Kickstarter Data Modeling

Georgia Barry, Daniel Blanco, Benedicte Kjaerran

8/1/2020

Modeling Process

The steps below are the processes we went through modeling the Kickstarter dataset. A poster will be provided that presents some of these findings in a more visually intuitive manner.

A sample of five years of Kickstarter data, from April 2020 to April 2015, was obtained to make inferences for Kickstarter campaigns.

The questions this report aims to answer are:
1. What factors are likely to increase percentage of a campaigns pledged amount over its goal (goal_ratio)?

2. What factors influence the success or failure of a project launched on Kickstarter?

3. How can the probability of success, or goal ratio, for a given Kickstarter campaign be assessed?

The answers to these questions are in the "Interpret Models" Section of this report

Reduce Indicator Variables

Country was changed to Continents to reduce the number of indicator variables. Continent contains much of the same information as country. Only continents that Kickstarter campaigns can be created in are included. Category (cat_parent) and subcategory were included in the data, but for the same reason as using continent, we only used cat_parent.

Data for LM model

We removed pledged because the idea was to predict goal_ratio, which is a feature engineered from pledged being divided by goal. ID was removed because it was not relevant to the model.
Target was removed because it indicates whether a campaign succeeded or failed, which is not known to a new Kickstarter campaign. Target was used for the binomial model.
Some features with heavy skews were logged. The features that were logged, say x, required log(x+1), since x may have been 0 in some instances.

Test Every Interaction for LM

We used Forward and backward to obtain the best models with BIC. **BIC is used because we are testing every interaction, which would be a large model. BIC is a conservative information criteria that will select a good fit with fewer features then AIC.**

PLS and LASSO were tested without the interactions just to see if these models have similar results to the Forward and Backward models.

It turned out that the residual plots for the models were not random. This lead to the conclusion that the normal distribution was not appropriate for the dataset. We then began trying different distributions.

Gamma Distribution

Gamma distribution is a more skewed tolerant distribution. Changing the 0's for goal_ratio is required for Gamma and Inverse Gaussian distributions. There was an attempt to change them to half of the minimum value in the data to prevent shifting the data, however, it produced errors. So instead we added 1 to each goal_ratio, which shifted all of the data. The error was related to starting values, which could be investigated in the future.

Change the Link Function?

When changing the link function from anything but log, we get this error.
Error: no valid set of coecients has been found: please supply starting values Finding good starting values is dicult. Finding good starting values may be investigated in the future.

Gamma Distro Results

The residual plot was better than LM, however, it still needed improvement.

Inverse Gaussian

Move on to a more skewed distribution to accommodate the skewness in the data.

Residuals Inverse Gaussian

Residuals plot look much better. We started with a set of variables that we believed to be relevant.

Remove some features

We removed some features to reduce the variance of the model and make it more simple. We did this using deviance's. It had no adverse effect on the residual plot.

Lets try all interactions

We tried a step with BIC on all interactions, the residual plot did not look better than the one we fitted. Omitting variables may make the model bias. Fortunately, the models bias caused the model to be more conservative and predict a lower goal ratio. A more conservative model is what someone starting a Kickstarter campaign would want, since they would not want unrealistic expectations. 78.3% of campaigns were predicted to have a goal_ratio less than or equal to 1, while only 44.29% were actually less than or equal to 1.

Final Gaussian Model

The final Inverse Gaussian model is demonstrated in the code below. The RMSE is 6.2410224.

Binominal model

For this model we are predicting target, which is whether a campaign succeeds or fails. We removed goal_ratio because it contained extra information that would not be available to a new campaign.

We used a step-wise approach using all interactions with BIC again to reduce the number of coefficients. The residual plot were strange, so we tried different approaches.

A model that was built without interactions and using Step AIC had similar performance with less features, meaning less variance. The residual plot was still strange. The model is good at discriminating with an accuracy above 94% and AUC above 98%.

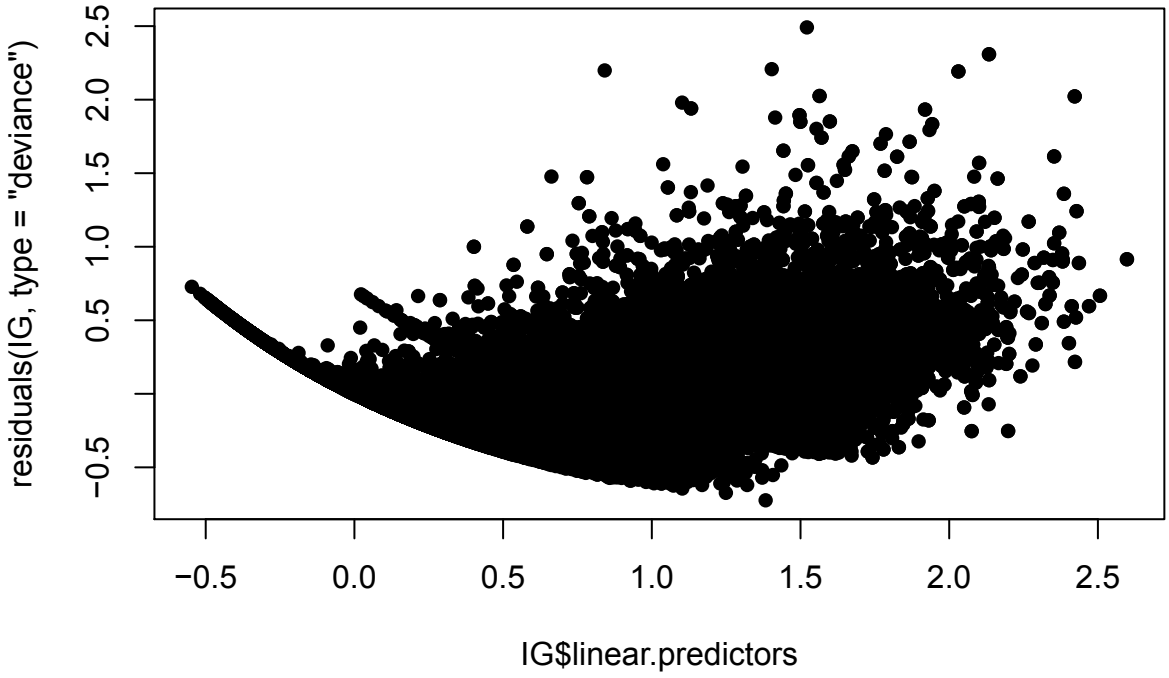
The residual plot makes the model not appear to be calibrated because the errors around 50% have high variances. However, 95.85% of the predictions are greater then 65% or less then 35% in probability ranges. The predictions outside of the 35-65% range have less errors, so the calibration issue is not likely to affect a campaigns.

Since this model was built using Step AIC, and interactions did not improve the residual plot, we expected it to be less bias. The results show that this model is less bias with 58% of the predictions being "successful" and the actual data having 56.67% "successful" campaigns.

