Open IIT 2021: Data Analytics

REPORT

- TEAM 36

PROBLEM STATEMENT

A major record label wants to purchase the rights to a music track. It does not want to encounter any losses with promotion and distribution of the track. It needs to decide on the royalties to be paid to the artists and composers.

OVERVIEW

GIVEN

- We are given with a dataset of around 12,000 different songs released on different years ranging from 1920 to 2021.
- The dataset contains many musical features like acousticness, danceability, energy, loudness, duration of the song etc..

	id	acousticness	danceability	energy	explicit	instrumentalness	key	liveness	loudness	mode	release_date	speechiness	tempo	valence	year	duration-min	popularity
0	2015	0.949	0.235	0.0276	No	0.9270	5	0.513	-27.398	Major	01-01-1947	0.0381	110.838	0.0398	1947	3.0	very low
1	15901	0.855	0.456	0.4850	No	0.0884	4	0.151	-10.046	Major	13-11-2020	0.0437	152.066	0.8590	2020	2.4	low
2	9002	0.827	0.495	0.4990	No	0.0000	0	0.401	-8.009	Minor	01-01-1950	0.0474	108.004	0.7090	1950	2.6	very low
3	6734	0.654	0.643	0.4690	No	0.1080	7	0.218	-15.917	Major	30-04-1974	0.0368	83.636	0.9640	1974	2.4	low
4	15563	0.738	0.705	0.3110	No	0.0000	5	0.322	-12.344	Major	01-01-1973	0.0488	117.260	0.7850	1973	3.4	average

AIM

- And from that our main goal is to predict the popularity of the song and use several models and features to test it.
- But maximum accuracy is not our goal here. Our goal is to maximize the revenue generated.

Popularity	Bid Price	Expected Revenue
very high	5	10
high	4	8
average	3	6
low	2	4
very low	1	2

- So we changed our evaluation metric. Instead of using the 'accuracy_score' function we coded a separate function called 'revmax' to generate the best possible income.

```
def revmax(pred,y_test):
  leftout=10000
  revenue=0
  for i,j in zip(pred,y_test):
    if i>=j:
      if leftout>=i:
          leftout=leftout-i
          revenue=revenue+ 2*j
  print(leftout)
  print(revenue)
```

- Analytically if you see this, our goal is to reduce the number of values that occur on the bottom side of the confusion matrix and also to balance the top part.

