Correlação entre o Retorno do Investimento em um Filme e a Avaliação de Críticos e do Público

Cleiton Pereira da Silva[1]* – Bacharel em Marketing pela Universidade Anhembi Morumbi, Pós graduando em Ciência de dados e Analise pela USP-Esalq - São Paulo – SP;
Leila Rabello de Oliveira[2] - Doutorado em Ciências Sociais pela PUC-SP e Mestrado em Ciência da Informação pela PUC-Campinas. Rua Dr. Alvaro Alvim, 90 – Vila Mariana – 04018-010 – São Paulo, SP, Brasil

*autor correspondente: cleitonps.mkt@hotmail.com

Correlação entre o Retorno do Investimento em um Filme e a Avaliação de Críticos e do Público

Resumo

No estudo avalia-se a correlação entre o retorno percentual do investimento financeiro em um filme e as outras variáveis envolvidas focando em analisar a correlação com as notas atribuídas pelos críticos especializados e pelo público em geral. A avaliação técnica de um filme é correlacionada com diversos fatores, como elenco, direção, roteiro, gênero, orçamento e duração, entre outros. Com o avanço das tecnologias de coleta e análise de dados, é possível explorar as relações entre essas variáveis e verificar se existem ou não padrões nestes dados e as notas atribuídas pelos críticos e o público a fim de traçar metas que ajudem os filmes a conseguir um retorno percentual do investimento financeiro. Aqui utiliza-se técnicas de data analytics e machine learning, que permitirão examinar um conjunto de dados abrangente de uma ampla lista de filmes, contendo informações como retorno percentual do investimento a partir do orçamento e da arrecadação de bilheteria, elenco, diretor, gênero, duração, juntamente com as notas atribuídas pelos críticos e pelo público. Faz-se uso de métodos estatísticos e algoritmos de regressão para identificar quais variáveis têm maior influência na determinação de um retorno satisfatório do investimento total. Este estudo visa contribuir para uma melhor compreensão das relações entre as variáveis de um filme e as avaliações de críticos e público, fornecendo possíveis insights para a indústria cinematográfica. Os resultados obtidos indicam que a tomada de decisões, como a seleção de elenco, direção e roteiro, não são estatisticamente relevantes para alcançar o sucesso financeiro no projeto. A recepção mais positiva pelos críticos especializados mostrou-se relevante para calcular as notas da avaliação do público em geral, contudo ambas as avaliações não se mostraram relevantes para traçar um perfil de filmes com sucesso financeiro de 300% de ROI (Return on investment) percentual.

Palavras-chave: Industria cinematográfica; Arvore de regressão; Machine learining; Bilheteria; Estatística.

Introdução

A indústria cinematográfica, ao longo dos anos, tornou-se um dos pilares da cultura global, cativando audiências e gerando bilhões em receita. No entanto, por trás das telas de cinema e das telas de nossos dispositivos, existe uma intricada teia de fatores que influenciam o sucesso de um filme. Entre esses fatores, destacam-se o orçamento estipulado a eles, as avaliações feitas pelos críticos especializados que intensificam a publicidade no lançamento

e as avaliações feitas pelo público em geral que também é utilizada para ampliar o alcance da notícia de lançamento de novos filmes. Esta relação complexa entre as características de um filme e seu desempenho financeiro, crítico e de público é o foco deste estudo.

Em um mundo onde o entretenimento audiovisual desempenha um papel fundamental na vida de muitas pessoas, entender se há um perfil definido para filmes bem-sucedidos é mais do que um mero exercício acadêmico; é uma questão de grande relevância para a indústria cinematográfica. O sucesso de um filme não se resume apenas à bilheteria que ele arrecada, mas também à forma como é percebido e avaliado pela crítica e pelo público, uma vez que hoje em dia também temos sua distribuição pelos meios digitais. É nesse contexto que surge o questionamento fundamental deste estudo: "Como as características de um filme, como elenco, direção, roteiro, gênero, orçamento e duração, se correlacionam com o seu retorno do investimento e as avaliações feitas por críticos e pelo público?".

A relação entre os elementos técnicos e criativos de um filme e sua recepção por parte da audiência é uma questão intrincada e multifacetada. Afinal, o que faz com que um filme seja bem avaliado e, ao mesmo tempo, lucrativo? Para abordar essa pergunta complexa, recorremos ao vasto campo do estudo do cinema e das teorias relacionadas, bem como às técnicas de análise de dados avançadas, como data analytics, arvore de regressão e machine learning via a ferramenta R Studio que utiliza a linguagem R.

O cinema, desde sua origem no final do século XIX, tem sido uma forma poderosa de expressão artística e entretenimento. Com o passar dos anos, a produção cinematográfica se tornou uma indústria altamente competitiva e lucrativa. Filmes são produtos de investimento significativo, que envolvem equipes talentosas, extensos recursos financeiros e longos períodos de produção. Portanto, é crucial entender como as escolhas criativas e técnicas feitas durante a criação de um filme se traduzem em seu sucesso financeiro e crítico.

Este estudo aproveita o avanço das tecnologias de coleta e análise de dados, que nos permitem explorar de forma abrangente as relações entre diversas variáveis e as avaliações atribuídas pelos críticos e pelo público. Para alcançar esse objetivo, reunimos um conjunto de dados abrangente, abrangendo uma ampla lista de filmes de diferentes gêneros, orçamentos e diretores. Esses dados incluem informações detalhadas sobre o elenco, direção, roteiro, gênero, orçamento, bilheteria, duração e, mais crucialmente, as notas atribuídas pelos críticos e pelo público.

A análise deste conjunto de dados é conduzida por meio de métodos estatísticos e algoritmos de regressão, que nos ajudarão a identificar quais variáveis possuem maior influência na determinação da nota média atribuída a um filme. Essa abordagem multidisciplinar, que combina teorias do cinema com técnicas de análise de dados, permitirá uma compreensão mais profunda das conexões entre a forma e o conteúdo cinematográfico e seu desempenho crítico e financeiro.

O objetivo central deste estudo é contribuir para uma melhor compreensão das relações intrincadas entre as características de um filme e as avaliações realizadas por críticos e pelo público. Além disso, almejamos fornecer insights valiosos para a indústria cinematográfica, que podem ser aplicados na tomada de decisões cruciais, como a seleção de elenco, direção e roteiro. Ao entender como essas escolhas impactam no retorno do investimento e na recepção do filme, a indústria poderá otimizar seus processos criativos e de produção, buscando alcançar tanto o sucesso financeiro quanto a satisfação do público.

Ao analisar aprofundadamente essas relações complexas e dinâmicas, esperamos contribuir significativamente para o avanço do conhecimento no campo do cinema. Este estudo não apenas busca desvendar os fatores que influenciam as avaliações críticas e o retorno financeiro dos filmes, mas também oferece uma visão mais clara sobre como a indústria cinematográfica pode se adaptar e evoluir em um cenário em constante mudança. À medida que mergulhamos na interseção entre arte, entretenimento e negócios, buscamos iluminar o caminho para um cinema mais envolvente e bem-sucedido, que continue a cativar e emocionar audiências em todo o mundo.

Material e Métodos

A base de dados

Obtém-se um conjunto de dados de filmes no sistema Data.World contendo informações sobre os filmes, como elenco, direção, roteiro, gênero, orçamento, arrecadação mundial, duração, notas atribuídas pelos críticos especializados (Indicada pelo site Metacritics) e pelo público em geral (Indicada pelo site IMDB) e a partir destas variáveis calcula-se o retorno do investimento percentual correlacionando a variável orçamento e a de arrecadação de bilheteria global.
Filtra-se as observações que indicam as variáveis de maior interesse neste estudo e obtém-se uma base de dados com 6760 observações com 13 variáveis pertinentes entre qualitativas

e quantitativas que tem-se seus impactos analisados como descrito abaixo na análise exploratória de cada uma das variáveis, procurando melhor compreender os dados disponíveis, para selecionar-se do Método Supervisionado de regressão apropriado e quais variáveis utiliza-se.

