Estudo exploratório de aplicação de técnicas de aprendizado de máquina e inteligencia artificial na industria cinematográfica.

Exploratory Study of Machine Learning and Artificial Intelligence Techniques Application in the Film Industry.

Patrick Dantas França dos Santos[1]
Roberto Angelo Fernandes Santos[2]

Resumo

O presente estudo aborda a aplicação de técnicas de aprendizado de máquina e inteligência artificial em contextos de inteligência de negócios de forma exploratória, em dados da industria cinematográfica, com o intuito de fornecer dados preditivos para geração de relatórios que possam apoiar o processo de decisão dos profissionais da área, desde a construção da programação de seus cinemas ao fechamento de contratos de exibição, esse estudo foi realizado com 130.000 registros utilizados em diversas técnicas como *Multilayer perceptron* (MLP) e Random Forest, em busca de possíveis recursos que poderiam ser aplicados em sistemas de inteligência de negócio. Os resultados demonstram os possíveis usos e melhoras que podem ser aplicados para aplicação do modelo em produtos de inteligencia de negócio.

Palavras-chave: Cinema. Inteligencia de Negócio. Random Forest. Aprendizado de Máquina. Industria Cinematográfica.

Abstract

The present study addresses the application of machine learning and artificial intelligence techniques in business intelligence contexts, specifically exploring data within the movie industry. The aim is to provide predictive data for generating reports that can support decision-making processes among industry professionals. This includes constructing movie theaters schedules and finalizing exhibition contracts. The study utilized 130,000 records across various techniques such as Multilayer Perceptron (MLP) and Random Forest, seeking potential resources applicable to business intelligence systems. The findings showcase potential uses and improvements that could be implemented in business intelligence product applications.

Keywords: Movie Theater. Business Intelligence. Random Forest. Machine Learning. Movie Industry.

[1] Graduando em Banco de dados na Faculdade Impacta. E-mail: PatrickDantas999@gmail.com
[2] Docente do curso de Banco de Dados da Faculdade Impacta. Mestre em Ciência da Computação. E-mail:

1 INTRODUÇÃO

Com o aumento do uso de inteligência de negócios para apoiar decisões, exibidores e distribuidores procuram por ferramentas que ajudem a compreender o mercado. Compreender o mercado possibilita aumentar lucros, reduzir gastos, fechar acordos melhores e evitar práticas fraudulentas.

A disseminação da inteligência artificial[1] requer que essas ferramentas se adaptem. A motivação por trás deste estudo é explorar como a inteligência artificial e o aprendizado de máquina podem analisar dados históricos da indústria cinematográfica. O objetivo é prever o número total de sessões semanais em 350 cinemas pelo Brasil, baseando-se apenas em informações do mercado brasileiro.

O estudo segue uma abordagem exploratória, reconhecendo que os resultados podem ser incertos devido à incerteza sobre a adequação dos dados para as técnicas testadas. A principal métrica usada para avaliar a qualidade dos modelos será a média absoluta de erros nas previsões, permitindo análises manuais para garantir previsões precisas e consistentes.

2 CARACTERIZAÇÃO DO PROBLEMA

O cinema brasileiro data desde 1896, tendo seus inicio apenas 6 meses após a primeira exibição realizada pelos irmãos Lumière em Paris[2]. Esse mercado se desenvolveu e mudou constantemente até se tornar o que é atualmente[3].

2.1 MERCADO CINEMATOGRÁFICO

O mercado do cinema é composto por distribuidores que são aqueles que produzem os filmes que serão exibidos nos cinemas, os exibidores são as redes de cinemas que por sua vez apresentam esse filmes, junto cada um ganha um porcentagem do ingresso, então esclarecido o mercado a decisão a ser tomada é, qual filme se deve exibir e em qual cinema deve ser exibido.

