CROWDFUNDING: PREDICTING KICKSTARTER PROJECT SUCCESS

Team Members:

Ayush Sharma 20UCC123
Sajal Sharma 20UCS168

PROBLEM STATEMENT

Predict whether the kickstarter project will succeed or fail in achieving the fundraising goal using information provided by the project launch.

DATA EXPLORATION

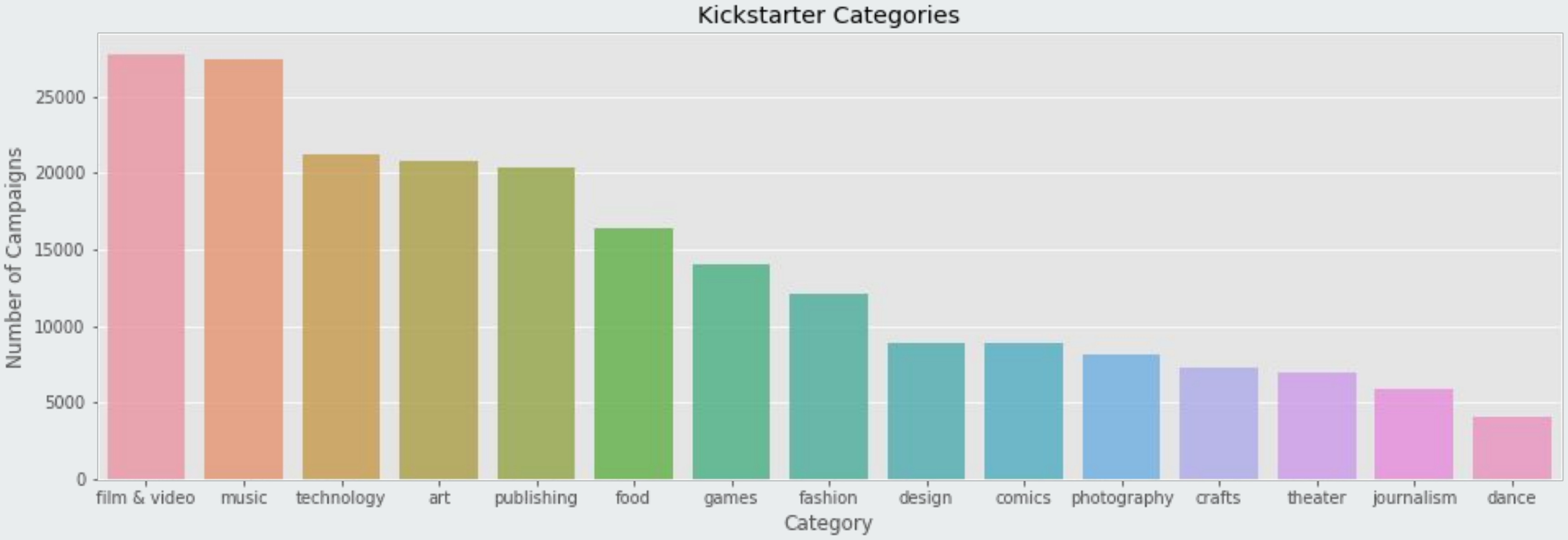
```
kickstarter.columns
```

```
Index(['Unnamed: 0', 'backers_count', 'blurb', 'category',
       'converted_pledged_amount', 'country', 'created_at', 'creator',
       'currency', 'currency_symbol', 'currency_trailing_code',
       'current_currency', 'deadline', 'disable_communication', 'friends',
       'fx_rate', 'goal', 'id', 'is_backing', 'is_starrable', 'is_starred',
       'launched_at', 'location', 'name', 'permissions', 'photo', 'pledged',
       'profile', 'slug', 'source_url', 'spotlight', 'staff_pick', 'state',
       'state_changed_at', 'static_usd_rate', 'urls', 'usd_pledged',
       'usd_type'],
      dtype='object')
```

