Term Project

Kinnick Fox

DSC550

Bellevue University

Introduction

The crowdfunding platform Kickstarter was the focus for this project. I have used Kickstarter for a few years now to fund some small projects that I believe should see a completed state. Kickstarter projects are often referred to as campaigns. Each campaign starts with a hard deadline at which time all funds must be accumulated from backers and meet or exceed the funding goal, also set at the beginning of the campaign. Backers are able to pledge any amount when backing a project and in some cases the creator of the project will set specific pledge amounts, also known as tiers, that will cause the backers of that amount or more to receive a physical or digital item from the campaign. There are many categories of projects on Kickstarter, many with subcategories that fall into certain niches. My goal with this project was to find which category and subcategory of project that has the highest likelihood of successfully funding during its campaign as well as creating a model with a high accuracy of predicting the success of a campaign.

Real World Implication

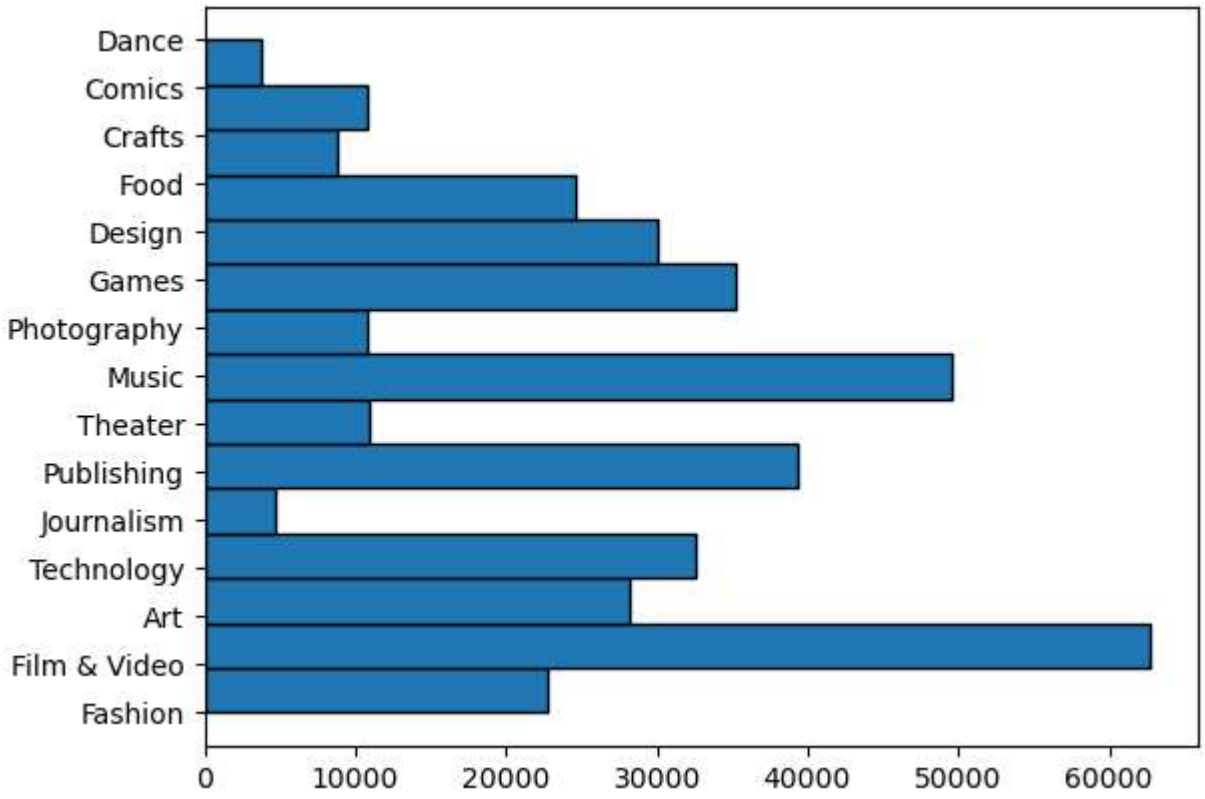
In the real world, this data could most likely be used to convince distributors and manufacturers to begin prototyping a project before campaign completion. These numbers could be used to show, with high certainty, that a project will be successful even if the funding goal has yet to be reached based purely on the categorical data used in creating the Kickstarter.

