Deducing Personal Traits from Music Listening History

Tejasvi Chebrolu - 2019114005

Prince Varshney - 2020121012

Motivation

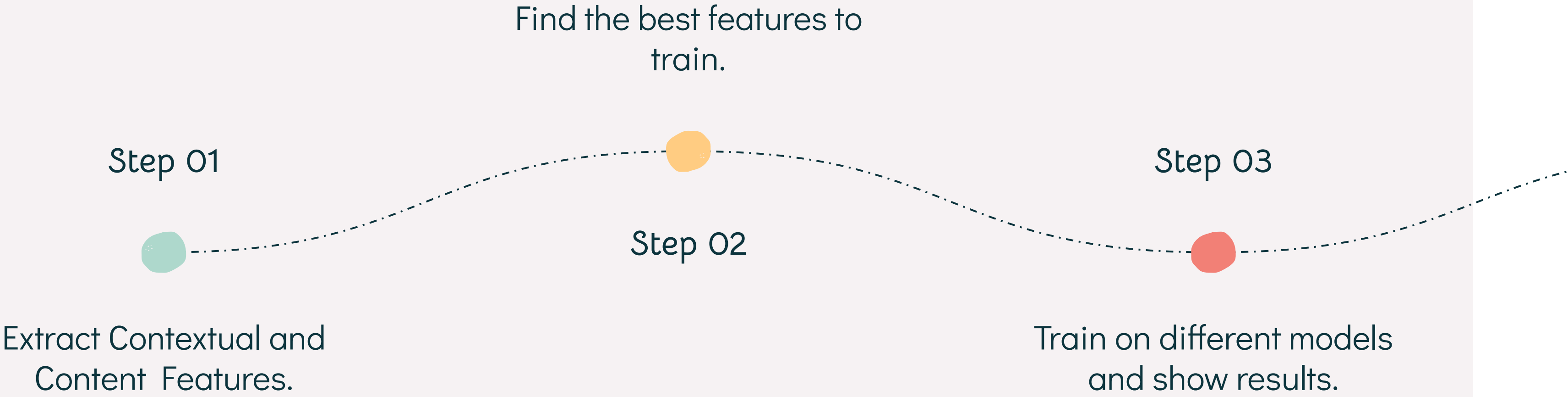
Music & Persona

Can we guess a person's personal features based on their music?

Privacy

Is anything you do on the internet safe?

Methodology



Step 01

Extract Contextual and Content Features.

Step 02

Find the best features to train.

Step 03

Train on different models and show results.

Background

Inferring Personal Traits from Music Listening History

Earliest work done with minimal understanding.

Predicting user demographics from music listening information

Newer work with a different dataset, model.

Predicting Personality Using Novel Mobile Phone-Based Metrics

Predicting personality instead of personal traits, and used a different kind of data.

Issues Faced

Dataset Size

The size of the dataset was quite small.

Dataset Accuracy

There were many NaN fields in the dataset.

Time-zones

The time-zones were not adjusted according to the geographical location.

Novel Work Done

Context

Definitions of context related features were tweaked to gain better accuracy.

Content

Definitions of content related features were tweaked to gain better accuracy.

Sessions

A new feature was defined to gain a better accuracy.

Models Trained

Support Vector Machine

A SVM was trained with different kernels to improve accuracy.

Linear Regression

A linear regression model was trained as well.

K- Nearest Neighbours

A KNN approach was also taken with K = 3.

Work Done

Contextual Features

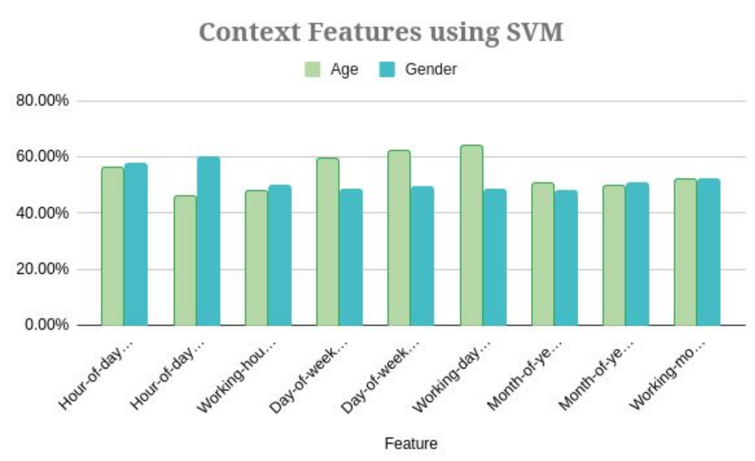
Can we use just contextual features for the prediction?

