# Kickstarter Dataset…

## … when a crowdfunding campaign will be profitable?

Powered by Apache Spark™

DDAM Course Project
A.Y. 2023/2024

Eva Manai
Paolo Andriani
Michele Velardita
Andrea Napolitano
Steffania Sierra Galvis

# THE TEAM

Andriani
Paolo

Computer Science

Manai
Eva

Computer Science

Napolitano
Andrea

Computer Science

Sierra Galvis
Steffania

Mathematics

Velardita
Michele

Computer Science

# Kickstarter Dataset

## Exploring the Platform Landscape

Kickstarter is a dynamic global crowdfunding platform dedicated to bringing creative projects to life.

With **$1.9 billion in pledges** from **9.4 million backers**, it has fueled **257,000 diverse projects**, from films to technology.

It serves as the driving force behind transforming creative visions into reality worldwide...

**... but not every project is able to reach the goal...**

# Kickstarter Dataset

## The Purpose

Only about **35%** of the total projects have raised **successful** fundings

Then, which projects can successfully achieve their goal?

Are there any **key project characteristics** that increases the chances of success?

# WORKFLOW

**01 Exploration**
Data characteristics

**02 Feature Extraction**
Increase descriptive power of data

**03 Multivariate Analysis**
Distribution, correlation analysis

**04 Title Analysis**
How title can contribute to success

**05 Preparation**
To run selected models

**06 Classification & Regression**
DT, RF, GBT, NN, LinReg, LogReg,

**07 Feature Importance**
Measured by models

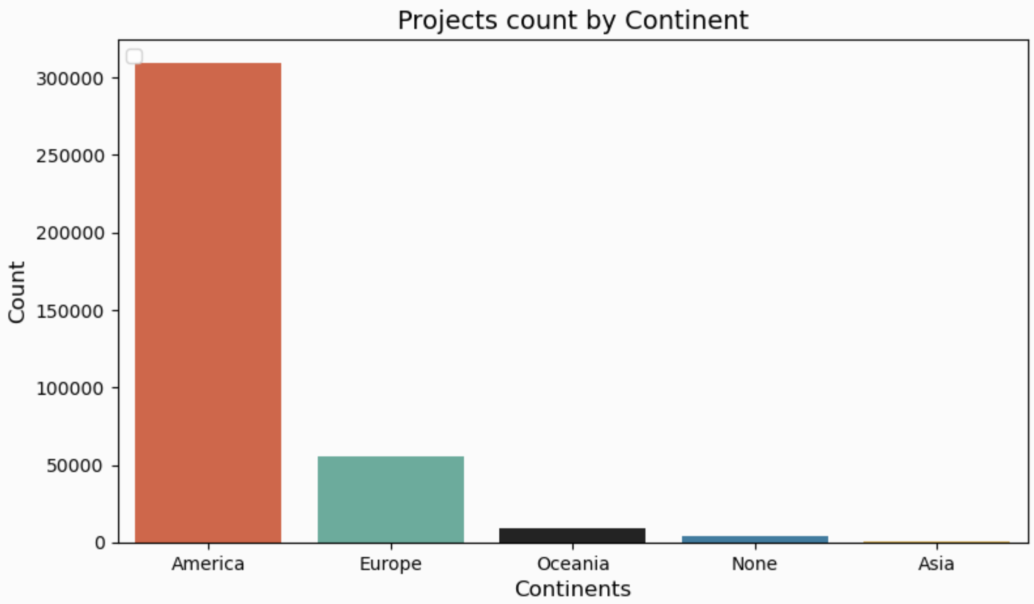
**08 Conclusions**
What we get

# Feature Extraction

- **Launch-Date**
  **Deadline**
  - year
  - month
  - day of launch
  - time interval
- **Title**
  - length
  - use of "?!"
- **Country**
  - continent

