Music Recommendation System

The Placeholders

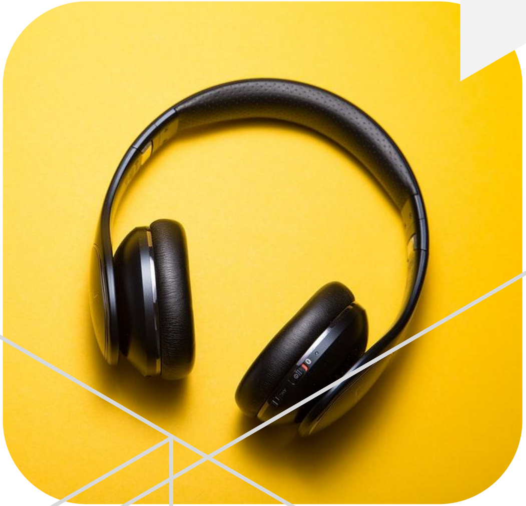
Aditi Sharma

Jyoti Prakash Maheswari

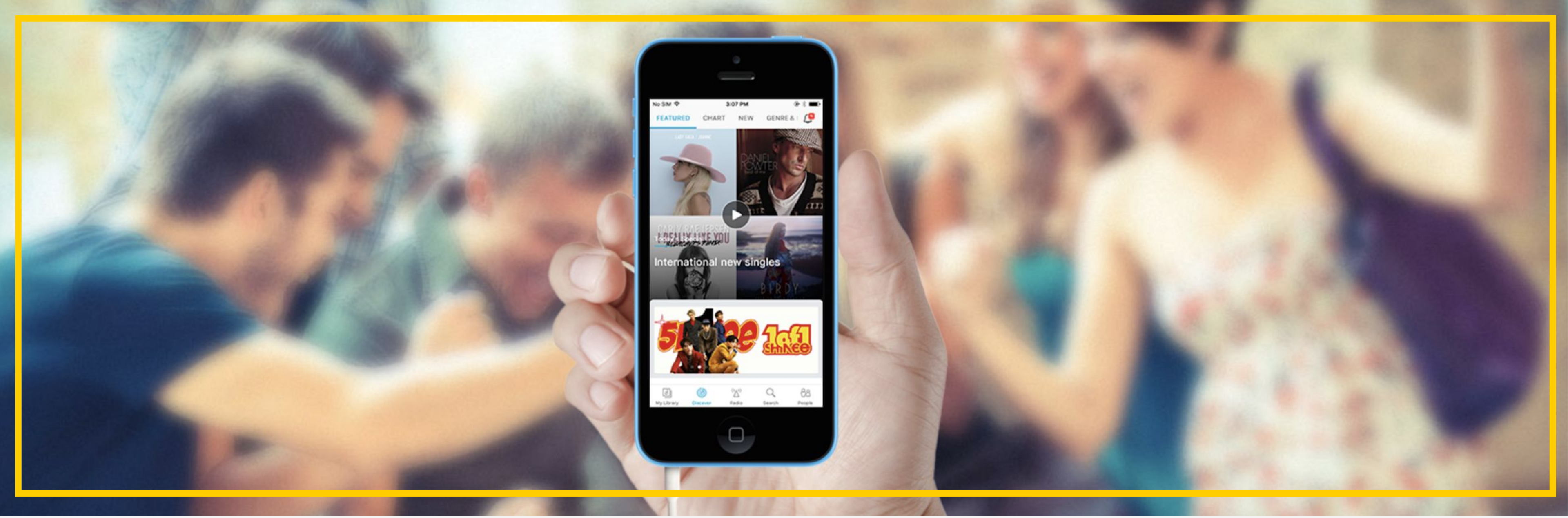
Joy Wenjing Qi

Xin Ke Sun

Zhe Yuan



MSDS621 Machine Learning - Final Project Presentation



kaggle

kkbox Music

A Music Streaming Service Mobile App

Asia's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 30 million tracks.

They want to build personalized recommendation algorithms to the users with unlimited streaming services.

Problem Statement

- ***What factors influence listener's decision to re-listen a given song***

- ***Build a model based on this understanding which predicts the chances of re-listening***

Our Process

1	2	3	4	5
Understand Data	**Data Processing**	**Build Models**	**Model Evaluation**	**Conclusion**
Summary Statistics	Data Cleaning	Modeling Methods	Confusion Matrix	Best Model
	Feature Engineering	Pipeline	ROC	Key Learnings
	Benchmark	Models		Demonstration

