TEAM 35

Understanding the Data & Problem Statement

12227 songs → Each song has several parameters like **acousticness**,**duration**,**year** etc → Each Song ranked on a scale of 1 to 5 according to popularity → Predict the popularity of each song

INSIGHTS

Release-Date and **Year column** are highly correlated

Imbalance in Dataset
The very high popularity constituted 3% of the total dataset

Random oversampling was used to treat the imbalance in dataset.This method randomly duplicate examples in the minority class.

- **Exploratory Data Analysis**
- Feature Selection
- Anomaly Detection
- Classification Models
- Comparative Analysis
- Conclusion

Understanding the Data & Problem Statement

DATA SUMMARY

Feature/Stat	Mean	Std	Min	25%	50%	75%	Max
Acousticness	0.430578	0.366893	0.000001	0.05895	0.354000	0.80500	0.996
Danceability	0.556353	0.175373	0.000000	0.43800	0.569000	0.68500	0.980
Energy	0.522129	0.262482	0.000020	0.30300	0.534000	0.73900	1.000
Instrumentalness	0.149321	0.297954	0.000000	0.00000	0.000115	0.05565	1.000
Key	5.205202	3.526954	0.000000	2.00000	5.000000	8.00000	11.000
Liveness	0.201365	0.173987	0.014700	0.09620	0.132000	0.25200	0.997
Loudness	-10.6686	5.506888	-43.738	-13.6560	-9.5840	-6.57150	1.006
Speechiness	0.097680	0.155895	0.000000	0.03470	0.045600	0.07890	0.968
Tempo	118.1674	30.200	0.000000	95.05050	116.9150	136.1085	216.843
Valence	0.525300	0.258205	0.000000	0.32100	0.532000	0.73700	1.000
Year	1984.517	25.9119	1920.00	1966.00	1987.00	2008.00	2021.000
DurationMin	3.888133	2.383133	0.200000	2.90000	3.600000	4.40000	2021.000

Exploratory Data Analysis

Feature Selection

Anomaly Detection

Classification Models

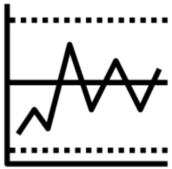
Comparative Analysis

Conclusion

Data Visualization and insights

CORRELATION MATRIX

Loudness and **Energy** of a song are **highly positively correlated.**



Acousticness is **highly negatively correlated** with the **energy** and **loudness** of the song

