 y 1 y se define como:

$$G = \left| 1 - \sum_{i=1}^{i=n-1} (X_{i+1} - X_i)\times(Y_{i+1} - Y_i) \right|$$

Donde X e Y corresponden a la proporción acumulada de la variable independiente y dependiente respectivamente. Está fuertemente asociada como la razón $\frac{a}{a+b}$ entre las áreas de la curva de Lorentz.



El índice de Gini puede utilizarse como una manera alternativa a Information Gain para seleccionar atributos (Robert, 2008).

Al realizar un árbol de decisiones con estos datos, el índice de Gini para un nodo D está dado por.

$$\text{Gini}(D)=1-\sum_{j} p_j^2$$

Donde p_j son las frecuencias relativas de la clase j en el nodo D. De la misma forma que Information Gain, el Indice de Gini se puede calcular para cada una de las variables y se ordenan en forma descendente, y se eligen los K primero atributos.

Árboles de Decisión

Los Árboles de Decisión son una técnica perteneciente al enfoque de aprendizaje supervisado, es decir, donde existe un patrón conocido dependiente de un conjunto de variables explicativas y se requiere conocer la relación entre dichas variables explicativas y la variable objetivo.

El Árbol de Decisión se caracteriza porque su construcción representa un árbol donde cada nivel agrupa nodos de valores de acuerdo a los atributos que se seleccionan; dichos atributos comúnmente son escogidos utilizando técnicas como Entropía relativa, Índice de Gini, Test Chi-Cuadrado, etc., según la implementación utilizada. Existen varias implementaciones de Árboles de Decisión, algunas de las más conocidas son ID3, C4.5 y CART.

Los Árboles de Decisión tienen la particularidad en que son fáciles de construir y simples de entender, admiten tanto atributos discretos como continuos (dependiendo de la implementación) y no tienen problemas con el manejo de missing values o atributos no significativos. Los árboles de decisión son fáciles de usar, algunos algoritmos admiten atributos discretos y continuos, y tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación.

Algunos de los algoritmos existentes para los arboles de decisión, son CART, CHAID, ID3 para atributos discretos y , C4.5 para atributos tanto discretos como continuos. El método más conocido para crearse árboles de decisión es el CHAID (Chi-square Automatic Interaction Detection) que examina la relación entre muchas variables categóricas o discretas y un objetivo categórico o medida de resultado. El criterio para elegir los atributos en ID3 y C4.5 es el "incremento de información" o "information gain", definiendo el concepto Information gain como la reducción de entropía como consecuencia de realizar una división de los datos. En otras palabras, el atributo que permita obtener la mayor ganancia de información será el seleccionado para dividir el nodo.

Fordward and Backward Feature Selection

Es una técnica para seleccionar atributos que se enmarca dentro de los métodos Wrapper, es decir, opera en conjunto con la técnica de Data Mining empleada.

La técnica Fordward se basa en un rankeo inicial (bajo algún criterio, por ejemplo ganancia de información, Test Chi-Cuadrado, Grado de Correlación, etc.) de las variables independientes respecto de la variable dependiente en forma descendiente. Se elige un grupo top de las variables de acuerdo al rankeo y ellas son utilizadas a través de una técnica de Data Mining para luego evaluar su efectividad.

La técnica Backward es justamente lo contrario a la anterior, es decir, va quitando variables que utiliza el modelo, de acuerdo a alguna regla con la finalidad de sustraer aquellas variables que no aportan al modelo, por el contrario, sólo generan ruido como mayor frecuencia de aparición en los errores tipo I y II.

Normalmente estos métodos se utilizan juntos de manera alternada, lo que permite obtener mejores resultados. Además a fin de realizar una medición de la efectividad de las variables que se van eligiendo y desechando en el modelo, es recomendable realizar una partición de la base de datos en dos, una de entrenamiento (training) y otra de testeo (test).

V de Crammer

Corresponde a una alternativa simple al Test Chi-Cuadrado para determinar el grado de asociación entre variables. Para su cálculo se utiliza la siguiente fórmula:

$$V = \sqrt{\frac{\chi^2}{N \times (k-1)}}$$

Donde :

- k es el mínimo entre el número de filas y el número de columnas
- N es el número total de observaciones

Además $0 \leq \mathrm{V} \leq 1$, el valor 1 indica asociación perfecta.

En la aplicación a la selección y extracción de atributos permite distinguir aquellas variables independientes que tienen alta asociación entre ellas a fin de descartar alguna(s) de ella(s), para evitar el sobre ajuste. Este coeficiente está acotado entre 0 y 1, y puede alcanzar ambas cotas, por lo tanto es el mejor de las medidas de asociación, por ser más fácil de interpretar.

