

Predicting Kickstarter Success

Data-driven decision support

*Machine Learning Group Project in AI Engineering Course, 02.02.2026
by Salvo, Kevin & Ivo*

Context - Kickstarter

What is Kickstarter?

- Online platform for funding projects
- Projects have a fixed funding goal and deadline

Examples



Peloton (Hometrainer)



Oura (Smart Ring)

A screenshot of a Kickstarter project page for "Handiheadset" by Stephen Moore. The page shows a video of a man wearing a headset, a progress bar indicating 9 backers and £256 pledged of £20,000 goal, and a "Funding Unsuccessful" message stating the goal was not reached on June 20. The page includes social sharing options and a contact button.

Handiheadset (unsuccessful)

Context - Business Problem

Key question

Can we predict the success of a kickstarter project?

Definition of success

Successful = when a project is funded

Not successful = if the business itself will be successful in the future

Dataset & Scope

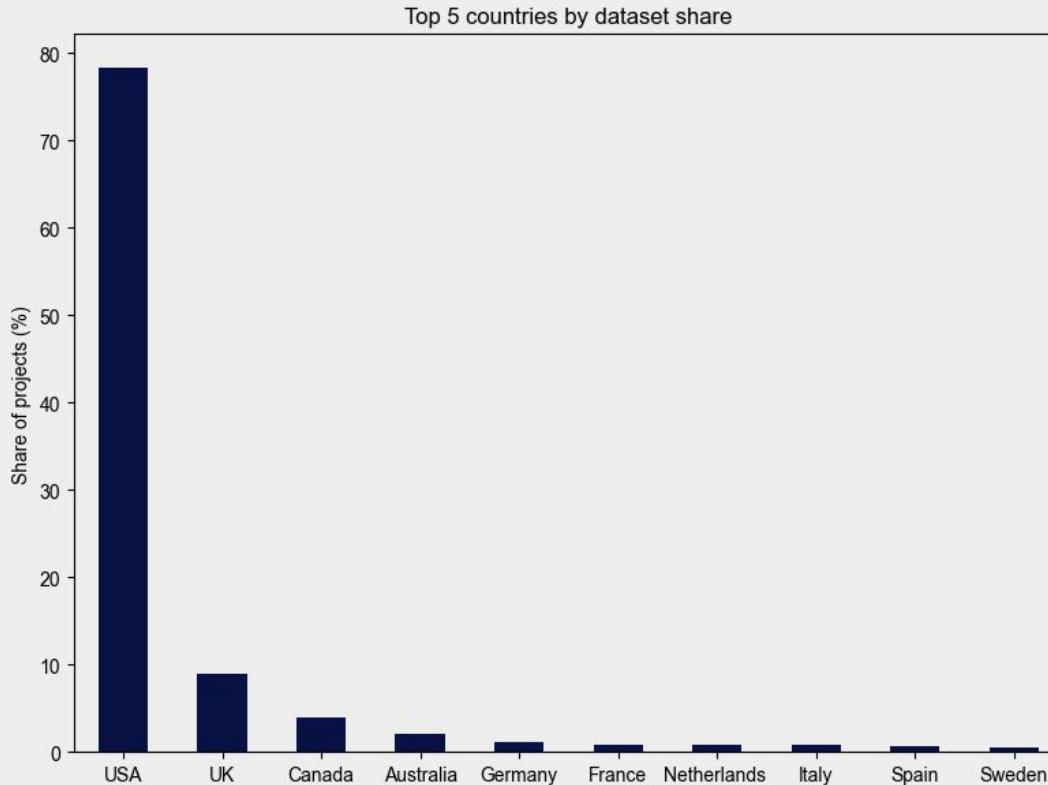
Dataset:

