YouTube channel ranking to predict possible career successes

Oliveira, Gabriel R.[1]

Universidade Federal de Lavras, Brasil
gabriel.ribolive@gmail.com

Resumo A postagem de vídeos em plataformas de streaming, cada vez mais, se torna comum entre muitos artistas no ramo musical. Com isso, se tornar uma pessoa influente, que tenha boa audiência se torna cada vez mais difícil nesse mercado. Pensando nisso, esse artigo aborda um meio de encontrar os principais parâmetros que alguém precisa forca para investir ou melhores, para se tornar um possível canal de sucesso na plataforma Youtube. Através de técnicas de mineração, utilizando regressão linear, esse artigo responde quais os principais fatores que são determinantes na definição de um canal de sucesso.

Keywords: Cover · music · Data Mining · Youtube · Cover Analytics · Linear Regression

1 Introdução

Com o passar dos anos o universo da música passou por diversos períodos e diversas mudanças, principalmente com o surgimento da tecnologia digital. Antigamente, profissionais da área, publicavam suas músicas através de gravadoras e mesmo assim, antes da realização dessas publicações, as músicas precisavam passar por um rigoroso critério de seleção para decidir se seriam aceitos pela gravadora ou não.

E com esse decorrer, mídias streaming começaram a surgir, assim começaram a nascer cantores que gravavam músicas, até mesmo amadores com câmeras pessoais, e conseguiam publicar nessas plataformas de streaming gratuito e assim atingir um publico considerável.

O Youtube, criado em fevereiro de 2005, ilustra bem esse meio digital, onde vemos todos os dias publicações de músicas novas e cantores surgindo. Muitas vezes, nos deparamos com pessoas, que ainda não foram introduzidas, profissionalmente, no ramo musical, lançando músicas e cantando em vídeos, sejam eles caseiros ou não.

Assim, gravadoras passaram a observar pessoas que poderiam ser possíveis sucessos, para conseguirem investir nelas antes de outros, gerando assim uma renda maior para essas empresas. Atualmente, além dos trabalhos cotidianos, nessa busca por novos talentos as empresas (no ramo) ficam limitadas a conhecer pessoas através de um 'single', uma música desse artista que chega até eles,

através de famosos ou então alguma influencia que já tenha algum sucesso até o momento em audiência com tais publicações.

Com o passar dos anos houve uma grande evolução na maneira como músicas e artistas publicam seu conteúdo, logo, podemos observar a evolução em como empresas da área encontram possíveis artistas. Mesmo com essa grande evolução, paramos para observar essa diferença e concluímos que, podemos utilizar um pouco mais das tecnologias para facilitar o trabalho dessas empresas. Com isso, esse artigo, sugere uma abordagem de técnicas de predição a fim de responder a seguinte questão:

- **Q1:** Quais os fatores que são determinantes na definição de um canal de sucesso?

O Youtube conta com uma API[4] para que os desenvolvedores interessados possam obter alguns recursos (outros somente com permissões do canal analisado) com técnicas de requisições. Para o acesso, a API exige uma chave de acesso para as requisições e uma segunda chave, sendo ambos colocadas em um arquivo, com outras informações para a realização de funções na ferramenta. Utilizando a linguagem de programação Python, foi capturado as informações de pessoas que gravaram, sendo possíveis sucessos no Youtube, para analise.

Esse artigo então utiliza os dados capturados dos vídeos publicados que foram encontrados através da busca em canais específicos, escolhidos através de uma pesquisa de campo, determinando quais canais de sucessos seriam a base para nossa analise. Com a utilização de regressões lineares, concluir os principais fatores, em um vídeos, que realmente determina o sucesso de sua audiência.

