DETERMINING KICKSTARTER PROJECT SUCCESS

Team #48

Tim Dufala tdufala3

Bui Thi Thu Giang gbui8

Rashmi Raju rraju6

Abhishek Surya asurya6

Fernanda Tello mtello3

Georgia Tech College of Computing

KICKSTARTER

Summary

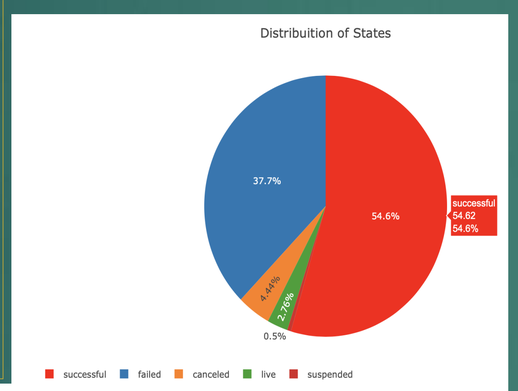
As of today more than 50% of the launched projects on Kickstarter fail to meet their funding goals. We wanted to know what factors make a Kickstarter project successful. **Kickstarter Project Success Predictor** is a web-based tool that uses predictive machine learning under the hood to highlight influential factors in a project campaign's success via interactive visualizations.

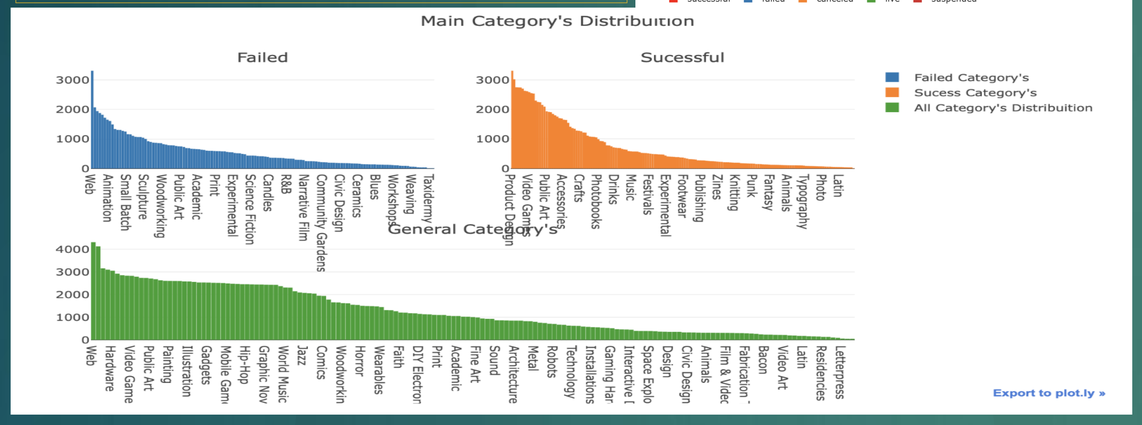
Approach

Our application allows users to upload data about their project and receive analysis and visualizations back. which can predict and suggest ways in which Kickstarter project can be improved. This will be successful because it will help the project creators focus on the weak variables to improve them.

Dataset

The Dataset used for machine learning prediction model and statistical visualization has been downloaded (18 October 2018) from webrobots.io. Web Robots scrapes Kickstarter data once-a-month. The raw dataset (~ 1 GB) was cleaned up by removing empty columns and rows. Size of dataset after data cleanup is (~ 23 MB) having in-total 186,337 project records.

