CAP5610: Machine Learning Final Project Presentation

Would Your Favorite Song Make it to Spotify's "Top 50 - Global" Playlist?

April 19, 2022

Bailey LaRea Nadine Rose Suma Marri

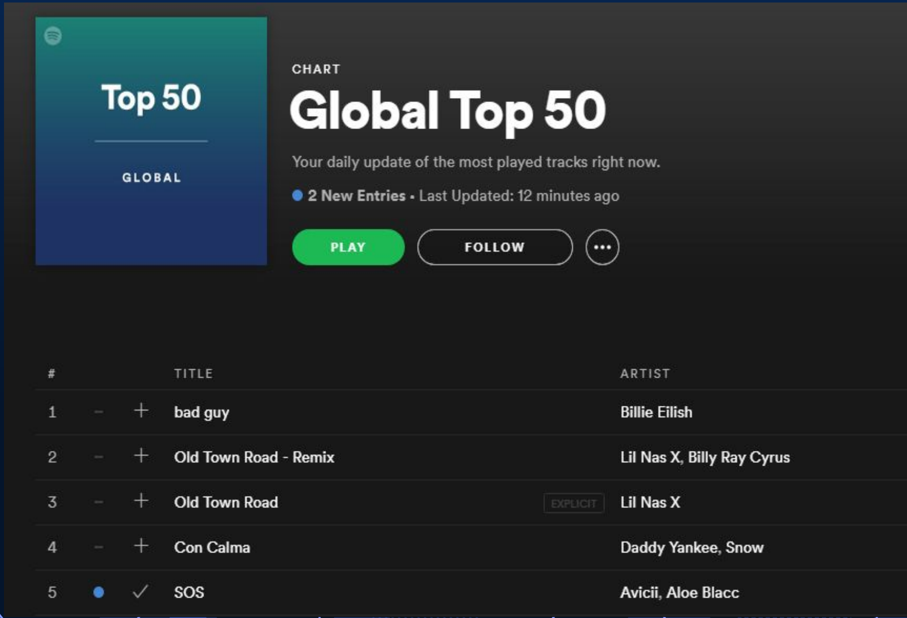
Agenda

Initial Project Questions
Description of the Data
Data Preparation
Visualizations for EDA
Revised Project Question
Baseline Model
Preliminary ML models
Next Steps

Initial Project Questions

Idea: Can we use Machine Learning techniques to help **predict** what songs will make it into Spotify's "Top 50 - Global" playlist?

Can specific track attributes be used to predict overall song popularity?



CHART

Global Top 50

Your daily update of the most played tracks right now.

● 2 New Entries • Last Updated: 12 minutes ago

#			TITLE		ARTIST
1	–	+	bad guy		Billie Eilish
2	–	+	Old Town Road - Remix		Lil Nas X, Billy Ray Cyrus
3	–	+	Old Town Road	EXPLICIT	Lil Nas X
4	–	+	Con Calma		Daddy Yankee, Snow
5	●	✓	SOS		Avicii, Aloe Blacc

Description of the Data

Top 50

CHART

Global Top 50

Your daily update of the most played tracks right now.

2 New Entries • Last Updated: 12 minutes ago

FOLLOWERS
13,618,894

#			TITLE	ARTIST	DAILY PLAYS
1	–	+	bad guy	Billie Eilish	5,961,324
2	–	+	Old Town Road - Remix	Lil Nas X, Billy Ray Cyrus	5,110,233
3	–	+	Old Town Road	Lil Nas X	4,015,651
4	–	+	Con Calma	Daddy Yankee, Snow	3,735,649
5	●	✓	SOS	Avicii, Aloe Blacc	3,501,801

There are 13 Track Attributes

TrackName

ArtistName

Genre

Beats Per Minute

Energy

Danceability

Loudness (dB)

Liveness

Valence

Length

Acousticness

Speechiness

Popularity

The Datasets

Top 50 Songs in 2019

- This dataset contains information for the top 50 songs in 2019.
- 50 rows of data

2019 Songs

- This dataset contains songs that were released in 2019.
- ≈ 11,000 rows of data

