SPOTIFY – SONG HIT PREDICTION

TEAM MEMBERS

- Indu Chahal
- Tushar Maheshwari
- Uday Kumar
- Venkatesh Gundu
- Zeba Anjum

MENTOR

Mr. Arpit Sharma

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

PROBLEM STATEMENT

THE MUSIC INDUSTRY IS BEST DEFINED AS A FLAWED WORK OF ART. INSTRUMENTAL AND VOCAL SOUND EXPRESSION IS PRIMARILY AN EXPLORATION PROCESS WITH NO RIGHT OR INCORRECT ANSWERS. THE LACK OF A CLEAR 'CORRECT' FRUSTRATES ALL PLAYERS IN THE INDUSTRY, INCLUDING ARTISTS, FANS, MANAGERS, AND RECORD LABEL EXECUTIVES.

WHAT IF YOU COULD TELL IF A SONG WILL BE A HIT BEFORE IT WAS EVEN RELEASED? ARTISTS WOULD REFRAIN FROM RELEASING SONGS THAT WERE NOT ECONOMICALLY SUCCESSFUL, AND MONEY WOULD BE USED MORE EFFICIENTLY IN THE PRODUCTION PROCESS.

AS A RESULT, THIS RESEARCH WILL ASSIST MUSICIANS AND RECORD LABELS. IT WILL NOT ONLY ASSIST IN DETERMINING HOW TO BEST GENERATE SONGS IN ORDER TO MAXIMIZE THEIR PROFITS, BUT IT WILL ALSO ASSIST IN DETERMINING HOW TO BEST PRODUCE SONGS IN ORDER TO MAXIMIZE THEIR CHANCES OF CREATING A HIT TRACK.

FURTHERMORE, IT WOULD HELP ARTISTS AND MUSIC LABELS DETERMINE WHICH SONGS ARE UNLIKELY TO BECOME BILLBOARD HOT 100 HITS.

WITH THIS STUDY WE AIM TO ACHIEVE THE FOLLOWING:

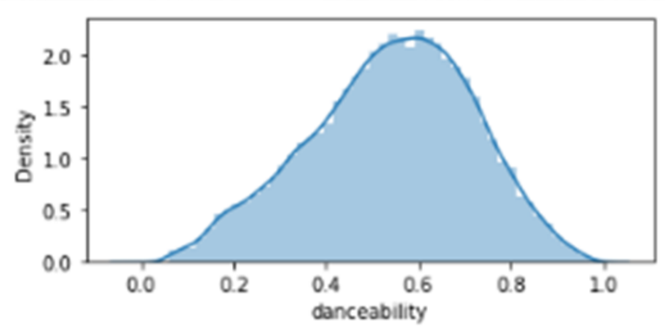
1.PREDICT IF THE SONGS WILL BE A HIT OR NOT.
2.WHAT SONGS WITH WHICH ARTIST ARE GETTING MORE HITS?
3.WHAT TYPE OF BEATS ARE IN THE HIT LIST.
4.WHAT KIND OF SONGS ARE POPULAR WITH RESPECT TO LYRICS?

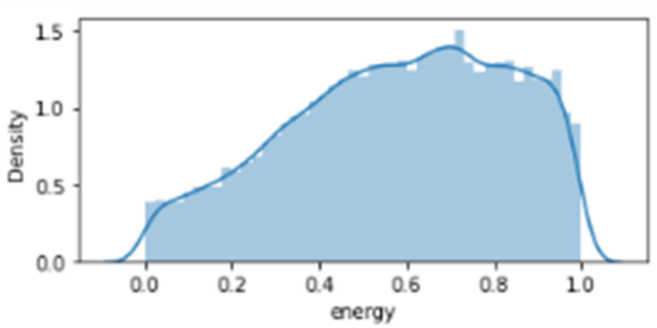
SR.NO	NAME	DATA TYPE
1	TRACK	OBJECT
2	ARTIST	OBJECT
3	DECADE	INT64
4	URI	OBJECT
5	DANCEABILITY	FLOAT64
6	ENERGY	FLOAT64
7	KEY	INT64
8	LOUDNESS	FLOAT64
9	MODE	INT64
10	SPEECHINESS	FLOAT64
11	ACOUSTICNESS	FLOAT64
12	INSTRUMENTALNESS	FLOAT64
13	LIVENESS	FLOAT64
14	VALENCE	FLOAT64
15	TEMPO	FLOAT64
16	DURATION_MS	INT64
17	TIME_SIGNATURE	INT64
18	CHORUS_HIT	FLOAT64
19	SECTIONS	INT64
20	TARGET	INT64