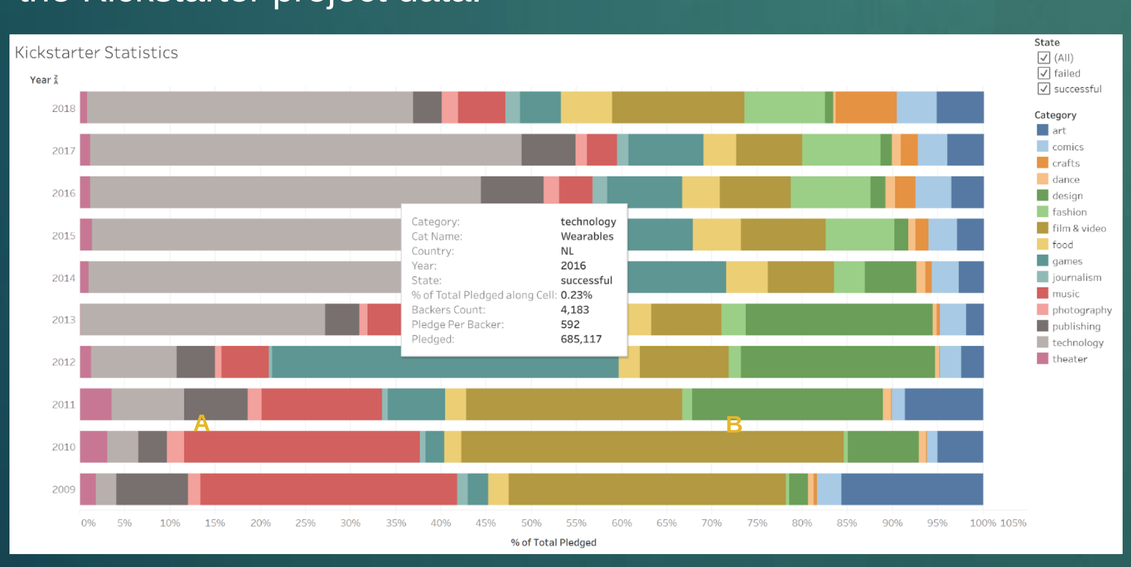




Statistical evaluation of dataset
A – Distribution of Categories across successful and failed projects
B – Distribution of all project states

Statistical evaluation & Visualization

We did a thorough evaluation of the dataset to generate an interactive horizontal bar chart in Tableau to give an overview of the Kickstarter project data.



Visualization#1 – Interactive statistical overview of Kickstarter projects using Tableau
(sources – webrobots dataset)

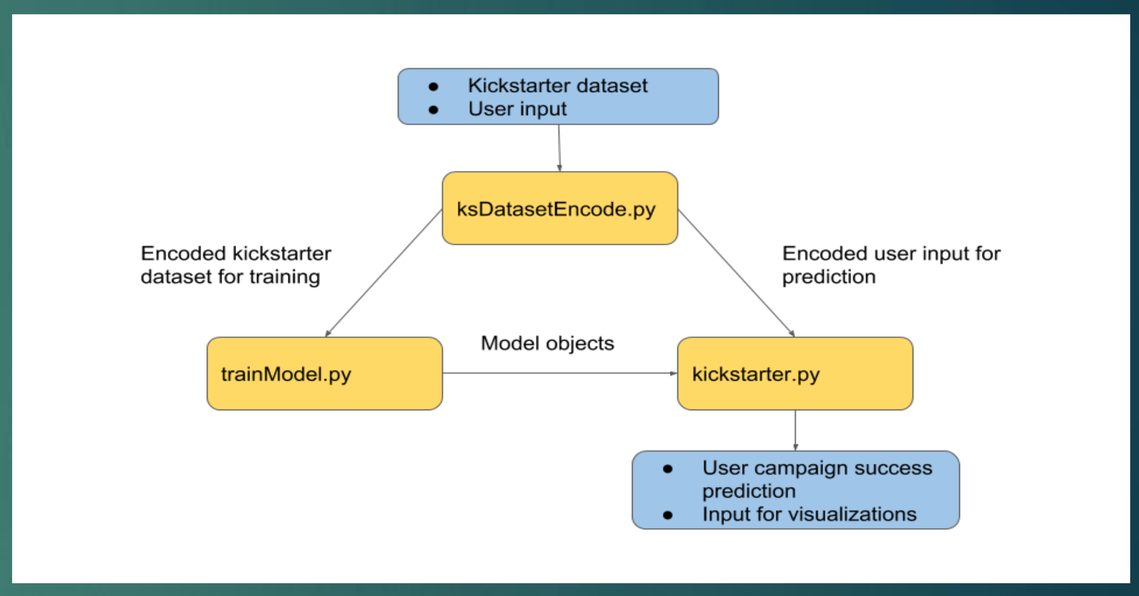
Min values		Mean values		Median values		Max values	
goal	0	goal	45633	goal	5000	goal	100000000
pledged	0	pledged	13600	pledged	1530	pledged	23343872
pledge_per_backer	0	pledge_per_backer	89	pledge_per_backer	46	pledge_per_backer	356374
duration	1	duration	34	duration	30	duration	91
backers_count	1	backers_count	144	backers_count	26	backers_count	105857

