PREDICTING SPOTIFY HITS

Abélia PETELLE
February, 2020

TABLE OF CONTENTS

01

MISSION STATEMENT

OUR COMPANY

DATALIVE CONSULTING GROUP

Datalive is an international consulting firm, specialized in analysing data for the the music industry since 2008.

OUR CLIENT

UNIVERSAL MUSIC GROUP

Universal MG is an American music label, one of the top three major record companies in the world.

PROBLEM VS. OUR SOLUTION

PROBLEM

Universal MG wants to put the track that has the best chance of becoming a hit at the first position of the album.

Expert opinions are not always convincing. They would like to have more certainty.

SOLUTION

Predict whether a song will become a hit using Machine Learning algorithms on Spotify's features tracks over the past 60 years.

02

PROCESS

PROCESS

DATA COLLECTION

- Data : Features for tracks fetched using Spotify's Web API
- Dataset : Smaller version of it from Kaggle

DATA CLEANING

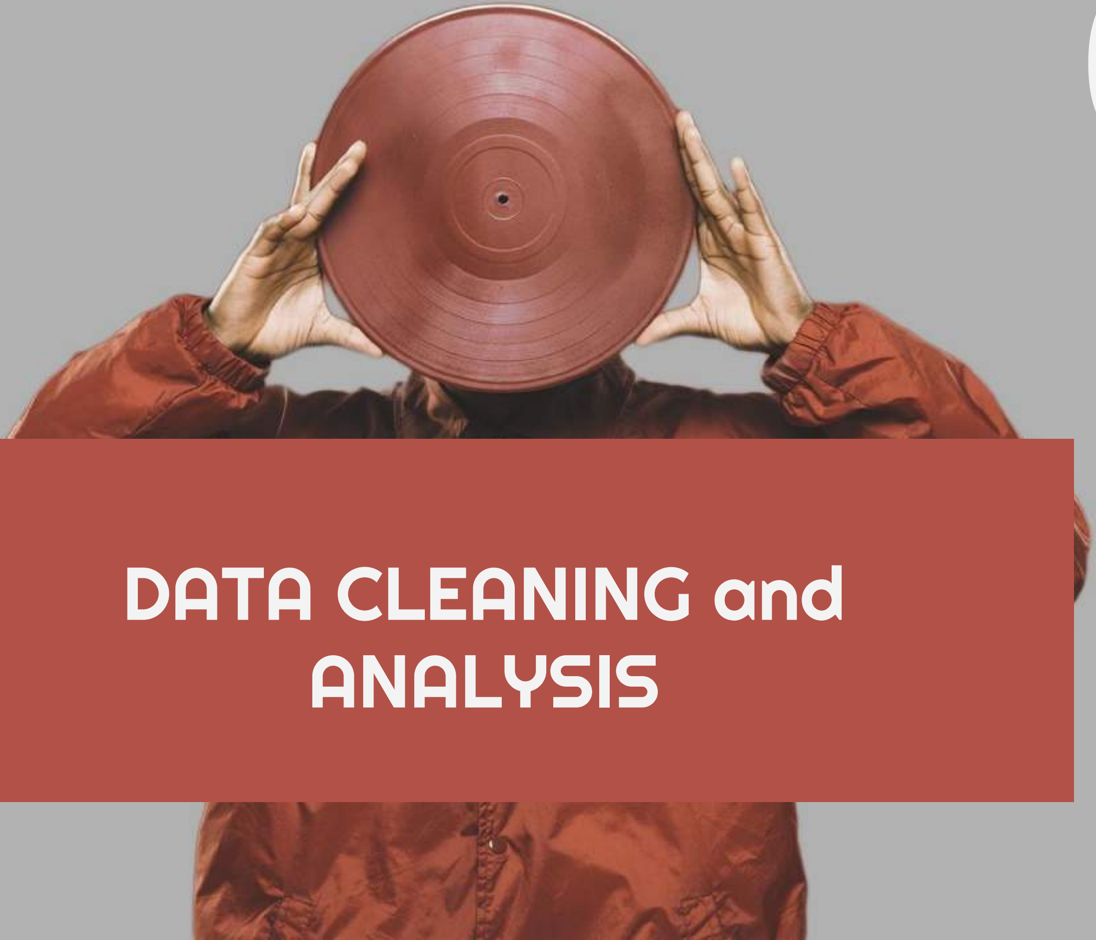
DATA EXPLORATORY

- Correlation
- Distribution

MODELING

1. Logistic Regression
2. K Nearest Neighbors
3. Support Vector Machine
4. Decision trees
5. Random forests
6. Naive Bayes
7. Catboost
8. With PCA

