Feature Engineering

Qué es Feature Engineering?

- Consiste en crear features (columnas)
- Es una tarea mayormente manual
- Es fundamental para el funcionamiento de los algoritmos de Machine Learning
- Es necesario evaluar cada feature creado (feature importance)
- Se debe documentar adecuadamente cada feature

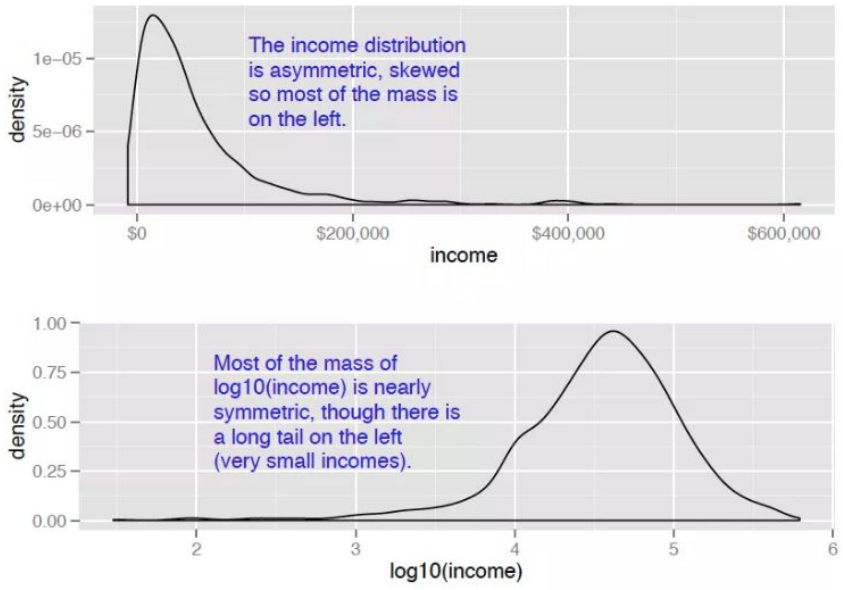
Feature Engineering

- Nuevos Features Individuales
- Transformación de Features
- Encoding de variables categóricas
- Interacción entre features

Nuevos Features Individuales

- Features basados en tiempo (lagged features)
- Features estadísticos (mean, median, std, max, min)
- Features basados en KNN (ej promedio de "x" para los "k" vecinos más cercanos, etc)
- Features basados en texto (contains, tf-Idf, etc)

Transformación de Features - Log



Transformación de Features - Otras

$$x = \sqrt{(x)}$$

$$x = x^{-\frac{1}{2}}$$

$$x = \sin^{-1} x$$

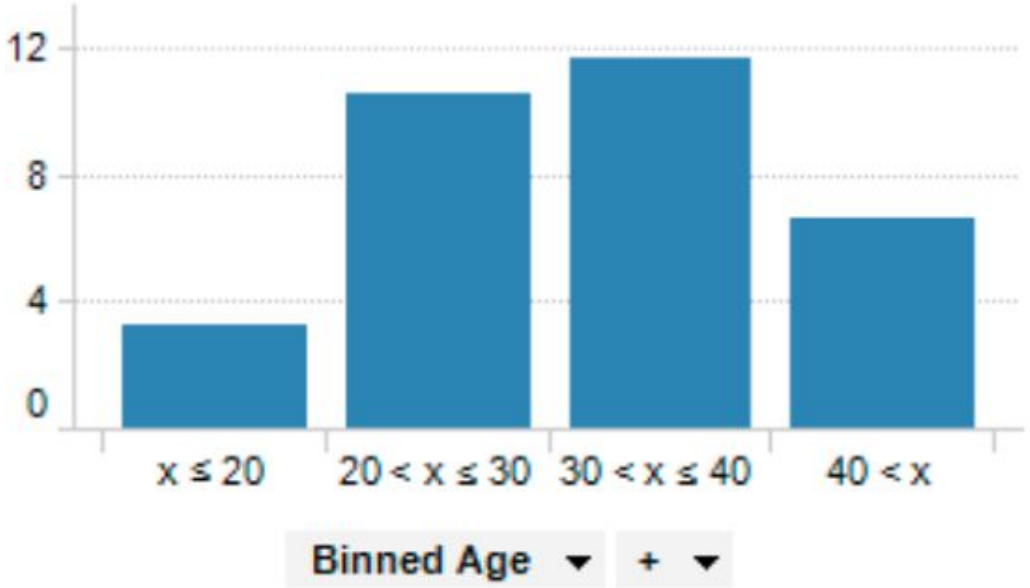
$$x = x^{-1}$$

Transformación de Features - Otras

- Normalizar la variable restando su promedio
- Normalizar y estandarizar (restar el promedio, dividir por std)
- Normalizar por rango (Ej: -1..1)

