

# Open IIT Data Analytics Report

## **Team 11**

---

## Content:

Serial Number	Heading	Page Number
1.	Overview	1-2
2.	Exploratory Data Analysis	2-4
3.	Data Summary	5
4.	Correlation Matrices	6
5.	Multi - Classification Models	6-12
6.	Conclusion	13-14

## Overview

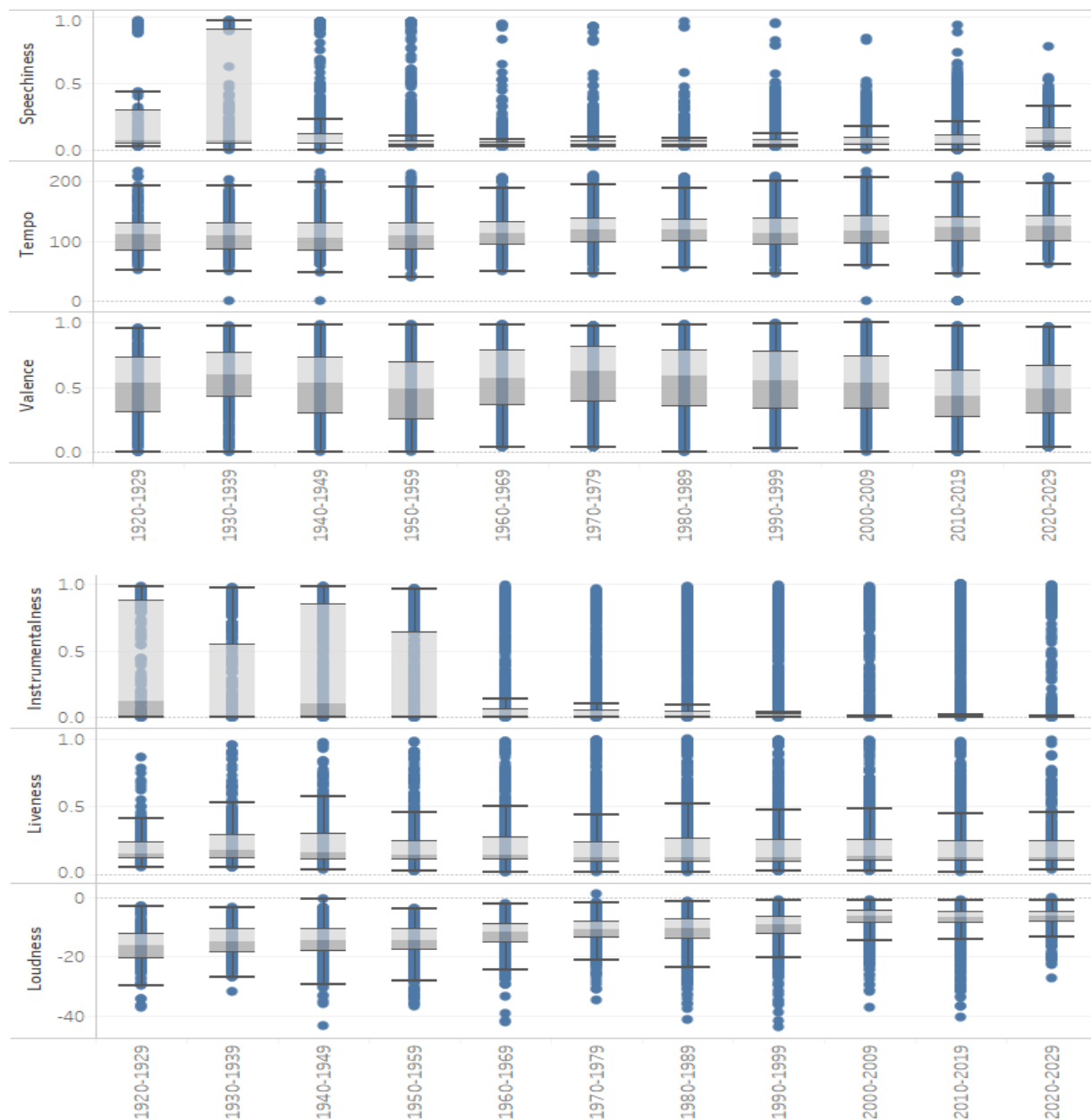
Music has played an important role in civilization and has a great value to people. Within cultures worldwide, music is tightly related to the fabric of life and plays an irreplaceable role in providing entertainment, enhancing mental peace, and communicating emotions. The music industry is well developed and has a revenue of more than 20 Billion USD every year.

Over the past few decades, the enhancement of technology has also revolutionized the music industry. The coming of electronic music and increasing dependence on audio mixing and various other technological influences have changed the music industry as we know it. From beat selection to final equalizing, all audios have various technological aspects to them and the rising innovation in the field of audio processing has led music companies to analyze the music on various parameters, which may influence the success of the music record.

