UNIVERSIDAD NACIONAL SAN AGUSTÍN

FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y SERVICIOS

ESCUELA PROFESIONAL DE CIENCIAS DE LA COMPUTACIÓN

Tópicos en Ciencias de Datos

Alumno:

Apaza Andaluz, Diego Francisco

AREQUIPA

2024

I. Titulo:

Análisis de videojuegos con loot boxes en Steam: Tendencias, géneros y comportamiento de usuarios a través de ciencia de datos

II. Introducción

En la última década, la industria de los videojuegos ha experimentado una transformación profunda, especialmente en sus modelos de monetización. El esquema tradicional de compra única ha sido reemplazado o complementado por estrategias basadas en microtransacciones, en particular dentro del modelo Free-to-Play (F2P). Entre estas, una de las más controversiales es la implementación de loot boxes: paquetes virtuales con recompensas aleatorias que pueden adquirirse con dinero real o moneda del juego. Estas mecánicas, presentes en una amplia variedad de títulos, han sido comparadas con los juegos de azar debido a su aleatoriedad, falta de transparencia y fuerte componente de refuerzo psicológico, especialmente en audiencias jóvenes [1], [2].

Diversos estudios han resaltado que las loot boxes pueden tener efectos negativos sobre el comportamiento de los jugadores. Wardle y Zendle encontraron una relación significativa entre la compra de loot boxes y síntomas de juego problemático, incluso tras controlar variables demográficas y de impulsividad [3]. Gibson et al. identificaron que las microtransacciones, incluyendo loot boxes, se asocian con un mayor gasto impulsivo y signos de adicción [4]. Desde una perspectiva computacional, investigaciones recientes como la de Kovačević et al. han demostrado que los comportamientos de compra dentro de videojuegos pueden ser modelados eficazmente usando técnicas de aprendizaje profundo como Transformers, lo que destaca el potencial de aplicar métodos computacionales al estudio de la monetización en videojuegos [5].

A pesar del creciente interés en las loot boxes desde el ámbito psicológico y ético, existe una carencia de estudios empíricos a gran escala que analicen su distribución y comportamiento desde un enfoque computacional. En particular, no se ha explorado sistemáticamente cómo se implementan las loot boxes en plataformas masivas como Steam, qué géneros o modelos de negocio las utilizan más, ni si su presencia se relaciona con métricas de éxito como la popularidad o el tiempo de juego.

Actualmente no existen análisis cuantitativos sistemáticos que caractericen computacionalmente la adopción de loot boxes como estrategia de monetización dominante en Steam, ni se ha evaluado su correlación con variables como género, modelo de negocio (F2P o P2P), año de publicación o métricas de popularidad (jugadores concurrentes y tiempo promedio jugado). El objetivo general de este proyecto es aplicar técnicas de ciencia de datos para analizar la presencia, evolución y patrones de comportamiento de los videojuegos que implementan loot boxes en Steam

Trabajos Relacionados:

Zendle et al. (2020) analizó la exposición de jugadores a loot boxes, microtransacciones cosméticas y "pay-to-win" en los juegos más populares de Steam (2010-2019) mediante joinpoint regression. Sus resultados mostraron un rápido crecimiento de loot boxes y cosméticas entre 2012-2014, alcanzando una alta prevalencia en 2019 (71.2% y 85.89%, respectivamente), mientras que el modelo "pay-to-win" tuvo menor adopción. Este estudio es relevante para el análisis de tendencias históricas en monetización, aunque no explora diferencias por género ni comportamiento individual.

Xiao et al. (2022) revisó las similitudes entre loot boxes y gambling, destacando su potencial adictivo y el marco regulatorio global. Clasificaron las loot boxes en categorías según su monetización y transferibilidad (e.g., Embedded-Embedded), encontrando correlaciones con problemas de juego en jóvenes. Aunque no utiliza datos cuantitativos, este trabajo proporciona un contexto crucial sobre los riesgos éticos y las políticas aplicables, útil para discutir implicaciones sociales en estudios de monetización.

