Kickstarter Success Prediction

Team 45: Joshua Ng, Kayla Hunt, Minjae Lee, Youngho Lim, Karishma Rana

Project Proposal video:

Youtube video

Introduction:

Since its inception in 2009, Kickstarter has been a popular way for entrepreneurs and creators to raise funds for their project. Anyone can become a backer for a project, pledging any amount of money and only paying if the project is deemed successful. Success on Kickstarter is defined as reaching the goal pledge amount within a predetermined time frame, at which point funds are collected and given to the creators. If the goal amount is not reached, the project is considered a failure. The current success rate of projects is relatively low, with only 44 percent of Kickstarter projects being considered successful (Yuan et al.). The most important information provided by our dataset is the category of the project, fundraising goal, how long the campaign was open, number of backers, amount pledged, country pledged from, and the outcome of the project.

Problem Definition:

What is our motivation behind this project?

The goal of our project is to predict the future success of a project put on Kickstarter. The motivation for this project is the number of projects that creators try to fund, with low levels of success. If we can accurately predict the success or failure of a project given how well they crowdfund (amount of backers, amount pledged, etc.), we can help creators to decide to pursue the project before the time, effort, and money is expended to create the campaign.

Our Dataset:

Kickstarter Projects

Data Features:

- Kickstarter ID
- Kickstarter name
- Category
- Main_category
- Currency
- Deadline
- Goal
- Pledged
- State

Dataset visualization

Pictured below are the distribution of our various categories.

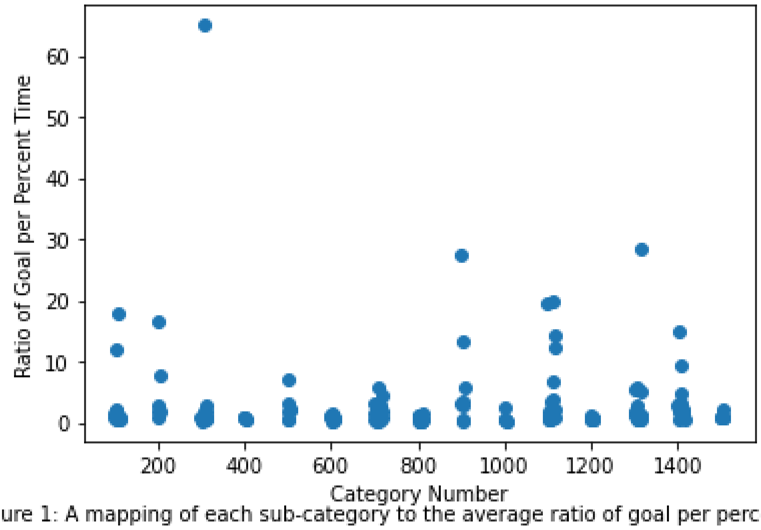


Figure 1: A mapping of each sub-category to the average ratio of goal per percent time

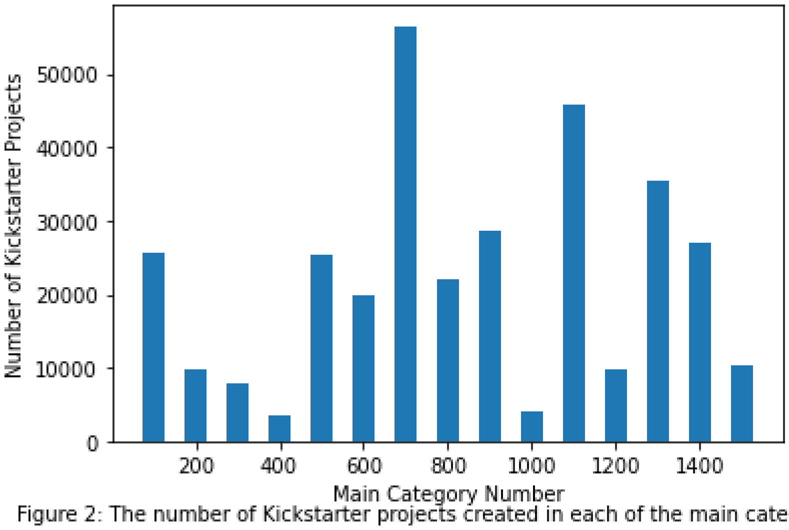


Figure 2: The number of Kickstarter projects created in each of the main categories.

