Kickstarter Projects - Data Analysis

By Sakina Zabuawala

Problem Statement

Kickstarter's All or Nothing policy and the requirement for the creators to give rewards to all their backers makes deciding the Goal Amount for their project very critical. So I wanted to create a model using Kickstarter's past project data to predict the amount of funding a campaign will get based on the information the creator puts on his/her project page. This would be a good tool for future entrepreneurs while setting their goal.

Data Collection

I scraped the Kickstarter website using both Selenium and Beautiful Soup. I focused my analysis on projects based out of USA only. Since the website would allow me to scrape only 2400 projects for each of its 15 categories, I ordered them based on the amount of funding they received. I collected the following information about each project: category, sub-category, goal amount, status, location, launch date, deadline, information about the various reward tiers, title, blurb and the number of backers.

Data Cleaning

I ended up scraping 36000 projects from the website. The data was pretty clean to begin with, I just had to do a few checks. This led to dropping the duplicates, removing projects which were not based out of USA and selecting projects that were either successful or failed. Finally, I was left with 35245 projects.

Exploratory Data Analysis

An initial look at the data gave the following information:

1. The dataset had ~92% successful and ~8% failed projects and only 4 of the 15 categories had those failed projects.
2. Some categories received significantly higher funds than others
3. The target variable (pledged amount) and the independent variable (goal amount) were highly left skewed
4. The correlation between the target variable and the independent variables was almost zero

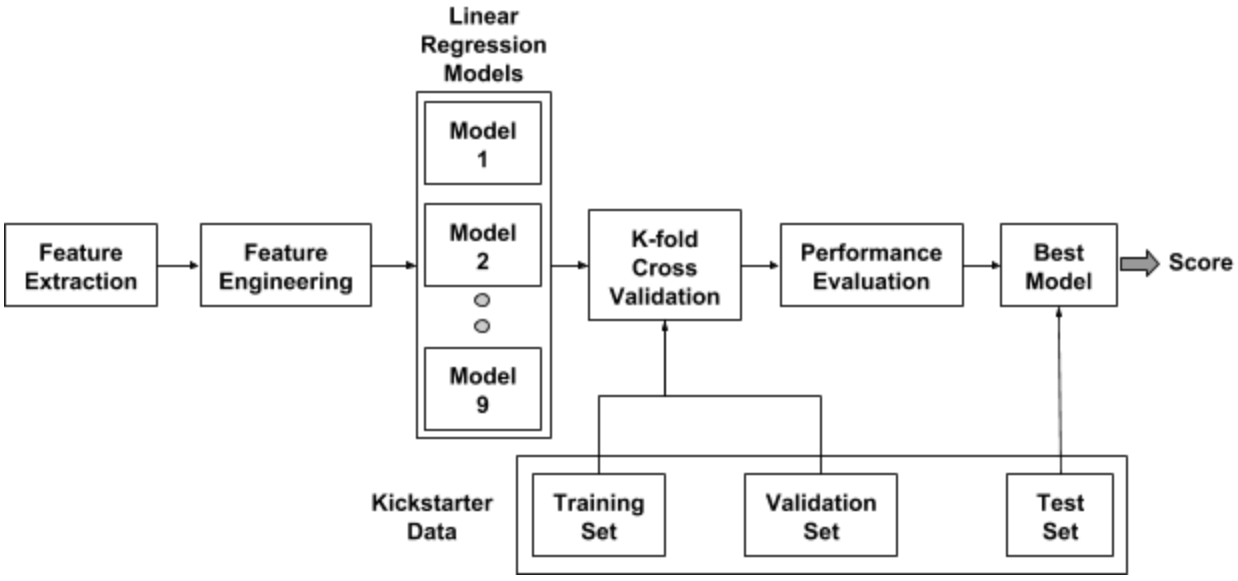
Linear Regression Models

Before trying out my models, I first divided the dataset into two sets - a Cross Validation set (28196 projects) and a Holdout set (7049 projects).

The following table shows the results of the 5-fold cross validation on each of the Linear Regression models that I tried.

Model	Target variable	Feature Set	Regularization	R-squared	P-values	Skew/ Kurtosis
1	pledged_ amount	'category', 'goal_amount', 'duration', 'rewards_count'	No	-0.152	Some >0.05	38.257/ 2861.675
			Yes	-0.152		
2	pledged_ amount (threshold=95 Percentile)	'category', 'goal_amount (with threshold = 95 Percentile)', 'duration', 'rewards_count'	No	0.567	duration >0.05	2.924/ 17.115
			Yes	0.567		
3	log (pledged_amount)	'category', 'log(goal_amount)', 'duration', 'rewards_count'	No	0.75	All <0.05	-0.299/ 11.864
			Yes	0.75		
4	pledged_ amount (threshold=95 Percentile)	'category', 'goal_amount (with threshold = 95 Percentile)', 'duration', 'rewards_count', 'goal_amount^2', 'duration^2', 'rewards_count^2'	No	0.568	Degree 2 features >0.05	2.941/ 17.222
			Yes	0.568		
5	log (pledged_amount)	'category', 'log(goal_amount)', 'duration', 'rewards_count', 'rewards_min', 'rewards_max', 'title_wordcnt', 'blurb_wordcnt', 'pledged_log'	No	0.751	All <0.05	-0.299/ 11.797
			Yes	0.751		
6	log (pledged_amount)	'category', 'log(goal_amount)', 'duration', 'log(rewards_count)'	No	0.757	All <0.05	-0.247/ 11.21
			Yes	0.757		
7	log (pledged_amount)	'category', 'log(goal_amount)', 'duration', 'log(rewards_count)', ‘state’	No	0.758	State features >0.05	-0.245/ 11.123
			Yes	0.758		
8	log (pledged_amount)	'category', 'log(goal_amount)', 'duration', 'log(rewards_count)', ‘launch_quarter’	No	0.757	Quarter features >0.05	-0.246/ 11.207
			Yes	0.757		
9	log (pledged_ amount (threshold=95 Percentile))	'category', 'log(goal_amount (with threshold = 95 Percentile))', 'duration', 'log(rewards_count)'	No	0.753	All <0.05	0.456/ 12.241
			Yes	0.753		

Workflow



Final Model

log(pledged_amount) = 9.4716 + 0.6793 * log(goal_amount) -0.0132 * duration
+ 0.2633 * log(rewards) + 0.0698 * Art + 0.0695 * Comics
- 0.2085 * Crafts - 0.1355 * Dance + 0.4379 * Design
+ 0.1304 * Fashion + 0.1934 * Film & Video + 0.1212 * Food
+ 0.4476 * Games - 0.7213 * Journalism + 0.1180 * Music
- 0.0700 * Photography + 0.1341 * Publishing
+ 0.3979 * Technology

The slope coefficient of 0.6793 means that on the margin a 1% change in goal amount is predicted to lead to a 0.67% change in pledge amount, with a compounding of this effect for larger percentage goal amount changes. Similarly, a 1% change in number of reward levels is predicted to lead to a 0.26% change in pledge amount. The -0.0132 coefficient shows that duration and pledged amount are negatively related.

Looking at the coefficients for the various categories, we notice that the categories (for example, Journalism) which had failed projects have negative values and the categories (for example, Games) which had the maximum overall funding have high positive values.

Results with Holdout Set

Finally, I tested my model on the Holdout set to see how well it performs on unseen data. The above model fit to my holdout data gave me a R-squared value of 0.7466 which is marginally worse than my cross validation R-squared value.