Criterio de Información de Akaike

Es una medida de nivel de ajuste de un modelo estadístico estimado basada en el concepto de entropía. Entrega una medida relativa de información perdida del modelo utilizado, de alguna manera describe el trade off entre el nivel del ajuste y la varianza en el modelo construido. Sirve para determinar el conjunto óptimo de variables explicativas y cuando presumiblemente una de ellas es el resultado de un conjunto de las variables explicativas. En nuestro caso serviría para determinar el conjunto óptimo de variables independientes categóricas que explican la variable objetivo. El procedimiento consiste en tomar un número inicial de variables categóricas y calcular el estadístico Akaike para cada uno de los pares "variable independiente-variable objetivo" que se tengan en el grupo inicial de variables considerado y con ello realizar un ranking en orden decreciente respecto a los valores del estadístico Akaike que se obtenga para cada par. A menor valor del estadístico más explicativa es la variable.

Coeficiente de Contingencia

El coeficiente de contingencia permite medir el grado de asociación entre variables cualitativas nominales que presentan más de una categoría, es decir, sirve como una alternativa cuando no es posible utilizar el Test Chi-Cuadrado debido a que alguna de las variables presenta más de dos categorías.

Además $0 \leq C \leq 1$, el valor 1 indica una alta asociación.

En la aplicación a la selección y extracción de atributos permite distinguir aquellas variables independientes que tienen alta asociación entre ellas a fin de descartar alguna(s) de ella(s).

Técnica Kernel-PCA para la extracción de atributos

Antes de explicar esta técnica es necesario saber en qué consiste técnica **PCA** por sí sola**.** PCA es una técnica de representación de datos que permite la reducción del número de variables extrayendo nuevas variables a partir de las originales. La técnica es capaz de realizar la reducción el espacio de atributos con la menor pérdida de información posible, lo que se consigue a través de la representación de los datos proyectándolos en ejes ortonormales (ortogonales de norma 1), dicha proyección se conoce como componentes principales, que se obtienen al realizar combinaciones lineales con las variables originales.

La técnica tiene dos supuestos básicos para su operación:

- Los datos de cada variable se distribuyen en forma Gaussiana;
- Los datos independientes e idénticamente distribuidos.

Un ejemplo introductorio es considerar una persona que debe reportar las medidas de calzado. Claramente existe un gran número de variables que el podría considerar, pero si miramos por

ejemplo un grafo en el cual presenta el largo del calzado y el ancho de este, podemos ver que ambas variables están altamente correlacionadas.

De esta forma, parece posible describir el número de calzado a través de una sólo variable que captura gran parte de la información de manera que al proyectar esta nueva variable sea factible obtener la información original.

El problema se plantea considerando la existencia de *p* variables y cada una de ellas con *n* valores. Para cada una de las variables originales se determina su valor promedio y éste se resta a cada uno de los valores en los datos. Observaciones centradas en el valor promedio (es decir, $\sum_{j}^{p} x'_j = 0$)

Luego podemos definir una matriz X de datos centrados (en la media), de dimensiones *n x p*. El problema se puede solucionar obteniendo un vector z_1 que corresponde a uno de los ejes donde se proyecta nuestra matriz X, supongamos por ahora que el vector v_1 es un vector de dimensión *px1*, cuyas componentes son los pesos de la proyección sobre z_1. Así la primera componente principal z_1 se obtiene de:

$$z_1 = Xv_1$$

Esta componente principal tiene las siguientes propiedades:

- Su media es cero;
- Su varianza es $Var(z_1) = \frac{1}{n} z_1^t z_1 = \frac{1}{n} v_1^t X^t X v_1 = v_1^t C v_1$

Donde $C = \frac{1}{n} X^t X = \frac{1}{n} \sum_i^n x_i x_i^t$ es la matriz de Varianza-Covarianza de la matriz X de datos centrados.

Como lo que se desea es que esta transformación genere la menor perdida de información, se puede definir el objetivo como que las componentes principales deben lograr maximizar la varianza.

Adicionalmente, a este vector de ponderación v_1 le exigiremos que tenga norma 1, es decir: $v_1^t v_1 = 1$.

De esta forma nuestro problema de optimización queda plateado por:

Max $v_1^t C v_1$ s.a. $v_1^t v_1 = 1$

Para resolverlo, aplicaremos Lagrange, de forma que la función objetivo relajada queda:

$$F = v_1^t C v_1 - \lambda\left(v_1^t v_1 - 1\right)$$

Para resolver, derivamos:

$$\frac{\partial F}{\partial v_1} = 2Cv_1 - 2\lambda v_1 = 0 \Rightarrow Cv_1 = \lambda v_1$$

Si generalizamos el resultado anterior para todos los componentes principales, se tiene que: $\lambda V = CV$, donde V es la matriz compuesta por los componentes principales $V = \left[v_1, v_{2,} \ldots\right]$ y λ son los valores propios asociados a dichos vectores propios.