03



DATA CLEANING and ANALYSIS

DATA

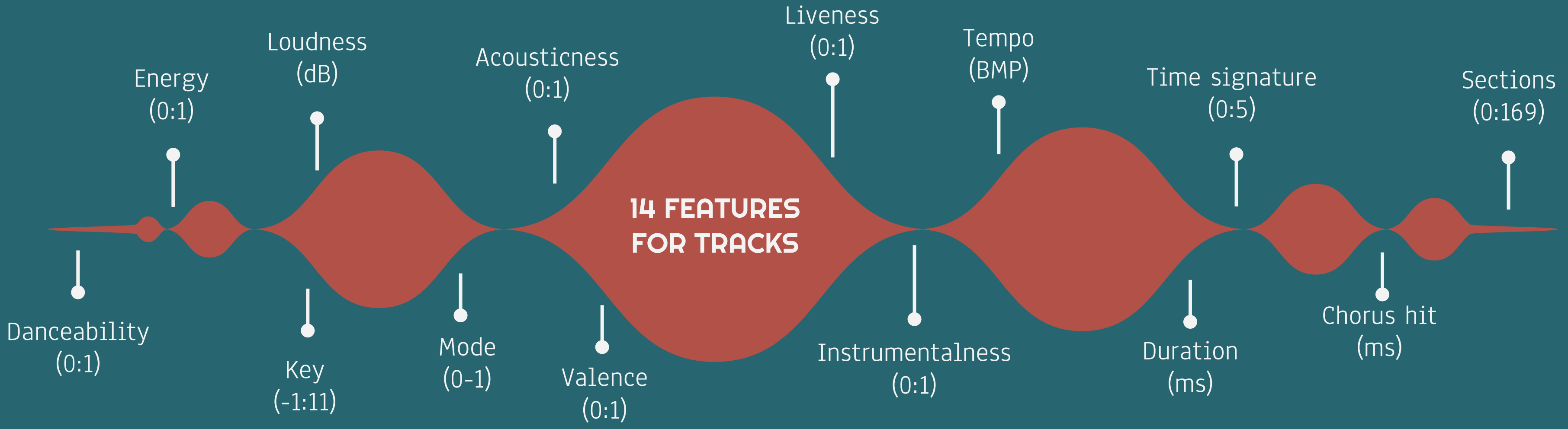
18 FEATURES

TRACK NAME

ARTIST NAME

TARGET :
- 1 : « Hit »
- 0 : « No hit »

SPOTIFY URI



DATA MANIPULATION

1. Adding « decade » column to classify each track from 1960s to 2010s

2. Removed the URI column (unique value, no valuable information)

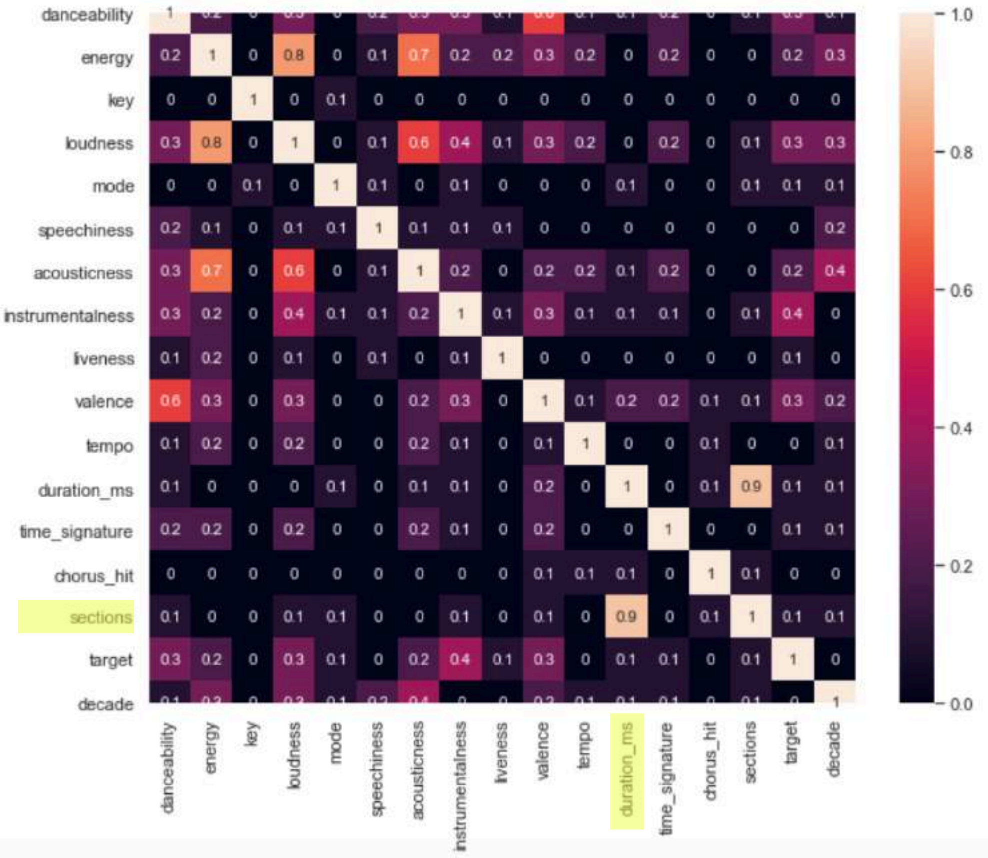
3. Creating dummies for long and short tracks