	acousticness	danceability	energy	explicit	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	valence	duration-min	release_day	release_month	release_year	popularity
acousticness	1	-0.28	-0.75	-0.23	0.27	-0.017	-0.025	-0.58	0.06	-0.028	-0.2	-0.17	-0.074	-0.2	-0.18	-0.56	-0.41
danceability	-0.28	1	0.21	0.25	-0.26	0.029	-0.12	0.29	-0.065	0.22	-0.021	0.49	-0.1	0.12	0.073	0.21	0.22
energy	-0.75	0.21	1	0.13	-0.22	0.033	0.13	0.78	-0.058	-0.078	0.23	0.33	0.049	0.2	0.17	0.49	0.33
explicit	-0.23	0.25	0.13	1	-0.15	0.01	0.023	0.18	-0.081	0.34	0.034	-0.055	-0.037	0.12	0.089	0.27	0.26
instrumentalness	0.27	-0.26	-0.22	-0.15	1	-0.018	-0.019	-0.38	-0.047	-0.11	-0.088	-0.2	0.086	-0.065	-0.06	-0.2	-0.32
key	-0.017	0.029	0.033	0.01	-0.018	1	-0.0092	0.026	-0.14	0.017	0.017	0.031	-0.0079	0.0069	-0.002	0.014	0.014
liveness	-0.025	-0.12	0.13	0.023	-0.019	-0.0092	1	0.047	0.0093	0.11	0.015	0.0012	0.029	0.0098	0.041	-0.046	-0.1
loudness	-0.58	0.29	0.78	0.18	-0.38	0.026	0.047	1	-0.034	-0.15	0.2	0.28	0.012	0.22	0.2	0.5	0.41
mode	0.06	-0.065	-0.058	-0.081	-0.047	-0.14	0.0093	-0.034	1	-0.037	0.0069	0.0092	-0.028	-0.055	-0.046	-0.066	-0.039
speechiness	-0.028	0.22	-0.078	0.34	-0.11	0.017	0.11	-0.15	-0.037	1	-0.008	0.042	-0.089	0.024	0.008	-0.16	-0.12
tempo	-0.2	-0.021	0.23	0.034	-0.088	0.017	0.015	0.2	0.0069	-0.008	1	0.14	-0.021	0.05	0.043	0.12	0.077
valence	-0.17	0.49	0.33	-0.055	-0.2	0.031	0.0012	0.28	0.0092	0.042	0.14	1	-0.15	-0.00036	-0.02	-0.091	-0.0053
duration-min	-0.074	-0.1	0.049	-0.037	0.086	-0.0079	0.029	0.012	-0.028	-0.089	-0.021	-0.15	1	-0.0087	-0.019	0.053	-0.0094
release_day	-0.2	0.12	0.2	0.12	-0.065	0.0069	0.0098	0.22	-0.055	0.024	0.05	-0.00036	-0.0087	1	0.54	0.31	0.2
release_month	-0.18	0.073	0.17	0.089	-0.06	-0.002	0.041	0.2	-0.046	0.008	0.043	-0.02	-0.019	0.54	1	0.27	0.17
release_year	-0.56	0.21	0.49	0.27	-0.2	0.014	-0.046	0.5	-0.066	-0.16	0.12	-0.091	0.053	0.31	0.27	1	0.64
popularity	-0.41	0.22	0.33	0.26	-0.32	0.014	-0.1	0.41	-0.039	-0.12	0.077	-0.0053	-0.0094	0.2	0.17	0.64	1