Una forma simple de obtener lo anterior es diagonalizar la matriz de varianza-covarianza C, y cada columna se descompone en su valor propio por un vector que forma una base de manera que tenga norma 1. Como todos las columnas forman una base y tienen norma uno, estos vectores son ortonormales. En la práctica lo anterior se logra estandarizando las variables antes de crear la matriz X.

Kernel PCA

Las técnicas Kernel sustituyen el espacio original de las observaciones X, que en general pertenecen a $\mathbb{R}^p$, por un espacio provisto de un operador de producto punto H mapeado a través de Φ, es decir:

$$\Phi: X \longrightarrow H$$

Partiendo de la misma forma que en el caso del PCA, tenemos que ahora los datos están centrados en H, llegamos a una nueva matriz de varianza-covarianza en el nuevo espacio.

$$C_H = \frac{1}{n}[\Phi(X)]^t[\Phi(X)] = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \Phi(\mathrm{x_i})\Phi\left(\mathrm{x_i^t}\right)$$

En este caso hay que encontrar valores propios no nulos ($\lambda > 0$) con sus respectivos vectores propios v_{II} que deben satisfacer un problema de optimización similar al del PCA, es decir:

$$\lambda V_H = C_H V_H$$

Por lo tanto para poder aplicar este método es necesario calcular la nueva matriz de varianza-covarianza, según la ecuación:

$$C_H = \frac{1}{n}[\Phi(X)]^t[\Phi(X)] = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \Phi(\mathrm{x_i})\Phi\left(\mathrm{x_i^t}\right)$$

Sin embargo, este método resulta un poco complicado y no elimina el problema de la centralización de los datos, para resolver esto, se define la matriz K de la siguiente forma:

$\mathrm{K} = \left(k_{ij}\right), \quad i \in \{1,2,..,n\}; \quad j \in \{1,2,\dots,p\}$ donde: $k_{ij} = \langle \Phi(x_i), \Phi\left(x_j\right)\rangle; \ \forall i; \ \forall j$

De esta forma, se puede escribir $C_H = \mathrm{K}^t\mathrm{K}$, con esta matriz se realiza de manera implícita la aplicación del operador Φ, además se pude resolver el problema de la centralización de los datos a través de reemplazar a la matriz K por su correspondiente versión centralizada:

$$\widetilde{\mathrm{K}} = \mathrm{K} - 1_n\mathrm{K} - \mathrm{K}1_n + 1_n\mathrm{K}1_n$$

Donde, $1_n = \left(\left(\frac{1}{n}\right)_{ij}\right)$ es una matriz cuadrada de $n \times n$ con elementos $\frac{1}{n}$.

La gran utilidad de la matriz K es que no es necesario hacer explícito el mapeo Φ, sino que en lugar de ello empleamos una función $\kappa: X \times X \to \mathbb{R}$, la cual representa al producto punto en H, es decir, $\kappa(x, x^t) = \langle \Phi(x), \Phi(x^t)\rangle$. Algunos Kernels comúnmente utilizados son:

Kernel Lineal: Corresponde al producto punto en el espacio inicial

$$\kappa(x, x^t) = \langle x, x^t \rangle$$

Kernel Polinomial : Representa la expansión a todas las combinaciones de monomios de orden *d*, y está dado por

$$\kappa(x, x^t) = \langle x, x^t \rangle^d$$

Kernel Gaussiano: Este kernel corresponde a una de las funciones radiales

$$\kappa(x, x^t) = exp\left(-\frac{\|x - x^t\|^2}{2\sigma^2}\right)$$

El Kernel-PCA tiene la misma aplicación en la extracción de atributos que el PCA, con la ventaja adicional que no requiere que se cumpla la condición de linealidad entre los atributos, es decir es una derivación no lineal de la técnica PCA.

La técnica de descomposición matricial SVD y su relación con PCA

La técnica singular value descomposition SVD presenta propiedades interesantes para el área de Data Mining en el sentido que entrega mayor información respecto de los datos, y que además "ordena" la información contenida en los datos de manera que, en términos simples, la "parte dominante" de la información se hace visible. (Eldén, 2007)

Las propiedades de aproximación presentes en la técnica SVD pueden ser usadas para elucidar la equivalencia entre SVD y PCA. Asumamos que $X \epsilon \mathbb{R}^{m \times n}$ es una matriz de datos, donde en este caso cada columna es una observación. Se asume que la matriz está centrada, es decir, el promedio de cada columna es cero.