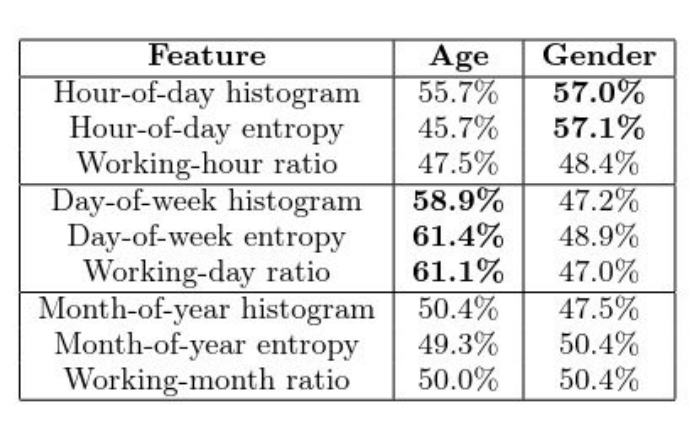
Work Done

Support Vector Machine

What are the predictions using a SVM?

Context Features





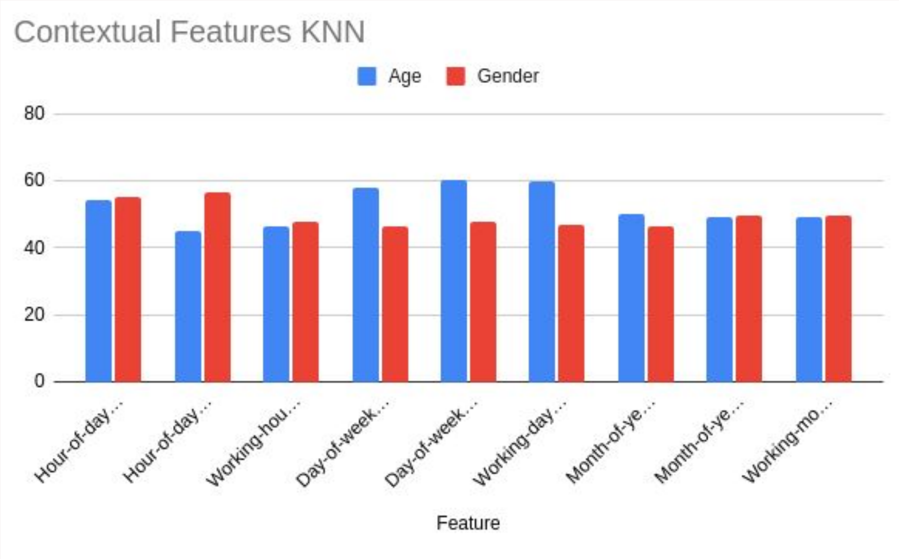
Feature	Age	Gender
Hour-of-day histogram	55.7%	**57.0%**
Hour-of-day entropy	45.7%	**57.1%**
Working-hour ratio	47.5%	48.4%
Day-of-week histogram	**58.9%**	47.2%
Day-of-week entropy	**61.4%**	48.9%
Working-day ratio	**61.1%**	47.0%
Month-of-year histogram	50.4%	47.5%
Month-of-year entropy	49.3%	50.4%
Working-month ratio	50.0%	50.4%

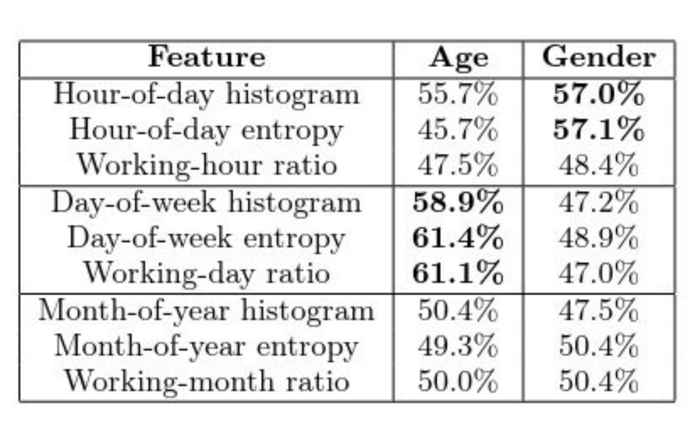
Work Done

K Nearest Neighbours

Is a KNN algorithm better?