- *~ 375k Samples , 11 Columns (incl. Target (State))*

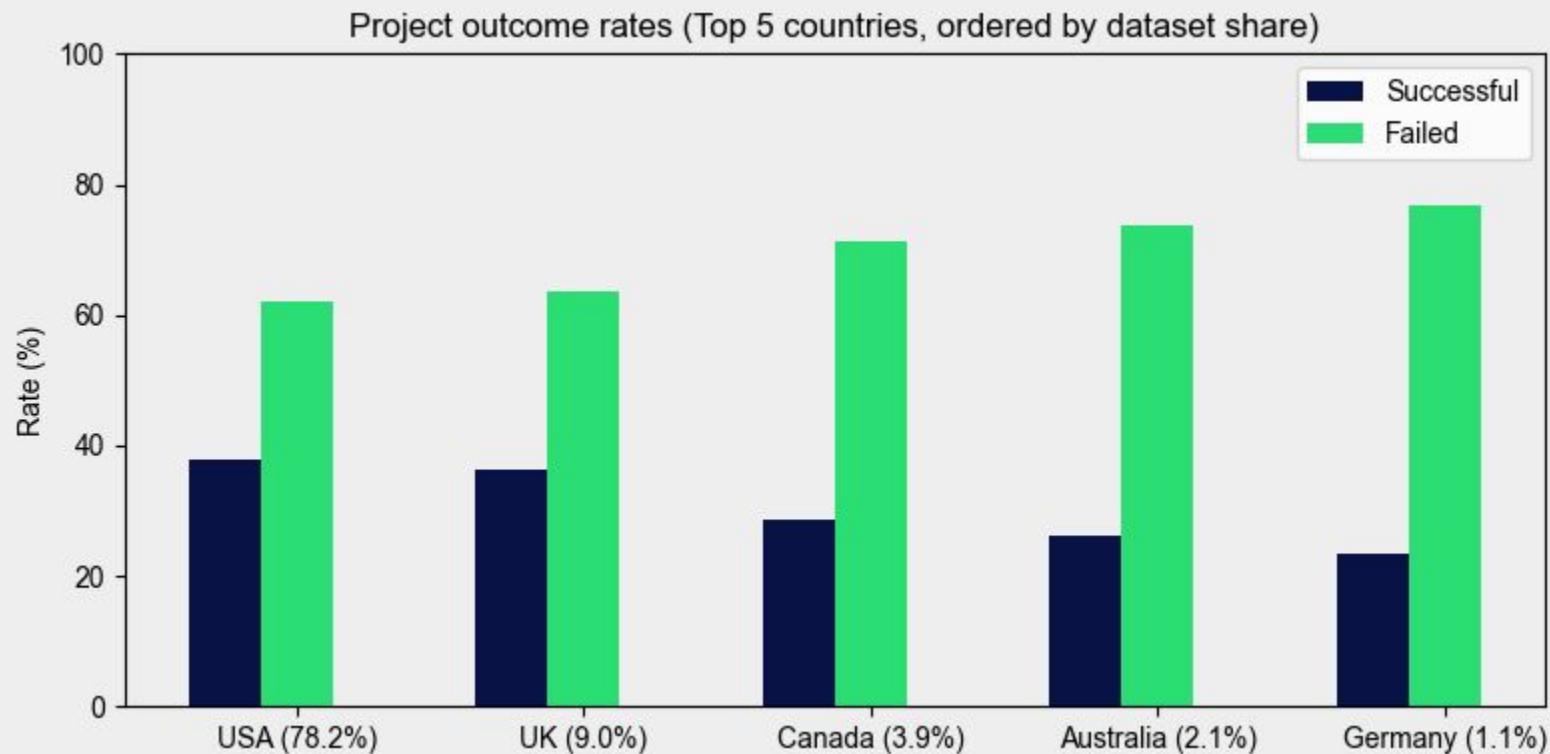
What data did we use?

- *Left out 5 of 11 Columns due to data leakage and no meaningful information*
- *Data Cleaning was very sparse*

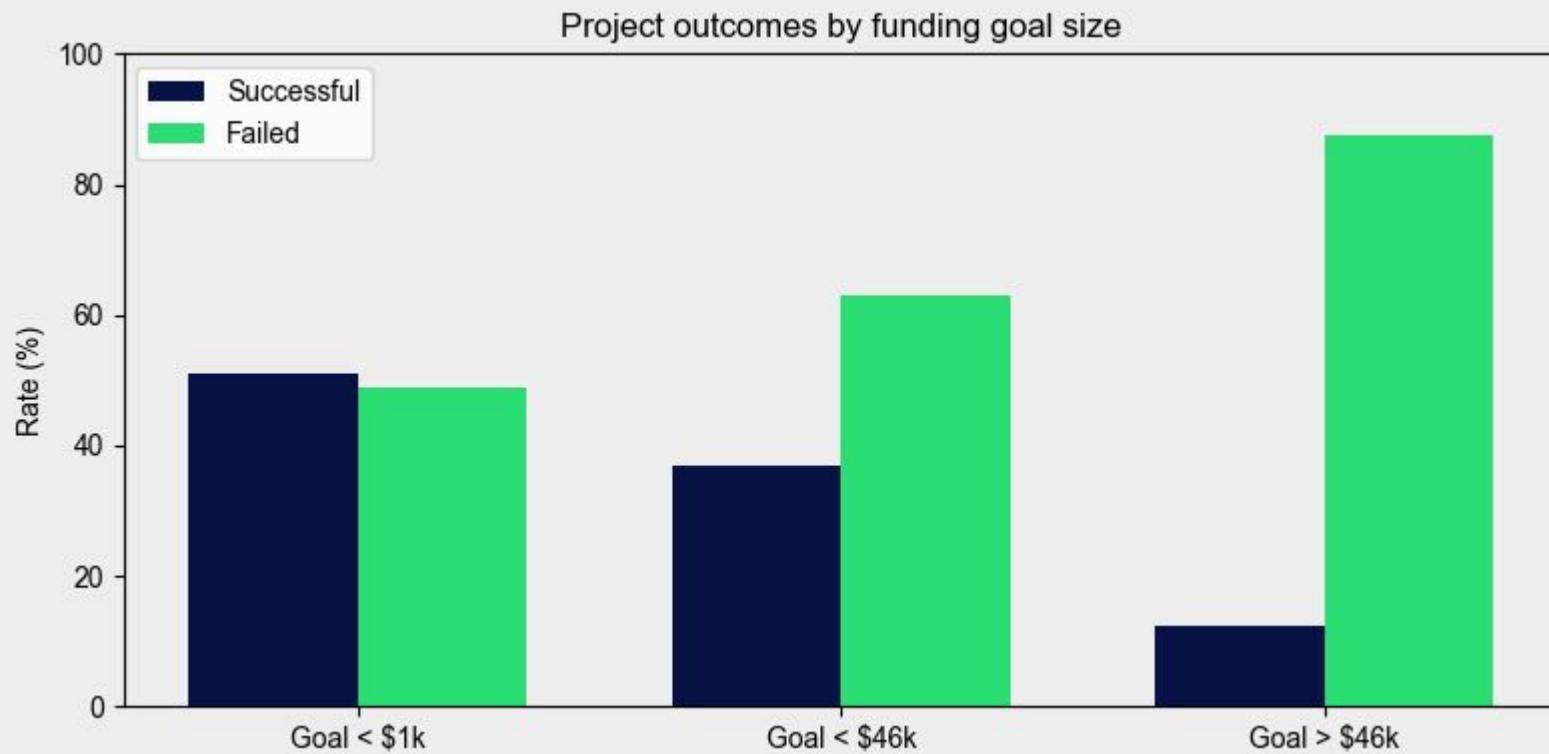
EDA - Visuals (Country Distribution)



