Kickstarter Campaigns

Success Factors and Analysis

Monica Boomgaarden
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Monica.Boomgaarden@colorado.edu

Rodrigo Garcia
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Rodrigo.Garcia-1@colorado.edu

Nicole Lincoln
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Nicole.Lincoln@colorado.edu

ABSTRACT

Kickstarter is a crowdfunding platform that acts as the foundation and economic engine for millions of community inspired projects worldwide. The ability to predict the success of a Kickstarter campaign would help many projects raise the funds they need to make their ideas become reality. In our analysis, we explored the work performed by other data scientists and attempted to develop our own Logistic Regression model, using python and open-source libraries, in an effort to identify an attribute or combination of attributes that can strongly predict a campaign's success.

1 INTRODUCTION

1.1 Problem Statement

Kickstarter is a website which first launched on April 28th, 2009 with the main purpose of bringing user inspired projects to life through community backing. It allows community members to provide financial support for a project's development and production. The final result of the project is then shared with the people who backed it. Since its inception, millions of people have provided support for the new ideas on Kickstarter's platform; resulting in approximately $5.3 Billion dollars being raised. The projects span a wide range of categories from Arts and Games to Dancing and Technology. Kickstarter's approach to project funding is "all or nothing". Therefore, if a project fails to achieve its community-based funding goals it receives no funding at all and each backer's financial contribution is returned. As a result, achieving the desired funding goals for the project is crucial. Therefore with our analysis we explore the questions: Can attributes be identified which impact the outcome of a Kickstarter campaign? If so, which factors contribute most to the success of a Kickstarter project's success or failure? Addressing these questions could provide useful data for future projects' success in achieving funding.

1.2 Project Plan

We obtained a relatively clean dataset so we anticipated doing very little cleaning other than dropping categories that were not relevant to our analysis. We planned on performing exploratory data analysis to determine what attributes provided interesting information. Then planned to build a logistic regression model using those attributes. After creating our model, we then planned on performing backwards stepwise attribute elimination to determine which attributes were strong predictors for our model. Attributes that were not would be eliminated not in order to arrive at our final predictive model. We would then analyze the results of the model to see what

it could tell us about predicting the success of a Kickstarter.

1.2 CCS Concepts

• Computing Methodologies • Machine Learning
• Data Science

1.3 Keywords

Kaggle, Machine Learning, Kickstarter, Success Prediction, Crowd Sourced Funding, Data Science, Logistic Regression

2 LITERATURE SURVEY

Kickstarter's results have been previously analyzed in other work by data scientists. We identified three articles that we thought would provide useful insights for our project. Each article discusses its own approach to analyzing the attributes present in its respective Kickstarter data, with the objective of developing a prediction model to determine the factors which contribute to a campaign's success. These articles were as follows:

- **[1] "We Analyzed 331,000 Kickstarter Projects. Here's What we Learned About Kickstarter Success"**: This article's primary goal was to look at the relationships between categories, subcategories and days of the week and having a successful funding campaign.
- **[2]** "**Kickstarter Analysis and Prediction"**: This paper looked at the category, goal, and country and attempted to create a predictive model of whether or not a project would meet its funding goals.
- **[3]** "**Using Machine Learning To Predict Kickstarter Success"**: This project attempted to create a predictive model for a Kickstarter project's success or failure by analyzing a wide range of attributes. They found the following attributes to have the most impact: goal, length of campaign, month campaign was started in, time of launch, country of origin, category, day of the week, and being a staff pick.

Of particular interest was the work performed by Lewis[3] . In her literature, a total of four different models were evaluated to determine which could provide the best prediction. The selected models were: Logistic Regression, Logistic Regression with PCA, Random Forests, and XG Boost. A brief overview of each methodology is described below:

PCA, short for Principal Component Analysis, is a dimensionality reduction method, which is typically used to reduce a large set of data into only its most important variables. This reduction can impact the accuracy of the results, but the trade off is a far more parsimonious model with fewer variable inputs.

Random Forests is an ensemble method that combines many decision trees into a single composite classification model. Boosting is also an ensemble method, which combines many weak learners to produce a strong learner. In this case, trees are built sequentially in such a way that each subsequent tree reduces the errors of the previous tree. XG Boost stands for eXtreme Gradient Boosting and uses gradient calculations to describe the steepness of the error function.