Variáveis qualitativas

Título original (Titulo_original)

A variável título original é uma variável nominal que apresentou a qual filme os dados se referem e foi classificada no R como Character. Na lista de nomes apresentada não havia respostas não disponíveis [NA].

Gênero (gênero)

A variável gênero é uma variável nominal que apresentou a qual filme os dados se referem e foi classificada no R como Character. Na lista de nomes apresentada não havia respostas não disponíveis [NA].
A lista original de possibilidade de gêneros se mostrou extensa e redundante em algumas subclassificações como mostra a tabela 1 abaixo:

Tabela 1 – frequência das observações de cada gênero, 25 maiores ocorrências.

Genero	Frequencia	Porcentagem.Freq
Drama	405	5.99112426
Comedy, Drama, Romance	312	4.615384615
Comedy, Drama	292	4.319526627
Comedy	259	3.831360947
Drama, Romance	222	3.284023669
Comedy, Romance	205	3.032544379
Action, Crime, Drama	170	2.514792899
Animation, Adventure, Comedy	162	2.396449704
Crime, Drama, Thriller	139	2.056213018
Action, Adventure, Sci-Fi	120	1.775147929
Action, Adventure, Comedy	107	1.582840237
Crime, Drama, Mystery	105	1.553254438
Horror, Mystery, Thriller	100	1.479289941
Action, Crime, Thriller	99	1.464497041
Action, Comedy, Crime	97	1.434911243
Biography, Drama, History	95	1.405325444
Horror, Thriller	89	1.316568047
Action, Adventure, Fantasy	88	1.301775148
Drama, Thriller	88	1.301775148
Crime, Drama	87	1.286982249
Comedy, Crime, Drama	80	1.183431953
Biography, Drama	77	1.139053254
Horror	76	1.124260355
Action, Adventure, Drama	73	1.079881657
Comedy, Crime	68	1.00591716

Fonte: Banco de dados original

Para possibilitar a transformação destas categorias em dummys, uma reclassificação foi aplicada para reduzir os gêneros avaliados, assim considera-se a primeira indicação de gênero de cada observação como a principal e única e obtém-se uma nova tabela de frequências como na tabela 2 para os gêneros compilados:

Tabela 2 - frequência das observações de cada gênero após a redução dos gêneros a suas indicações principais, juntamente com sua porcentagem de ocorrências.

Genero compilado	Frequencia	Porcentagem
Comedy	1808	26.74556213
Drama	1460	21.59763314
Action	1455	21.52366864
Crime	483	7.144970414
Biography	431	6.375739645
Adventure	349	5.162721893
Horror	333	4.926035503
Animation	326	4.822485207
Fantasy	33	0.48816568
Mystery	26	0.384615385
Thriller	15	0.221893491
Romance	11	0.162721893
Sci-Fi	9	0.133136095
Family	8	0.118343195
Western	5	0.073964497
Musical	4	0.059171598
War	2	0.029585799
Film-Noir	1	0.014792899
Music	1	0.014792899

Fonte: Banco de dados original

Tabela 3 - Variação da variavel ROI_percentual para cada gênero principal encontrado no dataframe, onde 100 indica que houve 100% de lucro sobre o orçamento para as gravações e -100 significa que todo o valor investido não foi recupero pela bilheteria nos cinemas.

genero_principal	ROI_percentual.max	ROI_percentual.mean	ROI_percentual.min
Action	29056	194	-100
Adventure	10834	259	-100
Animation	12237	369	-100
Biography	3945	165	-100
Comedy	36552	297	-100
Crime	4002	121	-100
Drama	17291430	12170	-100
Family	7452	1003	-82
Fantasy	895	137	-98
Film-Noir	-100	-100	-100
Horror	3.86E+08	1164277	-100
Music	-56	-56	-56
Musical	6516	1571	-100
Mystery	3603	320	-88
Romance	416	24	-100
Sci-Fi	563	73	-100
Thriller	901	0	-100
War	-52	-70	-87

Western	2400	993	-99

Fonte: Banco de dados original

Tabela 4 – 1º e 3º quatis da variavel ROI_percentual para cada gênero principal encontrado no dataframe, onde 100 indica que houve 100% de lucro sobre o orçamento para as gravações e -100 significa que todo o valor investido não foi recupero pela bilheteria nos cinemas.

genero_principal	ROI_percentual.Q1.25%	ROI_percentual.Q3.75%
Action	-25	243
Adventure	-32	305
Animation	36	389
Biography	-59	197
Comedy	-50	251
Crime	-79	159
Drama	-83	191
Family	-72	288
Fantasy	-64	217
Film-Noir	-100	-100
Horror	6	727
Music	-56	-56
Musical	-80	1586
Mystery	21	368
Romance	-99	93
Sci-Fi	-78	-4
Thriller	-100	-87
War	-78	-61
Western	-98	2004

Fonte: Banco de dados original

Língua (lingua)

A variável língua é categórica e indica as línguas usadas durante os diálogos nos filmes, esta variável foi classificada no R como numérica. Na lista de minutos por filme apresentada não havia respostas não disponíveis [NA].

Tabela 5 – frequência das trinta observações mais encontradas de cada língua utilizada nos diálogos dos filmes.

Lingua	Frequencia	Porcentagem.Freq
English	3986	58.96449704
English, Spanish	361	5.340236686
English, French	187	2.766272189
English, German	80	1.183431953
English, Italian	80	1.183431953
English, Russian	72	1.065088757
French	49	0.724852071
Spanish	41	0.606508876
English, Japanese	39	0.576923077
French, English	36	0.532544379
English, Mandarin	35	0.517751479
English, Arabic	32	0.473372781
English, Latin	29	0.428994083
English, French, Spanish	24	0.355029586
English, Ukrainian	24	0.355029586
Spanish, English	24	0.355029586
English, American Sign Lan	22	0.325443787
English, Hebrew	21	0.310650888
Japanese	20	0.295857988
English, French, German	19	0.281065089
English, Cantonese	17	0.25147929
English, German, French	16	0.236686391
English, Portuguese	16	0.236686391
Mandarin	16	0.236686391
English, French, Italian	15	0.221893491
German	15	0.221893491
Korean	13	0.192307692
English, Italian, Spanish	12	0.177514793
English, Spanish, French	12	0.177514793
Portuguese	11	0.162721893

Fonte: Banco de dados original

Diretor (diretor)

A variável diretor é categórica e indica os profissionais que comandaram as filmagens em cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de diretores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 6 – frequência dos dez diretores presentes em observações mais encontradas de cada trabalho responsável nos filmes.

Diretor	Frequencia	Porcentagem.Freq
Clint Eastwood	34	0.502959
Woody Allen	34	0.502959
Steven Spielberg	31	0.45858
Steven Soderbergh	26	0.384615
Martin Scorsese	24	0.35503

Ridley Scott	24	0.35503
Ron Howard	22	0.325444
Brian De Palma	18	0.266272
Renny Harlin	18	0.266272
Robert Zemeckis	18	0.266272
Tim Burton	18	0.266272

Fonte: Banco de dados original

Escritor (escritor)

A variável escritor é categórica e indica os profissionais que redigiram a história de cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de diretores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 7 – frequência dos dez escritores presentes em observações mais encontradas de cada trabalho responsável nos filmes.

