Kickerstarter modeling

Leander, Dipali, Dominik, 07.05.2024

Structure

- Kickstarter
- Data
- Data cleaning
- EDA
- Our models
- Error analysis of the best model

Kickstarter

Crowdfunding platform

- Present projects
  - Set funding goal
- People can pledge and become backer
  - Goal reached in funding period —> successful

Data overview

```
df.head()
```

✓ 0.0s

	id	name	category	subcategory	country	launched	deadline	goal	pledged	backers	state
0	1860890148	Grace Jones Does Not Give A F$#% T-Shirt (limi...	Fashion	Fashion	United States	2009-04-21 21:02:48	2009-05-31	1000	625	30	Failed
1	709707365	CRYSTAL ANTLERS UNTITLED MOVIE	Film & Video	Shorts	United States	2009-04-23 00:07:53	2009-07-20	80000	22	3	Failed
2	1703704063	drawing for dollars	Art	Illustration	United States	2009-04-24 21:52:03	2009-05-03	20	35	3	Successful
3	727286	Offline Wikipedia iPhone app	Technology	Software	United States	2009-04-25 17:36:21	2009-07-14	99	145	25	Successful
4	1622952265	Pantshirts	Fashion	Fashion	United States	2009-04-27 14:10:39	2009-05-26	1900	387	10	Failed