Each of the aforementioned models varied in the amount of time required to calculate their respective results: Logistic Regression w/PCA 48.56 min, Random Forests 72.2 min, and XG Boost (an eye watering) 14.5 hrs. No analysis time was provided for Logistic Regression alone. Ultimately, the results varied little from model to model, with each delivering a weighted average

F1 score of about 70%. These results clearly indicate that more complex and lengthy analysis methods delivered no additional predictive value and were therefore not considered for this project. Logistic regression by itself seemed to be the model that provided the best predictive capabilities and the least onerous runtime.

Overall, none of the articles that we reviewed identified any single attribute as a strong predictor of a Kickstarter project's success.

3 DATASET

Our data set can be found at https://www.kaggle.com/kemical/Kickstarter-projects. There are two versions of the Kickstarter data available. We will be using the 2018 version, since it includes more complete information and the dataset is more up to date. This is a multivariate dataset that provides information regarding the success and failure of recent Kickstarter projects. There are 15 attributes and 378,662 items in the dataset. The attributes include the following:

- **ID**: numerical, unique campaign identifier
- **Name**: nominal, describes the name of the projected
- **Category**: categorical, actually a subcategory of the Main_category
- **Main_category**: categorical, describes the overarching category of the campaign (e.g. Music, Crafts, Games, etc.)
- **Currency:** categorical, describes the currency types for the monetary columns
- **Deadline:** deadline for pledges to the campaign.
- **Goal:** numerical, ratio-scaled; monetary value that needed to be reached in order for campaign to be funded
- **Launched**: numerical, interval-scaled; date the campaign was launched
- **Pledged**: numerical, ratio-scaled; monetary amount that backers of the campaign agreed to pay
- **State**: Categorical, describes the current state of the project - successful, failed, or cancelled
- **Backers:** numerical; indicates the number of people who pledged support to the campaign
- **Country:** categorical; country location of the company that is seeking funding
- **usd pledged**: numerical; dollar amount of pledges converted into US dollars by Kickstarter
- **Usd_pledged_real:** numerical; dollar amount of pledges converted into US dollars by fixer.io
- **Usd_goal_real:** numerical; dollar amount of goal converted into US dollars by fixer.io

4 TECHNIQUES APPLIED

4.1 Preprocessing

Preprocessing for the Kickstarter dataset was straightforward. One of the first steps was to determine if any duplication of the data existed within the set. This was achieved by using the ID column since each Kickstarter is assigned a unique identifier. If any duplication did exist, it could be assessed and removed. Fortunately, no duplicate entries were found. The ID category was then dropped since it would not provide any further contributions to our analysis.

The remaining columns were then analyzed to remove irrelevant or redundant data. **Pledged**, **usd_pledged** and **goa**l, were eliminated since their values were already reflected in **usd_pledged_real** and **us_goal_real attributes.**

Next, we looked at the different states each Kickstarter reached. According to the data, each particular campaign wound up in one of six states: failed, cancelled, successful, live, undefined, and suspended. Figure 1 below shows the total number of Kickstarters tallied for each state.

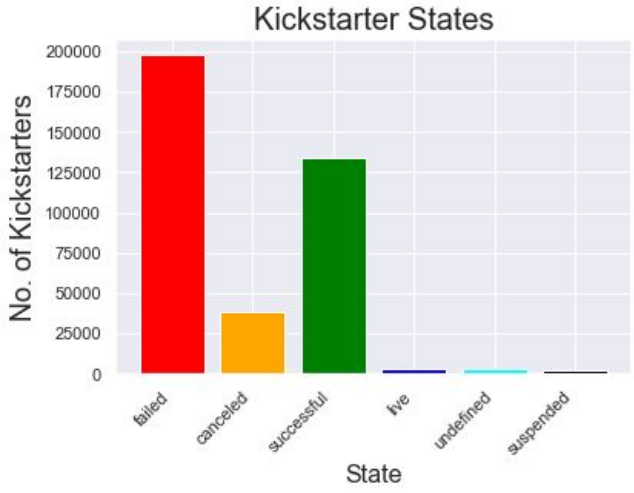

Figure 1.

A failed state indicates a Kickstarter was unable to achieve its funding goal. A cancelled state indicates that the individual who launched the project, for one reason or another, aborted the campaign before the Kickstarter ended. A live state indicates that the project was still in its funding round when the data was collected. The undefined state is for Kickstarters whose campaigns have yet to launch, and suspended are campaigns which were stopped by the Kickstarter platform.