Exploratory Data Analysis

Feature Selection

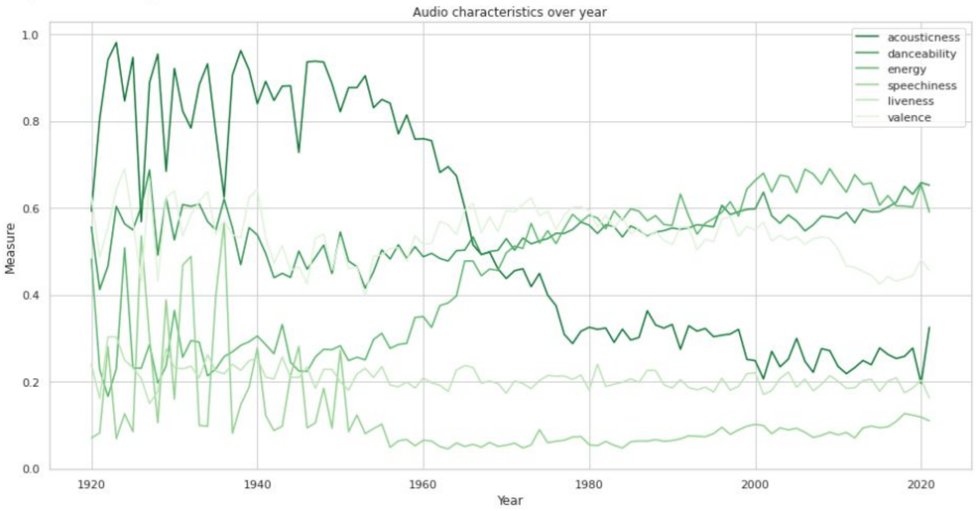
Anomaly Detection

Classification Models

Comparative Analysis

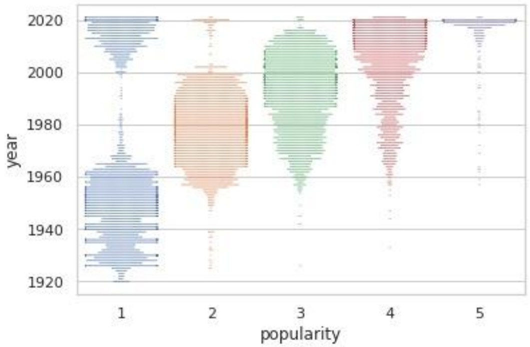
Conclusion

Data Visualization and insights



Acousticness in songs **decreased** over the years.

Energy increased over the years. Similarly, the **loudness** and the **tempo** of the songs **increased** over the years



We can see the majority of the songs having very **high popularity** are from **2015 onwards**

Feature Selection

Variance threshold

Any feature having variance less than a threshold is removed from the dataset

SelectFromModel

Selects a given number of features based on the importance weights

SelectKBest

Select features according to the k highest scores where the scoring function was ANOVA

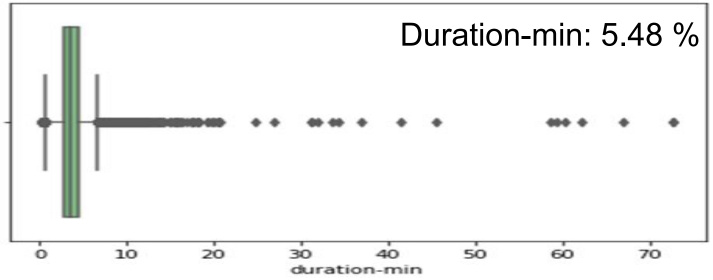
Greedy Feature Selection

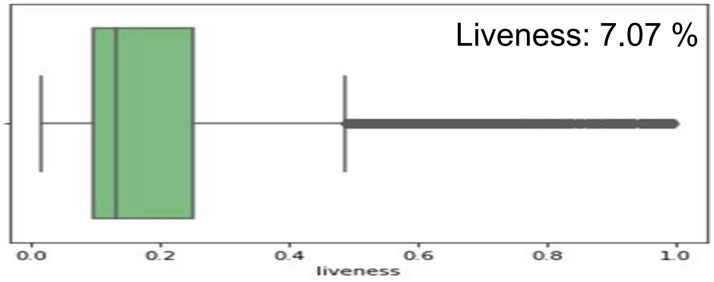
Selects features greedily one by one on the basis of which feature the evaluation metric increases the most.

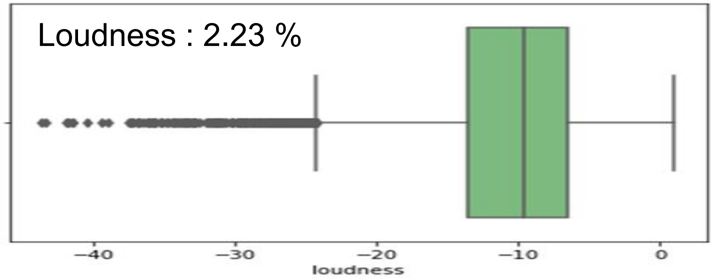
Top Features

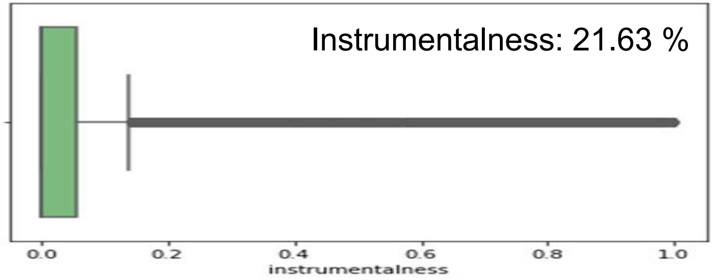
- Year
- Danceability
- Instrumentalness
- Duration-min
- Valence
- Acousticness
- Liveness

