Introducción a la Ciencia de Datos

Curso 2023

Tarea final

Grupo 8

Juan Manuel Varela - CI: 4.802.705-4

Nicolás Farías - CI: 4.143.102-6

Contenido

Introducción

Este documento es un informe sobre el trabajo realizado para la tarea final del curso Introducción a la Ciencia de Datos. Para esto se seleccionó un conjunto de datos con características de 4803 películas, el cual fue extraído de la siguiente página.

- https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

Se buscará explorar el conjunto de datos y plantear una pregunta a responder en base al mismo, además de describir el proceso a seguir para realizar esto.

Carga y Exploración Inicial

El conjunto de datos contiene las siguientes columnas:

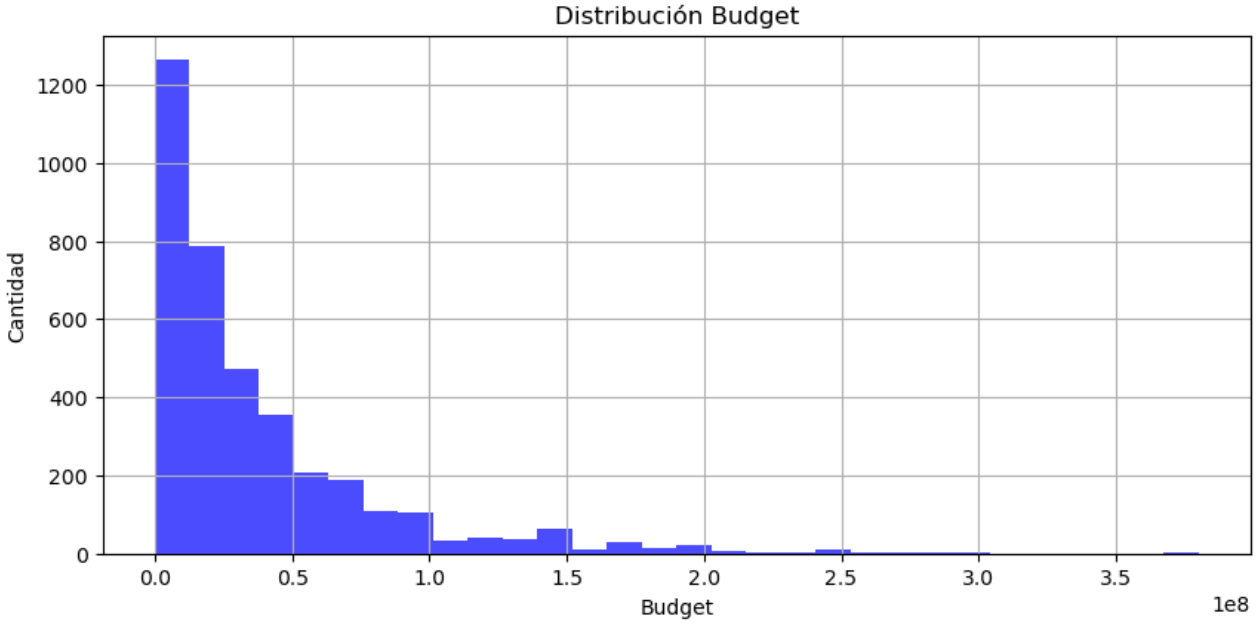
- budget: Presupuesto de la película.
- genres: Géneros de la película, en formato JSON.
- homepage: Página web oficial de la película.
- id: ID de la película.
- keywords: Palabras clave asociadas con la película, en formato JSON.
- original_language: Idioma original de la película.
- original_title: Título original de la película.
- overview: Resumen de la película.
- popularity: Popularidad de la película.
- production_companies: Compañías de producción de la película, en formato JSON.
- production_countries: Países de producción de la película, en formato JSON.
- release_date: Fecha de lanzamiento de la película.
- revenue: Ingresos de la película.
- runtime: Duración de la película en minutos.
- spoken_languages: Idiomas hablados en la película, en formato JSON.
- status: Estado de la película (p. ej., lanzada, en producción).
- tagline: Lema de la película.
- title: Título de la película.
- vote_average: Calificación promedio de la película.
- vote_count: Número de votos.

