ITBA
Instituto Tecnológicc
de Buenos Aires

82.05 | Análisis Predictivo

EXAMEN 2

AGUSTINA GONZALEZ CRESPO

ITBA

0.30944

Score obtenido

0	Id	977541 non-null	int64
1	averageRating	977541 non-null	float64
2	numVotes	977541 non-null	int64
3	titleType	977541 non-null	object
4	isAdult	977541 non-null	float64
5	startYear	977541 non-null	int64
6	endYear	977541 non-null	int64
7	runtimeMinutes	977541 non-null	int64
8	genres_x	977540 non-null	object
9	directors	977541 non-null	object
10	writers	977541 non-null	object
11	seasonNumber	438133 non-null	float64
12	episodeNumber	438133 non-null	float64
13	ordering	370842 non-null	float64
14	language	370842 non-null	object
15	attributes	370842 non-null	object
16	isOriginalTitle	370842 non-null	float64
17	adult	47305 non-null	object
18	budget	47305 non-null	float64
19	genres_y	47305 non-null	object
20	original_language	47294 non-null	object
21	overview	46512 non-null	object
22	popularity	47302 non-null	float64
23	production_companies	47302 non-null	object
24	production_countries	47302 non-null	object
25	release_date	47234 non-null	object
26	revenue	47302 non-null	float64
27	runtime	47096 non-null	float64
28	status	47229 non-null	object
29	tagline	23808 non-null	object
30	video	47302 non-null	object

Base de datos:

- 977541 registros
- 31 columnas

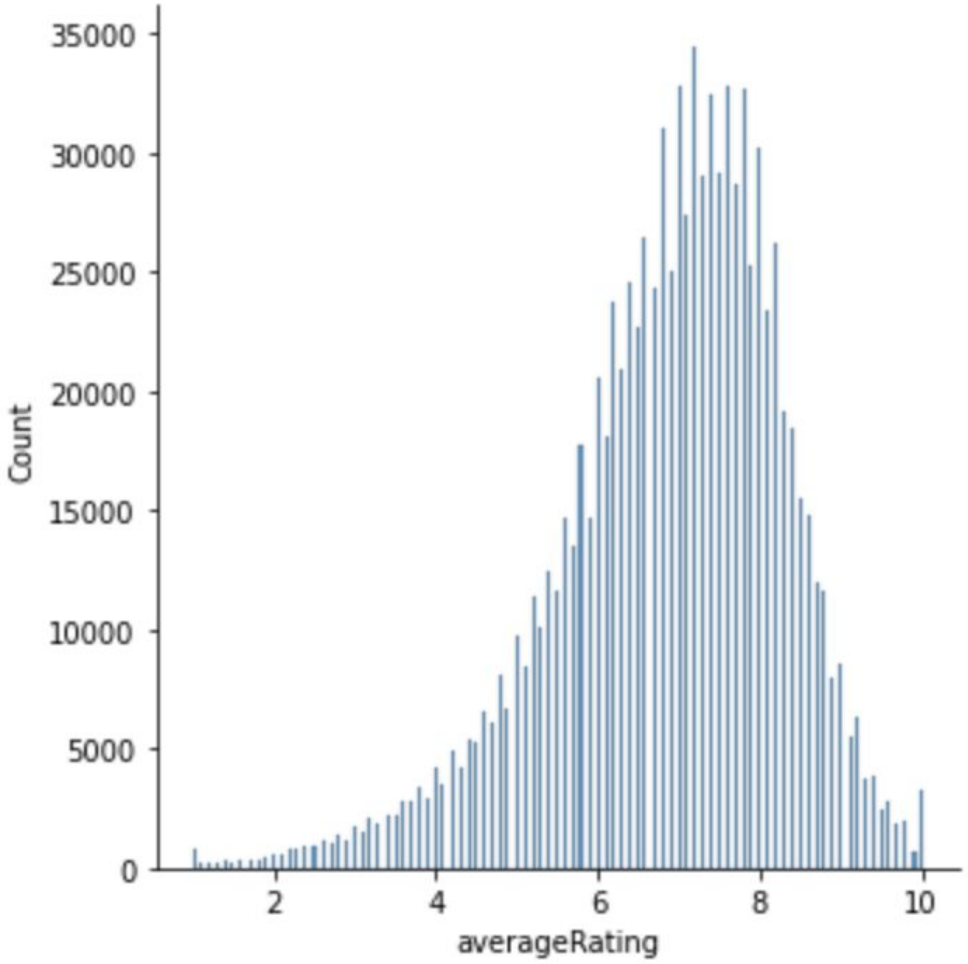
Variables numéricas:

- 15 variables

Variables categóricas:

- 16 variables

Distribución de variable target



AverageRating:

- promedio: **6,88**
- mediana: **7.1**

Variables Numéricas

	count	mean	std	min	25%	50%	75%	max
Id	977541.0	4.887700e+05	2.821919e+05	0.0	244385.000000	488770.000000	733155.000000	9.775400e+05
averageRating	977541.0	6.881284e+00	1.405724e+00	1.0	6.100000	7.100000	7.900000	1.000000e+01
numVotes	977541.0	1.618633e+03	2.464842e+04	5.0	9.000000	22.000000	92.000000	2.425542e+06
isAdult	977541.0	2.080527e-02	2.047549e+00	0.0	0.000000	0.000000	0.000000	2.020000e+03
startYear	977541.0	1.999362e+03	3.423928e+01	0.0	1992.000000	2008.000000	2015.000000	2.021000e+03
endYear	977541.0	5.798425e+01	3.358606e+02	0.0	0.000000	0.000000	0.000000	2.022000e+03
runtimeMinutes	977541.0	4.139097e+01	6.365649e+01	-29745.0	0.000000	27.000000	73.000000	1.440000e+04
seasonNumber	438133.0	4.035528e+00	9.920419e+00	0.0	1.000000	2.000000	4.000000	1.996000e+03
episodeNumber	438133.0	5.503868e+01	5.804034e+02	0.0	4.000000	8.000000	16.000000	1.576200e+04
ordering	370842.0	3.475165e+00	5.128331e+00	1.0	1.000000	2.000000	3.000000	1.620000e+02
isOriginalTitle	370842.0	1.887596e-05	4.344613e-03	0.0	0.000000	0.000000	0.000000	1.000000e+00
budget	47305.0	6.233607e+06	2.320708e+07	0.0	0.000000	0.000000	0.000000	3.800000e+08
popularity	47302.0	3.450543e+00	7.564531e+00	0.0	0.442414	1.282498	4.680878	5.474883e+02
revenue	47302.0	1.878803e+07	9.347703e+07	0.0	0.000000	0.000000	0.000000	2.787965e+09
runtime	47096.0	9.455020e+01	3.545469e+01	0.0	85.000000	95.000000	106.000000	1.140000e+03

Variables eliminadas:

- budget
- revenue

Missings

Variables	Imputación
SeasonNumber	0
EpisodeNumber	0
isOriginalTitle	0
popularity	0

SeasonNumber:

titleType	
movie	235881
short	111729
tvEpisode	15
tvMiniSeries	9751
tvMovie	40527
tvSeries	64913
tvShort	1895
tvSpecial	8785
video	55229
videoGame	10683

EpisodeNumber:

titleType	
movie	235881
short	111729
tvEpisode	15
tvMiniSeries	9751
tvMovie	40527
tvSeries	64913
tvShort	1895
tvSpecial	8785
video	55229
videoGame	10683

Outliers

RuntimeMinutes:

- se eliminó un valor negativo

SeasonNumber:

- se eliminaron los valores mayores a 100

isAdult:

- se eliminó un dato con un valor igual a 2020

Runtime vs RuntimeMinutes

	runtime	runtimeMinutes
33	93.0	93
66	92.0	92
68	100.0	100
72	84.0	84
112	99.0	99
157	119.0	119
161	94.0	94
199	170.0	170
225	1.0	1
230	110.0	110

Porcentaje de Missings:

- runtime: 95,18%
- runtimeMinutes: 0%

Como los valores eran los mismos, se decidió eliminar runtime debido a que tenía muchos missings

Ordering

```
db.loc[db['tagline']== "There won't be a dry seat in the house!" ]
```

	Id	averageRating	numVotes	titleType	isAdult	startYear	endYear	runtimeMinutes	genres_x	directors	writers	seasonNumber	episodeNumber	ordering	language
5929	5929	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	21.0	0
54907	54907	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	25.0	0
129282	129282	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	19.0	0
198721	198721	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	9.0	0

La variable ordering es un entero único que identifica la fila para un título en específico, por lo tanto fue eliminada

