CLEMSON UNIVERSITY

Project Name: Spotify Playlist Extender

CPSC 6300: APPLIED DATA SCIENCE

SEMESTER: SPRING 2020

Final report

May 1st, 2020

Course Instructor

Dr. Nina Hubig

Group Members

1. Shivam Panwar
2. Sundaresh Narayanan
3. Meghan Patil

INTRODUCTION

1. What is the main question your project seeks to answer?

Answer: Our Project also known as the “Spotify Playlist Extender” has the main motive to predict the songs that will be liked by the customers based on the most trending songs running in the market and considering the customer’s taste. Our project seeks to answer the question of what trending songs will be liked by the customers based on their playlist. Considering all the songs in a user’s playlist we will have to suggest him songs which will be best suited for him. Since its just a suggestion a user may like or dislike the suggestion. Our project tries to improve the extender so that majority of the customers like the suggested trending songs

2. Provide a brief motivation for your project question. Why is this question important? What can we learn from your project?

Answer: The current Spotify playlist extender predicts and suggests songs based on the song we listened to most recently or the song most played. Our project developed the motivation to design a predictor wherein we can give a miscellaneous list of suggestions based on the playlist (suggesting songs like pop, rock EDM etc.). This question is significant because it is important to deliver what the customer the right thing at the right moment. With better suggestions the application will be used more among many customers.

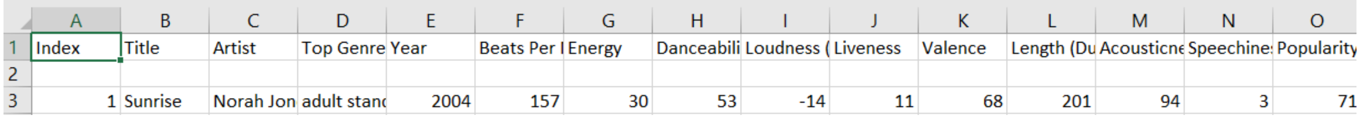
3. Briefly describe the data source(s) you have used in your project. Where is the data from? How big is the data in terms of data points and/or file size? If the data was not already available, how did you collect the data?

Answer: We have provided the dataset which is a .csv file. The dataset consists of the list of the trending songs from roughly 1956-2019. The dataset consists of 1994 rows and 14 columns. The dataset has the file size of 467 KB. We collected the data from the website Kaggle. We searched for the datasets in many websites and sources in github. We even tried to get it from the spotify website as they are known to provide free data to customers like Twitter.

SUMMARY OF EDA

1. What is the unit of analysis?

Answer:



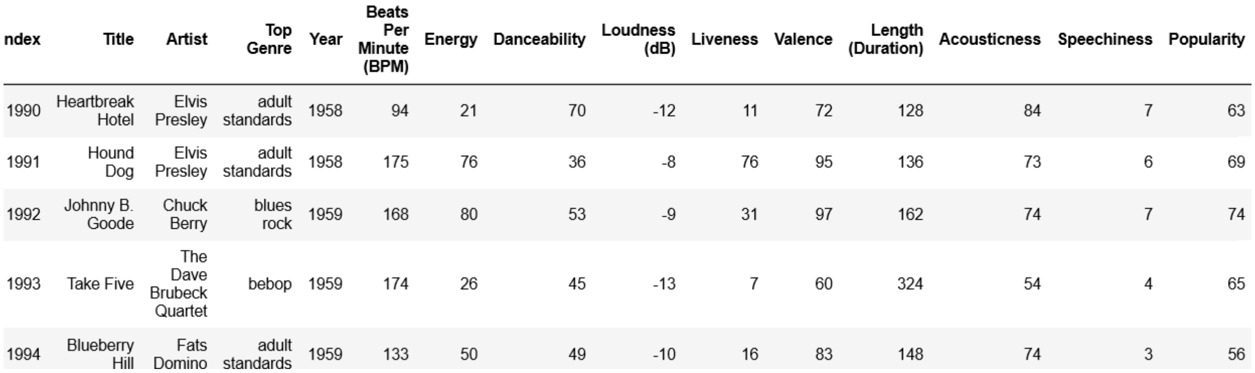
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Index	Title	Artist	Top Genre	Year	Beats Per I	Energy	Danceabili	Loudness (	Liveness	Valence	Length (Du	Acousticne	Speechine:	Popularity
2															
3	1	Sunrise	Norah Jon	adult stanc	2004	157	30	53	-14	11	68	201	94	3	71

Our row consists of 14 parameters just shown as the above image. The main aim of our project is to predict whether a certain song will be liked by the customer judging by the popularity of the song as well its artist.
So, the unit of analysis of our data set will be the column of popularity primarily.
We can measure the accuracy as
Accuracy = 1-(number of popular songs missed to predict/ number of actual popular songs liked by a person) * 100.

After studying and analyzing the dataset we found many parameters of not much use to our project and model designing. Therefore, we decided to trim our dataset and only keep the parameters which will be used in future.

2. How many observations in total are in the data set?

Answer:
The original data set before trimming
Number of rows: 1994
Number of columns: 14



ndex	Title	Artist	Top Genre	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
1990	Heartbreak Hotel	Elvis Presley	adult standards	1958	94	21	70	-12	11	72	128	84	7	63
1991	Hound Dog	Elvis Presley	adult standards	1958	175	76	36	-8	76	95	136	73	6	69
1992	Johnny B. Goode	Chuck Berry	blues rock	1959	168	80	53	-9	31	97	162	74	7	74
1993	Take Five	The Dave Brubeck Quartet	bebop	1959	174	26	45	-13	7	60	324	54	4	65
1994	Blueberry Hill	Fats Domino	adult standards	1959	133	50	49	-10	16	83	148	74	3	56

3. How many unique observations are in the data set?

Answer:
There aren't any particular unique values in our dataset, but we found one after analysis. We found the parameter Popularity to be quite unique as it had a level ranging from1-100 depending upon the number of people have added that song in their playlist or played that song multiple times. There are many other parameters which are unique, but we found the attribute popularity to be the most unique.

