Final Project Report

Team Members

Aisha Lalli - alalli@csumb.edu
Michelle Phung - mphung@csumb.edu
Hugo Rodriguez - hrodriguezgarcia@csumb.edu
Jeff Burk - jburk@csumb.edu

Introduction

The project was undertaken to help entrepreneurs and small businesses in making decisions concerning their goals of funding their projects. The Kickstarter platform includes data about various categories of projects to suit projects whether they be artistic, technological, culinary, or literary. This project is important because it could save those entrepreneurs and businesses from wasting time and money on unfruitful ventures.

The research question in this project was whether we could create a model to predict the pledged dollar amount for a given campaign based on the dataset of over 300k completed campaigns. We hypothesized that by using machine learning we could help users decide whether they should start a Kickstarter campaign or if it would be better for them to pursue other more traditional means of funding such as a bank loan or investor funding.

The purpose of the research was to find out if we could predict the outcomes of the Kickstarter campaigns based on the features given in the dataset. This can help determine if Kickstarter is the correct funding platform for a project, and help gain insight into how to optimize a campaign for success.

Selection of Data

This dataset was collected from the Kickstarter platform. The dataset contains information on over 300K Kickstarter projects from the company's launch in 2009 through 2017. The features of the dataset include ID, name, category, main_category, currency, deadline, goal, date launched, pledged amount, state of the project, number of backers, country, pledged amount and pledged amount in USD The dataset includes 15 columns of which 7 are numeric values and over 300k rows. The data was collected by a crowdfunding and data science enthusiast, Mickaël Mouillé, in 2018 and uploaded to Kaggle, an online community of data science enthusiasts. The dataset is currently available for download in Excel CSV format. The data's source URL is https://www.kaggle.com/kemical/kickstarter-projects and the date of download is June 7, 2022.

Methods

In this research, the Kickstarter data was downloaded from Kaggle and uploaded to Github for ease of access. We used Jupyter Notebook to process the data and run our training and tests. We also used Google Slides to create our video content. This research was conducted by a team consisting of Jeff Burk, Aisha Lalli, Michelle Phung, and Hugo Rodriguez. The project leader was Jeff Burk.

Results

Taking a look at our data we can see that the mean number of backers is around 116.5, and the min is 0, 25% of the applicants had 2 backers, 50% of the applicants had 15 backers and 75% had 63 backers. We noticed that the standard deviation was quite high in comparison to the mean at around 965.7 and this is probably due to the max being at 219382 which is skewing the data.
(please check this part highlighted).

We found some missing data in columns “USD pledged” and “name”; we found no need in these columns therefore we dropped them. In processing the data we found that projects could be in several states such as failed, successful, canceled, live, and suspended. We found only use in having failed and successful states; therefore, we dropped live, suspended, and canceled projects.

Using heatmap we saw that the columns 'usd_goal_real' and 'usd_pledged_real' were added by the data collector to account for current currency conversion rates. These columns are highly correlated to 'goal' and 'usd_pledged' and can therefore be dropped. Similarly 'usd_pledged' is highly correlated to 'pledged' and can be dropped. Additionally, the 'currency' column can also be dropped as we have the 'usd_pledged' values. Of the remaining numeric columns, The 'goal' column appears uncorrelated to 'backers' or 'pledged' The number of backers has about a 75% correlation to the amount pledged which is what we would expect. After preprocessing we were left with 10 columns to explore.

In exploring the data we were able to confirm that pledges and backers are correlated in determining the success of projects but we need to further explore which categories are getting the most pledged amounts. We can see that amounts were highest for Design, followed by Games, Technology and Film & Videos. It was also shown that the largest number of backers for a particular project was in Games. Using boxplot, we were able to infer that 2015 had the most amount of projects and the pledged amounts were roughly similar throughout the years. We found that most project campaigns were from the US and that the highest number of campaigns were from FIlm and Video, followed by Music and Publishing. We can see that Design had almost half the amount of campaigns as Film & Videos. The mean amount of dollars pledged for design is 30000 USD; while Film & Video is about 7000. Yet the number of backers was most for Film & Video, Music and Publishing. So, it seems like smaller, more numerous

donations are given to these categories while Design and other Tech fields get less numerous but bigger-sized pledges. Upon looking deeper we can see the subcategories with the highest number of backers per campaign. We see that the number of backers is highest for Tabletop Games, Production Design, and Video Games; with video games having the greatest amounts of projects with a high number of backers. Some of the least backed projects are Latin, Performances, Workshops, Crochet, and numerous other subcategories that have a low number of backers.

