Capstone Project 1

SONG POPULARITY PREDICTION

Ravinder Kumar Tanwar | Mentor – Siddharth Dixit | 02-11-2019

Problem Statement

Companies like YouTube , Spotify , Pandora , Amazon etc are constantly trying to improve their recommendation systems . One interesting application of this comes out to be popularity prediction. Problem is ; Given a song can we predict the popularity of the song if we have a database of songs containing features describing the song ?

WHO IS THIS USEFUL FOR ?

A general trend among Recording Companies and Record Labels is that they have been trying to constantly trying to improve the recommendation of their products and trying to predict user behavior especially now since there is no shortage of data.

On the same lines if we can make use of the data to find some insights from the past data and possibly predict the future. Spotify and Pandora could use this to recommend popular songs .

A record label could use this to release a single from an album that has a higher chance of being popular rather than randomly picking from the album.

If we can build a machine learning model that that can predict the popularity of a song that would greatly contribute to the profits of the company since it able to give them a peek into future as to what should be released into the outer world.

WHAT IS THE DATA USED FOR THE PROJECT? WHAT IS THE SOURCE OF THE DATA?

The data consists of around 250k songs along with various features like song's tempo , acousticness, loudness , key (major or minor) etc and finally a measure of popularity which ranges from 0-100 where 0 means not popular at all to 100 where 100 means popular all around the globe. The data is available on Kaggle and is available in csv(comma separated values) format

TECHNIQUES USED TO SOLVE THE PROBLEM

- Data Wrangling:- This step involves observing the missing values , outliers or incorrect/inconsistent values and cleaning the data for further use .
- Exploratory Data Analysis:- This step involves delving deep into the data to find the relation between the dependent variables and independent variable . This includes correlation matrices, various kinds of plots and in general finding the trends for data storytelling .
- Inferential Statistics :- This part includes checking the mean , mode ,median along with various hypothesis testing like t test , z test etc . Along with this the part includes drawing inference using various plots like pairplot , barplots and histograms .

- Machine Learning Model :- After the initial work on the dataset it would be time to build a Machine Learning Model to find the Popularity Prediction of the song . Accuracy will be tested by using various models like Logistic Regression , Random Forest , XGBoost .

DELIVERABLES

Following tools will be used to work on the project:-

- Jupyter Notebook for handling the programming aspect of the project.
- Python3 as the choice for programming language
- Google Doc for the submission of process progress reports
- Microsoft Powerpoint for Data Storytelling and presentation

REFERENCES

- Kaggle
- Springboard Curriculum material
- Datacamp courses
- Google, StackOverflow , Github etc
- Documentation and open source work