# CONTINENT EXTRACTION & ANALYSIS

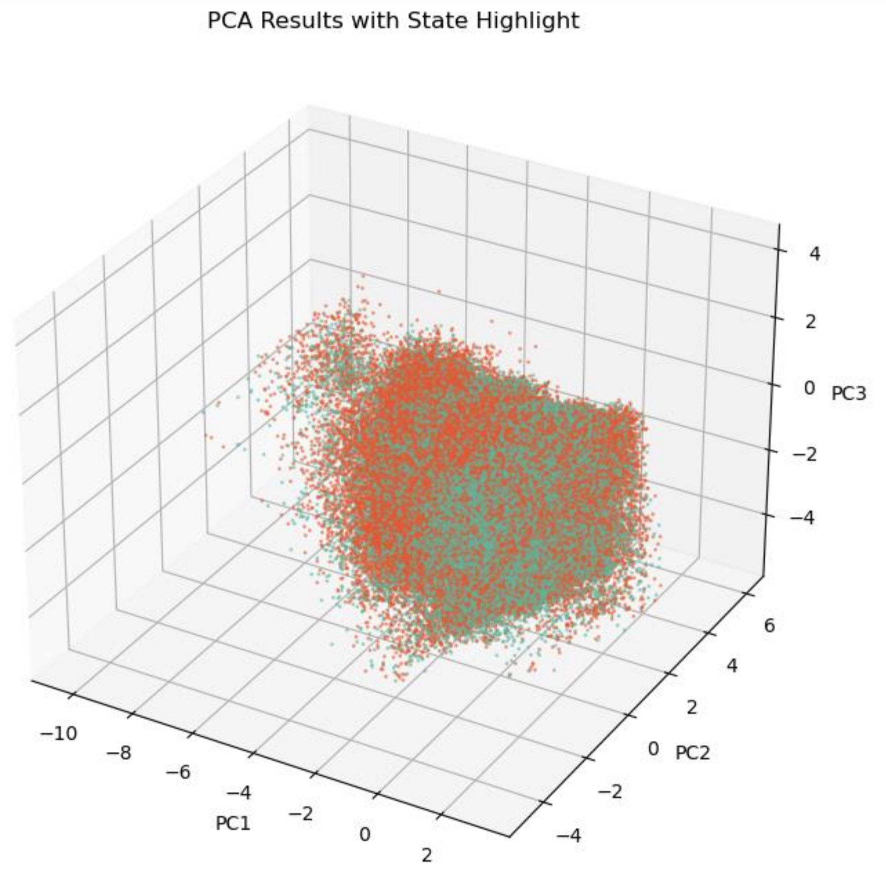
Looking for the continent with the highest number of projects




- America dominates over other continents
- Data are from Kickstarter, so these proportion don’t match reality

# PCA - 3 Components

**Features used**
Goal - Year - Month -
Day of week - Time interval -
Length of title - Use of ?! -
Category - Main Category -
Currency - Continent