Final Gaussian Model

```
IG<-glm(goal_ratio ~ + cat_parent+month+number_of_days+
     backers_count_log+continent+(staff_pick) * (goal_log),family=inverse.gaussian(link=log),IG_data)
```

```
plot(IG$linear.predictors,residuals(IG,type="deviance"),pch=16)
```




```
postResample(IG$linear.predictors, IG$data$goal_ratio)
```

```
     RMSE  Rsquared       MAE
6.2410224 0.1228689 1.8519327
```

```
length(kickstarter$goal_ratio[kickstarter$goal_ratio <= 1])
```

```
[1] 50679
```

```
length(IG$linear.predictors[IG$linear.predictors <= 1])
```

```
[1] 89611
```

Final Binomial Model

```
binominal_baseline <- glm(target ~., data = bi_data, family = binomial)
binominal_baseline <-step(binominal_baseline,trace=FALSE)
```

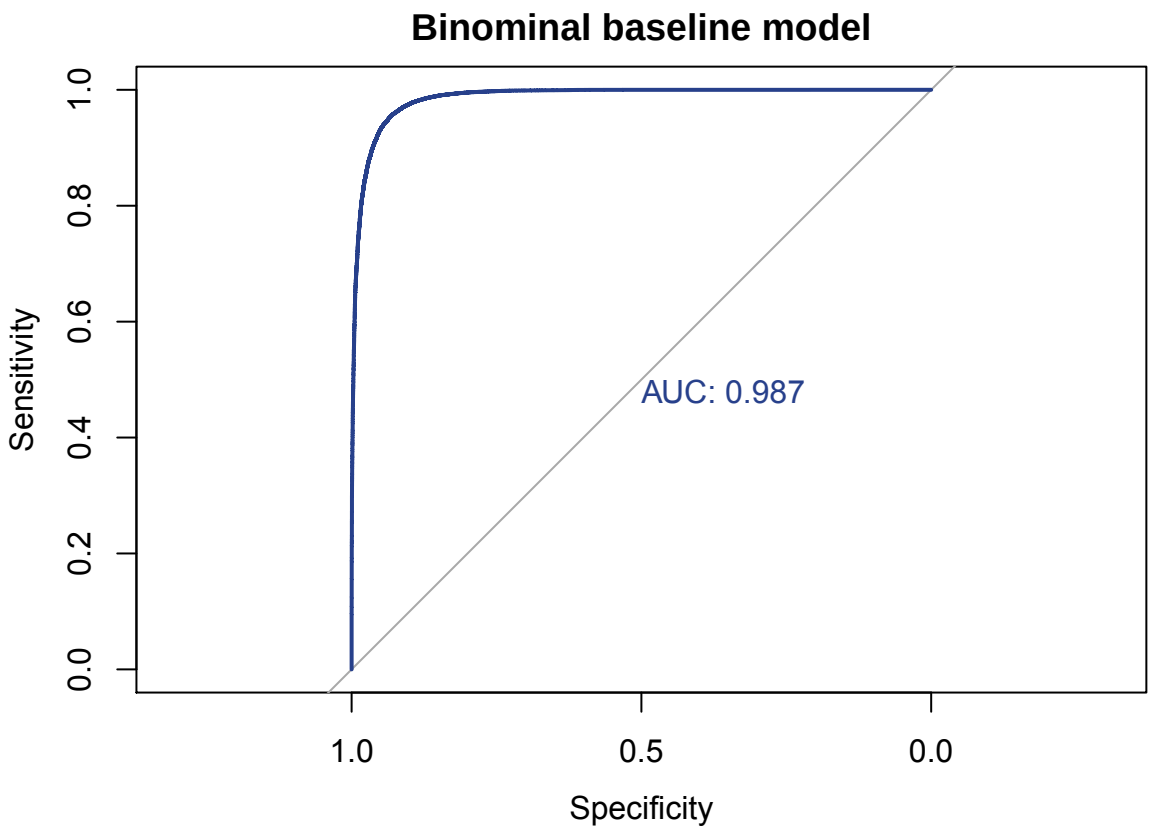
Confusion Table Stats

Table 1: Model performance in percentage

Accuracy	Sensitivity	Specificity	Precision
94.49	95.13	94.04	92.01

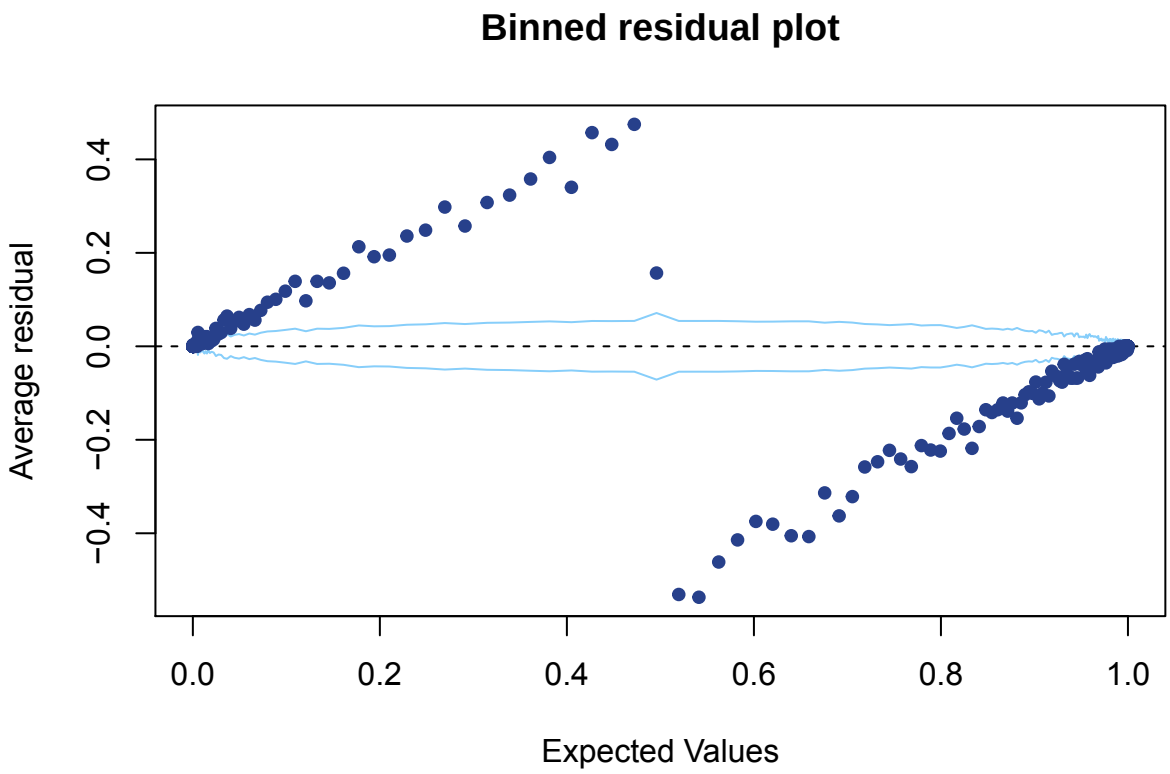
ROC

```
plot.roc(results$target,results$fit, col = "royalblue4", backgroundcol = "lightskyblue",
         main = "Binominal baseline model", print.auc = TRUE)
```



Binned Residuals

```
binnedplot(results$fit, as.numeric(results$target) - as.numeric(results$ClassPredict), col.pts = "royall
```




```
nrow(results)
```

```
[1] 114432
```

```
length(results$fit[results$fit>0.65 | results$fit<0.35 ])
```

```
[1] 109682
```

```
length(results$ClassPredict[results$ClassPredict=="successful"])
```

```
[1] 66472
```

```
length(results$target[results$target=="successful"])
```

```
[1] 64845
```

Interpret models

Ceteris Paribus - "all other things equal". Each individual features interpretation for the models is ceteris paribus.

