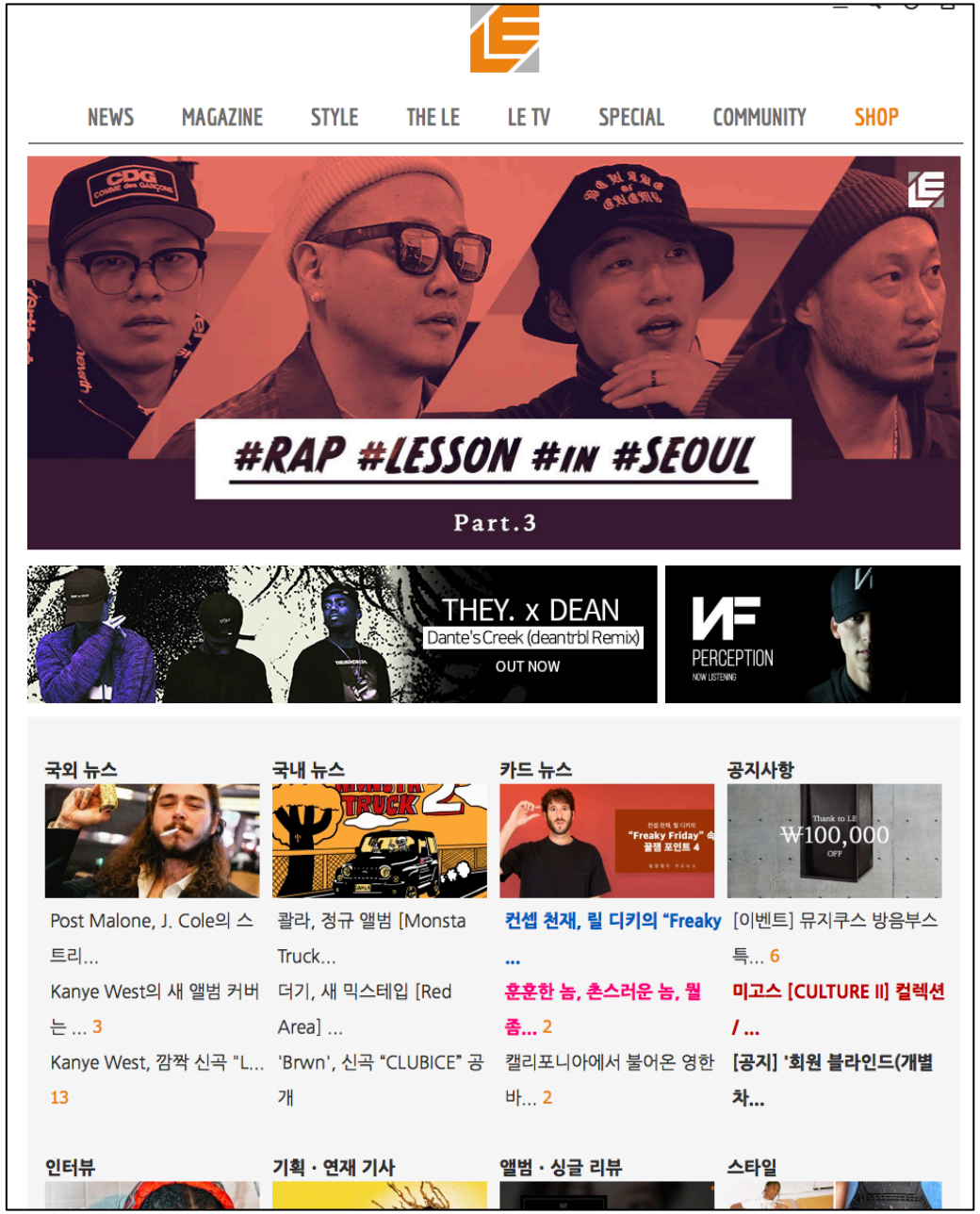
CONTENT-WORTH DEBUT ARTIST CLASSIFIER

Wonyoung Lucas Seo

INTRODUCTION

HIPHOPLE.com ??????????