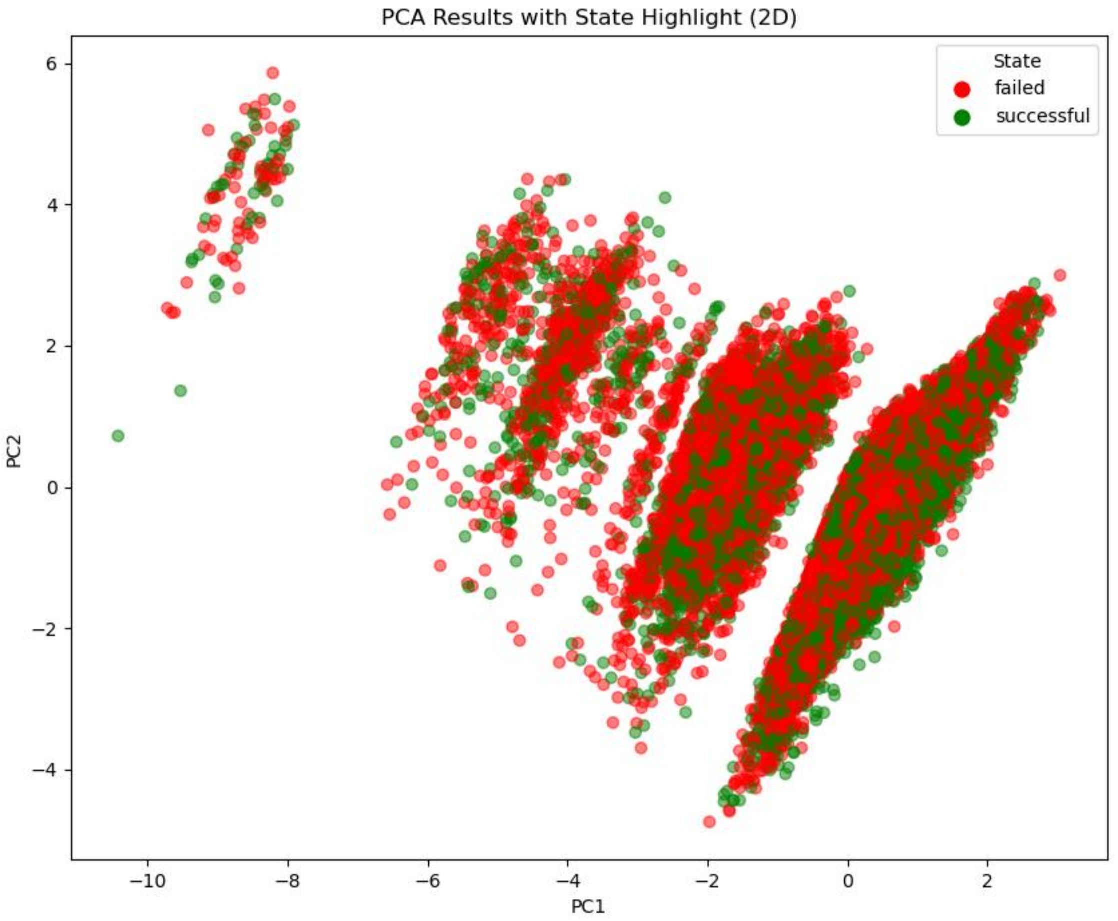

Fail

Success

**Not very interesting...**

# PCA - 2 Components

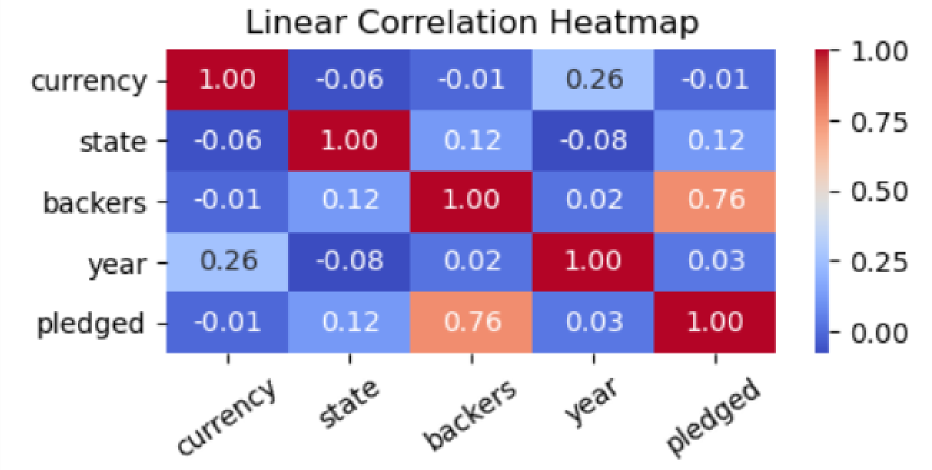
**Features used**
Goal - Year - Month -
Day of week - Time interval -
Length of title - Use of ?! -
Category - Main Category -
Currency - Continent



