

UCM - Minería de Datos

**TAREA 3 Credit Scoring (segunda tarea del
profesor Escot)**

GESTIÓN GLOBAL DEL RIESGO. SCORING

CAIO FERNANDES MORENO

1. Enunciado del trabajo

Esta es la segunda práctica que hay que entregar al profesor Escot. Se trata de construir una tarjeta de puntuación del riesgo de crédito para clientes de tarjeta de crédito. Y hacer previsiones sobre nuevos clientes ¿le daríais la tarjeta a esos nuevos clientes?.

En el Excel hay información sobre clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente). También hay clientes rechazados (Cardhldr=0) que son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na).

Por último al final del archivo hay 34 nuevos clientes (Cardhldr=na) individuos con identificador de cliente desde 1286 hasta 1319, que solicitan una tarjeta de crédito.

Tenéis que construir el modelo de Scoring siguiendo la metodología SAS Miner. Si queréis construir algún otro modelo (redes neuronales, árboles, o randomforest) también podéis hacerlo. En cualquier caso tenéis que seguir la metodología expuesta en clase, y en las presentaciones de Javier Monjas y Caridad Pavón (que esperemos que pueda venir en enero)

El resto de variables son las siguientes:

Age = Age n years plus twelfths of a year

Income = Yearly income (divided by 10,000)

Exp_Inc = Ratio of monthly credit card expenditure to yearly income

Avgexp = Average monthly credit card expenditure

Ownrent = 1 if owns their home, 0 if rent

Selfempl = 1 if self employed, 0 if not.

Depndt = 1 + number of dependents

Inc_per = Income divided by number of dependents

Cur_add = months living at current address

Major = number of major credit cards held

Active = number of active credit accounts

Os recuerdo las Fases

FASES DEL ANALISIS

- 1) Delimitación del estudio: objetivo, limitaciones y disponibilidad de datos
- 2) Análisis descriptivo exploratorio de los datos

¿Qué variables tenemos? ¿Cuál es la variable objetivo?

¿Qué porcentaje de buenos y malos tenemos en cada categoría?

¿Hay datos anómalos?

¿Hay datos perdidos?

¿Hay que imputar?

- 3) ¿Hay que tramificar variables?

- 4) Selección de variables ¿qué variables son las más importantes? ¿qué variables deben incluirse en el scorecard? disponibilidad, e importancia, estadísticos de selección (IV, correlación)

- 5) Muestreo:

Muestra de desarrollo/entrenamiento (70-80%) y de validación (30-20%)

Sobre muestreo de buenos/malos?

- 6) Estimación del primer modelo, diagnosis (Kolmogorov-Smirnov, c-statistic, Gini) y obtención del scorecard preliminar

- 7) Inferencia de denegados

- 8) Estimación del modelo definitivo y obtención del scorecard definitivo

- 9) Validación y Seguimiento

2. Ejecución del trabajo

Para el trabajo he utilizado en fichero Excel llamado **datospracticas_alumnos_OK.xls**.

Lo primero que hay que hacer es abrir el fichero **datospracticas_alumnos_OK.xls** y guardar el archivo en formato Excel 95 (**datospracticas_alumnos_OK_excel95.xls**) para poder importar los datos con SAS Base.

Manualmente he creado 3 archivos Excel:

CS_REJECTS.xls = Todos los clientes con Cardhldr=0 y default vacío, o sea clientes con tarjeta que ha sido denegada.

Contiene 291 observaciones.

CS_NEW_CLIENTS.xls = Todos los clientes que tienen default y Cardhldr vacío.

Contiene 34 observaciones.

CS_ACCEPTS.xls = Todos los clientes con Cardhldr=1 (aceptados para tener tarjeta de crédito) y default que no sea vacío.

Contiene 994 observaciones.

A continuación se detallará el código SAS para importar los datos y crear los datasets.

```

PROC IMPORT OUT= riesgo.cs_all
  DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumnos_OK_excel95.xls"
  DBMS=EXCEL5 REPLACE;
  GETNAMES=YES;
RUN;

PROC IMPORT OUT= riesgo.cs_all_without_new_clients
  DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumnos_OK_excel95_accepts_rejects_without_new_clients.xls"
  DBMS=EXCEL5 REPLACE;
  GETNAMES=YES;
RUN;

PROC IMPORT OUT= riesgo.cs_rejects
  DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumnos_OK_excel95_rejects.xls"
  DBMS=EXCEL5 REPLACE;
  GETNAMES=YES;
RUN;

PROC IMPORT OUT= riesgo.cs_accepts
  DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumnos_OK_excel95_accepts.xls"
  DBMS=EXCEL5 REPLACE;
  GETNAMES=YES;
RUN;
```

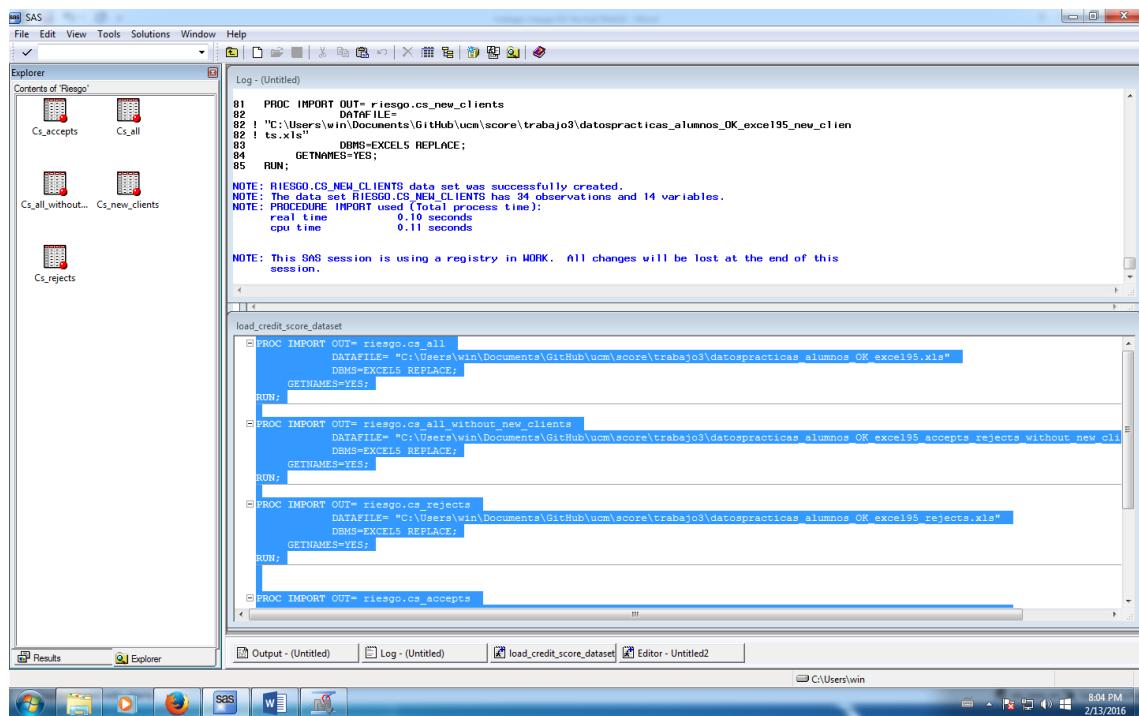
```

DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumn
os_OK_excel95_accepts.xls"
      DBMS=EXCEL5 REPLACE;
      GETNAMES=YES;
RUN;

PROC IMPORT OUT= riesgo.cs_new_clients
      DATAFILE=
"C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas_alumn
os_OK_excel95_new_clients.xls"
      DBMS=EXCEL5 REPLACE;
      GETNAMES=YES;
RUN;

```

Abajo la imagen de los 5 datasets creados.



1) Delimitación del estudio: objetivo, limitaciones y disponibilidad de datos

El objetivo del trabajo es construir el modelo de Scoring siguiendo la metodología SAS Miner, con tarjeta de puntuación del riesgo de crédito para clientes de tarjeta de crédito es posible predecir sobre los 34 nuevos clientes si se debe o no dar crédito.

Abajo hay informaciones importantes dadas en el enunciado del trabajo:

"En el Excel hay información sobre clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente). También hay clientes rechazados (Cardhldr=0) que son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na).

Por último al final del archivo hay 34 nuevos clientes (Cardhldr=na) individuos con identificador de cliente desde 1286 hasta 1319, que solicitan una tarjeta de crédito."

Notas:

Cardhldr = 1 = Ha recibido la tarjeta

Cardhldr = 0 = Clientes rechazados, no han recibido la tarjeta de crédito, tienen el default = na.

34 nuevos clientes con cardhldr = na de la línea 1286 hasta 1319 que solicitaron una tarjeta de crédito.

2) Análisis descriptivo exploratorio de los datos

¿Qué variables tenemos?

Hay un total de 14 variables existentes en el dataset y 1319 líneas (rows).

Son ellas:

- ID = Identificador
- Cardhldr = 1 ha recibido la tarjeta / 0 clientes rechazados
- Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)
- Age = Age n years plus twelfths of a year
- Income = Yearly income (divided by 10,000)
- Exp_Inc = Ratio of monthly credit card expenditure to yearly income
- Avgexp = Average monthly credit card expenditure
- Ownrent = 1 if owns their home, 0 if rent

- Selfempl = 1 if self employed, 0 if not.
- Depndt = 1 + number of dependents
- Inc_per = Income divided by number of dependents
- Cur_add = months living at current address
- Major = number of major credit cards held
- Active = number of active credit accounts

Abajo una tabla con las variables y el tipo de cada variable definido por mi.

