Bootcamp UDD
Ciencia de Datos & AI

Proyecto Final

Carlos Müller
Corte 5

Introducción

En el siguiente proyecto fue realizado según en la base de datos de reviews de aplicaciones de la Google Play Store. La primera parte de la base de datos muestra 9.960 aplicaciones distintas que se encuentran en Google Play Store. En cuanto a la segunda, contiene los reviews realizado por los usuarios por cada aplicación.

Las variables de la base de datos son:

- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres

Google Play

Objetivo

El objetivo de nuestro trabajo al analizar esta base de datos es el siguiente:

Actualmente una compañia cuenta con un aplicación de Juego en Google Play Store el cual ya tiene una trayectoria solida, pero en los últimos meses no ha obtenido los resultados deseados en cuanto a ingresos y números de descarga. No esta seguro si el precio de la aplicación es el correcto frente a la competencia.

Como analista de datos, mi objetivo será crear un modelo basado en la categoria de Juegos (“Game”) con el fin de poder predecir un precio estimado segpun las caracteristicas actuales de la aplicación del cliente.

Analisis Categoria Game

Categoria Games es la segunda más descargada

FAMILY	1746
GAME	1097
TOOLS	733
PRODUCTIVITY	351
MEDICAL	350
COMMUNICATION	328
FINANCE	323
SPORTS	319
PHOTOGRAPHY	317
LIFESTYLE	314

Por lo general la mayoria de las aplicaciones tiene sobre 100.000 descargas

Installs	
10.000.000	191
1.000.000	152
100.000.000	111
100.000	108
5.000.000	94
10.000	63
500.000	58
50.000.000	54
50.000	38
1.000	33