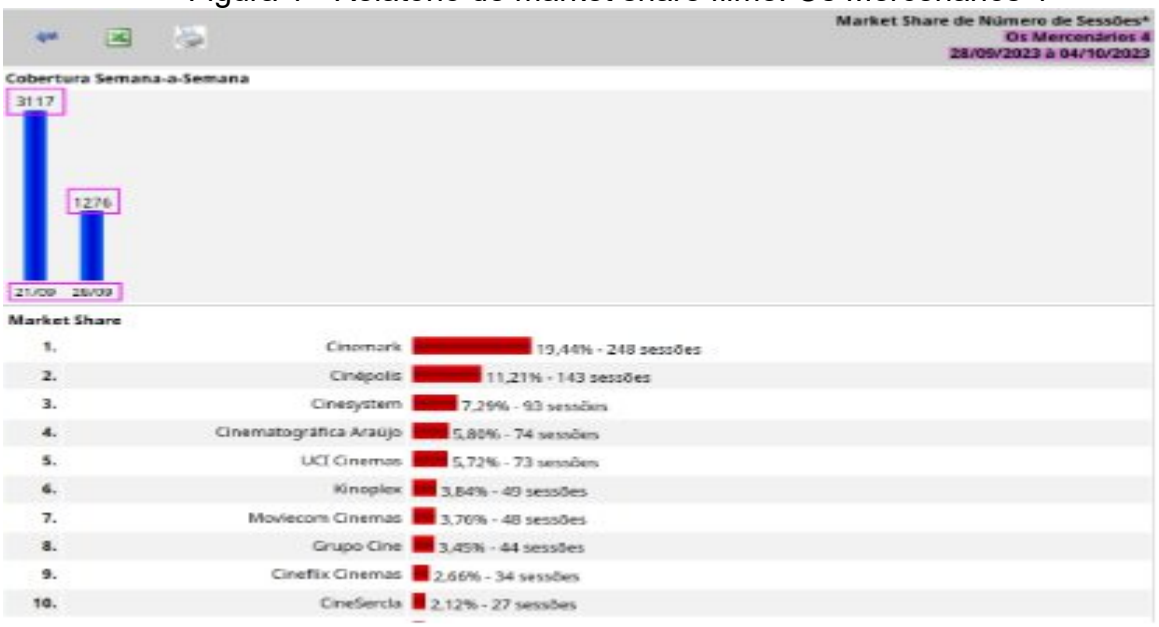
É comum distribuidores e exibidores fecharem acordos de exibição, como por exemplo o horários que um filme será exibido, quantas sessões, dias da semana, filmes rivais que passarão em outras salas, exclusividade de exibição entre diversos outros, que afetam o montagem da grade de programação semanal de um cinema e planos de cobertura e vendas dos distribuidores, avaliar o mercado para tomar a decisão é comum, exigências feitas para um exibidor que não é feita a outro, melhores ofertas, maior demanda, todos elementos que afetam como tal grade será montada.

Aplicações de Inteligencia de Negócio (BI), são uma fonte valiosa de informação, agregar dados de forma que permita se possuir um visão melhor do mercado, avaliar performance de filmes, volume de sessões concorrentes, distribuição de exibição do filme, comparar filmes entre diversos outros dados, que são usados para tomar as decisões dos próximos meses.

2.2 OBJETIVO

O foco deste estudo é o relatório de Market Share de filmes exibido na figura 1, que analisa a quantidade de sessões que um filme tem durante as semanas, identificando em qual exibidor ocorrem e quantas semanas em exibição.

Figura 1 - Relatório de market share filme: Os Mercenários 4



Fonte: https//www.equinoxprog.com.br

Com base na literatura[4], as técnicas de inteligência artificial e aprendizado de máquina, têm a capacidade de usar dados para criação de previsões de alta precisão e oferecer uma ferramenta valiosa para aqueles que dependem dessas informações para tomada de decisões.

O valor a ser previsto é a quantidade de sessões que o filme possui na semana que se deseja prever, de acordo com os dados que originam a Figura 1 temos a granularidade dos dados feitas pela quantia de sessões na semana, por filme em uma sala da unidade de cinema do exibidor no seguinte idioma podendo ser português(dublado), original(legendado) ou nacional apresentado com o vídeo em 2D ou 3D, portanto, o problema se enquadra como um problema de regressão.

3 MASSA DE DADOS

Os dados utilizados se originam do histórico de sessões realizadas no período entre 01 de julho de 2021 a 21 de agosto de 2023, obtidos a partir do serviço de venda de ingressos online conhecido como Velox[5]. A base de dados abrange um total de 84 redes de cinemas em todo o país, que compreendem um conjunto de 340 unidades distribuídas em território nacional. Essa abrangência resultou em uma coleta de aproximadamente 130.000 registros.