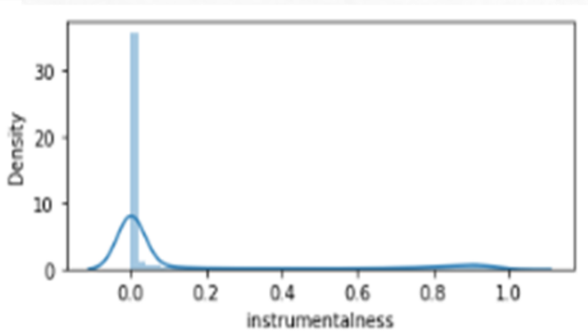
DATA INFORMATION

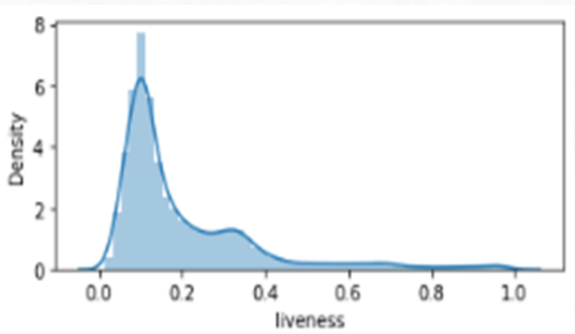
- FEATURES - 20
- ROWS - 41,106
- CATEGORICAL FEATURES - 8
- NUMERIC FEATURES - 12