Context Features





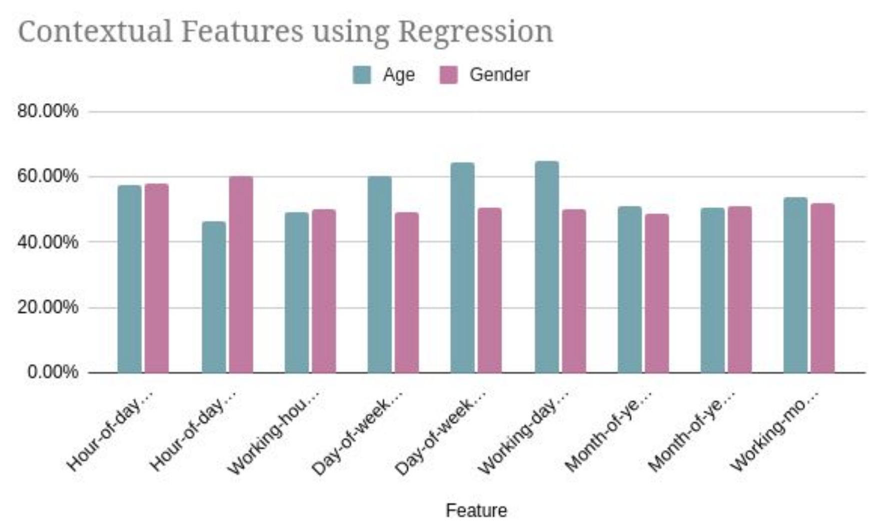
Feature	Age	Gender
Hour-of-day histogram	55.7%	**57.0%**
Hour-of-day entropy	45.7%	**57.1%**
Working-hour ratio	47.5%	48.4%
Day-of-week histogram	**58.9%**	47.2%
Day-of-week entropy	**61.4%**	48.9%
Working-day ratio	**61.1%**	47.0%
Month-of-year histogram	50.4%	47.5%
Month-of-year entropy	49.3%	50.4%
Working-month ratio	50.0%	50.4%

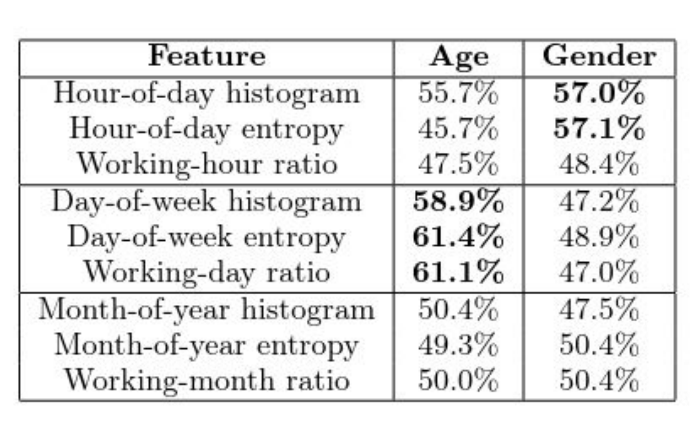
Work Done

Logistic Regression

Does using a Logistic Regression Model help?

Context Features





Feature	Age	Gender
Hour-of-day histogram	55.7%	**57.0%**
Hour-of-day entropy	45.7%	**57.1%**
Working-hour ratio	47.5%	48.4%
Day-of-week histogram	**58.9%**	47.2%
Day-of-week entropy	**61.4%**	48.9%
Working-day ratio	**61.1%**	47.0%
Month-of-year histogram	50.4%	47.5%
Month-of-year entropy	49.3%	50.4%
Working-month ratio	50.0%	50.4%