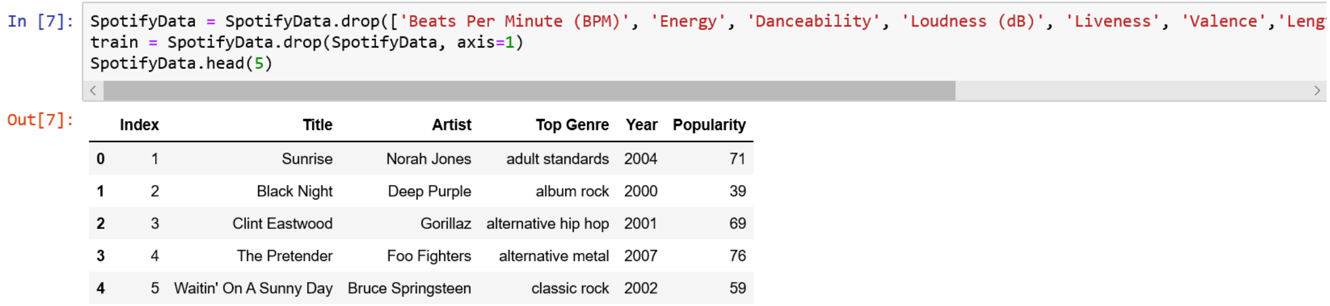
4. What time period is covered?

Answer: Songs from the year 1956 to 2019 are included in this dataset.

5. Briefly summarize any data cleaning steps you have performed

There were many null values in the column Acousticness and Loudness which were found irrelevant for us to predict a song liked by the customer or not. Therefore, we trimmed the dataset from having 14 columns to 5 columns.

We found 5 parameters to be very significant for our data set. We decided to trim the rest of the data as they had no relevance to what operations we are going to perform in designing the model.



```
In [7]: SpotifyData = SpotifyData.drop(['Beats Per Minute (BPM)', 'Energy', 'Danceability', 'Loudness (dB)', 'Liveness', 'Valence','Leng
        train = SpotifyData.drop(SpotifyData, axis=1)
        SpotifyData.head(5)
```

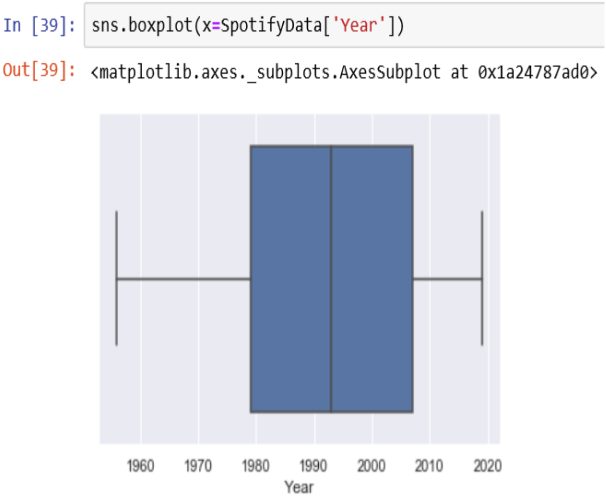
Out[7]:

	Index	Title	Artist	Top Genre	Year	Popularity
0	1	Sunrise	Norah Jones	adult standards	2004	71
1	2	Black Night	Deep Purple	album rock	2000	39
2	3	Clint Eastwood	Gorillaz	alternative hip hop	2001	69
3	4	The Pretender	Foo Fighters	alternative metal	2007	76
4	5	Waitin' On A Sunny Day	Bruce Springsteen	classic rock	2002	59

Data set after trimming
Number of rows: 1994
Number of columns: 5

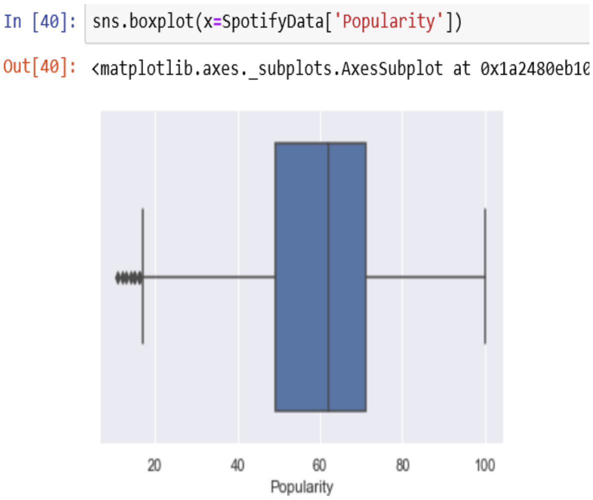
6. Visualization of the response with an appropriate technique.

Answer:



```
In [39]: sns.boxplot(x=SpotifyData['Year'])

Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x1a24787ad0>
```



```
In [40]: sns.boxplot(x=SpotifyData['Popularity'])

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2480eb1(
```

From the graphs we can clearly denote that most of the songs are from the year 1979-2008. We can study about the change in popularity over the years. Songs from 1980s are more famous than the ones created in recent years.

From the second graph we can understand the rate of popularity of songs. The rate mainly ranges between 50-70 (%). We can clearly see that there is not a single song which is liked by everyone. Judging by the rate of popularity we can predict whether the song will be liked by the customers.