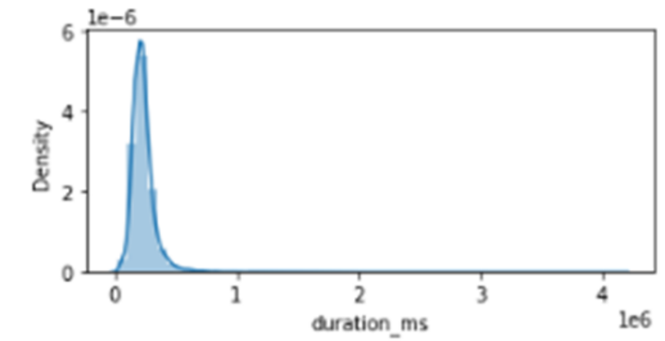
DATA PREPROCESSING

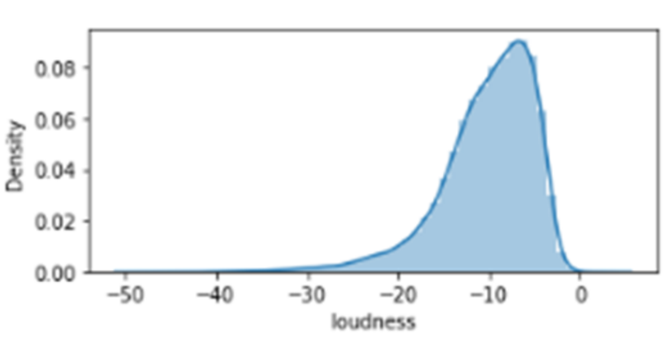


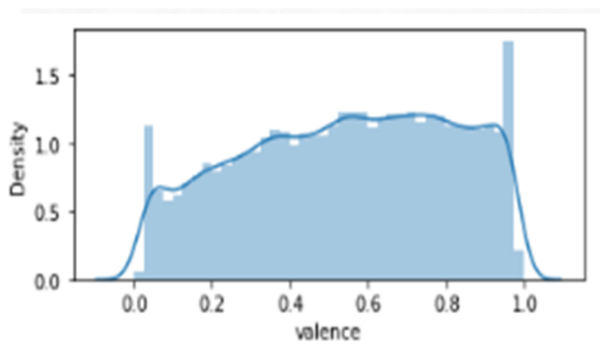


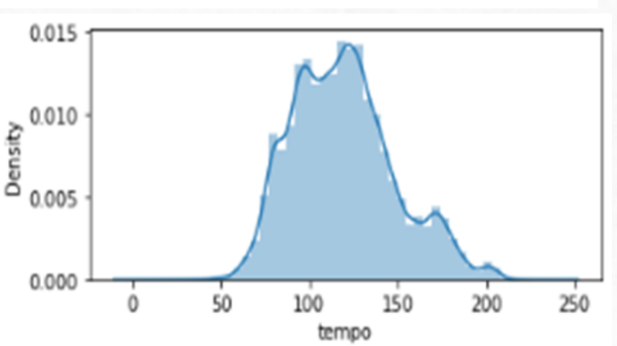


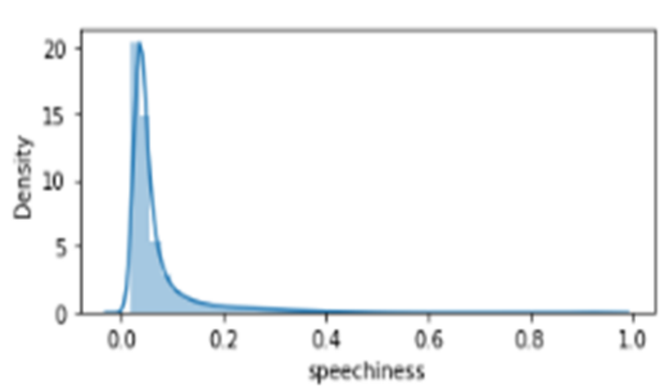