Data Source

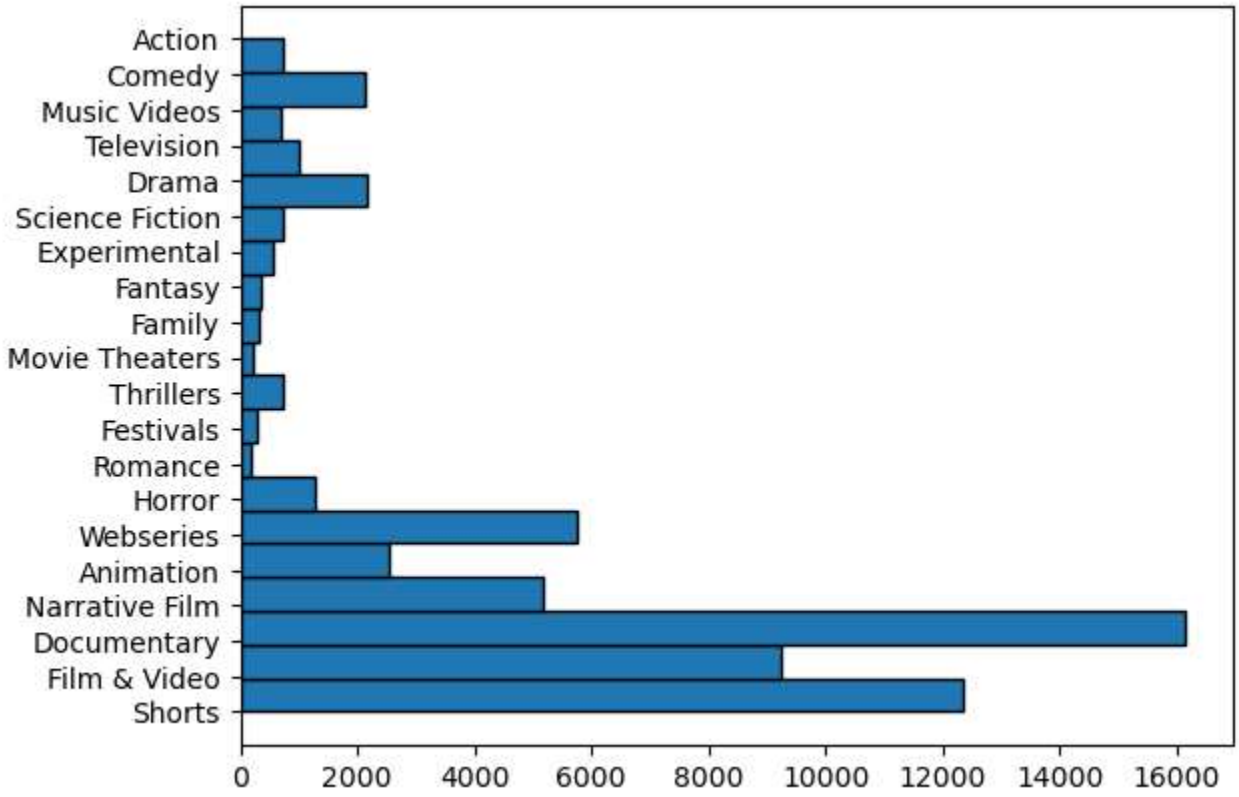
The data used in this project was obtained from Kaggle (https://www.kaggle.com/datasets/ulrikthygepedersen/kickstarter-projects). This dataset contains category, subcategory, start, end, funding goal, and final amount pledged data that was important for the completion of this project. The data ranges from the years 2009 to 2018 and possesses over 350,000 datapoints.

Milestone 1: Data Gathering

This milestone focused on the decision of which dataset to use as well as obtaining some basic information about the dataset. For starters, I wanted to see counts of the categories to ensure a usable distribution.