2 Referencial Teórico

A cada dia que se passa, mais pessoas buscam encontrar técnicas de mineração para otimizar processos, ou encontrar padrões. Em uma revisão literária realizada por uma revista de sistemas médicos[13], uma análise de artigos postado em um período de 10 anos, correlacionando a área de saúde em a área de mineração de dados, encontrando algoritmos importantes, como K-means, K-medoids e X-means, entre outros de decisões, estatísticos e vários outros. Entre eles, foi encontrado algoritmos de regressão lógica. Como citado em um artigo de descobertas e conhecimentos na área de mineração[7], também foi encontrado uma vasta informação de dados co-relacionados ao KDD, que são técnicas de descobertas para os bancos e as analises que define partes importantes para se seguir:

- Entendimento da Regra de Negócio
- Seleção do conjunto de dados
- Limpeza e pré-processamento
- Projeção e redução de dados
- Escolha do algoritmo
- Análise do modelo e hipótese

– Busca por padrão
– Interpretação de dados
– Possíveis Ações na base de dados

Assim conseguirmos observar um processo um pouco mais exemplificado e com possíveis passos á se seguir para a mineração e possíveis estudo em dados capturados.

Com o grande aumento do volume de dados, empresas cada vez mais abordam técnicas para aproximas melhor seu mercado e obter possíveis lucros[6], se destacando dos outros. Observa-se técnicas de aprendizado supervisionado e não-supervisionado, exemplificando à necessidade de uma pré-categorização em algoritmos supervisionados enquanto outro requer diversas técnicas para comparações e assim decidir a melhor situação para tal dado/informação.

As tarefas podem ser agrupadas em atividades preditivas e descritivas[8]. Com isso é identificado métodos importantes para a realização da analise, como redes neurais, algoritmos genéticos, lógica nebulosa, estatística e árvore de decisão. Algumas técnicas de classificação, agrupamento e de extração[5], trazem algumas abordagens de métodos para relacionar e tratar a onda de informações textuais que vêm sendo coletadas durante os anos. Formas de tratamentos como identificação em minerações de opiniões, formas de representações de textos e pré-processamentos, sendo de grande utilidade ao trabalho abordado, para possíveis analisas de sentimentos, concentrando ideias há trabalhos futuro com tratamentos de opiniões e sentimentos existentes em frases extraídas.

Diferentes métodos para a mineração de dados que são reconhecidos pela conferência IEEE-ICDM[12], entre eles, alguns de classificação e regressão demonstrando diversas maneiras para se chegar em uma solução. Diversos tipos de regressão podem ser analisados, como Regressões lineares, multivariadas, as polinomiais e árvores de decisão, assim é apresentado o algoritmo “CART” que utiliza uma árvore binária capaz de processar atributos contínuos e nominais.

Ao trabalharmos com dados relacionando com padrões multimídia, é recomendado que faça um estudo prévio de informações relacionadas a valores e informações que podem ser capturadas a partir de uma API [4] de uma rede de mídia social. O LARM (Lifetime Aware Regression Model)[10], para previsões em popularidade no Youtube, introduz um estudo de popularidade de vídeos utilizando métodos de regressão, relacionando diretamente à pergunta chave deste trabalho, que nada mais é estabelecer fatores que levam vídeos a sua popularidade.

A apresentação de algoritmos e regressões, empregadas a análises estatísticas e outros métodos. Um dos procedimentos de estimação mais utilizados na prática, consiste em, a partir de técnicas de diagnóstico, detectar e rejeitar os "outliers"[11], variáveis com valores muito grandes ou muito pequenas, em relação a média dos outros valores. A estilização de gráficos é de grande importância para visualização em regressões, através de métodos de uma ferramenta chamada "Matplotlib"[9][2], os gráficos e possíveis códigos de auxilio para estes são utilizados para melhor compreensão.

3 Metodologia

Nesta seção será apresentada a metodologia utilizada para o desenvolvimento do trabalho.

3.1 Coleta de Dados

A plataforma de streaming da Google, o Youtube, foi escolhido pela facilidade de encontrar canais e usuários que realmente gerassem conteúdos na plataforma. A API, fornecida pela própria ferramenta, disponibiliza diversas funções para a interação entre desenvolvedor e à aplicação.