- US HIPHOP/R&B based Web Magazine platform
- Major Contents
  - Translation
    - Lyric Translation(* this is the part I contribute)
    - Subtitled Music Video
    - News of HIPHOP / R&B Industry
  - Magazine
    - Feature Articles
    - Artist Interview
    - Album Review
  - Fashion / Life Style



Project Inspiration

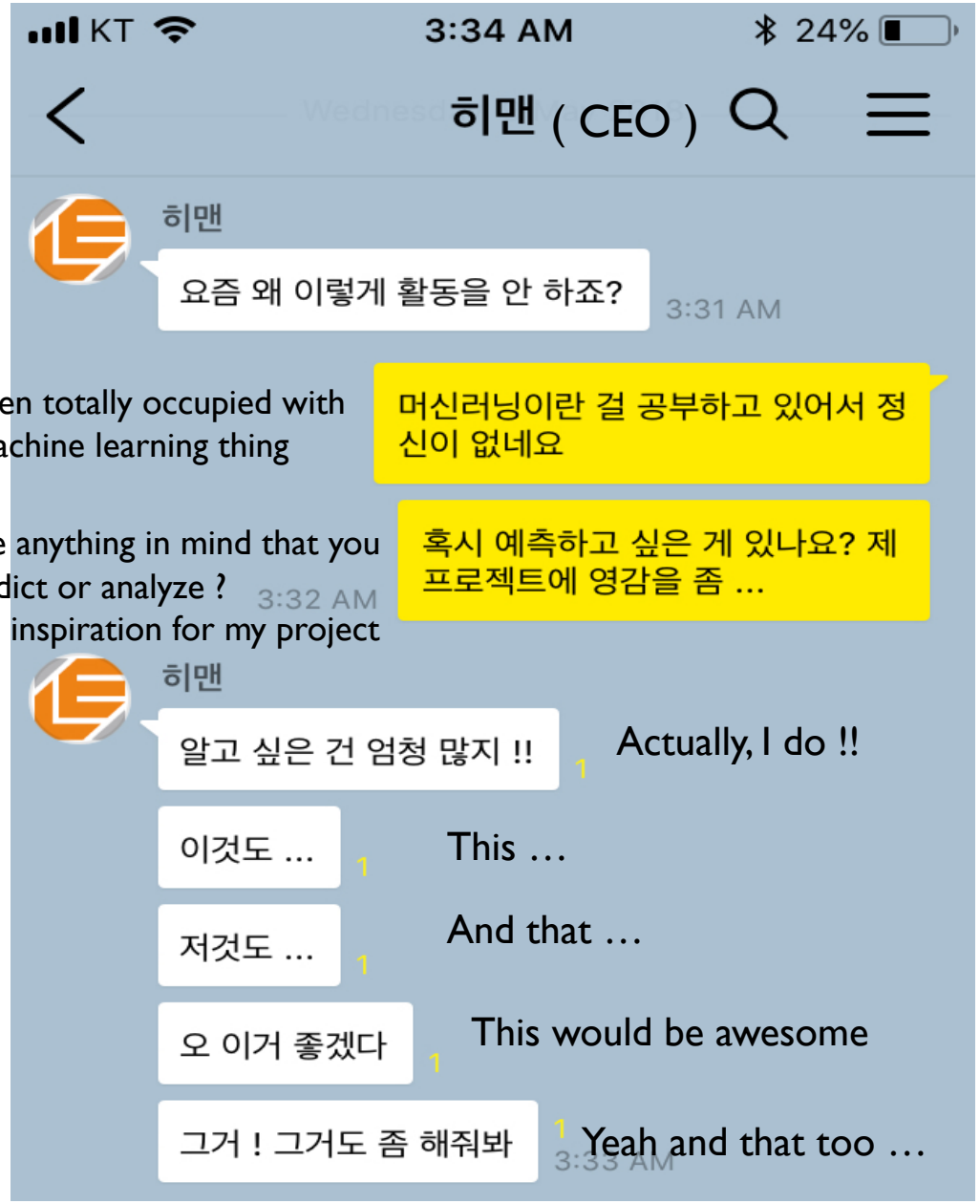
- **CEO's needs**

  - Public reputation / preference analysis for merchandise sales
  - K-Hiphop buzz analysis in South-East Asian region media
  - Super Rookie Artist prediction analysis
  - Messenger app chatbot
  - And many more …

Why don't I see you working these days?