1	2	3	4	5
0	37	7	1	0
0	11	21	0	0
0	4	138	28	1
0	3	79	513	89
0	1	11	115	950

PROCESS

01

STEP 1

Exploratory Analysis

STEP 2

Feature Engineering

02

03

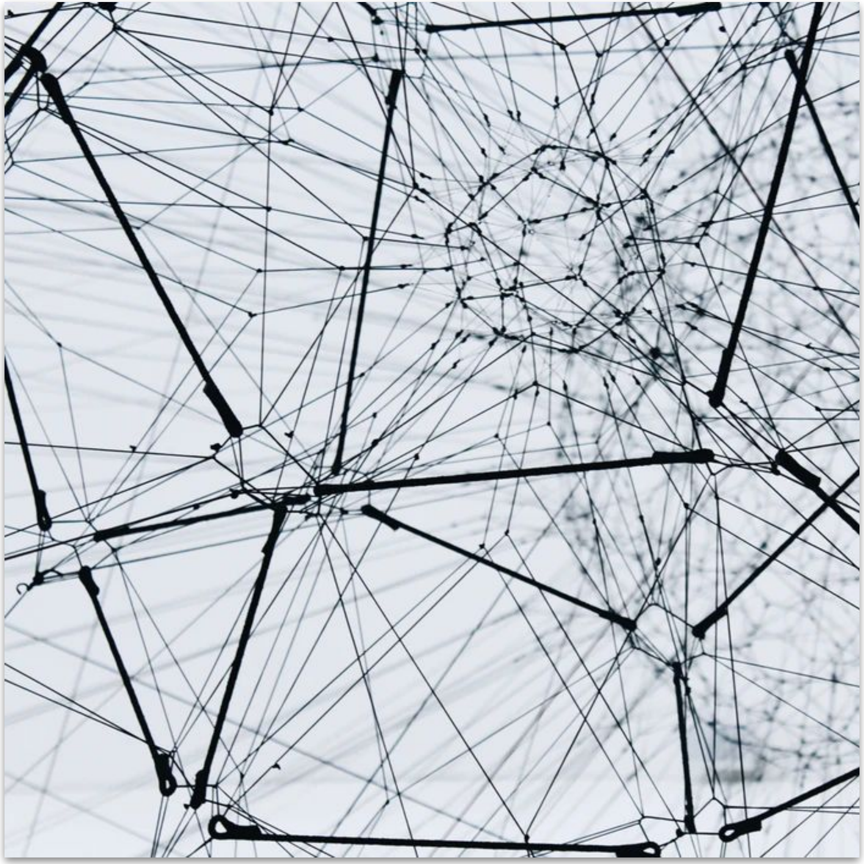
STEP 3

Model Selection

STEP 4

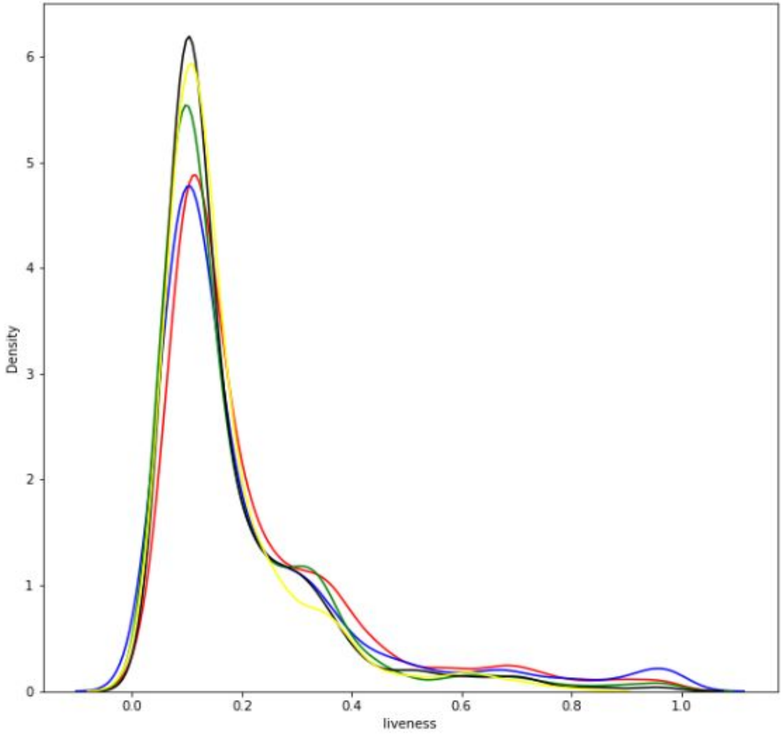
Model Flow

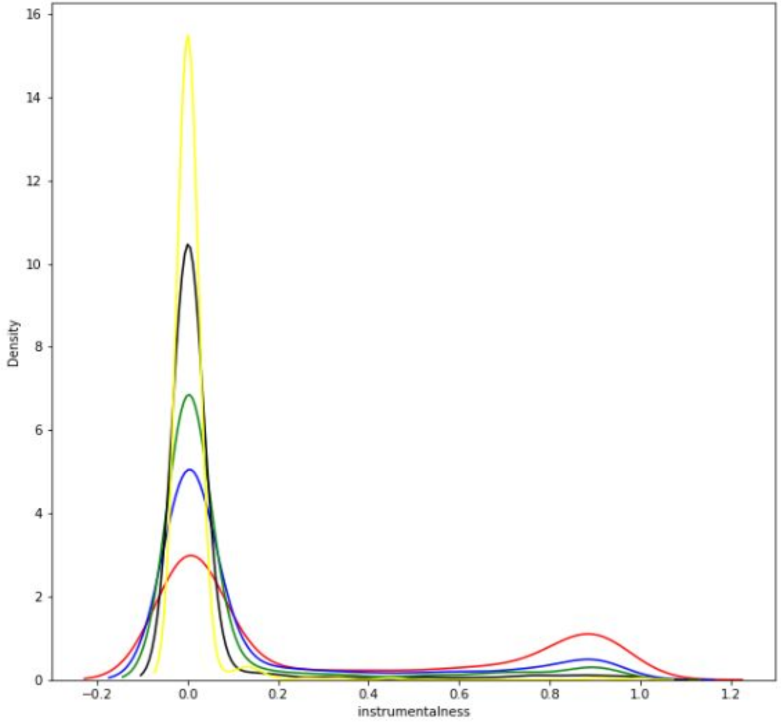
04



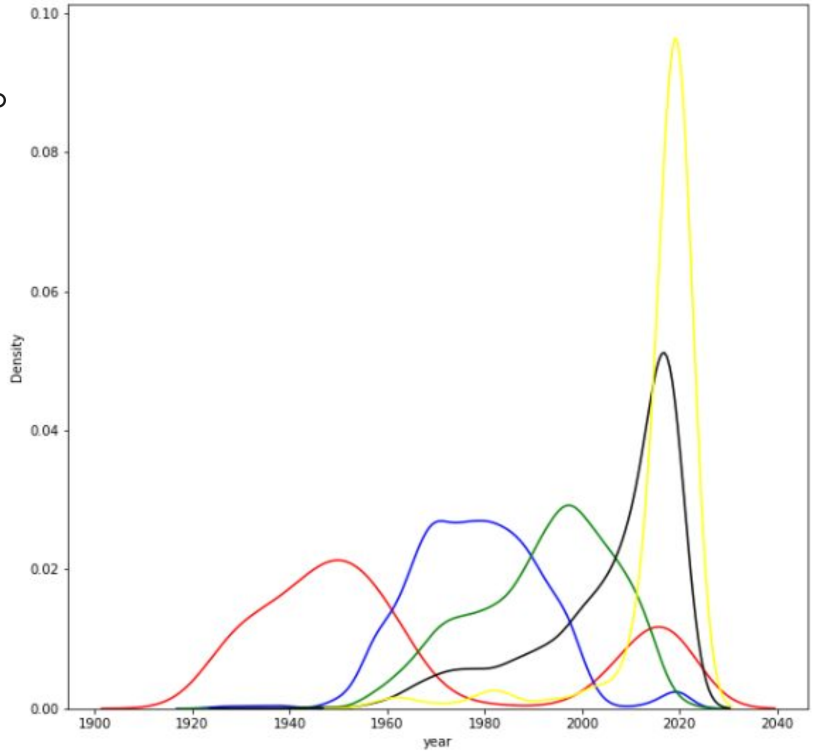
EXPLORATORY DATA ANALYSIS

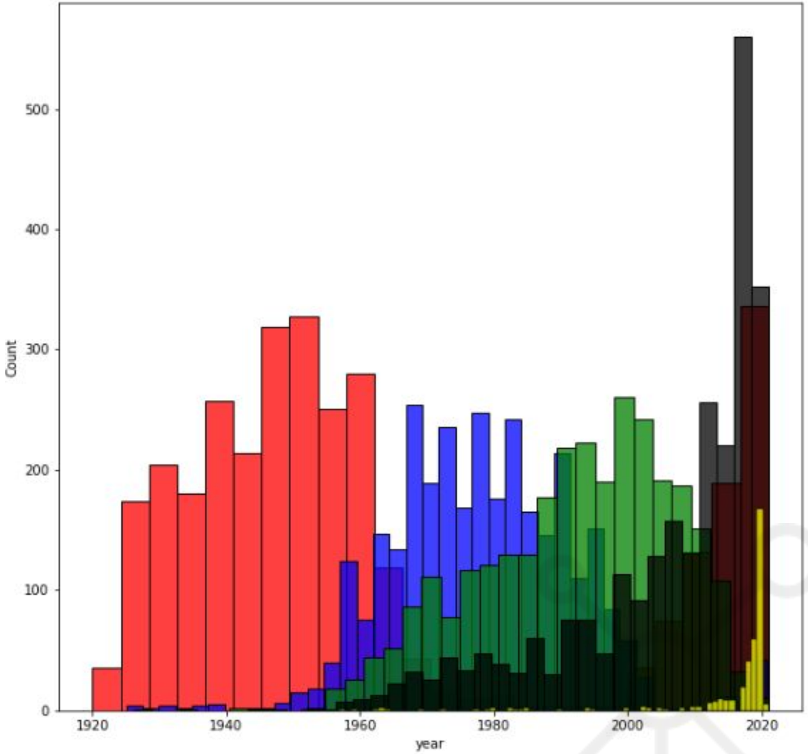
EXPLORATORY DATA ANALYSIS - 1



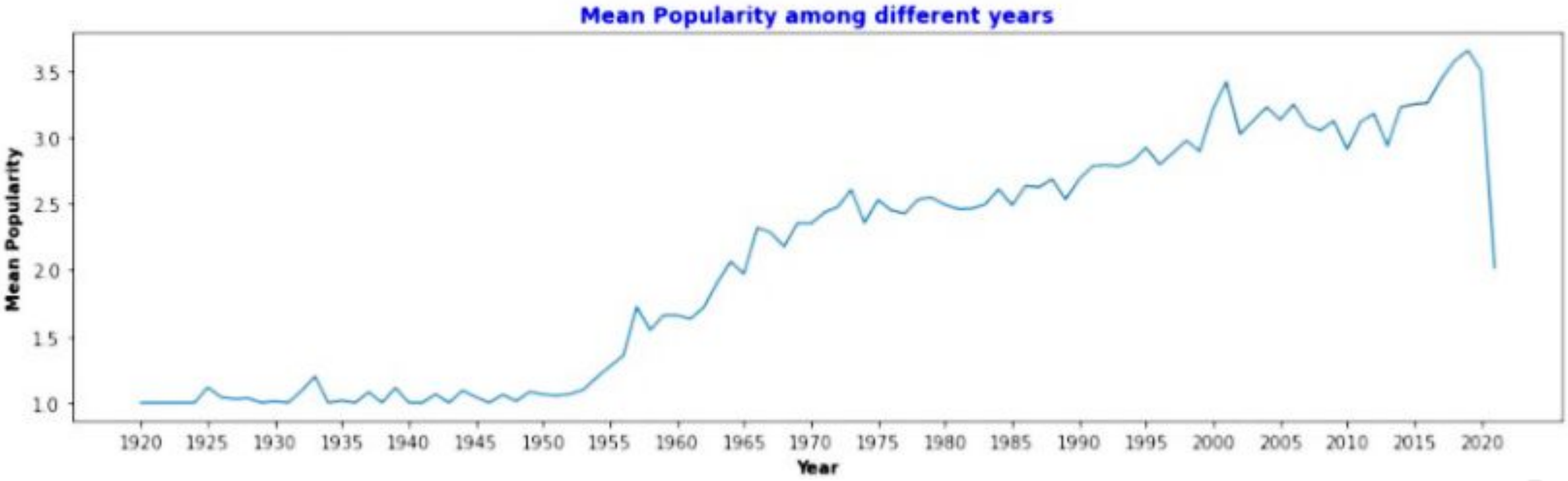


EXPLORATORY DATA ANALYSIS - 2

