KICKSTARTER

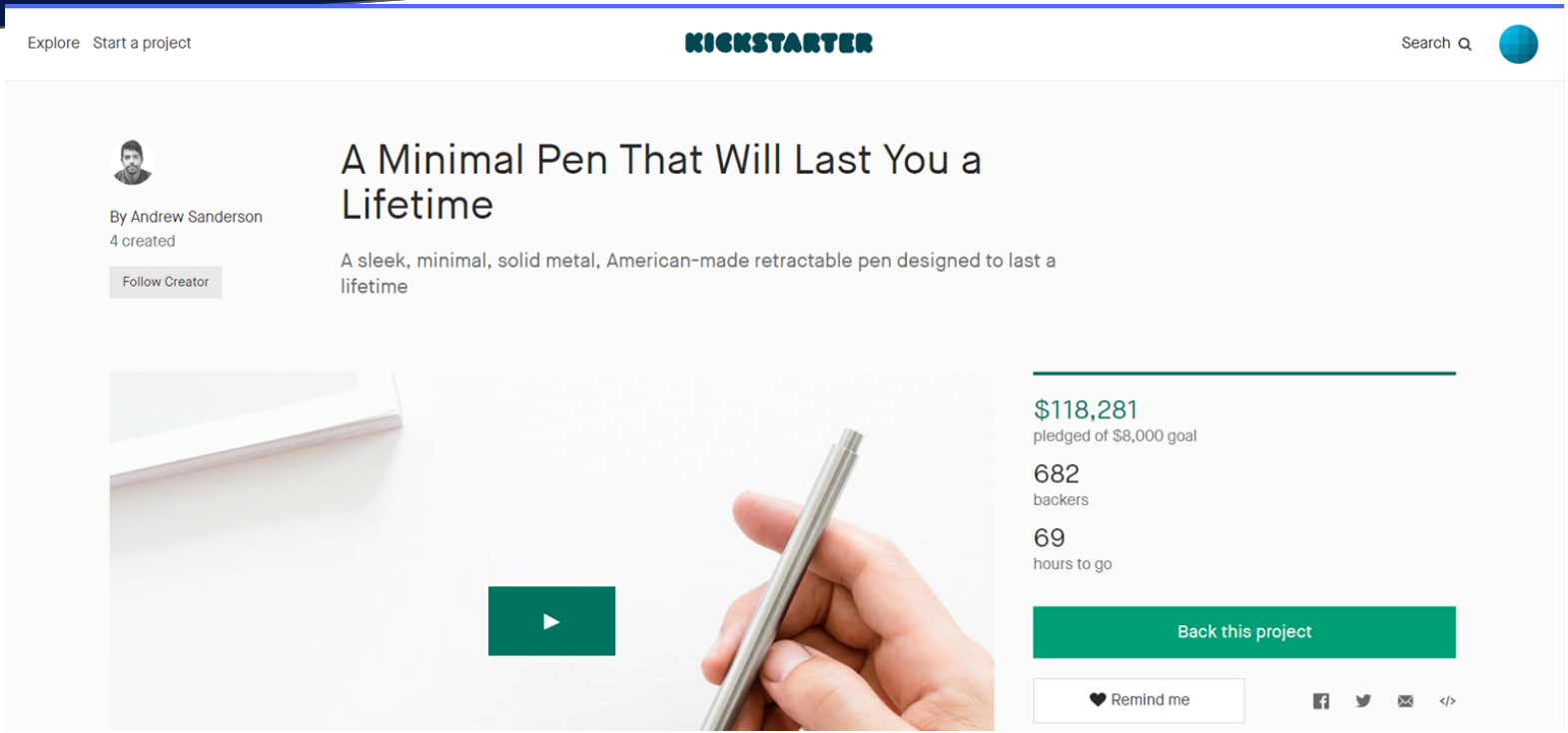
Team Kappa
Jeremy Bourne, Audrey Low and Aaron Liu

Machine Learning I

Agenda

- Business problem
- Data Cleaning
- Data Description
- Models - Classification
- Models - Regression
- Conclusions & Recommendations

The Business Problem



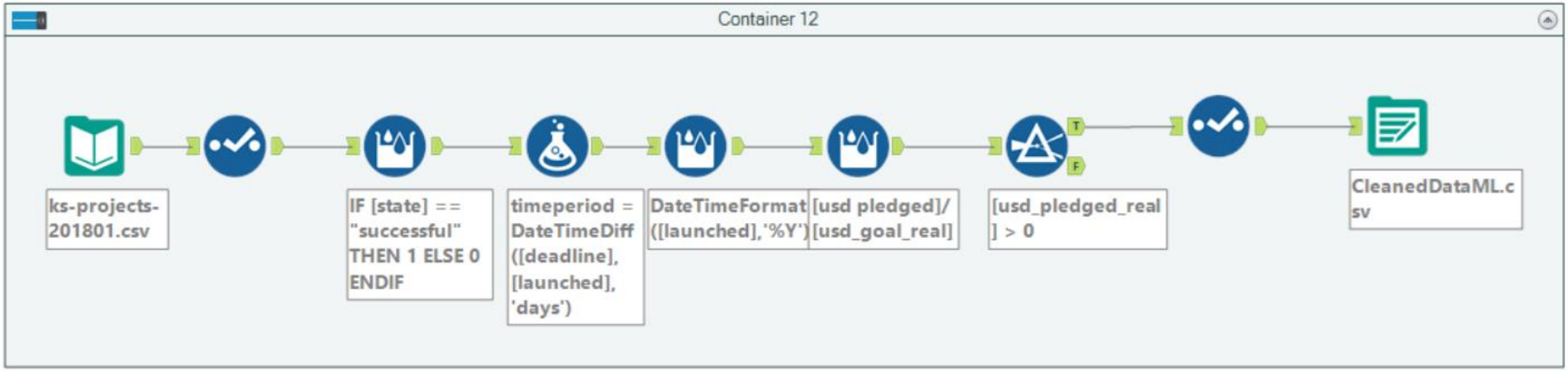
The Data

	# ID	A name	A category	A main_c...	A currency	deadline	# goal	launched	# pledged	A state	# backers	A country	# usd ple...
1	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09 11:36:00	1000	2015-08-11 12:12:28	0	failed	0	GB	0
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26 00:20:50	45000	2013-01-12 00:20:50	220	failed	3	US	220
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16 04:24:11	5000	2012-03-17 03:24:11	1	failed	1	US	1
4	1000011046	Community Film Project: The Art of Neighborhood Filmmaking	Film & Video	Film & Video	USD	2015-08-29 01:00:00	19500	2015-07-04 08:35:03	1283	canceled	14	US	1283