I've been totally occupied with this machine learning thing

Do you have anything in mind that you want to predict or analyze ?
I need some inspiration for my project



Predict Super Rookie ?? (… sounds cool. Maybe I should first focus on this one)

Setting More Specific Goal …

Let's predict Super Rookie!!!

Hiphople Magazine Team produces contents of successful debut artists

When producing contents on a debut artist … :

(i) Based on the domain/industrial knowledge , editors decide whether or not to produce a content on thie particular rookie artist.

(ii) Spend considerable amount of time to research them.

(iii) Or loose chance to become the first press to write about the aritst to other press.

How can we reduce research time by assigning priority list?
&
Improve decision making process?

Let's predict whether the rookie artist

is worth creating a content & publish /

distribute to our subscribers with Machine

Learning classification technique based on

the quantitative information

HIPHOPLE's contents =

Review + Interview + Feature Articles + SNS Account + @

[인터뷰] 디자이너 (Desiigner) 2017.12.29 20:18

LE_Magazine 조회 수 12432 추천 수 16 댓글 31

일약 스타덤에 올랐다는 표현이 정확하다. 이 97년생의 어린 래퍼는 첫 데뷔 싱글을 통해 전 세계를 가뿐히 정복했다. 빌보드 싱글 차트 1위를 차지하고 500만 장 이상의 높은 판 행하고, 머나먼 한국에까지 특유의 에너지를 (Desiigner)가 지난 12월 21일 아시아 투어의 통해 호평을 받으며 성공리에 무대를 마친 디자 기를 나누고 왔다. 디자이너의 유쾌한 에너지와

LE: 반갑다. 간단하게 인사를 부탁한다.

요! 때가 왔어. 디자이너야. 한국 팬들에게 샤라

LE: 먼저 당신의 대표곡에 대해서 이야기해보고

축복 그 자체야. 모든 게 가장 좋은 쪽으로 바뀌 게 해준 노래야. 비즈니스에 집중하게 됐고, 가 지금도 내 주변에 계속 함께 하는 이들이 있기도 말 감사해. 판다 판다 판다

[앨범 - 신보] [앨범] Daniel Caesar - Freudian 2017.09.09 02:20

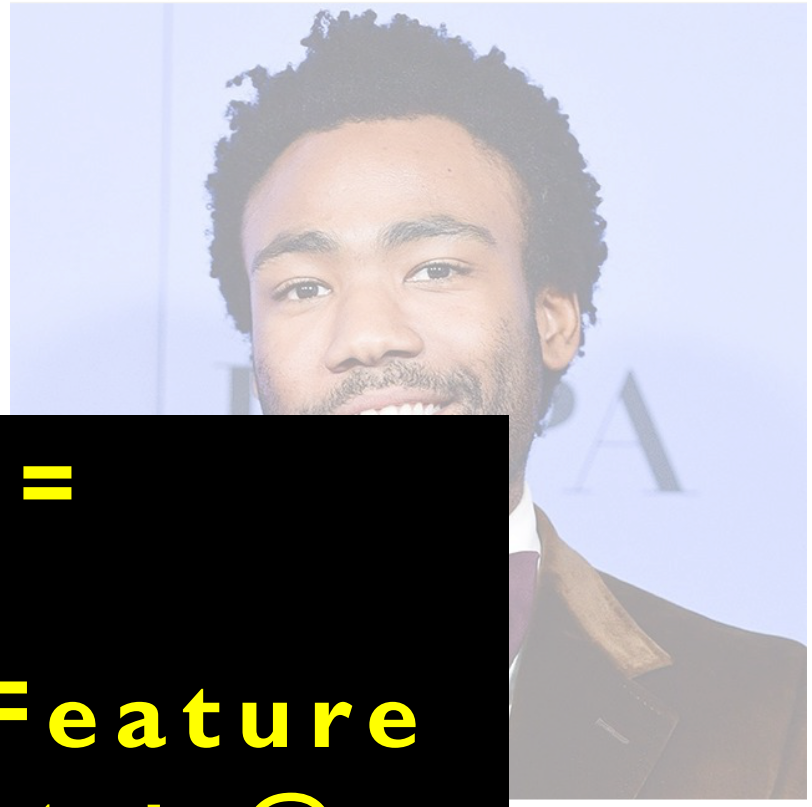
LE_Magazine

(124.5.114.105) 조회 수 3271 추천 수 5 댓글 3

발매일	2017/08/25
레이블	Independent
한줄평	내 영감의 원천은 오로지 당신.