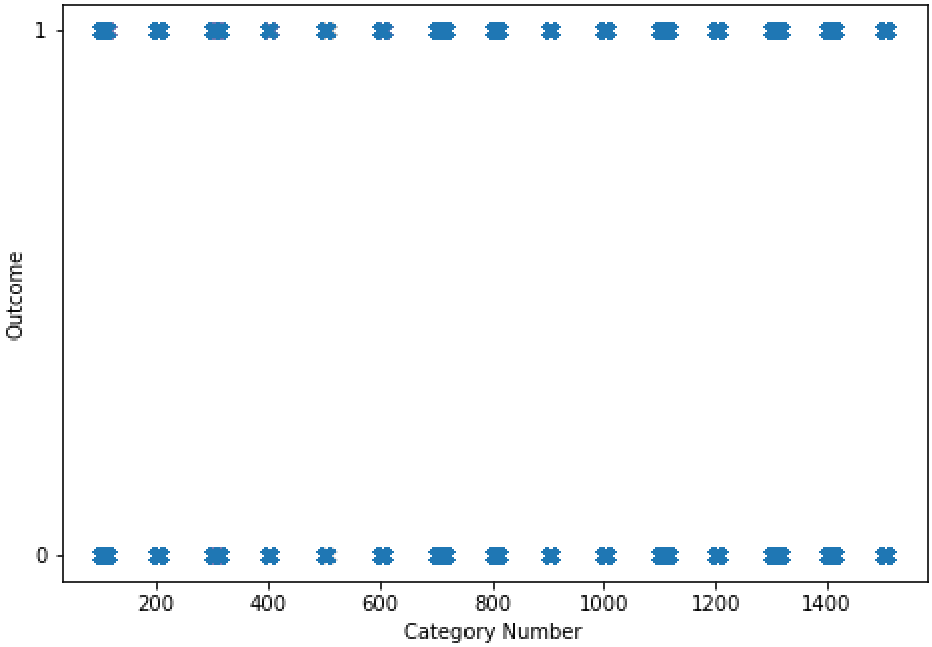


Figure 3: A mapping of Category Number to the campaign outcome

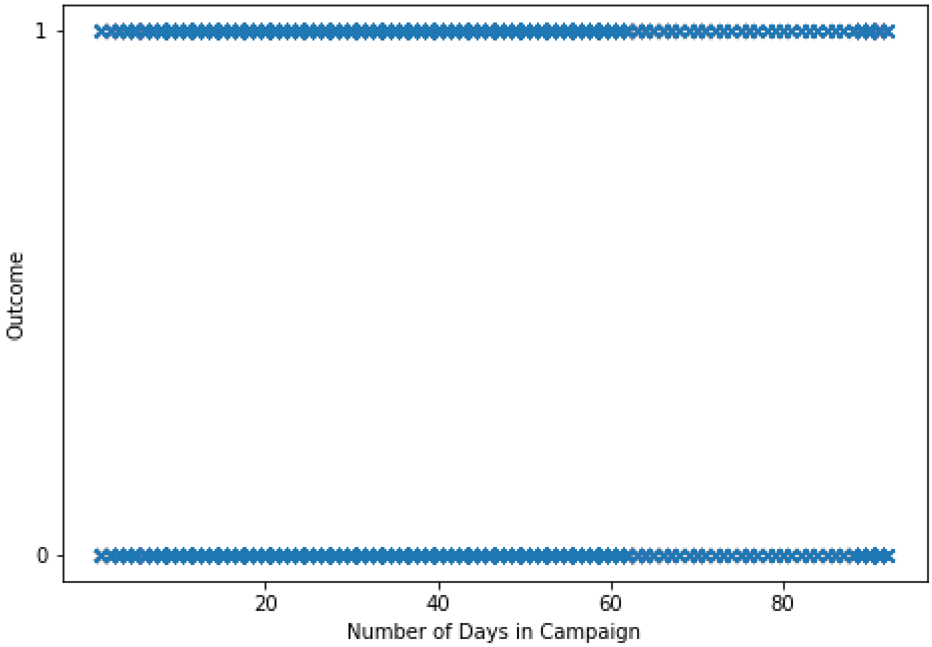


Figure 4: A mapping of Number of Days in Campaign to the campaign outcome

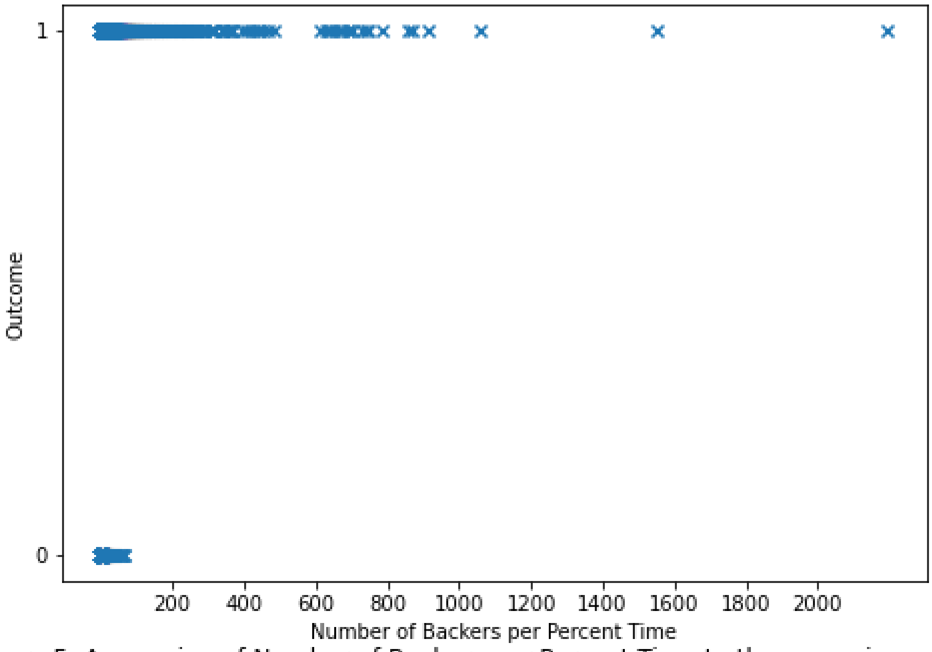


Figure 5: A mapping of Number of Backers per Percent Time to the campaign outcome

