Proposal for INgrooves

TIMOTHY LEE

PCTIMLEE@GMAIL.COM | (818) 808 - 7292

APRIL 17, 2017

Introduction

"[INgrooves] aspire[s] to be the most **transparent** and solution-driven partner for all of the labels & artists we work with." – INgrooves website, 'Who We Are – Our Story'

This proposal will focus on the idea of partner transparency and how to use data science to further INgrooves's efficacy in this area.

The basic idea is to leverage INgrooves's plethora of historical label partner data along with various streaming/social media data APIs to assemble a robust, predictive algorithm for potential label partners to see exactly what they're getting into when they partner with INgrooves.

The result is two-fold:

1. An accurate estimate of success (i.e. revenue) based on a potential label partner's current metrics
2. How that label partner can get to the next level by boosting or tweaking certain features

Lake Street Dive

This proposal will use Lake Street Dive's newest album, Side Pony, as an example.

The data will be sourced from various sources such as:

- Youtube API
- Spotify API
- Twitter API
- Internal databases

Spotify API


```
{
  "album_type" : "album",
  "artists" : [ {
    "external_urls" : {
      "spotify" : "https://open.spotify.com/artist/3nuc29fYGlQbIrwh4yrNWd"
    },
    "href" : "https://api.spotify.com/v1/artists/3nuc29fYGlQbIrwh4yrNWd",
    "id" : "3nuc29fYGlQbIrwh4yrNWd",
    "name" : "Lake Street Dive",
    "type" : "artist",
    "uri" : "spotify:artist:3nuc29fYGlQbIrwh4yrNWd"
  } ],
  "copyrights" : [ {
  "external_ids" : {
  "external_urls" : {
  "genres" : [ ],
  "href" : "https://api.spotify.com/v1/albums/1pXxKrTopCyALgXX7h5tmX",
  "id" : "1pXxKrTopCyALgXX7h5tmX",
  "images" : [ {
  "label" : "Nonesuch",
  "name" : "Side Pony",
  "popularity" : 52,
  "release_date" : "2016-02-19",
  "release_date_precision" : "day",
  "tracks" : {
  "type" : "album",
  "uri" : "spotify:album:1pXxKrTopCyALgXX7h5tmX"
}
```


GET https://api.spotify.com/v1/albums/1pXxKrTopCyALgXX7h5tmX?market=US

This API call is specifically for the Side Pony album metadata.

The Spotify API is unique in that it provides a ‘popularity’ metric that seems to be a result of their own algorithm. This is an excellent feature that can **reduce the dimensionality** of the feature set by taking many factors (e.g. plays on Spotify, plays per day, etc.) into account in one metric.

Twitter API (tweet text)


```
{
   "created_at":"Fri Apr 07 14:44:37 +0000 2017",
   "id":850358656042496001,
   "id_str":"850358656042496001",
   "text":"I've been mildly obsessed with the album Side Pony
      from Lake Street Dive over the past couple weeks.
      Such energy! https:\/\/t.co\/e2cRk0bXRa",
   "truncated":false,
   "entities":{
   "metadata":{
   "source":"\u003ca href=\"http:\/\/twitter.com\"
      rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e",
   "in_reply_to_status_id":null,
   "in_reply_to_status_id_str":null,
   "in_reply_to_user_id":null,
   "in_reply_to_user_id_str":null,
   "in_reply_to_screen_name":null,
   "user":{
   "geo":null,
   "coordinates":null,
   "place":null,
   "contributors":null,
   "is_quote_status":false,
   "retweet_count":0,
   "favorite_count":0,
   "favorited":false,
   "retweeted":false,
   "possibly_sensitive":false,
   "lang":"en"
},
```


GET https://api.twitter.com/1.1/search/tweets.json?q=lake%20street%20dive%20side%20pony

This API call searched for tweets that contain ‘lake street dive side pony’.

The ‘text’ object provides tweets that match this query. Performing **NLP sentiment analysis** on this tweets will allow for feature extraction that is directly related to consumers. This can measure if the label partner’s releases had a **positive or negative reception** and also possibly identifying certain user groups that particularly enjoy their style of content.

