Predicción de Ventas Globales de Videojuegos

Michael Steven Cuello

Esteban López Arévalo

Mateo Aldana Escobar

Universidad EIA

Proyecto final programación

05/06/2024

UNIVERSIDAD EIA
Ser, Saber y Servir

1. INTRODUCCIÓN

La industria de los videojuegos es un sector dinámico y en crecimiento constante, convirtiéndose en una de las mayores industrias del entretenimiento digital. Comprender los factores que influyen en las ventas de videojuegos es crucial para desarrolladores, publicistas y distribuidores. Este informe presenta un proyecto de analítica de datos cuyo objetivo es predecir las ventas globales de videojuegos.

Para este análisis, se utilizó un dataset con información de ventas de videojuegos hasta 2016, que incluye atributos como nombre, plataforma, año de lanzamiento, género, editor, ventas por región (Norteamérica, Europa, Japón y otras), puntuaciones de críticos y usuarios, número de críticas, desarrollador y clasificación del juego.

El análisis se estructuró en varias etapas. Primero, se definió el problema predictivo, estableciendo como objetivo principal la predicción de las ventas globales. Luego, se realizó un análisis exploratorio de los datos (EDA) para comprender la distribución de las variables y las relaciones entre ellas. Posteriormente, se llevó a cabo el preprocesado y limpieza de datos, eliminando datos faltantes y duplicados, y normalizando las variables.

En la fase de modelado, se seleccionaron y optimizaron dos algoritmos supervisados (Random Forest y Árboles de Decisión) utilizando técnicas de búsqueda de hiperparámetros. Los modelos se evaluaron mediante curvas de aprendizaje y métricas de desempeño, como el MSE y el R^2, identificando posibles problemas como el overfitting. Finalmente, se compararon los modelos para determinar el más efectivo en la predicción de ventas.

2. DEFINICÓN DEL PROBELMA

El propósito primordial de este proyecto es desarrollar un modelo predictivo capaz de estimar las ventas globales de videojuegos. La industria del entretenimiento interactivo es altamente dinámica y está influenciada por una variedad de factores. Por lo tanto, se requiere un análisis exhaustivo para comprender y predecir las tendencias de ventas.

El análisis se centrará en examinar múltiples variables que potencialmente influyen en las ventas de videojuegos. Estas variables incluyen, pero no se limitan a, la plataforma de lanzamiento del juego, el género del juego y las calificaciones otorgadas por críticos y usuarios. Cada uno de estos elementos puede desempeñar un papel crucial en la recepción y el éxito comercial de un videojuego.

La comprensión de estos factores y su interacción puede proporcionar información valiosa para los desarrolladores de videojuegos, distribuidores y otras partes interesadas en la industria. Al predecir las ventas globales con precisión, se puede mejorar la toma de decisiones estratégicas relacionadas con la inversión en desarrollo, la comercialización y la distribución de videojuegos.

Este proyecto no solo se enfoca en la aplicación de técnicas de aprendizaje automático para predecir las ventas, sino también en la interpretación de los resultados para identificar patrones significativos y tendencias emergentes en el mercado de videojuegos. La capacidad de anticipar la demanda y comprender las preferencias de los consumidores puede ofrecer una ventaja competitiva significativa en una industria tan competitiva y en constante evolución.

3. OBTENCIÓN DEL DATASET

El conjunto de datos utilizado en este proyecto fue adquirido de una fuente pública y se compone de más de 5000 instancias. Estas instancias abarcan un total de 15 atributos, que proporcionan una amplia gama de detalles sobre cada videojuego incluido en el conjunto de datos. Entre estos atributos se encuentran el nombre del juego, la plataforma para la que está disponible, el año de lanzamiento, el género, el editor responsable de su producción, las ventas registradas en diversas regiones del mundo, así como las puntuaciones otorgadas por críticos y usuarios, entre otros aspectos relevantes.

4. ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

Se llevó a cabo un análisis exploratorio exhaustivo con el propósito de comprender en profundidad la distribución de las variables y las relaciones entre ellas dentro del conjunto de datos. Este análisis abarcó un conjunto de técnicas estadísticas descriptivas y visualizaciones gráficas, incluyendo histogramas, diagramas de dispersión y mapas de calor.

Entre las observaciones destacadas durante el análisis exploratorio, se identificó que el género "Acción" mostraba consistentemente las mayores ventas en múltiples regiones, lo que sugiere una tendencia global hacia este tipo de juegos. Sin embargo, se notó una excepción notable en el mercado japonés, donde los juegos de rol predominaban sobre los de acción. Este hallazgo subraya la importancia de considerar las preferencias regionales al diseñar estrategias de lanzamiento y comercialización para maximizar el potencial de ventas en diferentes mercados.