Goal Ratio Prediction (Question 1)

1. What factors are likely to increase percentage of a campaigns pledged amount over its goal (goal_ratio)?

```
summary(IG)$coefficients[,1:2]
```

```
                            Estimate   Std. Error
(Intercept)               0.5127728804 0.0179106742
cat_parentComics         -0.0772154826 0.0090402833
cat_parentCrafts         -0.0113483669 0.0093186182
cat_parentDance          -0.0449010621 0.0127380325
cat_parentDesign          0.3381254852 0.0102741279
cat_parentFashion         0.0499993827 0.0079970065
cat_parentFilm & Video   -0.0161521436 0.0063305960
cat_parentFood           -0.0263589754 0.0066439581
cat_parentGames           0.0708505644 0.0080516094
cat_parentJournalism     -0.0194273292 0.0089016676
cat_parentMusic          -0.0254894423 0.0066364770
cat_parentPhotography    -0.0367333960 0.0087515027
cat_parentPublishing     -0.0383398641 0.0070903525
cat_parentTechnology      0.0735712356 0.0065077807
cat_parentTheater        -0.0293028318 0.0090037935
monthAug                  0.0141314018 0.0066060896
monthDec                  0.0190161160 0.0072131427
monthFeb                 -0.0011441451 0.0066812866
monthJan                  0.0243427626 0.0067732771
monthJul                  0.0072929802 0.0066052402
monthJun                  0.0006369510 0.0064518725
monthMar                  0.0053178349 0.0066418856
monthMay                 -0.0036382616 0.0063367117
monthNov                  0.0119907087 0.0065269607
monthOct                  0.0020460983 0.0064552586
monthSep                  0.0009236139 0.0066137514
number_of_days           -0.0002852526 0.0001155968
backers_count_log         0.2340420202 0.0008467952
continentEurope          -0.1673170965 0.0131203716
continentNorthAmerica    -0.1434220953 0.0129717965
continentPacific         -0.2067967058 0.0244518140
continentSouthAmerica    -0.1971919805 0.0159774218
staff_picktrue            0.8823737710 0.0388035465
goal_log                 -0.0594755768 0.0011341998
staff_picktrue:goal_log  -0.0992342290 0.0041922543
```

The model for predicting goal_ratio found that the most significant binary indicator for a high goal_ratio is if the campaign is a staff pick. If a campaign is a staff pick, goal_ratio prediction will increase by 88.23%.

The most significant category Design. A campaign with in Design category will have a goal_ratio prediction increase of 33.81%.

For every increase of backers count by 1%, goal ratio will increase 0.23%.

January is the best month to have a campaign, and May is the worse one.

What features has the model found will negatively impact goal_ratio?

Every continent, outside of Asia, effect the model negatively. The effects range from -14.34% in North America, to -20.68% in the Pacific continent. It is important to consider that the campaigns are not evenly distributed among the continents. Below is a table of the distribution of continents. It may be that the lack of campaigns in Asia has not created enough underfunded campaigns.

```
summary(kickstarter[c('continent')])
```

```
        continent
 Asia          : 2028
```

```
 Europe      :29006
 NorthAmerica:80795
 Pacific     :  495
 SouthAmerica: 2108
```

The model shows that Comics and Dance categories have the most adverse effect of the categories on goal_ratio.

```
summary(kickstarter$cat_parent)
```

```
         Art       Comics       Crafts        Dance       Design
        7372         5887         3019         1676         5135
     Fashion Film & Video         Food        Games   Journalism
        6545        15905        11036         8088         3168
       Music  Photography   Publishing   Technology      Theater
       12232         4137        10604        15396         4232
```

As expected, having a large goal (goal_log) has a negative effect on achieving that goal. This is increased with the interaction of staff pick being true and goal. In other words, each increase in 1% of goal decreases goal_ratio by 0.059%, with an additional decrease of 0.099% if staff pick is true. The implication is that staff_pick true will increase the overall goal_ratio by 88%, but for each percentage increase in goal_ratio, a decrease 0.158% will be applied to goal ratio. This effect will generally not outweigh the benefit of being a staff pick, unless the goal is very high.

Only 17% of the campaigns were staff pick from the five year sample of Kickstarter data obtained.

```
summary(kickstarter$staff_pick)[2]/summary(kickstarter$staff_pick)[1]
```

```
     true
0.1708156
```

Success or Failure Prediction (Questions 2)

2. What factors influence the success or failure of a project launched on Kickstarter?

```
summary(binominal_baseline)$coefficients[,1:2]
```

```
                          Estimate  Std. Error
(Intercept)             7.982860584 0.185742110
cat_parentComics       -1.256075478 0.090645995
cat_parentCrafts       -0.366583082 0.104404168
cat_parentDance         1.294478776 0.128492280
cat_parentDesign        0.852681424 0.093293824
cat_parentFashion       0.759063787 0.080377350
cat_parentFilm & Video  0.954874714 0.065966679
cat_parentFood         -0.113791431 0.071282727
cat_parentGames        -1.039447358 0.088974203
cat_parentJournalism   -0.747490627 0.110449669
cat_parentMusic         0.453087839 0.070754219
cat_parentPhotography   0.001077204 0.087862381
cat_parentPublishing    0.079112571 0.073256651
cat_parentTechnology   -0.127819881 0.070046858
cat_parentTheater       1.202259289 0.089384051
number_of_days         -0.008154855 0.001342114
name_length             0.009817536 0.005648891
goal_log               -2.016649396 0.021153059
continentEurope        -1.039292436 0.114997909
```

```
continentNorthAmerica  -0.815941730 0.113148681
continentPacific       -1.307987279 0.246654842
continentSouthAmerica  -1.303752721 0.153745055
backers_count_log       3.122941402 0.023905857
```

There are no serious conflicts between the discrete prediction model and binomial model, which is a good sign. The continents show similar results, with Asia being the only one without an adverse effect.

The most significant factor for increasing the likelihood of success is backers_count_log.

The main conflict was category Dance, which this model predicts increases odds more than any other category. It may be that dance categories succeed, but do not exceed their goal amounts very often.

Just as expected, having a higher goal reduces the probabilities of success. What was very strange was this model did not find staff pick to be useful for predicting probability of success.