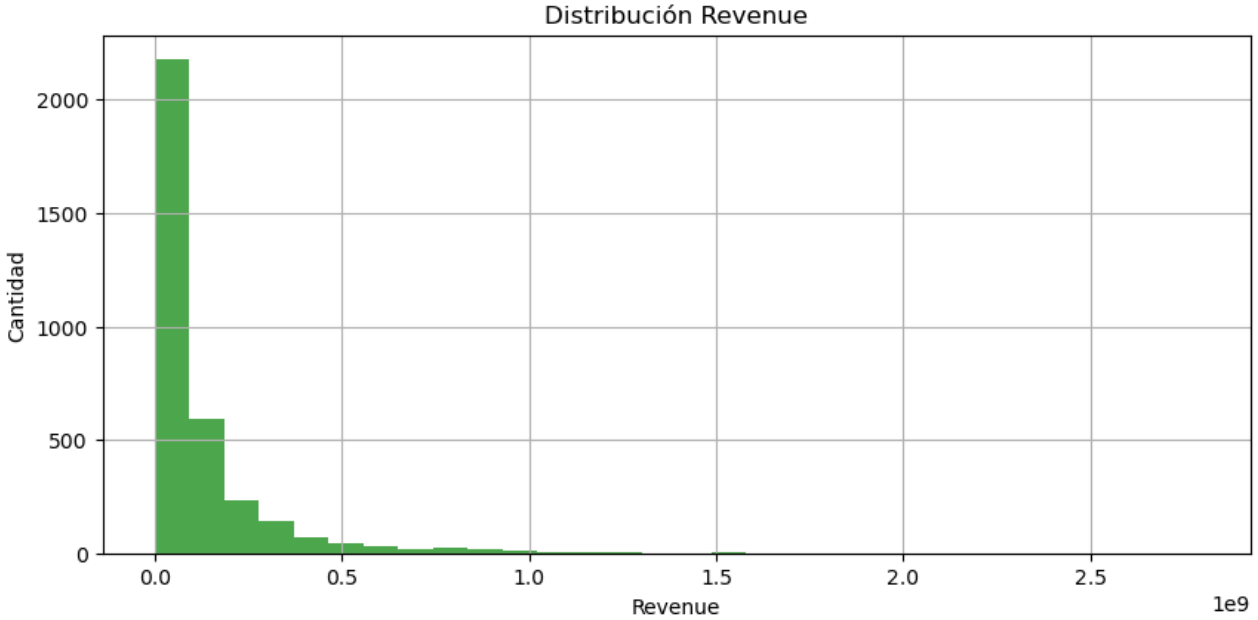
Algunos posibles problemas de calidad en este conjunto de datos podrían ser:

- Los valores JSON en las columnas genres, keywords, production_companies, production_countries y spoken_languages pueden ser difíciles de trabajar y requerirán un procesamiento adicional para extraer la información relevante.
- Se tiene valores faltantes en las columnas homepage (3091), overview (3), release_date (1), runtime (2) y tagline (844).
- La columna budget tiene 1037 valores que son cero, y la columna revenue tiene 1427 valores que son cero. Estos valores cero pueden representar datos faltantes o incorrectos y pueden requerir una limpieza adicional o una consideración especial en el análisis.
- En cuanto a la distribución, tanto budget como revenue tienen una alta desviación estándar, lo que indica una gran dispersión en los datos. Además, para ambas columnas,

la mediana es mucho menor que el promedio, lo que sugiere que la distribución de estos datos no es gaussiana, si no que tiene una inclinación hacia los valores bajos.

A continuación se muestra la distribución de estas columnas sin los valores nulos.





En general, este conjunto de datos parece tener suficiente información para realizar un análisis en profundidad, aunque puede requerir cierta limpieza y transformación de datos.