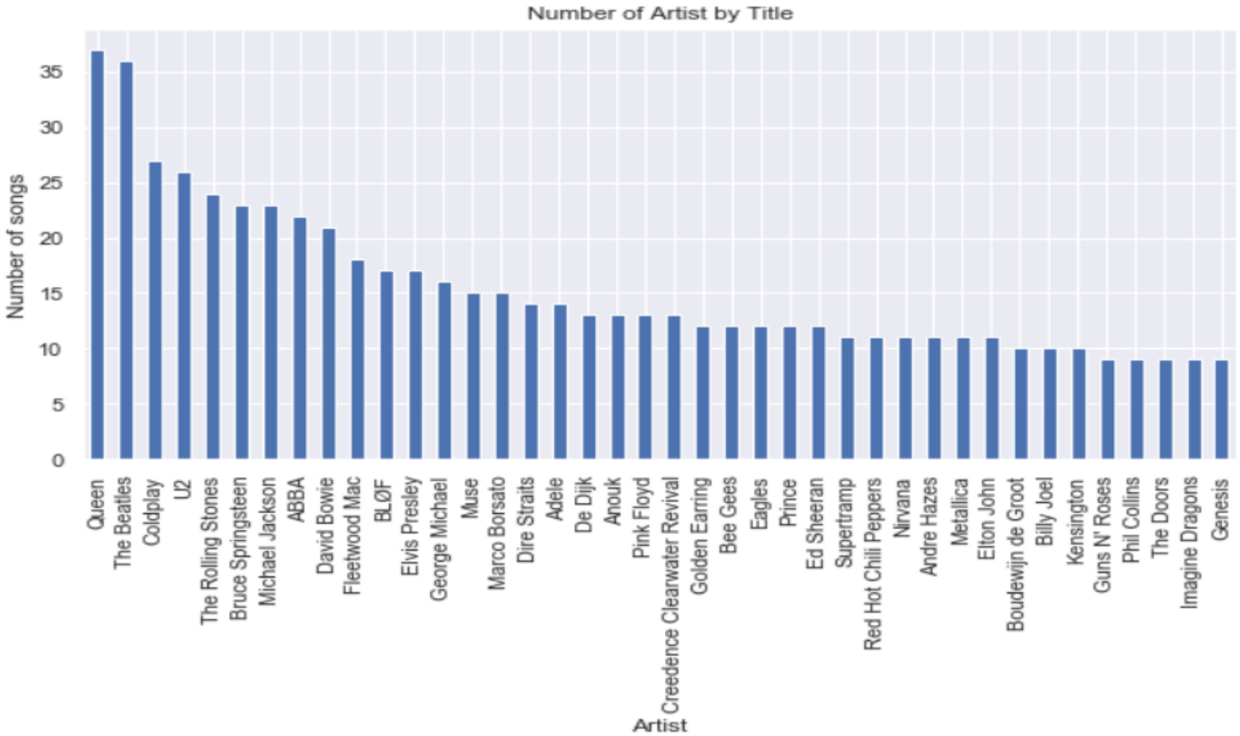


Fig: Histogram (Number of artists by title vs number of songs)

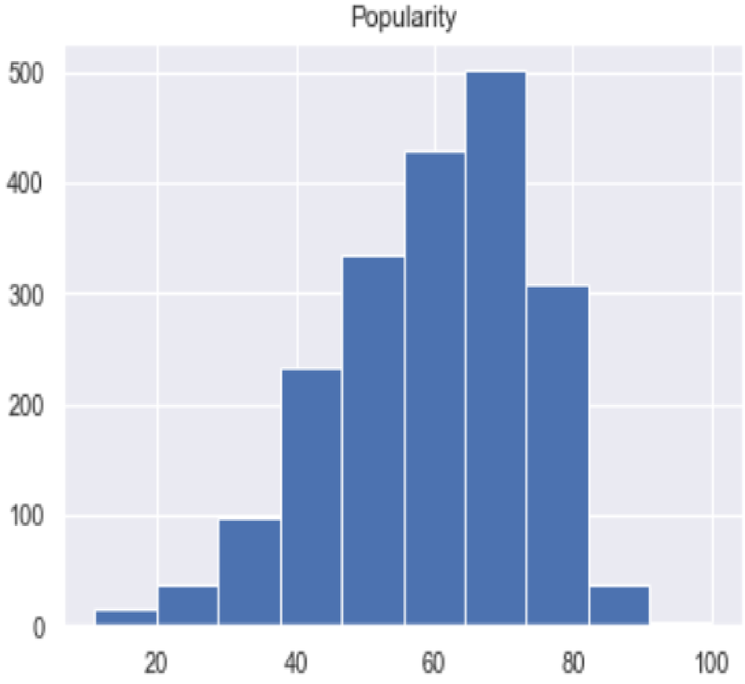


Fig: Histogram (Popularity vs number of songs)