41,106 ROWS

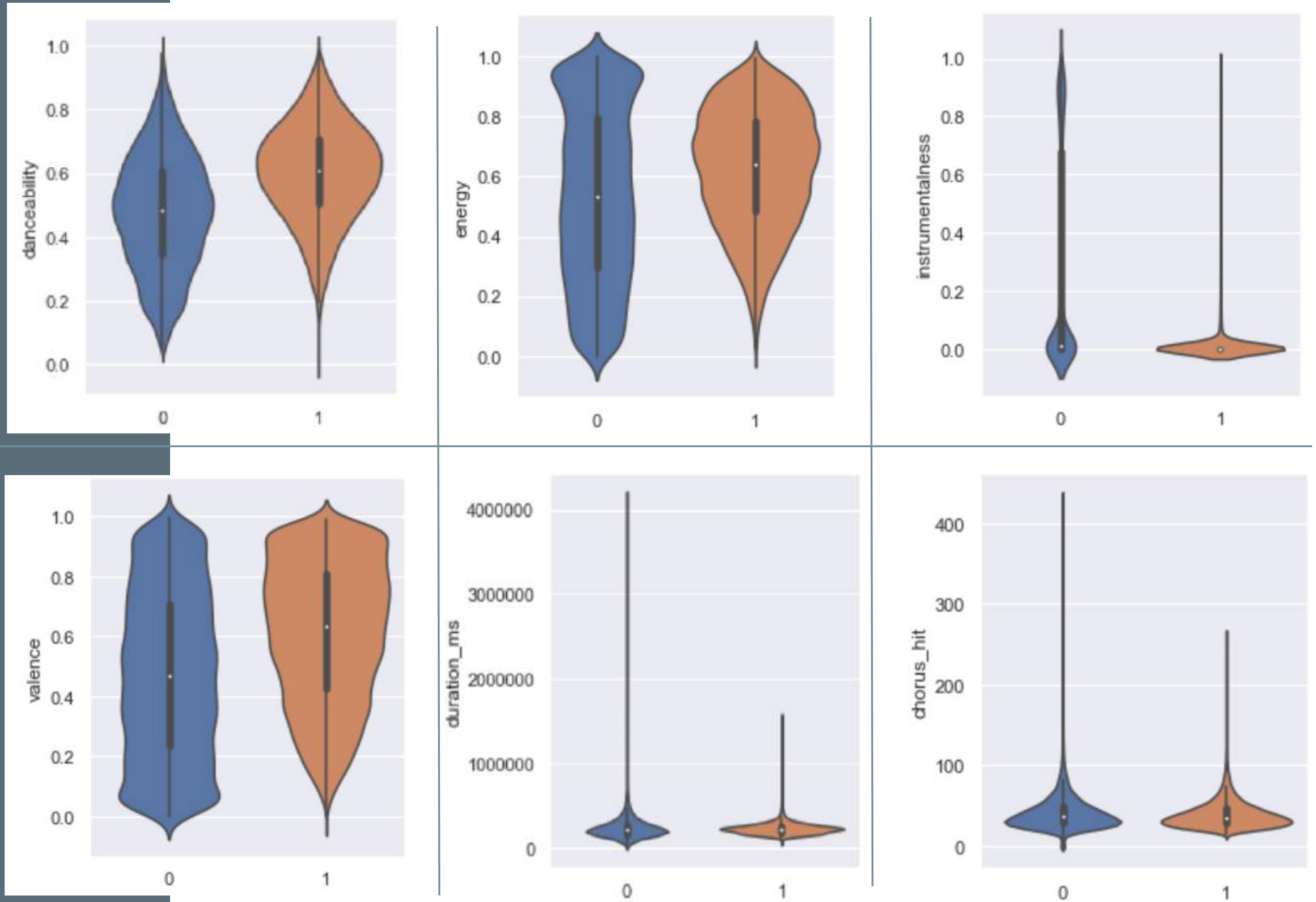
19 COLUMNS

HEATMAP

CORRELATION



→ Duration and sections are highly correlated



DISTRIBUTION

Compared to tracks that don't became hits, the hits :

- Have a higher level of danceability
- Have a higher level of energy
- Contain more vocals
- Have a higher level of positiveness
- Do not exceed 5min 33sec
- Their chorus start earlier

MODELING

04

MODELS

1. Logitic Regression
2. K Nearest Neighbors
3. Support Vector Machine
4. Decision trees
5. Random forests
6. Naive Bayes
7. Catboost

---

1. Logitic Regression with PCA
2. KNN with PCA
3. SVM with PCA

OBJECTIVE

TRUE NEGATIVE MODEL PREDICTED « NO HIT » AND IT IS NOT A HIT	**FALSE POSITIVE** MODEL PREDICTED « HIT » BUT IT NOT BECAME A HIT
FALSE NEGATIVE MODEL PREDICTED « NO HIT » BUT IT BECAME A HIT	**TRUE POSITIVE** MODEL PREDICTED « HIT » AND IT BECAME A HIT