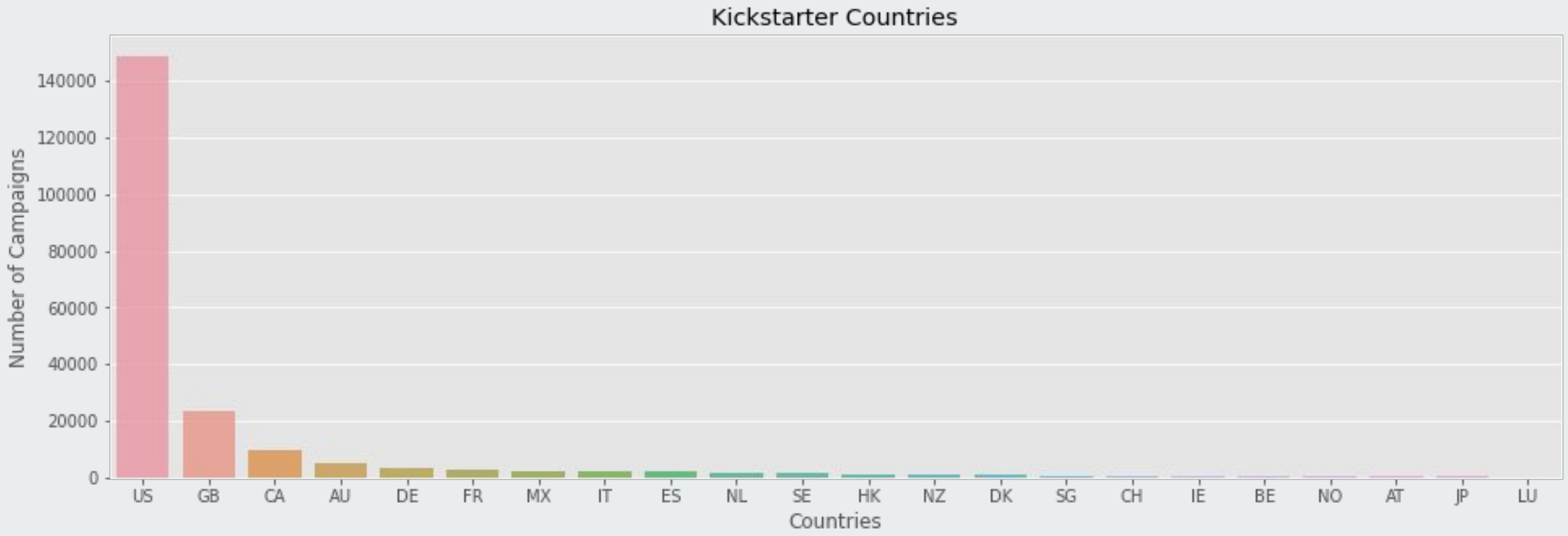
DATA PREPROCESSING STEPS

- Dropping unnecessary columns
- Expanding “state” column
- Categorical Encoding
- Reducing number of unique values for the “country” attribute
- Convert unix timestamp to datetime format for “created_at”, “deadline”, etc