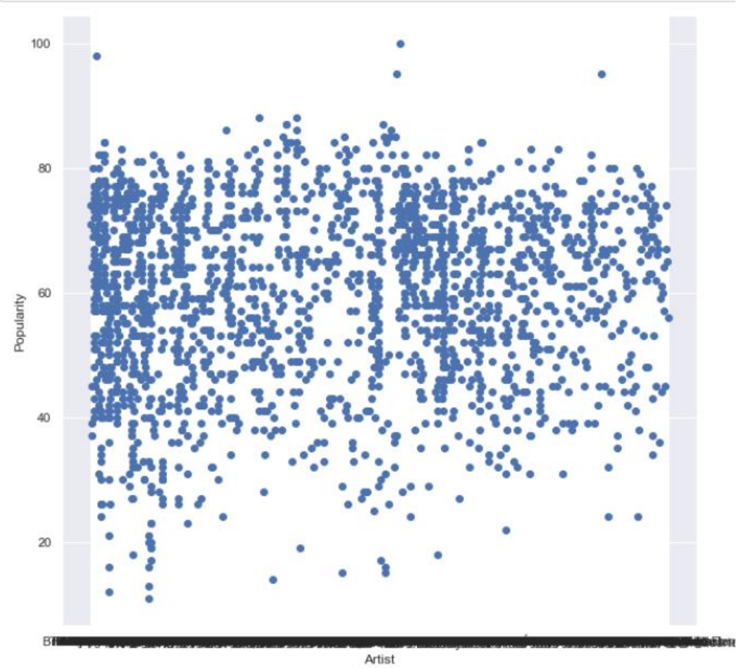


Fig: Scatter plot matrix (Artist vs Popularity)

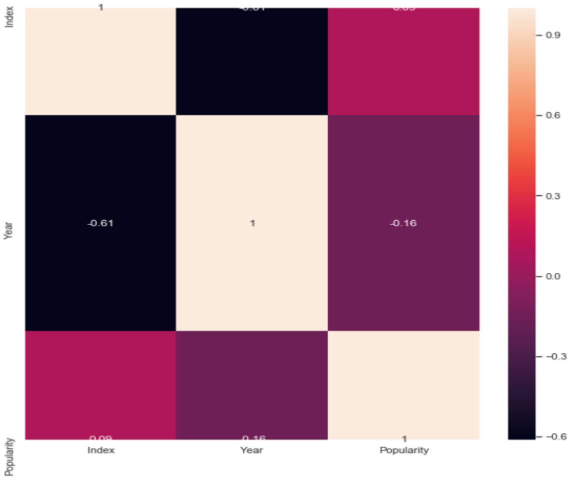


Fig: Heat map to find co-relation.

We come to know about the popularity of certain artists which can help us deduce the songs which may come out as the favorite of many people.
For Example: If the artist Imagine dragons is the most popular out of all we can predict that the songs of the same artist might come out as a favorite to many customers.

7. Visualization of key predictors against the response (e.g., scatterplot, boxplot, etc.). Pick one or two predictors that you think are going to be most important in explaining the response. Your selection of predictors can either be guided by your domain knowledge or be the result of your EDA on all predictors.

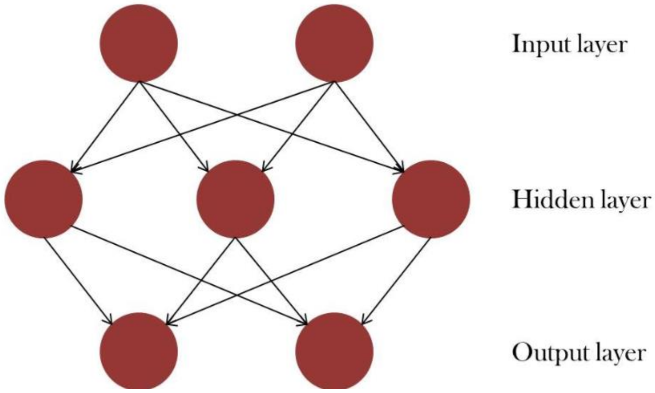
Answer:

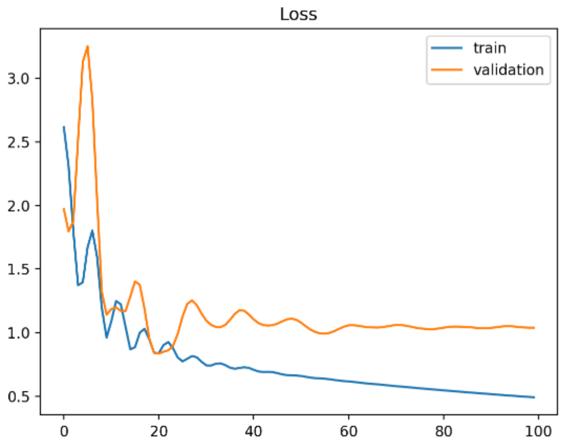
Type of predictors we are planning to choose and Why?

- Neural Networks: This algorithm is a self-learning method which thinks like a human and could prove useful for our project. Other algorithms need to be trained and still have a less accuracy ratio. This algorithm deduces the prediction with only those attributes that have an impact on the dataset that we have taken. For example: The mean age or salary will have no impact in a flight delay. Neural network automates itself by learning patterns from the data set and has a high prediction ratio than most of the traditional algorithms. With the nodes and neurons, it will be able detect the popularity of the song and can even deduce whether the upcoming song of a popular artist will be favorite of many.