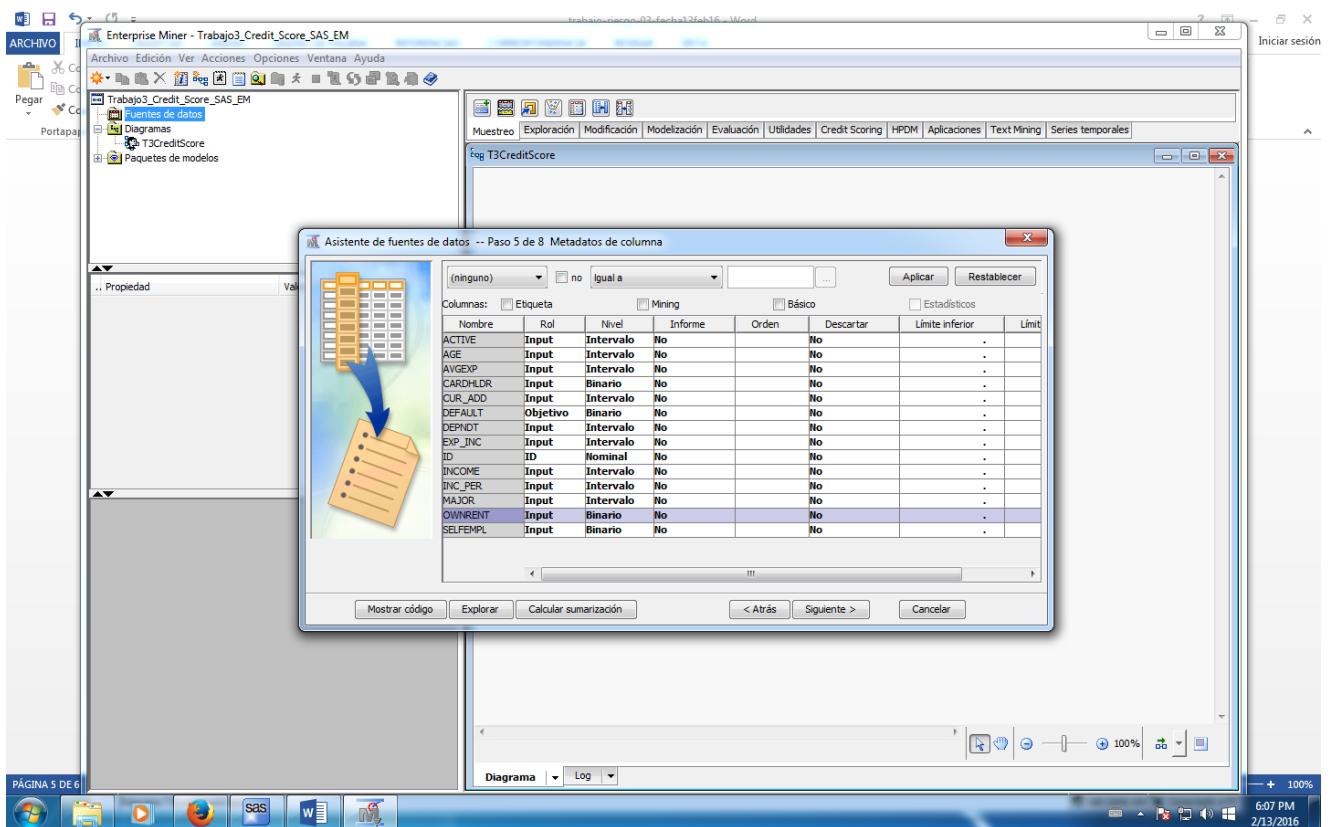
Variable	Tipo	Rol
ID	Nominal	ID
Cardhldr	Binaria	Input
Default	Binaria	Objetivo
Age	Intervalo	Input
Income	Intervalo	Input
Exp_Inc	Intervalo	Input
Avgexp	Intervalo	Input
Ownrent	Binaria	Input
Selfempl	Binaria	Input
Depndt	Intervalo	Input
Inc_per	Intervalo	Input
Cur_add	Intervalo	Input
Major	Intervalo	Input
Active	Intervalo	Input

¿Cuál es la variable objetivo?

La variable objetivo es la llamada **Default**.

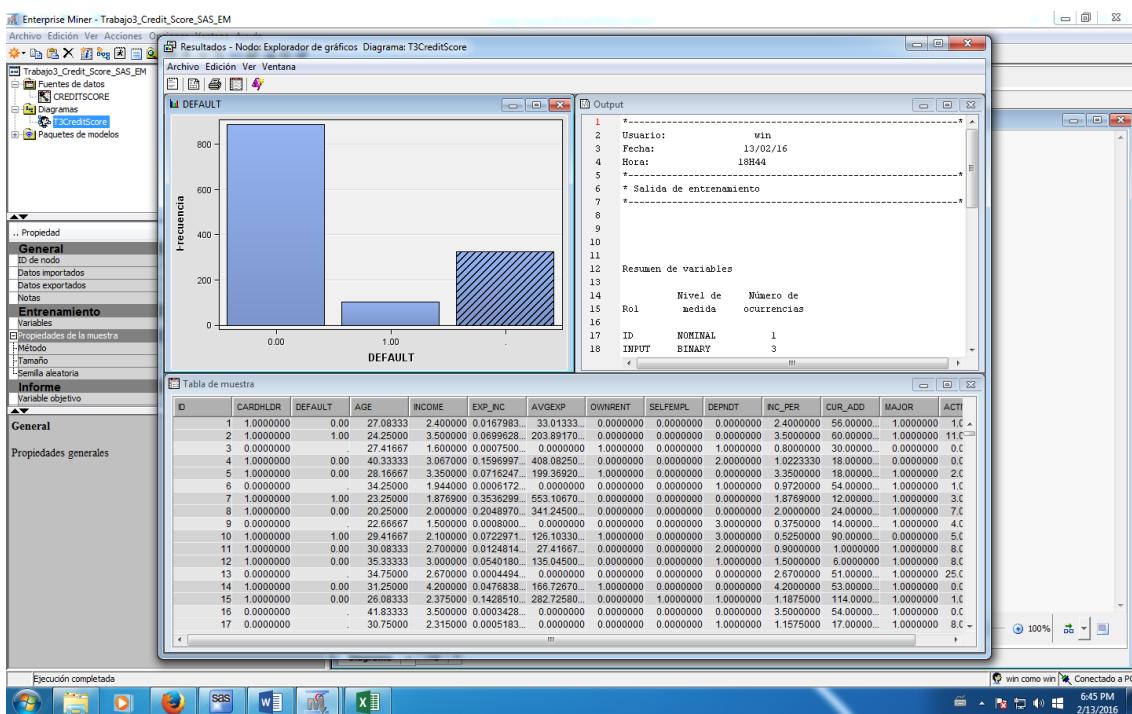
Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)

En la imagen abajo se puede ver como se han quedado las variables mapeadas en binarias, intervalo y nominales.



¿Qué porcentaje de buenos y malos tenemos en cada categoría?

En la imagen abajo se puede ver la cantidad de default = 0 (Buenos) y default = 1 (malos) y Default = Vacío. Estos datos son para todos los datos.



Los datos **buenos** son los que tienen default = 0 y hay 890 observaciones.

Los datos **malos** son los que tienen default = 1 y hay 104 observaciones.

Los con default vacío son 325 observaciones.

¿Hay datos anómalos?

La variable edad (age) se puede ver un poco rara, no tiene valores redondos, se recomienda transformar.

Si se puede ver en algunas variables una gran diferencia en los valores de mínimo, máximo y media.

En la imagen abajo se puede ver las estadísticas para tener una idea de los datos.

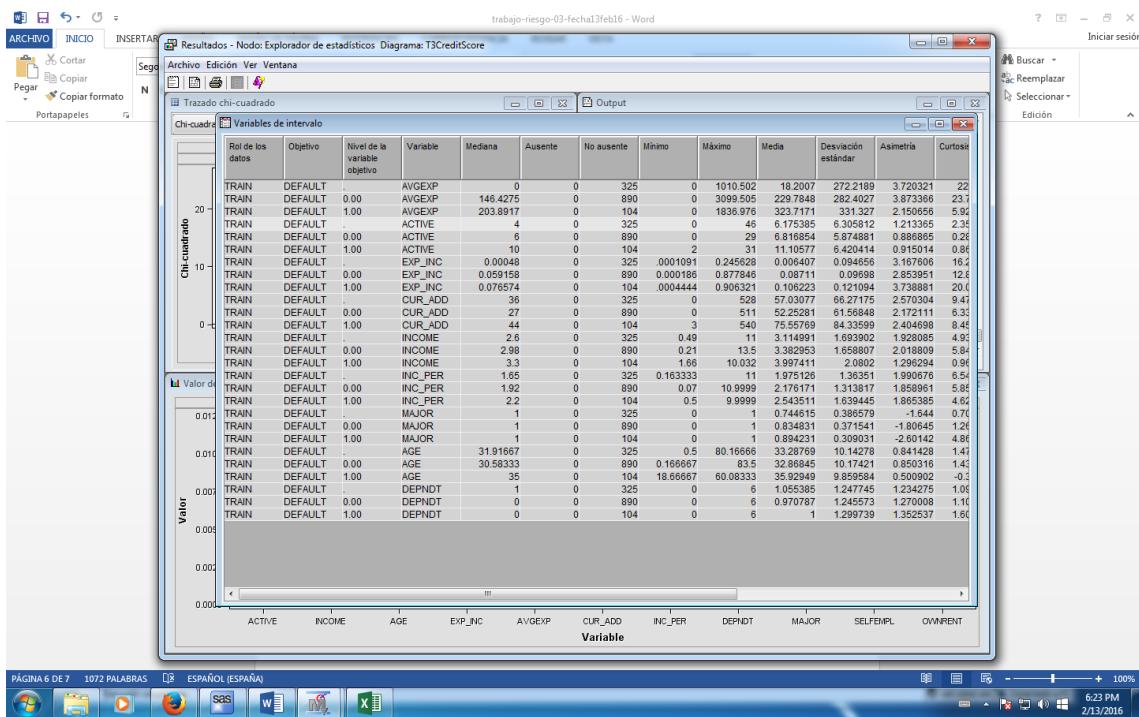
The screenshot shows a SAS Enterprise Miner interface with a 'Resultados' window open. The window displays statistical results for a dataset named 'T3CreditScore'. The table contains data for three roles: TRAIN, DEFAULT, and DEPNDT. The columns include: Rol de los datos, Objetivo, Nivel de la variable objetivo, Variable, Mediana, Ausente, No ausente, Mínimo, Máximo, Media, Desviación estándar, Asimetría, Curtosis, and Rol. The data shows various numerical values for variables like AGE, EXP_INC, and INC_PER, with some extreme values like 30.58333 and 31.91667 appearing in the AGE column for the TRAIN role.

¿Hay datos perdidos?

No hay datos nulos o perdidos en el dataset. Solo hay datos nulos en las variables Cardhldr y default pero está correcto.

Cuando el cliente es nuevo no hay datos en Cardhldr y default.

Cuando el cliente ha sido rechazado y no ha recibido la tarjeta de crédito no hay datos en la columna default.



¿Hay que imputar?

"En estadística, la imputación es la sustitución de valores no informados en una observación por otros.

A veces es un paso necesario para poder tratar los datos con determinadas técnicas estadísticas de análisis. Idealmente, este análisis debería tener en cuenta el hecho de que algunos de los datos no son observados sino que han sido imputados." (Wikipedia).

Yo he decidido no imputar datos.

3) ¿Hay que tramificar variables?

La técnica de tramificar variables es transformar los valores en tramos.