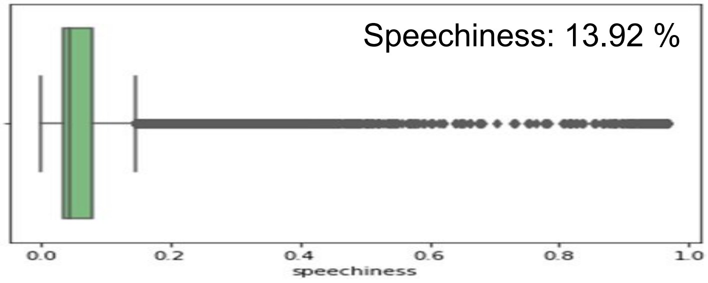
Anomaly Detection











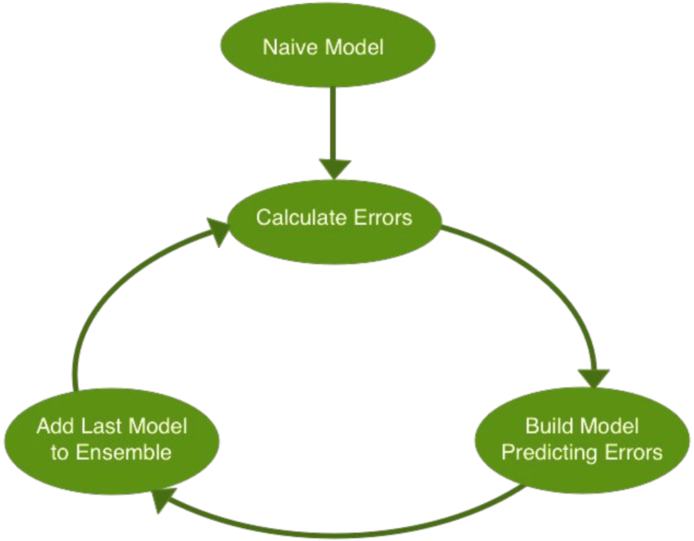
The outliers constituted **50%** of the dataset so removing them wasn't an option.

Did **not treat** the outliers as decision tree models were trained and outliers wouldn't affect the model much.

XGBoost (eXtreme Gradient Boosting)

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

It implements machine learning algorithms under the Gradient Boosting framework.

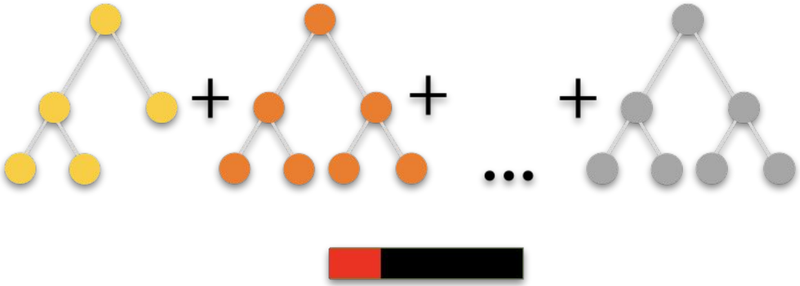


WHY XGBOOST?

RESULTS	
Test Accuracy	71.2
Training Accuracy	99.95
Bidding Value	7540
Revenue Collected	12488
F1-score	0.688

CatBoost

CatBoost grows oblivious trees, which means that the trees are grown by imposing the rule that all nodes at the same level, test the same predictor with the same condition, and hence an index of a leaf can be calculated with bitwise operations