Since the goal of the project is to predict success or failure of a Kickstarter campaign, any columns that did not explicitly meet these two states were dropped.

Figure 2.

The above figure illustrates the remaining data set after further filtering. Interestingly, almost 60% of the Kickstarters which launch fail to achieve their target goals, while only about 40% find the necessary financial support from their backers to launch their project.

With the preprocessing and data cleaning completed, the next step was to perform exploratory data analysis to get better insights into the dataset, to identify important attributes, transform the data into meaningful and useful representations, further detect and identify outliers/anomalies, and to begin to piece together the components that would be necessary for our logistic regression model.

4.2 Exploratory Data Analysis

Our Exploratory Data Analysis revealed some interesting differences between failed and successful projects. First we plotted the total count of projects launched by the Main Category attribute.

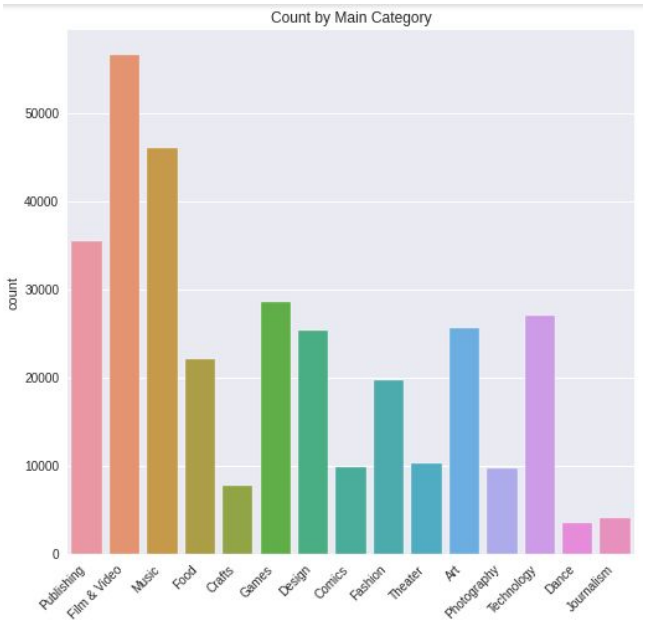

Figure 3.

In Figure 3, we can see that Film & Video projects are the most common, and Dance projects are the least common.

We also looked at the Main Category counts broken down by successful projects and failure projects which is shown in figure 4.

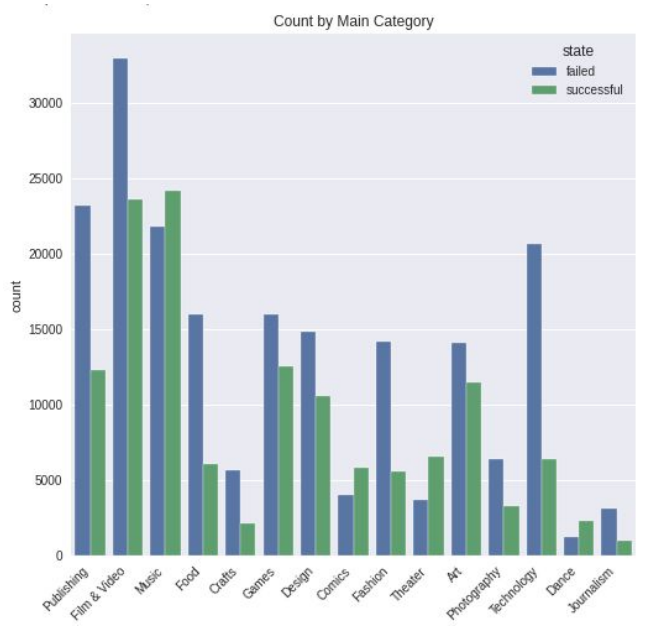

Figure 4.

The bar chart presented in Figure 4 provides more interesting detail because it shows us which Main Category was more likely to succeed. This chart shows that the only categories where success rate is above 50% are Music, Comics, Theater and Dance. Although Dance has the fewest number of projects launched, it has the highest success rate. In fact, it has over a 65% success rate. In contrast, Technology has the lowest success rate of 24%.