KICKSTARTER CATEGORIES



KICKSTARTER COUNTRIES



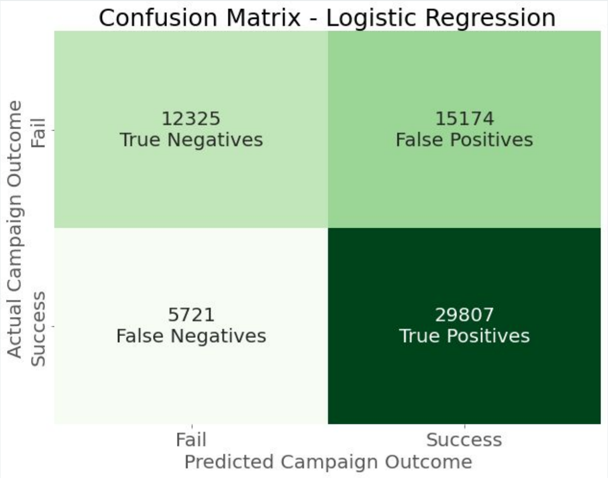
UNIX TO DATETIME

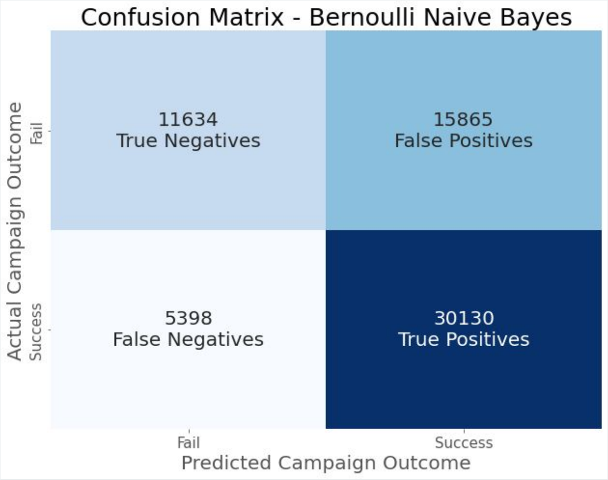
	created_at	deadline	launched_at	state_changed_at
0	2016-06-20 00:45:43	2016-07-12 00:00:00	2016-06-27 02:22:22	2016-07-12 00:00:01
1	2015-07-10 22:38:57	2015-09-12 23:15:03	2015-08-13 23:15:03	2015-09-12 23:15:15
2	2014-11-17 17:47:16	2015-03-16 15:44:46	2015-02-14 16:44:46	2015-03-16 15:45:13

