

CREDIT SCORING

SERVICIOS FINANCIEROS

Dr. Viterbo Berberena G.

Introducción

El credit scoring es el proceso de asignación de un puntaje de riesgo a una solicitud de crédito o a una cuenta de crédito existente, a través de un modelo estadístico.

Es el proceso de desarrollo de un modelo estadístico a partir de datos históricos.

Este término también se refiere a monitorear la precisión de uno de tales modelos estadísticos y monitorear el efecto que las decisiones basadas en los puntajes tienen sobre los indicadores fundamentales del negocio.

El credit scoring tiene tres aplicaciones fundamentales:

- Application scoring (para la aceptación o rechazo de las solicitudes de crédito).
- Behavioral scoring (para predecir la probabilidad de incumplimiento de los clientes que ya han sido aceptados).
- Collection scoring (para estimar el monto probable de deuda que el

En **application scoring** el CS se usa para optimizar la tasa de aprobación de las solicitudes. Permite a las organizaciones elegir un punto de corte óptimo de aceptación, de tal forma que se gana participación en el mercado mientras se mantiene la máxima rentabilidad. Los puntajes obtenidos en los clientes y los prospecto son esenciales para la personalización de los producto de crédito.

El **behavioral scoring** de los clientes existentes se usa para la detección temprana de cuentas de alto riesgo y permite a las organizaciones realizar acciones enfocadas a estos objetivos (por ejemplo, la reestructuración de la deuda). También forma las bases para cálculos más precisos del riesgo crediticio de todos los consumidores.

En el **collection score**, los recursos utilizados en el cobro de las deudas se pueden optimizar enfocando las actividades a objetivos concretos de recaudación (con alto puntaje de recaudación). También se usa para determinar el valor preciso del libro de deudas antes que este sea traspasado a una empresa recaudadora.

Construcción del Modelo de Crédito in-house

En algunas situaciones es apropiado comprar, a proveedores externos, modelos de crédito genéricos listos para usar, o tener modelos de créditos para propósitos específicos desarrollados por consultores externos. Sin embargo, mantener una práctica de construcción de modelos in-house ofrece ciertas ventajas:

- Ganancias a partir de economías de escala cuando se necesita construir muchos modelos específicos para una gran cantidad de segmentos.
- Consolidar una base de datos flexible y reutilizable, generar conocimientos y habilidades por si mismo, de forma que le sea fácil a la organización ser consistente en la interpretación de los resultados de los modelos y los reportes y en la propia metodología de modelación.

- Verificar la precisión y analizar las fortalezas y debilidades de los modelos de crédito adquiridos.
- Reducir el acceso de extraños a información estratégica y retener las ventajas competitivas con la creación de las mejores prácticas de la compañía.

Construcción del Modelo de Crédito con el Minero de Datos

El MD permite una serie de beneficios:

- El analista accede a las herramientas del minero a través de una interfase gráfica para crear los diagramas de flujo de proceso que sirven de estructura a las actividades de análisis.
- Los nodos que forman el diagrama de flujo del proceso están diseñados, de forma tal que el analista puede interactuar con los datos y el modelo

- Emplear a fondo su expertise, usando el software como un volante y no como piloto automático.
- Es ideal para probar nuevas ideas y experimentar con nuevos modelos de una manera eficiente y controlada. Esto incluye, por ejemplo, la creación y comparación de varios scorecards, modelos de árboles y

La Elección del Modelo Correcto

Con el EM es posible crear una variedad de modelos como los scorecards, árboles de decisiones o redes neuronales.

En la evaluación del modelo más adecuado para alcanzar la meta deseada se consideran criterios tales como:

- Facilidad en la aplicación del modelo.
- Facilidad para entenderlo.

Para cada modelo en particular es importante evaluar su capacidad predictiva, o sea, la precisión de los puntajes que este otorga a las solicitudes y las consecuencias de las decisiones de rechazar/aceptar que sugiere. Se usan una variedad de medidas de calidad relevantes para el negocio como:

- Las curvas de concentración.
- La curva de estrategia.
- Las curvas de ganancia.

El mejor modelo se determinará por el propósito para el que se usará el mismo y la estructura del conjunto de datos con el que se validará.

Los Scorecards

Es una tabla que contiene las preguntas que se formulan a un solicitante (llamadas características) y la lista de posibles respuestas (llamadas atributos).

Por cada respuesta el solicitante recibe una puntuación.

Esta puntuación es más alta si el atributo es de bajo riesgo y más baja si es de alto riesgo. Si el puntaje total de la solicitud excede el punto de corte, entonces se acepta.

Este modelo tiene ventajas con respecto a los árboles y las redes neuronales:

- Fácil de aplicar (en un papel frente al solicitante).
- Fácil de entender (escalas monótonas o lineales).

- Fácil de justificar la decisión (es posible revelar los grupos de características, donde el solicitante tiene potencial para mejorar el puntaje total y hacer esto en términos los suficientemente amplios para no poner en riesgo la manipulación de las solicitudes futuras).

Los Árboles de Decisión

Los AD pueden sobrepasar el desempeño de los scorecards en términos de precisión predictiva, porque a diferencia de los scorecards, estos detectan y explotan las interacciones entre las características. Cada respuesta que da el solicitante determina cuál es la siguiente pregunta que se debe hacer. De esta forma el modelo de AD consiste de un conjunto de reglas “if...then ...else”, que son aún más fáciles de

Las reglas de decisión son también fáciles de entender, aún más, cuando estas se basan en un puntaje total que se obtuvo de muchos componentes.

Aunque el AD es fácil de aplicar y entender puede ser difícil justificar en solicitudes que caen en la frontera entre dos segmentos.

Si el AD no se usara directamente para obtener los puntajes, este modelo todavía añade valor en diferentes formas:

- Puede dar una comprensión importante de la estructura de riesgo de la población, en la identificación de segmentos claramente definidos de solicitudes con un particular alto o bajo riesgo.
- En el monitoreo de los scorecards, se identifican segmentos de solicitudes donde los scorecard tiene un bajo desempeño

- Los AD frecuentemente pueden alcanzar un poder predictivo similar al scorecard con muchas menos características.
- Estos modelos de pocas características, llamados “short scores” se están haciendo populares en el contexto de las campañas y mercadotecnia de los productos de crédito. La desventaja de este enfoque es que disminuye la riqueza de la información que la organización puede recopilar de las

Las Redes Neuronales

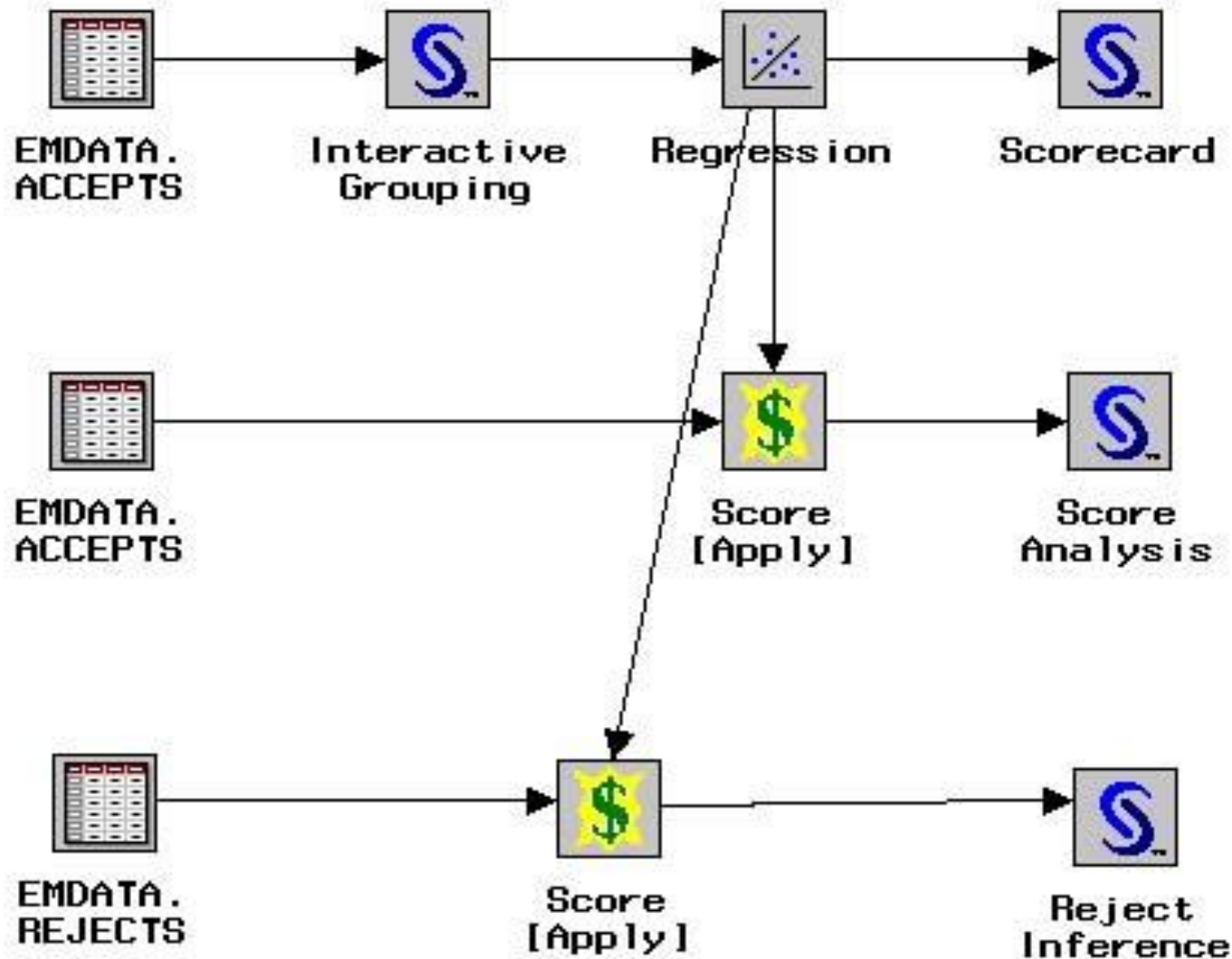
En los árboles de decisión, se puede observar que hay una cosa como la regla de decisión que es fácil de entender y por lo tanto invita al fraude. Irónicamente hablando no hay este peligro con las RN. Estas son modelos extremadamente flexible que reúnen combinaciones de características en diversas formas. Su precisión predictiva puede ser, por mucho, superior a los scorecards y no sufren por los abruptos

Sin embargo es virtualmente imposible explicar o entender de forma sencilla, los puntajes que se producen para una aplicación en particular. Puede ser difícil justificar una decisión hecha sobre la base de un modelo de red neuronal. En algunos países es un requerimiento legal explicar la decisión y tal justificación debe ser producida con métodos adicionales.

