Preprocesamiento de datos de “Video Game Sales with Ratings”

—

Daniel Cáceres Salas
Nicolas Larrañaga Cifuentes

El conjunto de datos

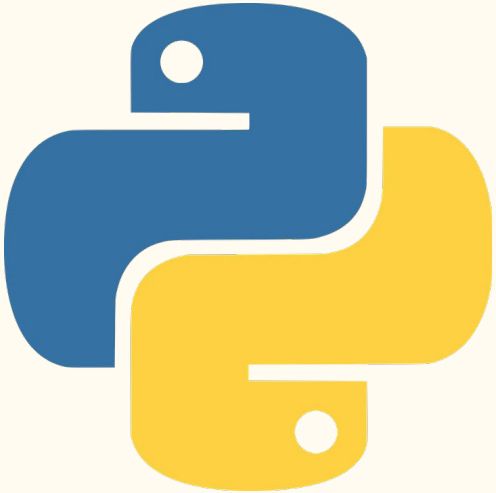
- Name
- Platform
- Year of the release
- Genre
- Publisher
- NA sales
- EU sales
- JP sales
- Other sales
- Global sales
- Critic Score
- Critic Count
- User Score
- User Count
- Developer
- Rating

Objetivo del proyecto

La idea de este proyecto es crear un modelo predictivo que permita a las empresas desarrolladoras de videojuegos obtener un estimado de ventas de un producto nuevo en las diferentes regiones (Japón, Norteamérica, Europa) con base al género, rating y recepción crítica del mismo.

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data

Preprocesamiento - Herramientas utilizadas



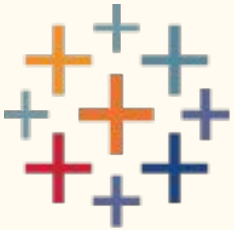
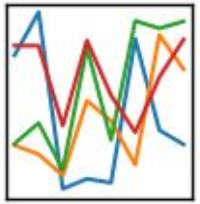


tableau SOFTWARE

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



https://github.com/larranaga/UNAL-data-mining/

Preprocesamiento - Pasos

1. Manejo de valores perdidos
2. Eliminación de ruido
3. Normalización
4. Muestreo
5. Reducción de dimensionalidad
6. Manejo de valores perdidos
7. Discretización
8. PCA

Preprocesamiento - Paso 1

Manejo de valores perdidos

Eliminación

- Name = null
- Developer = null && Publisher = null
- User_score = “tbd”

Inserción

- Rating: Moda por developer.
- Year: Media global.
- Developer o Publisher nulos.

Preprocesamiento - Paso 2

Eliminación de ruido

- User_score fuera del rango [0, 10]
- Critic_score fuera del [0,100]

Preprocesamiento - Paso 3

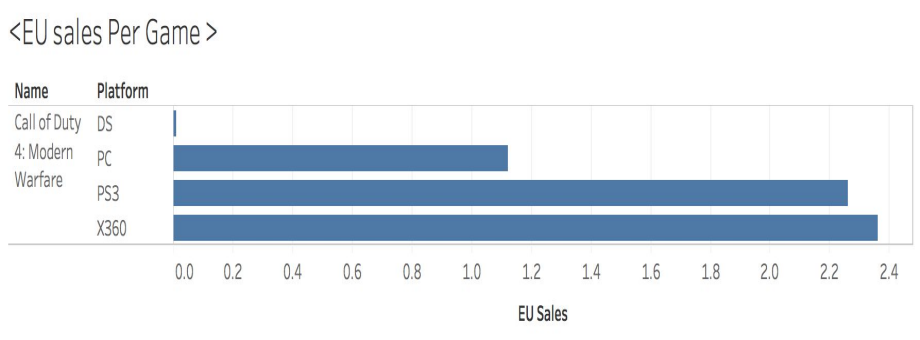
Normalización

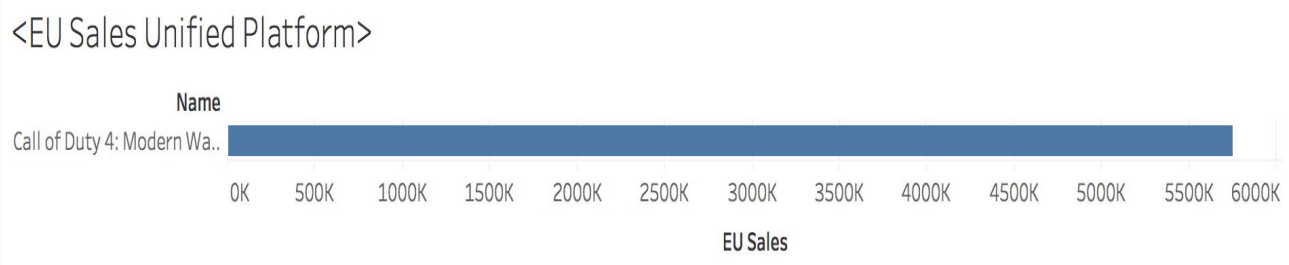
- Critic_score debe estar en el mismo rango que User_score.