Both datasets are sourced from Kaggle

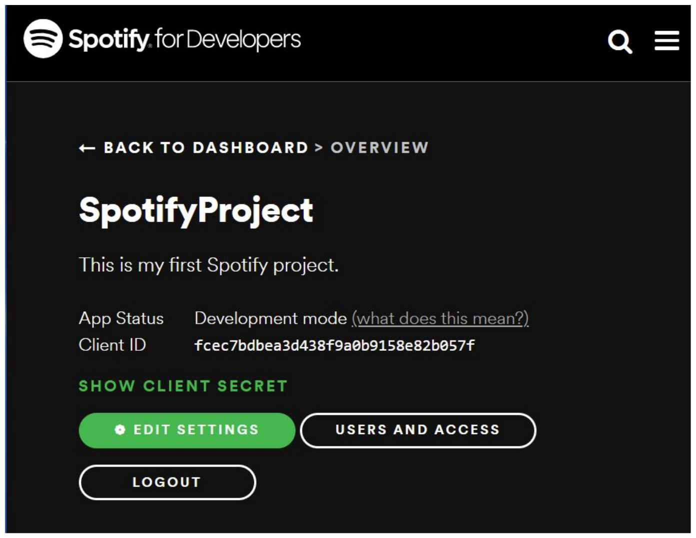
Data Preparation

Kaggle

- We used this for our initial dataset and for the top 50 songs.

Spotify for Developers - API

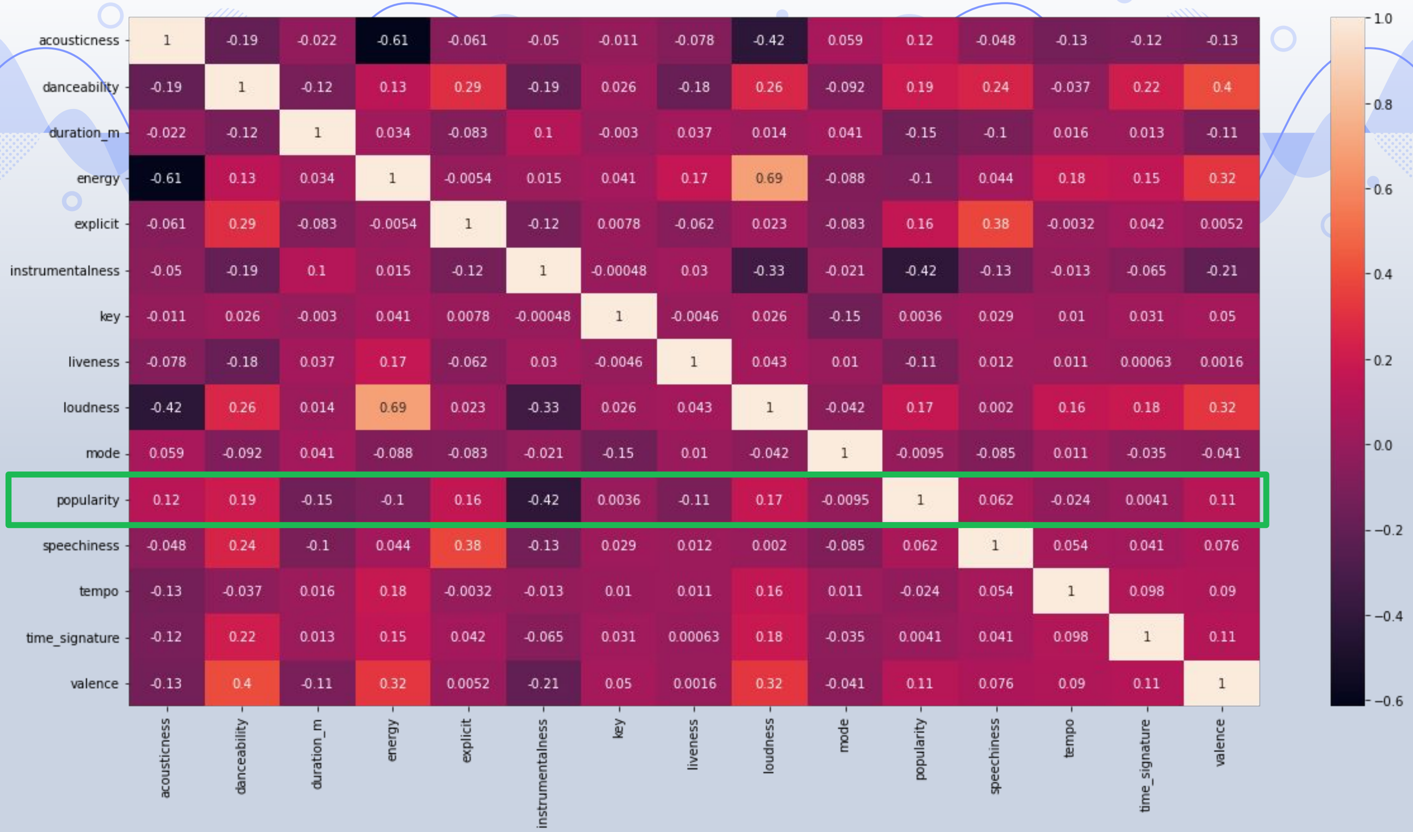
- The Spotify for Developers allowed us to extract and collect the data from Spotify
- We used the Client Keys to access the data on Spotify.
- After extracting the data, we had deleted and duplicate values and any records that had null values.