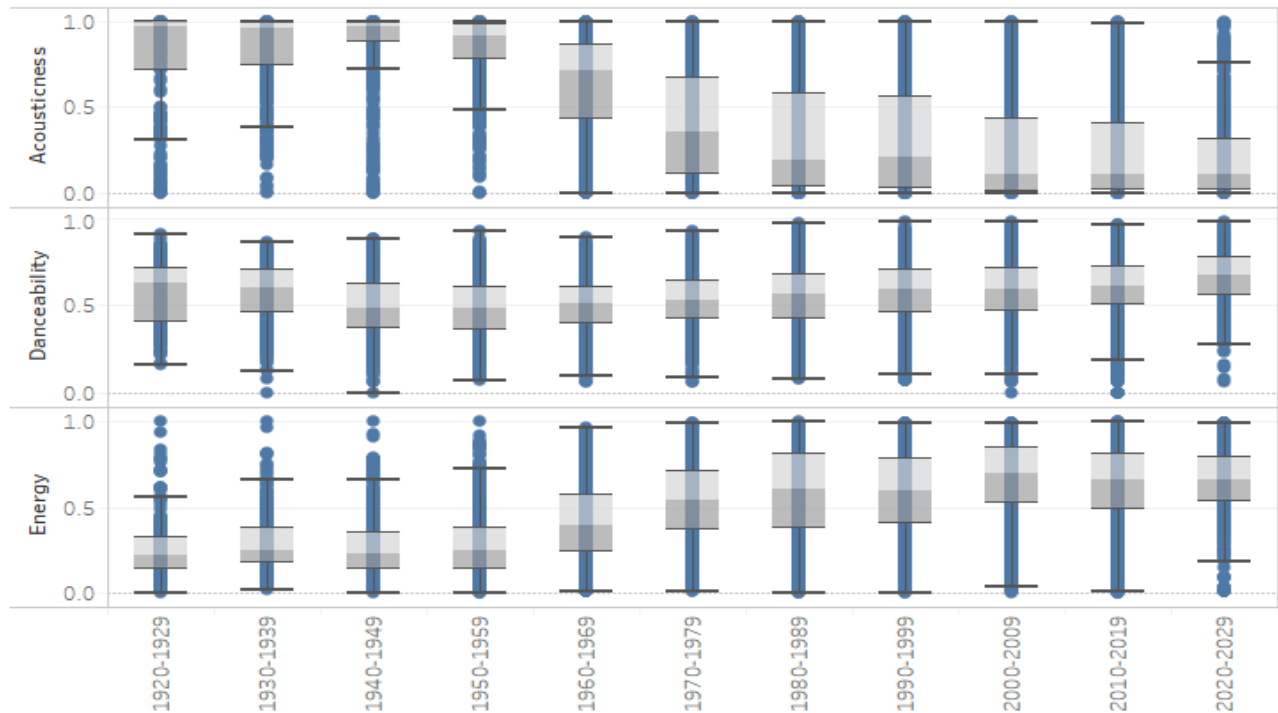
We aim to develop efficient classifier models for the record label and help them to better forecast the success of the tracks they aim to purchase. We have used data visualization tools to generate insights and developed intelligent features to support our classification models. We have developed various statistical, machine learning, and deep learning-based models, and evaluated

them on the validation set for model selection. Finally, we stacked results from the two best-performing models to produce robust classifications.

