Top 100 Tester & Popularity meter features for suppliers

Spotify

Index

- Project Overview.
- Business understanding.
- Data understanding.
- Data Modeling.
- Expected recommendations.

Project Overview

At the end of each year, Spotify compiles a playlist of the songs streamed most often over the course of that year.

What do these top songs have in common? Why do people like them? What does it take for these songs to become popular?

Our task:

- Look for patterns in the audio features of the songs. Why do people stream these songs the most?
- Create a Top 100 Tester
- Examine popularity meter based on audio features

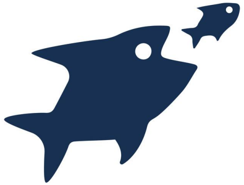
Business Understanding

Business Understanding - Industry

Music streaming industry

Contributed by **75%** to the reveunues of the Music industry in 2018.*

***MIDia research 2018**



Main players are:

- Spotify.
- Apple Music.
- SoundCloud.

Business Understanding - Company.

Founded in **2006**, this Swedish company has become the most popular streaming platform across the world.

In 2017, Spotify has reached **71.1** million subscribers.

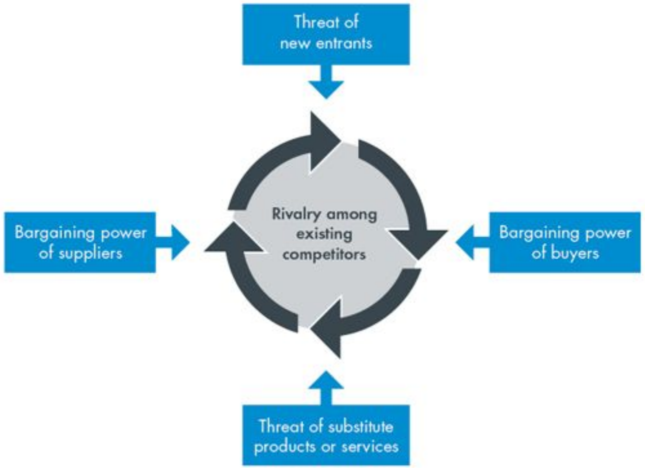
Spotify contols nearly **36%** of the global market.

Porter's Five Forces Framework

Supplier Power: In the music recording industry, the suppliers to the recording companies are the raw materials providers, artists, writers and producers. There is a large pool of talent, which is favorable because it gives the recording companies more negotiating power. It is evident that supplier power is **low** in the music industry.

Buyer Power: The threat the music recording industry faces from buyers is considered to be relatively **high**. Overall, buyers have significant power within the music industry as suppliers are forced to offer their products in various channels, especially online. Due to the high buyer power, revenues are decreased, costs are increased, and profits are decreased for the music industry.

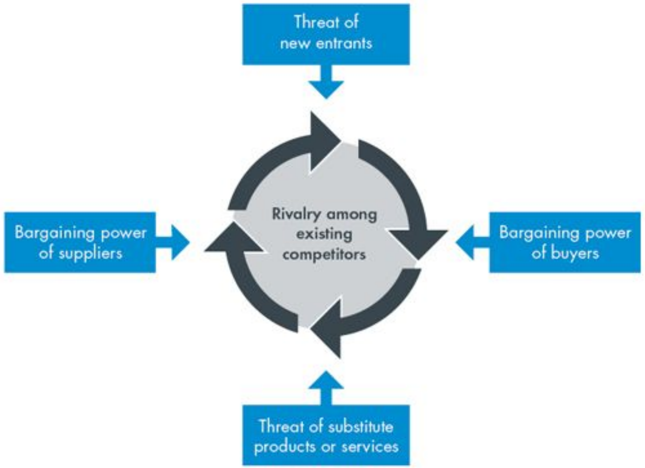
Threat of Substitutes: In the current music industry, music streaming services are confronted with various substitute products including physical records, digital media, TV, and radio channels playing 24 hour music, satellite radio, video streaming services, and piracy. The threat of substitutes is **high** and the industry experiences increased costs, decreased revenues, and decreased profits.



Porter’s Five Forces Framework

Threat of Entry: Independent artists and labels can now offer their products online at very low cost, and are able to skip several steps in the traditional value chain. Due to this change, the Big Three have reacted by building a network of resources and expertise to remain increasingly competitive in the industry. However, low product differentiation and moderate economies of scales still regards the threat of entry for the music industry as **high**.

Rivalry among existing competitors: There is a small number of large firms (Big Three-Universal, Sony, and Warner) that dominate the industry, and sales for each firm remain relatively high causing increased revenues and profits. However, there is low product differentiation because even though each label owns a selection of artists, the genres that they represent are common throughout the industry. Therefore, consumers are not dependent upon any one record company for a particular type of music. Overall, given that the music recording industry is dominated by a few, large competitors, has seen negative growth with increasing competition, and has low product differentiation, intra-industry rivalry is considered to be **high**.



Firm Description

Firm Description

- Universal Music Group is home to the most iconic and influential labels & brands in music. The firm provides recorded music, music publishing, and merchandising services. They develop, manufacture, market, sell, and distribute recorded music through a network of subsidiaries, joint ventures, and licensees. Universal Music Group serves customers worldwide
- 8319 employees at UMG as of 2018
- Global Firm: Australia, Germany, UK, China, Spain
- The line of business that is the subject of our analysis is the development and production of music.