캐나다 흑인음악 씬이 계속해서 심상치 않다. 시작은 드레이크(Drake)와 OVO 사운드(OVO Sound), 위켄드(The Weeknd)와 XO였다. 이전까지의 움직임이 아예 없었던 건 아니겠지만, 대중적인 관심이 쏠린 건 그 두 뮤지션이 활약하기 시작했던 2010년대 초반부터였다. 그들은 인기를 얻음과 동시에 자신의 레이블에 같은 연고인 온타리오 주의 뮤지션들을 영입했다. 그때까지만 하더라도 얼터너티브 알앤비라는 음악적 공통분모에 의한 결집 정도로 보였다. 하지만 이후에도 캐나다에서는 이것저것 섞고, 공간감을 살리고, 톤 다운된 감성을 내뿜는 아티스트들이 꾸준히 등장

hiphople



Liked by **shawna_le_**, **iandlals** and **799 others**

hiphople Childish Gambino, SNL에서 신곡 공개

_

올해 더 이상 컴백할 아티스트는 없다고 생각한 순간, 이번에는 Donald Glover가 Childish Gambino로서 신

PROJECT WORKFLOW

1. Define Variables / Input
2. Collect Data & Store AWS MySQL DB Server
3. Labeling Target Variable Classes
4. ORM Data Query
5. Merging Data
6. Fitting Models
7. Feature Engineering
8. Model Cross Validation and Optimization

DEFINE VARIABLES / INPUT

Target Variable

Binary Classification

1 : Worth-creating a content with this artist

0 : Ignore this artist

HIPHOPLE editors
participated in labeling process based
on their domain knowledge

Independent Variables (Features, Input)

Rows : **Debut Album** released from Jan 1st 2011 to Apr 18th 2018

- Genre

- Count of single album released before the official album release

- Count of articles published by Music / Culture Media Press

- Ratings from Music Media Press

- Number of SNS followers of each artist

COLLECTING DATA (CRAWLING)

Data Crawling → Stored in AWS EC2 MySQL DB

Crawling Data	Crawling Target	Crawling Python Library
Debut album by year	www.wikipedia.com	Requests, BeautifulSoup
Ratings	www.metacritic.com	Scrapy
	www.pitchfork.com	Requests
	www.albumoftheyear.com	Requests, BeautifulSoup
SNS followers of each artist	www.chartmetric.io	Selenium
Count of related article	www.billboard.com	Requests, BeautifulSoup
	www.genius.com	Requests, BeautifulSoup
	www.thesource.com	Requests, BeautifulSoup
	www.xxlmag.com	Requests, BeautifulSoup

TARGET VARIABLE CLASS LABELING

Labeling

- Editors participated in labeling the class as 1 , 0
- Labeling sheet was shared via Google Spreadsheet

