McGill

Introduction

This report encapsulates the findings and methodologies applied in analyzing Kickstarter campaign data. The objective was to identify key factors influencing the success of campaigns and segment them into meaningful clusters.

Preprocessing

The process began with the removal of rows containing missing values and entries with 'canceled' and 'suspended' statuses were filtered out. Also, all the variables with multicollinearity, are irrelevant and had little impact on the model were removed. The variable "goal" was changed to USD for uniformity. The dataset was then scaled using StandardScaler to normalize the distribution of numeric features. Categorical variables were transformed into numerical values through one-hot encoding, allowing the models to process and learn from these attributes effectively. Finally, an Isolation Forest algorithm was applied to detect and exclude outliers. The same preprocessing was performed on the grading data again to test the model.

Classification Model

The classification models were the evaluated and compared to predict the success of Kickstarter campaigns. The logistic regression, decision tree classifier, random forest classifier and gradient boosting classifier were tested to see the best model. The performance was evaluated for accuracy, precision, recall, and F1-score and the three models with the highest accuracy were the logistical, random forest and gradient boosting with accuracy around 76%. However, logistic regression demonstrated better adaptability on an external dataset, highlighting its utility in broader applications and giving around 75% accuracy score (almost same like with the training data) compared to around 60 % for the other models. This can be explained by its good balance of bias and variance compared to the overfitting done by the other models on the training set. Therefore, the logistical model is to be used to test the grading sample.

The features (feature importance) that affected the models the were also evaluated and can be seen in the appendix with goal being the highest followed by categories (Appendix figure 1).

Clustering Model

For the clustering, the data underwent a similar preprocessing routine. Numeric features were normalized using MinMaxScaler (it proved to be better than standard scalar in the silhouette score). Categorical attributes were dummified. A few extra variables were added compared to the classification model such as number of backers, dollars pledged and state as these are important to know what differentiated the project when clustering.

KMeans was the primary algorithm implemented to cluster the data. The optimal number of clusters was determined using the Elbow Method and the Silhouette Method (Appendix figures 2 and 3), which assesses cluster cohesion and separation. It was found to be 6 clusters with a silhouette score around 0.4 which is reasonably okay as more clusters will be hard to interpret the characteristics. Additionally, DBSCAN was employed as a complementary approach to KMeans. A PCA was utilized to facilitate the visualization and interpretation of the high-dimensional clustering outcomes allowing for the clear visualization of the distinct clusters formed by the KMeans and DBSCAN algorithms. As seen in figure 4 and 5 in the appendix, the PCA analysis from the KMeans shows better clustering for each cluster than the DBSCAN which mixes 2 clusters in1 so will use the KMeans cluster to analyze.

Visualizing the cluster info as found in table 1 in the appendix and the code we can get an idea what each cluster had. Cluster 0 represents a successful group, primarily in the 'Hardware' category, with an average pledged amount of $104,000 and a medium goal of around $34,000. The campaigns in this cluster are well-supported and tend to have a longer preparation phase, averaging 65 days from creation to launch meaning time was taken to prepare the product and not rushing it.

In contrast, Clusters 1, 2, 3, and 5 mainly consist of projects that didn't reach their funding targets, spread across categories like 'Web', 'Software', and 'Hardware'. Cluster 1, focusing on 'Web' projects, has a lower average backers count and sets a very high goal of around $130,000, reflecting the high monetary demand these projects need. Similarly, Clusters 3 and 5, with 'Software' and 'Hardware' projects, set high goals of $110,000 and $150,000 but face difficulties in fundraising as people might hesitate to invest when they see a very high goal. Also, Cluster 2, despite its size and category diversity, struggles to attract significant support as it also had a high average goal of $90,000.

On the other hand, Cluster 4 shows a pattern of success in almost all categories, achieving its funding goals with solid backing from the community. This cluster's success is partly due to its reasonable average goal of $16,000, indicating a well-calibrated approach to setting funding targets.

Even though clusters 2,3 and 5 had a very similar time to prepare the to launch the product (around 40 days) the high goal of the failed cluster group discouraged people from investing into the project compared to clusters 0 and 4 which had much more modest goal. And this can be shown when comparing clusters 0 and 5 which both are exclusively composed of the category hardware but cluster 0 having a much humbler goal.