Firm Description

Full Year 2018 Results - February 14, 2019

UNIVERSAL MUSIC GROUP

Key Figures

in euro millions	2017	2018	Δ organic (%)*
Revenues	**5,673**	**6,023**	**+10.0%**
Recorded music	4,559	4,828	+9.8%
Streaming and subscriptions	*1,971*	*2,596*	*+37.3%*
Other digital sales (mainly downloads)	*685*	*479*	*-26.6%*
Physical sales	*1,156*	*949*	*-16.1%*
License and Other	*747*	*804*	*+10.7%*
Music Publishing	854	941	+14.5%
Merchandising & Other	283	273	-1.5%
Intercompany Elimination	(23)	(19)	
Income from operations (IFO)	**798**	**946**	**+22.1%**
Income from operations margin	*14.1%*	15.7%	*+1.6pt*
Restructuring charges	(17)	(29)	
Share-based compensation plans	(9)	(4)	
Other special items excluded from IFO	(11)	(11)	
EBITA	**761**	**902**	**+22.1%**
EBITA margin	*13.4%*	*15.0%*	*+1.6pt*

* *At constant currency. See details on page 11*

Firm Description - SWOT Analysis

Strengths	Weaknesses	Opportunities	Threats
• Large global and local market • Strong Management • Brand Recognition • Artist Portfolio • Large Market Share • Influential Celebrity Power • Rich History • Artist Placement	• Piracy • File Sharing • Technology changing music trends (physical to digital shift) • Lack of discovery • Uncertainty regarding artist deals • Uncertainty with quality of content	• Diverse Consumer Base (global market) • Innovation Distribution channels • New Technologies • More fusion of genres • Festivals, concerts, events • Physical to Digital shift • Access to new talent	• Intra-Industry Competition • Government regulations (copyright laws) • Volatile costs • Individual artists • Music value to consumer (price)

Data Analysis and Understanding

Data Description

We used the following datasets for this analysis:

1. Top Spotify Tracks of 2017
2. Top Spotify Tracks of 2018
3. 19,000 Spotify Songs

Data Description

Audio Features:

song_name, song_popularity, song_duration_ms, acousticness, danceability, energy, instrumentalness, key, liveness, loudness, audio_mode, speechiness, tempo, time_signature, audio_valence

13070 records in our data set

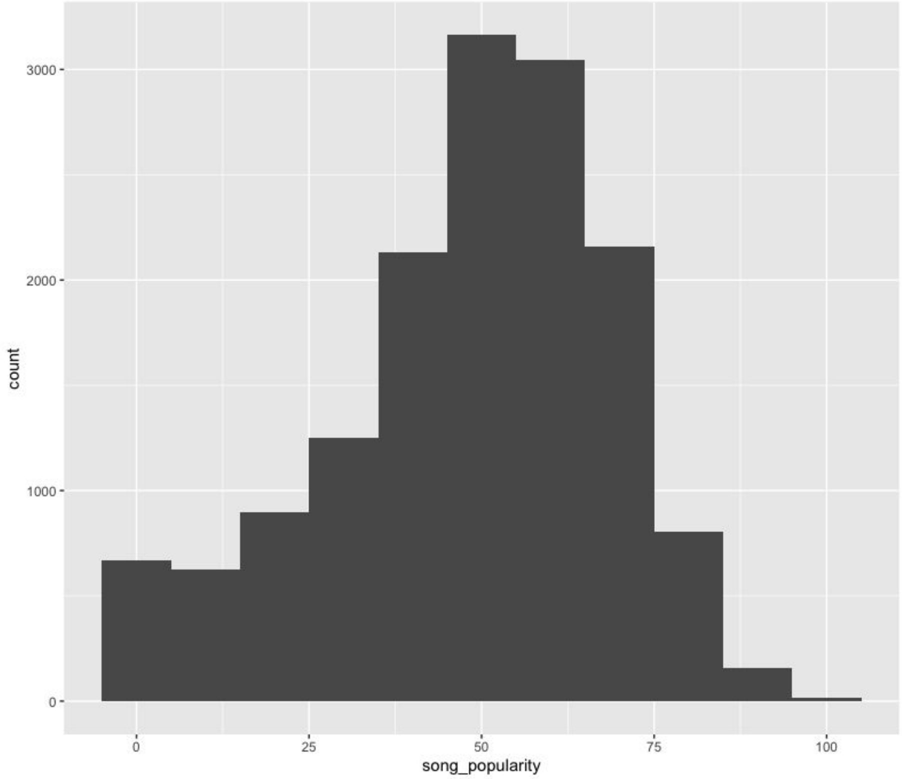
The data collection spanned 1 year

Target Variable: Song Popularity, Top Probability

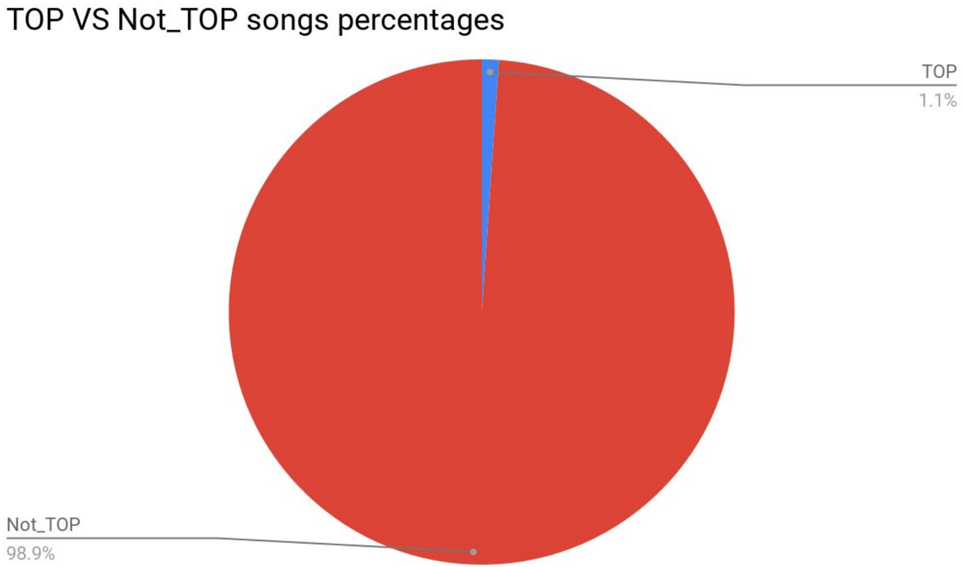
Data Summary

```
 song_popularity  song_duration_ms  acousticness        danceability       energy
 Min.   :  0.00   Min.   :  12000   Min.   :0.000001   Min.   :0.0000   Min.   :0.00107
 1st Qu.: 37.00   1st Qu.: 183944   1st Qu.:0.023600   1st Qu.:0.5240   1st Qu.:0.49600
 Median : 52.00   Median : 211846   Median :0.139000   Median :0.6360   Median :0.67200
 Mean   : 48.75   Mean   : 218950   Mean   :0.270452   Mean   :0.6245   Mean   :0.63976
 3rd Qu.: 63.75   3rd Qu.: 244720   3rd Qu.:0.458000   3rd Qu.:0.7400   3rd Qu.:0.81800
 Max.   :100.00   Max.   :1799346   Max.   :0.996000   Max.   :0.9870   Max.   :0.99900
 instrumentalness         key            liveness          loudness         audio_mode
 Min.   :0.0000000   Min.   : 0.000   Min.   :0.0109   Min.   :-38.768   Min.   :0.0000
 1st Qu.:0.0000000   1st Qu.: 2.000   1st Qu.:0.0930   1st Qu.: -9.389   1st Qu.:0.0000
 Median :0.0000208   Median : 5.000   Median :0.1220   Median : -6.750   Median :1.0000
 Mean   :0.0920668   Mean   : 5.301   Mean   :0.1804   Mean   : -7.677   Mean   :0.6319
 3rd Qu.:0.0051050   3rd Qu.: 8.000   3rd Qu.:0.2240   3rd Qu.: -4.991   3rd Qu.:1.0000
 Max.   :0.9970000   Max.   :11.000   Max.   :0.9860   Max.   :  1.585   Max.   :1.0000
  speechiness          tempo         time_signature   audio_valence
 Min.   :0.00000   Min.   :  0.00   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.03720   1st Qu.: 98.12   1st Qu.:4.000   1st Qu.:0.3320
 Median :0.05410   Median :120.02   Median :4.000   Median :0.5270
 Mean   :0.09942   Mean   :121.11   Mean   :3.953   Mean   :0.5270
 3rd Qu.:0.11300   3rd Qu.:139.94   3rd Qu.:4.000   3rd Qu.:0.7278
 Max.   :0.94100   Max.   :242.32   Max.   :5.000   Max.   :0.9840
```