Assessing sample campaigns (Question 3)

3. How can the probability of success, or goal ratio, for a given Kickstarter campaign be assessed?

Lets start by taking a sample from the Kickstarter data. We will modify that sample to demonstrate how different type of campaigns will have different predictions.

```
set.seed(19491)
sample <- kickstarter %>% sample_n(1)
high <- sample

high$staff_pick <- factor("true",levels=levels(high$staff_pick))
high$month <- factor("Jan",levels=levels(high$month))
high$cat_parent <- factor("Design",levels=levels(high$cat_parent))
high$backers_count_log <- log(500)
high$goal_log <- log(50000)

predict(IG,high)
```

```
       1
1.344621
```

```
predictLog(high, binominal_baseline, t="successful", f="failed")$fit
```

```
[1] 0.9956433
```

We can see that our model predicts a 99.56% chance of success of raising $50,000, and expects to raise 1.34 times that amount.

Lets try doubling the goal amount.

```
high$goal_log <- log(100000)
predict(IG,high)
```

```
       1
1.234612
```

```
predictLog(high, binominal_baseline, t="successful", f="failed")$fit
```

```
[1] 0.9826019
```

Even after doubling the goal, our models predict similar values

How about decreasing the number of backers 80%?

```
high$backers_count_log <- log(100)
predict(IG,high)
```

```
       1
0.8579359
```

```
predictLog(high, binominal_baseline, t="successful", f="failed")$fit
```

```
[1] 0.2704504
```

It seems like our model now predicts a failed campaign, and expects 85% of the goal amount to be raised.

Conclusion

We have produced models that answer the following questions:

1. What factors are likely to increase percentage of a campaigns pledged amount over its goal (goal_ratio)?
2. What factors influence the success or failure of a project launched on Kickstarter?
3. How can the probability of success, or goal ratio, for a given Kickstarter campaign be assessed?

The coalesced answers to these questions provide useful context for anyone considering Kickstarter as a funding option.

The goal_ratio predictions are much more conservative than the probability predictions, meaning goal_ratio will likely be understated.

Probability predictions from the binomial model were directionally correct 94.49% of the times. Area Under Curve (AUC) represents a models aggregate ability to discriminate between the two categories. The binomial models AUC of 0.987 confirms it is more than suitable for predicting "Successful" or "failure" of Kickstarter campaigns. Looking at the residual plot for the binomial model, campaigns near the 50% region should not be trusted due to increased variance. Further investigations found that probability predictions near the 50% mark are rare (5% of predictions) , so this should not be an issue fore most using the model.

Anyone starting a Kickstarter campaign should benefit from the conservative goal_ratio predictions with the probability predictions. Additionally, an understanding of the coefficients that increase the goal_ratio and probability of success improve planning for campaigns. The appendix of this report provides useful visualizations of the data that complement the findings of this report and provide more useful context for anyone considering a Kickstarter campaign. The visual aids convey understanding to the distributions of the data that may effect the models conclusions in regard to a campaigns. Additionally, a nice visualization of the models coefficient values, along with 95% confidence intervals, is provided. The coefficient values visual helps conceptualize what features have a more adverse or positive effect with less difficulty than looking at the raw numbers. Simply put, the larger coefficient values have a positive impact, and negative values have an adverse impact on probabilities or goal_ratio.

Notes

The models include backers count. It is true that this would not be known, but the models included it so that individuals starting a Kickstarter campaign would need to know the number backers to obtain a goal.

There may be factors that influence a project's outcome that were not included in our modeling data set. Factors, such as image quality, number of social media friends, project creators' past success on Kickstarter, project creators' background, choice of words, and promotional efforts were not factored in the model. Other factors that may lead to a successful campaign may not have been identified. The models and answers to the final report's questions should be treated as insight, not causal.

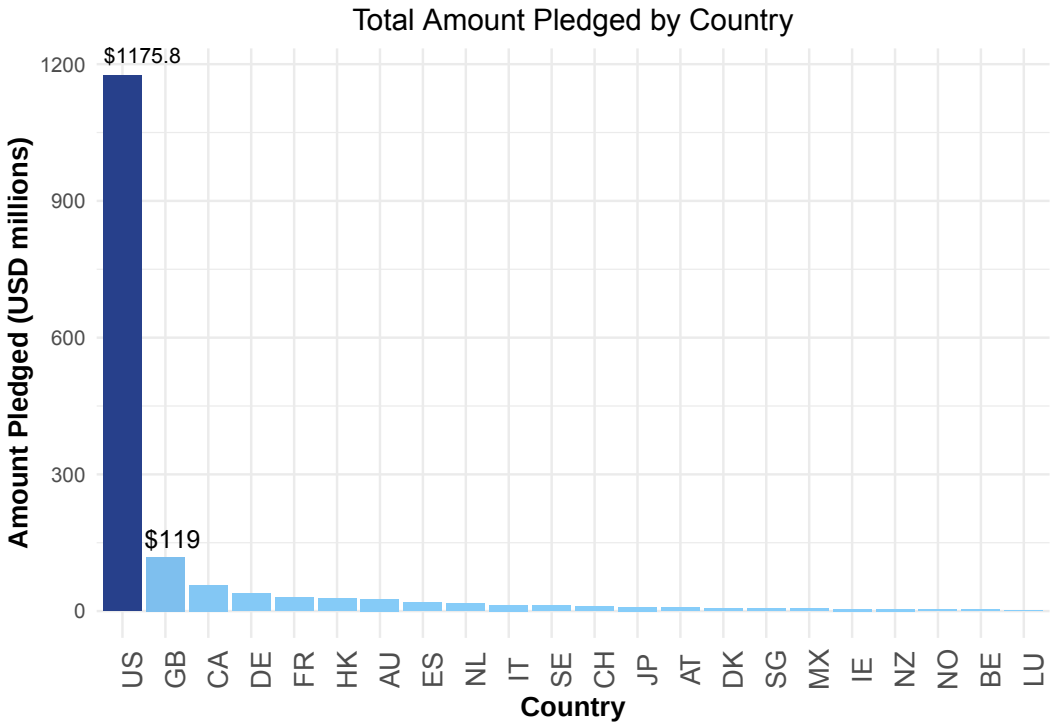
Areas of future research would include predicting the number of backers, correcting starting position errors related to different link functions and not shifting the goal_ratio by 1. The Rsquared below 13% indicates that the model may need more adjustments, and trying different link functions may improve it.