1

Understand Data

Summary Statistics

KKBOX Original Datasets

5 csv files

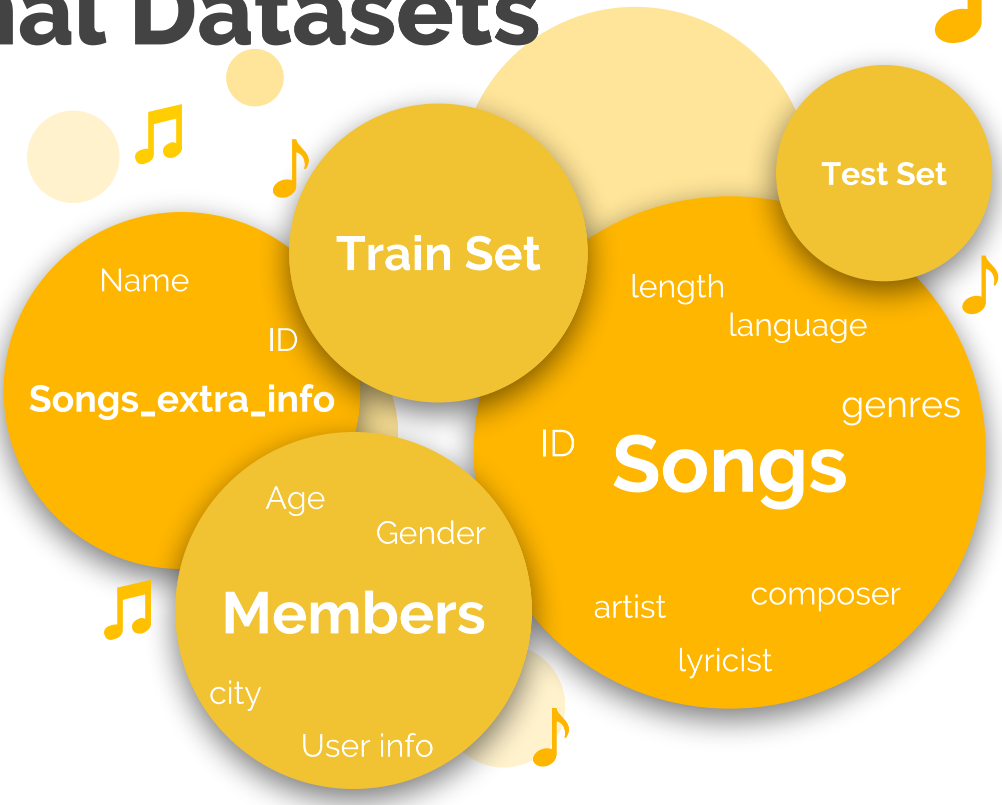
→ Songs.csv

→ Songs_extra_info.csv

→ Members.csv

→ Train.csv

→ Test.csv



Summary Statistics

Record

7 million records

360K unique songs

Music

2.3 million songs

11 distinct languages

192 distinct genres

220K distinct artists

320K distinct composers

100K distinct lyricists

User

34K unique users

Source

13 distinct source types

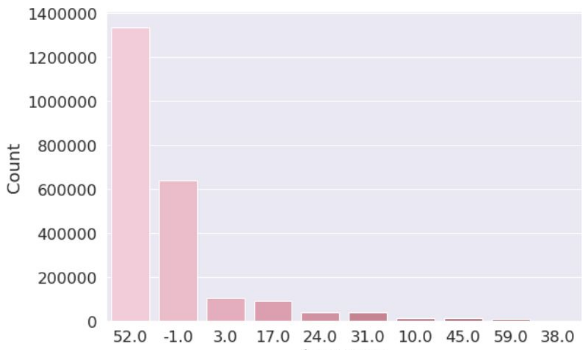
9 distinct source system tabs

21 distinct source screen

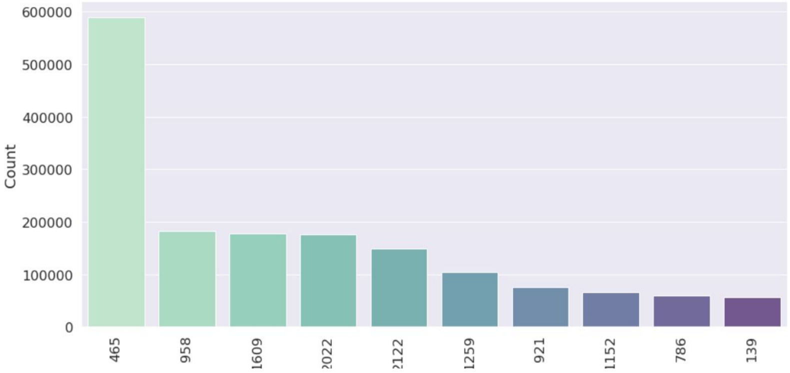
City

21 distinct cities

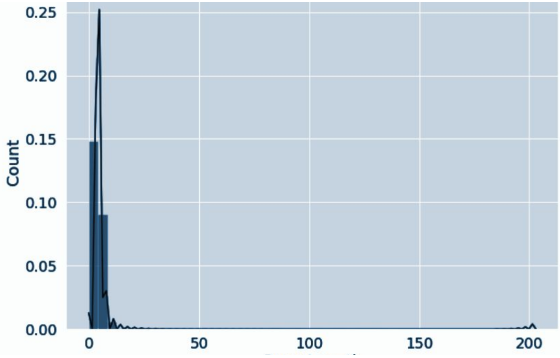