Data Summary



Data overview and challenges.

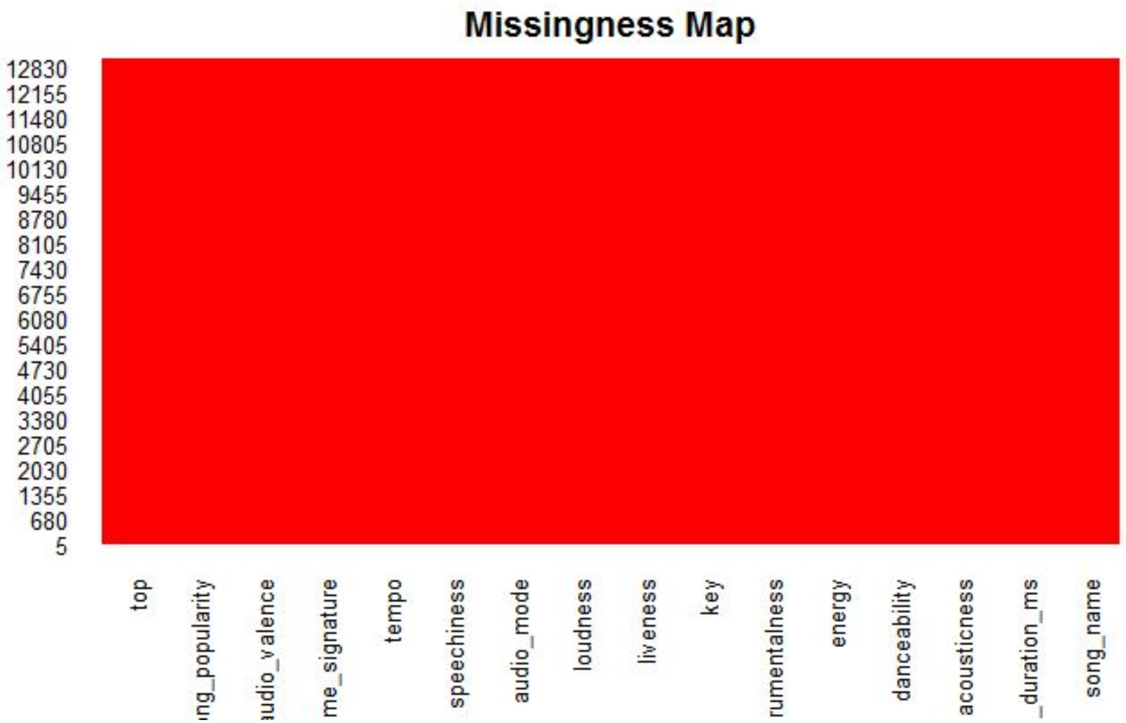


Data wrangling part:

We put **0** for songs which did not make it to the top 100 list in the last 2 years and **1** for everything else.

As we can find from the chart that positive values in the database is no more than **1.1 %** which would indicate that our database is unbalanced.

Data cleaning - Missing values

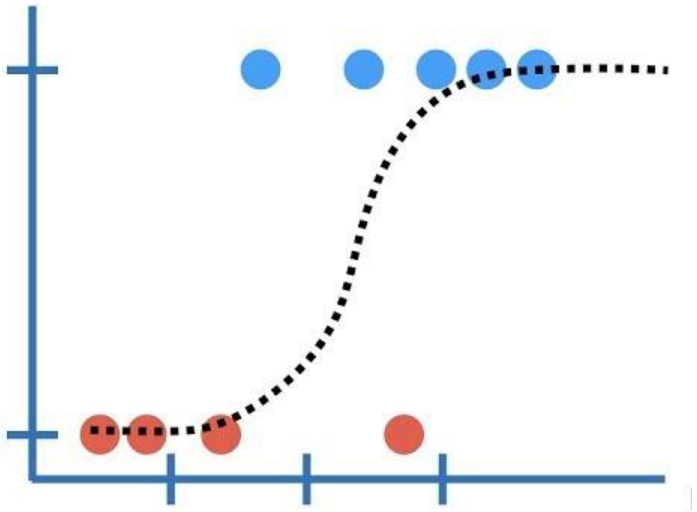


We are lucky that the database has no missing values.

At this stage we will proceed with the next step.

