Classification of Kickstarter's successful/failed projects based on data crawled by a scraper robot

Promotion février 2020
Gloria González Curto

Problem definition and goal

Kickstarter: a global crowdfunding platform focused on creativity.

Creators propose projects and fix an economic goal and a deadline.

If the goal is reached, the pledged money is collected from backers.

Goal

Predict if a project is going to succeed or fail its crowdfunding campaign based on data crawled by a web robot from the kickstarter site.

Interest of the project

- Guide creators to succeed in the set up of their project campaign.
- Guide creators to decide to launch a campaign on kickstarter based on their project's general subject

Description of raw data

- Raw data were collected from https://webrobots.io/kickstarter-datasets/
- Crawl executed on February 13 2020
- 57 CSV files with 3500 to 4000 rows each and 38 variables (redundant)
- 6/38 variables are JSON encoded
- Other variables are related to:
  - project text description, profile
  - location
  - goal, pledged, currency, index for conversion into USDs
  - dates (UNIX time)
  - state of the project (TARGET)

Dataset statistics (pandas profiling package)

- Number of variables 38
- Number of observations 206174
- Missing cells 824297
- Missing cells (%) 10.5%
- Duplicate rows 0
- Duplicate rows (%) 0.0%
- Total size in memory 1.2 GiB
- Variable types
  - CAT 18
  - NUM 12
  - BOOL 7
  - URL 1

Sources of data leaks:

- Rows(projects):
  - Errors parsing JSON:
    - Tried to fix errors with regexp and replaces
    - Filter out problematic rows (forecast of 22800 → 206174 rows)

  - Text in several languages
    - Translation
    - Detection of language and translation
    - Custom made filter to select English text (206174 → 122590 rows)

Sources of data leaks:

- Columns (features):
  - Variables with high amounts of missing values
  - Redundant or low information
  - High correlation to target
  - Incomplete information for ongoing projects

data pre-processing: new variables

Creation of variables:

- Dummification of categorical variables

- Date derived variables
  - Year, month, day (created, deadline, launched, state change at)
  - Initial duration of the project
  - Project set-up

- Profile (integer score accounting for profile completeness)

data pre-processing: new variables

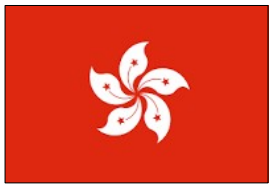
Creation of variables:

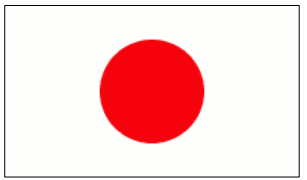
- Frequency score :
  - Selection of 200 keywords by category and project state, and its frequency (computed on training sets)
  - Add the frequency for occurrences of successful keywords
  - Subtract the frequency for occurrences of failed keywords
  - Normalize by text length

Country selection(198 levels): :

- More than 55% of successful projects and more than 50 projects

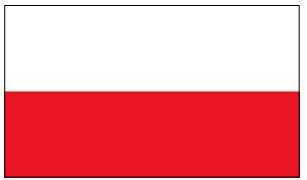

United Kingdom


Hong Kong

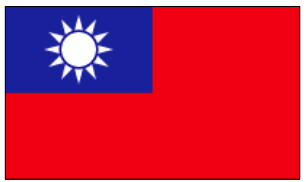

Japan

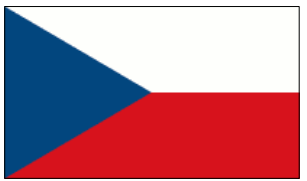

Singapore


China


Poland


Israel


Taiwan


Czech Republic


Greece


Indonesia


Argentina


Kenya


Iceland


Ghana


Portugal


Slovenia

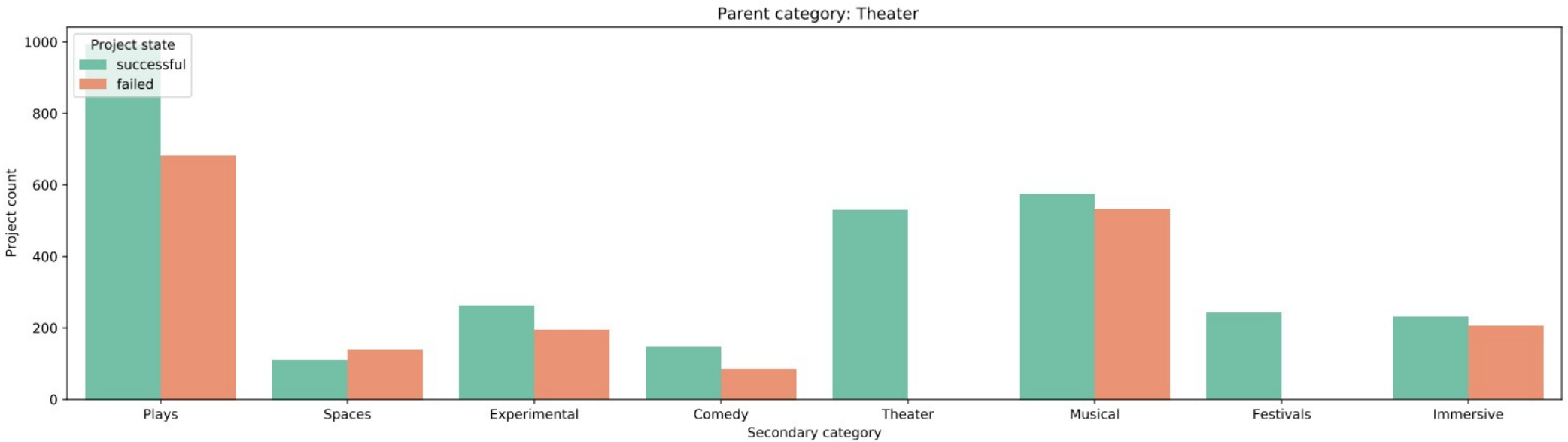

Finland

Selection of 18 countries

data pre-processing: feature selection

Secondary category selection (159 subcategories) :