Work Done

Content Features

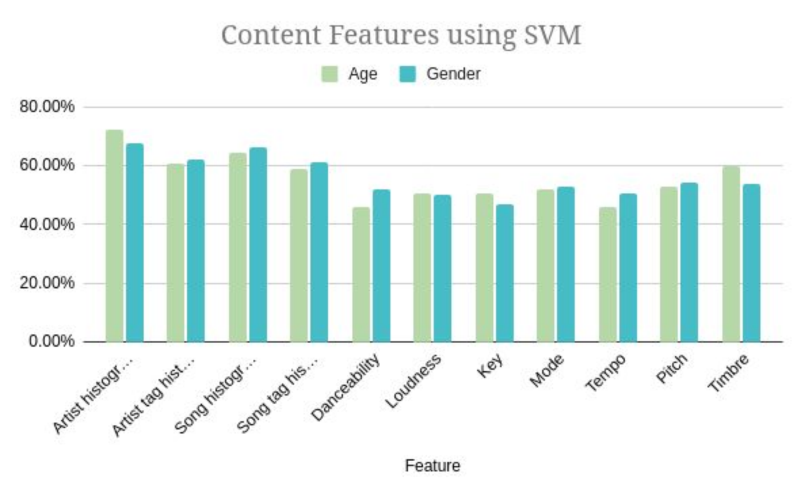
Can we use just content features for the prediction?

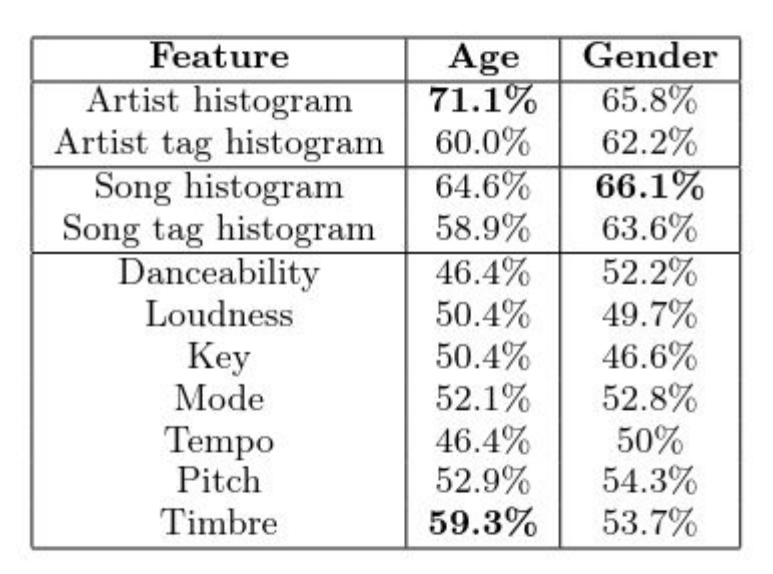
Work Done

Support Vector Machine

What are the predictions using a SVM?

Content Features





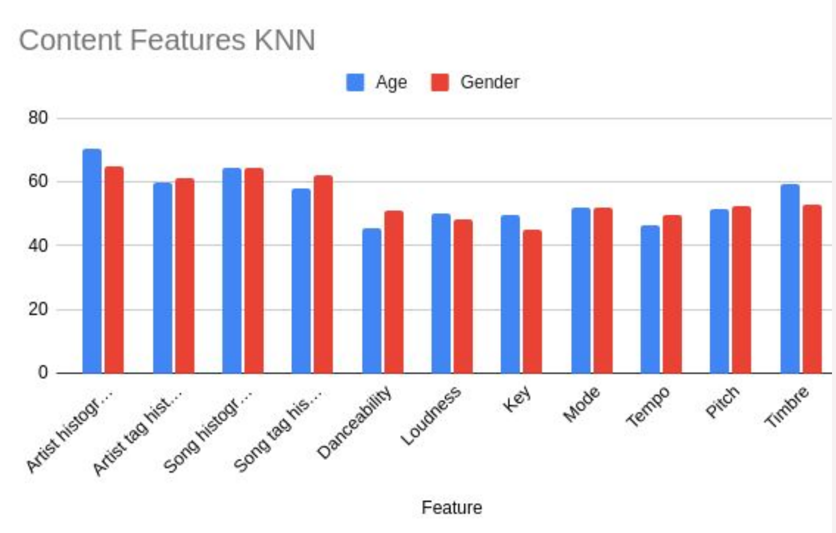
Feature	Age	Gender
Artist histogram	**71.1%**	65.8%
Artist tag histogram	60.0%	62.2%
Song histogram	64.6%	**66.1%**
Song tag histogram	58.9%	63.6%
Danceability	46.4%	52.2%
Loudness	50.4%	49.7%
Key	50.4%	46.6%
Mode	52.1%	52.8%
Tempo	46.4%	50%
Pitch	52.9%	54.3%
Timbre	**59.3%**	53.7%