## Exploratory Data Analysis



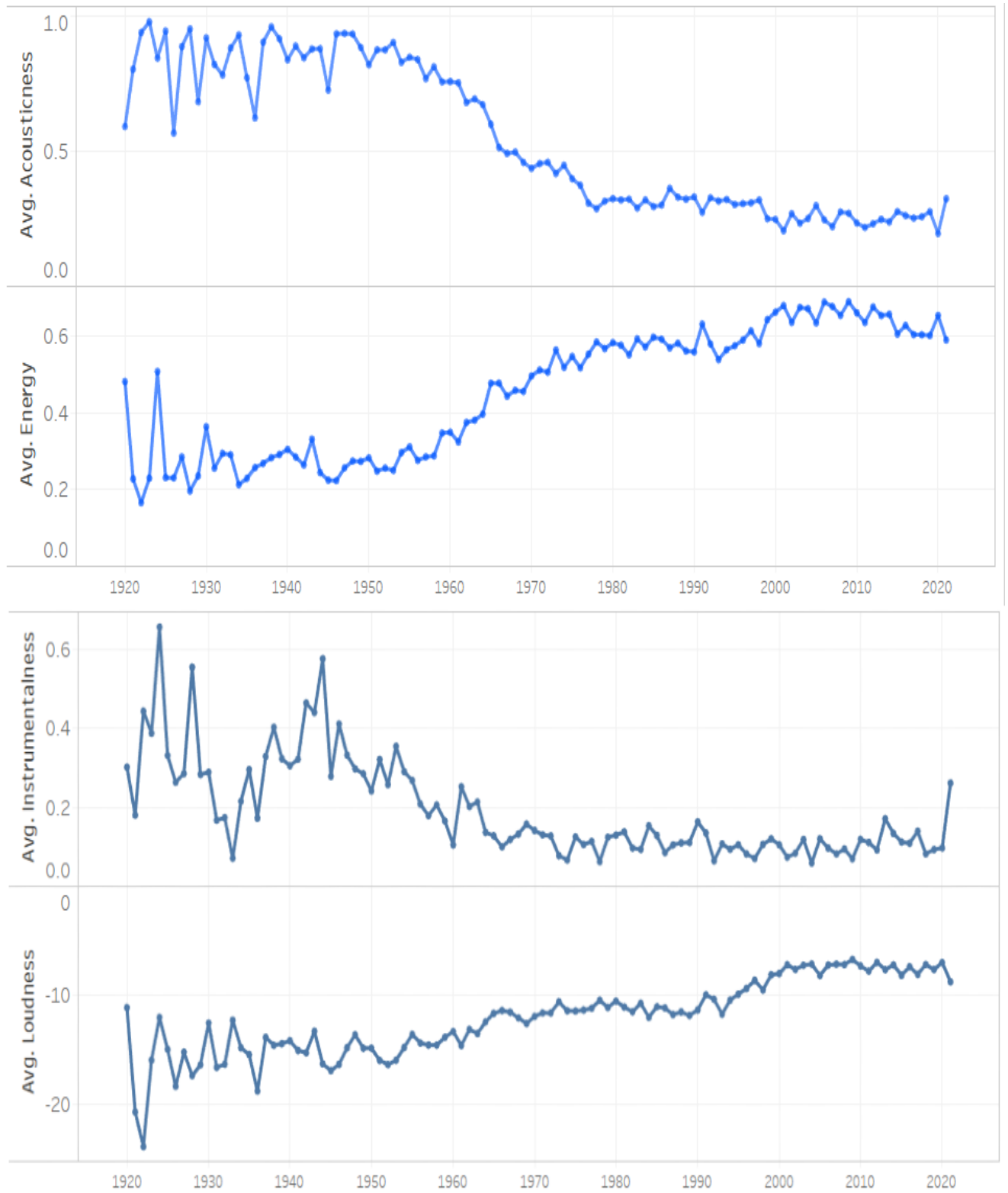
From the above plot it is quite evident that Loudness has increased with each decade, with a sharp decrease in Instrumentalness.



It can be observed that the acoustic quality of music tracks decreases and energy of music tracks increases with each passing decade.

## General Trends in Time Analysis:

Upon plotting the average values of Acousticness, Energy, Instrumentalness and Loudness for music tracks released each year, we can observe some general trends. We can see that values of acousticness and instrumentalness decrease while values of energy and loudness increase on an average over the course of 100 years.



## Data Summary

### Summary Statistics:

The data provided consists of different features important in determining the popularity of a music track from the year 1920 to the year 2021. The tables present give an aggregated summary statistics for the given training and testing dataset for all the different features, thus providing us an insight to how the training and test sets vary from each other. As we can observe from the mean and standard deviation, most of the tracks belong to the later years

Column1	id	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	speechiness	tempo	valence	year	duration-min
count	12227	12227	12227	12227	12227	12227	12227	12227	12227	12227	12227	12227	12227
mean	8094.03435	0.43057836	0.556352654	0.522128712	0.149320559	5.205201603	0.201364562	-10.66868651	0.097679807	118.1674949	0.525300073	1984.517298	3.888132821
std	4690.929822	0.366892892	0.175372545	0.262482291	0.297954314	3.526953879	0.173987492	5.506888135	0.155894608	30.20006382	0.258204698	25.91199777	2.383133109
min	1	1.04E-06	0	2.03E-05	0	0	0.0147	-43.738	0	0	0	1920	0.2
25%	4026	0.05895	0.438	0.303	0	2	0.0962	-13.656	0.0347	95.0505	0.321	1966	2.9
50%	8093	0.354	0.569	0.534	0.000115	5	0.132	-9.584	0.0456	116.915	0.532	1987	3.6
75%	12180	0.805	0.685	0.739	0.05565	8	0.252	-6.5715	0.0789	136.1085	0.737	2008	4.4
max	16227	0.996	0.98	1	1	11	0.997	1.006	0.968	216.843	1	2021	72.8