Distributions for Song data



Song Languages

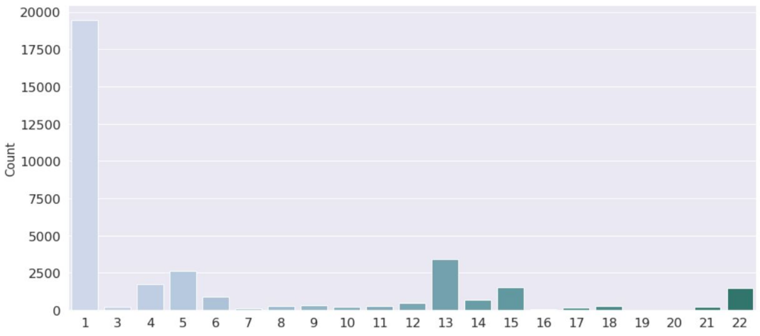


Song Genres

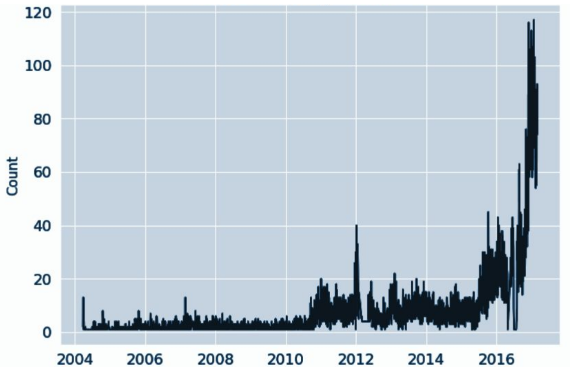


Song Lengths

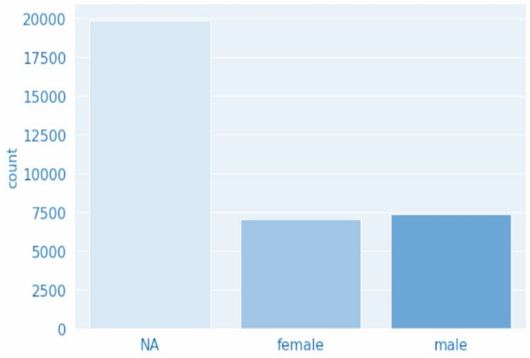
User data



City IDs



Registration Dates



Gender

Count missing values

Feature Name	NA%
source_screen_name	**0.33**
source_screen _name	**5.62**
source_type	**0.29**
gender	**58.84**
gender_ids	**4.09**
composer	**46.65**
lyricist	**84.71**
isrc	**5.94**

2

Data Processing

Data Cleaning, Feature Engineering

Data Cleaning & Feature Engineering Pipeline

Join Features

Merge & Split

Feature Engineering

Feature Extraction

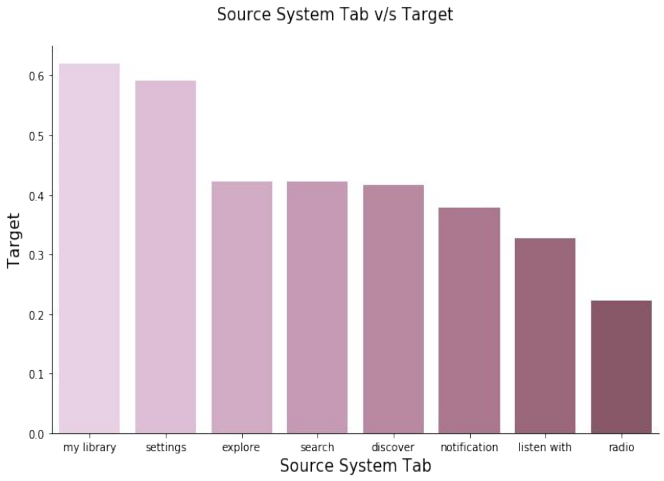
NA → 'nan'