Analisis Categoria Game

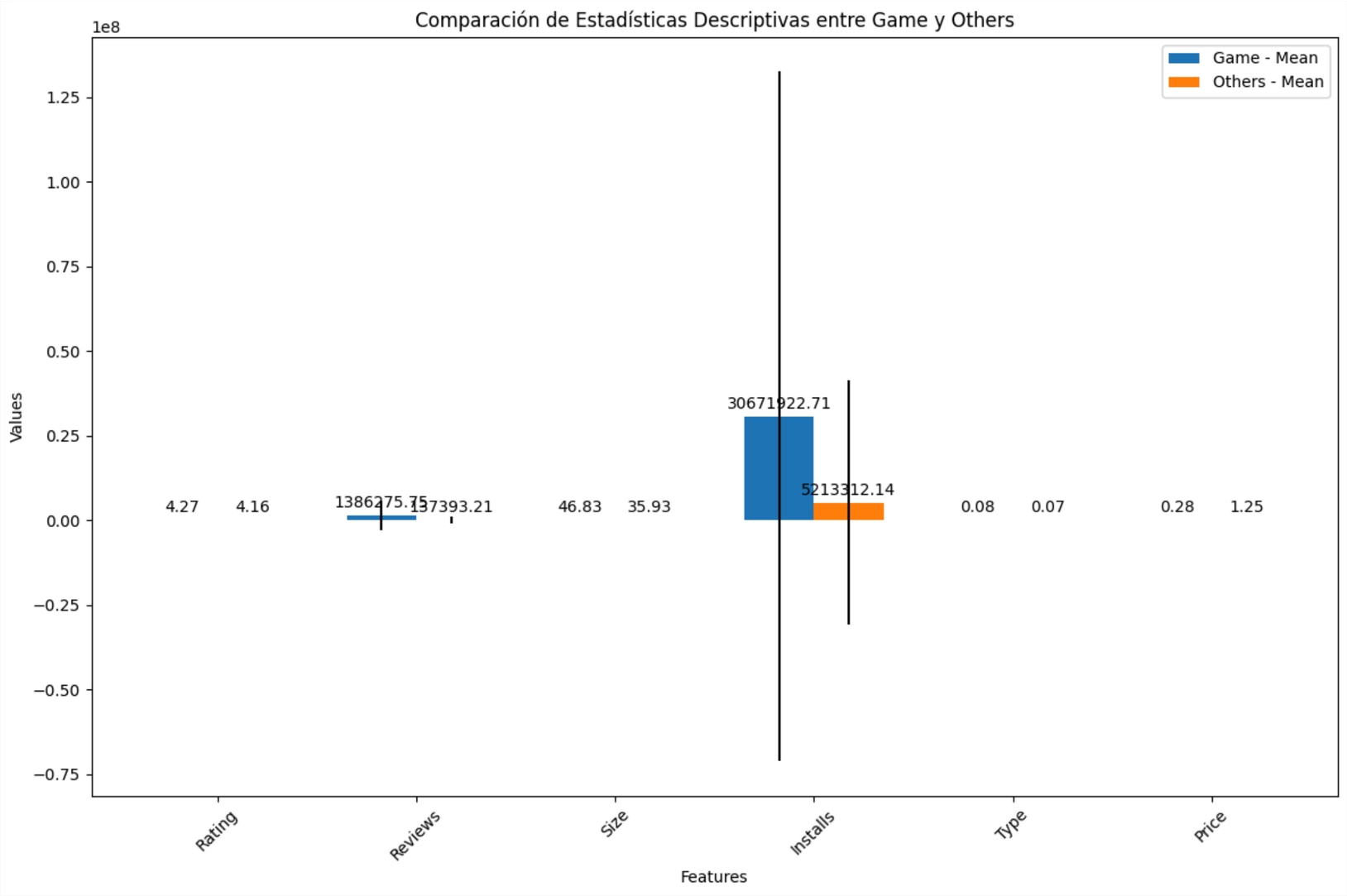
Revisamos número de descargas y Type

Installs	
10.000.000	191
1.000.000	152
100.000.000	111
100.000	108
5.000.000	94
10.000	63
500.000	58
50.000.000	54
50.000	38
1.000	33

Type	
Free	899
Paid	75

Analisis Categoria Game

Si comparamos Game con el resto de las categorías



- Game obtiene mayor puntiación de rating.
- Se descargan más veces.
- Tienen un valor promedio menor.

Creación del modelo

Para el modelo, vamos a trabajar con todas las aplicaciones de categoria=”GAME” y que sean de pago Type=1.

Además, trabajaremos con las siguientes variables X para redecir nuestra variable objetivo X=Price

Y = ['Rating', 'Reviews', 'Size', 'Installs', 'Content Rating'].

Tuning y Ensambles

Resultado de Tuning:
Random Forest: {'max_depth': None, 'n_estimators': 200}
Gradient Boosting: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}
AdaBoost: {'n_estimators': 200}

Ensamble:
Random Forest Mean Squared Error (MSE): 5243.348920231918 R-squared (R2): -0.015903491239917l9
Gradient Boosting Mean Squared Error (MSE): 5080.810097419081 R-squared (R2): 0.01558854945194743l
AdaBoost Mean Squared Error (MSE): 6111.225696880016 R-squared (R2): -0.18405538438607305

Gradient Boosting más bajo entre los tres, lo que significa que, en promedio, sus predicciones están más cerca de los valores reales que las de los otros modelos. Aunque el R^2 es muy bajo y aún negativo, el modelo está funcionando ligeramente mejor que Random Forest en términos de MSE.

Modelo seleccionado

Mejores parámetros para Gradient Boosting: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}

best_models = GradientBoostingRegressor(n_estimators=100, learning_rate=0.01, max_depth=3)
best_models.fit(X_train, y_train)

Predicción App cliente

La aplicación de nuestro cliente al que deseamos predecir Price son las siguientes:

datos = {
'Rating': [3.8],
'Reviews': [5200],
'Size': [48.0],
'Installs': [500000],
'Content Rating_Everyone 10+': [0],
'Content Rating_Mature 17+': [1],
'Content Rating_Teen': [0]
}

Predicciones para datos ficticios: [8.16123987]

El precio de la aplicación estima que deberia ser de $8.16

Muchas Gracias

Contáctanos si tienes alguna duda.

Carlos Müller C.

camullerco@gmail.com

+569 9289 7567

https://github.com/CarlosMullerC