Conclusion

In conclusion, the analysis of Kickstarter campaigns demonstrates the critical influence of setting realistic funding goals on campaign success. While clusters with higher funding targets generally struggled to attract sufficient support, those with more modest goals saw higher success rates. The Logistic Regression model, offering the right balance between complexity and interpretability, proved effective in predicting campaign outcomes. These insights can guide future campaign strategies, emphasizing the importance of careful goal setting and thorough preparation in the crowdfunding domain.

Appendix

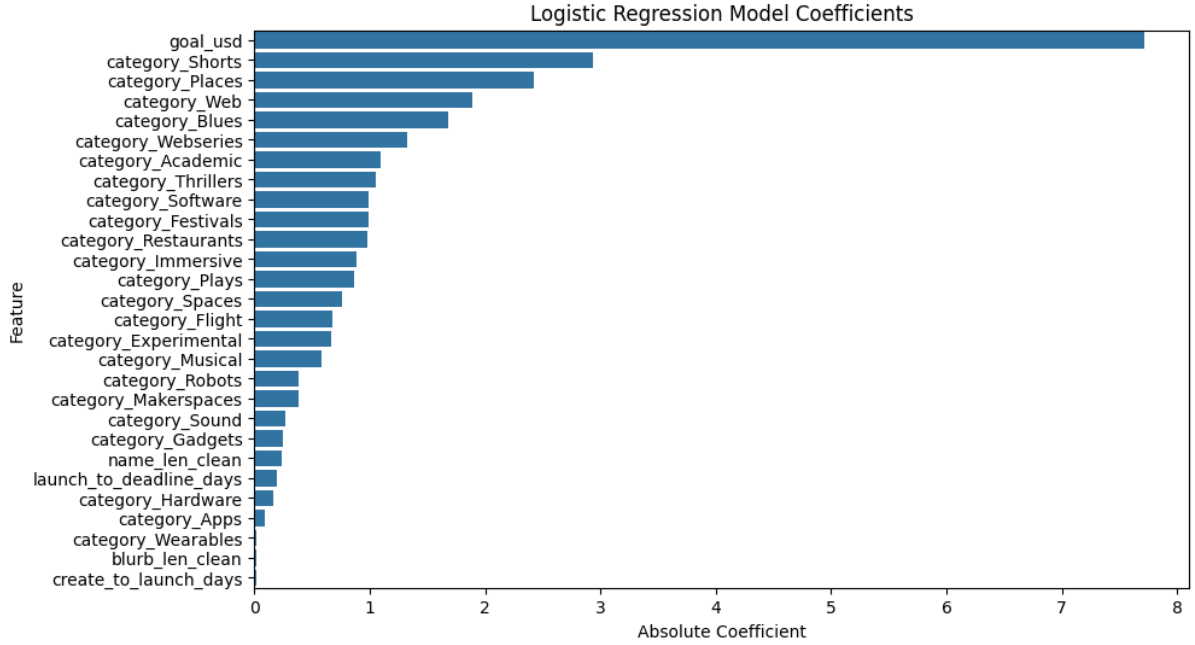

Figure 1 Logistic Regression Model Coefficients

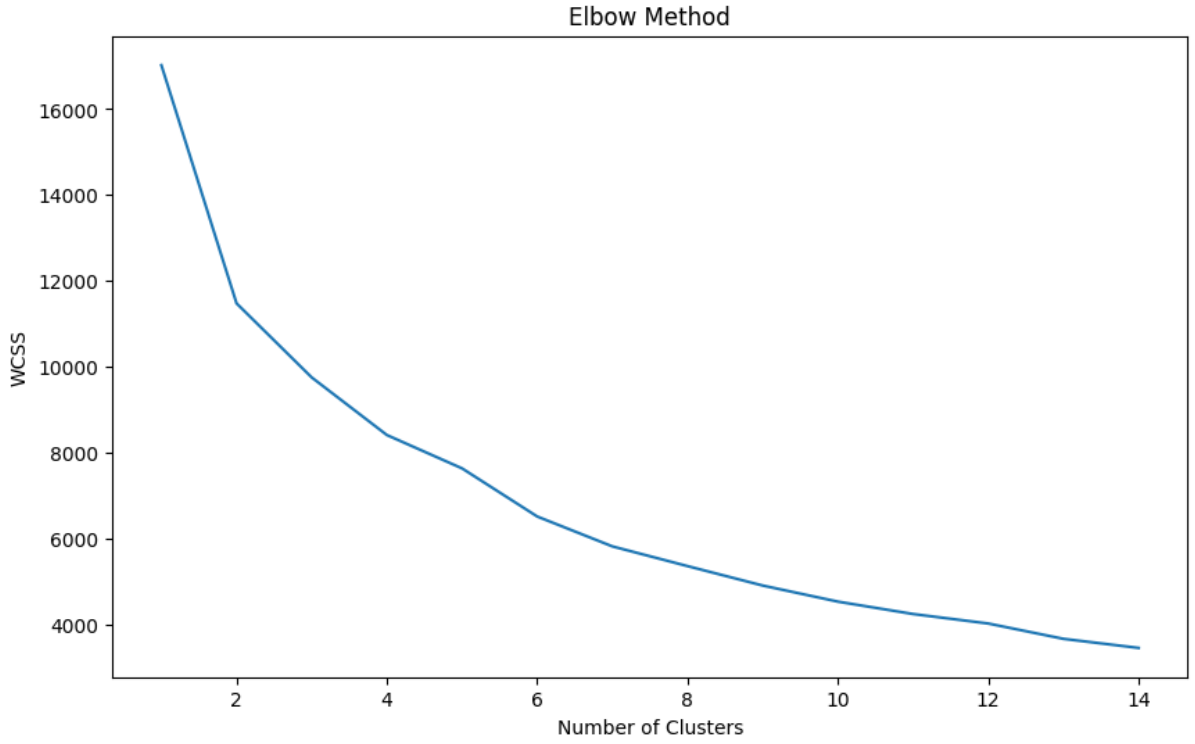

Figure 2 Elbow Method Kmeans