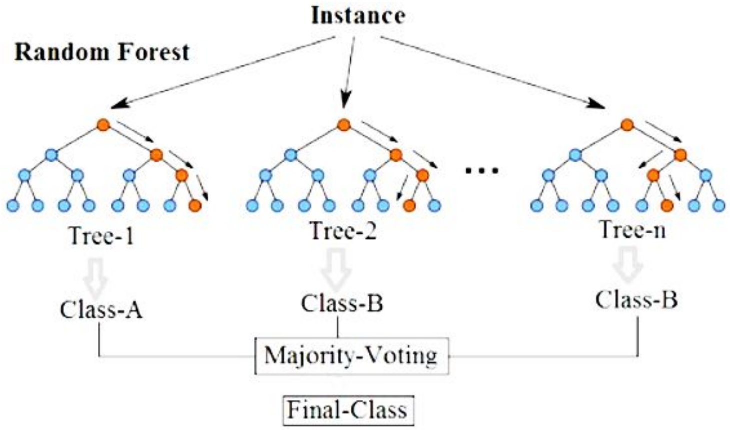
Why Catboost?

RESULTS

Test Accuracy	68.00
Training Accuracy	77.05
Bidding Value	7539
Revenue Collected	13972
F1-score	0.656

Random Forest

Random Forest operates as an ensemble of a large number of decision trees.
In this algorithm, all the trees spit out the prediction and the class with the most number of votes becomes the model's prediction.

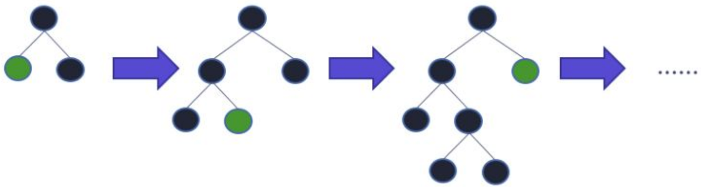


Why Random Forest?

RESULTS	
Test Accuracy	71.35
Training Accuracy	99.95
Bidding Value	7538
Revenue Collected	14086
F1-score	0.704

LightGBM

LightGBM is called "Light" because of its computation power and giving results faster.
Light GBM grows tree vertically i.e leaf-wise. It takes less memory to run and is able to deal with large amounts of data

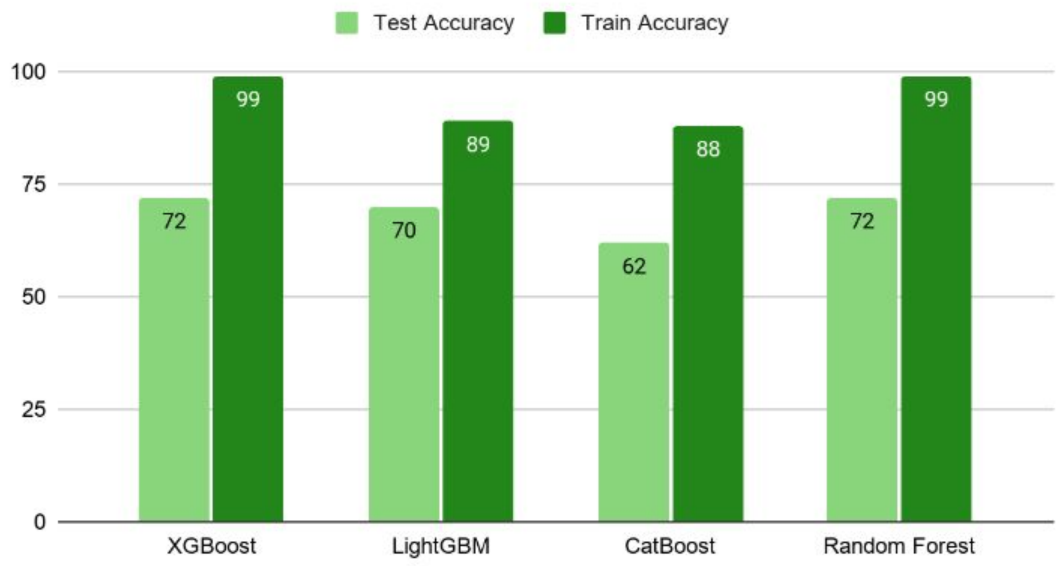


Why LightGBM?