Twitter API (user data)


```
"user":{
   "id":1895347069,
   "id_str":"1895347069",
   "name":"Brandon Post Music",
   "screen_name":"bpostmusic",
   "location":"Toronto, Canada.",
   "description":"Earthling. Human. Musician. Imago Dei.\n
      Work hard, stay humble, pursue passion.\n
      https:\/\/t.co\/peD2H9fxhz",
   "url":"https:\/\/t.co\/peD2H9fxhz",
   "entities":{
   "protected":false,
   "followers_count":137,
   "friends_count":281,
   "listed_count":7,
   "created_at":"Sun Sep 22 22:40:10 +0000 2013",
   "favourites_count":1148,
   "utc_offset":-25200,
   "time_zone":"Pacific Time (US & Canada)",
   "geo_enabled":false,
   "verified":false,
   "statuses_count":1246,
   "lang":"en",
```


This is the user object from the previous twitter API call. This provides **demographic** and general popularity (followers_count or friends_count) of the user who is tweeting about the content in question.

This data will allow for more granular analysis of tweets by being able to segregate sentiments by region (e.g. Toronto, Canada in this example).

YouTube API (channel comments)


```
"kind": "youtube#commentThread",
"etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/obD55z_hsG2Ztf_u2LFofP6HPhk\"",
"id": "z121dlhicnecu31au04ce1grwma4xldi2u00k",
"snippet": {
 "channelId": "UCuyL6pUjQT5DzQ-ct_DGt8w",
 "videoId": "KqEiWN44L3M",
 "topLevelComment": {
  "kind": "youtube#comment",
  "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/qZdlA9xQbsEZdz-KPjS9ZLLdrhU\"",
  "id": "z121dlhicnecu31au04ce1grwma4xldi2u00k",
  "snippet": {
   "authorDisplayName": "R0b Cruz",
   "authorProfileImageUrl": "https://yt3.ggpht.com/-5uPh3MWxCZw/
     AAAAAAAAAAI/AAAAAAAAAAA/VZqohjs52CE/s28-c-k-no-mo-rj-c0xffffff/photo.jpg",
   "authorChannelUrl": "http://www.youtube.com/channel/UCSicV6yAmckBvtjsRKUrTsw",
   "authorChannelId": {
    "value": "UCSicV6yAmckBvtjsRKUrTsw"
   },
   "channelId": "UCuyL6pUjQT5DzQ-ct_DGt8w",
   "videoId": "KqEiWN44L3M",
   "textDisplay": "That bass player is out of this world
     and damn that vocalist gives me chills.",
   "textOriginal": "That bass player is out of this world
     and damn that vocalist gives me chills.",
   "canRate": true,
   "viewerRating": "none",
   "likeCount": 0,
   "publishedAt": "2017-04-14T14:51:36.000Z",
   "updatedAt": "2017-04-14T14:51:36.000Z"
  }
 },
 "canReply": true,
 "totalReplyCount": 0,
 "isPublic": true
```


GET https://www.googleapis.com/youtube/v3/commentThreads?part=snippet%2Creplies&allThreadsRelatedToChannelId=UCuyL6pUjQT5DzQ-ct_DGt8w&key={YOUR_API_KEY}

This YouTube API provides comments on a particular artist's channel. Again, **NLP sentiment analysis** can also be used here to measure an artist's popularity. Additionally, mining these comments for keywords can also provide key insights into why a certain artist is doing well. In this example, the user comments on the exceptional bass player and a talented vocalist. Both of these keywords can be used to **generate more features** for the feature set.

Internal Data (from INgrooves)

After drawing together the data from outside sources, this data should be linked to all the internal data available on a particular partner such as:

- Genre
- Sales / marketing metrics
- Content subject
- Average age of artist/band
- Ethnicity
- Band size
- Style of play
- Featured instruments
- Singer gender/style

Data Transformation

The dataset will require a great deal of transformation to prepare it for classification/regression analysis.

For example, categorical columns will have to be converted to a Boolean column for each category. This creates a more balanced matrix for the algorithms to work with.

E.g.