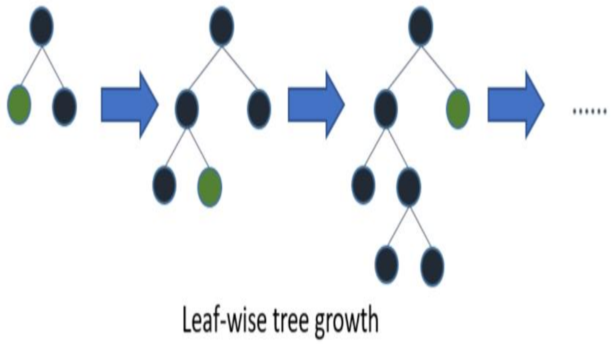
*y = function (w_1*x_1 + w_2*x_2 + c)*

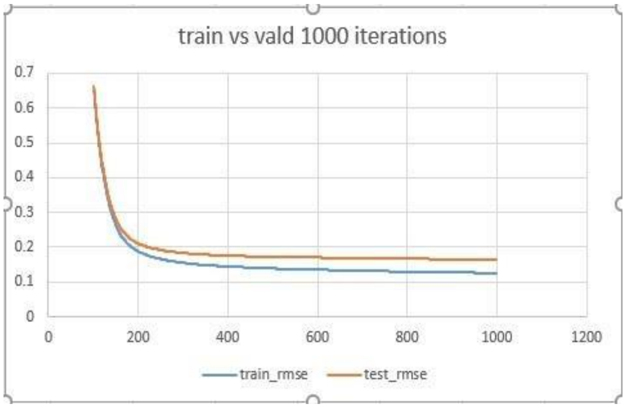
where f stands for 'a function of', x_n is the n-th input, w_n is the n-th weight of that input, and c is the constant for the node with output y.





LightGBM: It is known as one of the best classifiers among tree-based algorithms. LightGBM grows vertically whereas other algorithms grow horizontal level wise. This algorithm has leaf nodes which grow vertically to accumulate large datasets. It overfits small data set so it is not advisable for small datasets. In our dataset as we need to predict the choice of songs of customers according to the personal taste, we must deduce the factors that have an effect in the customer's choice. The attributes that come into consideration are artist popularity, song popularity, total likes for the song, the year the song was developed, length and album category. The study of the trend is necessary as in the early 60s people used to follow rock and gradually the taste has changed to pop and EDM (Electronic dance music). This algorithm can develop leaf nodes based on this attribute and help us increasing the accuracy ratio.





Based on the domain knowledge and Summary of EDA we diced to go with the above two predictors. We faced many complications while implementing the above two predictors.
In the model we have used the following predictors:

- K-neighbor
- Decision Tree classifier model
- ADA boost classifier model

SUMMARY OF MACHINE LEARNING MODELS

1. Justify your model choices based on how your response is measured and any observations you have made in your EDA.

Answer:

In the EDA we observed that the parameter popularity is the most important aspect of our model. The main reason we decided to go with th K-neighbor classifier is because it considers the objects which have similar attributes and properties.

- n *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

Now after considering the observation made from the EDA, a conclusion has been made to consider K nearest neighbor algorithm for model selection. The main reason why they had to choose this model is because this algorithm will consider all the nearby songs which is close to the most popular and trending song to suggest the same to the customers. All popular songs are clubbed together to form a group of most trending set. The other models we have considered is the LIGHTGBM and Neural network. For current instance we have considered these models to fit our data well and have a look how well it predicts other measures.

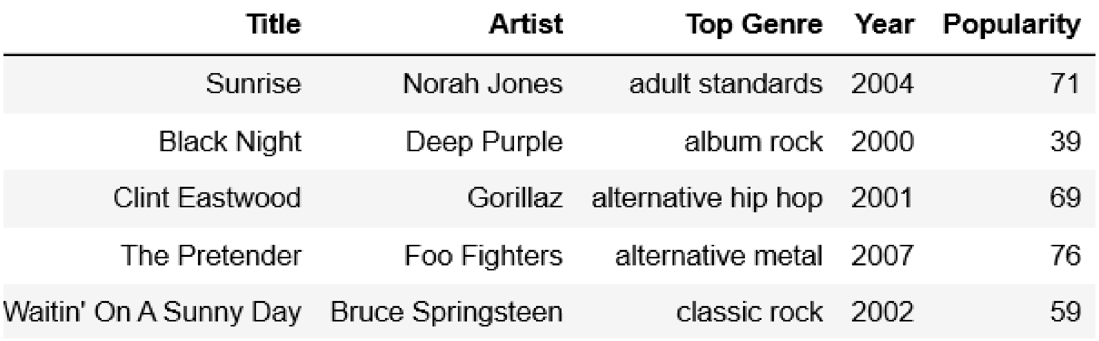
Decision Tree Classifier Model:

Decision tree plots the choice according to the choices we give to the model. After inducing certain conditions.

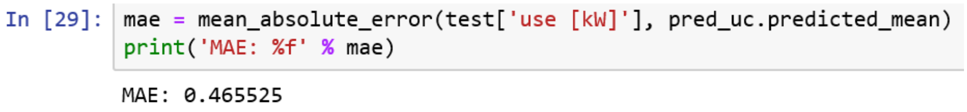
Using the decision algorithm, we start at the tree root and split the data on the feature that results in the **largest information gain (IG)** (reduction in uncertainty towards the final decision).
In an iterative process, we can then repeat this splitting procedure at each child node **until the leaves are pure**. This means that the samples at each leaf node all belong to the same class.
In practice, we may set a **limit on the depth of the tree to prevent overfitting**. We compromise on purity here somewhat as the final leaves may still have some impurity.

2. Report the results from at least two different models:
   - For each model, report the model's test error. Justify your choice.
   - For each model, discuss how well the model fits the data.

