INDENG 242 Project : Crowdfunding success prediction

Ghita Houir Alami

Zack Brodtman

Mehdi Badri

Berkeley
UNIVERSITY OF CALIFORNIA

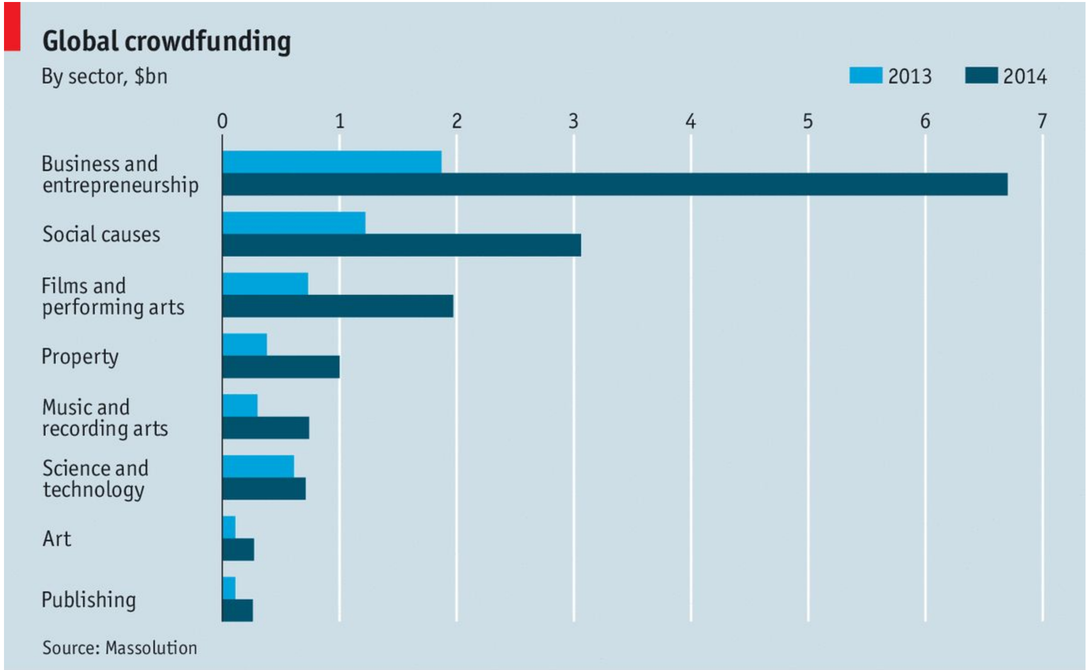
Contents

- Motivation and Data Overview
- Data Processing
- Models and Results

Section One:

Motivation and Data Overview

Problem Statement: How can we predict the outcome of a Crowdfunding campaign?