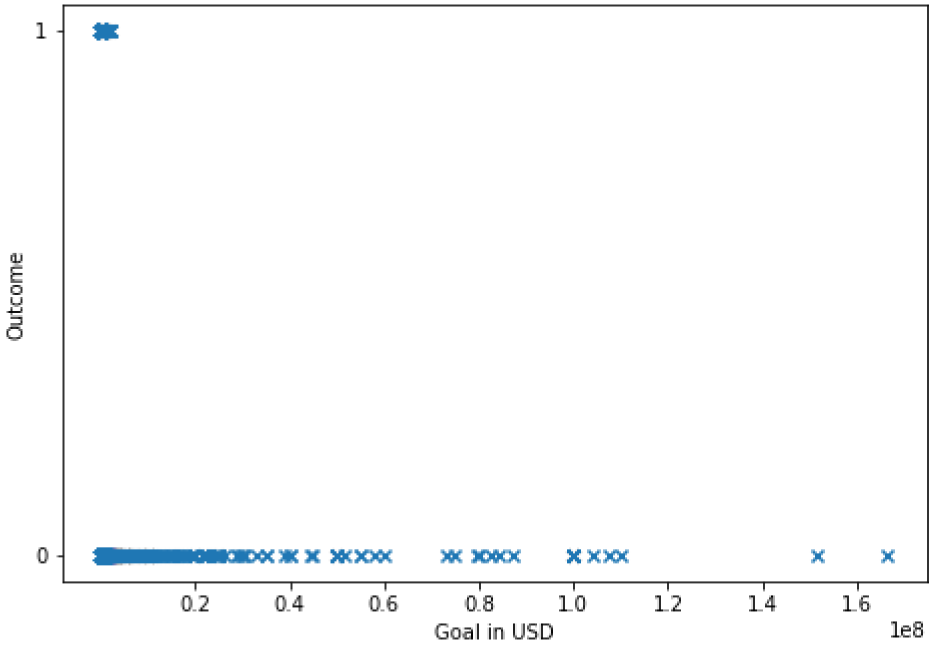


Figure 6: A mapping of Goal in USD to the campaign outcome

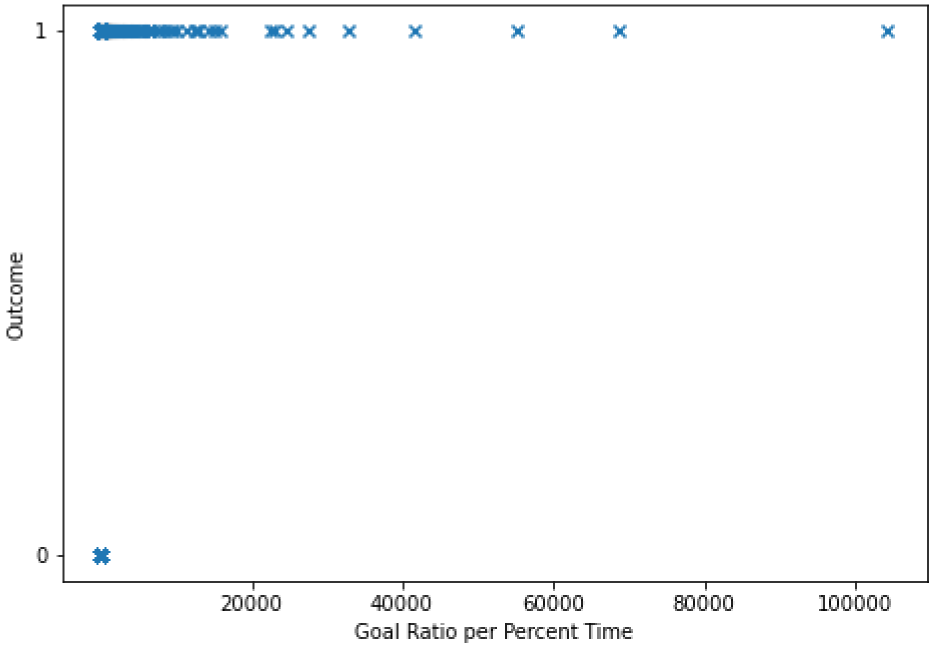


Figure 7: A mapping of Goal Ratio per Percent Time to the campaign outcome

There are 31 columns total, 12 string features, 8 decimal features, 4 date/time features, and 7 other.

Methods

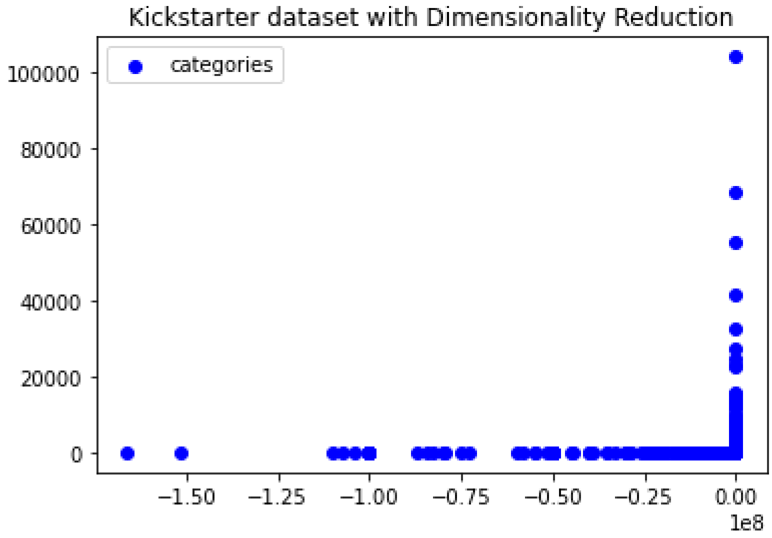
Currently, the algorithms that we have contemplated using to solve this problem are some combination of Naive Bayes, Logistic Regression, and Random Forests. All three of these potential algorithms are designed for supervised learning and classification problems, the type of problem that our dataset falls under. Additionally, we can utilize NLP algorithms like Neural Bag of Words on certain features such as project titles and categories to further help predict the projects' chances of success. Of course, the chosen algorithms are subject to change in the future, especially since the lectures have yet to cover supervised learning.

Predicted Results

Ideally, the result would be that the models we create would correctly label the data more than half of the time. That being said, based on the third-party research we conducted we found that previously created models averaged around a 65% to 75% accuracy (Chen et al.) (Greenburg et al.) (Yuan et al.). While it is worth noting that these models utilized different features than the ones our chosen dataset affords us, we still anticipate that creating a model with a similar accuracy will come with its challenges.

Feature Reduction (PCA)

Upon analyzing our data, there were some category numbers that did not have an intrinsic relationship with the success of the Kickstarter. To first approach our data, we first applied principal componenet analysis (PCA) to reduce our dataset. We reduced the dataset to 3 features from a total of 31 features and retained a 94% variance. Our new labelled categories and features can be foudn in the legend.txt file.