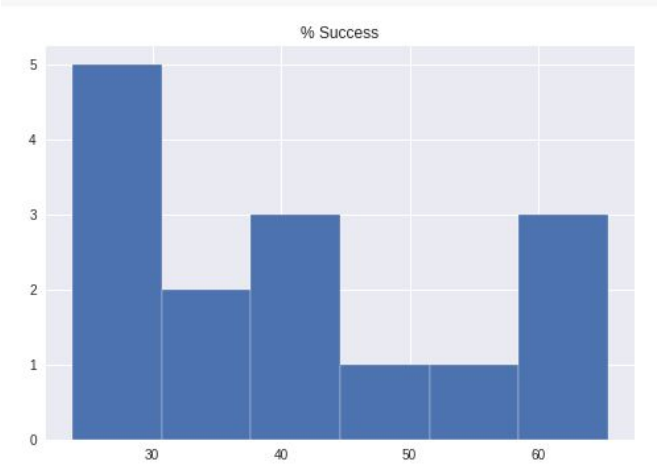

Figure 5.

As previously stated, only four main categories have a success rate above 50%. Figure 5 shows that of the remaining 11 projects, five have success rates under 30%.

To further analyze success and failure rates of projects we looked at their financial aspects. Projects with lower goals tended to have higher success rates than those with higher goals. As shown in Figure 6, the median goal for successful projects was slightly less than $4,000 while the median goal for failed projects was close to $7,500. We further noted that successful projects had more backers than unsuccessful projects, but that the average contribution per

backer was relatively consistent regardless of whether the project was successful or failed.

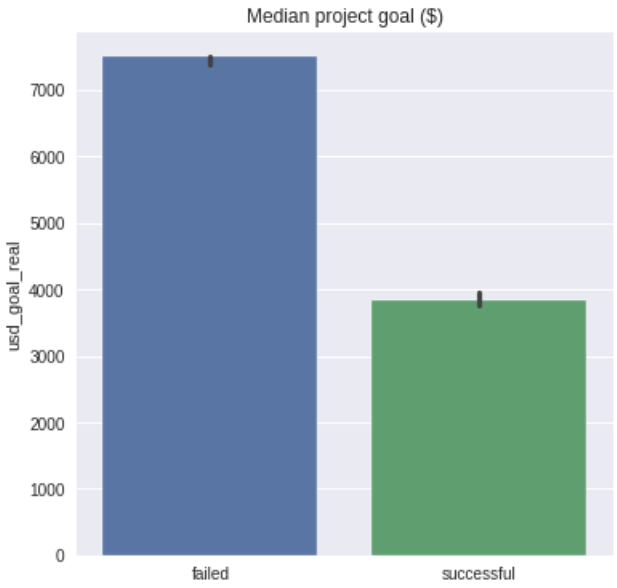

Figure 6.

We further broke down the median goal by main category, in order to see which categories had the highest median goal, compared to their median pledges. Figure 7 shows the median goal by main category. Technology had the highest median goal, of over $17,500, with Design and Music as the top two and three categories, both with median goals of $10,000. However, as shown in Figure 8, none of the Main Categories achieved a median a median amount pledged over $2,500.

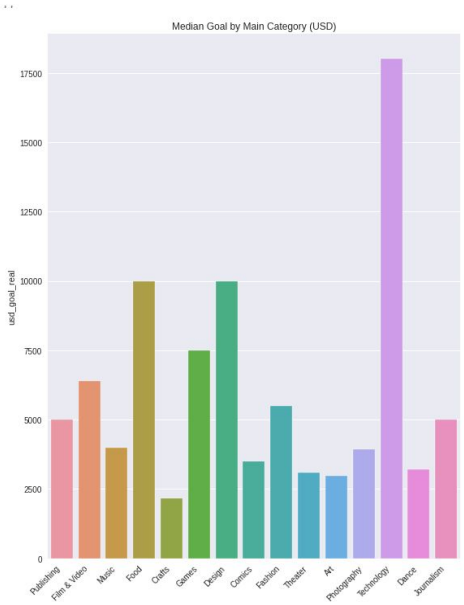

Figure 7.