Variables Categóricas

	count	unique	top	freq
titleType	977540	10	tvEpisode	438147
genres_x	977540	1932	Comedy	81475
directors	977540	241001	0	147406
writers	977540	398812	0	224376
language	370842	7	0	369213
attributes	370842	152	0	349264
adult	47305	2	False	47293
genres_y	47305	3679	[{'id': 18, 'name': 'Drama'}]	4706
original_language	47294	87	en	35382
overview	46512	36834	No overview found.	111
production_companies	47302	19203	[]	10940
production_countries	47302	2047	[{'iso_3166_1': 'US', 'name': 'United States o...	21701
release_date	47234	15913	2008-01-01	124
status	47229	5	Released	46918
tagline	23808	17086	There won't be a dry seat in the house!	22
video	47302	2	False	47217

Attributes

Se elimino ya que poseía valores distintos para un mismo title

Id	averageRating	numVotes	titleType	isAdult	startYear	endYear	runtimeMinutes	genres_x	directors	writers	seasonNumber	episodeNumber	ordering	language	attributes
5929	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	21.0	0	segment title
54907	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	25.0	0	segment title

Id	averageRating	numVotes	titleType	isAdult	startYear	endYear	runtimeMinutes	genres_x	directors	writers	seasonNumber	episodeNumber	ordering	language	attributes
551552	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	28.0	0	0
663509	4.7	515	movie	0.0	1978	0	84	Comedy	nm0588496	nm0695147,nm0975323,nm0588496,nm0032403	0.0	NaN	7.0	0	0

genres_x vs genres_y

	genres_x	genres_y
33	Drama	[{'id': 18, 'name': 'Drama'}]
66	Drama,Film-Noir,Thriller	[{'id': 18, 'name': 'Drama'}, {'id': 53, 'name...
68	Comedy,Crime,Thriller	[{'id': 35, 'name': 'Comedy'}, {'id': 53, 'nam...
72	Western	[{'id': 37, 'name': 'Western'}]
112	Crime,Drama,Thriller	[{'id': 80, 'name': 'Crime'}, {'id': 53, 'name...
157	Drama,Romance	[{'id': 10749, 'name': 'Romance'}, {'id': 18, ...
161	Comedy,Family,Romance	[]
199	Adventure,Fantasy,Horror	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...
225	Documentary,Short,Western	[{'id': 99, 'name': 'Documentary'}]
230	Crime,Drama,Mystery	[{'id': 80, 'name': 'Crime'}, {'id': 9648, 'na...

Porcentaje de Missings:

- genres_x: 0,0001%
- genres_y: 95,16%

Missings

Variables	Imputación
genres_x	“missing”
directors	“missing”
writers	“missing

Creación de variables

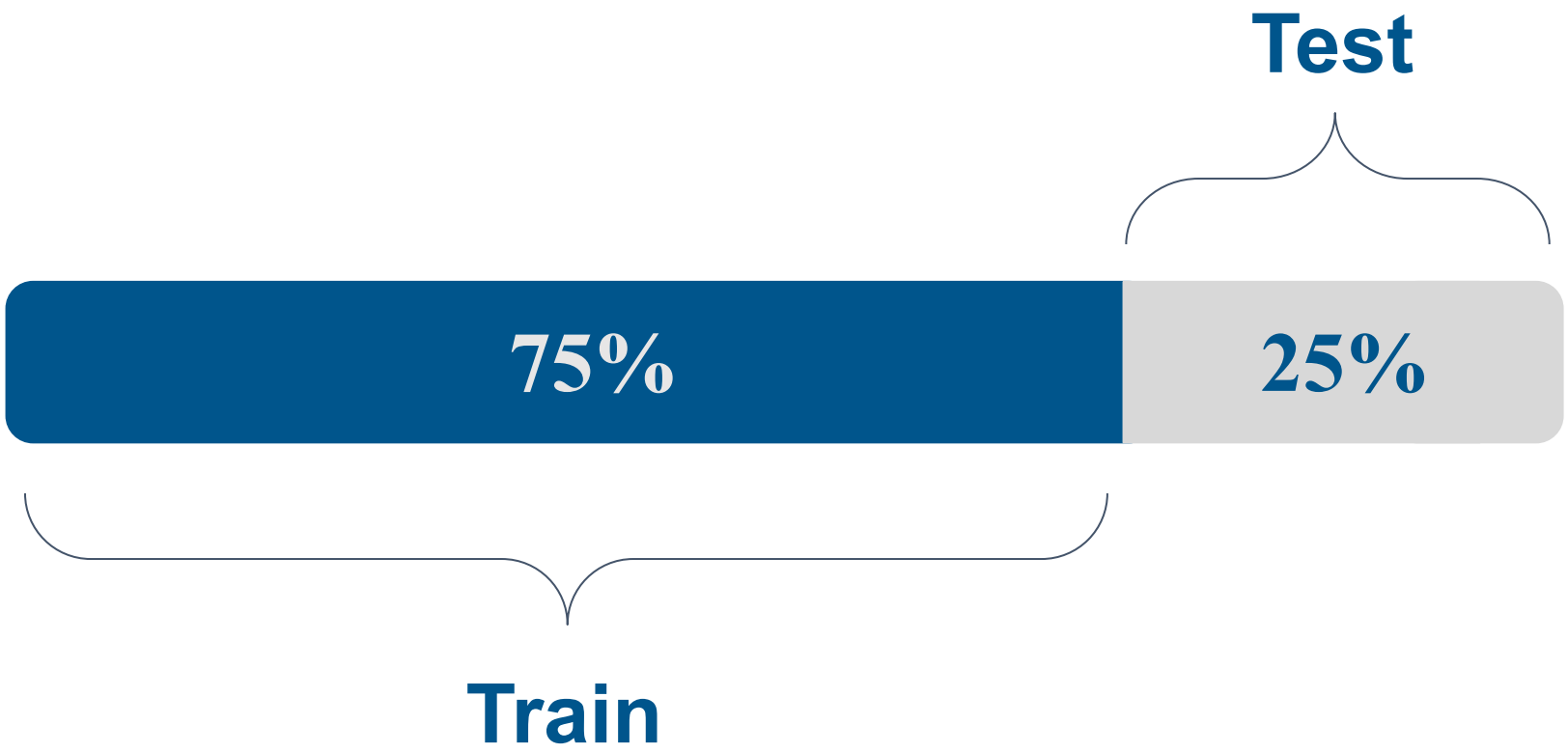
- genero → top 5 géneros
  - si alguno de los géneros del título está en el top 5 se le asigna un 1, de lo contrario un 0
- directores → top 20 directores
  - si alguno de los directores del título está en el top 20 se le asigna un 1, de lo contrario un 0
- escritores → top 20 escritores
  - si alguno de los escritores del título está en el top 20 se le asigna un 1, de lo contrario un 0
- variables dummies de titleType

Variables categóricas eliminadas:

- language
- adult
- original_language
- overview
- production_companies
- production_countries
- release_date
- status
- tagline
- video

Duplicados:

Luego de haber eliminado las variables **ordering** y **attributes** se encontraron **55832** registros duplicados que se **eliminaron**



Train: **691148** registros

Test: **230383** registros

Modelos

LinearRegression( )

- R2 score : 0,154743

DecisionTreeRegressor( )

- R2 score: 0,147078

RandomForestRegressor( )

- R2 score: 0,20044

XGBRegressor( )

- R2 score: 0,215522

HistGradientBoostingRegressor( )

- R2 score: 0,25458

GradientBoostingRegressor( )

- R2 score: 0,216121

LGBMRegressor( )

- R2 score: 0,253713

Modelo elegido:

LGBMRegressor( )

Light Gradient Boosting Machine (LightGBM) es un gradient boosting framework que utiliza algoritmos de aprendizaje basados en árboles donde se dividen las hojas del árbol con el “mejor ajuste”

Ventajas del modelo:

- alta velocidad de entrenamiento y mayor eficiencia
- mejor precisión
- menor uso de memoria
- capaz de manejar datos a gran escala

```
random_LGBM=lgb.LGBMRegressor(max_depth=25, min_split_gain=0.4,
              n_estimators=1000, num_leaves=200, reg_alpha=1.3, reg_lambda=1.2,
              subsample=0.7, subsample_freq=20)
```

Parámetros utilizados:

- max_depth: profundidad máxima del árbol
- min_split_gain: pérdida mínima de reducción requerida para realizar la siguiente partición en un nodo
- n_estimators: número de árboles que se deben crear
- num_leaves: número de hojas
- reg_alpha: regularización L1
- reg lambda: regularización L2
- subsample: porcentaje de filas utilizadas para crear el árbol en cada iteración
- subsample_freq: frecuencia del subsample

¡Muchas gracias!