In preparing the data for Machine Learning we created a new column for the number of days using columns deadline and launched dates. We also established functions to find the mean percentages of both main_category and category. We also converted the ‘state’ column into numerical form. Our predictors for this research were the features 'goal', 'backers', 'days', 'm_cat_perc', 's_cat_perc', and 'state'. While our target feature was ‘pledged’ since we want to know how much money, if any money, could be raised for a given type of Kickstarter campaign.

After converting categorical features to numeric ones we found that the best test RMSE was about 15% lower when converting the categorical columns manually to percentages versus using the get_dummies() method, with 6 predictors. In setting up a KNN regression model, a random small data selection was extracted from the large data set, to speed up the prediction process. We then scaled the data using these mini data samples for our training and testing sets.

We calculated the baseline performance of our test Root Mean Square Deviation at 66072.9, which we will use as a base comparison with the different machine learning algorithms. Using the KNeighborsRegressor with brute algorithm and default hyperparameters the test RMSE was much lower than the baseline at 52169.5. We tested odd k values from 1 until 29 and found that the best performance came under k= 15 at an RMSE of 41291.7. The training data’s RMSE was higher proportionally to the number of K neighbors. While the test RMSE declined with the increase in K. The Knn Regressor algorithm at k= 15 had better performance than both the baseline and default hyperparameter RMSE.

As for the LinearRegressor algorithm, we found that the slope of all features except for ‘m_cat_perc’ increases as the prediction of ‘pledged’ increases. In scatter plotting the predicted versus actual data, we can see our predictions stray off the actual line, giving us values that are either high above or below what is actual. For instance, at the point towards the far right, a prediction is at around 0.6 but the reality is at 2.0.

Since many predictions are off the line, we improve our results by fitting another model after splitting the training and test data. We find that the training data gives us similar results to our previous model. Next, we predicted our predicted versus actual using our test data and found that there are still many offline predictions (though noticeably less than our previous model). The RMSE using the test data results in an RMSE of 47526.27 which is a 13.118 increase from our lowest RMSE at Knn where

k=15. After scaling the results of our prediction do not change, but our result is still a better performer than our baseline model.

To better improve our model, we tried adding some second-degree polynomial features. Using all our features the new RMSE is 46128.66, and the RMSE of the first feature was way off at 109325.81. Upon experimenting with a different number of features, we found that up to 7 features, there was little improvement in our RMSE; however, using 8 features results in a much lower RMSE value of 46397.6 which is still higher than our best performer with Knn k= 15 at an RMSE of 41291.7. Our RMSE using Decision Tree Regression of 44574.83 did not beat our best performance with Knn k=15.

While testing various hyperparameters with the DecisionTreeRegressor model, adjusting the max_depth had the most impact on the RMSE. Setting the max_depth value to 5 provided the lowest RMSE of 44574.83, this was a big improvement when compared to the default max_depth value RMSE of 65859.07.

Our research question was whether we could create a model to predict the pledged dollar amount for a given campaign based on the dataset of over 300k completed campaigns. Our training data using Knn at k=15 gave us the lowest RMSE and is what we would use for any predictions that are to be made. It does not seem to give us the accuracy we were hoping for, there are many outliers. But if we are comparing it with our baseline model then it is a better model to predict with.

Discussion

It appears that the most funded campaigns are in the tech sector of Design, Games, and Technology. The more Tech the more likelihood of getting backers. It also seems that the tech field gets larger donations so they end up needing fewer backers, while others like the FIlm industry need more backers since the investment is usually much smaller. It makes sense that tech would be more enthusiastically funded because it has the potential to bring more returns that span years over. There is more use in technological things and design is needed for marketing and is essential in the economic sector. We do feel that the models we used did not achieve the goals we wanted. We wanted to be able to predict with better accuracy; however, the Knn model which won in terms of lowest RMSE did not do much more than we could have done while examining the data physically. Which is that tech does better and they have the greatest success, followed by the movie industry. Perhaps there is more that we should learn about in terms of machine learning that would help us reach our goals of being super helpful to those who consult the prediction of the Kickstarter campaign data.

Summary

While testing various hyperparameters with the three sets of algorithmic models, we optimized the RMSE to well below the baseline value. However, the lowest RMSE

achieved did not provide the accurate prediction we were looking for. We would consider exploring more complex models or try using another Kickstarter dataset with additional features to achieve a more accurate test RMSE. I would not say that we could not help in our prediction because there were clusters on the graphs that were pretty accurate however it is far from foolproof.