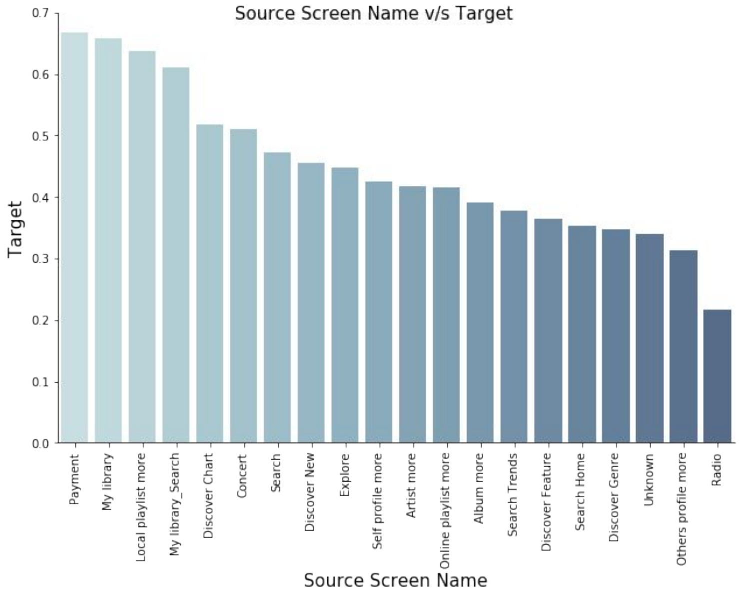
Imputation & Filling NAs

Encoding Data

Labels

Features vs Target





Possible New Features

- **Registration Date**
- **Expiration Date**
- **Counting Occurrences**
- **Listener's preference DIstribution**

Feature Sets

Feature Set 1

membership_days
day
year
month
is_featured
genre count
song_play_count

Feature Extraction
Counts for genres, songs, artists

Feature Set 2

month_start-end
composer_user_lev_c
Quarters
msno_genre_count
genre_columns

More Feature Extraction
User level distribution (Personalized)

Feature Engineering Functions

Counted Feature

genre_id_count
lyricist_count
composer_count
artist_count
is_featured

Data Imputation

cat_nan_list
cont_nan_list
fillna_nan

Feature Addition

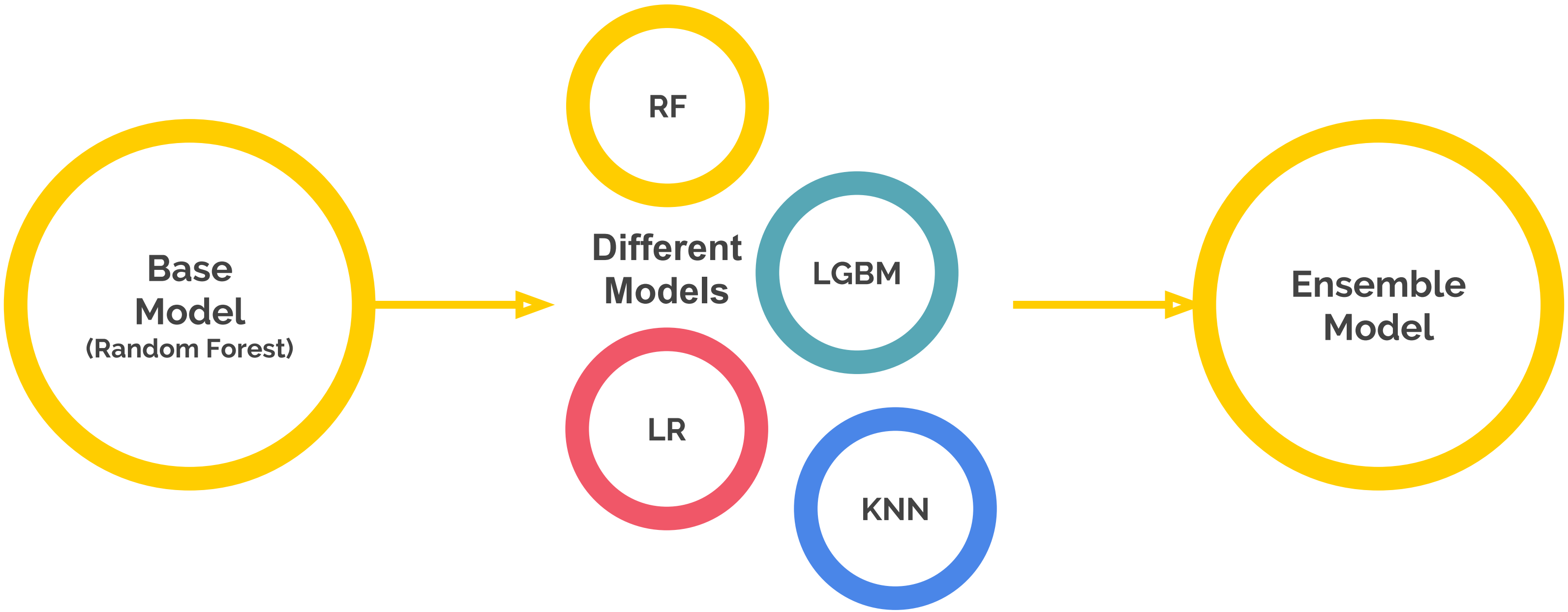
add_days_left
add_days_left
add_datepart_reg
add_datepart_exp
add_lyricist_count
add_composer_count
add_artist_count
add_featured_song
add_song_year
add_song_play_count
add_artist_played_count
add_msno_appear_count

