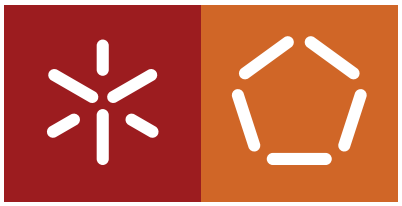
2016/2017



Universidade do Minho

Mestrado Integrado em Engenharia Informática

Unidade Curricular de Análise de Dados

IMDB 5000 Movie Dataset
Pré Relatório

Gil Gonçaslves - A67738
Luis Paulo Ferreira Pedro - A70415
José Pedro Santos Monteiro - A73014
Bruno Manuel Gonçalves Ribeiro - A73269

12 de Abril de 2017

Capítulo 1

Introdução

1.1 Contextualização

Com este trabalho pretendemos implementar de um sistema de aprendizagem automática, seguindo a metodologia CRISP-DM. A metodologia CRISP-DM tem com primeira fase o estudo do negócio (do Inglês Business understanding), que tem como função entender o que o cliente pretende realizar numa perspetiva de negócio. Na segunda fase está presente o estudo dos dados (do Inglês Data understanding), onde é realizada a aquisição e familiarização dos dados . Numa terceira etapa, é executada a preparação dos dados (do Inglês Data preparation), que é referente à decisão sobre o volume e o tipo de dados a usar no processo de análise, seguindo-se de modelação, avaliação e implementação. As seis fases foram pensadas para que pudessem ser aplicadas em qualquer área de negócio [CRISP-DM, 1999]. Neste primeiro pré relatório vamos nos focar nas primeiras 3 fases. O Internet Movie Database (também conhecido pelo acrónimo IMDb) é uma base de dados online de informação sobre música, cinema, filmes,programas e comerciais para televisão e jogos de computador. O nosso caso de estudo será baseado nos filmes e o *dataset* poderá ser encontrado em `https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset`. Este *dataset* contém 28 atributos descrevendo cerca de 5000 filmes.

1.2 Motivação e Objectivos

Temos como principal objetivo explorar o *dataset* que nos foi fornecido que contém informação de 5000 filmes, pretendemos então encontrar padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados. O processo consiste basicamente em 3 etapas: exploração, construção de modelo ou definição do padrão e validação/verificação.

Capítulo 2

Estudo do Negócio

A ideia principal é quantificar a trajetória e a experiência das pessoas por trás dos filmes e, em seguida, aplicar a métrica resultante para estimar o sucesso dos filmes que estão em fase de pós-produção. Assim sendo, precisamos de definir uma métrica para medir o sucesso de um filme. Em particular, estamos em primeiro lugar a considerar duas áreas principais:

- Prever se o filme vai ser bem recebido pela critica (imdb raking).
- Prever se o filme vai ser um sucesso a nível monetário.

A primeira foca-se no sucesso de qualidade, enquanto que a segunda se foca no sucesso a nível de bilheteira. Para ambas vamos utilizar o modelo de **Classification**, com o objetivo de prever com base no *dataset* se o filme será um sucesso ou não.

Durante o projeto tencionamos também responder a outras questões, nomeadamente:

- Analisar a tendência ao longo dos anos de medida a prever quais as tendências futuras.
- Conforme o género do filme, que ator e realizador escolher para ter sucesso.
- Encontrar filmes com impacto semelhante.

Na primeira questão pretendemos aplicar o modelo de **Clustering/Association Rules** de modo a agrupar as tendências e tentar perceber, ou até antever, qual será o passo a tomar pela pós produção. Quanto ao segundo ponto, temos como objetivo perceber a importância dos vários intervenientes para o sucesso do filme, e para tal pretendemos utilizar o modelo de **Association Rules** com o intuito de deduzir que resultado sucede com combinação de vários fatores. Por fim, pretendemos encontrar e agrupar filmes que tenham tido um impacto semelhante nas bilheteiras, e para tal iremos utilizar o modelo de **Clustering**.

Capítulo 3

Estudo dos Dados

Tendo como base as questões anteriormente explicitadas, é necessário então proceder à análise dos dados que temos a nosso dispor, a fim de posteriormente estarem preparados para lhes serem aplicados técnicas de *data mining*.

Os dados que irão ser utilizados foram retirados do site Internet Movie Database (IMDB) e encontram-se disponíveis no Kaggle. Este *dataset* é constituído por cerca de 5000 filmes e contém 28 atributos que descrevem cada filme em questão, desde o seu titulo até aos atores presentes neste. De seguida, é apresentada uma tabela que descreve cada um dos atributos presentes no *dataset*.

