Optimizing Kickstarter Campaign Success

James Geronimo, Emily Doami, Michael Florip

DATA 144 Final Project Presentation

AGENDA

1. Introduction
2. Dataset
3. Methods
4. Takeaways

01

Introduction

What is Kickstarter?

- **Crowdfunding platform** where individuals or organizations can raise funds for creative **projects**
- Projects range from personal art creations to large-scale film productions
- Campaigns are **time-bound** and **goal-driven**
    - **Success**: Achieved by **reaching the funding goal** within the set timeframe
- **All-or-Nothing**: No money is exchanged for failed campaigns
- Kickstarter earns a **5% commission** from successful projects only

How does Kickstarter work?

- **Discovering Campaigns**

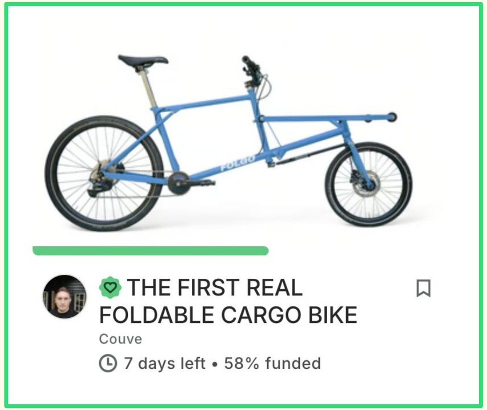
  Kickstarter showcases trending, recommended, and staff-picked campaigns

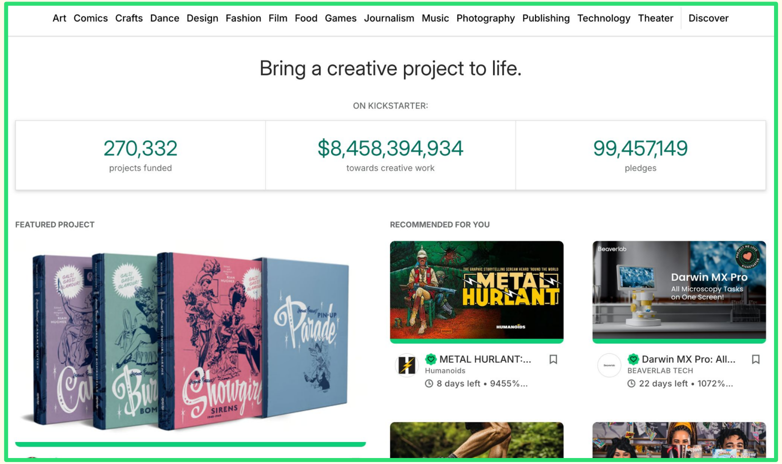
- **Backing a Campaign**

  Users pledge funds via credit card but are only charged if a campaign reaches its goal

- **Funding Process**

  All-or-Nothing Model: Campaign creators only receive funds if the goal is met by the deadline





Challenges for Campaign Success

- **Competition**

  Thousands of campaigns launch every month

- **Marketing**

  Reaching the right audience is crucial for success

- **Emotional Appeal**

  Titles and other information must resonate with potential backers

Project Goal

Identify features and qualities that make a campaign successful

Stakeholders

- **Creators**
  - Ensure that the funding campaign succeeds to acquire the pledged money
  - Maximize potential profits by optimizing project attributes they can set
- **Kickstarter**
  - Maximize campaign success to make 5% commission
  - Understand which attributes contribute most to successful campaigns, and emphasize them to users building campaigns