Wardle y Zendle (2020) investigaron la relación entre las loot boxes, el juego de azar y el juego problemático en jóvenes de 16 a 24 años mediante una encuesta transversal en línea a 3,549 participantes en Gran Bretaña. Los resultados revelaron que el 12.1% de los jóvenes había comprado loot boxes en el último año, mostrando una fuerte asociación con el juego problemático: en modelos no ajustados, la probabilidad de juego problemático fue 11.4 veces mayor entre compradores de loot boxes (IC 95%: 7.6–16.9). Esta asociación se atenuó pero persistió significativa (OR = 4.5, IC 95%: 2.6–7.9) al controlar por participación en otros juegos de azar, impulsividad y factores sociodemográficos. Los compradores de loot boxes presentaron mayor impulsividad, gasto semanal en apuestas (£19.20 vs. £5.50) y prevalencia de juego problemático (16.9% vs. 1.8%). El estudio sugiere que las loot boxes tienen una asociación con el juego problemático comparable a juegos de casino en línea, destacando la necesidad de regulación. Las limitaciones incluyen el diseño transversal (sin inferencia causal) y el sesgo de autoselección en muestras en línea. Los hallazgos respaldan políticas que consideren las loot boxes como productos de alto riesgo, especialmente para jóvenes.

Kovačević et al. (2024) desarrolló un modelo predictivo basado en Transformers para anticipar compras en juegos free-to-play, utilizando datos granulares de 977K jugadores. Su enfoque en secuencias temporales (tokenización de historiales de juego) superó a métodos tradicionales (Random Forest, XGBoost), logrando un F1-score de 0.85. Este estudio es pionero en aplicar deep learning al comportamiento de jugadores, ofreciendo metodologías transferibles para analizar engagement con loot boxes en plataformas como Steam.

Gibson et al. (2022) realizaron una revisión sistemática sobre la relación entre microtransacciones (especialmente loot boxes), el juego problemático y el gambling, analizando 19 estudios empíricos. Su objetivo fue evaluar el impacto psicológico y conductual de estos sistemas, destacando correlaciones significativas entre el gasto en loot boxes y el gambling problemático, con mayor prevalencia en jóvenes varones. A diferencia de mi enfoque en modelado predictivo con datos cuantitativos de Steam, su metodología se basó en síntesis cualitativas, identificando mecanismos como refuerzo intermitente y similitudes fisiológicas con el gambling. Sus hallazgos son relevantes para mi investigación al ofrecer un marco teórico sobre patrones de riesgo, aunque contrastan en enfoque: ellos priorizan implicaciones clínicas y regulatorias, mientras mi trabajo cuantifica tendencias comerciales y de engagement. Proponen regular loot boxes como gambling, una perspectiva útil para discutir implicaciones éticas en monetización.

Xiao (2025) investigó el cumplimiento regulatorio en la publicidad de loot boxes en redes sociales, analizando repositorios de anuncios de Meta y TikTok. El estudio reveló que solo el 7% de los anuncios cumplía con los requisitos de transparencia del Reino Unido y la UE, mientras que el 93% omitía información clave sobre la presencia de loot boxes. A través de un análisis empírico de 185 anuncios en Meta y 100 en TikTok, se demostró que estas prácticas engañosas alcanzaron más de 292 millones de impresiones en el Reino Unido, exponiendo repetidamente a los usuarios a publicidad no regulada. Aunque centrado en el marco legal británico, el trabajo ofrece evidencia crítica sobre las fallas en la autorregulación de la industria y la necesidad de una aplicación más estricta, proporcionando bases para discutir la protección al consumidor en sistemas de monetización de videojuegos.

III. Visión General

3.1. Descripción del dataset

El conjunto de datos utilizado en este estudio corresponde a una recopilación de información sobre videojuegos publicados en la plataforma Steam, una de las principales distribuidoras digitales de juegos para PC. El dataset fue construido a partir de múltiples fuentes públicas, incluyendo SteamSpy y SteamDB, y contiene información tabular sobre más de 100,000 videojuegos, con variables relacionadas a su lanzamiento, popularidad, monetización y características técnicas.

Cada registro del dataset representa un videojuego individual. El periodo de tiempo cubierto va aproximadamente desde 2005 hasta 2023. El conjunto original contiene más de 80 columnas, de las cuales se seleccionaron 25 atributos relevantes para esta investigación.

