Relatório 5 - Estatística p/ Aprendizado de Máquina

Edryck Freitas Nascimento

1. Introdução

O objetivo desta aula foi aprender sobre Estatística e Probabilidade e Prática de Python. A tarefa consistia em ler assistir as aulas da seção "*Statistics and Probability Refresher, and Python Practice*" do curso "*Machine Learning, Data Science and Deep Learning with Python*" e fazer um relatório com os principais conhecimentos adquiridos.

2. Desenvolvimento

Para a fixação do conteúdo adquirido no curso, tentei fazer uma análise de dados históricos de vendas de videogames, obtive um arquivo *.csv* onde contém o conjunto de dados de vendas, obtido no Kaggle[1].

O objetivo consistia em entender o comportamento das vendas globais e analisar a probabilidade de sucesso, na qual chamei de "Hit", de títulos, o foco foi na publicadora Nintendo. A análise foi realizada na linguagem Python utilizando as bibliotecas Pandas para a manipulação de dados e Matplotlib e Seaborn para visualização.

Inicialmente comparei a média com a mediana, pelo motivo de que normalmente em dados de vendas, a média geralmente é maior de que a mediana devido a *outliers* (são jogos que foram bastante vendido), observei uma discrepância entre a média e a mediana das vendas globais, isso indicava uma distribuição assimétrica. A maioria dos jogos vendem quantidades razoáveis, enquanto poucos "outliers" puxaram a média para cima.

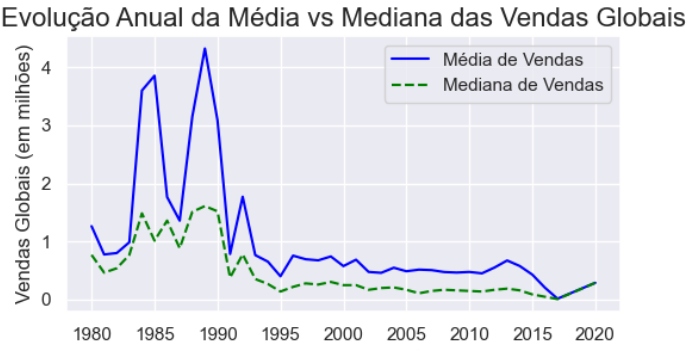

Gráfico 1. Evolução Anual de vendas globais. (autoria própria)

[1] https://www.kaggle.com/code/ahmedkhaledhemida/video-game-sales-dataset/input

Apliquei o Teorema de Bayes para responder à pergunta: Qual a probabilidade de um jogo ser da Nintendo, dado que ele é um sucesso de vendas?

Então calculei os seguintes eventos utilizando a fórmula:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

Onde:

P(A): é a chance de um jogo escolhido aleatoriamente no conjunto de dados ser da Nintendo.

P(B): é a chance geral de qualquer jogo no mercado ser considerado um sucesso.

P(B|A): é a probabilidade condicional, dado que um jogo é da Nintendo, qual a chance de ele ser um Hit (sucesso)? Basicamente esse valor reflete a taxa de acerto da Nintendo.

Após aplicar o Teorema de Bayes para analisar a dominância de mercado da Nintendo na amostra de jogos de alto desempenho, tivemos o resultado do cálculo probabilístico que confirmou pela prova real nos dados filtrados, ele resultou em uma probabilidade condicional P(A|B) ≈ 0,596.

Esse resultado de 59,6% é bem significativo, ele indica que, dado que um jogo atinge o patamar de mais de 10 milhões de cópias vendidas, há uma chance próxima de 60% de que ela seja da Nintendo, considerando que a Nintendo representa algo próximo de 4% de todos os jogos lançados no banco de dados, ela representar quase 60% dos jogos de mais sucesso demonstra uma eficiência excelente.

3. Conclusão

Após essa aplicação prática dos conceitos aprendidos no curso, mostrou que o sucesso no mercado de videogames não é uniformemente distribuído, através do Teorema de Bayes, foi possível quantificar a influência da publicadora no sucesso de um título, o que transforma as intuições de mercado em dados probabilísticos concretos.

Referências:

Origem do conjunto de dados utilizados:

Khaled, A. “VIDEO GAME SALES DATASET”, https://www.kaggle.com/code/ahmedkhaledhemida/video-game-sales-dataset/input

Material utilizado durante o curso:

Kane, F. “Course Materials: Machine Learning, Data Science, and Generative AI with Python”, https://www.sundog-education.com/machine-learning/