= LOST OF MONEY

LOST OF OPPORTUNITY =

The worst case
= **client losing money**

The main objective
= **decrease the False Positive**

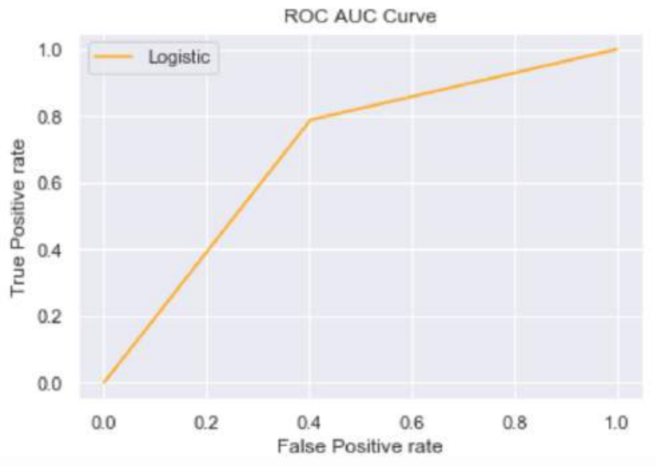
Metrics to increase
= **precision score**

1. LOGISTIC REGRESSION

First model

FEATURES

All except 'track','artist','target','sections'




```
The confusion matrix is:
 [[4052 2731]
 [1437 5345]]
The auc score is:  0.693
The accuracy score is:  0.693
The recall score is:  0.788
The precision score is:  0.662
F1 score is:  0.719
```

→ This first model is **not very convincing**:

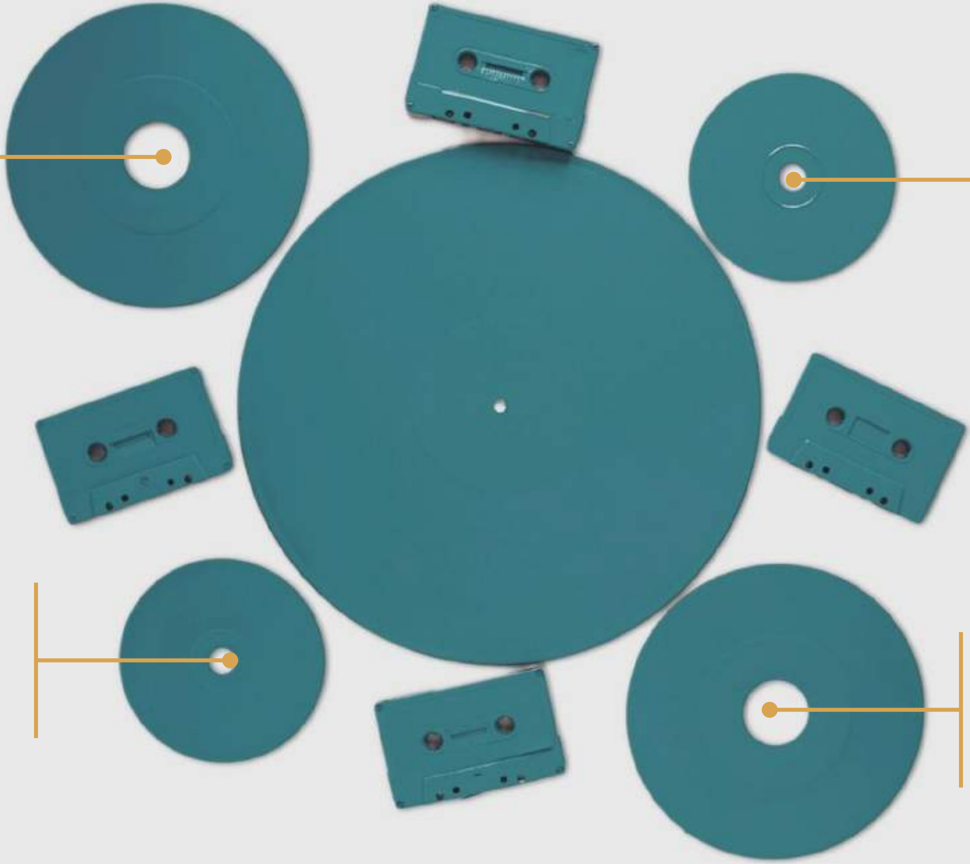
- false positive are very high
- all metrics are low and can be improved

