Análise de Sucesso de Filmes

Classificação dos Dados do "The Movie Database"

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

24/04/2023

FIA BUSINESS SCHOOL lab.data

Pós Graduação Lato Sensu MBA Analytics e Inteligência Artificial Data Science

Nome do Aluno:
Rafael de Queiroz Nunes

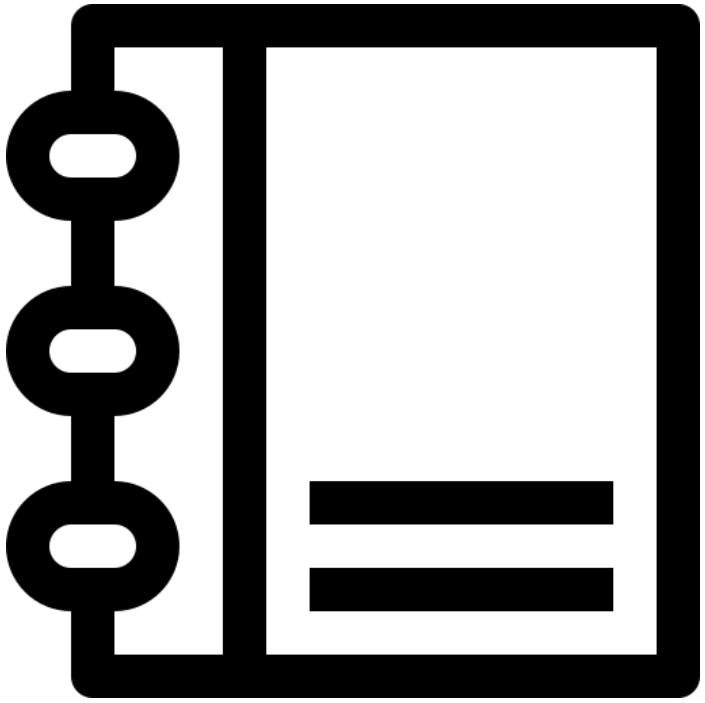
Coordenadores:
Prof.ª Dr.ª Alessandra de Álvila Montini
Prof. Dr. João Paulo da Costa Nogueira
Prof. Dr.ª Nina Maria Irma Soares Pinheiro Machado

Entrega:
24/04/2023

FIA BUSINESS SCHOOL lab.data

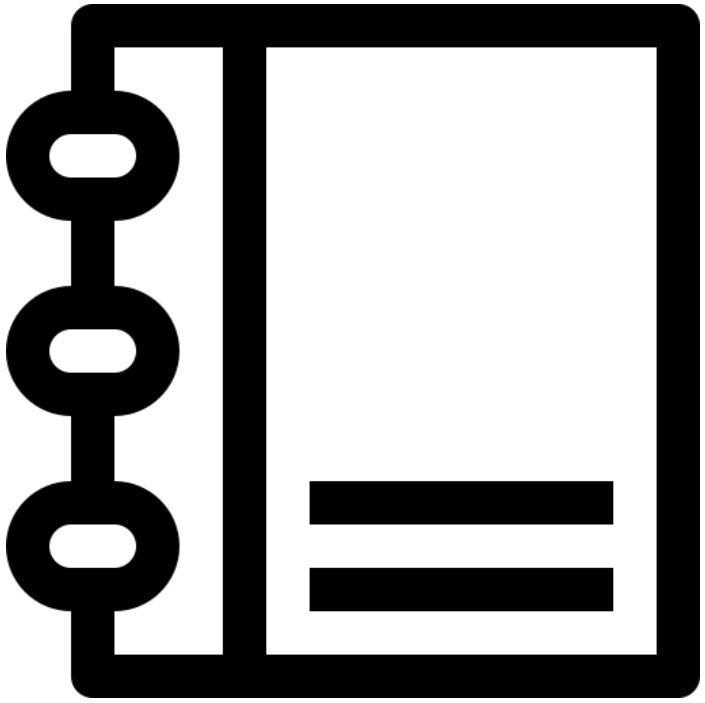
LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Agenda



1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    i. Bases Originais
    ii. Geração da *Analytical Base Table* (ABT)
    iii. Descrição das variáveis da ABT
4. Análise Exploratória de Dados (ABT)
5. Modelagem Estatística – Classificação
6. Conclusão
7. Próximos Passos (Entrega 3)

@2020 FIA LABDATA. Copyright all rights reserved.



@2020 FIA LABDATA. Copyright all rights reserved.

1. Objetivo do Trabalho

O objetivo do trabalho é **predizer se um filme consegue ou não fazer sucesso financeiro** com base seu orçamento e no faturamento de suas bilheterias.

A predição será realizada utilizando dados abertos do **"The Movie Database"**, que possui suas informações preenchidas pela comunidade do site. Para a predição deste trabalho, **serão aplicados modelos de classificação**, nos quais avaliação o sucesso do filme **se sua receita foi o dobro ou mais do que o seu orçamento**.

Desta forma, será analisado quais fatores são os mais determinantes, com base nas informações publicamente divulgadas, se um filme fez algum sucesso ou não.

@2020 FIA LABDATA. Copyright all rights reserved.

2. Contextualização do Problema

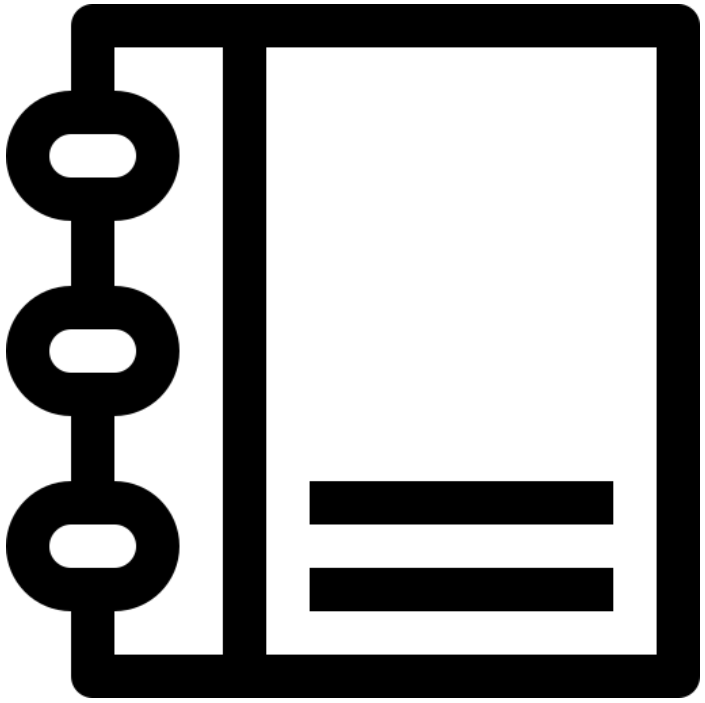
A indústria cinematográfica é uma das que mais investe dinheiro em projetos, com produção e marketing, o que torna a indústria muito arriscada. Portanto, é importante ter uma ideia precisa de quais filmes serão populares e rentáveis antes de investir.