La RN, con un poder predictivo superior, es más apropiada en los casos donde la precisión promedio de la predicción es más importante que profundizar en los puntajes para cada caso en particular. Las RN no se pueden aplicar manualmente como los scorecards o árboles de decisión, pues requieren de un software para calcular los puntajes de la solicitud. Una vez hecho esto su aplicación es tan simple como los otros modelos.

La Metodología de Credit Scoring

1. La carga de la base de datos.
2. El agrupamiento interactivo.
3. El ajuste del modelo de regresión logística.
4. El cálculo de los puntaje para cada atributo.
5. El análisis de los puntajes.
6. La inferencia de los rechazos.
7. La modelación en la base completa.



Process Flow Diagram – Accepts and Rejects

Caso de Estudio de GFM

Preparación de la Base de Datos

Manipulador de Datos de SAS

- Datos correspondientes al primer trimestre del 2001 (766,304 casos).
- Canales: Canal 1, Canal 2, Canal 3 y Canal 4.

Tamizado Inicial de las Variables

Se eliminaron los clientes que tenían fecha de cancelación.

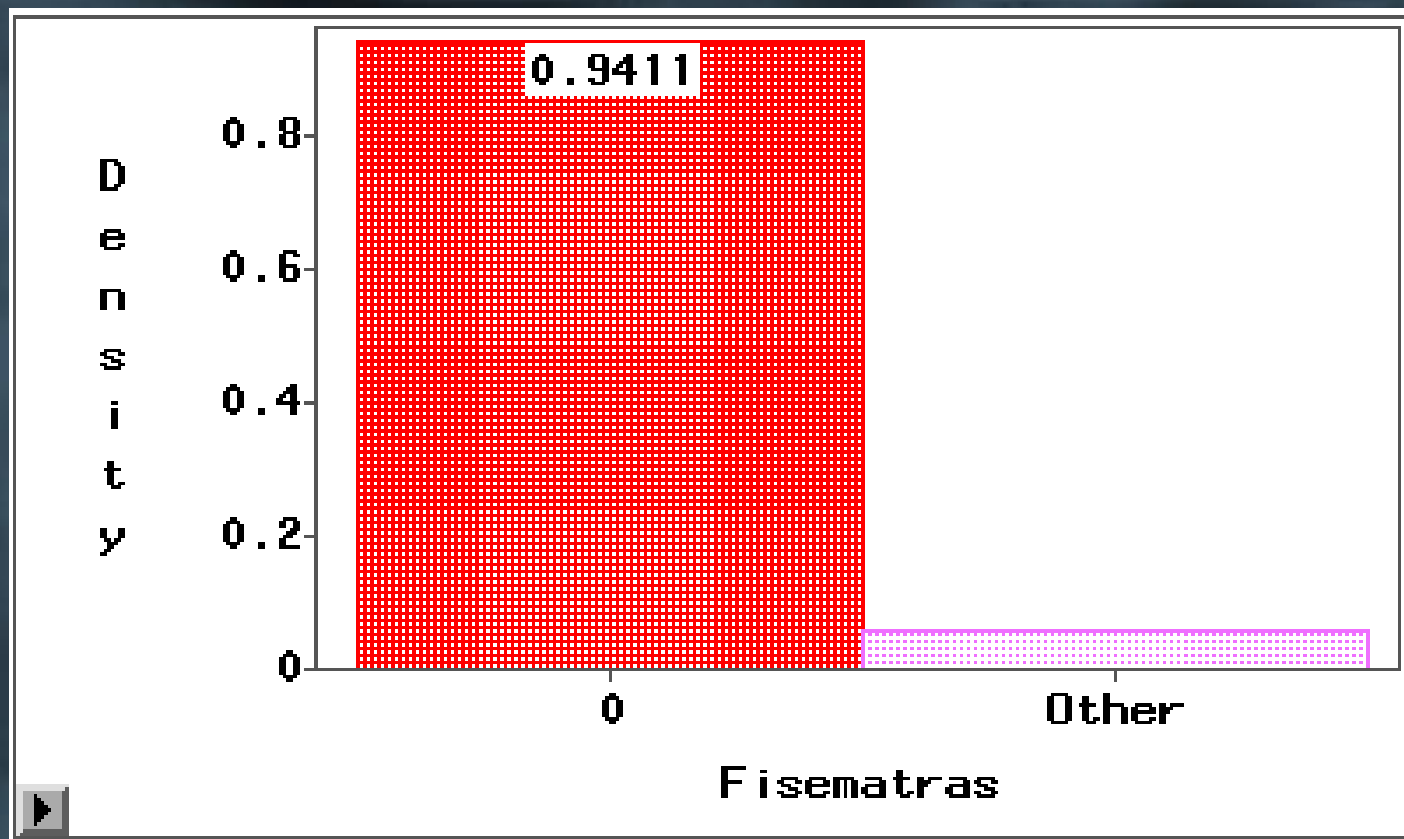
Se seleccionaron los datos con más información (menos valores perdidos), quedando principalmente datos del Canal Canal 1.

Una vez hecha esta selección los datos restantes abarcaron 144,675 clientes (filas) y 89 variables (columnas).

Creación de la Variable Objetivo

A partir de indicaciones de la Gerencia se usó la variable “**semanas de atrasos en los pagos**”, para la creación de la variable objetivo “**Good/Bad**”.

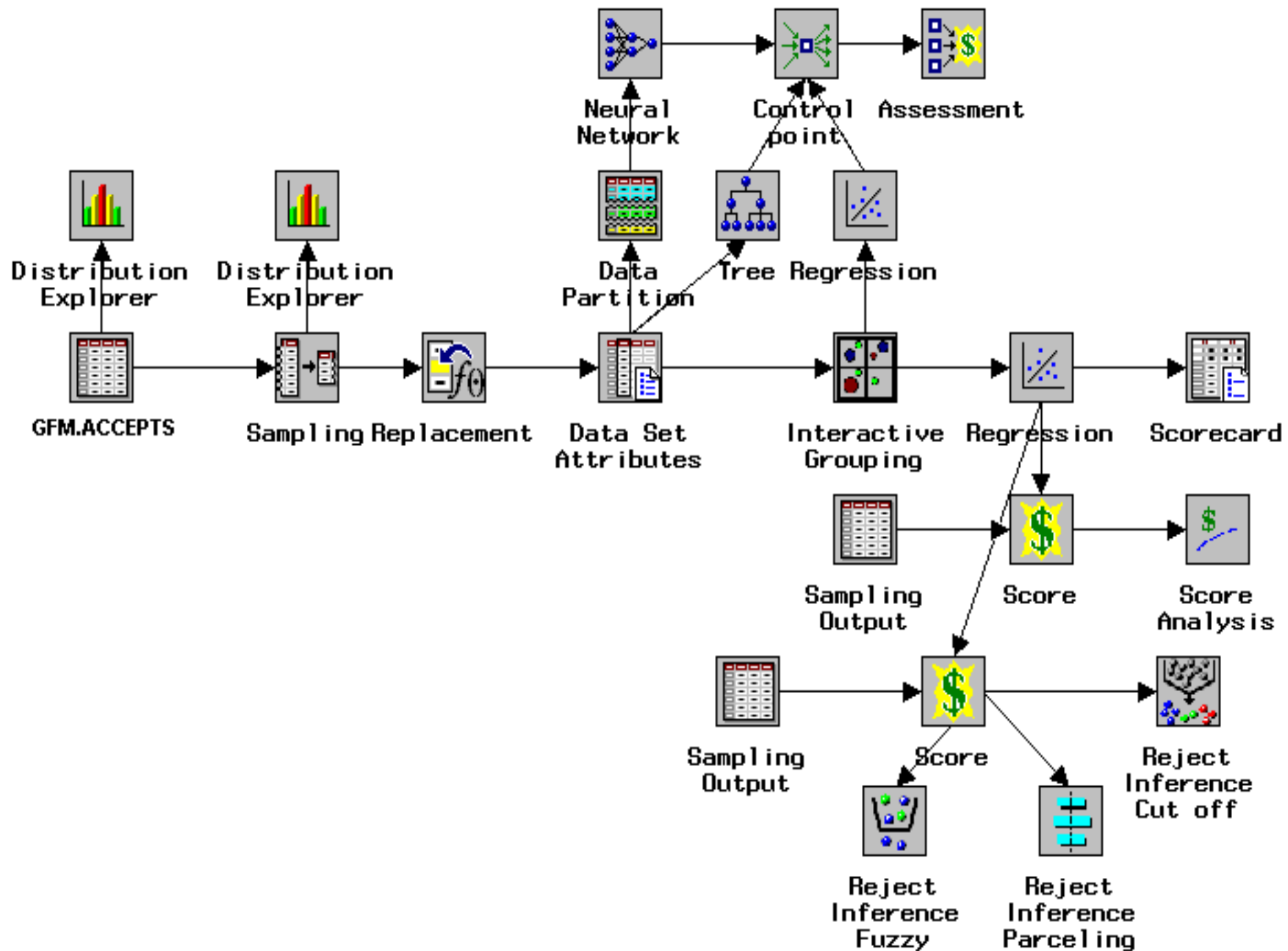
Se analizó la función de distribución (figura siguiente) y se asumió como buenos clientes, aquellos que no tenían ningún atraso y como malos clientes, lo que tenían algún tipo de atraso en sus pagos.