FEATURES

feature	count	mean	std	min	25%	50%	75%	max
danceability	41106	0.5	0.2	0	0.4	0.6	0.7	1
energy	41106	0.6	0.3	0	0.4	0.6	0.8	1
key	41106	5.2	3.5	0	2	5	8	11
loudness	41106	-10.2	5.3	-49.3	-12.8	-9.3	-6.4	3.7
mode	41106	0.7	0.5	0	0	1	1	1
speechiness	41106	0.1	0.1	0	0	0	0.1	1
acousticness	41106	0.4	0.3	0	0	0.3	0.7	1
instrumentalness	41106	0.2	0.3	0	0	0	0.1	1
liveness	41106	0.2	0.2	0	0.1	0.1	0.3	1
valence	41106	0.5	0.3	0	0.3	0.6	0.8	1
tempo	41106	119.3	29.1	0	97.4	117.6	136.5	241.4
duration_ms	41106	234877.6	118967.4	15168	172927.8	217907	266773	4170227
time_signature	41106	3.9	0.4	0	4	4	4	5
chorus_hit	41106	40.1	19	0	27.6	35.9	47.6	433.2
sections	41106	10.5	4.9	0	8	10	12	169
target	41106	0.5	0.5	0	0	0.5	1	1

→ Features that can be standardize : tempo, duration_ms, chorus_hit

FEATURE ENGINEERING



CHORUS HIT
Standardizing data

TEMPO
Standardizing data

DURATION
Converting millisecong to minute

DECADE
Creating a column that squared the decade to put more importance on the last decade

1. LOGISTIC REGRESSION

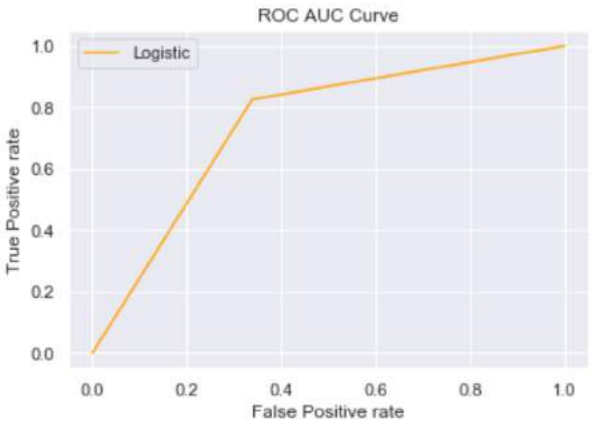
Second model

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Weight on the distance




```
The confusion matrix is:
 [[4487 2295]
 [1184 5599]]
The auc score is:   0.744
The accuracy score is:   0.744
The recall score is:   0.825
The precision score is:   0.709
F1 score is:   0.763
```

→ This second model is **better than the first one**: less false positive, precision have increase from 0.62 to 0.71

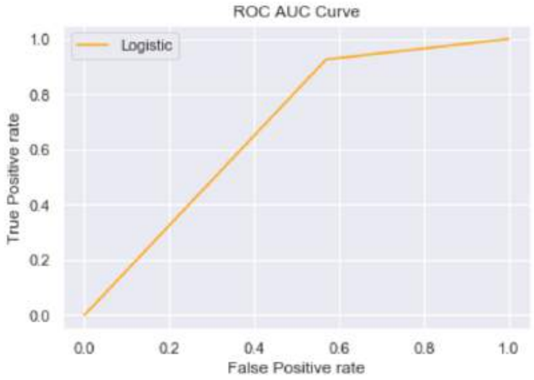
2. K NEAREST NEIGHBORS

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model
- Optimal number of neighbors = 99
- Weight on the distance




```
The confusion matrix is:
 [[2918 3864]
 [ 501 6282]]
The auc score is:  0.678
The accuracy score is:  0.678
The recall score is:  0.926
The precision score is:  0.619
F1 score is:  0.742
```

→ This third model is **worse than the previous one**: precision decreased from 0.71 to 0.62
The best model remains Linear Regression 1

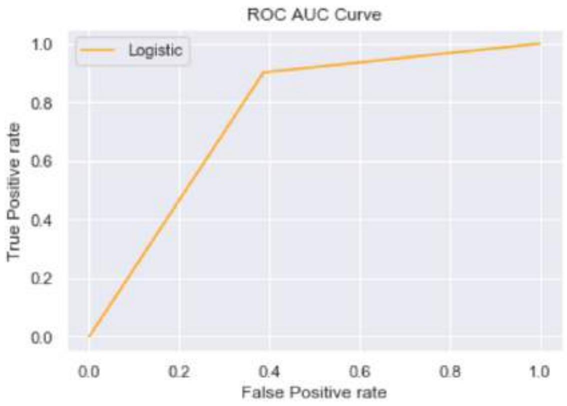
3. SUPPORT VECTOR MACHINE

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model
- Nu = 0.6
- Gamma = « scale »




```
The confusion matrix is:
 [[4160 2622]
 [ 666 6117]]
The auc score is:   0.758
The accuracy score is:  0.758
The recall score is:  0.902
The precision score is:  0.7
F1 score is:  0.788
```

→ The AUC score of this third model is **slightly better the Logistic Regression** model
But **not regarding the precision score**

