NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

MODELO DE PREVISÃO DA RENTABILIDADE DE UM FILME

Marta Sofia Proença Galvão

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

MODELO DE PREVISÃO DA RENTABILIDADE DE UM FILME

por

Marta Sofia Proença Galvão

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence

Orientador: Roberto Henriques

Coorientador: Gonçalo Ferreira

Novembro, 2017

RESUMO

Conseguir prever o lucro de um filme é cada vez mais uma necessidade emergente para os grandes estúdios mundiais da atualidade na hora da decisão do investimento relativo a uma determinada produção. Este estudo pretende prever a rentabilidade de um filme através da elaboração de um modelo preditivo, que recorre a diversas metodologias de Data Mining - redes neuronais, árvores de decisão e regressões - para obter a melhor antevisão das receitas de bilheteira.

Esta análise foi feita recorrendo a dados históricos relacionados com as variáveis mais significativas para os espectadores na hora da decisão da sua visualização, selecionadas consoante o seu valor e importância para o modelo. Foi então possível reunir um conjunto relevante de informações e características cinematográficas que permitiram, à *posteriori*, a análise e previsão das receitas de filmes ainda não classificados.

Neste modelo foram utilizadas três abordagens distintas da variável dependente (intervalar, binária e multi-classe), com o objetivo de estudar a diferença e influência preditiva que cada uma tem nos resultados da investigação.

Para determinar qual a previsão estatisticamente mais correta, foram ainda utilizadas métricas distintas: erro de classificação, índice ROC e medida-F para os modelos que utilizaram a variável dependente discreta e o erro quadrático médio e erro máximo absoluto para o caso da contínua.

Foi possível concluir, após este estudo, que os melhores resultados preditivos foram obtidos através das redes neuronais e que esta metodologia foi bastante superior em relação às outras abordagens. Apurou-se ainda relativamente à distinção representativa da variável dependente, que o modelo multi-classe apresentou uma taxa de erro bastante mais elevada do que as restantes, o que se justifica pelo aumento da dificuldade em prever acertadamente em nove classes distintas.

PALAVRAS-CHAVE

Filmes, Data Mining, Rentabilidade de um filme, Lucro de bilheteira, Análise preditiva, Redes neuronais, Árvores de decisão, Regressão

ÍNDICE

Índice de Figuras

Índice de Tabelas

Índice de Gráficos

1. Introdução

A maior indústria cinematográfica situa-se nos Estados Unidos da América, também conhecida por Hollywood, sendo a mais antiga do mundo e uma das mais lucrativas de sempre, tendo arrecadado cerca de 11.4 biliões de dólares só no decorrer do ano de 2016, com a estreia de 735 filmes (Mojo, 2016b). Porém, os prejuízos de um filme pouco lucrativo podem contribuir para o abalo parcial ou mesmo total do estado financeiro de um estúdio. Dados relativos a 2012 permitem estabelecer uma visão mais clara da importância da rentabilidade de uma produção, sendo que, apenas 10% dos filmes lançados foram responsáveis por mais de 68.8% dos lucros totais de bilheteira nesse ano (Ghiassi, Lio, & Moon, 2015).

Esta área é caracterizada por ser uma das mais arriscadas para os investidores devido à sua imprevisibilidade, sendo uma indústria multimilionária que gerou mais de 11 biliões de dólares de lucro no ano de 2016 e um crescimento de mais de 6 biliões de dólares nos últimos 11 anos (Nash Information Services, 2016) .

Entre 2000 e 2010 apenas cerca de 36% dos filmes norte americanos produzidos foram lucrativos relativamente ao orçamento de produção (Lash & Zhao, 2015). Face a esta constante incerteza monetária é imperativo conhecerem-se os componentes cinematográficos que mais influenciam as audiências de um filme para que o financiamento gasto possa ser justificado *à posteriori*.

Estima-se que 80% dos lucros da indústria na última década foram gerados a partir de apenas 6% dos filmes lançados (Im & Nguyen, 2011).

Alguns estudos defendem que o principal fator responsável pela rentabilidade de um filme é a opinião que os espectadores formam com base em conclusões e apreciações feitas pelos críticos (Boatwright, Basuroy, & Kamakura, 2007), enquanto outros afirmam que o êxito reside na popularidade das estrelas de Hollywood que nele participam (Karniouchina, 2010).

Esta teoria é também apoiada por alguns autores que defendem que a diversidade de um elenco é positivamente correlacionada com a receita de bilheteira, levando ao sucesso financeiro de um filme (Lash & Zhao, 2015). Porém, este sucesso não significa que o filme seja necessariamente rentável devido ao custo associado à participação de cada um dos atores.

O fenómeno de popularidade de cada filme baseia-se muito no comportamento que é gerado antes mesmo da sua estreia e está diretamente relacionado, não só com o *marketing* que é feito como também com a forma como cada um dos seus diversos componentes influencia o público.

Foi demonstrado que cerca de 53% das pessoas baseiam as suas escolhas cinematográficas em opiniões reveladas por outros espectadores (Karniouchina, 2010). Refira-se ainda, que um estudo recente que analisou *tweets* relacionados com alguns filmes, semanas antes da sua estreia, defende que esta popularidade gerada à volta de uma determinada produção é o principal fator para prever a sua rentabilidade (Redy, Kasat, & Jain, 2012).

São várias as variáveis constituintes de um filme que podem ser fatores essenciais para o seu posterior sucesso. Saber quais as mais relevantes e com maior peso na escolha do espectador, pode constituir uma mais-valia para os estúdios, na medida em que permitirá não só valorizar mais os seus filmes como também torná-los mais rentáveis. O aspeto fundamental e crítico para que esta indústria tenha sucesso e consiga gerar receitas é a audiência pois, sem esta, o negócio cinematográfico não seria sustentável nem vingaria no meio económico.

Face ao exposto, pretende-se com este estudo prever a rentabilidade de um filme, com base nessas mesmas variáveis e na influência que cada uma tem para o espectador na hora da sua decisão.

De modo a atingir o objetivo principal deste trabalho são exibidos de seguida os objetivos específicos para que a previsão seja bem-sucedida:

- Classificar os filmes mais lucrativos com base em diferentes variáveis;
- Identificar as variáveis que mais contribuem para o sucesso de um filme, eliminando as menos relevantes;
- Aplicar métodos preditivos - redes neuronais MLP, regressão e árvore de decisão - para estimar qual o valor ou classe final do lucro de um filme.

2. Revisão de literatura

Aliada ao grande desenvolvimento computacional que presenciamos hoje em dia, surgiu a urgente necessidade de obtenção de informação através de grandes quantidades de dados brutos. Este notório crescimento do volume de dados disponível é resultado da informatização da nossa sociedade e da rápida evolução de poderosas ferramentas para o seu armazenamento (Han, Kamber, & Pei, 2012).

A área do *Machine Learning* é uma das mais relevantes para o estudo preditivo e está em constante evolução. Ao longo do tempo foi-se tornando numa das ferramentas mais populares e utilizadas aquando da necessidade de extrair informações através de um grande conjunto de dados (Shalev-Shwartz & Ben-David, 2014).

Inicialmente focava-se apenas no reconhecimento de padrões, mas rapidamente escalou até ao patamar da inteligência artificial, automatizando o desenvolvimento de modelos analíticos baseados em operações lógicas ou binárias. Com esta ciência, as máquinas aprendem uma determinada tarefa baseando-se em exemplos extraídos a partir de dados históricos e potenciam uma rápida autonomia na classificação e previsão de grandes conjuntos de registos. A aprendizagem automatizada incide sobre pequenas regras, facilmente compreendidas, que tentam refletir as escolhas e decisões do raciocínio humano a uma máquina que irá estabelecer essas condições e abordagens autonomamente (Michie, Spiegelhalter, & Taylor, 1994).

Sendo considerado um dos principais instrumentos para o tratamento de dados, e incluído dentro das ferramentas de *Machine Learning*, o processo de *Data Mining* é atualmente sinónimo de maior competitividade e flexibilidade para responder às necessidades das grandes empresas mundiais. Através desta análise é possível descobrirem-se padrões e tendências que rapidamente se tornarão num pilar imprescindível para que esta indústria consiga vingar nos dias de hoje (García, Ventura, & Romero, 2007).

Atualmente os modelos preditivos já foram utilizados para os mais diversos objetivos tais como prever a falência de instituições bancárias, retornos económicos, tipo de cancro, transformações geográficas ou mesmo compra e venda de ações na bolsa de valores (Khan et al., 2001; Kimoto, Asakawa, Yoda, & Takeoka, 1990; Pijanowski, Brown, Shellito, & Manik, 2002; White, 1988; Zhang, Hu, Patuwo, & Indro, 1999).

São vários os estudos feitos com o objetivo de obter as melhores previsões para um orçamento mais viável e que envolva o mínimo de prejuízo possível, tendo como base as receitas cinematográficas (Fazzion, Casas, Gonçalves, Melo-Minardi, & Meira, 2013; Hunter, Smith, & Singh, 2016; Im & Nguyen, 2011; McKenzie, 2013; Song & Han, 2013).

O sucesso desta indústria está diretamente relacionado com o êxito de bilheteira que se reflete no lucro do distribuidor cinematográfico como principal indicador do triunfo comercial de um determinado filme (Wallace, Seigerman, & Holbrook, 1993), sendo que a semana de estreia é responsável por cerca de 40% do total de lucro de bilheteira (Einav, 2007). É considerado um negócio com uma grande variação na popularidade de cada produção, sendo que um único filme pode significar a diferença entre milhões de dólares de lucro ou de prejuízo para um estúdio (Simonoff & Sparrow, 2000).

Litman (1983) foi o grande pioneiro neste campo de pesquisa através da elaboração de um modelo de regressão múltipla capaz de prever o sucesso financeiro de um filme e concluiu que, de entre muitas variáveis, existiam algumas que desempenhavam um papel fundamental nessa previsão como o orçamento, a opinião dos críticos, género e prémios ganhos. Em 1989, com o auxílio de Kohl, Litman prosseguiu o seu estudo, recorrendo a novas variáveis como país de origem, sequela e concorrência direta no mercado. No entanto esta variação afetou a taxa de sucesso reduzindo-a face ao estudo anterior. Seguidamente Simonoff & Sparrow (2000), acrescentando algumas variáveis específicas, concluíram que existia uma grande relação entre o fator lucrativo do filme e as nomeações que arrecadou para a cerimónia dos Óscares nas categorias principais.

Em 2005 foi elaborado um estudo que desenvolveu uma nova abordagem face à variável dependente. Chang & Ki testaram três modelos de regressão distintos, com a variável dependente a variar entre o lucro de bilheteira total, lucro de bilheteira no primeiro fim-de-semana de estreia e o total de semanas em exibição. Concluíram que o segundo modelo era o mais preciso e o que apresentava uma taxa de sucesso superior (61.1%).

Delen & Sharda (2006) desenvolveram o mesmo estudo recorrendo, desta vez, a uma metodologia diferente: redes neuronais. Nesta abordagem o problema preditivo foi convertido num problema de classificação, sendo feita a divisão dos filmes em nove categorias distintas de acordo com as suas receitas de bilheteira, tendo a taxa de sucesso atingido apenas os 36.9%. Este valor foi melhorado em 2009, pelos mesmos autores, que desenvolveram outra investigação onde testaram vários modelos de previsão para além das redes neuronais, incluindo as árvores de decisão e uma amostra de filmes mais abrangente. O grande desafio desta pesquisa foi a utilização do método de fusão de informação, que combina duas ou mais metodologias de previsão de forma a gerar resultados mais precisos, tendo com isto aumentado a taxa de sucesso para 56.07%, percentagem muito superior à obtida anteriormente.

A crescente necessidade de aprofundar este tema motivou a existência de novos estudos, que permitiram um aumento exponencial da previsão do sucesso financeiro dos filmes, com a maior percentagem a fixar-se nos 94.1% (Ghiassi et al., 2015).

AUTORES	**METODOLOGIA (TAXA DE SUCESSO CORRESPONDENTE)**
(Litman, 1983)	Modelo de Regressão **(48.5%)**
(Litman & Kohl, 1989)	Modelo de Regressão **(30%)**
(Simonoff & Sparrow, 2000)	Modelo de Regressão **(50.6%)**
(Chang & Ki, 2005)	Modelo de Regressão **(61.1%)**
(Sharda & Delen, 2006)	Redes Neuronais (MLP) **(36.9%)**, Modelo de Regressão **(30.17%)**, Árvores de Decisão (CART) **(31.18%)**, Análise Discriminativa **(29.25%)**
(Terry, Cooley, & Zachary, 2008)	Modelo de Regressão **(74.8%)**
(Sharda & Delen, 2009)	Redes Neuronais(MLP) **(52.60%)**, Árvores de Decisão **(40.46%)**, SVM **(55.49%)**, *Random Forest* **(54.62%)**, *Boosted Trees* **(54.05%)**, Fusão de Informação **(56.07%)**
(L. Zhang, Luo, & Yang, 2009)	Redes Neuronais (MLBP) **(68.1%)**
(Im & Nguyen, 2011)	Modelo de Regressão **(42.9%)**
(Deniz & Hasbrouck, 2012)	Modelo de Regressão **(59.8%)**

(Song & Han, 2013)	Regressão Linear **(58.8%)**, *Random Forest* **(53.7%)**, *Gradient Boosting* **(60.4%)**
(Kaur & Nidhi, 2013)	Redes Neuronais **(93.3%)**
(Nithin, Pranav, Sarath, & Lijiya, 2014)	Regressão Linear **(50.7%)**, Regressão logística **(42.2%)**, SVM **(39%)**
(Mehta, Bhatt, & Desai, 2015)	Árvores de Decisão (ID3) **(66.67%)**
(Ghiassi et al., 2015)	Redes Neuronais (DAN2) **(94.1%)**
(Riwinoto, M.T., Zega, & Irlanda, 2015)	Redes Neuronais **(58.1%)**
(Rhee & Zulkernine, 2016)	Redes neuronais (MLBP) **(88.8%)**, SVM **(84.2%)**

Tabela 1 - Metodologias utilizadas em estudos anteriores

A maioria das variáveis foi adotada de estudos anteriores (Ghiassi et al., 2015; Sharda & Delen, 2006) na sua forma contínua, de modo a melhorar a sensibilidade dos métodos preditivos.

Considera-se da maior utilidade enunciar uma breve descrição de cada uma das variáveis independentes integradas nesta análise, enquadrando a respetiva importância segundo os autores que se consideram mais relevantes.

A escolha recaiu não só nas variáveis presentes nos estudos mais relevantes e com maior taxa de sucesso (sequela, ator, realizador, orçamento, nomeações, género, época, faixa etária, óscares), mas também noutras pouco ou nada utilizadas na revisão bibliográfica (críticos, duração, prémios, espectadores), de modo a que possa ser avaliado o seu impacto económico.

Sequela

As sequelas representam atualmente uma estratégia cada vez mais relevante na introdução de novos filmes na cultura cinéfila dos espectadores, através da capitalização do sucesso das produções anteriores. A probabilidade de redução do risco económico é o grande aliciante para que seja cada vez maior o número de estúdios a apostar em sequelas como mecanismo de sucesso no mercado cinematográfico.

Alguns estudos concluíram que as sequelas contribuem para um aumento do lucro de bilheteira (Ravid, 1999; Simonoff & Sparrow, 2000), o que pode também ser comprovado através de recentes pesquisas na área: os dados anuais relativos ao filme com mais lucro de bilheteira revelam que desde 1989, num total de 27 filmes, 17 são sequelas, o que representa a grande influência desta variável na rentabilidade de bilheteira (Mojo, 2016b).

MPAA

A *Motion Picture Association of America* (MPAA) é a entidade responsável pelo sistema de classificação de filmes por faixa etária no continente Americano. A sua organização baseia-se em 5 categorias distintas: G (acessível a todos), PG (Controlo Parental), PG-13 (Controlo Parental a partir dos 13 anos de idade), R (Restrito) e NC-17 (Sem crianças abaixo dos 17 anos de idade).

A classificação de um filme é um fator relevante para o seu sucesso, uma vez que certas categorias não são muito populares entre os espectadores e podem ser motivo de prejuízo no sucesso financeiro de uma nova produção.

Um estudo efetuado por Sawhney and Eliashberg (1996) revela que os filmes com uma classificação MPAA restrita irão obter menores rendimentos em bilheteira do que os restantes, conclusão também retirada por outros autores, como Sochay (1994).

Por outro lado, Ravid (1999) comprovou, recorrendo a um modelo de regressão linear, que as duas classificações que mais impacto positivo têm sobre o lucro de uma produção são o G e o PG.