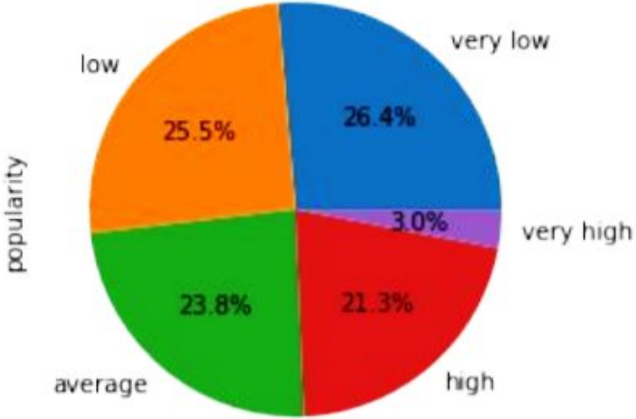




EXPLORATORY DATA ANALYSIS - 3



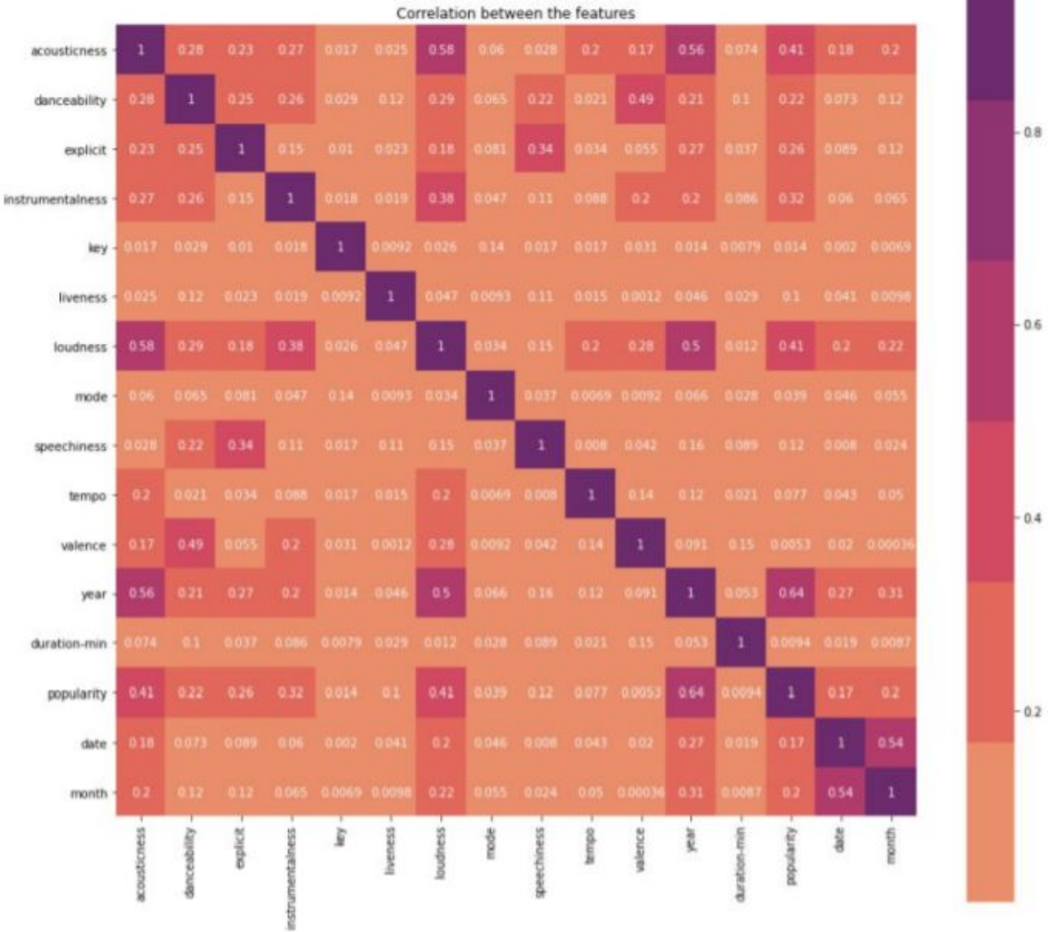
EXPLORATORY DATA ANALYSIS - 4

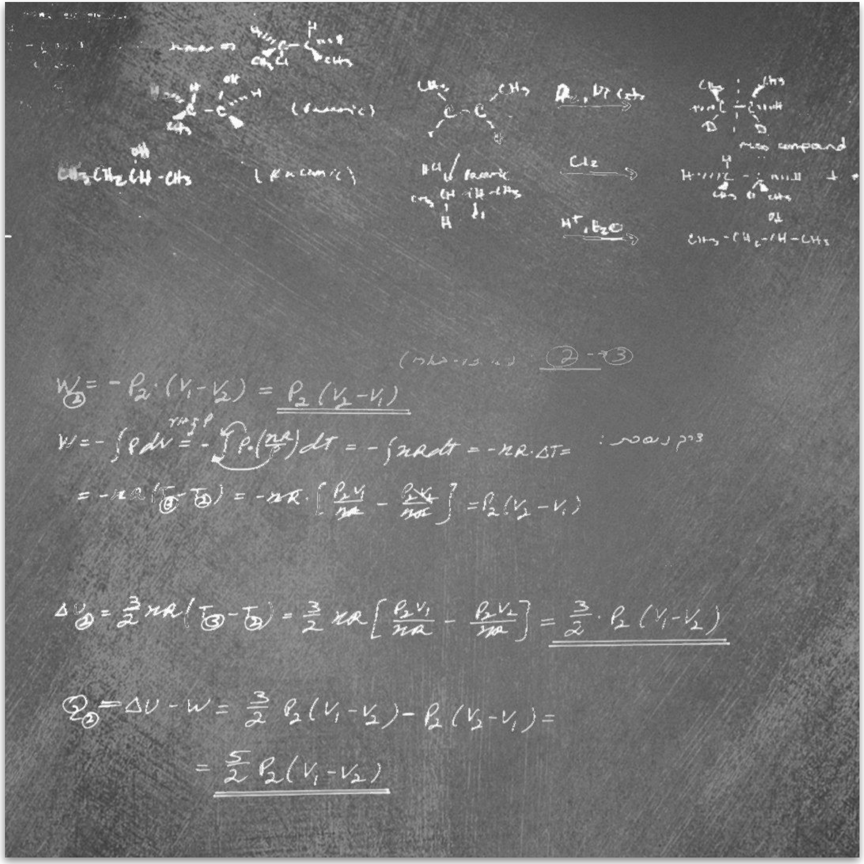


- As you can see from the graph on the right, the data points of each category are mostly equally distributed except in the case of very high category.
- This was reflecting while building our model - we saw in the confusion matrix that category 5 has very low accuracy when compared to other categories.

