Predict IMDB score low or high by decision tree

Puvit Pracharktam 6031830321

1 Data Preparation and Preprocessing

Amount of empty data or NaN in each rows
From 5043 datasets

director_name	104
num_critic_for_reviews	50
duration	15
director_facebook_likes	104
actor_3_facebook_likes	23
actor_2_name	13
actor_1_facebook_likes	7
gross	**884**
genres	0
actor_1_name	7
movie_title	0
num_voted_users	0
cast_total_facebook_likes	0
actor_3_name	23
facenumber_in_poster	13
plot_keywords	153
movie_imdb_link	0
num_user_for_reviews	21
language	12
country	5
content_rating	303
budget	**492**
title_year	108
actor_2_facebook_likes	13
imdb_score	0
aspect_ratio	329
movie_facebook_likes	0

APPROACH 1 num_df

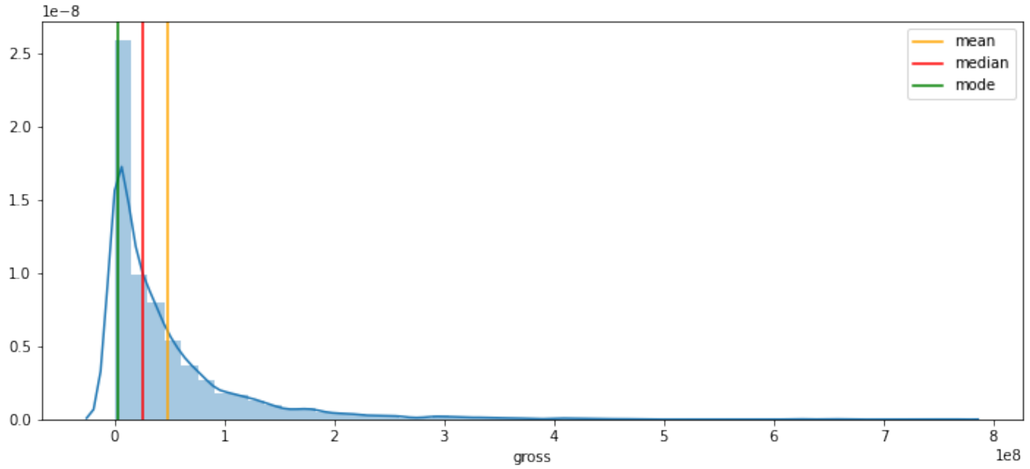
Select numerical data

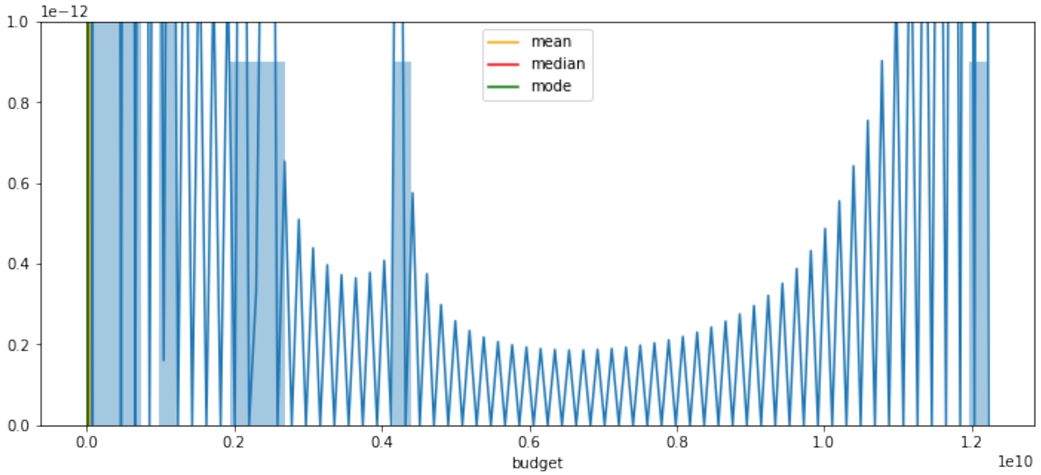
X =

num_critic_for_reviews	num_voted_users
duration	cast_total_facebook_likes
director_facebook_likes	facenumber_in_poster
actor_1_facebook_likes	num_users_for_reviews
actor_2_facebook_likes	content_rating
actor_3_facebook_likes	budget
gross	title_year
	movie_facebook_likes