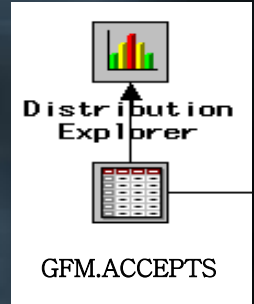
**Distribución de Frecuencias de la variable
“Semanas de Atrasos en los Pagos”**

El Credit Scoring con el SAS Enterprise Miner

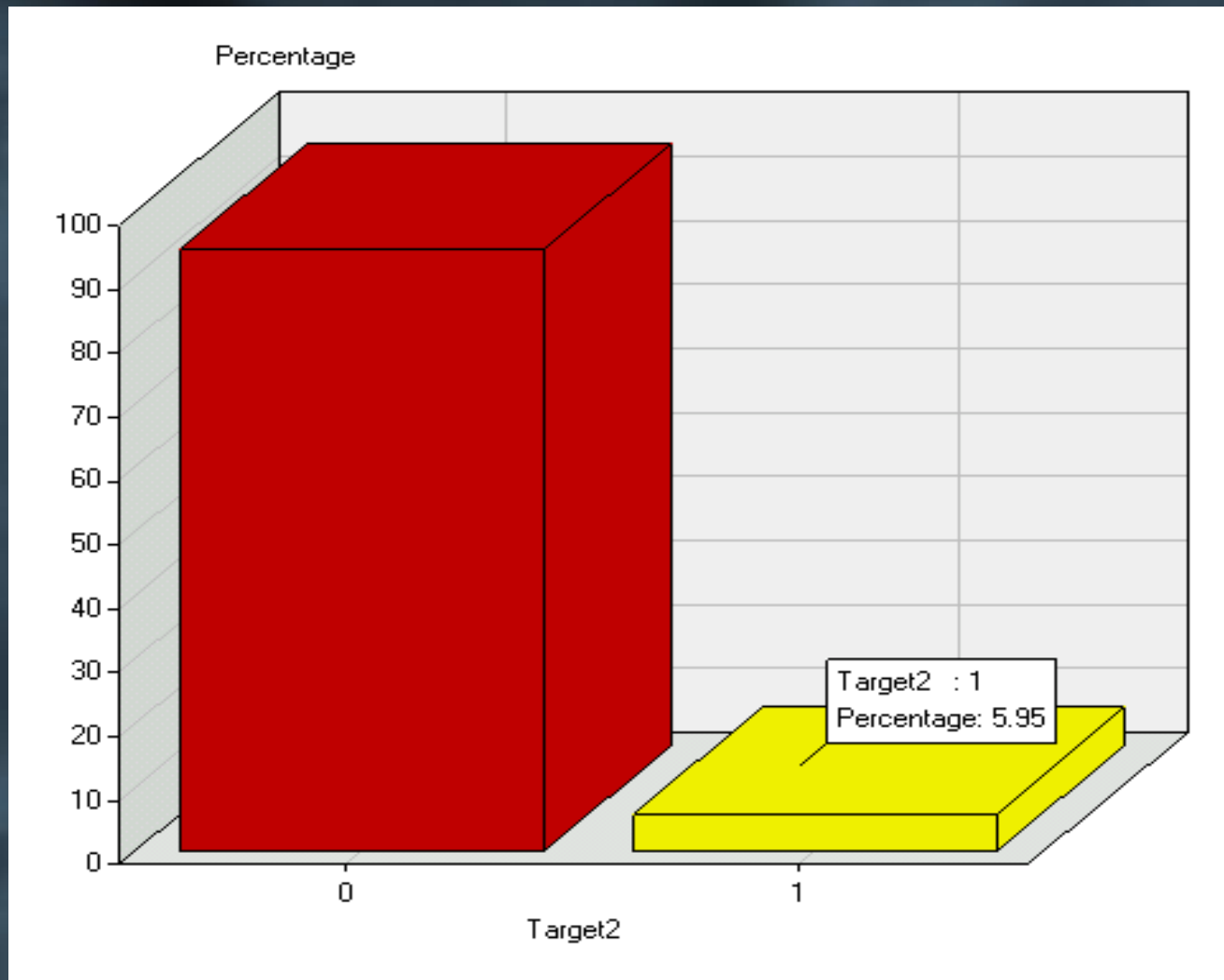
En la figura siguiente se presenta el Diagrama de Flujo del Proceso de la Metodología de Credit Scoring de SAS, aplicada al caso específico de la base de datos proporcionada por el Grupo Financiero Mexicano.



El Nodo Input Data Source



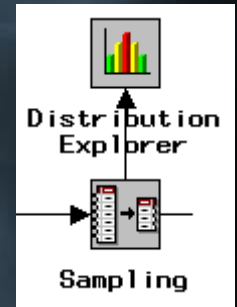
Se carga la tabla (base de datos del primer trimestre del 2001), que contiene los 144,675 filas (clientes) y 89 columnas (variables). Se selecciona la variable objetivo Good/Bad (Target2) y se verifica la distribución de la misma con un nodo de Distribution Explorer.



La variable objetivo (Bad/Good) es Target2. El evento igual a 0 es Good y el evento igual a 1 es Bad.

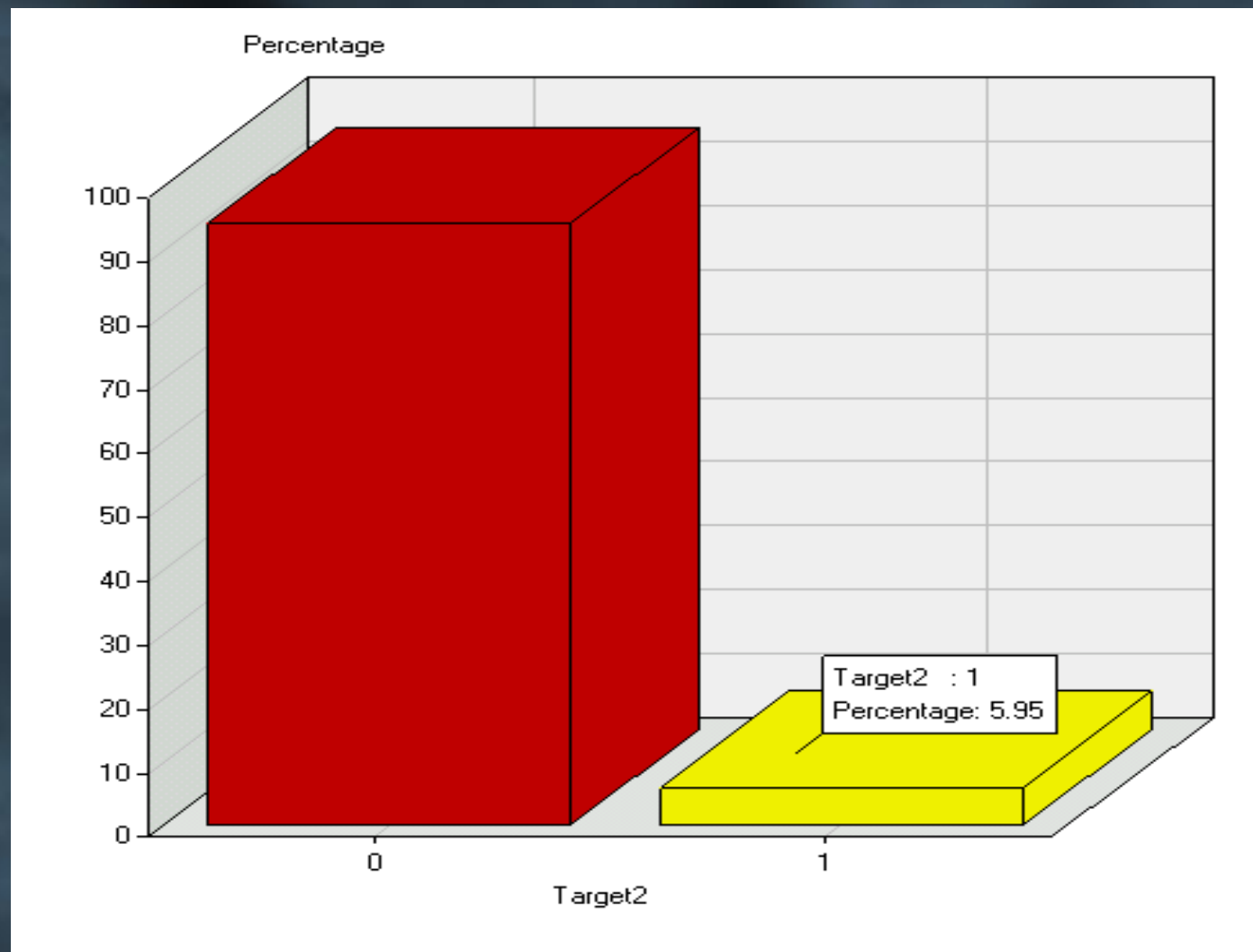
De las 89 variables, 48 están en escala de intervalos y 41 categóricas (nominales y ordinales). Se eliminan 11 variables (rejected). 5 por ser clasificadas como unary por el minero (una sola observación en la columna) y las 6 restantes por aportar muy poca información o no estar relacionadas con la variable de respuesta.

El Nodo Sampling



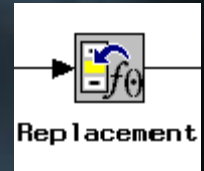
Se selecciona una muestra del 10% (14,468 casos) con el objetivo de construir los modelos y evitar el sobreajuste de la red neuronal (overfitting). Se utiliza el procedimiento de muestreo aleatorio simple (semilla del generador 12345).

También se verifica la distribución de la variable objetivo en la muestra obtenida (ver figura siguiente).



La variable objetivo (Bad/Good) es Target2. El evento igual a 0 es Good y el evento igual a 1 es Bad.

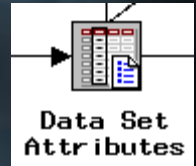
El Nodo Replacement



Se coloca este nodo con el objetivo reducir el ancho de algunos campos a menos de 16 caracteres para permitir que el nodo Interactive Grouping funcione adecuadamente.

También se hace la imputación de los missing con la media en el caso de las variables de intervalos y con la moda para el caso de las categóricas.

