



Ciencia de Datos Proyecto Final Google Play Store

Camila Plata

02/12/2022

Motivo del porqué elegí Play Store

Me pareció un buen dataset con buenos datos predictores en comparación con otros que encontré.

Predicción:

Hacer un modelo que diga la cantidad de descargas dándole todos los datos posibles.

Hipótesis:

Cuanto mayor rating tiene una app hay más descargas.

La categoría influye en cuanto si quieres descargar una app.

[HTTPS://WWW.KAGGLE.COM/DATASETS/LAVA18/GOOGLE-PLAY-STORE-APPS](https://www.kaggle.com/datasets/lava18/google-play-store-apps)

02
02

DATASET:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
(10841, 13)													

LIMPIEZA:

	Category	Rating	Reviews	Size	Installs	Type	Content Rating	Genres
0	ART_AND DESIGN	4.1	159	19M	10,000+	Free	Everyone	Art & Design
1	ART_AND DESIGN	3.9	967	14M	500,000+	Free	Everyone	Art & Design;Pretend Play
2	ART_AND DESIGN	4.7	87510	8.7M	5,000,000+	Free	Everyone	Art & Design
3	ART_AND DESIGN	4.5	215644	25M	50,000,000+	Free	Teen	Art & Design
4	ART_AND DESIGN	4.3	967	2.8M	100,000+	Free	Everyone	Art & Design;Creativity

▶ Problemas

Category	0
Rating	1474
Reviews	0
Size	0
Installs	0
Type	1
Price	0
Content Rating	1
Genres	0
Category	object
Rating	float64
Reviews	object
Size	object
Installs	object
Type	object
Price	object
Content Rating	object
Genres	object

Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
10472	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN February 11, 2018

Category	GAME
Rating	4.5
Reviews	10216538.0
Size	Varies with device
Installs	100000000.0
Price	0
Content Rating	Everyone 10+
Genres	Casual;Action & Adventure