» So we used sampling techniques like SMOTE and Randomized Under Sampling to get a more balanced dataset.

CORRELATION BETWEEN FEATURES





FEATURE ENGINEERING

DATE AND MONTH

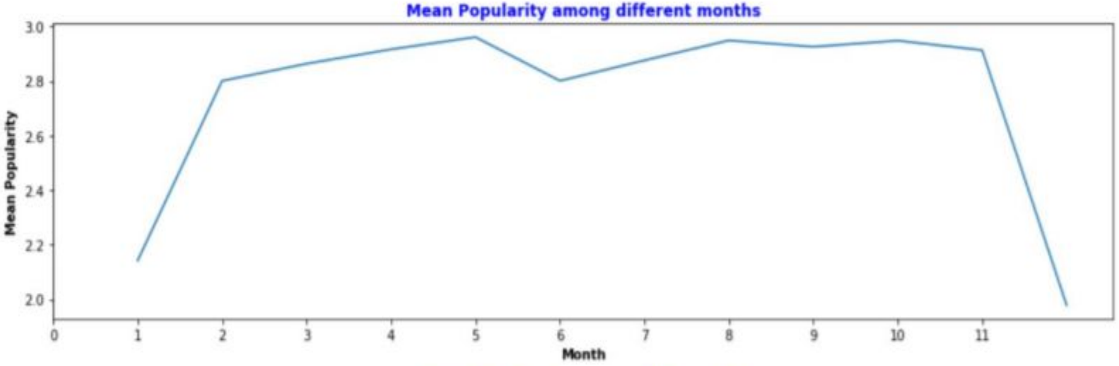
release_date
01-01-2014
01-04-1972
02-06-1998
08-09-1980

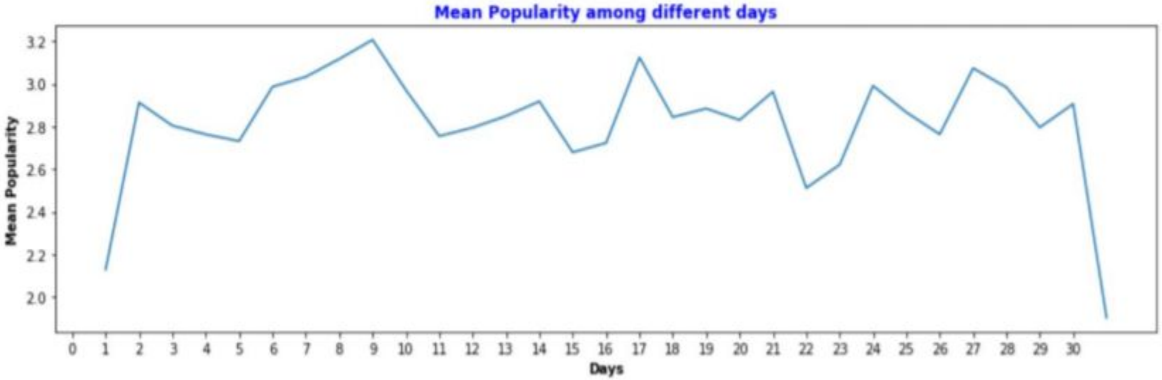
→

month	date
1	1
1	4
2	6
8	9

- From this we also created a new column ‘days’ which had the number of days passed from the starting of 1920. But latter it was found to have high correlation with the year column. So it was removed from the dataset.

