

Estadística III para Ingenieros de Sistemas

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Agenda

- **anuncios varios**
 - <https://forms.office.com/r/LeFxyg4rQ>
- **modelos de analitica (machine learning-ML) Supervisado**
 - **Regresión**
- **Práctica de regresión en Python**

Supervisado, Regresión ejemplo

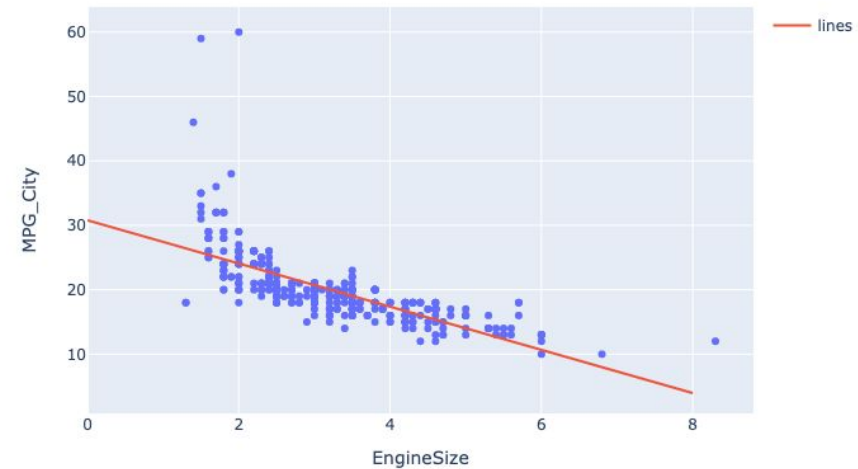
Tomando los datos de los carros, vamos a crear una regresión utilizando el tamaño del motor.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Autos Engine vs MPG



Medidas de error en los problemas de regresión:

$$MSE = \frac{1}{n} \sum_{i=0}^n (y_i - \tilde{y}_i)^2, MAPE = \frac{1}{n} \sum_{i=0}^n \frac{|y_i - \tilde{y}_i|}{y_i}$$

Ej $y=20$, $\tilde{y}=19$, $mse=1$, $mape = 0.05$

Supervisado, Regresión P-value (menor a 0.005)

Para conocer la relevancia de una variable se utilizan hipótesis test :

null hypothesis H0 :No existe relación entre las variables y el coeficiente es 0

H0 : $\beta_1 = 0$

alternative hypothesis Ha : Existe relación entre las variables

Ha : $\beta_1 \neq 0$,

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}, \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

Supervisado, Regresión ejemplo

MPG_City (y)	EngineSize(X)
18	2.5
18	3.8
21	2.0
29	2.0
18	4.3
25	2.0
13	5.6
16	3.0
16	3.2
29	1.8
17	4.6
16	5.0

TRAIN

TEST

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad MSE = \frac{1}{n} \sum_{i=0}^n (y_i - \tilde{y}_i)^2, \quad MAPE = \frac{1}{n} \sum_{i=0}^n \frac{|y_i - \tilde{y}_i|}{y_i}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\bar{x} = 30.2/10 = 3.2, \quad \bar{y} = 203/10 = 20.3$$

$$\beta_{1_num} = (2.5 - 3.2)(18 - 20.3) + (3.8 - 3.2)(18 - 20.3) + \dots + (1.8 - 3.2)(29 - 20.3) = \mathbf{-48.06}$$

$$\beta_{1_den} = (2.5 - 3.2)^2 + (3.8 - 3.2)^2 + \dots + (1.8 - 3.2)^2 = \mathbf{13.816}$$

$$\beta_1 = \mathbf{-48.06/13.816 = -3.47}$$

$$\beta_0 = 20.3 - (-3.47) * 3.2 = \mathbf{31.404}$$

$$\hat{y}_{11} = \mathbf{4.6 * (-3.47) + 31.404 = 15.442}$$

$$\hat{y}_{12} = \mathbf{5.0 * (-3.47) + 31.404 = 14.054}$$

$$MSE \text{ Train} = 11. = \sim 3.41$$

$$\mathbf{MSE} = ((17 - 15.44)^2 + (16 - 14.054)^2) / 2 = 3.110258, \text{ el error es } (3.11)^{0.5} = \mathbf{1.76}$$

$$\mathbf{MAPE} = (ABS(17 - 15.44) / 17 + abs(16 - 14.054) / 16) / 2 = \mathbf{0.10}$$