Before cleaning data

```
'data.frame':   378661 obs. of  15 variables:
 $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000011046
00023410 1000030581 1000034518 100004195 ...
 $ name            : Factor w/ 375765 levels "","    IT'S A HOT CAPPUCCINO NIGHT
135633 364946 344770 77274 206067 293430 69281 284103 290686 ...
 $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124
..
 $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 7 8 8 8 5
 $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14
.
 $ deadline        : Factor w/ 3164 levels "2009-05-03","2009-05-16",..: 2288 30
247 2463 1996 2448 1790 1863 ...
 $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125000 65
 $ launched        : Factor w/ 378089 levels "1970-01-01 01:00:00",..: 243292 36
57 235943 278600 187500 274014 139367 153766 ...
 $ pledged         : num  0 2421 220 1 1283 ...
 $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2 2 2 1 4 4 2
 $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
 $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23
 $ usd.pledged     : num  0 100 220 1 1283 ...
 $ usd_pledged_real: num  0 2421 220 1 1283 ...
 $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
```

Steps of Cleaning the Data in Alteryx



Steps of Cleaning the Data in R

- Change the categorical variables into numeric type for running classifiers
- Remove the outliers - top 1% and bottom 1%
- Convert campaign status from levels into binary
- Delete values that are null

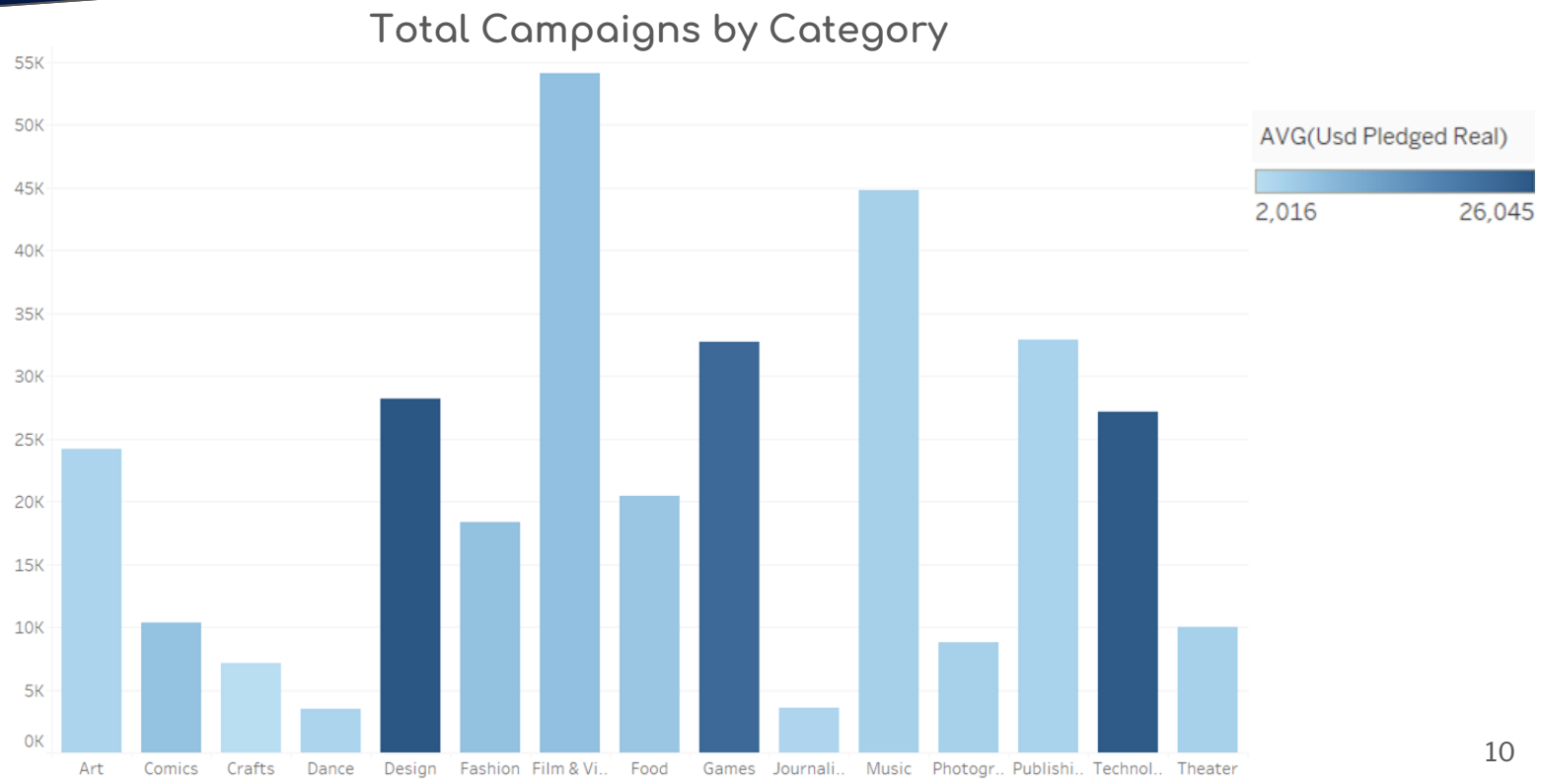
After Cleaning Data

```
> str(ks)
'data.frame':   322165 obs. of  18 variables:
 $ ID              : num  1e+09 1e+09 1e+09 1e+09 1e+09 ...
 $ name            : chr  "Greeting From Earth: ZGAC Arts Capsule For ET" "Where
nk?" "ToshiCapital Rekordz Needs Help to Complete Album" "Community Film Project:
Art of Neighborhood Filmmaking" ...
 $ category        : chr  "Narrative Film" "Narrative Film" "Music" "Film & Videc
 $ main_category   : num  13 13 19 13 14 14 14 10 13 19 ...
 $ currency        : chr  "USD" "USD" "USD" "USD" ...
 $ deadline        : chr  "2017/11/1" "2013/2/26" "2012/4/16" "2015/8/29" ...
 $ goal            : chr  "30000" "45000" "5000" "19500" ...
 $ launched        : chr  "2017/9/2" "2013/1/12" "2012/3/17" "2015/7/4" ...
 $ pledged         : chr  "2421" "220" "1" "1283" ...
 $ state           : chr  "failed" "failed" "failed" "canceled" ...
 $ backers         : num  15 3 1 14 224 16 40 58 43 100 ...
 $ country         : num  351 351 351 351 351 351 351 351 351 351 ...
 $ usd.pledged     : chr  "100" "220" "1" "1283" ...
 $ usd_pledged_real: num  2421 220 1 1283 52375 ...
 $ usd_goal_real   : num  30000 45000 5000 19500 50000 1000 25000 125000 65000 12
..
 $ timeperiod      : num  60 45 30 56 35 20 45 35 30 30 ...
 $ year            : num  2017 2013 2012 2015 2016 ...
 $ status_num1     : num  0 0 0 0 1 1 0 0 0 1 ...
```

