Segunda pre entrega del proyecto final

Google Play Store Apps

Alonso Latella, Melanie
Arana, Matías

Índice

CODERHOUSE

Motivación y audiencia

CONTEXTO

Hoy en día hay una extensa lista de aplicaciones disponibles tanto para iOS como para Android, por eso se debe realizar un amplio análisis de lo existente para poder saber a qué público apuntar y donde se centralizan las descargas.

AUDIENCIA

Cualquier persona que quiera iniciarse en el mundo de las aplicaciones. Le ayudará a saber por cual clase de aplicación arrancar y cuales son los nichos a desarrollar.

PROBLEMA COMERCIAL

El mayor problema comercial es ofrecer algo novedoso y atractivo. Hay mucha competencia en el mercado y por ello se debe realizar un amplio análisis de qué es lo que este demanda. También se debe lograr atraer al público para que realice compras dentro de la aplicación y así obtener una mayor ganancia.

Metadata

Se utiliza un archivo que contiene datos de más de 600K aplicaciones con 23 atributos.

Los atributos que consideramos más importantes a analizar son:

- ★ Categoría
- ★ Rating
- ★ Descargas
- ★ Tamaño
- ★ Compras en la app

CODERHOUSE

Hipótesis

Nuestras ***preguntas principales*** se orientan a saber:

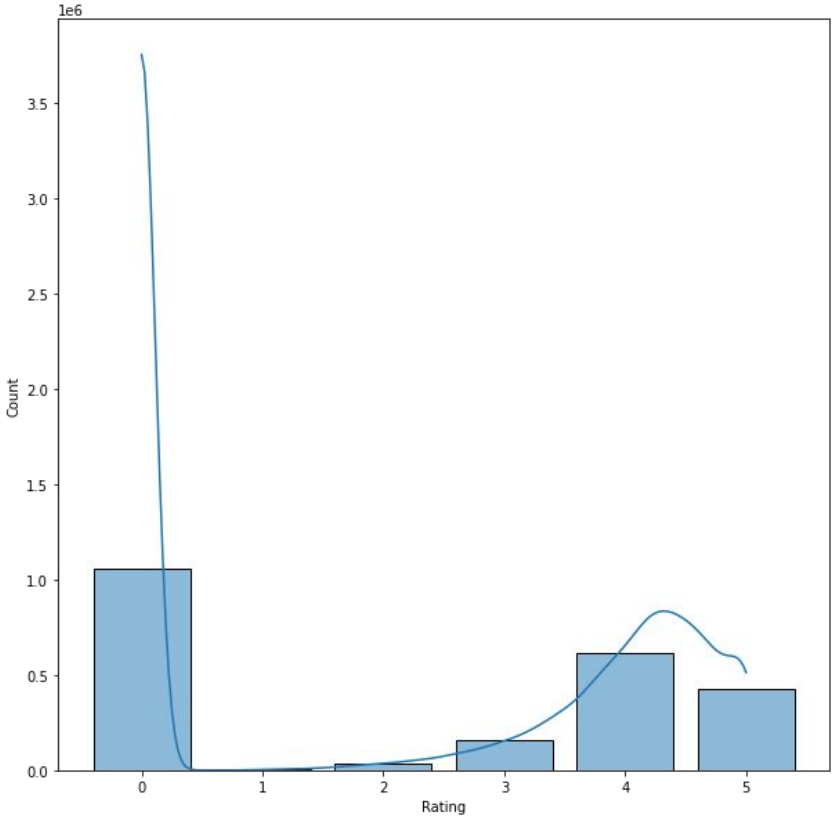
- ¿Qué clase de aplicación es la más bajada?
- ¿Se descargan más aplicaciones pagas o gratuitas?
- ¿El consumidor tiene en cuenta las calificaciones?

Para responder a esto, nos realizamos otras *preguntas de apoyo*:

- ¿Influye el tamaño de la aplicación?
- ¿Se pueden hacer compras dentro de la aplicación?
- ¿Qué clase de aplicación es la que más compras integradas tiene?