Top Tester

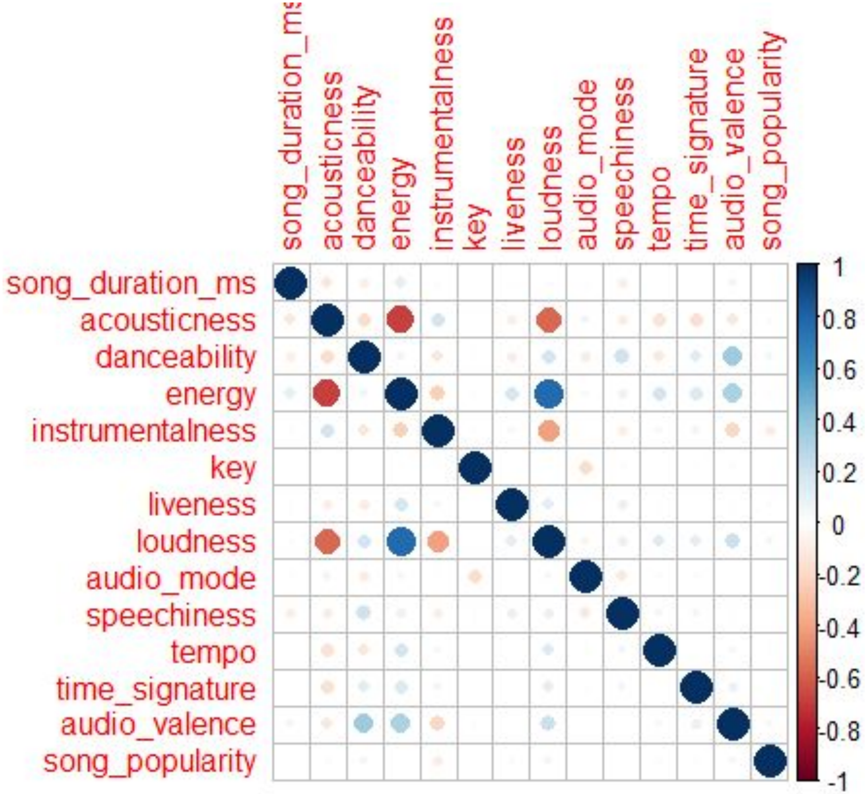
Model selection & challenges.



Choosing model:

As we can see here that the dependant variable which is the **“Top probability”** is a binomial one, so we chose the logistic regression.

Check for correlated variables



Checking correlated variables before start modeling is important in order to figure out which variables we will eliminate.

We can see that there are correlation between:

Energy, Acousticness and **Loudness**

Next step, we will check which has the least impact and remove it.

Sampling the data into training and validation sets

```
# Split the data to training and validation sets:
set.seed(123)
train_ind <- createDataPartition(rem_dupli$top, p = 0.75,list = FALSE)
train_set <- rem_dupli[train_ind,]
vali_set <- rem_dupli[-train_ind,]
```

We start with setting the seed in order to start with the same sample every run.

Then, we split the data into:

Training set with **75%** of the data.

Validation set with **25%** of the data.

Note: We split based on the positive values in order to have them well distributed into the two sets.

Start modeling and check the Pseudo R square

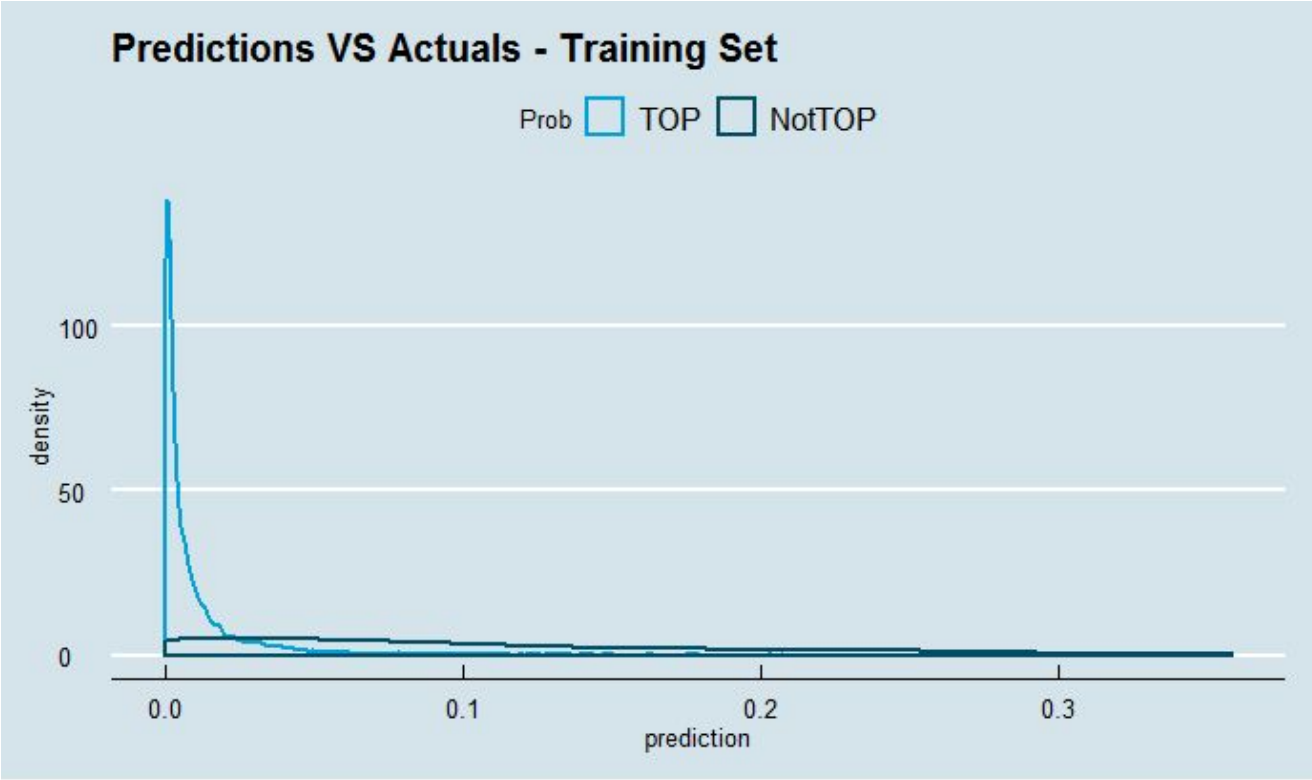
```r
# Logistic Regression modeling:
glm_fit <- glm(top ~ song_popularity + audio_mode + loudness+
                    liveness + energy + danceability +acousticness +
                    song_duration_ms + instrumentalness + key +
                    speechiness + tempo + time_signature +
                    audio_valence ,data = train_set,family = binomial("logit") )

# Check the Psuedo R and cooeffecents impact on the model:
summ_gfit <- summary(glm_fit)
list(summ_gfit$coefficients,round(1-(summ_gfit$deviance / summ_gfit$null.deviance),2))
```

```
[[1]]
                      Estimate   Std. Error     z value     Pr(>|z|)
(Intercept)      -1.047323e+01 2.471975e+00 -4.23678640 2.267416e-05
song_popularity   8.600091e-02 7.792904e-03 11.03579770 2.567628e-28
audio_mode        1.877855e-02 2.032973e-01  0.09236989 9.264042e-01
loudness          1.544610e-01 5.836308e-02  2.64655355 8.131663e-03
liveness         -1.162425e+00 8.541845e-01 -1.36085902 1.735582e-01
energy           -1.089891e+00 8.564242e-01 -1.27260689 2.031576e-01
danceability      2.306196e+00 8.037888e-01  2.86915635 4.115683e-03
song_duration_ms -2.641940e-06 2.257411e-06 -1.17034066 2.418639e-01
instrumentalness -1.033716e+00 1.047455e+00 -0.98688310 3.236999e-01
key               1.622213e-02 2.700201e-02  0.60077470 5.479901e-01
speechiness       3.001597e-01 9.856541e-01  0.30452843 7.607253e-01
tempo            -2.162501e-03 3.736636e-03 -0.57872945 5.627717e-01
time_signature    5.572984e-01 5.128489e-01  1.08667178 2.771819e-01
audio_valence    -7.256948e-01 4.771013e-01 -1.52104952 1.282474e-01

[[2]]
[1] 0.21
```