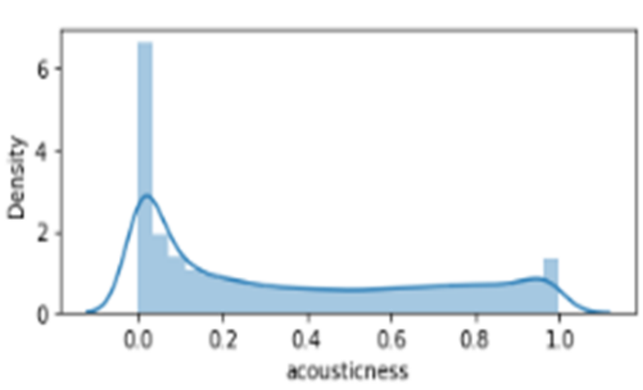


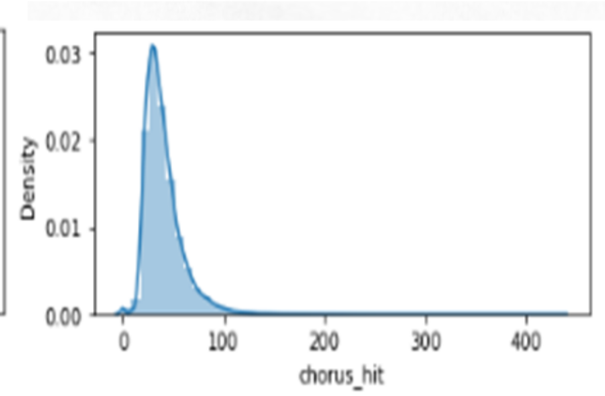


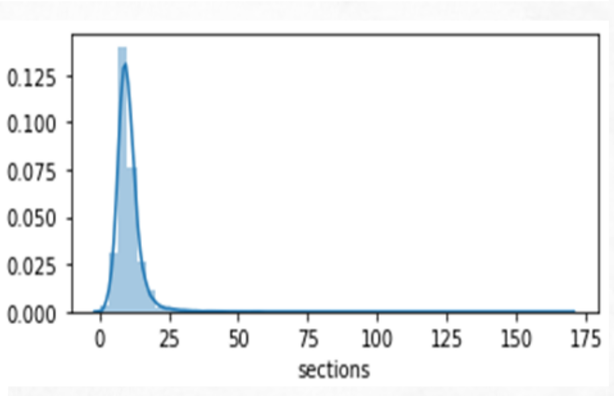




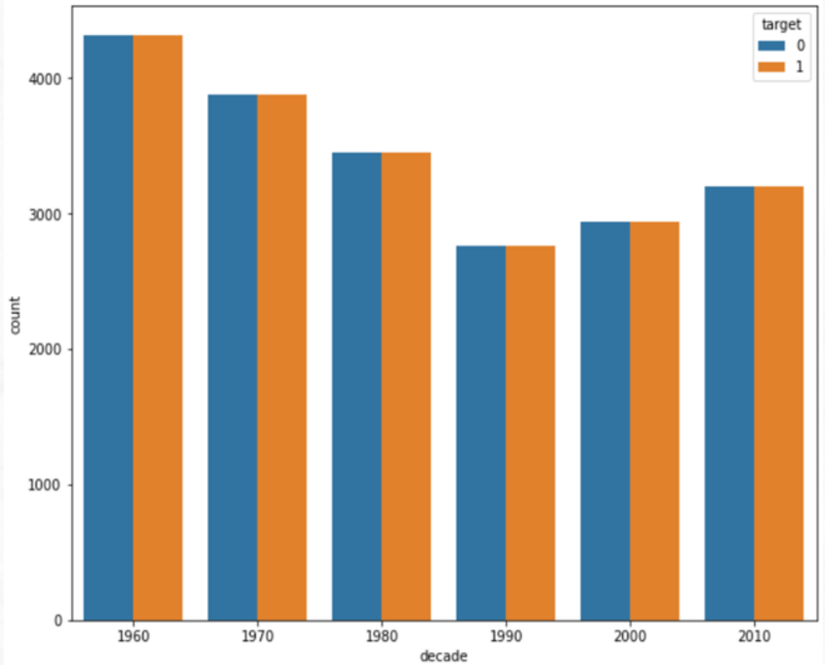


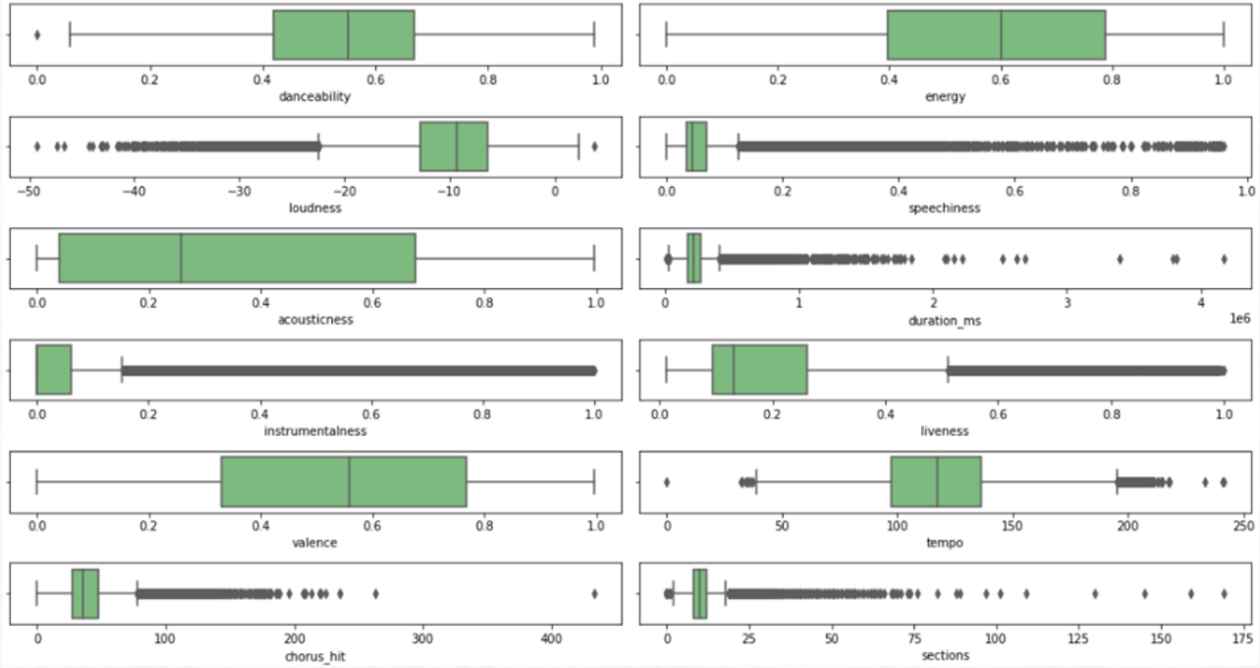


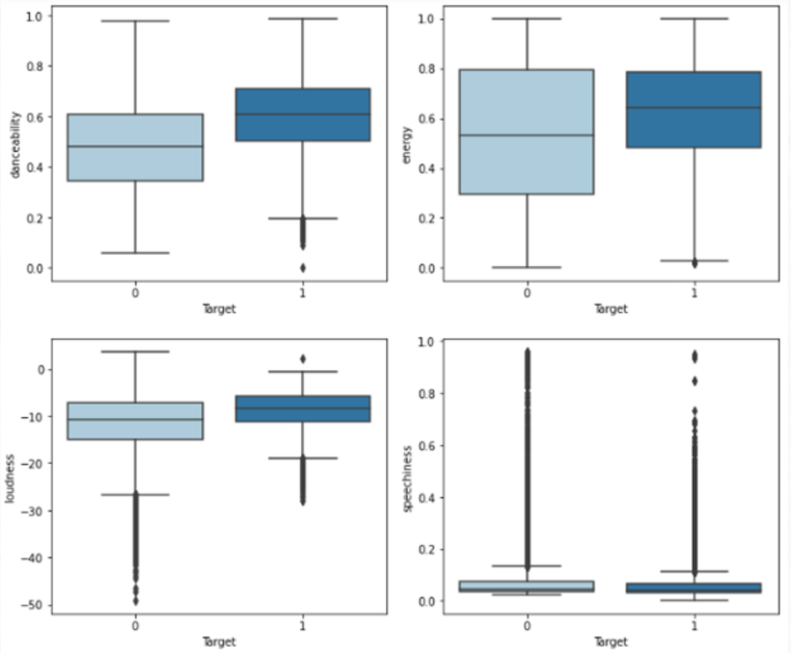