3

Building Models

Modeling Methods, Pipeline, Models

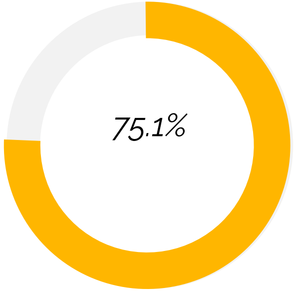
Modeling Method



Improved Accuracy

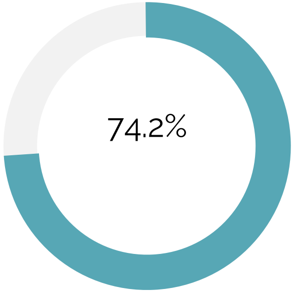
Models	Feature Set 1	Feature Set 2	More Data + Feature Set 2
Random Forest	56.7%	66.5%	75.1%
LightGBM	64.1%	68.5%	74.2%
KNN	53.8%	60.2%	69.3%
Logistic Regression	58.3%	63.0%	63.9%

Models selected



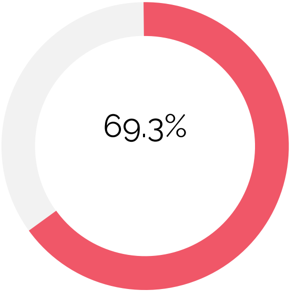
Mean Accuracy

Random Forest



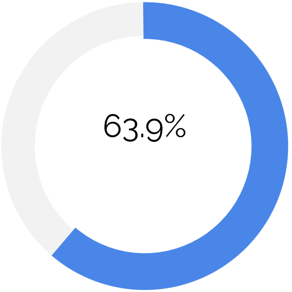
Mean Accuracy