Atributo	Descrição	Tipo	Dominio
color	Descreve se o filme é a cores ou a preto e branco	String	Color or Black and White
director_name	Identifica quem foi o realizador do filme	String	Ex: Christopher Nolan
num_critic_for_reviews	Número de análises feitas pelos críticos	Int	1-813
duration	Duração total do filme	Int	7-511
director_facebook_likes	Quantifica o número de likes no facebook do realizador	Int	0-23000
actor_3_facebook_likes	Quantifica o número de likes no facebook do terceiro ator (grau de importância no filme).	Int	0-23000
actor_2_name	Representa o nome do segundo ator do filme (grau de importância no filme).	String	Ex:Daniel Radcliffe
actor_1_facebook_likes	Quantifica o número de likes do ator principal.	Int	0-640000
gross	Quantifica o total de dinheiro adquirido com o filme	Int	162-760505847
genres	Lista os gêneros presentes naquele filme	String	Ex: Action $\|\|$ Adventure $\|\|Thriller$
actor_1_name	Representa o nome do ator principal	String	Ex: Johnny Depp

Atributo	Descrição	Tipo	Dominio
movie_title	Identifica o Nome do filme	String	Ex:Avatar
num_voted_users	Número de votos efetuados na plataforma, sendo esses votos que levam à classificação do filme.	Int	5-1689764
cast_total_facebook_likes	Número total de gostos no facebook do elenco	Int	0-656730
actor_3_name	Nome do terceiro ator (grau de importância no filme).	String	Ex:Scarlett Johansson
facenumber_in_poster	Número de caras presentes no poster do filme	Int	0-43
plot_keywords	Resumo do filme por palavras chaves.	String	Ex:avatar$\|\|future\|\|marine\|\|native\|\|paraplegic$.
movie_imdb_link	Link que redireciona para a pagina do filme no IMDB	String	EEx: http://www.imdb.com /title/tt049954 9/?ref_=fn_tt_tt_1
num_user_for_reviews	Número de análises feitas pelos utilizadores	Int	1-4667
language	Linguagem do filme.	String	Ex: English
country	País onde pertence o filme.	String	Ex: USA
content_rating	Indica qual faixa etária é adequada para exibir a referida mídia.	String	Ex: Approved
budget	Dinheiro despendido na realização do filme	Int	218-12215500000
title_year	Ano de lançamento do filme	Int	1916-2016
actor_2_facebook_likes	Quantifica o número de likes no facebook do terceiro ator (grau de importância no filme).	Int	0-137000
imdb_score	Quantifica de 1 a 10 a qualidade do filme, este ranking é atribuido pelos votos dos utilizadores	Float	1.2-9.5
aspect_ratio	A relação de aspecto de uma forma geométrica é a proporção de seus tamanhos em diferentes dimensões	Float	1.2 - 5.4
movie_facebook_likes	Número de likes que o filme tem na sua página de facebook	Int	0-349000

Tabela 3.1: Descrição dos atributos

Através da descrição dos vários atributos acima explicitados podemos verificar que este *dataset* possui um vasto conjunto de informações referentes aos filmes, que permitem assim tirar as conclusões inerentes a este projeto. Como tal, apesar de todos estes atributos estarem a descrever os vários filmes, alguns deles não irão ser de facto necessários para o desenrolar do projeto.

Através da exploração dos dados do *dataset*, podemos verificar que vários registos possuem valores nulos em alguns dos seus atributos, como nos atributos `country` e `budget`. Este caso é o mais complicado pois o `budget` representa um valor de despesa na moeda de um dado país, pelo que se o atributo `country` for nulo não é possível representar este valor numa dada moeda comum. Outro dos problemas encontrados prende-se ao atributo `genres` que possui uma lista de géneros em que o filme se enquadra, sendo preciso tratar deste atributo de a ser usado nos modelos.