	C	D	E	F	G
	genre	Yes / NO			
	hiphop	0			
	hiphop	0			
	hiphop	0			
	hiphop	0		왼쪽의 데이터는 2011년~현재 사이 발매 된 데뷔 스튜디오 앨범의 목록입니다.	
	hiphop	0		과거 힙합엘이에서 왼쪽의 앨범을 가지고 컨텐츠로 다루었던	
	hiphop	1		이력을 기재해주시면 됩니다.	
	hiphop	1		** 컨텐츠는 매거진팀 기획기사, 리뷰, SNS컨텐츠 등을 뜻합니다	
	hiphop	0		** 국외 뉴스, 발매 뉴스 등은 포함하지 않습니다	
ne Story	hiphop	1		** 너무 고민 마시고 술술 기재해주세요	
	hiphop	0			
	hiphop	1			
	hiphop	0		예시)	
	hiphop	0		과거 컨텐츠로 다룬 적이 있다 ---> 1	
	hiphop	0		기억이 안 나거나, 다룬 적 없지만, 다룰 만 하다 ---> 1	
	hiphop	1		컨텐츠를 만들 지 않았다 / 다룰 필요가 없다 ---> 0	
	hiphop	1			
	hiphop	1		예시)	
h Love II	hiphop	0		J. Cole - Sideline Story (hiphop) ---> 1	
	hiphop	0		Sevyn Streeter - Girl Disrupted ---> 0	
ver Told	hiphop	0		DRAM - Baby DRAM ---> 1	
	hiphop	1			
	hiphop	0			
	hiphop	1			

- **If you had created any contents on this artist/album please label it as " 1 " in roman numerals**
- **If you don't remember but still think its worth creating contents, please label it as " 1 "**
- **If you didn't created any contents related to this artist/album, please label it as " 0 "**

MERGE DATA

Data size : 1083 Rows

- Label (Target Variable)

Features : 15

Album Info

- Genre
- Count of Single Albums

Ratings

- Album of the Year
- Pitchfork
- Metacritic

Aritcle Counts

- Billboard
- Genius
- The Source
- XXL Magazine

SNS

- Twitter followers
- Instagram followers

- Facebook page Likes
- Youtube Channel subscribers

- Soundcloud followers
- Spotify followers

FITTING MODELS

Classification Algorithms: Scikit-Learn Libraries

Model	Classification Metrics (Class 1)				Comment
	Precision	Recall	F 1 Score	AUC	
KNN : K-Nearest Neighbours	0.79	0.43	0.55	0.83	
SGD : Steepest Gradient Decent	0.31	0.92	0.46	0.66	
Decision Tree	0.62	0.65	0.63	0.86	
Random Forest	**0.74**	**0.67**	**0.70**	**0.91**	★
Gradient Boosting	0.69	0.57	0.62	0.93	
XGBoost : Extra Gradient Boosting	0.78	0.64	0.70	0.93	

Reason for selecting Random Forest Classifier

Recall (Sensitivity) was major consideration for this project

- The impact of **False Positive (FP)** in case of considering **Precision**

$$Precision = \frac{TP}{TP + FP}$$

  - Incorrect prediction : answered Class 1 (prediction) to actual Class 0 data (Real)
  - Create content even though the artist & album not valued.
  - Spend more time and energy on incorrect class → Wasted internal resources

- The impact of False Negative (FN) in case of considering **Recall**

$$Recall\ /\ Sensitivity = \frac{TP}{TP + FN}$$

  - Incorrect prediction : answered Class 0 (prediction) to actual Class 1 data (Rea)
  - Ignore the artist & album that we shouldn't have ignored

    - **Increase possibility of other media creating content first hand**
    - **Increase possibility of losing potential subscribers**
    - **Increase possibility of losing opportunity of arranging management contract with the artist for management in South Korea**

FEATURE ENGINEERING

Feature Engineering Trials

Feature	Type	Descriptions	Performance Improvement
Genre	One-Hot Encoding	Create Dummy Variable with One-Hot Encoding	O
Ratings	Null Value Imputation	Fill in " 0 " rating score for the album that didn't received any ratings	O
Buzz Rating SNS Followers	Binning	- Binning the range of numeric values - Create Dummy Variables for each bin	X
SNS Followers	Ratio	- [Follower of Each SNS / Total Followers] - Calculate Ratio for each SNS platform	X
SNS Followers	Re-Grouping	- Private SNS (Twitter, Instagram) - Page & Channel based SNS (Facebook, Youtube) - Music based SNS (Soundcloud, Spotify)	X

IMPROVING PERFORMANCE & OPTIMIZATION

Recall Score Improvement

Imbalance Problem

- Number of a Class 1 is far less than Class 0 …
    - Classifier measures performance based on "Accuracy"
    - Then Classifier biased towards the major classes and hence show very poor classification rates on minor classes → still shows fairly good accuracy