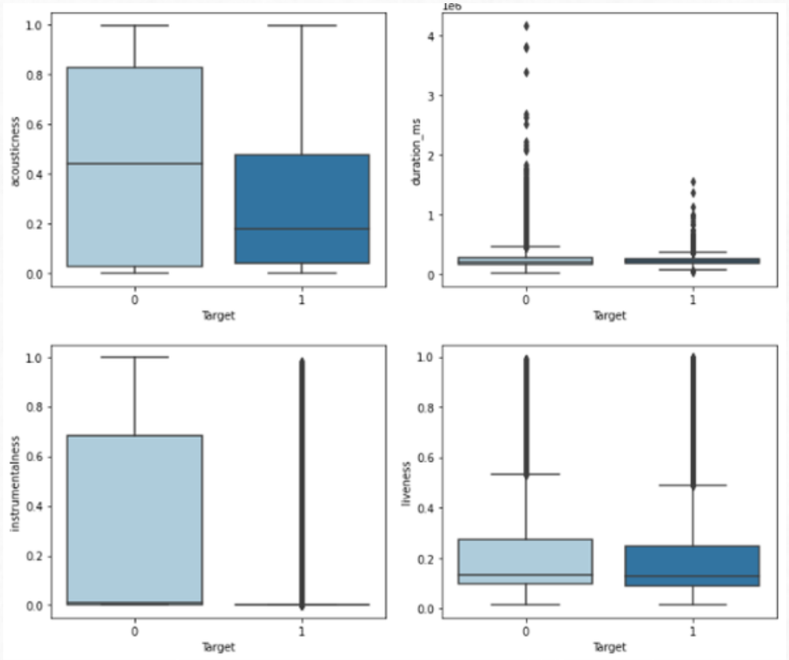


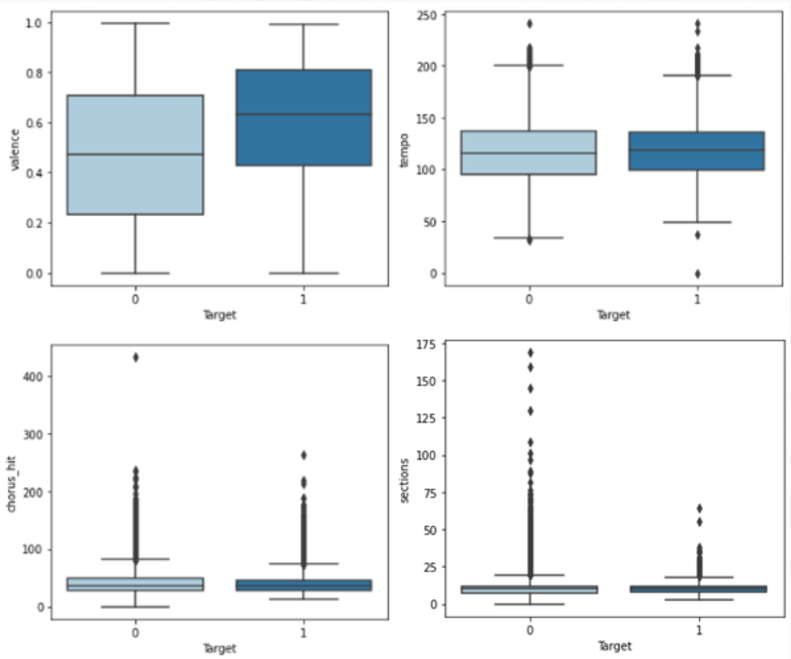
EXPLORATORY DATA ANALYSIS (EDA)











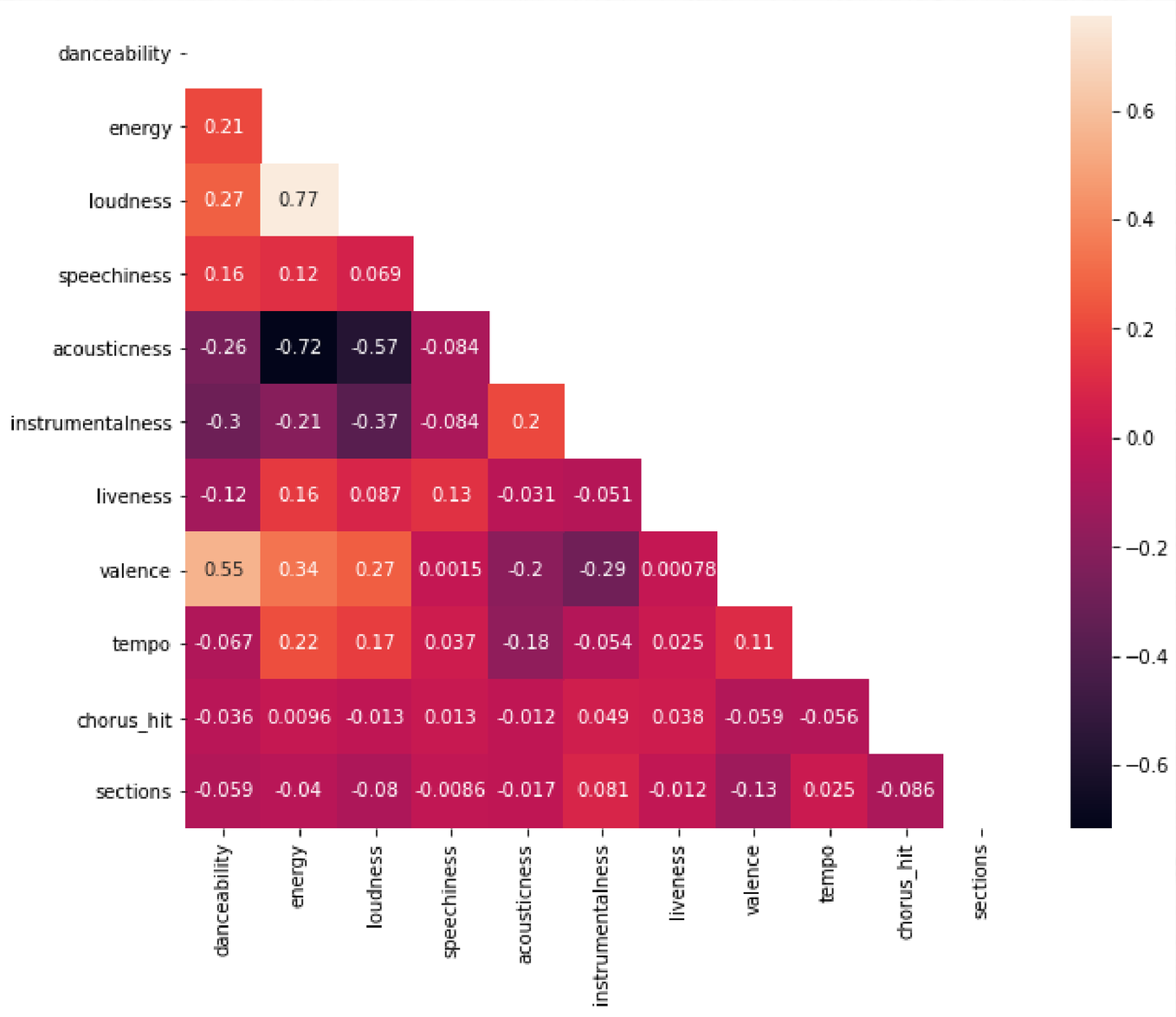
• FROM THE GIVEN HEAT-MAP ITS CLEAR THAT THERE EXISTS MULTICOLLINEARITY AMONG THE FEATURES.