5. PROCESADO Y LIMPIEZA DE DATOS

Se llevó a cabo un proceso exhaustivo de preprocesado y limpieza de datos con el objetivo de garantizar la calidad y la coherencia del conjunto de datos utilizado en el proyecto. Esto incluyó varias etapas, como la identificación y eliminación de datos faltantes y duplicados, así como la normalización de variables y la codificación de variables categóricas.

En primer lugar, se realizó una revisión exhaustiva para identificar y manejar los datos faltantes. En el caso de la variable 'Year_of_Release', por ejemplo, se optó por llenar los valores faltantes con la mediana de la columna, asegurando así la integridad de la información sin introducir sesgos significativos en el conjunto de datos.

Posteriormente, se llevó a cabo un proceso de normalización de variables para garantizar que todas las características del conjunto de datos estuvieran en una escala uniforme. Esto es fundamental para muchos algoritmos de aprendizaje automático, ya que ayuda a evitar que las características con escalas más grandes dominen las más pequeñas.

Finalmente, se codificaron las variables categóricas utilizando técnicas como la codificación one-hot, asegurando así que todas las variables fueran representadas de manera adecuada para su posterior análisis y modelado.

Este proceso de preprocesado y limpieza de datos es crucial para garantizar la fiabilidad y la eficacia de los modelos predictivos desarrollados en el proyecto, proporcionando una base sólida para la extracción de conocimientos y la toma de decisiones informadas.

6. SELECCIÓN Y OPTIMIZACIÓN DE ALGORITMOS PREDICTIVOS

En esta etapa del proyecto, se llevó a cabo un proceso de selección y optimización de algoritmos predictivos con el fin de desarrollar modelos robustos y precisos para la predicción de las ventas globales de videojuegos. Se optó por utilizar dos algoritmos supervisados ampliamente reconocidos: Random Forest y Árboles de Decisión.

La selección de estos algoritmos se basó en su capacidad para manejar conjuntos de datos complejos y su flexibilidad para capturar relaciones no lineales entre las variables predictoras y la variable objetivo. Además, ambos algoritmos ofrecen la ventaja de ser relativamente fáciles de interpretar, lo que facilita la comprensión de los factores que influyen en las predicciones.

Una vez seleccionados los algoritmos, se procedió a optimizar sus hiperparámetros mediante técnicas de búsqueda exhaustiva, específicamente utilizando GridSearchCV. Esta técnica permite explorar sistemáticamente el espacio de hiperparámetros y encontrar la combinación óptima que maximiza el rendimiento del modelo en términos de precisión predictiva.

El proceso de optimización se llevó a cabo con el objetivo de encontrar la configuración de hiperparámetros que maximizara la capacidad de generalización de los modelos, evitando el sobreajuste y garantizando su capacidad para realizar predicciones precisas en datos no vistos.

Este enfoque sistemático y riguroso en la selección y optimización de algoritmos es fundamental para asegurar la calidad y el rendimiento de los modelos predictivos desarrollados en el proyecto, proporcionando resultados confiables y significativos para la toma de decisiones en la industria de los videojuegos.

7. ENTRENAMIENTO Y EVALUACIÓN DE MODELOS

En esta fase del proyecto, se procedió a entrenar y evaluar los modelos predictivos desarrollados previamente utilizando los algoritmos Random Forest y Árboles de Decisión. Para evaluar la eficacia y la precisión de los modelos, se emplearon métricas comúnmente utilizadas en problemas de regresión, como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2).

El MSE proporciona una medida de la diferencia entre los valores predichos por el modelo y los valores reales de las ventas globales de videojuegos. Un valor de MSE más bajo indica un mejor ajuste del modelo a los datos observados.

Por otro lado, el coeficiente de determinación R^2 ofrece una medida de cuánto de la variabilidad en la variable objetivo puede explicar el modelo. Un valor de R^2 más cercano a 1 indica un mejor

ajuste del modelo a los datos, mientras que valores cercanos a 0 sugieren que el modelo no es capaz de explicar la variabilidad en los datos.

Además de evaluar las métricas de rendimiento, se llevaron a cabo análisis adicionales para diagnosticar posibles problemas como el sobreajuste (overfitting) de los modelos. Se realizaron curvas de aprendizaje para examinar la relación entre el rendimiento del modelo y el tamaño del conjunto de datos de entrenamiento. Esto permitió identificar si los modelos estaban aprendiendo de manera efectiva de los datos o si estaban sobreajustando, es decir, memorizando los datos de entrenamiento sin generalizar bien a nuevos datos.

