Kickstarter Project Report

Team: Group 26
Team Members: Amal Byju, Sreyas Sourav, Raghul Balamurugan

TABLE OF CONTENTS

Executive Summary:

Kickstarter.com is a crowdfunding platform that assists fundraisers in raising funds for their projects. It's where creators share new visions for creative work with the communities that will come together to fund them. Kickstarter's primary mission is to bring projects to life by providing resources and tools to the people. It is an organization concerned about artists and creative people having their talent and creativity come to reality.

We believe that one of the tools and resources that this organization can provide for the creative community is to help them understand how to successfully crowdfund their projects by being aware of the factors that will help them succeed. This is possible by utilizing predictive modeling in this business scenario. We have existing data that reflects several variables related to individual project fundraising. Among other things we have information about each project's creation and launch date, the project creator's profile, each project's description, and the fundraising goal. On top of that, we have three important variables that are meaningful to kickstarter's business case and also help the organization to give informative ideas for the fundraisers. The first one is the "success" variable which tells us whether a project fundraised more than the goal. The second variable is "backers_count" which contains the number of backers each project attracts. Finally, we have the variable "big_hit" which shows us whether a project earned a lot more money than the goal. These three variables are instrumental for our analysis and prediction.

In addition, we saw it fit to gather external data like the location of the project because we found that the actual location of the project doesn't necessarily correspond to the location of the user.
By using the above-mentioned variables as well as adding new ones, we want to assist Kickstarter to achieve its mission by providing data-backed information about the important factors that contribute to the success of a project as well as provide a validated model which predicts whether fundraising will succeed based on its existing profile.

Cleaning & Feature Engineering:

Existing Features/Target Variable	Task	Reason
success	Replace 'YES' with '1' & 'NO' with '0'	Target Variable - 0/1 Label encoding for modeling
region	Factorize	To understand the trend of which region has the most and least number of projects
category_parent	Factorize	To understand the trend of which region has the most and least number of projects
location_slug	Factorize, Grouping	Categorized locations (with <1500 projects) as 'Other Location Slug' - removing less significant locations for better model performance
category_name	Factorize, Grouping	Categorized category_name (with <3000 projects) as 'Other Category' - removing less significant categories for better model performance
deadline, created_at, launched_at	Change to 'date' type	Appropriate data type for modeling To calculate new dependent features like 'time_gap'
contains_youtube	Factorize	Categorical feature with 0's, 1's

New Features	Task	Reason
time_gap Time interval from project launch date to deadline	Find duration between deadline and launced_at	Numeric variable in days and it will be a possible predictor as the p-value was < 2e-16 in our logistic model
afinn_overall	Find difference between afinn_pos - afinn_neg	To find whether the entire description is positive or

		negative
extra_female_creators	Find difference between female_creator - male_creator	To find whether the creators are female-dominant or male-dominant
num_rewards	Parse reward_amounts column to count the number of rewards	To find the number of rewards for each project

Accuracy of the logistic model without the above new features was 0.6734

model_formula<-

success~goal+category_parent+region+ADV+VERB+CONJ+ADJ+avgsentencelength+smiling_creator+goal*region+goal*category_parent+contains_youtube+numfaces_project+location_slug+category_name+maxage_creator+minage_creator

However, after adding the new features the accuracy increased to **0.7078** with significant p-values.

model_formula <- success ~

goal+region+category_parent+num_words+grade_level+minage_creator+avg_wordlengths+**time_gap**+location_slug+category_name+**afinn_overall**+contains_youtube+avgsentencelength+numfaces_creator+sentence_counter+ADV+NOUN+ADP+PRT+DET+VERB+CONJ+**num_rewards**+launched_at+**extra_female_creators**+numfaces_project+minage_project+smiling_creator

To further improve the accuracy, we added more calculated fields and interactions between the terms. Interaction terms were chosen based on whether they improved the accuracy of the model or not. 'num_rewards' improved the accuracy of our model so we created the following new features based on reward_amounts again.