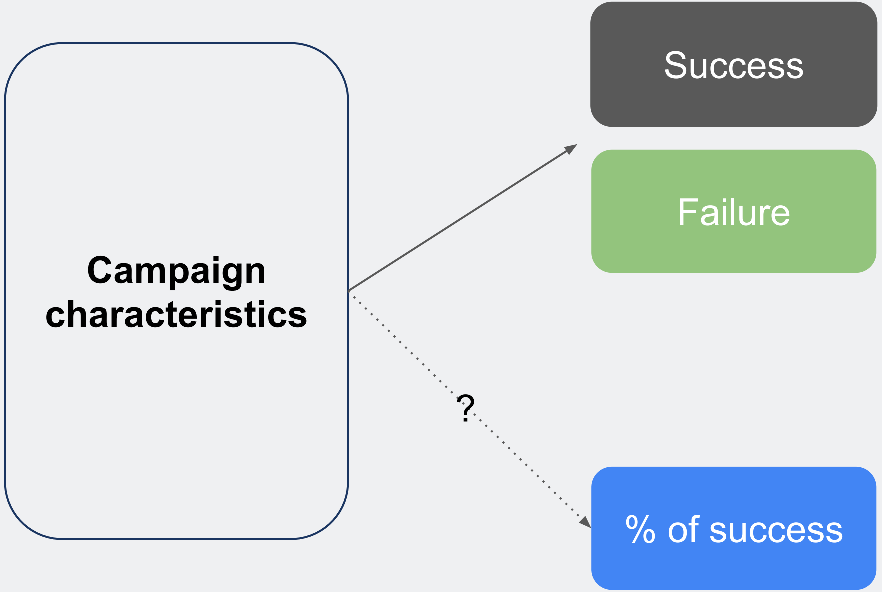


Source: Massolution

Economist.com

Source of the data

Objective



Campaign characteristics

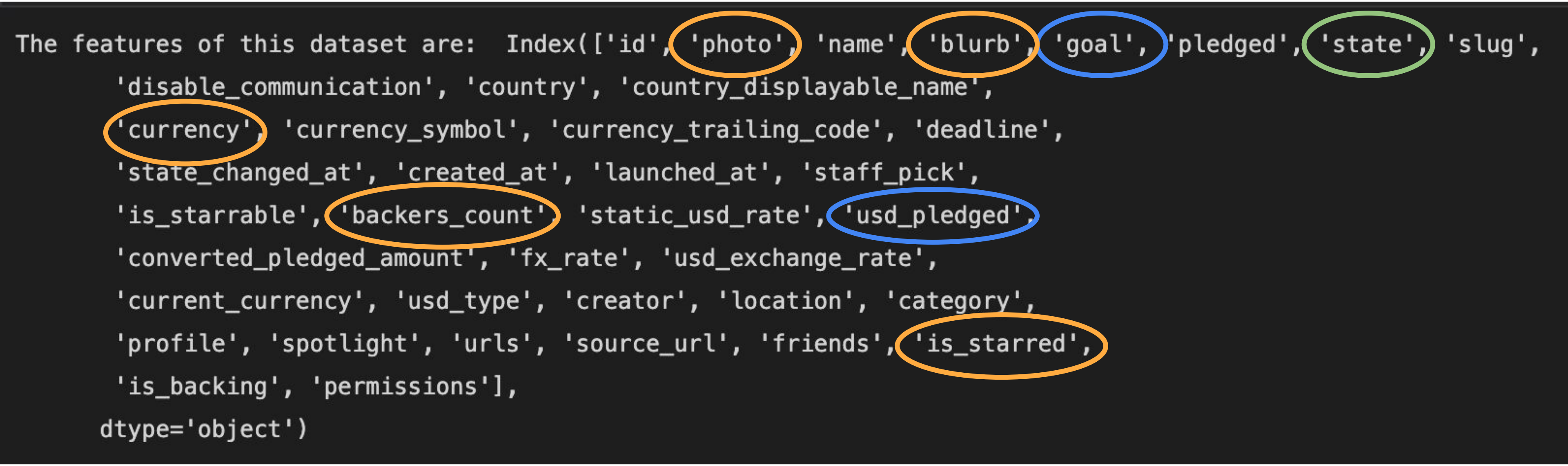
Success

Failure

?