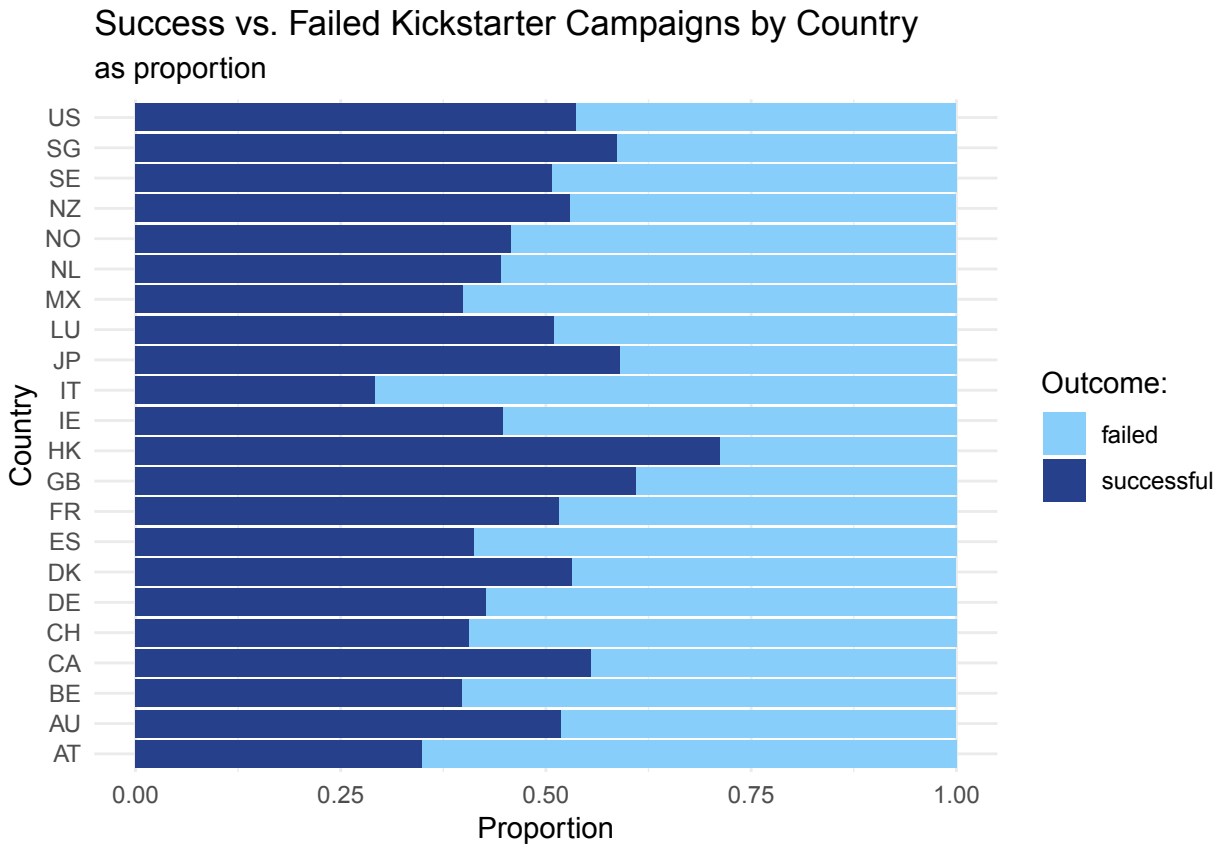
References

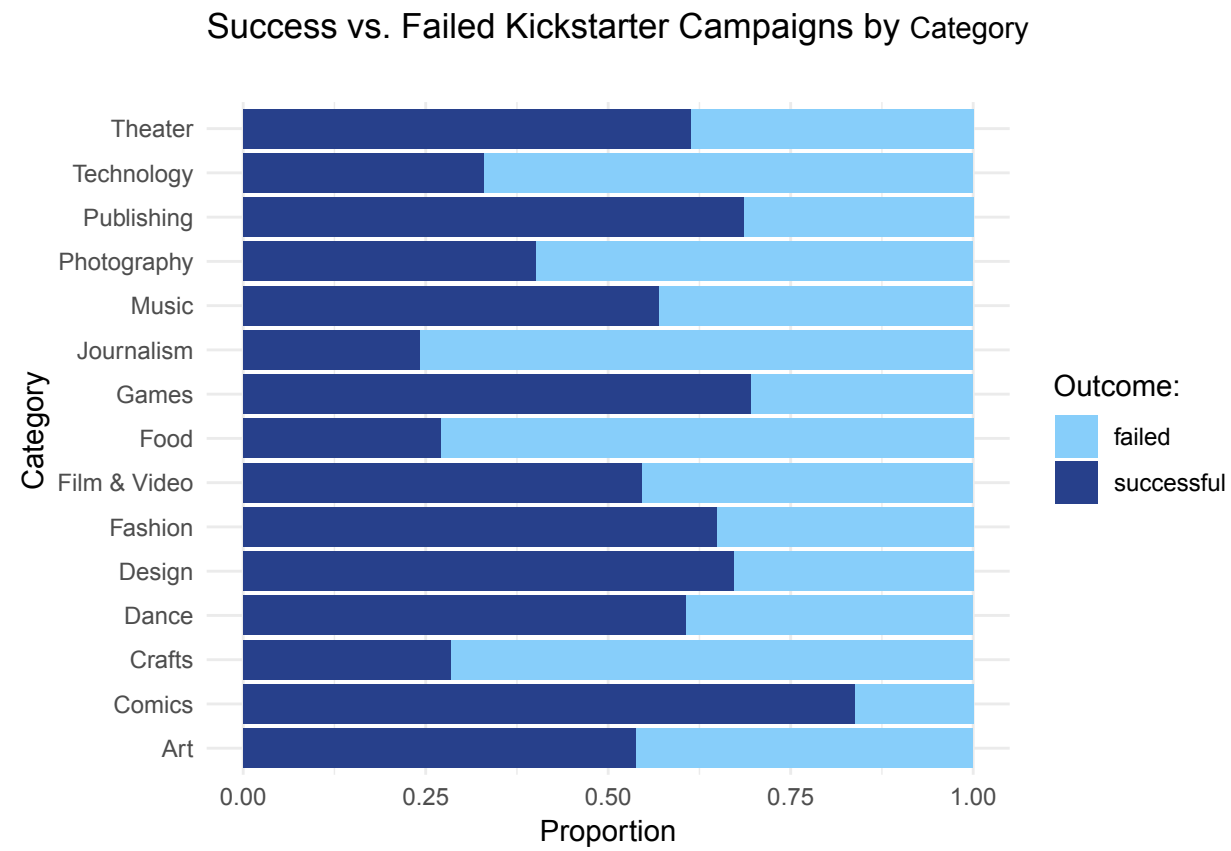
Kickstarter. (2020, July 20). Kickstarter Stats. Retrieved from Kickstarter: https://www.kickstarter.com/help/stats