min_reward	Find the minimum reward from	To find whether projects are

	reward_amounts	more successful when the minimum reward amount is low.
max_reward	Find the maximum reward from reward_amounts	To find whether projects are more successful when the maximum reward amount is high.
sd_reward	Find standard deviation between the reward_amounts	To find whether projects are more successful when there are bigger gaps between reward amounts.
proj_duration	Find difference between launched_at - created_at	To find the estimated duration of the project in days

proj_duration:goal	Interaction between project duration and goal
category_parent:goal	Interaction between parent category and goal
maxage_creator:goal	Interaction between the maximum age among the creators and goal
proj_duration:numfaces_creator	Interaction between the project duration and the number of creators
num_words:proj_duration	Interaction between the number of words and the duration of the project
contains_youtube:goal	Interaction between the goal and whether there is YouTube content
goal:max_reward	Interaction between the maximum reward amount and goal
grade_level:maxage_creator	Interaction between the age of the oldest creator and the grade level of the description
afinn_overall:category_parent	Interaction between how positive the description is with the parent category

grade_level:category_parent	Interaction between the grade level of the description and the parent category

Interaction terms were added based on whether they improved the accuracy of the model or not.

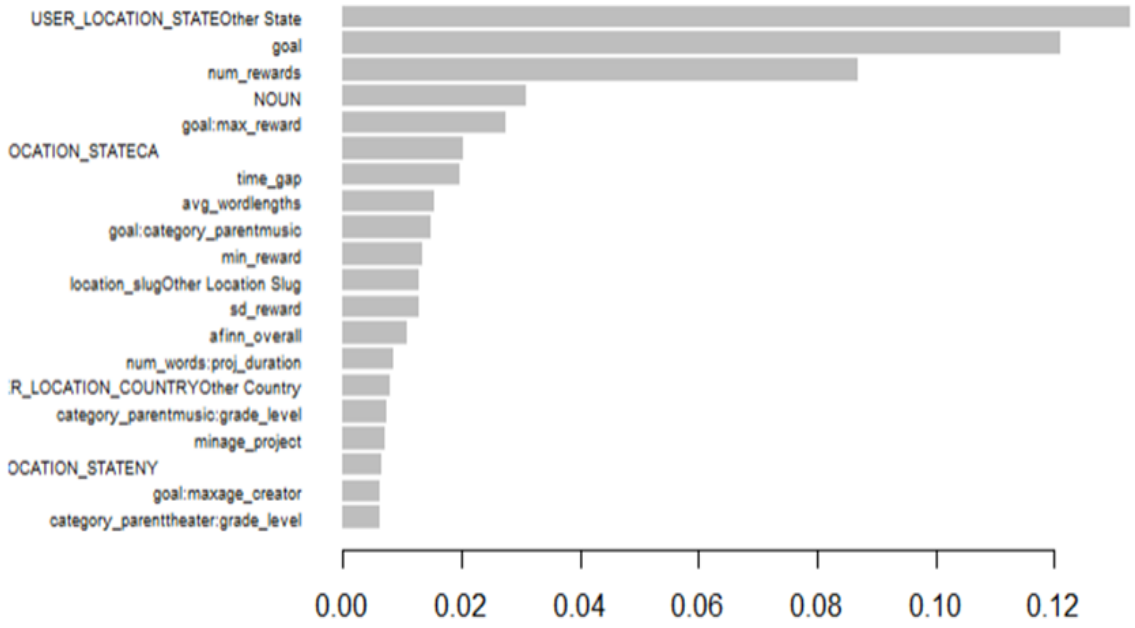
We also extracted another external dataset **(https://www.icpsr.umich.edu/web/ICPSR/studies/38050/datadocumentation)** to get 'USER_LOCATION_COUNTRY' and 'USER_LOCATION_STATE' which was merged to the kickstarter data on 'Project ID (PID)'. We found that this is an important feature since the actual location of the project doesn't correspond to the location of the user. Substantiating that, we observed almost 80k projects from the US, of which 24k projects were from California.

USER_LOCATION_COUNTRY	Factorize, Grouping	Categorized countries (with <10 projects) and null/missing values as 'Other Country' for better model performance
USER_LOCATION_STATE	Factorize, Grouping	Categorized states (with <10 projects) and null/missing values as 'Other State' for better model performance

Finally, we decided to perform text mining on five main columns - 'reward_descriptions', 'name' and 'blurb', 'captions' and 'tag_names'. In this step, we added unigrams, bigrams as evidence and performed stemming, removing punctuations/numbers. We also manually created a stop word list for each column vocabulary, set term_count_min = 20, doc_count_min = 10 to prune the data from more commonly occurring stop words and less frequently occurring terms, thereby preventing the model from overfitting.

We then created a DTM matrix for each of these pieces of evidence and added it to our data for model evaluation and selection which is our next step.