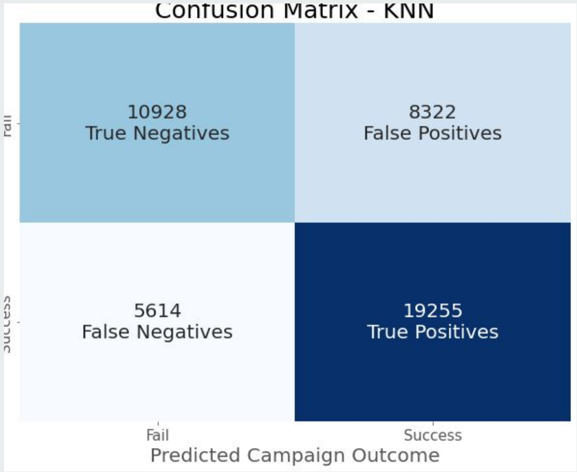
ATTRIBUTES AFTER PREPROCESSING

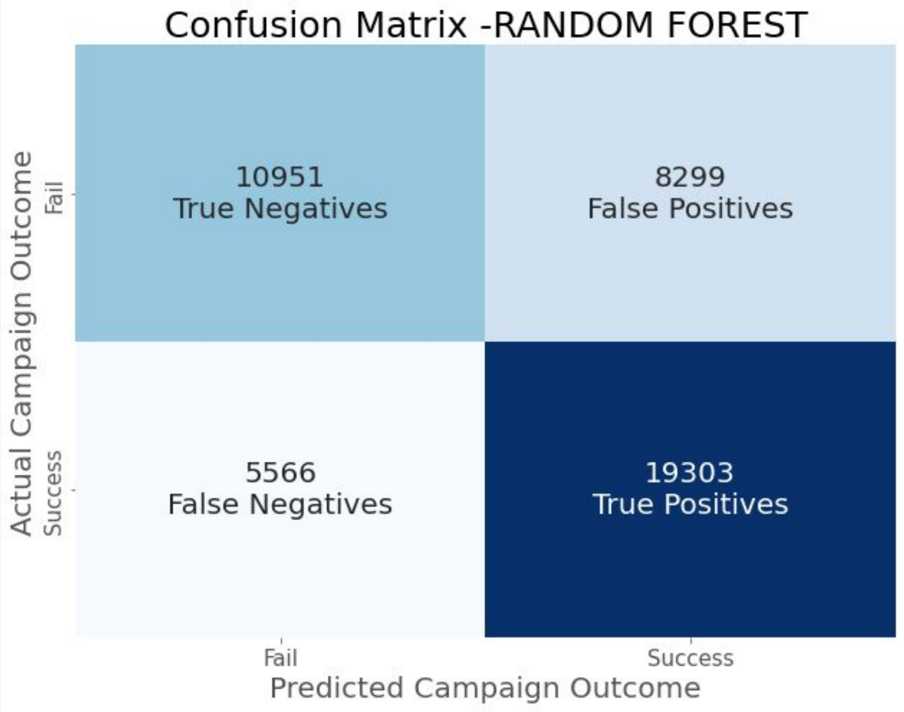
```
'backers_count',
'blurb',
'converted_pledged_amount',
'created_at',
'deadline',
'goal',
'id',
'launched_at',
'location',
'slug',
'state_changed_at',
'state_failed',
'state_live',
'state_successful',
'state_suspended',
'child_category',
'main_category_comics',
'main_category_crafts',
'main_category_dance',
'main_category_design',
'main_category_fashion',
'main_category_film & video',
'main_category_food',
'main_category_games',
'main_category_journalism',
'main_category_music',
'main_category_photography',
'main_category_publishing',
'main_category_technology',
'main_category_theater',
'country_CA',
'country_DE',
'country_ES',
'country_FR',
'country_GB',
'country_IT',
'country_MX',
'country_NL',
'country_OTHER',
'country_US',
'staff_pick_True',
'spotlight_True']
```

