Classificação de fatores de sucesso para novas produções cinematográficas

André Camacam[1], Cleverson Santos[2], Lucas Cyulik[3], Rafael Barreto[4]

[1,2,3,4]Faculdade de Administração e Economia (FAE) Business School, Brasil

Resumo – Cada vez mais as produções cinematográficas utilizam recursos homéricos para a gravação de novos filmes, seja ele pela necessidade de atores mais rentáveis ou pela criação de efeitos especiais. Por isso, é de extrema relevância aos investidores elementos e dados capazes de indicar uma maior probabilidade de acerto face à liberação de verbas requisitadas. Este artigo apresenta um estudo sobre a qualidade e retorno de investimento das produções cinematográficas antigas baseadas nos gêneros cinematográficos, atores, diretores e outros fatores.
Esses dados foram tirados da fonte IMDb, conhecido por ser um centro mundial de informações, críticas e usuários, e posteriormente foi-se aplicado métodos de processamento e algoritmos para formação de uma base para apresentar projeções baseada nos dados necessários para a tomada de decisão.

Palavras-chave: Crítica Cinematográfica; Atores; Investimento; Produção Cinematográfica; Retorno de Investimento

Abstract – More and more film productions use homeric resources for the recording of new films, be it by the need for more profitable actors or by the creation of special effects. Therefore, the investors need to be more certain when releasing the requested funds. This article presents a study on the quality and return on investment of cinematographic productions based on old cinematographic genres, actors, directors and other factors.
These data were taken from the IMDb source, known as a worldwide center of information, criticism and users, and afterwards we applied processing methods and algorithms to form a basis for presenting projections based on the data needed for decision making.

Keywords: Film Criticism; Actors; Investment; Film Production; Return Of Investiment

Introdução

O cinema é considerado a sétima arte e não é à toa. Os filmes são uma das formas mais populares de entretenimento em todo o mundo. Todos os anos são gastos milhões de dólares em produções das quais se esperam retornos enormes de investimento [1].

Visto isso, vislumbrou-se o uso de métodos de classificação por meio do método preditivo, em uma base com informações sobre filmes para encontrar fatores que são capazes de influenciar no sucesso de produções cinematográficas. Com efeito, buscou-se utilizar um algoritmo de árvore de decisão chamado C4.5.

A base de dados para processamento foi obtida na fonte do IMDb, (www.imdb.com), conhecido por ser um centro mundial de informações, de críticas e de usuários sobre filmes e séries.

Dados

O conjunto de dados utilizado para o desenvolvimento do presente artigo consiste em informações e avaliações de 5000 (cinco mil) produções cinematográficas provenientes do site de avaliações de filmes e séries IMDb.

As informações existentes de cada filme são: Título do filme; Cor; Linguagem; País de origem; Ano; Gênero; Duração; Proporção de imagem; Nome do diretor; Receita bruta; Orçamento; Censura; Classificação dos conteúdos; Quantidade de criticas; Nota no site IMDb; Conjunto de palavras-chave; Quantidade críticas feitas por usuários; Quantidade de votos dos usuários do site IMDb; Quantidade de "*likes*" no Facebook; Quantidade de "*likes*" do diretor no Facebook; Nome do ator um; Quantidade de "*likes*" do ator um no Facebook; Nome do ator dois; Quantidade de "*likes*" do ator dois no Facebook; Nome do ator três; Quantidade de "*likes*" do ator três no Facebook; Quantidade total de "*likes*" do elenco no Facebook; Quantidade de rostos presentes no pôster; e Endereço do filme na página do IMDb.

O conjunto de dados foi adquirido no website *Kaggle* [2], no formato .csv (*comma-separated values*).

Pré-processamento dos dados

O pré-processamento dos dados foi realizado utilizando-se a ferramenta Excel, distribuída no pacote Office da Microsoft.

A primeira ação tomada foi a identificação e tabulação dos dados nas respectivas colunas. Com os dados distribuídos de maneira correta foi realizada a identificação e eliminação de caracteres que prejudicariam a análise e impediriam o correto funcionamento da ferramenta de Data Mining.

Foram, então, eliminadas as colunas de dados que não apresentavam dados com grande significância para uma análise de investimento, que só havia para filmes mais novos e, também, foi feito a discretização de dados numéricos, que poderiam ser agrupados em clusters.

Foram eliminadas as seguintes colunas: Título do filme; Cor; Proporção de imagem; Censura; Classificação dos conteúdos; Quantidade de criticas; Nota no site IMDb; Conjunto de palavras-chave; Quantidade críticas feitas por usuários; Quantidade de votos dos usuários do site IMDb; Quantidade de "*likes*" no Facebook; Quantidade de "*likes*" do diretor no Facebook; Nome do ator um; Quantidade de "*likes*" do ator um no Facebook; Nome do ator dois; Quantidade de "*likes*" do ator dois no Facebook; Nome do ator três; Quantidade de "*likes*" do ator três no Facebook; Quantidade total de "*likes*" do elenco no Facebook; Quantidade de rostos presentes no pôster; e Endereço do filme na página do IMDb.