Shape: 331462, 11

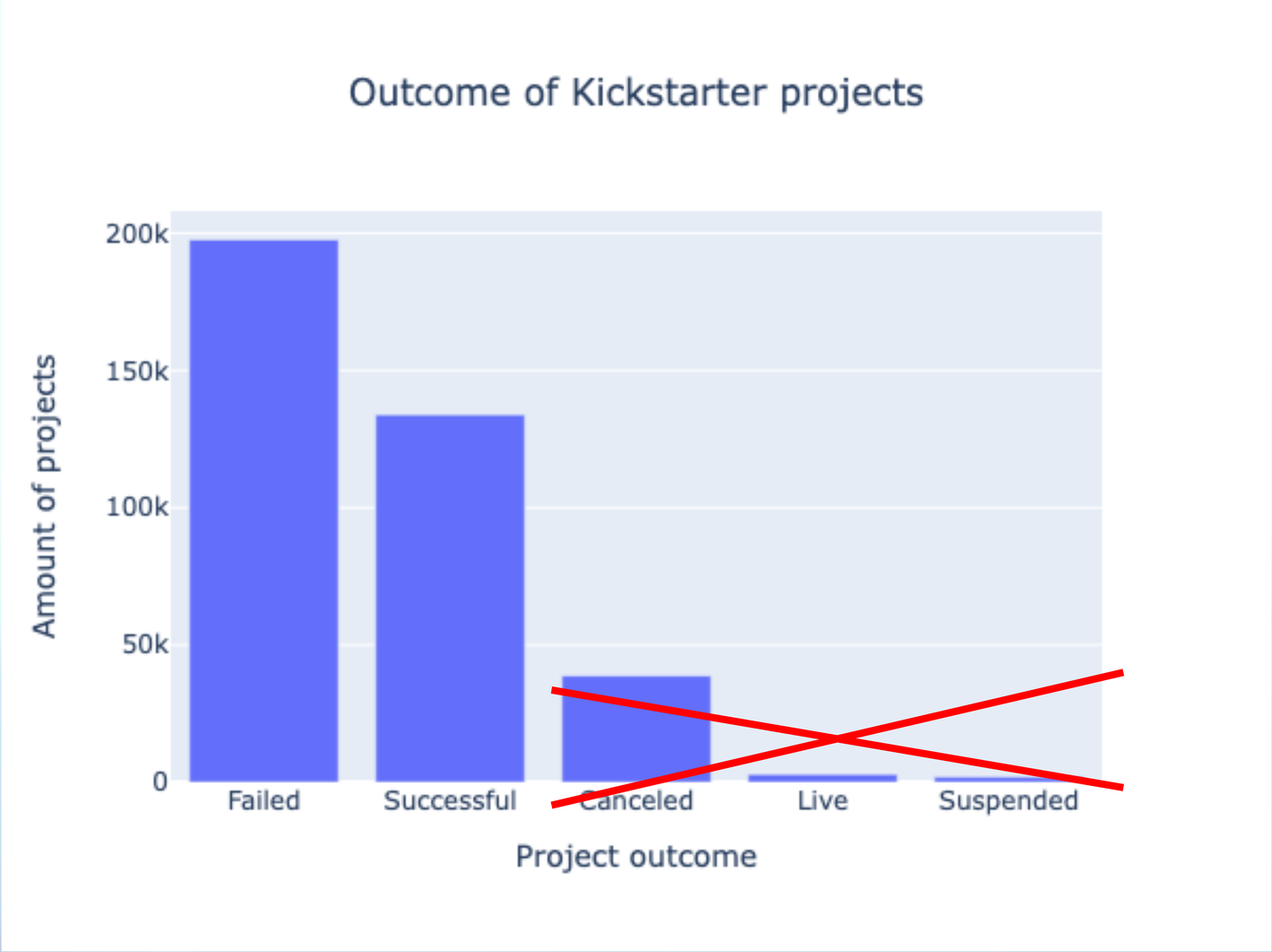
Target

Data cleaning

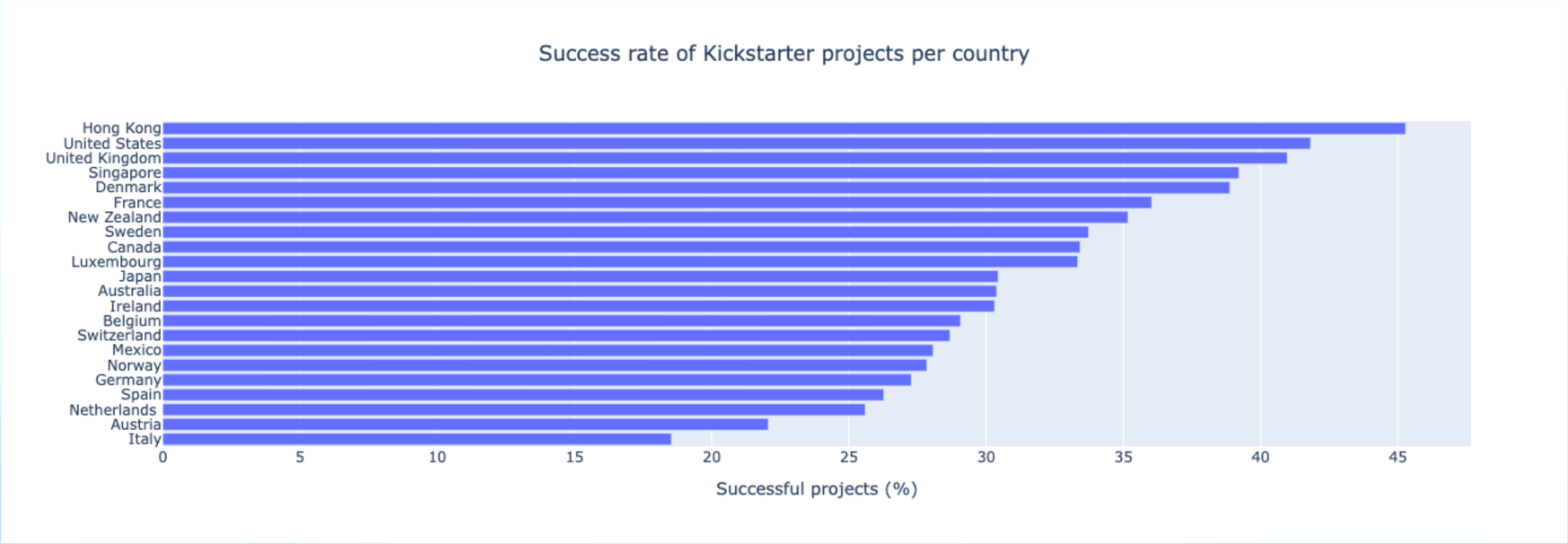
- Calculate funding period
- Mask for state ‘failed’ and ‘successful’
- Extract launch year
- One-hot-encode category, subcategory and country
- Drop unnecessary columns
- Shape of final dataset: 331462, 196