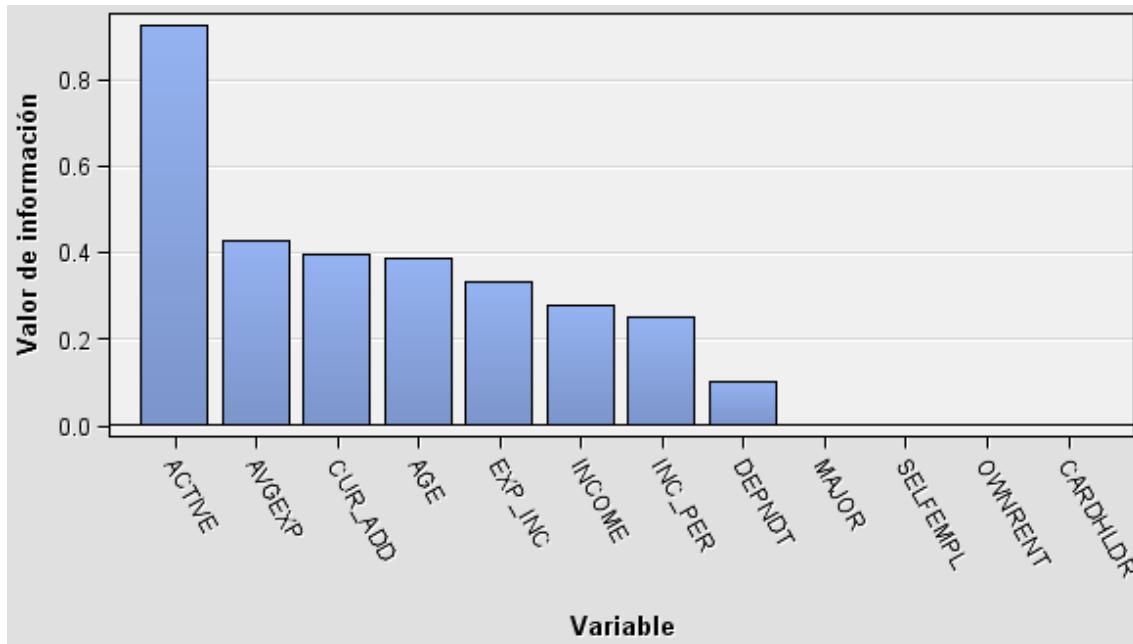
Ejemplo:

Original	Tramificado
1,2	1
3,8	4

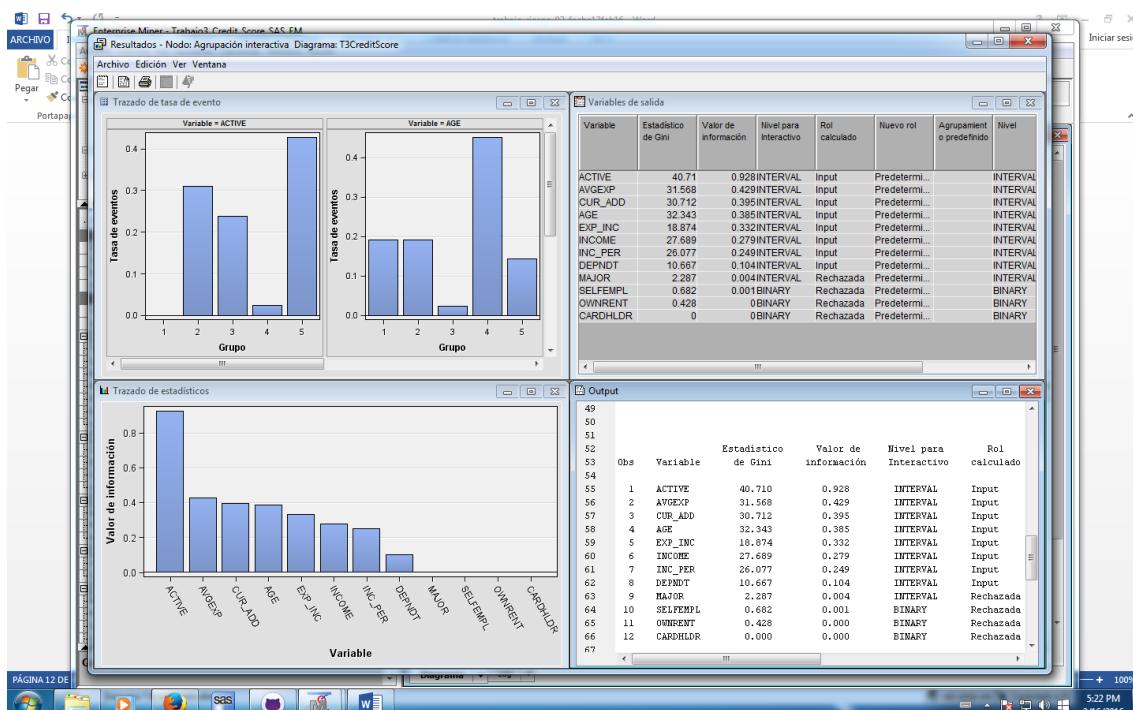
Yo no he tramificado ninguna variable del dataset, pero quizás se puede recomendar tramificar la variable edad (age) que tiene valores no enteros.

4) Selección de variables ¿qué variables son las más importantes?

Utilizando el nodo "Agrupación interactiva" se tiene las variables más importantes por valor de información como se puede ver en la imagen abajo.



En la imagen abajo se puede ver también las tablas con los estadísticos de Gini y valor de información.



¿qué variables deben incluirse en el scorecard?

Yo he puesto todas las variables en el scorecard, pero se podría quitar las que MAJOR, SELFEMPL, OWNRENT y CARDHLDLR.

Son ellas:

- ID = Identificador
- Cardhldr = 1 ha recibido la tarjeta / 0 clientes rechazados
- Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)
- Age = Age n years plus twelfths of a year
- Income = Yearly income (divided by 10,000)
- Exp_Inc = Ratio of monthly credit card expenditure to yearly income
- Avgexp = Average monthly credit card expenditure
- Ownrent = 1 if owns their home, 0 if rent
- Selfempl = 1 if self employed, 0 if not.
- Depndt = 1 + number of dependents
- Inc_per = Income divided by number of dependents
- Cur_add = months living at current address
- Major = number of major credit cards held
- Active = number of active credit accounts

Abajo una tabla con las variables y el tipo de cada variable definido por mí.

Variable	Tipo	Rol
ID	Nominal	ID
Cardhldr	Binaria	Input
Default	Binaria	Objetivo
Age	Intervalo	Input
Income	Intervalo	Input
Exp_Inc	Intervalo	Input
Avgexp	Intervalo	Input
Ownrent	Binaria	Input
Selfempl	Binaria	Input
Depndt	Intervalo	Input
Inc_per	Intervalo	Input
Cur_add	Intervalo	Input
Major	Intervalo	Input
Active	Intervalo	Input

Disponibilidad, e importancia, estadísticos de selección (IV, correlación)

Para ver la correlación en SAS:

```
data cs_accepts;
set riesgo.Cs_accepts;
run;
```



```
proc corr data=cs_accepts outp=cs_accepts_corr;
run;
```

The CORR Procedure

14 Variables:	ID CARDHLD R DEFAULT AGE INCOME EXP_INC AVGEXP OWNRENT SELFEMPL DEPNDT INC_PER CUR_ADD MAJOR ACTIVE
----------------------	---

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ID	994	643.06237	366.55474	639204	1.00000	1285	ID
CARDHLD R	994	1.00000	0	994.00000	1.00000	1.00000	CARDHLD R
DEFAULT	994	0.10463	0.30623	104.00000	0	1.00000	DEFAULT
AGE	994	33.18872	10.18014	32990	0.16667	83.50000	AGE
INCOME	994	3.44724	1.71689	3427	0.21000	13.50000	INCOME
EXP_INC	994	0.08911	0.09988	88.57550	0.0001860	0.90632	EXP_INC
AVGEXP	994	239.61272	289.15890	238175	0	3100	AVGEXP
OWNRENT	994	0.47988	0.49985	477.00000	0	1.00000	OWNRENT
SELFEMPL	994	0.06137	0.24013	61.00000	0	1.00000	SELFEMPL
DEPNDT	994	0.97384	1.25071	968.00000	0	6.00000	DEPNDT
INC_PER	994	2.21460	1.35528	2201	0.07000	10.99990	INC_PER
CUR_ADD	994	54.69115	64.67114	54363	0	540.00000	CUR_ADD
MAJOR	994	0.84105	0.36582	836.00000	0	1.00000	MAJOR
ACTIVE	994	7.26559	6.07455	7222	0	31.00000	ACTIVE

SAS - [Results Viewer - SAS Output]

File Edit View Go Tools Solutions Window Help

ACTIVE 994 7.26559 6.07455 7222 0 31.00000 ACTIVE

Pearson Correlation Coefficients, N = 994
Prob > |r| under H0: Rho=0

	ID	CARDHLDR	DEFAULT	AGE	INCOME	EXP_INC	AVGEXP	OWNRENT	SELFEMPL	DEPNDT	INC_PER	CUR_ADD	MAJOR	ACTIVE
ID	1.00000	-0.01821	0.03903	0.01337	-0.02885	-0.00988	0.01833	0.02518	0.01555	-0.01064	0.03387	-0.01938	-0.02776	
ID	0.5664	0.2084	0.6738	0.3978	0.7556	0.5638	0.4277	0.6243	0.7375	0.2860	0.5420	0.3820		
CARDHLDR	-	-	-	-	-	-	-	-	-	-	-	-	-	
CARDHLDR	-	-	-	-	-	-	-	-	-	-	-	-	-	
DEFAULT	-0.01821	1.00000	0.09208	0.10960	0.05880	0.09948	0.03350	0.03858	0.00715	0.08300	0.11035	0.04972	0.21621	
DEFAULT	0.5664	0.0037	0.0037	0.0048	0.0017	0.2913	0.8218	0.0088	0.0005	0.1172			<0.001	
AGE	0.03993	-	0.09208	1.00000	0.35963	-0.14424	0.01891	0.37650	0.12520	0.23717	0.06562	0.44811	-0.01017	
AGE	0.0037	-	<0.001	<0.001	0.0004	<0.001	<0.001	<0.001	0.0388	<0.001	0.7488		<0.001	
INCOME	0.01337	-	0.10960	0.35963	1.00000	-0.11281	0.29814	0.33674	0.13100	0.35628	0.46210	0.12688	0.10801	
INCOME	0.6738	-	0.0005	<0.001	-	0.0004	<0.001	<0.001	<0.001	<0.001	<0.001	0.0006	<0.001	
EXP_INC	-0.02885	-	0.05880	-0.14424	-0.11281	1.00000	0.81104	-0.09175	-0.08029	-0.08404	-0.03206	-0.06669	0.00752	
EXP_INC	0.3978	-	0.0004	<0.001	0.0004	-	<0.001	0.0038	0.0113	0.0080	0.3127	0.0355	0.8128	
AVGEXP	-0.00988	-	0.09948	0.01891	0.29814	0.81104	1.00000	0.04762	-0.02949	0.08359	0.14734	-0.03334	0.04451	
AVGEXP	0.7556	-	0.0017	0.5515	<0.001	<0.001	-	0.1338	0.3531	0.0084	<0.001	0.2936	0.1609	
OWNRENT	0.01833	-	0.03350	0.37650	0.33674	-0.09175	0.04762	1.00000	0.07322	0.34066	-0.03249	0.26120	0.03206	
OWNRENT	0.5638	-	0.2913	<0.001	<0.001	0.0038	0.1338	-	0.0210	<0.001	0.3081	<0.001	0.3126	
SELFEMPL	0.02518	-	0.03585	0.12520	0.13100	-0.08029	0.02849	0.07322	1.00000	0.04894	0.09005	0.08384	0.00798	
SELFEMPL	0.4277	-	0.2588	<0.001	<0.001	0.0113	0.3531	0.0210	-	0.1231	0.0045	0.0082	0.8016	
DEPNDT	0.01556	-	0.00715	0.23717	0.35628	-0.08404	0.08358	0.04066	0.04894	1.00000	-0.54569	0.09124	0.00851	
DEPNDT	0.6243	-	0.8218	<0.001	<0.001	0.0080	0.0084	<0.001	0.1231	-	<0.001	0.0040	0.7887	
INC_PER	-0.01064	-	0.08300	0.06562	0.46210	-0.03206	0.14734	-0.03249	0.08005	-0.54569	1.00000	-0.03933	0.07790	
INC_PER	0.7375	-	0.0008	0.0366	<0.001	0.3127	<0.001	0.3061	0.0045	<0.001	0.9016	0.0140	0.1525	
CUR_ADD	0.03387	-	0.11035	0.44811	0.12888	-0.06669	-0.03334	0.26120	0.08384	0.09124	-0.03933	1.00000	-0.03473	
CUR_ADD	0.2860	-	0.0005	<0.001	<0.001	0.0355	0.2936	<0.001	0.0082	0.0040	0.9016		0.2740	
MAJOR	-0.01936	-	0.04972	-0.10170	0.10801	0.00752	0.04451	0.03206	0.00798	0.00851	0.07790	-0.03473	1.00000	
MAJOR	0.5420	-	0.1172	0.7488	0.0006	0.8128	0.1609	0.3126	0.8016	0.7887	0.0140	0.2740	<0.001	
ACTIVE	-0.02776	-	0.21621	0.18349	0.20150	-0.07278	0.03214	0.27273	0.03576	0.13700	0.04541	0.12771	1.04002	
ACTIVE	0.3820	-	<0.001	<0.001	<0.001	0.0217	0.3115	<0.001	0.2600	<0.001	0.1525	<0.001	<0.001	

Results Explorer Log Editor Results Viewer - SAS ...