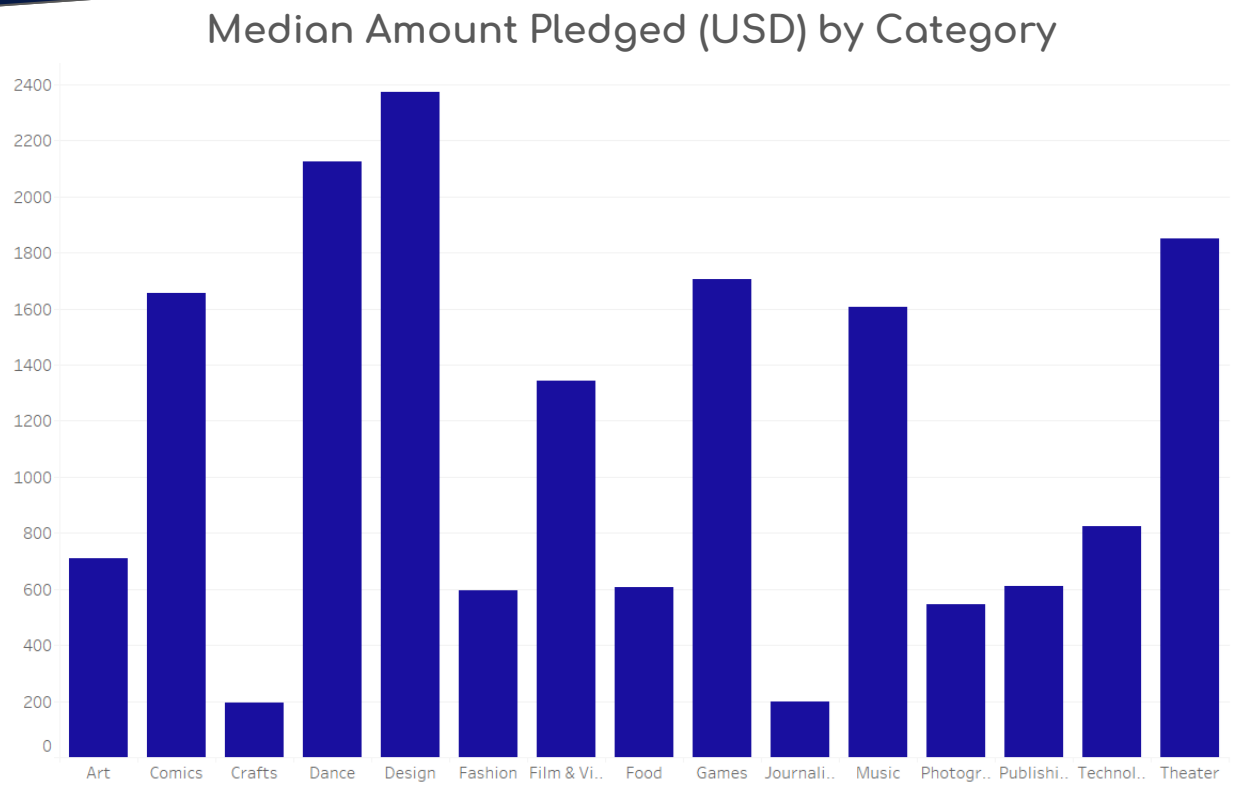
After Cleaning Data - Updated

```
> str(ks)
data.frame':   321775 obs. of  18 variables:
$ ID              : int  1000003930 1000004038 1000007540 1000011046 1000014025
1000034518 100004195 100005484 ...
$ name            : chr  "Greeting From Earth: ZGAC Arts Capsule For ET" "Where
al Rekordz Needs Help to Complete Album" "Community Film Project: The Art of Nei
..
$ category        : chr  "Narrative Film" "Narrative Film" "Music" "Film & Vide
$ main_category   : Factor w/ 15 levels "Art","Comics",..: 7 7 11 7 8 8 8 5 7 1
$ currency        : chr  "USD" "USD" "USD" "USD" ...
$ deadline        : chr  "2017/11/1" "2013/2/26" "2012/4/16" "2015/8/29" ...
$ goal            : chr  "30000" "45000" "5000" "19500" ...
$ launched        : chr  "2017/9/2" "2013/1/12" "2012/3/17" "2015/7/4" ...
$ pledged         : chr  "2421" "220" "1" "1283" ...
$ state           : chr  "failed" "failed" "failed" "canceled" ...
$ backers         : num  15 3 1 14 224 16 40 58 43 100 ...
$ country         : Factor w/ 23 levels "AT","AU","BE",..: 23 23 23 23 23 23 23
$ usd.pledged     : chr  "100" "220" "1" "1283" ...
$ usd_pledged_real: num  2421 220 1 1283 52375 ...
$ usd_goal_real   : num  30000 45000 5000 19500 50000 1000 25000 125000 65000 1
$ timeperiod      : int  60 45 30 56 35 20 45 35 30 30 ...
$ year            : int  2017 2013 2012 2015 2016 2014 2016 2014 2014 2013 ...
$ status_num1     : num  0 0 0 0 1 1 0 0 0 1 ...
```