EDA

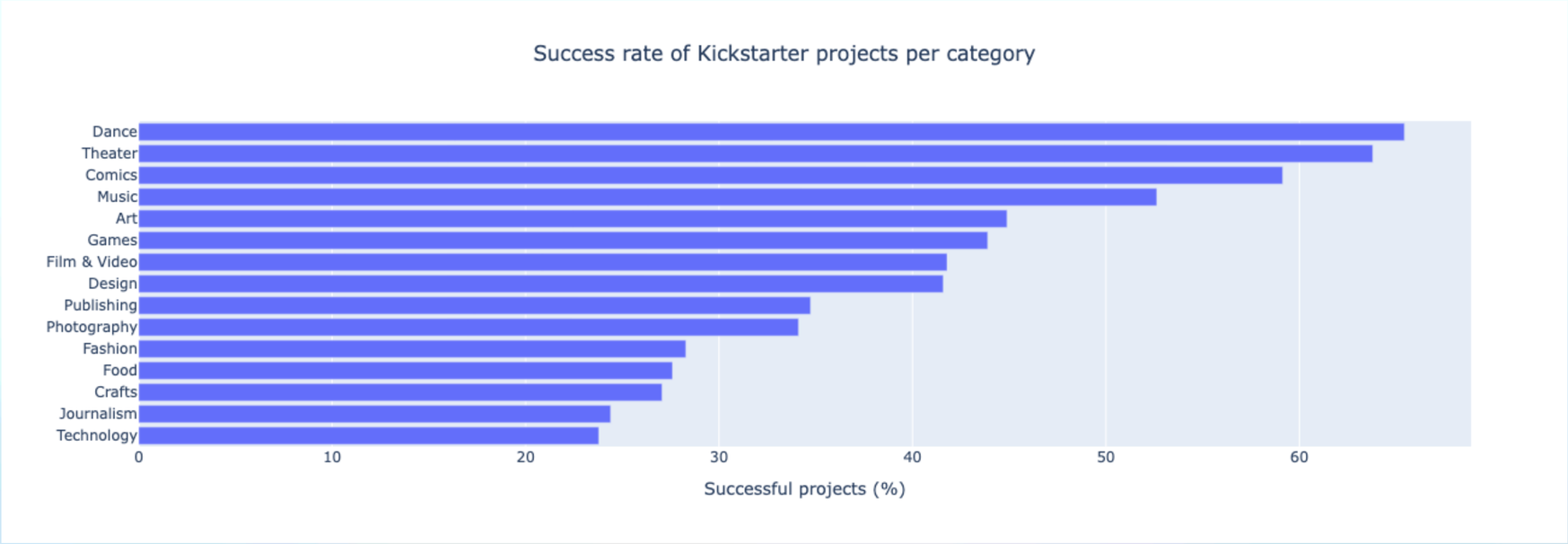
Target distribution is not balanced



Project origin country influences success



Project category influences success



Modeling

Our models

- Metric to optimize: Precision
- Optimisation by
  - Feature engineering
  - RandomSearch with cv

Model	Train precision	Test precision
Baseline (logistic regression)	0,113	0,116
KNN	0,458	0,409
Random Forest	0,54	0,52
AdaBoost	0,882	0,828
XGBoost	0,659	0,656