This graphic clearly shows that film/video was the most common category and journalism and dance are the least common. That said, both journalism and dance had approximately 3000 datapoints each which I did not believe was a small enough number for them to be considered outliers. Next, I wanted to see how many subcategories were present in film/video category to get a sense of how many a single category could possess.



This visualization shows that the category containing the most data within the dataset contained twenty subcategories. Some of the subcategories were lacking in data but this will be addressed later.

Milestone 2: Data Preparation

This milestone focused on cleaning and transforming the data features to suite the project. Any campaigns lacking categorical data were dropped from the dataframe. Any “Live” projects were dropped, these projects had not yet hit their deadline and thus could not be discerned whether or not they would be successful. Canceled projects where also dropped. When a project is cancelled, funding stops immediately which could cause issues in analyzing the total time of a project. Average pledge amount per backer and time elapsed columns were added to the dataframe using mathematical formulas. Last, but certainly not least, the state, category, subcategory, and county features were replaced by dummy variables representing the same data. All said and done, this left the dataframe with a little over 330,000 datapoints to work with.

Milestone 3: Data Modeling

Work for this milestone was focused on the creation and rating of data models. Due to the target outcome of the data being a Boolean variable, logistic regression was chosen as the modeling solution. An early blunder was caught when the model returned an accuracy of over 99%. This was due to the presence of the amount pledged and average amount pledged per backer features and the model simply comparing these features to the goal feature. Once removed, the model maintained an accuracy of nearly 90% which is still fairly accurate and, more importantly, realistic. The most influential features were also identified by scoring their coefficients. Ordering

features	coef
Category_Film & Video	[0.6352877552199053]
Category_Music	[0.4924212955062913]
Subcategory_Shorts	[0.31640266899439684]
Category_Theater	[0.3082825086246921]
Subcategory_Theater	[0.20699222548149687]
...	...
Subcategory_Tabletop Games	[-0.20020391727679468]
Category_Design	[-0.25250698667577404]
Subcategory_Video Games	[-0.30223269187941654]
Category_Publishing	[-0.333614894269565]
Category_Games	[-0.6512075391766338]

features	coef
Subcategory_Chiptune	[1.6376316386278986]
Subcategory_Residencies	[1.471203592948301]
Subcategory_Anthologies	[1.3005773519361845]
Subcategory_Dance	[1.2586507862895115]
Subcategory_Indie Rock	[1.2313833170714126]
...	...
Subcategory_Software	[-1.236392936604073]
Subcategory_Food Trucks	[-1.2443019702909552]
Subcategory_Web	[-1.4736244297951198]
Subcategory_Mobile Games	[-1.5326639024161217]
Subcategory_Apps	[-1.9439462329455302]