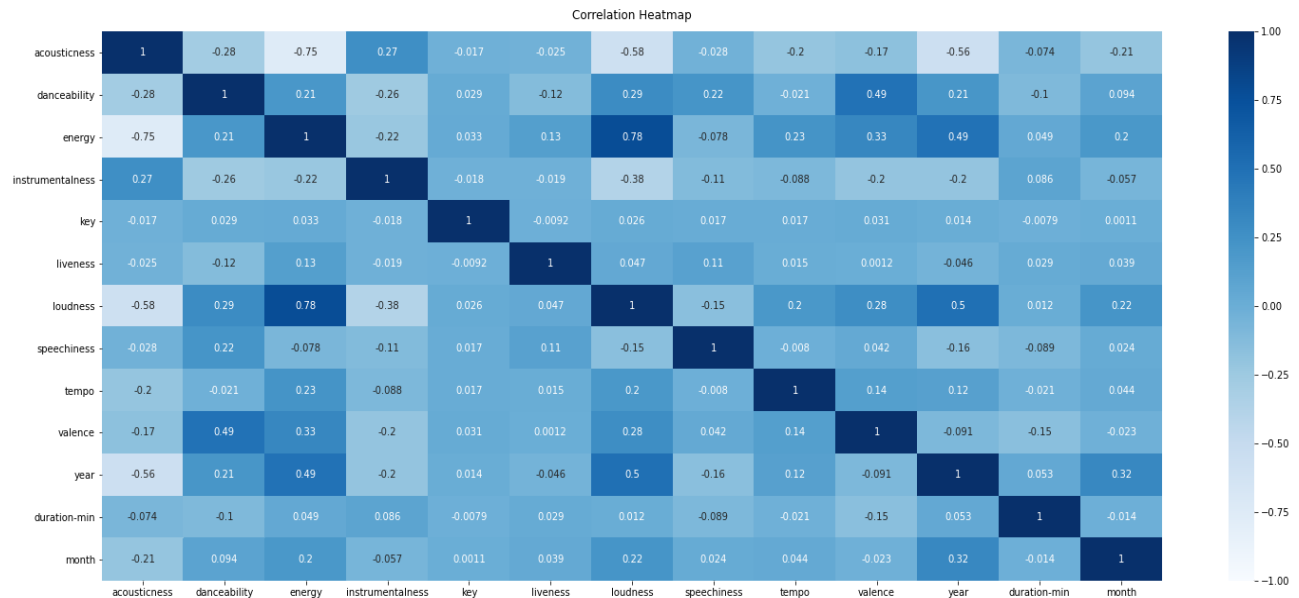
### \*Training Data Summary

Column1	id	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	speechiness	tempo	valence	year	duration-min
count	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000
mean	8175.03	0.427859723	0.555744225	0.529122438	0.145706233	5.186	0.2029935	-10.5767665	0.0940926	117.7419722	0.52948885	1984.4245	3.87085
std	4664.746749	0.367669395	0.17458893	0.26518421	0.295532417	3.538295336	0.173700286	5.423012693	0.151436185	29.54031501	0.259133368	25.79517683	2.386807384
min	2	1.73E-06	0	0.00167	0	0	0.0199	-42.66	0	0	0	1920	0.5
25%	4137.75	0.058925	0.435	0.306	0	2	0.095875	-13.5715	0.0346	95.0175	0.321	1965	2.9
50%	8195	0.345	0.565	0.538	9.44E-05	5	0.132	-9.546	0.0456	116.2605	0.542	1987	3.5
75%	12143.75	0.805	0.686	0.752	0.0418	8	0.26	-6.5335	0.07645	135.2445	0.74	2008	4.4
max	16226	0.996	0.966	0.999	0.994	11	0.99	-0.007	0.966	217.913	0.991	2021	80

### \*Testing Data summary

From the mean, 75% and the maximum readings of the meters, it can be seen that the **dataset consists of outliers** that have to be treated for better prediction.

## Correlation Matrices




This is the Pearson's correlation plot. The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. With only the numerical variables being taken into consideration, this plot justifies the choice of variables in the final model. The categorical variables like Explicit and Mode are also quantified and are further proved to be a good measure of evaluation.

We have calculated the correlation matrix across all the features. We found out that acousticness, energy, loudness and year of release are **fairly correlated**. Danceability has a decent positive correlation with valence. Instrumentalness has a decent negative correlation with loudness.

## Multi-Classification Models

There are a variety of Multi-Class Classification models available and are used in a variety of fields. We will be using these models to build an efficient and robust model to predict the popularity of the music tracks. We applied various Classification models from different classes such as Statistical (Linear discriminant analysis), Machine Learning based (XGBoost Regressor, Random Forest Regressor) as well as Deep Learning-based (Multi-Layer Perceptron) models. We



have presented below the results for the best models and our intuition of why these models performed the way they have performed.

Next, we have explored the problem of class imbalance in the given dataset. We have experimented with various sampling techniques, and have presented the ones that deem to be suitable for the model. We ultimately found Synthetic Minority Oversampling Technique (SMOTE) to be the best among other sampling techniques and used this to balance the labels.

We compared various evaluation metric values for the validation set and have presented them below for each model. We observed that over-sampling generally brought a positive impact on the model prediction. You can see the various metrics and graphs of the results of the same below. First, we will describe the sampling technique used as a preprocessing step in all of our models.

## Synthetic Minority Oversampling Technique (SMOTE)

A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. One way to solve this problem is to oversample the examples in the minority class.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then  $k$  of the nearest neighbors for that example are found (typically  $k=5$ ). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

This procedure can be used to create as many synthetic examples for the minority class as are required. The approach is effective because new synthetic examples from the minority class are created that are plausible, that is, are relatively close in feature space to existing examples from the minority class.