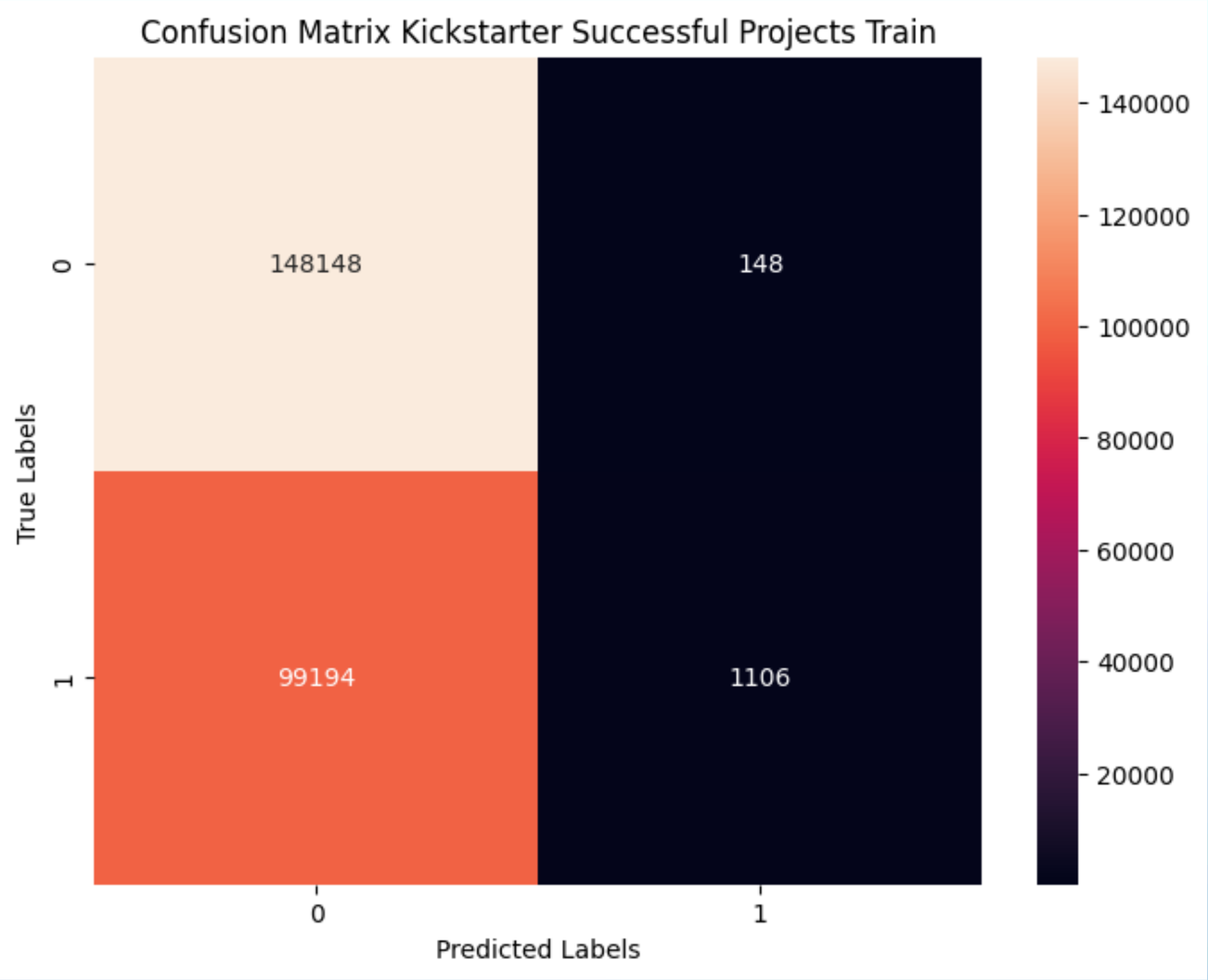
Error analysis

Why not AdaBoost?