CODERHOUSE

Visualizaciones

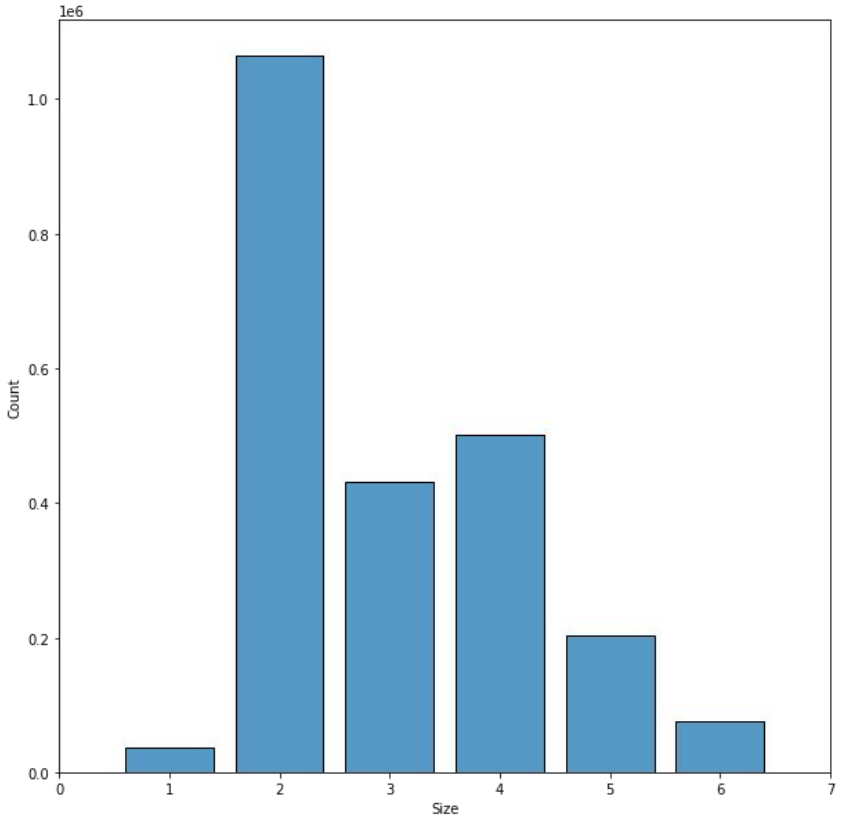


Se observa un gran número de aplicaciones calificadas entre 4 y 5 puntos.

Pero existe un gran número de apps no calificadas, sería interesante ver qué relación existe entre estas y la cantidad de descargas.

CODERHOUSE

Visualizaciones

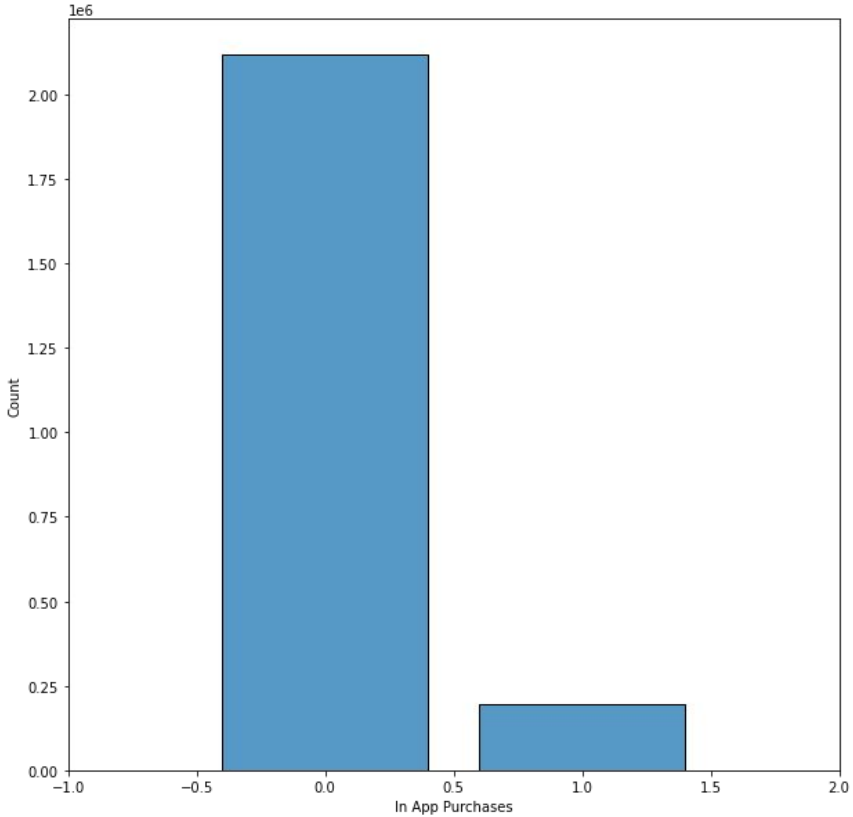


En cuanto al tamaño de las aplicaciones, los mayores porcentajes se encuentran en:

- ★ Grupo 2 (<10M)
- ★ Grupo 3 (<20M)
- ★ Grupo 4 (<50)