Finalizado 7:33 PM 2/16/2016

Enterprise Miner - Trabajo3_Credit_Score_SAS_EM

Archivo Edición Ver Acciones Opciones

Trabajo3_Credit_Score_SAS_EM

Fuentes de datos CREDITSCORE Diagramas T3CreditScore Paquetes de modelos

Efectos en el modelo

R cuadrado secundario

Efecto

R2 Valores

Efecto

Selección de variables

Nombre de la variable Rol Nivel de medida Tipo Etiqueta Motivo del rechazo

ACTIVE Input Intervalo Numérico ACTIVE

AGE Rechazado Intervalo Numérico AGE VarSelvalor...

AVGEXP Input Rechazado Intervalo Numérico AVGEXP VarSelvalor...

CARDHLDR Rechazado Input Intervalo Numérico CARDHLDR VarSelvalor...

CUR_ADD Rechazado Input Intervalo Numérico CUR_ADD VarSelvalor...

DEPNDT Rechazado Intervalo Numérico DEPNDT VarSelvalor...

EXP_INC Rechazado Intervalo Numérico EXP_INC VarSelvalor...

INCOME Rechazado Intervalo Numérico INCOME VarSelvalor...

INC_PER Input Intervalo Numérico INC_PER VarSelvalor...

MAJOR Rechazado Intervalo Numérico MAJOR VarSelvalor...

OWNRENT Rechazado Binario Numérico OWNRENT VarSelvalor...

SELFEMPL Rechazado Binario Numérico SELFEMPL VarSelvalor...

Output

1 -----

2 Usuario: win

3 Fecha: 13/02/16

4 Hora: 18H42

5 -----

6 * Salida de entrenamiento

7 -----

8 -----

9 -----

10 -----

11 -----

12 Resumen de variables

13 -----

14 Nivel de medida Número de ocurrencias

15 Rol BINARY 3

16 INPUT INTERVAL 9

17 INPUT IMPLICIT

18 INPUT INTERVAL

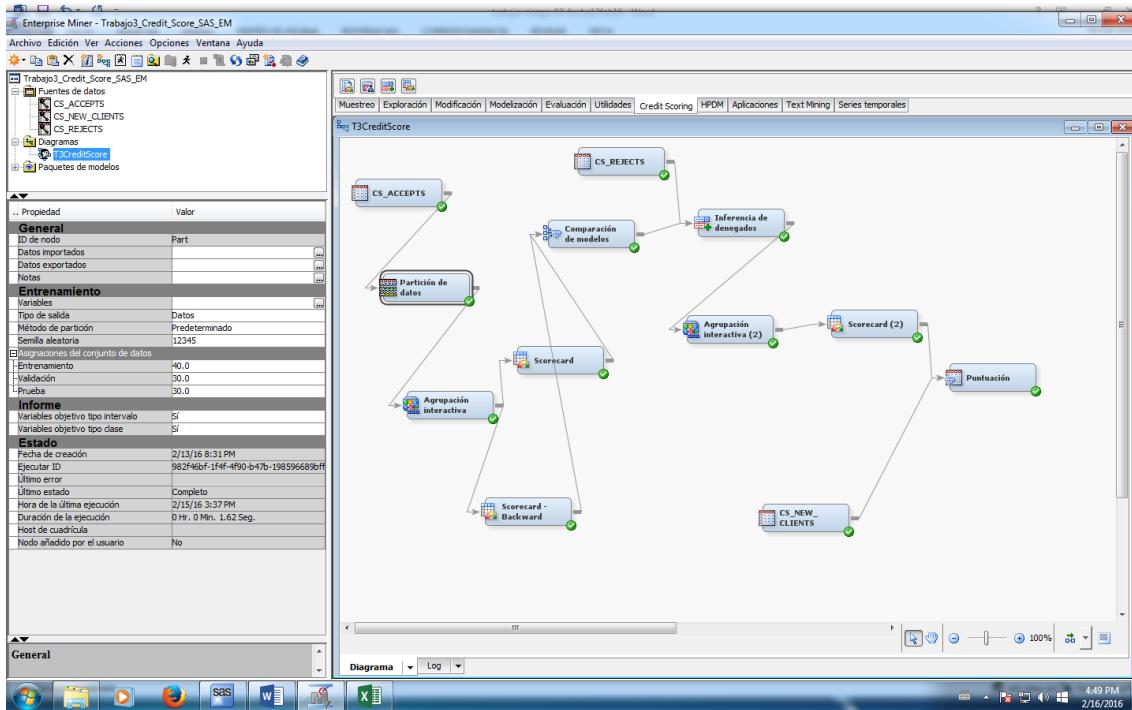
Ejecución completada

wm como win Conectado a PC 6:43 PM 2/13/2016

5) Muestreo:

Muestra de desarrollo/entrenamiento (70-80%) y de validación (30-20%)

Sobre muestreo de buenos/malos?



The screenshot shows the properties for the 'Partición de datos' node. The 'Output' tab is selected, displaying the following descriptive statistics:

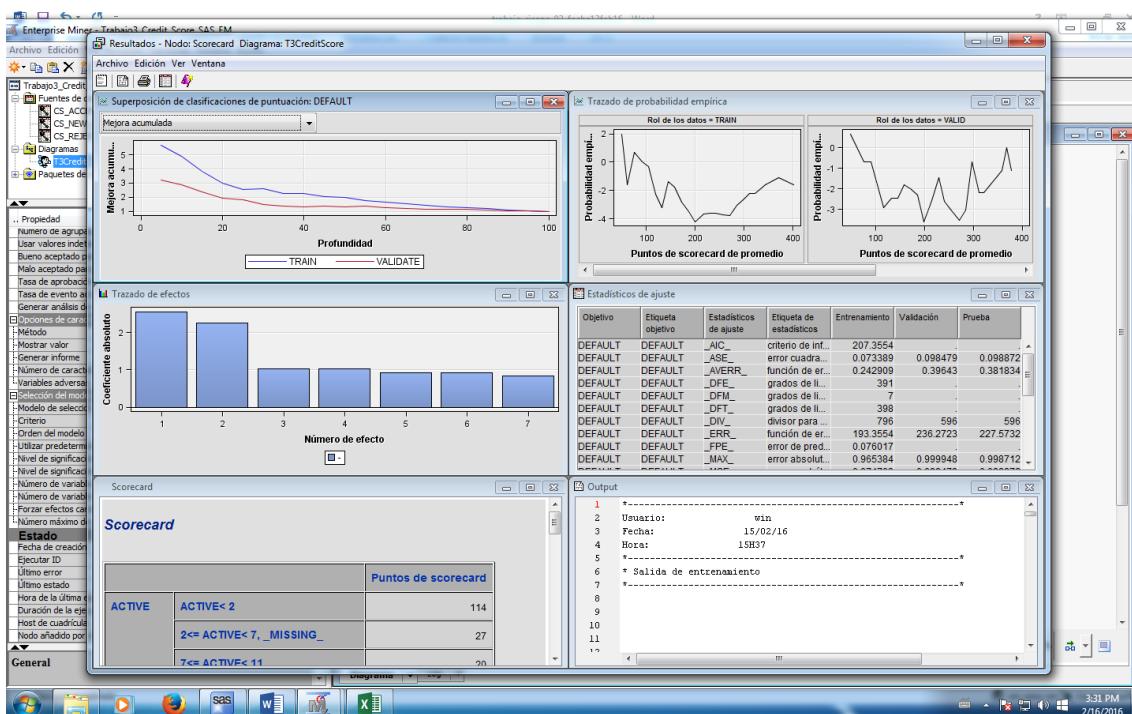
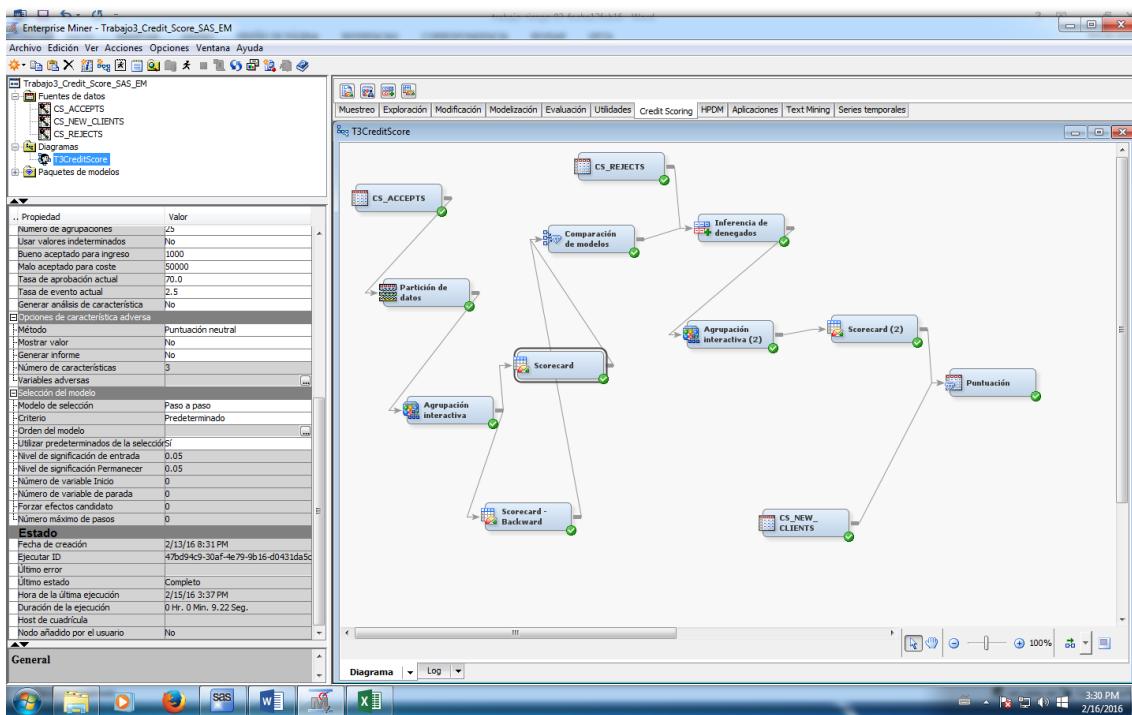
	Valor	Valor	Número de	Porcentaje	Etiqueta
Variable	númerico	formateado	ocurrencias		
46					
47					
48	Estadísticos descriptivos para las variables objetivo de clase				
49	Datos=DATA				
50					
51					
52					
53					
54					
55	DEFAULT	0	0.00	890	89.5372
56	DEFAULT	1	1.00	104	10.4628
57					
58					
59	Datos=TEST				
60					
61					
62	Variable	Valor	Valor	Número de	
63	númerico	formatado	formateado	ocurrencias	
64	64	0	0.00	267	89.5973
65	DEFAULT	1	1.00	31	10.4027
66					
67					
68	Datos=TRAIN				
69					
70	Variable	Valor	Valor	Número de	
71	númerico	formatado	formateado	ocurrencias	
72	73	DEFAULT	0	0.00	356
73	74	DEFAULT	1	1.00	42
74	75				10.5526
75					DEFAULT
76					
77	Datos=VALIDATE				
78					
79	Variable	Valor	Valor	Número de	
80	númerico	formatado	formateado	ocurrencias	
81	82	DEFAULT	0	0.00	267
82	83	DEFAULT	1	1.00	31
83	84				10.4027
84					DEFAULT