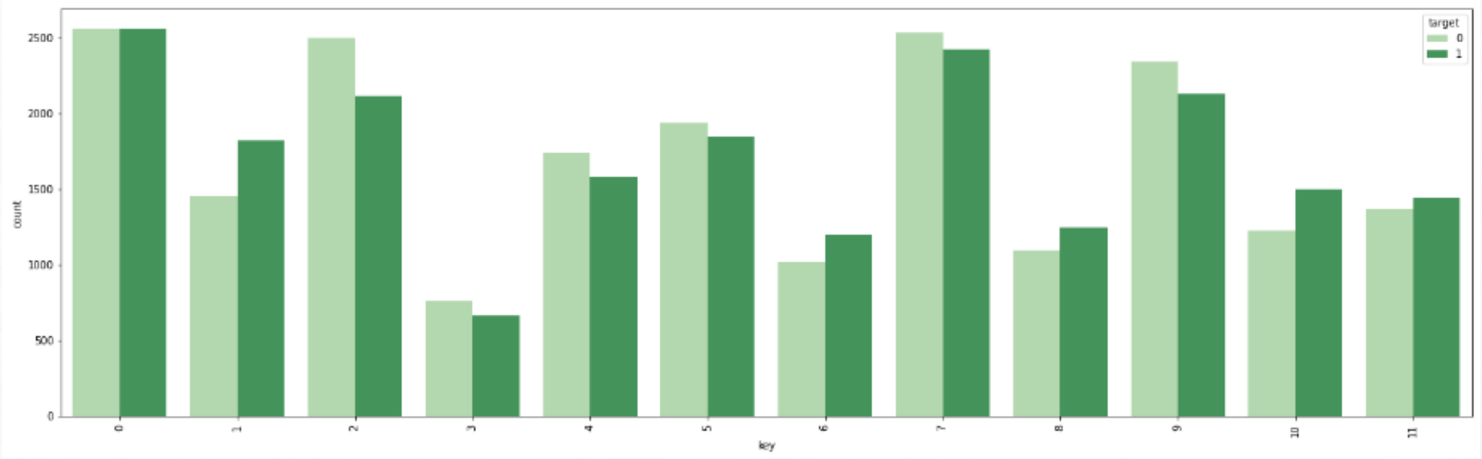
•DANCEABILITY HAS A HIGH CORRELATION WITH VALENCE

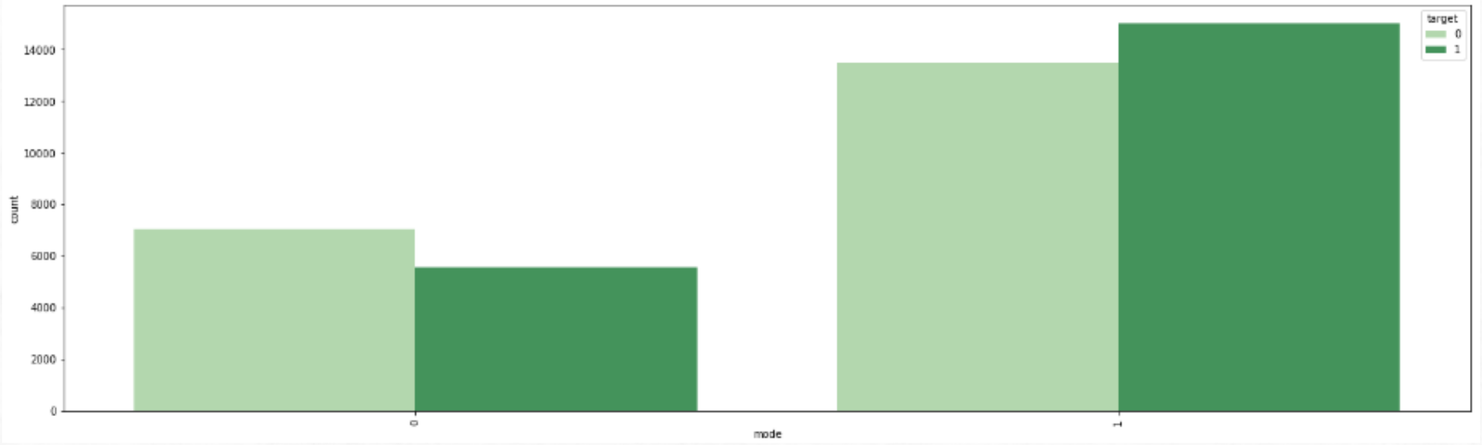
•ENERGY HAS A VERY HIGH CORRELATION WITH LOUDNESS

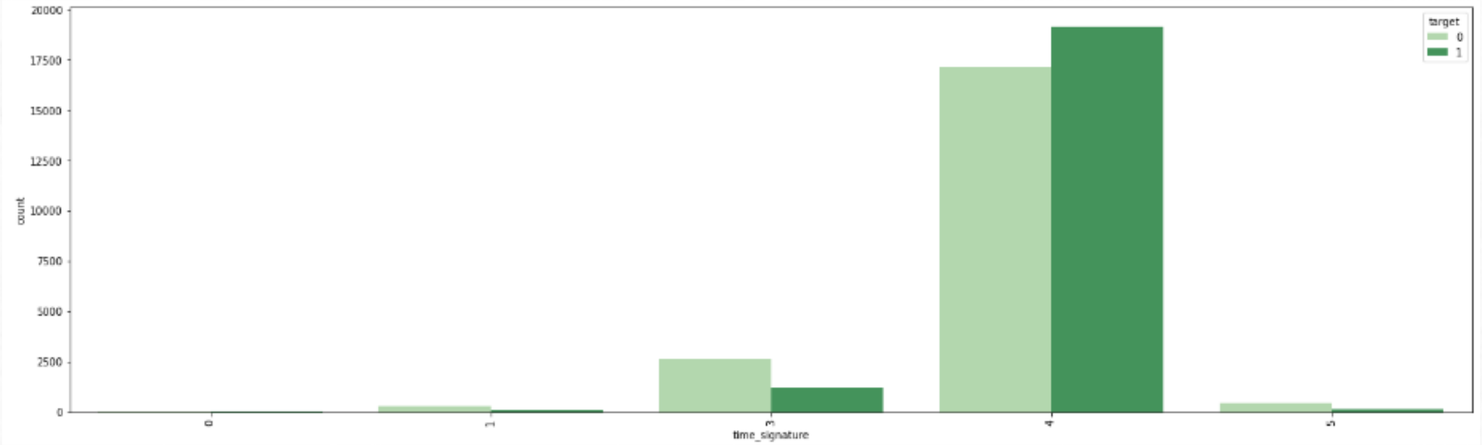
•ACCOUSTICNESS HAS A HIGH NEGATIVE CORRELATION WITH LOUDNESS AND ENERGY

• DUE TO THIS REASON A PCA MODEL COULD PROVIDE A MORE ACCURATE PREDICTIONS.









•HERE KEY REFERS TO THE VARIOUS KEY NOTES USED IN THE SONG . IT CAN BE SEEN THAN THE KEY-NOTES 1,8,10 AND 11 HAVE MUCH IMPACT IN DETERMINING A HIT SONG.

•THE MODE FEATURE REFERS TO SCALE OF THE DATA(MAJOR/MINOR). IT CAN BE INFERRED THAN MODE WHEN 1 IE MAJOR MODE CAN BE CONSIDERED AS ONE OF THE FACTORS IN PRODUCING A HIT SONG.

•THE TIME_SIGNATURE FEATURE,INDICATES HOW MANY BEATS ARE THERE IN EACH MEASURE OF A PIECE OF MUSIC.ALTHOUGH, TIME_SIGNATURE OF MEASURE 4 IS RELATIVELY HIGH ,BUT HIT SONGS IN GENERAL,SEEM TO HAVE HIGH TIME_SIGNATURE OF 4 THAN NON-HIT SONGS.

•OUR TARGET VARIABLE IS CATEGORICAL, AND IT CAN BE SEEN THAT IT IS EQUALLY WELL BALANCED WITH NUMBER OF HITS AND NON-HITS

.

STATISTICAL TESTS

Chi2_Contingency

	features	pvalue
0	decade	1.000000e+00
1	key	1.547053e-27
2	mode	1.306112e-58
3	time_signature	1.063178e-212

Skewness

danceability	-0.251762
energy	-0.320168