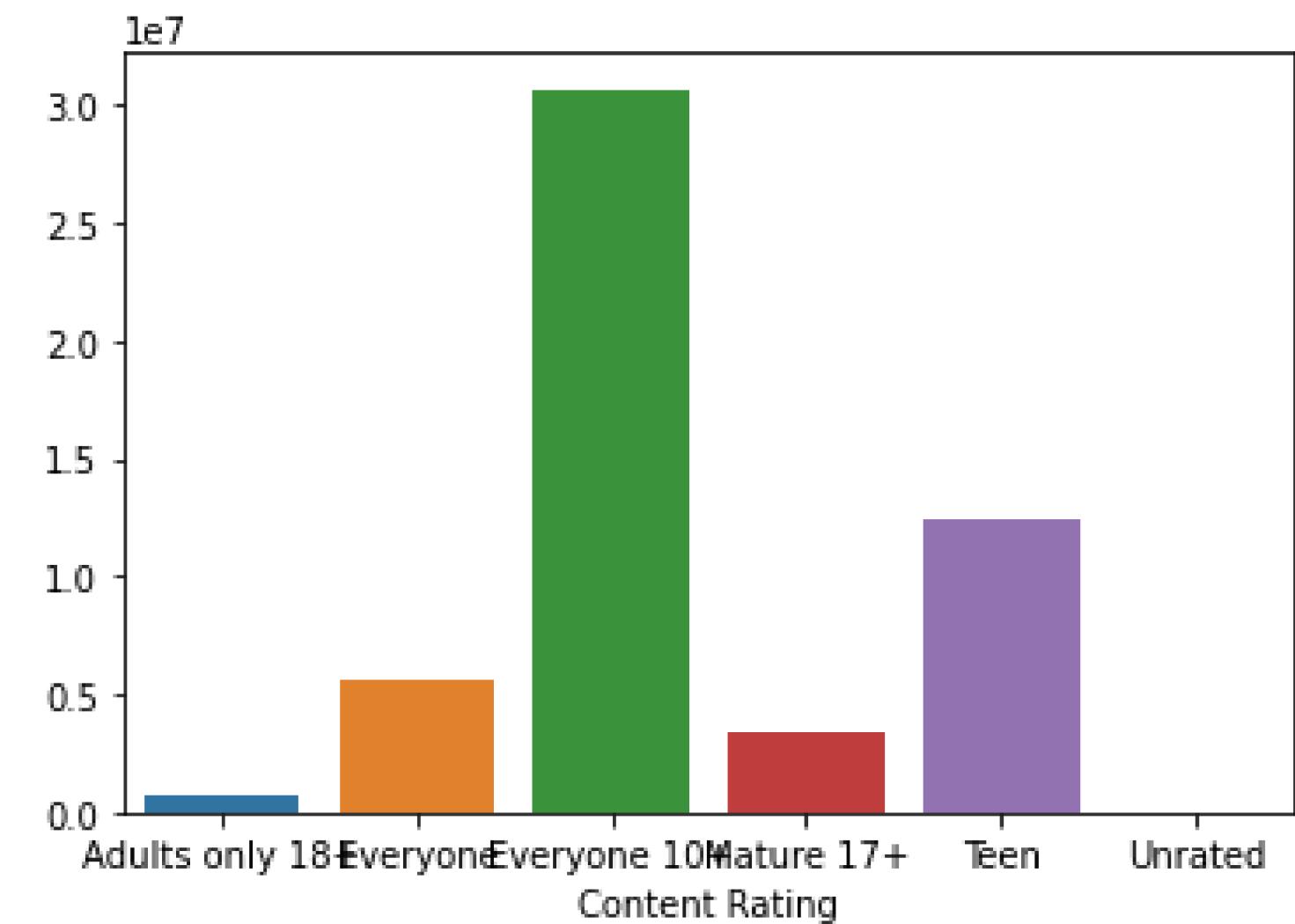
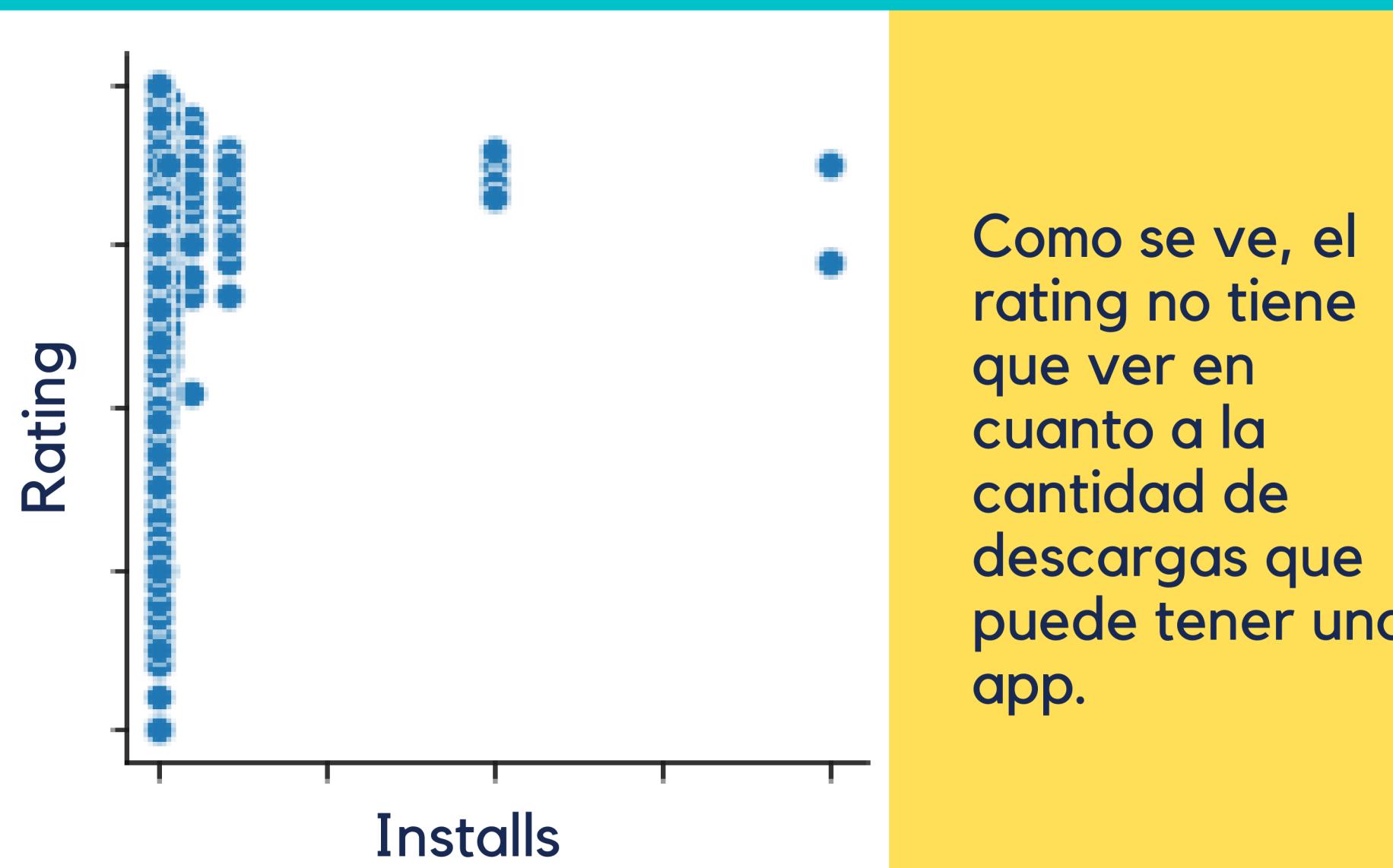
04