RESULTS

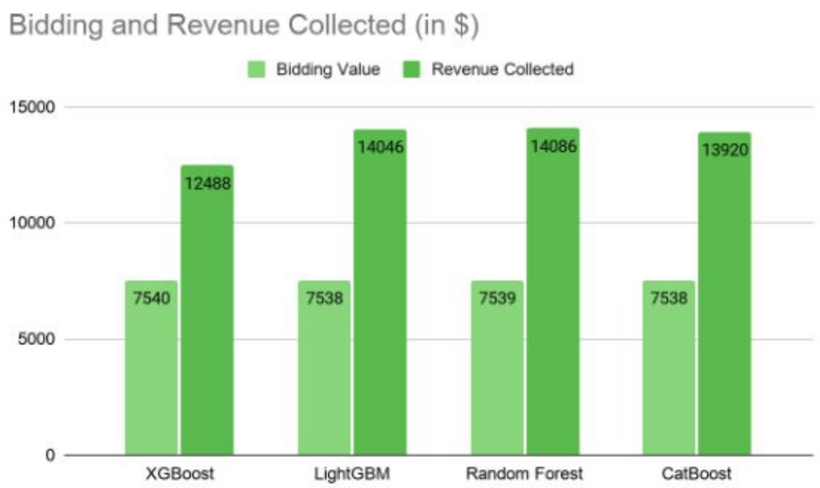
Test Accuracy	70.05
Training Accuracy	89.95
Bidding Value	7538
Revenue Collected	14046
F1-score	0.686

Comparative Analysis



Since both XGBoost and Random Forest models are **overfitting**, and the score of CatBoost is less, we are going to choose **LightGBM** as our final model.

Conclusion



Class	Precision	Recall	F1-score
Very Low	0.56	0.68	0.62
Low	0.84	0.81	0.82
Average	0.47	0.46	0.47
High	0.67	0.47	0.55
Very High	0.90	0.98	0.97

The bidding total of the model on our validation set of size 3016 rows is **$7538**, and the revenue collected is **$14046**.

Our model has good predictions for **Very High and Low** popularity, which is visible in our classification report of the LightGBM model

THANK YOU

XGBoost

Class	Precision	Recall	F1-score
Very Low	0.85	0.78	0.81
Low	0.56	0.67	0.61
Average	0.45	0.44	0.44
High	0.68	0.56	0.61
Very High	0.95	1.00	0.97

LightGBM

Class	Precision	Recall	F1-score
Very Low	0.56	0.68	0.62
Low	0.84	0.81	0.82
Average	0.47	0.46	0.47
High	0.67	0.47	0.55
Very High	0.90	0.98	0.97

CatBoost

Class	Precision	Recall	F1-score
Very Low	0.84	0.81	0.82
Low	0.56	0.66	0.60
Average	0.46	0.47	0.46
High	0.63	0.40	0.49
Very High	0.85	0.98	0.91

Random Forest

Class	Precision	Recall	F1-score
Very Low	0.86	0.79	0.83
Low	0.57	0.70	0.63
Average	0.47	0.48	0.48
High	0.70	0.54	0.61
Very High	0.95	1.00	0.97

