# Predicting the success factors of a song

Sophie De Becker
Yao Di
Mattia Gallese
Giacomo Martiriggiano

11th May, 2020

# PRESENTATION OVERVIEW

SONY®

1. Business Context
2. Data Collection and Exploration
3. Classification Problem
4. Regression Approach
5. Conclusions

# The Global Music Industry

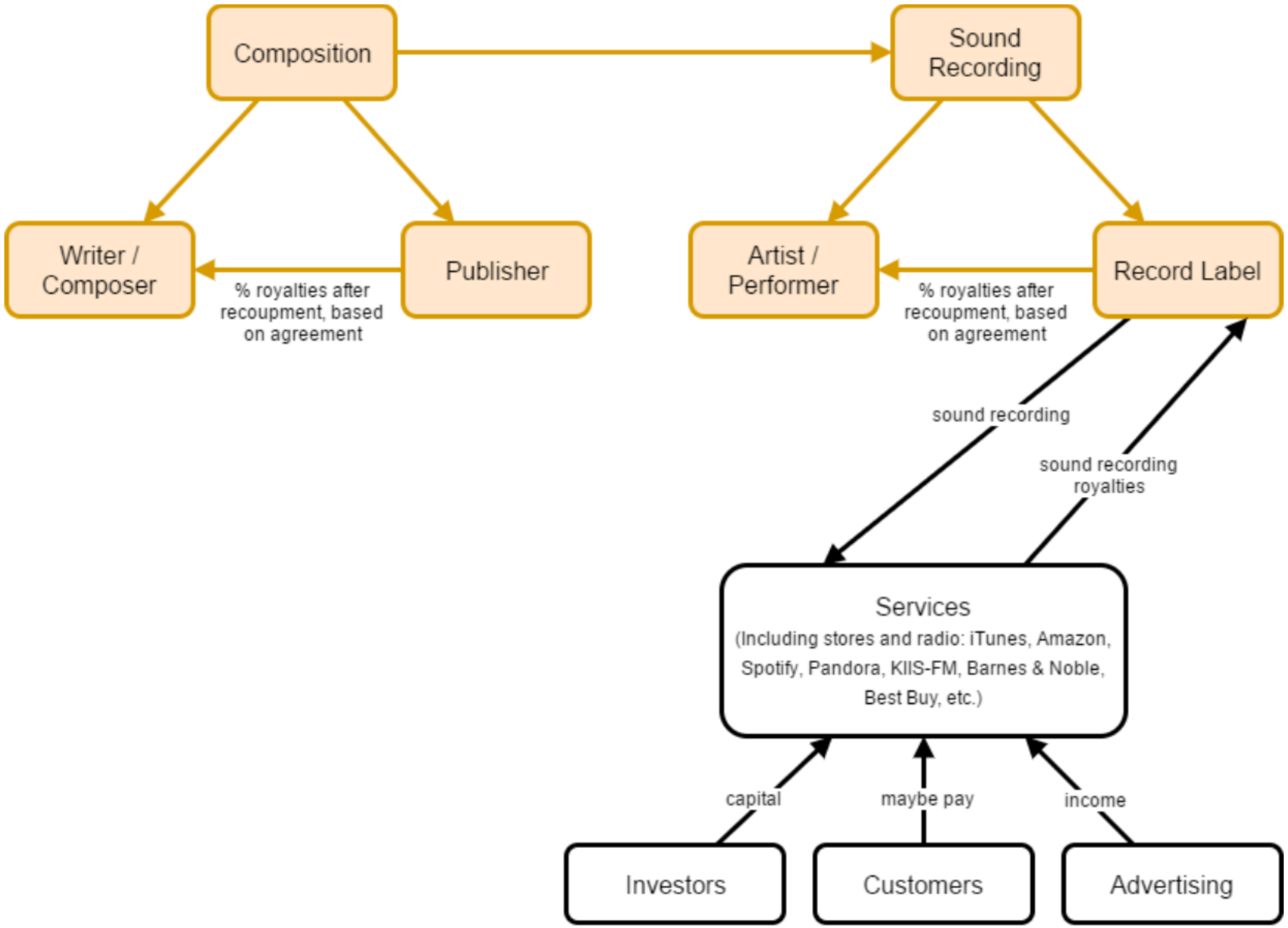
## Industry Figures

19.1 B USD in 2018
Highest Value Ever

59% of Value from
Digital (11.2 B USD)