A continuación se describen algunos de los atributos clave utilizados:

- Name: nombre del videojuego (categórica, no nula).
- Release date: fecha de lanzamiento (temporal, algunas fechas faltantes).
- Estimated owners: rango de copias estimadas vendidas (ordinal, categórica).
- Peak CCU: cantidad máxima de jugadores concurrentes (numérica, continua).
- Price: precio base del juego en dólares estadounidenses (numérica, continua).
- Genres: lista de géneros del juego (categórica multivaluada).
- Tags: etiquetas descriptivas de jugabilidad o estética (texto libre, categórica multivaluada).
- FreeToPlay: indicador binario si el juego es gratuito o no (booleana).
- HasMicrotransactions: indicador si el juego incluye microtransacciones (booleana).
- HasLootboxKeywords: variable binaria derivada que identifica si el juego contiene términos relacionados a loot boxes en sus descripciones, etiquetas o categorías.

El atributo Release Year fue derivado de la fecha de lanzamiento, y se utiliza para análisis temporal. Las variables de monetización (como FreeToPlay y HasLootboxKeywords) fueron generadas mediante reglas heurísticas y búsqueda de patrones de texto en campos como Tags, Categories y About the Game.

3.2. Pre-procesamiento de datos

El pre-procesamiento del conjunto de datos incluyó varios pasos esenciales para asegurar su calidad y compatibilidad con técnicas de análisis exploratorio y aprendizaje automático:

a) Normalización de tipos de datos:
Las columnas de fecha se transformaron al tipo datetime. Las variables categóricas se mantuvieron como texto, y las variables booleanas fueron convertidas a True/False. Las columnas de texto multivaluadas (como Tags y Genres) se separaron por comas y se limpiaron los espacios y duplicados.

b) Creación de variables derivadas:
Se generaron columnas adicionales como Release Year (año de publicación), FreeToPlay (a partir de la columna Price y categoría), HasLootboxKeywords (mediante búsqueda de términos como “loot box”, “gacha”, “crates”, etc.), y HasMicrotransactions (con términos como “in-app purchase”, “store”, “DLC”, etc.).

c) Manejo de valores nulos:
Se eliminaron registros con fecha inválida, y en campos como Genres, Tags y Categories, los valores faltantes se reemplazaron por cadenas vacías. Para campos numéricos como Peak CCU, se utilizaron cero o el promedio como estrategia de imputación, según el caso.

d) Reducción dimensional:
Se descartaron columnas irrelevantes para el análisis (por ejemplo, enlaces, imágenes o soporte técnico) y se mantuvo un subconjunto optimizado para visualización y modelado.

e) Balanceo y codificación:
En etapas posteriores se aplicó codificación categórica (one-hot encoding) a variables como género para facilitar su uso en modelos computacionales.

IV. Diseño del Sistema:

Estructura General

El dashboard está organizado en dos áreas principales:

- Barra lateral (sidebar): Contiene todos los controles de filtrado interactivos
- Área principal: Muestra visualizaciones y métricas basadas en los filtros aplicados
- Sección de Filtros (Sidebar)

1. Filtro por Tipo de Juego

- Elemento: Radio buttons ("Todos", "Solo F2P", "Solo P2P")
- Encoding: Selección categórica que filtra el dataset por modelo de negocio
- Propósito: Permitir comparar juegos Free-to-Play vs Pay-to-Play

2. Filtro por Géneros

- **Elemento:** Multiselect dropdown
- **Encoding:** Lista de géneros extraídos del campo "Genres" (excluyendo categorías genéricas)
- **Propósito:** Analizar patrones específicos por género

3. Filtro por Etiquetas (Tags)

- Elemento: Multiselect dropdown
- Encoding: Lista de tags extraídos del campo "Tags"

- Propósito: Identificar características comunes en juegos con lootboxes

4. Filtro Temporal

- Elemento: Slider de rango
- Encoding: Años de lanzamiento (Release Year) con escala lineal
- Propósito: Analizar tendencias temporales

5. Filtro por Popularidad

- Elemento: Slider de rango para Peak CCU (Concurrent Players)
- Encoding: Número máximo de jugadores concurrentes con escala lineal
- Propósito: Filtrar por popularidad del juego

6. Filtro por Precio

- Elemento: Slider de rango
- Encoding: Precio en USD con escala lineal
- Propósito: Analizar relación entre precio y presencia de lootboxes

7. Filtro por Edad Requerida

- Elemento: Slider de rango
- Encoding: Edad mínima recomendada con escala lineal
- Propósito: Estudiar distribución por clasificación etaria