02

Dataset

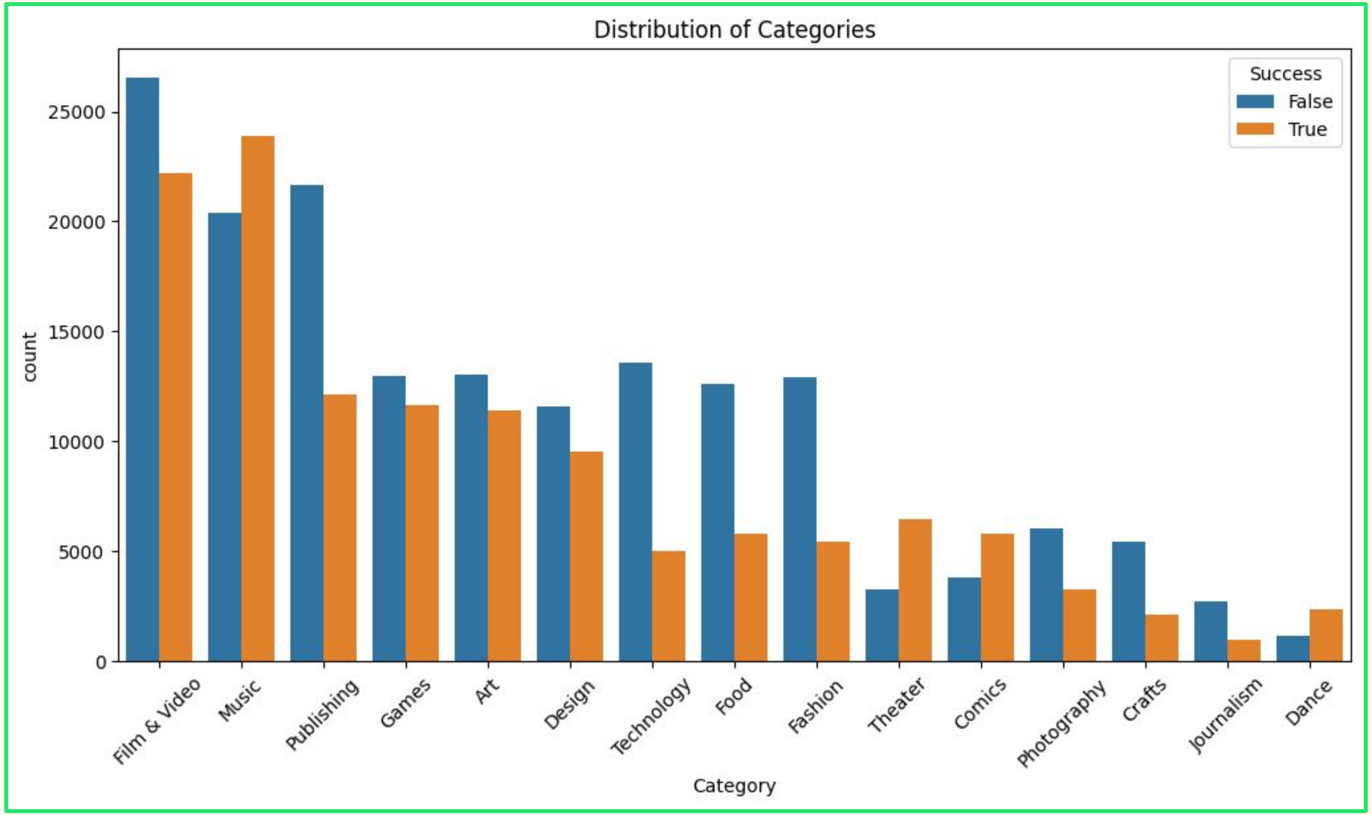
Description

- **Source**: Kaggle (Link)
- **Features (11)**: ID, Name, Category, Subcategory, Country, Launched, Deadline, Goal, Pledged, Backers, State
- **Timeframe**: Apr 2009 - Jan 2018
- **File Size**: 49.22 MB
- **# of Rows**: 370,000+
- **Kaggle Usability Score**: 10.00 / 10.00