Supervisado, Regresión P-value (menor a 0.005) ejemplo

Para conocer la relevancia de una variable se utilizan hipótesis test :

null hypothesis H0 :No existe relación entre las variables y el coeficiente es 0

H0 : $\beta_1 = 0$

alternative hypothesis Ha : Existe relación entre las variables

Ha : $\beta_1 \neq 0$,

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}, \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\sigma^2 = (y - \hat{y})^2 = ((18-22.7)^2 + (18-18.18)^2 + \dots + (29-25.14)^2) = 116.50/10 = 11.65$$

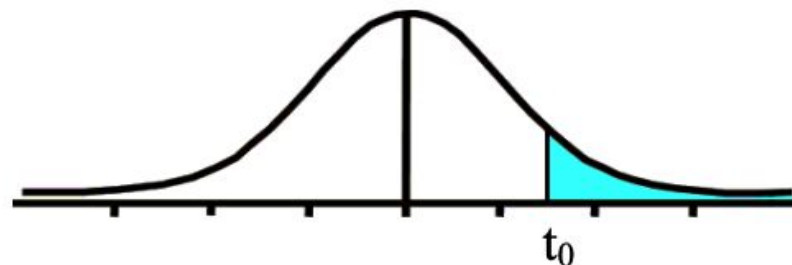
$$SE(\beta_1) = \sigma^2 / (x - \bar{x})^2 = 116.50 / ((2.5 - 3.2)^2 + (3.8 - 3.2)^2 + \dots + (1.8 - 3.2)^2) = 11.65 / 13.8 = 0.84$$

$$t = (-3.47 - 0) / 0.84^{0.5} = -3.7860$$

grados de libertad son el numero de observaciones -1 - el numero de parámetros que es el numero de betas.

Supervisado, Regresión P-value (menor a 0.005) ejemplo

Tabla t-Student



Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693

Supervisado, Regresión

MPG_City (y)	EngineSize(X)
18	2.5
18	3.8
21	2.0
29	2.0
18	4.3
25	2.0
13	5.6
16	3.0
16	3.2
29	1.8
17	4.6
16	5.0

Calcular el R-squared utilizando la regresión del ejemplo.

$$\beta_1 = -48.06 / 13.816 = -3.47$$

$$\beta_0 = 20.3 - (-3.47) * 3.2 = 31.404$$

$$\hat{y} = x * \beta_1 + \beta_0$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

$$TSS = \sum (y_i - \bar{y})^2$$

Supervisado, Regresión

Resultado de una regresión, R-squared es la proporción de la varianza de las millas por galón que es explicada por el motor .

OLS Regression Results

Dep. Variable:	MPG_City	R-squared:	0.503
Model:	OLS	Adj. R-squared:	0.502
Method:	Least Squares	F-statistic:	431.7
Date:	Thu, 16 Mar 2023	Prob (F-statistic):	9.86e-67
Time:	17:17:56	Log-Likelihood:	-1165.8
No. Observations:	428	AIC:	2336.
Df Residuals:	426	BIC:	2344.
Df Model:	1		
Covariance Type:	nonrobust		
	coef	std err	t P> t [0.025 0.975]
const	30.7772	0.546	56.388 0.000 29.704 31.850
EngineSize	-3.3523	0.161	-20.778 0.000 -3.669 -3.035
Omnibus:	447.615	Durbin-Watson:	1.310
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25957.663
Skew:	4.519	Prob(JB):	0.00
Kurtosis:	40.066	Cond. No.	11.1