Escritor	Frequencia	Porcentagem.Freq
Woody Allen	29	0.428994
Joel Coen, Ethan Coen	13	0.192308
John Hughes	11	0.162722
Kevin Smith	10	0.147929
M. Night Shyamalan	10	0.147929
Jon Lucas, Scott Moore	9	0.133136
Christopher Markus, Stephen McFeely	8	0.118343
Luc Besson, Robert Mark Kamen	8	0.118343
Patrick Melton, Marcus Dunstan	8	0.118343
Tyler Perry	8	0.118343

Fonte: Banco de dados original

Atores principais e coadjuvantes

Divide-se duas variáveis indicando os Atores principais e coadjuvantes relacionados as filmagens de cada um dos filmes, esta variável foi classificada no R como Factor. Na lista de atores por filme apresentada não havia respostas não disponíveis [NA].

Tabela 8 – frequência dos dez atores principais mais encontrados em observações de cada filme descrito.

Ator_principal	Frequencia	Porcentagem
Nicolas Cage	45	0.665680473
Robert De Niro	36	0.532544379
Bruce Willis	35	0.517751479

Clint Eastwood	33	0.48816568
Johnny Depp	32	0.473372781
Tom Hanks	32	0.473372781
Tom Cruise	31	0.458579882
Adam Sandler	30	0.443786982
Denzel Washington	28	0.414201183
John Travolta	28	0.414201183

Fonte: Banco de dados original

Tabela 9 – frequência dos dez atores coadjuvantes mais encontrados em observações de cada filme descrito.

Ator_coadjuvante	Frequencia	Porcentagem
Gene Hackman	14	0.207192541
Morgan Freeman	14	0.207192541
Robert Downey Jr.	14	0.207192541
Annette Bening	13	0.192393074
Danny DeVito	13	0.192393074
Diane Keaton	13	0.192393074
Samuel L. Jackson	13	0.192393074
Julianne Moore	12	0.177593607
Nicole Kidman	12	0.177593607
Robert De Niro	12	0.177593607

Fonte: Banco de dados original

Variáveis quantitativas

Duração (duração)

A variável duração está em minutos e foi classificada no R como numérica. Na lista de minutos por filme apresentada não havia respostas não disponíveis [NA].

Tabela 10 – Analise descritiva de 6760 observações da variável duração em minutos

Variavel Duração em minutos	Valores
Min.	63
1st Qu.	95
Median	104
Mean	107.6419
3rd Qu.	117
Max.	321

Fonte: Banco de dados original

Nota média IMDB (Nota_média_IMDB)

A variável nota média IMDB mostra a nota média de 0 a 10 dadas pelo publico em geral ao filme, quanto maior o número de votos na variável número de votos IMDB maior a estabilidade desta média e a garantia que ela representa a opinião de diferentes espectadores de diferentes culturas e regiões geográficas.

Tabela 11 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original.

	Nota_média_Imdb
Minimo	1.4
1° quartil	5.8
Média	6.4
3° quartil	7.1
Maximo	9.3

Fonte: Banco de dados original

Número votos IMDB (Número_votos_IMDB)

A variável número votos IMDB mostra o número de votos IMDB recebidos por diferentes espectadores em cada filme da base. Quanto maior o número de votos recebidos, maior a estabilidade da nota média de um filme e maior a garantia que ela representa a opinião de diferentes espectadores de diferentes culturas e regiões geográficas.

Tabela 12 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original

	Número_votos_IMDB
Minimo	100
1° quartil	10,166
Média	90,120
3° quartil	98,731
Maximo	2,159,628

Fonte: Banco de dados original

Nota média Metacritcs (Nota_média_metascore)

A variável nota média Metacritics mostra a nota média de 0 a 100 dadas por críticos de cinema ao filme, quanto maior o número de votos na variável número avaliações críticos de cinema maior a estabilidade desta média e a garantia que ela representa a opinião de

diferentes profissionais do ramo cinematografico de diferentes culturas e regiões geográficas.

Tabela 13 – Analise descritiva de 6760 observações da nota média Metacritics dos filmes da base de dados original.

	Nota_média_metascore
Minimo	1
1° quartil	41
Média	55
3° quartil	68
Maximo	100

Fonte: Banco de dados original

Número avaliações críticos de cinema (Número_avaliações_criticoscinema)

A variável número avaliações críticos de cinema mostra o número de votos no sistema Metacritics, especializado em identificar avaliações de filmes feitas por profissionais do ramo, avaliações preparadas por diferentes criticos para cada filme da base. Quanto maior o número de avaliações recebidas, maior a estabilidade da nota média de um filme e maior a garantia que ela representa a opinião de diferentes profissionais do ramo cinematografico de diferentes culturas e regiões geográficas.

Tabela 14 – Analise descritiva de 6760 observações do número de avaliações de críticos de cinema para os filmes da base de dados original.

	Número_avaliações_criticoscinema
Minimo	1
1° quartil	55
Média	144
3° quartil	192
Maximo	987

Fonte: Banco de dados original

Número avaliações Metacritics (Número_avaliações_Metacritics)

A variável número de avaliações Metacritics mostra o número de avaliações indicadas no sistema por diferentes espectadores em cada filme da base. Por não ser a especialidade do site Metacritics, é possível perceber que quanto maior a popularidade do

filme em questão, maior o número de avaliações. Estas avaliações não contam para montar a média Metascore pois ela utiliza somente opiniões de profissionais do ramo.

Tabela 15 – Analise descritiva de 6760 observações da nota média IMDB dos filmes da base de dados original

	Número_avaliações_Metacritics
Minimo	1
1° quartil	69
Média	288
3° quartil	327
Maximo	8302

Fonte: Banco de dados original

Orçamento USD USA (Orçamento_USD_USA)

A variável orçamento USD USA indica qual foi o montante utilizado para a a produção de divulgação de cada filme, está classificada como variável numérica no R e auxiliou a criar a variável ROI percentual.

Tabela 16 – Analise descritiva de 6760 observações do valor do orçamento de cada filme indicado no banco de dados.

	Orçamento_USD_USA
Minimo	$2.00
1° quartil	$5,000,000.00
Média	$29,414,267.18
3° quartil	$35,000,000.00
Maximo	$356,000,000.00

Fonte: Banco de dados original

Renda bruta mundial USD (Renda_bruta_mundial_USD)

A variável renda bruta mundial USD indica qual foi o montante arrecadado em bilheteria após a comercialização de cada filme, a mesma está classificada como variável numérica no R e auxiliou a criar a variável ROI percentual.

Tabela 17 – Analise descritiva de 6760 observações do valor arrecadado em bilheteria de cada filme indicado no banco de dados.

	Renda_bruta_mundial_USD
Minimo	$77.00
1° quartil	$3,231,421.50

Média	$84,368,142.36
3° quartil	$86,298,548.50
Maximo	$2,797,800,564.00

Fonte: Banco de dados original

ROI percentual (ROI_percentual)

A variável ROI percentual indica a porcentagem de lucro ou prejuízo após a comercialização de cada filme, a mesma está classificada como variável numérica no R e foi criada a partir das variáveis Renda Bruta mundial e Orçamento USD USA pois pode-se estimar aproximadamente o ROI (Return on investment) de um projeto a partir das despesas e receitas totais do mesmo.

Tabela 18 – Analise descritiva de 6760 observações do ROI percentual calculado de cada filme indicado no banco de dados.