Na discretização dos dados, os anos de lançamento dos filmes foram agrupados por década.

Ao final do pré-processamento restaram 3758 (três mil setecentos e cinquenta e oito) filmes, dos 5000 (cinco mil) que haviam orginalmente.

Processamento dos Dados

Para o processamento da base de dados foi utilizado o *Weka* [3], um pacote de software *opensource* para mineração de dados, criado pela Universidade de Waikato – Nova Zelândia. A versão do software utilizada para realização deste artigo foi o *Weka* 3.8.1.

Para classificação foi escolhido um método preditivo de árvore de decisão chamado C4.5 [4] ou J48 (nomenclatura utilizada pela ferramenta *Weka*).

Após realizar a carga dos dados pelo *Weka* e ter selecionado o J48, iniciou-se o processo através de algumas interações de teste e o valor de 0.6 para o grau de confiança foi encontrado.

O atributo alvo para a geração da árvore de decisão foi o score do IMDb. Como está apresentado na Figura 1, foram identificadas duas classes para classificação dos filmes, uma que representava os filmes de sucesso (BOM), e outra que, por sua vez, representava os filmes que não foram bem-sucedidos (RUIM).

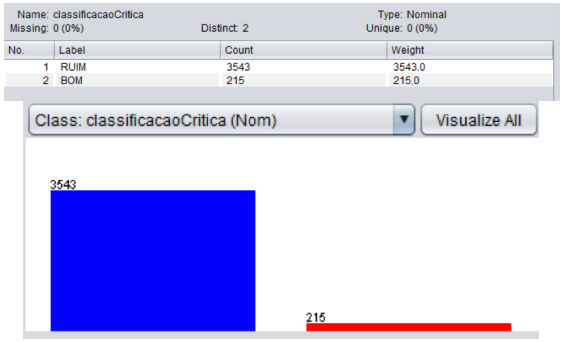

Figura 1 – *Classes (Fonte: Criada pelo autor)*

O processamento gerou uma árvore de tamanho 3523 (três mil quinhentos e vinte três) com 3501 folhas (três mil quinhentos e um). Tem-se como exemplo do processamento um trecho da árvore:

```
budget <= 117000000
|   Decada = D90: RUIM (694.0)
|   Decada = D80: RUIM (209.0)
|   Decada = D70: RUIM (47.0)
|   Decada = D60: RUIM (22.0)
|   Decada = D50: RUIM (4.0)
|   Decada = D40: RUIM (4.0)
|   Decada = 2K
|   |   director_name = Edward Burns: RUIM (1.0)
|   |   director_name = Robert Rodriguez: RUIM (9.0)
|   |   director_name = Garry Marshall: RUIM (6.0)
|   |   director_name = Steven Soderbergh
|   |   |   genres = Comedy: RUIM (3.0)
|   |   |   genres = Action: RUIM (1.0)
|   |   |   genres = Drama: BOM (1.0)
|   |   |   genres = Adventure: RUIM (0.0)
|   |   |   genres = Crime
|   |   |   |   budget <= 39000000: BOM (2.0)
|   |   |   |   budget > 39000000: RUIM (4.0)
|   |   |   genres = Horror: RUIM (0.0)
```

A árvore é composta de cinco elementos, orçamento (budget), década (Decada), nome do diretor (diretor_name), gênero (genres) e receita bruta (gross).

Conclusão

Com base nas análises gráficas, as produções cinematográficas estão com um custo cada vez mais elevado. Deste modo, com base na tendência obtida, foi possível perceber uma elevação constante destes valores, além das seguintes informações:

Por um lado, pode-se apresentar como resultados claros, que, nos casos em que os budgets sejam maiores do que 117 milhões de dólares, a quantidade de variáveis envolvidas é bem menor.

Por outro lado, com relação às bilheterias, que obtiveram valores inferiores a 65%, independentemente do budget, o filme tende ao fracasso. Se a bilheteria for maior, depende do diretor do filme.

Além disso, foi possível concluir, com base no processamento realizado, que não se pode restringir a uma melhor opção, até porque haveria o risco de engessamento da indústria cinematográfica. Pode-se, assim, apresentar uma lista de opções, conforme a árvore de decisão apresentada para que se possa chegar a uma decisão o interesse em produzir determinado filme.

Referências

[1] Simonton, D. K. (2005). **Cinematic creativity and production budgets: Does money make the movie?**. The Journal of Creative Behavior, 39(1), 1-15.

[2] **IMDB 5000 Movie Dataset**. Acessado em 25 de julho de 2017. Disponível em: https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

[3] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). **The WEKA data mining software: an update.** ACM SIGKDD explorations newsletter, 11(1), 10-18.

[4] Quinlan, J. R. (2014). **C4.5: programs for machine learning**. Elsevier.