- More than 55% of successful projects
- Information non redundant to principal category (15)

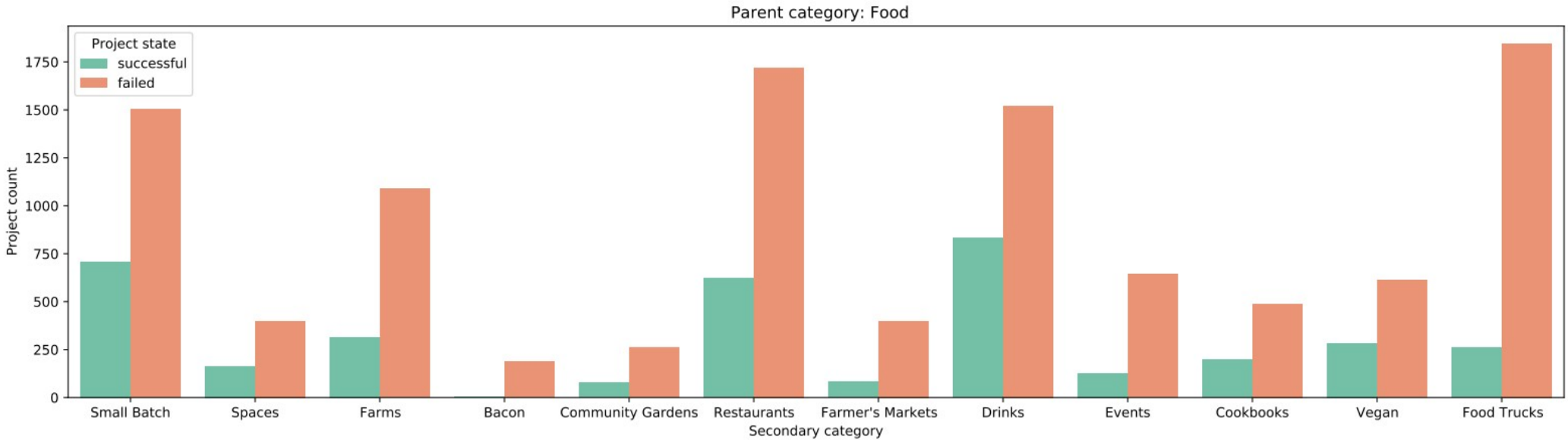


Selection of 46 subcategories

data pre-processing: feature selection

Secondary category selection (159 subcategories) :

- More than 55% of successful projects
- Information non redundant to principal category (15)