SpotifyProject

This is my first Spotify project.

App Status Development mode (what does this mean?)
Client ID fcec7bdbea3d438f9a0b9158e82b057f



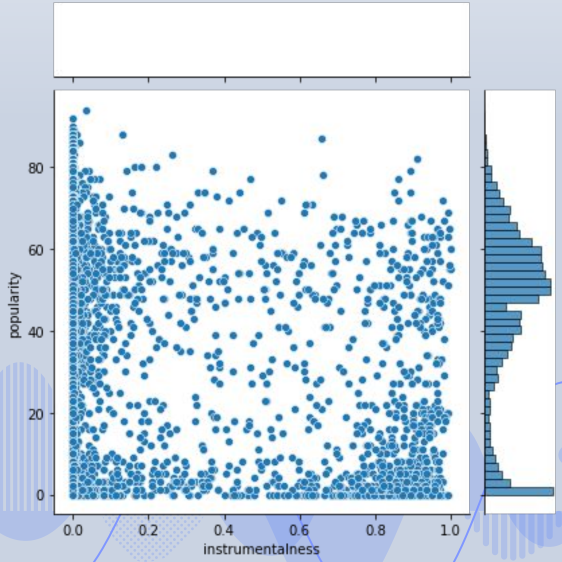
	acousticness	danceability	duration_m	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	time_signature	valence
acousticness	1	-0.19	-0.022	-0.61	-0.061	-0.05	-0.011	-0.078	-0.42	0.059	0.12	-0.048	-0.13	-0.12	-0.13
danceability	-0.19	1	-0.12	0.13	0.29	-0.19	0.026	-0.18	0.26	-0.092	0.19	0.24	-0.037	0.22	0.4
duration_m	-0.022	-0.12	1	0.034	-0.083	0.1	-0.003	0.037	0.014	0.041	-0.15	-0.1	0.016	0.013	-0.11
energy	-0.61	0.13	0.034	1	-0.0054	0.015	0.041	0.17	0.69	-0.088	-0.1	0.044	0.18	0.15	0.32
explicit	-0.061	0.29	-0.083	-0.0054	1	-0.12	0.0078	-0.062	0.023	-0.083	0.16	0.38	-0.0032	0.042	0.0052
instrumentalness	-0.05	-0.19	0.1	0.015	-0.12	1	-0.00048	0.03	-0.33	-0.021	-0.42	-0.13	-0.013	-0.065	-0.21
key	-0.011	0.026	-0.003	0.041	0.0078	-0.00048	1	-0.0046	0.026	-0.15	0.0036	0.029	0.01	0.031	0.05
liveness	-0.078	-0.18	0.037	0.17	-0.062	0.03	-0.0046	1	0.043	0.01	-0.11	0.012	0.011	0.00063	0.0016
loudness	-0.42	0.26	0.014	0.69	0.023	-0.33	0.026	0.043	1	-0.042	0.17	0.002	0.16	0.18	0.32
mode	0.059	-0.092	0.041	-0.088	-0.083	-0.021	-0.15	0.01	-0.042	1	-0.0095	-0.085	0.011	-0.035	-0.041
popularity	0.12	0.19	-0.15	-0.1	0.16	-0.42	0.0036	-0.11	0.17	-0.0095	1	0.062	-0.024	0.0041	0.11
speechiness	-0.048	0.24	-0.1	0.044	0.38	-0.13	0.029	0.012	0.002	-0.085	0.062	1	0.054	0.041	0.076
tempo	-0.13	-0.037	0.016	0.18	-0.0032	-0.013	0.01	0.011	0.16	0.011	-0.024	0.054	1	0.098	0.09
time_signature	-0.12	0.22	0.013	0.15	0.042	-0.065	0.031	0.00063	0.18	-0.035	0.0041	0.041	0.098	1	0.11
valence	-0.13	0.4	-0.11	0.32	0.0052	-0.21	0.05	0.0016	0.32	-0.041	0.11	0.076	0.09	0.11	1

