TESTE ELEFLOW

Projeto: Teste para o processo seletivo

Rafael Rangel - Cientista de Dados

São Paulo

2023

Sumário

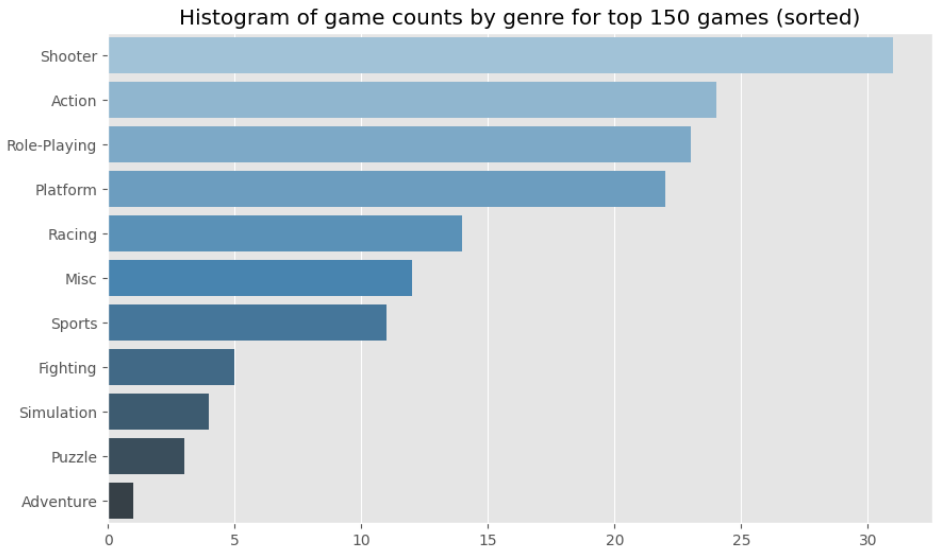
1. Introdução

O relatório apresenta uma análise do mercado de videogames e uma investigação sobre o desempenho de vendas, gêneros de jogos, e ranking das maiores empresas do setor. O relatório cobre os primeiros 150 títulos de jogos mais bem classificados, lançamentos de jogos nos últimos 10 anos e as cinco maiores empresas em termos de vendas nos Estados Unidos. Além disso, em um desafio adicional, a análise abordou também um conjunto de dados da plataforma de streaming Netflix para prever as notas de filmes e séries disponíveis na plataforma.

2. Visualização de dados

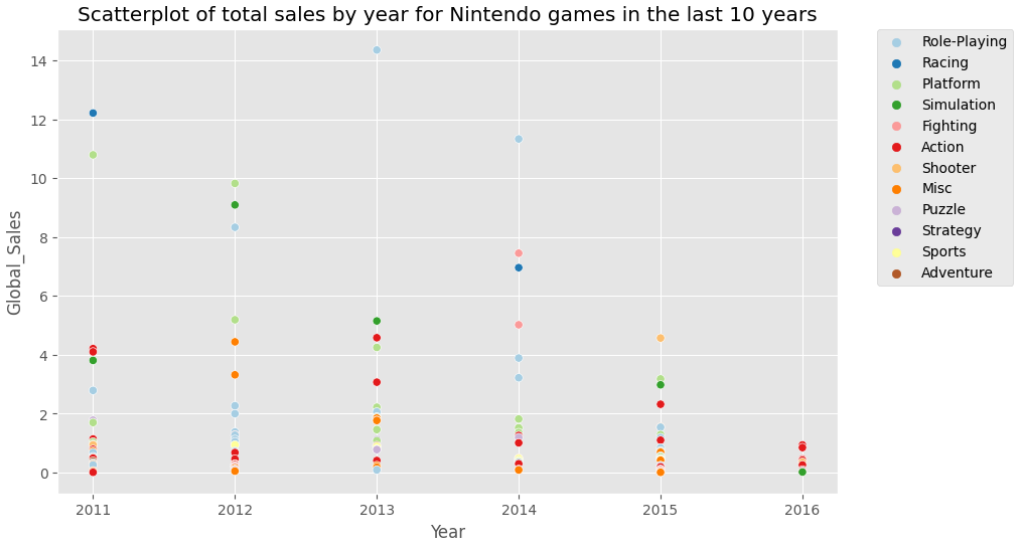
Sobre o dataset "Games", o relatório mostrou que os jogos de "shooter" são os mais predominantes, seguidos dos gêneros de ação e role-playing. A Nintendo foi a empresa que obteve destaque nas vendas na última década por lançar jogos voltados para o público familiar. As cinco maiores empresas em termos de vendas nos Estados Unidos nos últimos 10 anos são, em ordem decrescente: Nintendo, Electronic Arts (EA), Activision, Sony e Ubisoft.

