Teste de Relação Entre Taxa de Aprovação dos Jogos da Steam com seu Preço e Gênero

Israel de Melo Pedrosa
2017014790

1 Resumo

Steam é a maior plataforma de comércio de jogos no mundo. Ela oferece um catálogo com mais de 60.000 jogos além de servir como rede social e possuir features como a avaliação dos jogos, permitindo que os usuários avaliem positiva ou negativamente, e que escrevam uma critica sobre o jogo, além disso a plataforma também separa os jogos em diversos marcadores (ou gêneros) facilitando a busca por um jogo que atenda o gosto do usuário.

2 Introdução

O objetivo deste projeto é observar como o preço e o gênero de cada jogo pode nos ajudar a entender o quão bem esse título é avaliado. A motivação para isso está no fato de que determinados estilos de jogos são mais apreciados do que outros enquanto alguns gêneros só fazem sucesso dentro de determinados nichos. Para tentar entender esse relacionamento foram coletados dados de mais de 1.000 jogos, dentre os dados estão o título, o preço, as avaliações positivas, as avaliações totais, e os três principais marcadores para cada jogo.

	Titulo	Preco	ReviewTotal	ReviewPositivo	Tags	index	Mágica	Fantasia	Mundo Aberto	Terror	...
0	Hogwarts Legacy	R$ 249,99	207740	191016	('Mágica', 'Fantasia', 'Mundo Aberto')	0	1	1	1	0	...
1	Dead by Daylight	R$ 49,99	671948	537518	('Terror', 'Terror de Sobrevivência', 'Multijo...	100	0	0	0	1	...
2	Baldur's Gate 3	R$ 90,99	61960	54582	('Acesso Antecipado', 'RPG', 'Você Decide')	200	0	0	0	0	...
3	Marvel's Spider-Man Remastered	R$ 249,90	67934	65592	('Super-Heróis', 'Mundo Aberto', 'Ação')	300	0	0	1	0	...
4	Sea of Thieves 2023 Edition	R$ 29,00	286085	257989	('Aventura', 'Multijogador', 'Mundo Aberto')	400	0	0	1	0	...
5	Tom Clancy's Rainbow Six® Siege	R$ 59,99	1148734	990987	('Tiro em Primeira Pessoa (FPS)', '', 'eSports')	500	0	0	0	0	...

3 Metodologia

3.1 Coleta de dados

Utilizando a biblioteca Selenium em um python notebook, parti da página de buscas da steam, limitando a busca apenas a jogos para evitar a coleta de conteúdos adicionais dos jogos como DLCs e outras ferramentas que também estão à venda na plataforma.
Da página de buscas eu coletei os links de aproximadamente mil e quinhentos jogos, e utilizando esses links o próximo passo era acessar a página do jogo e retornar o título, o preço, as avaliações positivas e totais e os três marcadores principais daquele jogo.

3.2 Limpeza e utilização dos dados

Por algum erro da plataforma no período da coleta dos dados alguns jogos estavam sem informação de preço, dessa forma decidi remover da base de dados todos os jogos que possuíam informações faltando, e para facilitar a utilização dos dados para a provável regressão linear múltipla que deveria ser feita para entender o relacionamento observado transformei a coluna da base de dados que possuía uma tupla com os três gêneros principais dos jogos em cem colunas binárias que representavam os marcadores que o jogo poderia ter, sendo assim se um jogo possuísse os marcadores para ação, aventura e exploração por exemplo, as colunas referentes as esses gêneros seriam preenchidas com o valor 1 e as outras colunas ficariam preenchidas com o valor 0.

```
X2 = (X2-X2.mean())/X2.std()
y = (y-y.mean())/y.std()
```

```
def derivada_reg(theta, X, y):
    return -2 * ((y - (X @ theta)) @ X)
```

```
def grad(theta, X, y, l=0.0001, tolerancia=0.00001):
    theta = theta.copy()
    erro = 1

    while True:
        grad = derivada_reg(theta, X, y)

        theta_novo = theta - l * grad

        erro_novo = ((y - X.dot(theta)) ** 2).mean()
        if np.abs(erro - erro_novo) <= tolerancia:
          break

        theta = theta_novo
        erro = erro_novo

    return theta
```

4 Análise dos dados

4.1 Regressão linear múltipla

Com os dados preparados comecei a efetuar alguns dos métodos vistos em sala de aula para deixá-los prontos para a regressão linear múltipla, primeiro peguei os valores de avaliações positivas e avaliações totais e calculei a taxa de avaliações positivas para cada jogo armazenando-as em um vetor e em seguida separei a coluna de preço e as colunas com os marcadores dos jogos em uma matriz X. para observar melhor

a relação entre os dados z-normalizei eles previamente para que uma possível diferença de escala não interfira na análise a ser realizada, adicionei em X a coluna do intercepto, e então gerei um vetor theta aleatório. Com isso é possível aplicar o método do gradiente descendente para retornar os valores de theta que minimizam a função de perda da soma dos erros quadrados. Com theta calculado é possível gerar um modelo para tentar explicar o relacionamento dos dados, mas o mais importante é conseguir aferir o quão bem esse mesmo modelo exerce sua função.

4.2 Avaliando a qualidade do modelo

Para avaliar a qualidade do modelo calculei o valor do coeficiente de determinação (ou R^2), que indica o quão melhor o modelo representa a variável observada (taxa de avaliações positivas) do que o modelo que gera os erros apenas utilizando a média. O valor do coeficiente de determinação calculado foi de 0.2132…, o que me leva ao fato de que o modelo gerado pela regressão é minimamente melhor do que utilizar a média para prever o valor da taxa de avaliações positivas, isso nos mostra que muitos outros fatores interferem em como o jogo será avaliado pelos usuários da plataforma.

```
X2 = np.nan_to_num(X2)
```

```
def erro(theta, X, y):
    return y - X@theta
```

```
def R_sq(theta, X, y):
    sse = sum(erro(theta, X, y)**2)
    sst = sum((y - np.mean(y))**2)
    return 1.0 - (sse / sst)
```

```
R_sq(theta, X, y)
```

```
0.21320185229392707
```

```
Taxa normalizada / Taxa prevista

0.6720348431028998 / 0.47537997901506535
-0.42586604723664784 / -0.2731521447082959
0.317818081266342 / -0.06924105199165018
1.0947350933236628 / -0.0975879572404988
0.5094549479583927 / -0.2587819269327182
0.15026497038103592 / 0.7291552163339107
0.9790964866905568 / -0.8984001688498174
0.4865506973927067 / -0.5718679830524707
0.2157336300026823 / -0.6893937182962586
0.2454347434284187 / -0.8241048292176743
1.2481286522272925 / 0.2352697914112167
0.8708017727249352 / -0.9002375685279314
1.15571286578358 / 0.5535108865596842
0.8343770378590376 / -0.2887144231633398
0.7892068140312677 / -0.22207036968935265
0.8319674279113494 / -0.7360394141371174
```

5 Conclusão

Com as análises e os dados apresentados nesse artigo é possível concluir que o relacionamento esperado inicialmente entre o preço e os principais marcadores de um determinado jogo, e o quão boa é sua taxa de avaliações positivas, é mais fraco do que o previsto, muitos outros fatores como orçamento, tempo necessário para zerar e até mesmo o marketing podem interferir em como os usuários avaliam o jogo.