Dada a seguinte base, temos como alvo a variável QtdSessoes apresentada na Tabela 1 que representa a quantia de sessões, todas as outras

variáveis apresentadas na Tabela 1 podem ser utilizadas como entrada para o treinamento do modelo ou geração de variáveis artificiais.

Tabela 1 - Dicionario de dados

Variável	Tipo	Descrição	Objeto
NomeBrasil	string	Nome do filme no Brasil	Filme
NomeOriginal	string	Nome original do filme	Filme
ELENCO	string	Atores que atuaram no filme	Filme
DIRETOR	string	Diretor(es) do filme	Filme
PAIS	string	País(es) de origem do filme	Filme
Brasileiro	bool	Indica se o filme é brasileiro	Filme
GENERO	string	Gênero(s) do filme	Filme
ESTREIA	string	Data de estreia do filme	Filme
ClassificacaoIndicativa	int	Classificação etária do filme	Filme
SINOPSE	string	Descrição do filme	Filme
Duracao	int	Duração em minutos do filme	Filme
Roteirista	string	Roteirista(s) do filme	Filme
Producao	string	Equipe de produção do filme	Filme
Executiva	string	Equipe de produção executiva do filme	Filme
Comercial	bool	Indica se o filme é comercial ou cultural	Filme
F_Dolby	bool	Indica suporte ao áudio Dolby Atmos	Filme
F_3D	bool	Indica suporte ao formato 3D	Filme
F_IMAX	bool	Indica suporte ao formato IMAX	Filme
Trailer	string	Código do trailer do filme no YouTube	Filme
Distribuidora	string	Empresa responsável pela distribuição do filme	Filme
Estudio	string	Estúdio que produziu o filme	Filme
Cinema	string	Nome do cinema	Cinema
Exibidor	string	Nome da rede de cinemas	Cinema
CIDADE	string	Cidade onde o cinema está localizado	Cinema
ESTADO	string	Estado onde o cinema está localizado	Cinema
QTD_SALAS	int	Quantidade de salas do cinema	Cinema
SessaoCinema	int	Código do cinema	Sessão
SessaoLegenda	string	Indica se o filme foi exibido com dublagem, legenda ou idioma original	Sessão
SessaoVideo	string	Indica o formato da exibição (2D, 3D, IMAX, etc.)	Sessão
CinesemanaInicio	string	Início da semana de exibição	Sessão
CinesemanaFim	string	Fim da semana de exibição	Sessão
QtdSessoes	int	Quantidade de sessões exibidas (Variável Alvo)	Sessão

Na base temos variáveis originadas de objetos com filme, cinema e sessão, que indicam representatividade do dado, é importante saber quais

variáveis devem ser utilizadas e a importância delas para o modelo, a utilização dos objetos deve ser homogênea para garantir uma previsão de boa qualidade, se não teria-se problemas se um filme claramente de baixa performance tenha previsões altas pois o modelo priorizou somente aquelas referente a cinema.

Dentre os objetos, o de filme é o mais problemático já que algumas variáveis não possuem grande utilidade pois apesar de se possuir dados dos filmes, não é possível se obter valores válidos de análise de sentimento da sinopse por não existir uma biblioteca de boa performance em português que é o idioma em que o dado se encontra[6].

Junto a sinopse temos elenco, diretor e roteirista que não possuem uso efetivo uma vez que a relevância desses indivíduos é muito volátil e vária de acordo com o gênero do filme, opinião publica e fatores externos, exemplifica-se com os ocorridos com o ator Johnny Depp[7], e se agrava já que não se possui a informação do papel interpretado pelo ator, assim não se pode julgar um filme A onde ator A foi protagonista e filme B onde ator A foi vilão. Algo que pode causar mudança de tal cenário é com a criação de modelos dedicados a tais valores para gerar variáveis sintéticas para o atualmente estudado.

4 PROCEDIMENTO METODOLÓGICO

Inicialmente analisou-se a correlação das variáveis numéricas e booleanas apresentadas na Tabela 1 com o alvo através de algorítimos como Pearson, Kendall e Spearman[8], amplamente utilizado no entendimento da correlação das variáveis de uma base entre si, o resultado dessa correlação permite que se obtenha um entendimento dos valores atuais da base.