8. Filtro por Tiempo de Juego

- Elemento: Slider de rango (minutos)
- Encoding: Tiempo promedio de juego con escala lineal
- Propósito: Analizar engagement de los jugadores

9. Filtro por Plataformas

- Elemento: Checkbox multiselect
- Encoding: Plataformas soportadas (Windows/Mac/Linux)
- Propósito: Comparar disponibilidad entre sistemas

Área Principal de Visualización

1. Métricas Principales

Encoding:

- Número absoluto de juegos que cumplen los filtros
- Comparación con el total general de juegos con lootboxes
- Propósito: Dar contexto sobre el alcance de los filtros aplicados

2. Visualización de Precios

Gráfico: Histograma + KDE

Encoding:

- Eje X: Precio en USD (escala lineal)
- Eje Y: Frecuencia (escala lineal)
- Color: Azul claro ("skyblue")
- Propósito: Mostrar distribución de precios y posibles agrupaciones

3. Hipótesis 1: Comparación F2P vs P2P

Gráfico: Bar plot

Encoding:

- Eje X: Categorías F2P/P2P (nominal)
- Eje Y: Peak CCU promedio (escala lineal)
- Color: Verde claro (F2P) vs Salmón (P2P) - codificación categórica
- Propósito: Validar si los juegos F2P atraen más jugadores

4. Hipótesis 2: Evolución Temporal

Gráfico: Line plot

Encoding:

- Eje X: Año (escala temporal)
- Eje Y: Porcentaje de juegos con lootboxes (escala lineal)
- Color: Naranja (énfasis)
- Marcadores: Puntos circulares
- Propósito: Mostrar tendencia histórica de adopción de lootboxes

5. Hipótesis 3: Géneros Populares

- Gráfico: Bar plot horizontal
- Encoding:
- Eje X: Peak CCU acumulado (escala lineal)
- Eje Y: Géneros (nominal, ordenado por valor)
- Color: Púrpura medio
- Propósito: Identificar qué géneros tienen mayor engagement

6. Tabla de Resultados

Encoding:

- Filas: Juegos individuales
- Columnas: Atributos clave (Nombre, Géneros, Tags, etc.)
- Orden: Descendente por Peak CCU
- Propósito: Permitir exploración detallada de los resultados

Principios de Visualización Aplicados

1. **Jerarquía Visual:** Las métricas clave aparecen primero, seguidas de hipótesis y finalmente datos crudos

2. **Comparabilidad:** Gráficos side-by-side para comparación directa (F2P vs P2P)
3. **Consistencia:** Mismo esquema de color en todo el dashboard
4. **Interactividad:** Todos los filtros actualizan todas las visualizaciones
5. **Responsividad:** Diseño adaptable a diferentes tamaños de pantalla

Referencias:

1. Xiao LY, Henderson LL, Nielsen RKL, Grabarczyk P, Newall PWS. Loot boxes: Gambling-like mechanics in video games. In: Handbook of Digital Games and Entertainment Technologies. Springer; 2022. https://doi.org/10.1007/978-3-319-08234-9_459-1

2. Zendle D, Meyer R, Ballou N. The changing face of desktop video game monetisation: An exploration of exposure to loot boxes, pay to win, and cosmetic microtransactions in the most-played Steam games of 2010–2019. PLoS ONE. 2020;15(5):e0232780. https://doi.org/10.1371/journal.pone.0232780

3. Wardle H, Zendle D. Loot boxes, gambling, and problem gambling among young people: Results from a cross-sectional online survey. Cyberpsychol Behav Soc Netw. 2021;24(4):267–274. https://doi.org/10.1089/cyber.2020.0299

4. Gibson E, Griffiths MD, Calado F, Harris A. The relationship between videogame micro-transactions and problem gaming and gambling: A systematic review. Comput Human Behav. 2022;131:107219. https://doi.org/10.1016/j.chb.2022.107219

5. Kovačević MA, Pešović MD, Petrović ZZ, Pucanović ZS. Predictive analytics of in-game transactions: Tokenized player history and self-attention techniques. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.0429000

6. Xiao, L. Y. (2025). Illegal loot box advertising on social media? An empirical study using the Meta and TikTok ad transparency repositories. Computer Law & Security Review, 56, 106069. https://doi.org/10.1016/j.clsr.2024.106069