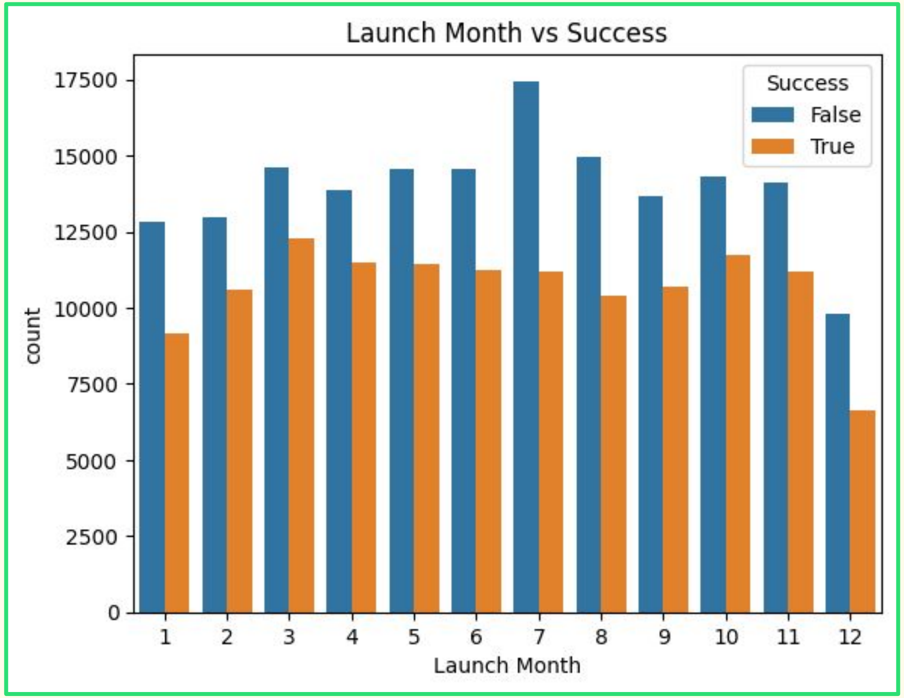
Data Cleaning

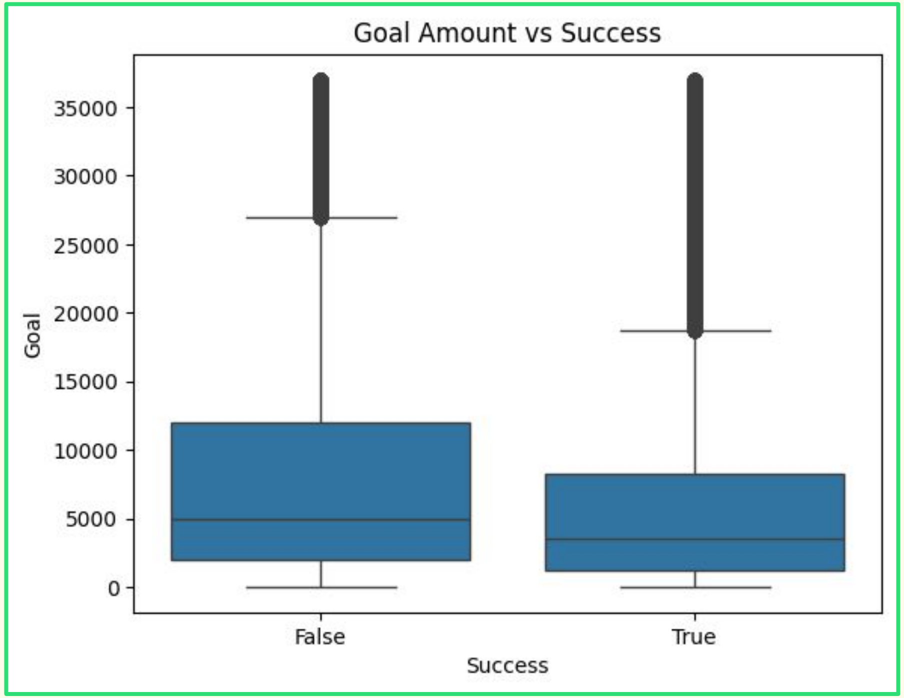
- Removed “**Live**,” “**Suspended**,” and “**Canceled**” Campaigns
  - Only analyzing campaigns that were **Successful** or **Failed**
  - Cannot make assumptions about future donor activity, why a campaign was suspended, or why a campaign was canceled
- Removed “**Goal**” **Outliers**
  - Used the **Interquartile Range (IQR)** method for removing outliers
  - Avoids biasing analysis toward campaigns with significantly smaller/larger goals
- **NaN Values**
  - Dataset had no NaN values to remove