Género

São vários os estudos que comprovam a influência do género cinematográfico no rendimento de bilheteira (Simonoff & Sparrow, 2000; Vany & Walls, 2002; (W. David Walls, 2005) W. David Walls, 2005).

Certos autores defendem que o único género relevante que demonstra uma relação positiva é o de ficção científica (Litman, 1983), enquanto que outros afirmam, através das suas pesquisas, que os filmes de ação e *thriller* são os mais significativos e populares para os espectadores na hora da sua decisão. Os filmes românticos são os menos requeridos (Neelamegham & Chintagunta, 1999).

Orçamento

O orçamento de um filme está fortemente correlacionado com a previsibilidade do lucro de bilheteira (Deniz & Hasbrouck, 2012). Contudo, alguns autores concluíram que o nível das receitas não corresponde necessariamente ao retorno pretendido (John, Ravid, & Sunder, 2003). Esta variável é das decisões mais controversas para os estúdios cinematográficos devido à inexistência de uma "receita" correta para o sucesso de um filme, uma vez que existem bastantes exemplos de filmes com pequenos orçamentos que se tornaram financeiramente bem-sucedidos nas grandes salas de cinema. Litman (1983) argumenta que grandes orçamentos refletem maior qualidade e maior popularidade nas bilheteiras e garante que, ainda associado ao orçamento de produção de um filme, estão também os salários excessivos para estrelas, significativos atrasos de produção ou gestão ineficiente, o que inflaciona os valores finais que não traduzem, na sua totalidade, a qualidade da produção.

Óscares, Prémios e Nomeações

Os prémios da academia representam um importante e prestigioso marco para o cinema mundial. Uma produção vencedora de um óscar é sinónimo do aumento do lucro do mesmo em bilheteira. É importante compreender que a distinção monetária destes prémios não se limita aos filmes vencedores, sendo apenas a sua nomeação suficiente para gerar receitas muito superiores aos restantes (Deuchert, Adjamah, & Pauly, 2005).

Um dos primeiros estudos sobre a influência comercial desta variável na rentabilidade de um filme foi realizado por Litman (1983), que conseguiu demostrar que uma nomeação ao óscar nas categorias de melhor ator/atriz e melhor filme acrescentavam, a nível económico, um lucro de cerca de 7,34 milhões de dólares, e a conquista desse prémio valeria mais de 16 milhões de dólares. Este estudo foi apoiado, anos mais tarde, por outros autores (Nelson, Donihue, Waldman, & Wheaton, 2011).

Representando um impacto extremamente positivo nas receitas, é compreensível que a maioria dos grandes estúdios aposte fortemente em filmes capazes de integrarem a lista dos candidatos a estes prémios tão significativos e benéficos para o seu próprio sucesso e reconhecimento**.**

Estrelas (Ator e Realizador)

O fenómeno "estrela" que um ator atinge pela sua popularidade em Hollywood parece favorecer as produções. Atualmente, segundo alguns autores, é uma importante variável para a rentabilidade das mesmas (Prag & Casavant, 1994; Wallace et al., 1993). O custo associado para integrar uma "estrela" no elenco é bastante elevado, porém, poderá trazer lucros muito favoráveis a longo prazo. Estima- se que o impacto médio de um ator de relevo seja um aumento de 6,5 milhões de dólares no lucro final (W. D. Walls, 2009)(W. D. Walls, 2009).

Sochay (1994) concluiu que a presença de estrelas no elenco de um filme é sinónimo de uma maior lucratividade do filme em questão, defendendo que promover um ator bastante conhecido pelos espectadores é mais fácil do que promover um outro que seja desconhecido do mundo cinéfilo.

Época festiva

A data de estreia de um filme é uma variável influenciadora do lucro de bilheteira, uma vez que a afluência dos espectadores aos cinemas aumenta significativamente em férias ou épocas festivas. Foram vários os autores que comprovaram a sua relação direta com o lucro cinematográfico (Basuroy, Chatterjee, & Ravid, 2003; Chang & Ki, 2005; Simonoff & Sparrow, 2000).

Litman (1983) considerou que existiam três picos de audiências ao longo do ano: Novembro/Dezembro (Natal), Março/Abril (Páscoa) e Junho/Julho/Agosto (Verão). Defendeu também que a altura mais favorável para a estreia de um filme seria a época natalícia, enquanto que Sochay (1994) comprovou que seria no verão que os lucros cinematográficos atingiriam o seu auge.

Espectadores e críticos

As avaliações recebidas antes da visualização de um filme são sempre importantes para o espectador, sejam dadas por outros espectadores ou sejam atribuídas por críticos profissionais. Apesar de serem distintas, estas duas variáveis estão bastante relacionadas. Os "críticos amadores", ou seja, os espectadores comuns formam uma opinião após a visualização de um filme e partilham-na com um determinado grupo de pessoas informalmente, enquanto que os críticos profissionais, que se dedicam exclusivamente à visualização de novas produções e formam uma crítica muito mais complexa e fundamentada, partilham-na a nível mundial.

Para a indústria cinematográfica esta aposta é uma mais valia em termos de *marketing*, uma vez que possibilita a previsão, a longo prazo, do sucesso ou insucesso de um filme (Chakravarty, Liu, & Mazumdar, 2010; Duan, Gu, & Whinston, 2008).

Nos Estados Unidos da América cerca de um terço da população confia totalmente na opinião de outras pessoas antes da sua decisão de ver ou não uma determinada produção (Basuroy, Chatterjee, & Ravid, 2003). Estes autores analisaram algumas questões relacionadas com o papel influenciador da opinião de outros, entre as quais o papel da crítica, no sucesso económico das bilheteiras traduzido na diferença do impacto entre opiniões positivas e opiniões negativas.

As suas conclusões recaíram numa clara influência dos críticos sobre o lucro de bilheteira e num maior impacto das opiniões negativas sobre um filme por comparação com as positivas.

Em 1997, Eliashberg e Shugan propuseram uma inovadora teoria que consiste no facto de existirem dois tipos diferentes de críticos: os influenciadores e os previsores. Os primeiros são pessoas que são consideradas influentes num determinado grupo ou meio sobre um assunto em particular, neste caso, cinema. Desta forma, todos os elementos do grupo definem como verdadeira a opinião do influenciador e através dela seguem padrões comportamentais. Por outro lado, um previsor constrói a sua opinião com base em técnicas preditivas muito mais credíveis.

Na tabela seguinte (tabela 2) estão representadas as variáveis utilizadas pelos estudos mais relevantes nesta área, assim como a respetiva taxa de sucesso.

AUTORES	Género	MPAA	Estrelas	Orçamento	Distribuidor	Data	Nomeações	Óscar ganho	Críticos	País	Sequela	Nº Ecrãs	Espectadores	Concorrência	E. Especiais	Marketing	TAXA DE SUCESSO
(Litman, 1983)	x	x	x	x	x	x	x	x	x								**48.5%**
(Litman & Kohl, 1989)	x	x	x	x	x	x			x	x	x	x				x	**30.0%**
(Simonoff & Sparrow, 2000)	x	x	x	x		x	x	x	x	x	x	x					**50.6%**
(Chang & Ki, 2005)	x	x	x	x	x	x			x		x	x	x				**61.1%**
(Sharda & Delen, 2006)	x	x	x								x	x		x	x		**36.9%**
(Terry et al., 2008)	x	x		x				x	x		x						**74.8%**
(Sharda & Delen, 2009)	x	x	x								x	x		x	x		**56.1%**
(L. Zhang et al., 2009)	x		x			x				x		x		x		x	**68.1%**
(Im & Nguyen, 2011)	x	x		x	x	x						x	x				**42.9%**
(Deniz & Hasbrouck, 2012)	x	x	x	x		x	x	x									**59.8%**
(Song & Han, 2013)	x		x		x	x					x			x			**60.4%**
(Kaur & Nidhi, 2013)			x	x	x	x										x	**93.3%**
(Nithin et al., 2014)	x		x	x									x				**50.7%**
(Mehta et al., 2015)	x								x				x				**66.7%**
(Ghiassi et al., 2015)	x	x	x								x	x		x	x		**94.1%**
(Riwinoto et al., 2015)	x	x	x								x	x		x	x		**58.1%**
(Rhee & Zulkernine, 2016)			x	x		x					x		x				**88.8%**

Tabela 2 - Variáveis utilizadas em estudos anteriores

No presente estudo proponho a criação de um sistema preditivo, recorrendo a redes neuronais, árvores de decisão e regressões, que consiga prever a rentabilidade de um filme, tendo como variável dependente os lucros mundiais de bilheteira e baseando-se em variáveis independentes distintas das maioritárias nas principais pesquisas sobre o tema (críticos, duração, prémios recebidos, classificação dos utilizados).

O seu principal objetivo é comparar a taxa de sucesso deste modelo face aos estudos já publicados.

Conseguir prever a lucratividade de uma produção consoante os diversos componentes que a envolvem trará vantagens para todos os intervenientes desta indústria sendo, portanto, uma necessidade emergente no atual mercado competitivo.

3. Metodologia

A metodologia utilizada neste projeto tem por nome SEMMA (*Sample, Explore, Modify, Model, Assess*), processo de *Data Mining* desenvolvido pelo SAS Institute (REF para SEMMA), responsável pela criação do software utilizado: SAS *Enterprise Miner*. É caracterizado por cinco etapas fundamentais, responsáveis por um melhor desempenho, simplificando e agilizando todas as etapas requeridas no modelo preditivo. É um esquema cíclico, como se pode verificar na representação seguinte (figura 1).

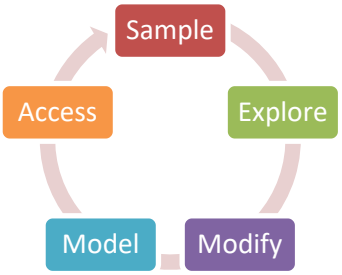

Figura 1 - Esquema da metodologia SEMMA

A figura 2 representa a metodologia utilizada neste estudo. A coleção de dados proveio de quatro fontes distintas – *Opus Data, IMDB, Mojo e Metacritic* – que foram posteriormente tratadas e complementadas entre si na ferramenta *Microsoft Office Excel 2016*. No *SAS Enterprise Miner* procedeu-se à análise gráfica e estrutural dos dados importados e consequentemente à redução de dimensionalidade das variáveis integrantes do modelo, que foi complementada com o tratamento de *outliers* e a utilização de partições distintas. Foram utilizadas três ferramentas preditivas – redes neuronais, árvore de decisão e regressão – que originaram vários modelos com taxas de erro distintas obtidas através da avaliação do modelo.

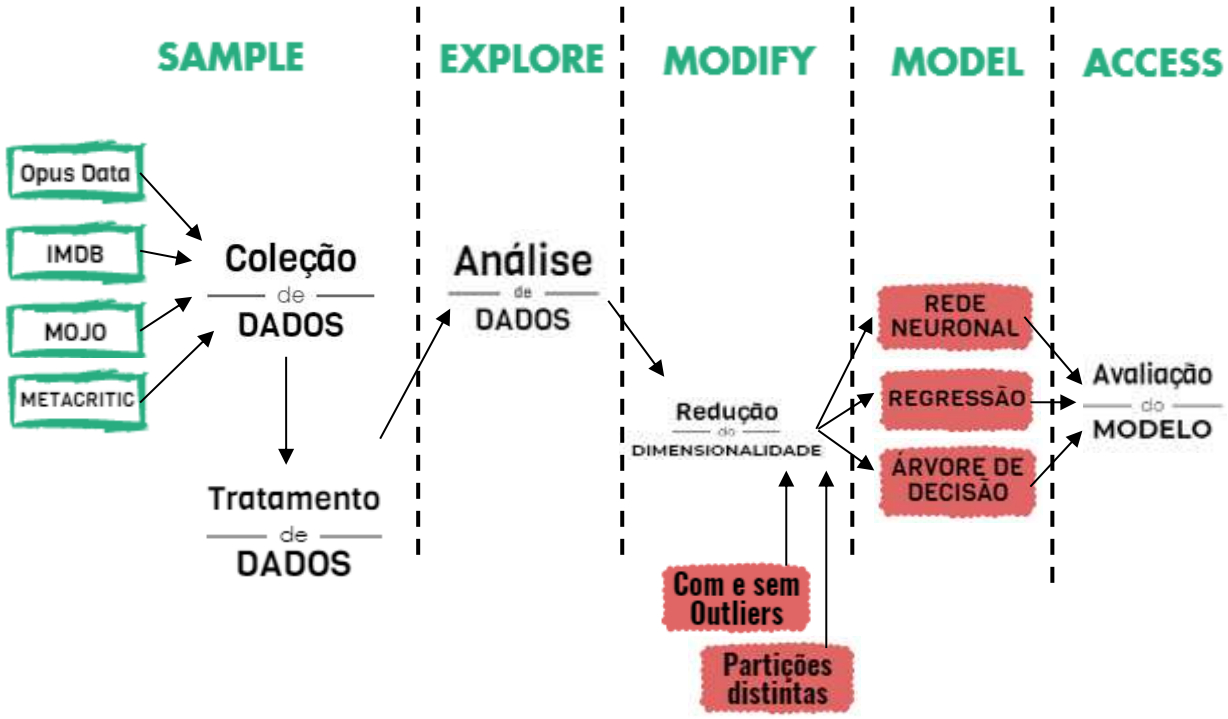

Figura 2 - Metodologia utilizada no estudo

3.1 *Sample*

A etapa da amostra é caracterizada pelo conjunto dos dados representativos da população. Engloba todo o processo executado para formar a base de dados utilizada no estudo, as variáveis que a compõem, a normalização a que foram sujeitas e por fim, o esquema partitivo utilizado no modelo.

3.1.1 *Data Collection*

O conjunto de dados utilizado neste projeto foi extraído através do *OpusData*, um serviço de dados pertencente à *Nash Information Services*, uma poderosa plataforma informática que se dedica ao fornecimento de dados e serviços analíticos na indústria cinematográfica para diversas entidades relacionadas.

Posteriormente estes dados foram complementados com informações em falta, recorrendo desta vez, a um dos maiores arquivos virtuais da atualidade, que inclui dados relativos a música, cinema, séries televisivas ou mesmo videojogos, *Internet Movie Database* (IMDB).

A base de dados utilizada neste modelo inclui 1920 filmes compreendidos entre o ano de 2000 e 2016, e diz respeito ao lucro de bilheteiras mundial, que inclui o mercado doméstico e o internacional.

Algumas das variáveis utilizadas não provieram da base de dados principal, sendo necessário o auxílio de outras plataformas digitais:

- A variável "Espectadores" foi obtida através de um *rating* criado pelo IMDB, que é calculado através da média ponderada de todos os votos efetuados pelos utilizadores do site;
- A variável "Críticos", foi obtida através do site *Metacritic*, que reúne todas as avaliações dos mais importantes críticos de filmes do mundo num só valor, que irá ilustrar a decisão global que caracteriza uma certa produção;
- As variáveis "Ator" e "Realizador", foram classificadas segundo a lista dos atores e realizadores mais lucrativos de sempre (Mojo, 2016a).

O pré-processamento de dados inclui duas fases distintas:

- A coleção de dados através de diversas plataformas com informações específicas para cada necessidade do modelo. A base principal assentou nos dados fornecidos pelo *OpusData* e as restantes serviram de complemento para variáveis ou dados em falta que, posteriormente irão fornecer grande valor preditivo.
- A base de dados principal, fornecida pelo *OpusData,* que se encontra em formato *Microsoft Office Excel* 2016, e completada manualmente com os dados das restantes plataformas.

O poderoso software de *Data Mining*, SAS *Enterprise Miner*, utilizado neste projeto, recebe os dados diretamente do *Excel* com o objetivo de serem trabalhados e analisados, originando informação e conhecimento prático para o futuro da indústria cinematográfica.

3.1.2 Variáveis

Como demonstrado no quadro abaixo, o modelo é constituído por 3 variáveis nominais ("Género", "Faixa Etária", "Época"), 3 variáveis binárias ("Sequela", "Ator", "Realizador") e 7 variáveis intervalares ("Prémios", "Óscares", "Orçamento", "Nomeações", "Espectadores", "Críticos", "Duração").

A variável dependente deste estudo, "Lucro", será tratada de forma distinta, com o objetivo de chegar a resultados mais precisos: na sua forma contínua adotará um tipo intervalar e na sua forma discreta um tipo ordinal e binário.