Transformación de Features - Binning

- Convertir una variable numérica en categórica



Encoding de variables categóricas - One Hot Encoding

Ciudad	Gasto	Ad
Moscu	100	0
Moscu	20	1
Paris	105	1
Moscu	50	0
Roma	120	0
Paris	40	0
Roma	80	0
Londres	50	1

Gasto	Ad	Moscu	Paris	Roma	Londres
100	0	1	0	0	0
20	1	1	0	0	0
105	1	0	1	0	0
50	0	1	0	0	0
120	0	0	0	1	0
40	0	0	1	0	0
80	0	0	0	1	0
50	1	0	0	0	1

Encoding de variables categóricas - Binary Encoding

Ciudad	Gasto	Ad
Moscu	100	0
Moscu	20	1
Paris	105	1
Moscu	50	0
Roma	120	0
Paris	40	0
Roma	80	0
Londres	50	1

Gasto	Ad	C2	C1	C0
100	0	0	0	1
20	1	0	0	1
105	1	0	1	0
50	0	0	0	1
120	0	0	1	1
40	0	0	1	0
80	0	0	0	1
50	1	1	0	0

Moscu = 001

Paris = 010

Roma = 011

Londres = 100

Encoding de variables categóricas - Count Encoding

Ciudad	Gasto	Ad
Moscu	100	0
Moscu	20	1
Paris	105	1
Moscu	50	0
Roma	120	0
Paris	40	0
Roma	80	0
Londres	50	1

Gasto	Ad	Ciudad-Count
100	0	3
20	1	3
105	1	2
50	0	3
120	0	2
40	0	2
80	0	2
50	1	1

Encoding de variables categóricas - Mean (Target) Encoding

Ciudad	Gasto	Ad
Moscu	100	0
Moscu	20	1
Paris	105	1
Moscu	50	0
Roma	120	0
Paris	40	0
Roma	80	0
Londres	50	1

Gasto	Ad	Ciudad-Mean
100	0	0.33
20	1	0.33
105	1	0.5
50	0	0.33
120	0	0
40	0	0.5
80	0	0
50	1	1

Peligro: se filtra información de los labels a los features de entrenamiento.

El cálculo no debe utilizar los datos de test.

Encoding de variables categóricas - Mean Encoding

- **Smoothing**

$$\mu = \frac{n \times \bar{x} + m \times w}{n + m}$$

n: cantidad de valores para esa categoría
w: promedio general
m: peso para el promedio general

- **CV Mean Encoding**
- **Expanding Mean (CatBoost Encoding)**
- **Leave One Out Encoding**

Encoding de variables categóricas - Ordinal Encoding

- Útil cuando las categorías poseen un orden intrínseco.
- Label Encoding con orden correcto.

Experiencia
Baja
Media
Alta

Experiencia
1
2
3

Interacción entre features - Categóricas

- Crear un nuevo feature que sea la concatenación de los features y encodear este nuevo feature.
- Encodear cada feature individualmente y luego generar el producto entre los encodings.

Interacción entre features - Numéricas

- Multiplicación
- Suma
- División
- Diferencia

Interacción entre features - Elección

- Con N variables hay N*(N-1) interacciones posibles (solo tomadas de a dos)
- Se debe seleccionar solo las interacciones útiles.
- Feature selection.

Feature Selection

- En ocasiones conviene quedarnos solo con las K features más importantes.
- Necesitamos obtener las K features más importantes:
  - Univariate Feature Selection: medir cuánto depende el target de cada feature.
  - L1 regularization (Lasso Regression): penalización al aumentar la complejidad del modelo.
  - Random Forest: obtener los features más utilizados en los árboles.

Código...