Partner	Genre
Artist 1	Jazz
Artist 2	Country/Rock
Artist 3	Pop
Artist 4	Rock

Partner	Jazz	Country	Pop	Rock
Artist 1	1	0	0	0
Artist 2	0	1	0	0
Artist 3	0	0	1	0
Artist 4	0	1	0	1

Data Transformation

Other non-trivial data matters to address:

- How to convert keywords from tweets and YouTube comments into features
  - How many mentions will a keyword require to be considered significant?
  - Suggestion: Find 5 most keywords that are most common in describing artists and search for these for in partner's social media

- How to store positive versus negative reception in social media (sentiment analysis)
  - Suggestion: store as ratio
    - E.g. Positive: .62, Negative: .20, and neutral sentiments are left out

Implementation of Algorithm

The dataset should be analyzed in two ways:

Regression – Applying supervised learning by using the data as a training set to predict a partner's revenue generated from partnering with INgrooves.

Feature Analysis – Analyze trained model to see which features are most important and the ways in which a partner can improve its success

Regression: Random Forest

The suggested model is a Random Forest Regressor from the sklearn.ensemble module.

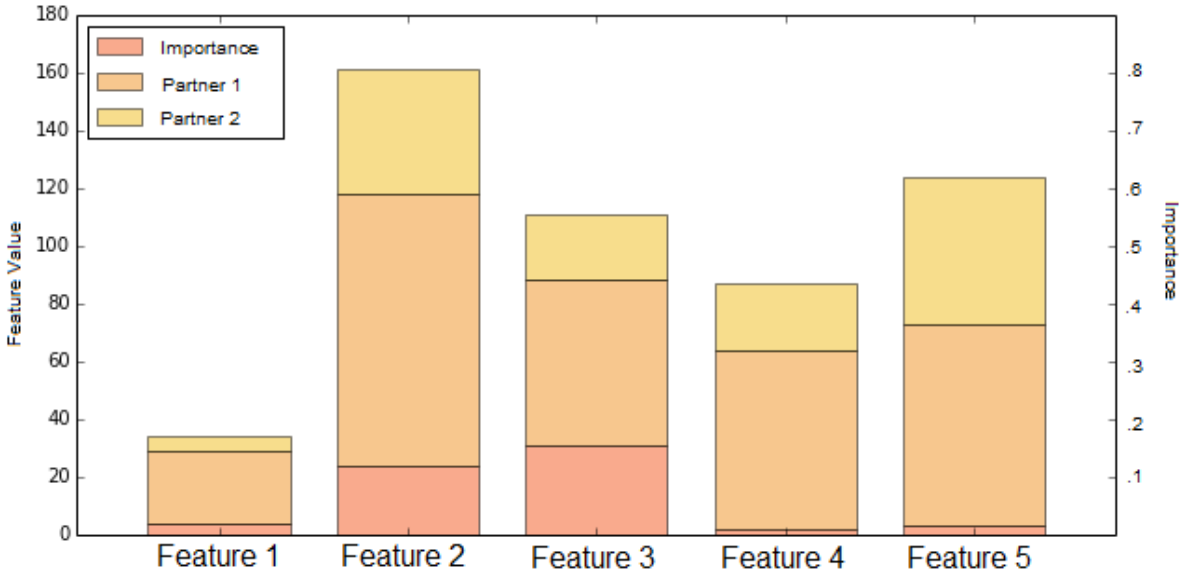
This model builds multiple decision trees on subsamples of the data and features and averages the results to optimize accuracy and control overfitting.

Before fitting the data to the regressor, the data is split into training and validation sets using sklearn.cross_validation.train_test_split. This will allow for validation of the trained model using an independent dataset.

When instantiating the regressor, two parameters to pay attention to are n_estimators (the number of trees in the forest) and max_depth (the maximum depth of a tree).

Essentially, this model will be trained to predict a partner's revenue generated from partnering with INgrooves.

Feature Analysis



Once the model is trained, the feature importance can be extracted and graphed.

These relative importances indicates how heavily a specific feature contributes towards predicting the outcome. This will be useful to determine which features a partner needs to work on first.

On this graph, partner 1 and partner 2's feature values are stacked on top of the importances. Partner 1 represents a partner trying to reach partner 2's success.

Graphing feature importance vs. partner 1 vs partner 2 will give an insightful look into how far a partner is from getting to the 'next level' and which features to prioritize in doing this.

Conclusion

Using API calls to various social media and digital streaming services can provide a robust set of features to apply Machine Learning algorithms to.

NLP sentiment analysis on users' comments and tweets can further organize this data.

After fitting the data to the Random Forest regressor, one can predict revenue generated from partnering INgrooves.

The trained model also provides feature importances, which can be used to suggest to partners where to prioritize improvement and/or changes for greater success.