Os principais fatores para o sucesso de um filme são:

- **Qualidade da produção**: a qualidade de roteiro, elenco, direção e edição do filme;
- **Marketing**: divulgação e promoção. Uma campanha de marketing mal elaborada pode resultar em baixa bilheteria, mesmo que o filme seja bom;
- **Competição**: há diversos filmes, muitos de gêneros similares, sendo lançados a cada semana. Portanto, é importante saber quais filmes têm maior probabilidade de se destacar em meio à concorrência;
- **Tendências**: análises de sentimentos, por exemplo, pode-se determinar se a trama, franquia, personagens ou cinematografia por exemplo são mais valorizados pelos espectadores. Isso pode ser útil para orientar a produção de filmes mais atraentes e bem-sucedidos.

A predição de machine learning pode ajudar a determinar quais filmes terão maior aceitação pelo público, com base em dados de bilheteria, gênero, críticas, elenco, entre outros. Pode também determinar quais elementos dos filmes são mais importantes para a satisfação do público.

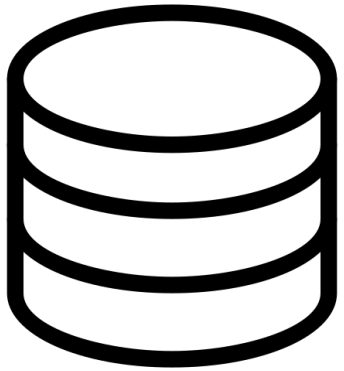
@2020 FIA LABDATA. Copyright all rights reserved.



@2020 FIA LABDATA. Copyright all rights reserved.

3.i Base Original

3. BASE DE DADOS



Origem

- Os dados dos filmes para este trabalho foram extraídos da API do The Movie Database (TMDB), através do pacote do Python **tmdbv3api**, onde os dados são abertos e alimentados pela comunidade do site.

Base de Dados Bruta Gerada

- Através do pacote e de sua extração filme a filme, foram gerados as seguintes bases iniciais para este trabalho:

Tabela	Descrição
df_movies_list	Base principal de filmes, que contém suas principais características
df_movie_belongs_to_collection	Base de coleção de filmes classificada no TMDB
df_movie_genres	Base que relaciona os filmes aos seus gêneros
df_movie_production_companies	Base que relaciona as empresas produtoras do filme
df_movie_production_countries	Base que relaciona os países de produção destes filmes
df_movie_spoken_languages	Base que relaciona os idiomas falados nos filmes

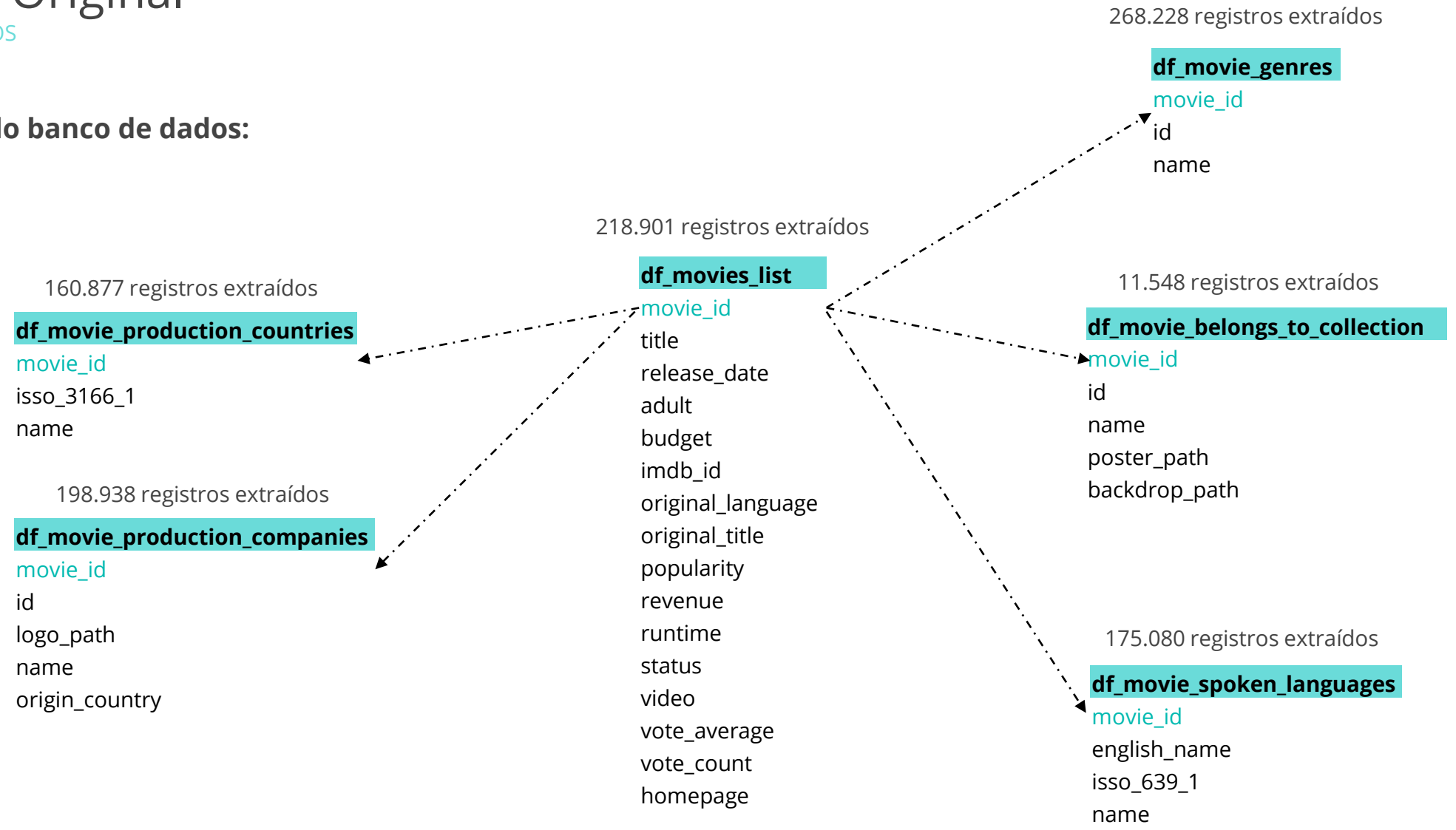
- O relacionamento de todas as bases é pelo campo 'movie_id' centralizado na base 'df_movies_list';
- Uma amostra da base está foi gerada por conta do limite de downloada da API;
- Valores de filmes válidos com 'movie_id' extraídos:
  - Entre 1 a 200.000;
  - Entre 1.000.001 até 1.113.517;
  - Lembrando que não há filmes para todos as correspondências no intervalo mostrado.