Como resultado dessa correlação se destaca duas variáveis que apresentam relevância mediana, Comercial e QTD_SALAS de acordo com com Kendall e somente QTD_SALAS de acordo com Pearson, mas, em todos os três algorítimos os valores obtidos foram elevados, se destacando entre os outros, o que indica que ambas possuem peso que as técnicas propostas podem utilizar para prever o alvo.

Ainda sim se faz necessário o uso de mais variáveis, já que as encontradas se tratam somente sobre os objetos filme e cinema, assim se realizou análises simples pela distribuição da frequência sob a mediana[9] das variáveis restantes que apresentam relevância de acordo com a experiencia de mercado, assim criando classe manuais como demonstrado na Tabela 2.

Tabela 2 – Recorte demonstração de agrupamento em classes da variável Exibidor.

Exibidor	Frequência	% Mediana	Classe
Centerplex	12493	133,33%	A
Cineart	10670	144,44%	A
Grupo Cine	15778	100%	B
Moviecom Cinemas	15564	100%	B
Cine A	12356	77,78%	B
Espaço Itaú	8430	144,44%	C

Cinemais	7266	155,56%	C
Circuito Cinemas	3370	155,56%	C
Cineplus Cinemas	3077	133,33%	C
CineX	2790	155,56%	C
PMC Cinemas	2624	155,56%	C
Multicine Cinemas	5242	77,78%	D
Cine Bom Vizinho	2085	88,89%	D

Cada classe será uma dummy de sua respectiva variável, assim se tem estabelecido um dos processo de criação de variáveis artificiais. As variáveis que passaram por esse processo são: Cinema, Exibidor, QTD_SALAS, GENERO e Distribuidor.

Criou-se também as variáveis temporais de ano e semana de exibição das sessões, mês e ano de estreia do filme sendo exibido, a quantia de gêneros e dias em exibição. A base final que seria utilizada para criação dos modelos em diversas técnicas de aprendizado de máquina e redes neurais[9] ficou com as variáveis apresentadas na Tabela 3.

Tabela 3: Dicionario de Dados da base pós-processamento.

Variável	Descrição	Tipo de Dados
Brasileiro	Indica se o filme é brasileiro	Booleano
ClassificacaoIndicativa	Classificação etária do filme	Inteiro
Duracao	Duração em minutos do filme	Inteiro
Comercial	Indica se o filme é comercial ou cultural	Booleano
Cinema_A	Indica presença ou ausência do Cinema A	Booleano
Cinema_B	Indica presença ou ausência do Cinema B	Booleano
Cinema_C	Indica presença ou ausência do Cinema C	Booleano
Cinema_D	Indica presença ou ausência do Cinema D	Booleano
Cinema_E	Indica presença ou ausência do Cinema E	Booleano
Cinema_F	Indica presença ou ausência do Cinema F	Booleano
Exibidor_A	Indica presença ou ausência do Exibidor A	Booleano
Exibidor_B	Indica presença ou ausência do Exibidor B	Booleano
Exibidor_C	Indica presença ou ausência do Exibidor C	Booleano
Exibidor_D	Indica presença ou ausência do Exibidor D	Booleano
salas_A	Quantidade de salas no Cinema A	Booleano
salas_B	Quantidade de salas no Cinema B	Booleano
salas_C	Quantidade de salas no Cinema C	Booleano
salas_D	Quantidade de salas no Cinema D	Booleano
Genero_A	Indica presença ou ausência do Gênero A	Booleano
Genero_B	Indica presença ou ausência do Gênero B	Booleano
Genero_C	Indica presença ou ausência do Gênero C	Booleano
Genero_D	Indica presença ou ausência do Gênero D	Booleano
Genero_E	Indica presença ou ausência do Gênero E	Booleano
Genero_F	Indica presença ou ausência do Gênero F	Booleano