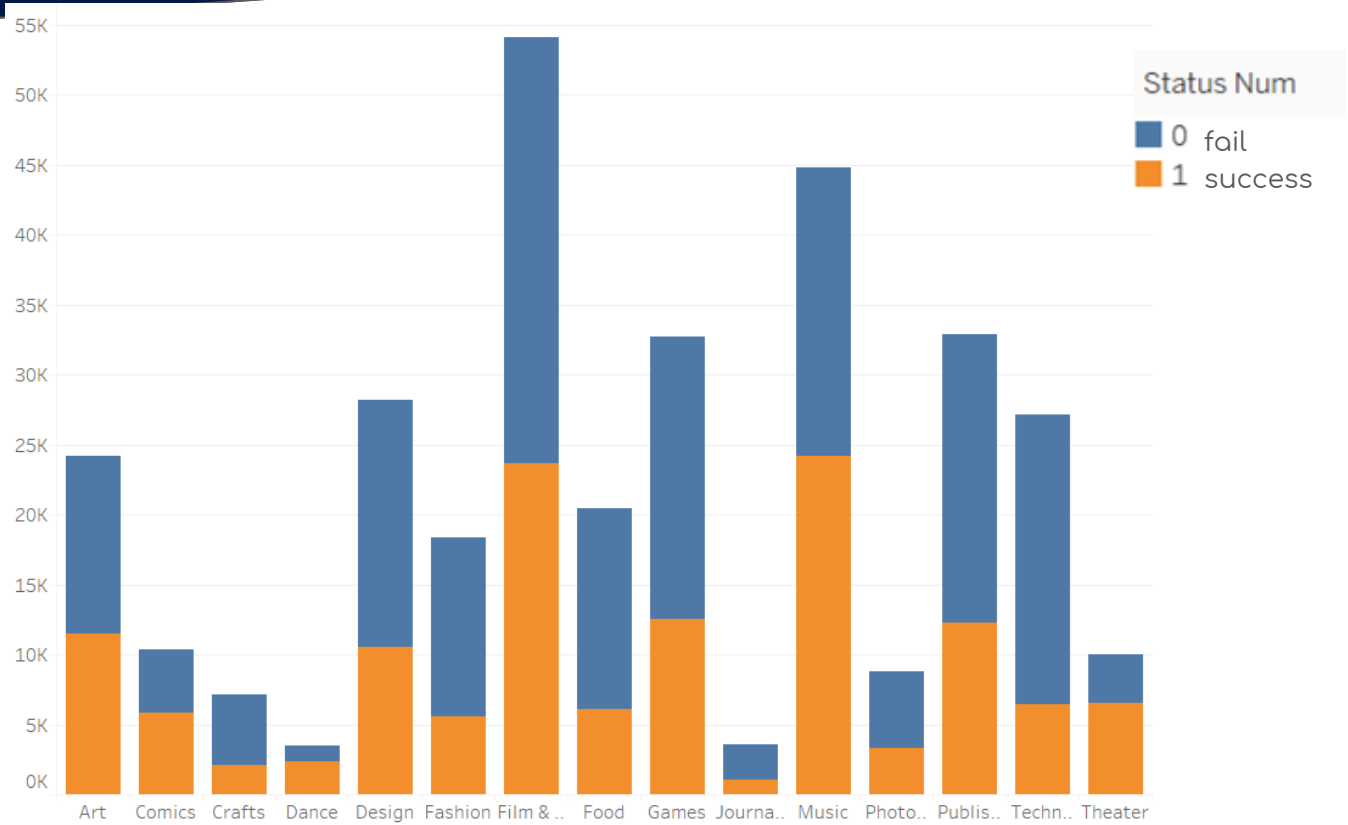
Data Exploration



Median Amount Pledged by Category



Success by Category



All Possible Predictors

- Main_category
- Country
- Backers
- Usd_goal_real
- Timeperiod
- Year

Classification

Selecting Predictors for Classification

```
> glmulti.summaryglm$bestmodel
[1] "ks.sub1$status_num1 ~ 1 + main_category + backers + country + "
[2] "    usd_goal_real + timeperiod"
```