TERMINADO:

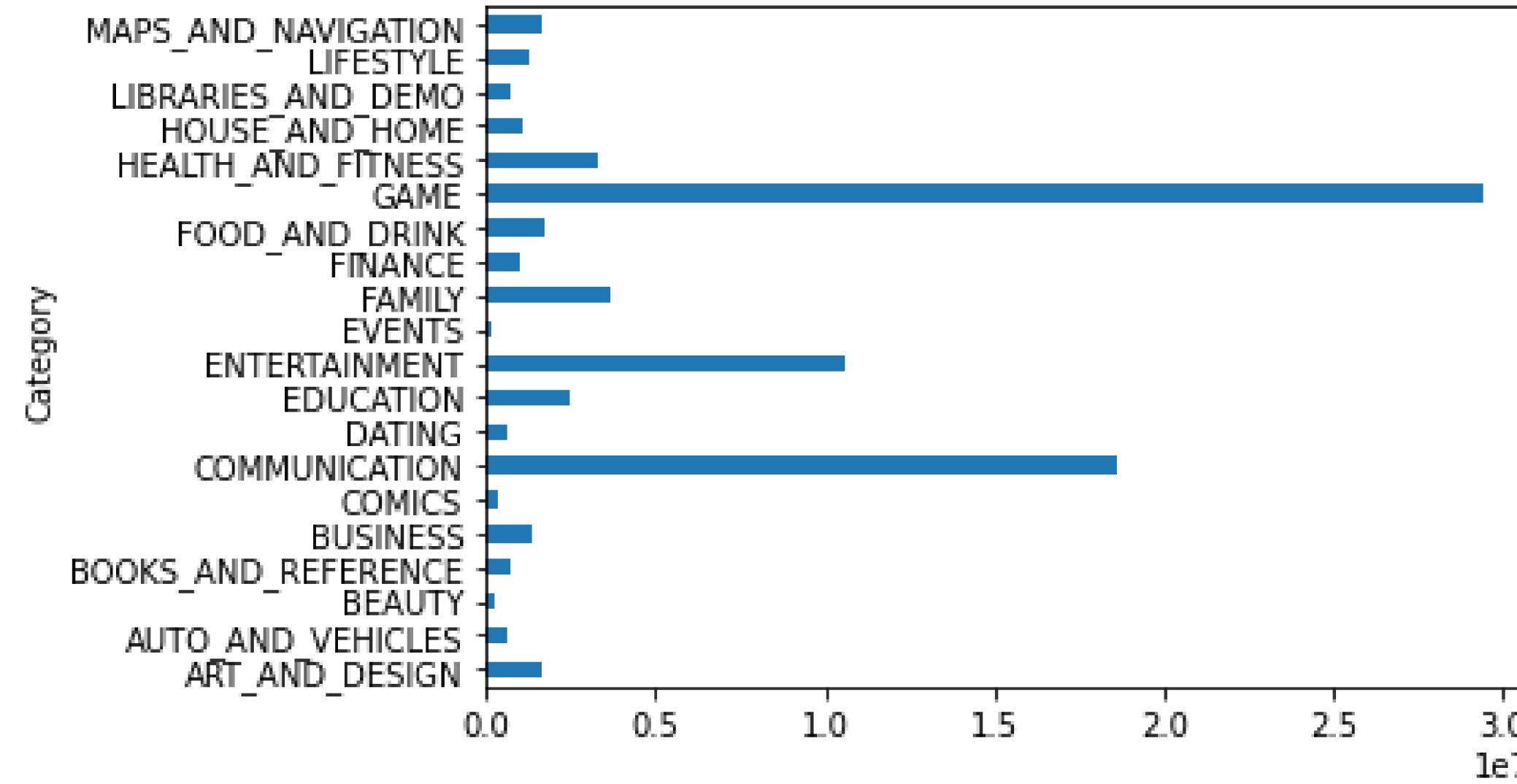
	Category	Rating	Reviews	Installs	Type	Content Rating	Valor_recuperado
0	ART_AND DESIGN	4.100000	159.0	10000.0	0	Everyone	19000000.0
1	ART_AND DESIGN	3.900000	967.0	500000.0	0	Everyone	14000000.0
2	ART_AND DESIGN	4.700000	87510.0	5000000.0	0	Everyone	8700000.0
3	ART_AND DESIGN	4.500000	215644.0	50000000.0	0	Teen	25000000.0
4	ART_AND DESIGN	4.300000	967.0	100000.0	0	Everyone	2800000.0
...
10835	BUSINESS	4.193338	0.0	10.0	0	Everyone	9600000.0
10836	FAMILY	4.500000	38.0	5000.0	0	Everyone	53000000.0
10837	FAMILY	5.000000	4.0	100.0	0	Everyone	3600000.0
10838	MEDICAL	4.193338	3.0	1000.0	0	Everyone	9500000.0
10840	LIFESTYLE	4.500000	398307.0	10000000.0	0	Everyone	19000000.0