Distribution of Categories Based on Success of the Project



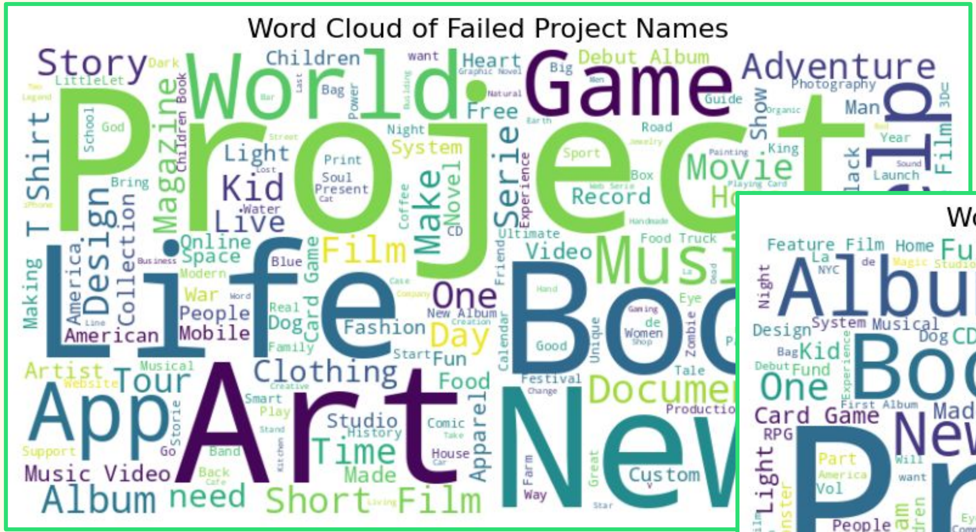
Launch Month and Goal Amount vs Success

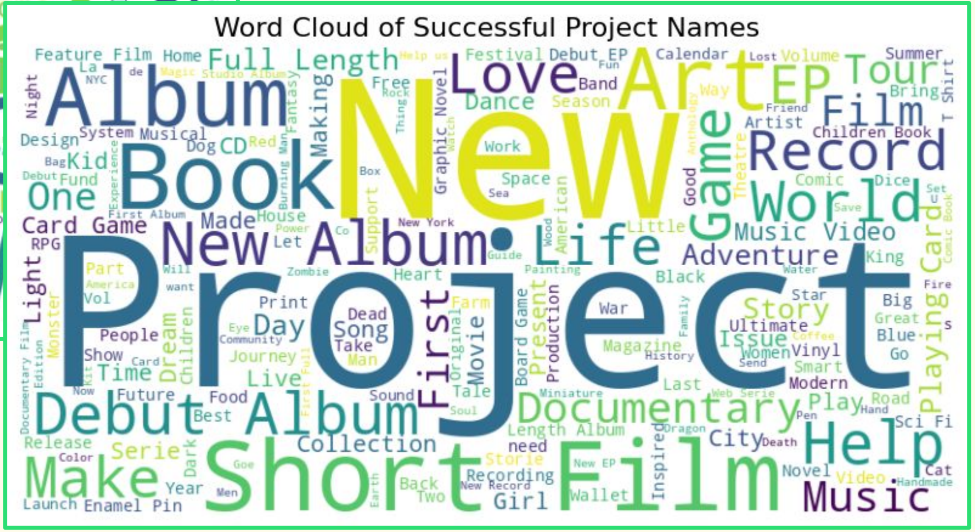




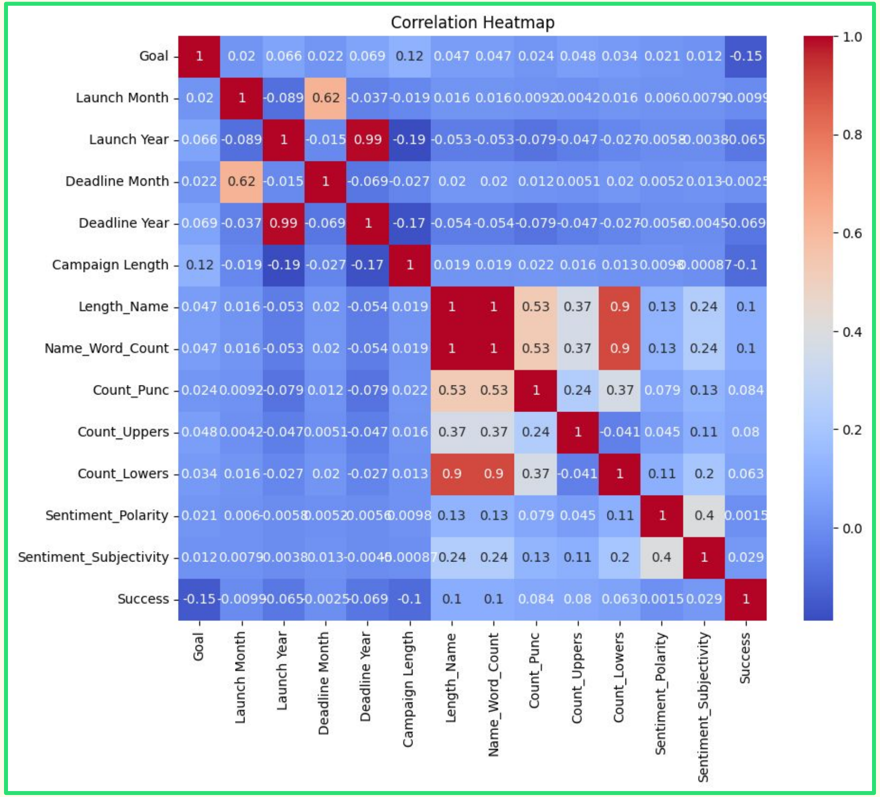
Project Name Word Cloud

Word Clouds for Successful vs. Failed Projects





Correlation Heatmap of Features



Feature Engineering

- **One-Hot-Encoded Categorical Variables**

  To include "Country," "Category," "Subcategory" features in predictive models

- **Created "Campaign Length" feature using "Launched" and "Deadline"**

  For exploration into how campaign length affects chance of success

- **Converted "Launched" and "Deadline" into Datetime type**

  To be able to include both variables into models

- **Changed "State" feature to only "Successful" or "Failed"**

  Some campaigns were mis-classified (Ex: Campaign that didn't reach goal was "Successful")

Feature Engineering: NLP on “Name” Feature

- **TF-IDF Vectorization** (Term Frequency - Inverse Document Frequency)