6) Estimación del primer modelo, diagnosis (Kolmogorov-Smirnov, c-statistic, Gini) y obtención del scorecard preliminar

En la imagen abajo se puede ver el diagnosis del primer modelo.

The screenshot displays two windows from the SAS Enterprise Miner interface:

- Results - Node: Comparison of models Diagram: T3CreditScore**: This window shows a table titled "Estadísticos de ajuste" (Adjustment statistics) comparing various statistical methods. The columns include metrics like Entrenar: estadístico Kolmogorov-Smirnov de dos factores basado en variables categorizadas, Entrenar: corte de probabilidad, Valida: índice Roc, Valida: coeficiente de Gini, etc. The table has two rows of data, both showing values such as 0.562, 0.079, 0.645, 0.29, 0.254, 0.34, 0.227, 0.338, 0.625, 0.251, 0.238, 0.07, 0.235, 0.104, 385.6548, and 188.
- Enterprise Miner - Trabajo3_Credit_Score_SAS_EM**: This window shows the flow of the data mining process. It includes a sidebar with project files like CS_ACCEPTS, CS_NEW_CLIENTS, CS_REJECTS, and a node titled "Scorecard". The main area shows a flowchart where data from CS_ACCEPTS and CS_REJECTS feeds into a "Comparación de modelos" (Comparison of models) node. This node then connects to "Inferencia de denegados" (Inference of denegated), "Scorecard", and "Agrupación interactiva (2)". The "Scorecard" node connects to "Scorecard (2)", which then connects to "Puntuación". A "CS_NEW_CLIENTS" node also feeds into the "Scorecard (2)" node.



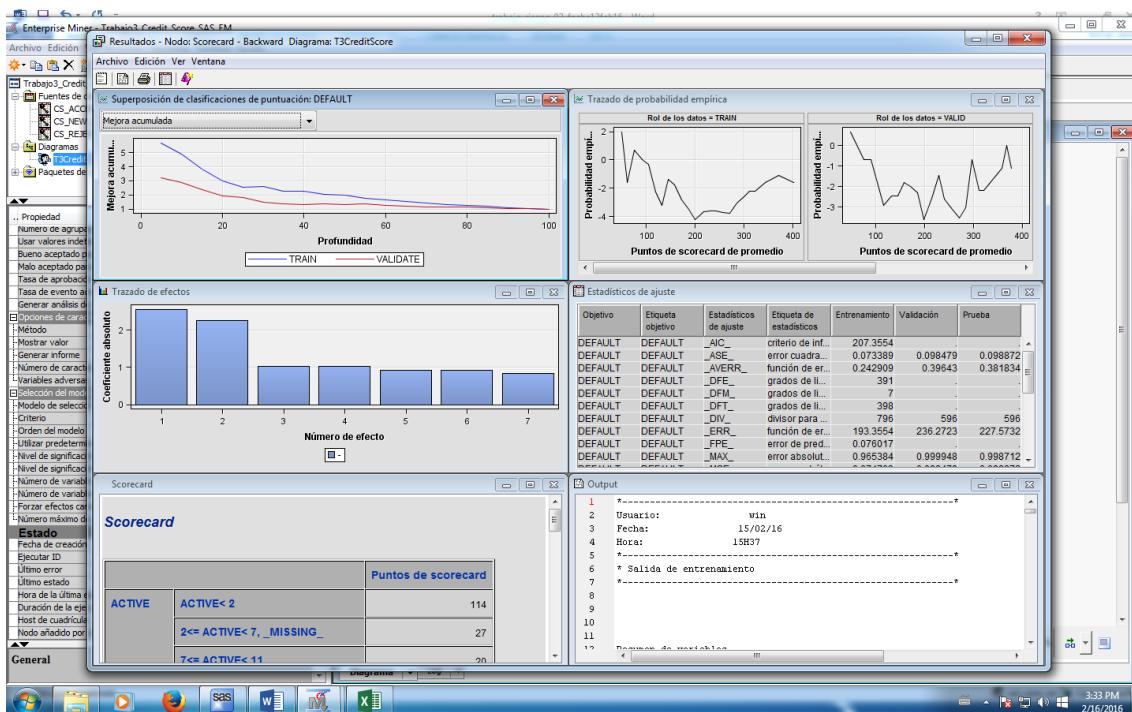
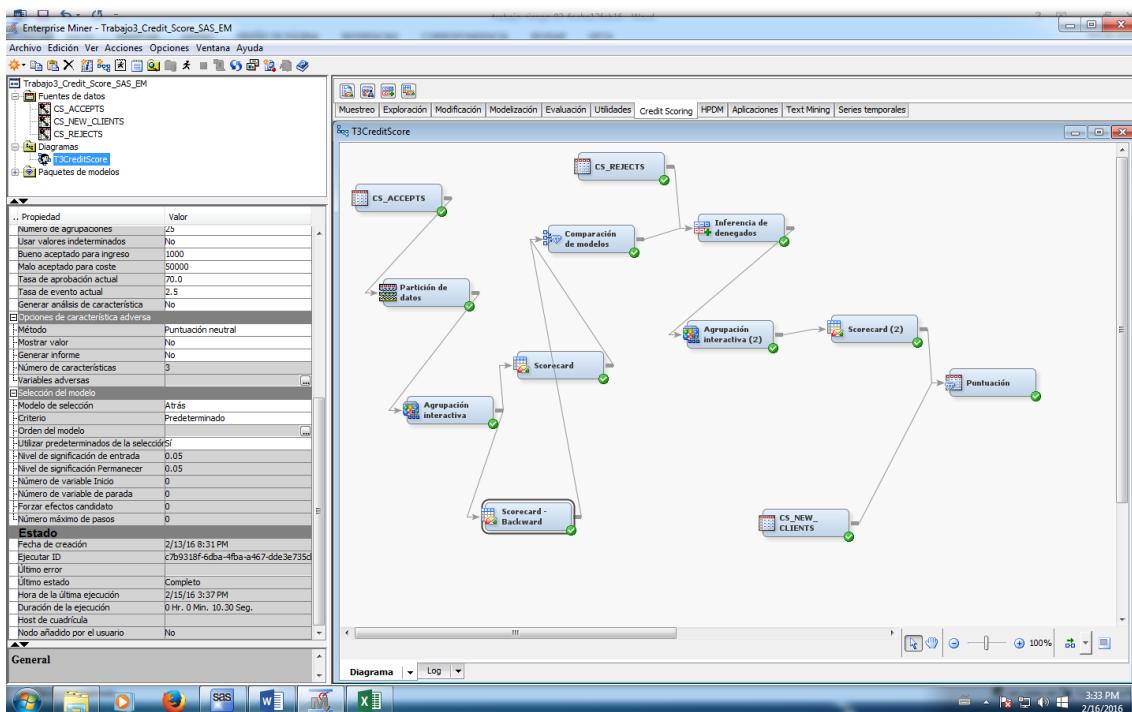
Scorecard

		Puntos de scorecard
ACTIVE	ACTIVE< 2	114
	2<= ACTIVE< 7, _MISSING_	27
	7<= ACTIVE< 11	20
	11<= ACTIVE< 12	51
	12<= ACTIVE	1
AGE	AGE< 27.75, _MISSING_	43
	27.75<= AGE< 29.83	6
	29.83<= AGE< 32.33	65
	32.33<= AGE< 44.92	13
	44.92<= AGE	26
AVGEXP	AVGEXP< 87.15, _MISSING_	17
	87.15<= AVGEXP< 132.85	40
	132.85<= AVGEXP< 175.44	91
	175.44<= AVGEXP< 420.58	33
	420.58<= AVGEXP	3
CUR_ADD	CUR_ADD< 6	59

Scorecard

		Puntos de scorecard
AVGEXP	AVGEXP< 87.15, _MISSING_	17
	87.15<= AVGEXP< 132.85	40
	132.85<= AVGEXP< 175.44	91
	175.44<= AVGEXP< 420.58	33
	420.58<= AVGEXP	3
CUR_ADD	CUR_ADD< 6	59
	6<= CUR_ADD< 12	18
	12<= CUR_ADD< 24	52
	24<= CUR_ADD< 94, _MISSING_	14
	94<= CUR_ADD	30
DEPNDT	DEPNDT< 1, _MISSING_	15
	1<= DEPNDT< 2	24
	2<= DEPNDT< 3	40
	3<= DEPNDT< 4	19
	4<= DEPNDT	142
INCOME	INCOME< 1.68	47
	1.68<= INCOME< 2.1	5
	2.1<= INCOME< 3.11, _MISSING_	27
	3.11<= INCOME< 5.92	40
	5.92<= INCOME	1