VARIÁVEL	TIPO DE MEDIDA	VALORES OBSERVADOS	MÉDIA	SIGNIFICADO
ATOR	**Binária**	**0 - 1**		**Variável binária que indica se o ator está entre os mais rentáveis (1) ou não (0)**
REALIZADOR	**Binária**	**0 - 1**		**Variável binária que indica se o realizador está entre os mais rentáveis (1) ou não (0)**
ESPECTADORES	**Intervalar**	**0 - 9**	**6.40151**	**Avaliação atribuída pelos espectadores ao filme**
CRÍTICOS	**Intervalar**	**9 - 98**	**52.28646**	**Avaliação atribuída pelos críticos ao filme**
DURAÇÃO	**Intervalar**	**0 - 219**	**109.7568**	**Duração do filme**
ÉPOCA	**Nominal**	**Natal, Normal, Páscoa, Verão**		**Período do ano em que o filme estreou**
FAIXA ETÁRIA	**Nominal**	**G, M/6, NC-17, PG, PG-12, PG-13, R**		**Classificação etária atribuída ao filme**
GÉNERO	**Nominal**	**Ação, Animação, Aventura, Biografia, Comédia, Crime, Documentário, Drama, Familiar, Fantasia, Ficção científica, História, Horror, Mistério, Musical, Thriller, Western**		**Género cinematográfico atribuído ao filme**
LUCRO	**Ordinal**	**1 - 9**	**6.054688**	**Variável dependente que atribui uma das 9 classes a cada filme consoante o lucro obtido em bilheteira**
	Intervalar	**87793-2.7839E9**	**1.5754E8**	**Variável dependente que representa o valor monetário (dólares) do lucro de bilheteira do filme**
	Binária	**0 - 1**		**Variável dependente binária que indica se o filme foi lucrativo (1) ou não (0)**
NOMEAÇÕES	**Intervalar**	**0 - 549**	**26.79271**	**Número de nomeações que o filme obteve**

ORÇAMENTO	**Intervalar**	**10350000 – 425000000**	**54839263**	**Valor monetário (dólares) do orçamento gasto com filme**
ÓSCARES	**Intervalar**	**0 - 11**	**0.140104**	**Número de óscares que o filme obteve**
PRÉMIOS	**Intervalar**	**0 - 234**	**8.152083**	**Número de prémios que o filme obteve**
SEQUELA	**Binária**	**0 - 1**		**Variável binária que indica se o filme é sequela (1) ou não (0)**

Tabela 3 - Variáveis utilizadas no modelo

Para que seja retirada informação mais clara e completa dos dados, e com o objetivo de tornar as variáveis mais estruturadas e relevantes para o modelo, procedeu-se a algumas transformações significativas no que diz respeito ao seu teor prático.

Relativamente às variáveis "Ator" e "Realizador", foi atribuído o grau de "estrela" aos primeiros 150 da tabela classificativa, isto é, aos atores ou realizadores economicamente mais influentes, foi conferido a constante "1" e aos restantes a constante "0", de forma facilitar o processo de influência das variáveis face ao lucro final do filme em bilheteira. Esta transformação binária foi também utilizada na variável "sequela" que toma o valor de "1" se o filme em causa tiver um argumento consequente de um outro e "0" caso contrário.

A variável dependente desta investigação é o lucro de bilheteira, que irá ter três representações distintas no modelo:

- A primeira baseia-se na metodologia de Sharda e Delen (2006, 2009), em que a variável contínua é transformada numa variável discreta, mais concretamente em 9 classes distintas como se pode comprovar na tabela 4. A classe mais reduzida (classe 1) representa os filmes que foram um fiasco nas salas de cinema, enquanto que os sucessos são exibidos pela classe mais elevada (classe 9);

- A segunda consiste na sua forma binária, onde tomará o valor "1" se o seu lucro de bilheteira for igual ou superior ao dobro do orçamento e "0" caso contrário; por fim, será também representada a nível intervalar, sendo que o objetivo se estabelece em determinar o valor exato do lucro de uma determinada produção.

CLASSE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Intervalo (milhões de dólares)	<1 (Fiasco)	[1,10]	[10,20]	[20,40]	[40,65]	[65,100]	[100,150]	[150,200]	>200 (Sucesso)

Tabela 4 - Receitas de bilheteiras Sharda e Delen (2006,2009)

Com este estudo propõe-se analisar e comparar os resultados destes três tipos de variável dependente, recorrendo a três metodologias preditivas distintas de forma a determinar qual deles é mais preciso na determinação e previsão do lucro associado.

3.1.3 Normalização

A normalização dos dados num modelo preditivo visa melhorar a precisão e eficácia dos registos nele englobados, atingindo assim resultados mais reais e corretos. Quando o valor de uma certa variável é apresentado em unidades menores, existirá inevitavelmente um aumento significativo do seu alcance e valor para o modelo, igualando o peso entre as diversas variáveis.

Existem vários métodos utilizados para proceder a uma boa normalização sendo os mais conhecidos a normalização *min-max* e a normalização *z-score*, a qual foi utilizada neste estudo.

A normalização *z-score* é uma das mais populares na estandardização dos dados de um modelo e baseia-se no valor da média e desvio-padrão de uma amostra, ou seja, um determinado valor específico é normalizado com base nessas duas medidas estatísticas.

$$\text{zscore} = \frac{x - \mu}{\delta}$$

Onde:
x – Valor
μ – Média da população
δ – Desvio-Padrão da população

Todas as variáveis intervalares presentes neste modelo foram devidamente normalizadas, recorrendo ao método do z-score, antes de serem importadas para o *SAS Enterprise Miner*. Este tratamento normativo foi executado no Microsoft Office Excel, na fase de pré processamento.

3.1.4 *Data Partition*

A partição de dados é uma etapa essencial não só garantir a qualidade e eficácia do modelo como também para evitar a sua sobreaprendizagem, que consiste no fenómeno dos novos dados se adaptarem demasiado aos dados utilizados para treinar o modelo, impedindo assim uma boa classificação.

Ling e Li (1998), defendem que um dos processos a ser executado com vista a este objetivo consiste em dividir a base de dados em duas partes distintas - conjunto de treino e conjunto de teste -, desenvolver o modelo sobre o conjunto de treino e aplicá-lo igualmente sobre o conjunto de teste, com a finalidade de avaliar se os resultados obtidos traduzem uma correta apreciação.

Os dados do conjunto de treino são utilizados pelas ferramentas preditivas de forma a servirem de base à aprendizagem do modelo. Para que essa aprendizagem seja executada de forma correta e bem-sucedida é necessário avaliar o desempenho do classificador, neste caso da rede neuronal, árvore de decisão ou regressão, num conjunto de dados independente do inicial (conjunto de teste), de modo a ser possível calcular a taxa de erro do modelo preditivo.

Neste estudo foi também utilizado um conjunto de dados intermédio denominado de "conjunto de validação", que permite otimizar os parâmetros do modelo antes de ser utilizado no conjunto de teste.

Com o intuito de agilizar o processo e retirar do modelo a previsão mais correta foram testadas duas partições distintas:

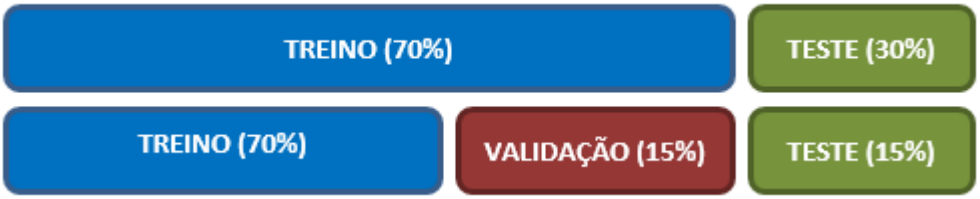

Figura 3 - Partições utilizadas no modelo preditivo

Em termos práticos, os dados do *data set* inicial foram divididos da seguinte forma:

Partição	Tipo	Nº Observações
70/15/15	Treino	1337
	Validação	288
	Teste	295
70/30	Treino	1337
	Teste	583

Tabela 5 - Divisão das observações do Data Set

3.2 *Explore*

A segunda etapa consiste na exploração dos dados, onde é feita uma caracterização das variáveis integradas no estudo e a redução de dimensionalidade a que irão ser sujeitas para benefício de uma melhor *performance* por parte do modelo preditivo.

3.2.1 Análise exploratória

Como já foi referido anteriormente, a variável dependente, lucro cinematográfico, irá ser apresentada de três formas distintas, com o intuito de maximizar e melhorar a futura previsão. O gráfico seguinte (gráfico 1) representa a variável dependente multi-classe após a sua categorização e como é possível observar, dentro da amostra utilizada neste modelo, a classe mais numerosa é a 9 (lucros superiores a 200 milhões de dólares) representando cerca de 22.6% da amostra total, enquanto que a menos popular é a classe 1 (menos de 1 milhão de dólares). Podemos afirmar que é uma amostra bastante diversificada, o que trará apenas vantagem na análise e previsão a que este estudo se propõe.

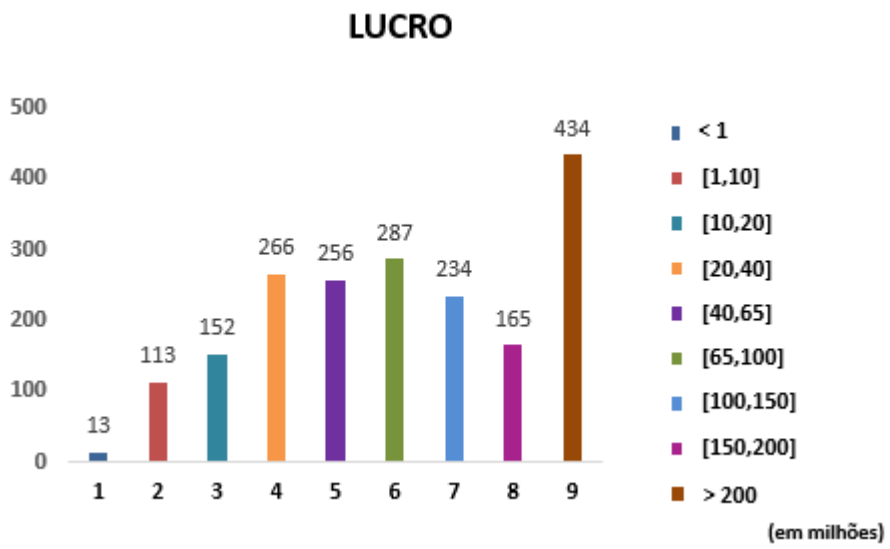

Gráfico 1- Variável dependente multi-classe

Na sua natureza binária, a variável dependente, irá dividir-se em filmes lucrativos e filmes não lucrativos. Esta classificação é atribuída através de um cálculo matemático que relaciona o lucro cinematográfico com o orçamento despendido pela produtora.

Não se preconiza como correta a avaliação da lucratividade de um filme tendo em conta apenas o valor monetário conseguido em bilheteira, devido ao facto de existirem muito mais variáveis em jogo no que diz respeito às receitas globais que uma produção pode arrecadar, como o *marketing*, eventos ou mesmo os próprios *DVD's* que começam a circular momentos após o fim da sua exibição. Porém, neste caso, estando apenas a estudar o efeito no lucro de bilheteira podemos relacioná-lo com o orçamento gasto e chegar a conclusões lucrativas ou mesmo desastrosas para o estúdio responsável, tendo apenas em conta esta variável.

Desta forma, e sabendo que um filme não se torna particularmente rentável se o valor de bilheteira exceder apenas uns dólares do orçamento, estipulou-se que um filme seria considerado lucrativo, caso o lucro de bilheteira tomasse valores iguais ou superiores ao dobro do orçamento: se essa inequação se verificar o lucro toma o valor de "1", caso contrário é representado por um "0".

É possível comprovar pela visualização do gráfico 2, que a amostra cinéfila selecionada é representada por uma frequência significativa de filmes, que seguindo o critério acima descrito, é considerada lucrativa: cerca de 64% do total de produções que se encontram englobadas neste estudo apresentam um lucro igual ou superior ao dobro do orçamento.

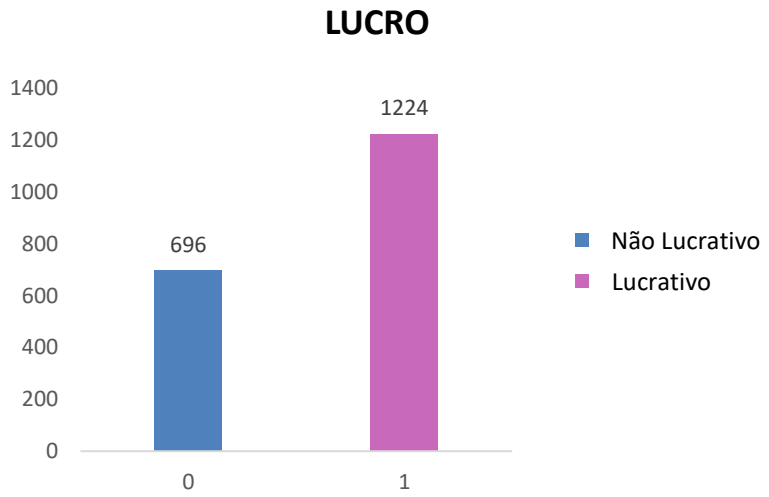

Gráfico 2 - Variável dependente binária

Para a análise comparativa das variáveis recorreu-se aos gráficos bivariados, devido à sua grande utilidade na análise estatística, que permite determinar a relação empírica que existe entre uma qualquer variável em estudo e a variável dependente, sendo assim possível verificar o quão explicativa é para o modelo.

No seguimento da sua maior simplicidade e clareza nesta análise foi utilizada a variável dependente binária, tornando-se assim o pilar fundamental nas análises que se seguem. Saber se uma determinada variável é um fator influenciador para o lucro de bilheteira é um marco essencial e um dos objetivos deste estudo, sendo analisadas apenas as que maior relevo e importância apresentam para a variável independente e consequentemente para o modelo.

Como é possível analisar na imagem seguinte (gráfico 3), existem alguns géneros que contribuem para uma maior afluência dos espectadores às salas de cinema. Foi calculado o rácio de filmes lucrativos face ao total de filmes dentro de cada género, de forma a ser possível fazer uma análise mais correta e perceber a influência que cada um dos géneros tem no lucro de bilheteira. Os mais relevantes e que apresentam maior relação com os filmes lucrativos são “Animação” (65%), “Aventura” (61%), “Ficção científica” (58%) e “Horror” (76%), que é o que maior percentagem apresenta tendo em consideração o seu total no *data set*.

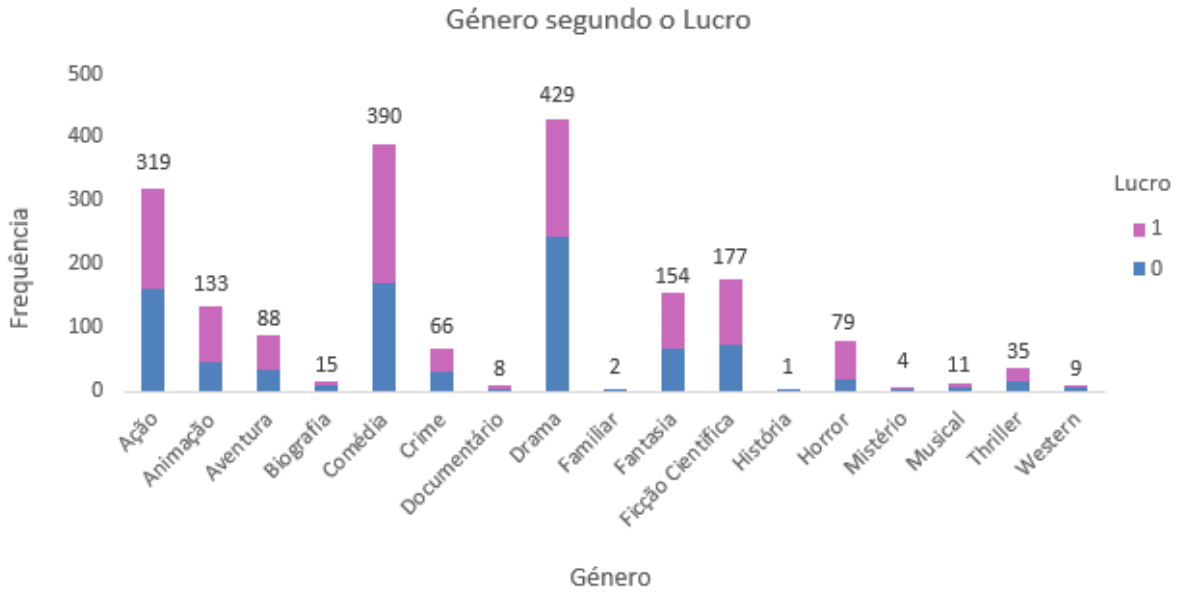

Gráfico 3 - Influência do Lucro na variável "Género"

Em relação à variável "Realizador" (gráfico 4), é possível comprovar a clara relação entre o sucesso cinéfilo de um filme e o realizador que o produz. Quando o realizador pertence à classe dos mais rentáveis (1), a percentagem de filmes lucrativos ronda os 67%, valor bastante relevante comparativamente aos apenas 47% na classe dos realizadores menos rentáveis (0). Podemos então concluir que um dos fatores que contribui para o lucro de bilheteira é o realizador que o produz.

