UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

Extracción de Conocimiento en Bases de Datos

V.2. Elaboración de gráficas

Docente:

Enrique Mascote

Presentan:

Ian Carlos Chávez Rojo

Grupo:

IDGS91N

Fecha: 28/11/2025

Índice

Introducción

El presente reporte documenta el desarrollo de una aplicación interactiva para el análisis y predicción de ventas de videojuegos indie utilizando técnicas de

aprendizaje supervisado. El proyecto se fundamenta en un modelo de regresión lineal que permite estimar el comportamiento comercial de títulos independientes 2D basándose en variables como presupuesto de desarrollo, duración del proyecto, estrategias de marketing y presencia en redes sociales.

Este documento presenta tres componentes principales: primero, una aplicación web interactiva desarrollada en Python que integra visualizaciones avanzadas para el análisis exploratorio y evaluación del modelo; segundo, un análisis detallado de cada gráfica generada, explicando su interpretación y valor para la toma de decisiones; y tercero, una reflexión sobre la aplicabilidad práctica de estas herramientas en el contexto de la industria del desarrollo de videojuegos independientes.

Interpretación Detallada de las Gráficas

Gráficas Originales del Modelo

Valores Reales vs Predichos (Scatter Plot)

Esta visualización presenta una comparación directa entre las ventas observadas y las estimadas por el modelo. Los puntos azules representan cada videojuego del conjunto de prueba, mientras que la línea roja punteada marca el escenario ideal donde las predicciones coincidirían perfectamente con la realidad.

Interpretación:

- La mayoría de los puntos se agrupa cerca de la línea ideal, lo que indica que el modelo tiene una capacidad aceptable de predicción
- Se observa una ligera dispersión en valores superiores a 60k unidades, sugiriendo que el modelo tiene mayor dificultad para predecir ventas de juegos altamente exitosos
- No se identifican patrones sistemáticos de sobreestimación o subestimación, lo cual es positivo
- **Aplicación práctica:** Los desarrolladores pueden confiar en las estimaciones del modelo para juegos con ventas esperadas entre 30k y 60k unidades

Gráfico de Residuos

Los residuos representan la diferencia entre el valor real y el predicho. Esta gráfica distribuye estos errores respecto a las predicciones realizadas.

Interpretación:

- Los residuos se distribuyen de manera relativamente aleatoria alrededor de la línea cero (roja punteada), lo que indica que no hay sesgos sistemáticos graves
- La varianza de los residuos parece relativamente constante a lo largo del rango de predicciones (homocedasticidad), cumpliendo uno de los supuestos de la regresión lineal
- La ausencia de patrones curvos o en forma de embudo sugiere que la relación lineal es apropiada
- **Aplicación práctica:** El modelo no presenta sesgos importantes que pudieran llevar a decisiones empresariales incorrectas

Distribución de Errores Absolutos (Histograma)

Esta gráfica muestra la frecuencia de errores de diferentes magnitudes, independientemente de si fueron sobreestimaciones o subestimaciones.

Interpretación:

- La mayor frecuencia de errores se concentra en el rango de 0-5k unidades, lo que indica buena precisión general
- La distribución muestra una cola hacia la derecha, con menos casos de errores grandes (15-20k)
- Aproximadamente el 70% de las predicciones tienen errores menores a 10k unidades
- **Aplicación práctica:** Los desarrolladores pueden establecer márgenes de error realistas al planificar presupuestos y estrategias de lanzamiento

Nuevas Gráficas Implementadas

D) Importancia de Variables (Gráfico de Barras Horizontal)

Esta visualización muestra los coeficientes normalizados del modelo, representando cuánto contribuye cada variable a la predicción de ventas.

Interpretación:

- **Presupuesto de desarrollo** (0.45) y **Puntuación en reseñas** (0.38) son los factores más determinantes, explicando juntos gran parte de la variabilidad en ventas

- **Marketing pagado** (0.32) y **Seguidores en redes sociales** (0.28) tienen un impacto medio pero significativo
- Variables como **Duración del desarrollo** (0.15) y **Mes de lanzamiento** (0.08) tienen influencia limitada
- **Aplicación práctica:**
    - Los estudios indie deberían priorizar conseguir financiamiento adecuado y buscar buenas reseñas críticas
    - La inversión en marketing y construcción de comunidad antes del lanzamiento tiene retorno medible
    - Extender el tiempo de desarrollo no garantiza mayores ventas por sí solo

E) Análisis de Ventas por Género (Gráfico de Barras + Pastel)

Compara el rendimiento comercial promedio según el género del videojuego, complementado con la distribución de la muestra.

Interpretación:

- Los **juegos RPG** lideran con 61k unidades promedio, aunque representan solo el 15% de la muestra (3 juegos)
- Los **juegos de Plataforma** muestran buen rendimiento (52k) y tienen mayor representación (30% de la muestra)
- Los **juegos Puzzle** tienen el menor rendimiento (35k unidades promedio)
- **Aplicación práctica:**
    - Desarrolladores deberían considerar el género como factor estratégico en la planificación
    - Los RPG indie pueden tener nichos de mercado muy rentables pero requieren mayor inversión
    - Los juegos de plataforma ofrecen un equilibrio entre riesgo y retorno
    - Para estudios pequeños, los puzzle pueden ser una entrada de bajo riesgo pero con expectativas comerciales moderadas

F) Precisión del Modelo por Rango de Ventas (Gráfico de Líneas Dual)

Evalúa cómo varía el desempeño del modelo según diferentes niveles de ventas, mostrando tanto la precisión como el error promedio.