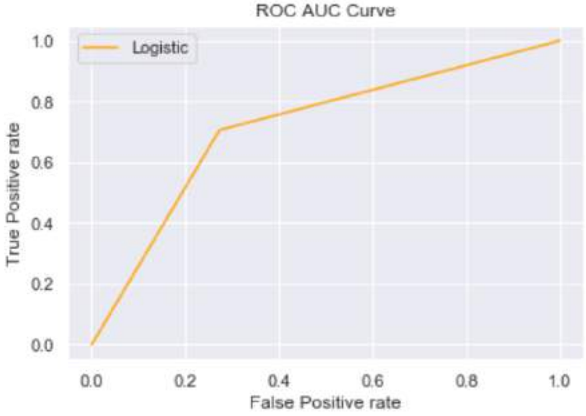
4. DECISION TREE

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model
- With weight on the distance




```
The confusion matrix is:
 [[4932 1850]
 [1995 4788]]
The auc score is:  0.717
The accuracy score is:  0.717
The recall score is:  0.706
The precision score is:  0.721
F1 score is:  0.714
```

→ This fourth model have **the best precision score** of all models but it is also known for sometimes **overfitted**

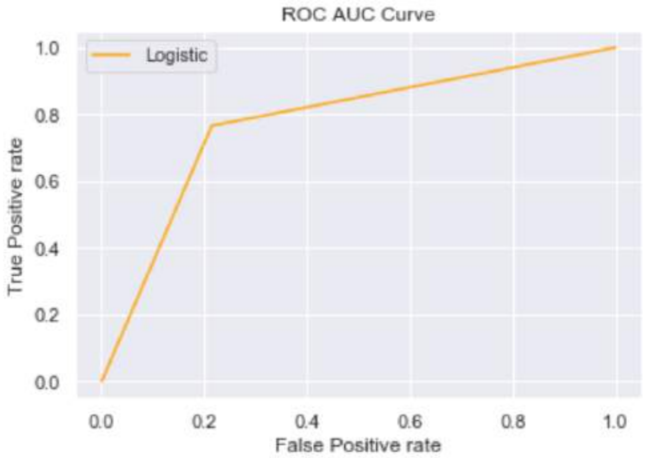
5. RANDOM FOREST

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model




```
The confusion matrix is:
 [[5327 1455]
 [1587 5196]]
The auc score is:  0.776
The accuracy score is:  0.776
The recall score is:  0.766
The precision score is:  0.781
F1 score is:  0.774
```

→ The AUC score from Random Forest is higher than the AUC score from Decision Tree : **Decision Tree was not overfitted**
This model becomes the **best model** since we have a AUC score and a precision score at 0.78

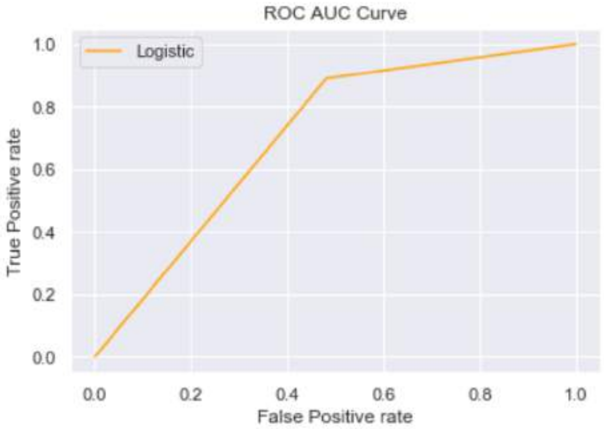
6. NAIVE BAYES

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model




```
The confusion matrix is:
 [[3518 3264]
 [ 743 6040]]
The auc score is:  0.705
The accuracy score is:  0.705
The recall score is:  0.89
The precision score is:  0.649
F1 score is:  0.751
```

→ From the previous model, **all metrics decreased**

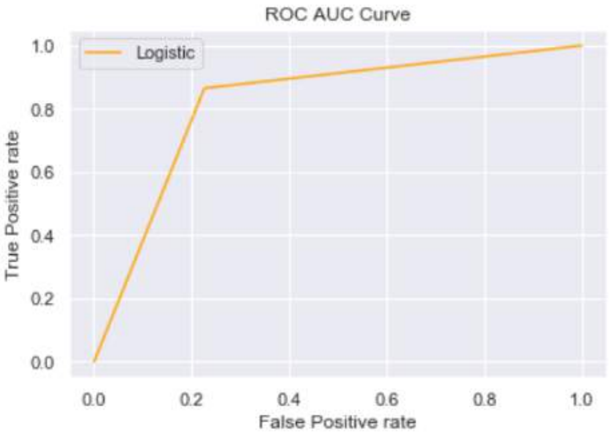
FEATURES

All except 'track','artist','target','sections'