@2020 FIA LABDATA. Copyright all rights reserved.

3.i Base Original

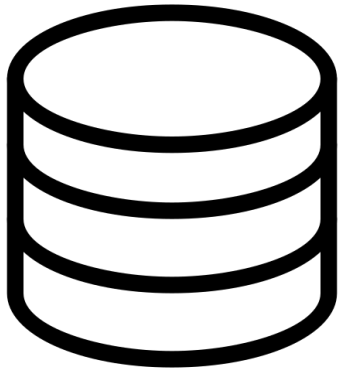
3. BASE DE DADOS

Visão geral do banco de dados:



@2020 FIA LABDATA. Copyright all rights reserved.

3.ii Geração da Analytical Base Table (ABT)



Filtros executados

- Filtrado filmes apenas com campo 'released' na tabela 'df_movies_list';
- Filtrado filmes apenas com campo 'budget' maior que zero na tabela 'df_movies_list';
- Filtrado filmes apenas com campo 'revenue' com ao menos 1% do valor de 'budget', ambos da 'df_movies_list';
- Filtrado filmes que possuem data de lançamento divulgada, 'release_date' da tabela 'df_movies_list'.

Números preliminares da *Analytical Base Table* (ABT) gerada das bases iniciais

- 6.596 registros;
- 32 campos;
- 19 campos derivados da base 'df_movies_genres' com valores binários de gêneros;
- Variável chave da ABT: "movie_id";
- Variável target da nossa classificação: "financial success".

@2020 FIA LABDATA. Copyright all rights reserved.

3.iii Descrição das variáveis da ABT

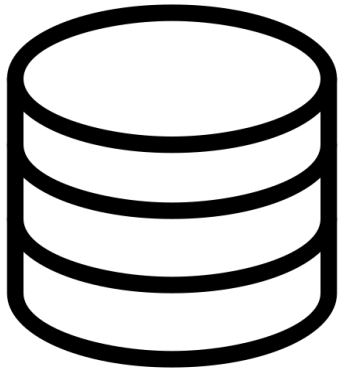
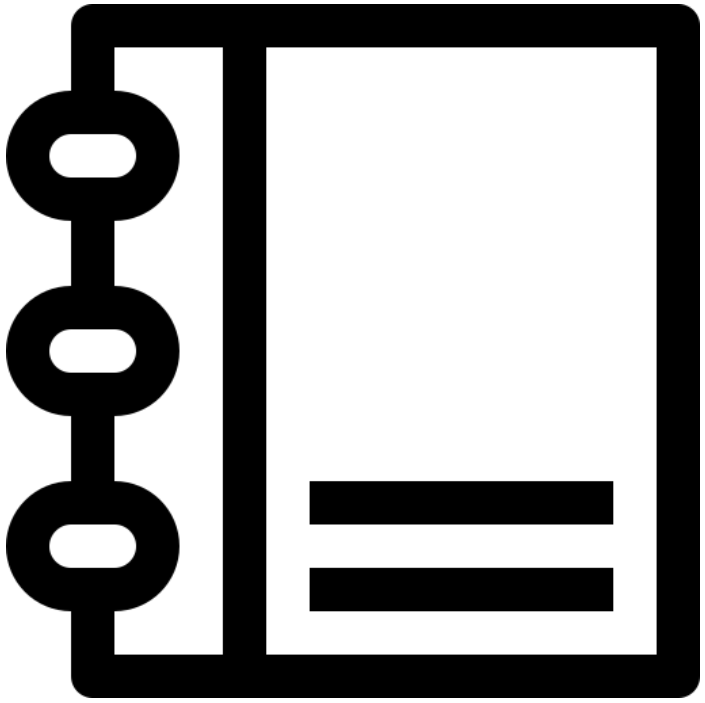


Tabela descritiva das variáveis que serão

Campo	Tipo de Dado	Variável Qualitativa ou Quantitativa	Descrição
movie_id	Inteiro (chave primária)	Qualitativa	ID dos filmes da TMDB
adult	Binário	Qualitativa	Classifica se o filme é adulto ou não
budget	Inteiro	Quantitativa	Orçamento informado no site
original_language	Texto	Qualitativa	Idioma de origem do filme
runtime	Decimal	Quantitativa	Tempo de filme em minutos
vote_average	Decimal	Quantitativa	Média de votos do público na plataforma TMDB
vote_count	Inteiro	Quantitativa	Quantidade de votos do público no TMDB
nmonth_release	Inteiro	Qualitativa	Mês de lançamento do filme
has_collection	Binário	Qualitativa	Possui alguma coleção classificada no TMDB
spoken_languages	Inteiro	Quantitativa	Quantidade de idiomas de falados originalmente no filme
genre ... *(19 campos)*	Binário	Qualitativa	Traz a classificação binária do gênero do filme, em que um filme pode estar classificado em mais de um gênero. São 19 campos classificados em: ação, aventura, animação, comédia, crime, documentário, drama, família, fantasia, história, horror, música, mistério, romance, ficção científica, filme para tv, suspense, guerra, faroeste
financial_success	Binário (target)	Qualitativa	Classificação se o filme fez ou não sucesso, considerando que a receita deve ser o dobro do orçamento. Se sim, será um, caso contrário, zero.