## Custom Encoding

More than two-thirds of Billboard's top 100 number one songs that charted in 2017 feature explicit lyrics, a new study shows based on Spotify API data. Data shows that up till 2001, only five songs with explicit lyrics made it to number one, but since then, explicit chart-toppers have rocketed by 833 percent. Owing to this study, we decided to perform exclusive ordinal encoding on the Explicit variable, taking into consideration that explicit songs are more likely to be popular.



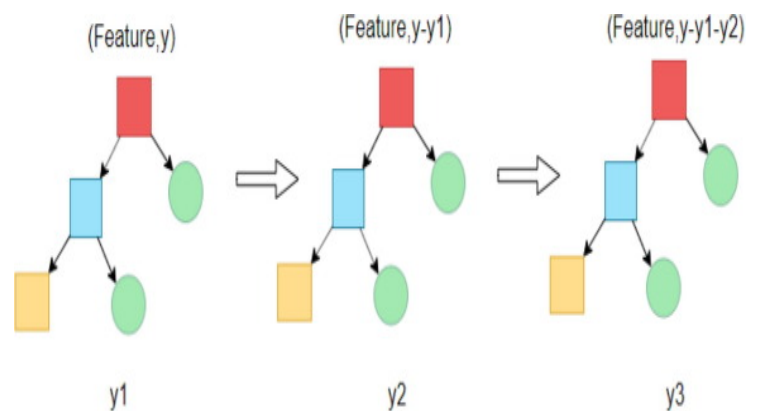
## XGBoost (eXtreme Gradient Boosting)

### Theory

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. It is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

### Why XGBoost?

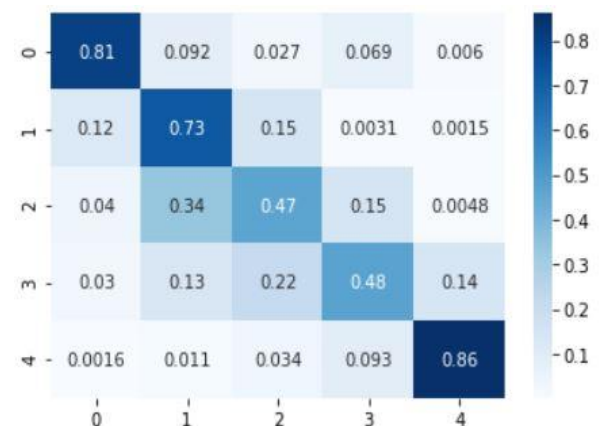
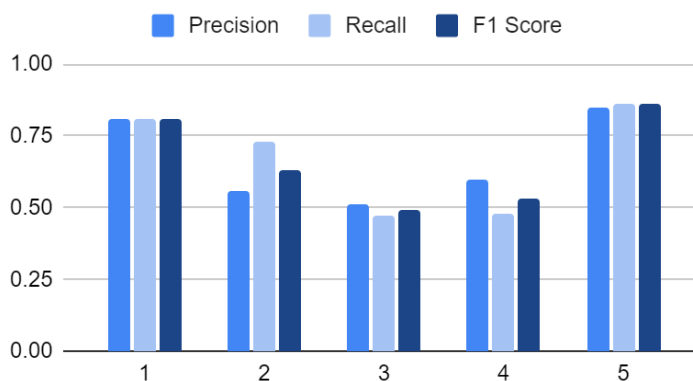
- **Execution Speed:** It is really fast when compared to other implementations of gradient boosting.
- **Model Performance:** It dominates structured or tabular datasets on classification and regression predictive modelling problems.
- Finally, the evidence is that it is the go-to algorithm for competition winners on the Kaggle competitive data science platform.



### Results

We fine-tuned the XGBoost model and achieved an accuracy of 64.2%, which is quite low compared to the other models. However, upon using SMOTE, it jumped up to 67.54%.

### Evaluation on the Validation Set



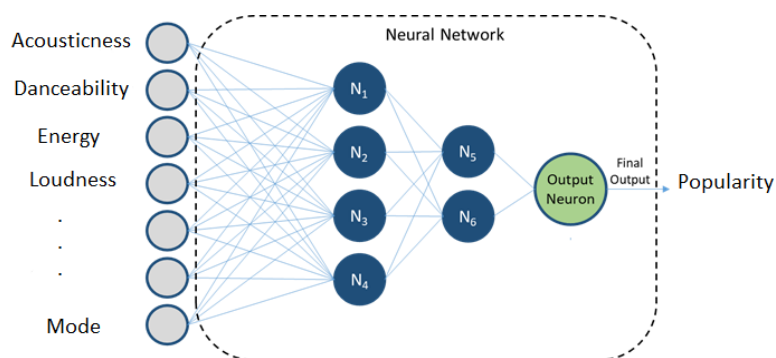
## Multi-Layer Neural Network

### Theory

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. Neural networks are broadly used, with applications for financial operations, enterprise planning, trading, business analytics and product maintenance. Neural networks have also gained widespread adoption in business applications such as forecasting and marketing research solutions, fraud detection and risk assessment.