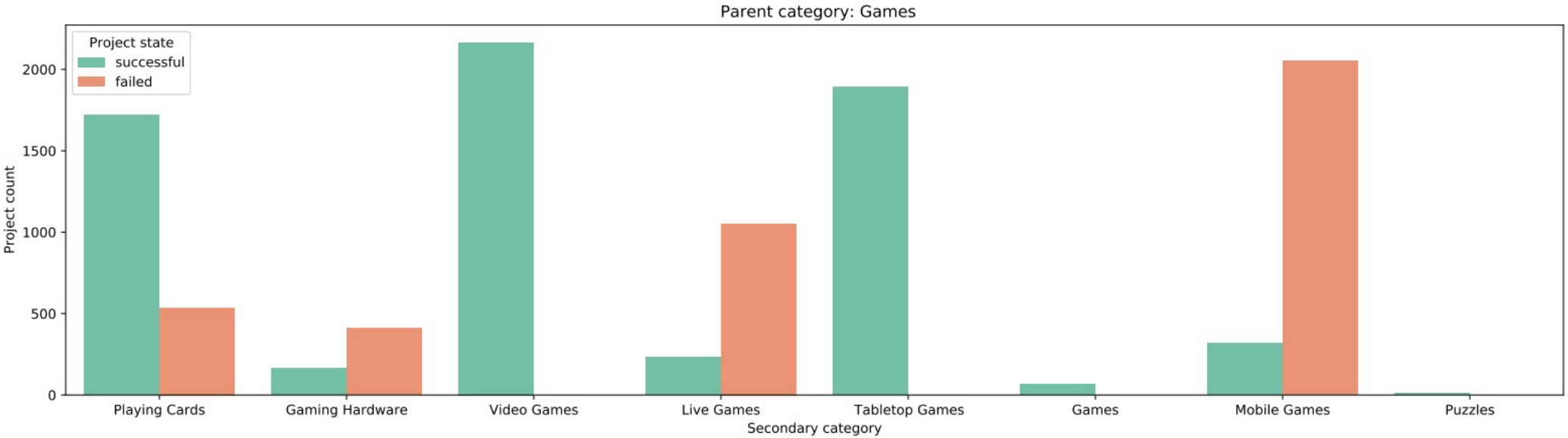


Selection of 46 subcategories

data pre-processing: feature selection

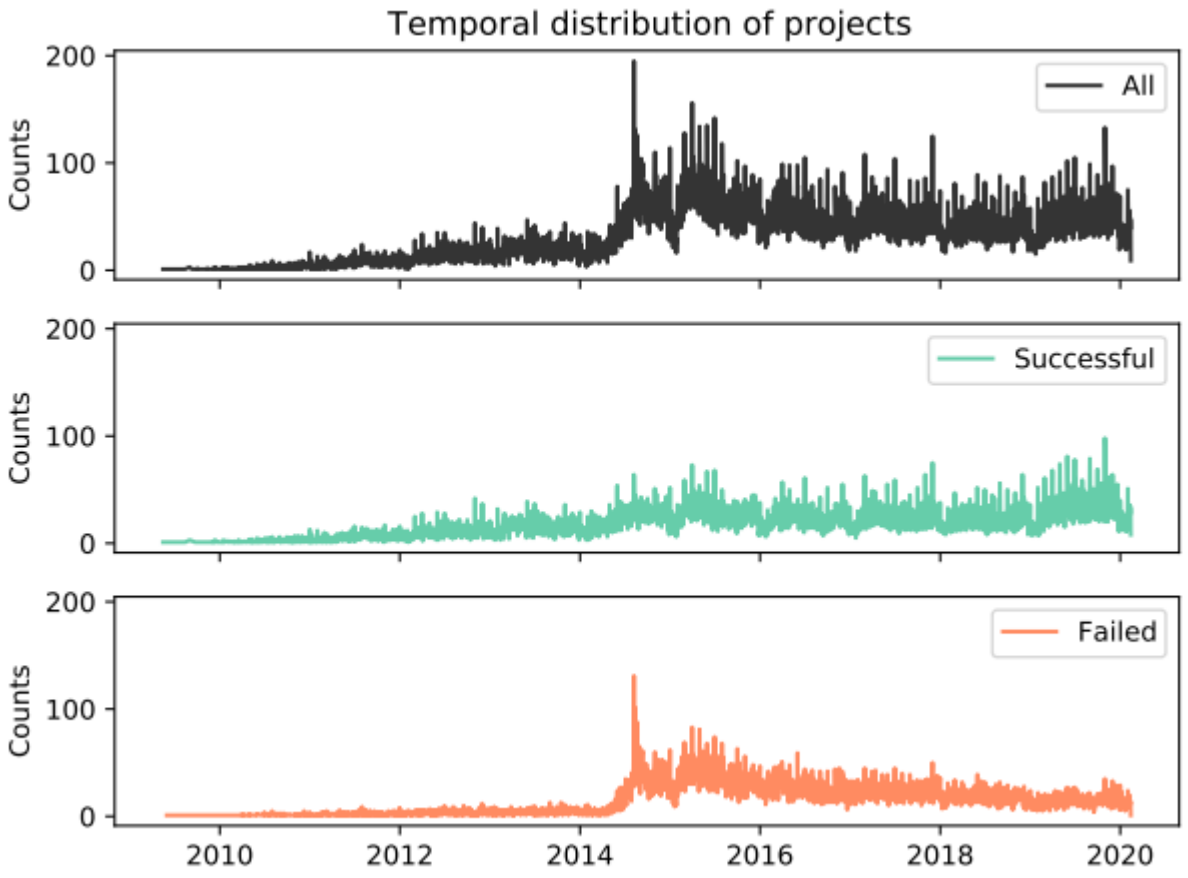
Secondary category selection (159 subcategories) :

- More than 55% of successful projects
- Information non redundant to principal category (15)