@2020 FIA LABDATA. Copyright all rights reserved.



@2020 FIA LABDATA. Copyright all rights reserved.

4.i Principais estatísticas (variáveis quantitativas)

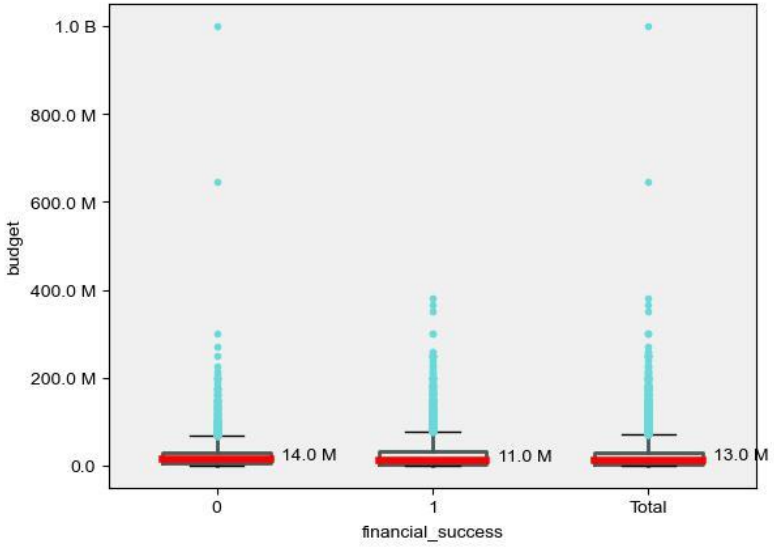
Principais estatísticas dos campos quantitativos.

	budget	runtime	vote_average	vote_count	spoken_languages
Média	26.064.225,4	109,4	6,4	1.603,2	1,5
Desvio Padrão	39.265.014,3	22,8	1,0	2.968,3	0,9
Coeficiênte de Variação	0,7	4,8	6,5	0,5	1,6
Mínimo	103.000,0	0,0	0,0	0,0	1,0
Quartil inferior (P25)	3.800.000,0	95,0	5,8	128,0	1,0
Mediana (P50)	13.000.000,0	105,0	6,4	512,0	1,0
Quartil superior (P75)	30.625.000,0	120,0	7,0	1.637,5	2,0
Máximo	999.999.999,0	339,0	10,0	33.473,0	11,0

@2020 FIA LABDATA. Copyright all rights reserved.

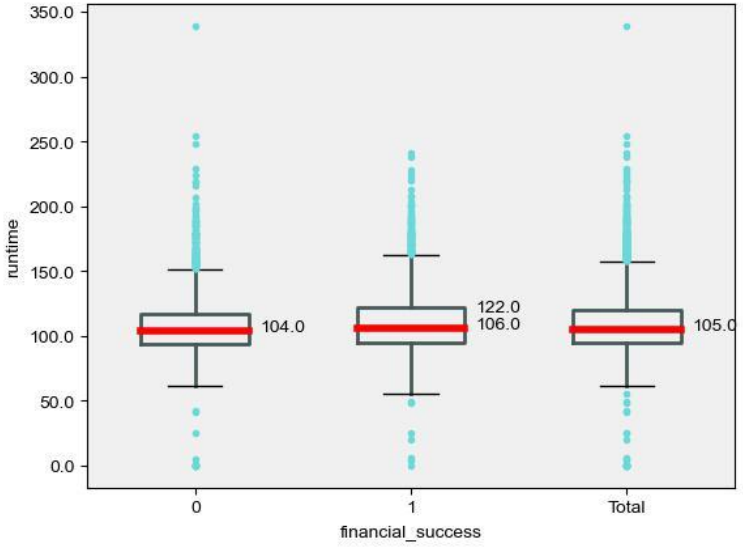
4.i Box-plots (variáveis quantitativas)

budget



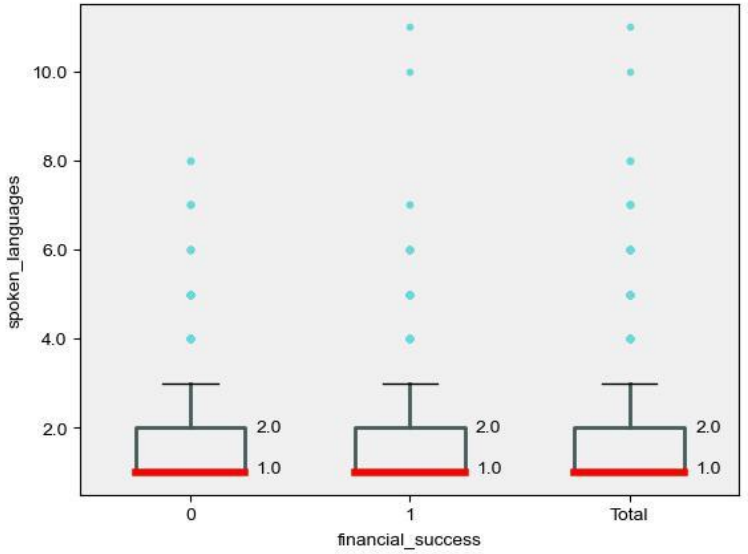
- A grande maioria dos orçamentos giram na média de 26 milhões, com mediana de 13;
- No entanto, percebe-se vários outliers com filmes de grandes orçamentos;
- Não se percebe uma diferença nos boxplots em relação à variável 'financial_success".

runtime



- A amostra possui média e mediana, em torno de 105, similares, sem outliers muito chamativos;
- Não se percebe uma diferença nos boxplots em relação à variável 'financial_success".