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

$$TSS = \sum (y_i - \bar{y})^2$$

Supervisado, Regresión con más variables

Calcular Betas o w para muchas variables

$$\hat{y} = w^T \mathbf{x} + b = \sum_{i=1}^p w_i x_i + b$$

$$\min_w \mathcal{L}(w) = \frac{1}{2} \|\mathbf{X}w - \mathbf{Y}\|^2 + \frac{\lambda}{2} \|w\|^2$$

$$w = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_D \right)^{-1} \mathbf{X}^T \mathbf{Y}$$

Supervisado, Regresión

La predicción es el valor de la función $f(x)$ con los nuevos dato \hat{y}

$$\hat{y} = w^T \mathbf{x} + b = \sum_{i=1}^p w_i x_i + b \equiv \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,D} \\ x_{2,1} & x_{2,2} & \dots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,D} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} \sum_d x_{1,d} w_d \\ \sum_d x_{2,d} w_d \\ \vdots \\ \sum_d x_{N,d} w_d \end{bmatrix}}_{\hat{\mathbf{Y}}} \approx \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{Y}}$$

Estadística III para Ingenieros de Sistemas

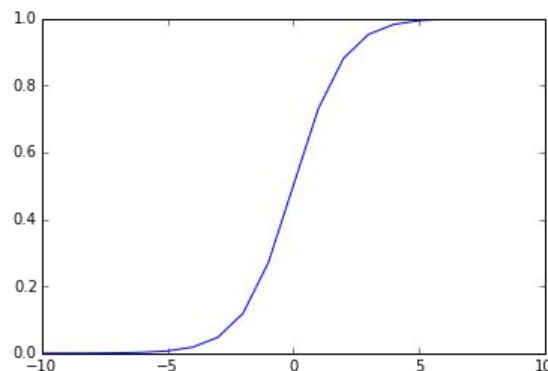
Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Supervisado, Regresión y regresión Logística

Regresión logística, encontrar los coeficientes W de una función, objetivo de reducir los errores en la clasificación.

$$\tilde{y} = \frac{1}{1 + e^{-\sum_{i=0}^p w_i x_i + b}}$$

La función **Sigmoid**, tiene forma de “S” y valores entre $[0,1]$. Por defecto retorna 1 si $\tilde{y} > 0.5$



Ejemplo: utilizando datos históricos del año anterior. Predecir si un alumno pasará el curso este año. ¿Cuál es el porcentaje de personas que el modelo predice que pasaron el curso?

Supervisado, Regresión y regresión Logística

Regresión logística, los resultados de la regresión logística son similares a los de la regresión. Sin embargo, en los problemas de clasificación se utilizan otras métricas para medir el error:

Labels \ Predicción	No Paso	Paso
No Paso	4 (True Negative)	1 (False Positive)
Paso	1 (False Negative)	14 (True Positive)

