Springboard

Capstone 2: Final Project Report

Problem Identification Overview

The focus of this project is to identify the key factors that classify the genre of a song and create an algorithm that will define the genre based on those factors.

In this project, we will use a Spotify dataset already wrangled from Kaggle that will allow us to search the data for genre classifications. The data that we were able to capture includes a selection of music from the 1960's to 2010's.

Unsupervised or Supervised Classification or Regression:

Supervised Classification

Deliverables:

- Jupyter notebooks
- Final report
- Presentation slide deck

Data Preprocessing Steps of Notes

Luckily, the data set captured was very clean and didn't require a lot of pre-processing. By looking at the scatter plots and correlation heatmaps, it was clear that you could use certain features to isolate particular genres indicating that a decision tree was going to be a strong model selection.

First, we took a look at the data set that was available via data wrangling. There were no missing values and a number of features we had to chose from. We had mostly numerical features with four non-numerical features. Two were fairly irrelevant to the analysis, track name and artist. The other two were important to note but ultimately, we only utilized one, the genre column which was what we were trying to use for identify the classes being used in the model.

The next step was exploratory data analysis to get a better idea of what the data looked like which would help us in defining what the model would need to look at. When looking at the available genre's we found an imbalanced classification set with six genres of various sizes. The largest sample size was pop at 16,460 tracks and the smallest class was EDM at 1,515 tracks. We also looked at the count of tracks per decade to see if there was an equal spread of tracks over time. The spread over decades was imbalanced but not as severely imbalanced as the classes. The initial thought was that the definition of genres may change over time but the results of the decision tree analysis proved that the decade the track was released was irrelevant to the analysis.

Next, we looked at tracks per artist. While there were a number of artists that hand multiples of tracks listed, the artist appeared to be irrelevant. There were hundreds, if not thousands, of artists included in the dataset showing there to be too many variances to be valuable in a decision tree.

Next, we looked at the numerical features that appeared to be the most valuable. The data was already scaled by the API which was very helpful. Several variables such as speechiness and energy showed to have one genre that stood out above the other genres which was the first indicator that a decision tree may be the most beneficial to define the relationship of the features to the genre. This was further visualized by the scatter plots at the end of the notebook.

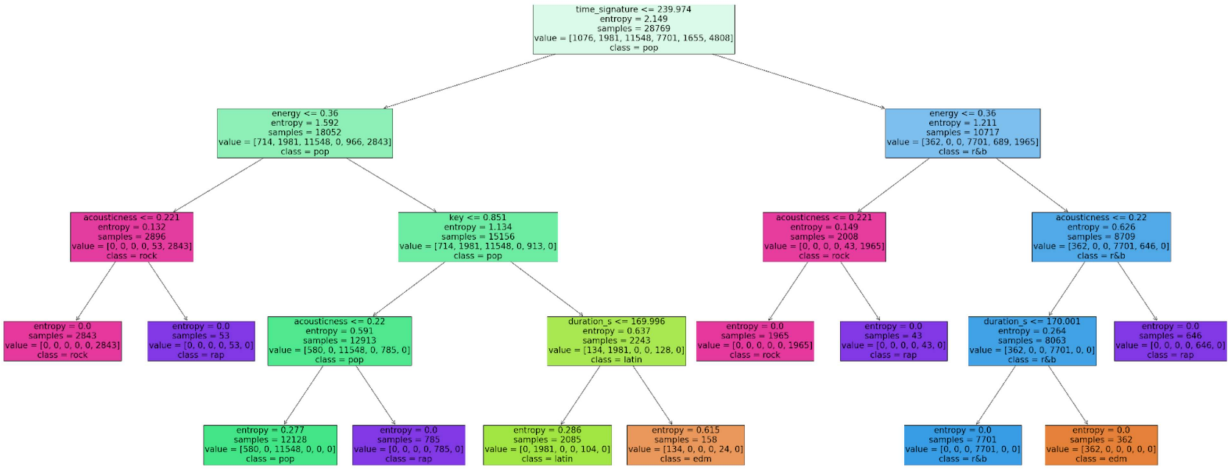
The last piece of data that we looked at before preparing for the model, was the correlation between the variables. We only showed two paring of variables with high correlation to each other: energy to loudness and duration to sections. Had we chosen another model, this would have been important to note for excluding highly correlated variables. But as we were looking to use decision models first, it was highly likely that the automation built into the decision tree classifier would automatically pull out the overly correlated variables.

Model Description

Input data size and features: 41,098 tracks

Model Algorithm and Parameters:

tree.DecisionTreeClassifier(criterion = 'entropy', max_depth = 4)



Along with the genre classifier, the data set included track, popularity, sections, chorus hit, time signature, duration, tempo, valence, liveness, instrumentalness, acousticness, speechiness, mode, loudness, key, energy, danceability, artist, and decade features.

Model Performance

	Precision	Recall	F-1 Score	Support
EDM	0.96	0.48	0.64	439
Latin	0.95	1.0	0.97	889
Pop	0.96	1.0	0.98	4,912
R&B	1.0	1.0	1.0	3,339
Rap	1.0	0.92	0.96	709
Rock	1.0	1.0	1.0	2,042
Accuracy			0.98	12,330
Macro Average	0.98	0.90	0.92	12,330
Weighted Avg.	0.98	0.98	0.97	12,330

Model Findings

While we had access to 19 different features, only time signature, energy, accousticness, key, and duration features appeared to have a significant impact on predicting genre. With those few features, we were able to predict genre with 98% accuracy.

Features for Pop

- Class Size: 16,460
- Time Signature: less than 240
- Energy: greater than .36
- Key: less than .85
- Acousticness: less than .22

Features for Rock

- Time Signature: less than 240
- Acousticness: less than .22

Features for Rap:

- Acousticness: greater than .22

Features for Latin:

- Time Signature: less than 240
- Energy: greater than .36
- Key: greater than .85
- Duration (in seconds): less than 170

Features for EDM

- Energy: greater than .36
- Key: greater than .85
- Duration (in seconds): greater than 170

Features for R&B

- Time Signature: greater than 240
- Energy: greater than .36
- Acousticness: less than .22
- Duration (in seconds): less than 170

Next Steps

The model performs well and there is easy access to the data via the Spotify API. The next step to the process would be to set up a link to the API and get the model into production.