Projeto Orientado em Computação 2

Prevendo a popularidade de uma música

Lucas Augusto Leone

Universidade Federal de Minas Gerais

Departamento de Ciência da Computação

Orientador: Adriano Alonso Veloso

lucasaugustoleone@gmail.com , *adrianov@dcc.ufmg.br*

Resumo

A música tem um espaço grande na nossa vida e, além dessa relação emocional, é um mercado lucrativo. Neste projeto, busquei entender quais são as características que podem determinar a popularidade de uma música. Avaliei algoritmos de classificação e regressão para prever a popularidade de uma música e determinar quais são os principais atributos que compõem essa fórmula do sucesso em um dataset autoral construído a partir da coleta dos dados do aplicativo Spotify no qual contém dados relevantes sobre as canções, como a sua popularidade, a popularidade do artista, data de lançamento e informações musicais.

Introdução

A música tem um espaço grande na nossa vida e, além dessa relação emocional, é um mercado lucrativo. Segundo relatório da Pró-música, uma associação de produtores fonográficos associados, em 2021 o mercado da música cresceu 32% em 2021 no Brasil. Em questões numéricas, em 2021 R$ 2,1 bilhões foram gerados pela indústria musical no Brasil, desses, R$ 1,8 bilhão veio do streaming.
Dessa forma, a fim de melhorar cada vez mais esses resultados, a indústria da música busca aperfeiçoar suas produções, campanhas de marketing e divulgação de lançamentos para conseguir lançar um novo hit que caia nas graças do povo. Para isso, cada vez mais pesquisas e desenvolvimentos na área de inteligência artificial e machine learning estão sendo realizadas para aumentar a chance de lançar um novo hit, como por exemplo análise de sentimento de letras de musica, análise de atributos musicais e também o uso de inteligência artificial generativa para a produção de melodias.

Nesse contexto, a aplicação de algoritmos de aprendizado de máquina, em um conjunto de dados sobre músicas coletadas do Spotify emerge como uma abordagem inicial promissora. Os algoritmos de machine learning possibilitam uma interpretação automatizada e um entendimento não óbvio dos dados, fazendo com que conseguimos traçar uma possível relação entre os atributos musicais e a popularidade da música e assim conseguir prever essa métrica para uma música recém lançada ou que estar para lançar, auxiliando os produtores a focar em pontos que possam melhorar esse score.

Por meio da análise automatizada de dados musicais, pode-se capturar nuances entre os dados, mostrando possíveis correlações ou não entre as features, auxiliando assim a prever a popularidade. A alimentação desses dados em algoritmos de machine learning permite a criação de modelos preditivos que se aperfeiçoam com o tempo e tem o potencial de se tornarem cada vez mais precisos e mais utilizados na indústria. Isso não apenas auxilia os produtores no processo de criação da música, mas também facilita a alocação de recursos de forma mais certeira em músicas que possuem maior potencial de se tornar um hit, ou seja, serem bastante populares.

Tal cenário demonstra vasta aplicabilidade das técnicas desenvolvidas, indo além apenas da pós produção das músicas e abrangendo áreas como cinema, criação de vídeo, marketing digital, entre outras. Dessa forma, o potencial impacto dessas abordagens se estende para além do âmbito comercial, contribuindo para soluções mais abrangentes e inovadoras em variados contextos.

Neste estudo, abordarei essa temática, com o objetivo de prever a métrica de popularidade de uma música no principal aplicativo de streaming, o Spotify.

Esse trabalho foi dividido nas seguintes partes:

- Coleta dos dados
- Escolha dos modelos
- Avaliação dos modelos
- Importância das Features

Referencial Teórico

Nesta seção, apresentaremos os conceitos e ferramentas relevantes para compreender o contexto do nosso trabalho de pesquisa, que envolve a aplicação de Machine Learning (ML) e Importância das Features para a comparação de modelos na tarefa de prever a popularidade de uma música. A seguir, descrevemos os principais conceitos e soluções relacionados.

Popularidade

O que é a popularidade de uma música?