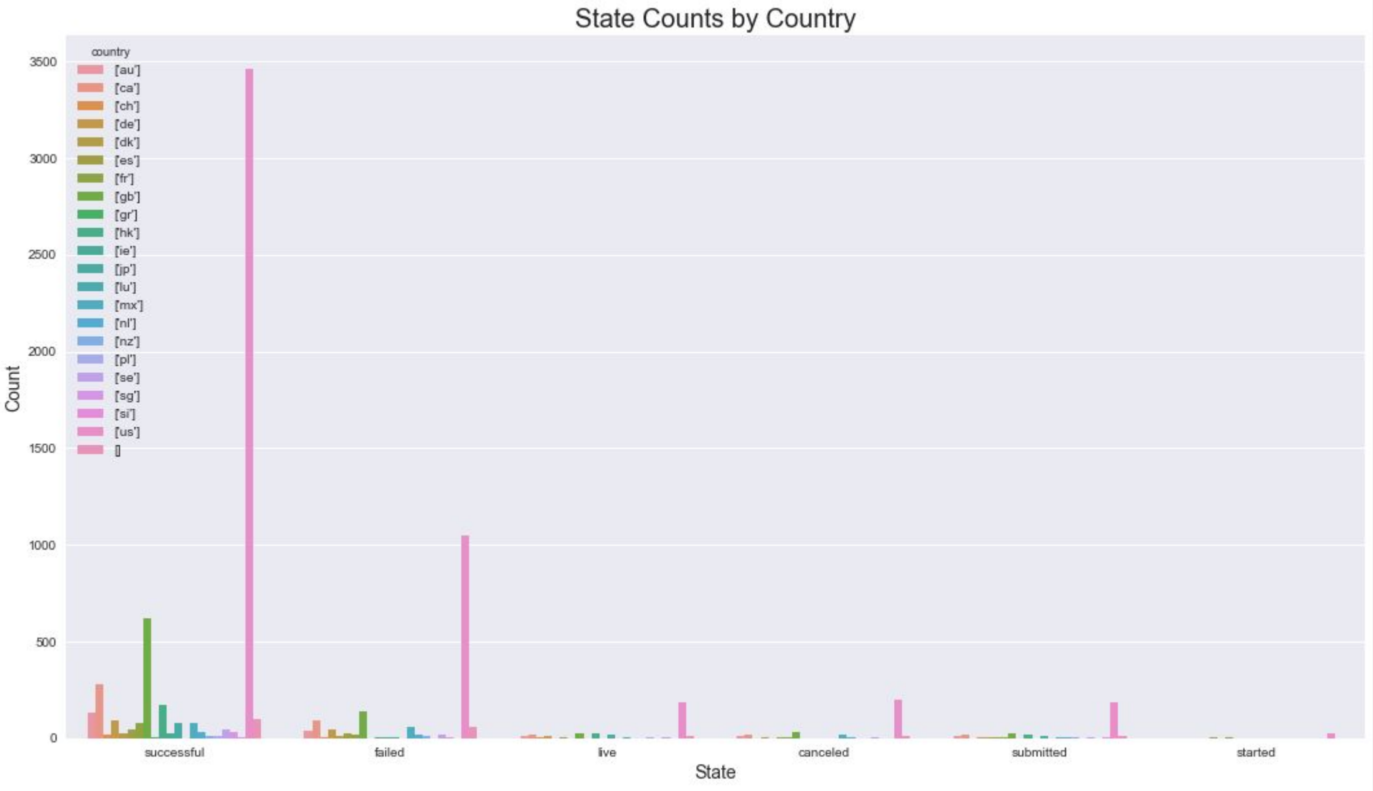
% of success

Data and Features

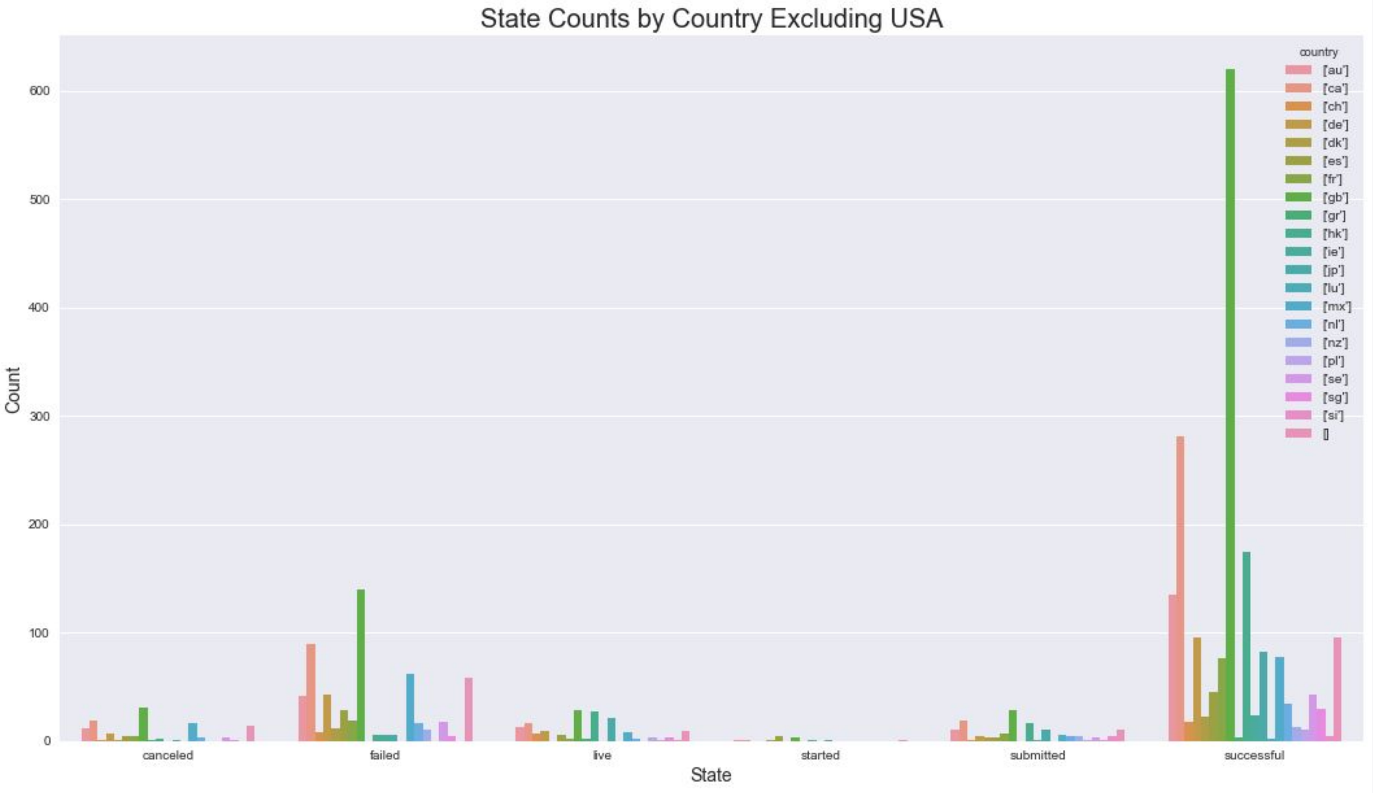


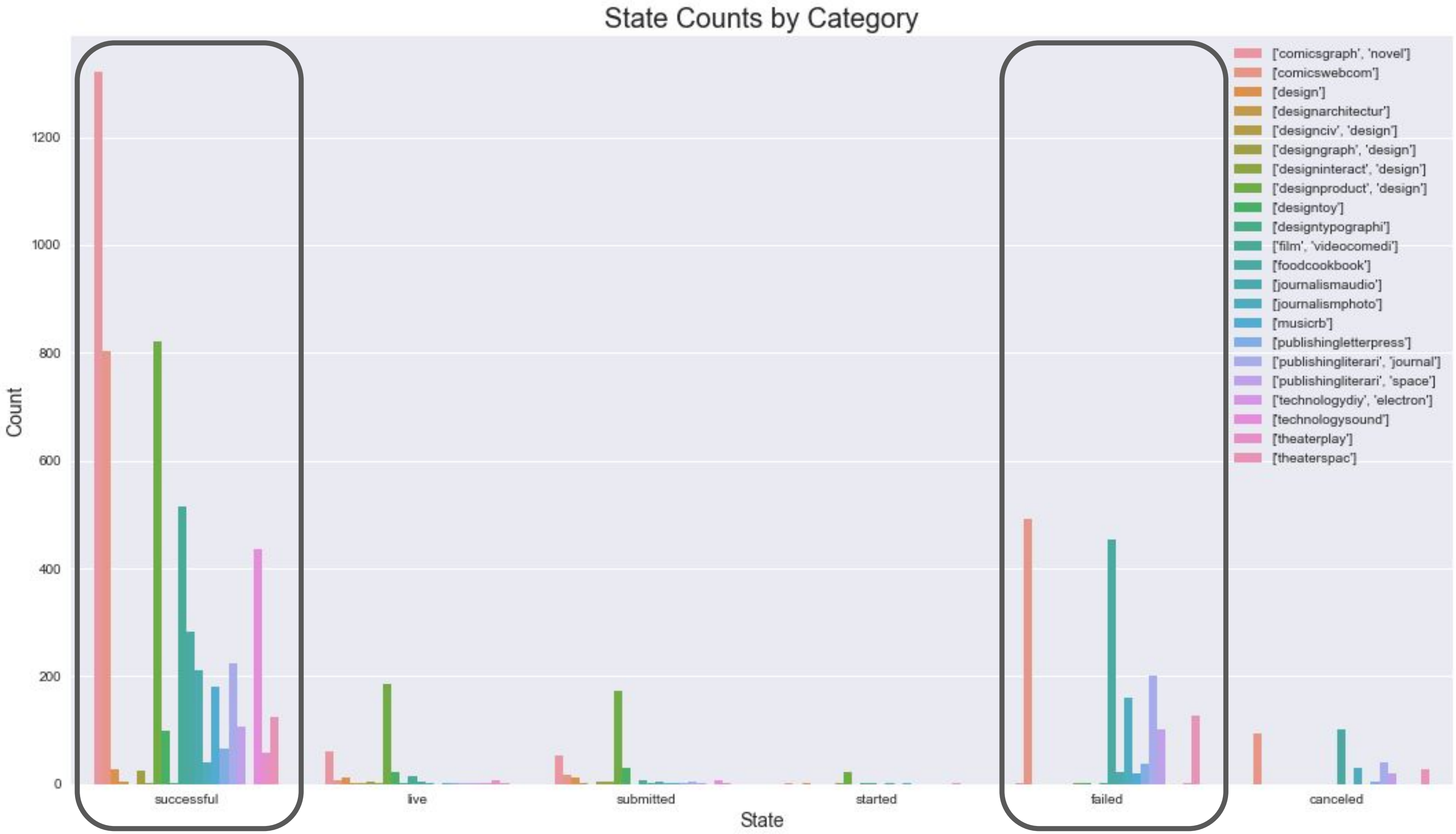
```
The features of this dataset are:  Index(['id', 'photo', 'name', 'blurb', 'goal', 'pledged', 'state', 'slug',
       'disable_communication', 'country', 'country_displayable_name',
       'currency', 'currency_symbol', 'currency_trailing_code', 'deadline',
       'state_changed_at', 'created_at', 'launched_at', 'staff_pick',
       'is_starrable', 'backers_count', 'static_usd_rate', 'usd_pledged',
       'converted_pledged_amount', 'fx_rate', 'usd_exchange_rate',
       'current_currency', 'usd_type', 'creator', 'location', 'category',
       'profile', 'spotlight', 'urls', 'source_url', 'friends', 'is_starred',
       'is_backing', 'permissions'],
      dtype='object')
```