El Nodo Data Set Attributes



Es un nodo que permite la eliminación de un grupo de variables con muy poca información o muy poco relacionadas con la variable de respuesta, de forma tal que de las variables anteriores solo quedan las siguientes 24.

Este es un tamizado fino antes de empezar el proceso de modelación y transformación de variables.

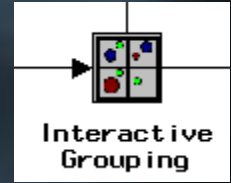
Variables que quedaron

1. Ingresos mensuales del cliente.
2. Ingreso mensual cónyuge.
3. Ingreso mensual fiador.
4. Importe del ingreso de otras fuentes.
5. Ingreso solidarios del cliente.
6. Edad del cliente (a partir de la fecha nacimiento).
7. Teléfono.

8. Identificador del tipo de trabajo.
9. Estado civil del cliente.
10. Ocupación del cliente.
11. Sexo del cliente.
12. Estado del domicilio del cliente.
13. Tipo de trabajo.
14. Identificador de la persona Física/Moral.
15. Sucursal.

16. Tipo del departamento del pedido.
17. Importe total de la venta.
18. Saldo capital del pedido.
19. Plazo del pedido.
20. Cantidad de artículos.
21. Importe del enganche pagado.
22. Importe del abono de los pagos del pedido.
23. Ingresos del cónyuge.
24. Target2 (variable Good/Bad).

El Nodo Interactive Grouping



El objetivo de este nodo es la clasificación, que no es más que el proceso automático y/o interactivo de redistribución y agrupamiento de las variables de intervalos, ordinales y nominales con el propósito de:

- Manejar el número de atributos (nuevos niveles de la variable) por característica (la variable).

- Mejorar el poder predictivo de la característica.
- Seleccionar los predictores (características con poder de predicción).
- Crear las variables WOE (weights of evidence) para lograr que los puntajes del scorecard varíen suavemente (de forma monótona) o linealmente a través de los atributos.

La cantidad de puntos que representa el valor del atributo en el scorecard se determina por dos factores:

- El riesgo de un atributo con relación a otros atributos de la misma característica (se determina por los weights of evidence).
- La contribución relativa de la característica al puntaje total (se establece por los coeficientes de regresión logística).

El WOE de un atributo se define como el logaritmo de la relación entre la proporción de goods y la de bads en el atributo:

$$WOE = \log \frac{P_{attribute}^{good}}{P_{attribute}^{bad}}$$

$$P_{attribute}^{good} = \frac{\# goods_{attribute}}{\# goods_{character.}}$$

$$P_{attribute}^{bad} = \frac{\# bads_{attribute}}{\# bads_{character.}}$$

Valores grandes negativos de la variable corresponden a un alto riesgo y alto valores positivos se relacionan con bajo riesgo.

Partiendo de que la cantidad de puntos en el scorecard es proporcional al el WOE, el proceso de clasificación determina cuánto vale un atributo con respecto a los otros dentro de una misma característica.

Después que la clasificación ha definido los atributos de las características hay que determinar el poder predictivo de las mismas, es decir, su capacidad para separar los altos riesgos de los bajos riesgos. Esto se realiza con la ayuda de una medida llamada Information Value (IV). El IV es la suma ponderada de las WOE de los atributos. El factor de ponderación es la diferencia entre la proporción de buenos y la de malos en el atributo respectivo:

$$IV = \sum \left(p_{attribute}^{good} - p_{attribute}^{bad} \right) \times WOE_{attribute}$$

El valor de IV debe ser mayor que .02 para que una característica sea considerada para incluirla en el scorecard. IV por debajo de .1 se consideran débiles, menores a .3 medios y menores a .5 fuertes. Si el IV es mayor de .5, la característica está sobrepredicha (overpredicting), lo que significa que de alguna forma está trivialmente relacionada con la variable objetivo (target).

Grouping

Scorecode

Log

Output Groupings

Options

Target Variable TARGET2

Variable to Categorise EDAD

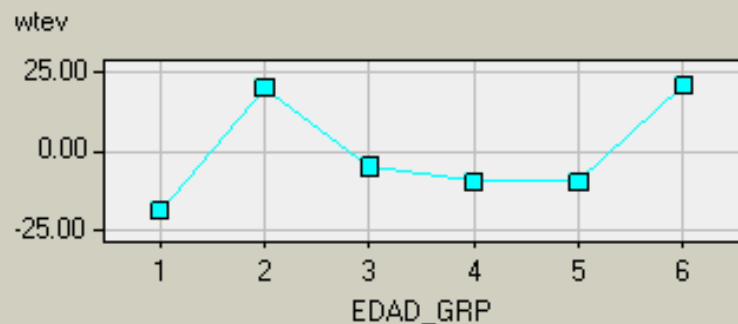
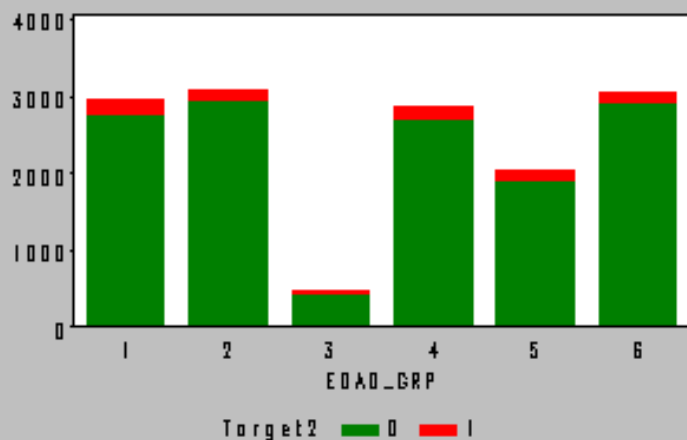
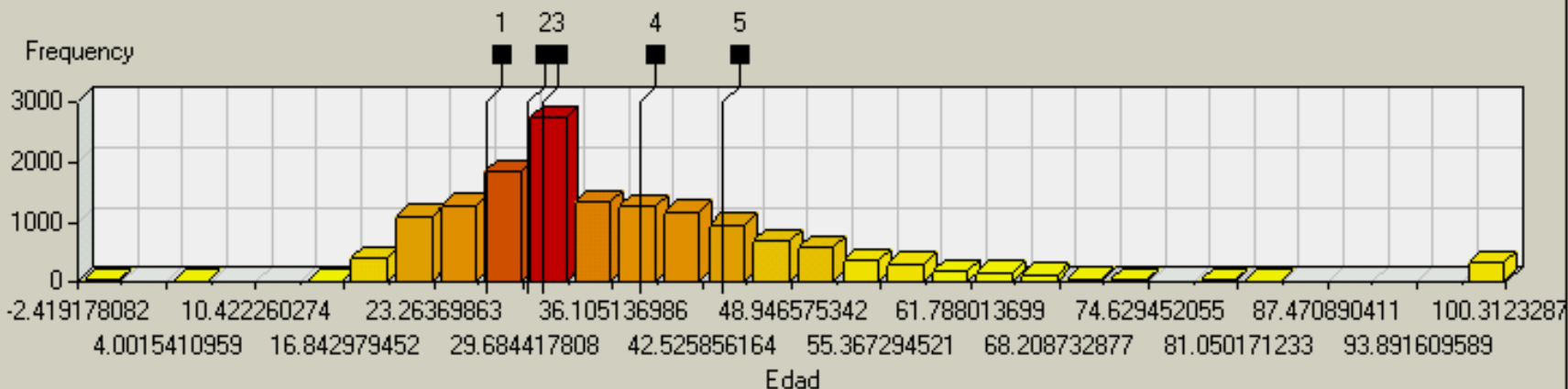
Alter Reference Line

Number of Bins 6

Binning Method

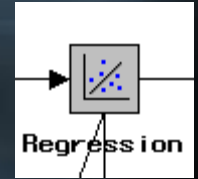
Buckets

Quantile



Clasificación de la Variable Edad

El Nodo Regression



Una vez que ha sido cuantificado el riesgo relativo de los atributos en una misma característica, un análisis de regresión logística determina el peso relativo de las características unas con otras. Este nodo recibe una variable de entrada por cada característica. Esta contiene como valores los WOE de los atributos de la característica.

Se pueden utilizar diferentes métodos de regresión para eliminar las características redundantes. Se recomienda hacer un estudio con el método stepwise y luego proseguir con el nodo de scorecard.

A continuación se muestran los valores de los coeficientes de regresión que sirven de base para el cálculo del scorecard.

The DMREG Procedure

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	4791.989	3070.400	.
SBC	4799.310	3143.609	.
-2 LOG L	4789.989	3050.400	1739.588 with 9 DF (p<.0001)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square	Standardized Estimate	exp(Est)
Intercept	1	-2.7535	0.0618	1986.22	<.0001	.	0.064
EDAD_WOE	1	-0.0153	0.00309	24.39	<.0001	-0.146948	0.985
FCCTEESTADO_WOE	1	-0.00859	0.00111	59.97	<.0001	-0.217144	0.991
FILABIDTIPOTRABAJO_W	1	-0.00359	0.000905	15.73	<.0001	-0.176437	0.996
FITIPODEPTO_WOE	1	-0.00952	0.000567	282.02	<.0001	-0.490532	0.991
FNCTEINGMEN_WOE	1	-0.00584	0.000930	39.42	<.0001	-0.362425	0.994
FNCYINGRESO_WOE	1	-0.00320	0.00150	4.54	0.0331	-0.059664	0.997
FNIMPTOTVTA_WOE	1	-0.00601	0.00141	18.21	<.0001	-0.119703	0.994
FNINGMENSUAL_WOE	1	-0.00933	0.00350	7.09	0.0077	-0.090297	0.991
FPLAZO_WOE	1	-0.0104	0.00171	36.58	<.0001	-0.217614	0.990