A popularidade de uma música refere-se ao seu alcance e aceitação pelo público, frequentemente medida pelo número de reproduções, vendas, posições em paradas musicais, e presença em plataformas de streaming. Ela reflete o impacto cultural e comercial da música, influenciada por fatores como marketing, qualidade artística, e tendências sociais. Contudo neste trabalho, utilizamos o conceito de popularidade definido pela plataforma de streaming Spotify, que consiste em três fatores

1. Total de reproduções
2. Recência das reproduções
3. Frequência das reproduções

Machine Learning

Machine Learning é uma área da inteligência artificial que envolve o treinamento de algoritmos em volumes de dados para reconhecer padrões e tomar decisões com base nesses dados. Exemplos notáveis incluem redes neurais e algoritmos de aprendizado supervisionado e não supervisionado. Esses métodos são fundamentais para tarefas como reconhecimento de imagem, análise de dados e previsão de tendências.

Árvores de decisão

As árvores de decisão são modelos de machine learning que dividem os dados em subconjuntos baseados em condições de características, criando uma estrutura hierárquica semelhante a uma árvore. Cada nó interno representa uma condição em uma característica, cada ramo representa o resultado dessa condição, e cada nó folha representa uma classe ou valor final. Elas são usadas para tarefas de classificação e regressão, sendo intuitivas e fáceis de interpretar.

Random Forest

O Random Forest é um método de ensemble learning que utiliza múltiplas árvores de decisão para melhorar a precisão e a robustez do modelo. Ele gera diversas árvores de decisão durante o treinamento e combina suas previsões para obter um resultado final mais confiável e estável. Este método é eficaz para evitar o overfitting e é amplamente utilizado em tarefas de classificação e regressão, oferecendo alta precisão e capacidade de generalização.

Algoritmos de Regressão

Os algoritmos de regressão são métodos de machine learning usados para prever valores contínuos com base em dados históricos. Eles modelam a relação entre variáveis independentes e uma variável dependente. Exemplos incluem a regressão linear, a regressão polinomial e as redes neurais. Esses algoritmos são essenciais para tarefas como previsão de preços, análise de tendências e estimação de risco.

Soluções Existentes na Literatura/Mercado

Modelos baseados em aprendizado profundo e aprendizado de máquina têm sido extensivamente estudados para tarefas de processamento de dados musicais, seja para conseguir criar uma música ou para analisá-la de forma mais assertiva, entretanto, ainda a poucos estudos relacionados a área de predição de popularidade, já que é uma parte bem específica no geral, que depende de dados precisos e de difícil mapeamento por serem bastantes subjetivos.

Como meu dataset foi coletado a partir da plataforma Spotify, busquei por encontrar trabalhos com um foco similar ao meu, dessa forma, em [*3*], o autor realiza a comparação de cinco modelos na tarefa de prever a popularidade de uma música. Os modelos utilizados foram: Regressão logística, Support Vector Machines, Multilayer Perceptron, Linear Discriminant Analysis e Quadratic Discriminant

Analysis

O segundo trabalho relacionado [4] utiliza uma abordagem diferente e foca mais nas features do que nos modelos em si, já que ele testa apenas um único modelo baseado em Support Vector Machines e faz as predições com base em cenários como: utilizando todas as features, utilizando apenas as features acústicas, utilizando apenas as features nao acústicas.

Temos também um terceiro trabalho relacionado [5] no qual se assemelha bastante a minha pesquisa ao fazer a comparação entre modelos de classificação e regressão, contudo foca mais nas limitações do trabalho, como uma baixa quantidade de dados, features não tão expressivas e a flutuação da popularidade de uma música ao longo do tempo.