	ROI_percentual
Minimo	-100
1° quartil	-56
Média	60.158
3° quartil	249
Maximo	385.711.840

Fonte: Banco de dados original

Resultados e Discussão

Correlação Variáveis Numéricas

Com a função Cor que faz parte da base do R studio entende-se qual a correlação entre as variáveis numéricas e como seus comportamentos estão ou não interligados. No cálculo apresenta-se números de -1 a 1 onde 1 são correlações diretamente proporcionais e -1 é apresentado para correlações inversamente proporcionais. Os cálculos que apresentam números abaixo de 0,5, sejam positivos ou negativos, indicam que não há uma correlação significativa entre as variáveis, já cálculos que apresentam números entre 0,5 e 0,7, sejam positivos ou negativos, indicaram uma correlação moderada para estas variáveis.
Ainda que uma correlação forte seja algo significativo entre variáveis, não pode-se atribuir o peso de causa e efeito as variáveis correlacionadas pois muito frequentemente o comportamento de uma variável é ditado pelo comportamento de um conjunto de outras variáveis que podem ou não compor o banco de dados do estudo e que influenciam o resultado final, assim, precisamos inclui-las ao cálculo de um modelo de regressão para

entendermos qual a magnitude do impacto de cada variável independente no resultado final de uma variável de interesse.

Após o cálculo obtem-se as seguintes matrizes de correlação, onde as correlações fortes e moderadas foram destacadas em vermelho.

Tabela 19 – Matrix de correlação 1

	Nota_media_Imdb	Numero_votos_IMDB
ROI_percentual	0.007566378	0.020786152
Duração	0.393848365	0.322905304
Nota_media_Imdb	1	0.429060951
Numero_votos_IMDB	0.429060951	1
Orçamento_Usd_USA	0.03638646	0.445612336
Renda_bruta_mundial	0.187968565	0.60745703
Nota_media_metascore	0.737522875	0.294331177
Numero_avaliações_Metacritics	0.285854836	0.740869616
Numero_avaliações_criticoscinema	0.321291417	0.607526423

Fonte: Banco de dados original

Tabela 20 – Matrix de correlação 2

	Duração	Numero_avaliações_Metacritics
ROI_percentual	-0.020449276	0.05914556
Duração	1	0.318545567
Nota_media_Imdb	0.393848365	0.285854836
Numero_votos_IMDB	0.322905304	0.740869616
Orçamento_Usd_USA	0.276429593	0.493707867
Renda_bruta_mundial	0.25113298	0.621451303
Nota_media_metascore	0.277902281	0.200792944
Numero_avaliações_Metacritics	0.318545567	1
Numero_avaliações_criticoscinema	0.25564011	0.608298185

Fonte: Banco de dados original

Tabela 21 – Matrix de correlação 3

	Nota_media_ metascore	Numero_avaliações_ criticoscinema
ROI_percentual	0.024015817	0.028568535
Duração	0.277902281	0.25564011
Nota_media_Imdb	0.737522875	0.321291417
Numero_votos_IMDB	0.294331177	0.607526423
Orçamento_Usd_USA	-0.039414471	0.505259314
Renda_bruta_mundial	0.121124117	0.53785828
Nota_media_metascore	1	0.307194319

Numero_avaliações_Metacritics	0.200792944	0.608298185
Numero_avaliações_criticoscinema	0.307194319	1

Fonte: Banco de dados original

Tabela 22 – Matrix de correlação 4

	Orçamento_Usd_USA	Renda_bruta_mundial
ROI_percentual	-0.016100116	0.018882105
Duração	0.276429593	0.25113298
Nota_media_Imdb	0.03638646	0.187968565
Numero_votos_IMDB	0.445612336	0.60745703
Orçamento_Usd_USA	1	0.748274837
Renda_bruta_mundial	0.748274837	1
Nota_media_metascore	-0.039414471	0.121124117
Numero_avaliações_Metacritics	0.493707867	0.621451303
Numero_avaliações_criticoscinema	0.505259314	0.53785828

Fonte: Banco de dados original

Tabela 23 – Matrix de correlação 5

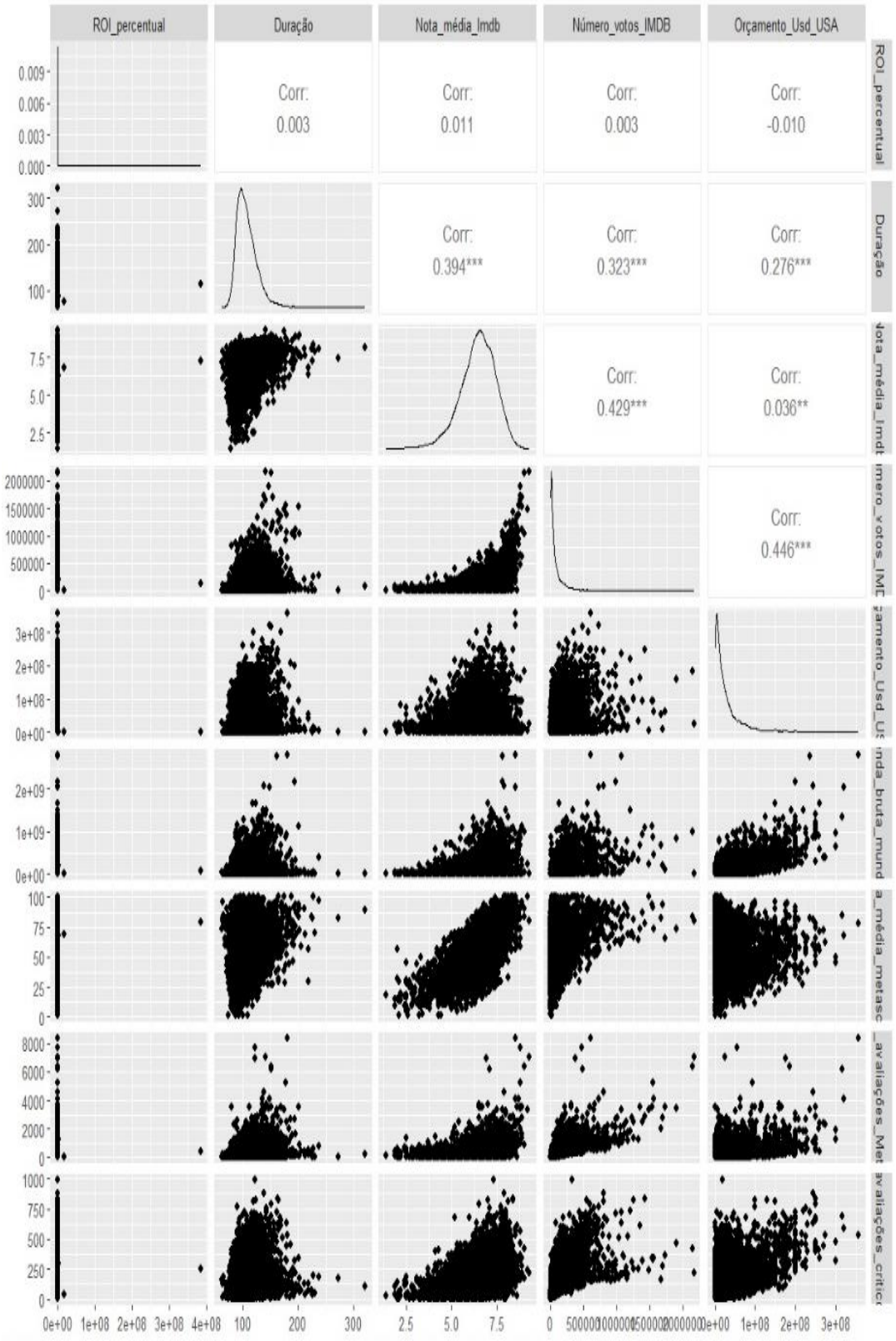
	ROI_percentual
ROI_percentual	1
Duração	-0.020449276
Nota_media_Imdb	0.007566378
Numero_votos_IMDB	0.020786152
Orçamento_Usd_USA	-0.016100116
Renda_bruta_mundial	0.018882105
Nota_media_metascore	0.024015817
Numero_avaliações_Metacritics	0.05914556
Numero_avaliações_criticoscinema	0.028568535