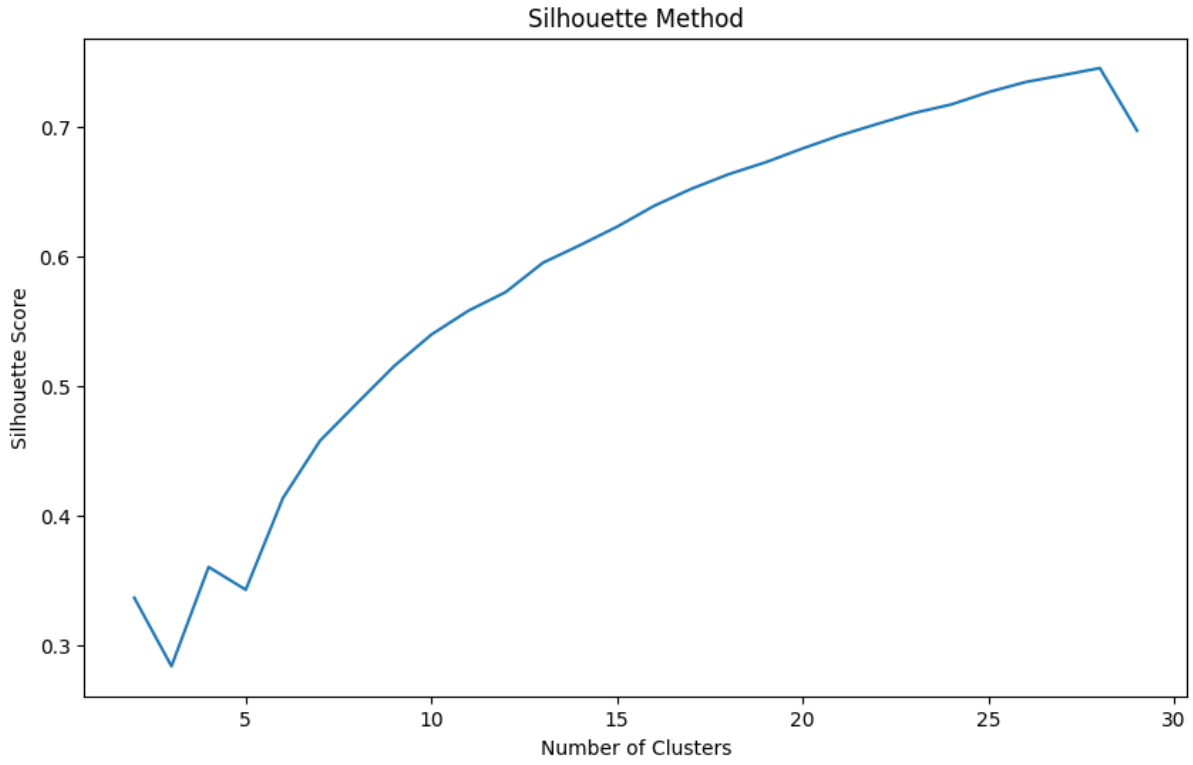

Figure 3 Silhouette Method

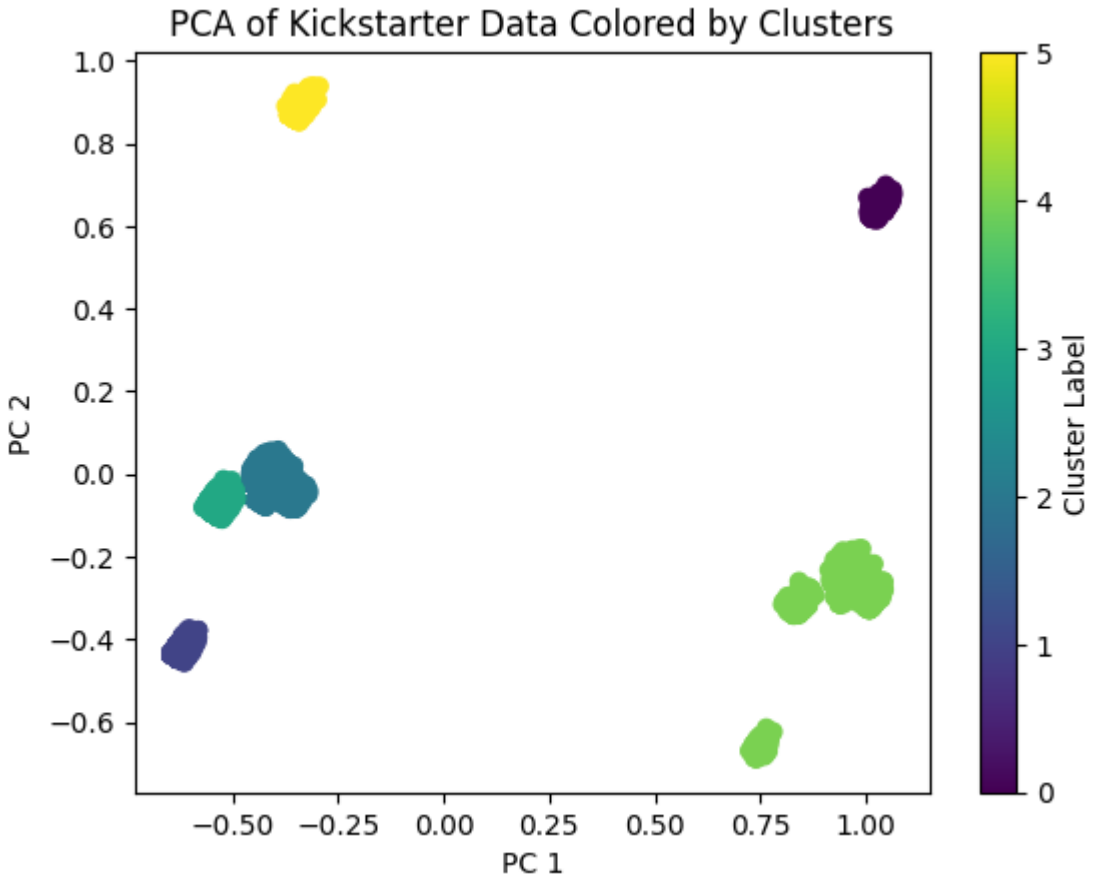

Figure 4 Kmeans PCA Clusters of 6

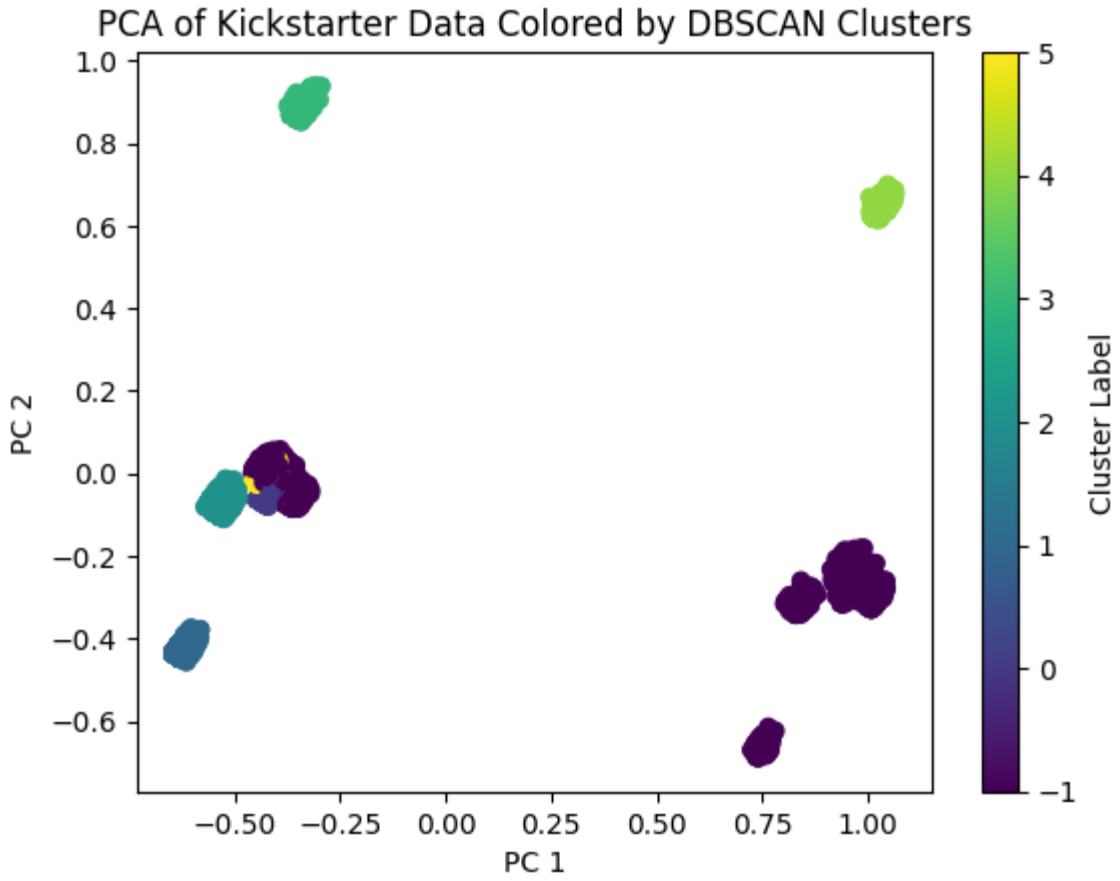

Figure 5 DBSCAN PCA Clusters of 6

Cluster	Backers Count (Mean)	USD Pledged (Mean)	Name Length (Mean)	Creation to Launch Days (Mean)	Launch to Deadline Days (Mean)	Launch to State Change Days (Mean)	Goal USD (Mean)	Failed Projects	Successful Projects
0	830.37	103755.6	6.18	64.73	34.07	34.07	33991.86	0	926
1	6.12	538.49	4.11	32.78	35.44	35.44	129985.9	1967	0
2	24.61	2866.13	4.98	44.75	35.13	35.13	90119.65	3288	0
3	14.05	1008.97	4.66	42.04	36.1	36.1	109234.8	1542	0
4	385.58	40989.61	5.55	43.56	32.07	32.07	15986.81	0	3036
5	53.14	7107.03	5.64	70.16	35.81	35.81	150863.9	1421	0

Table 1 Cluster Characteristics