It’s All Funds & Games

Milestone Presentation

Brian Tunnell, Mark Giannini, Patrick Tinsley

Motivation (Refresher)

- We want to be able to predict the ultimate fate of a Kickstarter campaign
  - final_status column → failure = 0, success = 1
- Data comes from Kaggle (108,129 projects)
- Train-test Split: 66.6% - 33.3% (via train_test_split function from sklearn.model_selection)
  - Training Data: 72,446 projects
  - Testing Data: 35,683 projects

Feature Engineering / Data Integration

- Here are some functions that we found very useful:
  - to_datetime (pandas library): converts unix time (1/1/1970) to datetime object
    - created_at, launched_at, state_changed_at, deadline, ...
  - get_dummies (pandas library): dummy codes categorical data
    - country, launch_hour, launch_day, launch_month, deadline_day, …
- Expected Duration = deadline - launched_at
- Actual Duration = state_changed_at - launched_at

Data Transformation Chart (Original)

```
Columns:
['name' 'desc' 'goal' 'keywords' 'disable_communication' 'country'
 'currency' 'deadline' 'state_changed_at' 'created_at' 'launched_at'
 'backers_count' 'final_status']
Data Types:
name                     object
desc                     object
goal                    float64
keywords                 object
disable_communication      bool
country                  object
currency                 object
deadline                  int64
state_changed_at          int64
created_at                int64
launched_at               int64
backers_count             int64
final_status              int64
dtype: object
Shape:
(108129, 13)
```

Data Transformation Chart (Intermediate)

```
Columns:
['name' 'desc' 'goal' 'keywords' 'disable_communication' 'country'
 'backers_count' 'final_status' 'expected_duration' 'actual_duration'
 'launch_year' 'launch_month' 'launch_day' 'launch_hour' 'deadline_year'
 'deadline_month' 'deadline_day']
Data Types:
name                      object
desc                      object
goal                     float64
keywords                  object
disable_communication       bool
country                   object
backers_count              int64
final_status               int64
expected_duration          int64
actual_duration            int64
launch_year                int64
launch_month               int64
launch_day                 int64
launch_hour                int64
deadline_year              int64
deadline_month             int64
deadline_day               int64
dtype: object
Shape:
(108129, 17)
```

Data Transformation Chart (‘Final’)

```
Columns:
['name' 'desc' 'goal' 'keywords' 'disable_communication' 'backers_count'
 'final_status' 'expected_duration' 'actual_duration' 'country_AU'
 'country_CA' 'country_DE' 'country_DK' 'country_GB' 'country_IE'
 'country_NL' 'country_NO' 'country_NZ' 'country_SE' 'country_US'
 'launch_hour_0' 'launch_hour_1' 'launch_hour_2' 'launch_hour_3'
 'launch_hour_4' 'launch_hour_5' 'launch_hour_6' 'launch_hour_7'
 'launch_hour_8' 'launch_hour_9' 'launch_hour_10' 'launch_hour_11'
 'launch_hour_12' 'launch_hour_13' 'launch_hour_14' 'launch_hour_15'
 'launch_hour_16' 'launch_hour_17' 'launch_hour_18' 'launch_hour_19'
 'launch_hour_20' 'launch_hour_21' 'launch_hour_22' 'launch_hour_23'
 'launch_day_1' 'launch_day_2' 'launch_day_3' 'launch_day_4' 'launch_day_5'
 'launch_day_6' 'launch_day_7' 'launch_day_8' 'launch_day_9'
 'launch_day_10' 'launch_day_11' 'launch_day_12' 'launch_day_13'
 'launch_day_14' 'launch_day_15' 'launch_day_16' 'launch_day_17'
 'launch_day_18' 'launch_day_19' 'launch_day_20' 'launch_day_21'
 'launch_day_22' 'launch_day_23' 'launch_day_24' 'launch_day_25'
 'launch_day_26' 'launch_day_27' 'launch_day_28' 'launch_day_29'
 'launch_day_30' 'launch_day_31' 'launch_month_1' 'launch_month_2'
```

Classifier Evaluation

Model	Accuracy	Precision	Recall	F-1 Score	AUC
DT (Gini)	0.75985	0.49497	0.67057	0.56954	0.72907
LogReg	0.80063	0.44643	0.86852	0.58973	0.82787
RF (Gini)	0.85387	0.79193	0.76212	0.77674	0.83095
RF (Entropy)	0.85135	0.78905	0.75781	0.77312	0.82806
NB (Bernoulli)	0.62870	0.46651	0.42805	0.44645	0.58235
NB (Gaussian)	0.49995	0.87322	0.37894	0.52852	0.61133

Progress

- Right now, we have classifiers built for the data **without** the columns that include text
- Ultimately, we want to cluster the projects into categories based on the text columns (name, description, keywords)
- We can then use the cluster/category as another feature
    - We expect to have to use stop-words, stemming, tokenization, and tf-idf
    - Tf-idf = “term frequency - inverse document frequency” : measures how significant a word is document or collection of documents