Learned the hart way not to trust precision alone

Train precision: 0.882

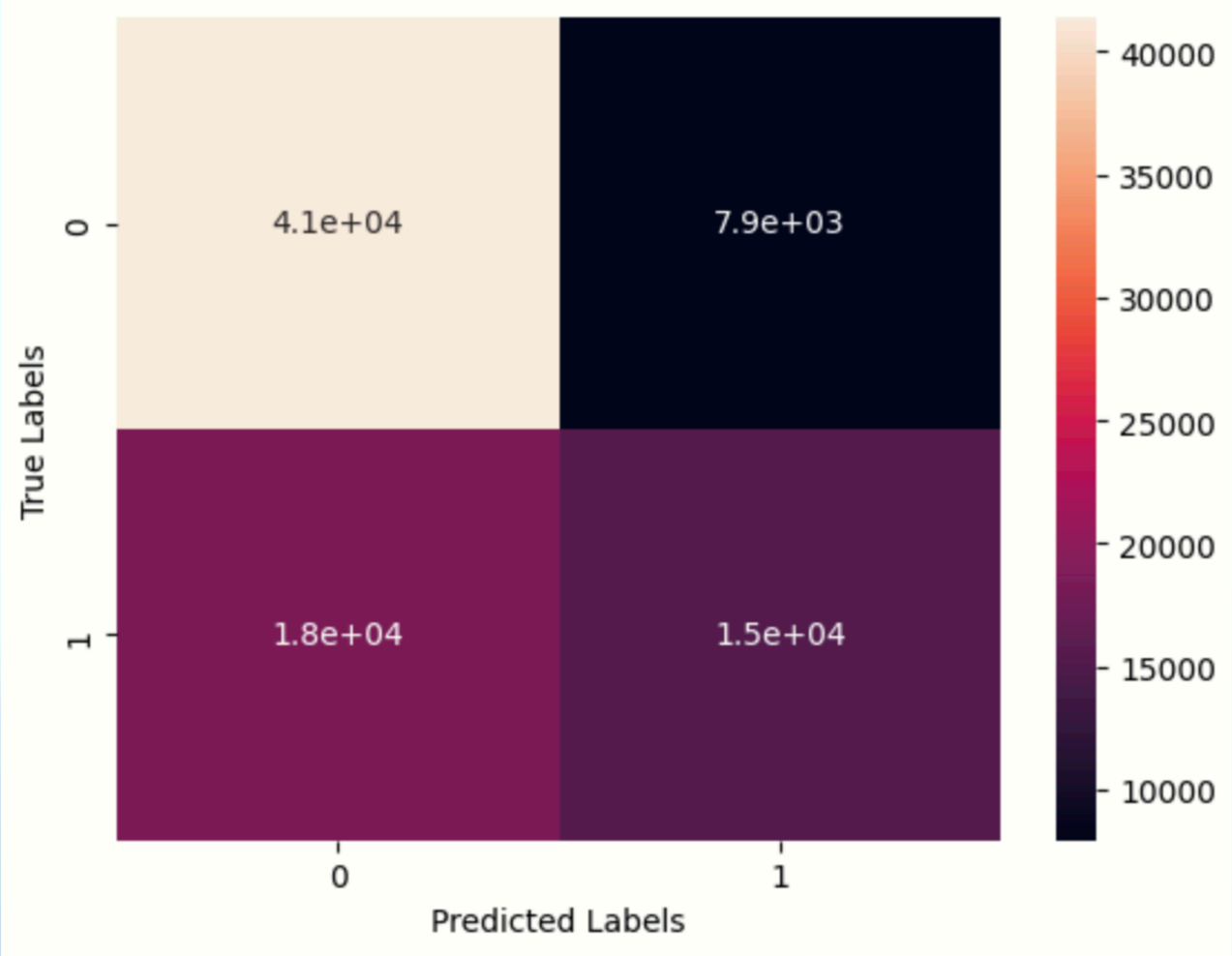
Test precision: 0,828



XGBoost Confusion Matrix

Train precision: 0.659