Scorecard Backward

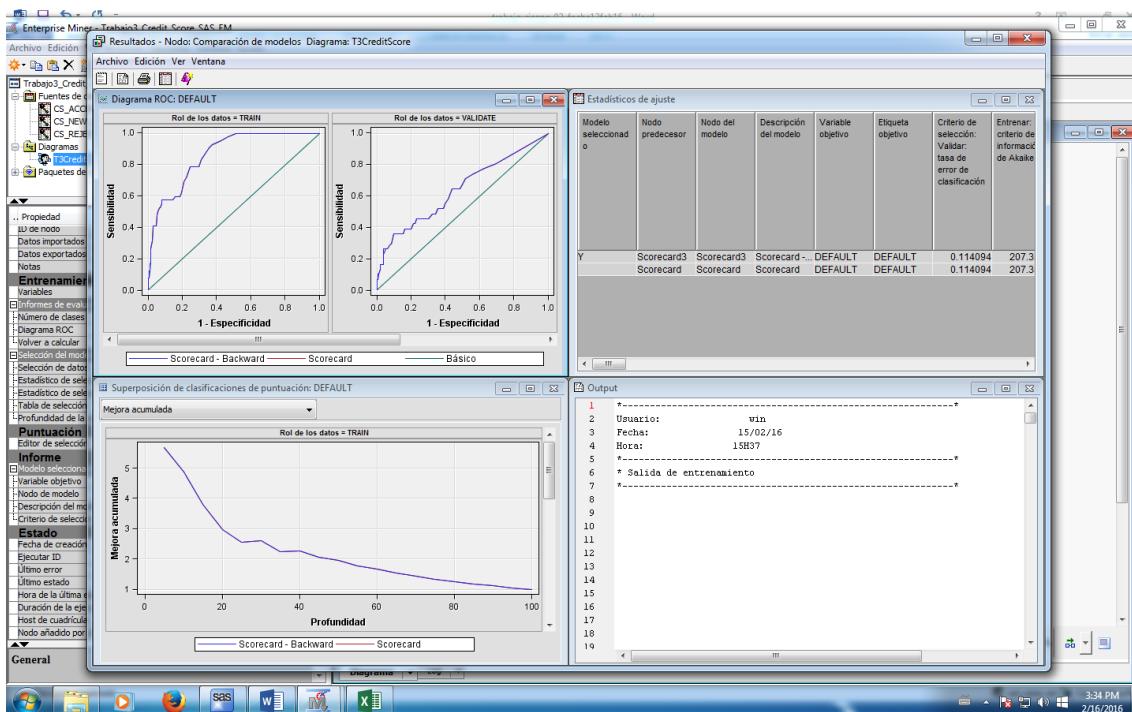


Scorecard

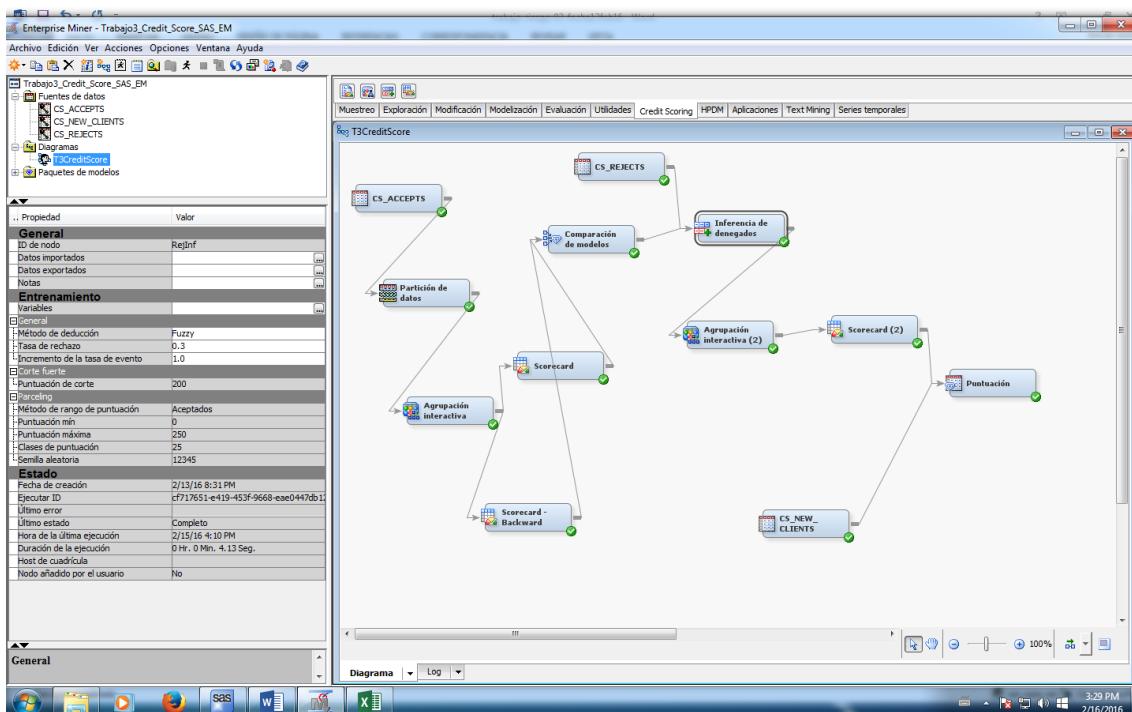
		Puntos de scorecard
ACTIVE	ACTIVE< 2	114
	2<= ACTIVE< 7, _MISSING_	27
	7<= ACTIVE< 11	20
	11<= ACTIVE< 12	51
	12<= ACTIVE	1
AGE	AGE< 27.75, _MISSING_	43
	27.75<= AGE< 29.83	6
	29.83<= AGE< 32.33	65
	32.33<= AGE< 44.92	13
	44.92<= AGE	26
AVGEXP	AVGEXP< 87.15, _MISSING_	17
	87.15<= AVGEXP< 132.85	40
	132.85<= AVGEXP< 175.44	91
	175.44<= AVGEXP< 420.58	33
	420.58<= AVGEXP	3
CUR_ADD	CUR_ADD< 6	59

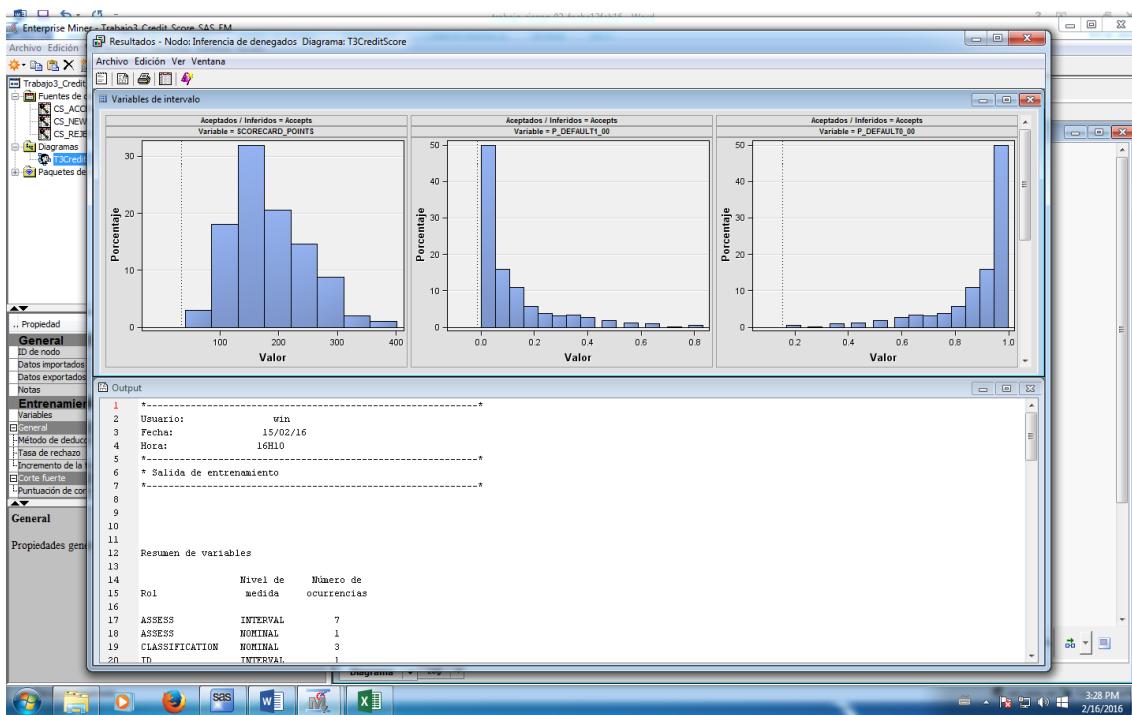
Scorecard

		Puntos de scorecard
AVGEXP	AVGEXP< 87.15, _MISSING_	17
	87.15<= AVGEXP< 132.85	40
	132.85<= AVGEXP< 175.44	91
	175.44<= AVGEXP< 420.58	33
	420.58<= AVGEXP	3
CUR_ADD	CUR_ADD< 6	59
	6<= CUR_ADD< 12	18
	12<= CUR_ADD< 24	52
	24<= CUR_ADD< 94, _MISSING_	14
	94<= CUR_ADD	30
DEPNDT	DEPNDT< 1, _MISSING_	15
	1<= DEPNDT< 2	24
	2<= DEPNDT< 3	40
	3<= DEPNDT< 4	19
	4<= DEPNDT	142
INCOME	INCOME< 1.68	47
	1.68<= INCOME< 2.1	5
	2.1<= INCOME< 3.11, _MISSING_	27
	3.11<= INCOME< 5.92	40
	5.92<= INCOME	1