Tabla de estadígrafos de bondad de ajuste del modelo de regresión logística

Effect T-scores Estimates Table

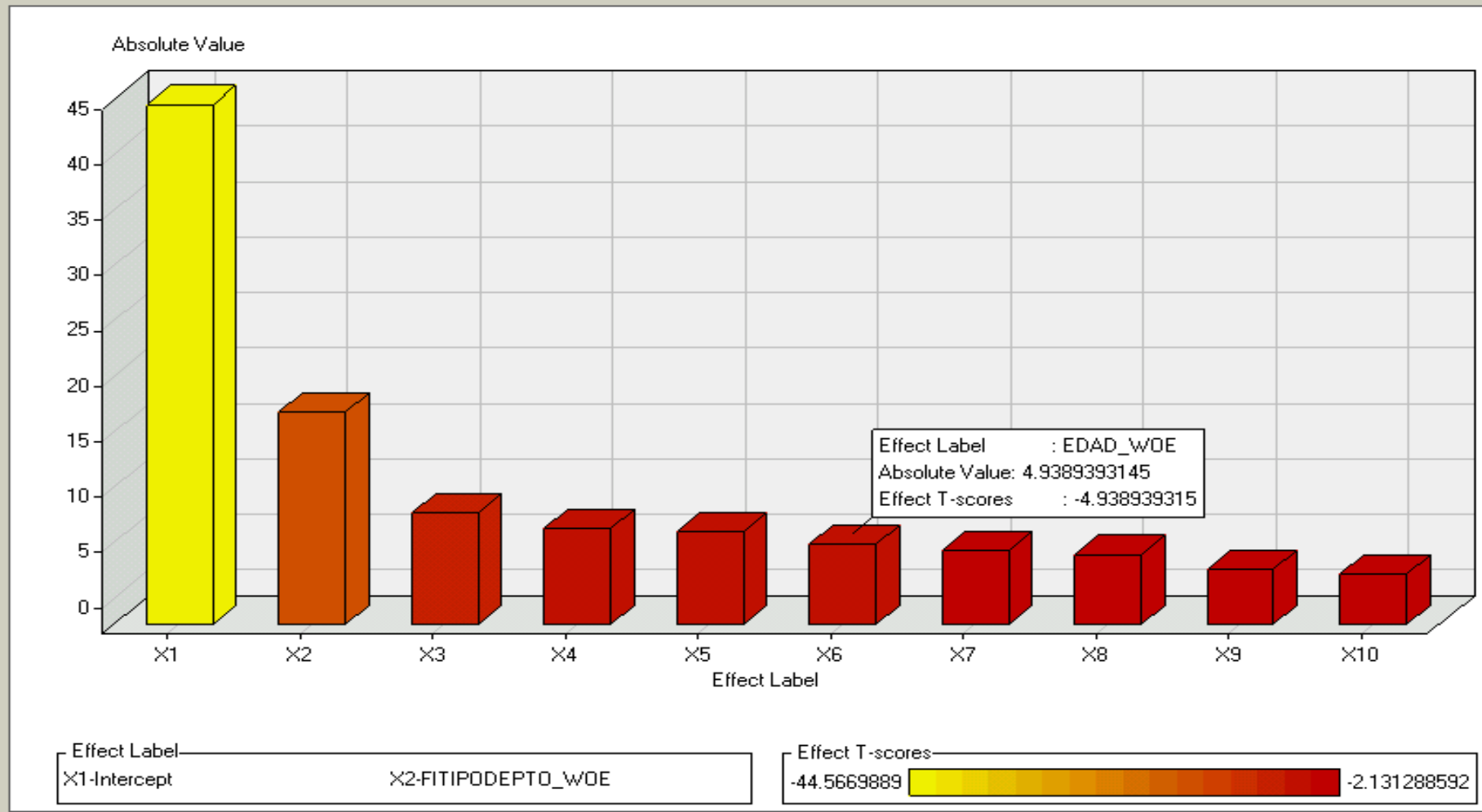
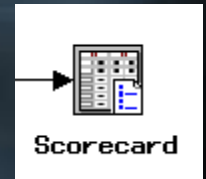


Gráfico de los estimados de los coeficientes de la regresión logística

El Nodo Scorecard



El WOE de cada atributo es multiplicado por el coeficiente de regresión de su característica para obtener el puntaje del scorecard del atributo. El puntaje total del solicitante es proporcional al logaritmo de la probabilidad de su razón good/bad estimada. A continuación se presenta la ecuación para el cálculo de los puntajes del scorecard para cada solicitante.

$$score = \log(odds) * factor + offset =$$

$$\left(-\sum_{i=1}^n (woe_i * \beta_i) + a\right) * factor + offset =$$

$$\left(-\sum_{i=1}^n \left(woe_i * \beta_i + \frac{a}{n}\right)\right) * factor + offset =$$

$$\sum_{i=1}^n \left(-\left(woe_i * \beta_i + \frac{a}{n}\right) * factor + \frac{offset}{n}\right)$$

Los puntajes están en escala lineal de enteros conforme a las normas de esta industria.

Se tomó una escala de puntajes tal que el valor de 600 corresponde a una relación good/bad de 50/1 y que un incremento en el puntaje de 20 unidades coincide con el doble de la relación good/bad. Para la obtención de la regla de escalamiento que transforme los puntajes de cada atributo se usan las ecuaciones:

$$600 = \log(50) * factor + offset$$

$$620 = \log(100) * factor + offset$$

$$\textit{factor} = 20 / \log(2)$$

$$\textit{offset} = 600 - \textit{factor} * \log(50)$$

El scorecard resultante es una tabla, opcionalmente en formato HTML que se muestra a continuación. Se aprecia como los puntajes de las características cubren diferentes rangos.

Characteristic Name	Attribute	Scorecard Points
EDAD	. -> 27	56
EDAD	27 -> 30	70
EDAD	30 -> 31	61
EDAD	31 -> 38	60
EDAD	38 -> 44	60
EDAD	44 -> .	70
ESTADO	"CHIAPAS", "COAHUILA", "JALISCO", "NAYARIT", "NL", "SINALOA", "SONORA"	49
ESTADO	"CHIHUAHUA", "GUERRERO", "SLP"	55
ESTADO	"AGS", "DF"	56
ESTADO	"GUANAJUATO", "MICHOACAN", "PUEBLA", "QUERETARO", "YUCATAN", "ZACATECAS"	67
ESTADO	"HIDALGO", "MEXICO", "TAMAULIPAS"	70
ESTADO	"BCALIFORNIA", "CAMPECHE", "DURANGO", "MORELOS", "OAXACA", "QUINTANAROO", "TABASCO", "TLXCALA", "VERACRUZ"	80
TIPO TRABAJO	"0", "8"	56
TIPO TRABAJO	"1", "2"	74
TIPO DEPTO	. -> 4	76
TIPO DEPTO	4 -> .	-102

INGRESOS CTE	. -> 55	54
INGRESOS CTE	55 -> 11950	91
INGRESOS CTE	11950 -> .	55
INGRESOS CONYUGE	. -> 28	66
INGRESOS CONYUGE	28 -> .	59
IMPORTE VENTA	. -> 2173	73
IMPORTE VENTA	2173 -> 2840	69
IMPORTE VENTA	2840 -> 3818	63
IMPORTE VENTA	3818 -> 4646	64
IMPORTE VENTA	4646 -> 6596	65
IMPORTE VENTA	6596 -> 7963	63
IMPORTE VENTA	7963 -> .	52
INGRESO FIADOR	. -> 1701	75
INGRESO FIADOR	1701 -> .	61
PLAZO	. -> 28	87
PLAZO	28 -> 37	69
PLAZO	37 -> 41	60
PLAZO	41 -> .	50

El Nodo Score Analysis



Este nodo tiene de predecesor un nodo Score y este a su vez un Input Data Source. El Input Data Source en este caso está leyendo la salida del Nodo Sampling. El node de Score debe estar conectado al nodo Regression.

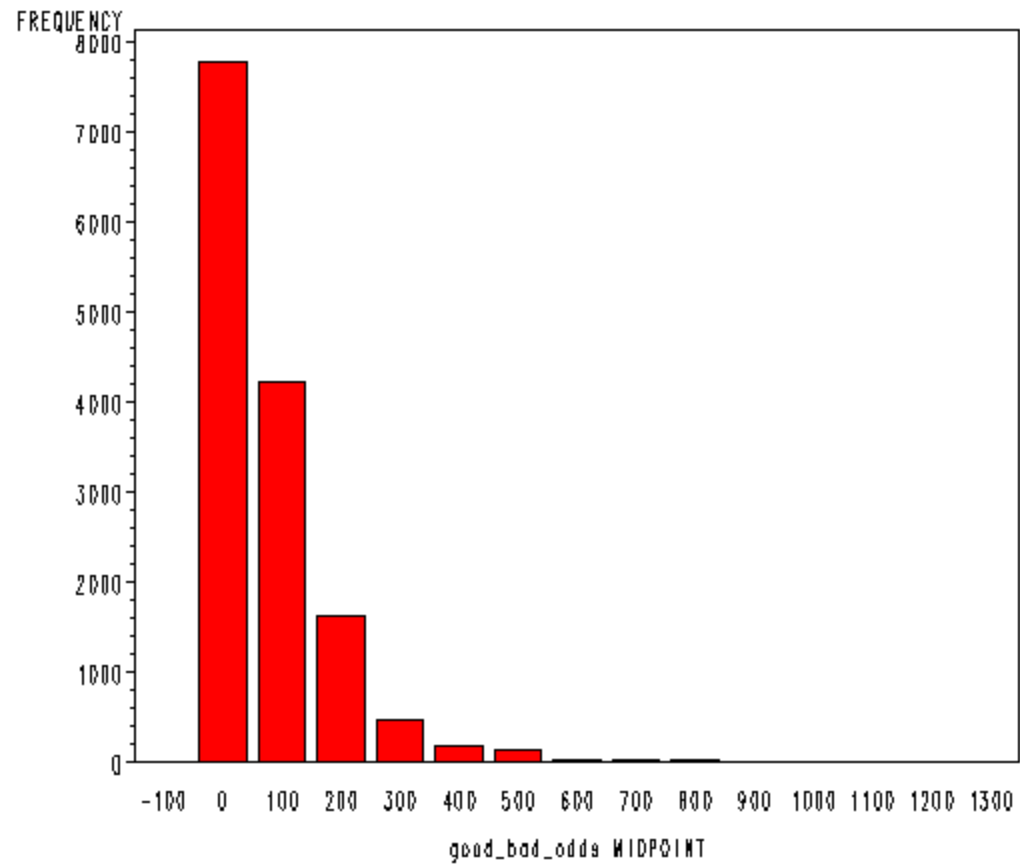
La función del nodo Score Analysis es:

- Visualizar la distribución de varios puntajes con relación a estadígrafos.