Best model is

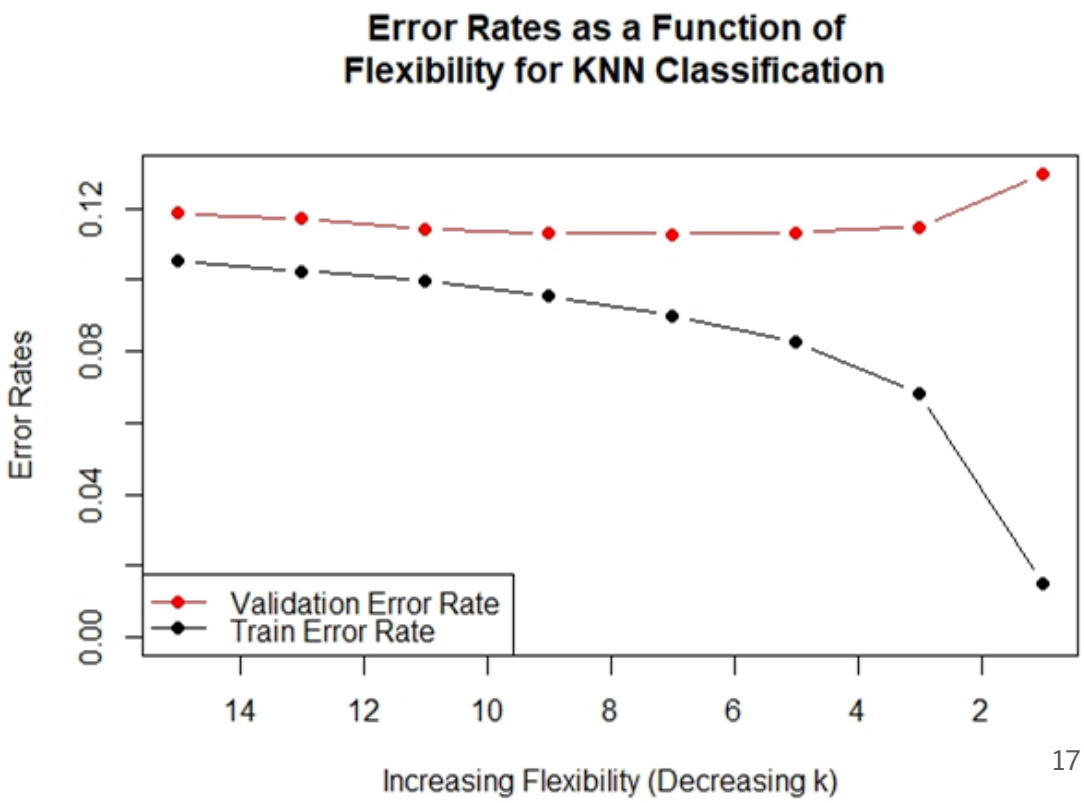
- Main_category
- Country
- Backers
- Usd_goal_real
- Timeperiod

Classifiers

- KNN
- Logistic Regression
- LDA
- QDA
- Random Forest

KNN Classifier

Choose k = 5 to make the prediction & avoid overtraining in order to balance the tradeoff between bias & variance



Accuracy of Classifiers - Comparison

- Main_category
- usd_goal_real
- country
- timeperiod
- backers

	Accuracy	Power	Precision
Random Forest	0.9155557	0.9082003	0.8849551
KNN	0.8829226	0.8652838	0.8471237
Logistic	0.8830568	0.7827278	0.914305
LDA	0.6271631	0.115174	0.7338412
QDA	0.4993249	0.950783	0.4430744

Accuracy of Classifiers - Comparison (Updated)

- Main_category
- usd_goal_real
- country
- timeperiod
- backers

	Accuracy	Power	Precision
Logistic	0.8895191	0.8182515	0.9002698
LDA	0.6588144	0.3655659	0.6387308
QDA	0.5123767	0.9169029	0.450334
Random Forest	0.9149561	0.909202	0.8848985

Accuracy of Classifiers - Question

Why are the accuracies of LDA and QDA much lower than others ?

- Low-dimension datasets v.s. high-dimension datasets
- The observations of each class (predictor) should be normally distributed.

Accuracy of Classifiers

- backers
- main_category

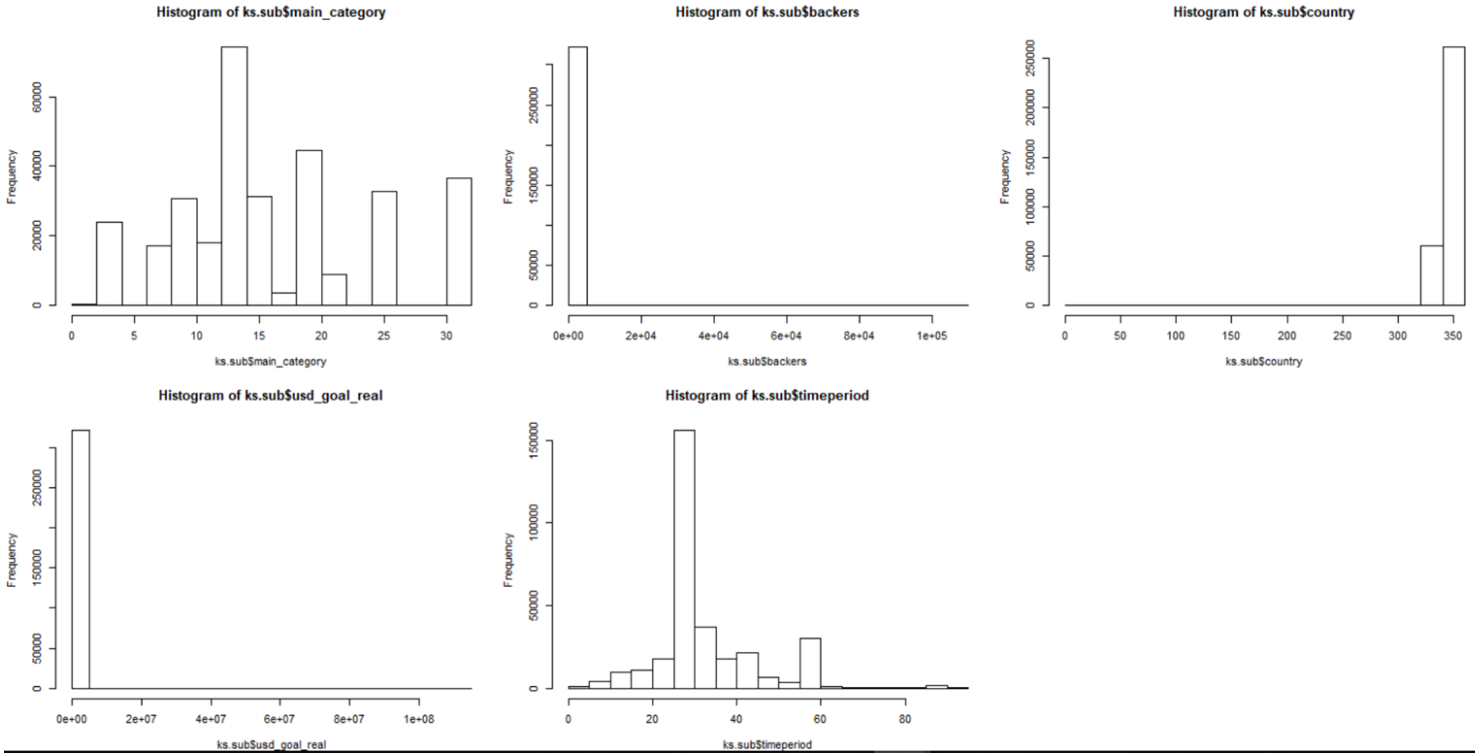
New:

	Accuracy	Power	Precision
LDA	0.6143125	0.0434313	0.9566695
QDA	0.6545714	0.1578724	0.9061324

Old:

LDA	0.6271631	0.115174	0.7338412
QDA	0.4993249	0.950783	0.4430744

Histogram of the Predictors



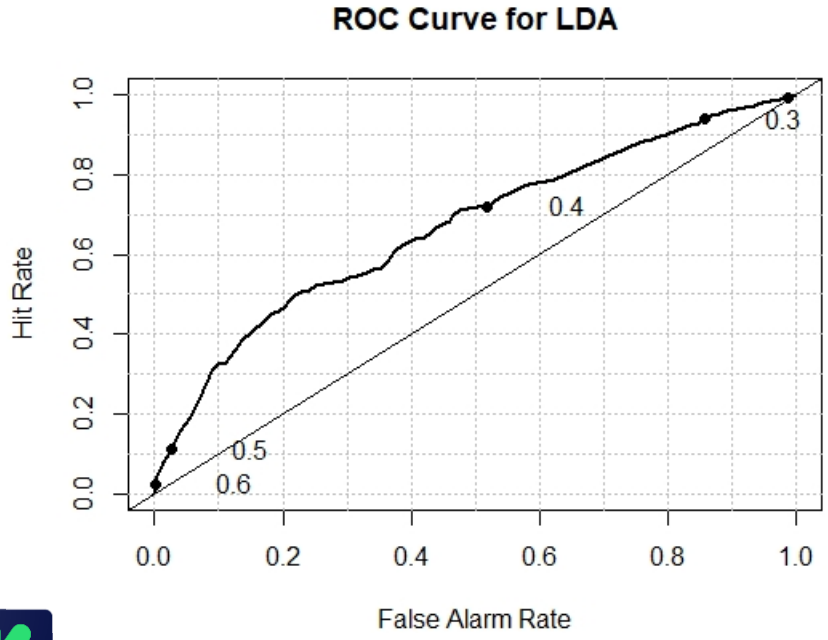
Error Rates of Classifiers - Comparison

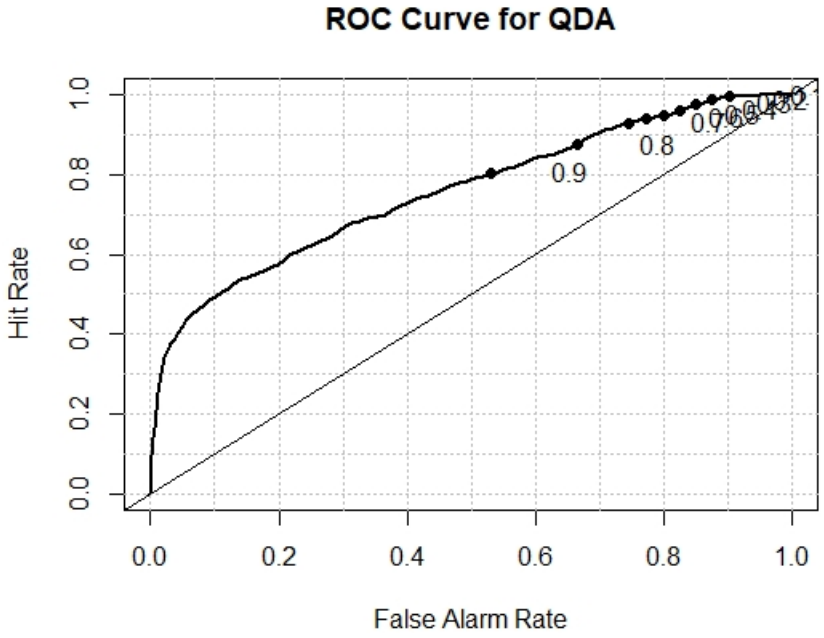
	Total.Error	Type1.Error	Type2.Error
Logistic	0.1169432	0.04939362	0.2172722
LDA	0.3728369	0.02812476	0.884826
QDA	0.5006751	0.8046329	0.049217
Random Forest	0.08444431	0.07949204	0.09179974
KNN	0.1170774	0.1051948	0.1347162