LightGBM



Mean Accuracy

KNN



Mean Accuracy

Logistic Regression

Random Forest

Training Dataset

Featured Set 2
- 72 features
- 5M data

Baseline Model

Built Baseline
- 100K data points

66.5%

Regularization Parameter tuning

Best Features
- 32 features

Grid Search
- n_estimator
- max_feature
- min_sample_split

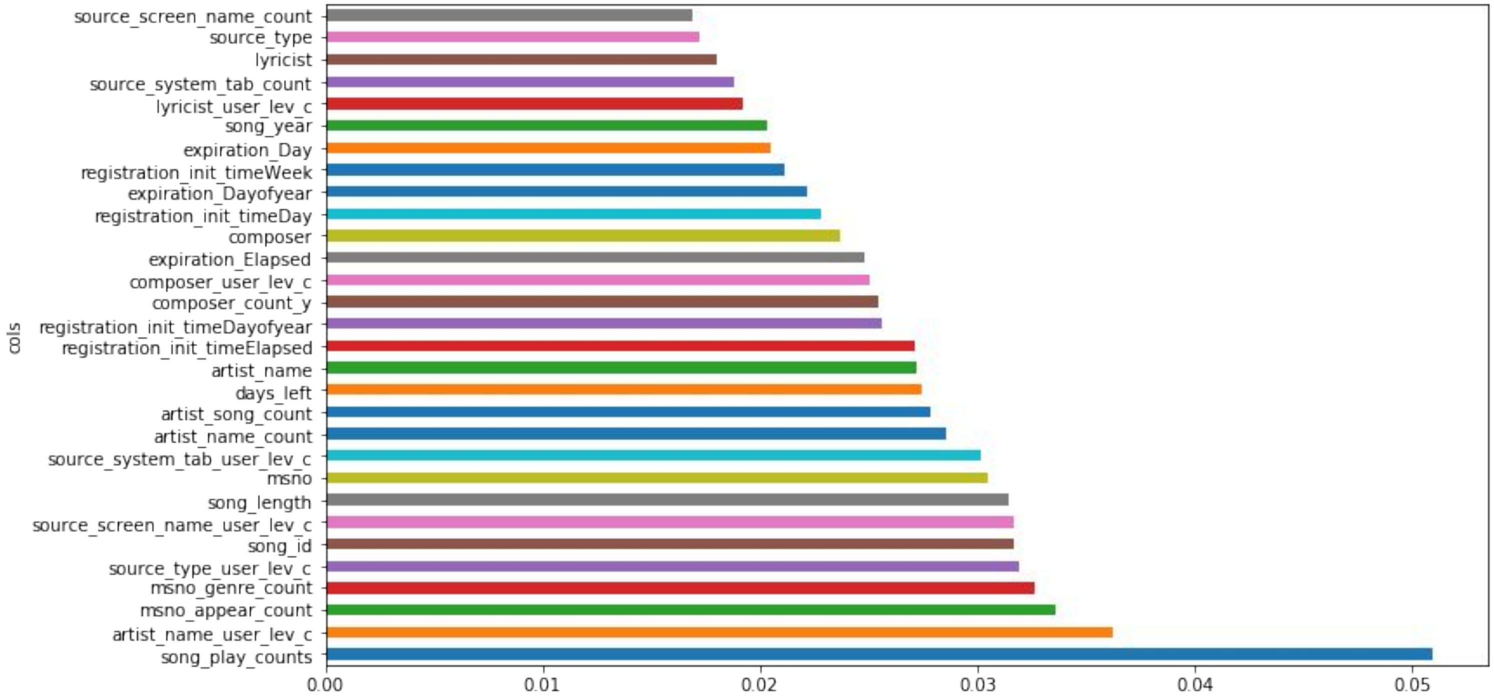
66.7%

Build Model
(With Best Params + 5M Data)

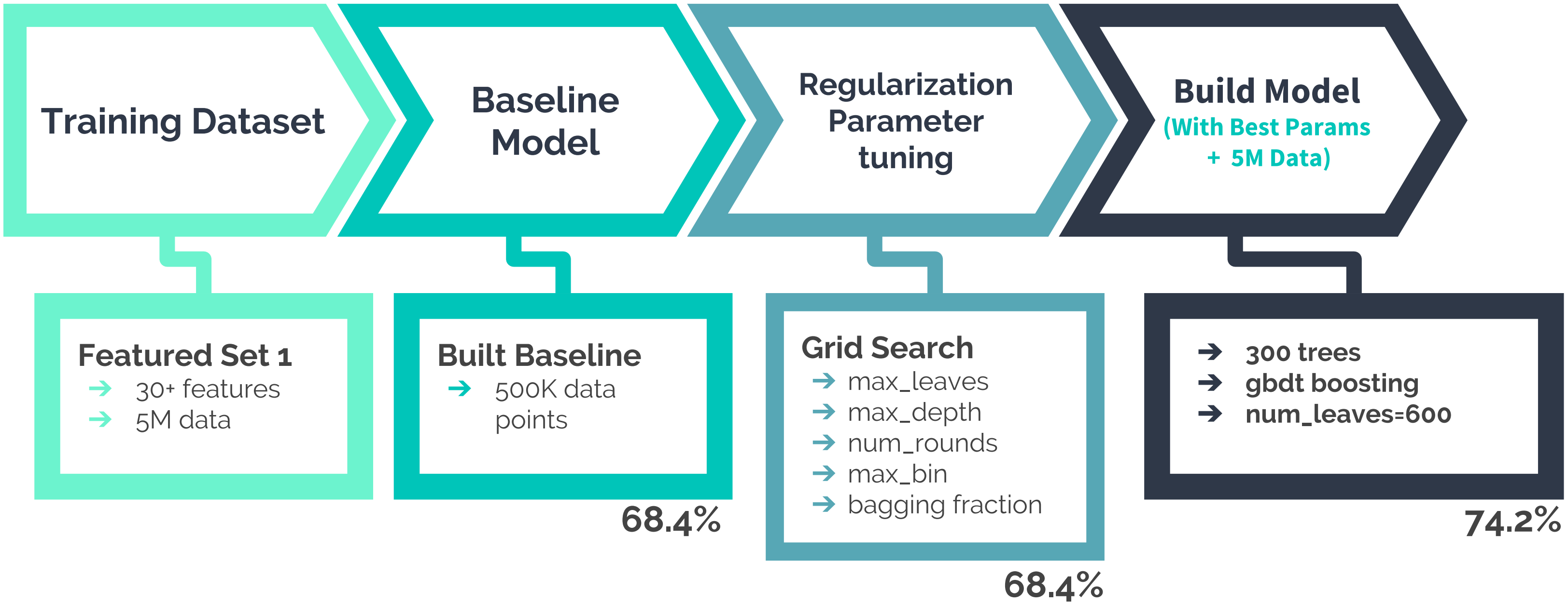
- 100 trees
- Sqrt feature selection
- min_sample_split=2