Sea la transformación SVD de la matriz $X = U\Sigma V^t$, donde la matrices $U \in \mathbb{R}^{m\times m}$ y $V \in \mathbb{R}^{n\times n}$ son ortogonales y $\Sigma \in \mathbb{R}^{n\times n}$ es una matriz diagonal, es decir, $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_n)$, tal que $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$. Por otro lado, las columnas de las matrices U y V son llamados vectores singulares (el equivalente de los vectores propios con la particularidad de tener norma 1) y los elementos de Σ son denominados valores singulares.

Dada la transformación anterior $X = U\Sigma V^t$ se tiene que los vectores singulares de la derecha v_i son denominados *Principal Components Directions* de X. Luego el vector $z_1 = Xv_1 = \sigma_1 u_1$

Tiene la mayor varianza entre todas las combinaciones lineales de las columnas de X:

$$Var(z_1) = Var(Xv_1) = \frac{1}{m}\sigma_1^2$$

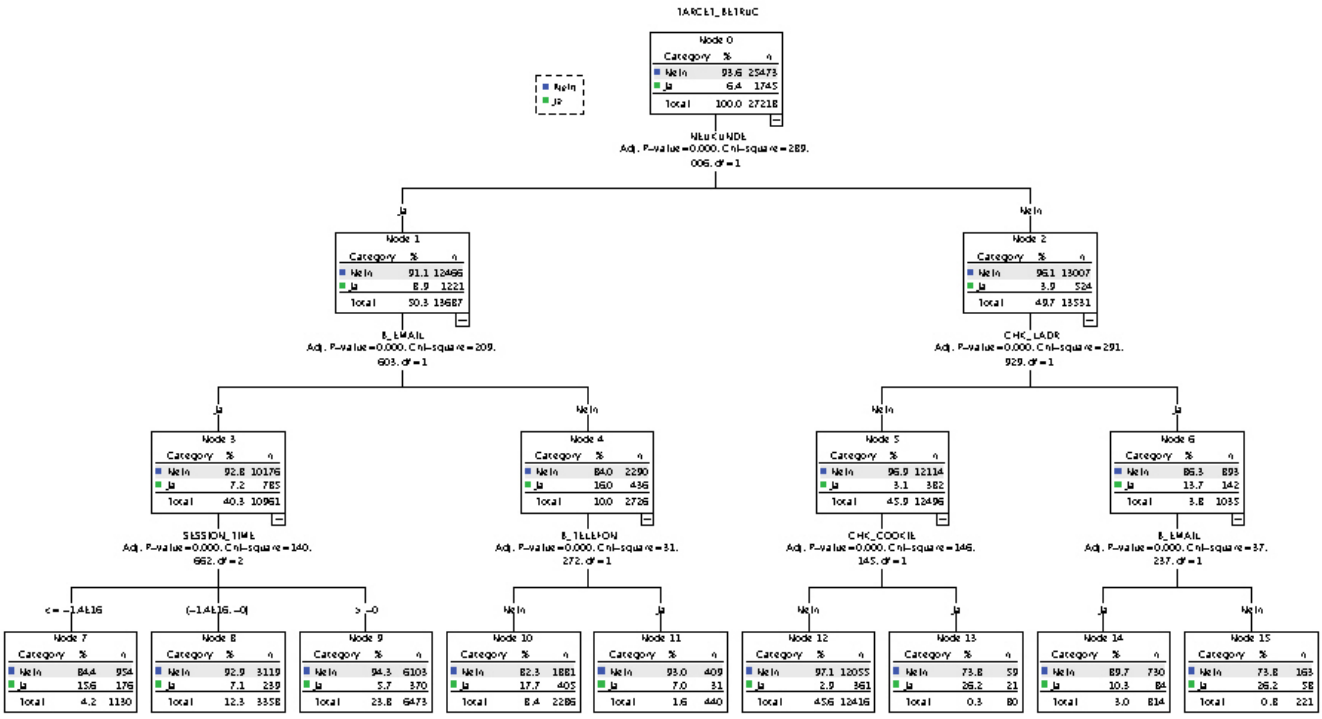
La variable normalizada $u_1 = \left({}^{1}/_{\sigma_1}\right)Xv_1$ es conocido como el Primer Componente Principal Normalizado de X.

Habiendo determinado el vector con la mayor varianza muestral z_1, se obtiene el segundo vector que es ortogonal al primero calculando el vector de mayor varianza sobre la matriz reducida $X - \sigma_1 u_1 v_1^t$. De esta manera se obtienen los siguientes componentes principales en forma ordenada.