Fonte: Banco de dados original

Cálculos de significância estatística

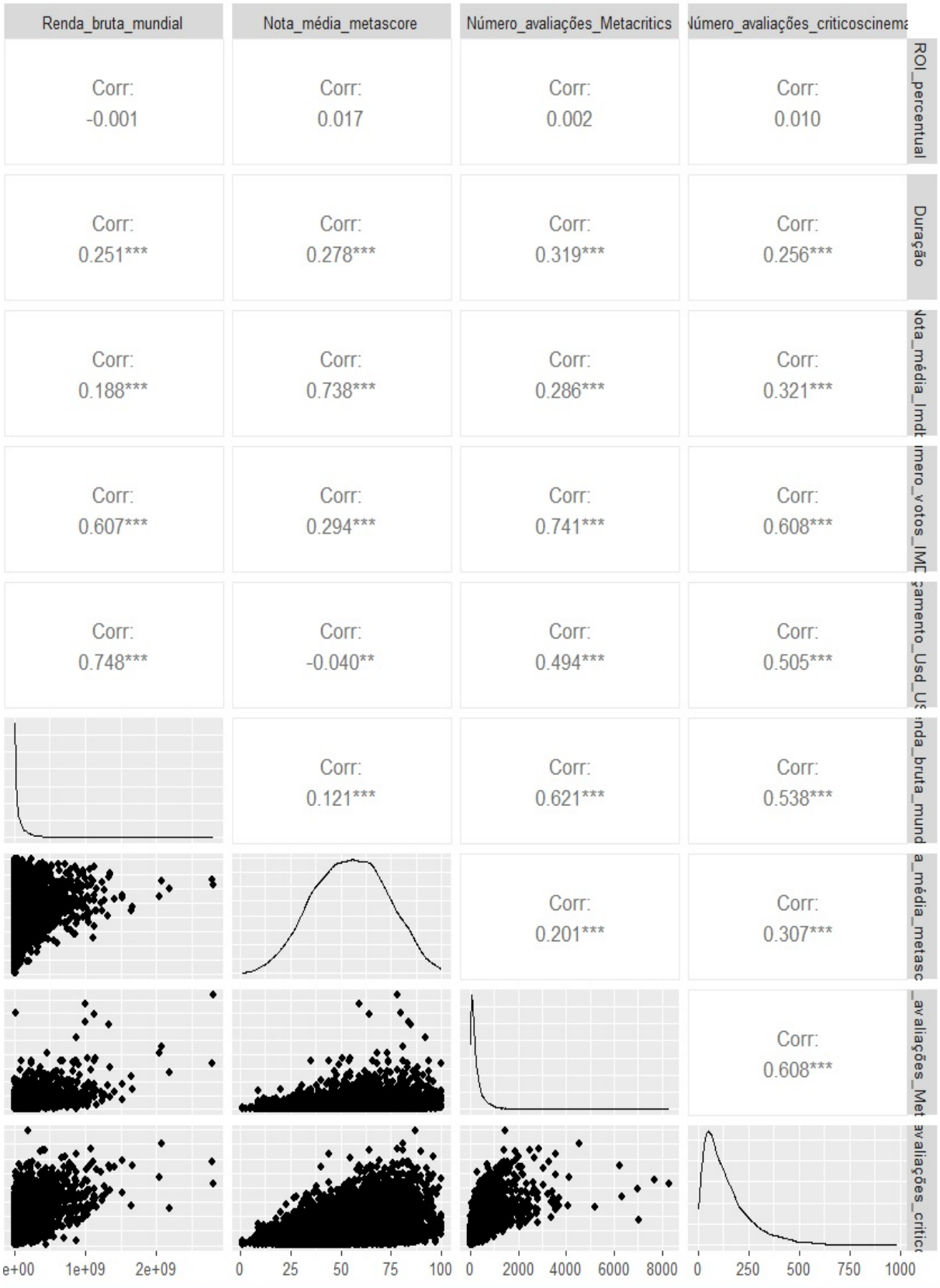
Agora utiliza-se o pacote Ggally do R studio para calcular as variáveis numéricas do banco de dados, deste calculo obtém-se a figura 1, que indica o grau de significância estatística de cada correlação entre as variáveis numéricas, bem com os gráficos de dispersão entre elas e o comportamento linear proporcional entre algumas delas.
Os asteriscos indicados em cada correlação mostram se esta correlação tem baixa probabilidade de ser obtida por acaso, onde um “*” significa até 5% de chance de ser uma correlação ao acaso, dois “**” significam até 1% e três “***” significam menos de 1% de acaso na correlação.

Figura 1 – Primeira parte dos gráficos de dispersão, de frequência e significância estatística das variáveis numéricas:



Fonte: Banco de dados original

Figura 2 – Segunda parte dos gráficos de dispersão, de frequência e significância estatística das variáveis numéricas:



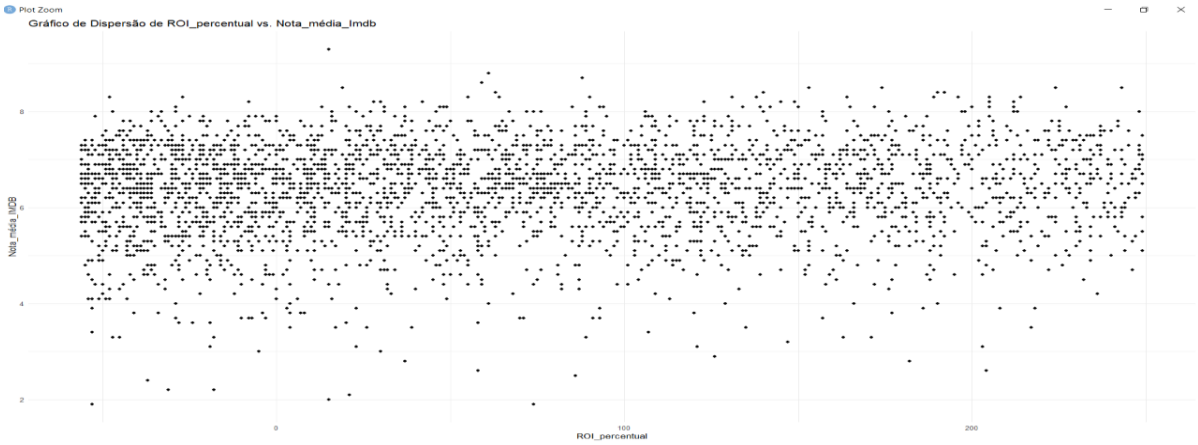
Fonte: Banco de dados original

Na primeira linha da matriz das figuras 1 e 2 fica expressa a baixa correlação estatística entre o Retorno do Investimento percentual (ROI_percentual) e as demais variáveis numéricas do banco de dados original.