No que diz respeito à variável "Sequela" (gráfico 5), existe igualmente um grande domínio de produções rentavelmente económicas. Cerca de 81% dos filmes incluídos neste estudo são sequelas (1) e apresentaram uma boa performance na semana de estreia, contra os 49% dos restantes que não o são.

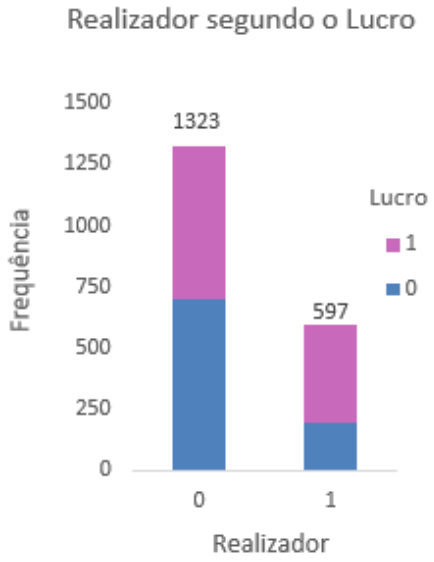

Gráfico 4 - Influência do Lucro na variável "Realizador"

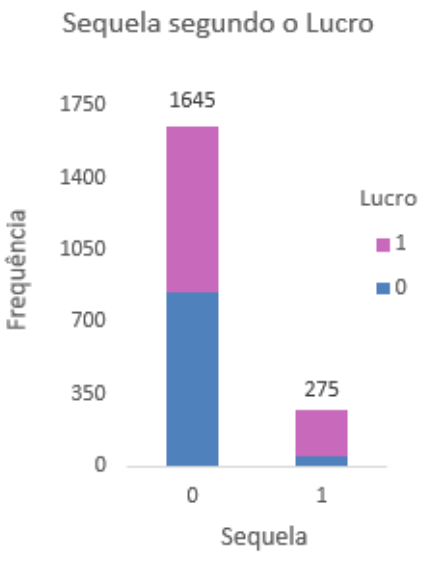

Gráfico 5 - Influência do Lucro na variável "Sequela"

3.3 *Modify*

Esta fase existe para que os registos e variáveis possam ser manipulados e transformados consoante o seu objetivo específico para o estudo, sendo que, como já foi referido, a variável dependente irá tomar três formas distintas: contínua, multi-classe e binária.

Esta variável foi introduzida, na sua forma intervalar, pelo que para ser possível a execução da análise nas suas outras duas configurações, se recorreu a código SAS, de forma a executar essas transformações.

3.3.1 Variável dependente discreta – multi-classe

O código SAS seguidamente representado (figura 4), transforma a variável dependente intervalar "Lucro" numa variável discreta constituída por 9 classes distintas, consoante o lucro que apresenta. Cada classe, representada na tabela 4, abrange um determinado intervalo lucrativo que permitirá a execução de um processo preditivo mais ágil.

A normalização dos parâmetros intervalares foi feita diretamente neste código através do método z-score, tendo sido utilizada a média e desvio padrão da população.

Training Code

```
DATA &EM_EXPORT_TRAIN;
   SET &EM_IMPORT_DATA;

media=157535793.9;
desvio_padrao=217889990.2;
If LUCRO<((1000000-media)/desvio_padrao) then LUCRO=1;
Else If LUCRO>=((1000000-media)/desvio_padrao) and LUCRO<((10000000-media)/desvio_padrao) then LUCRO=2;
Else If LUCRO>=((10000000-media)/desvio_padrao) and LUCRO<((20000000-media)/desvio_padrao) then LUCRO=3;
Else If LUCRO>=((20000000-media)/desvio_padrao) and LUCRO<((40000000-media)/desvio_padrao) then LUCRO=4;
Else If LUCRO>=((40000000-media)/desvio_padrao) and LUCRO<((65000000-media)/desvio_padrao) then LUCRO=5;
Else If LUCRO>=((65000000-media)/desvio_padrao) and LUCRO<((100000000-media)/desvio_padrao) then LUCRO=6;
Else If LUCRO>=((100000000-media)/desvio_padrao) and LUCRO<((150000000-media)/desvio_padrao) then LUCRO=7;
Else If LUCRO>=((15000000-media)/desvio_padrao) and LUCRO<((200000000-media)/desvio_padrao) then LUCRO=8;
Else LUCRO=9;
Run;
```

Figura 4 -Código SAS para variável dependente multi-classe

3.3.2 Variável dependente discreta – binária

O código SAS seguidamente representado na figura 5, transforma a variável dependente intervalar "Lucro" numa variável discreta binária.

Como já foi referido, o processo executado na sua transformação binária baseou-se na relação entre a receita cinéfila e o respetivo orçamento despendido, sendo um filme considerado lucrativo caso o lucro de bilheteira igualasse ou ultrapassasse o dobro do orçamento, representado pelo valor "1", e caso contrário por "0".

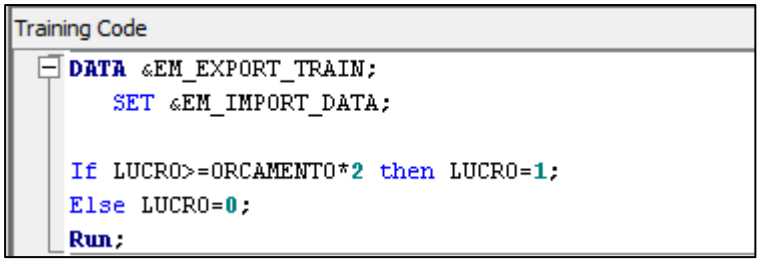

Figura 5 -Código SAS para variável dependente binária

3.3.3 *Outliers*

Algumas variáveis intervalares apresentam valores bastante homogéneos sendo que alguns deles foram considerados como *outliers*, definição para os dados que se encontram bastante distanciados dos restantes, e que poderão criar um grande impacto na interpretação dos resultados.

Neste caso específico um *outlier* pode representar filmes com uma grande quantidade de prémios arrecadados ou mesmo um orçamento elevadíssimo. Desta forma, optou-se por testar o modelo com e sem *outliers*, uma vez que estes, dada a sua natureza explicativa, poderiam trazer valor preditivo para o estudo em si.

3.2.2 Redução da Dimensionalidade

A complexidade que o modelo apresenta pode ser prejudicial a uma fácil e clara compreensão da previsão requerida. De forma a evitar que esta situação se verifique é necessário proceder a uma redução das variáveis em estudo, sendo que as menos relevantes e com menor importância para a variável dependente (lucro), serão descartadas em prol de uma análise mais sintetizada e menos complexa.

É imprescindível que as variáveis explicativas do modelo se encontrem correlacionadas com a variável dependente e não excessivamente correlacionadas entre si.

Alguns autores defendem que, para as variáveis mostrarem um grau de relevância elevado, é necessário que os seus valores variem com as diferentes classes (Gennari, Langley, & Fisher, 1989). Deste modo consegue-se evitar a redundância e reduzir a dimensionalidade do modelo preditivo.

Com o objetivo de tornar o modelo mais relevante na sua missão preditiva recorreu-se a três técnicas distintas para a eliminação de variáveis:

- **Chi-quadrado e coeficiente de determinação**

É necessário e fundamental observar a distribuição estatística de cada uma das variáveis. No que diz respeito à variável dependente discreta, o qui-quadrado (X^2) é um teste numérico que mede o desvio da distribuição esperada, sendo que as variáveis estudadas são independentes do valor da classe, isto é, existe um estudo das variáveis independentes face à variável dependente, que neste caso é o valor da receita de bilheteira (Ikran & Cherukuri, 2016).

Yang e Pedersen (1997) defenderam que este teste tem uma *performance* bastante elevada no que diz respeito a dados multi-classe, que traduz um dos problemas em estudo, uma vez que, uma das variáveis dependentes é constituída por 9 classes distintas.

Quando se lida com uma variável dependente contínua o teste do chi-quadrado deixa de ser relevante e apropriado, dando então lugar ao coeficiente de determinação (R^2), que se caracteriza por ser uma medida estatística responsável pela percentagem de ajustamento que a variável apresenta face ao modelo linear.

- **Matriz de Correlação de *Spearman* e *Pearson***

É imprescindível analisar a matriz de correlação, evitando assim integrar variáveis altamente correlacionadas entre si na análise. O coeficiente de correlação de *Spearman* é uma estatística de classificação não paramétrica, caracterizada por ser a mais adequada em dados ordinais, ao contrário do que acontece com o método de correlação de *Pearson* que, não só é uma estatística de classificação paramétrica, como também é apenas aplicável a dados intervalares (Chok, 2010).

Duas variáveis são classificadas como fortemente correlacionadas se o seu coeficiente de correlação for equivalente ou superior a 0.75 (Marôco, 2014). Assim, quanto maior for esse coeficiente, mais forte é a associação entre duas variáveis.

- **Coeficiente de regressão**

Através da análise dos *p-values*, é possível rejeitar variáveis cujo coeficiente de regressão não seja significativo para o modelo. O valor aceitável para esta medida assenta nos 0.05, nível de significância usual, o que significa que todas as variáveis que apresentem um registo superior serão estatisticamente insignificantes e dispensáveis para modelo.

De forma a ser obtida a melhor previsão do modelo foram testadas múltiplas combinações práticas, não só através do tratamento ou não de *outliers*, como também da utilização das três técnicas distintas, acima referidas, para a eliminação de variáveis prescindíveis em termos de desempenho preditivo para modelo.

3.2.2.1 Variável dependente discreta – multi-classe

Primeiramente foi executada uma redução de dimensão do modelo, cuja variável dependente é proveniente do estudo de *Sharda* e *Delen* (2006, 2009), categorizada em nove classes distintas.

O gráfico de importância, ferramenta utilizada nesta análise, ordena as diferentes variáveis de acordo com o seu peso na previsão da variável dependente através do índice *Gini*, característico do algoritmo CART e utilizado em árvores de decisão. Desta forma foi descartada a variável "**Ator**" por não apresentar grande valor significativo para o lucro cinéfilo.

- **Com *Outliers* (qui-quadrado)**

Este modelo preditivo multi-classe (nove classes monetárias) contém *outliers* e recorre ao teste estatístico qui-quadrado para a eliminação de variáveis pouco significativas para o modelo.

Excluíram-se as variáveis que apresentavam valores mais reduzidos quando submetidas ao teste estatístico do qui-quadrado: "**Duração**", "**Óscares**" e "**Sequela**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (*Spearman*)**

Este modelo preditivo multi-classe (nove classes monetárias) contém *outliers* e recorre à matriz de correlação de *Spearman* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar na tabela 6, as variáveis "Espectadores" e "Críticos" encontram-se fortemente correlacionadas apresentando um coeficiente de correlação de aproximadamente 0.76, assim como as variáveis "Prémios" e "Nomeações", que fixaram o seu valor em aproximadamente 0.88.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CRITICOS (1)	1						
DURAÇÃO (2)	0.348509	1					
ESPECTADORES (3)	0.762796	0.467909	1				
NOMEAÇÕES (4)	0.641934	0.388765	0.608807	1			
ORÇAMENTO (5)	0.026558	0.219425	0.058304	0.247598	1		
OSCARES (6)	0.344774	0.204209	0.332884	0.398546	0.054428	1	
PRÉMIOS (7)	0.601566	0.368661	0.583729	0.876307	0.168358	0.406179	1

Tabela 6 - Matriz de correlação de *Spearman* relativa ao lucro categorizado (com *outliers*)

Ao recorrer ao gráfico de importância das variáveis (gráfico 6), optou-se por prescindir das variáveis "**Críticos**" e "**Prémios**" por apresentarem um grau de importância inferior comparativamente com aquelas com as quais possuíam uma grande correlação ("Espectadores" e "Nomeações" respetivamente).

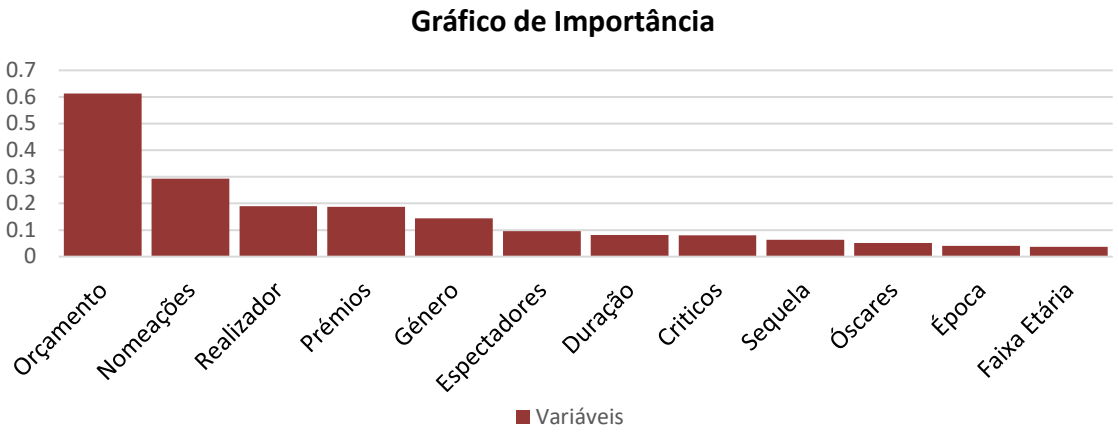

Gráfico 6 - Gráfico de importância relativo ao lucro multi-classe (com *outliers*)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (coeficiente de regressão)**

Este modelo preditivo multi-classe (nove classes monetárias) contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 7, existem três variáveis ("**Críticos**", "**Óscares**" e "**Prémios**"), cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	*P-Value*
Críticos	**0.1193**
Duração	0.0004
Época	0.0388
Espectadores	<0.0001
Faixa Etária	0.0004
Género	<0.0001
Nomeações	<0.0001
Orçamento	<0.0001
Óscares	**0.9545**
Prémios	**0.1649**
Realizador	<0.0001
Sequela	<0.0001

Variáveis rejeitadas

Tabela 7 - Coeficientes de regressão relativos ao lucro categorizado (com *outliers*)

- **Sem *Outliers* (qui-quadrado)**

Este modelo preditivo multi-classe (nove classes monetárias), não contém *outliers* e recorre ao teste estatístico qui-quadrado para a eliminação de variáveis pouco significativas para o modelo.

Foram excluídas as variáveis que apresentavam valores mais reduzidos quando submetidas ao teste estatístico do qui-quadrado: "**Faixa Etária**" e "**Óscares**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (*Spearman*)**

Este modelo preditivo multi-classe (nove classes monetárias) não contém *outliers* e recorre à matriz de correlação de *Spearman* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar na tabela 8, as variáveis "Espectadores" e "Críticos" encontram-se fortemente correlacionadas com um coeficiente de correlação de aproximadamente 0.75, assim como as variáveis "Prémios" e "Nomeações" que fixaram o seu valor em aproximadamente 0.87.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CRITICOS (1)	1						
DURAÇÃO (2)	0.288397	1					
ESPECTADORES (3)	0.745589	0.436431	1				
NOMEAÇÕES (4)	0.626113	0.344992	0.576404	1			
ORÇAMENTO (5)	0.018482	0.202989	0.031838	0.260697	1		
OSCARES (6)	0.294457	0.155362	0.277437	0.343704	0.036254	1	
PRÉMIOS (7)	0.584541	0.354936	0.559292	0.869569	0.183915	0.351855	1

Tabela 8 - Matriz de correlação de Spearman relativa ao lucro categorizado (sem outliers)

Ao recorrer ao gráfico de importância das variáveis (gráfico 7), optou-se por prescindir das variáveis "**Críticos**" e "**Prémios**", por apresentarem um grau de importância inferior comparativamente com aquelas com as quais possuíam uma grande correlação ("Espectadores" e "Nomeações" respetivamente).