Selección de atributos en la base de datos Retail

En primer lugar se evaluó qué técnicas podrían utilizares, y de acuerdo a ellos se decidió analizar de acuerdo a tres métodos: Árbol de decisión, componentes principales y Kernel PCA.

El primer método ejecutado debido a su simplicidad y capacidad explicativa fue el de Árbol de Decisión. El tipo de árbol seleccionado fue del tipo CHAID. El resultado obtenido por el árbol indica que la variable que genera mayor separación en primer lugar es NEUKUNDE, luego B_MAIL y CHK_LADR y, posteriormente, SESSION_TIME,B_TELEFON,CHK_COOKIE y B_MAIL como la figura:



Este árbol nos da intuición sobre el comportamiento de los clientes que comenten fraude. Considerando las casillas donde aumenta la probabildad de descubrir un fraude, tenemos por un lado que aumenta la probabilidad de fraude si son clientes nuevos. Si son antiguos, aumenta la probabilidad de fraude cuando se ve que el cliente hace reiteradas órdenes dentro de 3 días en la misma dirección ya sea física o virtual (CHK_LADER y CHK_COOKIE), y si es física, los fraudes no agregan un correo de contacto junto con la órden.

Luego, se realizó un análisis por componentes principales, donde de acuerdo a las comunalidades de extracción se fueron seleccionando variables iterativamente de forma manual a medida que mejorara el nivel de ajuste del modelo. El criterio de selección fue una comunalidad de extracción >0,5 en las dos primeras iteraciones y luego se fue probando con 0,6 y 0,5. Se testeó mediante el test de ajuste KMO y el de esfericidad de Barlett, se colocó la variable TARGET_BETRUG como filtro del análisis priorizando el valor de los Fraudes.

Variables descartadas:

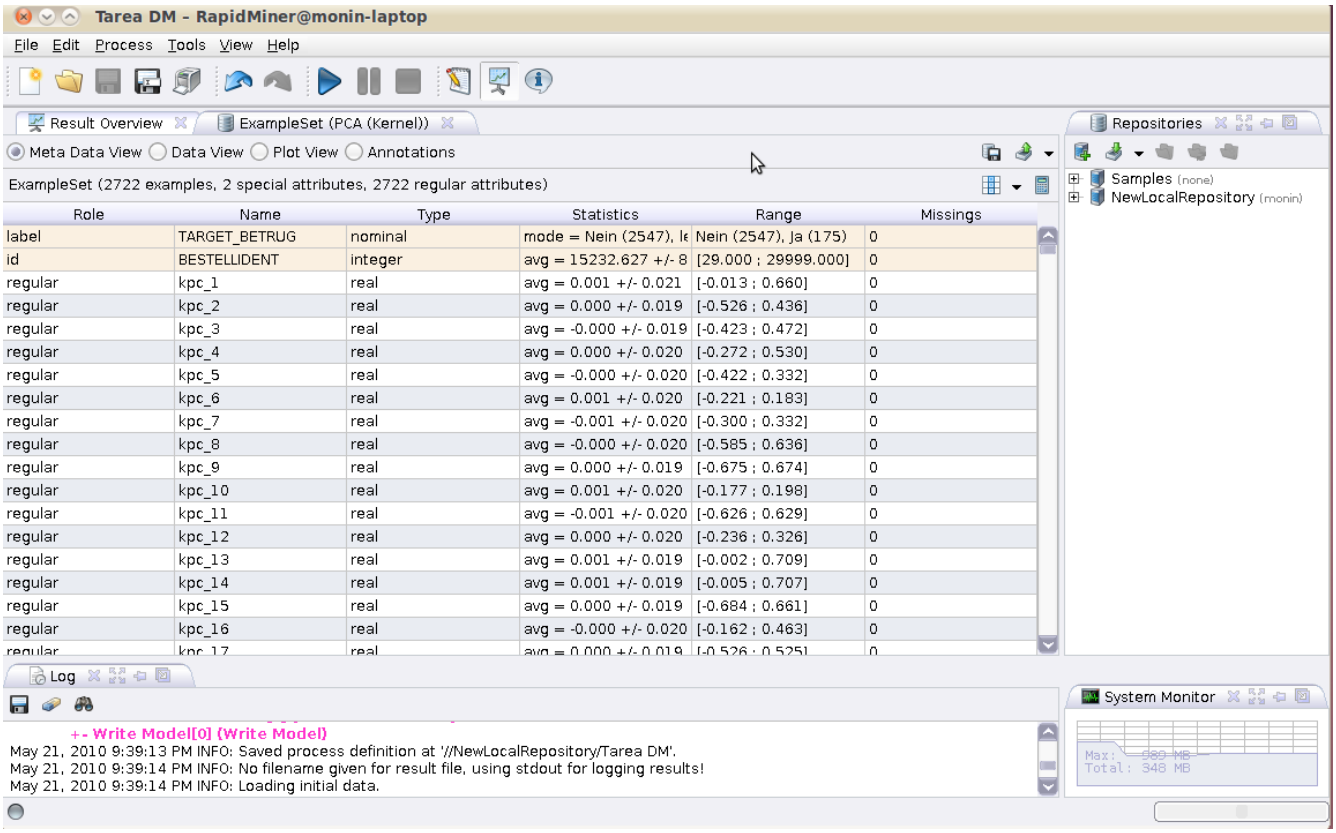
- Iteración 1: CHK_KTO, CHK_IP,FAIL_LPLZ, FAIL_RPLZORTMATCH, PRODUCTO_403950
- Iteración 2: PRODUCTO_402845.
- Iteración 3: TAG_BEST, ANZ_BEST_GES, CHK_LADR, CHK_RADR, CHK_CARD, FAIL_LORT, FAIL_LPLZORTMATCH.
- Iteración 4: Z_JAR.
- Iteración 5: FAIL_RORT.
- Iteración 6: TIME_BEST.