Contribuição

Escolha do Dataset

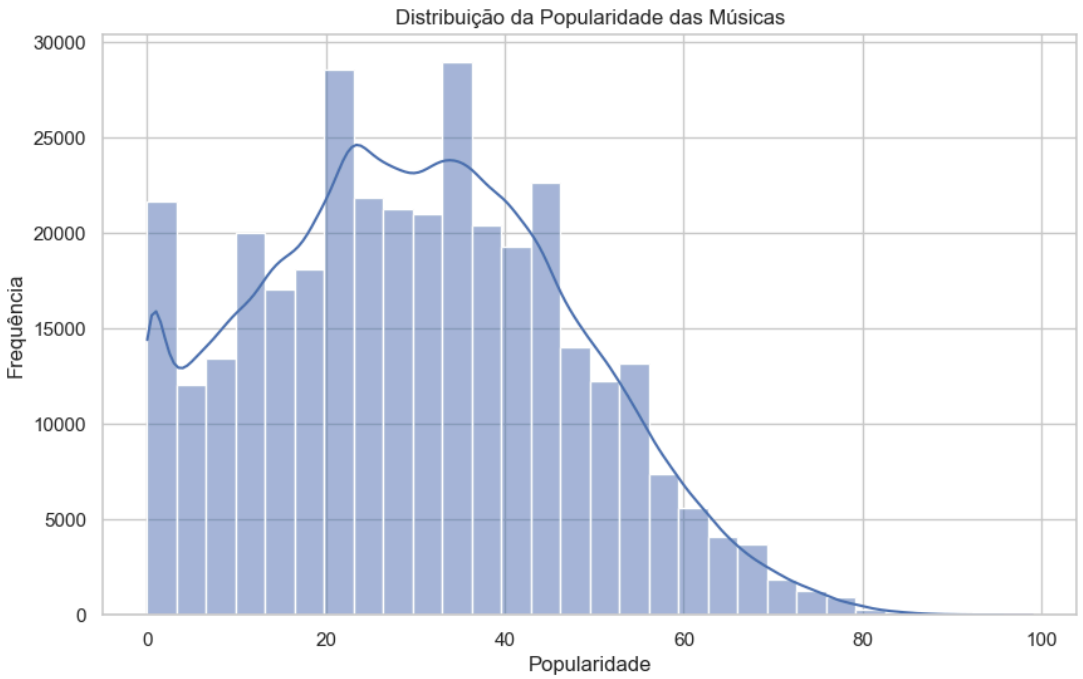
A primeira etapa desse projeto foi a coleta dos dados para o treinamento dos modelos, para isso, foi utilizada a biblioteca em python Spotipy para fazer chamadas na API do aplicativo de streaming de músicas Spotify. Foram coletadas ao longo do tempo mais de trezentas mil músicas (300K) e 22 features, na qual 14 features foram escolhidas para o treinamento do modelo.

Balanceamento das classes:

- Rock-> 145759 entradas
- Pop-> 132463 entradas
- Jazz-> 23793 entradas
- Folk-> 13403 entradas
- Funk-> 10578 entradas
- Classical-> 9500 entradas
- Rap-> 5732 entradas
- Reggae-> 3764 entradas
- Country-> 2987
- Rnb-> 2759
- Edm-> 462

Tabela de Features

popularity	int
explicit	int
danceability	float
energy	float
key	int
loudness	float
mode	int
speechiness	float
acousticness	float
instrumentalness	float
liveness	float
valence	float
tempo	float
genre	object
popularidade_artista	float



Modelos selecionados

A fim de encontrar o melhor modelo para executar essa tarefa, foi testado uma grande quantidade de modelos, sendo eles: Regressão Linear simples, Ridge, Random Forest, XGBoost, Ada boost e Deep learning.

Os modelos foram avaliados utilizando o Erro Médio Absoluto (MAE) que constitui da seguinte fórmula.

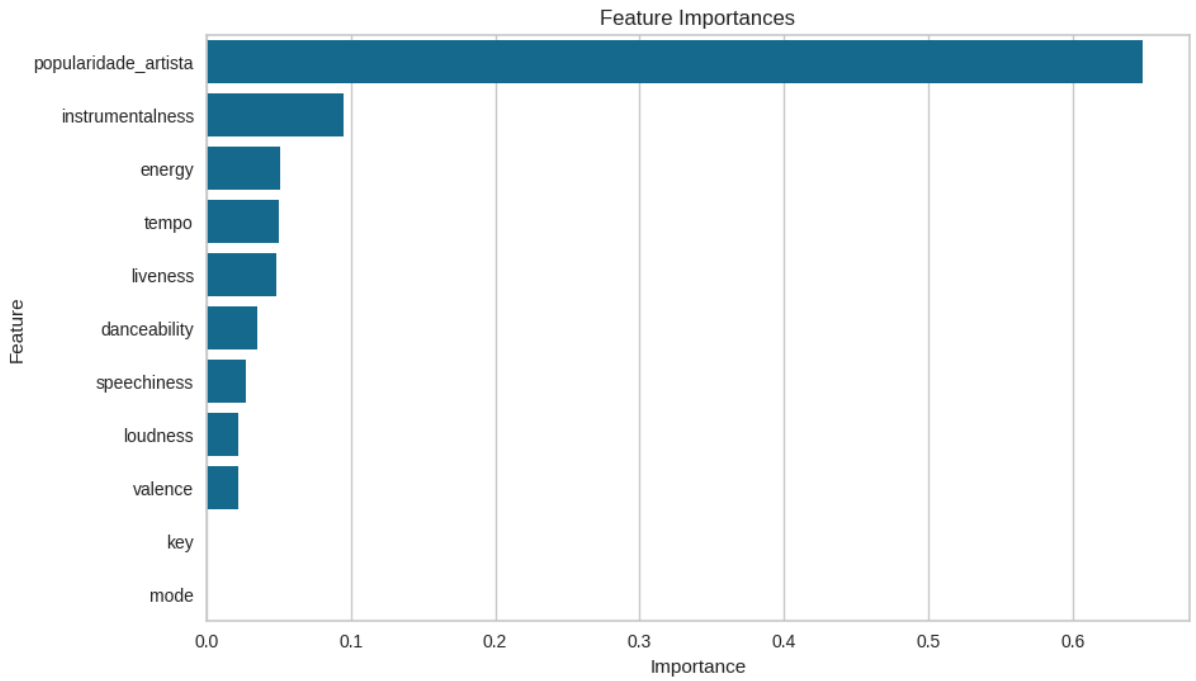
$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