Answer: K-neighbor classifier:

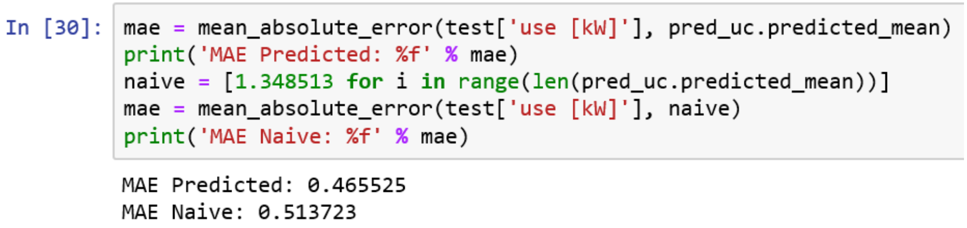


Title	Artist	Top Genre	Year	Popularity
Sunrise	Norah Jones	adult standards	2004	71
Black Night	Deep Purple	album rock	2000	39
Clint Eastwood	Gorillaz	alternative hip hop	2001	69
The Pretender	Foo Fighters	alternative metal	2007	76
Waitin' On A Sunny Day	Bruce Springsteen	classic rock	2002	59



```
In [29]: mae = mean_absolute_error(test['use [kW]'], pred_uc.predicted_mean)
         print('MAE: %f' % mae)

         MAE: 0.465525
```

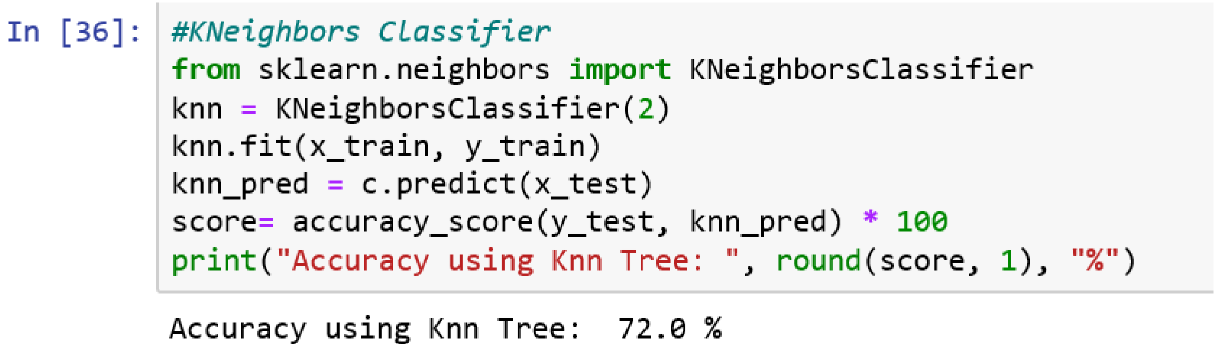


```
In [30]: mae = mean_absolute_error(test['use [kW]'], pred_uc.predicted_mean)
         print('MAE Predicted: %f' % mae)
         naive = [1.348513 for i in range(len(pred_uc.predicted_mean))]
         mae = mean_absolute_error(test['use [kW]'], naive)
         print('MAE Naive: %f' % mae)

         MAE Predicted: 0.465525
         MAE Naive: 0.513723
```

In the lecture, we discussed about finding the mean error using the method and formula of mean absolute error. Here, in the dataset as well we implemented the same to find out the error rate associated with the data and considering how well the model fits in our data. It must be made sure that the selected model does not overfit to provide poor predictions in other data sets.

You can see the result in the above screenshot that the MAE is 0.465525 which can be considered as a good result. We can change our prediction model if we get a minimum error rate in other models like the LIGHTGBM and Neural network.



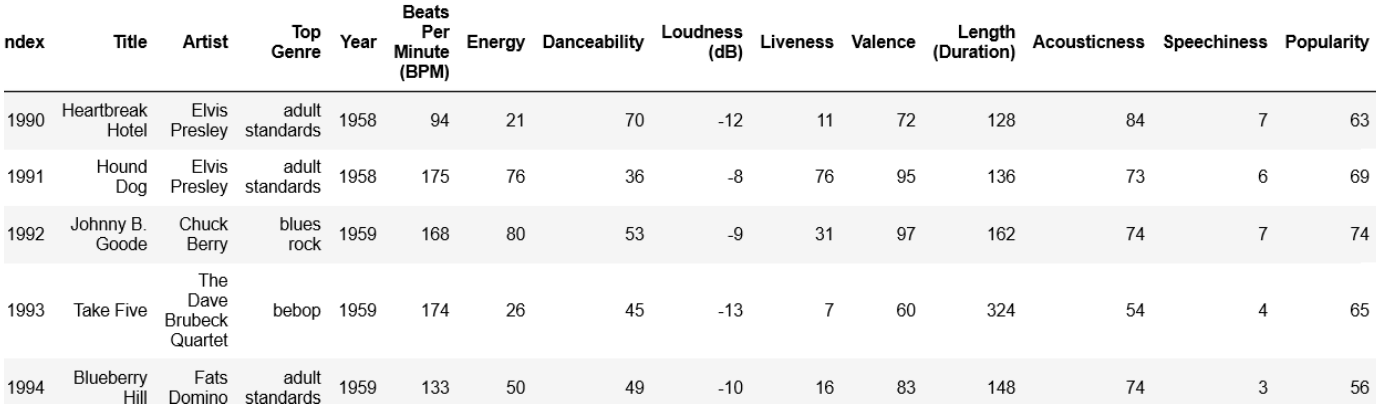
```
In [36]: #KNeighbors Classifier
         from sklearn.neighbors import KNeighborsClassifier
         knn = KNeighborsClassifier(2)
         knn.fit(x_train, y_train)
         knn_pred = c.predict(x_test)
         score= accuracy_score(y_test, knn_pred) * 100
         print("Accuracy using Knn Tree: ", round(score, 1), "%")

         Accuracy using Knn Tree:  72.0 %
```