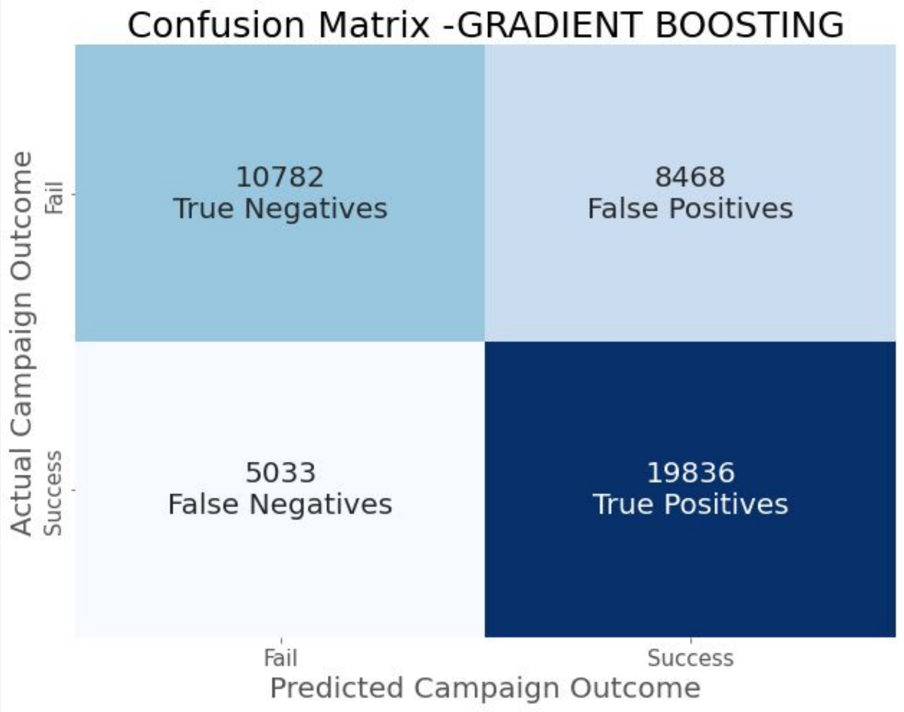
CONFUSION MATRICES







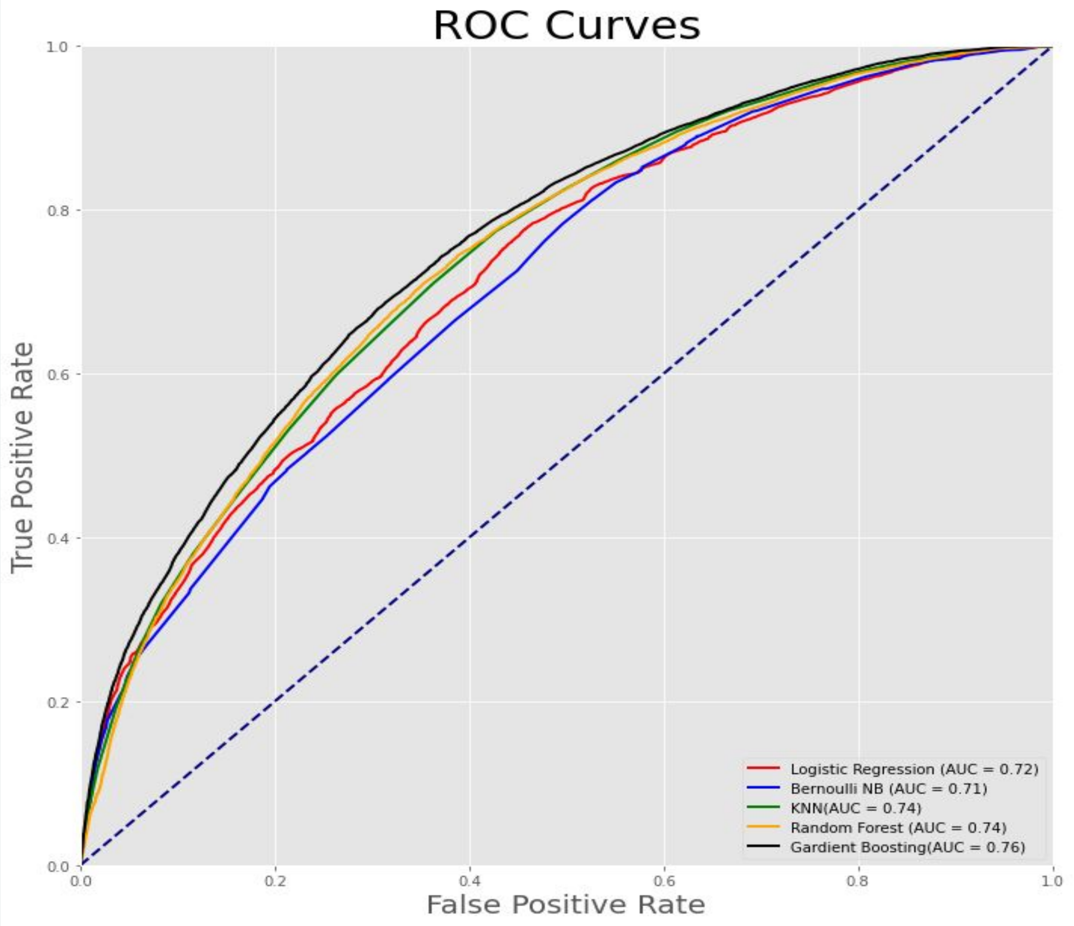

Confusion Matrix -RANDOM FOREST
10951
True Negatives
8299
False Positives
5566
False Negatives
19303
True Positives
Actual Campaign Outcome
Fail
Success
Fail
Success
Predicted Campaign Outcome


Confusion Matrix -GRADIENT BOOSTING
10782
True Negatives
8468
False Positives
5033
False Negatives
19836
True Positives
Actual Campaign Outcome
Fail
Success
Fail
Success
Predicted Campaign Outcome