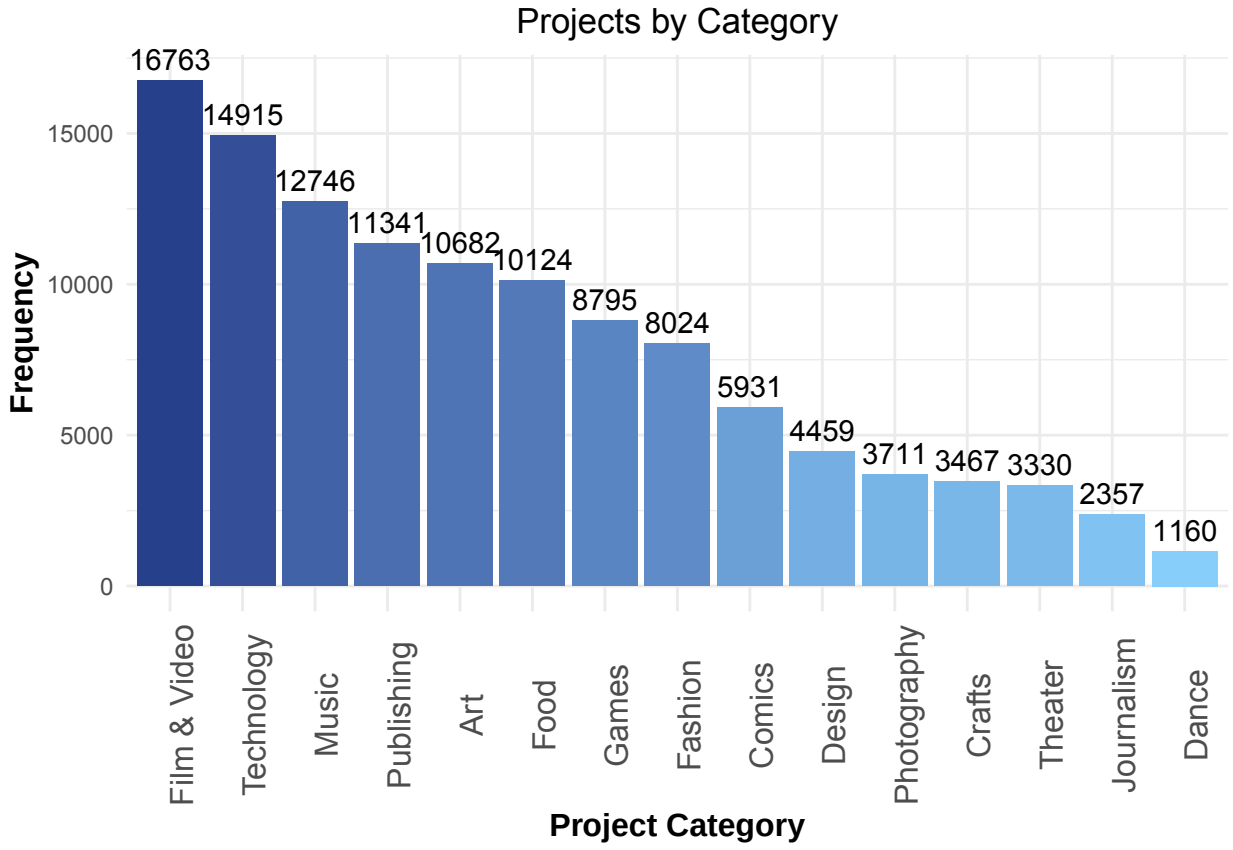
Web Robots. (2020, July). Kickstarter Datasets. Retrieved from Web Robots: https://webrobots.io/kickstarter-datasets/