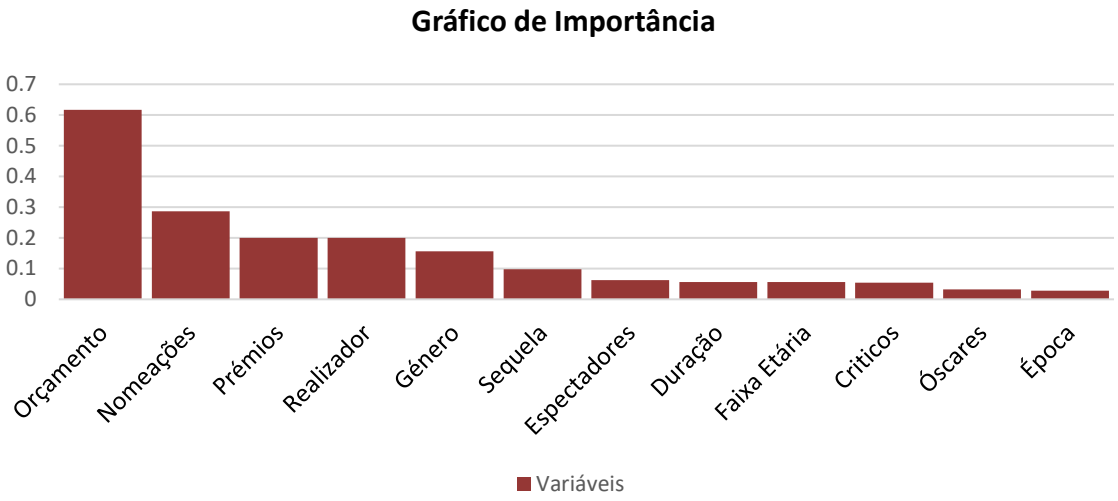

Gráfico 7 - Gráfico de importância relativo ao lucro multi-classe (sem *outliers*)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (coeficiente de regressão)**

Este modelo preditivo multi-classe (nove classes monetárias) não contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 9, existem duas variáveis ("**Óscares**" e "**Prémios**"), cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	***P-Value***
Críticos	0.0258
Duração	0.0334
Época	0.0092
Espectadores	<0.0001
Faixa Etária	<0.0001
Género	<0.0001
Nomeações	<0.0001
Orçamento	<0.0001
Óscares	**0.3839**
Prémios	**0.7828**
Realizador	<0.0001
Sequela	<0.0001

Variáveis rejeitadas

Tabela 9 - Coeficientes de regressão relativos ao lucro categorizado (sem *outliers*)

3.2.2.2 Variável dependente discreta – binário

Seguidamente procedeu-se à redução de dimensão das variáveis do modelo cuja variável dependente assumia a sua forma binária.

Recorrendo novamente ao gráfico de importância, optou-se por descartar a variável "**Óscares**", devido à sua reduzida relação com o lucro cinematográfico.

- **Com *Outliers* (qui-quadrado)**

Este modelo preditivo binário contém *outliers* e recorre ao teste estatístico qui-quadrado para a eliminação de variáveis pouco significativas para o modelo.

Foram excluídas as variáveis que apresentavam valores mais reduzidos quando submetidas ao teste estatístico do qui-quadrado: "**Prémios**", "**Ator**", "**Realizador**" e "**Sequela**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (*Spearman*)**

Este modelo preditivo binário contém *outliers* e recorre à matriz de correlação de *Spearman* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar na tabela 10, as variáveis "Espectadores" e "Críticos" encontram-se fortemente correlacionadas com um coeficiente de correlação de aproximadamente 0.76, assim como as variáveis "Prémios" e "Nomeações" que fixaram o seu valor em aproximadamente 0.87.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)
CRITICOS (1)	1					
DURAÇÃO (2)	0.346774	1				
ESPECTADORES (3)	0.756342	0.468303	1			
NOMEAÇÕES (4)	0.635525	0.389172	0.595468	1		
ORÇAMENTO (5)	0.021704	0.220504	0.043363	0.250157	1	
PRÉMIOS (6)	0.598552	0.370026	0.571293	0.873955	0.173532	1

Tabela 10 - Matriz de correlação de Spearman relativa ao lucro binário (com outliers)

Ao recorrer ao gráfico de importância das variáveis (gráfico 8), optou-se por prescindir das variáveis "**Espectadores**" e "**Prémios**", por apresentarem um grau de importância inferior comparativamente com aquelas com as quais possuíam uma grande correlação ("Críticos" e "Nomeações" respetivamente).

Gráfico 8 - Gráfico de importância relativo ao lucro binário (com *outliers*)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (coeficiente de regressão)**

Este modelo preditivo binário contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 11, existem sete variáveis ("**Ator**", "**Críticos**", "**Época**", "**Faixa Etária**", "**Género**", "**Nomeações**" e "**Prémios**"), cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	*P-Value*
Ator	**0.6114**
Críticos	**0.6833**
Duração	0.0335
Época	**0.5875**
Espectadores	0.0001
Faixa Etária	**0.0813**
Género	**0.7065**
Nomeações	**0.1398**
Orçamento	<0.0001
Prémios	**0.7603**
Realizador	0.0282
Sequela	<0.0001

Variáveis rejeitadas

Tabela 11 - Coeficientes de regressão relativos ao lucro binário (com *outliers*)

- **Sem Outliers (qui-quadrado)**

Este modelo preditivo contínuo não contém *outliers* e recorre ao teste estatístico qui-quadrado para a eliminação de variáveis pouco significativas para o modelo.

Foram excluídas as variáveis que apresentavam valores mais reduzidos quando submetidas ao teste estatístico do qui-quadrado: "**Duração**", "**Ator**", "**Realizador**" e "**Prémios**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (*Spearman*)**

Este modelo preditivo binário não contém *outliers* e recorre à matriz de correlação de *Spearman* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar (tabela 12), as variáveis "Prémios" e "Nomeações" encontram-se fortemente correlacionadas com um coeficiente de correlação de aproximadamente 0.87.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)
CRITICOS (1)	1					
DURAÇÃO (2)	0.317142	1				
ESPECTADORES (3)	0.741576	0.436858	1			
NOMEAÇÕES (4)	0.624501	0.375796	0.578851	1		
ORÇAMENTO (5)	0.009104	0.216768	0.016798	0.248671	1	
PRÉMIOS (6)	0.580041	0.356279	0.544615	0.868758	0.162564	1

Tabela 12 - Matriz de correlação de Spearman relativa ao lucro binário (sem outliers)

Ao recorrer ao gráfico de importância das variáveis (gráfico 9), optou-se por prescindir da variável "**Prémios**", pois apresenta um grau de importância inferior comparativamente com a variável com a qual apresentava um forte coeficiente de correlação, "Nomeações".

Gráfico 9 - Gráfico de importância relativo ao lucro binário (sem outliers)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (coeficiente de regressão)**

Este modelo preditivo binário não contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 13, existem sete variáveis ("**Ator**", "**Críticos**", "**Duração**", "**Época**", "**Género**", "**Nomeações**" e "**Realizador**"), cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	*P-Value*
Ator	**0.7777**
Críticos	**0.5571**
Duração	**0.2109**
Época	**0.2005**
Espectadores	0.0001
Faixa Etária	0.0044
Género	**0.5234**
Nomeações	**0.9072**
Orçamento	<0.0001
Prémios	0.0103
Realizador	**0.0864**
Sequela	<0.0001

Variáveis rejeitadas

Tabela 13 - Coeficientes de regressão relativos ao lucro binário (sem *outliers*)

3.2.2.3 Variável dependente contínua

Por fim, a redução de dimensão foi executada nas variáveis do modelo cuja variável dependente assumia a sua forma intervalar e contínua.

Recorrendo novamente ao gráfico de importância, optou-se por descartar a variável "**Ator**", devido à sua reduzida relação com o lucro cinematográfico.

- **Com *Outliers* (coeficiente de determinação)**

Este modelo preditivo contínuo contém *outliers* e recorre ao coeficiente de determinação para a eliminação de variáveis pouco significativas para o modelo.

Foram excluídas as variáveis que apresentavam um coeficiente de determinação mais reduzido: "**Prémios**", "**Críticos**" e "**Óscares**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (*Pearson*)**

Este modelo contínuo contém *outliers* e recorre à matriz de correlação de *Pearson* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar na tabela 14, as variáveis "Espectadores" e "Críticos" encontram-se fortemente correlacionadas com um coeficiente de correlação de aproximadamente 0.75, assim como as variáveis "Prémios" e "Nomeações" que fixaram o seu valor em aproximadamente 0.94 e ainda as variáveis "Prémios" e "Óscares" com um valor correspondente a 0.76.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CRÍTICOS (1)	1						
DURAÇÃO (2)	0.342512	1					
ESPECTADORES (3)	0.747489	0.45314	1				
NOMEAÇÕES (4)	0.592641	0.365406	0.483352	1			
ORÇAMENTO (5)	0.093433	0.329214	0.131677	0.163548	1		
OSCARES (6)	0.321836	0.26449	0.275219	0.671189	0.098787	1	
PRÉMIOS (7)	0.525895	0.309922	0.424891	0.940047	0.103613	0.754664	1

Tabela 14 - Matriz de correlação de *Pearson* relativa ao lucro intervalar (com outliers)

Ao recorrer ao gráfico de importância das variáveis (gráfico 10), optou-se por prescindir das variáveis "**Nomeações**", "**Óscares**" e "**Críticos**", por apresentarem um grau de importância inferior comparativamente com aquelas com as quais possuíam uma grande correlação ("Prémios" e "Espectadores").

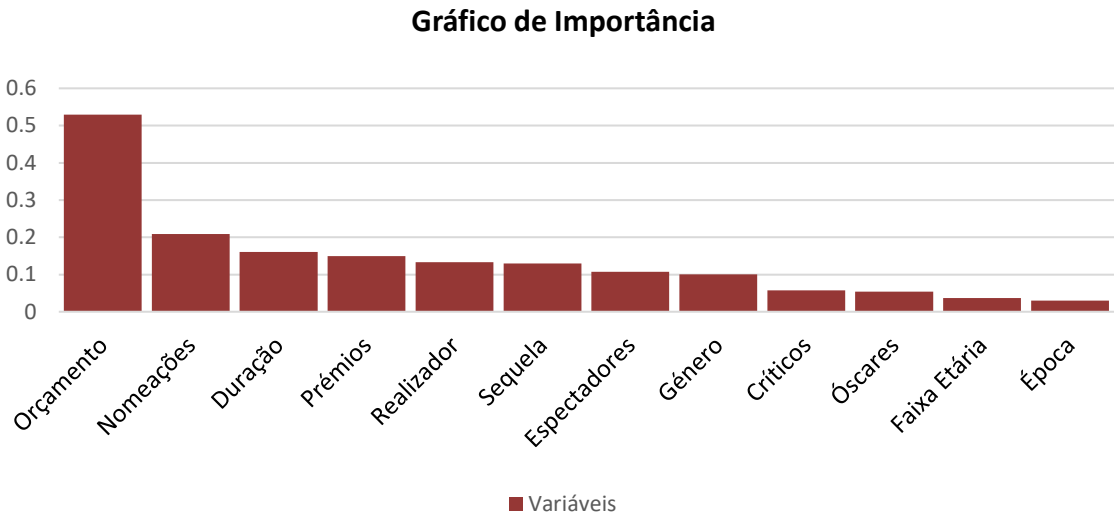

Gráfico 10 - Gráfico de importância relativo ao lucro intervalar (com outliers)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Com *Outliers* (coeficiente de regressão)**

Este modelo preditivo contínuo contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 15, existem seis variáveis ("**Críticos**", "**Duração**", "**Época**", "**Faixa Etária**", "**Óscares**" e "**Prémios**"), cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	*P-Value*
Críticos	**0.7712**
Duração	**0.7546**
Época	**0.1969**
Espectadores	<0.0001
Faixa Etária	**0.1508**
Género	0.0033
Nomeações	0.0263
Orçamento	<0.0001
Óscares	**0.3998**
Prémios	**0.2507**
Realizador	<0.0001
Sequela	<0.0001

Variáveis rejeitadas

Tabela 15 - Coeficientes de regressão relativos ao lucro intervalar (com outliers)

- **Sem *Outliers* (coeficiente de determinação)**

Este modelo preditivo contínuo não contém *outliers* e recorre ao coeficiente de determinação para a eliminação de variáveis pouco significativas para o modelo.

Foram excluídas as variáveis que apresentavam um coeficiente de determinação mais reduzido: "**Duração**", "**Faixa Etária**" e "**Género**", que são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (*Pearson*)**

Este modelo preditivo contínuo não contém *outliers* e recorre à matriz de correlação de *Pearson* para a eliminação de variáveis muito correlacionadas e pouco significativas para o modelo.

Como é possível comprovar na tabela 16, as variáveis "Prémios" e "Nomeações" encontram-se fortemente correlacionadas com um coeficiente de correlação de aproximadamente 0.94.

Quando duas variáveis se encontram altamente correlacionadas é imperativo que uma delas seja eliminada para não existir redundância no modelo.

VARIÁVEL	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CRITICOS (1)	1						
DURAÇÃO (2)	0.323711	1					
ESPECTADORES (3)	0.740038	0.436951	1				
NOMEAÇÕES (4)	0.599056	0.351934	0.507656	1			
ORÇAMENTO (5)	0.084673	0.303689	0.101562	0.144773	1		
OSCARES (6)	0.321285	0.214976	0.271221	0.656431	0.085928	1	
PRÉMIOS (7)	0.544908	0.311525	0.456293	0.937844	0.077526	0.695917	1

Tabela 16 - Matriz de correlação de *Pearson* relativa ao lucro intervalar (sem outliers)

Ao recorrer ao gráfico de importância das variáveis (gráfico 11), optou-se por prescindir da variável "**Prémios**", pois apresenta um grau de importância inferior comparativamente com a variável com a qual apresentava um forte coeficiente de correlação, "Nomeações".

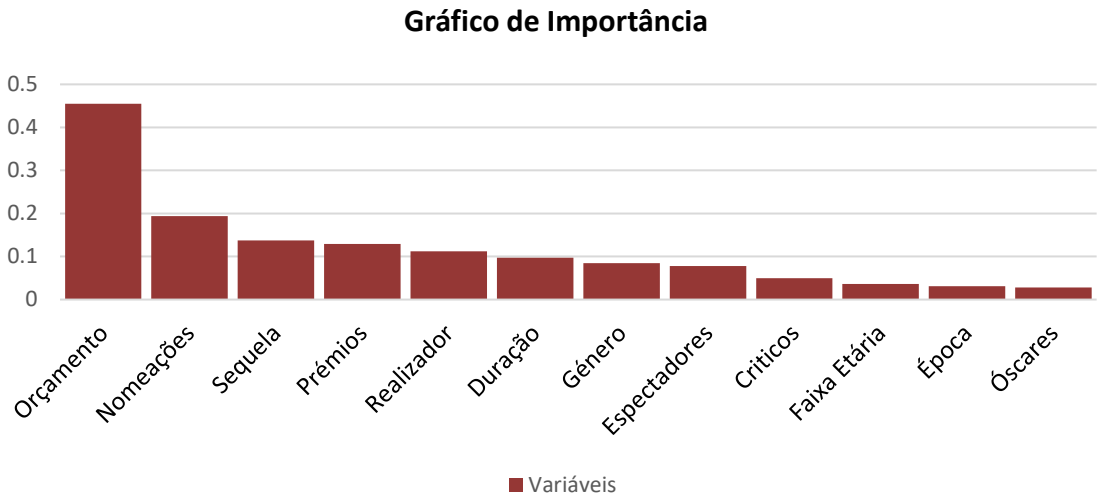

Gráfico 11 - Gráfico de importância relativo ao lucro intervalar (sem outliers)

Estas variáveis são consequentemente rejeitadas e eliminadas do modelo, de forma a não influenciarem negativamente o resultado da previsão.

- **Sem *Outliers* (coeficiente de regressão)**

Este modelo preditivo contínuo não contém *outliers* e recorre ao coeficiente de regressão para a eliminação de variáveis pouco significativas para o modelo.

Através da análise dos *p-values*, é possível perceber a irrelevância de algumas variáveis neste contexto, que apenas influenciarão o modelo negativamente, inflacionando os resultados preditivos.

Como podemos conferir na tabela 17, existem duas variáveis ("**Críticos**" e "**Duração**") cujo valor é superior aos 0.05 de referência desta medida estatística, que não trazem qualquer valor explícito para o modelo, sendo assim rejeitadas posteriormente.