Y = imdb_score

Has NaN data more than 10% of dataset

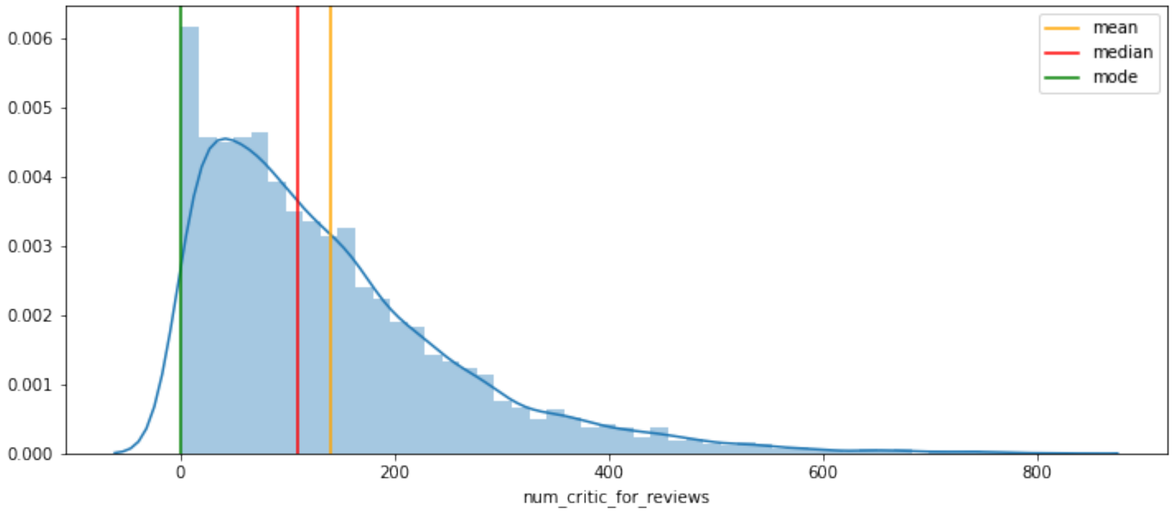




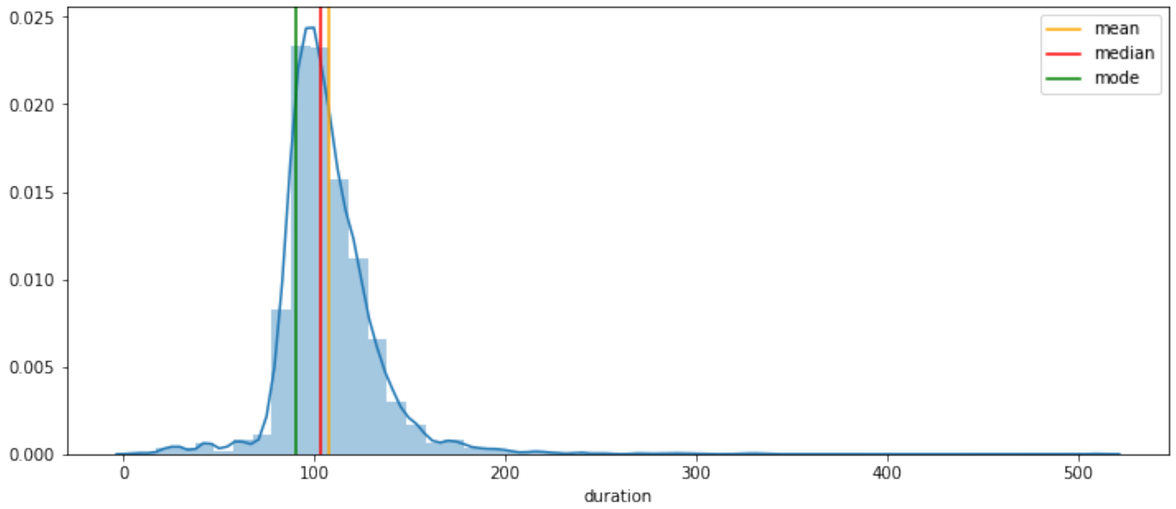
The graph shows that it is very hard to impute mean or median