Test precision: 0,656



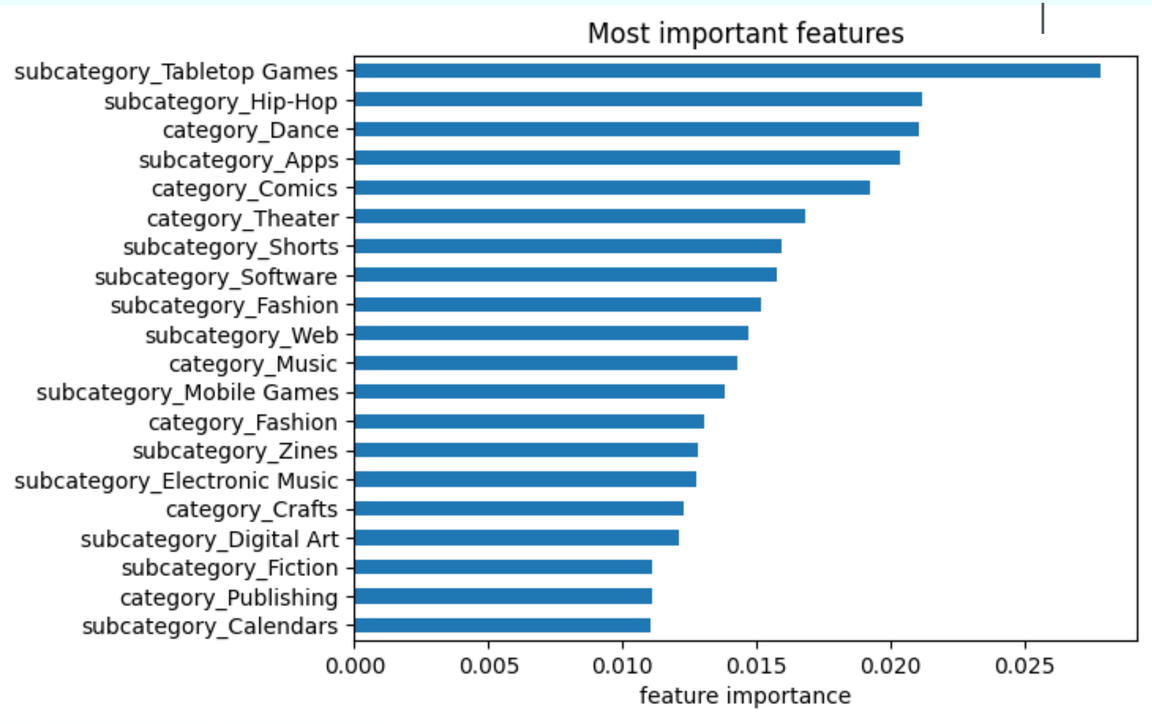
High bias,

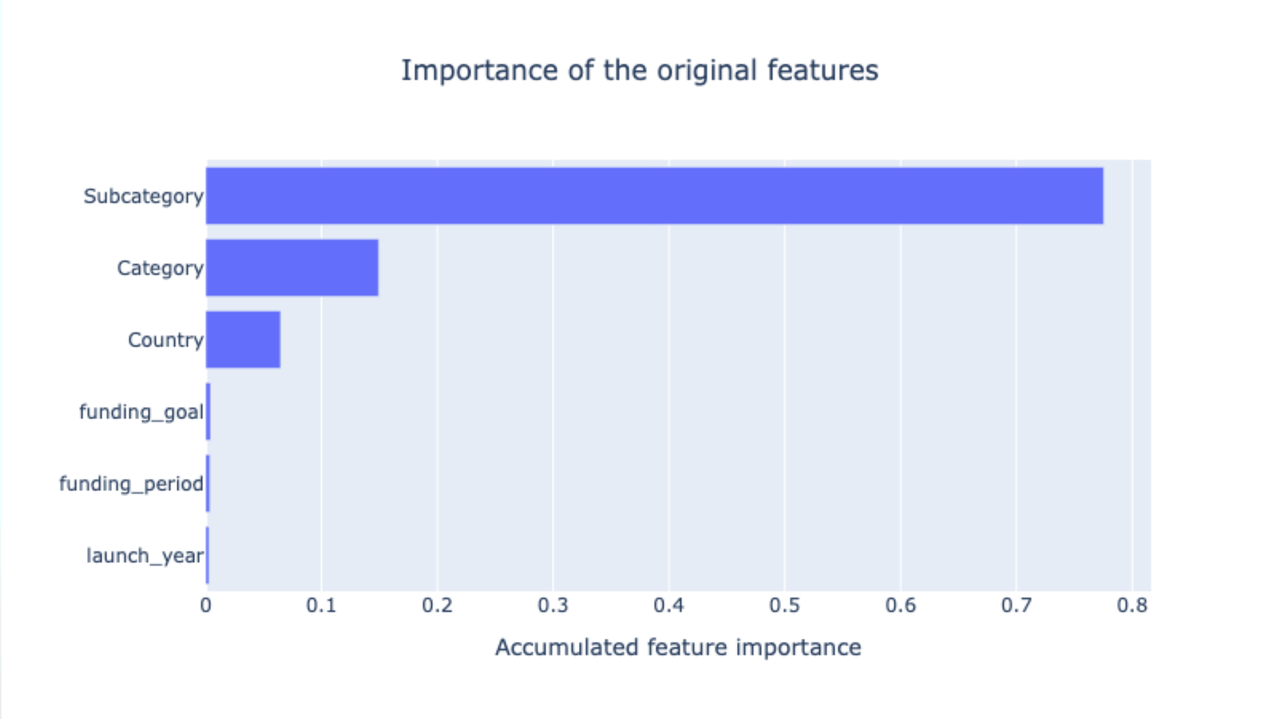
low variance

(Sub-)Categories are most important features