Variável	*P-Value*
Críticos	**0.1738**
Duração	**0.8581**
Época	0.0428
Espectadores	<0.0001
Faixa Etária	0.0234
Género	0.0008
Nomeações	0.0089
Orçamento	<0.0001
Óscares	0.0001
Prémios	0.0454
Realizador	<0.0001
Sequela	<0.0001

Variáveis rejeitadas

Tabela 17 - Coeficientes de regressão relativos ao lucro intervalar (sem outliers)

3.4 *Model*

O modelo preditivo utilizado neste estudo é caracterizado por uma aprendizagem supervisionada - redes neuronais, árvores de decisão e regressão - ,que se destaca pela sua maior precisão (Lee, Booth, & Alam, 2005), e em termos práticos, recebe um conjunto de dados de entrada onde a sua função consiste em decifrar os relacionamentos existentes entre eles. O objetivo principal é conseguir que o modelo seja o máximo realista e conciso nas previsões comportamentais de observações futuras, tendo como base observações históricas.

Para que a análise seja corretamente efetuada é essencial que se observem duas premissas indispensáveis para o modelo.

Em primeiro lugar é necessário a existência de um conjunto de dados adequado e devidamente ajustado para que o modelo possa aprender e posteriormente identificar o padrão pretendido.

É também imperativo que os dados sejam válidos e representativos da realidade que está a ser estudada, sendo que o mais pequeno erro pode ser responsável por uma má previsão.

Existe um grande número de algoritmos utilizados em *Data Mining*. Porém, consoante o problema e o tipo de dados apresentados, uns revelam-se mais adequados do que outros, o que não significa que resolvam o problema de uma forma mais precisa e rápida.

De uma forma geral, não existe uma solução simples e completamente correta para a previsão de um determinado requisito.

Uma das decisões essenciais num modelo preditivo é saber qual o método de aprendizagem a ser utilizado. Neste projeto a escolha recaiu sobre redes neuronais, árvores de decisão e regressão.

3.4.1 Redes Neuronais

As redes neuronais são uma inovadora ferramenta preditiva que tem sido aplicada a diversas áreas científicas, com o objetivo de treinar certos modelos para que sejam capazes de responder aos mais variados problemas (Craven & Shavlik, 1997). Asseguram não só, um grande poder no processo de extração de informação útil para posteriores decisões de negócio, como também, uma grande simplicidade no decorrer de todo o processo, sendo também consideradas representações matemáticas da arquitetura neurológica humana (Amato et al., 2013).

Uma rede neuronal artificial caracteriza-se pela sua inspiração na estrutura das redes neuronais do cérebro humano, que são compostas por uma quantidade elevadíssima de neurónios que se conectam através de uma extensa e elaborada rede de comunicação que permite assim realizar todas as tarefas do dia-a-dia, das mais simples até às mais complexas (Shalev-Shwartz & Ben-David, 2014).

Um estudo conduzido na área defende que as redes neuronais artificiais são mais eficazes quando as relações subjacentes são não-lineares e que melhoram a precisão da previsão dos lucros de bilheteira em 32.8% em relação aos modelos existentes (Ghiassi, Lio, & Moon, 2015).

Atualmente existem inúmeros e variados tipos de redes neuronais artificiais, distinguindo-se cada um deles através das suas regras de aprendizagem e topologias. Neste caso concreto, recorreu-se às redes neuronais *Multi-Layer* (MLP), umas das mais prestigiadas e bem-sucedidas arquiteturas utilizadas em problemas preditivos e de classificação, para aprender complexas funções não lineares, de modo a que possa ser feita uma previsão correta dos dados.

Complementado com esta rede neuronal surge também o algoritmo de *backpropagation*, bastante característico deste tipo de inteligência artificial, que se caracteriza pelo seu decrescente gradiente ao longo de toda a rede, responsável por minimizar o erro quadrático médio do output do modelo. As redes neuronais artificiais que utilizam este algoritmo são as mais utilizadas e consideradas fundamentais para uma maior taxa de sucesso nesta área preditiva (Basheer & Hajmeer, 2000). Apresentam não só uma forte capacidade de aprendizagem e de generalização, essenciais para uma performance exemplar nos muitos domínios onde são utilizadas, especialmente para classificação e previsão, como também eliminam as limitações do método de regressão e estabelecem o mapeamento com precisão entre as variáveis de entrada e de saída (Kak, 2002).

Este tipo de redes neuronais atua com base no mapeamento de conjuntos de dados de entrada em conjuntos de dados de saída e é constituído por uma ou mais camadas ocultas que recorrem a funções *sigmoid*. A combinação destas funções aliada a uma só camada oculta é capaz, por si só, de resolver qualquer problema preditivo ou de classificação.

Como é possível observar na figura 6, as redes neuronais MLP encontram-se divididas em camadas: camada de entrada, camadas escondidas e a camada de saída.

- A primeira camada que é apresentada, é caracterizada pela receção dos dados com os quais o modelo irá ser treinado e tem a função de os armazenar na rede.

- Seguidamente as camadas ocultas responsabilizam-se pelos processamentos que irão atuar ao longo de todo o sistema e podem variar numericamente consoante as necessidades emergentes do modelo.

- Por fim encontra-se a última camada ou camada de saída, que determina o output de todo o modelo, ou seja, discrimina em que classe monetária se encontra cada produção cinéfila que foi sujeita à classificação da rede neuronal apresentada.

As camadas encontram-se todas ligadas através de sinapses (ligações) com a camada anterior, à exceção da primeira, onde se inicia todo o processo.

A escolha do número de camadas escondidas é um fator complexo e contornado muitas vezes pelo exercício de testar várias redes com um número de camadas intermédias distinto, o que permitirá escolher a melhor através da percentagem de erro de cada uma delas, sendo selecionada a que menor valor apresentar.

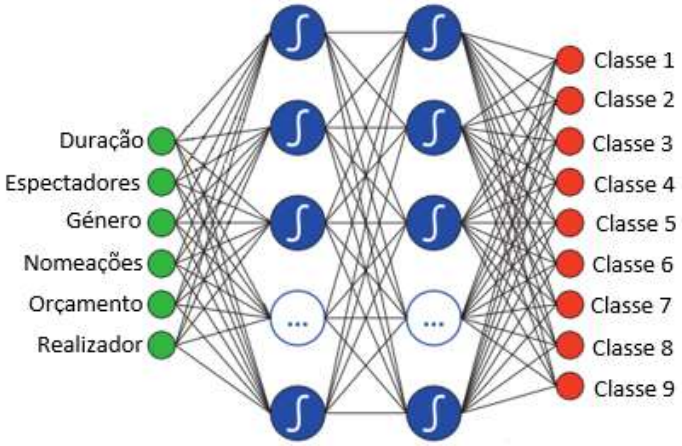

Figura 6 - Representação gráfica de uma rede neuronal MLP com duas camadas ocultas

O processo das redes neuronais MLP começa com a distribuição inicial e aleatória dos pesos associados e definidos logo no início da aprendizagem.

Cada variável da camada inicial é treinada e sintetizada pelas camadas ocultas e posteriormente os valores obtidos na camada de saída são comparados com os reais, terminando o processo com o cálculo do valor de erro que a rede obteve. Este erro é depois reduzido graças ao algoritmo de *backpropagation* característico destas redes neuronais, alterando e ajustando gradualmente o peso das sinapses.

Um dos principais fatores responsáveis pela popularidade das redes neuronais é a sua flexibilidade, ou seja, a capacidade de modelar uma grande quantidade de funções (Williams, 1998), de forma a generalizar o modelo e a classificar novos dados.

3.4.2 Árvores de decisão

Assim como as redes neuronais, acima descritas, também as árvores de decisão representam uma eficaz ferramenta preditiva com algumas particularidades específicas: utilizam um gráfico idêntico a uma árvore onde estão representadas decisões e posteriores consequências, onde uma certa estratégia é escolhida de forma a maximizar a previsão.

O uso deste modelo de aprendizagem é cada vez mais popular no meio das ferramentas preditivas uma vez que, para além de representar regras bastante simples de serem interpretadas por qualquer pessoa em qualquer linguagem, produz grande valor mesmo para conjuntos de dados diversificados e pouco rígidos.

Na figura 7 está representada o inicio da árvore de decisão utilizada neste modelo. A variável inicial é a que apresenta maior valor significativo para o modelo e é a partir dela que a árvore começa a ser construída.

O processo resume-se à tomada de decisões e respetivas consequências no que diz respeito aos valores das diferentes variáveis cinematográficas. Neste caso, os dados irão adotar o caminho que lhes diz respeito, consoante as regras com as quais se irão confrontar, ou seja, se o valor do orçamento de um determinado filme for inferior a 122500000 dólares, significa que o filme adotará o caminho da esquerda em prol do da direita.

O esquema original progride continuamente no sentido descendente, passando por todas as variáveis integrantes, de forma a treinar o modelo da melhor forma possível, sendo no final encontrado o "caminho vencedor" que servirá de base à aprendizagem requerida.

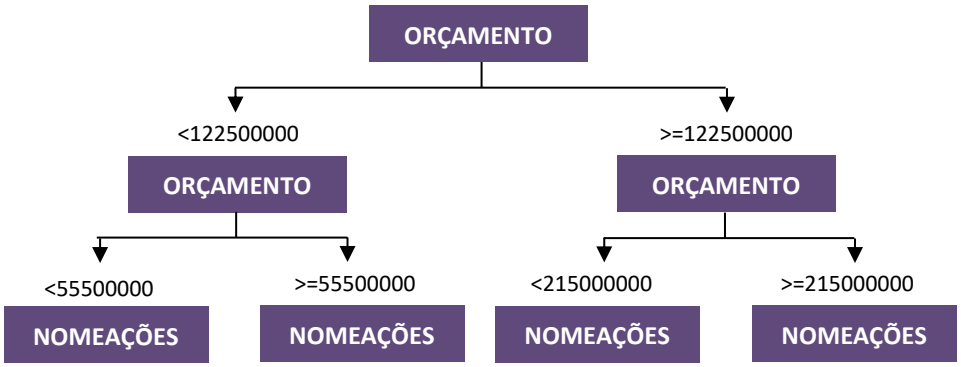

Figura 7 - Representação gráfica de uma árvore de decisão

3.4.3 Regressão

A regressão tem sido uma das abordagens de eleição no que diz respeito a modelos preditivos que englobem a relação entre múltiplas variáveis independentes e uma variável dependente. É maioritariamente utilizada para previsão numérica contínua, porém abrange igualmente identificação de tendências discretas (Han, Kamber, & Pei, 2012).

Sendo uma das técnicas estatísticas mais utilizadas neste meio, permite modelar a relação entre as diversas variáveis intervenientes (Cerrito, 2008).

Neste estudo foram utilizados dois tipos distintos de regressão:

- Regressão linear para um lucro cinéfilo contínuo;
- Regressão logística para quando este se apresenta em forma binária ou ordinal.

A regressão logística, utilizada em variáveis dependentes categóricas, neste caso ordinais e binárias, baseia-se no algoritmo da estimativa por máxima verossimilhança, que defende que os coeficientes devem ser selecionados de forma a maximizar a probabilidade dos dados observados e apresenta uma distribuição binomial da variável dependente.

A regressão linear, que pressupõe a linearidade da relação entre uma variável dependente contínua e as restantes variáveis independentes, baseia-se no algoritmo da estimativa dos mínimos quadrados, que defende que os coeficientes de regressão são estabelecidos com o objetivo de minimizar a soma dos quadrados das diferenças entre os valores estimados e os valores observados. Apresenta uma distribuição normal ou gaussiana da variável dependente.

3.5 *Assess*

É essencial que se meça o erro associado a uma qualquer previsão. Estar ciente da diferença e desvio da previsão face ao valor real é imprescindível não só para se proceder a uma avaliação crítica do modelo, como também para determinar a veracidade global dos dados previstos e saber se se encontram dentro dos limites razoáveis. Para tal, existem métricas responsáveis pela avaliação da ferramenta preditiva utilizada, como o erro de classificação e o erro quadrático médio.

3.5.1 Erro de classificação

O erro de classificação (EC) é um erro preditivo utilizado apenas em variáveis discretas (binárias, multi-classe), que recorre à matriz de confusão como auxílio para o seu cálculo.

A Matriz de Confusão é uma representação tabular para variáveis dependentes discretas, onde as linhas reproduzem a classe real e as colunas a classe prevista. A diagonal central engloba os valores do conjunto de teste que foram corretamente classificados, ou seja, onde a classe prevista é equivalente à classe real. Todos os restantes conjuntos, com exceção do total, representam valores que foram mal classificados aquando da execução do modelo.

		PREVISTO		
		Positivo	**Negativo**	**TOTAL**
REAL	**Positivo**	**VP**	FP	VP+FP
	Negativo	FN	**VN**	FN+VN
	TOTAL	VP+FN	FP+VN	**VP+FN+FP+VN**

Esta métrica é utilizada para avaliar o modelo ou mesmo para decidir entre modelos distintos, quanto menor o erro de classificação associado, melhor será o modelo.

$$EC = \frac{VP + VN}{VP + FN + FP + VN}$$

Onde:
VP – Verdadeiros positivos
VN – Verdadeiros negativos
FN – Falsos Negativos
FP – Falsos Positivos

3.5.2 Precisão, Exatidão e Medida-F

Utilizando a matriz de confusão é possível calcular também outras duas medidas relativas à variável dependente binária: precisão e exatidão.

A precisão é a métrica que calcula os valores preditivos positivos, sendo a exatidão a que traduz a sensibilidade do modelo:

$$Precisão = \frac{VP}{(VP + FP)} \qquad Exatidão = \frac{VP}{(VP + FN)}$$

Onde:
VP - Vetor dos valores reais
FP - Vetor dos valores previstos
FN - Vetor dos valores previstos

A combinação destas duas medidas recai sobre uma terceira métrica denominada de "Medida-F" que traduz aproximadamente a média das anteriores para que seja possível atingir resultados mais precisos e analíticos na avaliação do modelo.

$$Medida\ F = 2 \times \frac{PRC \times EXT}{PRC + EXT}$$

Onde:
PRC - Precisão
EXT - Exatidão

3.5.3 Curva ROC e a área abaixo da curva

A análise da curva ROC é um processo simples e eficaz, relativo a variáveis dependentes discretas, que permite estudar a variação da sensibilidade e especificidade para diferentes valores de corte. Em termos geométricos, esta curva é representada por um gráfico cartesiano que combina os valores relativos à taxa dos falsos positivos (1-especificidade) com a taxa dos verdadeiros positivos

(sensibilidade).

A precisão deste teste estatístico é medida através da área abaixo da curva ROC, que consoante o seu valor - que varia entre 0.5 e 1, - indica quão bom é o modelo em estudo.

3.5.4 Erro quadrático médio

O erro quadrático médio (EQM) é utilizado em variáveis contínuas e representa a média da diferença entre o valor do estimador e do parâmetro ao quadrado.

Uma previsão é 100% correta quando o valor do erro quadrático médio assume o valor de "0".

Quanto maior o valor do erro quadrático, pior será o modelo e a previsão associada.

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(xi - \tilde{x}i)^2$$

Onde:
xi - Vetor dos valores reais
$\tilde{x}i$ - Vetor dos valores previstos
n – Número de elementos da população

3.5.5 Erro máximo absoluto

O erro máximo absoluto advém do erro absoluto (EA), e é caracterizado por ser o maior valor registado pelo módulo da diferença entre o valor real e o valor previsto.

Quanto mais elevado for o valor apresentado pior será o modelo em causa.

$$EA = |x - x^*|$$

Onde:
x - Valor real
x^* - Valor previsto

4. Resultados e discussão

Este estudo propunha prever a rentabilidade de um filme recorrendo a um modelo preditivo, onde a variável dependente assumia três morfologias distintas, e comparar os respetivos resultados não só entre si, mas também com outros estudos já realizados na área.

Lucro categórico

Em relação à variável dependente composta por nove classes distintas, a tabela 18 representa todas as combinações que foram executadas de modo a obter o mínimo erro de classificação (EC) e o máximo índice ROC no modelo preditivo. Foram conjugados fatores como a partição dos dados (treino,

validação e teste ou treino e teste), tratamento de *outliers* (com *outliers* ou sem *outliers*) e método de redução da dimensionalidade (qui-quadrado, matriz de correlação de *Spearman* ou coeficiente de regressão). Para cada uma dessas opções foram ainda testadas três metodologias preditivas distintas: redes neuronais, regressão e árvore de decisão.