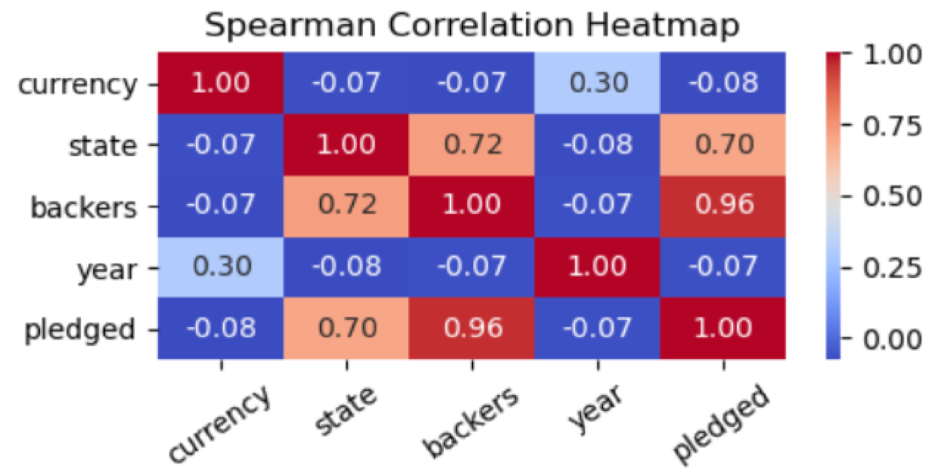

We can notice a **clear separation** between some areas of the points

Maybe some clustering technique can explain them better

# Correlation Analysis

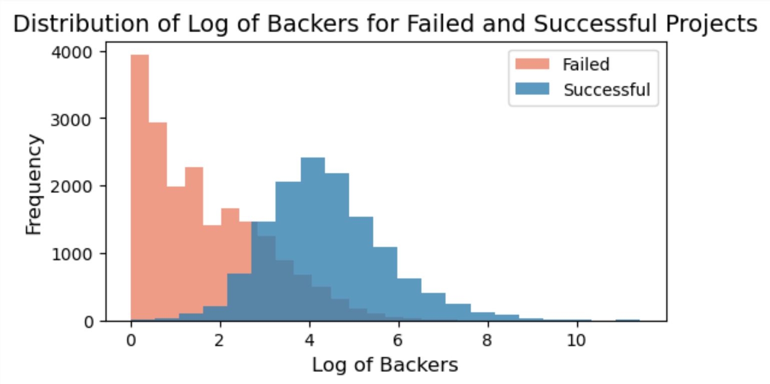


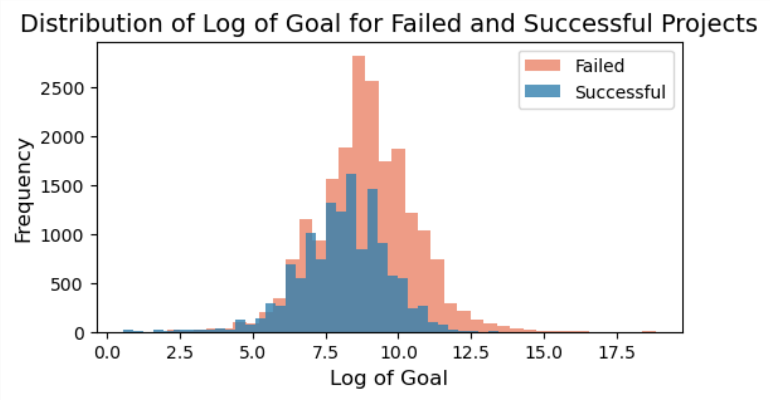

Which variables are **linearly correlated**?




Can we find other variables **non-linear correlated**?

# Correlation Analysis - Insight




**0.72** correlation with **number of backers**

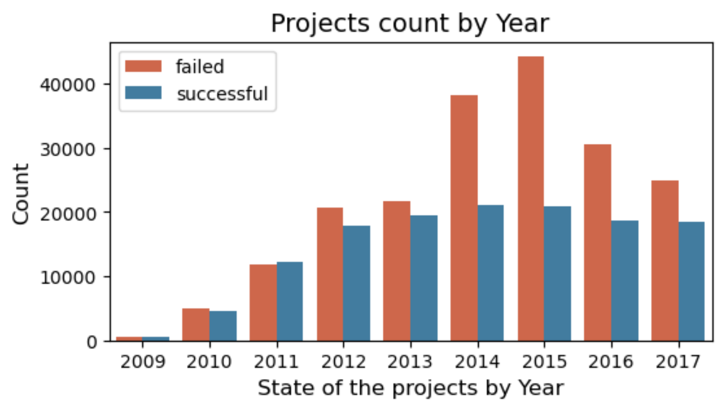
More supporters raise the probability of success

**-0.22** correlation with the **goal**

Very high fundraising goals are difficult to achieve

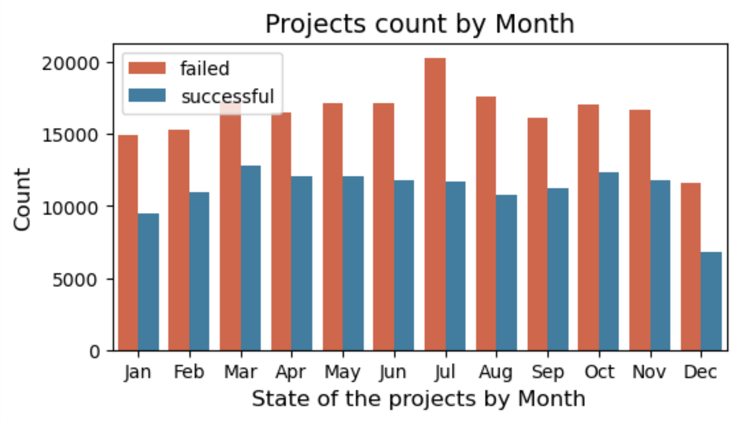
# Temporal Analysis

How the **period** and the **duration** of a campaign influence the outcome?