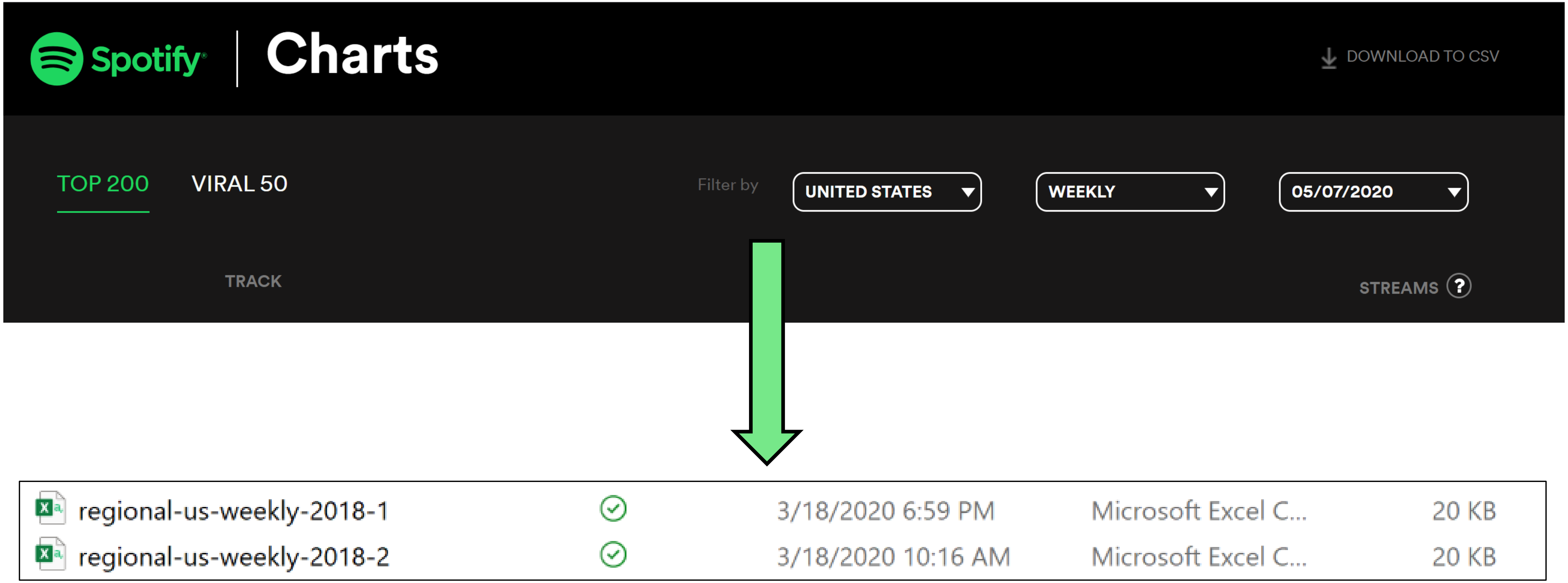
Large Social Impact

# Stakeholder's Interactions


Composition
Sound Recording
Writer / Composer
Publisher
% royalties after recoupment, based on agreement
Artist / Performer
Record Label
% royalties after recoupment, based on agreement
sound recording
sound recording royalties
Services
(Including stores and radio: iTunes, Amazon, Spotify, Pandora, KIIS-FM, Barnes & Noble, Best Buy, etc.)
capital
maybe pay
income
Investors
Customers
Advertising