spoken_languages

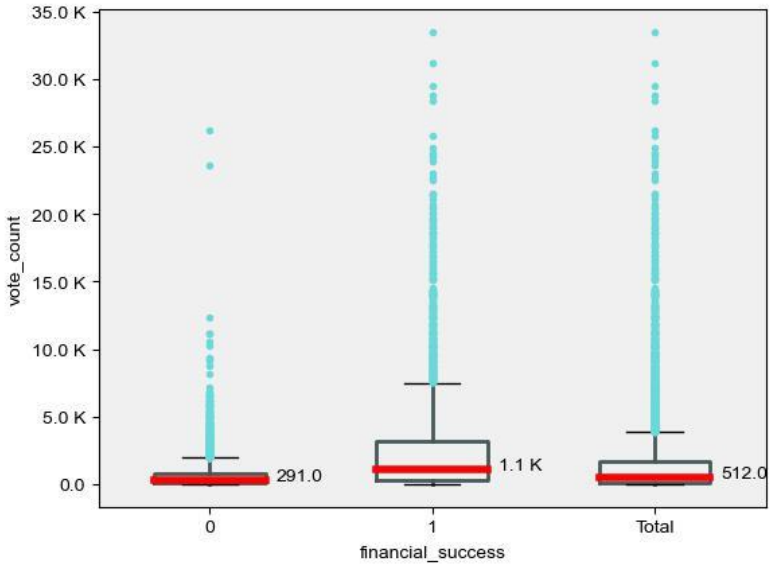


- Existem alguns outliers nos idiomas falados, mostrando que há exceções na questão dos idiomas que são falados no filme;
- Praticamente não há diferença nos boxplots em relação à variável 'financial_success".

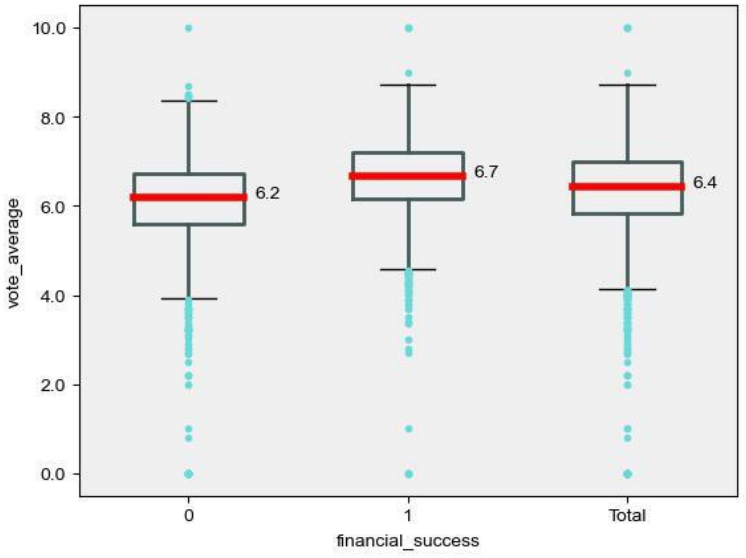
@2020 FIA LABDATA. Copyright all rights reserved.

4.i Box-plots (variáveis quantitativas)

vote_count



vote_average



- A média de votos possui inúmeros outliers superiores, o que pode refletir a popularidade desses filmes em si na plataforma TMDB;
- Temos uma clara distinção pelos boxplots por 'financial_success', em que filmes com 'vote_count' maiores tende a ter mais chanves de sucesso.

- Média de votos com 6.4, com vários outliers para abaixo, e alguns poucos superiores;
- Não se percebe uma diferença nos boxplots em relação à variável 'financial_success".

@2020 FIA LABDATA. Copyright all rights reserved.

4.ii Tabela de Frequência e Análise (variáveis qualitativas)

Tabela distribuição de frequência da variável target 'financial_success'

Campo	Valor	Quantidade Absoluta	Quantidade Relativa (%)
financial_success	0	3.388	51,4
	1	3.207	48,6

Tabela distribuição de frequência das variáveis binárias

Campo	Valor	Quantidade Absoluta	Quantidade Relativa (%)	% 'finantial_success' = 1*
adult	0	6.591	99,9	48,6%
	1	4	0,1	50,0%
has_collection	0	5.065	76,8	41,3%
	1	1.530	23,2	72,9%
genre Action	0	4.986	75,6	49,1%
	1	1.609	24,4	47,1%
genre Adventure	0	5.488	83,2	47,8%
	1	1.107	16,8	52,8%
genre Animation	0	6.271	95,1	48,4%
	1	324	4,9	53,4%
genre Comedy	0	4.155	63,0	47,2%
	1	2.440	37,0	51,1%
genre Crime	0	5.526	83,8	49,0%
	1	1.069	16,2	46,7%
genre Documentary	0	6.518	98,8	48,6%
	1	77	1,2	54,5%
genre Drama	0	3.475	52,7	52,4%
	1	3.120	47,3	44,4%
genre Family	0	5.930	89,9	47,9%
	1	665	10,1	55,3%

**"% 'financial_success' = 1" refere-se ao percentual de valores igual a um da variável target 'financial_success'.*

- Percebe-se que a variável target 'financial_success' possui dados bem distribuidos.

- Apenas a variável binária 'has_collection' apresentou mostrar um diferencial na quantidade de filmes com sucesso, mostrando que filmes categorizados com coleção no TMDB possui mais propensão a terem sucesso;
- As outras variáveis binárias, 'adult' e as de gênero ('genre ...') não mostraram ter variações significativas no percentual de sucesso, mostrando uma possível baixa correlação e assim possivelmente não influenciariam no modelo como deveriam.

@2020 FIA LABDATA. Copyright all rights reserved.

4.ii Tabela de Frequência e Análise (variáveis qualitativas)

Tabela distribuição de frequência das variáveis binárias

Campo	Valor	Quantidade Absoluta	Quantidade Relativa (%)	% 'finantial_success' = 1*
adult	0	6.591	99,9	48,6%
	1	4	0,1	50,0%
has_collection	0	5.065	76,8	41,3%
	1	1.530	23,2	72,9%
genre Action	0	4.986	75,6	49,1%
	1	1.609	24,4	47,1%
genre Adventure	0	5.488	83,2	47,8%
	1	1.107	16,8	52,8%
genre Animation	0	6.271	95,1	48,4%
	1	324	4,9	53,4%
genre Comedy	0	4.155	63,0	47,2%
	1	2.440	37,0	51,1%
genre Crime	0	5.526	83,8	49,0%
	1	1.069	16,2	46,7%
genre Documentary	0	6.518	98,8	48,6%
	1	77	1,2	54,5%
genre Drama	0	3.475	52,7	52,4%
	1	3.120	47,3	44,4%
genre Family	0	5.930	89,9	47,9%
	1	665	10,1	55,3%

- As outras variáveis binárias, ‘adult’ e as de gênero (‘genre ...’) não mostraram ter variações significativas no percentual de sucesso, mostrando uma possível baixa correlação e assim possivelmente não influenciariam no modelo como deveriam.