Some Exploratory Data Analysis



Some Exploratory Data Analysis




State Counts by Category
Count
0
200
400
600
800
1000
1200
State
successful
live
submitted
started
failed
canceled
['comicsgraph', 'novel']
['comicswebcom']
['design']
['designarchitectur']
['designciv', 'design']
['designgraph', 'design']
['designinteract', 'design']
['designproduct', 'design']
['designtoy']
['designtypographi']
['film', 'videocomedi']
['foodcookbook']
['journalismaudio']
['journalismphoto']
['musicrb']
['publishingletterpress']
['publishingliterari', 'journal']
['publishingliterari', 'space']
['technologydiy', 'electron']
['technologysound']
['theaterplay']
['theaterspac']

Section Two:

Data Processing

Feature selection & Feature Engineering

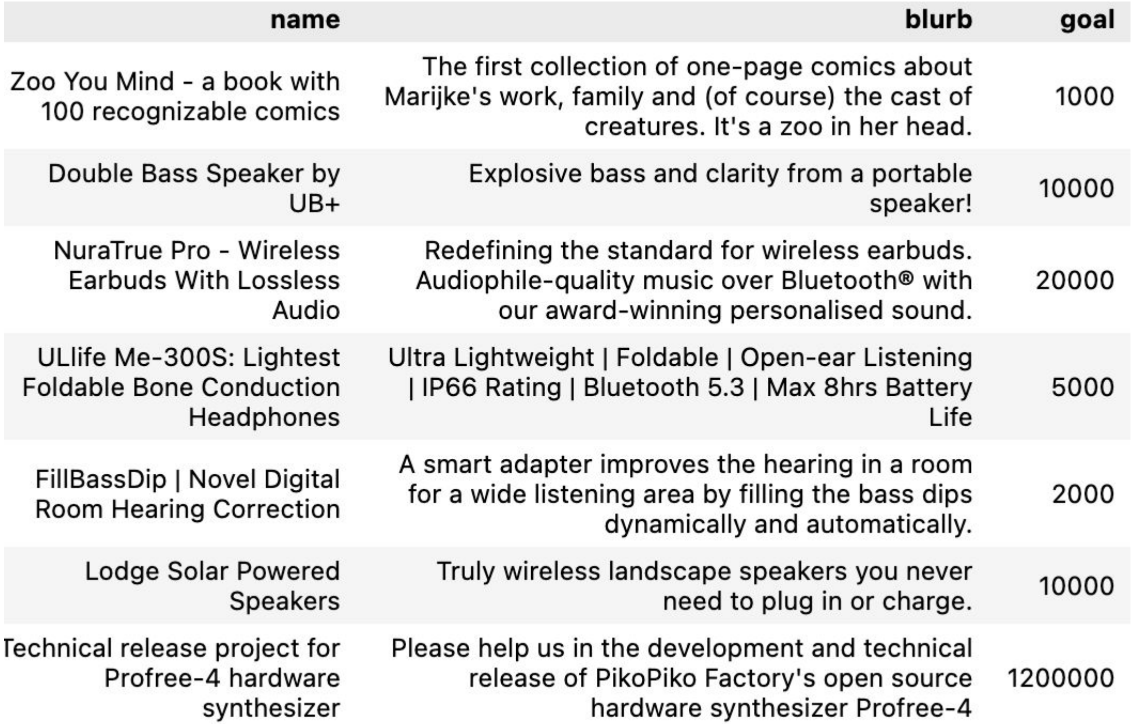
previous_launcher: Boolean feature

percentage_of_success: Numerical feature

pledge_per_backer: Numerical feature

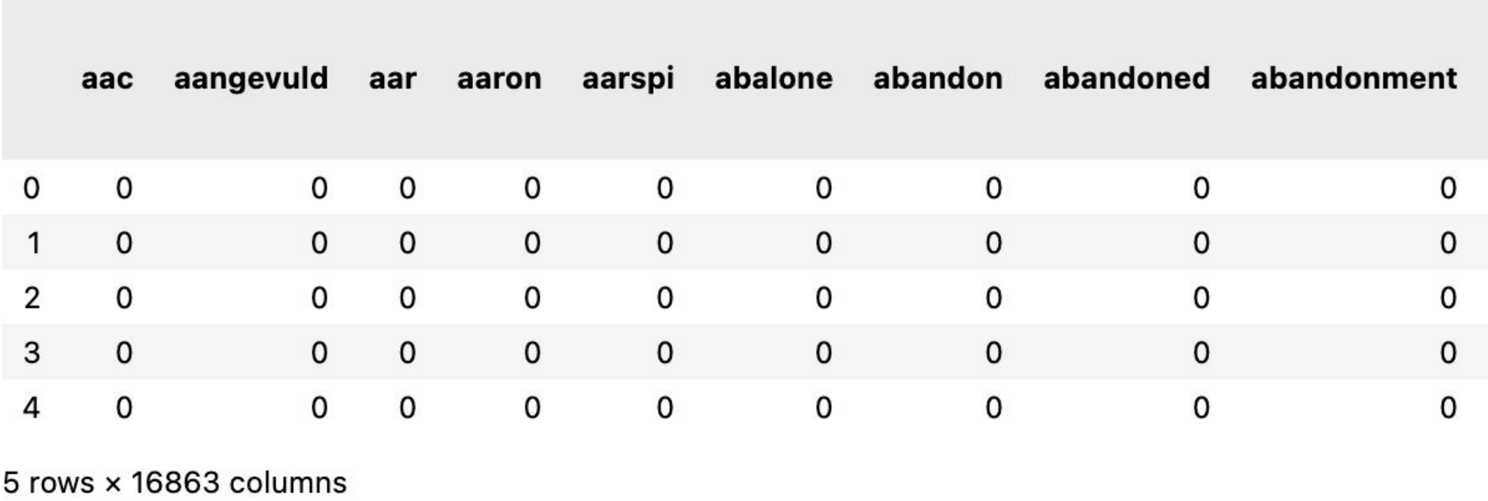
converting all amounts to usd

Working with text data



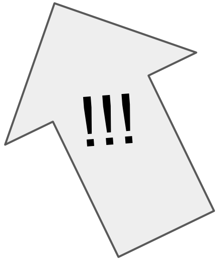
name	blurb	goal
Zoo You Mind - a book with 100 recognizable comics	The first collection of one-page comics about Marijke's work, family and (of course) the cast of creatures. It's a zoo in her head.	1000
Double Bass Speaker by UB+	Explosive bass and clarity from a portable speaker!	10000
NuraTrue Pro - Wireless Earbuds With Lossless Audio	Redefining the standard for wireless earbuds. Audiophile-quality music over Bluetooth® with our award-winning personalised sound.	20000
ULlife Me-300S: Lightest Foldable Bone Conduction Headphones	Ultra Lightweight | Foldable | Open-ear Listening | IP66 Rating | Bluetooth 5.3 | Max 8hrs Battery Life	5000
FillBassDip | Novel Digital Room Hearing Correction	A smart adapter improves the hearing in a room for a wide listening area by filling the bass dips dynamically and automatically.	2000
Lodge Solar Powered Speakers	Truly wireless landscape speakers you never need to plug in or charge.	10000
Technical release project for Profree-4 hardware synthesizer	Please help us in the development and technical release of PikoPiko Factory's open source hardware synthesizer Profree-4	1200000

Working with text data



	aac	aangevuld	aar	aaron	aarspi	abalone	abandon	abandoned	abandonment
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