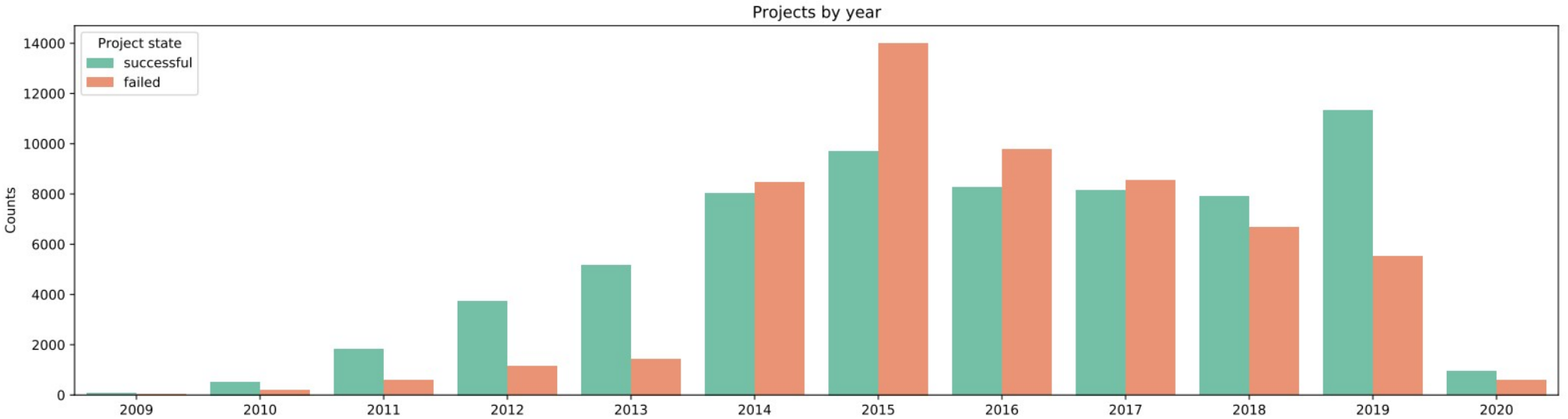


Selection of 46 subcategories

Exploratory data visualization: temporal dimension



Exploratory data visualization: temporal dimension



Exploratory data visualization: wordclouds

Category: Comics

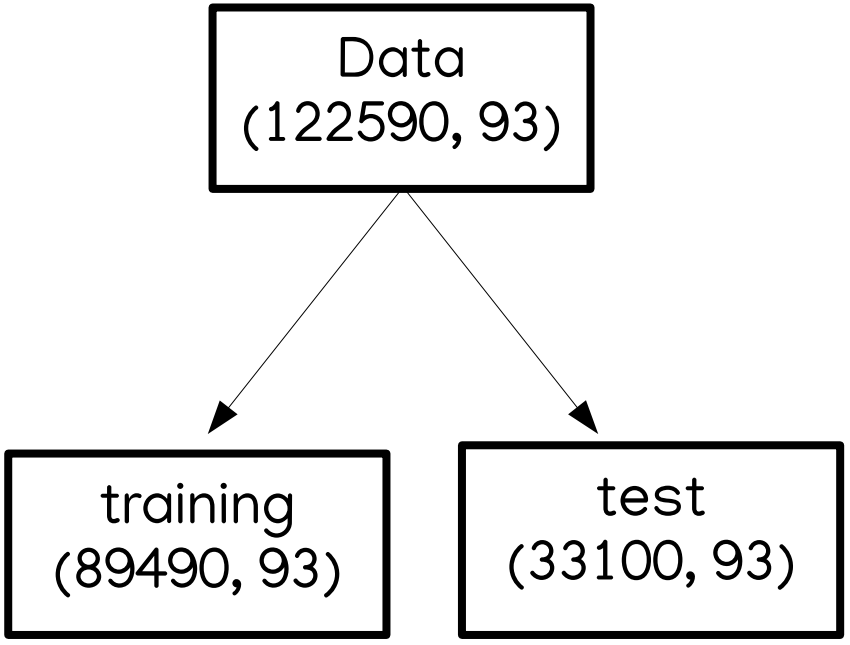
Exploratory data visualization: wordclouds

Category: Fashion

Exploratory data visualization: wordclouds

Category: Food

data pre-processing:
split train and test



XGBoost binary classification: bayesian hyperparameter optimization

Parameters for crossvalidation:

```
'objective': 'binary:logistic',
'max_depth': int(max_depth),
'gamma': gamma,
'learning_rate':learning_rate,
'subsample': subsample,
'eval_metric': 'auc'
```

```
num_boost_round=100
nfold=5
early_stopping_rounds=80
as_pandas=True
seed=37
```

Hyper parameter space:

```
'max_depth': (3, 8), # default 6
'gamma': (0, 5), # default 0
'learning_rate':(0, 1), #default 0.3
'subsample':(0, 1) # default 1
```

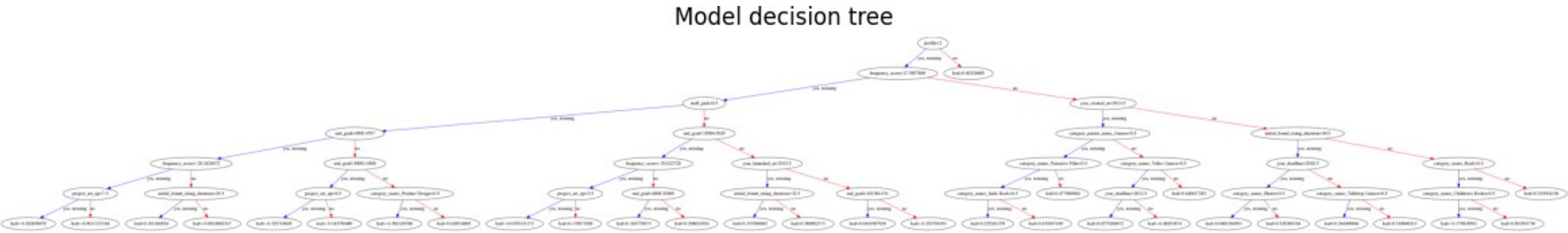
XGBoost binary classification: Evaluation metrics on test set

	precision	recall	f1-score	support
0	0.83	0.89	0.86	15423
1	0.90	0.84	0.87	17677
accuracy			0.86	33100
macro avg	0.86	0.87	0.86	33100
weighted avg	0.87	0.86	0.86	33100

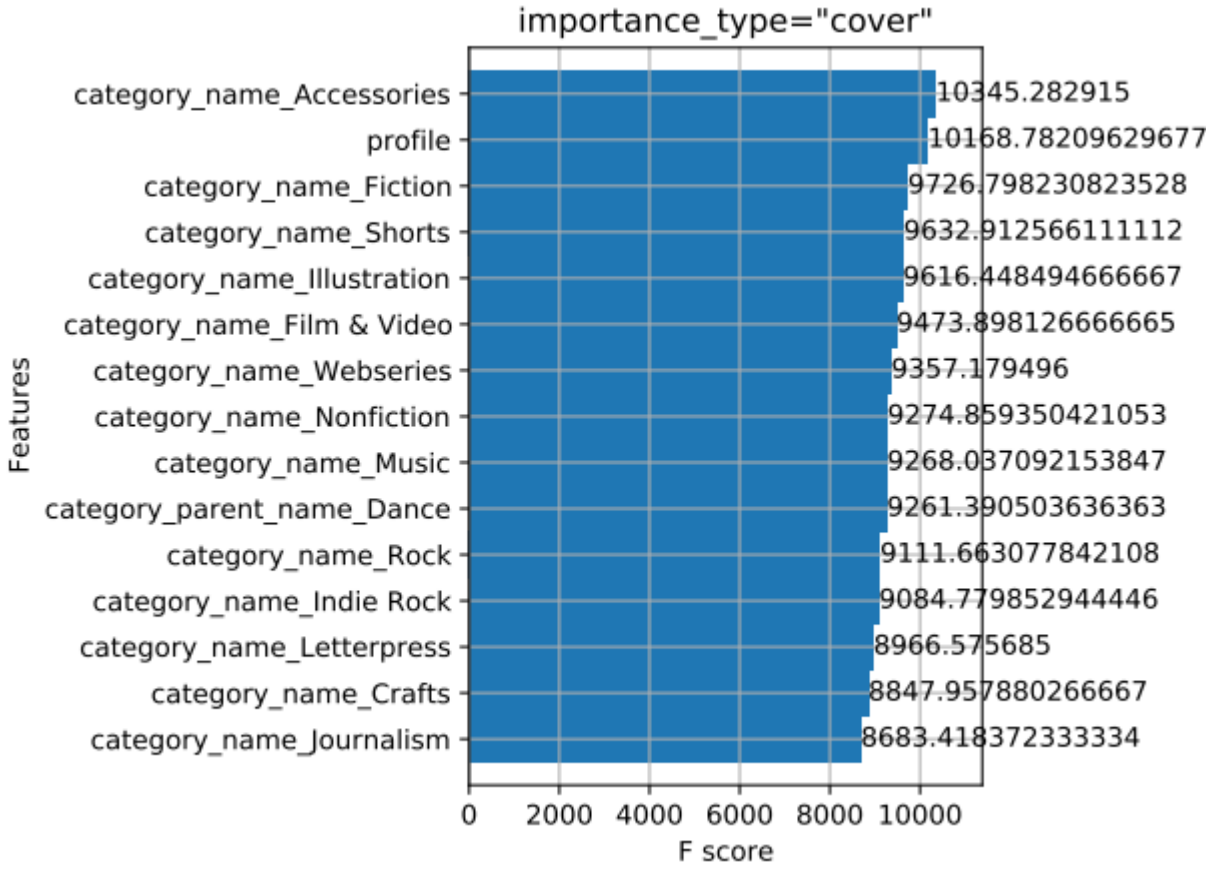
Confusion matrix

```
      0     1
0 [13763  1660]
1 [ 2844 14833]
```