### Why Neural Networks?

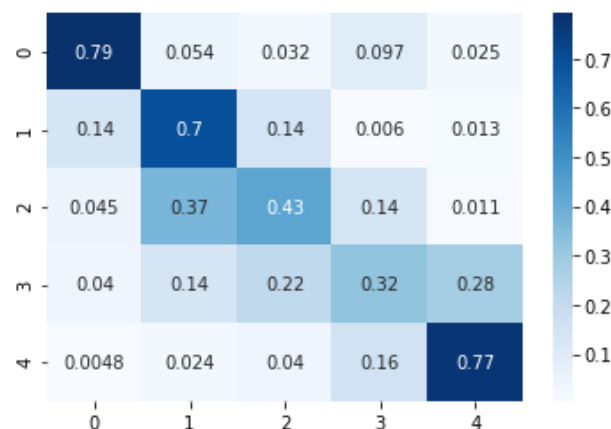
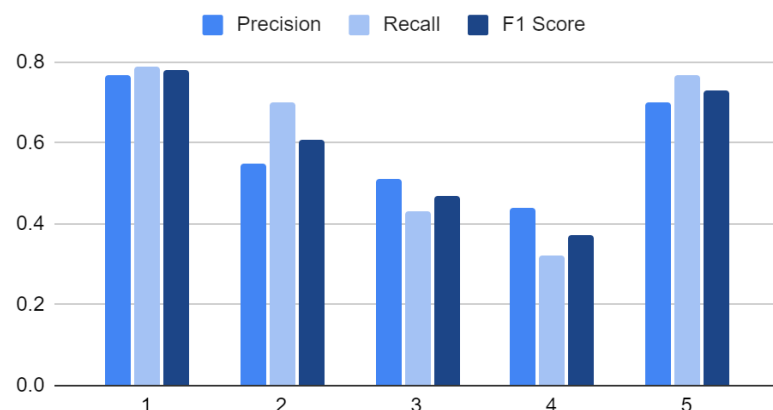
Neural networks have been shown to outperform a number of machine learning algorithms in many industry domains. They keep learning until it comes out with the best set of features to obtain a satisfying predictive performance. It requires less feature engineering from our side as the network itself learns complex functions to predict the label.



### Results

Though it requires a lot of hyperparameter tuning, we were able to take advantage of modern libraries like PyCaret for tuning. We achieved an accuracy of 58.67%, which is significantly lower than other models. However, upon using SMOTE, it increased to 60.24%.

### Evaluating on the Validation Set



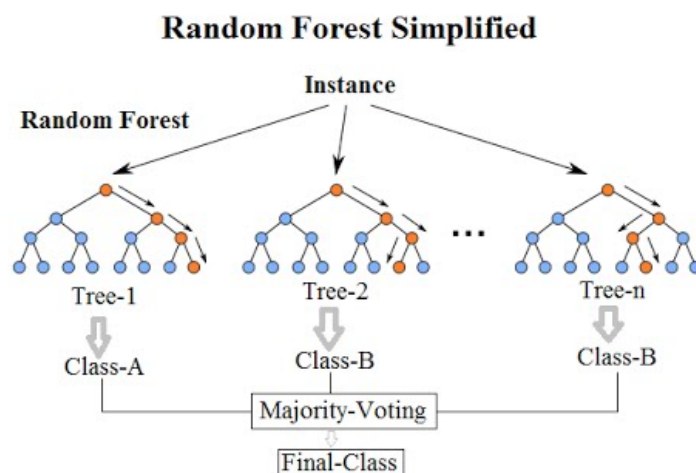
## Random Forest

### Theory

Random forest is a supervised learning algorithm which is used for both classification as well as regression. However, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