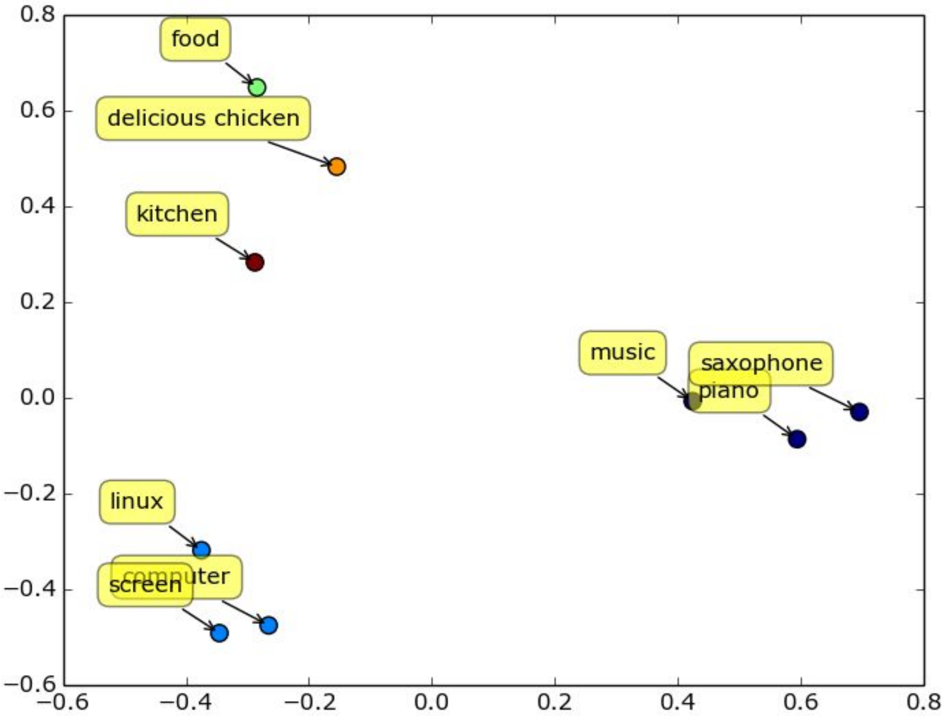
5 rows × 16863 columns



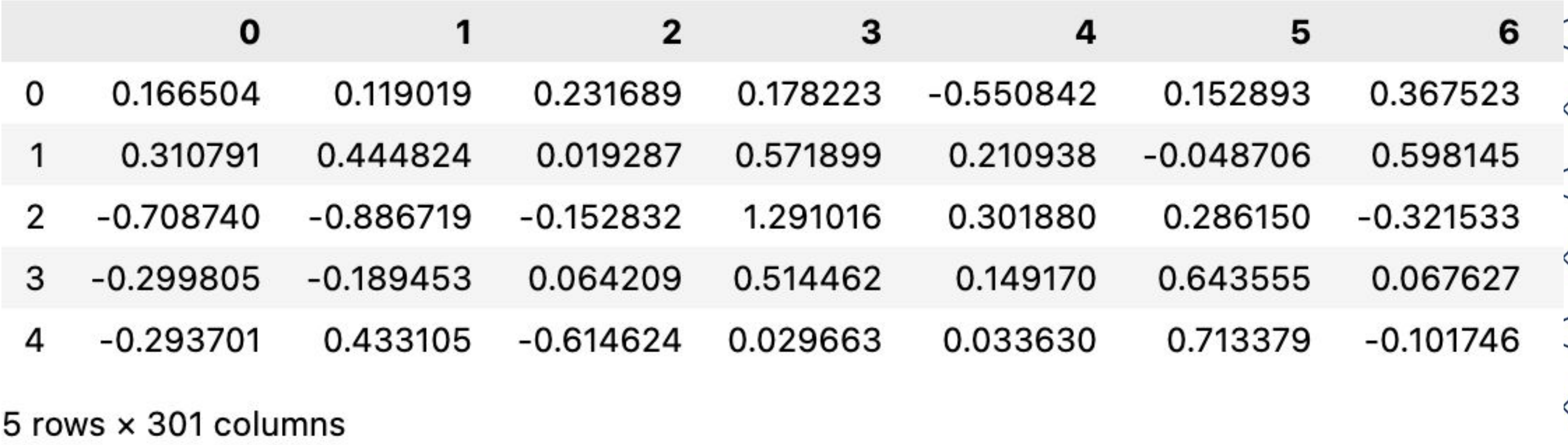
Word2Vec

- Tomas Mikolov, Google 2013
- Became the most popular NLP technique but now considered outdated when compared with transformers
- Neural networks trained on large datasets to capture semantics of words
  - E.g. King - Man + Woman = ?
- Reduce dimensionality

Word2Vec

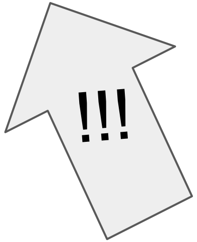

food
delicious chicken
kitchen
music
saxophone
piano
linux
computer
screen
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Working with text data



	0	1	2	3	4	5	6
0	0.166504	0.119019	0.231689	0.178223	-0.550842	0.152893	0.367523
1	0.310791	0.444824	0.019287	0.571899	0.210938	-0.048706	0.598145
2	-0.708740	-0.886719	-0.152832	1.291016	0.301880	0.286150	-0.321533
3	-0.299805	-0.189453	0.064209	0.514462	0.149170	0.643555	0.067627
4	-0.293701	0.433105	-0.614624	0.029663	0.033630	0.713379	-0.101746

5 rows × 301 columns



Interpretable Predictors

- Goal USD
- Original Currency
- Staff Pick
- Previous Launcher
- Category

= 30 Encoded Features

Non-Interpretable Predictors

- Blurb: Textual data

= 300 Embeddings

Which is more useful for predicting the state of a project?

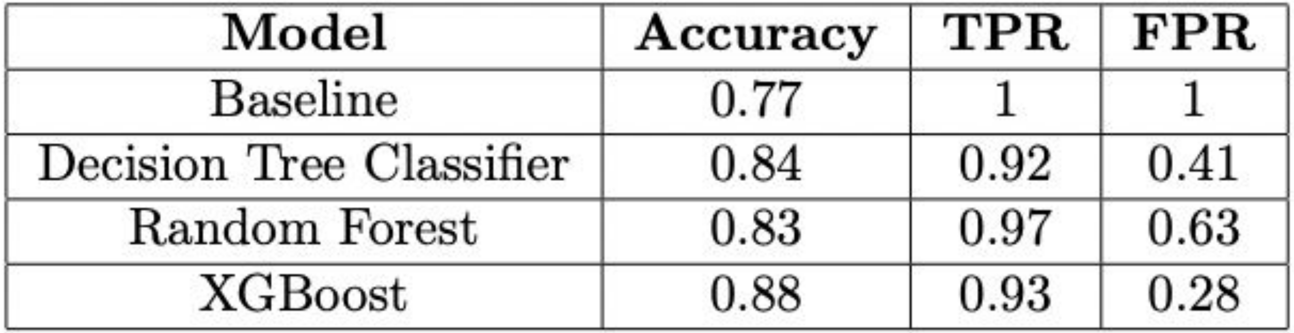
Section Three:

Models and Results

4 Different models :

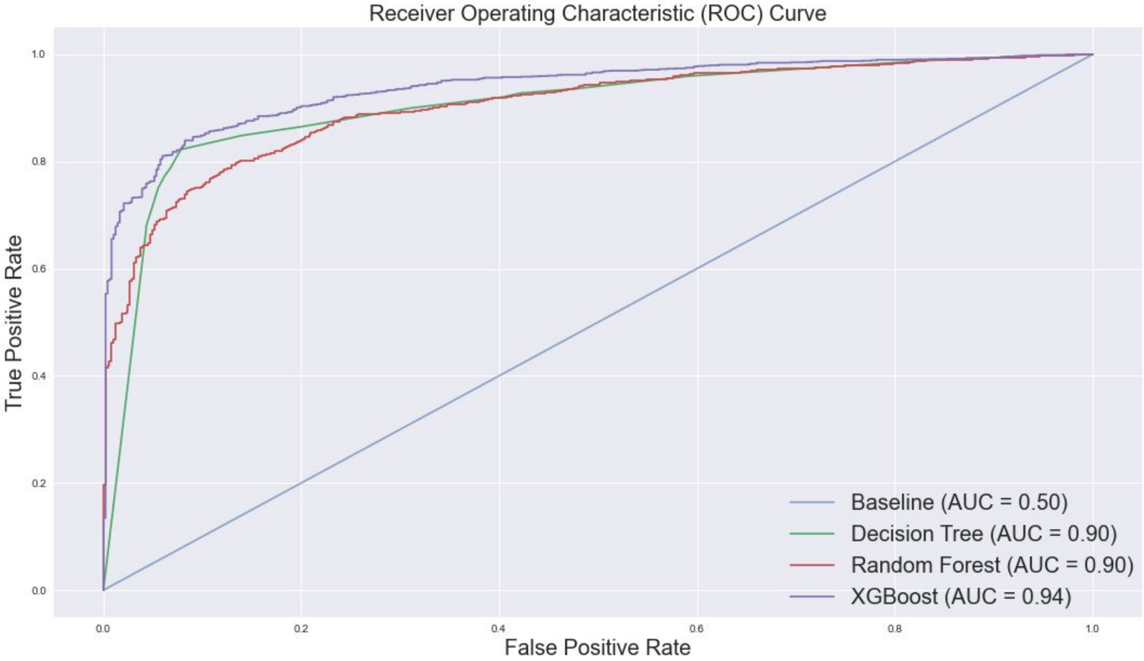
- Baseline
- Decision Tree Classifier
- Random Forest
- XGBoost

4 Different models :



Model	Accuracy	TPR	FPR
Baseline	0.77	1	1
Decision Tree Classifier	0.84	0.92	0.41
Random Forest	0.83	0.97	0.63
XGBoost	0.88	0.93	0.28

4 Different models :



Our Winner

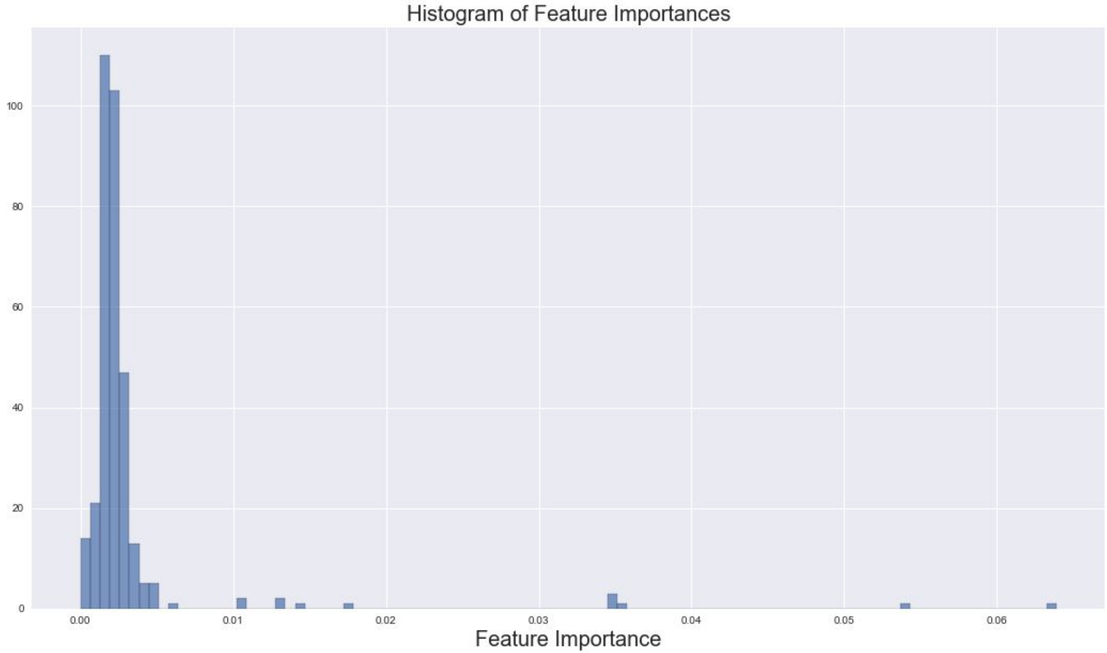
XGBoost overperforms !