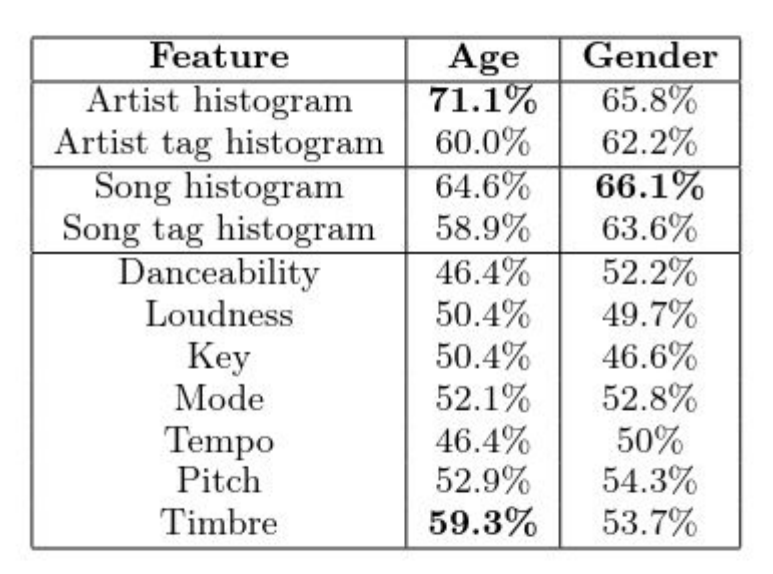
Work Done

K Nearest Neighbours

Is a KNN algorithm better?

Content Features





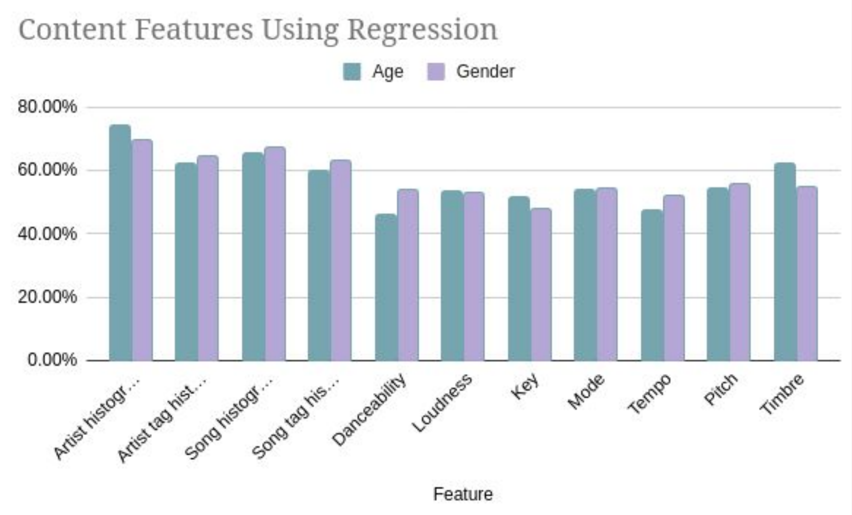
Feature	Age	Gender
Artist histogram	**71.1%**	65.8%
Artist tag histogram	60.0%	62.2%
Song histogram	64.6%	**66.1%**
Song tag histogram	58.9%	63.6%
Danceability	46.4%	52.2%
Loudness	50.4%	49.7%
Key	50.4%	46.6%
Mode	52.1%	52.8%
Tempo	46.4%	50%
Pitch	52.9%	54.3%
Timbre	**59.3%**	53.7%