- Applied **Under Sampling and** adjust Class ratio to 50:50
- Trade-off Precision → Focus on Recall Score

Under Sampling : Imblearn Library

Model	Classification Metrics (Class 1)				Comment
	Precision	Recall	F 1 Score	AUC	
ClusterCentroids	0.45	0.95	0.61	0.86	
Random Under Sampler	**0.67**	**0.88**	**0.89**	**0.94**	★
CondensedNearestNeighbour	0.61	0.80	0.69	0.92	
AllKNN Metrics	0.54	0.89	0.67	0.92	
InstanceHardnessThreshold	0.60	0.89	0.72	0.92	
NearMiss	0.32	0.95	0.47	0.74	
NeighbourhoodCleaningRule	0.60	0.80	0.69	0.93	
OneSidedSelection	0.76	0.73	0.75	0.92	
TomekLinks Metrics	0.75	0.71	0.73	0.95	

Optimization : Scikit-Learn GridSearchCV

- **Apply GridSearch Result**
- **Tune Hyper Parameter**

```
%%time
gs_result = gs.fit(X_resampled_rus, y_resampled_rus)
```

```
CPU times: user 1h 32min 3s, sys: 47.4 s, total: 1h 32min 50s
Wall time: 1h 37min 8s
```

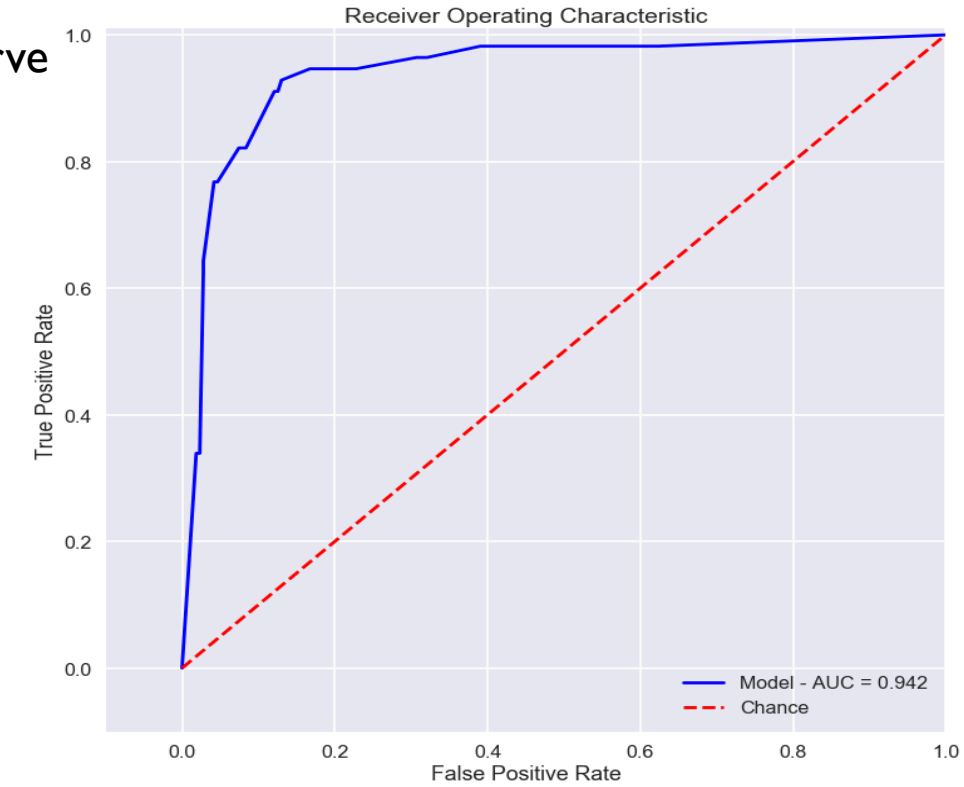
```
gs.best_params_
```

```
{'bootstrap': True,
 'class_weight': None,
 'criterion': 'entropy',
 'max_depth': 10,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 4,
 'min_samples_split': 3,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 10,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Classification Report

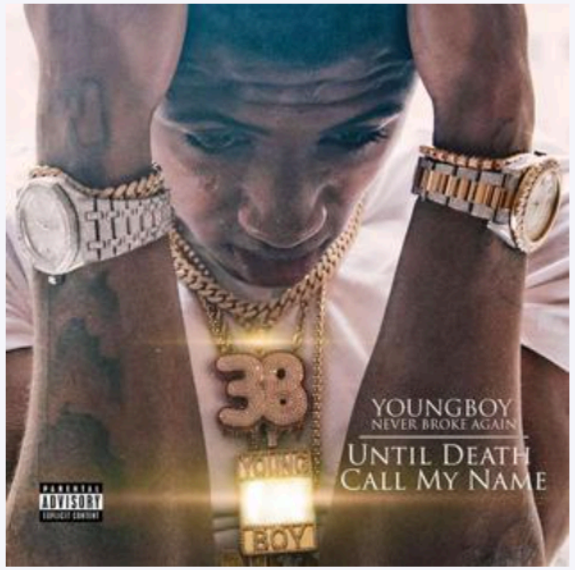
	Precision	Recall	F1 Score
Class 0	0.98	0.87	0.92
Class 1	0.65	0.93	0.76
Avg / Total	0.91	0.88	0.89

ROC Curve AUC



TEST NEW DATA !

Until Death Call My Name



Studio album by YoungBoy Never Broke Again

Released	April 27, 2018
Recorded	2017–2018
Genre	Hip hop · trap
Label	Never Broke Again · Atlantic
Producer	1040 · Ben Billions · Big Korey · Bighead · CashMoneyAP · DJ Montay · DJ Swift · DMacTooBangin · Drumma Boy · Dubba-AA · Judge · Kenny Beats · Mook · Schife Karbeen · TM88 · Wheezy

Youngboy Never Broke Again – Until Death Call My Name

- **Release date : Apr 27th 2018 (Genre : Hiphop)**
- **Debut Studio Album**

*** Answer * HIPHOPLE hasn't produced any contents for this artist and album as of May 9th**

Then … how would the model predict on this album?