$$\text{accuracy} = (TN + TP)/(TN+FP+FN+TP) = 18/20$$

$$\text{precisión} = TP/(TP+FP)=14/15 \text{ (cols)}$$

$$\text{recall} = TP/(FN + TP)=14/15 \text{ (rows)}$$

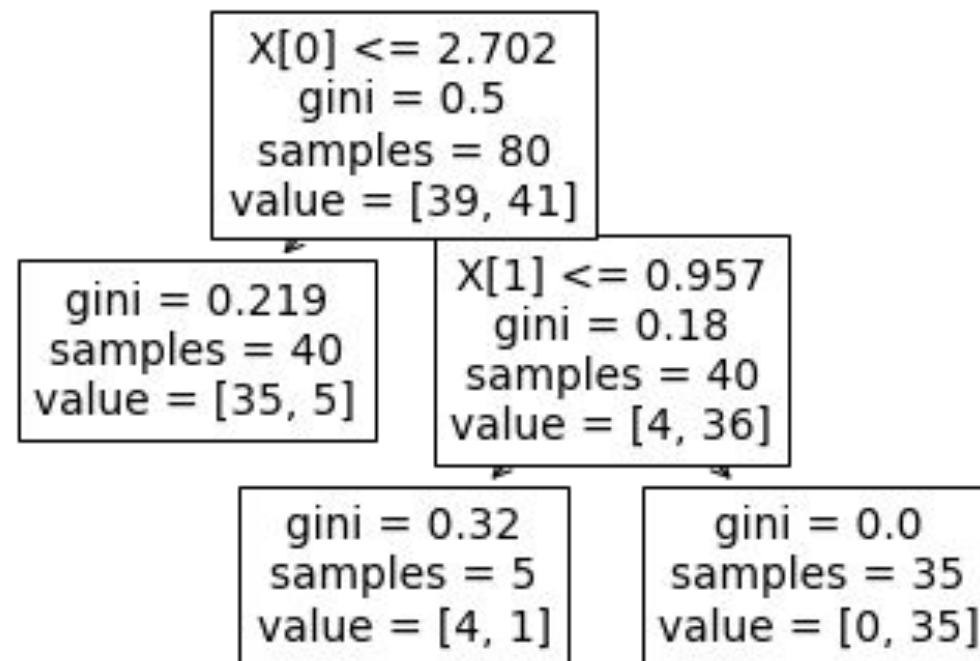
Supervisado, Árboles

Árboles, son modelos que separan los datos basados en la capacidad de una variable de dividir los datos. Funcionan muy bien con datos categóricos. Si la variable objetivo es:

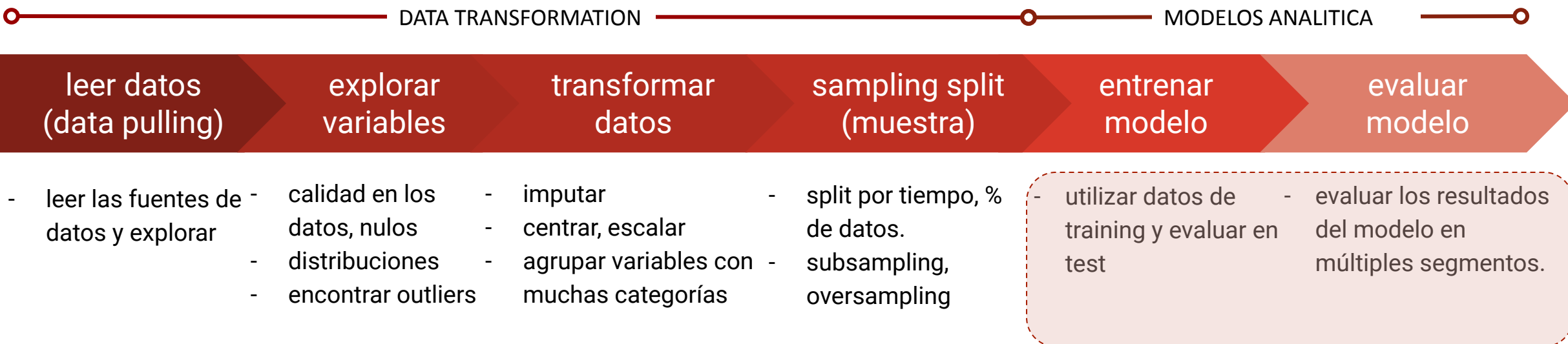
- **continúa**: la predicción es el promedio de los vecinos.
- **categórica**: la predicción es la variable objetivo más común

Los parámetros a definir son la profundidad del árbol, numero mínimo de samples entre otros.

Ejemplo: utilizando datos históricos del año anterior. Predecir si un alumno pasará el curso este año.