EDA - Visuals (Project Outcome)



EDA - Visuals (Goal/Success)



Success Definition

Target Variable

Project outcome: Successful vs. Failed

- *Successful = reaching funding goal*
- *Binary decision problem*

Modeling - Evaluation Metric

Founder's question

"If I launch my project like this, how likely is it to succeed?"

Why *precision* matters:

- founders decision about whether to launch or not
 - launching a project requires time, money, reputation
- **When we predict success, we want to be highly confident**
- **Evaluation metric: *Precision***

Modeling - Approach

Tested Model

Baseline (Logistic Regression)

Decision Tree

Random Forest

KNN

Logistic Regression

XGBoost

Approach for all models

- *Feature Engineering*
- *Preprocessors: GridSearch (cross-validated)*
- *Hyperparameter Tuning: RandomizedSearch (cross-validated)*
- *Optimized for: precision*

Modeling - Evaluation

Tested Model	Precision	Recall
Baseline (Logistic Regression)	0.58	0.30
Decision Tree	0.89	0.91
Random Forest	0.89	0.64
KNN	0.90	0.89
Logistic Regression	0.90	0.91
XGBoost	0.90	0.91

Modeling - Evaluation

What the model does well	What the model does not do well
<ul style="list-style-type: none">+ <i>If the model predicts a successful funding, ~9 out of 10 succeed</i>	<ul style="list-style-type: none">- only reproduction, no focus on innovation
<ul style="list-style-type: none">+ <i>Features have similar distributions for right/wrong predictions</i>	<ul style="list-style-type: none">- no visibility on why false predictions occur

Modeling - Limitations & Risks

- **Data Coverage:**
 - strong USA bias
 - limited to pre-launch data
 - unclear, if data is already systematically biased
- **Feature Scope:**
 - no information about project quality
- **Applicability:**
 - Emotional factors not covered

Application

- None
- Model will predict based on historically promising features:
 - country
 - category/subcategory
 - goal



Next Steps

More beneficial features:

- psychometric information
- language
- founder track record (previous projects, success history, etc.)
- value proposition

Enhance model to analyze:

- project name
- project description
- design

Thank you for your attention

Main part

slide 1

- begin the story: founder wants to know if his project will be fully pledged TODO choose method

slide 2

- Chose Logistic Regression as Baseline Model (quick / low energy use)
 - baseline model had precision 1.0 → Needed to cut out "pledged Column" because it leaked the outcome
 - after cutting "pledged Column" → 0.57
- show the models (little insight into model building) TODO table
 - Tested Models :
 - Decision Tree , Random Forest , KNN , Logistic Regression , XGBoost
 - Tested Different Preprocessors
 - GS for Preprocessor
 - CV for validating best Preprocessors
 - Hyperparameter Tuning
 - RandomizedSearch for Hyperparameters
 - CV for validating best Hyperparameters
- explain the chosen model: TODO: → LG
 - As good Outcome as XGBoost for less Cost

slide 3: performance of the data TODO: separate data and analyze

- small “eda” of the false flagged data (because these could be interesting for founders)
 - show if there occur patterns (which categories perform well)
 - show if the model misses specific/ performs bad on specific projects

Outlook

slide 1:

- different perspective
- show the prediction of the model for a founder

slide 2:

- limitations: 80% USA (how meaningful is the dataset for other countries?), False Negatives ignored (to be checked)
- assumption of additional & valuable features (no information/features about projects themselves)
- talk about applicability

To think about

- story/founder view
 - reality check: check the founders project with our model

vs

- more technical/performance oriented
 - reality check: test high goal + successful projects with the model
 - false predicted data eda
- build NLP to analyze names and design
- use more psychometric features