© 2017 TheMusicMaze.com

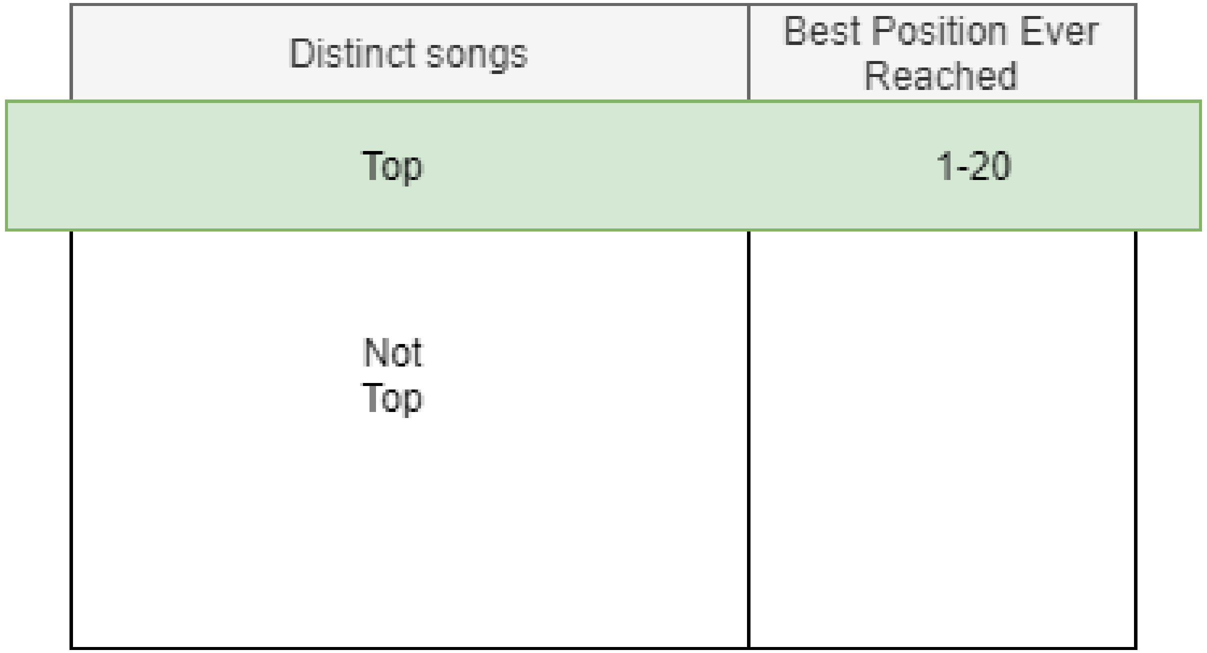
# Data Collection & Error Handling



Filter by UNITED STATES ▼ WEEKLY ▼ 05/07/2020 ▼

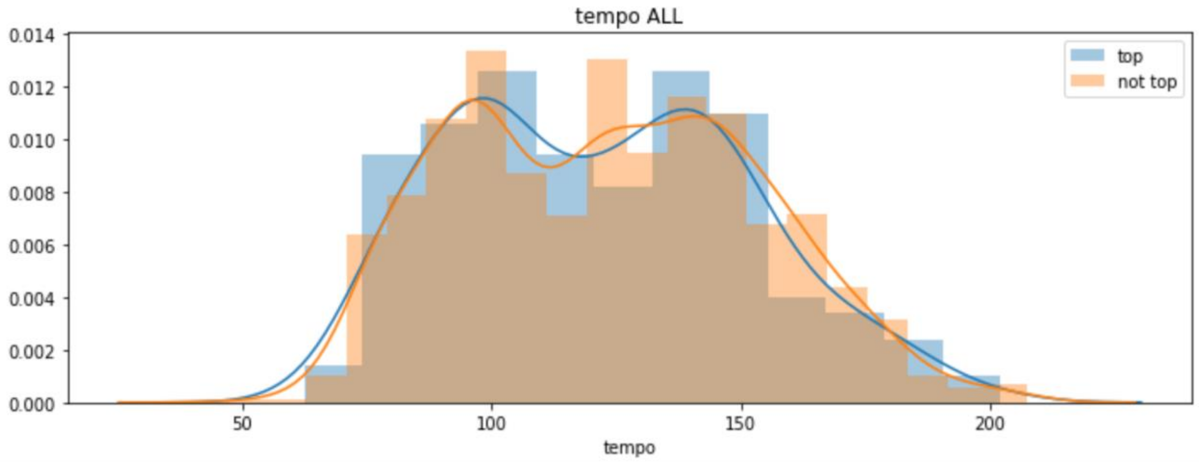
| | | | | |
|---|---|---|---|---|
| regional-us-weekly-2018-1 | ✓ | 3/18/2020 6:59 PM | Microsoft Excel C... | 20 KB |
| regional-us-weekly-2018-2 | ✓ | 3/18/2020 10:16 AM | Microsoft Excel C... | 20 KB |

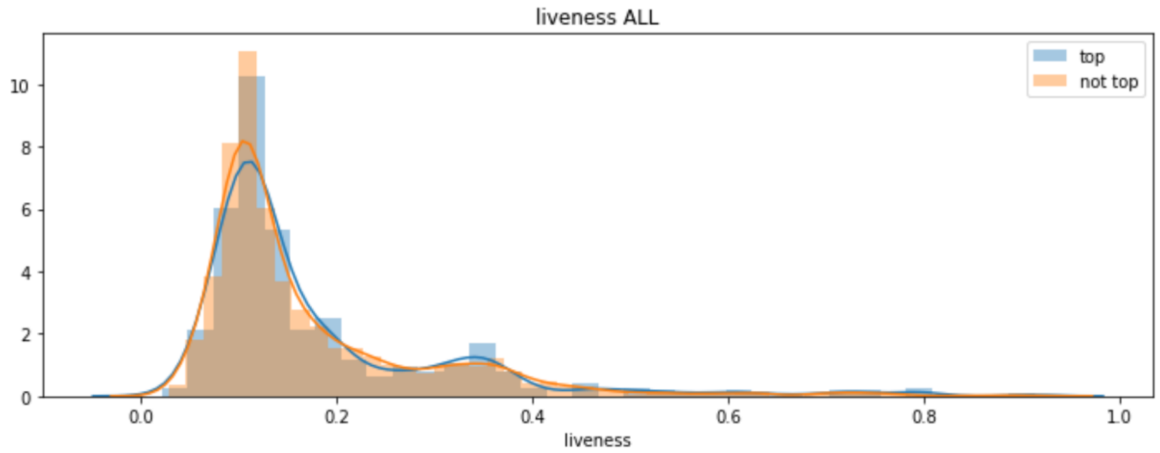
# Success Definition

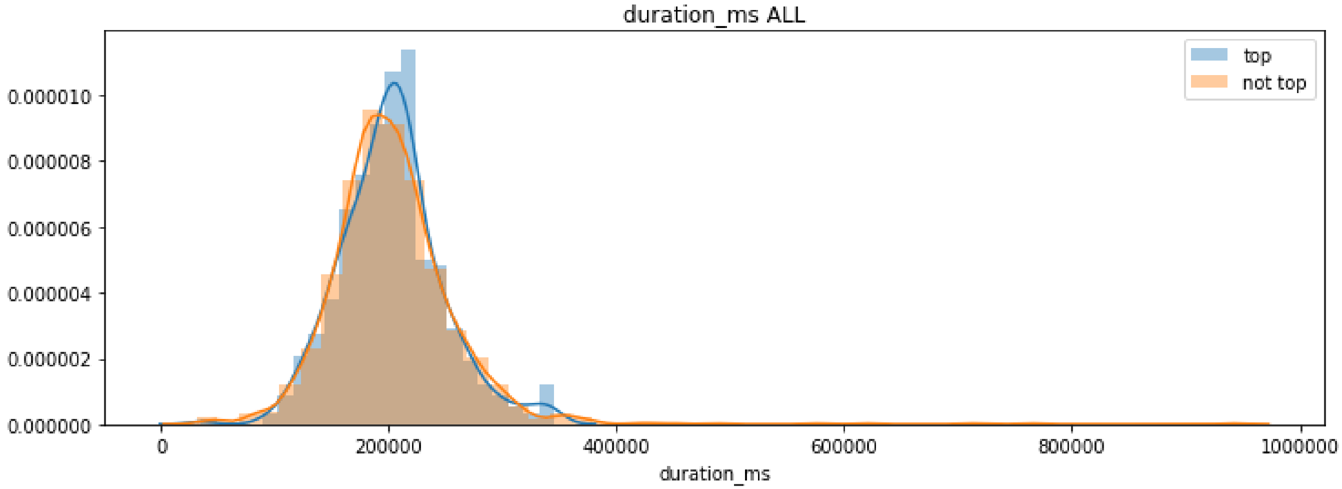


| Distinct songs | Best Position Ever Reached |
|---|---|
| Top | 1-20 |
| Not Top | |