Modelagem de uma regressão linear simples

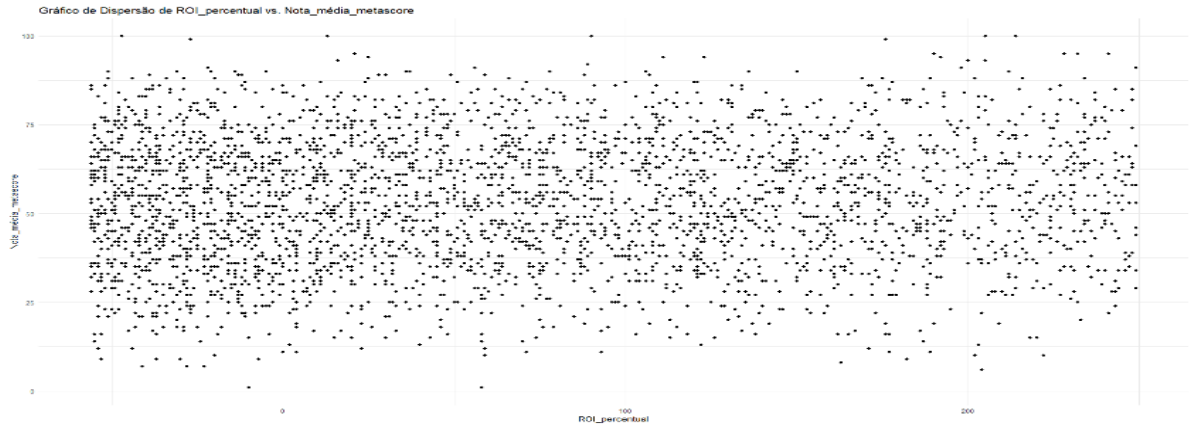
Avalia-se o comportamento da relação das variáveis de interesse ROI_percentual, Nota_média_imdb e Nota_média_metascore nas figuras 3, 4 e 5 abaixo:

Figura 3 – Gráfico de dispersão do primeiro ao terceiro quartil da variável ROI percentual vs Nota média Imdb



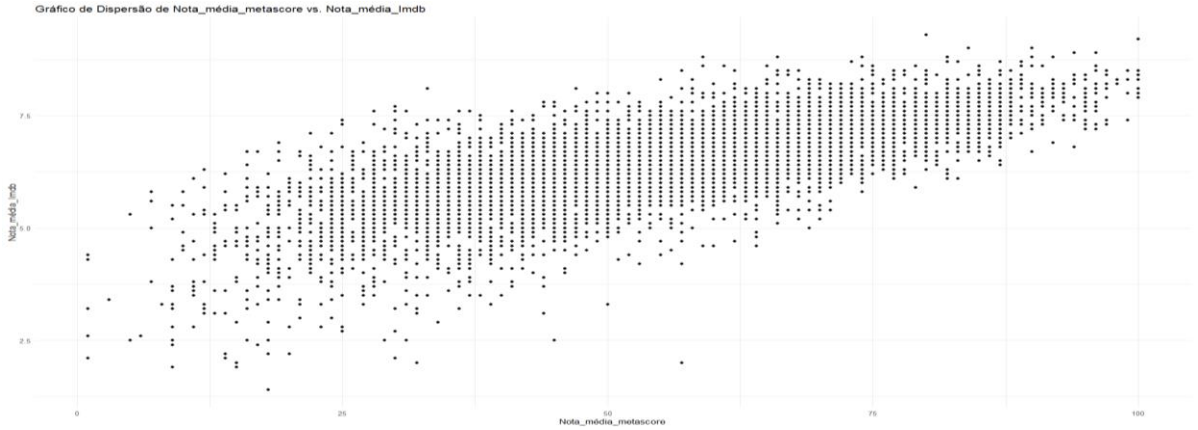
Fonte: Banco de dados original

Figura 4 – Gráfico de dispersão do primeiro ao terceiro quartil da variável ROI percentual vs Nota média Metascore



Fonte: Banco de dados original

Figura 5 – Gráfico de dispersão das variáveis Nota média Imdb vs Nota média Metascore



Fonte: Banco de dados original

Ao avaliar os gráficos, percebe-se que apenas o comportamento das duas últimas variáveis tem correlação estatística de intensidade moderada entre si e podem ser modelados afim de estimar os resultados de uma com dados do outra, assim modela-se uma estimativa para a Nota_média_imdb, que apresenta a recepção do público em geral, a partir da variável Nota_média_metascore, que apresenta a recepção média dos profissionais da área, pois a variável independente é composta com em média 144 opiniões de profissionais do ramo distribuídos em diferentes culturas de regiões geográficas, assim uma média mais fácil de se obter antes de disponibilizar o filme para o público geral e gerar receita a partir das bilheterias. O modelo calculado retornou os seguintes detalhes:

Call:

lm(formula = Nota_média_Imdb ~ Nota_média_metascore, data = Dadosfiltrados)

(um modelo de regressão linear simples com "Nota_média_Imdb" como a variável de resposta e "Nota_média_metascore" como a variável independente.)

Residuals:

Min	1Q	Median	3Q	Max
-4.5181	-0.3801	0.0210	0.4243	2.5472

(Esta parte fornece estatísticas resumidas sobre os resíduos do modelo)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.225449	0.025819	163.66	<2e-16 ***
Nota_média_metascore	0.040222	0.000448	89.79	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Aqui, encontra-se informações sobre os coeficientes do modelo, que representam a relação entre a variável independente e a variável de resposta.

Estimate: Este é o valor estimado do coeficiente. Neste caso, o coeficiente estimado para "Nota_média_metascore" é 0.040222.
Std. Error: O erro padrão do coeficiente estimado.
t value: O valor t é a estatística t, que mede o quão distante o coeficiente estimado está de zero em termos de erros padrão.
Pr(>|t|): Este é o valor-p associado ao teste t. Indica a probabilidade de observar um valor t tão extremo quanto o observado, assumindo que não há relação entre as variáveis. Quanto menor o valor-p, mais significativo é o coeficiente.
Signif. codes: São códigos que indicam o nível de significância estatística. No seu caso, "***" significa altamente significativo (p-value < 0.001).)

Residual standard error: 0.6717 on 6760 degrees of freedom
Multiple R-squared: 0.544,
Adjusted R-squared: 0.5439
F-statistic: 8062 on 1 and 6758 DF,
p-value: < 2.2e-16

Neste caso, o modelo parece significativo, com valores de p muito baixos (menos de 2.2e-16), indicando que a variável "Nota_média_metascore" é significativa na previsão da "Nota_média_Imdb". Além disso, o R^2 múltiplo de 0.544 sugere que cerca de 54.4% da variabilidade na "Nota_média_Imdb" pode ser explicada via a correlação "Nota_média_metascore".
Ao calcular o erro percentual em relação à variável de resposta real foi obtido o valor de erro percentual médio de 7.96%, este valor foi calculado a partir do erro do Erro Médio Absoluto (MAE - Mean Absolute Error), onde ele foi dividido pela média da variável resposta e multiplicado por 100. Assim, pode-se indicar que o modelo erra a predição em aproximadamente 8% de suas tentativas dentro da mesma base analisada.

Modelagem comparando base de treinamento e base teste

Após avaliar o cálculo da regressão linear simples anterior, divide-se a base de 6760 observações em uma base de treino e outra de teste, para avaliar a capacidade de estimação do modelo para dados futuros.
Assim modela-se uma estimativa na base de treino para a Nota_média_imdb, que apresenta a recepção do público em geral, a partir da variável Nota_média_metascore, que apresenta a recepção média dos profissionais da área, e comparada com a capacidade de predição na base de teste. O MAPE (Mean Absolute Percentage Error) erro percentual médio estimado deste novo modelo foi de 9,23%, assim, pode-se dizer que uma estimativa de nota 7,0 no IMDB a partir da Nota_média_metasocre, terá uma margem de erro de +/- 0,32, o que manteria a avaliação final deste filme na mesma faixa de avaliação de um bom filme segundo o público em geral de diferentes culturas e regiões geográficas pois a variação de 0,32 não impacta significativamente na nota média final do IMDB predita.

Potencial de sucesso no retorno do investimento

Uma vez que tem-se a composição da variável Nota_média_metascore a partir das avaliações individuais de críticos de cinema que são contabilizadas na variável Numero_avaliações_críticoscinema, para se ter uma nota média metascore confiável é preciso alcançar a média indicada na Tabela 13, que é 144 avaliações de profissionais do ramo cinematrografico, ou ao menos número de avaliações do primeiro quartil desta varável, 55 avaliações (Tabela 13), avaliações estas de autores com diferentes perspectivas, formações acadêmicas, idades e costumes, sendo todas as notas atribuídas de 1 a 100 com o mesmo peso entre si para compor a média e sem viés e envolvimento do avaliador com a produtora e distribuidora do filme sob análise. Quanto maior este número de avaliações, mais perto a média técnica real o número está e mais precisa é a estimativa da recepção do público final.