Exploratory Data Analysis

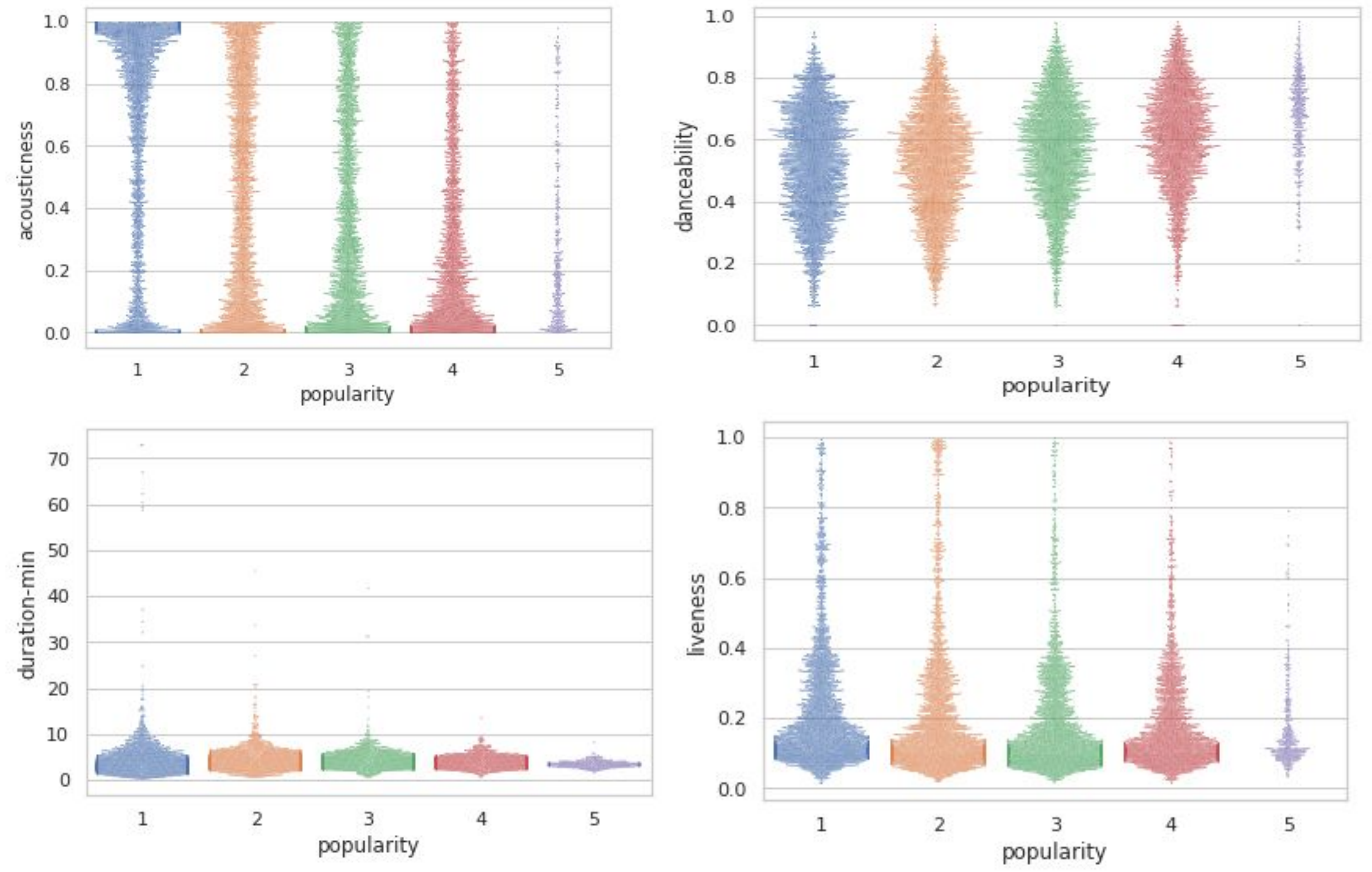
Feature Engineering

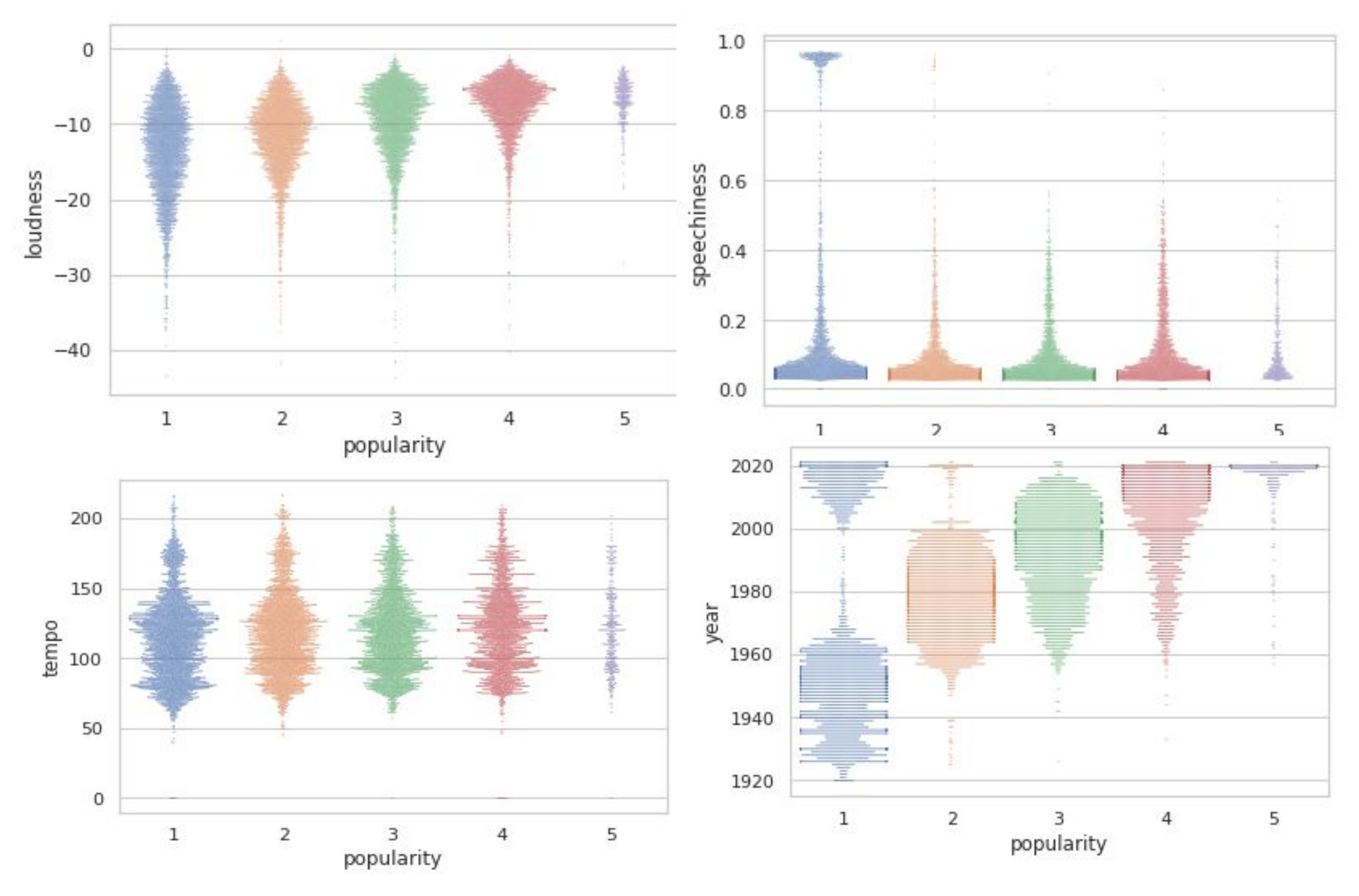
Anomaly Detection

Classification Models

Comparative Analysis

Conclusion




0
−10
−20
−30
−40
loudness
1
2
3
4
5
popularity
1.0
0.8
0.6
0.4
0.2
0.0
speechiness
200
150
100
50
0
tempo
popularity
2020
2000
1980
1960
1940
1920
year
popularity