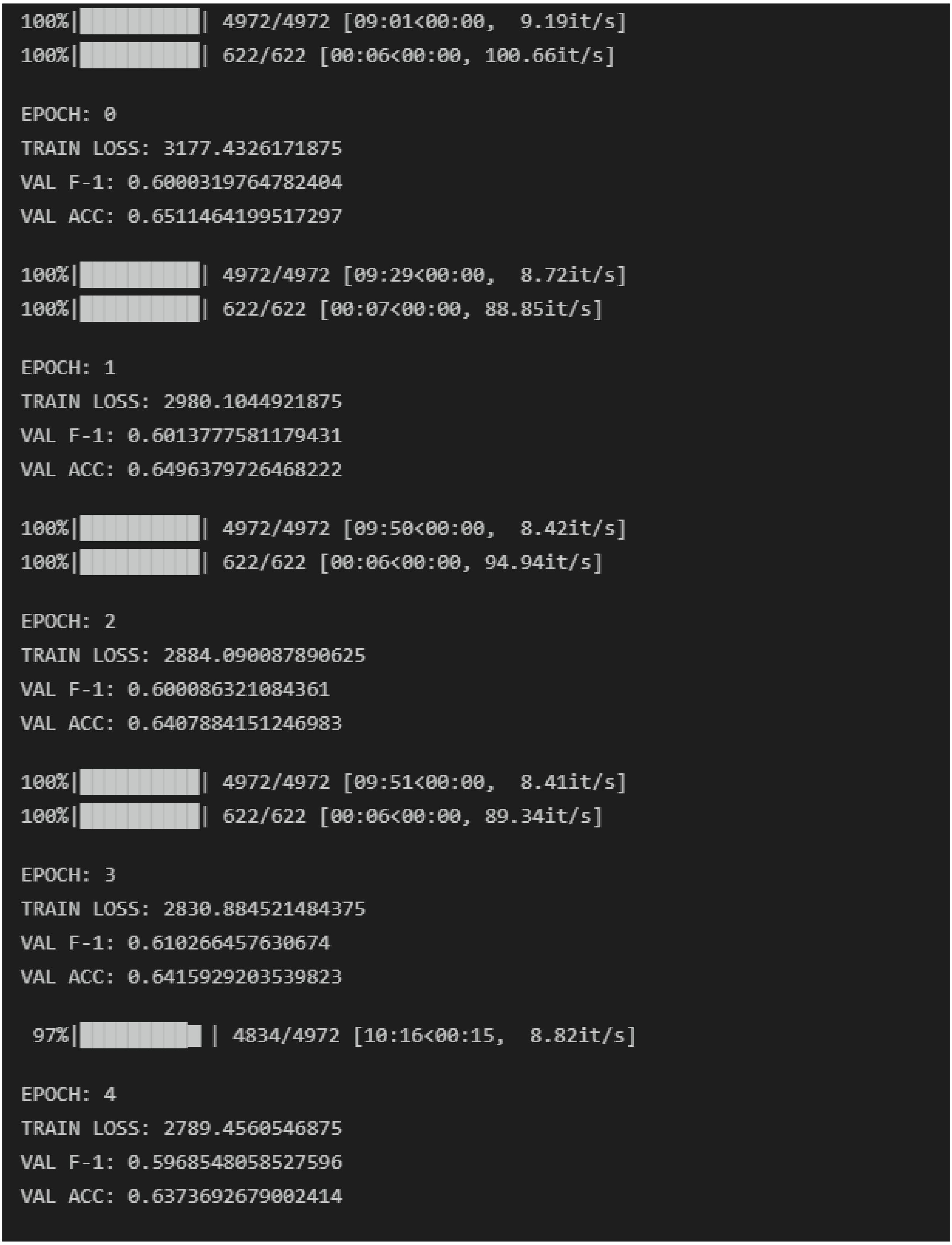
The different groups of datapoints on the scatterplot represent the different grouping of similar features. There is difficulty distinguishing between different category numbers as well as USD for each goal, so there is not a clear trend for our reduced data.

Naive Bayes

After reducing our data dimensions and training the model, we end up with an accuracy of 59.2%. This less than ideal accuracy is a result of the shortcomings of PCA. Typicaly , PCA does not accept categorical data. We had to take out the categorical data, run PCA, and add the catogrical data back to our reduced dimensions, which could explain the reduction in accuracy. PCA also has difficulty determining order if variables are not correlated. I believe that there is not a strong correlation between our data for USD raised and category, so PCA was unable to determine a correlation. Because this was the case, PCA ordered our variables according to variances, which heavily affected our accuracy.