- Ayudar a la determinación del puntaje de corte óptimo a través de la creación de una tabla de ganancia, las curvas de equilibrio y el análisis de las características.
- Sacar los códigos de programación de la aplicación.

Score Analysis

Elektro
Score Distributions



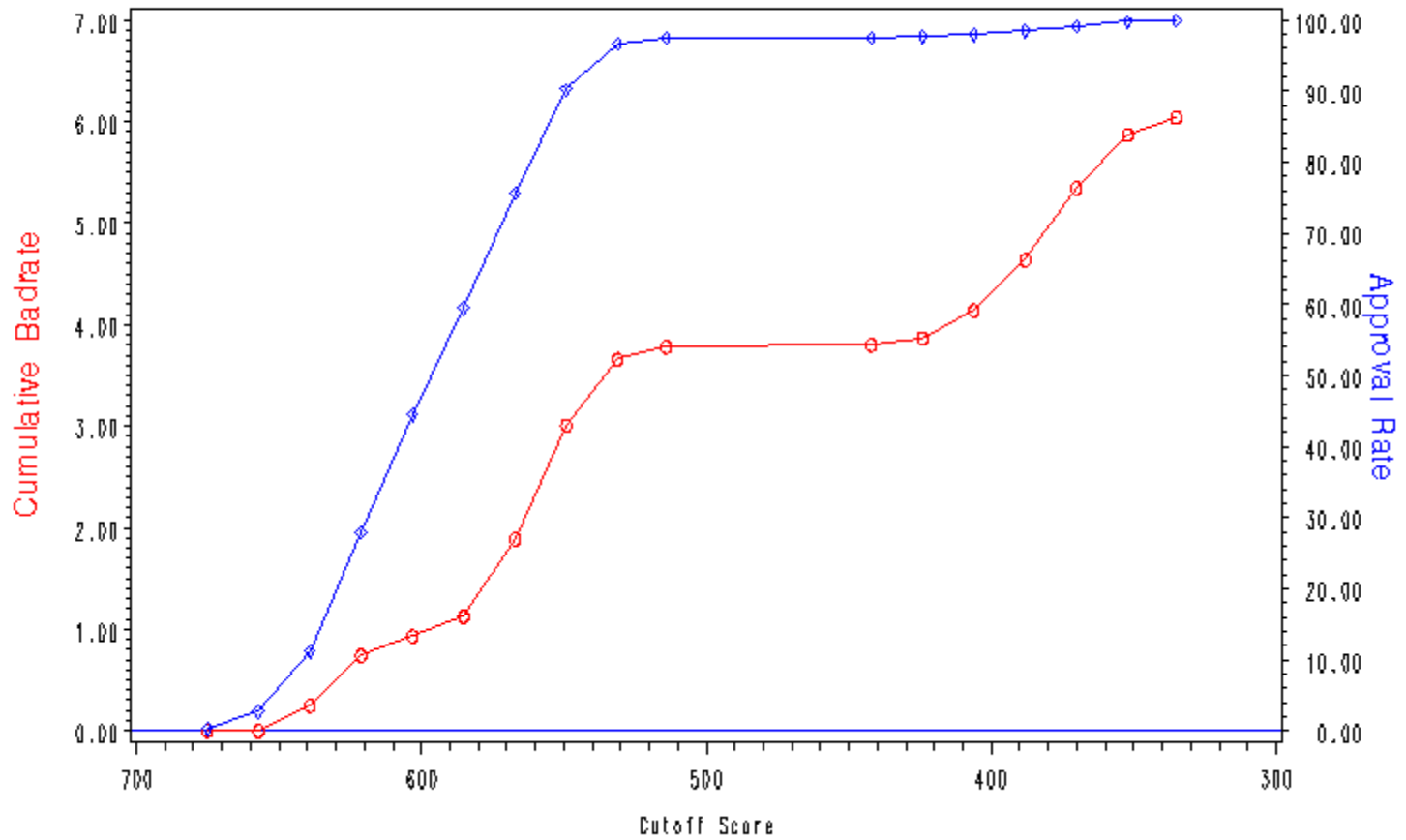
09MAY02:17:31:07

Gains Table

Score Range	Count	Cumulative Count	Number of Goods	Cumulative Number of Goods	Number of Bads	Cumulative Number of Bads	Marginal Badrate	Cumulative Badrate	Approval Rate	Average Marginal Profit	Average Total Profit
675<=Score<693	43	43	43	43	0	0	0.00	0.00	0.30	1000.00	2.97
667<=Score<675	347	390	347	390	0	0	0.00	0.00	2.70	1000.00	26.96
639<=Score<667	1220	1610	1216	1606	4	4	0.33	0.25	11.13	987.68	109.90
621<=Score<639	2427	4037	2401	4007	26	30	1.07	0.74	27.90	962.84	268.68
603<=Score<621	2392	6429	2362	6369	30	60	1.25	0.93	44.44	953.34	423.62
585<=Score<603	2170	8599	2133	8502	37	97	1.71	1.13	69.43	943.60	660.82
567<=Score<585	2360	10949	2240	10742	110	207	4.68	1.89	75.68	905.47	685.24
549<=Score<567	2111	13060	1925	12667	186	393	8.81	3.01	90.27	849.54	766.86
531<=Score<549	943	14003	823	13490	120	513	12.73	3.66	96.79	816.82	790.57
514<=Score<531	113	14116	62	13682	21	534	18.58	3.78	97.57	810.85	791.13
442<=Score<460	3	14119	0	13682	3	537	100.0	3.80	97.59	809.83	790.30
424<=Score<442	11	14130	1	13683	10	547	90.91	3.87	97.66	808.44	787.60
406<=Score<424	42	14172	2	13685	40	587	95.24	4.14	97.95	792.90	776.68
388<=Score<406	77	14249	3	13688	74	661	96.10	4.64	98.49	768.05	766.43
370<=Score<388	110	14359	3	13691	107	768	97.27	5.35	99.25	732.57	727.05
352<=Score<370	83	14442	3	13694	80	848	96.39	5.87	99.82	708.41	705.14
335<=Score<352	26	14468	0	13694	26	874	100.0	6.04	100.00	697.95	697.95