# Exploratory Data Analysis – Continuous Variables

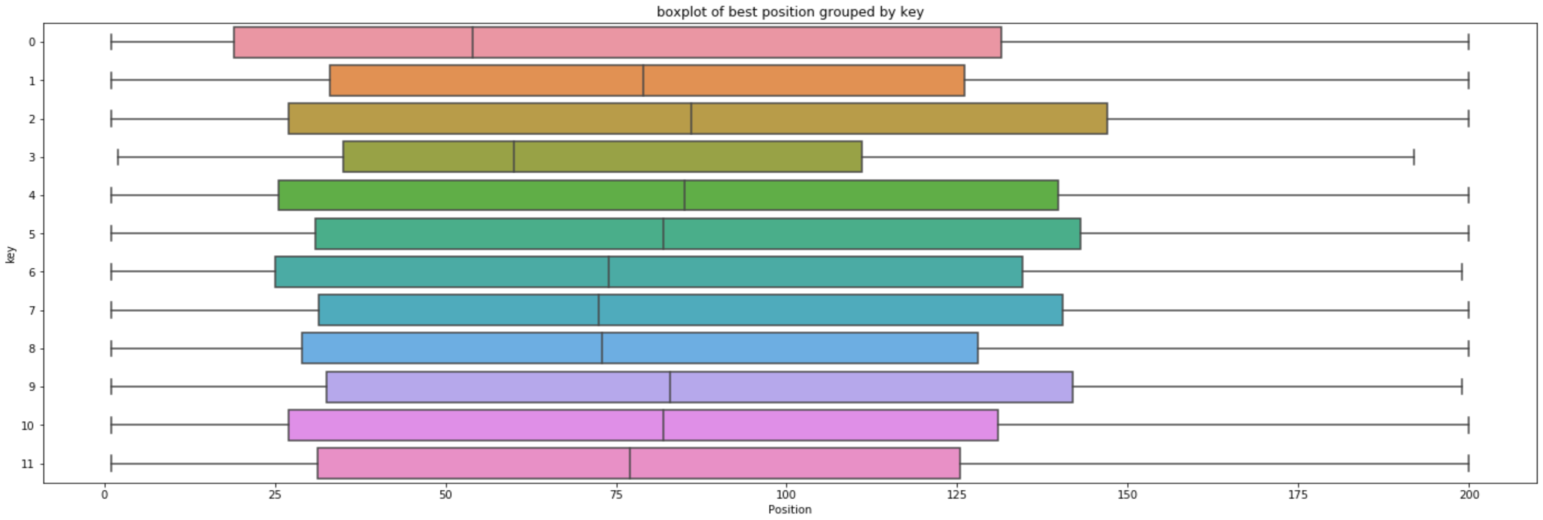




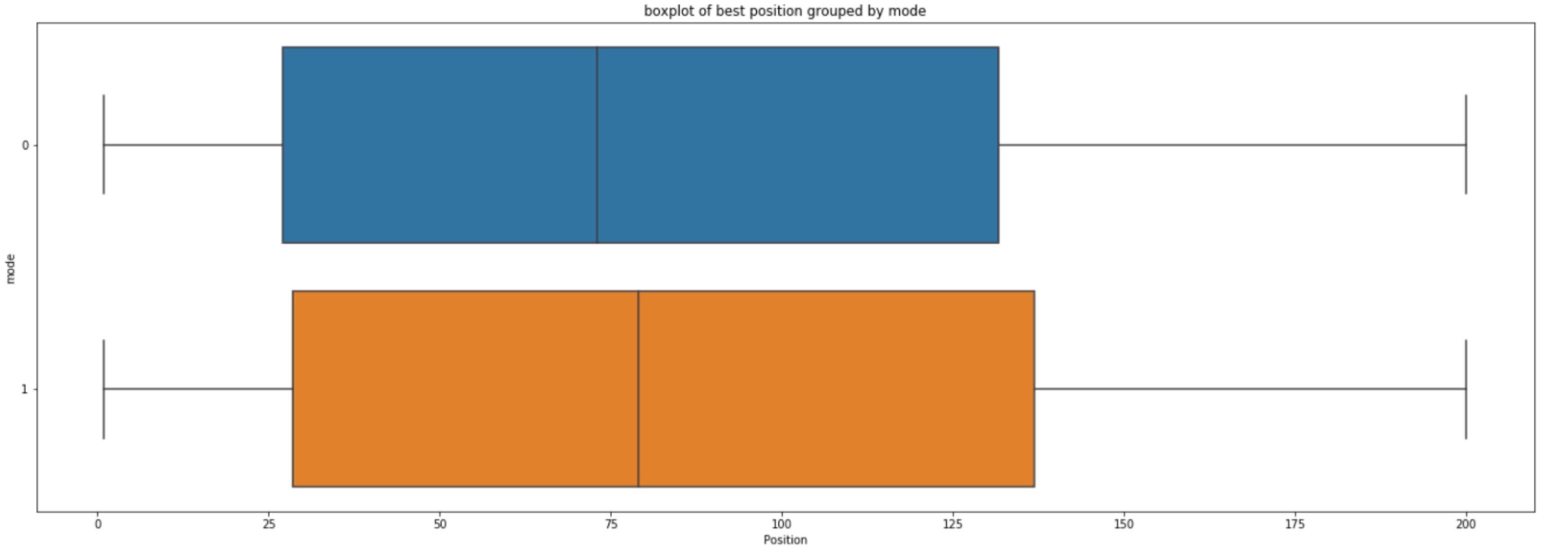


# Exploratory Data Analysis – Discrete Variables

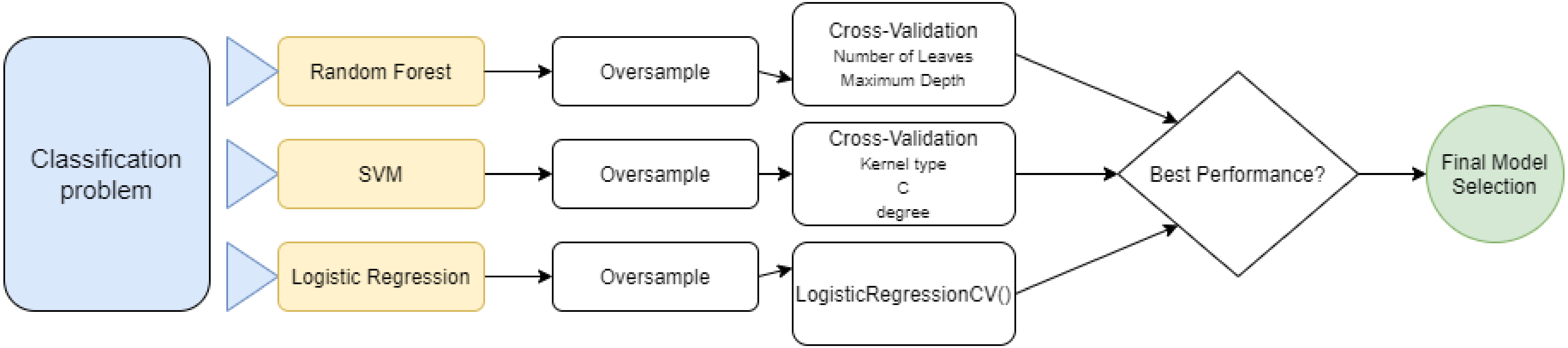
Key has an **impact**



Mode has **no significant** impact



# Classification Problem Process Flow



Classification problem

Random Forest

SVM

Logistic Regression

Oversample

Oversample

Oversample

Cross-Validation
Number of Leaves
Maximum Depth

Cross-Validation
Kernel type
C
degree

LogisticRegressionCV()

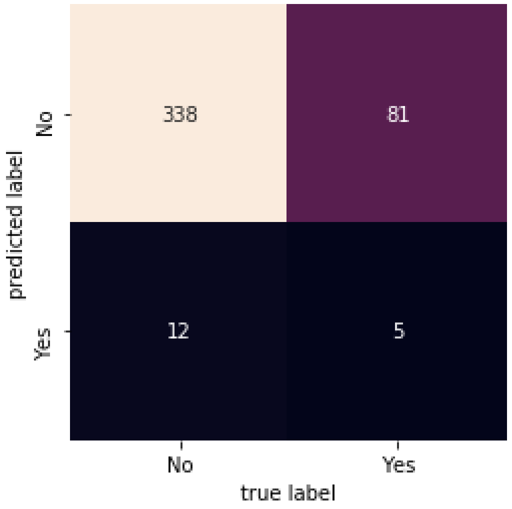
Best Performance?