  - Creates a vector for each word in the corpus and its relative frequency in each name
  - For exploration of specific word’s impact on the success of a campaign
- **Counts** of Characteristics of the Names
  - Amount of **Upper/Lowercase** Letters
  - **Length** of Names Based on **Characters** and **Words**
  - Amount of **Punctuation**

Feature Engineering: Sentiment Analysis

- Identify **emotional tone** in the campaign titles
- **Polarity**: Measure of how **positive** or **negative** the sentiment is (-1 to +1)
  - Ex: "Beautiful Horizon Art"
- **Subjectivity**: Measure of how **subjective** or **objective** the text is (0 to +1)
  - Ex: "Best Game Ever" vs. "Website for Short Story Author"

Intuition

- Polarizing titles draw attention → incentivizes viewers to donate to the project
- Similar idea with subjectivity

03

Methods

Models Used

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Neural Network

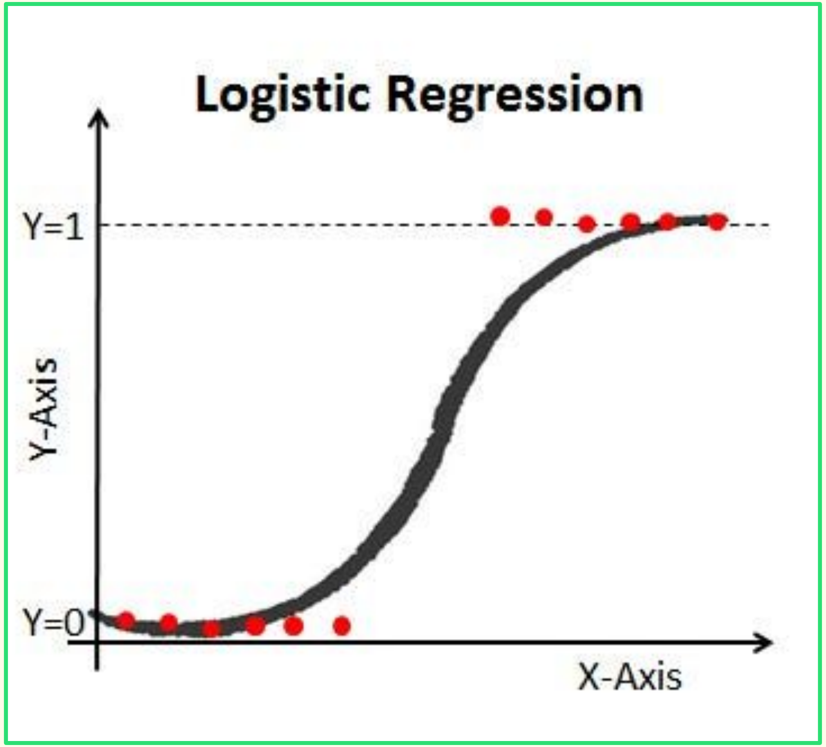
Logistic Regression

How it works:

Combines weights and features to separate instances into categories

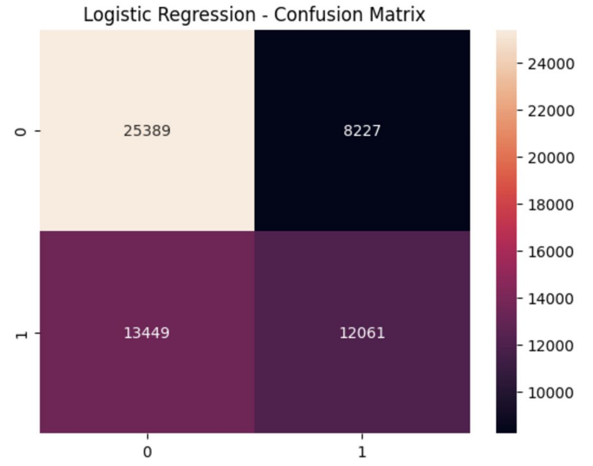
Best use case:

Most compatible with simpler binary classification where relationships are linear



Logistic Regression

	precision	recall	f1-score	support
False	0.65	0.76	0.70	33616
True	0.59	0.47	0.53	25510
accuracy			0.63	59126
macro avg	0.62	0.61	0.61	59126
weighted avg	0.63	0.63	0.63	59126



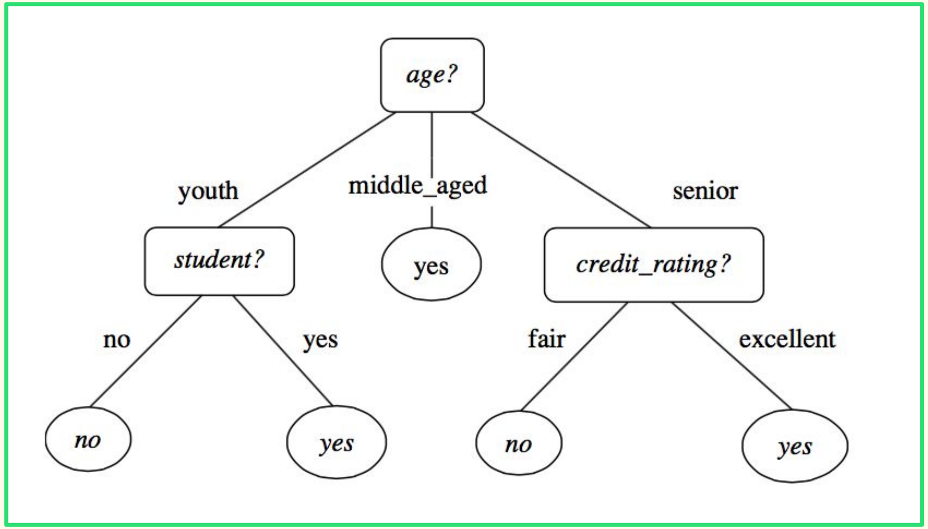
Decision Tree

How it works:

Series of binary questions based on feature values that lead to classifications

Best use case:

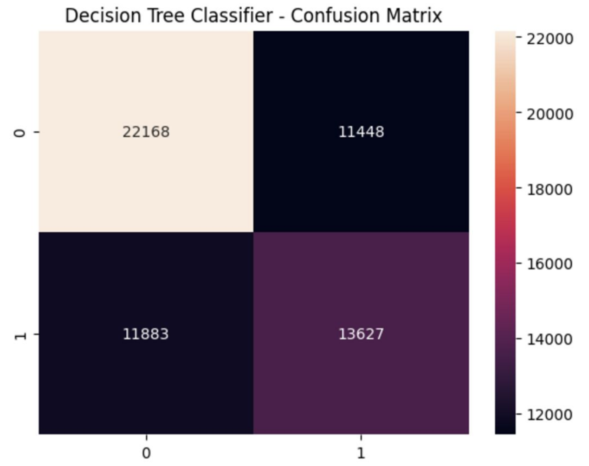
Easy interpretability, efficient, and useful with simpler datasets