“workflow” Flujos de trabajo para crear un modelo



Supervisado, entrenar modelo... antes seleccione la muestra

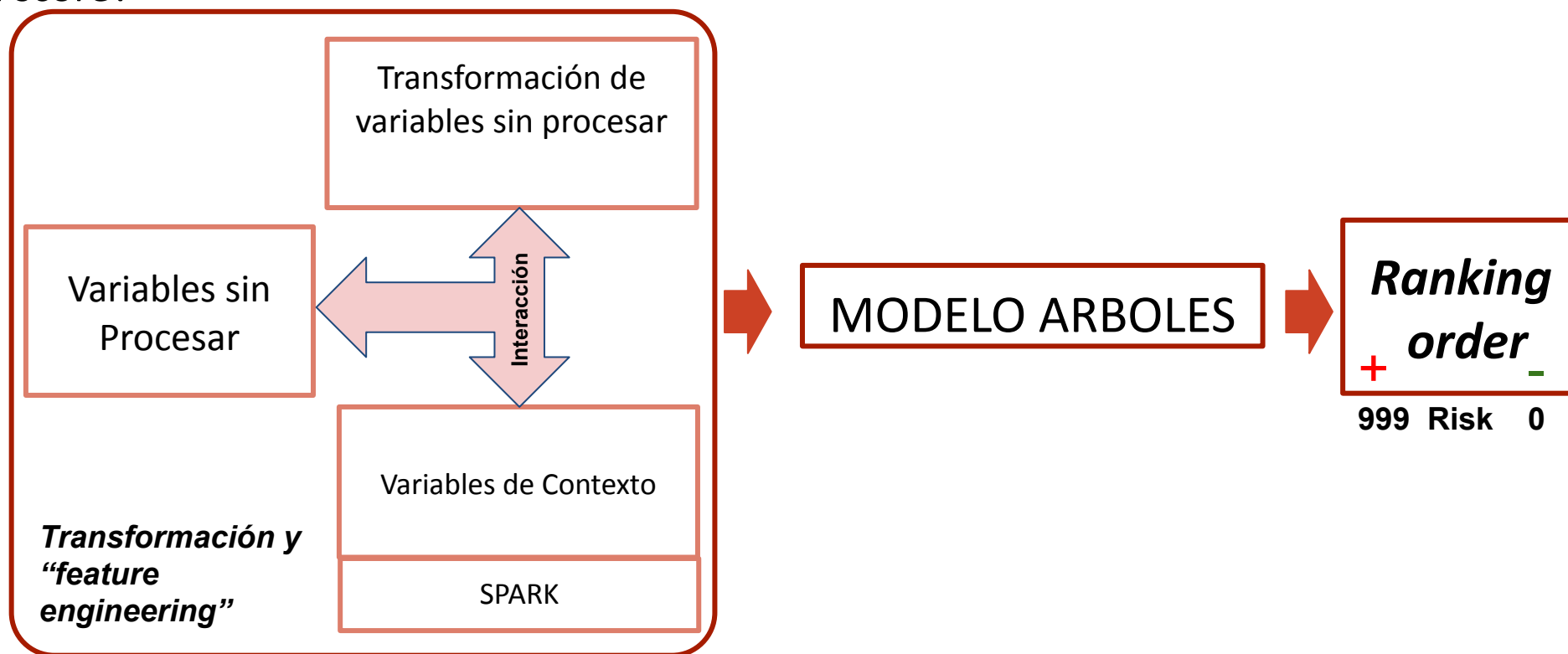
El muestreo se realizó de la siguiente manera:

1. Se tomaron los 40 millones de datos y se dividieron en 2: 10% para pruebas (test) y 0.9 para entrenamiento (training). Se hace estratificado para tener la misma proporción de fraudes
2. A su vez el 0.9 de los datos de entrenamiento (training) se dividieron en 2: 0.8 para entrenamiento y 0.2 para validación dentro del modelo.

Test=0.1*All	
0.9 All	Train 0.8 (Marca=1 or Rand<0.1)
	Validation 0.2 (Marca=1 or Rand<0.1)

Supervisado, entrenar modelo... transformar datos

¿Es posible detectar anomalías?Cuál es la importancia del contexto socioeconómico y demográfico y ¿cómo representar las variables de contexto como “features”? ¿Por qué se quiere utilizar la probabilidad del modelo como un score?