Top 20 features with their importance



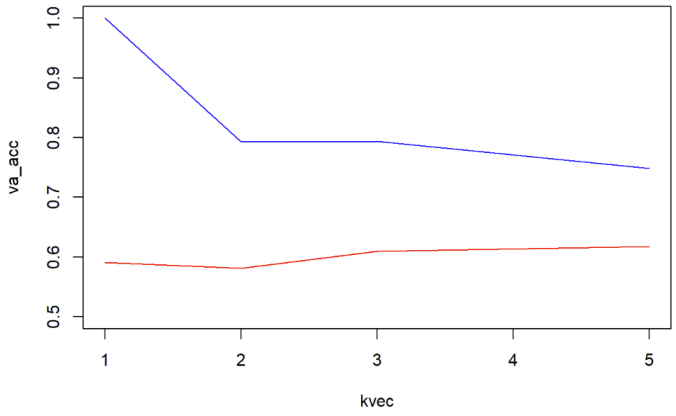
Attached is the entire csv with features and gains.

Model Evaluation & Selection Methodology:

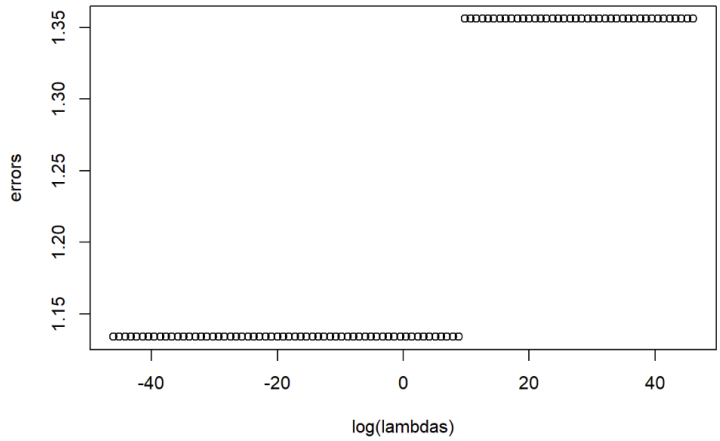
Model Name	Accuracy
Logistic regression	0.7305824
KNN with k= 5	0.6179087
Lasso with 5-fold cross validation	0.6851434
Ridge with 5-fold cross validation	0.685246
XG Boost	0.7846438

Fitting curves for different models we tried (Red - Validation, Blue - Training)