Decision Tree

Decision Tree Classifier:

	precision	recall	f1-score	support
False	0.65	0.66	0.66	33616
True	0.54	0.53	0.54	25510
accuracy			0.61	59126
macro avg	0.60	0.60	0.60	59126
weighted avg	0.60	0.61	0.60	59126



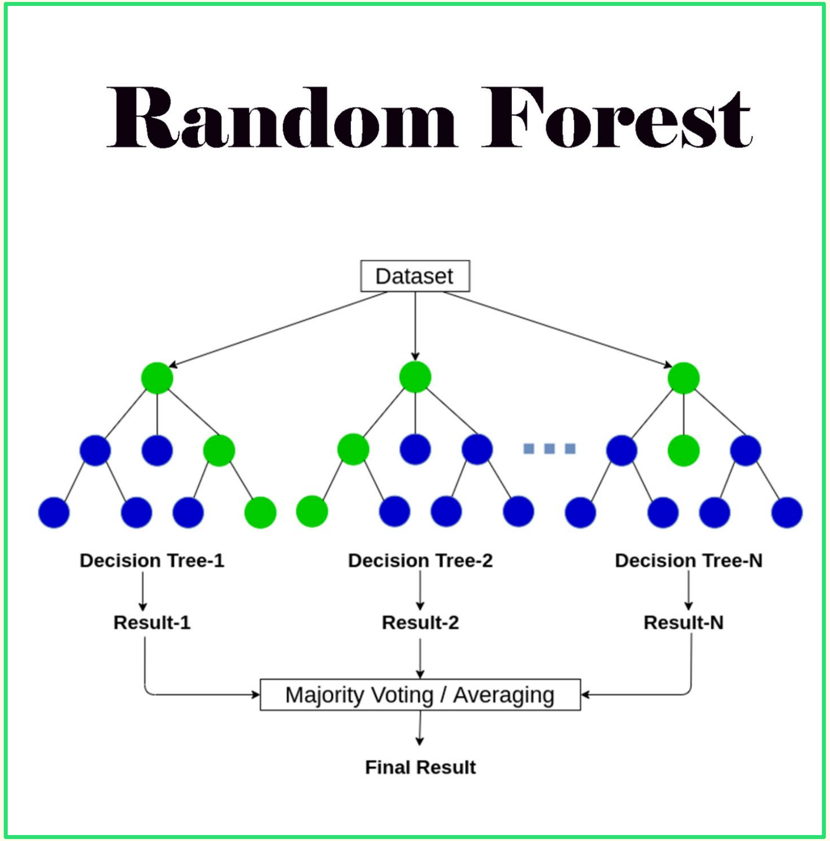
Random Forest

How it works:

Ensemble learning method that uses a collection of decision trees

Best use case:

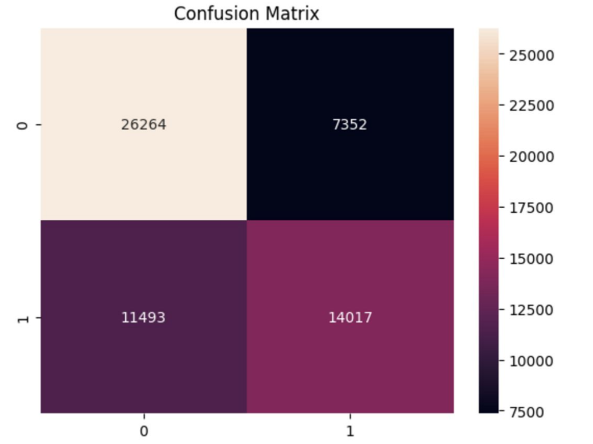
Able to identify more complex patterns–improving accuracy while reducing overfitting



Random Forest

Random Forest Classifier:

	precision	recall	f1-score	support
False	0.70	0.78	0.74	33616
True	0.66	0.55	0.60	25510
accuracy			0.68	59126
macro avg	0.68	0.67	0.67	59126
weighted avg	0.68	0.68	0.68	59126



Random Forest

GridSearchCV:

Best Parameters:

```
{'max_features':
'auto',
'min_samples_leaf': 4,
'min_samples_split':
10, 'n_estimators':
200}
```

Best CV Score: ~0.701

	feature	importances
0	Goal	0.077100
5	Campaign Length	0.052861
10	Count_Lowers	0.048781
6	Length_Name	0.048430
7	Name_Word_Count	0.047858
9	Count_Uppers	0.041197
3	Deadline Month	0.039136
1	Launch Month	0.038940
2	Launch Year	0.026424
4	Deadline Year	0.026400