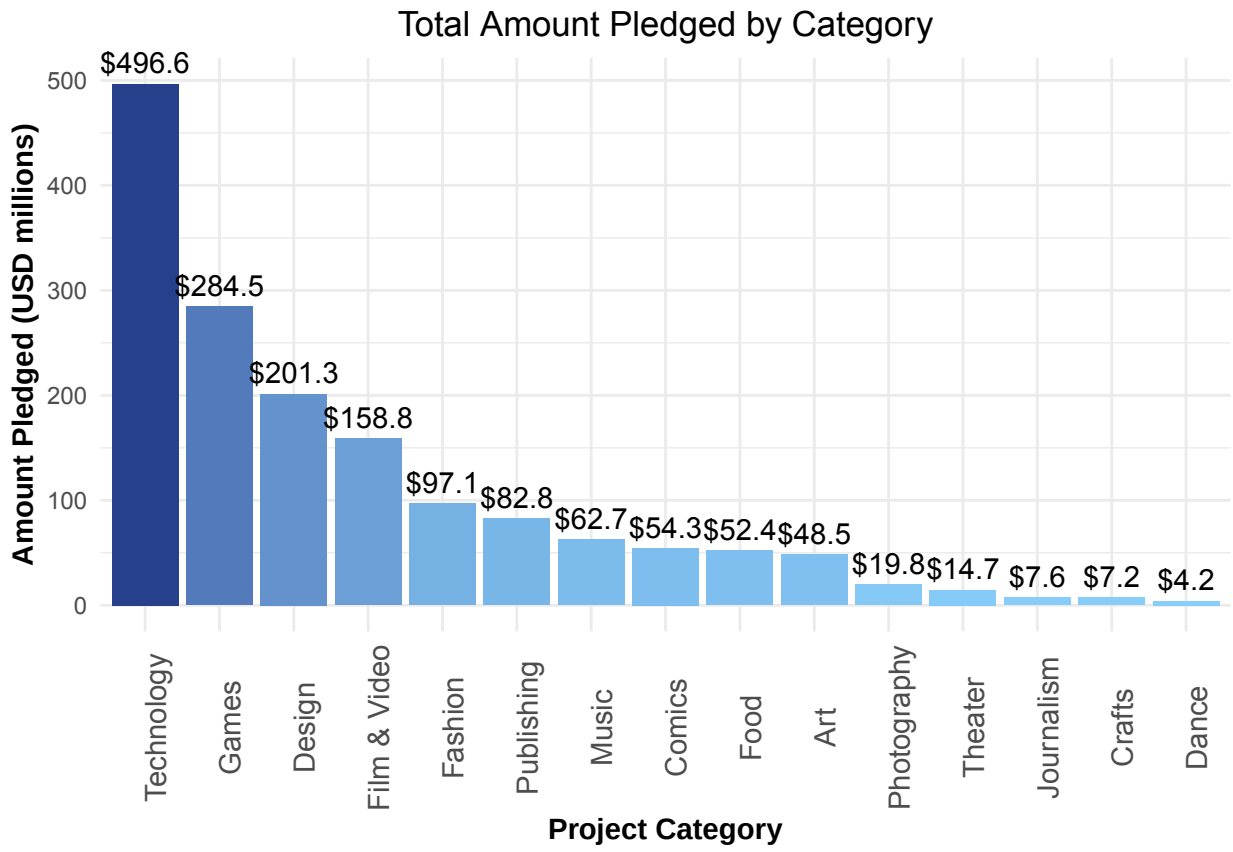
Appendix

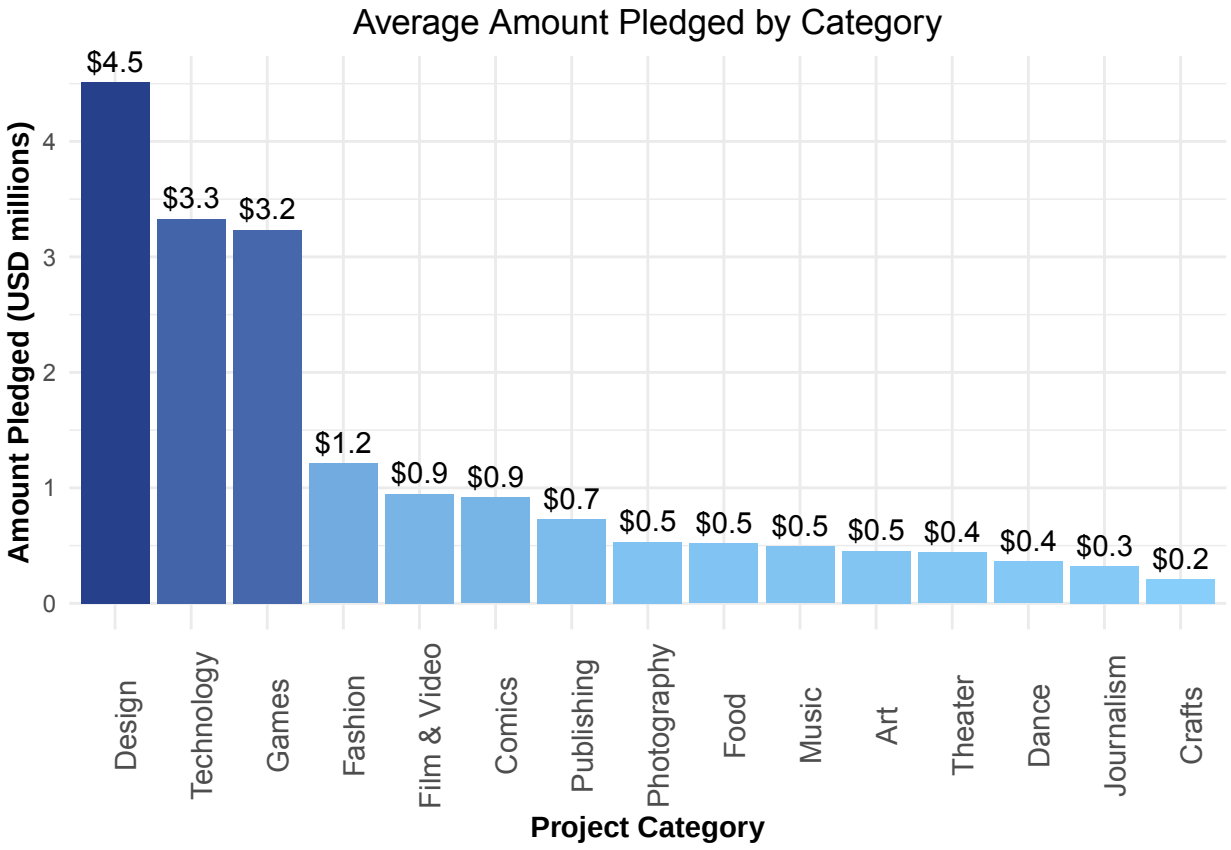




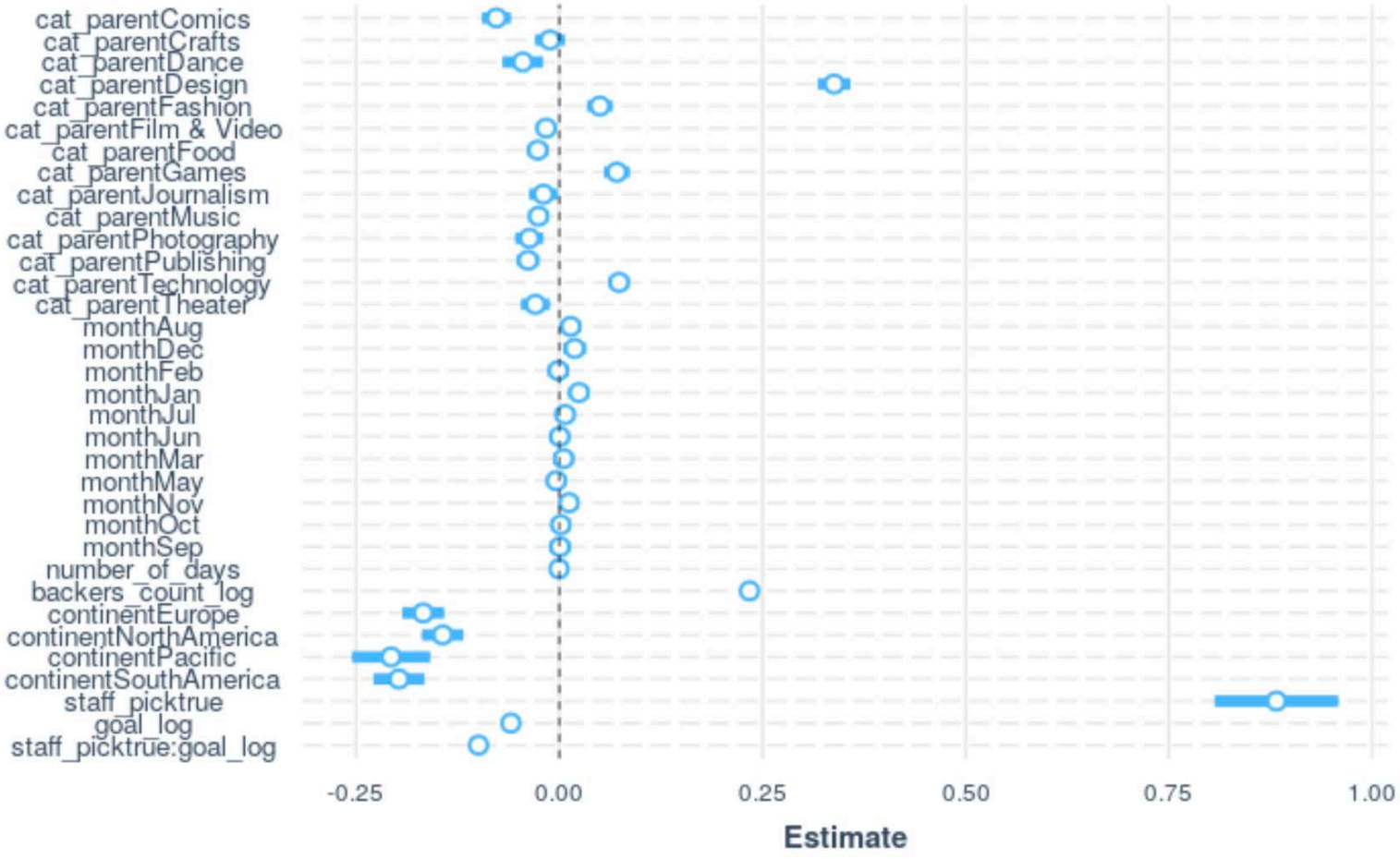

Success vs. Failed Kickstarter Campaigns by Category
Theater
Technology
Publishing
Photography
Music
Journalism
Games
Food
Film & Video
Fashion
Design
Dance
Crafts
Comics
Art
Category
0.00
0.25
0.50
0.75
1.00
Proportion
Outcome:
failed
successful