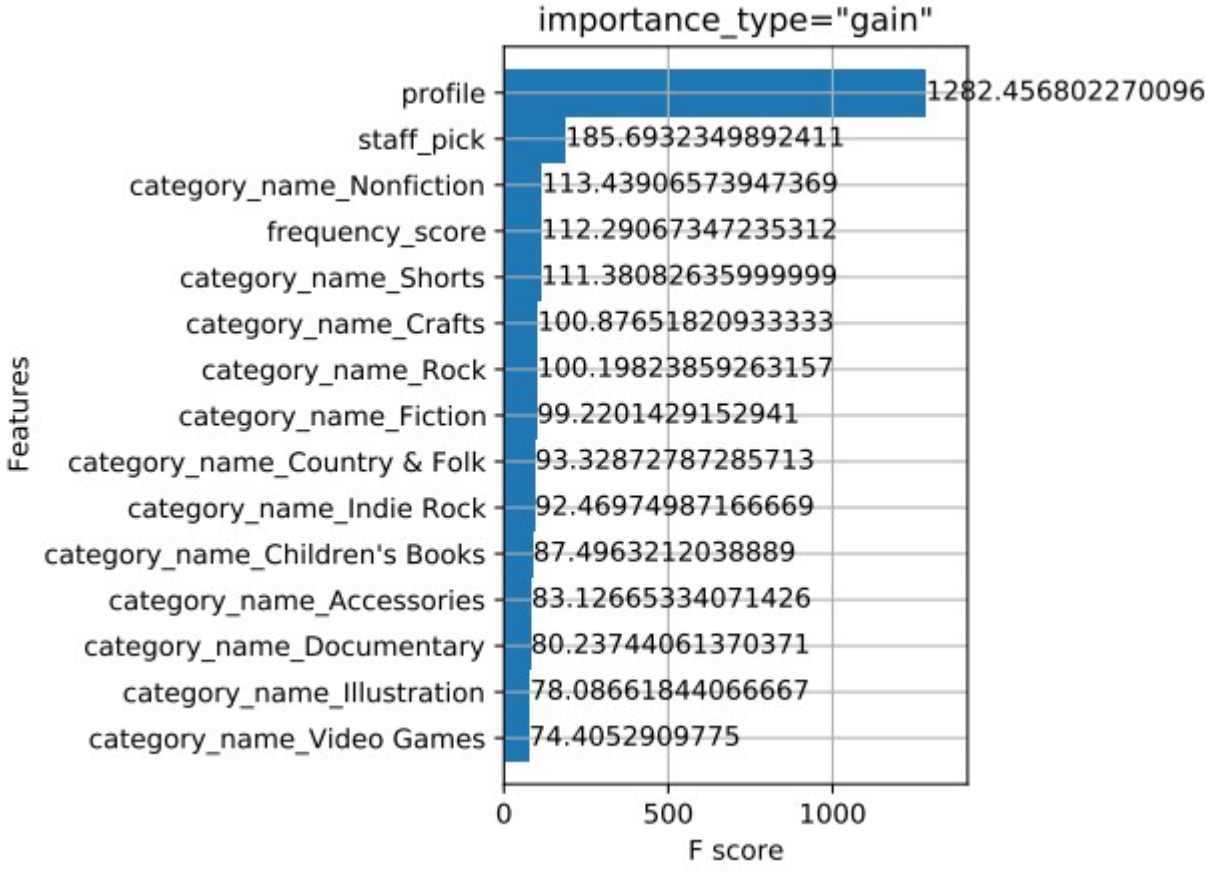
XGBoost binary classification: model interpretation