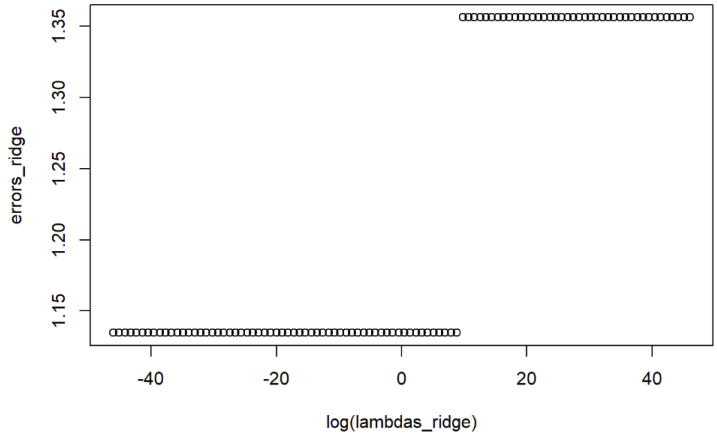
1) **KNN**



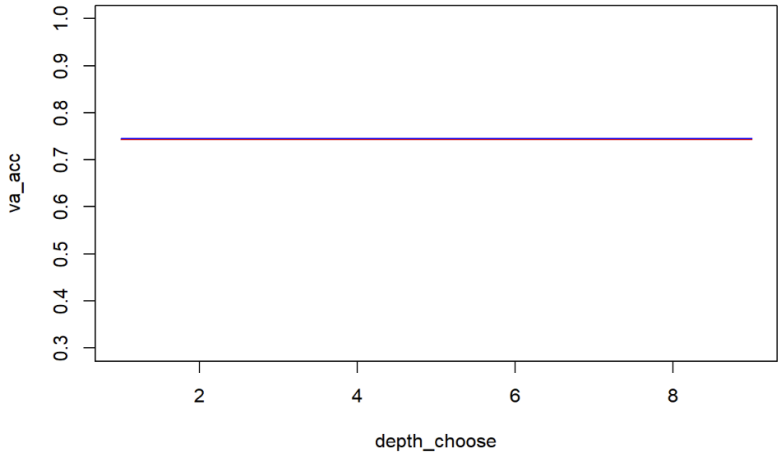
2) Lasso

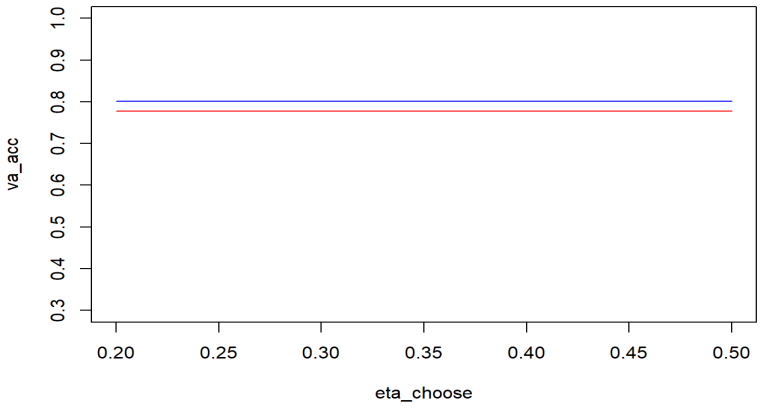


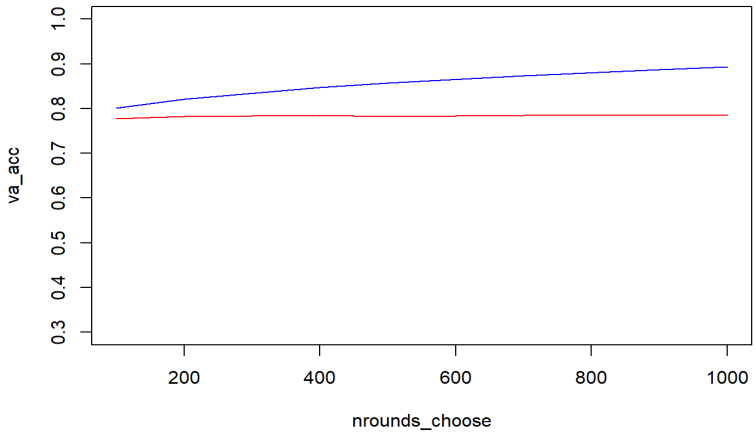
3) Ridge



4) XG boost (Parameters depth - 4 , eta - 0.2 , n_rounds - 900)







As described in the above table, XG boosting has the highest accuracy as well as the best AUC performance with 0.8679449.

Conclusion:

After experimenting with different models like logistic regression, KNN, Lasso and ridge, and finally XG boosting, we were finally able to reach a maximum accuracy of 0.7846438 with XG boosting. This model is very helpful for Kickstarter to make predictions for fundraising campaigns. On top of that we have found out the top three factors that have more impact in predicting success of a fundraising campaign are the user's location or state, goal of the campaign in US$ and number of rewards offered by Kickstarter.

All these efforts were possible due to our group's hard work and close coordination. We did incredibly well at delegating tasks effectively. The project was underway from the first week of starting the project. On top of this, we split the work evenly and attended meetings regularly. Each member contributed to the final model and helped in the report building process. Some of our challenges we faced in this process were the size of the dataset, cleaning and categorizing them. We believe we have done a good job but if there was one thing we would have done differently, it would have been to try to find more features before hyperparameter tuning. And, if we had more time to work on this project, more features would have been added and also tried more models and tuned more hyperparameters.

Finally, we recommend other students to begin working on this project early, analyze the dataset a bit more and perhaps experiment with unsupervised modeling as well.

Appendix (Member's contribution)

S.No.	Team Member	Contribution
1.	Amal Byju	● Added interaction variables in feature engineering. ● Wrote the code for the linear regression model. ● Wrote some text mining code for the XGBoost model.

2.	Raghul Balamurugan	● Added interaction variables in feature engineering. ● Wrote the code for ensemble, logistic regression models. ● Completed the final report.
3.	Sreyas Sourav	● Wrote the code for the KNN, Ridge and Lasso. ● Worked on the R markdown file and the report. ● Tuned the hyperparameters for each model.

References

https://www.kickstarter.com/charter?ref=hello

Ref: External Data-Set

https://www.icpsr.umich.edu/web/ICPSR/studies/38050/datadocumentation