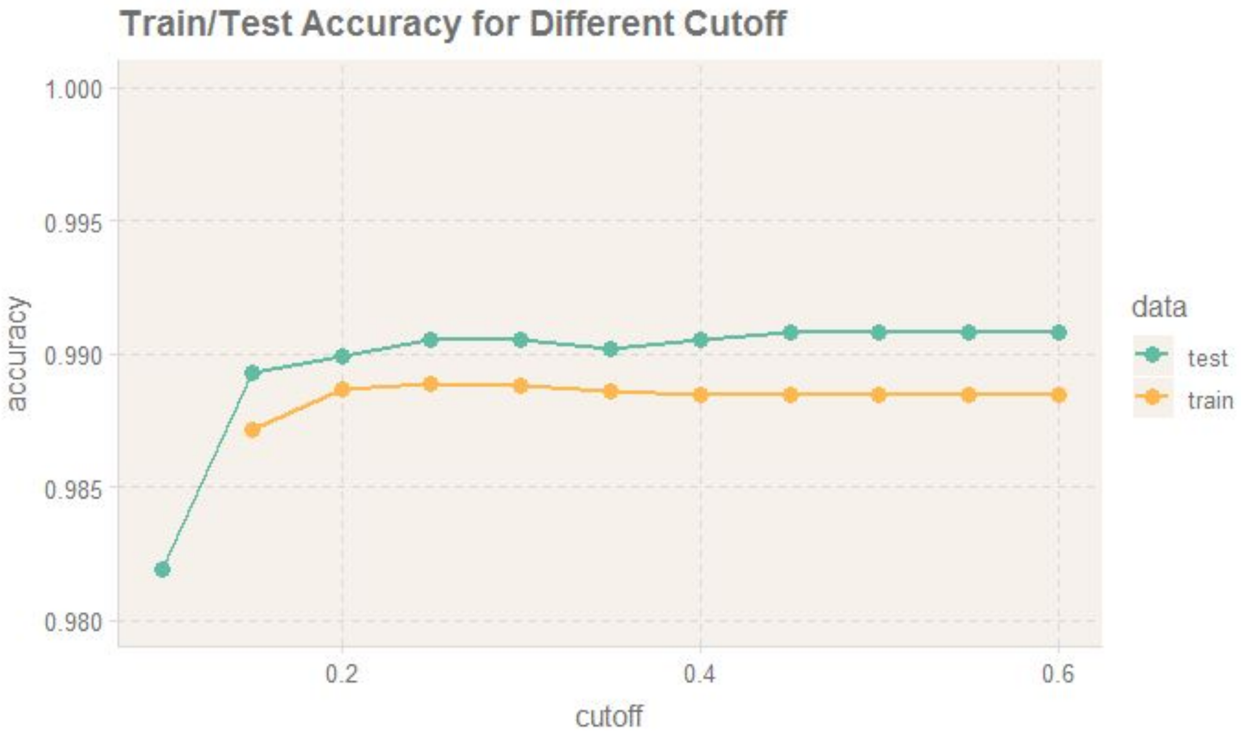
- We choose the uncorrelated independent variables as indicated in the above photo.
- After that we list the **impact of every independent variable** on the model as indicated in the second photo.
- Lastly, we checked the **Pseudo R square** which will explain **how much variability** is explained with our model which is so little; **21%.**

We conclude from here that the model is not stable, but will continue any way.