75.1%

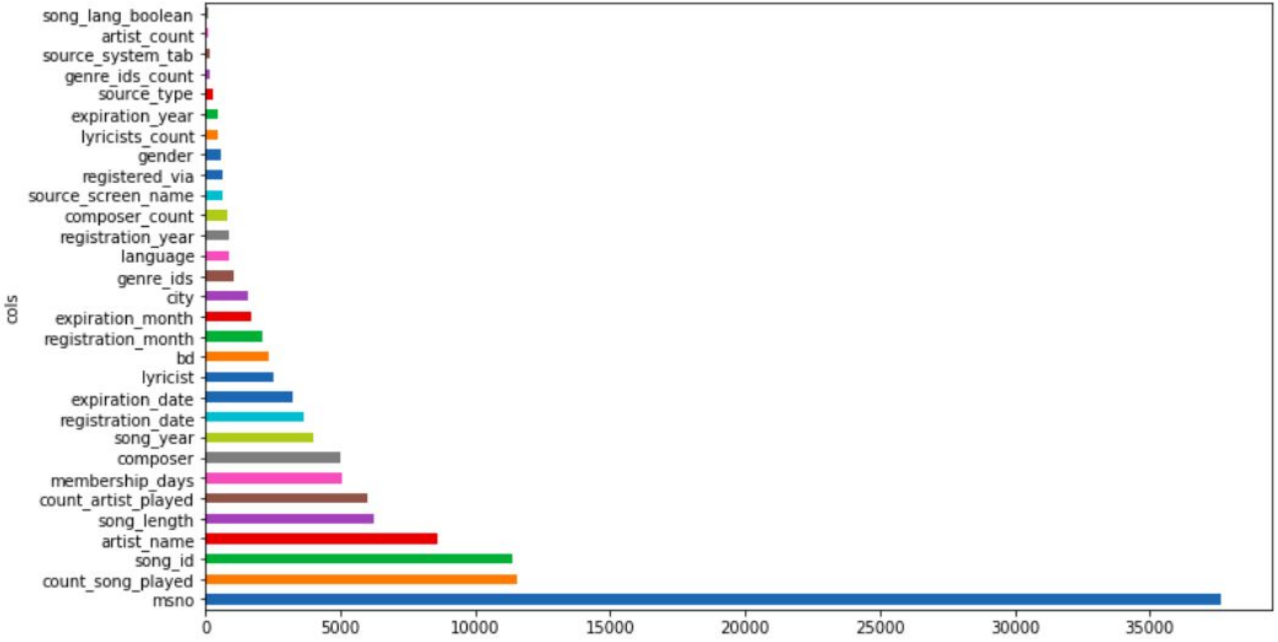
Important Features Random Forest



LGBM Light Gradient Boosting Machine



Important Features LightGBM

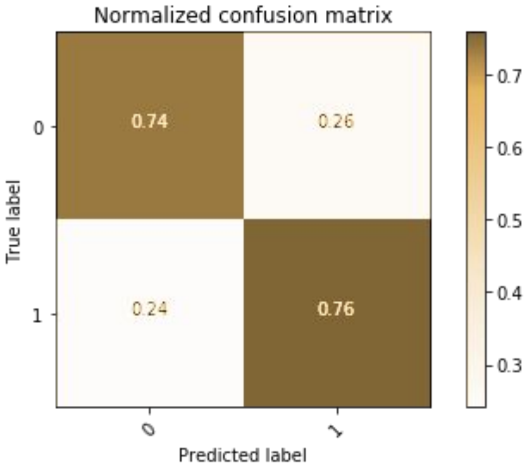


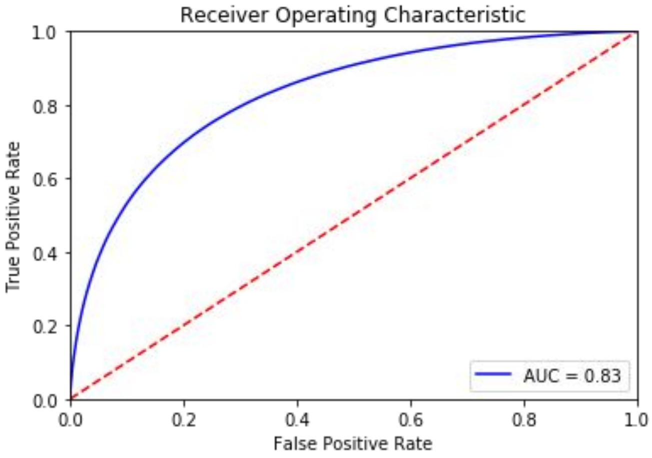
4

Model Evaluation

Confusion Matrix, ROC

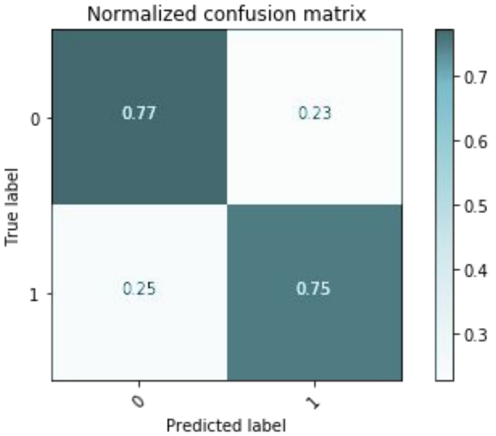
Evaluation Metrics Random Forest

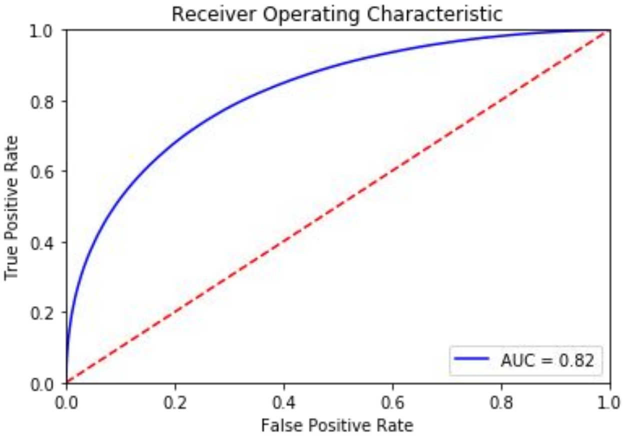




Metrics	Score
Precision	**75.2%**
Recall	**75.3%**
f-score	**75.2%**
Accuracy	**75.1%**

Evaluation Metrics LGBM





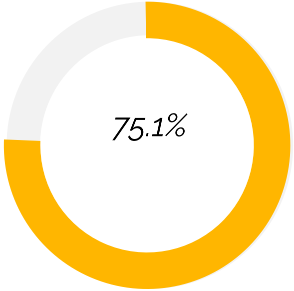
Metrics	Score
Precision	**74.7%**
Recall	**73.8%**
f-score	**74.3%**
accuracy	**74.1%**

5

Conclusion

Best Model, Key Learnings, Demonstration

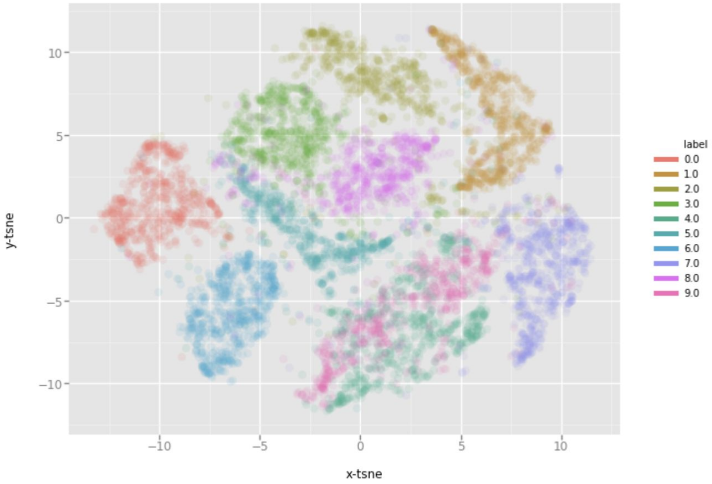
Random Forest BEST MODEL