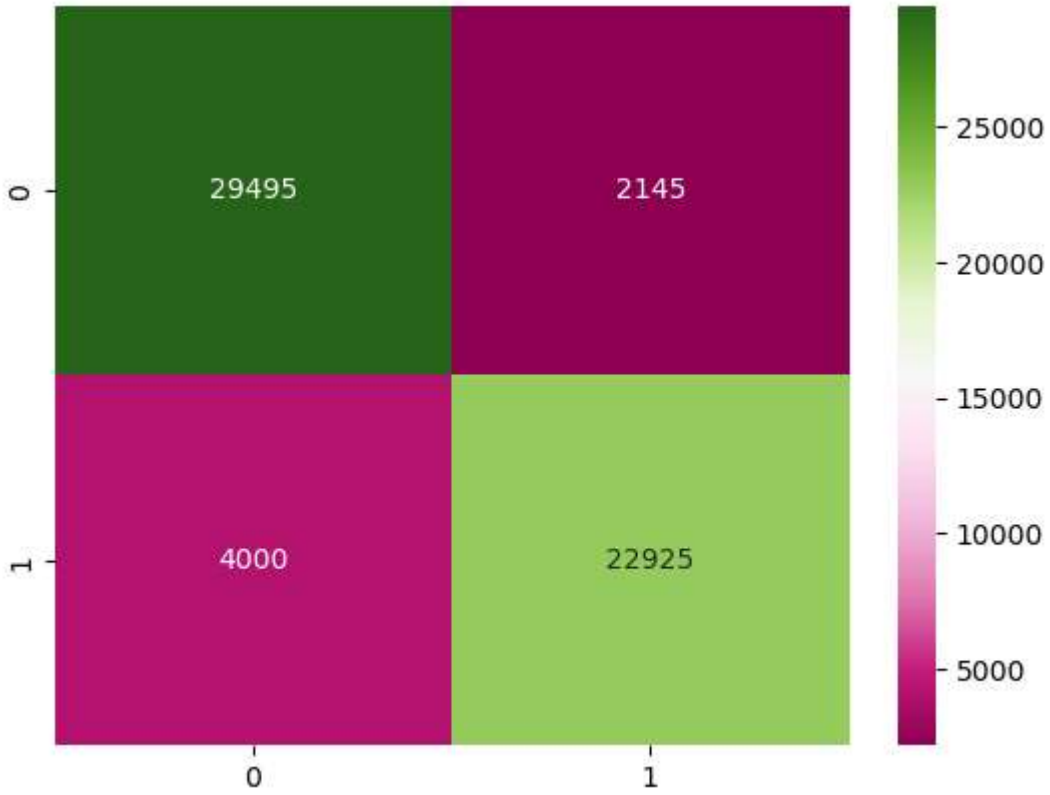
features	coef
Category_Dance	[0.9880341439311862]
Category_Theater	[0.9267404899225129]
Category_Comics	[0.6098947830775179]
Category_Music	[0.5505937782364076]
Category_Art	[0.2060338562229546]
Category_Film & Video	[0.0805920369904562]
Category_Games	[0.020616216620302576]
Category_Design	[-0.07692441078133919]
Category_Publishing	[-0.2086985873287951]
Category_Photography	[-0.23181474293901952]
Category_Fashion	[-0.4844117101977026]
Category_Food	[-0.5729772831794934]
Category_Crafts	[-0.5742986246925822]
Category_Journalism	[-0.6244302084668731]
Category_Technology	[-0.7775665004795673]

the features by their coefficients revealed the top five most positive and most negative features. This led to similar models being built for categories, subcategories, and countries to attempt to see their influence on the success of a project in greater detail.

features	coef
Country_Hong Kong	[0.5242733333173674]
Country_United States	[0.3950943015194249]
Country_United Kingdom	[0.37437140293865356]
Country_Denmark	[0.30921070145279467]
Country_Singapore	[0.22202742005100698]
Country_France	[0.19188442634736758]
Country_New Zealand	[0.1063233247981489]
Country_Canada	[0.08584082292988794]
Country_Sweden	[0.06303972394450169]
Country_Luxembourg	[-0.004593331015587865]
Country_Japan	[-0.02407818942271372]
Country_Mexico	[-0.075594836946959]
Country_Australia	[-0.07925026562891495]
Country_Ireland	[-0.0973075280093377]
Country_Switzerland	[-0.16508374527765765]
Country_Norway	[-0.1686091081672948]
Country_Belgium	[-0.19038657734855619]
Country_Germany	[-0.22893238776609576]
Country_Spain	[-0.23271739709897704]
Country_Netherlands	[-0.2919372439038888]
Country_Austria	[-0.5266199133743545]
Country_Italy	[-0.6873673742926518]

These findings were very interesting and led to the inclusion of a confusion matrix for the final week of the project.

Conclusion



This visualization shows the actual outcome of campaigns on the x-axis as 0 being a failed campaign and 1 being a successful campaign with the predicted outcome of campaigns on the y-axis is attempting to predict the outcome based on all of the categorical data gathered. As you can see it is fairly accurate, but tends to predict false positive (bottom-left quadrant) about twice as often as it predicts false negatives (top-right quadrant). The outcome of this model was better than expected and fulfilled the original goal of the project. In a future iteration, end of

day totals could be added for live project monitoring to more accurately predict the projected end state of the project in conjunction with the insight generated with this model.