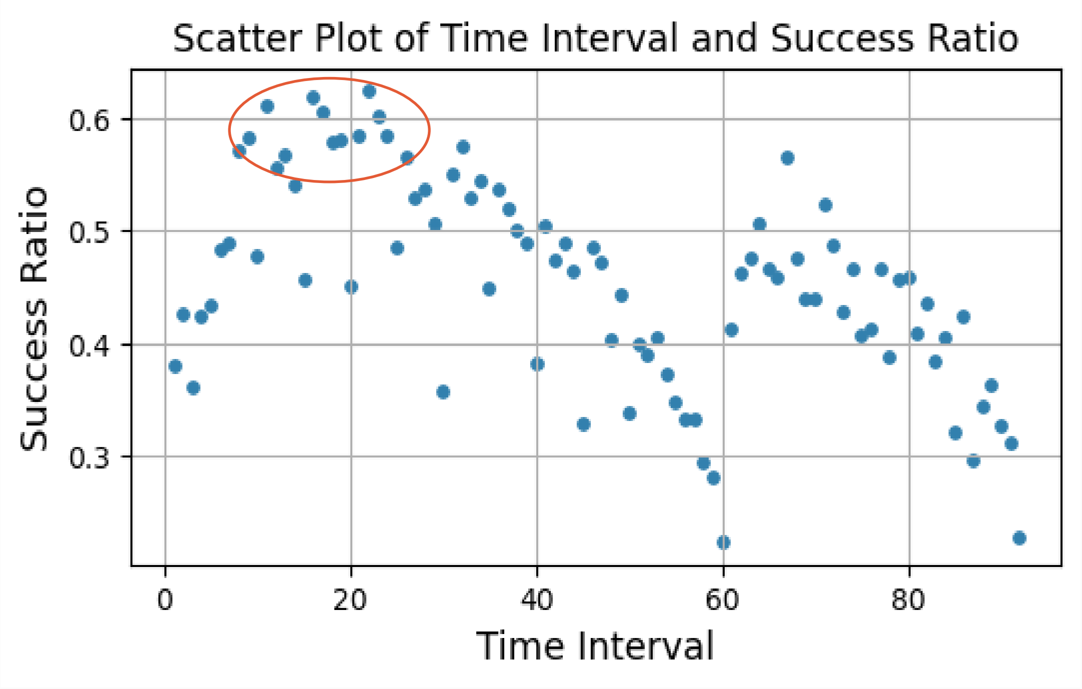

Crowdfunding platforms have followed a growing trend

Less interesting information can be extrapolated from a seasonal analysis



# Temporal Analysis

Relation between the **duration** and the **success rate**




- The bivariate distribution shows a bimodal behavior
- A **short-medium** interval seems to be the best choice

# Categories Analysis - A Big Picture

Best and worst categories for success/fail ratio

| Rank | Category | Success/Fail Ratio |
|---|---|---|
| 1 | Dance | 1.9 |
| 2 | Theater | 1.8 |
| ... | ... | ... |
| 14 | Journalism | 0.32 |
| 15 | Technology | 0.31 |

niche but very attractive
(3500 projects)

more frequent but less attractive
(27000 projects)

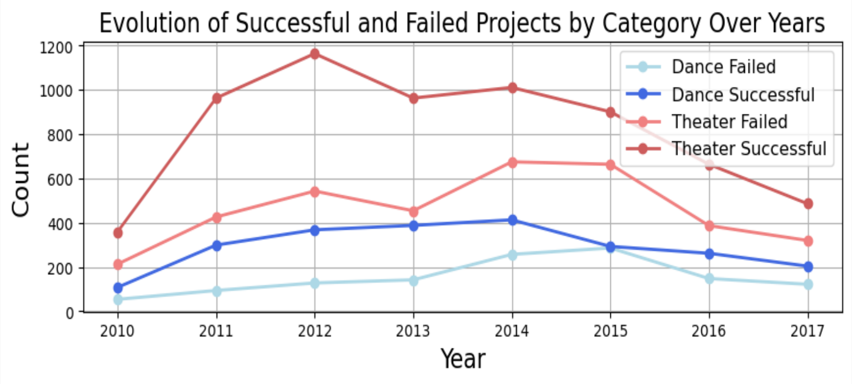
Niche categories gather more support while the more widespread ones have a lower success rate

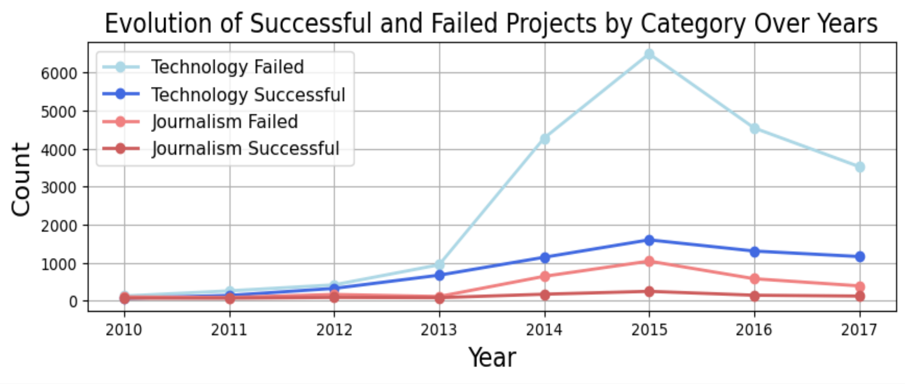
# Categories Analysis - Trend

Examining category success growth over the years, distinguishing successful from unsuccessful, two patterns emerge:

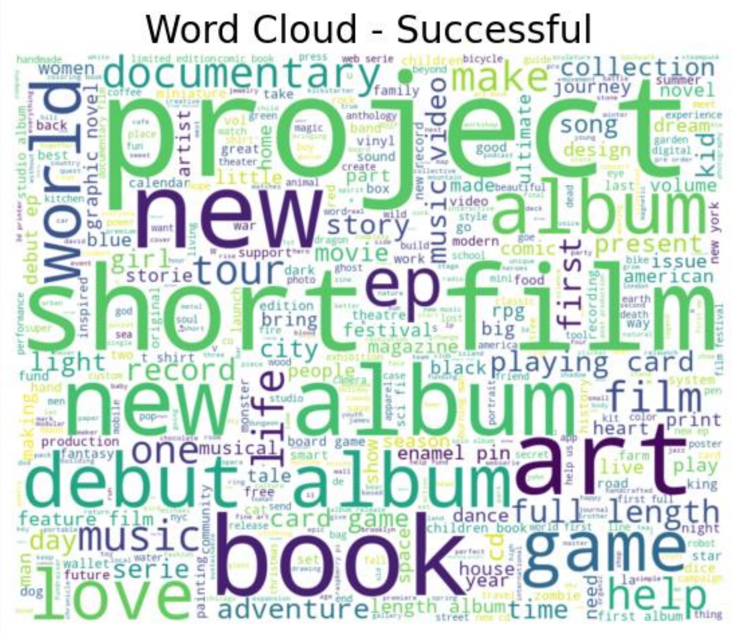
Less frequent categories remain stable over time, maintaining a consistent success/failure ratio

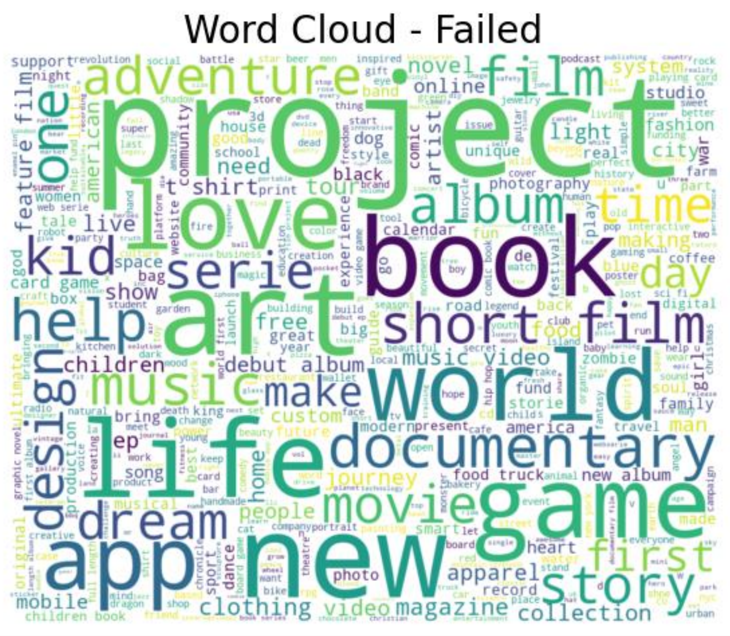
Popular categories align with the overall trend each year: failures exceed successes in growth