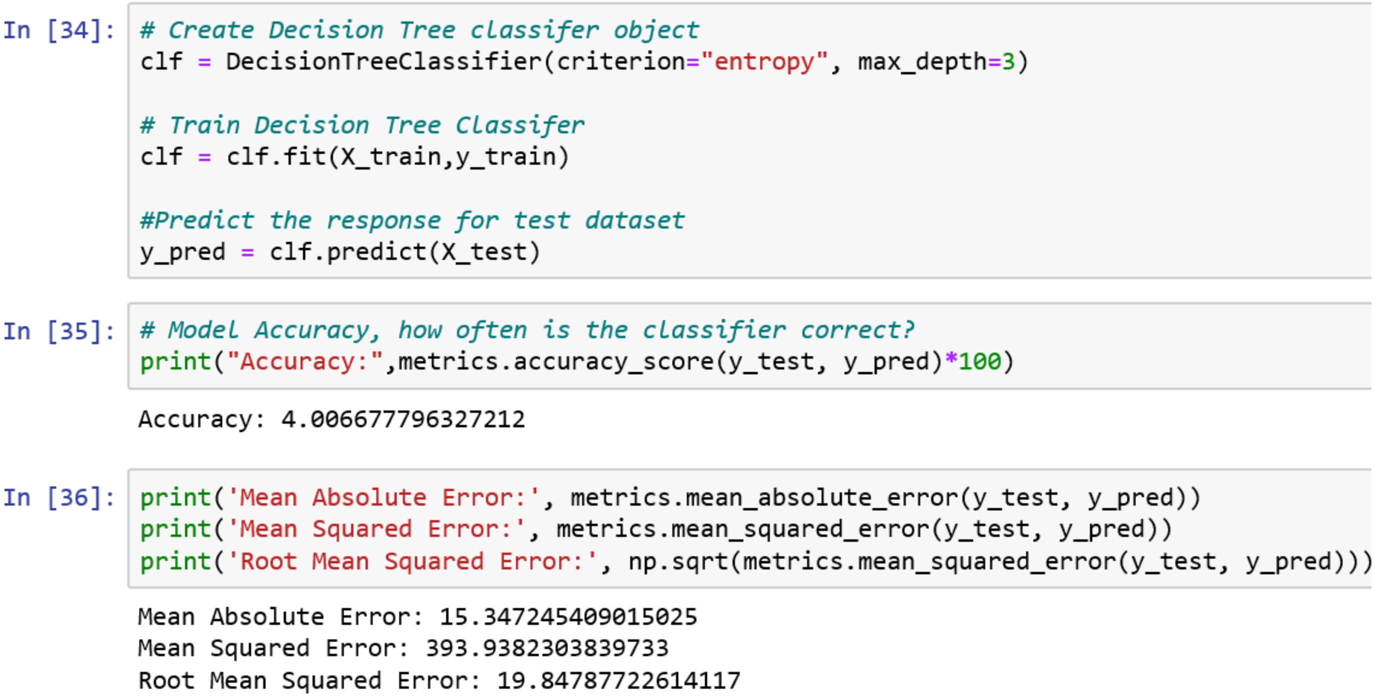
The test error rate has been discussed above with the diagram. Here, we can observe the accuracy rate of the model to new dataset. Since, our model does not overfit it provides a decent result considering that it will be much more optimal and efficient in future datasets. The lesser the error rate and more the accuracy rate can be an example of overfitting in given data set, but it is the opposite in the case of test data set. The training data set should only give the training model a pattern of information which can be used to classify the trending songs in a separate set.

Decision Tree Classifier:

Prediction of values based on the attribute popularity



ndex	Title	Artist	Top Genre	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
1990	Heartbreak Hotel	Elvis Presley	adult standards	1958	94	21	70	-12	11	72	128	84	7	63
1991	Hound Dog	Elvis Presley	adult standards	1958	175	76	36	-8	76	95	136	73	6	69
1992	Johnny B. Goode	Chuck Berry	blues rock	1959	168	80	53	-9	31	97	162	74	7	74
1993	Take Five	The Dave Brubeck Quartet	bebop	1959	174	26	45	-13	7	60	324	54	4	65
1994	Blueberry Hill	Fats Domino	adult standards	1959	133	50	49	-10	16	83	148	74	3	56



```
In [34]: # Create Decision Tree classifer object
         clf = DecisionTreeClassifier(criterion="entropy", max_depth=3)

         # Train Decision Tree Classifer
         clf = clf.fit(X_train,y_train)

         #Predict the response for test dataset
         y_pred = clf.predict(X_test)
```

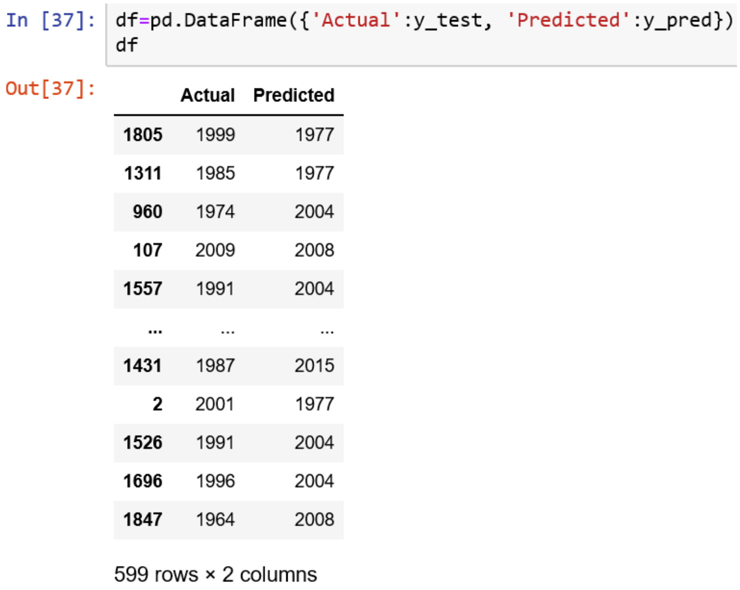
```
In [35]: # Model Accuracy, how often is the classifier correct?
         print("Accuracy:",metrics.accuracy_score(y_test, y_pred)*100)

         Accuracy: 4.006677796327212
```

```
In [36]: print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
         print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
         print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

         Mean Absolute Error: 15.347245409015025
         Mean Squared Error: 393.9382303839733
         Root Mean Squared Error: 19.84787722614117
```