Final Model Selection

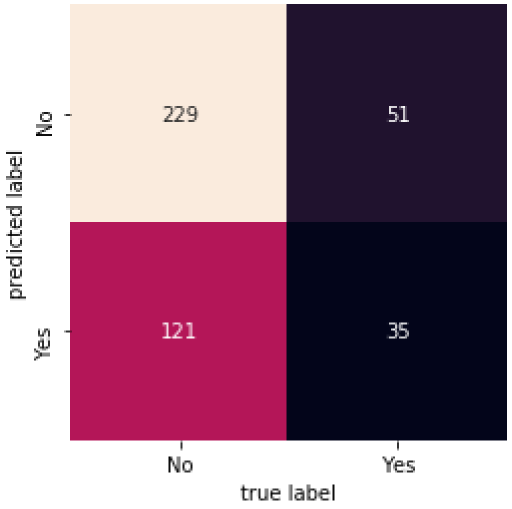
# Classification Problem Results

Final model selection

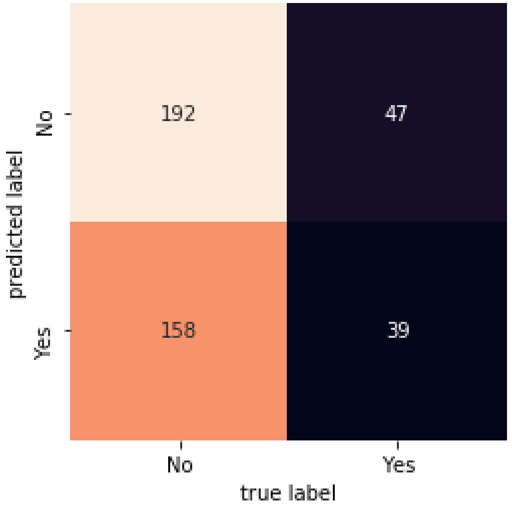
Random Forest



SVM



Logistic Regression

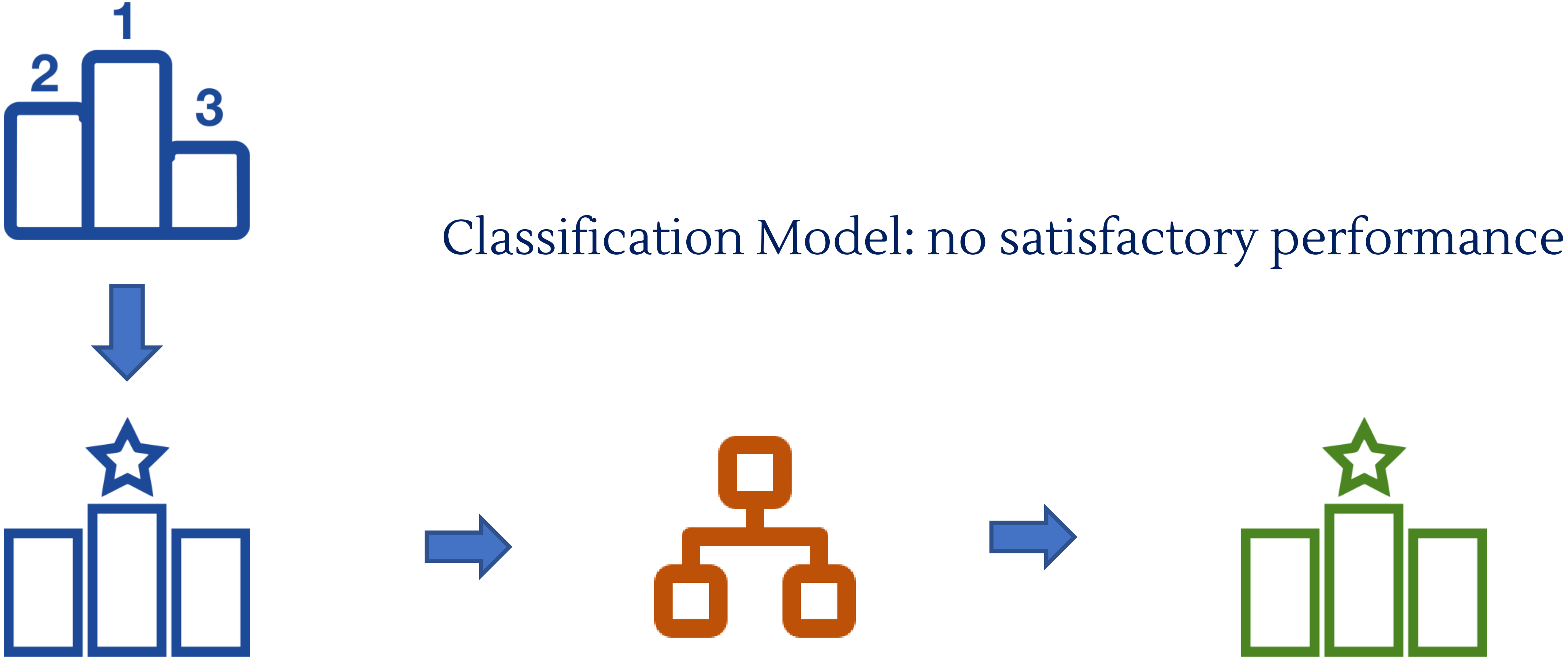


