Proyecto de Aprendizaje Automático de Maquina

Análisis de desempeño y preferencias de videojuegos

Juan Manuel Ramirez

Para la realización del proyecto se escogieron dos bases de datos similares. De las variables utilizadas las bases contenían los mismos datos en sus columnas de nombre (del juego) y genero indicando que se en ambos casos se usaron los mismos juegos. Esto es verificado mediante una función donde se filtran aquellos juegos que eran compartidos. Debido a que se tenían los mismos juegos analizados en el 2016 se juntaron los datos de NA_players, EU_players, JP_players, Other_Players, NA_sales, EU_sales, JP_sales, Other_sales, Critic_Score, User_Score. El proyecto se realizó con dos enfoques; identificar cual es la mejor métrica de desempeño de videojuegos e intentar definir preferencias de géneros de videojuegos basándose en la cantidad de jugadores, ventas y puntajes de opinión.

Inicialmente se realizó un procesamiento de las 16719 unidades estadísticas del cual se resultó en una base con 6825 al remover datos con valores vacíos o no válidos. Se revisó la correlación de entre los datos identificando, mediante la baja correlación del puntaje de los críticos con respecto a las demás variables, pruebas iniciales de una hipótesis planteada para el desarrollo del trabajo; los puntajes de críticos son inviables como una medida de desempeño de un juego. Sin embargo, de este mismo análisis se genera una hipótesis adicional de que tampoco son los puntajes de los usuarios una métrica viable.

Con estos datos se intentó detectar preferencias mediante la predicción del género del juego basándose en las métricas de desempeño planteadas usando redes neuronales. La predicción de este modelo fue de 0.3018 para la fase de entrenamiento y 0.3048 para la fase de validación. Este resultado indica una definición correcta del modelo, pero una dificultad de alcanzar la predicción de los resultados deseados. De esto se concluyó que las preferencias de los juegos no se enfocan en un solo género en ninguno de los mercados demostrando que para todo juego que este bien realizado existe un mercado dispuesto a comprarlo y jugarlo.

Para graficar los resultados, y hacer un análisis adicional de los datos, se usó PCA sobre los datos de donde se obtuvieron 2 PCA con un total de 71.57% de la varianza explicada. El primer vector de PCA resulto en una fuerte relación con respecto a las métricas de opiniones y una ligera oposición a las métricas distintas al mercado japones. Por el lado del segundo vector es fácilmente interpretable como las métricas del mercado japones.

Posteriormente se empleó el uso de clustering para identificar la mejor forma de agrupar estos datos intentando identificar si existe alguna característica común entre

los distintos mercados considerados. Para esto se calcularon métricas para identificar la cantidad de agrupaciones optimas, siendo el método del codo inconcluyente y el método de Silhouette indicador de dos grupos se optó por usar dos grupos. El resultado de esta separación se mostró gráficamente como un limite donde se diferenciaban aquellos juegos con un desempeño en el mercado japones correlacionado con las opiniones de críticos y usuarios contra aquellos juegos que presentaban disparidades entre estos pares de métricas. Esta diferenciación se puede confirmar comparando las medias resultantes de los datos relacionados con el mercado japones. Los datos se separaron por genero sin obtener un resultado significativo por la baja cantidad de resultados en el segundo grupo. Al imprimir el resultado del segundo grupo es interesante considerar la alta presencia de juegos de Nintendo, una empresa japonesa.