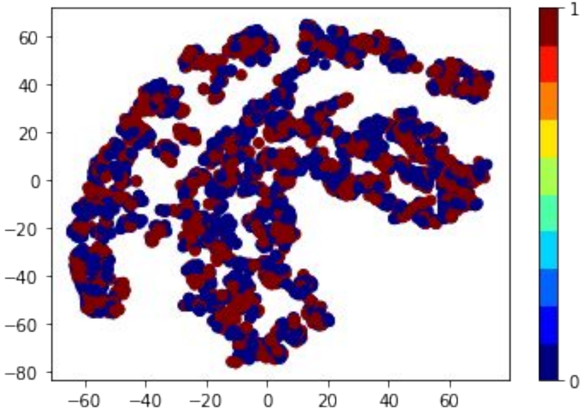


Mean Accuracy

Random Forest

tSNE dimensions colored by digit





Let's **review** our learnings

Machine Learning	Business Domain

Data

- → More data is better
- → Sample the data for Grid Search

Feature Engineering is important

- → Dropping features manually for regularization
- → Different feature engineering for different model **(No Free Lunch)**

Ensembling

- → Ensembling works if models are different

Feature

- → Even counts are important
- → User level Personalization is required

Model

- → Find model performance correlations
- → Select fewer important parameters
- → Feature importance helps interpretation
- → Make hypothesis and **demonstrate**

Demonstration

Thank You

Any questions?

Presented by ***The Placeholders***