Our Winner

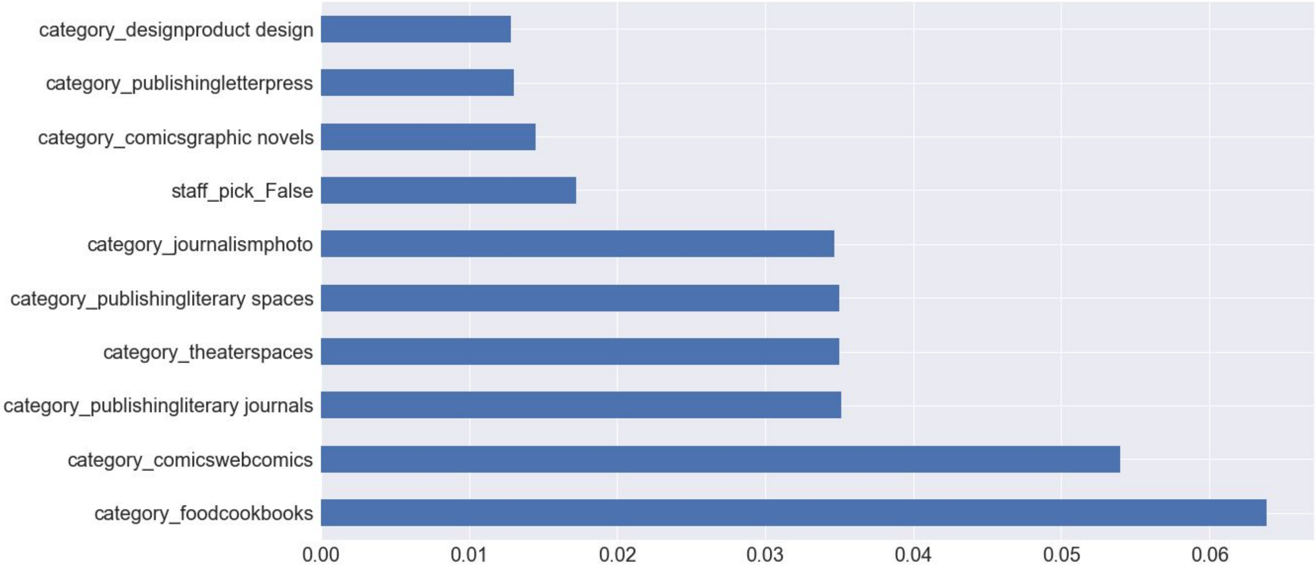
XGBoost overperforms !

But let’s take a look at the feature importances

Feature Importance

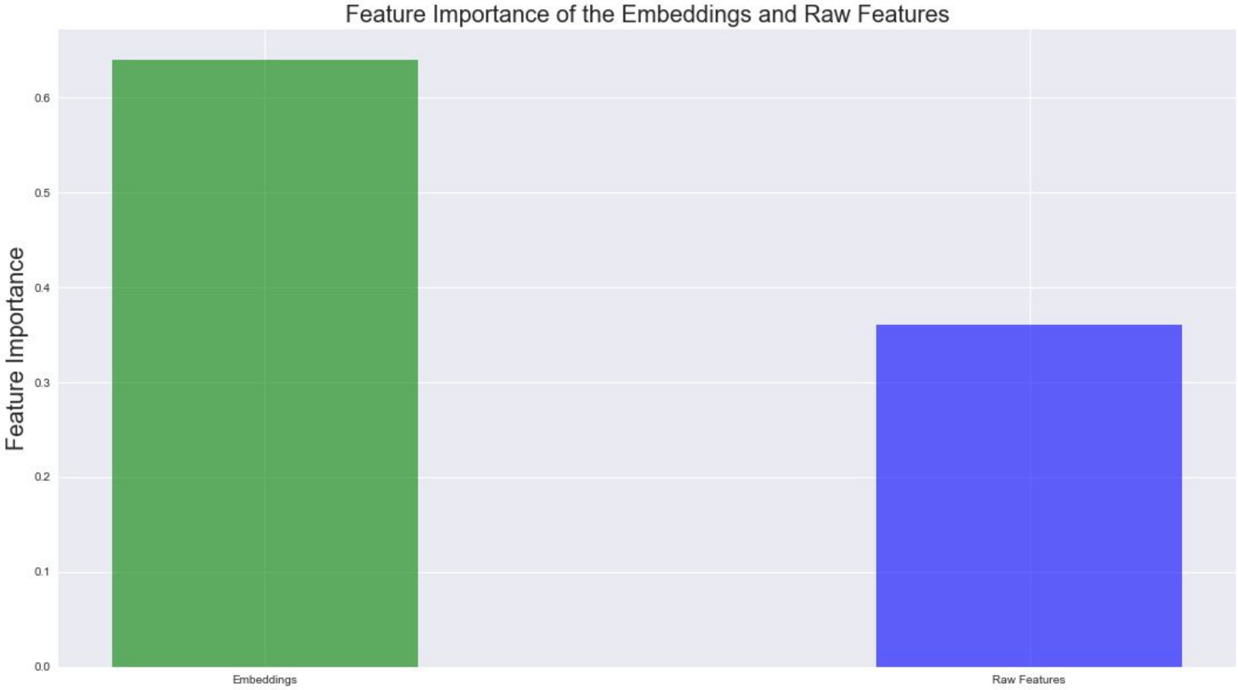


Feature Importance



Average Feature Importance : 0.3%

Feature Importance



Thank you for your Attention