Model Building & Prediction

Tools/Libraries: Python ScikitLearn (model building and training), Pandas and Numpy (dataframe manipulation), Pickle (data persistence for i/o transfer), Matplotlib (generate an image of most important features)

Models:

Logistic Regression	Random Forest Classifier	Gradient Boosting Machine	KMeans Clustering
• L2 ridge regression: avoids overfit and considers all features • LIBLINEAR: classifies large data quickly • Test Accuracy: 96% • Result: Not used for prediction. Random forest provides higher accuracy	• 20 estimators: largest amount to avoid overfitting • Cross Validation 10 folds • Test Accuracy: 98% • Result: used to predict success of campaign	• Tried variations of depth and # estimators • Test Accuracy:100% • Result: not used due to risk of overfitting	• 50 buckets: trains fairly evenly distributed values for 'backers_count', 'goal', and 'pledge_per_backer' features • Result: average values provide thresholds and ultimately, suggestions to drive visual#2



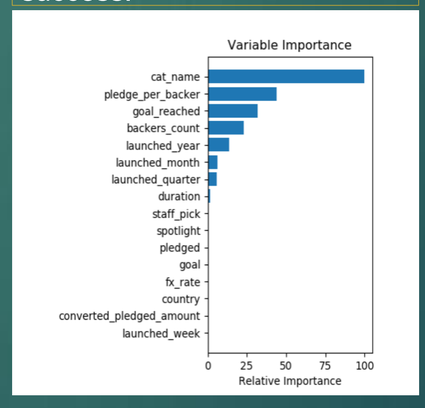
Approach: Overall Backend Model Building flow

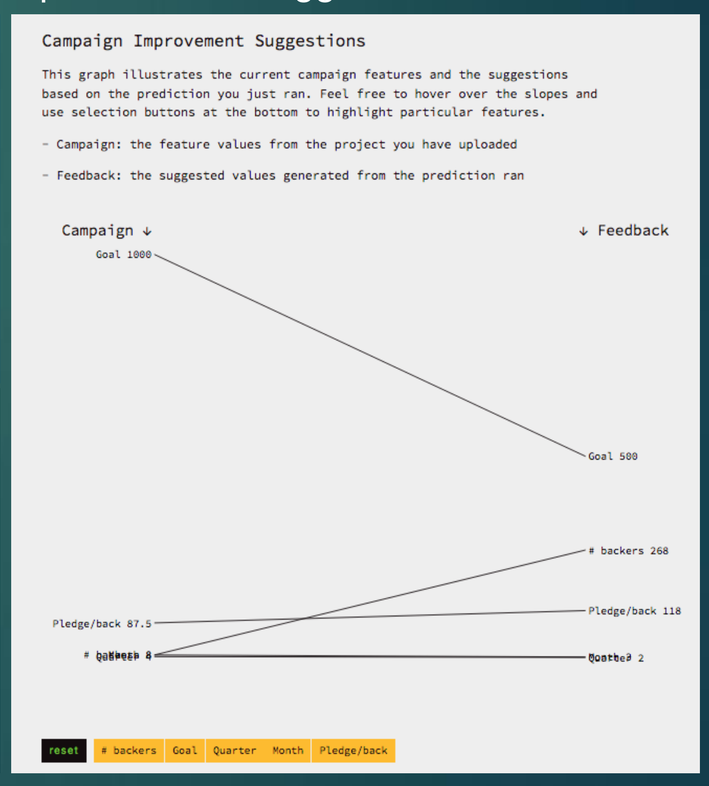
Evaluation Results

Prediction model gives overall 'Success'/'Failure' verdict for the provided project data. Output includes suggested feature values

Visualizations

Besides the overall prediction result, an interactive visualization (B) in the form of slope-graph is provided to show comparison between existing feature data and recommended feature parameters for project success.





Visualizations
left – Relative feature importance of given Kickstarter project using Python Matplotlib library
right – Campaign feedback for improvement using D3.js

Conclusion

Our product makes use of the best of both worlds: Visualization and Data Analytics to provide an enhanced UX. It takes advantage of the dataset statistics with improved visualization driven by the campaign success/failure prediction using the Random Forest with prediction of ~98% (greater than 85% we initially aimed) and KMeans Clustering. Furthermore, our diverse set of visualizations allows the user to understand their campaign success in various dimensions. Our application will be further validated with usability, unit and regression testing combined to cover frontend, backend, and model capabilities.