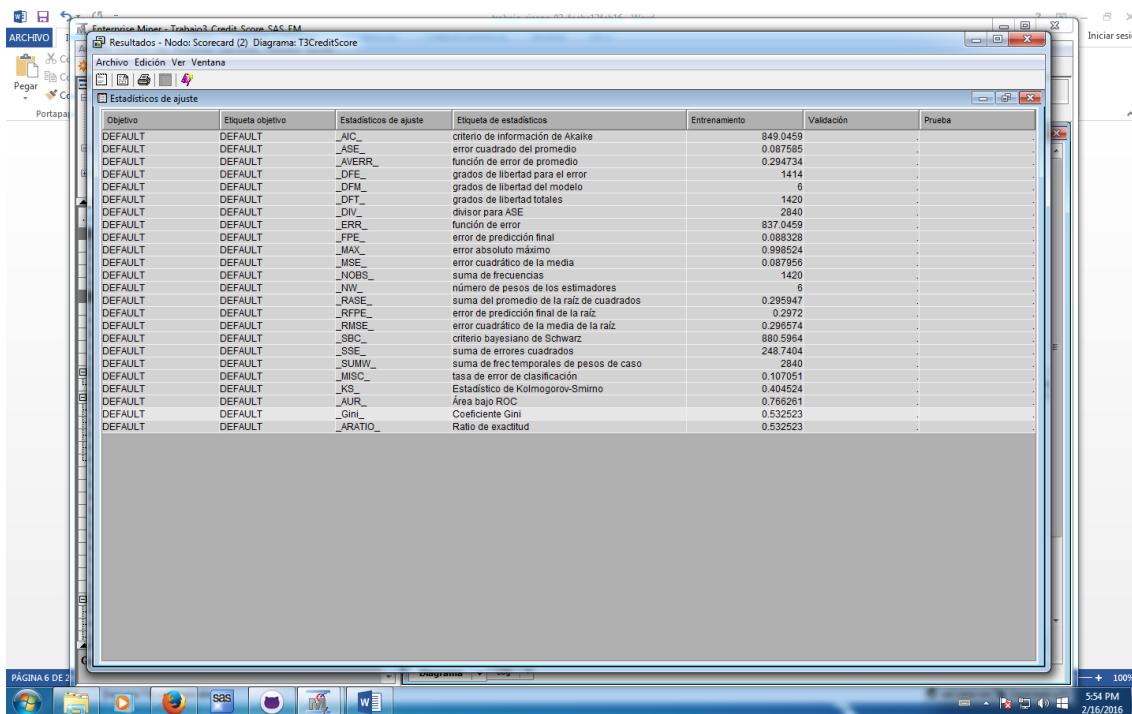


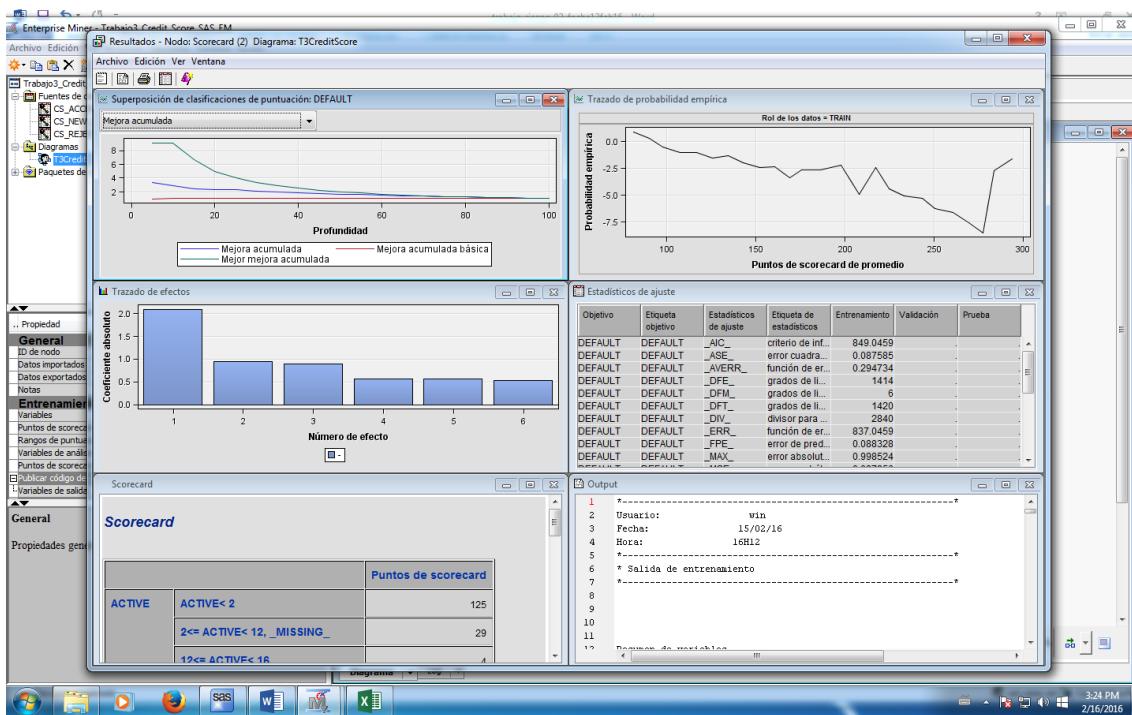
7) Inferencia de denegados





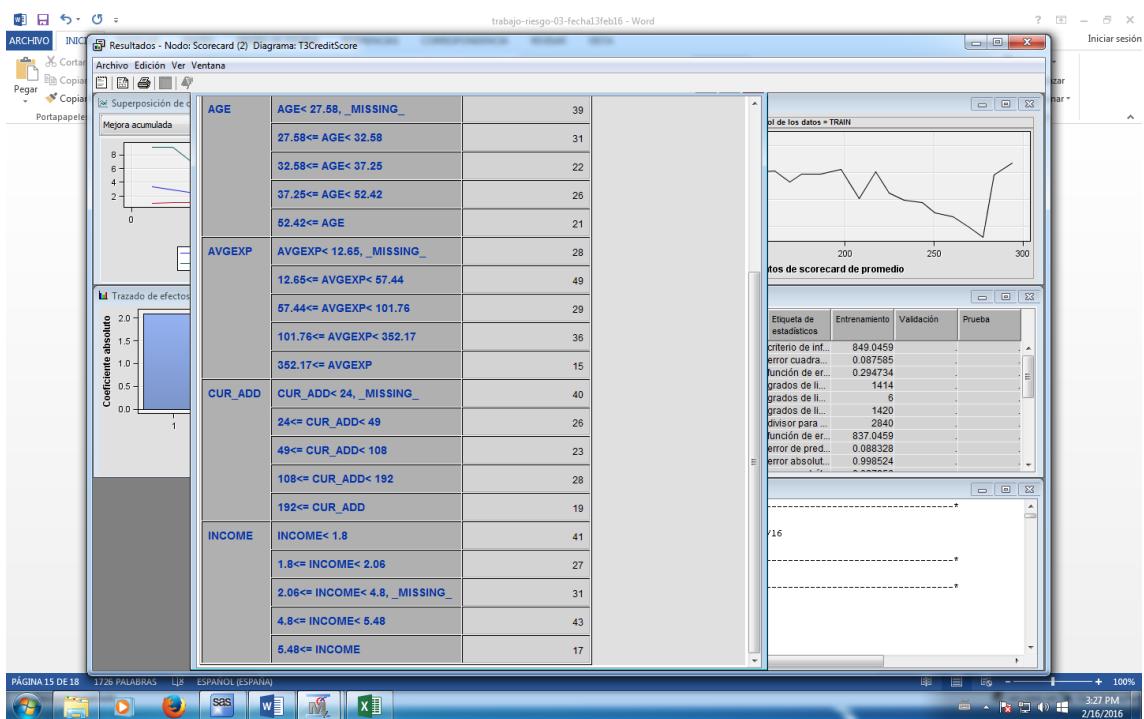
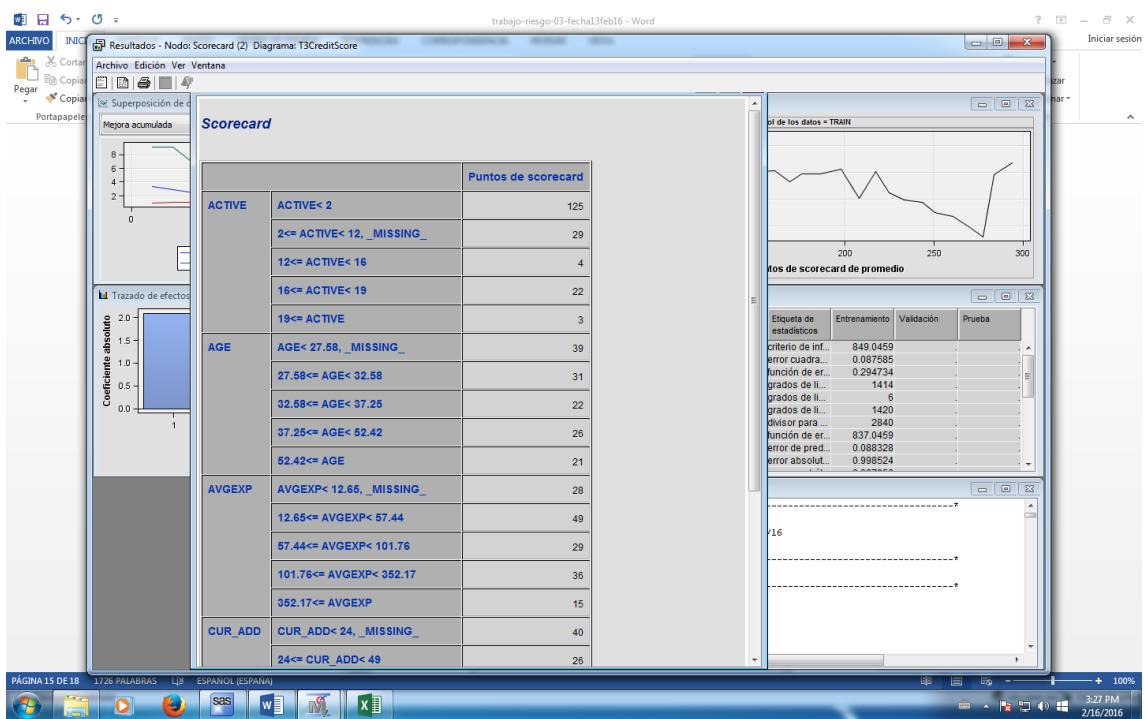
8) Estimación del modelo definitivo y obtención del scorecard definitivo





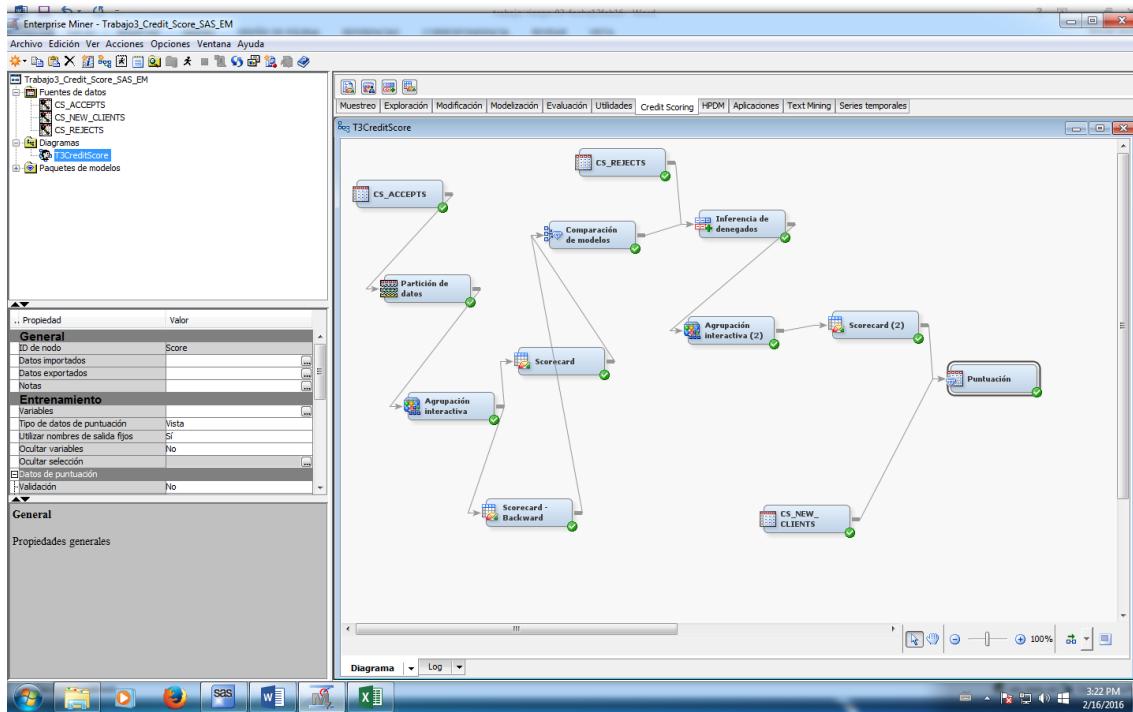
Scorecard definitivo

Scorecard		Puntos de scorecard
ACTIVE	ACTIVE< 2	125
	2<= ACTIVE< 12, _MISSING_	26
	12<= ACTIVE< 16	4
	16<= ACTIVE< 19	35
	19<= ACTIVE	4
AGE	AGE< 27.58, MISSING	35
	27.58<= AGE< 32.58	35
	32.58<= AGE< 37.25	32
	37.25<= AGE< 52.42	26
	52.42<= AGE	26
AVGEXP	AVGEXP< 12.65, _MISSING_	35
	12.65<= AVGEXP< 57.44	35
	57.44<= AVGEXP< 101.76	35
	101.76<= AVGEXP< 352.17	35
	352.17<= AVGEXP	15
CUR_ADD	CUR_ADD< 24, _MISSING_	45
	24<= CUR_ADD< 49	26
	49<= CUR_ADD< 108	26
	108<= CUR_ADD< 192	26
	192<= CUR_ADD	16
INCOME	INCOME< 1.8	41
	1.8<= INCOME< 2.06	27
	2.06<= INCOME< 4.8, _MISSING_	31
	4.8<= INCOME< 5.48	45
	5.48<= INCOME	17