2.1 Histogramas de quantos jogos cada gênero possui nos primeiros 150 títulos do rank



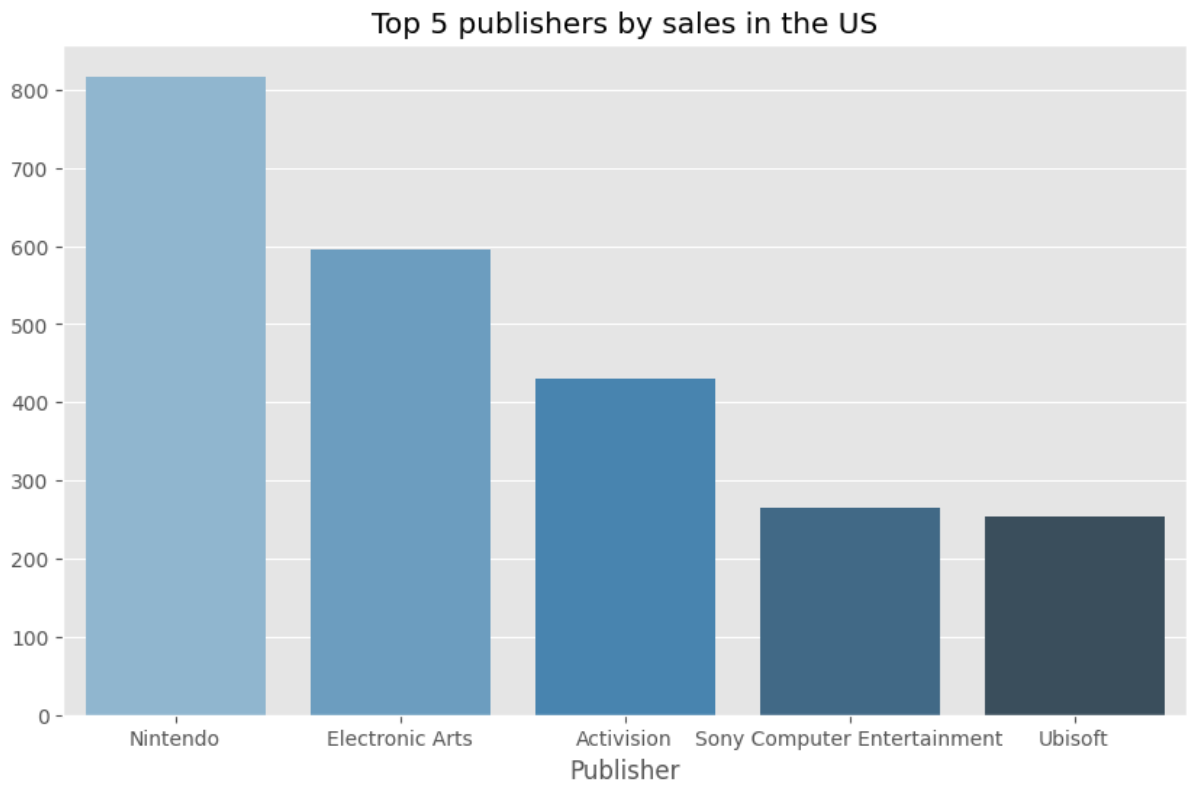
O histograma apresenta a distribuição de gêneros entre os 150 títulos de jogos mais bem classificados. Observa-se que os jogos de "shooter" predominam, seguidos pelos gêneros de ação e role-playing. Vale ressaltar que a categorização dos jogos pode não ser precisa, uma vez que alguns títulos podem se enquadrar em múltiplos gêneros, o que pode dificultar a avaliação e definição exata de suas categorias. No entanto, é evidente que os jogos de ação e shooter exercem grande influência no mercado global.

2.2 Um gráfico de dispersão entre o ano da publicação e o total de vendas da Nintendo nos últimos 10 anos



O gráfico de dispersão exibe os lançamentos de jogos da Nintendo e suas respectivas vendas globais ao longo da última década (2010 a 2020). A empresa é conhecida por sua ênfase em jogos voltados para o público familiar e de nicho, tais como 'Zelda' e 'Super Mario', e lança um menor número de títulos no mercado em comparação a outras grandes companhias do setor. Essa característica está associada à estratégia da Nintendo de focar principalmente em suas franquias consolidadas. Nota-se que picos de vendas geralmente relacionam-se ao lançamento de jogos dessas franquias populares ou de consoles inovadores, como é o caso do Wii e seus revolucionários controles de movimento.

2.3 As 5 maiores “publishers” em vendas nos Estados Unidos



As cinco maiores empresas em termos de faturamento nos Estados Unidos nos últimos dez anos (2010 a 2020) são, em ordem decrescente: Nintendo, Electronic Arts (EA), Activision, Sony e Ubisoft. A Nintendo lidera o ranking com um faturamento superior a 800 milhões de dólares, resultado de sua estratégia bem-sucedida de explorar suas franquias exclusivas e mantê-las atreladas aos seus consoles. A EA, em segundo lugar, se destaca com uma diversidade de franquias, incluindo a popular série FIFA. A Activision, por sua vez, ocupa a terceira posição com seus aclamados jogos da franquia Call of Duty. A Sony, reconhecida por títulos exclusivos como God of War, figura na quarta posição. Por fim, em quinto lugar, encontra-se a Ubisoft, responsável pela extensa franquia Assassin's Creed.