A API de dados oferecido pela ferramenta, juntamente, com a linguagem de programação Python, é utilizada para a captura das informações dos canais escolhidos, para assim, prosseguir com um estudo aprofundado e definir tais parâmetros seriam de fato os selecionados.

Para os dados capturados, foi então processado e reorganizado em uma linha temporal de informações referente ao gosto de outros usuários em relação aos mesmos. A partir dessa linha temporal, foi realizado uma linha temporal das diferenças entre as capturas, e com isso o algoritmo de regressão realizar os cálculos definindo uma lista com os principais parâmetros que influencia o sucesso de um canal.

A coleta dos dados, através da API, foi realizada algumas vezes por dia durante um período de 14 dias (duas semana). O período capturado havia sido escolhido através de uma analise empírica em cima de publicações dos canais selecionados. A seleção destes canais, listados logo abaixo, foi realizada com base em buscas em fóruns determinando canais de sucessos que tivessem postagens constantes na plataforma. A partir dessa análise foi observado que a maioria dos canais buscavam publicar no mínimo um (1) vídeo a cada duas semanas. Então, para garantir que será selecionado uma base de dados relevante para análise esse período foi estabelecido.

- Boyce Avenue
- Mariana Nolasco
- Ana Gabriela
- Sofia Karlberg
- Cimorelli the band
- Gabi Luthai
- Daniela Sings
- Joana Castanheira
- Carina Mennitto

Para cada canal selecionado, os dados capturados tem as seguintes características: a partir de uma busca simples de informações do canal, obtemos informações necessárias gerais sobre o canal, logo em seguida é feito uma busca de vídeos a partir de uma ID do canal, pre-selecionado anteriormente (canais listados acima). A API permite escolher grupos para retorno de informações,

ou seja, selecionando quais possíveis grupos de valores seriam os melhores para serem utilizados, como exemplo no campo "part", um campo que permite vários argumentos, como exemplo temos o "id"e "statistics", para retornar informações de um determinado ID e suas estatísticas, entre vários outros argumentos. Na tabela 1 vemos alguns que foram capturados.

Tabela 1. Tabela busca de canais

Propriedades por Canal	
Propriedade	**Descrição**
kind	O tipo do recurso da API. O valor será youtube#channel.
etag	A Etag deste recurso.
id	O ID que o YouTube usa para identificar o canal de forma exclusiva.
snippet	objeto snippet contém detalhes básicos sobre o canal, como seu título, sua descrição e suas imagens em miniatura.
snippet.title	O título do canal.
snippet.description	A descrição do canal.
snippet.publishedAt	A data e a hora em que o canal foi criado. O valor é especificado no formato ISO 8601 (YYYY-MM-DDThh:mm:ss.sZ).
contentDetails	O ID da playlist que contém os vídeos enviados do canal. Use o método videos.insert para enviar novos vídeos e o videos.delete para excluir vídeos enviados anteriormente. (método: relatedPlaylists.uploads)
statistics (s)	O objeto statistics encapsula estatísticas para o canal.
s.viewCount	Quantidade de visualizações do canal.
s.commentCount	Quantidade de comentários do canal.
s.subscriberCount	Quantidade de inscritos do canal.
s.videoCount	Quantidade de vídeos enviados para o canal.
topicDetails	O objeto topicDetails encapsula informações sobre tópicos Freebase associados ao canal.

É realizado a busca de vídeos para cada canal selecionado, para cada estilo de requisição, a API desconta em sua cota diária para essas capturas, por exemplo, uma busca geral na requisição 'vídeo_search' gasta 100, enquanto as outras referente a um canal e um vídeo gastam 5 (cada). Fazendo com que a captura de dados se limitassem a quase que uma por dia (para a quantidade de vídeos e canais selecionados).