Interpretación:

- El modelo es más preciso (85%) para juegos con ventas bajas (<30k), con errores promedio de solo 3.2k unidades
- La precisión disminuye progresivamente en rangos superiores, llegando a 72% para juegos con ventas >70k
- Existe una relación inversa clara entre el nivel de ventas y la precisión del modelo
- **Aplicación práctica:**
    - Los estudios indie que desarrollan proyectos modestos pueden confiar más en las predicciones
    - Para proyectos ambiciosos con expectativas de altas ventas, se debe considerar un mayor margen de incertidumbre
    - Esta información es crucial para la gestión de riesgos y planificación financiera

G) Análisis Radar Comparativo (Radar Chart)

Compara el rendimiento actual del modelo contra métricas objetivo en cinco dimensiones: precisión, velocidad, interpretabilidad, escalabilidad y generalización.

Interpretación:

- **Fortalezas identificadas:**
    - Velocidad de procesamiento (95%): El modelo entrena y predice rápidamente, ideal para análisis iterativos
    - Interpretabilidad (88%): Los coeficientes son claros y explicables a stakeholders no técnicos
- **Oportunidades de mejora:**
    - Precisión (79% vs 90% objetivo): Gap de 11 puntos que podría cerrarse con más datos o regularización

    - Generalización (76% vs 85% objetivo): El modelo podría beneficiarse de técnicas de validación cruzada
    - Escalabilidad (70% vs 75% objetivo): Rendimiento aceptable pero podría optimizarse
- **Aplicación práctica:**
    - El modelo actual es adecuado para entornos de producción donde se requiera rapidez y transparencia
    - Para aplicaciones críticas que requieran máxima precisión, se recomienda considerar ensemble methods
    - La arquitectura del modelo permite su uso en equipos multidisciplinarios donde no todos tienen experiencia técnica avanzada

Conclusiones

El desarrollo de este proyecto de análisis supervisado ha demostrado la viabilidad y utilidad práctica de aplicar técnicas de machine learning al sector del desarrollo de videojuegos indie. A través de la implementación de un modelo de regresión lineal complementado con siete tipos diferentes de visualizaciones, se logró crear una herramienta integral que no solo predice ventas, sino que proporciona insights accionables para la toma de decisiones estratégicas.

Los resultados obtenidos, con un coeficiente de determinación R^2 de 0.79, indican que aproximadamente el 79% de la variabilidad en las ventas puede explicarse mediante las variables consideradas, lo cual representa un rendimiento sólido para un modelo base. Las métricas MAE (8.45) y RMSE (11.23) confirman que los errores de predicción se mantienen dentro de rangos manejables para la planificación empresarial. Más importante aún, el análisis de importancia de variables reveló que factores controlables como presupuesto, calidad del producto (reflejada en puntuaciones) y estrategias de marketing tienen impacto medible y significativo en el éxito comercial.

Las visualizaciones avanzadas implementadas presupuesto, género y rango de ventas proporcionan una comprensión multidimensional del fenómeno estudiado. Estas herramientas permiten a los desarrolladores indie identificar oportunidades de mercado, evaluar riesgos y optimizar la asignación de recursos limitados. El análisis por género, por ejemplo, evidencia que los RPG indie pueden generar mayores retornos aunque requieren inversiones más sustanciales, mientras que los juegos de plataforma ofrecen un equilibrio favorable entre riesgo y recompensa.

Desde una perspectiva técnica, el proyecto también expone las limitaciones inherentes de la regresión lineal cuando se enfrenta a relaciones complejas y no lineales. La disminución de precisión en rangos de ventas superiores a 70k unidades sugiere que fenómenos como efectos de red, viralidad o timing de mercado no se capturan adecuadamente mediante relaciones lineales simples. Esto señala claramente la necesidad de explorar arquitecturas más sofisticadas para futuras iteraciones.

Finalmente, este trabajo subraya la importancia de democratizar el acceso a herramientas analíticas avanzadas en la industria del videojuego independiente. Al transformar datos históricos en conocimiento predictivo presentado de manera visual e intuitiva, se empodera a pequeños estudios para competir en igualdad de condiciones con desarrolladores establecidos. La metodología presentada es escalable, reproducible y adaptable a otros contextos de la industria creativa

digital, estableciendo un precedente valioso para la integración de data science en procesos creativos y comerciales.

Referencias

1. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://web.stanford.edu/~hastie/ElemStatLearn/
2. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer. https://www.statlearning.com/
3. **Géron, A.** (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. [Guía práctica integral sobre implementación de algoritmos de ML]
4. **VanderPlas, J.** (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/
5. **Kuhn, M., & Johnson, K.** (2013). *Applied Predictive Modeling*. Springer. [Referencia esencial sobre evaluación y selección de modelos predictivos]
6. **McKinney, W.** (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media. [Autor de pandas, fuente autorizada sobre manipulación de datos]
7. **Raschka, S., & Mirjalili, V.** (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2* (3rd ed.). Packt Publishing.
8. **Müller, A. C., & Guido, S.** (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media. [Enfoque práctico específico para scikit-learn]
9. **Bruce, P., Bruce, A., & Gedeck, P.** (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python* (2nd ed.). O'Reilly Media.
10. **Wickham, H., & Grolemund, G.** (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. https://r4ds.had.co.nz/ [Principios de visualización aplicables a cualquier lenguaje]