**”% ‘financial_success’ = 1” refere-se ao percentual de valores igual a um da variável target ‘financial_success’.*

@2020 FIA LABDATA. Copyright all rights reserved.

4.ii Tabela de Frequência e Análise (variáveis qualitativas)

Tabela distribuição de frequência da variável 'original_language', com os top 10 valores mais frequentes

original_language	Quantidade Absoluta	Quantidade Relativa (%)	% 'finantial_success' = 1*
en	5.542	84,0	49,7%
hi	184	2,8	54,9%
fr	155	2,4	31,6%
ru	101	1,5	31,7%
ja	83	1,3	45,8%
ta	74	1,1	47,3%
es	68	1,0	42,6%
it	48	0,7	35,4%
de	40	0,6	40,0%
zh	38	0,6	50,0%

- 'original_language' possui praticamente todos os valores 'en', lingua inglesa, representando 84% de todos os valores
- No top 10 valores da variável, não se verifica grandes correlações com a variável target.

Tabela distribuição de frequência da variável 'nmonth_release'

nmonth_release	Quantidade Absoluta	Quantidade Relativa (%)	% 'finantial_success' = 1*
1	393	6,0	45,5%
2	441	6,7	46,7%
3	532	8,1	47,4%
4	502	7,6	44,6%
5	518	7,9	50,8%
6	552	8,4	59,6%
7	509	7,7	55,8%
8	590	8,9	45,8%
9	709	10,8	38,4%
10	628	9,5	44,3%
11	508	7,7	51,6%
12	713	10,8	54,4%

- 'nmonth_release' mostra um bom equilíbrio nos valores, verificando que a amostra tem valores bem distribuidos de períodos de lançamento;
- Também não se verifica grandes correlações com a variável target, aparentemente não há meses de maior sucesso nas bilheterias.

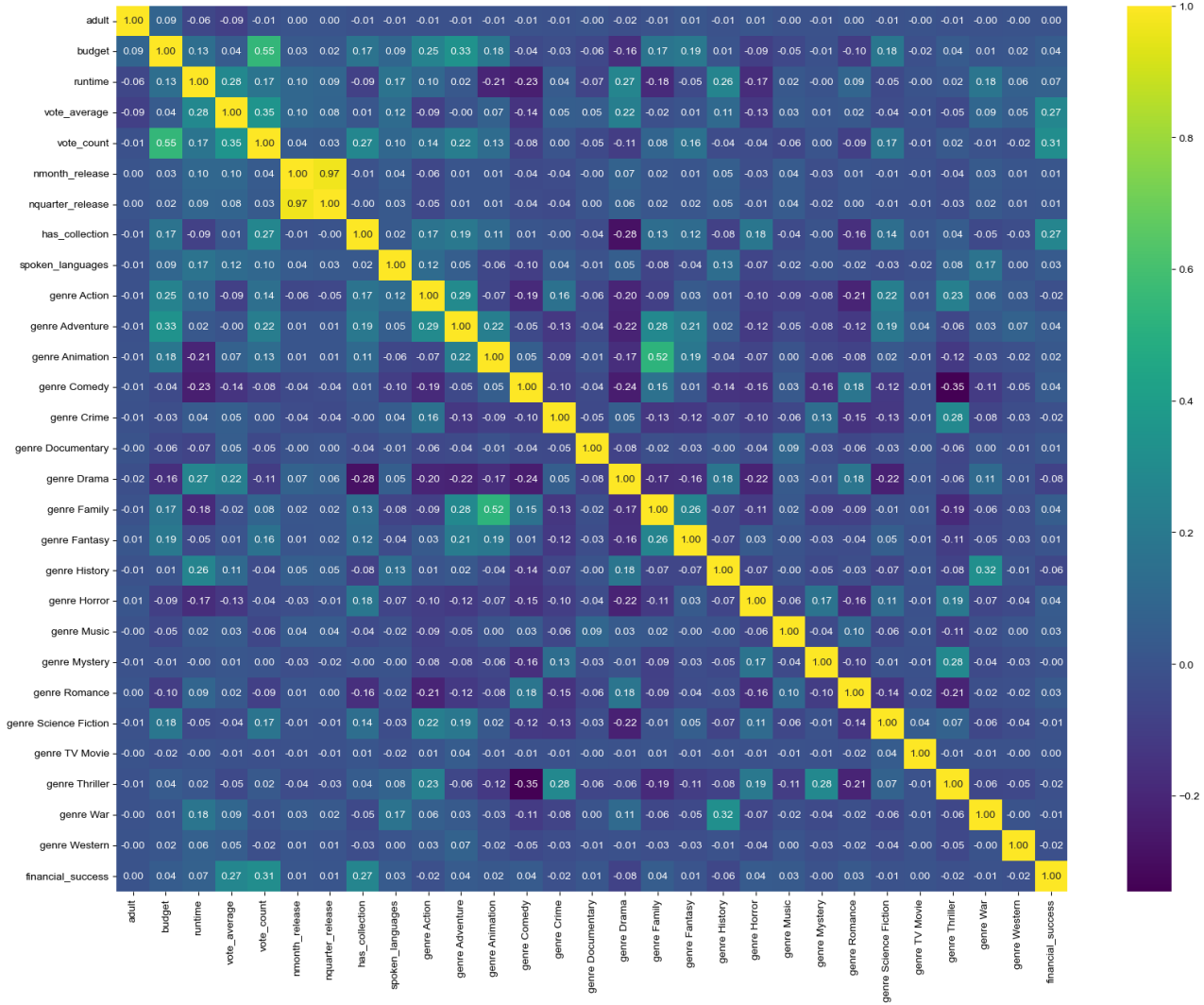
"% 'financial_success' = 1" refere-se ao percentual de valores igual a um da variável target 'financial_success'.