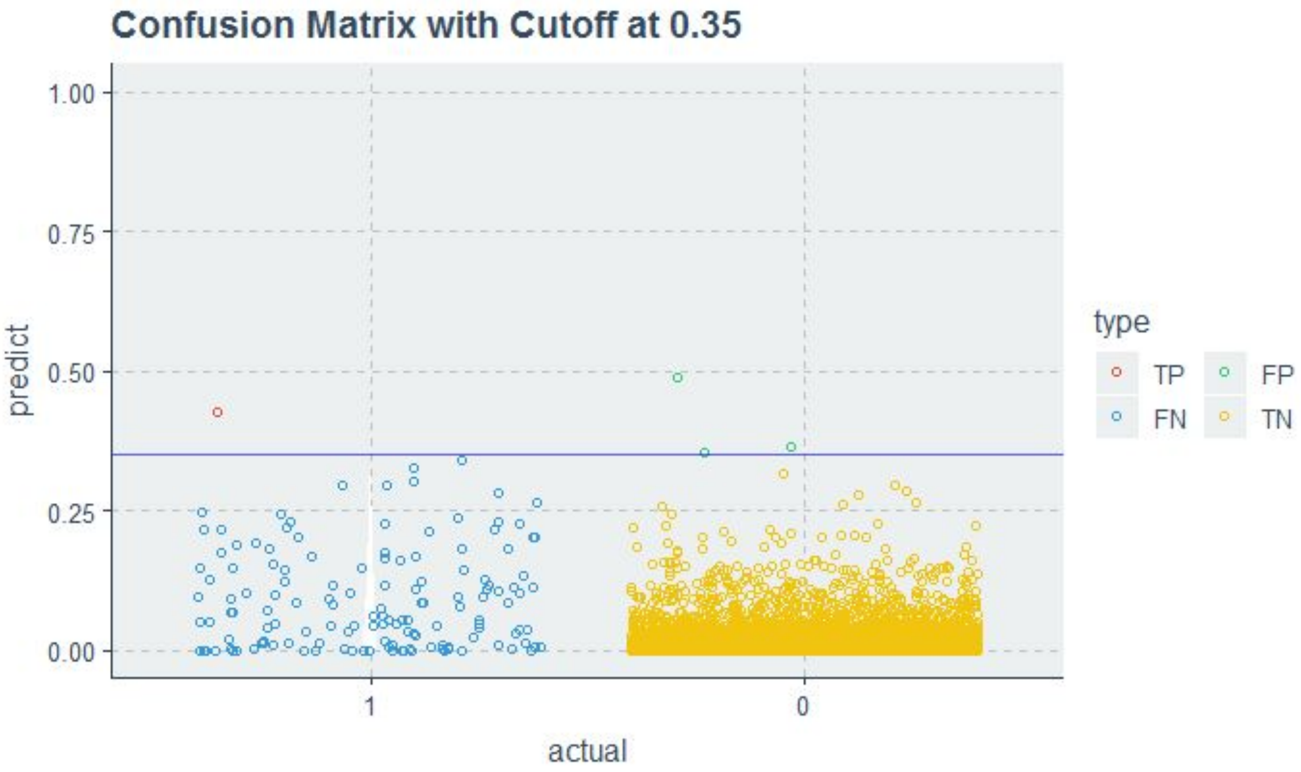
Skewed double density plot indicate that accuracy is not the best way to judge this model