Visualizations - Heatmap

Visualizations - Scatterplots

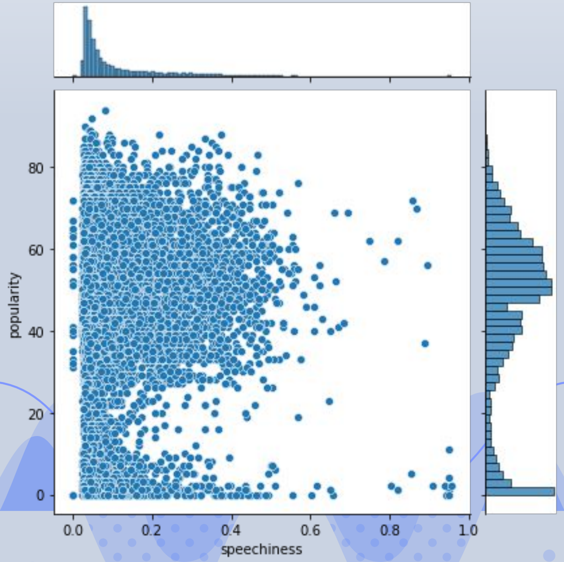
Instrumentalness vs. Popularity

We can see that there is a **negative correlation** between both variables.



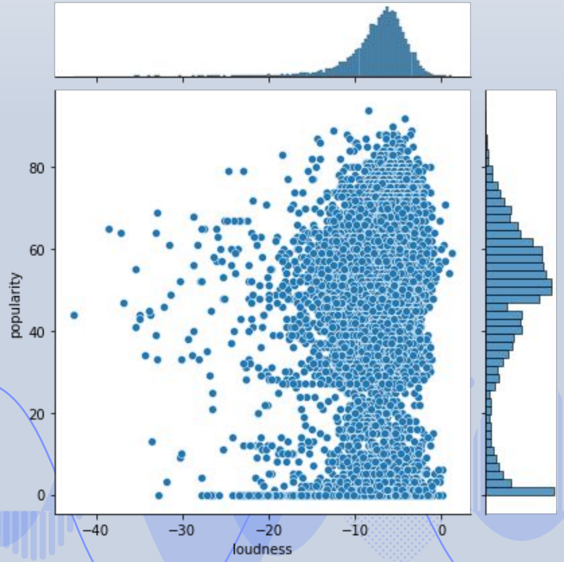
Speechiness vs. Popularity

We can see that there is **no correlation** between both variables.