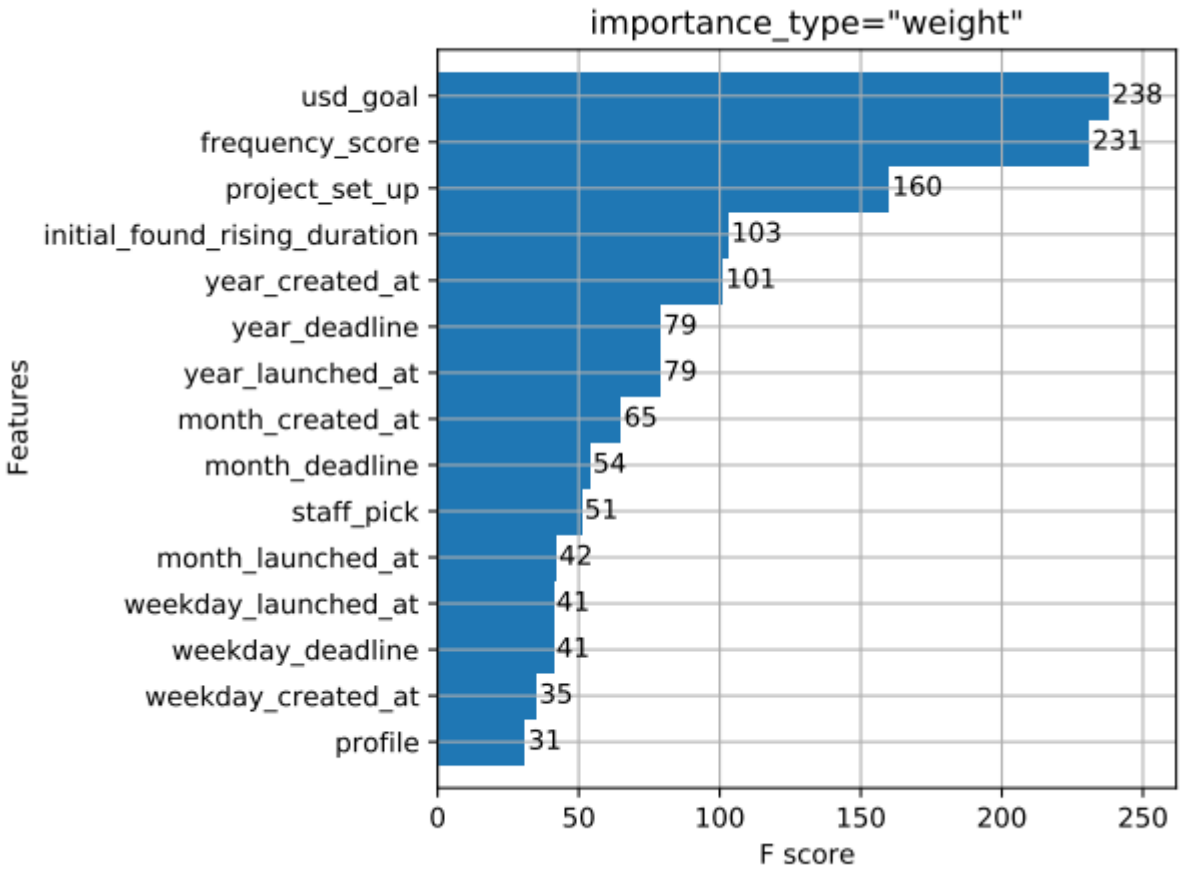
XGBoost binary classification: model interpretation



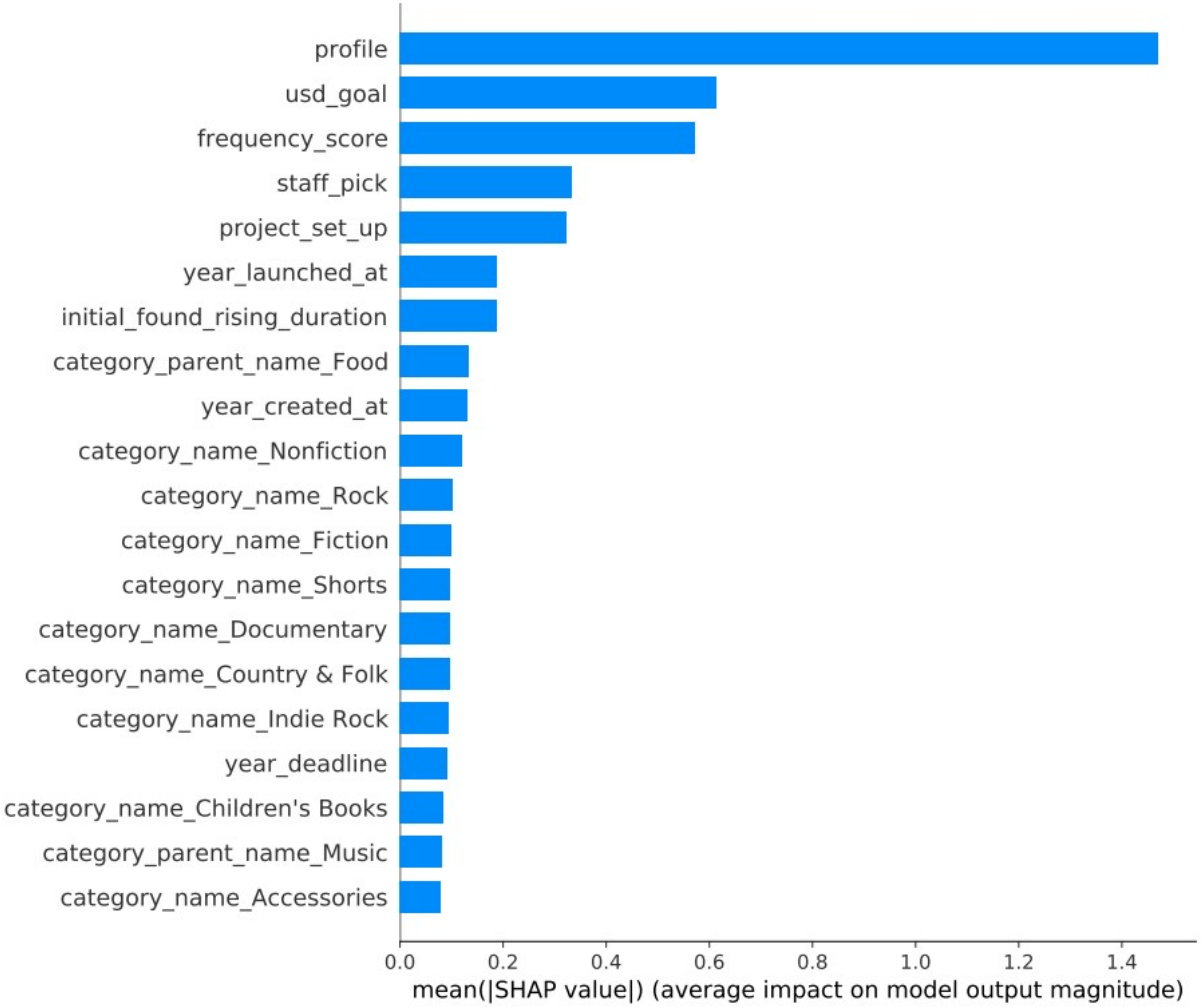
XGBoost binary classification: model interpretation