Has NaN data more than 10% of dataset

Drop ‘gross’ and ‘budget’



Safe to impute **num_critic_for_reviews**
By **mean**



Safe to impute **duration** by **median**

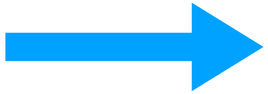
Impute **the rest** by **ffill**

If row n null impute row n by row n-1

APPROACH 2

num_with_genres_df = **Numerical data + genres**

0	Action|Adventure|Fantasy|Sci-Fi
1	Action|Adventure|Fantasy
2	Action|Adventure|Thriller
3	Action|Thriller
4	Documentary
...	...
5038	Comedy|Drama
5039	Crime|Drama|Mystery|Thriller
5040	Drama|Horror|Thriller
5041	Comedy|Drama|Romance
5042	Documentary



based on web series	estate	reference to jesus christ	retrograde narrative	futuristic city	interpol	haunted	female surfer	film starts with text	based on web series
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0

Split each genres to Boolean and append to **num_df**

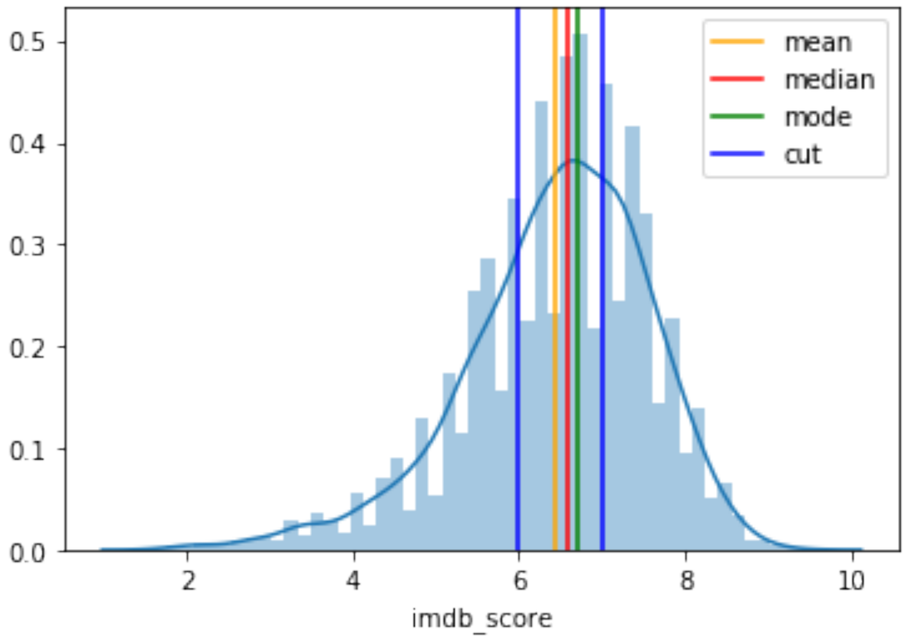
APPROACH 3

num_with_genres_and_keywords = Numerical data + genres + plot keywords

```
0                  avatar|future|marine|native|paraplegic
1       goddess|marriage ceremony|marriage proposal|pi...
2                     bomb|espionage|sequel|spy|terrorist
3       deception|imprisonment|lawlessness|police offi...
4                                                     NaN
                              ...
5038               fraud|postal worker|prison|theft|trial
5039         cult|fbi|hideout|prison escape|serial killer
5040                                                  NaN
5041                                                  NaN
5042    actress name in title|crush|date|four word tit...
Name: plot_keywords, Length: 5043, dtype: object
```

based on web series	estate	reference to jesus christ	retrograde narrative	futuristic city	interpol	haunted	female surfer	film starts with text	based on web series
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0