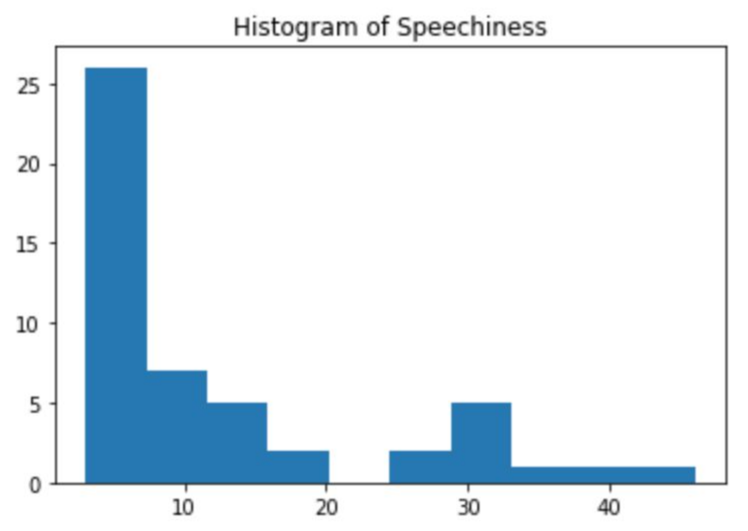
Loudness vs. Popularity

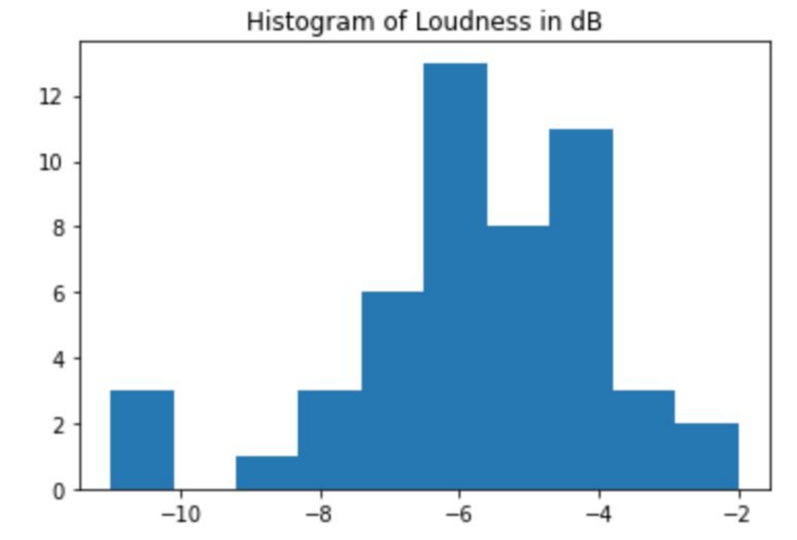
We can see that there is a **positive correlation** between both variables.