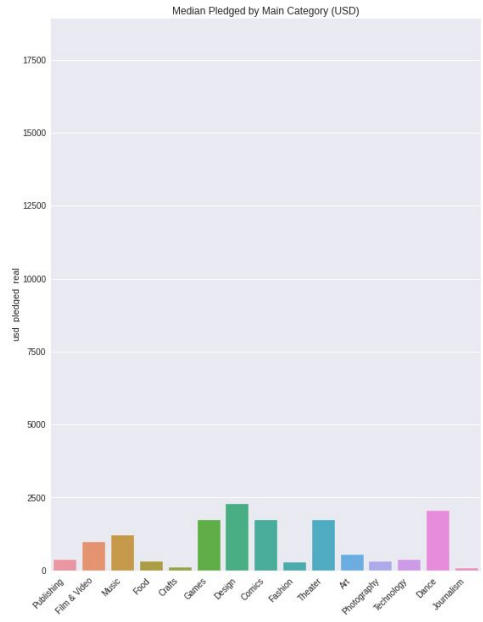

Figure 8.

Figure 9 and Figure 10 below show median goal and pledge information for only those campaigns that were successful. When only looking at successful campaigns, the median amount pledged exceeds the median goal, often by as much as twice the amount.

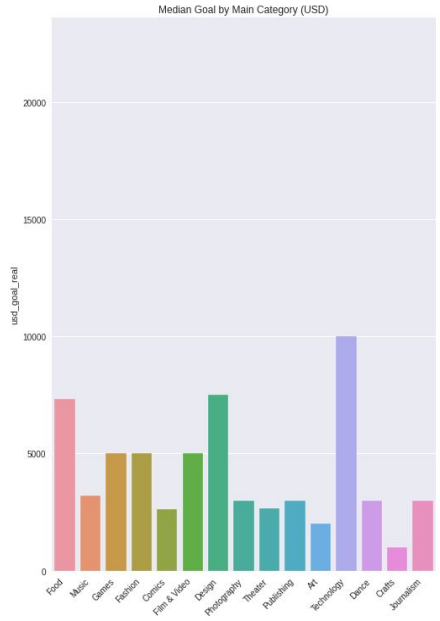

Figure 9.

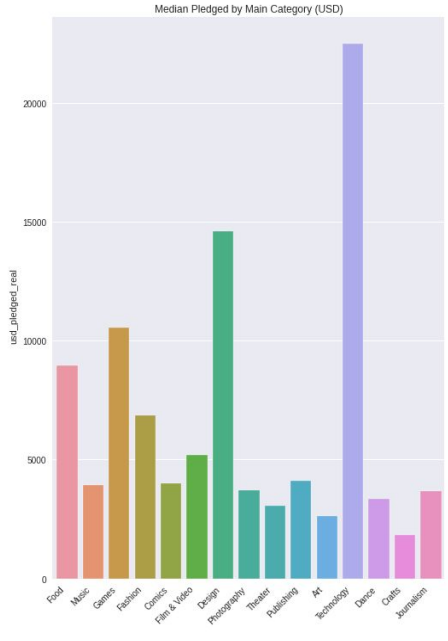

Figure 10.

In contrast, when looking at failed campaigns only, the median goal amount is significantly higher than the amount pledged (Figure 11 And Figure 12). This is to be expected since failed campaigns by definition do not meet their funding goals. However, this demonstrates the large gap by which failed campaigns typically fail.

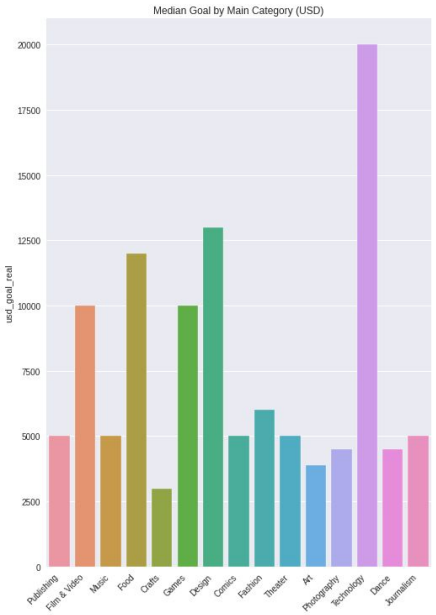

Figure 11.

Figure 12.

We did some further exploration of the goal metrics by successful and failed campaigns.