Genero_G	Indica presença ou ausência do Gênero G	Booleano
Genero_H	Indica presença ou ausência do Gênero H	Booleano
Distribuidor_A	Indica presença ou ausência do Distribuidor A	Booleano
Distribuidor_B	Indica presença ou ausência do Distribuidor B	Booleano
Distribuidor_C	Indica presença ou ausência do Distribuidor C	Booleano
Distribuidor_D	Indica presença ou ausência do Distribuidor D	Booleano
QtdGenero	Quantidade de Gêneros do Filme	Número Inteiro
semana	Semana de Exibição	Número Inteiro
ano	Ano de Exibição	Número Inteiro
MesEstreia	Mês de Estreia do Filme	Número Inteiro
AnoEstreia	Ano de Estreia do Filme	Número Inteiro
QtdSessoes	Quantidade de Sessões Exibidas(Alvo)	Número Inteiro
DiasEmExibicao	Dias em Exibição	Número Inteiro

5 AVALIAÇÃO

Como principal métrica de avaliação será utilizado o erro médio absoluto, devido aos dados em si serem agregações para relatórios, possibilitando uma avaliação mais direta da distribuição dos erros pela base, como pontos onde esses erros ocorrem com maior frequência, o objetivo é ter uma previsão com erros homogêneos, e não casos onde temos grupos onde a previsão é muito precisa e outros onde a previsão é péssima, se faz necessário que o modelo erre de forma parecida pela base.

Dentre as técnicas utilizadas temos a rede neural *Multilayer perceptron (MLP)*, junto as técnicas de aprendizado de máquina *Random Forest* e *XGBoost* , a Tabela 4 indica os resultados.

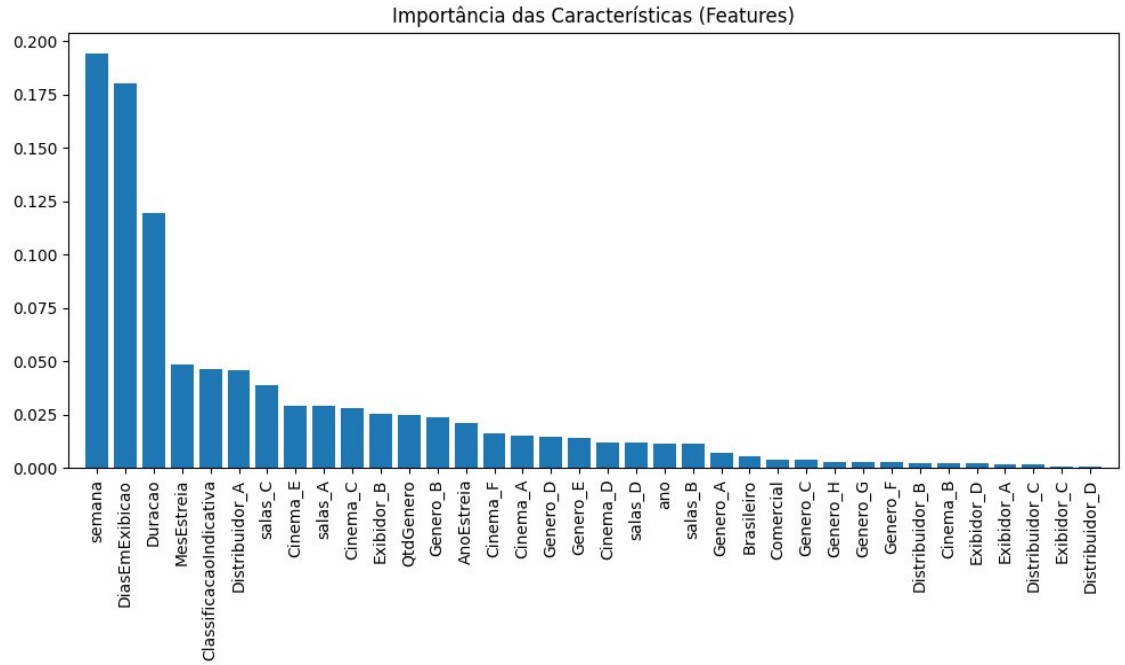
Tabela 4 – Resultados das técnicas.

Técnica	Erro Médio Absoluto
RandomForest	4.896
MLP	6.669
XGBoost	6.268

A técnica de aprendizado de máquina[10] *Random Forest* apresentou melhores resultados, que após análise da distribuição desses observa-se valores homogêneos em sua distribuição nas variáveis de mes, classificacaoIndicativa, Comercial, Estado e Distribuidora indicando que não a locais isolados ou regiões afetadas de forma extrema pelos erros.