A rede neuronal, com três neurónios na camada oculta, foi a que apresentou menor erro de classificação (0.6), assim como um índice ROC bastante elevado (0.933) e, por conseguinte, a que melhor poder preditivo apresenta quando atua no conjunto de teste. O valor do erro, apesar de ser o mais reduzido dentro de todas as metodologias utilizadas, é também, em termos representativos, muito elevado para que o modelo possa ser considerado favorável na resposta à questão primordial deste estudo, o que se deve ao grande número de classes da variável dependente.

Quanto maior for o número de classes a ser previsto pelo modelo, maior irá ser a percentagem de erro e a dificuldade em prever acertadamente todas as classes.

			Erro	RN1	RN2	**RN3**	RN4	RG	AD
OUTLIERS	**70/15/15**	X^2	EC	0.6746	0.6373	0.6475	0.6678	0.6610	0.6644
			ROC	0.919	0.934	0.922	0.918	0.921	0.894
		Spearman	EC	0.6440	0.6260	0.6081	0.6073	0.6350	0.6103
			ROC	0.928	0.925	0.94	0.933	0.931	0.895
		Regressão	EC	0.6373	0.6509	**0.6**	0.6475	0.6509	0.6712
			ROC	0.93	0.929	**0.933**	0.94	0.931	0.897
	70/30	X^2	EC	0.6467	0.6501	0.6535	0.6329	0.6570	0.6621
			ROC	0.926	0.935	0.937	0.924	0.926	0.9
		Spearman	EC	0.6518	0.6467	0.6106	0.6467	0.6604	0.6587
			ROC	0.93	0.922	0.943	0.927	0.931	0.882
		Regressão	EC	0.6604	0.6467	0.6501	0.6415	0.6604	0.6621
			ROC	0.928	0.929	0.929	0.933	0.93	0.878
SEM *OUTLIERS*	70/15/15	X^2	EC	0.6444	0.6655	0.6373	0.6444	0.6655	0.6620
			ROC	0.923	0.922	0.923	0.928	0.917	0.897
		Spearman	EC	0.6585	0.6585	0.6197	0.6197	0.6585	0.6796
			ROC	0.919	0.918	0.922	0.914	0.917	0.898
		Regressão	EC	0.6725	0.6514	0.6338	0.6444	0.6444	0.6796
			ROC	0.92	0.915	0.917	0.923	0.917	0.897
	70/30	X^2	EC	0.6454	0.6348	0.6454	0.6383	0.6472	0.6507
			ROC	0.923	0.931	0.929	0.935	0.919	0.911
		Spearman	EC	0.6401	0.6472	0.6153	0.6188	0.6525	0.6614
			ROC	0.92	0.922	0.931	0.93	0.917	0.91
		Regressão	EC	0.6578	0.6578	0.6418	0.6472	0.6472	0.6614
			ROC	0.921	0.921	0.932	0.938	0.919	0.91

Tabela 18 - Erros de classificação do modelo preditivo para variável dependente multi-classe

Na figura 19 é apresentada a matriz de confusão correspondente a este modelo particular que divide a variável dependente em nove classes distintas.

		CLASSE PREVISTA									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**TOTAL**
CLASSE REAL	**1**	**0**	1	0	1	0	0	0	0	0	2
	2	0	**4**	1	7	2	4	0	0	0	18
	3	0	1	**1**	12	4	4	2	0	0	24
	4	0	2	0	**23**	4	7	2	0	3	41
	5	0	1	1	17	**7**	10	2	0	2	40
	6	0	1	0	8	4	**16**	4	0	11	44
	7	0	0	0	3	2	13	**8**	0	9	35
	8	0	0	0	1	1	9	5	**0**	9	25
	9	0	0	0	2	0	3	2	0	**59**	66
	TOTAL	0	10	3	74	24	66	25	0	93	**295**

Tabela 19- Matriz de Confusão do modelo preditivo para variável dependente multi-classe

Como é possível observar, das 295 observações do conjunto de teste, apenas 118 foram corretamente classificadas, o que traduz bem a questão acima descrita que relaciona a fraca percentagem de sucesso na classificação com o número elevado de classes a serem previstas.

Desta forma, foi alargada a previsão da amostra, sendo que, para uma percentagem bem-sucedida não é necessário acertar apenas na classe real, mas também nas duas classes inferiores e superiores: isto é, se o modelo classificar um filme, que apresenta uma classe monetária 6, com a classe 4 significa que a aprendizagem foi bem-sucedida para este registo, uma vez que apesar de não ter sido prevista a classe real foi prevista uma classe que se encontra dois dígitos abaixo, o que para este estudo se converte num sucesso preditivo.

Foram analisados os resultados da previsão do conjunto de teste com o intuito de ser feito um reajustamento do erro de classificação, tendo por base a nova abordagem acima descrita: observou-se uma clara diminuição desse erro, de aproximadamente 45% (tabela 20).

Critério de sucesso	Erro de Classificação
Acertar na classe real	0.60
Acertar na classe real com uma margem de erro de duas classes (inferior/superior)	0.15

Tabela 20 – Comparação de abordagens distintas para a variável dependente multi-classe

Foram ainda avaliadas duas outras representações da variável dependente, binária e intervalar, que aumentaram e melhoraram os resultados da previsão recorrendo às mesmas ferramentas preditivas utilizadas para o modelo multi-classe.

Lucro Binário

Em termos de resultados da variável dependente binária, é possível observar na tabela 21, a diferença exponencial face ao modelo anterior.

Com um erro de classificação extremamente reduzido (0.0669), e com um índice ROC (0.979) e uma Medida-f (0.9091) bastante elevados, obtidos através da rede neuronal com três neurónios na camada oculta, recorrendo, quer ao tratamento de *outliers*, com uma partição de três conjuntos distintos (conjunto de treino, conjunto de validação e conjunto de teste), quer ao coeficiente de regressão para a redução da dimensionalidade, comprova-se que o número de classes com o qual se categoriza a variável dependente faz toda a diferença no sucesso do modelo preditivo.

			Erro	RN1	RN2	**RN3**	RN4	RG	AD
OUTLIERS	70/15/15	X^2	EC	0.1138	0.1104	0.1104	0.1035	0.1207	0.1
			ROC	0.947	0.96	0.949	0.959	0.948	0.933
			ME-F	0.8507	0.8572	0.8546	0.8612	0.8326	0.8664
		Spearman	EC	0.1138	0.1172	0.1104	0.1104	0.1	0.1
			ROC	0.952	0.965	0.958	0.964	0.955	0.933
			ME-F	0.8507	0.8509	0.8519	0.8546	0.8572	0.8664
		Regressão	EC	0.1069	0.1069	0.1035	0.1104	0.1241	0.1069
			ROC	0.944	0.957	0.964	0.959	0.944	0.93
			ME-F	0.8610	0.8610	0.8649	0.8572	0.8286	0.8692
	70/30	X^2	EC	0.0917	0.1021	0.0882	0.0865	0.0986	0.0900
			ROC	0.961	0.966	0.966	0.962	0.959	0.944
			ME-F	0.8759	0.8638	0.8771	0.8821	0.8606	0.8738
		Spearman	EC	0.0917	0.1055	0.0917	0.1125	0.0900	0.0865
			ROC	0.961	0.966	0.969	0.946	0.961	0.945
			ME-F	0.8741	0.8610	0.8759	0.8464	0.8700	0.8780
		Regressão	EC	0.0848	0.0865	0.0951	0.0952	0.0986	0.1021
			ROC	0.965	0.972	0.973	0.969	0.964	0.952
			ME-F	0.8858	0.8854	0.8736	0.8753	0.8599	0.8692
SEM *OUTLIERS*	**70/15/15**	X^2	EC	0.0880	0.1302	0.0845	0.0810	0.0986	0.0845
			ROC	0.973	0.965	0.978	0.979	0.957	0.936
			ME-F	0.8838	0.8212	0.8857	0.8899	0.8614	0.8835
		Spearman	EC	0.0810	0.0986	0.0880	0.0704	0.0986	0.0845
			ROC	0.975	0.971	0.976	0.974	0.959	0.936
			ME-F	0.8921	0.8628	0.8804	0.9039	0.8614	0.8835
		Regressão	EC	0.0845	0.0704	**0.0669**	0.0951	0.0775	0.0810
			ROC	0.977	0.981	**0.979**	0.976	0.976	0.922
			ME-F	0.8879	0.9039	**0.9091**	0.8683	0.8900	0.8940
	70/30	X^2	EC	0.094	0.1201	0.1060	0.0866	0.0919	0.1025
			ROC	0.971	0.967	0.968	0.968	0.959	0.931
			ME-F	0.8739	0.8325	0.8551	0.8825	0.8706	0.8505
		Spearman	EC	0.0848	0.1025	0.0972	0.0972	0.0901	0.1025
			ROC	0.972	0.965	0.971	0.97	0.96	0.931
			ME-F	0.8863	0.8592	0.8662	0.8669	0.8735	0.8505
		Regressão	EC	0.0830	0.0795	0.0671	0.0936	0.0830	0.0989
			ROC	0.973	0.978	0.978	0.976	0.974	0.929
			ME-F	0.8894	0.8936	0.9095	0.8735	0.8839	0.8593

Tabela 21 - Erros de classificação do modelo preditivo para variável dependente binária

Analisando novamente a matriz de confusão, é possível comprovar-se que, num total de 295 observações, apenas 20 foram mal classificadas, valor bastante favorável para a creditação e avaliação deste modelo.

		REAL		
		0	**1**	**TOTAL**
PREVISTO	**0**	**102**	12	114
	1	8	**173**	181
	TOTAL	110	185	**295**

Tabela 22 *- Matriz de Confusão do modelo preditivo para variável dependente binária*

Lucro intervalar

Por último, foi ainda testado o modelo com a variável dependente intervalar. Os resultados estão apresentados na tabela 23, sendo a rede neuronal com quatro neurónios na camada oculta a que melhor traduz a previsão requerida, com um erro quadrático médio (EQM) de 0.2361 e um erro máximo absoluto (EMA) de 3.0232, valor obtido através de uma amostra sem *outliers* com uma partição em três conjuntos distintos (conjunto de treino, validação e teste), e recorrendo ao coeficiente de determinação para a redução da dimensionalidade.

			Erro	RN1	RN2	RN3	**RN4**	RG	AD
OUTLIERS	70/15/15	R^2	EQM	0.3137	0.3059	0.3053	0.2859	0.3241	0.3961
			EMA	2.8373	3.0023	3.0408	3.0499	3.1934	3.5308
		Pearson	EQM	0.3153	0.2776	0.2753	0.2707	0.3226	0.4957
			EMA	2.7300	2.8091	2.7863	2.4189	3.0341	3.5169
		Regressão	EQM	0.3062	0.3320	0.3109	0.2888	0.3298	0.4018
			EMA	3.0119	3.1249	3.2251	2.9244	3.2247	3.5308
	70/30	R^2	EQM	0.3472	1.4697	0.3210	0.3092	0.3725	0.3593
			EMA	3.9688	24.73	3.1241	3.5430	4.2417	4.0247
		Pearson	EQM	0.3516	0.3177	0.289	0.2972	0.3741	0.4627
			EMA	3.9337	3.6607	3.9271	3.6733	4.3204	4.3729
		Regressão	EQM	0.3480	0.3568	0.4699	0.2950	0.3763	0.3711
			EMA	4.0432	4.3193	5.6024	3.6208	4.2657	4.0247
SEM *OUTLIERS*	**70/15/15**	**R^2**	EQM	0.2825	0.2449	0.2620	**0.2361**	0.3264	0.3034
			EMA	3.6149	3.1051	3.7863	**3.0232**	4.2774	3.7132
		Pearson	EQM	0.2600	0.2850	0.2683	0.2694	0.3258	0.3036
			EMA	3.6300	4.1375	3.5293	3.6567	4.3395	3.7132
		Regressão	EQM	0.3140	0.2850	0.2853	0.3180	0.3215	0.3992
			EMA	2.8960	2.6436	3.1141	3.3745	3.1453	3.5308
	70/30	R^2	EQM	0.2935	0.2361	0.2663	0.3030	0.3352	0.3301
			EMA	3.6588	3.1029	3.9165	5.3770	4.2774	4.3871
		Pearson	EQM	0.2683	0.3883	0.2614	0.2676	0.3314	0.3310
			EMA	3.6300	4.8289	3.5167	3.4580	4.3395	4.3871
		Regressão	EQM	0.2680	0.2593	0.2514	0.2437	0.3278	0.3268
			EMA	3.6463	3.9334	3.8259	3.3968	4.3827	4.3871

Tabela 23 - Erros quadráticos médios do modelo preditivo para a variável dependente intervalar

5. Conclusão

Este projeto teve como principal objetivo o desenvolvimento de um modelo capaz de prever o sucesso financeiro de bilheteira de um determinado conjunto de filmes, através de variáveis específicas e dados históricos.

Foi possível concluir que, a percentagem de sucesso da previsão das receitas cinematográficas diferencia-se bastante consoante a tipologia da variável dependente utilizada no estudo.

O modelo empírico demonstrou bons resultados estatísticos aquando da utilização da variável dependente binária e intervalar. Porém, no que diz respeito à previsão multi-classe, os resultados ficaram muito aquém da realidade, influenciando negativamente o modelo.

Na tabela 24 estão representados os erros classificativos finais para cada um dos modelos que utilizaram a variável dependente discreta (multi-classe e binária) e respetivas ferramentas preditivas. Por sua vez, a tabela 25 representa o modelo contínuo que apresentou o menor erro quadrático médio.

Variável dependente	Erro de Classificação	Metodologia Preditiva
Multi-Classe	0.6	Rede Neuronal 3
Binária	0.0669	Rede Neuronal 3

Tabela 24 - Erros de classificação do modelo preditivo para a variável dependente multi-classe e binária

Variável dependente	Erro Quadrático Médio	Metodologia Preditiva
Intervalar	0.2361	Rede Neuronal 4

Tabela 25 - Erro quadrático médio do modelo preditivo para a variável dependente intervalar

Destaque-se ainda, o elevado erro de classificação da variável multi-classe, que apresenta uma percentagem de sucesso de cerca de 40%, menos 12.6% comparativamente com a taxa obtida por Sharda e Delen em 2009.

Já a variável binária apresenta um erro de classificação bastante favorável, uma vez que, um dos estudos com maior percentagem de sucesso recorrendo a redes neuronais e à mesma tipologia binária (Rhee & Zulkernine, 2016), obteve uma *performance* positiva de 88.8%, cerca de 4.5 valores percentuais abaixo do resultado deste estudo.

Por outro lado, o erro quadrático médio alcançado no modelo contínuo obteve resultados muito favoráveis e superou, significativamente, o obtido por um estudo semelhante que utilizou a mesma tipologia (Hunter, Smith, & Singh, 2016), e que não conseguiu atingir valores de erro inferiores a 0.427.

Em relação às metodologias utilizadas, as redes neuronais foram as que melhor poder preditivo forneceram, apresentando uma clara homogeneidade em qualquer um dos três modelos desenvolvidos, o que não surpreende já que, da investigação subjacente ao presente estudo, se concluiu que as percentagens de sucesso mais elevadas e notórias foram obtidas exatamente a partir desta importante ferramenta de *Data Mining* (Ghiassi et al., 2015; Kaur & Nidhi, 2013; Rhee &

Zulkernine, 2016).

No que diz respeito à influência das variáveis presentes no estudo foi possível concluir que, algumas delas não são relevantes para um maior lucro de bilheteira, tendo por isso sido afastadas. A variável "Ator", no caso particular do lucro multi-classe e intervalar, não apresentou grande valor explicativo comparativamente com as restantes, situação que ocorreu simultaneamente com a variável dependente binária em relação à variável "Óscares".

Por outro lado, existe um grupo de variáveis bastante explicativas do modelo e que contribuíram de uma forma inequívoca para o seu sucesso preditivo: "Orçamento", "Realizador" e "Sequela".

6. Limitações e recomendações para trabalhos futuros

Foram identificadas algumas limitações no modelo preditivo que o impediram de alcançar taxas de erro mais reduzidas. Uma das principais, que já foi descrita anteriormente, foi sem dúvida a utilização de uma variável dependente caracterizada por noves classes. Esta escolha foi justificada pela mesma utilização num dos principais estudos feitos na área (Sharda & Delen, 2006, 2009), onde a percentagem de sucesso mais elevada atingiu os 56.1%, valor bastante superior comparativamente ao obtido neste estudo (40%).