3. Machine Learning

Qual é a diferença entre o aprendizado de máquina supervisionado e não supervisionado?

Dê exemplos de algoritmos para cada um.

A aprendizagem de máquina pode ser classificada principalmente em dois tipos: supervisionado e não supervisionado.

3.1 Aprendizado Supervisionado:

No aprendizado supervisionado, os algoritmos de aprendizado de máquina são treinados usando conjuntos de dados rotulados. Isso significa que a resposta ou o resultado desejado já é conhecido. Durante o treinamento, o algoritmo procura padrões nos recursos que estão associados aos rótulos. Uma vez treinados, estes modelos são então aplicados a novos dados para prever resultados.

Exemplos de algoritmos de aprendizado supervisionado incluem:

* Regressão linear e logística
* Máquinas de vetores de suporte (SVM)
* Árvores de decisão
* Florestas aleatórias
* Redes neurais profundas (quando possuem dados rotulados)

3.2 Aprendizado Não Supervisionado:

No aprendizado não supervisionado, os algoritmos de aprendizado de máquina são treinados usando conjuntos de dados não rotulados. O algoritmo tenta identificar padrões e relações nos dados, mesmo que não saibamos quais são as respostas corretas.

Em outras palavras, ele aprende os padrões nos dados e, com base nesses padrões, ele pode agrupar ou associar novos dados.

Exemplos de algoritmos de aprendizado não supervisionado incluem:

* Clustering (por exemplo, algoritmo K-means, Clustering Aglomerativo Hierárquico)
* Análise de componentes principais (PCA)
* Regras de associação (por exemplo, Apriori, FP-growth)
* Autoencoders em redes neurais (um tipo especial de rede neural usada para representação de dados e detecção de anomalias).

Após a execução de dois algoritmos diferentes de machine learning, são obtidas as seguintes matrizes de confusão:

Modelo 1:

	Previsão Negativa	Previsão Positiva
Negativo	5740	519
Positivo	1119	9413

Modelo 2:

	Previsão Negativa	Previsão Positiva
Negativo	6751	705
Positivo	2005	7330

Analisando apenas essas matrizes, qual modelo você consideraria o melhor para ser utilizado? Justifique sua resposta.

Ao analisar as matrizes de confusão, é possível extrair métricas como precisão, recall e acurácia. Tendo em conta que não existe mais contexto sobre a importância relativa ao classificar corretamente casos positivos e negativos, vamos nos basear nessas métricas para decidir qual modelo é o melhor.

Calculando a precisão, recall e acurácia para ambos os modelos:

* Modelo 1: Precisão = 0.948, Recall = 0.894, Acurácia = 0.902
* Modelo 2: Precisão = 0.912, Recall = 0.785, Acurácia = 0.839

Considerando essas métricas, o Modelo 1 apresenta maior precisão (menos chances de falsos positivos), recall (menos chances de falsos negativos) e acurácia geral. Portanto, caso estas métricas sejam importantes para nosso problema, o Modelo 1 seria o preferido. No entanto, é importante ressaltar que esta conclusão depende da importância que damos a cada tipo de erro em nosso problema específico.

4. Desafio

Para o dataset "Netflix", abordou-se a aprendizagem de máquina, com diferenças entre o aprendizado supervisionado e não supervisionado, explanando exemplos de algoritmos para cada um. Os resultados das análises avançaram para o desafio de previsão das notas de filmes usando base de dados e abordagens de machine learning. Conclui-se que a disponibilidade dos dados, a maneira como eles são tratados e utilizados para a construção das características têm grande influência na criação do modelo de previsão, necessitando de aprimoramentos contínuos para aumentar a assertividade dos resultados.

Um maior detalhamento do processo encontra-se no Jupyter notebook.

TESTE ELEFLOW.ipynb

Referências

- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
  Website: https://scikit-learn.org/stable/about.html#citing-scikit-learn

- PyCaret: An open-source Python library for Machine Learning
  Website: https://pycaret.org/citation/

- Seaborn: statistical data visualization, Michael Waskom.
  Website: https://seaborn.pydata.org/citing.html

- Travis E, Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).
  Website: https://numpy.org/citing-numpy/

- Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
  Website: https://pandas.pydata.org/about/citing.html

- Wikipedia Contributors. "Confusion matrix." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia,
  Website: https://en.wikipedia.org/wiki/Confusion_matrix

- Wikipedia Contributors. "F1 score." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia,
  Website: https://en.wikipedia.org/wiki/F1_score