Após ter capturado os dados da busca, eles foram processados e armazenados de acordo com a sua relevância (algo que poderia ser utilizado na regressão),

como na tabela 2 e então foi realizado esse processo para todos os canais selecionados. Para cada canal, foi capturado varias informações sobre um vídeo e, assim como para os canais, foram capturados para analise somente os que pareciam ser relevantes, veja na tabela 3.

Tabela 2. Tabela de propriedades de um canal

Atributo	Descrição
ID	Um código único para representar o canal
Inscritos	Quantidade de inscritos no canal no momento da captura
Quantidade de vídeos	Quantidade de vídeos já capturados
Quantidade de visualizações	Quantidade total de visualizações no canal até o momento
Título	Identidade (nome) do canal

3.2 Limpeza dos dados

Após uma analise dos dados obtidos e observar um pouco sobre o funcionamento do algoritmo de regressão, foi realizado uma exclusão no armazenamento de alguns destes dados das tabelas 2 e 3. Somente os dados a seguir foram selecionados:

- Id do canal
- Quantidade de inscritos
- Quantidade de vídeos pulicados
- Quantidade total de visualizações

Após a coleta dos dados de um canal, com uma segunda analise sobre quais dados realmente fariam sentido aumentar a quantidade da audiência, foram selecionados os seguintes atributos para cada vídeo capturado:

- Id do vídeo
- Quantidade de visualizações
- Quantidade de curtidas (likes)
- Quantidade de 'não curtidas' (dislikes)
- Quantidade de comentários
- Id do canal

Com esse pré-processamento, uma redução considerável pode ser observada melhorando o armazenamento e assim uma analise dos estudos mais concretos.

Tabela 3. Tabela de propriedades de um vídeo

Atributos	Descrição
ID Vídeo	Um ID único representando o vídeo
ID Canal	Um ID único representando o canal daquele vídeo
ID categoria	Categoria que aquele video se encaixa
Título	Título do vídeo
Quantidade de comentarios	Quantidade de comentários no momento da captura
Descrição	Descrição presente no vídeo
Quantidade de curtidas	Quantidade de likes no momento da captura
Quantidade de 'não curtiu'	Quantidade de dislike no momento da captura
Quantidade de favoritados	Quantidade de usuários que favoritaram o vídeo
Quantidade de visualizações	Quantidade de visualizações do vídeo
Titulo do canal	Título do canal, que aquele vídeo pertence
Linguagem padrão	Uma linguagem padrão que esta o vídeo
Broadcast conteudo ao vivo	Dados sobre os conteúdo gravado
Publicação	Data da publicação daquele vídeo
Tags	As tag's presente no vídeo

Obtendo uma lista de vídeos para cada canal, com as informações acima, evitando cálculos desnecessários com as informações.

Como um vídeo é sempre co-relacionado com um canal, foi possível, de maneira fácil, obter acesso aos dados seja ele um vídeo ou o canal.

Mesmo com esse pré-processamento, para a utilização dos dados no algoritmo de regressão, foram utilizados todos os dados armazenados de um vídeos e para um canal, foi utilizado somente o valor total de inscritos na ultima captura.

3.3 Servidor

A computação e armazenamento foi realizado através de um servidor hospedado virtualmente. Este servidor ficou responsável de coleta e o armazenamento os dados referente as requisição, durante todo o período da coleta. Como a base analisada não teria uma grande quantidade de informações armazenada, a quantidade de armazenamento disponível no servidor foi mais do que suficiente.

O servidor selecionado é hospedado pela Digital Ocean [1] com recursos disponíveis para estudantes por um determinado período, suficiente para a analise. Contando com suporte de memoria e armazenamento para a realização deste estudo.