# Title Analysis – most used words



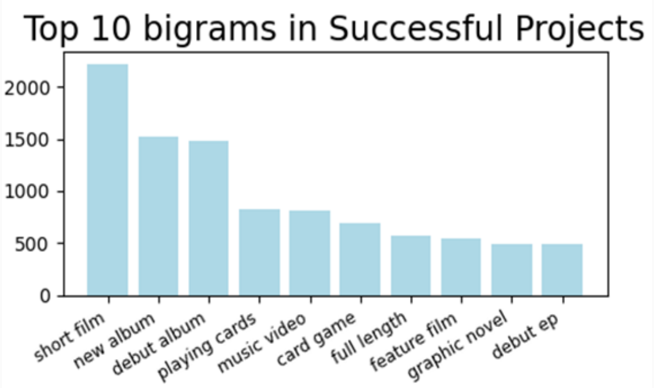



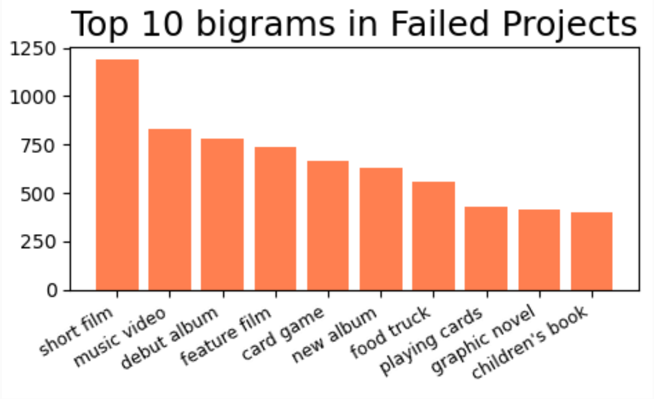
# Title Analysis - Bigrams

Bigrams are two consecutive words
that appear in a sentence

Similar bigrams but different frequencies

| Bigram | Success rate | Fail rate |
|---|---|---|
| new album | 70.8% | 29.2% |
| debut album | 65.4% | 34.6% |
| short film | 65.1% | 34.9% |



# Classification - performance

- Hyperparameter tuning: Grid search with Cross Validation (4 folds)
- **Backers, Pledged** are a posteriori info: excluded from input features

| MODEL | F1-score | AUROC | Accuracy |
|---|---|---|---|
| Decision Tree | 0.56 | 0.65 | 0.67 |
| Gradient Boosting Trees | 0.68 | 0.75 | 0.69 |
| Neural Network | 0.63 | 0.68 | 0.64 |
| Logistic Regression | 0.59 | 0.65 | 0.62 |

# Classification - Insight

**Neural Network**

Lack of flexibility in hidden's activation functions

**DT & GBT**

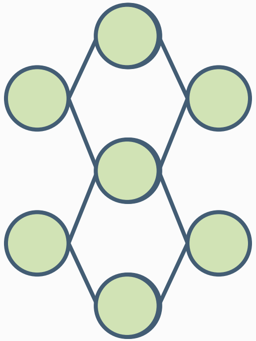
Few descriptive features.
Shallow Tree structure