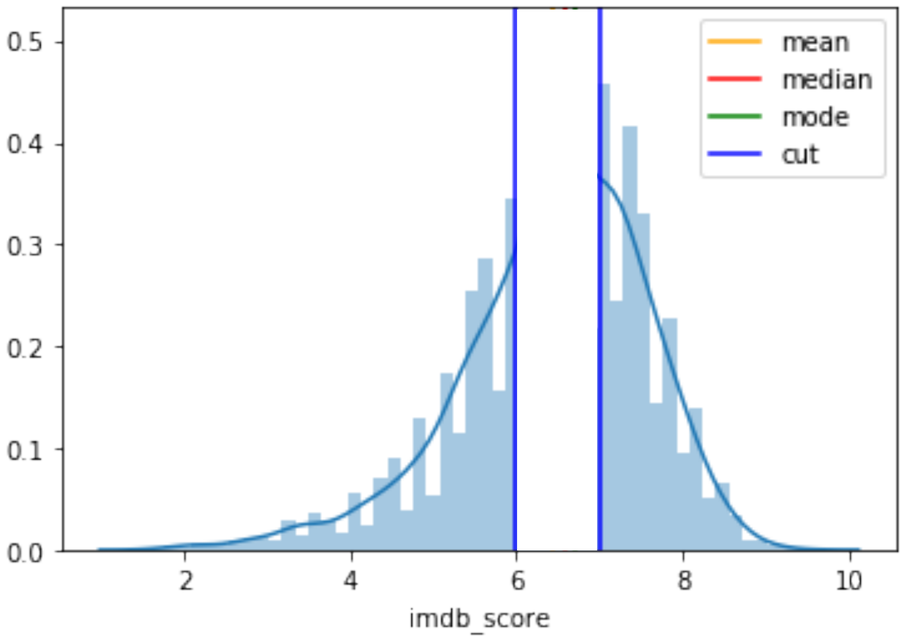
2 Visualize & Analysis



Mean : 6.44
Median : 6.6
SD : 1.2

CUTOFF 6-7

Define target variable



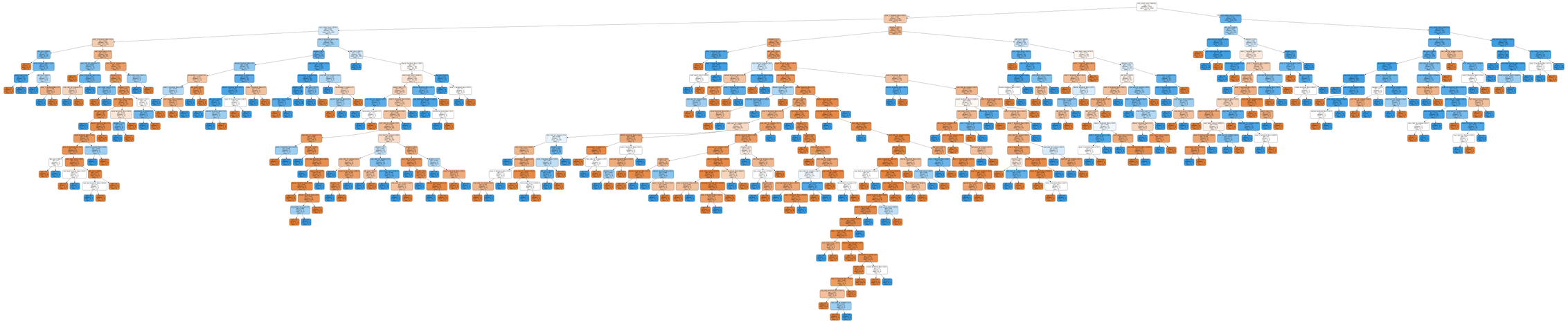
< 6.0
is_score_high = 0

> 7.0
is_score_high = 1

3 Modeling by linear regression compare to decision tree

3A APPROACH 1 num_df

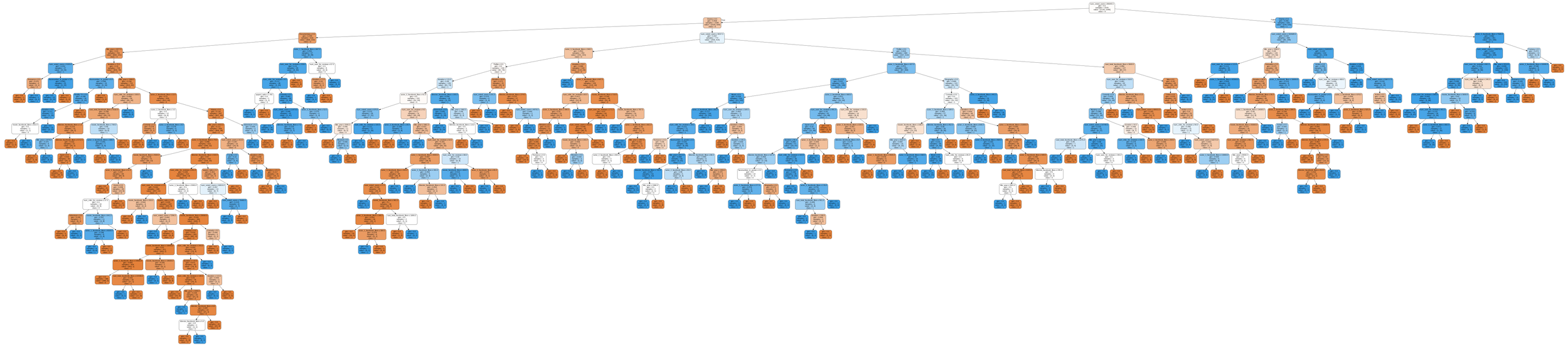
Decision Tree Result



bit.ly/bigdatahw7-6031830321

3A APPROACH 2 num_with_genres_df

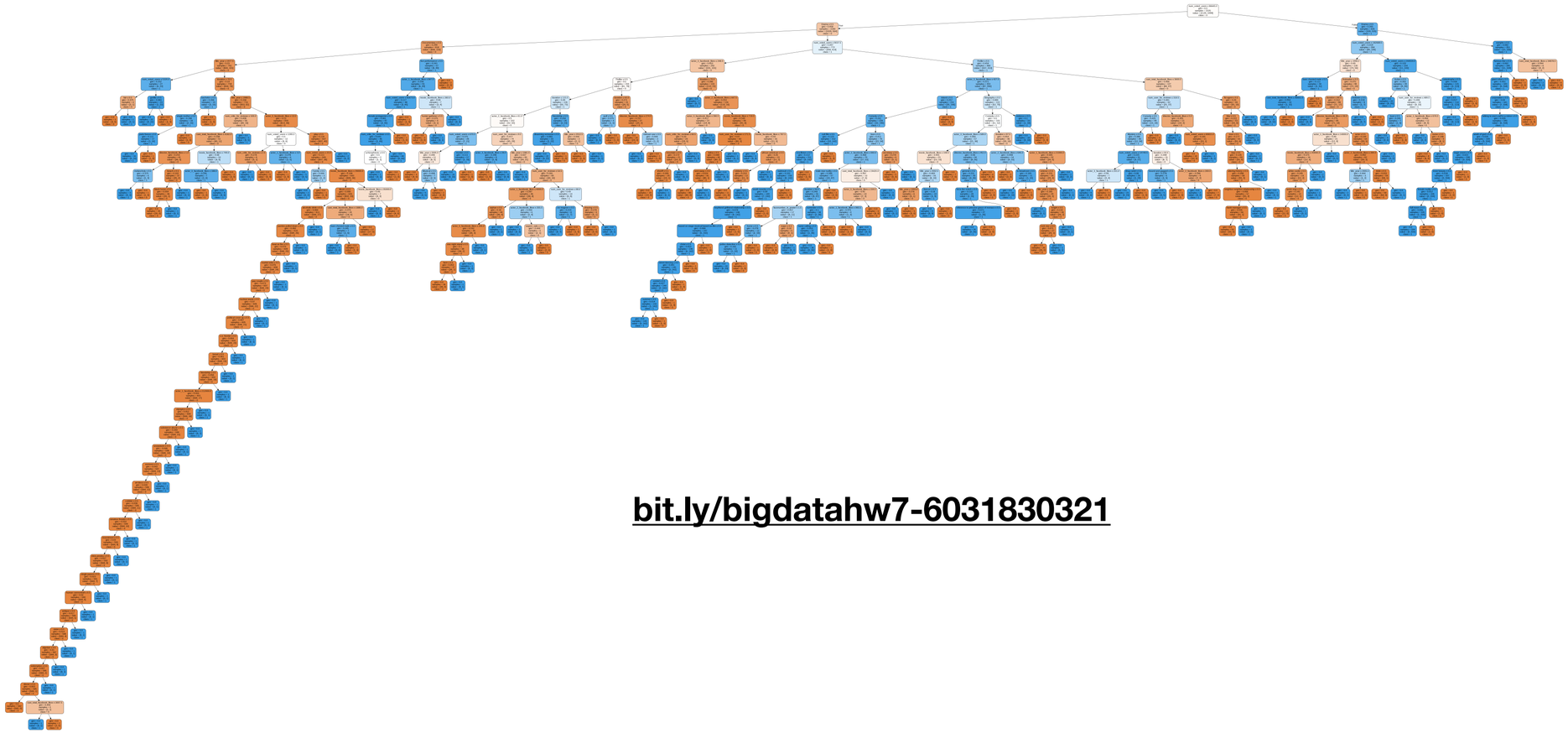
Decision Tree Result



bit.ly/bigdatahw7-6031830321

3A APPROACH 3 num_df

Decision Tree Result