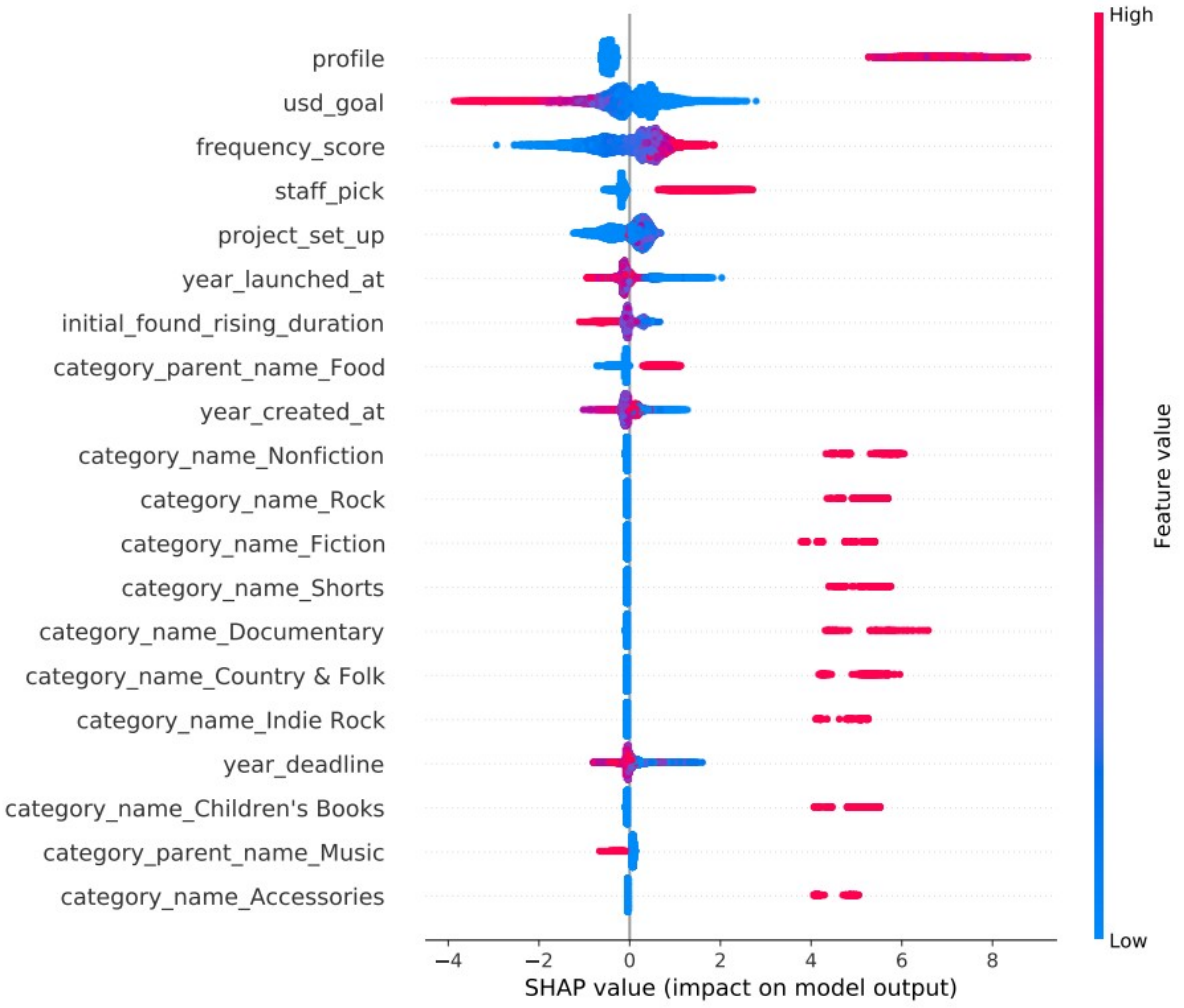
XGBoost binary classification: model interpretation