3.4 Base de dados

Nenhum estudo parecido com o atual foi encontrado na literatura, no momento da escrita deste artigo, logo, nenhum dado pode ser previsto antes que fosse realizado a busca destes, fazendo com não fosse possível obter uma média de gastos computacionais seriam necessários. Para algum futuro trabalho, em uma analise de duas semanas, foram armazenados menos de 0,5 MB de dados.

Para o armazenamento, durante as requisições diárias, foram armazenado no banco as informações da seguinte forma:

1. Tabela de **vídeos**
   - Id do canal
   - Quantidade de inscritos
   - Quantidade de vídeos publicados
2. Tabela de **canais**
   - Id do vídeo
   - Quantidade de visualizações
   - Quantidade de curtidas (likes)
   - Quantidade de 'não curtidas' (dislikes)
   - Quantidade de comentários
   - Id do canal

A partir da busca, seleção, pré-processamento e transformação, os dados utilizados, como comentado anteriormente, foram colocados como parâmetros para o algoritmo de regressão de forma a melhorar o processamento e buscando melhores resultados.

Lembrando que esses passos são essenciais, principalmente o pré-processamento, para que não seja armazenado informações não uteis para o problema abordado, como as gravações dos vídeos, imagens, alguns dados de perfis, entre outros.

3.5 Regras e padrões

Com a quantidade relevante de conteúdo publicado por pessoas, é complicado indicar pessoas para analisar e assim gerar as métricas de qualquer canal. Com isso a ideia central do algoritmo de regressão é a captura de informações para a realização de um estudo dos parâmetros processados.

Para o algoritmo de regressão linear, foi realizado um pré-processamento dos dados para deixa-los de forma a melhorar o desempenho e os resultados obtidos.

Para as informações dos vídeos capturados, foram processados a fim de obter uma média por vídeo para o algoritmo conseguir trabalhar com os vídeo de forma individual, podemos ver o passo a passo realizado a seguir:

1. Para cada canal capturado foi armazenado um dicionario de vídeos (por ID), para cada vídeo uma lista de dias e para cada dia uma lista das capturas daquele vídeo nesse dia.
2. Para cada um desses dias foi realizado uma média dos valores capturados e armazenado na forma de um vídeo por dia. Com a realização da média, foi capturado o desvio padrão existente para cada dado e armazenado, assim como as médias.
3. Foi realizado uma Linearização dos valores desses vídeos para o algoritmo de regressão trabalhar com as grandezas dos números e não com os valores brutos deles (Linearização realizada com o Logaritmo do valor na base 10).
4. Ainda pensando no desempenho e realização da regressão, foi realizado uma normalização (deixando-os entre 0 e 1).
5. Foi capturado a quantidade de inscritos, dos valores armazenados para os canais e acrescentado nesse conjunto de dados, também linearizado e normalizado, para servir como o atributo de foco para determinar a audiência no algoritmo regressão.

Vamos na tabela 4 os atributos, após o novo processamento deles, que foi utilizado como entrada ('input') para o algoritmo de regressão. Para conseguir obter os resultados desse algoritmo, foi utilizado a ferramenta Scikit-learn [3] com o algoritmo de regressão linear presente nele.

Tabela 4. Tabela de atributos para o algoritmo de regressão

Atributos	Descrição
Id do vídeo	Representa qual vídeo tem as informações abaixo
Média visualizações	Média de visualizações
Desvio padrão visualizações	Desvio padrão de visualizações
Média like	Média de like
Desvio padrão like	Desvio padrão de like
Média não like	Média de não like
Desvio padrão não like	Desvio padrão de não like
Média comentários	Média de comentários
Desvio padrão comentários	Desvio padrão de comentários
Inscritos no canal	Quantidade de inscritos no canal (na ultima captura)

4 Resultados

Nesta seção será apresentada resultados obtidos e como foram utilizados os métodos para obter esses resultados.

4.1 Instancia de teste

Após alguns testes e observações nos resultados, a instancia de teste, detalhada na seção anterior, foi dividida em duas partes, sendo 20% dela para testes e os outros 80% para o treinamento do modelo para os resultados apresentados nessa seção.