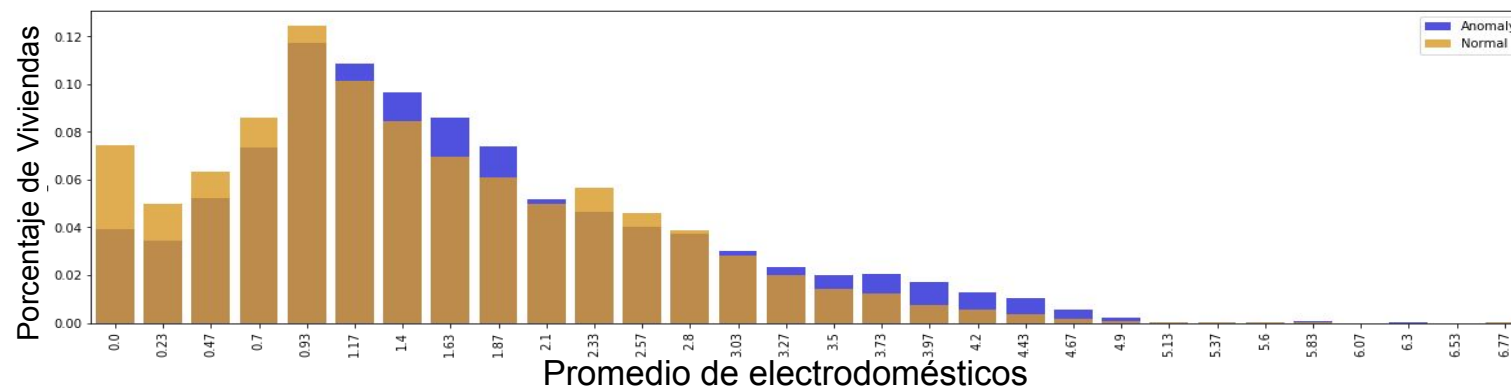
Supervisado, entrenar modelo... transformar datos

¿Cómo funcionan las interacciones entre variables en los modelos de árbol ?

1. Variable Individual

Electrodomésticos = computador portatil + TV + lavadora + nevera + horno + aire acondicionado + equipo de sonido