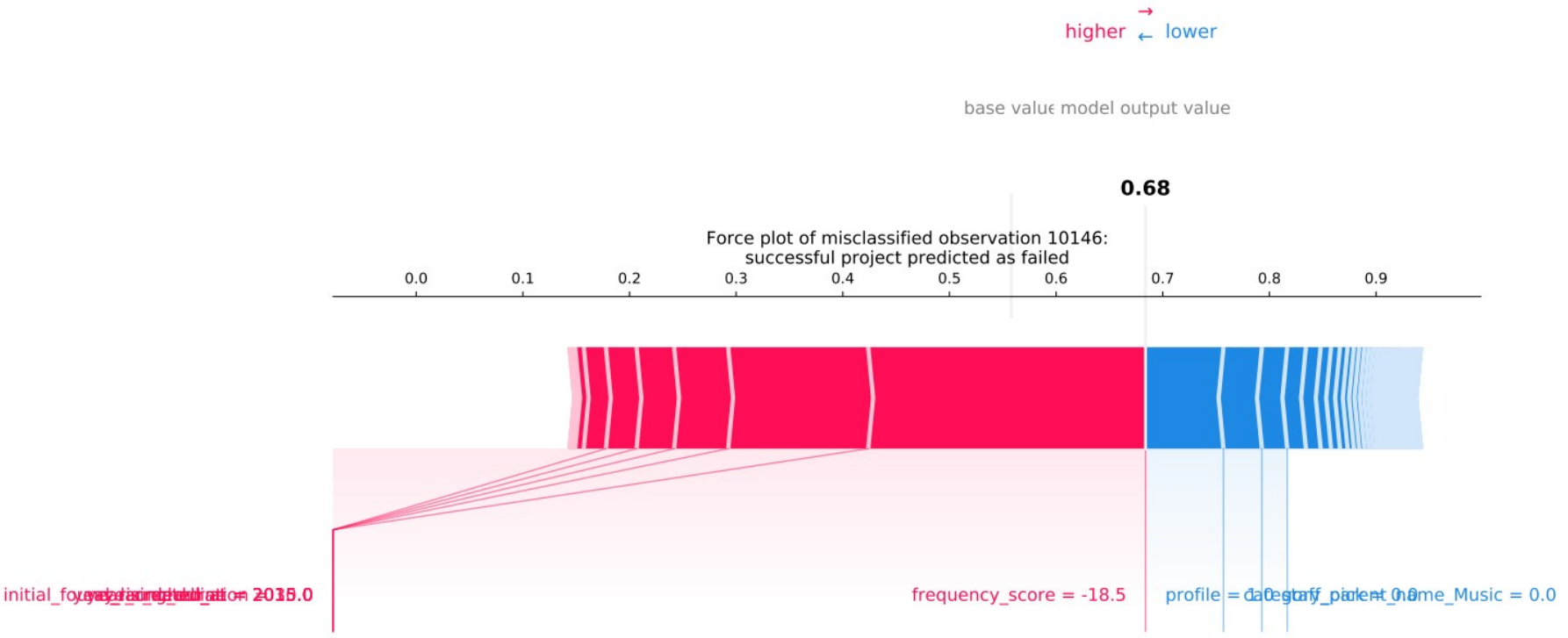
SHAP: model interpretation



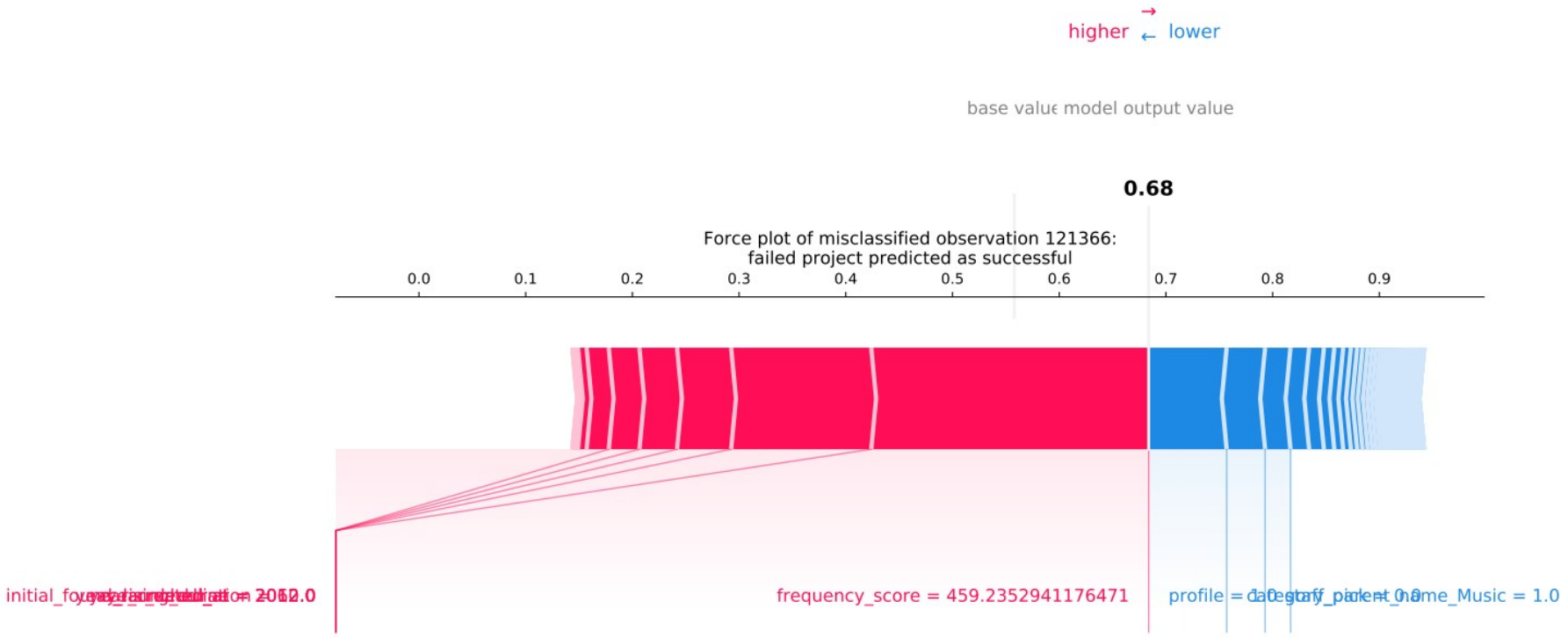
SHAP: model interpretation



SHAP: Error evaluation



SHAP: Error evaluation



Features important for the success of a project:
Category
Complete profile

Saturation tendency (increased rate of failed projects)

Perspectives:
- Perfect model:
  - Complete evaluation of wrong predicitions
  - PCA or Temporal Split
  - Change XGBoost booster (dart booster)
  - Other classifiers
  - New features (google trends)

Development tools

Visual Studio Code

Git + GitHub