Na figura 1 vemos as estatísticas da instancia separada para teste e logo em seguida, na figura 2 as estatísticas da instancia de treino. Lembrando que os valores foram linearizados, e, assim, foram exibidos nas figuras, momentos antes da normalização em seus valores.

	likes_mean	likes_std	view_mean	views_std	comment_mean	comment_std	dislike_mean	dislike_std
count	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	1.268300	0.307652	2.487344	0.161509	1.536319	0.275510	1.393958	0.292355
std	1.156293	0.207655	1.399088	0.063620	0.877960	0.228328	0.749104	0.281548
min	0.097852	0.156259	0.478111	0.123842	0.233614	0.132547	0.236798	0.132547
25%	0.555087	0.185963	1.796190	0.132557	0.767791	0.136011	0.804937	0.136011
50%	0.744819	0.248850	2.145416	0.134250	1.850667	0.175940	1.578260	0.212660
75%	1.830711	0.321895	3.293438	0.151148	2.191747	0.334188	2.029471	0.271852
max	3.827388	0.864543	4.858724	0.332698	2.671450	0.874454	2.238078	1.065252

Figura 1. Estatísticas da instancia de teste

	likes_mean	likes_std	view_mean	views_std	comment_mean	comment_std	dislike_mean	dislike_std
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	1.399707	0.319706	2.648543	0.157816	1.412569	0.577786	1.071337	0.536342
std	1.085706	0.242300	1.384837	0.079931	0.628829	0.566534	0.654434	0.436257
min	0.097923	0.000000	0.170066	0.000000	0.170066	0.000000	0.160626	0.000000
25%	0.506549	0.167413	1.853373	0.125225	0.836639	0.155377	0.527693	0.220060
50%	1.107316	0.271096	2.487884	0.133673	1.394592	0.303801	0.820515	0.341980
75%	1.875971	0.350437	3.845341	0.166830	2.047135	0.832965	1.758277	0.908291
max	3.661888	1.079770	4.954846	0.361364	2.361796	1.903434	2.179236	1.493635

Figura 2. Estatísticas da instancia de treino

4.2 Regressão Linear

Para rodar o algoritmo de regressão linear os dados foram apresentados de forma transposta, em sua matriz original, para o funcionamento melhor do algoritmo de regressão.

Após a realização do algoritmo de regressão na instancia de entrada, utilizando todos os possíveis atributos de entradas (input labels), observamos que ao comparar a base de treino com a base de teste, mesmo obtendo taxas de erros baixas, a média de erros calculados, de inscritos estipulados, foram valores consideráveis. Cerca de 480000 inscritos em média, em uma base onde, em média, conta com 3.5 milhões de inscritos.

Mesmo com valores altos, o algoritmo conseguir estipular com valores muito bons, considerando a quantidade média de audiência. Porem, esses valores foram mais para mostrar a relevância do algoritmo de regressão utilizado.

Após essa analise, conseguimos obter valores para a importância de cada atributo dado como entrada. Com o funcionamento do algoritmo de regressão, ele acha o melhor valor de coeficiente para cada atributo, assim, colocando um grau de importância para cada um em sua função objetivo. Na figura 3 vemos os 5 principais atributos estipulados pelo regressor. Ignorando a primeira coluna, pois ela representa a posição referente na tabela (utilizados em algoritmos para próximos resultados), podemos observar que alguns valores de desvio padrão foram mais importantes para a regressão do que valores das médias.