	Successful	Failed
mean	9532.85	63174.74
std	27961.44	1435682.63
min	0.01	0.15
25%	1301.92	2500.00
50%	3837.74	7500.00
75%	10000.00	20000.00
max	2015608.88	166361390.71

Figure 13.

There are significant differences between each of the statistics shown in Figure 13. Most notably, the mean goal for a failed campaign is over six times that of a successful campaign. The max value of the successful campaigns in our dataset is just over $2 Million, however, the max value of an unsuccessful campaign is over $166 Million. Lastly, the median goal for a successful campaign is only about half of what the median goal is for a failed campaign.

4.3 Modeling

4.3.1 Model Selection

Since a Kickstarter either succeeds or fails, predicting the outcome of a Kickstarter is a classification problem.

As mentioned above, we selected binary logistic regression to model this classification problem. Logistic regression is a statistical model that uses a logistic equation to model a binary dependent variable, in this case the success or failure of a Kickstarter. Logistic regression models can accept multiple independent variables. Independent variables may be binary or continuous. The outcomes of the logistic model represent the probability of the outcome of the dependent variable being one. Anything with odds below .5 is interpreted to be a 0 and anything with odd above .5 is interpreted to be a 1 when analyzing the outputs from the model.

4.3.2 Test and Training Data

Our data was skewed toward failed projects, so prior to training our model, we balanced the data set by randomly choosing 100,000 data points from each state. When running the model with the imbalanced data, the model was unable to predict successful models correctly because it's predictions were outweighed by the distribution of the dataset. Balancing the data helped this significantly.

We also scaled the monetary Goal category. We used min-max normalization to scale the goal between [0.0, 1.0]. We used min-max scaling to minimize the effects of large values. As mentioned above there is a large difference between the maximum goal value for a failed campaign and a successful campaign, and scaling helped reduce the effect of the large variance in goal values.

For the campaign duration, we transformed the duration into categorical variables by using buckets for 15-day periods of time. We did this because, in the event that duration was an important feature, we felt it would be a more useful piece of information to tell someone that the ideal duration of a campaign was a timeframe, rather than just broadly saying that duration was relevant to the success of a campaign. We then used one-hot encoding on duration and the rest of the categorical variables, in order to use these in our model.

4.3.3 Attribute Selection

Prior to running our model, we also dropped several more attributes, as we felt they would not add to our model based on the EDA, or because

we had transformed the attribute into a more useful piece of information.

We fit our logistic regression model, and received a score of approximately 0.60. We then performed feature importance analysis to determine which attributes were impacting our model the most.

We used sklearn's RFE method to analyze feature importance. We also used the coefficient method to extract the coefficients for each of the important features. From this, we found that 'goal' was the most important feature, with a -5.5620 coefficient, indicating a very strong negative correlation, compared to the remaining features. The next closest was the 'main_category: Dance' feature, with a positive 1.0136 coefficient. We refit our model using the top ten features, including 'goal', eight of the 'main_category' features, and a 'duration' of 46-60 days.

5 KEY RESULTS AND FUTURE WORK

5.1 Key Results

Our final model predicted the success of a Kickstarter campaign with a R^2 score of around 0.59.

Figure 14 shows the confusion matrix for our results.

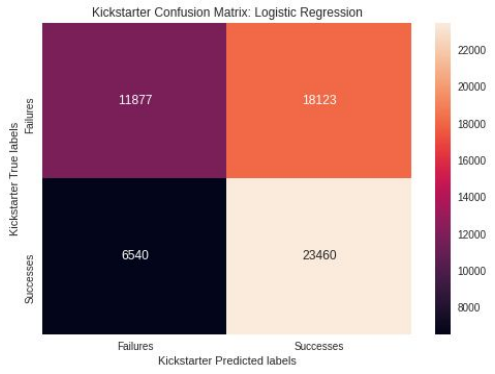

Figure 14.

Out of 60,000 test samples, the model accurately predicted 35,337 campaigns correctly, giving us an accuracy score of approximately 0.59. Our model correctly predicted 23,460 successful campaigns, for a recall score of approximately 0.78. The model also incorrectly predicted 18,123 failed campaigns as successful, giving the model a precision score of approximately 0.56. Figure 15 shows the ROC curve for our model.