Communalities[a]

	Initial	Extraction
ALTER	1,000	,608
ZCV_MONAT	1,000	,591
WERT_BEST	1,000	,715
SESSION_TIME	1,000	,573
Z_CARD_ART = VISA	1,000	,944
Z_CARD_ART = EUROCARD	1,000	,947
Z_CARD_ART = KUNDENKARTER	1,000	,683
Z_CARD_ART = AMEX	1,000	,900
B_EMAIL	1,000	,568
B_TELEFON	1,000	,660
FLAG_LRIDENTISCH	1,000	,788
FLAG_NEWSLETTER	1,000	,686
CHK_COOKIE	1,000	,722
FAIL_LPLZ	1,000	,734
NEUKUNDE	1,000	,632
PRODUCTO_409513	1,000	,559
Z_METHODE = Rechnung	1,000	,930
Z_METHODE = Lastshrift	1,000	,926
MAHN_AKT = MAHN_0	1,000	,868
MAHN_AKT = MAHN_1	1,000	,833
MAHN_AKT = MAHN_3	1,000	,898
MAHN_AKT = MAHN_2	1,000	,704
MAHN_HOECHST = MAHN_0	1,000	,856
MAHN_HOECHST = MAHN_1	1,000	,810
MAHN_HOECHST = MAHN_2	1,000	,783
MAHN_HOECHST = MAHN_3	1,000	,886
ANZ_BEST	1,000	,720

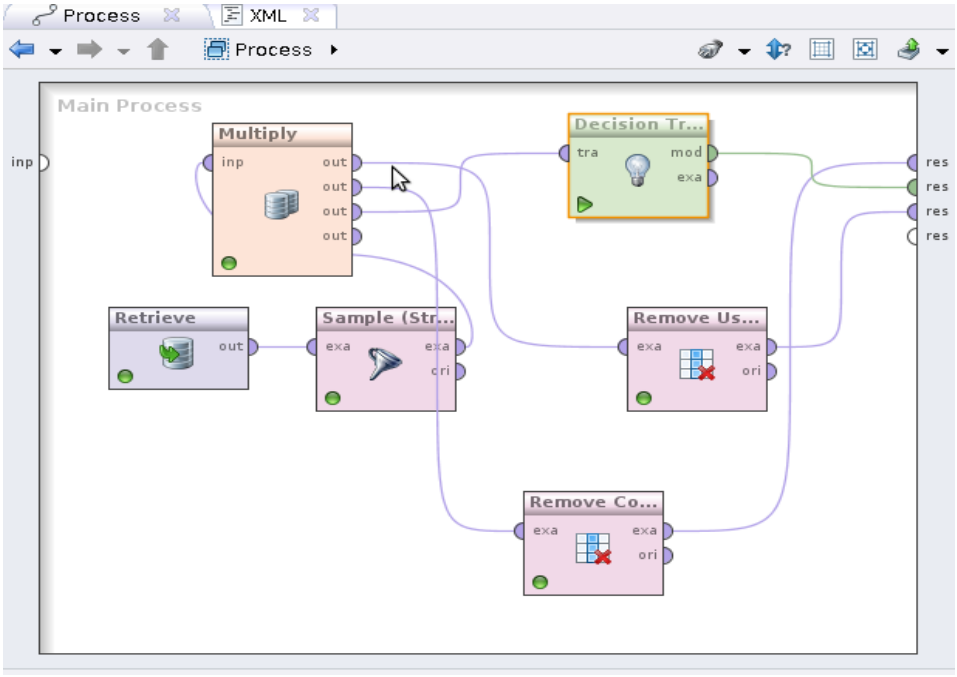
Luego, tomando el Análisis Factorial, se generan 14 factores que explicarían el 76% de la varianza del modelo, lo cual y dado la dificultad de interpretar cada uno de los factores decidimos que la cantidad de atributos entregados es apropiado para trabajar con modelos predictivos y no contradice lo entregado por el árbol, salvo por CHK_LADER que será agregada al análisis y los dos Productos debido a que en el análisis de árbol anterior, ayudaban a predecir con un porcentaje aceptable los fraudes.