@2020 FIA LABDATA. Copyright all rights reserved.

FIA BUSINESS SCHOOL lab.data

4.iii Tabela de Correlação das variáveis

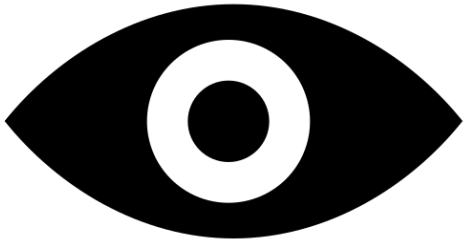
Tabela de correlação das variáveis do modelo de classificação de filmes



- A variavel target (‘financial_success’) mostrou uma correlação maior com as variáveis: ‘vote_average’, ‘vote_count’ e ‘has_collection’;
- Destaque também para uma correlaçao grande para ‘budget’ e ‘vote_count’, o que indica que filmes que possuem maior orçamento chamam a atenção para mais votos;
- Existem correlações mais altas para as variáveis de gênero (‘genre ...”), mas elas não trazem grandes preocupações à nossa análise.

@2020 FIA LABDATA. Copyright all rights reserved.

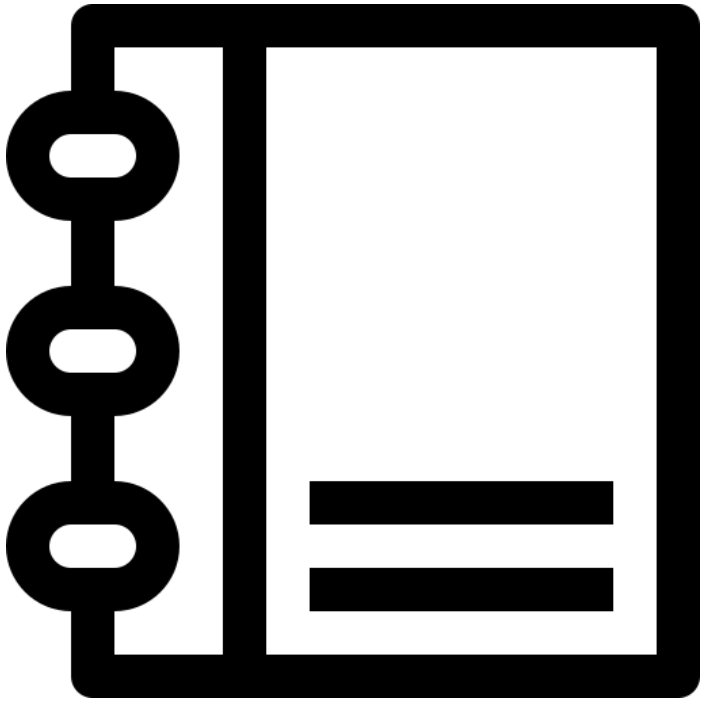
4.iv Principais pontos da análise exploratória



- As variáveis em geral aparentam ter um equilíbrio na distribuição de seus valores para a modelagem preditiva, o que é positivo para resultados mais fidedignos;
- Na análise exploratória bivariável com a variável target 'financial_success', apenas as variáveis 'vote_count' e 'has collection' possuem diferenças em medianas conforme visto no boxplot, e tem possui uma correlação maior segundo a tabela de correlação;
- 'vote_average' aparece como uma correlação maior com 'financial_success' na sua tabela de correlaçao.

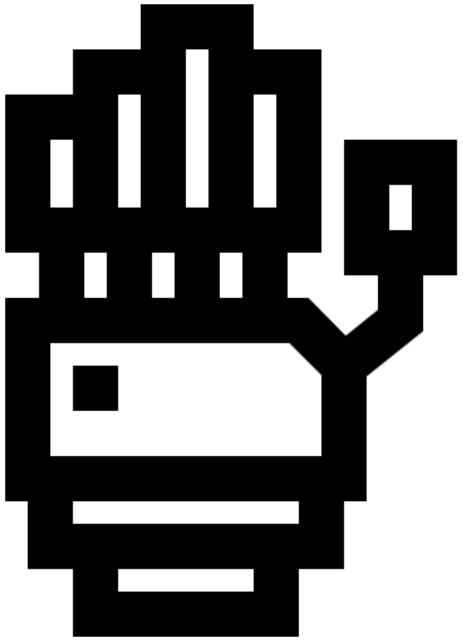
*"% 'financial_success' = 1" refere-se ao percentual de valores igual a um da variável target 'financial_success'.

@2020 FIA LABDATA. Copyright all rights reserved.



@2020 FIA LABDATA. Copyright all rights reserved.

Explicação da Modelagem de Classificação



O modelo de classificação foi trabalhado da seguinte forma:

- Foi dividido entre treino e teste, sendo 70% e 30% respectivamente;
- Serão testados vários modelos de classificação, nos quais sofreram Cross Validation para determinar qual o mais adequado para a modelagem
- A metrica usada para a seleção é a de precisão;
  - A métrica de precisão foi escolhida nesse pois é preciso maximizar as chances de encontrarmos filmes que terão sucesso de bilheteria;
- O melhor modelo será usado para ser rodado e comparado com a base de teste enfim.

@2020 FIA LABDATA. Copyright all rights reserved.

FIA BUSINESS SCHOOL lab.data

FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO | LABDATA

Cross-Validation, Treino e Teste

5. MODELAGEM ESTATÍSTICA – CLASSIFICAÇÃO

Resultados das métricas médias dos testes de Cross Validation nos modelos de classificação testados

Modelo de ML	Tipo	Precisão	Acurácia	AUC	Recall	F1
Gradient Boosting Classifier	Árvore	75,1%	74,6%	82,5%	71,7%	73,3%
Cat Boost Classifier	Árvore	74,8%	74,7%	82,9%	72,7%	73,7%
LGBM Classifier	Árvore	74,2%	74,2%	82,4%	72,2%	73,2%
Random Forest	Árvore	73,8%	73,5%	81,2%	71,1%	72,4%
Regressão Logísitca	Linear	73,2%	71,1%	78,3%	64,5%	68,6%
XGBoost	Árvore	72,4%	72,8%	80,6%	71,5%	71,9%
Support Vector Machines	Linear	70,7%	67,3%	74,8%	56,4%	62,7%
Árvore de Decisão	Árvore	67,4%	68,1%	68,1%	67,5%	67,4%