Month_curse & days_curse





DECADE OF THE SONG

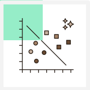
```
year_new
         40s
         10s
         40s
         70s
         70s
```

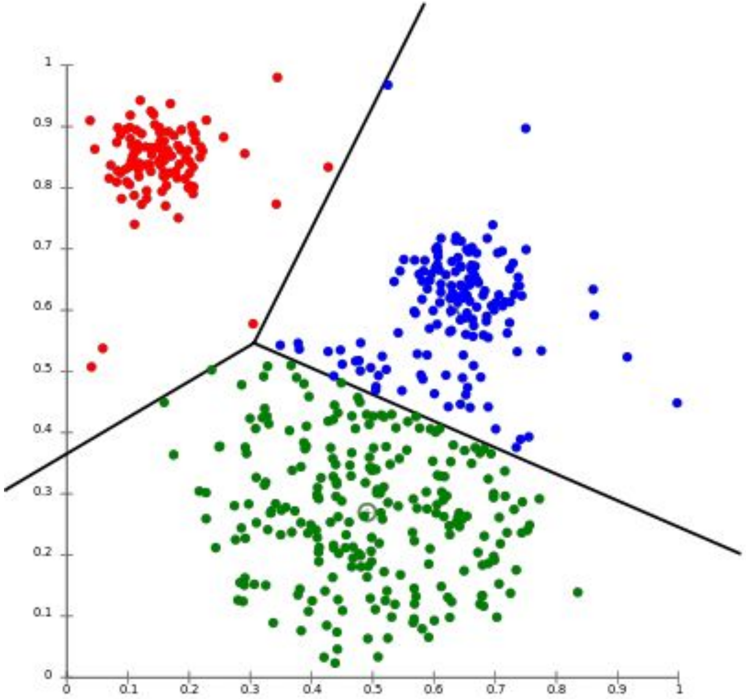
- A new categorical column was also created to keep an eye of the decade in which the song is published.
- During the model evaluation, we found out that this was creating the same impact as the year column. So we removed the year column and used this for our evaluation.

TEMPO CATEGORY

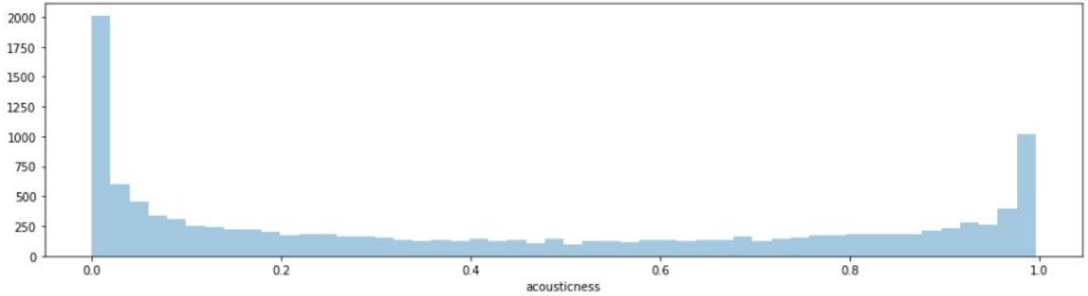
Larghissimo	very, very slow, almost droning (20 BPM and below)
Grave	slow and solemn (20-40 BPM)
Lento	slowly (40-60 BPM)
Largo	the most commonly indicated “slow” tempo (40-60 BPM)
Larghetto	rather broadly, and still quite slow (60-66 BPM)
Adagio	another popular slow tempo, which translates to mean “at ease” (66-76 BPM)
Adagietto	rather slow (70-80 BPM)
Adante moderato	A bit slower than andante
Andante	A popular tempo that translates as “at a walking pace” (76-108 BPM)

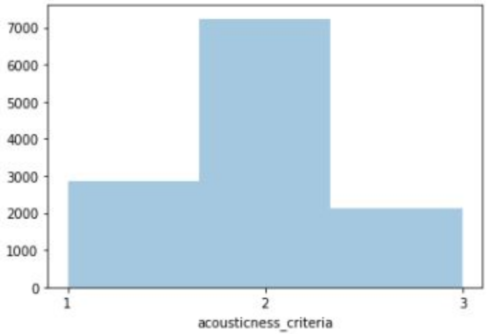
K - MEANS CLUSTERING





ACOUSTICNESS CRITERIA





Similarly we did for instrumentalness column too.

FEATURE SCALING

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Most of the columns in the dataset were lying in the range of 0 to 1, so there was no need to scale them. But there are some columns with high values - like duration-min, tempo, loudness etc.. These columns are scaled using Min-Max Scaler to make sure the values between 0 and 1.