Esta diferença tão significativa pode justificar-se pela utilização de uma amostra um pouco mais diversificada e com um maior número de registos, onde a escolha das variáveis foi também distinta e mais seletiva, o que representa outra das limitações encontradas: o número elevado de variáveis pouco significativas para o modelo.

Inicialmente foram introduzidas catorze variáveis independentes, que foram sujeitas uma redução de dimensionalidade recorrendo a técnicas distintas que por vezes chegaram a rejeitar sete variáveis. Se tivesse sido feita uma avaliação à *priori* mais aprofundada da relação das variáveis utilizadas no estudo com o lucro cinematográfico, ter-se-ia evitado a utilização das mesmas e sido incluídas outras com maior valor preditivo.

Como referência e sugestão para trabalhos futuros é proposto um modelo preditivo que responda às necessidades e falhas deste estudo.

Foram testadas três abordagens da variável dependente, sendo a binária a mais bem-sucedida. No entanto, em termos práticos e monetários para os estúdios cinematográficos, adotar um modelo preditivo intervalar mais preciso, proporcionará uma maior vantagem competitiva no mercado atual.

Desta forma é recomendado um modelo caracterizado por uma variável dependente contínua, onde a escolha das variáveis deverá recair sobre as que maior valor apresentaram no modelo correspondente deste estudo ("Orçamento", "Nomeações", "Prémios", "Realizador" e "Sequela"), acrescentando outras interessantes, como o orçamento e lucro obtido com o *marketing* do filme, a opinião "boca-a-boca" através das redes sociais, recorrendo a ferramentas de *Text Mining* e os competidores diretos (filmes lançados no mesmo mês e caracterizados pelo mesmo género).

Seria também interessante limitar este estudo a um único mercado, como por exemplo os Estados Unidos da América, de forma a conseguir restringir os dados e resultados a um único país. Desta forma, seria possível estudar outro tipo de variáveis como os feriados nacionais e dias comemorativos, o número de salas que exibem os filmes em cada Estado, ou mesmo quais as regiões mais propícias a contribuírem monetariamente para um determinado género cinematográfico.

7. Bibliografia

Amato, F., López, A., Peña-Méndez, E., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, *11*(2), 47–58. https://doi.org/10.2478/v10136-012-0031-x

Azevedo, A., & Santos, M. (2008). KDD , SEMMA and CRISP-DM : a parallel overview. *IADIS European Conf. Data Mining*, (8), 182–185.

Basheer, I., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, *67*(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692

Bhalla, D. (2014). Difference between linear regression and logistic regression. Retrieved from http://www.listendata.com/2014/11/difference-between-linear-regression.html

Boatwright, P., Basuroy, S., & Kamakura, W. (2007). Reviewing the reviewers: The impact of individual film critics on box office performance. *Quantitative Marketing and Economics*, *5*(4), 401–425. https://doi.org/10.1007/s11129-007-9029-1

Boyacioglu, M. A., Kara, Y., & Baykan, ??mer Kaan. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, *36*(2 PART 2), 3355–3366. https://doi.org/10.1016/j.eswa.2008.01.003

Cerrito, P. B. (2008). The Difference Between Predictive Modeling and Regression, 1–19.

Chakravarty, A., Liu, Y., & Mazumdar, T. (2010). The Differential Effects of Online Word-of-Mouth and Critics ' Reviews on Pre-release Movie Evaluation. *Forthcoming at Journal of Interactive Marketing*.

Chang, B.-H., & Ki, E.-J. (2005). Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property. *Journal of Media Economics*, *18*(4), 247–269. https://doi.org/10.1207/s15327736me1804

Chok, N. (2010). Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. *Graduate School of Public Health*, 1–53. https://doi.org/10.1017/CBO9781107415324.004

Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, *13*(2–3), 211–229. https://doi.org/10.1016/S0167-739X(97)00022-8

Deniz, B., & Hasbrouck, R. B. (2012). What Determines Box Office Success of a Movie in the United States? *Proceedings for the Northeast Region Decision Sciences Institute*, (757), 447.

Deuchert, E., Adjamah, K., & Pauly, F. (2005). For Oscar Glory Or Oscar Money? *Journal of Cultural Economics*, *29*(3), 159–176. https://doi.org/10.1007/s10824-005-3338-6

Duan, W., Gu, B., & Whinston, A. (2008). The Dynamics of Online Word-of-Mouth and Product Sales – An Empirical Investigation of the Movie Industry. *Forthcoming at Journal of Retailing*.

Einav, L. (2007). Seasonality in the U . S . motion picture industry. *The Rand Journal of Economics*, *38*, 127–145.

Elberse, A. (2007). The power of stars: do star actors drive the success of movies? *Journal of Marketing*, *71*(4), 102–120. https://doi.org/10.1007/978-1-4419-6803-6

Eliashberg, J., & Shugan, S. M. (1997). Film critics : Influencers or predictors ? *Journal of Marketing*, *61*(2), 68.

Fazzion, E., Casas, P., Gonçalves, G., Melo-Minardi, R., & Meira, W. (2013). Open Weekend and Rating Prediction Based on Visualization Techniques. *Proc. IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper)*, (October), 1–2. https://doi.org/10.13140/RG.2.1.2793.2329

García, E., Ventura, S., & Romero, C. (2007). Data mining in course management systems : Moodle case

study and tutorial. *Computers & Education*. https://doi.org/10.1016/j.compedu.2007.05.016

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, *40*(1–3), 11–61. https://doi.org/10.1016/0004-3702(89)90046-5

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, *42*(6), 3176–3193. https://doi.org/10.1016/j.eswa.2014.11.022

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and techniques* (3rd ed.).

Hunter, S., Smith, S., & Singh, S. (2016). Predicting Box Office from the Screenplay: An Empirical Model. *Journal of Screenwriting*, *7*(2). https://doi.org/10.1386/josc.7.2.135

Ikran, S. T., & Cherukuri, A. K. (2016). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences*, 1319–1578. https://doi.org/10.1016/j.jksuci.2015.12.004

Im, D., & Nguyen, M. T. (2011). Predicting Box-Office Success of Movies in the U . S . Market, 1–5.

John, K., Ravid, S. A., & Sunder, J. (2003). The role of termination in employment contracts: Theory and evidence from film directors' careers. *SSRN Electronic Journal*.

Kak, S. (2002). A class of instantaneously trained neural networks. *Information Sciences*, *148*, 97–102.

Karniouchina, E. V. (2010). Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing*, *28*(1), 62–74. https://doi.org/10.1016/j.ijresmar.2010.10.001

Kaur, A., & Nidhi, A. P. (2013). Predicting Movie Success Using Neural Network. *International Journal of Science and Research (IJSR)*, *2*(9), 69–71.

Khan, J., Wei, J., Ringnér, M., Saal, L., Ladanyi, M., Westermann, F., … Meltzer, P. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–9. https://doi.org/10.1038/89044

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990). Stock market prediction system with modular neural networks. *1990 IJCNN International Joint Conference on Neural Networks*, 1–6 vol.1. https://doi.org/10.1109/IJCNN.1990.137535

Kumar, G. R., Kongara, V. S., & Ramachandra, D. G. . (2013). An Efficient Ensemble Based Classification Techniques for Medical Diagnosis. *IJLTEMAS*, *2*(8), 2278–2540.

Lash, M., & Zhao, K. (2015). Early Predictions of Movie Success: the Who, What, and When of Profitability. *Social Computing, Behavioral-Cultural Modeling, and Prediction Lecture Notes in Computer Science*, *9021*, 345–349. Retrieved from http://arxiv.org/abs/1506.05382

Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*, *29*(1), 1–16. https://doi.org/10.1016/j.eswa.2005.01.004

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. *KDD-98*, 73–79. https://doi.org/10.1.1.332.1803

Litman, B. R. (1983). Predicting Succes of Theatrical Movies: An Empirical Study. *Journal of Popular Culture*, 159–175.

Litman, B. R., & Kohl, L. S. (1989). Predicting Financial Success of Motion Pictures: The '80s Experience. *Journal of Media Economics*, *2*(2), 35–50. https://doi.org/10.1080/08997768909358184

Marôco, J. (2014). *Análise Estatística com o SPSS Statistics*. (ReportNumber, Ed.).

McKenzie, J. (2013). Predicting box office with and without markets: Do internet users know anything? *Information Economics and Policy*, *25*(2), 70–80. https://doi.org/10.1016/j.infoecopol.2013.05.001

Mehta, S., Bhatt, H., & Desai, P. D. (2015). A Compendium for Prediction of Success of a Movie Based Upon Different Factors. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(12), 297–300. https://doi.org/10.17148/IJARCCE.2015.41268

Michie, D., Spiegelhalter, D., & Taylor, C. (1994). Machine Learning , Neural and Statistical Classification. *Ellis Horwood Limited*.

Mojo, B. O. (2016a). People index by gross. Retrieved from http://www.boxofficemojo.com/people/?view=Actor&sort=sumgross&p=.htm

Mojo, B. O. (2016b). Yearly Box Office. Retrieved from http://www.boxofficemojo.com/yearly/?view2=domestic&view=releasedate&p=.htm

Nash Information Services, L. (2016). Annual Ticket Sales. Retrieved from http://www.the-numbers.com/market/

Neelamegham, R., & Chintagunta, P. (1999). A Bayesian Model to Forecast New Product Performance in Domestic and International Markets. *Marketing Science*, *18*(2), 115–136. https://doi.org/10.2307/193212

Nelson, R., Donihue, M., Waldman, D., & Wheaton, C. (2011). What â€™s an Oscar Worth. *Economic Inquiry*. https://doi.org/10.1111/j.1465-7295.2001.tb00046.x

Nithin, V., Pranav, M., Sarath, P., & Lijiya, A. (2014). Predicting Movie Sucess Based on IMDB Data. *International Journal of Data Mining Techniques and Applications*, *3*, 365–368.

Pijanowski, B., Brown, D., Shellito, B., & Manik, G. (2002). Using neural networks and GIS to forecast land use changes: A Land Transformation Model. *Computers, Environment and Urban Systems*, *26*(6), 553–575. https://doi.org/10.1016/S0198-9715(01)00015-1

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143. https://doi.org/10.1017/S0962492900002919

Prag, J., & Casavant, J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *Journal of Cultural Economics*, *18*(3), 217–235. https://doi.org/10.1007/BF01080227

Qiang, L., Yuan, Y., Chun, M., & Dong, S. (2006). BP neural network prediction of the mechnical properties of porous NiTi shape memory alloy prepared by thermal explosion reaction. *Materials Science & Engineering A*, *419*, 214–217. https://doi.org/10.1016/j.msea.2005.12.027

Ravid, S. (1999). Information, Blockbusters and Stars: A Study of the Film Industry. *Journal of Business*, *72*.

Redy, A., Kasat, P., & Jain, A. (2012). Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining. *International Journal of Computer Applications*, *56*(1). https://doi.org/10.5120/8852-2794

Rhee, T., & Zulkernine, F. (2016). Predicting Movie Box Office Gross: A Neural Network Approach. *15th IEEE International Conference on Machine Learning and Applications*, 665–670. https://doi.org/10.1109/ICMLA.2016.138

Riwinoto, M.T., Zega, S., & Irlanda, G. (2015). Predicting animated film of box-office success with neural networks. *Jurnal Teknologi*, *23*, 77–82.

Sawhney, M. S., & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, *15*(2), 113. https://doi.org/10.1287/mksc.15.2.113

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: from theory to algorithms*.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018

Sharda, R., & Delen, D. (2009). Predicting the financial success of hollywood movies using an information fusio approach, 30–38.

Simonoff, J., & Sparrow, I. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance-Berlin Then New*, *13*(3), 40. https://doi.org/10.1080/09332480.2000.10542216

Sochay, S. (1994). Predicting the Performance of Motion Pictures. *Journal of Media Economics*, *7*(4), 1–20. https://doi.org/10.1207/s15327736me0704_1

Song, J., & Han, S. (2013). Predicting Gross Box Office Revenue for Domestic Films. *Communications for Statistical Applications and Methods*, *20*(4), 301–309. https://doi.org/10.5351/CSAM.2013.20.4.301

Terry, N., Cooley, J. W., & Zachary, M. (2008). The determinants of foreign box office revenue for English language movies. *Journal of International Business and Cultural Studies*, 1–12. Retrieved from http://aabri.com/manuscripts/09274.pdf

Topf, P. (2010). Examining Success at the Domestic Box-Office in the Motion Picture Industry. *Honors Projects*. Retrieved from http://digitalcommons.iwu.edu/econ_honproj/110

Vany, A., & Walls, D. (2002). Movie stars, big budgets, and wide releases. Empirical analysis of the blockbuster strategy. *Latin American Meeting of the Econometric Society*.

Wallace, W., Seigerman, A., & Holbrook, M. (1993). The role of actors and actresses in the success of films: How much is a movie star worth? *Journal of Cultural Economics*, *17*(1), 1–27.

Walls, W. D. (2005). Modeling Movie Success When "Nobody Knows Anything": Conditional Stable-Distribution Analysis Of Film Returns. *Journal of Cultural Economics*, *29*(3), 177–190. https://doi.org/10.1007/s10824-005-1156-5

Walls, W. D. (2009). Screen wars, star wars, and sequels. *Empirical Economics*, *37*(2), 447–461. https://doi.org/10.1007/s00181-008-0240-z

White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns. *Neural Networks, 1988., IEEE International Conference on*, 451–458. https://doi.org/10.1109/ICNN.1988.23959

Williams, C. K. I. (1998). Prediction with Gaussian processes: from linear regression to linear prediction and beyond. *Learning and Inference in Graphical Models*.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco: Inc., Morgan Kaufmann Publishers.

Yang, Y., & Pedersen, J. (1997). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Proceedings of the 14th International Conference on Machine Learning (ICML)*, 412–420. https://doi.org/10.1093/bioinformatics/bth267

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *European Journal of Operational Research*, *116*(1), 16–32. https://doi.org/10.1016/S0377-2217(98)00051-4

Zhang, L., Luo, J., & Yang, S. (2009). Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications*, *36*, 6580–6587. https://doi.org/10.1016/j.eswa.2008.07.064

8. Anexos

Anexo 1 - Precisão e Exatidão da variável dependente binária

			Erro	RN1	RN2	RN3	RN4	RG	AD
OUTLIERS	70/15/15	X^2	PRC	0.8868	0.9057	0.8868	0.8774	0.8208	0.8868
			RCC	0.8174	0.8136	0.8246	0.8455	0.8447	0.8469
		Spearman	PRC	0.8868	0.9151	0.8680	0.8868	0.8208	0.8868
			RCC	0.8174	0.7951	0.8364	0.8246	0.8970	0.8469
		Regressão	PRC	0.9057	0.9057	0.9057	0.9057	0.8208	0.9717
			RCC	0.8205	0.8205	0.8276	0.8136	0.8365	0.7863
	70/30	X^2	PRC	0.8905	0.8905	0.8667	0.8905	0.8381	0.8571
			RCC	0.8618	0.8386	0.8879	0.8738	0.8844	0.8911
		Spearman	PRC	0.8762	0.9	0.8905	0.8524	0.8286	0.8571
			RCC	0.8720	0.8253	0.8618	0.8404	0.9158	0.9
		Regressão	PRC	0.9048	0.9191	0.9048	0.9191	0.8333	0.9333
			RCC	0.8676	0.8540	0.8444	0.8355	0.8883	0.8133
SEM *OUTLIERS*	70/15/15	X^2	PRC	0.9135	0.8173	0.8942	0.8942	0.8365	0.875
			RCC	0.8559	0.8252	0.8774	0.8857	0.8878	0.8922
		Spearman	PRC	0.9135	0.8462	0.8846	0.9039	0.8365	0.875
			RCC	0.8716	0.88	0.8762	0.9039	0.8878	0.8922
		Regressão	PRC	0.9135	0.9039	0.9135	0.8558	0.8558	0.9327
			RCC	0.8636	0.9039	0.9048	0.8812	0.9271	0.8584
	70/30	X^2	PRC	0.9034	0.8164	0.8551	0.8889	0.8454	0.7971
			RCC	0.8462	0.8492	0.8551	0.8762	0.8974	0.9116
		Spearman	PRC	0.9034	0.8551	0.8599	0.8647	0.8502	0.7971
			RCC	0.8698	0.8634	0.8726	0.8690	0.8980	0.9116
		Regressão	PRC	0.9130	0.9130	0.9227	0.8841	0.8647	0.8261
			RCC	0.8670	0.875	0.8967	0.8632	0.9040	0.8953

Tabela 26 - Valores de precisão e exatidão da variável dependente binária