### Why Random Forest ?

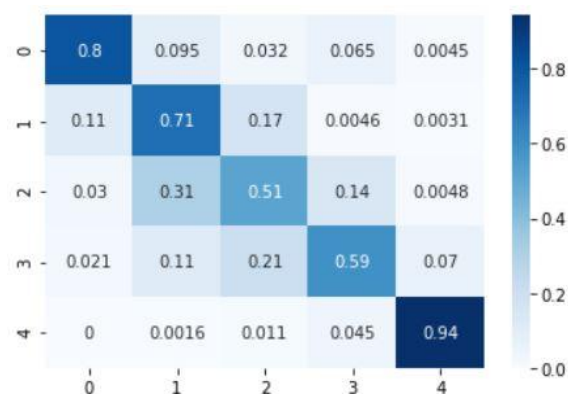
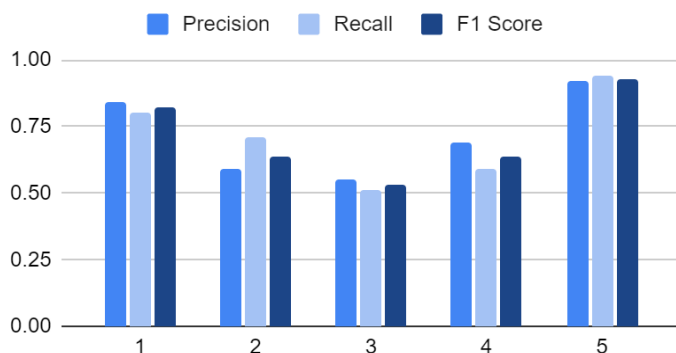
Random Forests are mainly used because it overcomes the problem of overfitting by averaging or combining the results of different decision trees. Random forests work well for a larger range of data items than a single decision tree does. They are very flexible and possess very high accuracy. Scaling of data is not required in a random forest algorithm as it maintains good accuracy even after providing data without scaling. Random Forest algorithms maintain good accuracy even if a large proportion of the data is missing. Also they work best for class-imbalanced data, as in our case.



### Results

We had to fine-tune the RandomForest model as the model performance depends a lot on its complexity. After fine-tuning the model we achieved an accuracy of 62.09%, which is comparable to other models. However, upon using SMOTE, it increased to 72.28% which is the highest validation accuracy we achieved during our experimentation.

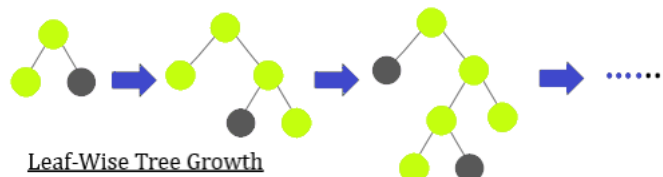
### Evaluation on the Validation Set



## LightGBM ( Light Gradient Boost Machine )

### Theory

LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks . LightGBM grows trees vertically while other tree based learning algorithms grow trees horizontally. It means that LightGBM grows tree leaf-wise while other algorithms grow level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, leaf-wise algorithms can reduce more loss than a level-wise algorithm.



### Why LightGBM ?

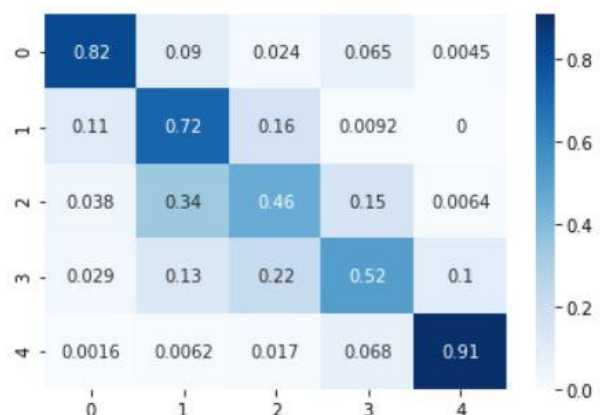
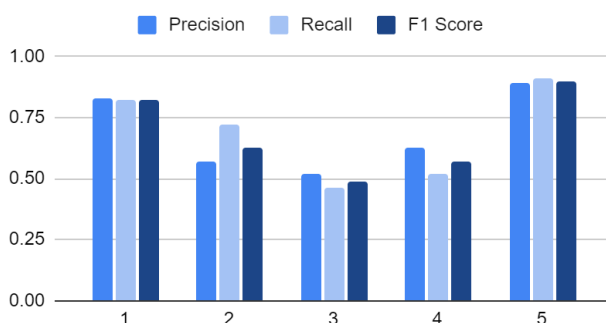
LightGBM as we already know is a gradient boosting framework that makes the use of tree-based learning algorithms. It works with significantly lower memory consumption on both efficiency and accuracy, and can outperform other existing boosting frameworks. Adding on, LightGBM by using multiple machines for training in specific settings can achieve a linear speed-up. It is designed with the following advantages :

- Higher efficiency as well as faster training speed
- Usage of lower memory
- Better accuracy
- Supports Parallel and GPU learning
- Data of large-scale can be handled

### Results

We fine-tuned the XGBoost model and achieved an accuracy of 61.37%, which is comparable to the other models. However, upon using SMOTE, it jumped up to 69.52% which is one of the highest accuracies we achieved.