Os seguintes resultados foram obtidos, sendo válido ressaltar que a métrica foi ajustada para refletir o erro em uma nota de 0 a 100.

Model	MAE
LR	11,02
Ridge	11,02
DL	11,1
Random Forest	10.3
XGB	10,5
AdaBoost	11,51

Portanto o modelo mais performático em nossos testes foi o Random Forest, seguido pelo XGBoost e depois por modelos simples de Regressão Linear, apontando a direção de que modelos ensembles trazem uma maior robustez a solução e que podem potencialmente ser melhor explorados. Além de que modelos de Deep Learning são promessas, devido a sua interpretabilidade de dados não rotulados, como para o nosso caso de músicas, temos o exemplo de ondas sonoras, tons da música, notas musicais entre outros.

A seguir vemos quais são as features mais importantes para nossos modelos

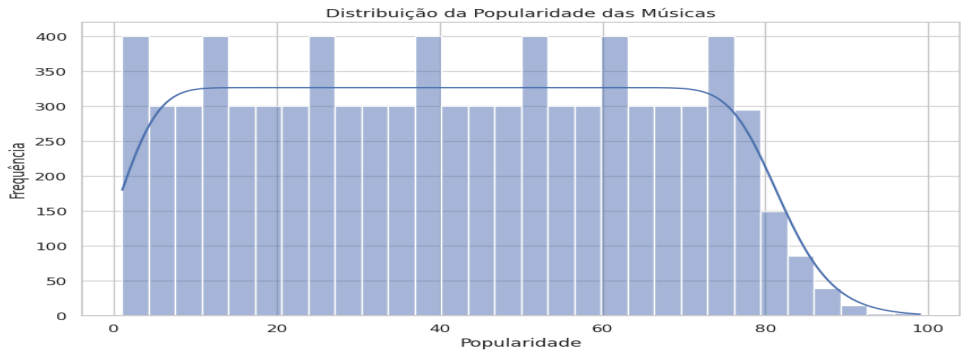


Fica escancarado que a feature com maior importância para os modelos foi a popularidade do artista, feature a qual foi utilizada para representar o nome de um artista de forma mensurável, já que muitas vezes sabemos que um artista é popular pelo seu nome. Portanto é impossível separar a popularidade de uma música da popularidade do cantor o que implica na dificuldade de prever hits para cantores pouco conhecidos e provavelmente não conseguiriamos prever que musicas como Balada de Gustavo Lima e Call Me Maybe de Carly Rae Jepsen que não eram conhecidos quando as lançaram seriam grandes sucessos.

Apesar dos resultados promissores obtidos, fiquei insatisfeito com a minha abordagem para o problema e resolvi simplificar o escopo da pesquisa a fim de obter resultados mais expressivos e que me convencessem mais. Dessa forma medidas foram tomadas buscando esse objetivo.
Reduzimos nosso dataset a apenas músicas do gênero Pop, será criada uma distribuição mais uniforme das popularidades para obtermos uma maior representatividade para cada classe. Sendo assim, vamos novamente realizar o treinamento dos modelos, sendo estes também impactados, já que novos modelos foram considerados para a análise e os modelos de Deep Learning e Regressão Linear foram descartados.