Ao se analisar a forma como o modelo resultante da técnica de *Random Forest* utilizou as variáveis geradas(Tabela 3) na Figura 2, temos a confirmação que todos os objetos descritos na Tabela 1 possuem importância para o modelo, garantindo assim que as previsões geradas não som causadas pelo acaso ou vícios durante o desenvolvimento.

Figura 2 - Gráfico de Importância de Características



Todavia a variável de distribuidora apesar de em grande parte estar homogênea, alguns indivíduos estão causando distorção do resultado, com previsões acima do erro médio calculado como apresentado na Tabela 5, já que setenta e dois dos oitenta e quatro exibidores apresentam um erro médio absoluto entre dois e menos dois.

Tabela 5 – Trecho de avaliação de distribuição de erros.

Distribuidor	Erro médio Absoluto
Movimento Cinema Bruto	13,65
Plural Filmes	5,20
Livres Filmes	4,85
Espaço Filmes	4,14
Riofilme	3,97
20th Century Fox	3,44
Lumière	2,66
Vilabela Produções Artísticas	2,29
Kinoscópio	1,95

6 CONSIDERAÇÕES FINAIS

Como resultado do estudo temos como confirmar a possibilidade do uso de técnicas de aprendizado de máquina no problema apresentado, já que foi possível obter previsões com margens de erro médio medianas, e com uma distribuição geral dos erros homogênea, o que indica que o modelo não esta apenas adivinhando os valores, mas realmente prevendo o comportamento dos dados.

Os principais pontos de melhora que oferecem melhor confiabilidade do modelo apresentado é a adição de dados de filmes na língua inglesa, já que sua eficiência se apresenta como possibilidade para adição de variáveis sintéticas.

Recomenda-se a integração com serviços externos de catalogação de filmes, como IMDB e TMDB. Tais plataformas possuem dados de popularidade, notas, relevância, comentários do publico entre diversos outros, o que possibilita a geração de novas variáveis para melhoria do modelo.

Outro ponto possível de melhora é a criação de modelos intermediários para os objetos abordados, um modelo que permita gerar um escore para os filmes, junto a um modelo que possa trabalhar com os atores e acompanhar as mudanças da percepção do publico a eles, já que grande parte das variáveis relacionada a filmes apresentaram baixa relevância para as previsões.

REFERENCIAS

[1] Souza Junior, Edelvicio . (2023). Embrapii. “Desafios para o uso da inteligência artificial no Brasil”. Disponível em: https://embrapii.org.br/desafios-uso-inteligencia-artificial-brasil/.

[2]Rist, Peter. “A Brief Introduction to Brazilian Cinema”. Disponível em: https://offscreen.com/view/intro_braziliancinema.

[3]Chalréot, Fernando . (2016). ILOS. Distribuição Cinematográfica: como um filme chega ao cinema? Disponível em: https://ilos.com.br/distribuicao-cinematografica-como-um-filme-chega-um-cinema-perto-de-voce/.

[4]BERNARDA LUDERMIR, TERESA, Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências, Universidade Federal de Pernambuco, Centro de Informática, Recife, Pernambuco, Brasil, 2021.

[5]Velox Tickets.(2023). Disponível em: veloxtickets.com.

[6]SILVA, T. A., & VIEIRA, R. (2013). Análise de sentimentos em artigos de opinião. Revista de Informática Teórica e Aplicada, 20(2), 179-198.

[7] ARKIN, DANIEL.(2022).NBC News. “Inside the key allegations in the Johnny Depp-Amber Heard trial”. Disponível em: ttps://www.nbcnews.com/pop-culture/pop-culture-news/key-allegations-johnny-depp-amber-heard-trial-rcna30147.

[8]FIORIN, D. V., MARTINS, F. R., SCHUCH, N. J., & PEREIRA, E. B. (2011). Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. Revista Brasileira de Ensino de Física, 33(1), 1309. Disponível em: SciELO.

[9]ROSSI, R. M., MARTINS, E. N., LOPES, P. S., & SILVA, F. F. (2019). Efeito do uso de bioestimulante na qualidade de sementes de soja.

[10]NICOLA, M. J., & DE OLIVEIRA, J. P. M. (2021). Adoção de random forest e regressão linear para previsão de demanda de energia elétrica: um estudo comparativo com redes neurais MLP e XGBoost. Disponível em: teses.usp.br.