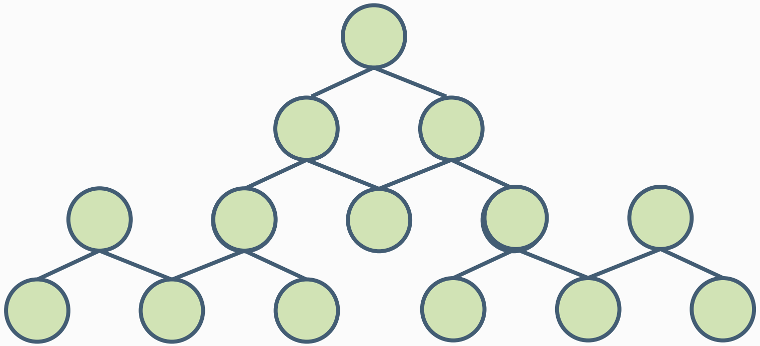
parameters

Hidden layers: (6,4,2)
stepsize = 0.01

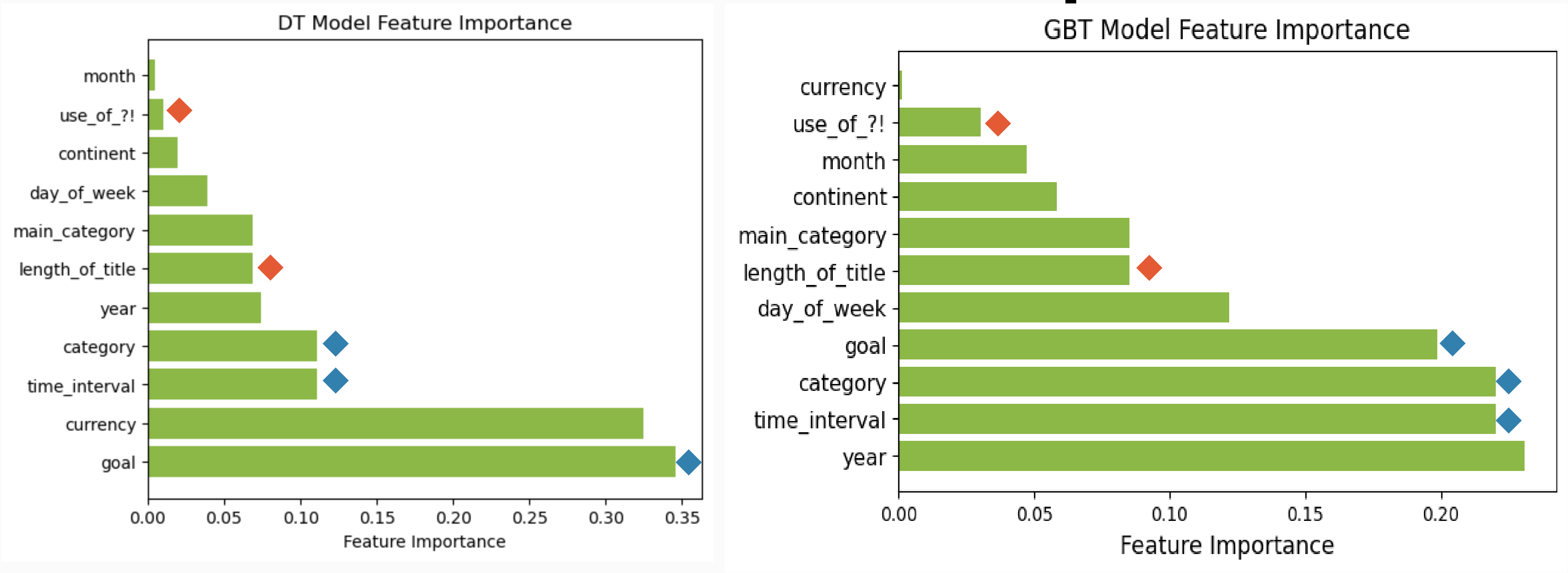
DT
depth = 15
MinInstancePerNode = 80

GBT
depth = 5
stepsize = 0.2





# Classification - Feature Importance




- **goal, time_interval, category** are the most important for both models
- **Feature extracted from title** don't influence the outcome of a project

# Linear Regression

Features not considered:
**continent** and **backers**

**Performance**

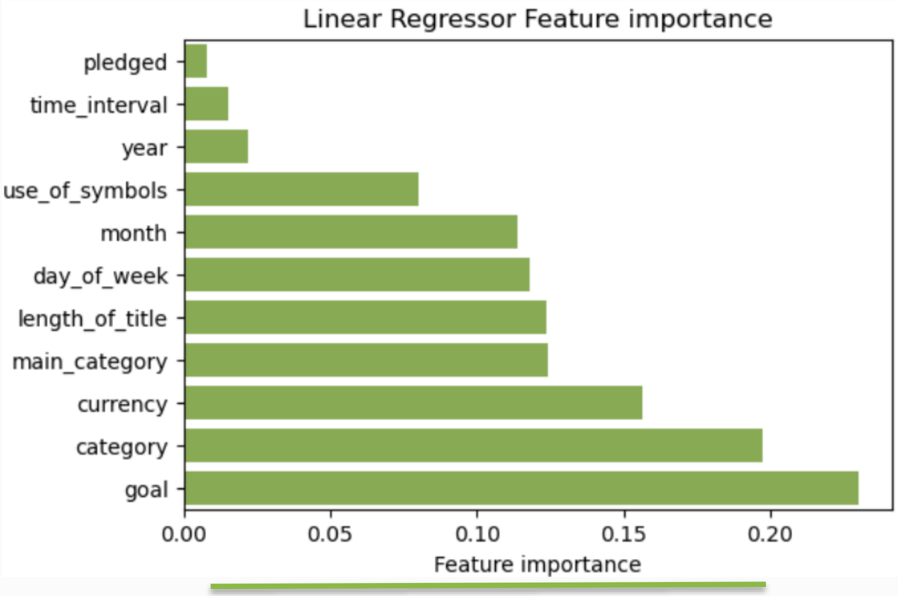
| | |
|---|---|
| RMSE | 0.36 |
| R^2 | 0.48 |

Features not considered:
**continent, backers** and **pledged**

Additional parameter:
**regularization value**.

**Performance**

| | |
|---|---|
| RMSE | 0.49 |
| R^2 | 0.004 |




**Feature importance**

# Conclusion

Our analysis tells an insightful story about crowdfunding projects, from 2011 to 2018

- **Annual and seasonal trends** reveal significant shifts in the ratio between fundraisers and backers
- Some **categories outperform**, particularly niche ones (Dance, Theater and Comics)
- The title words and punctuation do not influence donors
- Fundraiser tip: **Choose goal and duration wisely**

However, training a model to **predict** campaign outcomes in advance is **challenging**, due to the **limited** availability of project **informations**

# Thanks!

# Questions?