Com base nos dados obtidos, a aprovação tanto do público geral como dos críticos do ramo não está estatisticamente correlacionada a um percentual especifico do retorno do investimento indicado na variável ROI_percentual, assim, é mais plausível correlacionar a aprovação dos críticos apenas ao objetivo de conseguir aprovação minimamente aceitável do publico em geral, pois ainda que não se tenha uma correlação estatística entre o retorno do investimento e a aprovação do público, para se intensificar o efeito boca a boca de publicidade do lançamento e ampliar a receita total alcançada pelas bilheterias, é desejável que as primeiras impressões do filme sejam positivas. Na tabela abaixo é possível perceber uma variação entre a média do ROI_percentual de filmes de diferentes faixas de aceitação, o que

possivelmente é um indicio do fortalecimento ou enfraquecimento do efeito boca a boca de divulgação após a construção do pré-conceito de que um filme atende ou não atende as expectativas do público.

Tabela 24 – Media ROI_percentual de cada faixa de avaliação no IMDB.

Avaliação	Faixa_Nota_Imdb	Média_ROI_percentual
Antipatia esmagadora	0 a 1.9	159
Geralmente não favoravel	2 a 3.9	46
Mista ou média	4 a 6.9	4427
Geralmente favoravel	7 a 8.9	181816
Aclamação universal	9 a 10	1182

Fonte: Banco de dados original

A faixa de avaliação “Antipatia esmagadora” deve-se notificar o baixo orçamento médio dedicado a cada filme, como na tabela abaixo, o que facilita o alcance da média de 150% de retorno no investimento ainda que as avaliações indiquem antipatia ao filme. Outro ponto é que 150% de ROI não é considerado um sucesso de bilheteria pelo ramo cinematográfico, o sucesso financeiro só é apontado em filmes que alcançam 300% de ROI, assim custeando seus custos por arrecadar quatro vezes o valor de seu orçamento total, assim restando os 300% de ROI.

Tabela 25 – Media Orçamento_Usd_USA de cada faixa de avaliação no IMDB.

Avaliação	Faixa_Nota_Imdb	Média_Orçamento_usd_USA
Antipatia esmagadora	0 a 1.9	13,500,000
Geralmente não favoravel	2 a 3.9	24,052,870
Mista ou média	4 a 6.9	29,496,867
Geralmente favoravel	7 a 8.9	29,519,536
Aclamação universal	9 a 10	57,250,000

Fonte: Banco de dados original

Tabela 26 – Número de filmes em cada faixa de avaliação no IMDB.

Avaliação	Faixa_Nota_Imdb	Filmes por faixa
Antipatia esmagadora	0 a 1.9	3
Geralmente não favoravel	2 a 3.9	123
Mista ou média	4 a 6.9	4,503
Geralmente favoravel	7 a 8.9	2,127
Aclamação universal	9 a 10	4

Fonte: Banco de dados original

Arvore de regressão preditora para avaliação do publico

Montou-se uma primeira arvore de regressão para predizer a variável Nota_média_Imdb agora também contando com uma das variáveis categóricas, além da variável que indica a nota dos críticos, para se avaliar o quanto as variáveis categóricas impactam nas predições. No total, as seguintes variáveis foram utilizadas no cálculo:

modelo_arvore_Kfold10 <- rpart(Nota_média_Imdb ~ Nota_média_metascore + genero_principal + Duração + Orçamento_Usd_USA, data = Dadosfiltrados, xval=10,

control = rpart.control(cp = 0, maxdepth = 30))

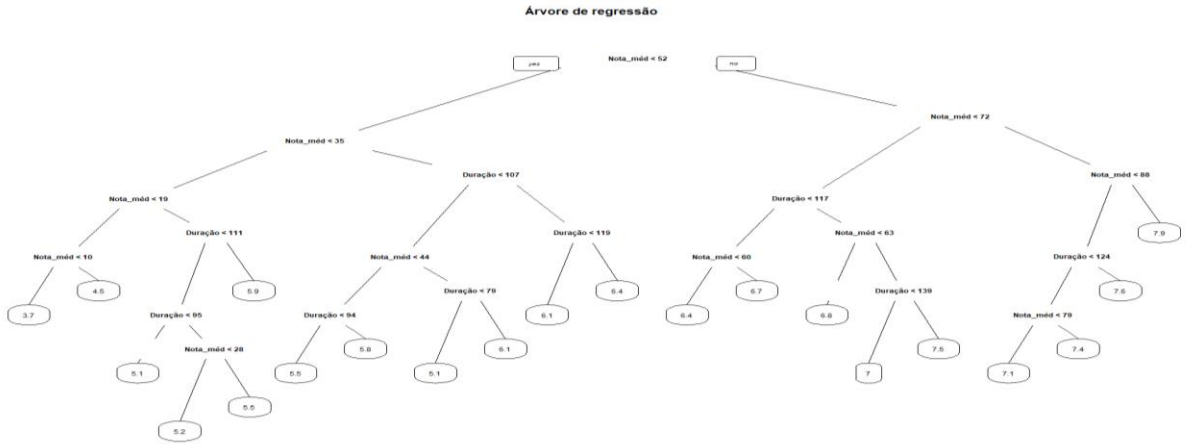
O modelo acima criou dez partições com o comando xval=10 para uma validação cruzada entre os modelos independentes. Deste modelo foi possível calcular o menor valor do custo de complexidade e obter uma arvore com este hiperparametro ajustado. Ao se utilizar o comando summary (modelo_arvore_Cpmin) foi possível verificar as seguintes importâncias percentuais de cada variável no calculo

Nota_média_metascore 81%
Duração 12%
genero_principal 6%
Orçamento_Usd_USA 1%

Calculo do SQE (Soma dos Quadrados de Erro): 2770.365
Calculo do SST (Soma dos Quadrados Total): 6685.724
Calculo do QMT (Quadrado Médio do Tratamento): 0.5792807
Calculo do R-quadrado: 0.5856297

Esta arvore preditora ilustrada ficou no seguinte formato:

Figura 6 – Formato arvore de decisão 1



Fonte: Banco de dados original

A arvore acima segue a taxa de erro de aproximadamente 0,3 para mais ou para menos e avaliou todos os critérios indicado.

Ao predizer a nota do IMDB a partir das variáveis acima para o filme do ano de 2023 “Mission: Impossible – Dead Reckoning Part One” que não consta na base de treinamento, nem na base de teste deste estudo por elas abrangerem filmes somente até o ano de 2019, verificou-se:
Nota média metascore = 81 / Genero principal = “Action” / Duração = 164 / Orçamento Usd USA = 291.000.000 / Predição nota Imdb arvore 1 = 7.60
A nota média real da avaliação do público no sistema IMDB está em 7.9 após 139 mil opiniões listadas até 22/09/2023.