Posteriormente, realizamos en RapidMiner el análisis de Kernel PCA, el cual dio la siguiente salida:



De acuero a ello, notamos que el procedimiento de cálculo es mucho más complejo que PCA. En la prática sólo pudimos correr este método con una muestra estratificada al 10% y tomó alrededor de 21 minutos obtener resultados en un computador con sistema operativo Ubuntu Lucid, 2 MB Ram procesador Intel Core Duo. Los rsultados obtenidos no son dan intuición respecto a cada componente por lo que decidimos descartar este método del análisis debido a la difilcultad para interpretarlo y los problemas computacionales a los cuales nos enfrentamos.

Finalmente y de manera exploratoria, decidimos revisar qué variables eliminan los métodos incluidos en la aplicación Rapidminer. Se probó un árbol de decisición, Selección de Variables Inútiles y Variables Correlacionadas.

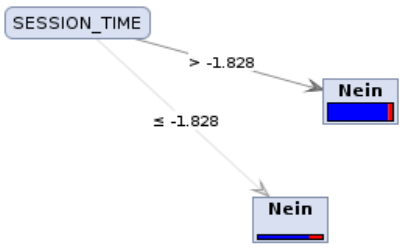


Los resultados obtenidos fueron los siguientes:

Árbol de Decisión: probamos con los 4 criterios disponibles (Information Gain, Gini Index,Gain_Ratio y varias configuraciones de los otros atributos, pero el equipo se quedaba sin memoria para análisis más específicos y también tuvimos problemas con el tipo de datos soportados. Además de los problemas de memoria que impidieron el cálculo de un árbol útil.

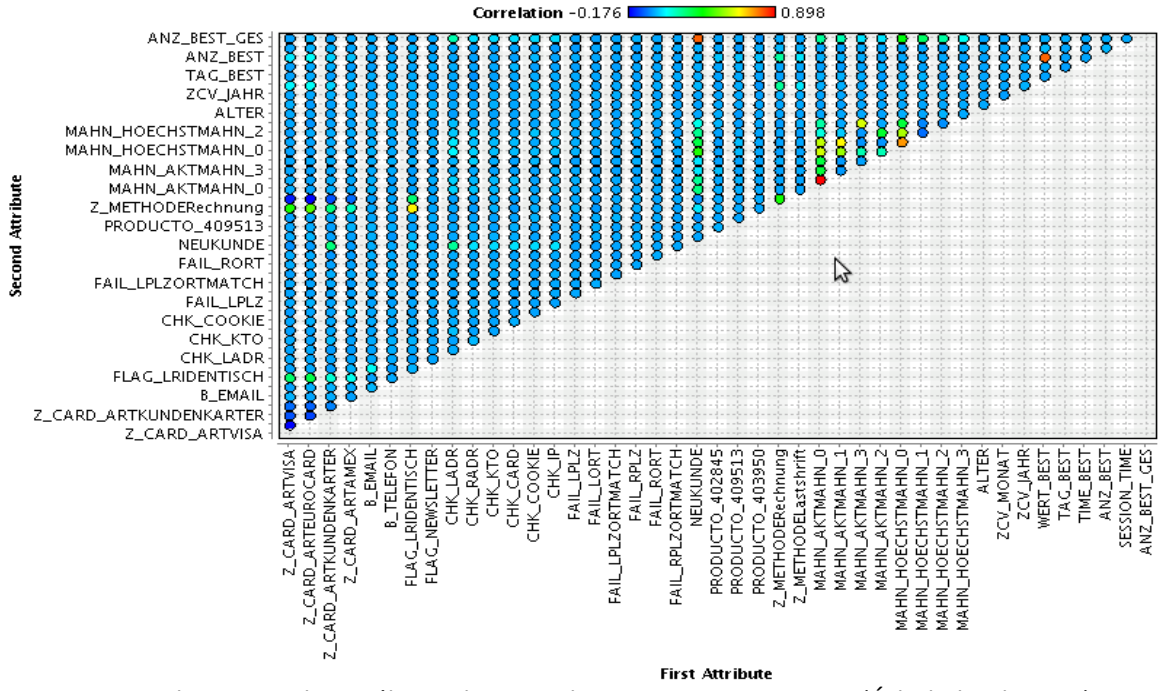
Uno de los resultados que pudimos obtener fue el siguiente:

- SESSION_TIME > -1.828: Nein {Nein=24867, Ja=1600}
- SESSION_TIME ≤ -1.828: Nein {Nein=606, Ja=145}



De lo cual rescatamos que SESSION_TIME puede ser una variable interesante al momento de analizar el problema.