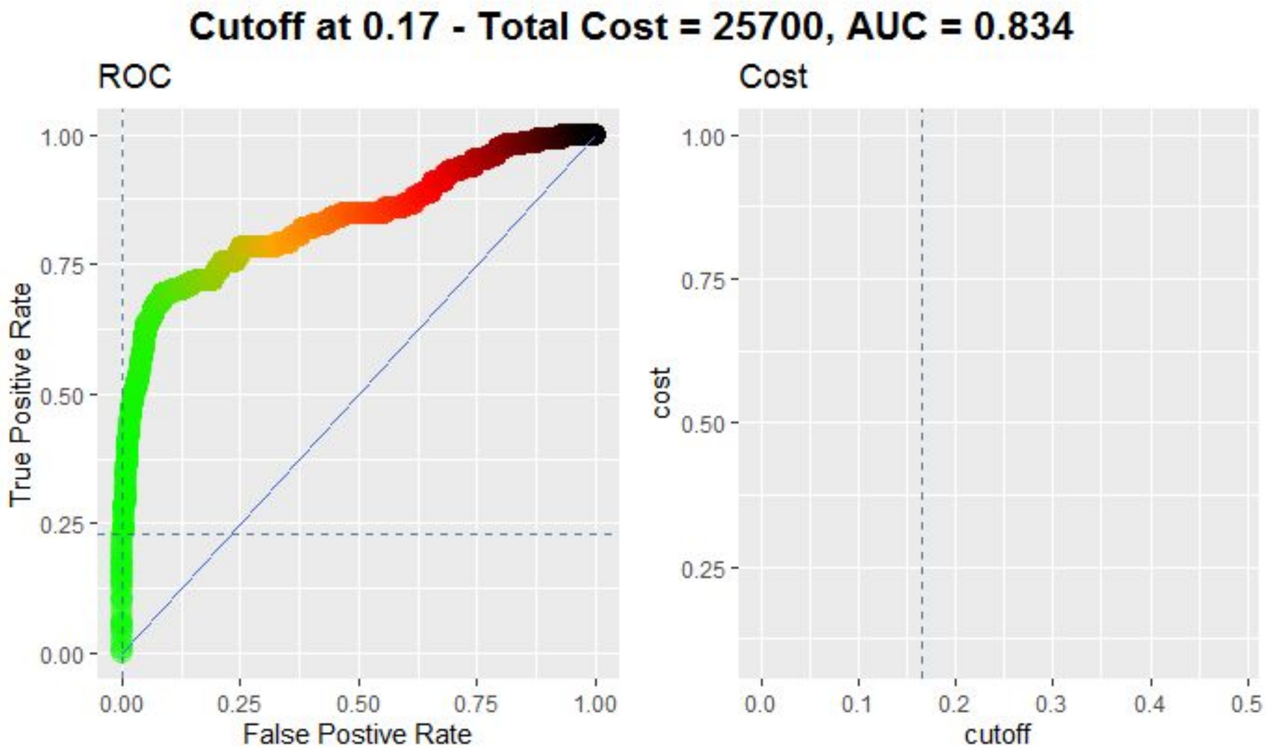
We can find that the best one 0.35



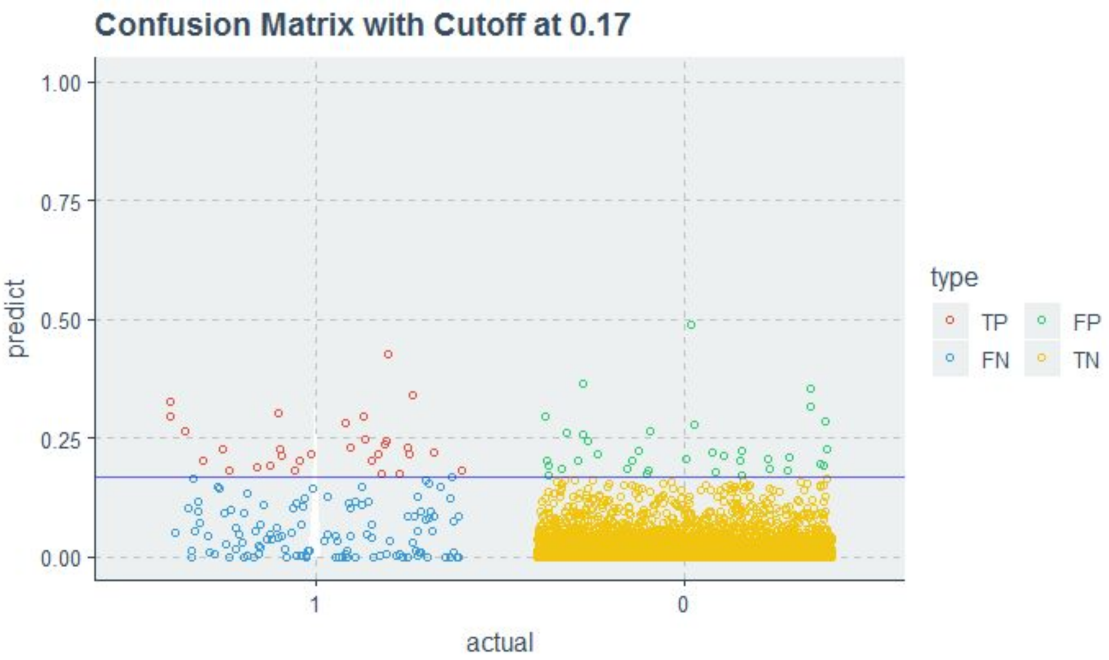
In majority class problems we will find so much FN comparing to TP.



Assigning cost for FP and FN will help getting the best cutoff to improve the model.



Finally we could improve the number of TP at cutoff of 0.17



Logistics Regression insights and conclusions

Technical insights:

As we saw in the previous slides that we are facing Majority class challenge in our data.

Based on that we figured out that:

- **Accuracy** is not the best indicator to use when judging datasets with **majority class issue**.
- **Using LOGIT** to model the majority class datasets will return **many FN** and **less TP.**

Hence,

Logistic regression is not the best model in case of Majority class datasets, maybe decision trees is a better.

Business recommendations:

Spotify can add “TOP meter” to the songs’ producers’ platform (“Supplier side”) in order to help them understand how they can improve their songs’ ranking year over year.

On another hand, **Spotify** can some sort of early predictions on the top 100 list which will help in:

- Contracts negotiations.
- Advertisements planning.

Popularity meter

Popularity

Goals and steps

Opportunity description:

Every single song on Spotify has the following:

- Music features (Danceability, Valence , etc).
- Popularity score (1:100) [1 is the least popular].

Business need:

Business can add “popularity meter” feature to suppliers’ platform in order to follow up on their songs performance.

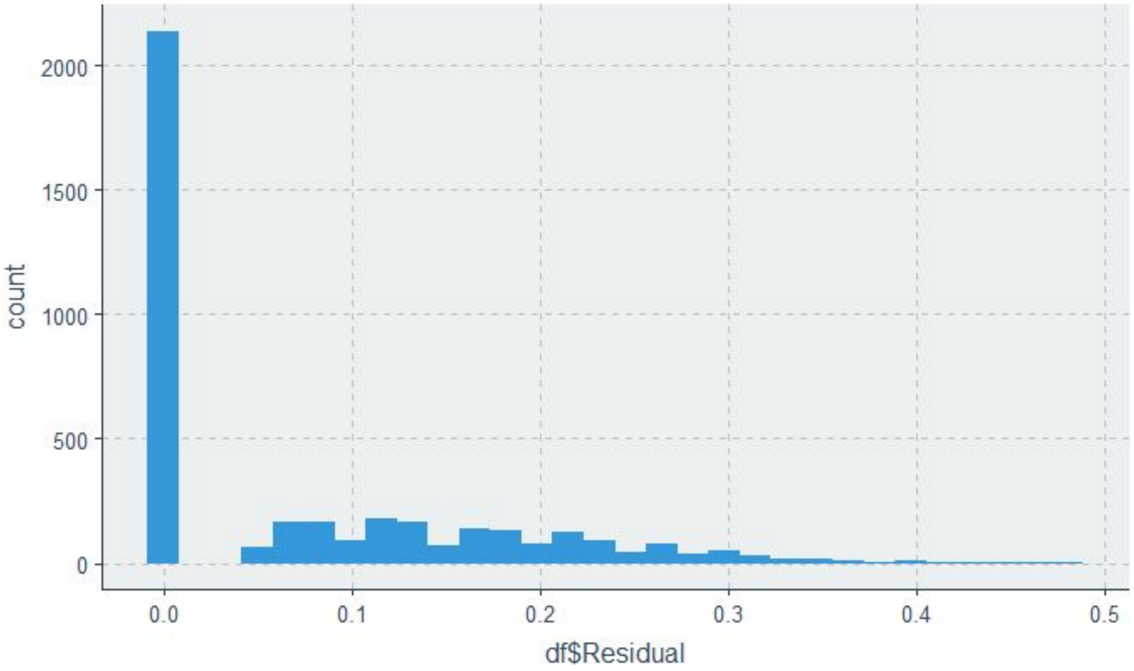
Technical approach:

As we know that the popularity variable ranges from 1 to 100, hence:

We used linear regression modeling over 14 musical features.

Over the upcoming slides we will remove not useful variable.