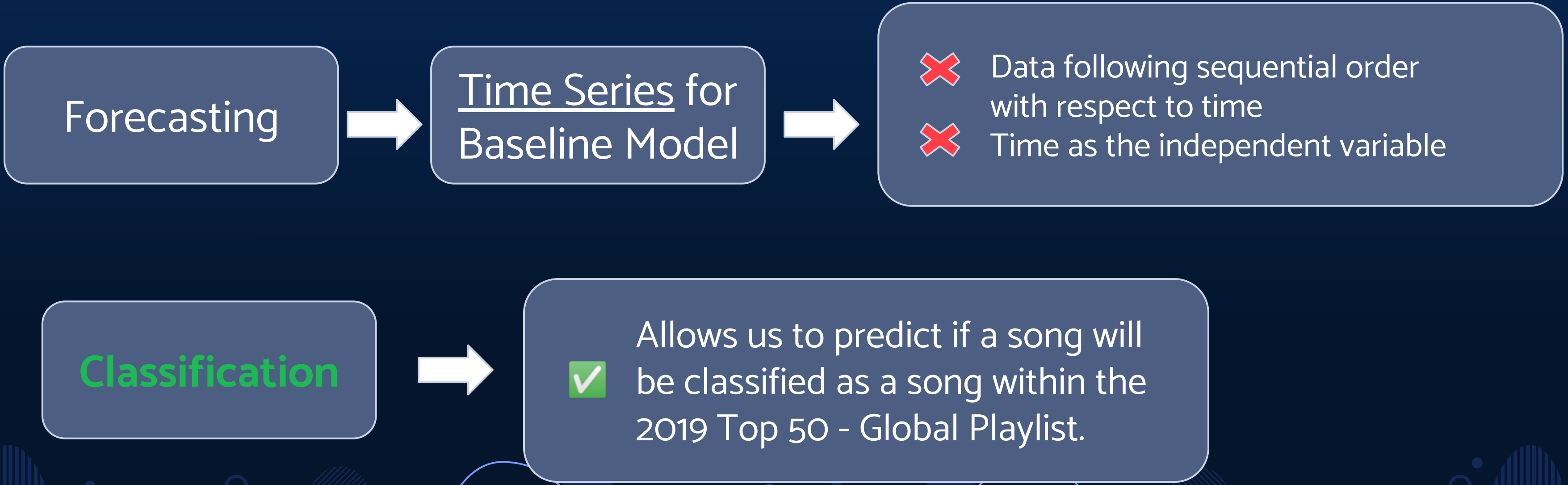
Visualizations - Histogram

Here are the following distributions of Speechiness and Loudness from another dataset. They follow distributions that align with the prior slides.





Revised Project Question



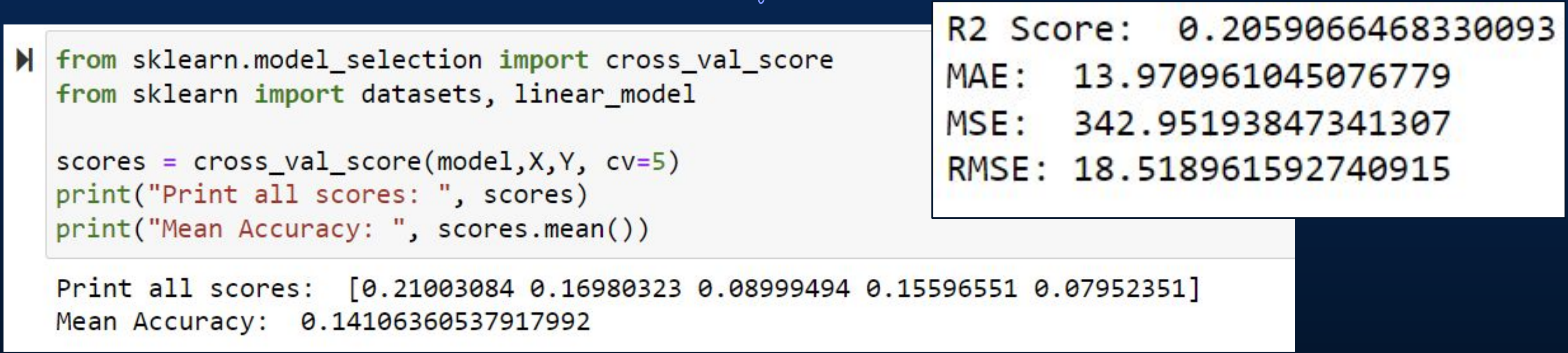
Forecasting → Time Series for Baseline Model →

- ❌ Data following sequential order with respect to time
- ❌ Time as the independent variable

Classification →

✅ Allows us to predict if a song will be classified as a song within the 2019 Top 50 - Global Playlist.

Baseline Model: Linear Regression



```
from sklearn.model_selection import cross_val_score
from sklearn import datasets, linear_model

scores = cross_val_score(model,X,Y, cv=5)
print("Print all scores: ", scores)
print("Mean Accuracy: ", scores.mean())
```