Error Rates of Classifiers - Updated

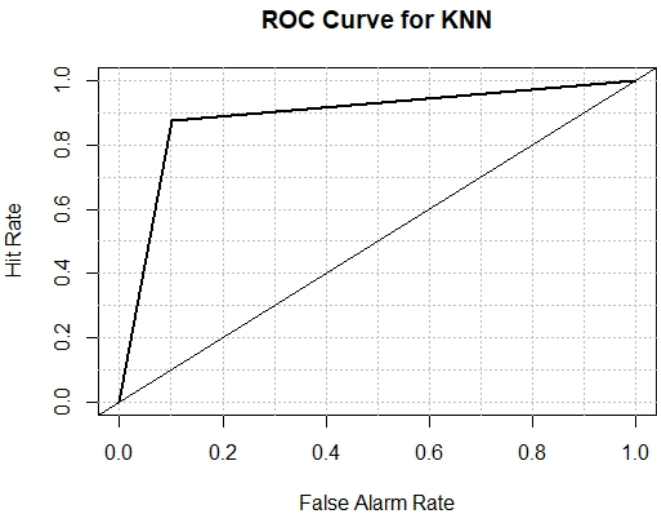
	Total.Error	Type1.Error	Type2.Error
Logistic	0.1104809	0.06185136	0.1817485
LDA	0.3411856	0.141087	0.6344341
QDA	0.4876233	0.7636525	0.08309708
Random Forest	0.0850439	0.08109807	0.09079797

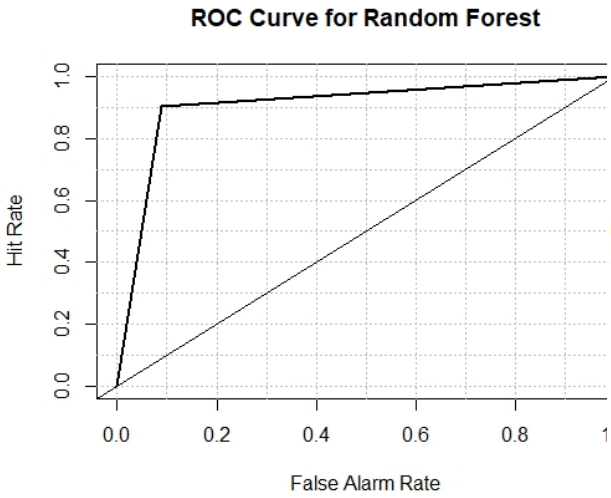
Compare ROC Curve between LDA and QDA Classifier

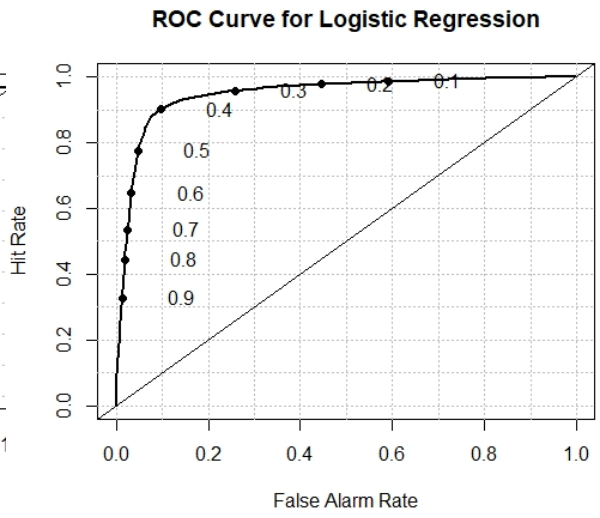




Compare ROC Curve Between KNN, Random Forest and Logistic Regression





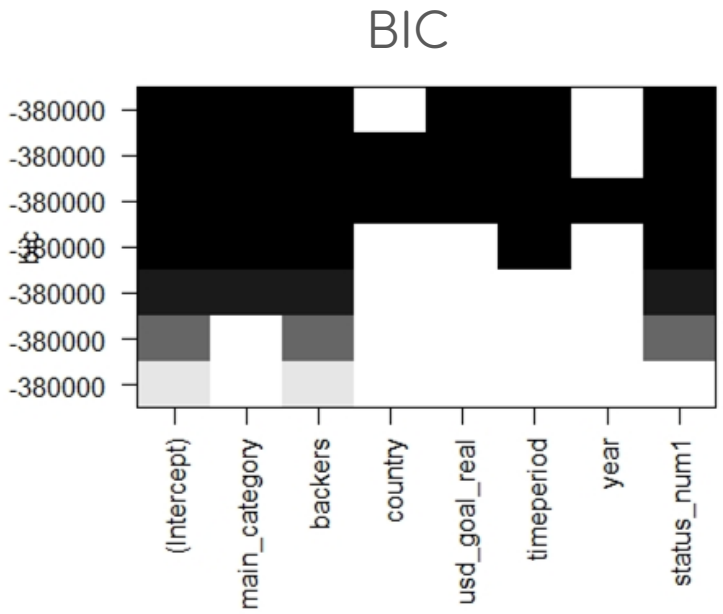


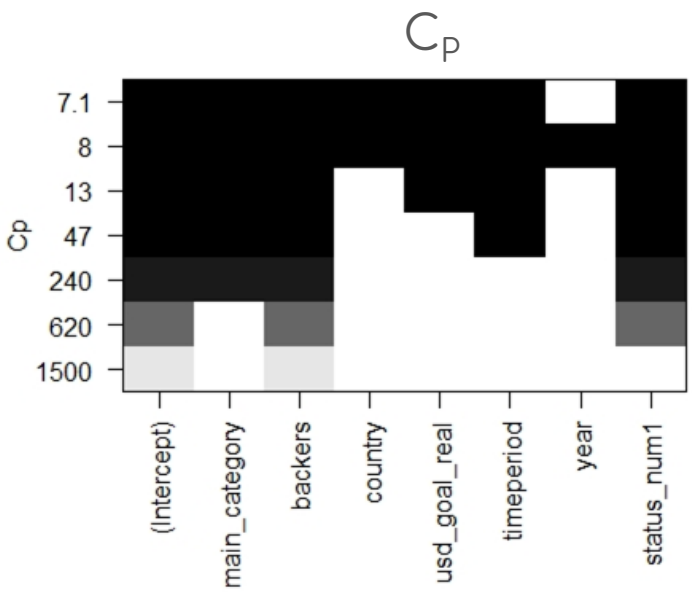
Predictions for Classification

- Based on comparison of accuracy and ROC curve, we choose Random Forest classifier to predict the future campaign.
- We choose five predictors, which are
  - Main_category
  - Country
  - Backers
  - Usd_goal_real
  - Timeperiod