MODEL EVALUATION

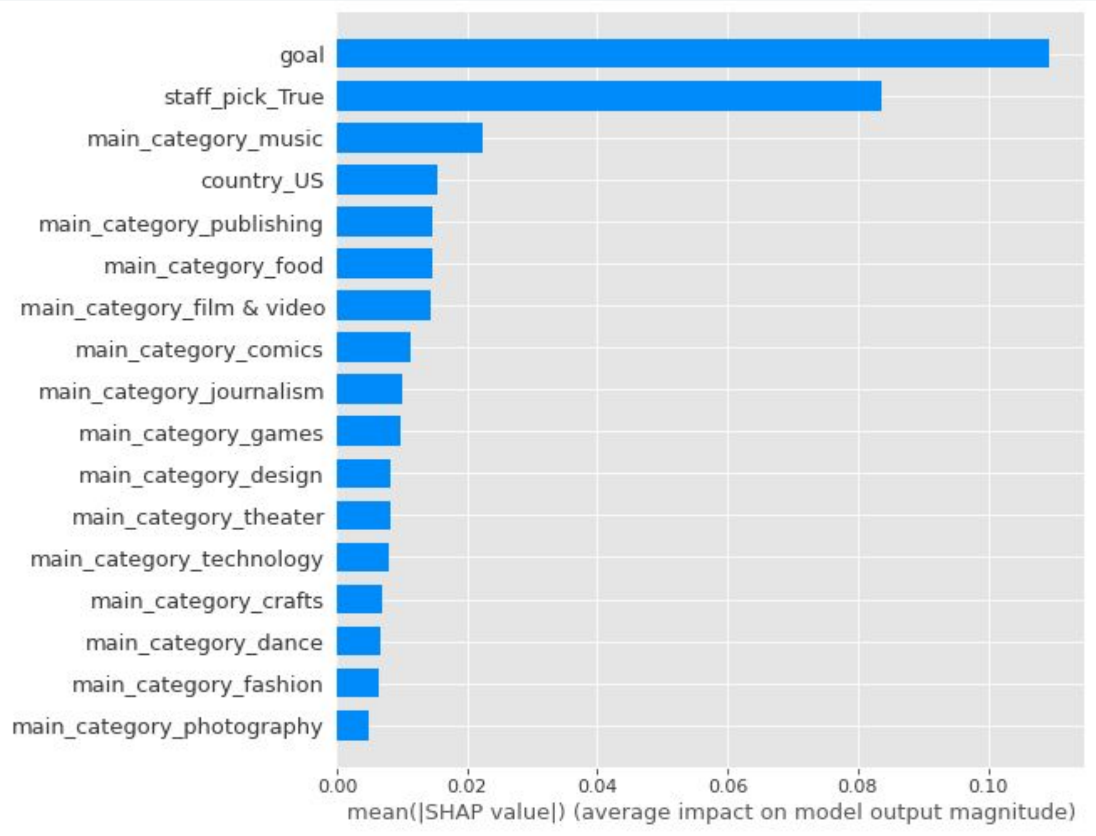
Model	LR	NB	KNN	RF	XGBoost
F-1	0.74	0.74	0.73	0.74	0.75
Recall	0.84	0.85	0.77	0.78	0.80
Precision	0.66	0.66	0.70	0.70	0.70
Accuracy	0.67	0.66	0.68	0.69	0.69

ROC CURVE



ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.72)
Bernoulli NB (AUC = 0.71)
KNN(AUC = 0.74)
Random Forest (AUC = 0.74)
Gardient Boosting(AUC = 0.76)

FEATURE COMPARISON (FOR XGBOOST)



FEATURE IMPORTANCE (FOR XGBOOST)

- Setting a low goal benefits the campaign
- Shorter campaigns do better
- Kickstarter endorsement improves chance of success
- Choice of category matters

LIMITATIONS

- High time complexity for large datasets of algorithms like SVM
- Manual feature selection
- Poor performance of some features in a particular algorithm.

THANK YOU