2. Variable de Contexto



3. Interact. helps GBM

Predicción=1

Si

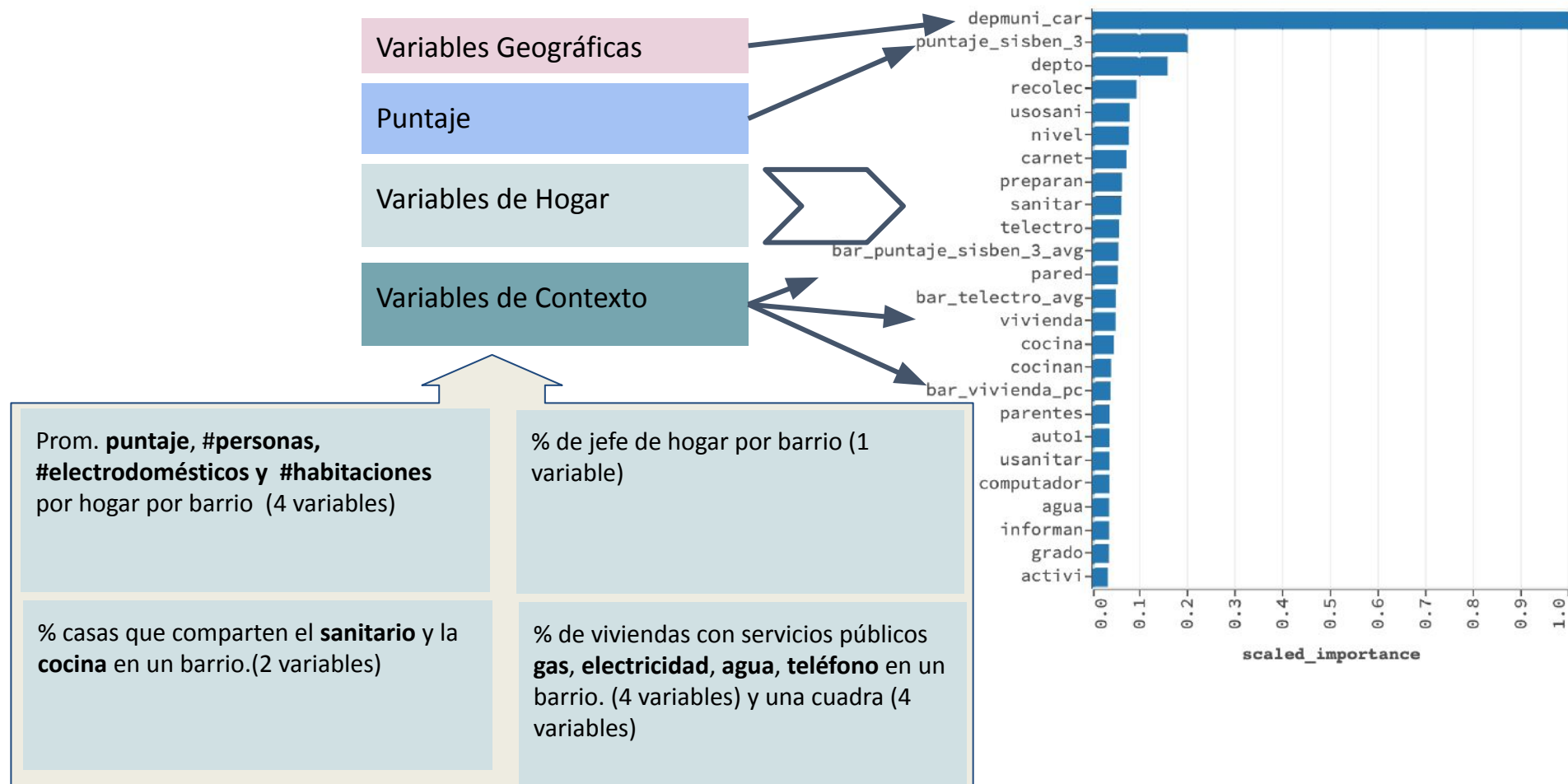
Mi num_electro
<2 y prom.
vecinos >4

No

Predicción=0

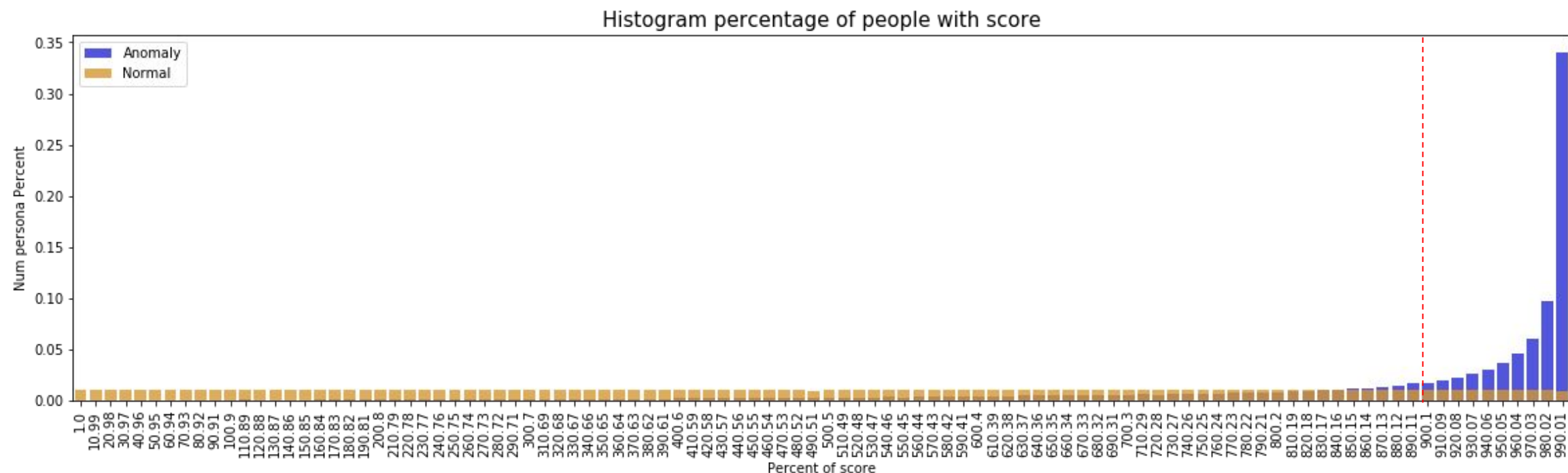
Supervisado, entrenar modelo... importancia de variables

¿Variables más importantes en la detección de fraude para el Sisben?