Position	Parametro	Coeficiente para regressão
2	view_mean	0.3732200360586011
0	likes_mean	0.1678438286517744
7	dislike_std	0.15496454397202333
6	dislike_mean	0.05756773420383723
1	likes_std	0.027823830633219138

Figura 3. Os 5 principais parâmetros estipulados pelo regressor

Até o momento, caso finalizássemos as pesquisas, poderíamos dizer que os parâmetros principais para predizer o quanto um canal está do sucesso, ou seja, o quão próximo esta da linha de regressão, seriam:

1. Média quantidade de visualizações
2. Média quantidade de curtidas (likes)
3. Desvio padrão quantidade de "não gostei"(dislikes)
4. Média quantidade de "não gostei"(dislikes)
5. Desvio padrão quantidade de curtidas (likes)

Porem após realizar outros testes, foi possível minimizar esses valores a 2 itens principais. Foi capturado esses informações e realizado uma nova regressão linear,

porem limitando os valores de entrada somente para as médias dos 3 principais valores encontrados, ou seja, média de visualizações, de curtidas (likes) e de 'não curtidas' (dislikes).

Foi realizado a regressão e exibida em gráficos, mostrando os erros para cada uma dessas regressões. Vamos na figura 4 a regressão realizada para todos valores de entrada, na figura 5 vemos a regressão somente com valores de visualizações, na figura 6 vemos a regressão com os valores de visualizações e de curtidas e na figura 7 vemos a regressão com os valores de visualizações, curtidas e de não gostei (dislikes).

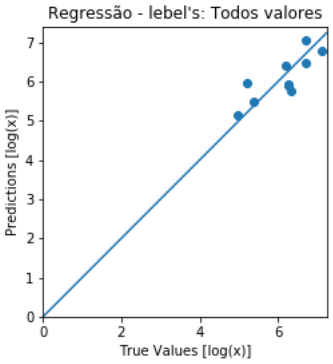

Figura 4. Regressão com valores gerais

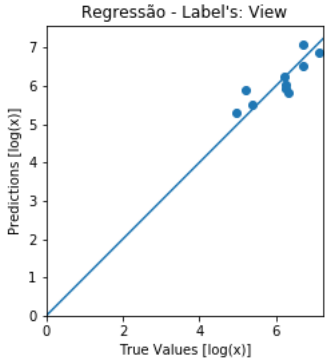

Figura 5. Regressão: dados de visualização

Vemos que visualmente, não podemos notar tanta diferença, mas ao observarmos os erros médios, na figura 8, utilizando as métricas MAE e MSE, vemos

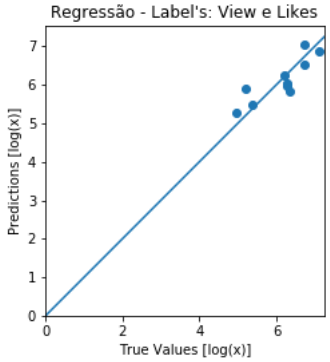

Figura 6. Regressão: dados de visualização e de curtidas

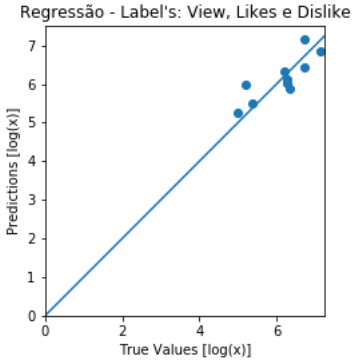

Figura 7. Regressão: dados de visualização, curtidas e não gostei

que temos resultados melhores quando utilizamos somente os atributos de visualizações e de curtidas.

Base utilizada	MAE	MSE
Geral	0.3507371548651297	0.3977645810891414
Views	0.3099424225394416	0.3607402872847538
Views e Likes	0.29458746570548594	0.34720609048117634
Vies, likes e dislikes	0.31518495805187347	0.3707967124521461

Figura 8. Médias de erros (métricas MAE e MSE)

Apesar de serem resultados bem próximos, ao acrescentar mais atributos na regressão linear, a regressão apresentou uma performance um pouco pior do que para somente os dois atributos encontrados.

5 Discussões

Nesta seção será apresentada discussões e futuros trabalhos.