Problema propuesto

Un posible enfoque de análisis podría ser predecir el éxito de una película en función de sus características disponibles antes del lanzamiento. Podemos definir "éxito" en varios sentidos, como el ingreso bruto, la popularidad o la calificación promedio de la votación.

Los enfoques posibles son los siguientes:

- Definir el éxito en función del ingreso bruto, lo que significa que el interés principal es el rendimiento financiero de una película. Este enfoque es útil desde la perspectiva de un productor de cine o un estudio de cine.
- Definir el éxito en función de la calificación promedio de los votos significa que el interés principal es la recepción crítica de una película. Este enfoque puede ser útil desde la perspectiva del público, los actores, el director o los críticos de cine.
- También se puede definir un enfoque combinado, donde se considera tanto el ingreso bruto como la calificación. Este enfoque combinado puede proporcionar una visión más completa del éxito de una película.

Por lo tanto la pregunta propuesta es: ¿Se puede predecir el éxito de una película antes de su lanzamiento en función de características como el presupuesto, los géneros, la compañía de producción, el idioma y el país de producción? Esto se puede traducir a un problema de aprendizaje supervisado de tipo regresión.

Metodología propuesta

Para resolver el problema propuesto, se seguirían los siguientes pasos.

1. Limpieza y procesamiento de los datos: El conjunto de datos requiere una limpieza significativa. Se necesita resolver los valores faltantes y cero, y extraer información útil de las columnas JSON. Para los valores faltantes, se podrían llenar con un valor apropiado o eliminar las filas/columnas afectadas, dependiendo de la cantidad de datos faltantes y la importancia de la columna. Para los valores cero en las columnas budget y revenue, se podría realizar un análisis más detallado para decidir cómo manejarlos, ya que podrían representar datos faltantes, errores o simplemente películas de bajo presupuesto o ingresos bajos. Para las columnas JSON, se podría extraer la información relevante en nuevas columnas, por ejemplo, una columna para cada género o compañía de producción, utilizando la codificación *one hot encoding*. De acuerdo al enfoque para medir el éxito que se haya tomado, puede ser necesario crear una columna en el dataset con un nuevo valor calculado.

2. Análisis exploratorio de los datos (EDA): Este paso implica comprender más a fondo las variables y su relación con el éxito de la película. Se puede utilizar visualizaciones como histogramas, boxplots y scatterplots, así como estadísticas descriptivas. También se puede investigar correlaciones entre las variables. Se puede aplicar la técnica PCA para graficar en un plano las características (nube de variables). Esto puede ayudar a buscar las correlaciones de manera visual.

3. Selección de características: Basándose en el EDA, se seleccionarían las características que parecen tener el mayor impacto en el éxito de la película. También se podría considerar la ingeniería de nuevas características, en caso de parecer útil.

4. Modelado: Con las características seleccionadas, se podría probar varios modelos para predecir el éxito de la película. Posibles opciones podrían ser la regresión lineal, árboles de decisión o random forest. Se dividirían los datos en un conjunto de entrenamiento y un conjunto de prueba, se entrenaría los modelos en el conjunto de entrenamiento y luego se evaluaría su rendimiento en el conjunto de prueba.

5. Evaluación y ajuste del modelo: Se evaluaría el rendimiento de los modelos utilizando métricas apropiadas, como la raíz del error cuadrático medio (RMSE) para la regresión. Basándose en la evaluación, se ajustarían los hiperparámetros de los modelos, y se seleccionaría el modelo que ofrezca el mejor rendimiento. Para esto se podría separar un conjunto de datos para validación o usar la validación cruzada.

6. Entrenamiento y evaluación del modelo final: Se entrenaría el modelo seleccionado sobre todo el conjunto de entrenamiento. Finalmente se evaluaría su rendimiento sobre el conjunto de evaluación.

7. Por último, para comunicar los hallazgos, se crearían visualizaciones que muestren las características más importantes para predecir el éxito de una película, y cómo el modelo seleccionado las utiliza para hacer sus predicciones.