Score Analysis

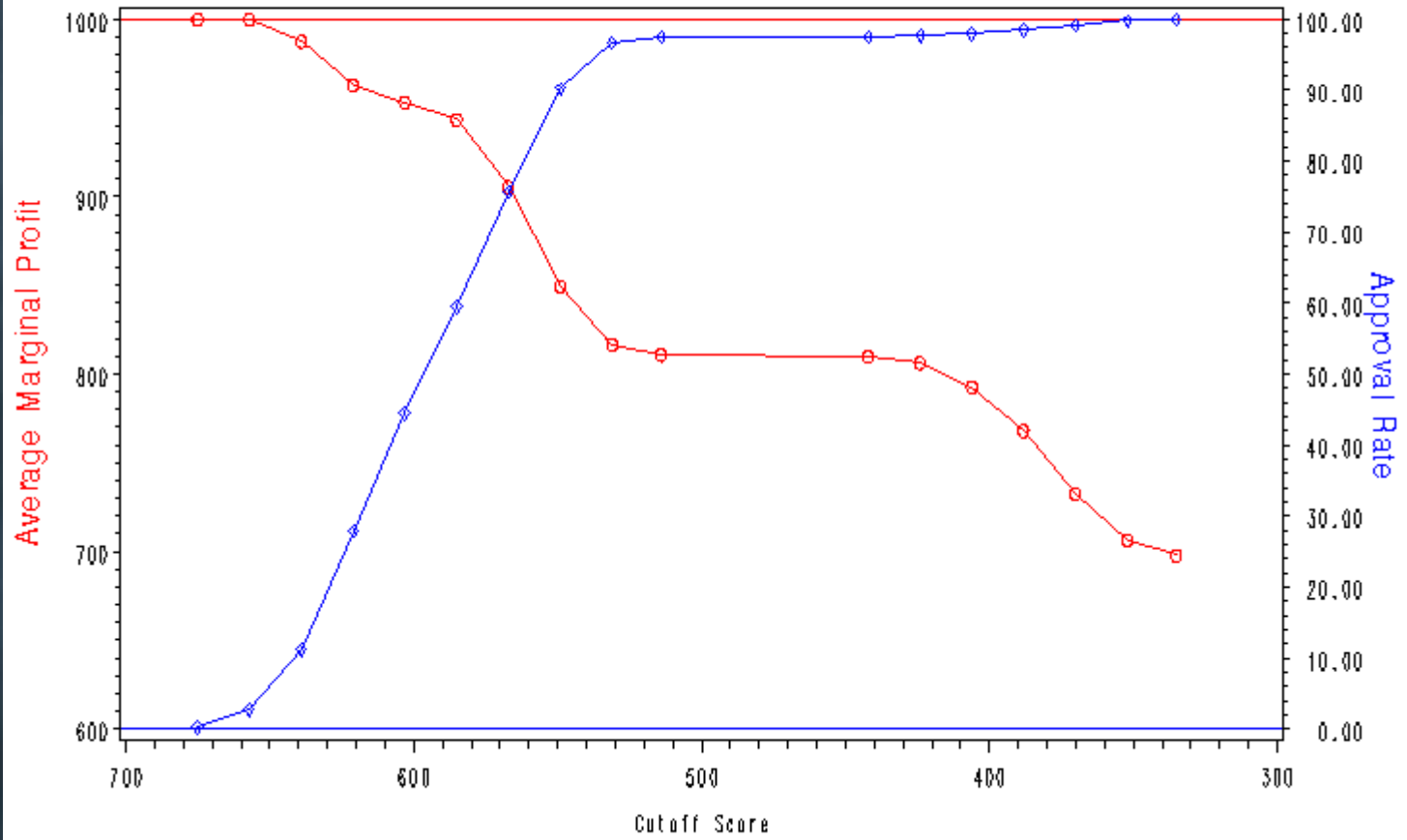
Elektro
Trade-Off Chart



09MAY02:17:31:07

Score Analysis

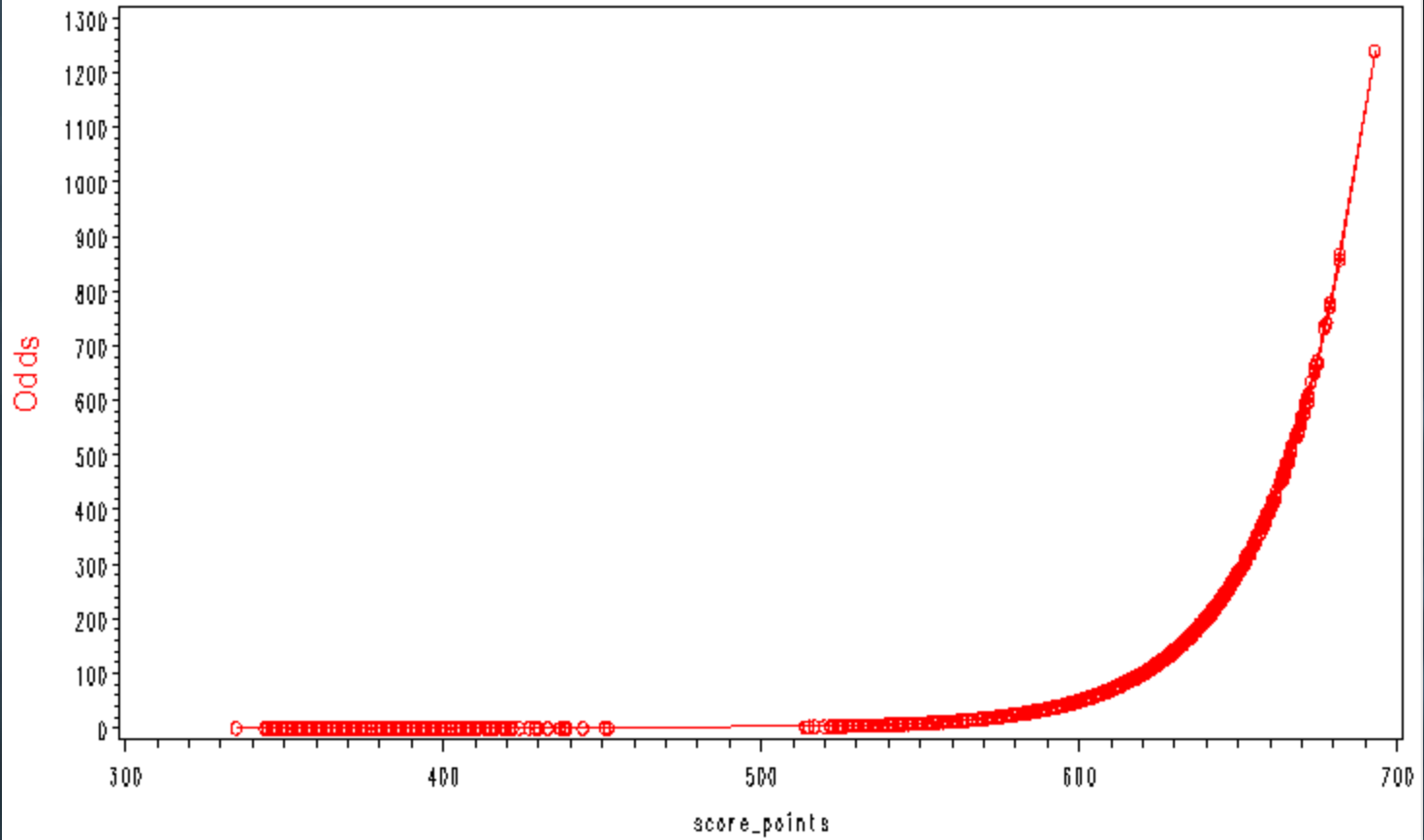
Elektra
Trade-Off Chart



00/11/00, 12:31:02

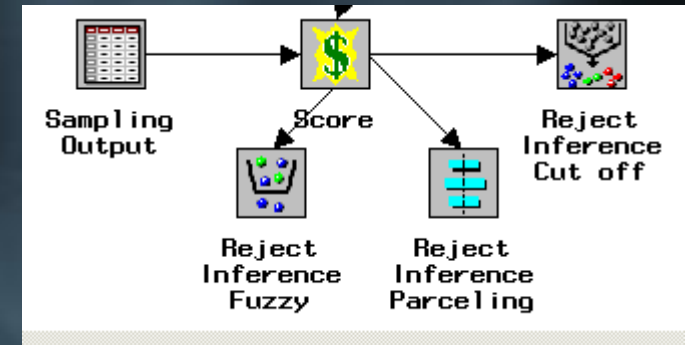
Score Analysis

Elektra
Strategy Curve



09MAY02-17-31-07

El Nodo Reject Inference



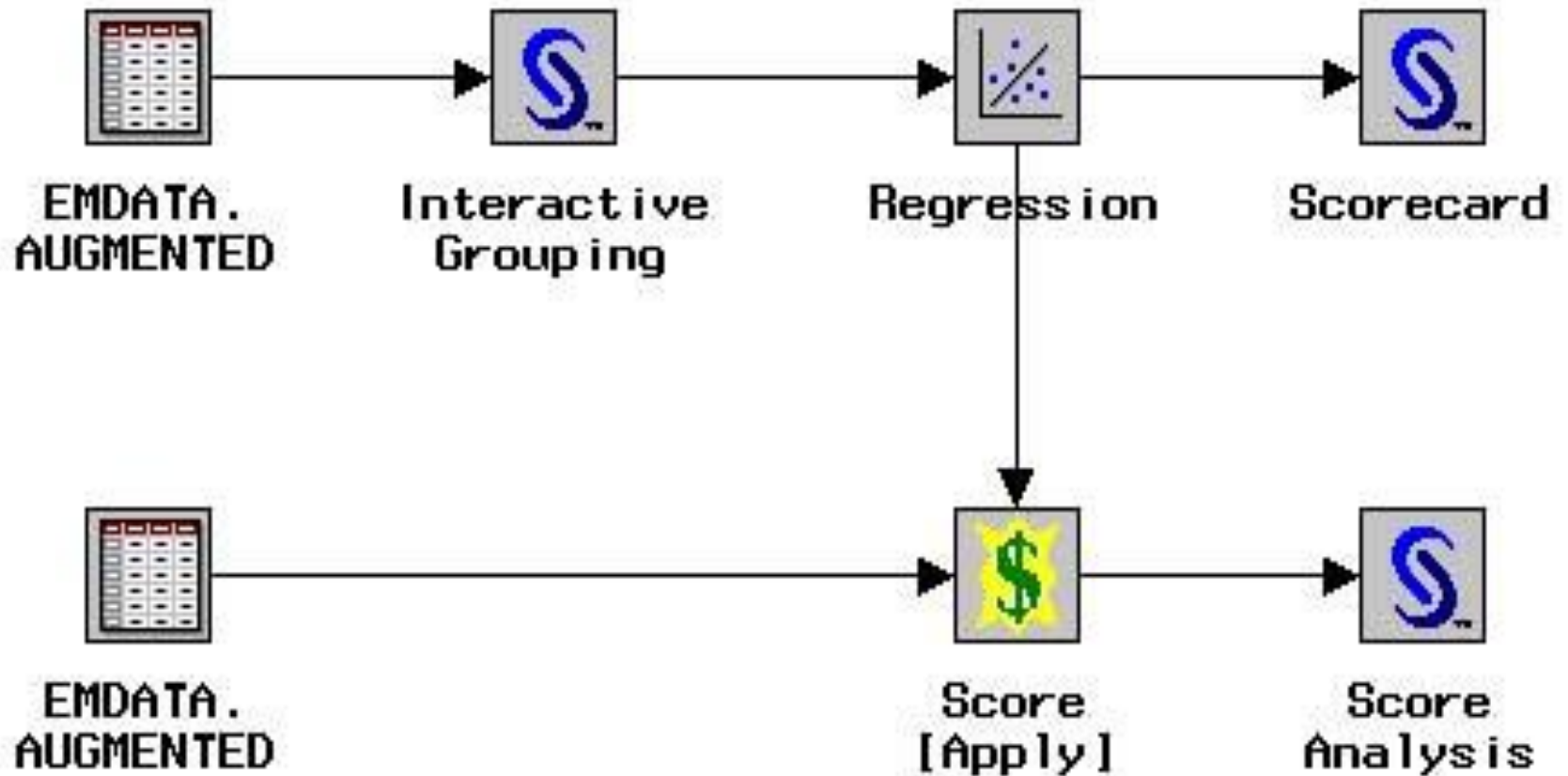
La función de este nodo es llevar a cabo la inferencia good/bad en una muestra de solicitudes denegadas y crear una muestra de solicitud “through the door” con indicadores de desempeño. Esto generalmente se hace en dos etapas:

- **Clasificación:** Donde a las solicitudes denegadas se les asigna una clasificación good or bad.
- **Integración:** Donde las solicitudes ya clasificadas se agregan a una muestra de good/bad conocidos para crear una muestra unificada de todos los good/bad (augmented data set).

Esto es necesario porque el modelo de evaluación que tenemos hasta el momento sufre de un sesgo esencial. Se ha creado basado en una población que es estructuralmente diferente a la que se supone que se aplicará.

La idea para resolver este problema es asignar puntajes a los datos de solicitudes denegadas con el modelo que fue creado sobre los datos de solicitudes aceptadas.