Regression

Selecting Predictors for Regression by BIC and CP





Selecting Predictors for Regression

- Main_category
- Backers
- Usd_goal_real
- Timeperiod

Multiple Regression

- Created 15 models for each main category with only the successful cases
- Predicted the amount pledged in USD for each main category

Film & Video

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.803e+03  3.086e+02  -5.844 5.17e-09 ***
backers         4.434e+01  2.353e-01 188.444  < 2e-16 ***
usd_goal_real   8.252e-01  6.057e-03 136.231  < 2e-16 ***
timeperiod     -6.614e+00  8.879e+00  -0.745    0.456
```

Fashion - the exception

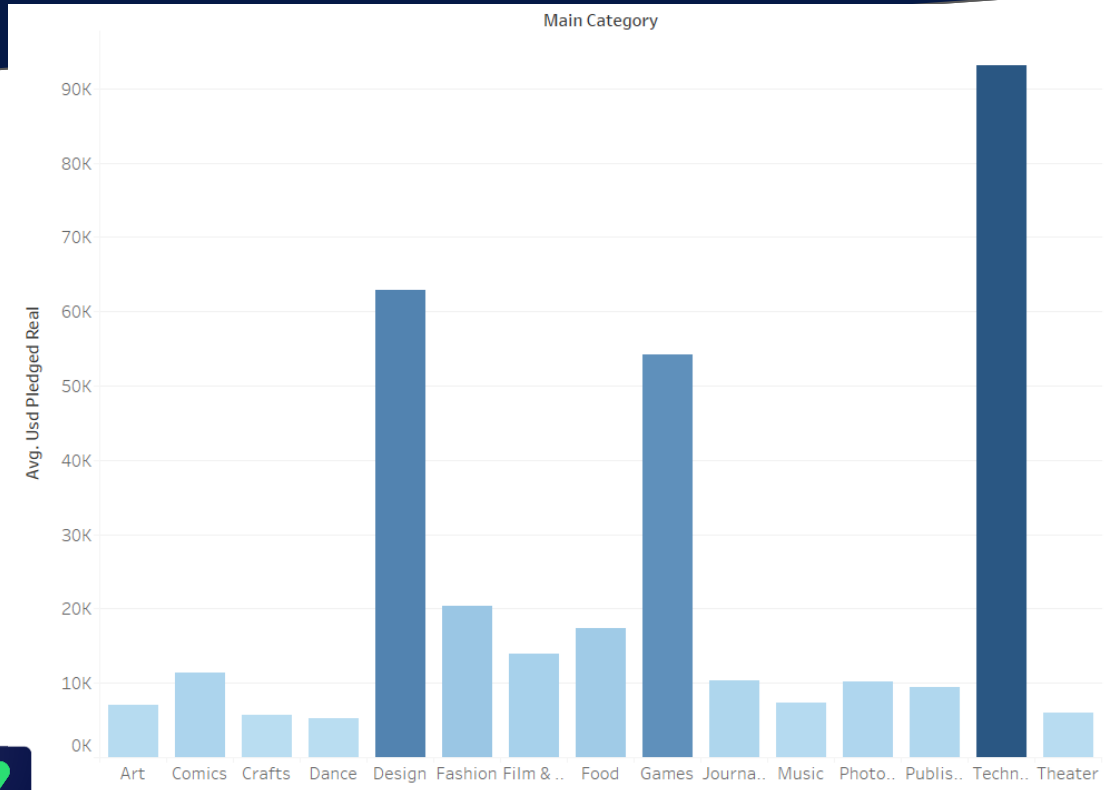
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.237e+06  2.545e+05  -4.862 1.20e-06 ***
backers         4.065e+01  6.717e-01  60.510  < 2e-16 ***
usd_goal_real   1.010e+00  1.985e-02  50.890  < 2e-16 ***
timeperiod      9.922e+01  2.372e+01   4.183 2.92e-05 ***
```

- Timeperiod = Duration of campaign

Regression with Adjusted R square

Main Category	Adjusted R square	p_value > 0.05
Dance	0.97	time
Journalism	0.95	time
Publishing	0.86	time
Film & Video	0.93	time
Music	0.84	time
Crafts	0.86	time
Games	0.81	time
Comics	0.85	time
Theater	0.92	time
Design	0.68	time
Art	0.69	time
Photography	0.67	time
Technology	0.7	time
Fashion	0.7	

Predictions for Regression



	pred.table
Crafts	5083.386
Dance	5198.276
Theater	5983.587
Art	6194.568
music	7251.713
Publishing	8733.607
Journalism	10387.726
comics	10475.914
film	13974.106
Food	15502.178
Fashion	15935.095
Design	36374.576
Games	36902.322
Technology	63685.800

Conclusions & Recommendations

- **As Kickstarter:**
  - Collaborate with initiatives to optimize chances of success
- **As a potential Kickstarter campaign:**
  - Chances of success
  - Expected amount pledged

Thank you

Technical Slides

Problems encountered

1. Why the accuracy is so different between logistic, random forest, KNN and lda, qda?
2. How do these models learn the data?
3. When running data in logistic regression, we found that the confusion matrix would be like this sometimes if we only use one predictor to predict our independent variable - status.