Projects by Category
16763
14915
12746
11341
10682
10124
8795
8024
5931
4459
3711
3467
3330
2357
1160
15000
10000
5000
0
Frequency
Film & Video
Technology
Music
Publishing
Art
Food
Games
Fashion
Comics
Design
Photography
Crafts
Theater
Journalism
Dance
Project Category

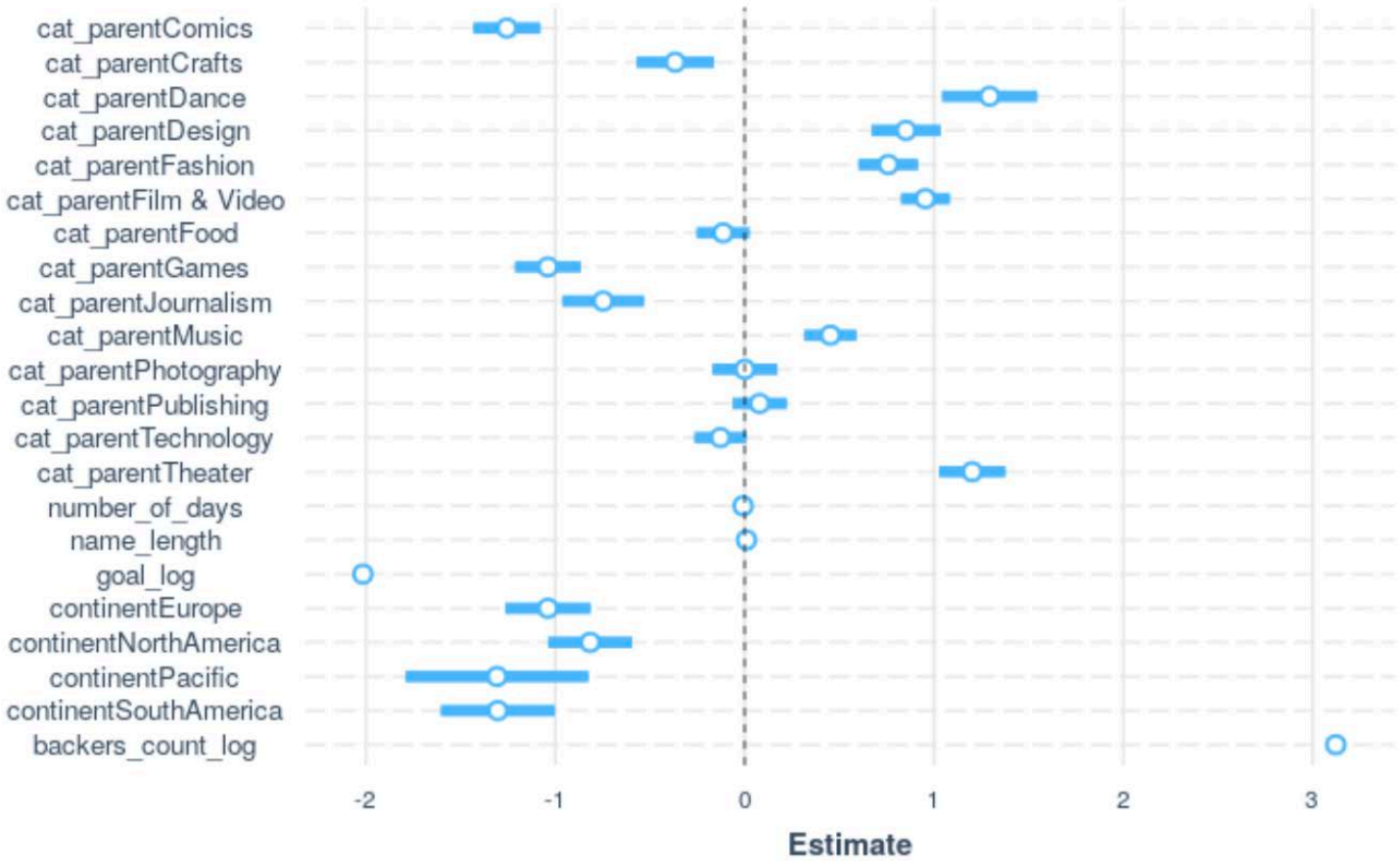

Total Amount Pledged by Category
Amount Pledged (USD millions)
500
400
300
200
100
0
$496.6
$284.5
$201.3
$158.8
$97.1
$82.8
$62.7
$54.3
$52.4
$48.5
$19.8
$14.7
$7.6
$7.2
$4.2
Technology
Games
Design
Film & Video
Fashion
Publishing
Music
Comics
Food
Art
Photography
Theater
Journalism
Crafts
Dance
Project Category


Average Amount Pledged by Category
Amount Pledged (USD millions)
4
3
2
1
0
$4.5
$3.3
$3.2
$1.2
$0.9
$0.9
$0.7
$0.5
$0.5
$0.5
$0.5
$0.4
$0.4
$0.3
$0.2
Design
Technology
Games
Fashion
Film & Video
Comics
Publishing
Photography
Food
Music
Art
Theater
Dance
Journalism
Crafts
Project Category

Goal Ratio (GLM) Model Coefficient Values with 95% Confidence Intervals



Successful / Failure (Binomial) Model Coefficient Values with 95% Confidence Intervals




```
     goal          staff_pick           target             cat_parent
Min.   :   1000   false:97737   failed    :49587   Film & Video:15905
1st Qu.:   3000   true :16695   successful:64845   Technology  :15396
Median :   6700                                    Music       :12232
Mean   :  20928                                    Food        :11036
3rd Qu.:  17000                                    Publishing  :10604
Max.   :1000000                                    Games       : 8088
                                                   (Other)     :41171
    month       number_of_days  blurb_length    name_length        pledged
Oct    :10720   Min.   : 1.0   Min.   : 2.00   Min.   : 2.000   Min.   :       0
May    :10658   1st Qu.:30.0   1st Qu.:16.00   1st Qu.: 4.000   1st Qu.:     141
Apr    :10342   Median :30.0   Median :21.00   Median : 6.000   Median :    2862
Nov    :10102   Mean   :33.5   Mean   :19.32   Mean   : 5.952   Mean   :   18810
Jun    : 9984   3rd Qu.:35.0   3rd Qu.:23.00   3rd Qu.: 8.000   3rd Qu.:   10298
Mar    : 9946   Max.   :61.0   Max.   :36.00   Max.   :20.000   Max.   :12969608
(Other):52680
  goal_ratio              continent     backers_count
Min.   :  0.0000   Asia         : 2028   Min.   :    0.0
1st Qu.:  0.0174   Europe       :29006   1st Qu.:    4.0
Median :  1.0162   NorthAmerica:80795   Median :   39.0
Mean   :  1.4895   Pacific      :  495   Mean   :  197.6
3rd Qu.:  1.2295   SouthAmerica: 2108   3rd Qu.:  125.0
Max.   :540.5412                        Max.   :88887.0
```