F1 score:  0.097

F1 score: 0.289

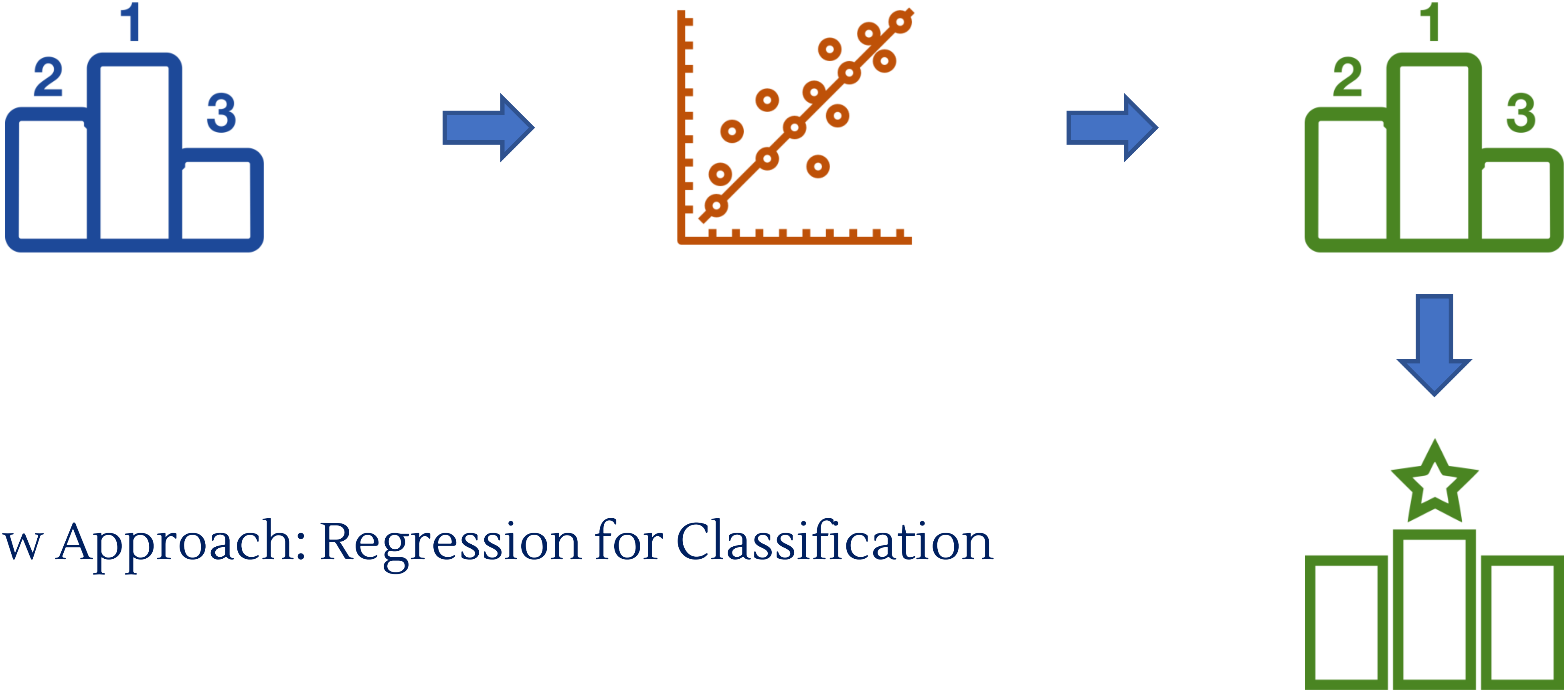
F1 score: 0.276

# Success Prediction - Another Approach



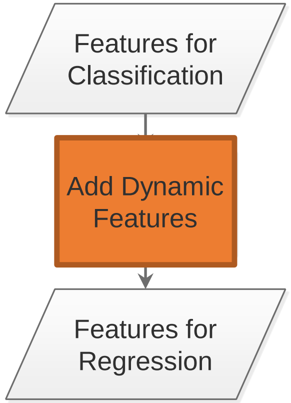
Classification Model: no satisfactory performance