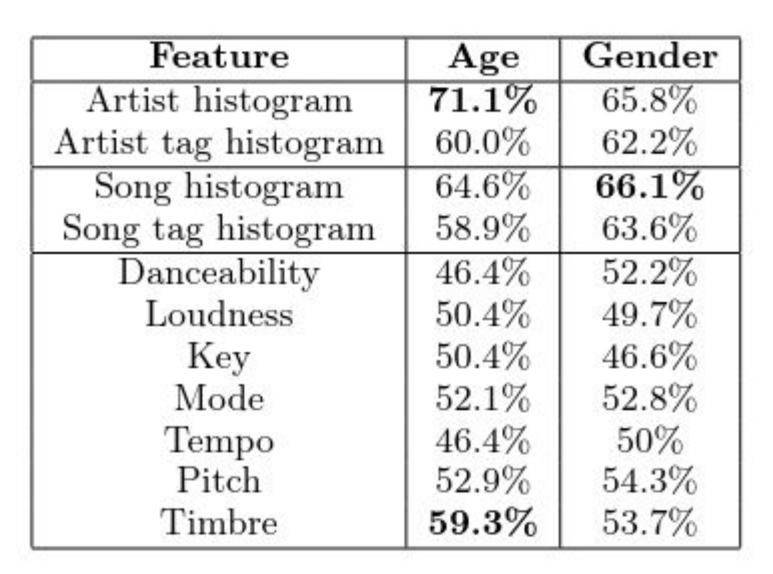
Work Done

Logistic Regression

Does using a Logistic Regression Model help?

Content Features





Feature	Age	Gender
Artist histogram	**71.1%**	65.8%
Artist tag histogram	60.0%	62.2%
Song histogram	64.6%	**66.1%**
Song tag histogram	58.9%	63.6%
Danceability	46.4%	52.2%
Loudness	50.4%	49.7%
Key	50.4%	46.6%
Mode	52.1%	52.8%
Tempo	46.4%	50%
Pitch	52.9%	54.3%
Timbre	**59.3%**	53.7%

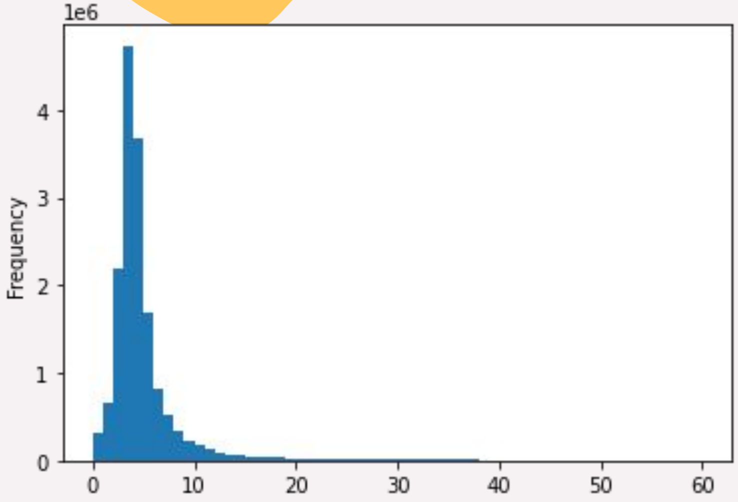
Work Done

Sessions

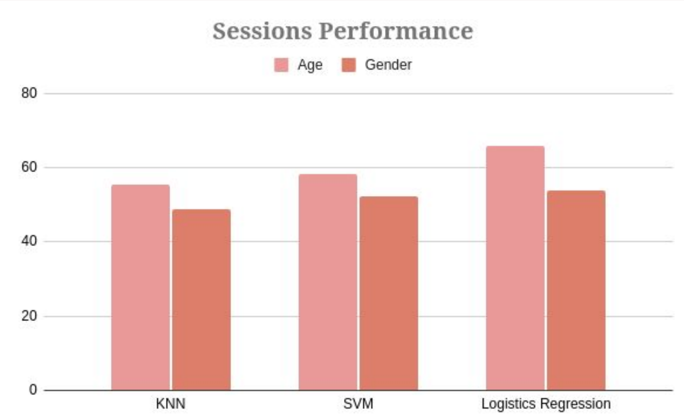
What are sessions? Is it helpful?

Session is defined as the sequence of songs that user listened to such that difference between any two songs is below the threshold.

How to find threshold?



Sessions Performance



Inferences

Models

Logistic Regression performed the best.

Context vs Content

Content seemed to perform better than Context.

Sessions

It performed better because of the definition.

General

1. Hard to use such data to make conclusive claims because of variations.
2. More data would have helped.
3. Genre tags, maybe?

Future Work

Thank You