FEATURES USED

CATEGORICAL FEATURES

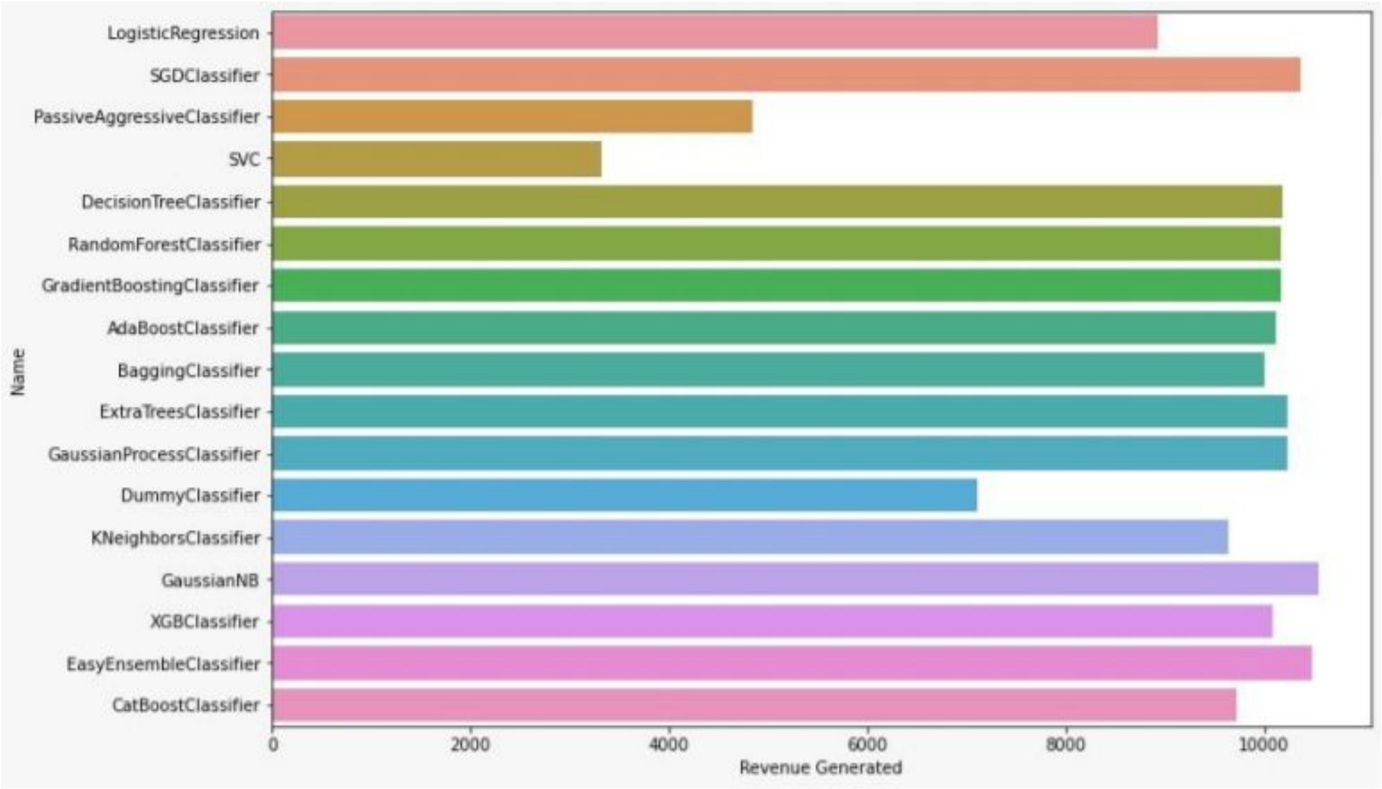
- Explicit
- Mode
- Year Class
- Month Curse
- Date Curse
- Tempo Category
- Month
- Date
- Acoust criteria
- Instru criteria
- K-MEANS cluster

NUMERICAL FEATURES

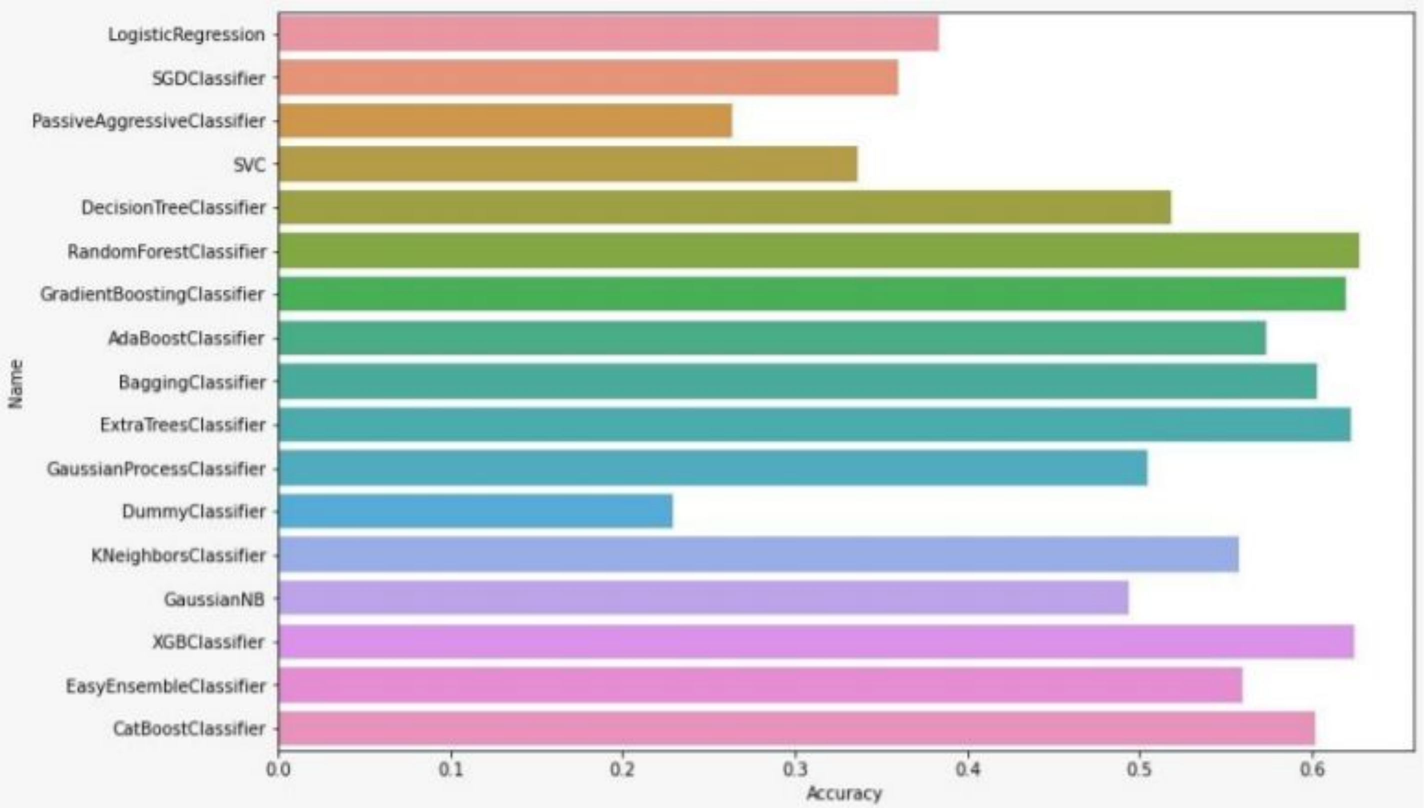
- Acousticness
- Danceability
- Energy
- Instrumentalness
- Key
- Liveness
- Loudness
- Speechiness
- Tempo
- Valence
- Duration-min
- Year
- Release date
- Days