7. CATBOOST

PARAMETERS

- Same feature engineering as the second Logistic Regression model




```
The confusion matrix is:
 [[5248 1534]
 [ 915 5868]]
The auc score is:  0.819
The accuracy score is:  0.819
The recall score is:  0.865
The precision score is:  0.793
F1 score is:  0.827
```

→ This last model have the **highest AUC and precision score** compared to all the other models

COMPARAISON

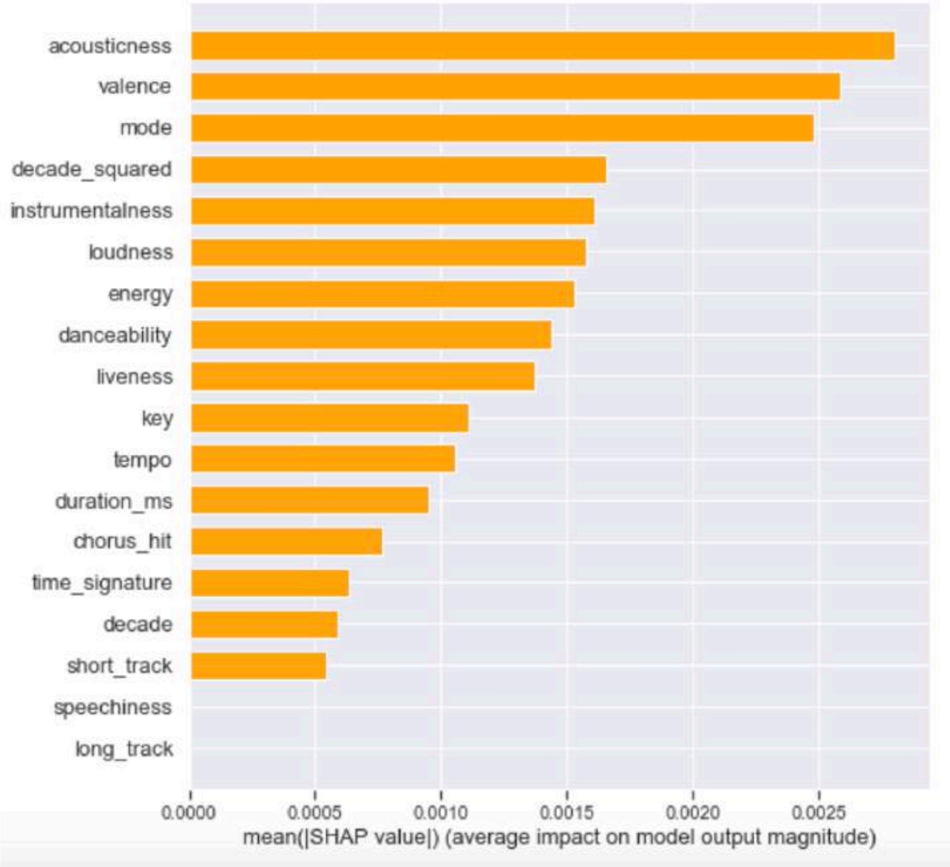
Model	Roc Auc	Accuracy	Recall	Precision	F1 score
Logistic Regression1	0.6	0.6	0.63	0.59	0.61
Logistic Regression2	0.74	0.74	0.83	0.71	0.76
K Nearest Neighbors	0.68	0.68	0.93	0.62	0.74
SVM	0.76	0.76	0.9	0.7	0.79
Decision Tree	0.72	0.72	0.71	0.72	0.71
Random Forest	0.78	0.78	0.77	0.78	0.77
Naive Bayes	0.7	0.7	0.89	0.65	0.75
Catboost	0.82	0.82	0.87	0.79	0.83

So far, the best model to predict hits on Spotify is **Catboost** regarding the AUC score and the precision score

SINCE OUR BEST MODEL IS CATBOOST,
LET'S ANALYSE THE
CONTRIBUTION OF THE FEATURES
FOR THIS MODEL

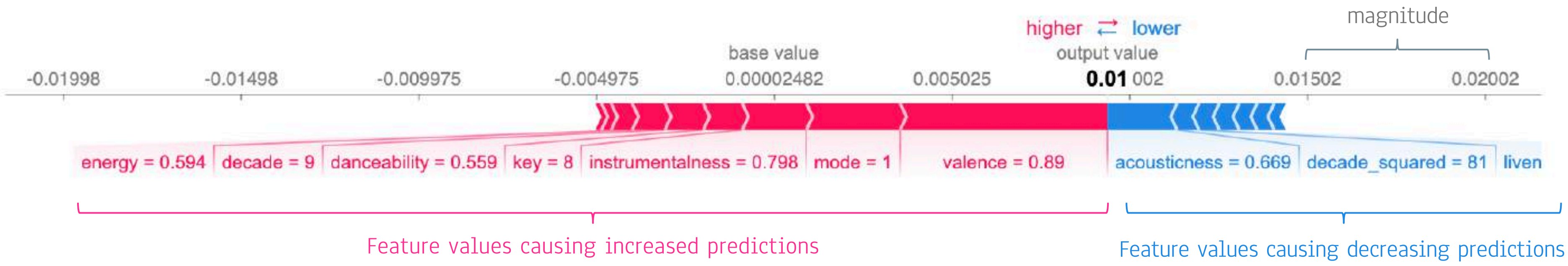
SHAP VALUE

Features that have a **high impact** for predicting a hit with catboost model are :



FORCE PLOT

How much was a prediction driven by the fact that the valence was equal to 1, instead of other values of valance ?