Evaluation on the Validation Set



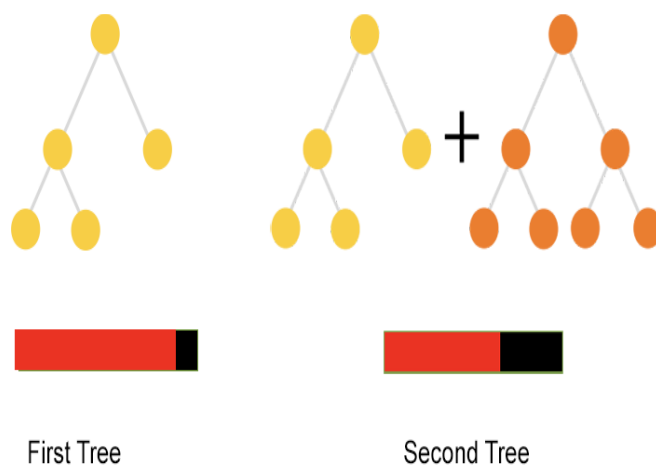
# CatBoost

## Theory

CatBoost is an open-source machine learning algorithm from Yandex which is based on Gradient Boosting and can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. CatBoost yields state-of-the-art results without extensive data training typically required by other machine learning methods and provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems.

## Why CatBoost?

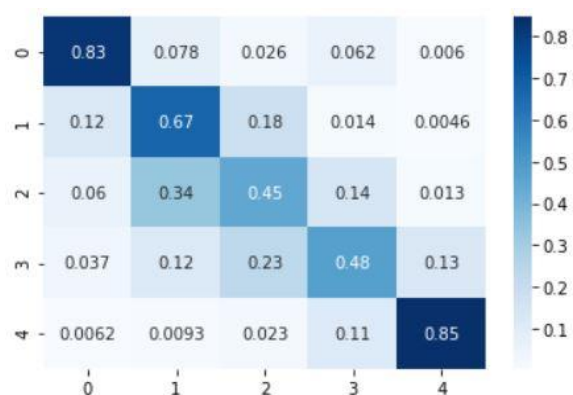
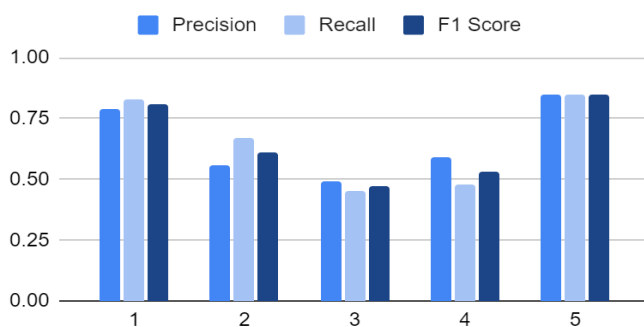
CatBoost provides state of the art results and it is competitive with any leading machine learning algorithm on the performance front. We can use CatBoost without any explicit pre-processing to convert categories into numbers. CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features. It reduces the need for extensive hyper-parameter tuning and lowers the chances of overfitting also which leads to more generalized models.



## Results

After fine-tuning some parameters like number of trees, tree depth, we were able to achieve an accuracy of 61.60% which is comparable to other models. However, upon using SMOTE, it increased to 69.34% which outperforms some of the models we used.

Evaluation on the Validation Set



## Conclusion


On a final note, we present our journey through using different models, their accuracies and thus the inherent reasons for exploring further classification models and feature engineering techniques. The main motive behind finalizing and fine tuning the models was to maximise the bidding profit and the number of successful bets placed, rather than the maximizing the classification accuracy.

### Data preprocessing:

Usual preprocessing techniques such as encoding and scaling were implemented. Further, we employed several visualisation tools to get an intuitive sense of the importance of some data over others (some of these insights are provided above). Custom weights were given to the popularity data to get a higher revenue and bidding accuracy.

### Modelling:

Model	Accuracy(%)	Remarks
Multi-Layer Neural Network	60.24	An immediate crude neural net fitted with all features as it is
Catboost	69.34	Gradient boosting on full decision trees (depth of 2) run for 100 iterations
XGBoost	67.54	Slight hyperparameter tuning (especially max-depth of local tree) gave us the sited increase
LightGBM	69.52	Another gradient boosting algorithm (with vertically/leaf-wise growing decision trees) working with random state of 5
Random Forest (with iterations over the random state)	72.28	A heuristic decision making algorithm was implemented to find the optimal random state by going over all locally optimal decision trees



A particular area of interest in our modulation is the huge jump in accuracy we observed for almost all models with the introduction of over-sampling (SMOTE). The most rudimentary implementations of SMOTE in effect could simply duplicate the minority classes, and still a boost in classification accuracy is observed. Following its success, we attempted to implement several hyperparameter tunings to refine this sample generation. Indeed, over-sampling can easily over-fit on test data and mislead a designer into believing that the real-world data could also be modelled likewise. Speaking of real-world, we have also tried to find out the bidding accuracy resulting from said classification.

### Bidding accuracy (making the client happy):

The model based on Random Forest Algorithm generates the highest possible revenue based on the predictions. For 10000(in 10k \$) initial investment to place bids on the 4000 music tracks, our model generates 83.71 % successful bids and a revenue of \$15648 (in \$10k) which amounts to 163% Return on Investment for the record label.

### Ending note:

Through this report, we have portrayed our efforts towards smashing new sales records! Since all technicalities have already been discussed at length, we now sit back and try out our new model as much as possible! We would love your opinion on the same! Would you recommend this book to someone else?