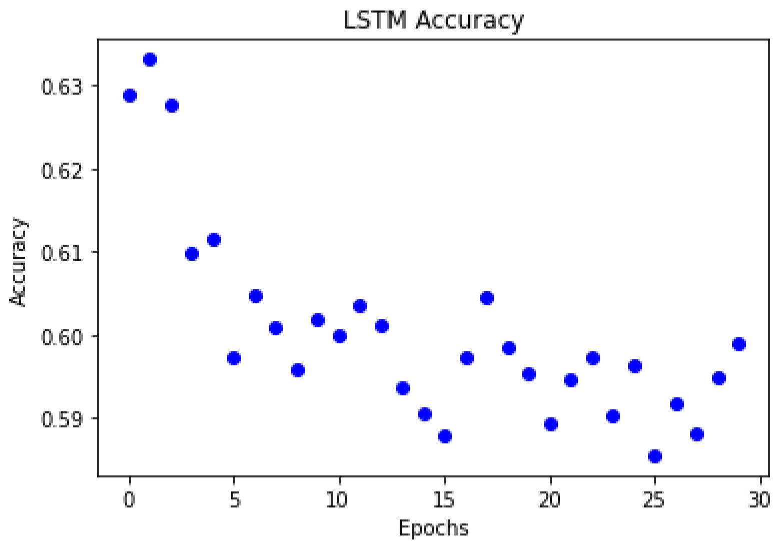
Natrual Language Processing

In addition to Naive Bayes, we implemented the NLP Bag of Words approach on our list of Kickstarter titles and categories. Initially training the model with 20% of the dataset, we get a result of 64% accuracy and an F-score of 0.6. However, when we increase the size of the dataset we are sampling, as well as use better hardware, our accuracy increases. I tested with 30% of the dataset, and as seen below, we yield a greater F-1 test and accuracy.



```
100%|██████████| 4972/4972 [09:01<00:00,  9.19it/s]
100%|██████████| 622/622 [00:06<00:00, 100.66it/s]

EPOCH: 0
TRAIN LOSS: 3177.4326171875
VAL F-1: 0.6000319764782404
VAL ACC: 0.6511464199517297

100%|██████████| 4972/4972 [09:29<00:00,  8.72it/s]
100%|██████████| 622/622 [00:07<00:00, 88.85it/s]

EPOCH: 1
TRAIN LOSS: 2980.1044921875
VAL F-1: 0.6013777581179431
VAL ACC: 0.6496379726468222

100%|██████████| 4972/4972 [09:50<00:00,  8.42it/s]
100%|██████████| 622/622 [00:06<00:00, 94.94it/s]

EPOCH: 2
TRAIN LOSS: 2884.090087890625
VAL F-1: 0.600086321084361
VAL ACC: 0.6407884151246983

100%|██████████| 4972/4972 [09:51<00:00,  8.41it/s]
100%|██████████| 622/622 [00:06<00:00, 89.34it/s]

EPOCH: 3
TRAIN LOSS: 2830.884521484375
VAL F-1: 0.610266457630674
VAL ACC: 0.6415929203539823

 97%|█████████▋| 4834/4972 [10:16<00:15,  8.82it/s]

EPOCH: 4
TRAIN LOSS: 2789.4560546875
VAL F-1: 0.5968548058527596
VAL ACC: 0.6373692679002414
```

The final approach we used was long term short memory. The difference between bag of words and LSTM is that there are feedback connections present. After running 30 epochs, our results are shown below.



The drop in accuracy can be explained by the nature of our data: the categories make it difficult to differeniate data and approach it with either bag of words or LSTM.

References

- Yuan, Hui, et al. “The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach.” Decision Support Systems, North-Holland, 6 Aug. 2016, https://www.sciencedirect.com/science/article/pii/S0167923616301373.
- Chen, Kevin, et al. “Courses.cms.caltech.edu.” KickPredict: Predicting Kickstarter Success, http://courses.cms.caltech.edu/cs145/2013/blue.pdf.
- Greenberg, Michael D, et al. “Crowdfunding Support Tools: Chi '13 Extended Abstracts on Human Factors in Computing Systems.” Crowdfunding Support Tools: Predicting Success & Failure, 1 Apr. 2013, https://dl.acm.org/doi/pdf/10.1145/2468356.2468682.