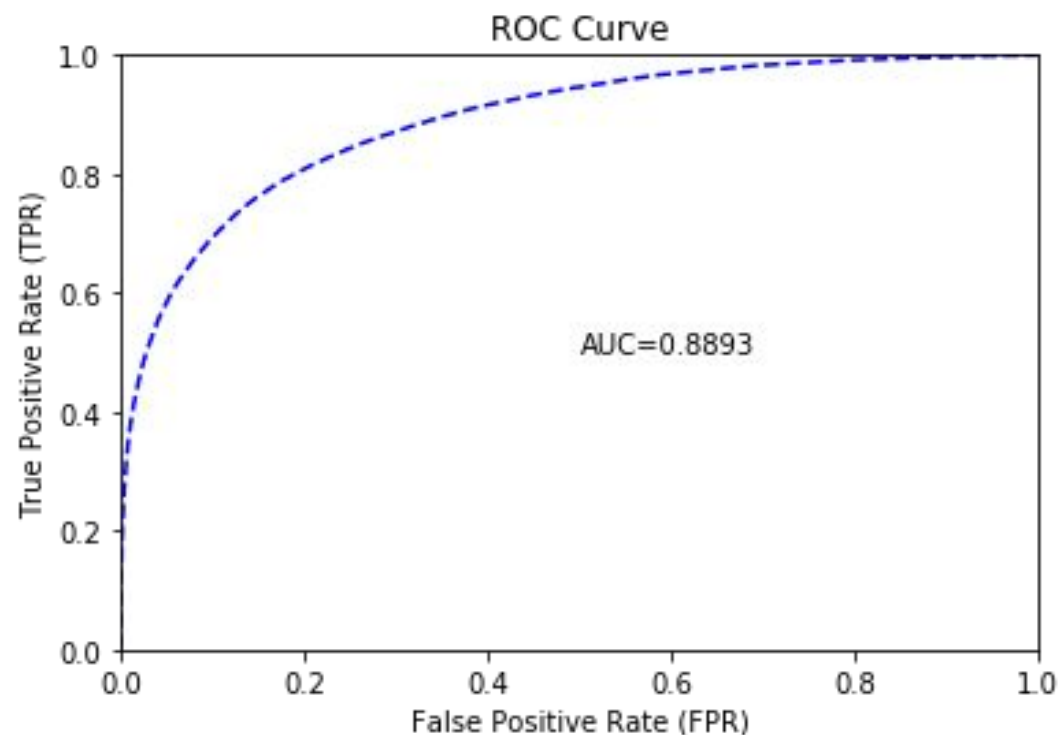
Supervisado, entrenar modelo... validar la predicción

Utilizando test, predecir la probabilidad de fraude. Después crear un histograma para fraude y otro para no fraude. Las probabilidades más altas deben tener más fraude si el modelo funciona como se espera.



Supervisado, evaluar modelo...

En casos con mucho desbalance como el caso del SISBEN, es preferible utilizar otras métricas como AUC. Para comparar el nuevo modelo con otros. Se debe seleccionar el numero máximo de falsos positivos desde el negocio



	AUC	% fraude en 4 Millones
GBM (Arboles)	0.88	0.71

Conclusiones

Para el caso de uso del Sisben se utilizó un modelo llamado GBM, el resultado es la capacidad de capturar el 70% de los fraudes visitando el 10% de los hogares, en términos financieros trae un beneficio de 1:8. Los modelos que utilizan árboles tiene los siguientes beneficios:

- son modelos con mucha flexibilidad
- funcionan bien con variables categóricas
- aunque la regresión logística tiene mucha interpretabilidad , los árboles también tienen interpretabilidad, es posible ver los árboles y validar la importancia de los variables