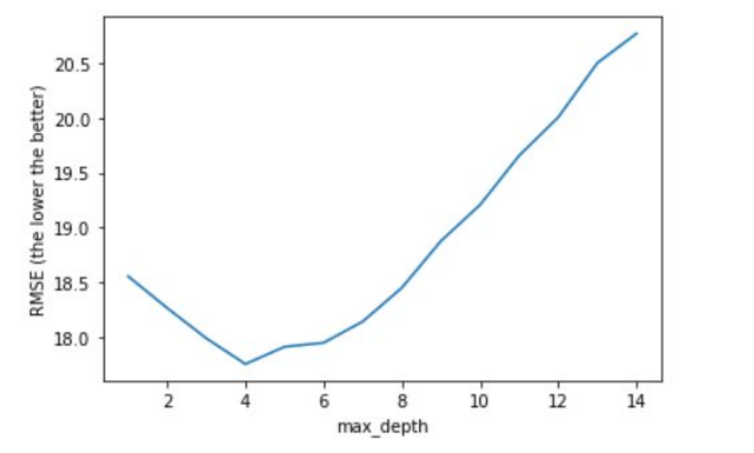
```
Print all scores:  [0.21003084 0.16980323 0.08999494 0.15596551 0.07952351]
Mean Accuracy:  0.14106360537917992
```

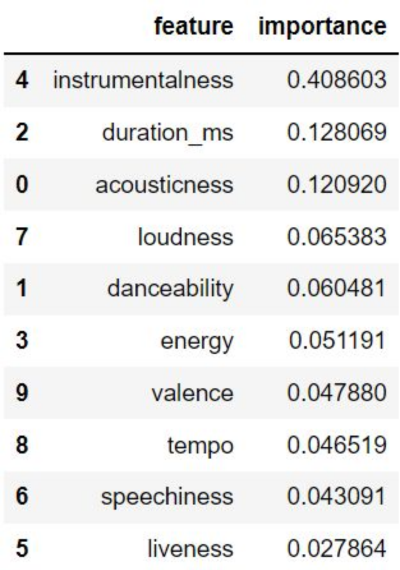
```
R2 Score:  0.2059066468330093
MAE:  13.970961045076779
MSE:  342.95193847341307
RMSE: 18.518961592740915
```

With a mean accuracy score of only ~14%, some changes need to be made!

Preliminary Version of ML Models

Decision Tree - Plot of RMSE





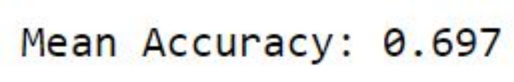
	feature	importance
4	instrumentalness	0.408603
2	duration_ms	0.128069
0	acousticness	0.120920
7	loudness	0.065383
1	danceability	0.060481
3	energy	0.051191
9	valence	0.047880
8	tempo	0.046519
6	speechiness	0.043091
5	liveness	0.027864

There is a little bit of improvement.... However, there must be a **better model** out there?

Random Forest & KNN are also used!

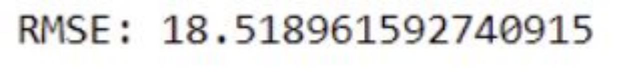
Random Forest: Classification

Random Forest

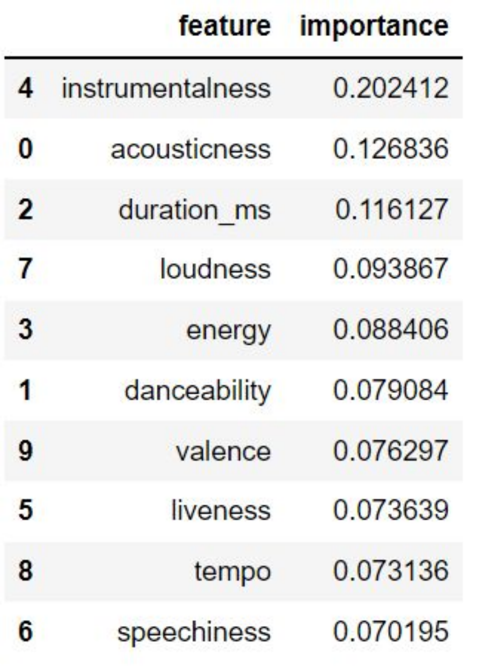


```
Mean Accuracy: 0.697
```

Baseline: LR



```
RMSE: 18.518961592740915
```

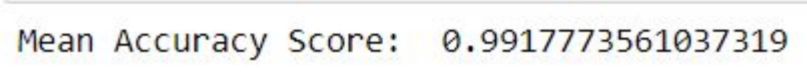


	feature	importance
4	instrumentalness	0.202412
0	acousticness	0.126836
2	duration_ms	0.116127
7	loudness	0.093867
3	energy	0.088406
1	danceability	0.079084
9	valence	0.076297
5	liveness	0.073639
8	tempo	0.073136
6	speechiness	0.070195

There is great improvement.... However, is there a **better model** out there?