Hipótesis 1:

Pero en este grafico vemos el content rating en base a descargas promedio.



Hipótesis 2:



Este gráfico muestra la categoría en el dataset en base a las descargas promedio.

07

MODELOS:

Árbol de Decisión

9332	1000.0
8614	0.0
3868	1000000.0
7196	10000.0
1615	100000.0
	...
8833	100.0
4174	50000.0
5338	500000.0
2960	100000.0
4006	10000.0

Datos predichos con datos de entrenamiento

12.360721259031239

14785018.590798246

Error medio cuadrado con los datos de entrenamiento

Error medio cuadrado con los datos de testeo

PROBLEMA:
La cantidad de preguntas que hizo el modelo para poder predecir los datos.

51 preguntas

Overfitting

MODELOS:

Random Forest

SIN HIPERPARAMETROS

4097880.624297988

Error medio cuadrado con los datos de entrenamiento

15318904.984784273

Error medio cuadrado con los datos de testeo

CON HIPERPARAMETROS

Error medio cuadrado con los datos de entrenamiento

38872177.39762643

Error medio cuadrado con los datos de testeo

32042253.54115564

PROBLEMA:

El dato que queremos predecir no es posible porque es posible que los datos no sean buenos predictores. A pesar de haber puesto hiperparametros.

CAMBIO DEL DATO A PREDICIR: RATING

9332	3.100000
8614	4.193338
3868	4.200000
7196	4.500000
1615	4.600000
	...
8833	4.193338
4174	3.800000
5338	4.500000
2960	4.300000
4006	1.700000

Datos predichos con datos de entrenamiento

0.013908766507167844

Error medio cuadrado con los datos de entrenamiento

0.6493675995064627

Error medio cuadrado con los datos de testeo

ARBOL DE DECISION:
A pesar de haber hecho muchas preguntas se puede ver que el modelo es mejor que antes.

47 preguntas



Sin overfitting

CAMBIO DEL DATO A PREDICIR: RATING

0.49257656194925714

0.4838682595215794

Error medio cuadrado con los datos de entrenamiento

Error medio cuadrado con los datos de testeo

RANDOM FOREST:
Como se puede ver, ahora ambos modelos predicen mejor al haber cambiado el dato a predecir.



CONCLUSIONES:

Si bien este dataset no parecía difícil hubo momentos complicados. Como el pasar los datos a numéricos o tratar con registros corridos. También esta el tema de que hubo overfitting al momento de predecir las descargas con los modelos de Árbol de Decisión y Random Forest. Pero al cambiarlo por rating no hubo ese problema. Podemos concluir que no todos los datos pueden ser predictivos o a ayudar a predecir otro dato.