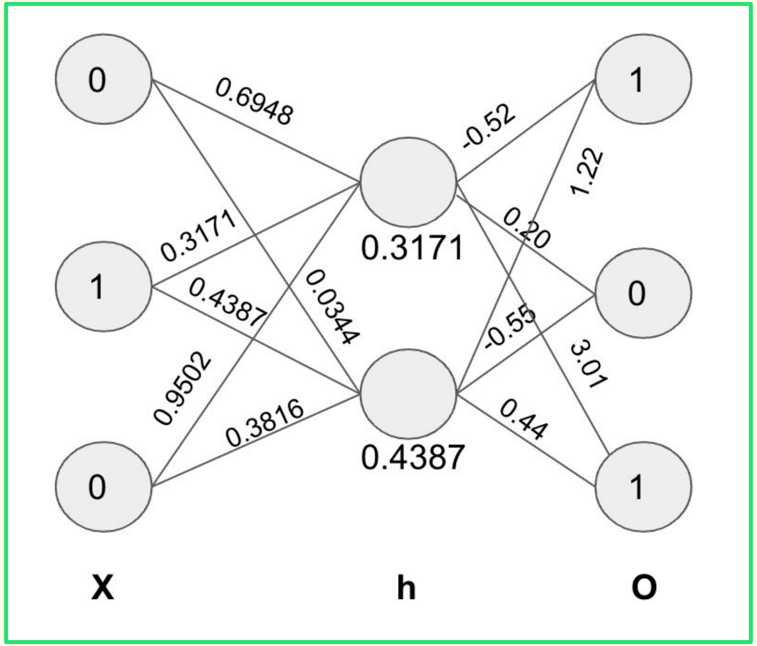
Neural Network - MLP Classifier

How it works:

Mimics brain neurons by using multiple hidden layers and utilizes backpropagation to optimize weights and biases

Best use case:

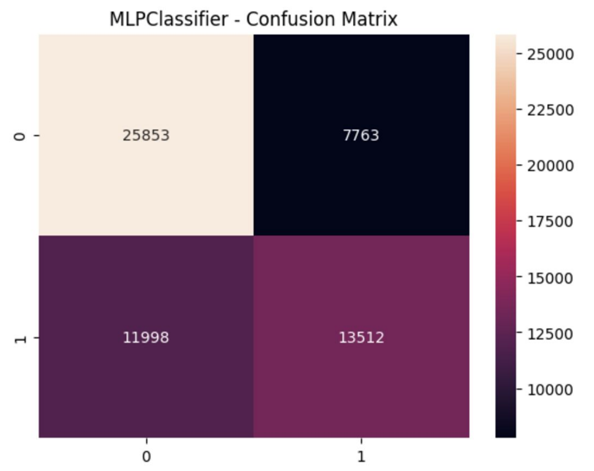
Can fit to more complex/non-linear relationships



Neural Network - MLP Classifier

MLPClassifier:

	precision	recall	f1-score	support
False	0.68	0.77	0.72	33616
True	0.64	0.53	0.58	25510
accuracy			0.67	59126
macro avg	0.66	0.65	0.65	59126
weighted avg	0.66	0.67	0.66	59126



04

Takeaways

Findings

- Slight positive correlation between **success** and **name length**
  - Longer names may be more effective in conveying the campaign goal
- Slight negative correlation between **success** and **campaign length**
  - Campaign length (days) shouldn't be too long, potentially decreases sense of urgency backers have to contribute
- Difficult to **predict campaign success** based on selected features
- Difficult to **quantify subjectivity** in human decision-making
  - Emotional appeal, alignment to personal values, etc.

Missing Features Impact Interpretability

- **Visual Content**: Images and videos not included
- **Campaign Descriptions**: More details about the project that influence backer decision-making were not included in the dataset
- **External Influences**: Social media reach, creator credibility, and marketing strategies cannot be captured

Thank you!

Appendix

05

Relevant Sources

https://www.kaggle.com/datasets/ulrikthygepedersen/kickstarter-projects/data

https://www.kickstarter.com/about

https://www.kickstarter.com/help/stats

https://www.kickstarter.com/help/handbook/funding