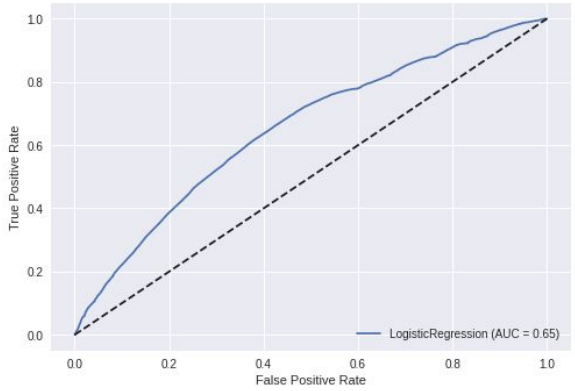

Figure 15.

As previously discussed, we used the coef_ method to extract the coefficients for each of the important features that were used in the final model. From this, we found that 'goal' was the most important feature, with an approximately -5.56 coefficient, indicating a strong negative correlation, compared to the remaining features. The next closest was the 'main_category: Dance' feature, with a positive 1.01 coefficient.

Additionally, even though we dropped backers as an attribute during the EDA process, we decided to try running our initial model with it included once just to gauge its impact. This resulted in an R^2 score exceeding 0.80 which is a significantly better predictive model. This relationship is worth noting. However, since the number of backers a campaign has is not something that is controllable when developing and launching a

campaign, we decided to leave it out of our final model like we had originally planned.

Overall, the model is not as successful as we had hoped it would be at predicting what drives whether a Kickstarter campaign will succeed or not. However, as discussed below, there are some takeaways from our analysis and model that can be used by future campaigns.

5.2 Future Work

Our model showed that the biggest predictor of the success of a Kickstarter was the size of the fundraising goal. An interesting topic to explore in future work would be how chances of success change with the size of the goal, for example is it a linear relationship or is there a threshold at which success rates drop significantly? Is there a range within which success rates are both positive and relatively constant? This information could help users choose the goal that best meets their needs while also having the best chance of success.

We also noted in our EDA that Kickstarters which achieve their goal typically end up exceeding their goal. This again shows that picking a fundraising goal with a high probability of success should be the primary focus. As previously discussed, if a Kickstarter does not meet its goal it receives nothing, however, if it meets it's goal it typically exceeds it. This makes stretch goals an interesting topic for future analysis as well.

Lastly backers, an attribute we decided to exclude from our model, would be worth analyzing further due to its significant impact on success rates. It makes sense that more backers most likely means more money, and therefore a better chance of achieving the Kickstarter's goal. Therefore, an area of research that may be interesting to explore in the future is how does the number of backer's correlate with social media presence. Are most Kickstarters finding backers solely through the Kickstarter website? Or are they attracting backers through social media exposure or other advertising means employed by the campaigns creators?

6 APPLICATIONS

Based on our model and EDA the two attributes that have the biggest impact on success of a Kickstarter are goal and backers. The takeaway from this is that when setting up a Kickstarter the user should select the lowest goal possible that still meets their funding needs. Choosing the lowest goal makes sense because it has a higher success rate and Kickstarter's that meet their goals typically exceed them so it is likely that even more money will be raised. Selecting a higher goal has a higher chance of failure and therefore, the Kickstarter risks receiving no funding at all. Additionally, most pledges are small with a range of approximately $10 to $120 so attracting a high number of backers appears to be a key factor in the success of a Kickstarter campaign. This also shows that Kickstarter is not the place to launch a project if you are looking for an angel investor or funding from a private equity group.

Lastly, category, months, and duration did not seem to significantly impact the outcome of a Kickstarter. Thus, when creating a Kickstarter campaign users should feel relatively confident picking any category, launch month, or duration without it significantly affecting the success of their campaign.

REFERENCES

[1] Daniel Kupka, 2019. “We Analyzed 331,000 Kickstarter Projects. Here’s What We Learned About Kickstarter Success”. *https://medium.com/@daniel.kupka/we-analyzed-331-000-Kickstarter-projects-here-s-what-we-learned-about-crowdfunding-success-63b341b025ac*

[2] Justin S Gage, 2017. “Kickstarter Analysis and Prediction”. *https://www.kaggle.com/gagejustins/Kickstarter-analysis-and-prediction*

[3] Laura Lewis, 2019 “Using Machine Learning To Predict Kickstarter Success”,. https://towardsdatascience.com/using-machine-learning-to-predict-Kickstarter-success-e371ab56a743