→ The biggest impact comes from **valence** being 0.89

→ Though the **accousticness** value has a meaningful effect, it decreasing the prediction

IF WE DO MORE FEATURE ENGINEERING : DOEST OUR MODEL CAN BE IMPROVED ?

Let’s test with Principal Component Analysis approach

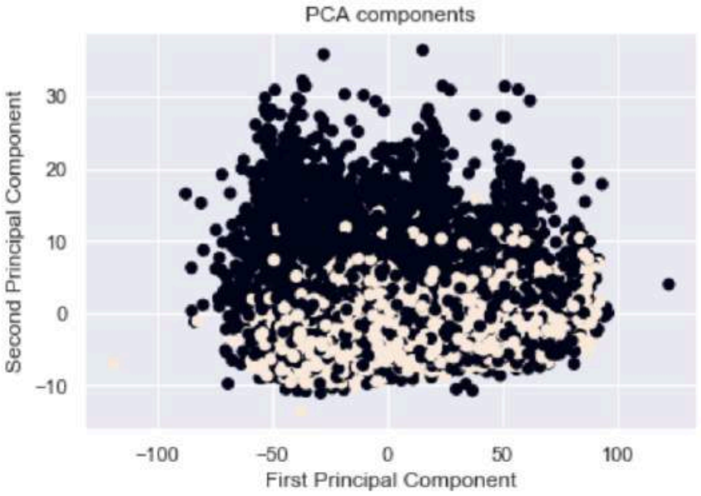
PCA

FEATURES

All except 'track','artist','target','sections'

PARAMETERS

- Same feature engineering as the second Logistic Regression model but without "squared decade" column



→ We can see that the **separation between the hits and non hits is not very clear**

COMPARISON

	Model	Roc Auc	Accuracy	Recall	Precision	F1 score
0	Logistic Regression1	0.598	0.598	0.629	0.592	0.610
1	Logistic Regression2	0.744	0.744	0.829	0.708	0.764
2	K Nearest Neighbors	0.678	0.678	0.926	0.619	0.742
3	SVM	0.758	0.758	0.902	0.700	0.788
4	Decision Tree	0.717	0.717	0.706	0.721	0.714
5	Random Forest	0.776	0.776	0.766	0.781	0.774
6	Naive Bayes	0.705	0.705	0.890	0.649	0.751
7	Catboost	0.819	0.819	0.865	0.793	0.827
8	Logistic Regression with PCA	0.596	0.596	0.645	0.587	0.615
9	K Nearest Neighbors with PCA	0.607	0.607	0.715	0.588	0.645
10	SVM with PCA	0.605	0.605	0.735	0.584	0.651

As, predicted, modeling with PCA is not convincing at all

06

SUMMARY

IMPROVMENTS

ADD THE NUMBER OF HITS AN ARTIST HAD FROM THE PAST

SUMMARY

- With Catboost, we are able to predict the success of a track by 82%, which is not that bad
- But False Positive errors remain high

- This dataset brings together tracks of all styles, and from all eras but without specifying the genre, or year
- With that kind of information, we could have specialized our model

→ We have an universal model, necessarily moderately successful over the entire world music catalogue

THANKS

Does anyone have any questions?

GLOSSARY

- **Track** : name of the song
- **Artist** : name of the artist
- **Uri** : resource identifier for the track
- **Danceability** (0:1) : how suitable a track is for dancing based on a combination of musical elements
- **Energy** (0:1) : represents a perceptual measure of intensity and activity
- **Key** (-1:[0:?]): the estimated overall key of the track (If no key was detected, the value is -1)
- **Loudness** : overall loudness of a track in decibels (dB)
- **Mode** (0-1) : Major is represented by 1 and minor is 0
- **Speechiness** (0:1): Speechiness detects the presence of spoken words in a track
- **Acousticness** (0:1) : whether the track is acoustic
- **Instrumentalness** (0:1): predicts whether a track contains no vocals
- **Liveness** (0:1) : the presence of an audience in the recording (live)
- **Valence** (0:1) : the musical positiveness conveyed by a track
- **Tempo** : the overall estimated tempo of a track in beats per minute (BPM)
- **duration_ms** : the duration of the track (in milliseconds)
- **time_signature** : notational convention to specify how many beats are in each bar (or measure)
- **chorus_hit** : estimate of when the chorus would start for the track (in milliseconds)
- **Sections** : the number of sections the particular track has
- **Target** : 1 : the song has featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is therefore a 'hit' and 0 : Implies that the track is not a hit