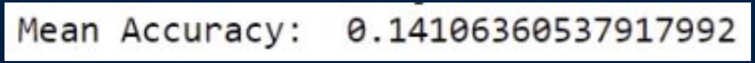
K Nearest Neighbor (KNN)

KNN



```
Mean Accuracy Score:  0.9917773561037319
```

Baseline: LR



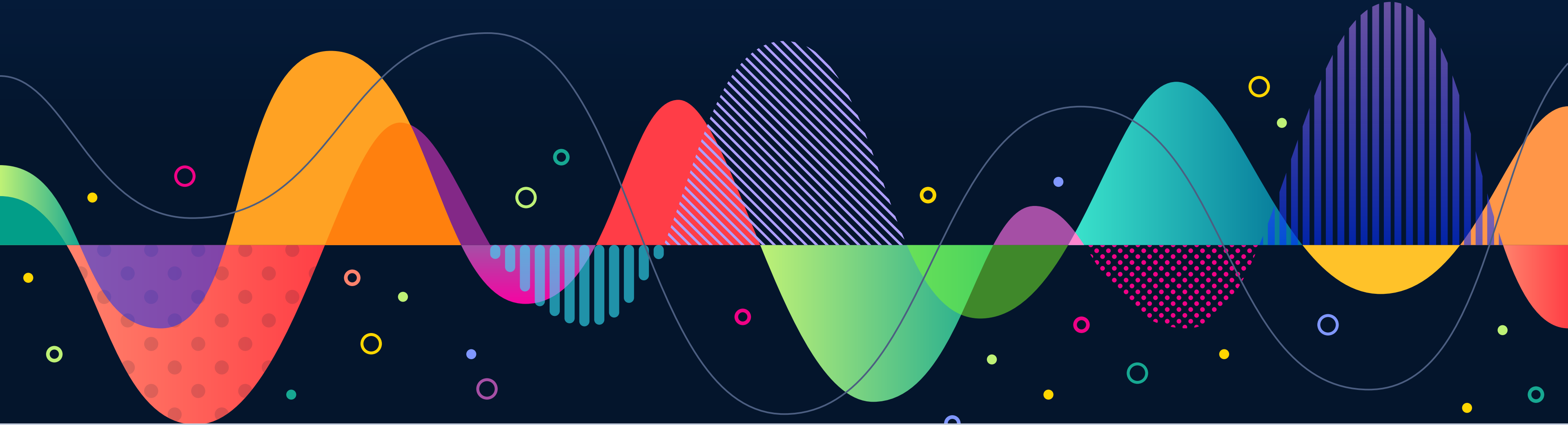
```
Mean Accuracy:  0.14106360537917992
```

There is significant improvement.... However, there is the risk of **overfitting**!

Next Steps

- Investigate potential overfitting of KNN model
- Use the finalized ML models to see if successful classification is achieved
- Data Preparation

Questions?



References

- https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset?select=dataset-of-00s.csv
- https://www.kaggle.com/leonardopena/top50spotify2019
- https://towardsdatascience.com/predicting-popularity-on-spotify-when-data-needs-culture-more-than-culture-needs-data-2ed3661f75f1

Notes from Dr. Yousefi

- RMSE is used for regression models
- Mean Accuracy is used for classification models
- Need to use more regression models to compare with the baseline of linear regression
- Set a baseline for classification models and have at least 2 other models to compare to baseline
- For overfitting, it is suggested to use diagnostic plots
- All learning models should have at least 0.5 mean accuracy, check linear model again
- Decision Tree is defined as regression model in our code not classification
- Good to give the statistics of the data in the report
- 94% is good and usually not overfitting
- Regression = predict
- Argue why we are choosing the features in the analysis