Ao predizer a nota do IMDB a partir das variáveis acima para o filme do ano de 2023 “Barbie” que não consta na base de treinamento, nem na base de teste deste estudo por elas abrangerem filmes somente até o ano de 2019, verificou-se:
Nota média metascore = 80 / Genero principal = “Comedy” / Duração = 114 / Orçamento Usd USA = 145.000.000 / Predição nota Imdb arvore 1 = 7.37
A nota média real da avaliação do público no sistema está em 7.1 após 323 mil opiniões listadas até 22/09/2023

Arvores de regressão preditora a partir de dados técnicos

Ao comandar a linha abaixo no R studio cria-se uma arvore de regressão utilizando como variáveis preditoras tanto as variáveis categóricas como as numéricas do banco de dados, exceto as avaliações dos criticos:

```
modelo_arvore_IMDB_categoricas1 <- rpart(Nota_média_Imdb ~ Ator_principal +
Ator_coadjuvante + Escritor + Produtora + genero_principal + Duração + Pais + Lingua +
Diretor + Orçamento_Usd_USA, data = Dadosfiltrados, xval=10,
                control = rpart.control(cp = 0,
                                        minsplit = 2,
                                        maxdepth = 30))
```

Utiliza-se o comando which.min para encontrar o menor xerror e seu CP (custo de complexidade) desta arvore e entender até qual nó ela pode ser estendida para predizer a variável Nota média IMDB de filmes fora da base de treinamento.
Esta arvore se estabeleceu com apenas um nó de divisão, que evidencia que não é possível predizer as avaliações do publico em geral, sem correlaciona-la a variável nota média Metascore com as avaliações dos críticos. Adicionar mais “nós” apenas criaria uma estimativa melhor para os dados da própria base de treinamento.

Ao deixar o modelo se super ajustar aos dados da base com o comando:
modelo_arvore_IMDB_categoricas_2 <- rpart(Nota_média_Imdb ~ Ator_principal + Ator_coadjuvante + Escritor + Produtora + genero_principal + Duração + Pais + Lingua + Diretor + Orçamento_Usd_USA, data = Dadosfiltrados, xval=10, control = rpart.control(cp = 0, minsplit = 2, maxdepth = 5))
Temos as importâncias de cada variável para esta previsão de valores dentro da própria base.
Variável e porcentagem de importância:
Escritor 27%
Ator_coadjuvante 20%
Diretor 19%
Ator_principal 17%
Produtora 13%
Pais 3%
Lingua 1%
Assim, podemos indicar que o cargo mais importante dentro da base de 6760 filmes para predizer a nota média imdb é o de Escritor, para um R-quadrado de 0.9644815 super ajustado a base de treinamento, seguido de Ator_coadjuvante, Diretor e Ator_principal.

Ao montar uma arvore com o mesmo formato, mas agora para predizer as notas dos críticos de cinema para os filmes dentro da própria base de treinamento com o comando:
modelo_arvore_Metascore_categoricas <- rpart(Nota_média_metascore ~ Ator_principal + Ator_coadjuvante + Escritor + Produtora + genero_principal + Duração + Pais + Lingua + Diretor + Orçamento_Usd_USA, data = Dadosfiltrados, xval=10, control = rpart.control(cp = 0, minsplit = 2, maxdepth = 3))

Obtem-se as seguintes importâncias para cada variável:
Escritor 26%
Ator_coadjuvante 19%
Diretor 18%
Ator_principal 16%
Produtora 14%
Lingua 6%
Pais 1%
Reforçando a importância de cada cargo para a composição da nota dos críticos para os 6760 filmes de dentro da base de treinamento.

Arvore de regressão preditora para ROI_percentual

Ao comandar a linha abaixo cria-se uma arvore de regressão preditora para o ROI-percentual utilizando como variáveis preditoras tanto as variáveis categóricas como as numéricas do banco de dados:

modelo_arvore_ROI <- rpart(ROI_percentual ~ Ator_principal + Ator_coadjuvante + Escritor + Produtora + genero_principal + Duração + Pais + Lingua + Diretor + Nota_média_Imdb + Orçamento_Usd_USA + Nota_média_metascore, data = Dadosfiltrados, xval=10, control = rpart.control(cp = cp_min, minsplit = 2, maxdepth = 30))

Verifica-se que o valor do erro de validação cruzada de dez partições para de reduzir já no primeiro ponto, assim esta predição não se mostra confiável e útil para observações de fora da amostra utilizada para o treinamento, pois ao adicionar mais nós nesta arvore, ela apenas se ajusta e melhora sua predição aos dados já obtidos, diminuindo sua taxa de acerto para dados desconhecidos.

Abaixo os dados de complexidade do nó citada:

CP	nsplit	rel error	xerror	xstd
0.997982	0	1	0	0

Node number 1: 6760 observations

mean=60157.71, MSE=2.204882e+13

Fonte: Base de dados original

Considerações Finais

No estudo mostrou-se, estatisticamente, que é uma boa pratica das produtoras buscar avaliações de profissionais do ramo cinematográfico antes de lançar seus filmes, afim de uma previsão da recepção do público geral. Mostrou-se que diferentes configurações de orçamento, elenco e equipe técnica podem alcançar o sucesso financeiro ao se produzir um filme, já que o ROI percentual não está estatisticamente ligado a categorias de filmes, equipes técnicas e volumes de orçamento específicos. Mostrou que mesmo filmes bem avaliados pelo público podem gerar prejuízo e que filmes mal avaliados podem gerar lucro, ainda que a bibliografia especializada indique que más avaliações geram um impacto negativo no lançamento e na publicidade boca a boca, impacto negativo que provavelmente é reversível a partir de campanhas de divulgação para o público alvo correto para os filmes em questão, pois este ignora avaliações previas, sejam boas ou más.

O estudo ainda pode ser complementado por uma abordagem mais profunda das variáveis categóricas do banco dados a fim de indicar sua importância para a nota média do público geral e dos críticos de cinema, pois estas variáveis se mostraram relevantes somente nestas predições. Em testes, o alto número de categorias em cada uma das variáveis categóricas foi reduzido após ser aninhado segundo suas avaliações do público em geral, mas impossibilitou a predição para dados de fora do banco de dados atual, uma vez que uma mesma categoria foi relacionada a diferentes agrupamentos e o alto número de variáveis categóricas neste formato multiplicou esta complexidade na predição. Nenhuma das variáveis categóricas foi impactante para predizer o ROI de filmes de fora do banco de dados, aninhadas ou não.
Ainda é possível perceber na tabela número 21 que a variável renda bruta mundial em USD está correlacionada ao aumento do valor da variável orçamento usd usa, contudo, a mesma influencia não é observada no percentual de ROI, assim, produções de orçamento maior tendem a arrecadar mais nas bilheterias mundiais ao mesmo tempo que seguem com sua alta variação do percentual de retorno no investimento.

Agradecimentos

Gostaria de agradecer a Cleverson Pereira da Silva, meu irmão, graduando em Ciências da computação, que com sua paciência me inspirou a ser persistente no aprendizado e na utilização das ferramentas e das técnicas de machine learning. Sem o exemplo dele, não teria me mantido engajado para completar este passo de minha pós-graduação.

Referências

Arnold, Martin; Gerber, Alexander; Hanck, Christoph; Schmelzer, Martin. Introduction to Econometrics with R.

Belfiore, Patricia; Favero, Luiz Paulo. Data Science for Business and Decision Making. 2019.

Belfiore, Patricia; Favero, Luiz Paulo. Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 2017. Elsevier, Rio de Janeiro, RJ, Brasil.

Elberse, Anita. Blockbusters: Hit-making, Risk-taking, and the Big Business of Entertainment. 2013.

Epstein, Edward Jay. The Hollywood Economist 2.0: The Hidden Financial Reality Behind the Movies. 2012.

Fritz, Ben. The Big Picture: The Fight for the Future of Movies. 2018.

Harrison, Matt. Machine Learning – Guia de Referência Rápida: Trabalhando com Dados Estruturados. 2019.

James, G.; Hastie, T.; Tibshirani, R.; Witten, D. An Introduction to Statistical Learning: With Applications in R. Editora Springer. 2021.

McKeon, J. Machine Learning Predicts Dialysis, Death in COVID-19 Patients. 2021.

McNulty, Keith. Handbook of Regression Modeling in People Analytics: With Examples in R and Python. 2021.

Peter, A.; Peter, B. Estatística prática para cientistas de dados: 50 conceitos essenciais. 2019.