MODEL SELECTION

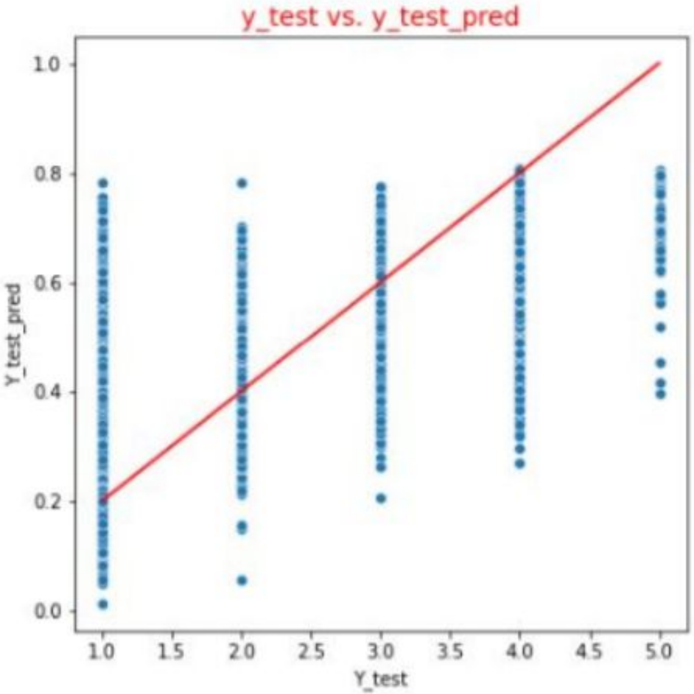
MODEL SELECTION

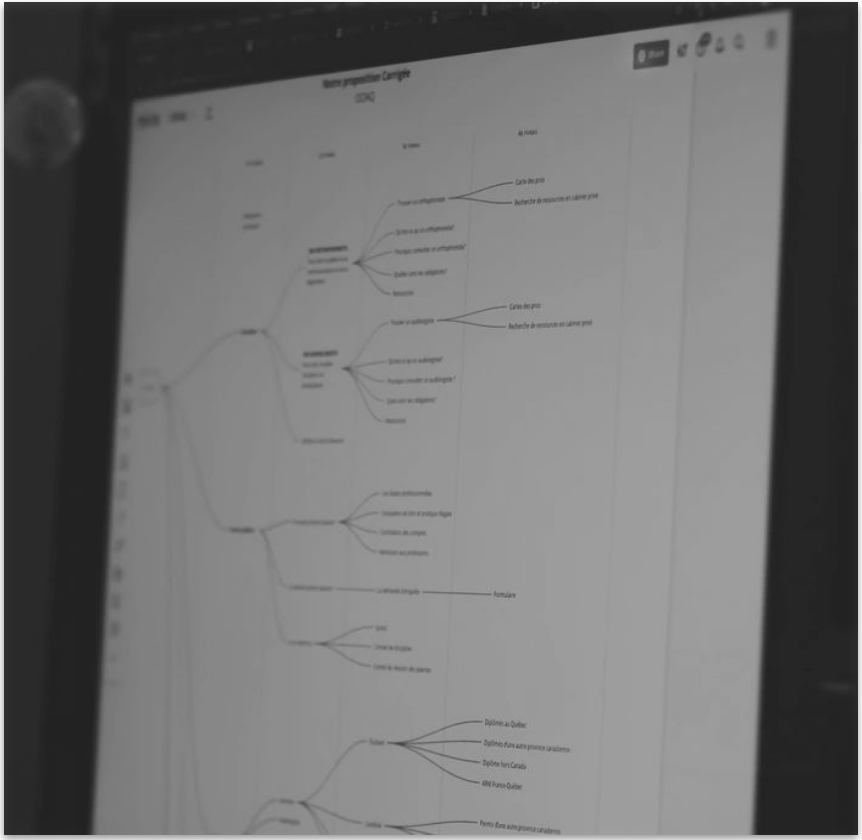


MODEL SELECTION



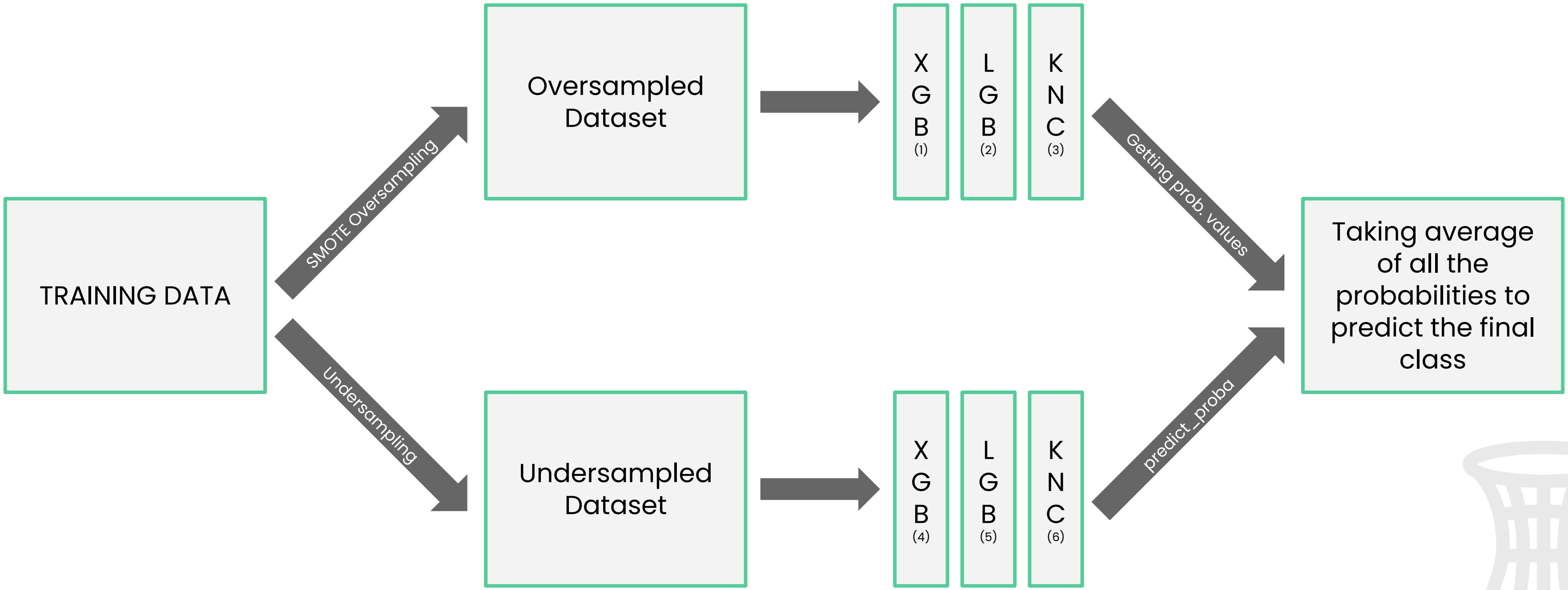
TRYING IT AS A REGRESSION PROBLEM



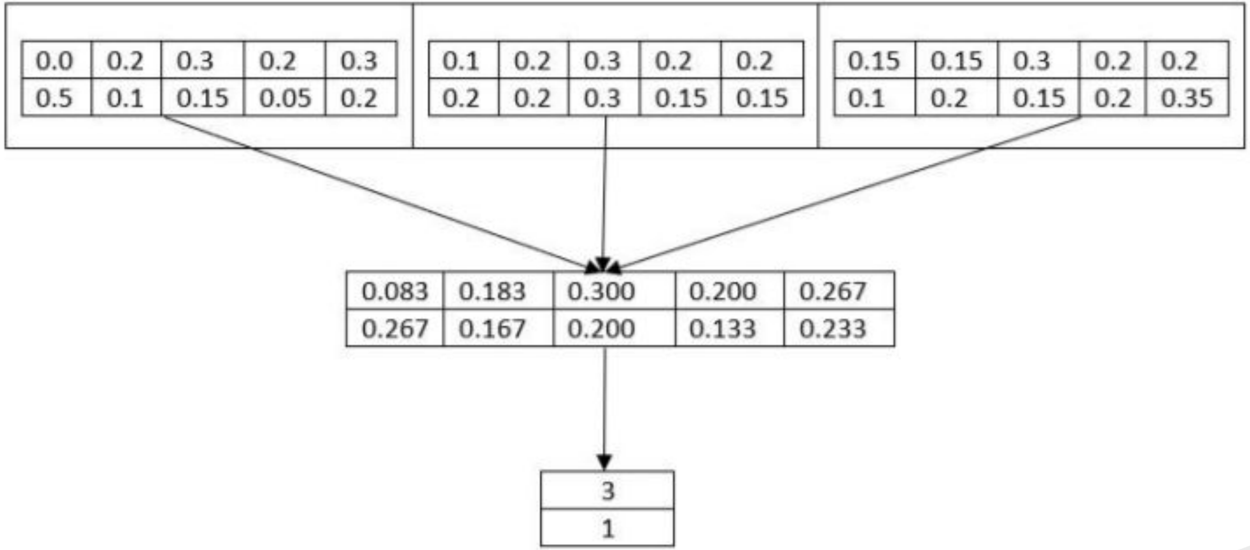


MODEL FLOW

MODEL FLOW


TRAINING DATA
SMOTE Oversampling
Undersampling
Oversampled Dataset
Undersampled Dataset
XGB (1)
LGB (2)
KNC (3)
XGB (4)
LGB (5)
KNC (6)
Getting prob. values
predict_proba