# Success Prediction - Another Approach



New Approach: Regression for Classification

# Feature Engineering for Regression Model



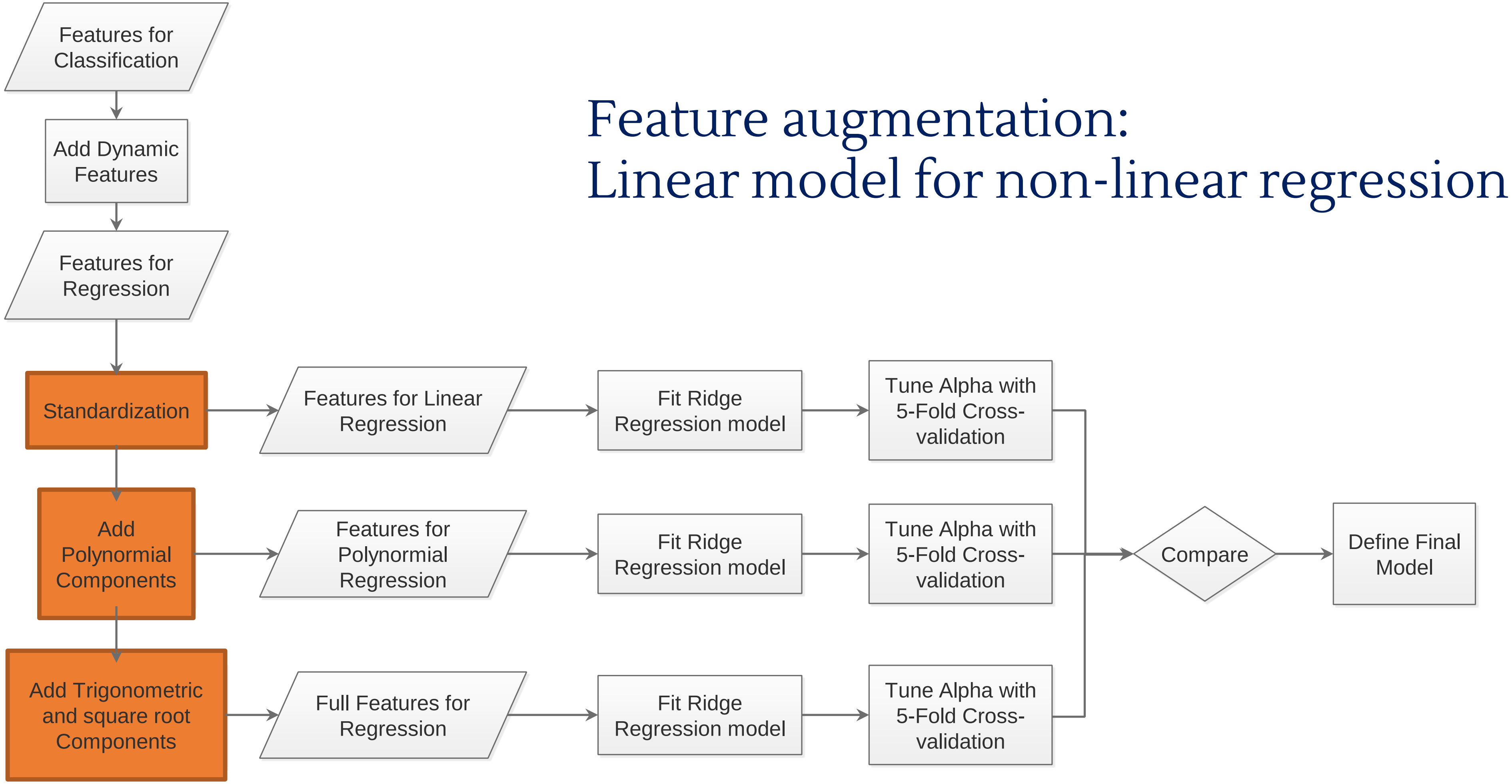
Features for Classification

Add Dynamic Features

Features for Regression

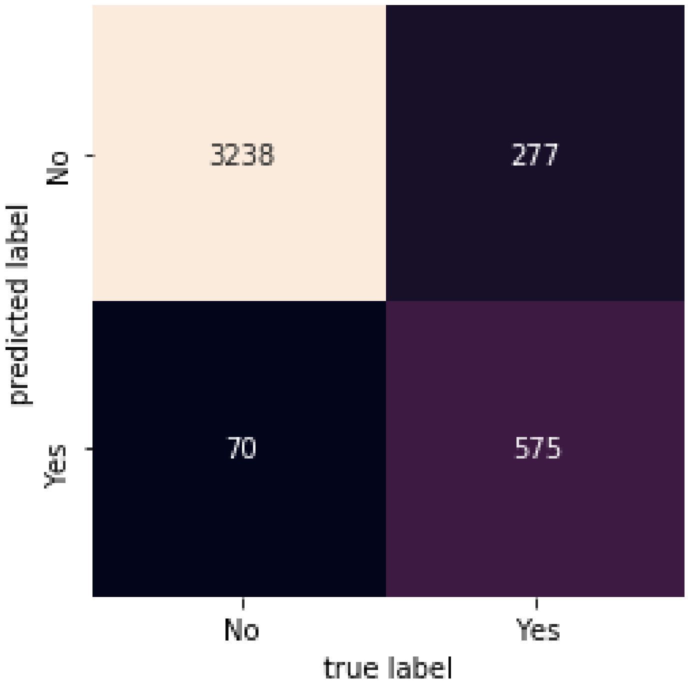
Dynamic features: impact of historical data

# Feature Engineering for Regression Model


Features for Classification
Add Dynamic Features
Features for Regression
Standardization
Add Polynormial Components
Add Trigonometric and square root Components
Features for Linear Regression
Features for Polynormial Regression
Full Features for Regression
Fit Ridge Regression model
Fit Ridge Regression model
Fit Ridge Regression model
Tune Alpha with 5-Fold Cross-validation