Load Model

In [2]:
```
model = pickle.load(open("rfc_rus_gs.plk", "rb"))
```

Load Model

Input New Data

In [3]:
```
data = {
    'single_count': [2],
    'genre_funk': [0],
    'genre_hiphop': [1],
    'genre_pop': [0],
    'genre_rnb': [0],
    'genre_soul': [0],
    'freq_billboard': [1],
    'freq_genius': [0],
    'freq_theSource': [0],
    'freq_xxl': [42],
    'rating': [69],
    'twitter': [912000],
    'instagram': [3900000],
    'facebook': [581939],
    'youtube': [2170000],
    'soundcloud': [384000],
    'spotify': [854874]
}
```

Input data from the artist and album

In [4]:
```
test = pd.DataFrame.from_dict(data)
```

Predict

In [5]:
```
model.predict(test)
```

Out[5]: array([0])

Class Predict Result:

0 → Ignore this artist and album

Predict Probability

In [6]:
```
model.predict_proba(test)[:, 1]
```

Out[6]: array([0.3])

Probability Prediction Result:

30% → Low Probability → Low Priority

OH - YEAH

HOW CAN I IMPROVE FURTHER?

List of To-Do's

1. Research other method to tackle Imbalanced Problem to also increase Precision

2. Try XGBoost Algorithm and tune its best Hyper Parameter

3. Small Dataset

    - Accumulate more data for debut albums in future and reflect on model
    - Crawl input data from more media press
    - Text sentiment analysis would be a plus

4. Dash-based visualization web application

    - Automation of Crawling Input Data + Prediction
        - Crawl Feature Data → update dataset
        - Load Pickle filed model → Predict and return result

For more information on this project …

Please check the links below :D

Github Repository Source Code
https://github.com/lucaseo/content-worth-debut-artist-classification-project

Data Visualization Web Application
http://le-rookie-clf.herokuapp.com

And about me …

Please visit the links below :)

Github
http://www.github.com/lucaseo

LinkedIn
http://www.linkedin.com/in/lucaseo

Github Blog
http://www.lucaseo.github.io

THANK YOU !