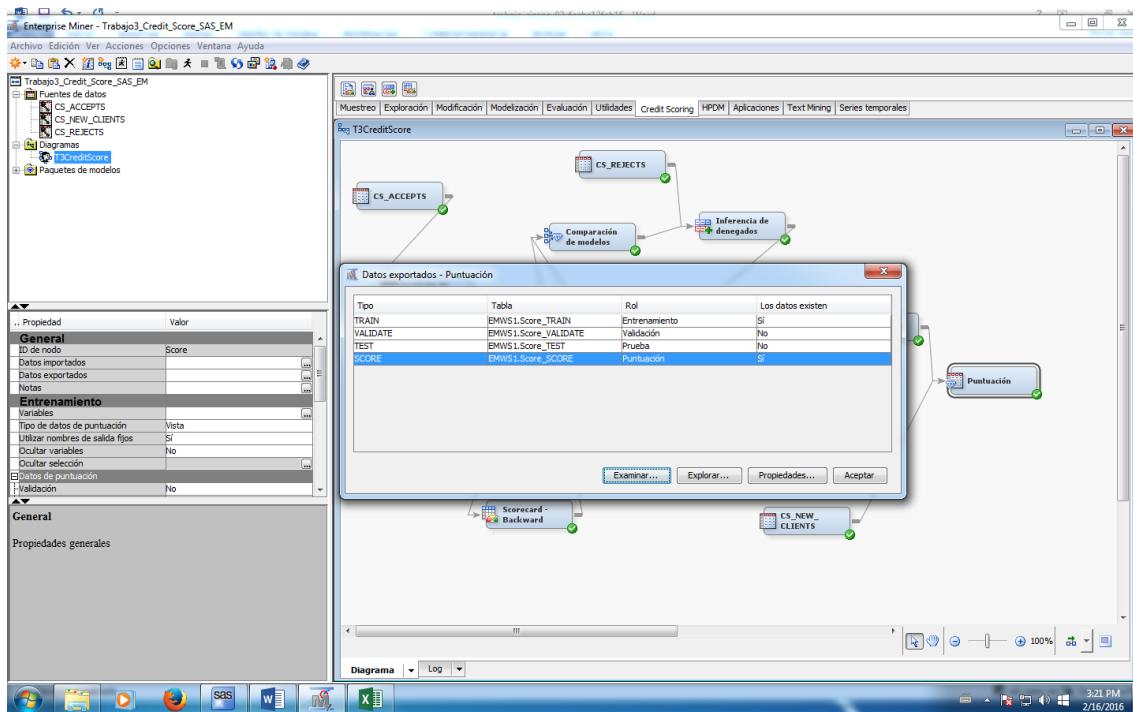


9) Validación y Seguimiento

Abajo se puede ver todo el flujo completo para generar la tabla de puntuación con SAS para los 34 clientes.



Para ver los resultados de la tabla de puntuación en SAS hay que ir en datos exportados y pinchar en SCORE > Examinar.



La tabla de puntuación generada por SAS para los 34 clientes ha sido:

ID	SCORECARD_POINTS	SCR_ACTIVE	SCR_AGE	SCR_AVGEXP	SCR_CUR_ADD	SCR_DEPDNT	SCR_INCOME	SCORECARD_BIN	b_DEFAULT	Segmento	Probability for level 1.00 of DEFAULT	Probability of Classification	Prediction for DEFAULT
1	252,0	125,0	39,0	29,0	28,0	15,0	31,0	20,0	17,0	20,0	0,00344270731309443	0,9965572926869055	0,00
2	115,0	29,0	31,0	15,0	23,0	15,0	17,0	5,0	9,0	5,0	0,2735939077224464	0,726640692277553	0,00
3	184,0	4,0	31,0	36,0	40,0	40,0	43,0	9,0	16,0	9,0	0,0893072841139831	0,910692159566017	0,00
4	264,0	125,0	39,0	29,0	40,0	15,0	31,0	21,0	17,0	21,0	0,002236100283152039	0,997763899716848	0,00
5	121,0	4,0	31,0	36,0	23,0	15,0	27,0	6,0	12,0	6,0	0,2366227627657738	0,763337723732422	0,00
6	123,0	29,0	22,0	29,0	26,0	15,0	17,0	6,0	4,0	6,0	0,2291091501225125	0,770909498775487	0,00
7	151,0	29,0	22,0	29,0	40,0	19,0	31,0	9,0	9,0	9,0	0,1011417912570153	0,8988582087429847	0,00
8	125,0	4,0	26,0	36,0	28,0	40,0	31,0	6,0	9,0	6,0	0,21733573554521018	0,762644244547699	0,00
9	230,0	125,0	26,0	36,0	26,0	40,0	17,0	15,0	14,0	18,0	0,00713673656708452	0,9928632363432915	0,00
10	90,0	4,0	21,0	29,0	19,0	24,0	17,0	2,0	9,0	2,0	0,4786951576438101	0,8213943232361899	0,00
11	123,0	29,0	22,0	15,0	26,0	24,0	31,0	6,0	6,0	6,0	0,2278825660826797	0,772117633391732	0,00
12	136,0	3,0	26,0	36,0	40,0	40,0	31,0	7,0	15,0	7,0	0,15477359327787688	0,845226406722121	0,00
13	145,0	29,0	26,0	36,0	23,0	24,0	31,0	8,0	10,0	8,0	0,1219134256751978	0,878865743248022	0,00
14	150,0	4,0	39,0	36,0	40,0	15,0	31,0	9,0	10,0	9,0	0,1041497714615249	0,895850235384751	0,00
15	246,0	125,0	26,0	36,0	28,0	15,0	31,0	19,0	17,0	19,0	0,00417751542015174	0,995822494579848	0,00
16	140,0	29,0	26,0	28,0	26,0	40,0	31,0	8,0	4,0	8,0	0,13807737082465973	0,861522629175340	0,00
17	242,0	125,0	31,0	15,0	40,0	15,0	31,0	19,0	18,0	19,0	0,004955910771787973	0,9954340892262121	0,00
18	167,0	29,0	39,0	28,0	40,0	15,0	31,0	11,0	11,0	11,0	0,059201073426457924	0,9407982626573542	0,00
19	108,0	4,0	22,0	28,0	23,0	24,0	31,0	4,0	1,0	4,0	0,3267952636599662	0,6732047436440388	0,00
20	263,0	125,0	39,0	28,0	40,0	40,0	31,0	21,0	20,0	21,0	0,002241407861538319	0,997758921384616	0,00
21	137,0	29,0	22,0	36,0	19,0	15,0	31,0	7,0	14,0	7,0	0,1542813723738267	0,8457187627266174	0,00
22	176,0	29,0	39,0	49,0	28,0	15,0	31,0	12,0	10,0	12,0	0,045494250191815	0,9545105749601055	0,00
23	171,0	29,0	39,0	36,0	26,0	24,0	41,0	11,0	6,0	11,0	0,0533919404367493	0,94691816595569325	0,00
24	112,0	4,0	26,0	28,0	23,0	24,0	31,0	5,0	1,0	5,0	0,29560156736145576	0,70439641326365442	0,00
25	154,0	29,0	26,0	28,0	40,0	142,0	31,0	9,0	20,0	9,0	0,08884321481668837	0,9111567951833117	0,00
26	141,0	22,0	22,0	28,0	28,0	15,0	41,0	8,0	1,0	8,0	0,13074597923661236	0,8692540207613877	0,00
27	120,0	29,0	22,0	15,0	23,0	15,0	31,0	5,0	4,0	5,0	0,24651948539156068	0,753480514608439	0,00
28	145,0	29,0	26,0	36,0	23,0	15,0	31,0	8,0	5,0	8,0	0,1219134256751978	0,878865743248022	0,00
29	249,0	125,0	39,0	28,0	26,0	15,0	31,0	20,0	16,0	20,0	0,00367721109544508	0,9963227889804555	0,00
30	137,0	29,0	21,0	28,0	24,0	31,0	7,0	9,0	7,0	7,0	0,1528782397414194	0,847121760258581	0,00
31	132,0	29,0	31,0	15,0	26,0	40,0	31,0	7,0	11,0	7,0	0,17430007830978203	0,825699921690218	0,00
32	180,0	29,0	31,0	49,0	40,0	40,0	31,0	12,0	7,0	12,0	0,0381242692580626	0,9616757737041937	0,00
33	150,0	29,0	26,0	36,0	28,0	24,0	31,0	9,0	11,0	9,0	0,105153131571701	0,894946868428299	0,00
34	265,0	125,0	31,0	28,0	40,0	40,0	41,0	22,0	20,0	22,0	0,0199361631046571	0,988066383689534	0,00

Abajo se puede ver una tabla con los 34 nuevos clientes.

ID = ID del cliente

% = Es la probabilidad calculada

Aceptado o no el crédito = Es un campo creado por mí para decir si acepto o no.

La regla es dar crédito para cualquier cliente donde él % sea igual o más grande que 85%. La decisión es muy personal mía, en épocas donde hay necesidad de dar crédito a más gente se puede bajar el % y así aceptar más clientes y también tener más riesgos en la operación de dar crédito, también se puede subir para 90% o 95% y así quedar con menos clientes pero con mucho mejor potencial de pagar todo.

ID	%	Aceptado o no el crédito
1319	0.99800664	SI
1289	0.9977639	SI
1305	0.99775859	SI
1286	0.99655729	SI
1314	0.99632279	SI
1300	0.99582248	SI
1302	0.99543409	SI
1294	0.99286324	SI
1317	0.96187577	SI

1307	0.95451057	SI
1308	0.94691806	SI
1303	0.94079893	SI
1310	0.91115679	SI
1288	0.91069272	SI
1292	0.89885821	SI
1299	0.89585023	SI
1318	0.89484687	SI
1298	0.87808657	SI
1313	0.87808657	SI
1311	0.86925402	SI
1301	0.86192263	SI
1315	0.84712172	NO
1306	0.84571876	NO
1297	0.84522641	NO
1316	0.82569992	NO
1293	0.78266426	NO
1296	0.77211763	NO
1291	0.77089085	NO
1290	0.76333772	NO
1312	0.75348051	NO
1287	0.72664061	NO
1309	0.70439843	NO
1304	0.67320474	NO
1295	0.52139483	NO