Remove Useless Atributes y Remove Correlated Atributes: con un 0,7 de correlación ninguno de los dos métodos llevaron a cabo eliminaciones. Corriendo la matriz de correlaciones obtuvimos lo siguiente:



Luego, nos quedamos con los análisis seleccionados en primera instancia (Árbol Chaid y PCA), ambos entregados por la aplicación SPSS. Con lo que la base de datos queda con 27.218 registros y 32 variables incluyendo la variable objetivo.

Es importante recalcar que el tiempo en obtener los resultados y la facilidad en el manejo del tipo de datos varía notablemente entre las distintas aplicaciones y los distintos sistemas operativos. Parte de la complejidad de llevar a cabo el análisis fue preparar los datos en el formato soportado por una de las aplicaciones.

También creemos de gran relevancia las opciones para desplegar los resultados que posee cada una de las herramientas, ya que es muy importante entender el comportamiento de los datos en relación al problema y las razones por qué se descarta una u otra variable.

Finalmente, el pre procesamiento y transformación de los datos es una etapa muy compleja y clave para la siguiente fase, ya que es en este proceso donde se logra comprender el problema y donde, además, se adquiere la intuición sobre qué resultados o hipótesis podemos obtener en un análisis más profundo de predicción o clasificación. Paralelamente, si en esta etapa seleccionamos datos o variables poco significantes, todo el análisis posterior también lo será.

Trabajos citados

Allison, P. D. (2001). *Missing Data* (Vol. 136). Thousand Oaks, CA: SAGE.

Andrew, M. (2003). Recuperado el 08 de 05 de 2010, de http://www.autonlab.org/tutorials/infogain.html

Barnard, M. J. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research , 8*, 17-36.

Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques - For Marketing, Sales, and Customer Relationship Management* (Segunda Edición ed.). Indianapolis, United States of America: Wiley Publishing, Inc.

Bramer, M. (2007). *Principles of Data Mining.* London: Springer.

Chawla, N. V., Japkowicz, N., & Ko lcz, A. (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. *Sigkdd Explorations , Volume 6, Issue 1*, 1-6.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), , 39* (1), 1-38.

Eldén, L. (2007). Matrix Methods in Data Mining and Pattern Recognition. En L. Eldén, *Matrix Methods in Data Mining and Pattern Recognition* (pág. 57; 66). SIAM.

Farhangfara, A., Kurganb, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition , 41* (12), 3692-3705.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence* , 37-54.

Meng XL, B. J. (1999). Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. *Statistical Methods in Medical Research , 8* (1), 17-36.

Parr Rud, O. (2001). *Data Mining Cookbook - Modeling Data for Marketing, Risk, and Customer Relationship Management.* New York, United States: John Wiley & Sons, Inc.

Pearson, R. K. (2005). Mining Imperfect Data. En R. K. Pearson, *Mining Imperfect Data* (págs. 2-31; 69-86). SIAM.

Pearson, R. K. (2005). *Mining Imperfect Data: Dealing with Contamination and Incomplete Records.* Philadelphia: SIAM: Society for Industrial and Applied Mathematics.

Pei, J., Kamber, M., & Han, J. (2005). *Data Mining: Concepts and Techniques* (2 edition ed.). Morgan Kaufmann.

Robert, S. (2008). Recuperado el 09 de 05 de 2010, de http://www.csc.liv.ac.uk/~azaroth/courses/current/comp527/lectures/comp527-08.pdf

Rousseeuw, P. J., & Driessen, K. V. (2002). *Computing LTS Regression for Large Data Sets.*

Scheffer, J. (2002). Dealing with Missing Data. *Research Letters in the Information and Mathematical Sciences , 3*, 153-160.

Wagstaff, K. L., & Laidler, V. G. (2005). Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy. *Astronomical Data Analysis Software and Systems XIV P2.1.25* (págs. 1-5). P. L. Shopbell, M. C. Britton, and R. Ebert, eds.

Wayman, J. C. (2003). Multiple Imputation For Missing Data: What Is It And How Can I Use It? *Annual Meeting of the American Educational Research Association*, (pág. 16). Chicago, IL.