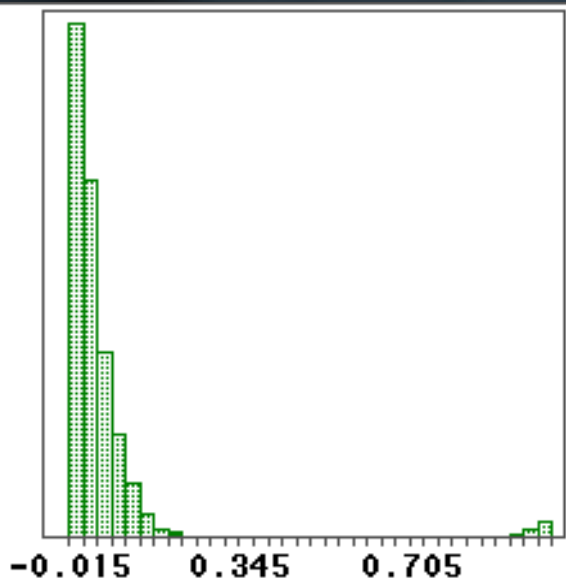
A cada solicitud denegada se le infiere un good o un bad y se añade a una muestra actual conocida de goods/bads. Los datos resultantes integrados sirven como base de datos para una segunda modelación donde se reajustan las clases y se recalculan los coeficiente de regresión. La lógica que hay detrás de este enfoque es que un punto de corte rígido para inferir la no aceptación de una solicitud es una decisión arbitraria que sesga la base de datos integrada (augmented data set).



Process Flow Diagram – Augmented

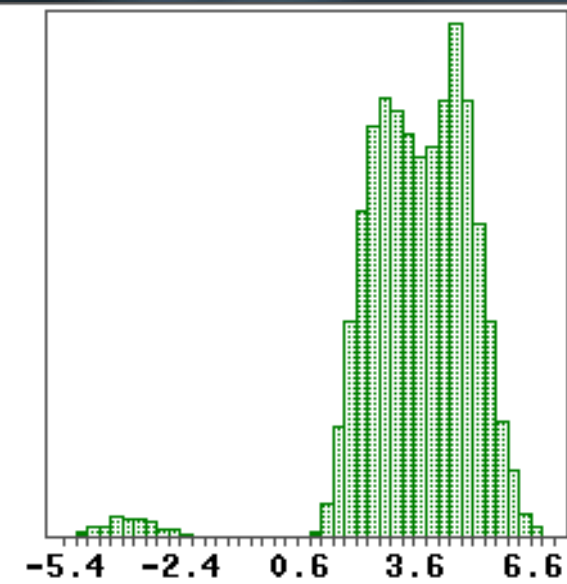
A continuación se presentan varias funciones de distribución que fueron obtenidas por la inferencia pero de la base de datos de solicitudes aceptadas. Desgraciadamente no tenemos la base de datos de solicitudes denegadas con la hubiera sido muy interesante para comparar los resultados y determinar cuán sesgado está el modelo de scorecard obtenido.

F
r
e
q
u
e
n
c
y



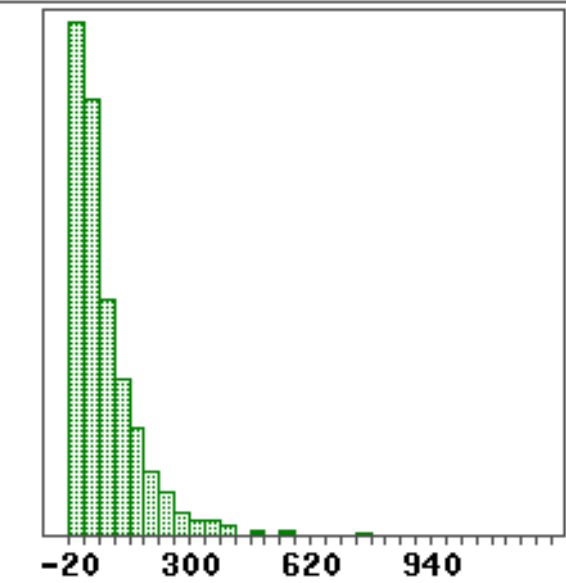
P_Target21

F
r
e
q
u
e
n
c
y



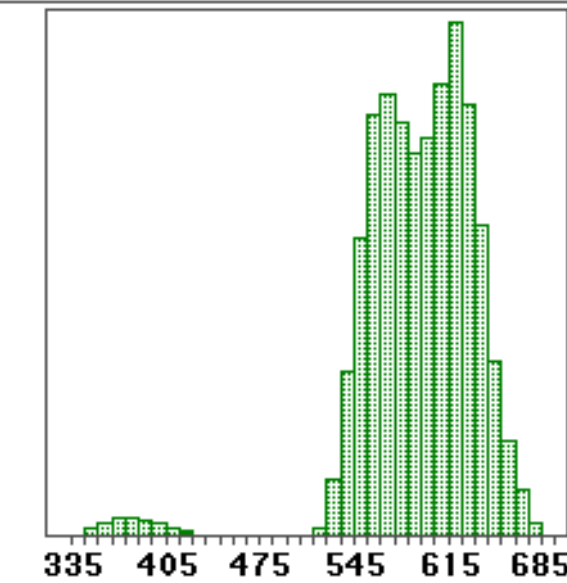
log_good_bad_odds

F
r
e
q
u
e
n
c
y



good_bad_odds

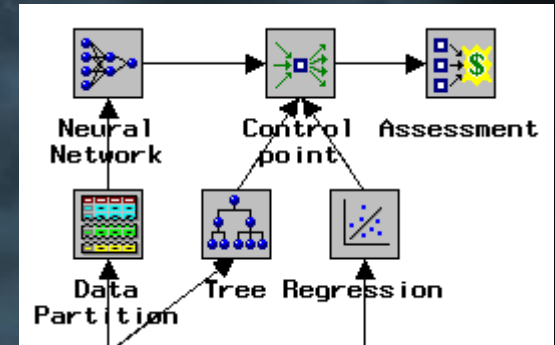
F
r
e
q
u
e
n
c
y



score_points

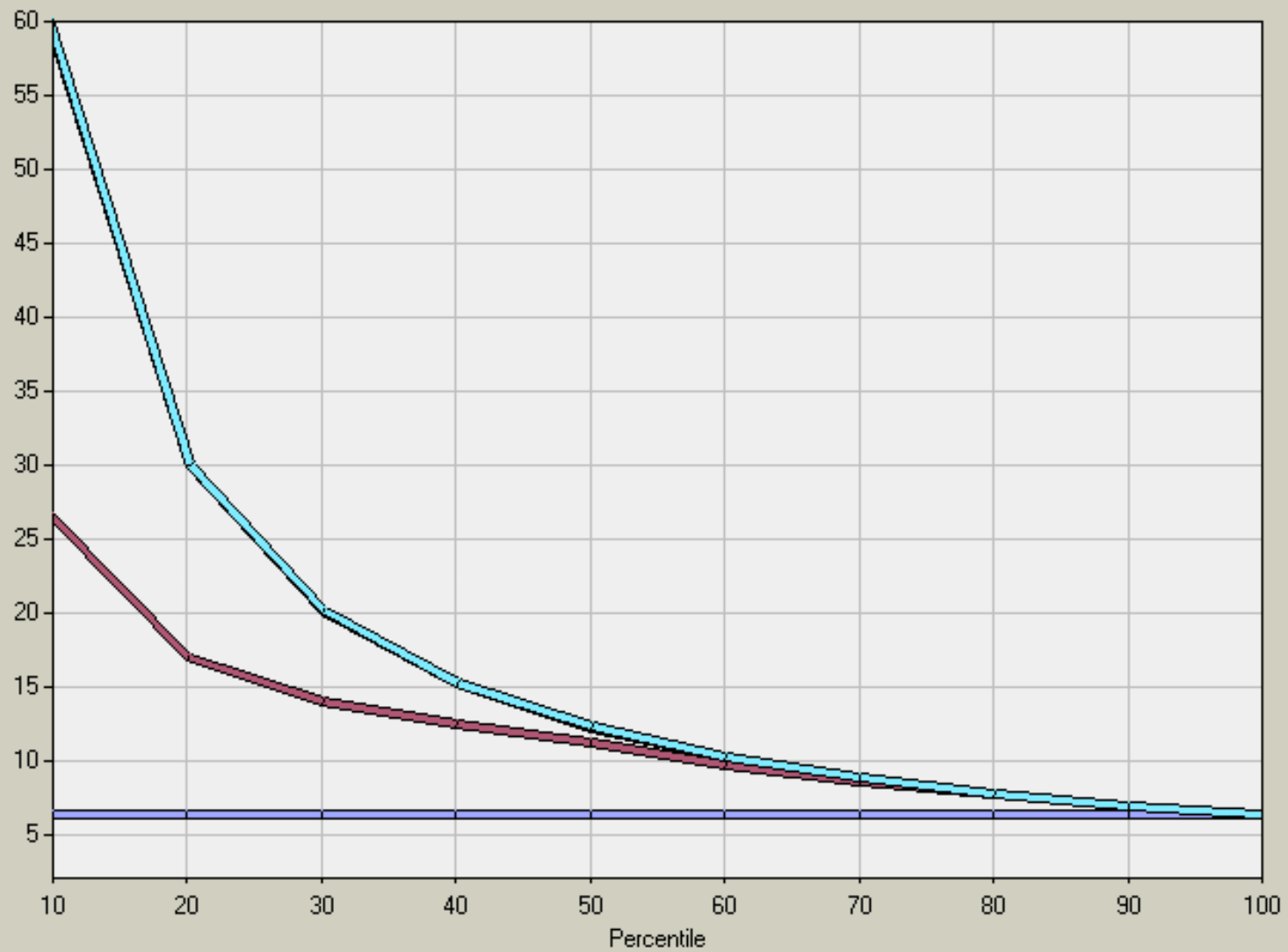
Este nodo genera una tabla de inferencia y para cada solicitud denegada se le infieren ambos un good y un bad, dando al inferido como bad un peso proporcional a la probabilidad de bad (p_{bad}) y al inferido good un peso proporcional a la probabilidad de good ($p_{\text{good}}=1-p_{\text{bad}}$).

Comparación del modelo de regresión con el de árbol y la red neuronal



Con el objetivo de comparar la bondad de ajuste del modelo de regresión, se cálculo un árbol de decisión y una red neuronal con con la arquitectura de un Multilayer Perceptron. En la figura siguiente se muestran los resultados.

%Response



Tool Name

Baseline Reg Tree Neural