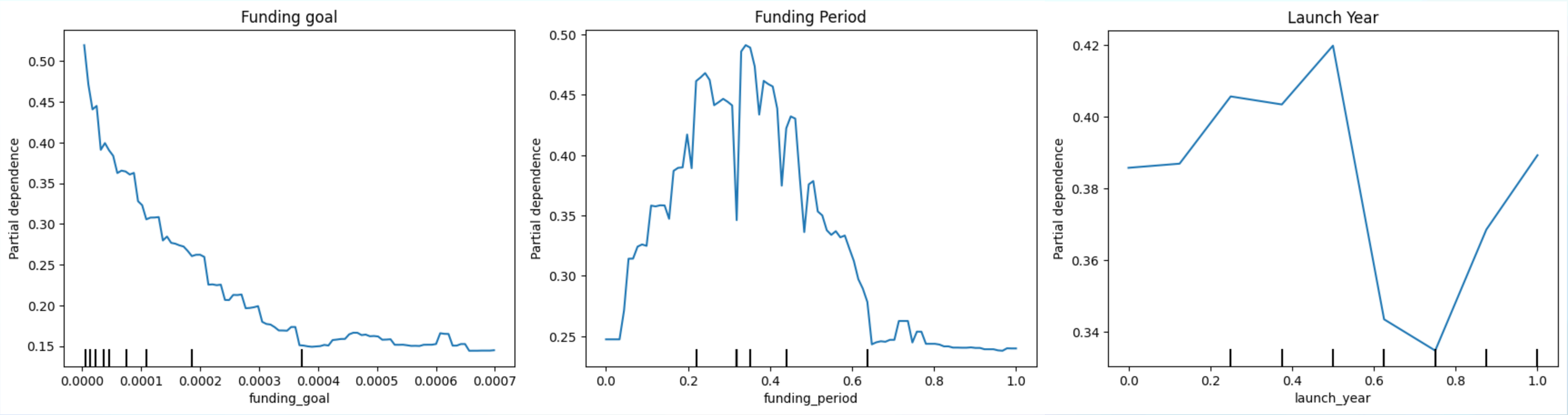
Top 20 feature importance

Dominik is happy





Partial Dependency Plots

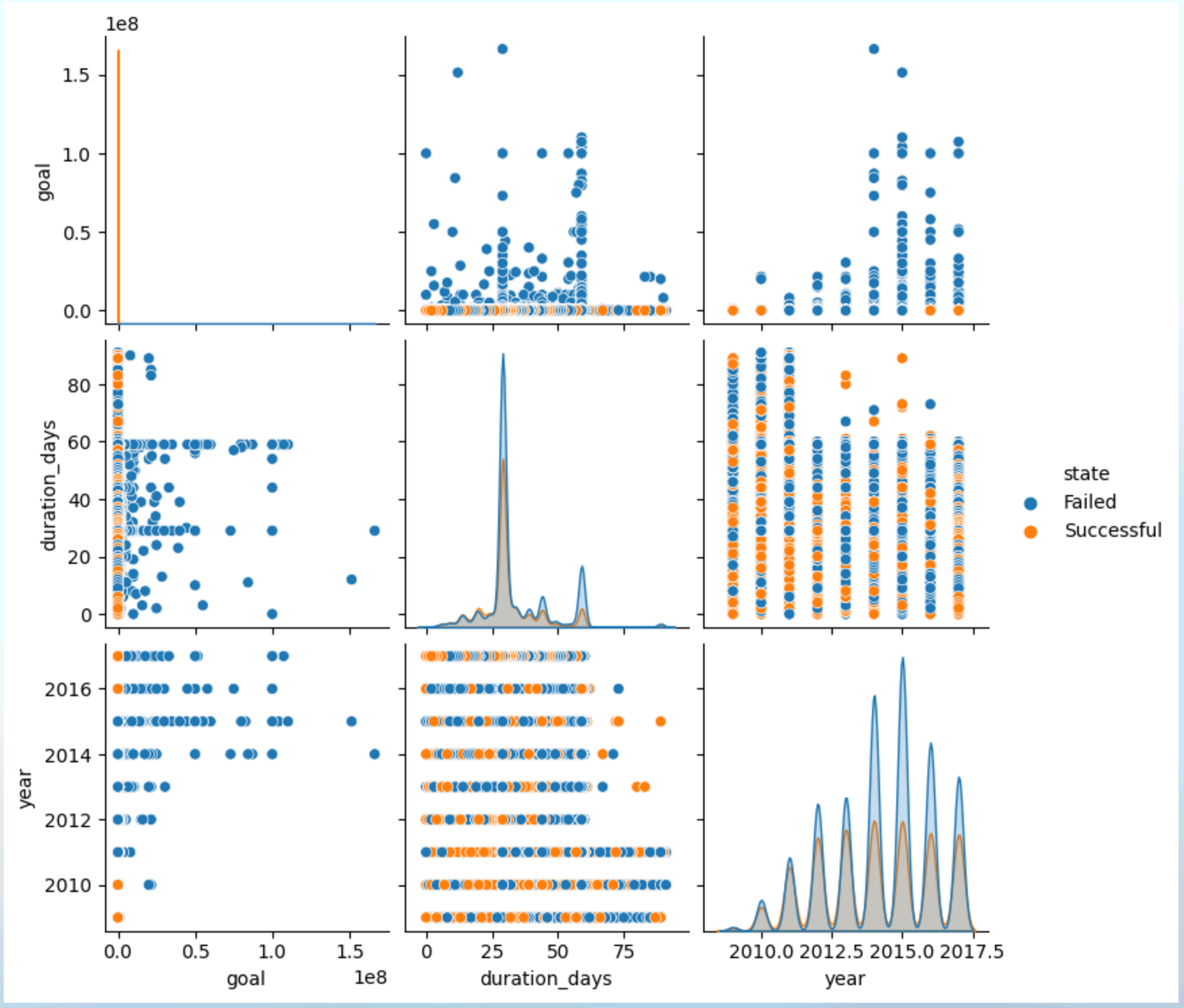


What could have been done better?