loudness	-1.415109
speechiness	4.573376
acousticness	0.493360
instrumentalness	1.745277
liveness	2.123818
valence	-0.179745
tempo	0.485278
chorus_hit	2.215338
sections	6.053588

Kruskal-Wallis Test

	K.Features	pvalue
0	danceability	0.000000e+00
1	energy	1.013489e-198
2	loudness	0.000000e+00
3	speechiness	1.598041e-153
4	acousticness	2.450126e-254
5	instrumentalness	0.000000e+00
6	liveness	5.740819e-29
7	valence	0.000000e+00
8	tempo	7.827986e-15
9	duration_ms	1.303689e-05
10	chorus_hit	5.228208e-11
11	sections	5.221666e-06

MODEL BUILDING

- WE CREATED DIFFERENT MODELS TO IMPROVE THE DIFFERENT METRICS THAT HELP DETERMINE THE MOST EFFICIENT MODEL.
- FROM OUR BUSINESS PERSPECTIVE, WE BELIEVE IT IS VITAL TO REDUCE FALSE NEGATIVES & FALSE POSITIVES AND THEREFORE WE USED "ACCURACY" AND "F1-SCORE" AS OUR TARGET METRICS
- WE HAVE CREATED USER-DEFINED FUNCTION PLOT_CONFUSION_MATRIX THAT SHALL CREATE A CONFUSION MATRIX AND A USER-DEFINED FUNCTION PLOT_ROC THAT SHALL PLOT THE AUC SCORE ON A CHART AND DISPLAY ITS VALUE BASED ON THE MODEL AND THE TEST & PREDICTED VALUES
- AFTER EACH MODEL, THE RELEVANT METRICS ARE CALCULATED AND ADDED TO A DATA FRAME NAMED SCORE CARD WHICH CONTAINS THE NAME OF THE MODEL, ACCURACY SCORE, F1-SCORE, AND THE COHEN-KAPPA VALUE.