Preprocesamiento - Paso 4

Muestreo

- Compresión de registros eliminando la variable Platform.





Preprocesamiento - Paso 5

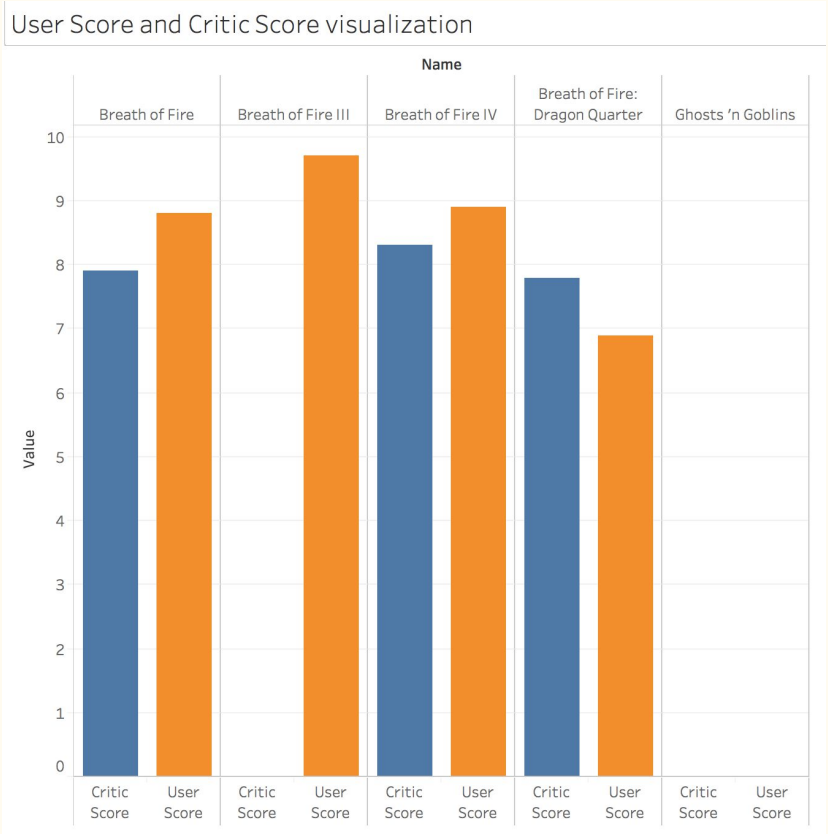
Reducción de dimensionalidad

- Al comprimir los registros ignorando la plataforma, es posible eliminar la dimensión sin ningún problema.

Preprocesamiento - Paso 6

Manejo de valores perdidos

- 5000 datos nulos en User o Critic Score
- Agrupación por Clases (Usando Developer)
- Asignación de media de la unión de las dimensiones.
- Reducción a solo 1200 valores nulos



Preprocesamiento - Paso 7

Discretización

- Discretización sobre las variables ‘Sales’ (NA, EU, JP y Other).
- Método: K-Bins.
- Herramienta: Tableau

Preprocesamiento - Paso 8

PCA

- Separación de la variable Name del conjunto de datos.
- Conversión de las variables de tipo String (Rating, Publisher, Developer, Genre) a tipo entero.
- Cabe resaltar que esta transformación no cambia el hecho de que las variables sean de categoría nominal.
- Herramienta: Python, librería sklearn

Preprocesamiento - Paso 8

PCA

- Cantidad de dimensiones a las que se redujo:

```
0.9458753748078372
0.9774095185442682
0.9957856275326887
0.9999999265821357
0.99999998549735
0.9999999996067076
0.9999999998755031
0.9999999999934568
0.9999999999978755
0.9999999999994034
0.9999999999996987
0.9999999999999404
0.9999999999999999
0.9999999999999999

Process finished with exit code 0
```

¿Preguntas?