CODERHOUSE

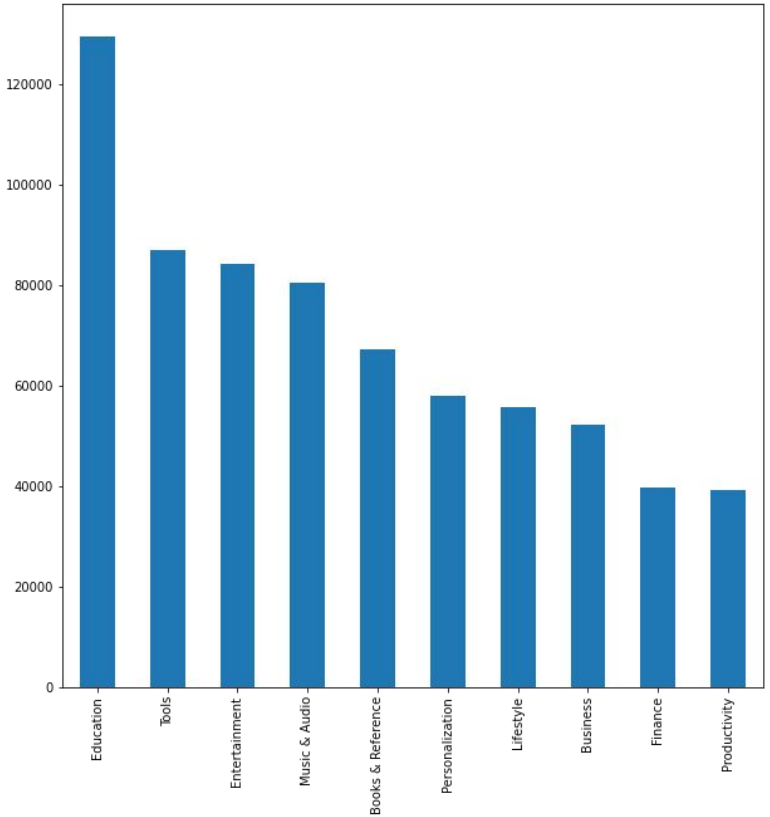
Visualizaciones



Encontramos que la mayoría de las aplicaciones no contienen compras incluidas.

CODERHOUSE

Visualizaciones



Se selecciona el TOP 10 de las categorías de aplicaciones descargadas.
Se observa que entre las categorías más presentes está la de educación, liderando por amplio margen

CODERHOUSE

Visualizaciones

Las variables más correlacionadas son el tamaño y las compras incluidas, así como el rating y las compras.

CODERHOUSE

Análisis univariado

Se analizó cada una de las variables por separado para entender mejor el conjunto de datos.
Se puede decir que una aplicación posiblemente orientada a ser de las más exitosas debe cumplir con:

- ★ La categoría más descargada es Educación y Música.
- ★ Estar dentro del Grupo 2 (<10M)
- ★ No hace falta tener compras incluidas

De este modo obtendrá buenas calificaciones, lo cual provocará que continúen bajando la aplicación.

CODERHOUSE

Análisis Multivariado

Empleamos este tipo de análisis para evaluar la relación de múltiples factores al mismo tiempo.
Se observó que las variables más correlacionadas son:

- Tamaño y compras incluidas
- Rating y compras incluidas

La matriz de correlación indicó que la correlación lineal es la indicada, por eso usamos pairplot.

CODERHOUSE

Limpieza de datos

- Como los valores del máximo de instalaciones está muy disperso, se acota los datos al 25% y 75%
- Para armonizar los datos, se divide por mil y redondea a un decimal. Siendo 1 equivalente a 1000 installs.
- Se convierten las variables categóricas en numericas.
- Se definen entradas y salidas
- Normalizamos columnas

CODERHOUSE

Entrenamiento, testeo y optimización

A través de un Random forest se intentaron crear varios árboles que trabajen en conjunto, pero el error que se presenta es bastante grande. Por este motivo, se pasó a la optimización de hiperparametros.
Se realizó la validación empleando Out-of-bag, k-cross-validation y neg-root-mean-squared-error.
Una vez identificados los mejores hiperparametros, se re-entrenó el modelo indicando los valores óptimos en sus argumentos. Si el GridSearchCV indica refit=True, este re-entrenamiento se hace automáticamente.

CODERHOUSE

Conclusión

Partiendo de las preguntas planteadas y debido al análisis realizado a lo largo del proyecto, con las distintas herramientas otorgadas en el curso, se pudieron tomar diversas conclusiones.
Las tres aplicaciones más descargadas según las categorías son Education, Tools y Entretainment.
Las aplicaciones con compras incluidas son un parámetro muy importante a tener en cuenta a la hora de las descargas.

Y uno de los parámetros más fuertes es la puntuación (o rating) a la hora de las descargas, teniendo el problema de que son muchas las aplicaciones que tienen descargas sin ninguna clase de clasificación.

Un gran problema que encontramos es que, a pesar de la limpieza realizada y las columnas agregadas, por la mala calidad de caudal de datos en el DataSet no se pudo llegar a entrenar el modelo. Se realizó todo el plan de entrenamiento, pero el mismo no funcionó. Por tal motivo, no se pudo predecir cuál sería la aplicación ideal para lanzar al mercado, siendo conocido que el mercado es muy variable y quizás la ambición del proyecto era muy alta.

CODERHOUSE

Futuras líneas

La idea de futuras líneas es implementar nuevos algoritmos para la predicción de valores, así como también, realizar un trabajo más exhaustivo explorando nuevos modelos y optimizaciones.

Es fundamental ampliar el dataset, con mejor calidad y caudal de información.

Creemos que con un dataset más sólido, sería posible que funcione el modelo y predecir cuál es el tipo de aplicación que el mercado solicita. Ayudando así a los desarrolladores donde orientarse a la hora de querer crear una aplicación atractiva para la audiencia.

CODERHOUSE

Gracias

CODERHOUSE