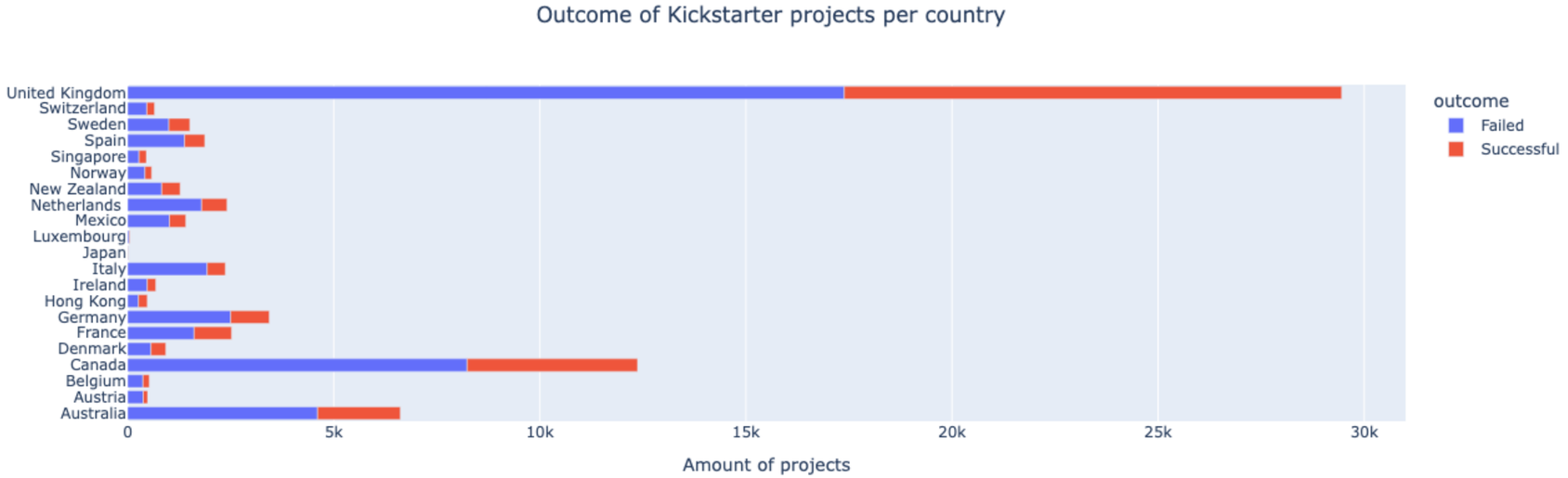
- Treat years as category? (One-hot-encode)
- Dimensionality reduction
- Exclude outliers
- Choose different metric to optimise (f1 score or ROC AUC)
- Do stacking
- Why the high bias despite a lot of hyper parameter optimisation?
  - Not the necessary input data available?
  - Domain knowledge: advertisement budget/strategy or campaign design more important then origin country or launch year

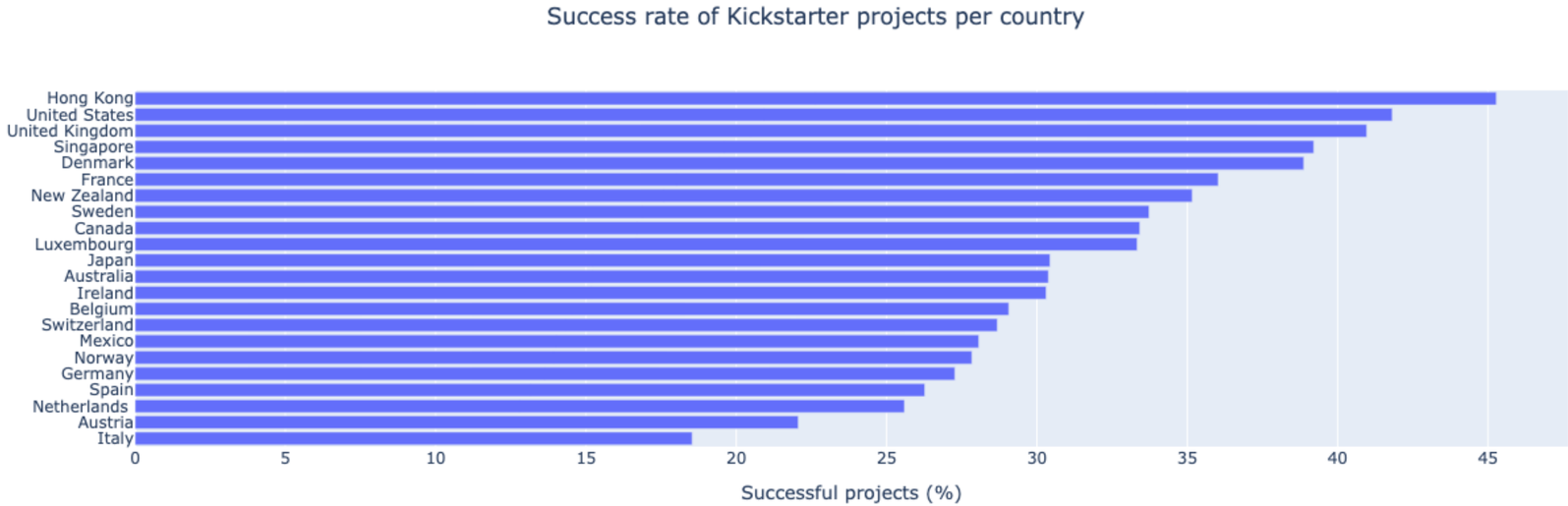
Pairplot

- More success with lower funding goal?
- No real influence of the duration days



Project origin country influences success





Project category influences success