Tune Alpha with 5-Fold Cross-validation
Tune Alpha with 5-Fold Cross-validation
Compare
Define Final Model

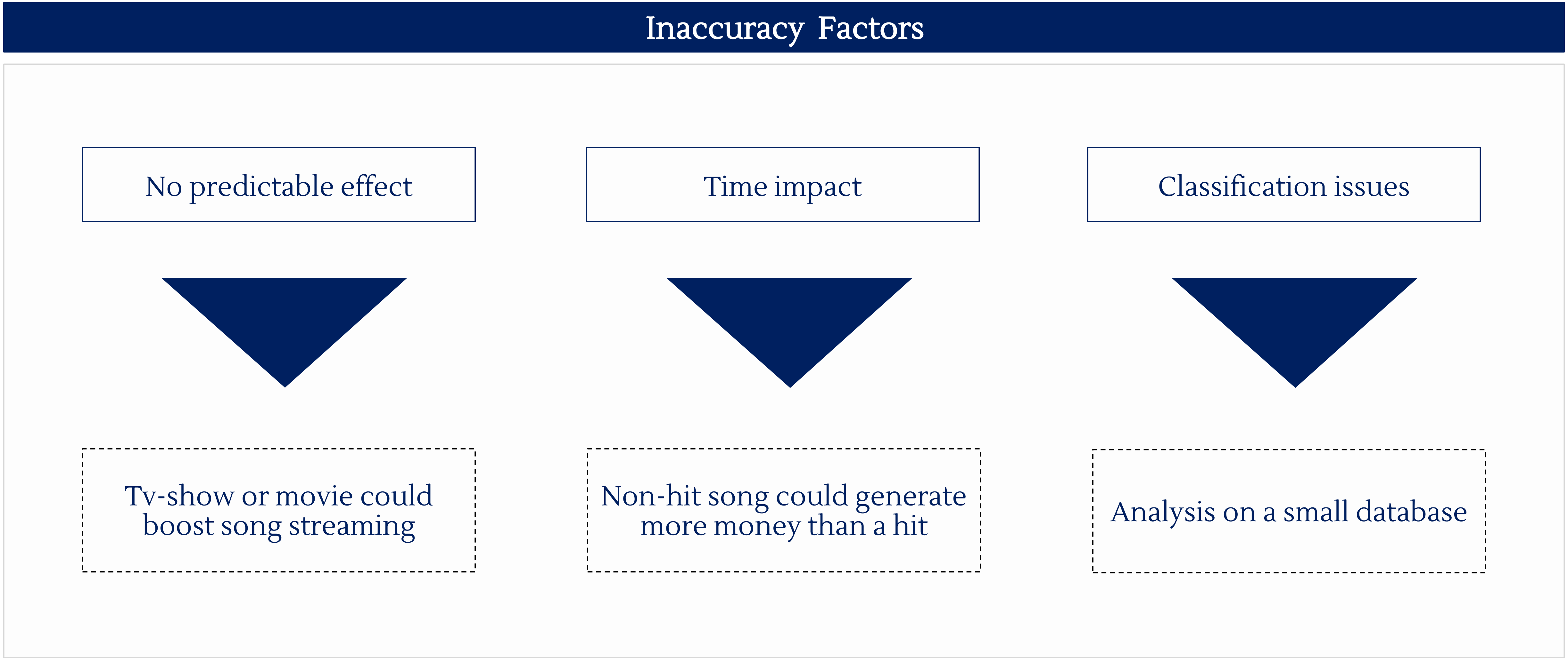
## Feature augmentation:
## Linear model for non-linear regression

# Result for Regression Model



| Metric Type | Metrics Score |
|---|---|
| Accuracy | 0.92 |
| Recall | 0.67 |
| Precision | 0.891 |
| F1 | 0.77 |

# Business Insights



Inaccuracy Factors

No predictable effect

Tv-show or movie could boost song streaming

Time impact

Non-hit song could generate more money than a hit

Classification issues

Analysis on a small database

# Conclusion

1. Difficult Industry to Model (As Expected)

2. Easier to Predict Evolution Based on Historical Data

3. There is Room for Improvement