Resultados das métricas na modelagem do melhor modelo

Gradient Boosting Classifier (final)	Base de Treino	Base de Teste
Precisão	79,6%	74,1%
Acurácia	79,1%	74,4%
ROC/AUC	79,0%	74,4%
Recall	76,9%	72,4%
F1	78,2%	73,2%

- O modelo de **Gradient Boosting Classifier** foi o modelo com a maior precisão;
- Teve uma das maiores AUC e acurácia próximos dos valores do Cat Boost Classifier;
- O Cat Boost Classifier poderia ser uma boa escolha, devido a outros bons parâmetros como ter as maiores AUC e acurácia. Mas seguindo o critério estabelecido, foi feita a modelagem final com o primeiro modelo;
- Na modelagem final, percebe-se que tivemos quase 80% de precisão no treino, mas houve uma queda de 74% com a base de teste, que é uma queda baixa no modelo, assim ainda o consideramos confiável para novas predições.

@2020 FIA LABDATA. Copyright all rights reserved.

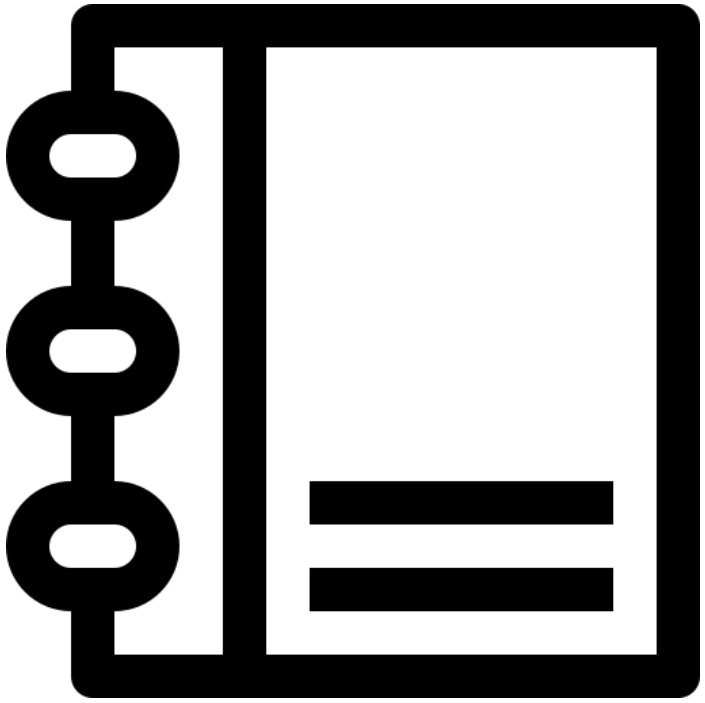
Importância das Features na Modelagem

% de importância das features / campos na modelagem do Gradient Boosting Classifier

Feature / Campo	Importância
vote_count	43,0%
budget	23,0%
has_collection	10,0%
vote_average	9,0%
runtime	3,0%
genre Romance	1,0%
original_language_fr	1,0%
genre Comedy	1,0%
genre Science Fiction	1,0%
genre Drama	1,0%

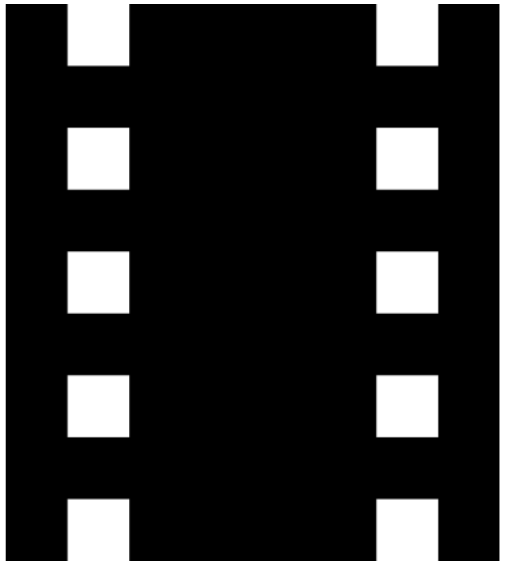
- No treinamento do modelo, foi avaliado a importância da variáveis, para se obter um feedback melhor deste treinamento e preparar melhorias para a entrega 3;
- Nota-se que a importância das variáveis seguiu em grande medida como esperado, onde houve uma grande importante para 'vote_count', 'has_collection' e 'vote_average';
- 'budget' foi uma surpresa em relação ao que foi analisado na análise exploratória, sendo a segunda variável mais importante. Considerando o caso, há sentido na importância já que grandes investimentos de filmes se esperam grande retornos de bilheteria;
- Após a variável 'runtime', não há variáveis com alguma importância de grande relevância no modelo. O que é concordante com o apresentado na análise exploratória.

@2020 FIA LABDATA. Copyright all rights reserved.



@2020 FIA LABDATA. Copyright all rights reserved.

6. Conclusão



- A base do TMDB possui diversos filmes, mas infelizmente não foram tantos adequados para a modelagem (de 218.901 para 6.596 filmes);
- A base possui as informações mais essenciais de filmes, mas ainda sim as variáveis não são ideias, seria necessário testar informações complementares;
- O modelo de **Gradient Boosting Classifier** foi o melhor modelo com base na medida 'precisão', e fez uma boa modelagem, apesar de serem apenas poucas variáveis as relevantes para o modelo;
- As variáveis relacionadas ao orçamento e aos votos e às classificações da comunidade do TMDB foram as mais importantes para se analisar o sucesso financeiro de um filme.

@2020 FIA LABDATA. Copyright all rights reserved.

7. Próximos Passos (Entrega 3)

- Obter toda a base do TMDB para poder fazer o modelo com 100% da base disponível;
- Cruzar com outras bases de dados se viável, como o banco de dados do IMDB;
- Fazer outras modelagens de inteligência artificial, como projetar a receita com base nas informações tratadas;
- Utilizar modelagem avançada de redes neurais com o modelo mostrado e também com novas variáveis nas quais não se podia fazer a modelagem (ex.: imagens, nome dos filmes, videos etc.)

@2020 FIA LABDATA. Copyright all rights reserved.