Estos análisis de entrenamiento y evaluación son fundamentales para comprender la capacidad predictiva de los modelos desarrollados, así como para identificar áreas de mejora y optimización en el proceso de modelado. Los resultados obtenidos proporcionan información valiosa para la iteración y refinamiento de los modelos, con el objetivo de mejorar su rendimiento y su capacidad para realizar predicciones precisas en situaciones del mundo real.

8. COMPARACION DEL DESEMPEÑO DE LOS MODELOS

En esta etapa del análisis, se procedió a comparar el desempeño de los modelos desarrollados en función de su capacidad predictiva para estimar las ventas globales de videojuegos. Se evaluaron varios modelos, incluyendo Random Forest, Árboles de Decisión y Regresión Lineal.

Los resultados obtenidos indicaron que el modelo de Random Forest exhibió el mejor desempeño predictivo en comparación con los otros modelos evaluados. Este hallazgo sugiere que la complejidad y la capacidad de adaptación del algoritmo Random Forest fueron beneficiosas para capturar las relaciones no lineales presentes en los datos y para realizar predicciones más precisas.

En segundo lugar, se encontraron los modelos basados en Árboles de Decisión, los cuales también mostraron un rendimiento considerablemente bueno, aunque ligeramente inferior al de Random Forest. Los Árboles de Decisión son conocidos por su capacidad para manejar relaciones complejas entre variables predictoras y la variable objetivo, lo que los hace una opción sólida para problemas de regresión como este.

Por último, se evaluó un modelo de Regresión Lineal, que mostró un desempeño inferior en comparación con los otros modelos. Aunque la Regresión Lineal es un enfoque más simple y fácil de interpretar, su capacidad para capturar relaciones no lineales puede ser limitada en comparación con algoritmos más complejos como Random Forest y Árboles de Decisión.

Es importante destacar que, a pesar de las diferencias en el desempeño entre los modelos, todos mostraron un coeficiente de determinación (R^2) relativamente bajo. Esto sugiere que, aunque los modelos son capaces de hacer predicciones, aún existe una cantidad significativa de variabilidad en los datos que no están siendo capturados por los modelos desarrollados.

Estos resultados subrayan la importancia de continuar iterando y refinando los modelos, así como de explorar técnicas adicionales para mejorar su capacidad predictiva. Además, destacan la necesidad de considerar otros factores y características que puedan influir en las ventas globales de videojuegos, con el fin de desarrollar modelos más precisos y robustos en futuras investigaciones.

9. CONCLUSIONES Y RECOMENDACIONES

Basándonos en los resultados obtenidos de este estudio, se llega a la conclusión de que se necesitan modelos más sofisticados y una mayor cantidad de características predictivas para mejorar el rendimiento en la predicción de las ventas globales de videojuegos.

A pesar de los esfuerzos realizados en la selección y optimización de los modelos, los resultados muestran que aún queda margen para mejorar la precisión de las predicciones. Esto sugiere que el conjunto actual de características predictoras puede no ser suficiente para capturar toda la complejidad y variabilidad presente en los datos.

Como recomendación para futuras investigaciones, se sugiere explorar modelos más avanzados como Gradient Boosting Machines y XGBoost, que han demostrado ser altamente efectivos en una variedad de problemas de regresión. Estos modelos tienen la capacidad de aprender de los errores de los modelos anteriores y mejorar gradualmente su rendimiento, lo que los hace una opción prometedora para mejorar la precisión de las predicciones en este contexto.

Además, se recomienda realizar una búsqueda más exhaustiva de hiperparámetros para encontrar la combinación óptima que maximice el rendimiento de los modelos. La optimización de hiperparámetros es un proceso crucial en el desarrollo de modelos de aprendizaje automático, ya que puede tener un impacto significativo en la capacidad predictiva y generalización de los modelos.

Por último, se sugiere explorar la aplicación de transformaciones adicionales de datos, como la ingeniería de características y la selección de características, para mejorar la calidad de los datos de entrada y aumentar la capacidad de los modelos para capturar las relaciones subyacentes en los datos.

En resumen, si bien este estudio proporciona información valiosa sobre la predicción de las ventas globales de videojuegos, se necesitan esfuerzos adicionales para mejorar la precisión de las predicciones y desarrollar modelos más robustos y confiables en este dominio.