5.1 Ameaças a validade e futuros trabalhos

A quantidade de valores capturados, durante as duas semanas, talvez possa não ter sido bons parâmetros para a entrada de dados. Sendo assim, uma analise de dados com maiores tempo de captura pode ser que obtenha melhores resultados, ou encontre diferentes tipos de padrões, além dos citados neste artigo.

A analise empírica de dados importantes, com base em literaturas e conhecimentos musicais, ao mesmo tempo que pode ter ajudado o modelo, pode ser que tenha alterado, parcialmente, o reconhecimento de parâmetros.

Analisar uma quantidade maior de canais, além dos de sucesso, para ver o comportamento da regressão pode vir a ser relevante para o modelo em questão.

O algoritmo utilizado, de regressão linear, como trabalha diretamente com algo linear, talvez outros algoritmos para múltiplas variáveis possa vir a resultar em outros padrões, não reconhecidos nesse artigo.

5.2 Conclusões

Observando os valores nas tabelas de erros, para as métricas utilizadas, concluímos que os principais fatores para determinar um canal de sucesso são as visualizações e as curtidas, tendo como base a analise de audiência ligada diretamente a quantidade de inscritos no canal.

Pois ao acrescentar parâmetros para o algoritmo de regressão, vemos um aumento na quantidade de erros existente, para a nossa base de teste e treinamento.

Foi realizado alguns outros testes, sem o armazenamento dos dados destes, para chegar na partição de 20% para teste e 80% para treinamento, como colocado no trabalho apresentado. Oque fez com que fosse observado que a captura de dados durante um período maior seria importante para o treinamento destes dados.

Com pequenas alterações no algoritmo criado, seria possível ter como entrada novo valor conseguindo informações co-relacionando o canal em questão com os canais analisados, como um novo teste para o modelo, conseguindo dizer qual ponto aquele artista precisa focar em evoluir, seja com técnicas de marketing digital ou outros meios. Logo, conseguir predizer os parâmetros que fara desse canal um de sucesso.

Referências

1. Digital ocean. `https://www.digitalocean.com`. Acessado dia 06/12/2019.
2. Matplot lib. `https://matplotlib.org/`. Acessado dia 06/12/2019.
3. Scikit learn. `https://scikit-learn.org/stable/`. Acessado dia 06/12/2019.
4. Youtube analytics api. `https://developers.google.com/youtube/v3/docs?hl=pt-br`. Acessado dia 13/09/2019.
5. Seyedamin e Assefi Mehdi e Safaei-Saied e Trippe Elizabeth D e Gutierrez Juan B e Kochut Krys Allahyari, Mehdi e Pouriyeh. Uma breve pesquisa sobre mineração de texto: técnicas de classificação, agrupamento e extração.
6. Cássio Oliveira Camilo and João Carlos da Silva. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, pages 1–29, 2009.
7. G e Smyth P Fayyad, U e Piatetsky-Shapiro. Da mineração de dados ao conhecimento de descoberta em bancos de dados. *Revista da AI*, 3(17):37–54.
8. Noemi Dreyer Galvão and Heimar de Fátima Marin. Técnica de mineração de dados: uma revisão da literatura. *Acta Paulista de Enfermagem*, 22(5):686–690, 2009.
9. John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90, 2007.
10. Changsha Ma, Zhisheng Yan, and Chang Wen Chen. Larm: A lifetime aware regression model for predicting youtube video popularity. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 467–476. ACM, 2017.
11. Elisa Maria Caetano dos Santos et al. Estimadores ls, drls e'tau'no modelo de regressão linear: estudo comparativo por simulação. 1989.
12. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.
13. Patricia e Marinov Miroslav e Pena-Hernandez Keila e Gopidi Rajitha e Chang Jia-Fu e Hua Lei Yoo, Illhoi e Alafaireet. Mineração de dados em saúde e biomedicina: um levantamento da literatura. *Journal of medical systems*, 36(4):2431–2448.