1. Introduction

Through the use of a website called Kickstarter, people and organisations may generate money for various endeavours, such developing a new good or service, by asking many people for donations. Giving supporters reward such as a working prototype of the product under development or a limited-edition of the final product is a common practise in Kickstarter projects. A well-liked platform for artists, inventors, and entrepreneurs to seek capital for their ideas, Kickstarter is renowned for assisting in the launch of several creative and inventive projects.

2. Data Preprocessing

The data pre-processing was started by understanding the meaning of different attributes and the relationships among them. For the purpose of implementing classification techniques, new attributes were created that would increase the accuracy of making the right prediction. The tasks performed were:

1. Creating a new attribute NewGoal (NewGoal = goal*static_usd_rate): This was done to make sure that the difference of currency used in different projects would not impact the true value of amount asked.

2. Creating a new attribute launchTIme (launchTime = (launched_at_month-1)*30+launched_at_day) : This will track on which day of a year the project was launched. The main intention behind it is to track the seasonality throughout the year that can influence the chances of a project's success.

3. NULL values in category were replaced with "others". No significant change was observed when we removed them, so we kept them.

4. Rows with NULL values from all other attributes were removed.

5. Only projects with state = 'failed' or 'successful' were kept as mentioned in the project.

Only those attributes that were available at the time of starting the project were considered for prediction. all others (eg No. of backers) were ignored.

The attributes used for testing different models were : static_usd_rate,name_ len_clean,create_to_launch_days, launch_to_deadline_days ,NewGoal ,launchTime,launched_at_yr,category (Dummified)

3. Classification Model

After reviewing the various classification models available, testing the dataset, and reviewing the preliminary results generated, the gradient-boosted model was chosen because it performed the best. The reasoning behind it could be that they are more accurate than traditional decision trees, can handle complex relationships between input and output variables, and are robust to outliers.

The process of building and selecting the model is as follows:

1) Split the data into testing and training set (80/20 split was performed)
2) Build the final models using the attributes that were decided by various parameter tuning methods and checking the feature importance of different attributes used to make a better model.

Accuracy of gradient boosting: 0.7428571428571429
Classification Report:

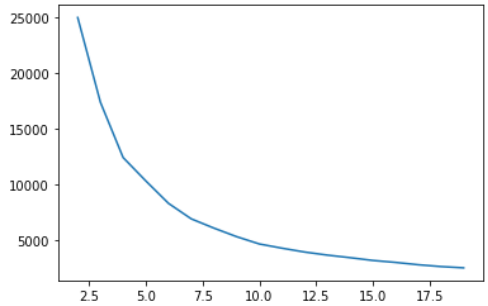
	precision	recall	f1-score	support
failed	0.76	0.87	0.81	222
successful	0.70	0.52	0.59	128
accuracy			0.74	350
macro avg	0.73	0.69	0.70	350
weighted avg	0.74	0.74	0.73	350

Based on the results obtained, the gradient boosting model had better performance and was selected as the final model. The results are mentioned on right.

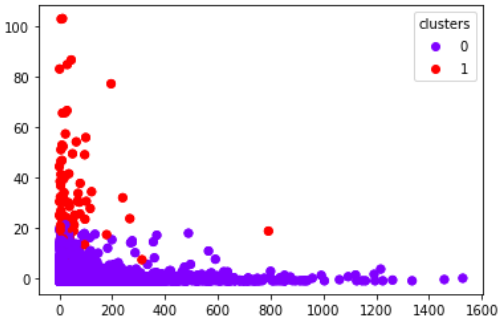
4. Clustering Model

KKN was used to find hidden patterns in data. Because KKN doesn't work well on binary values, finding relation between successful and other attributes was difficult.

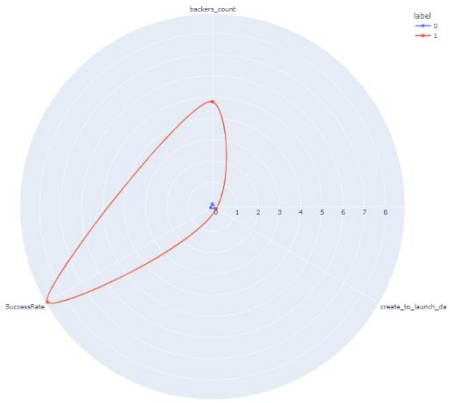
So we created a new attribute ‘SuccessRate’. which is the relative value of (pledged - goal)/goal. Its significance is the that now we don’t have to deal with categorical variable 'state' and its amount is -ve if the project failed and + if the project was a success and the value about 1 will tell by how many times the project was able to outdo the goal, therefore showing the scope of success of project



An attempt to find relation between ‘SuccessRate’,’ NewGoal’ and ‘NewGoal’ was done. To remove outliers in these attributes ,Box plot was created and values of each attribute that were outside of 3 standard deviations were removed. By Creating Elbow Method Graph and Checking silhouette_score (0.9141418), No. of cluster was decided as 2.



By Looking at Graph in the right (It Shows the relation between Create to launch (X-axis) and Success Rate (Y axis), we can conclude that projects which gave return of more than 20 times were generally created in less time (less than 200) (But also main of them failed). By looking at the Diagram on the bottom right ,We can also conclude that there is a good correlation between Success rate and backers count.



The Median of Attributes in each cluster and their cluster center as mentioned below.

We can see in cluster one that SuccessRate is higher as no of backers is higher than the backers of cluster zero and this can also be attributed to the fact that create_to_launch days were higher than in other cluster.

	Cluster	create_to_launch_days	SuccessRate	backers_count
Median	0	13	-0.9278	12
Median	1	21.5	27.248791	2857
Cluster Center	0	-0.000355	-0.064415	-0.034903
Cluster Center	1	0.048684	8.824166	4.781367