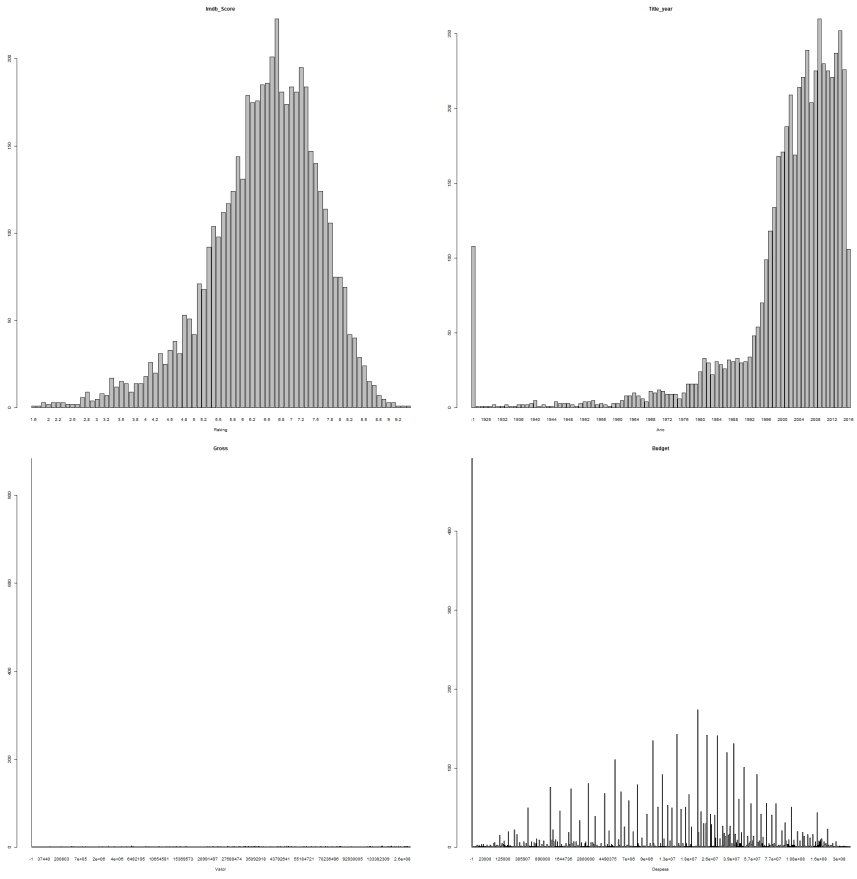


Figura 3.1: : Distribuição dos dados

Na figura a cima está presente a distribuição de valores dos atributos *imdb_score*,*title_year*,*gross* e *budget*. Como são valores numéricos positivos utilizou-se o valor **-1** para representar os valores desconhecidos.

Capítulo 4

Preparação dos Dados

Neste etapa do projeto irá ser feita uma preparação dos dados recolhidos de forma a podermos utilizar estas nos nossos modelos de *data mining*. Sendo a qualidade dos dados um fator relevante para a construção destes modelos, é ao longo deste capítulo que vamos demonstrar de que forma é que esta foi alcançada.

4.1 Seleção dos Dados

Em relação à seleção dos dados, nem todos os filmes podem ser selecionados para os nossos modelos por causa de possuírem atributos com valor nulo. Assim, apenas os registos que puderem ser tratados em casos de anomalias irão ser utilizados.

Com base nas questões inerentes a este projeto, verifica-se que existem atributos que não serão necessários para as análises que queremos efetuar, e como tal devem ser descartados os seguintes:

Atributo
Color
Duration
Facenumber_In_Poster
Plot_Keywords
Movie_Imdb_Link
Content_Rating
Aspect_Ratio

Tabela 4.1: Lista de atributos eliminados.

4.2 Limpeza dos Dados

Os principais problemas encontrados em relação aos atributos foi a existência de campos com valores nulos. Se estivermos a utilizar campos como o `budget` ou o `gross` e se estes tiverem valores nulos, não é possível utilizar esses registos nos nossos modelos, pois fazer uma estimativa para estes campos introduziria erro nestes. No entanto, se for possível estimar valores para alguns campos sem afetar a correção do modelo, esses registos poderão ser usados. De seguida é apresentada a lista de atributos que possuem campos de valor nulo:

Atributo
Actor_1_Facebook_Likes
Actor_1_Name
Actor_2_Facebook_Likes
Actor_2_Name
Actor_3_Facebook_Likes
Actor_3_Name
Budget
Nome_Director
Country
Director_Facebook_Likes
Gross
Language
Num_Criti_For_Reviews
Num_User_For_Reviews
Title_Year

Tabela 4.2: Lista de atributos com campos nulos.

Outro caso que se deve ter em conta é a unidade monetária dos campos `gross` e `budget`, na qual o valor destes campos está expresso na unidade monetária do pais correspondente ao filme. Assim, como os Estados Unidos é o pais mais frequente, iremos proceder à conversão da moeda para dólares. No entanto, é preciso ter em conta que por vezes o campo `country` é nulo, pelo que é necessário aquando da criação dos modelos verificar se a utilização destes valores mesmo sem a conversão da moeda afetam negativamente o modelo, e em caso afirmativo descarta-se o registo.

4.3 Construção de Novos Dados

O atributo `genres` contém a informação dos vários géneros em que está inserido o filme separados por um delimitador. De forma a podermos usar os vários géneros de cada filme nos nossos modelos, optámos por separar este atributo por vários atributos género.