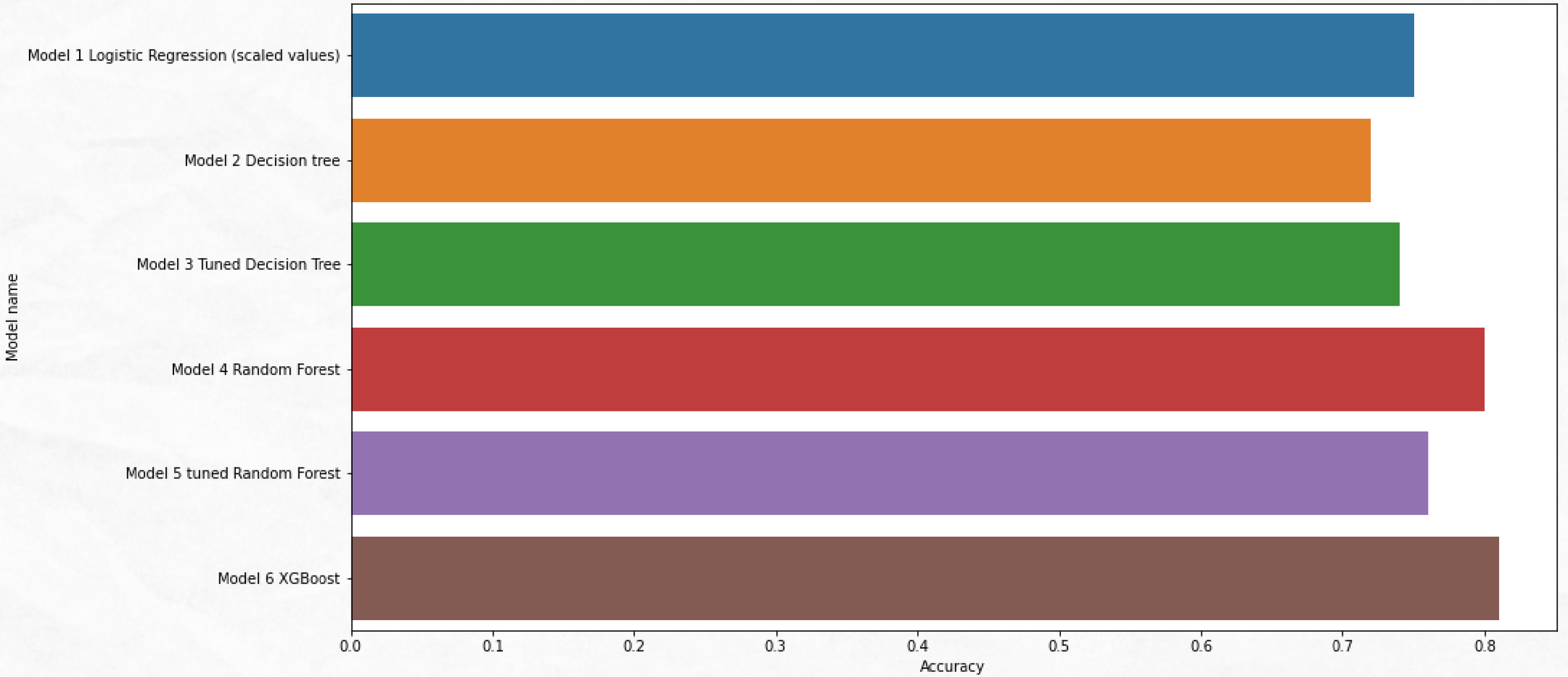
SCORE CARD FOR MODEL COMPARISON

	Model name	Accuracy	F1-Score	Cohen-Kappa
0	Model 1 Logistic Regression (scaled values)	0.75	0.76	0.50
1	Model 2 Decision tree	0.72	0.72	0.43
2	Model 3 Tuned Decision Tree	0.74	0.76	0.48
3	Model 4 Random Forest	0.80	0.81	0.60
4	Model 5 tuned Random Forest	0.76	0.78	0.53
5	Model 6 XGBoost	0.81	0.82	0.62

DESPITE HYPER TUNING THE PARAMETERS OF THE RANDOM FOREST MODEL, THE MODEL HAS NOT GIVEN BETTER RESULTS COMPARED TO THE FIRST RANDOM FOREST MODEL.

THE XGBOOST MODEL PROVIDES BETTER RESULTS WITH AN ACCURACY OF 0.81 AND AN F1-SCORE OF 0.82 BUT THE DATA STILL CONTAINS OUTLIERS AND THIS BOOSTING TECHNIQUE IS SENSITIVE TO OUTLIERS.

CONCLUSION



SINCE THE MAIN REQUIREMENT IS TO MAXIMIZE THE ACCURACY AND F1-SCORE, IT'S IDEAL TO CONSIDER THE RANDOM FOREST BASE MODEL AS FINAL MODEL.

BUSINESS JUSTIFICATION

- BETTER SONGS COULD BE SUGGESTED TO THEIR USERS
- HELP INCREASE THE USE OF THE APP BY ITS CUSTOMERS
- MOTIVATE ITS USERS TO PURCHASE THE SPOTIFY PREMIUM
- POSSIBLE COLLABORATIONS WITH MEMBERS OF THE MUSIC INDUSTRY.

REFERENCES

- HTTPS://WWW.KAGGLE.COM/DATASETS/THEOVERMAN/THE-SPOTIFY-HIT-PREDICTOR-DATASET
- HTTPS://SCHOLAR.SMU.EDU/CGI/VIEWCONTENT.CGI?ARTICLE=1204&CONTEXT=DATASCIENCEREVIEW
- HTTP://CS229.STANFORD.EDU/PROJ2018/REPORT/16.PDF

THANK YOU