The accuracy of decision tree classifier model is very low than we expected. It has a mean absolute error of 15.347
The expectations we had set from this model did not come out as planned. Additionally, it started predicted wrong values when we compared what the actual outputs should be.



```
In [37]: df=pd.DataFrame({'Actual':y_test, 'Predicted':y_pred})
         df
```

Out[37]:

	Actual	Predicted
1805	1999	1977
1311	1985	1977
960	1974	2004
107	2009	2008
1557	1991	2004
...	...	...
1431	1987	2015
2	2001	1977
1526	1991	2004
1696	1996	2004
1847	1964	2008

599 rows × 2 columns

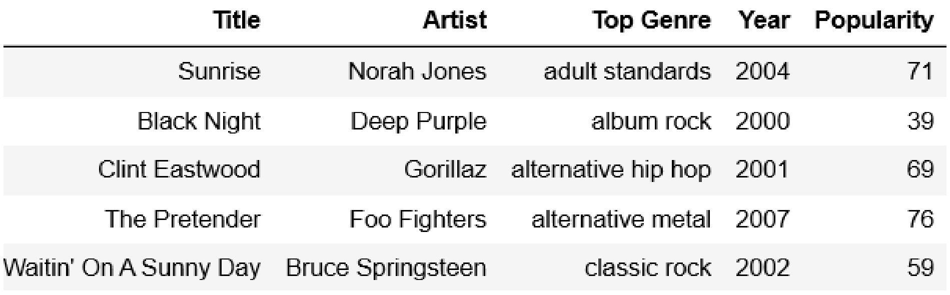
3. Briefly discuss which model fits the data better.

Answer:

The test error rate has been discussed above with the diagram. Here, we can observe the accuracy rate of the model to new dataset. Since, our model does not overfit it provides a decent result considering that it will be much more optimal and efficient in future datasets. The lesser the error rate and more the accuracy rate can be an example of overfitting in given data set, but it is the opposite in the case of test data set. The training data set should only give the training model a pattern of information which can be used to classify the trending songs in a separate set.

We can hereby conclude that the model K-neighbor fits the model better than the others.

4. For the model that fits the data best, make predictions for at least three cases of interest. One option is to show changes in predicted outcomes for changes in one of the predictors, holding all other predictors constant. Another option is to calculate predicted outcomes for cases of interest from the data set, or for hypothetical cases that are of interest.



Title	Artist	Top Genre	Year	Popularity
Sunrise	Norah Jones	adult standards	2004	71
Black Night	Deep Purple	album rock	2000	39
Clint Eastwood	Gorillaz	alternative hip hop	2001	69
The Pretender	Foo Fighters	alternative metal	2007	76
Waitin' On A Sunny Day	Bruce Springsteen	classic rock	2002	59

The above diagram shows the prediction for 5 cases from the test data based on the attribute popularity. The most trending songs of the whole dataset if considered as a single playlist can be determined that these 5 songs are the most popular and the customer has a high chance of liking the songs provided. The classifier model also considers the popularity of the artist as well. If there is a song of the artist who has enlisted all the famous songs, then they will also be displayed in the list.

SUMMARY AND CONCLUSION

Summary:

In our group project of the “Spotify” data set our model must predict the most upcoming trending songs which will be liked by the customers in future or give them suggestions for their playlist. After the Exploratory Data analysis, we found out that popularity is the main parameter that needs to be considered along with the type of song (Rock, Pop, EDM) and artist. With the help of these three parameters we need to select a model that fits well with the data set and predicts accurately with the prediction of upcoming trending songs.
From our project we have successfully learnt to plot graphs of a given dataset. We learnt how to do the exploratory data analysis. We could choose from a plethora of existing classifiers the one which would suit our data the most and give the most optimal results. We learned about the classifiers.

1. Going back to the question that has motivated your project, how would you answer that question given the results of your analysis?

   Answer:

   The model we implemented has a decent accuracy but not the best. We feel that neural networks with convolution layers and feature maps would have done much better than our current model. The NN model would have understood the pattern of trending songs from the dataset and without even much error it would have presented a better result than our current model. We are happy about our model that it did not overfit the training data. If NN could have been implemented it could be considered in real time application as well because deep learning can be very productive in such extender applications.

2. Think about domain experts in the field you have analyzed. What can they learn from your project? How could the results of your analysis inform their work?

   Answer:

   The traditional extender only suggests the songs that are have being listened to recently by the customers or have been played most the times in a playlist. Our deployed model introduces functions which can capture songs with different genre as well. Since our model does not overfit the training data it produces optimal results with the test data.

3. Identify one way that your project could be improved if you had more time and resources to work on this project. For example, what additional data would you gather? What alternative data cleaning decisions would you make? What additional models would you estimate?

   We could have deployed the model of neural networks if we were given additional time. We would have trained the NN model with more than 100 datasets of spotify songs. NN results could have excelled our model by miles as we would have achieved excellent results of prediction. Keras libraries and tensor flow convolution layers might have uplifted our project with optimal and better results than our current existing model.