O novo dataset é composto por 8191 músicas que possuem a seguinte distribuição

de popularidade, bem mais comportada que a anterior



Os seguintes resultados foram obtidos

Model	MAE
Ridge	13,3
Random Forest	11,8
XGB	12,4
AdaBoost	14,9
LGBM	11,9
CatBoost	11,9

Após a redução do escopo a fim de buscar melhores resultados para um gênero específico obtive um resultado não esperado, a maior parte dos modelos testados obteve um resultado pior do que quando consideramos todos os gêneros. Dessa forma, duas questões são levantadas; A quantidade de dados pode ter sido um fator relevante para a depreciação dos modelos, já que ao reduzir o escopo, reduzimos em mais de 90% da quantidade de dados utilizados para treinar os primeiros modelos. Outro ponto é que provavelmente o gênero não deve ser um fator que impacte nas métricas, isso pelo fato de que independente do gênero, o modelo consegue aprender a partir dos atributos musicais e fazer inferências, ou seja, a partir dos atributos musicais de gêneros similares, deve ser possível inferir a popularidade de uma música de um gênero diferente, podemos utilizar o Folk e o Country como exemplos de gêneros diferentes mas que possuem acústica similar.

Conclusão

Neste trabalho, propusemos avaliar a possibilidade de se prever a popularidade de uma música em diversos gêneros musicais diferentes, utilizando aprendizado de máquina supervisionado e avaliando o erro médio das previsões. Foram testados mais de cinco modelos diferentes, em dados retirados da plataforma de streaming Spotify, que se baseiam em classificação e regressão para conseguir dizer se uma música será popular ou não em uma escala de 0 a 100. Os modelos obtiveram um erro médio absoluto na casa dos 12 pontos na escala o que demonstrou um indício promissor para tal tarefa. Contudo, ao buscarmos melhores resultados restringindo o escopo e analisando apenas o gênero musical Pop, foi constatado uma piora nos resultados obtidos, ou seja, os modelos tiveram um erro médio absoluto maior nessas condições. Além do mais, evidenciamos a notória importância da popularidade do cantor para a popularidade da música, ou seja, cantores mais populares possuem músicas mais populares e vice-versa. Contudo também foi demonstrado que a feature de instrumentalness teve um impacto significativo nos resultados e pode ser um possível caminho a se explorar para aumentarmos a precisão dos modelos

Para futuros trabalhos, busco coletar mais informações sobre músicas com o intuito de obter uma maior representatividade de um certo fator para a popularização de uma música. Devo também atentar a sazonalidade musical além de considerar a fluidez da popularidade, ou seja, utilizarmos uma feature dinâmica que represente essas alterações que podem acontecer de forma diária e espontânea. Por fim, uma abordagem com dados não rotulados como ondas sonoras, tons musicais entre outros e a utilização de modelos de aprendizado profundo pode ser uma solução mais interessante para a solução deste problema do que a utilização de modelos tradicionais de aprendizado de máquina.

Referências

1 Welcome to Spotipy! — spotipy 2.0 documentation. Disponível em: <https://spotipy.readthedocs.io/en/2.21.0/#>. Acesso em: 17 abr. 2024.

2 SPOTIFY. Web API | Spotify for Developers. Disponível em: <https://developer.spotify.com/documentation/web-api>.

3 PHAM, J.; KYAUK, E. Predicting Song Popularity. [s.l: s.n.]. Disponível em: <https://cs229.stanford.edu/proj2015/140_report.pdf>.

4 ARAUJO, C.; CRISTO, M.; GIUSTI, R. Predicting Music Popularity on Streaming Platforms. [s.l: s.n.]. Disponível em: <https://sol.sbc.org.br/index.php/sbcm/article/download/10436/10303/>. Acesso em: 23 abr. 2024.

5 HARRIMAN SAMUEL SARAGIH. Predicting song popularity based on spotify's audio features: insights from the Indonesian streaming users. Journal of Management Analytics, v. 10, n. 4, p. 693–709, 27 jul. 2023.

6 OLIVEIRA, G. P.; PAULA, A.; MORO, M. M. What makes a viral song? Unraveling music virality factors. 21 maio 2024.

7 SCIKIT-LEARN. scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation. Disponível em: <https://scikit-learn.org/stable/index.html>.