bit.ly/bigdatahw7-6031830321

4 Evaluation Model

4A APPROACH 1 num_df

Decision Tree

classification_report

		precision	recall	f1-score	support
	0.0	0.83	0.79	0.81	478
	1.0	0.80	0.83	0.82	476
accur	acy			0.81	954
macro	avg	0.81	0.81	0.81	954
weighted	avg	0.81	0.81	0.81	954

Confusion_matrix

```
[[380  98]
 [ 79 397]]
```

4A APPROACH 1 num_df

Logistic Regression

classification_report

precision			recall	f1-score	support
	0.0	0.90	0.82	0.86	501
	1.0	0.82	0.90	0.86	453
accur	acy			0.86	954
macro	avg	0.86	0.86	0.86	954
weighted	avg	0.86	0.86	0.86	954

Confusion_matrix

```
[[412  89]
 [ 47 406]]
```

4A APPROACH 2 num_with_genres_df

Decision Tree

classification_report

		precision	recall	f1-score	support
	0.0	0.83	0.81	0.82	473
	1.0	0.82	0.84	0.83	481
accur	acy			0.82	954
macro	avg	0.82	0.82	0.82	954
weighted	avg	0.82	0.82	0.82	954

Confusion_matrix

```
[[382  91]
 [ 77 404]]
```

4B APPROACH 2 num_with_genres_df

Logistic Regression

classification_report

		precision	recall	f1-score	support
	0.0	0.88	0.81	0.84	495
	1.0	0.81	0.88	0.84	459
accur	acy			0.84	954
macro	avg	0.85	0.85	0.84	954
weighted	avg	0.85	0.84	0.84	954

Confusion_matrix

```
[[382  91]
 [ 77 404]]
```