After modeling we could reach over 60% accuracy

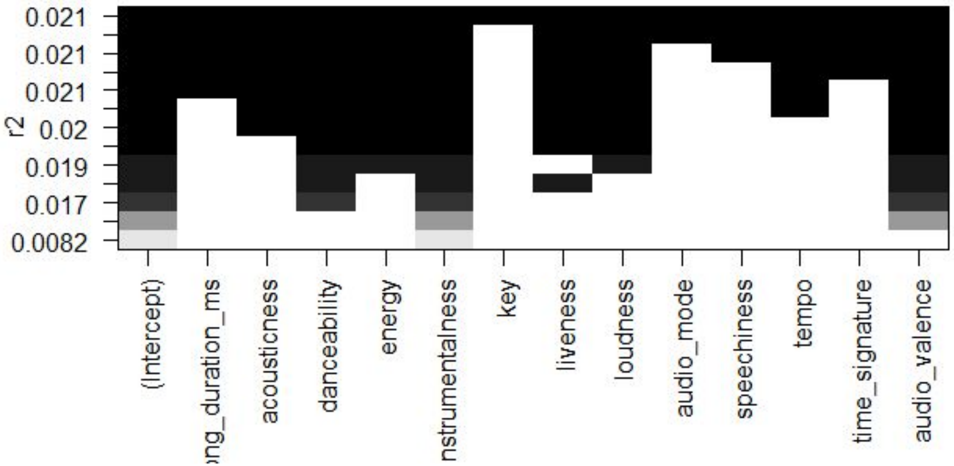


Testing the model and removing unnecessary variables would improve the model.

```
Call:
lm(formula = song_popularity ~ audio_mode + loudness + liveness +
    energy + danceability + acousticness + instrumentalness +
    tempo + audio_valence, data = train_set_lr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57316 -0.11032  0.02905  0.14276  0.49777

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.438842   0.025892  16.949  < 2e-16 ***
audio_mode        0.003299   0.004345   0.759  0.44773
loudness          0.151350   0.035522   4.261 2.06e-05 ***
liveness         -0.037853   0.014352  -2.638  0.00836 **
energy           -0.073506   0.018196  -4.040 5.40e-05 ***
danceability      0.065305   0.014957   4.366 1.28e-05 ***
acousticness     -0.023484   0.009764  -2.405  0.01619 *
instrumentalness -0.064571   0.009419  -6.856 7.56e-12 ***
tempo            -0.036927   0.017867  -2.067  0.03878 *
audio_valence    -0.053417   0.009696  -5.509 3.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1989 on 9141 degrees of freedom
Multiple R-squared:  0.02072,	Adjusted R-squared:  0.01975
F-statistic: 21.49 on 9 and 9141 DF,  p-value: < 2.2e-16
```

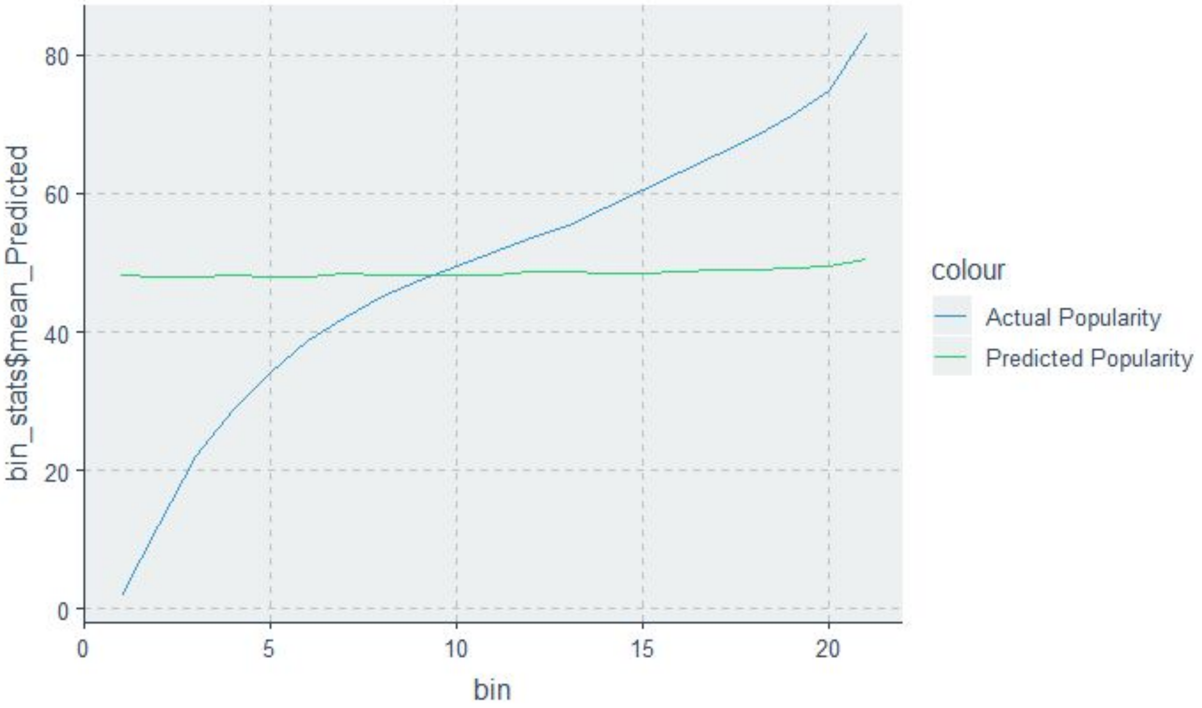
We started with **14 variables** in the model but ended up using **only 9** most important with **R-square** **= 20.7**

After searching for the best 4 or 5 combinations of predictors we reached this result



Now, we can use only 4 variable to deliver the same or better accuracy for the model.

The accuracy is more than 60% but we can improve it by deep diving in feature selection.



Evaluate Model Performance

Logistic Regression: We managed to get better accuracy through choosing the best cutoff value. In the end, we concluded that logistic regression is not the most efficient in a dataset with majority class issues.

Linear Regression: For this type of dataset, a linear regression model would result in the most accuracy. However, in the future, it would be essential to dive in deeper and implement more feature selection.

Recommendations

1. Using the popularity meter, Universal Music Group implement the meter to improve the ranking of the charts
2. The Top Tester will help UMG to evaluate their song portfolio in order to know which songs will break the Top 100 charts and try to improve the features as much as they can
3. When producing and signing new artists, UMG can help structure their songs strategically using predictive analytics to boost the success of the track

References

- [Rcode].
- [Datasets].
- [Research paper].

Thanks....