Taking average of all the probabilities to predict the final class

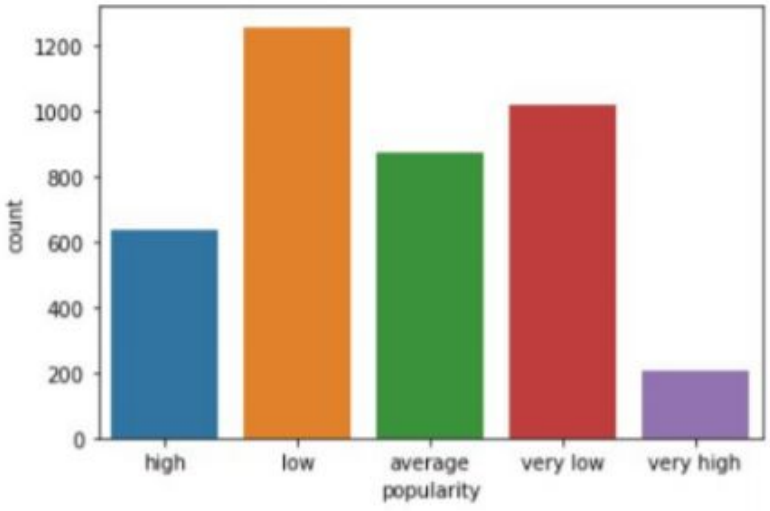
FINAL PREDICTIONS



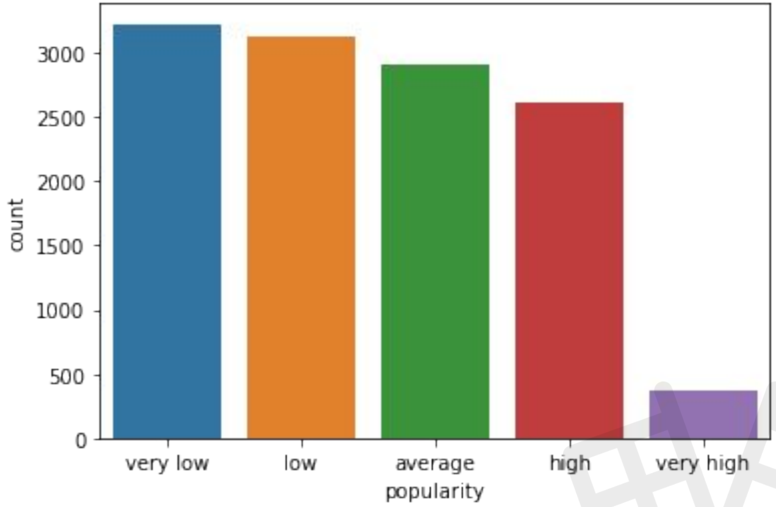
- Final predictions = np.argmax((prob1+prob2+prob3+prob4)/4.0))+1

DISTRIBUTION OF FINAL PREDICTIONS

TEST DATA DISTRIBUTION



TRAINING DATA DISTRIBUTION



Let’s Dive In Deeper!
Have any questions?

Recap.

- → Exploratory Analysis
- → Feature Engineering
- → Model Selection
- → Model Flow

THANK YOU!