INSY - 662 - Data Mining & Visualization

INDIVIDUAL PROJECT

Executive Summary

The Kickstarter Project analysis aimed to explore and derive insights from a dataset containing information about various Kickstarter campaigns. The primary goals were to build a classification model to predict project outcomes and perform clustering to uncover patterns among different projects.

Choosing the Predictors: Dates play a key role in determining the success or failure of a project. Considering that I chose create_to_launch_days and launch_to_deadline_days because the difference between the days can play a significant role. The weekdays can also affect it considering that weekends might attract more investors. It is because of this that I chose deadline_weekday, created_at_weekday and launched_at_weekday. The length of the project name and their description might also be a factor. So, I chose name_len_clean and blurb_len_clean. Also, certain categories and from certain countries might attract more funds. The number of backers for a project and the spotlight might also affect its success. It would also help with the project's success if the investors communicated with the owners. Finally, I engineered a new column amount_usd by multiplying goal and static_usd_rate.

Data Preprocessing: There were no duplicate values in the columns I selected. There were 1392 missing values in the Category column, but I chose to drop them because replacing them with the most common value or other logic might have affected the classification outcome as some projects in a certain category might be more popular. I converted the categorical columns using

one-hot encoding and then used StandardScaler on numeric columns to scale the values on the same scale.

Task 1: Classification Model

Model Training involved employing TrainTestSplit to divide the dataset into 80-20 splits and initializing and testing the Random Forest Classifier.

Results and Recommendations:

Class	Precision	Recall	F1-Score	Support
Canceled	0.50	0.03	0.05	334
Failed	0.84	0.99	0.91	1674
Successful	1.00	1.00	1.00	781
Suspended	1.00	1.00	1.00	27
Accuracy			0.88	2816
Macro Avg	0.83	0.76	0.74	2816
Weighted Avg	0.84	0.88	0.83	2816

The overall accuracy of the model is 0.88, which means that it correctly classified 88% of the samples. In contrast, if the model had classified all the projects in the Failed class, the accuracy of the model would have been 59%. This would prove that the model is performing well. For the Canceled class, the precision is very low, at 0.03. This means that the model often predicts that samples belong to the Canceled class when they belong to another class. I also created and printed a DecisionTree with depth 5 (Figure 1). I found out that the predictors 'spotlight' and 'disable_communication' play a major role in classification.

Business Use: The classification model for the Kickstarter dataset serves to predict project outcomes, aiding project creators, backers, and the platform itself. It facilitates risk assessment, targeted marketing, decision support, guiding stakeholders in making informed choices, and enhancing platform performance.

Task 2: Clustering Model

I used the Elbow Method (Figure 2) and Silhouette Score to determine the optimal number of clusters. The Elbow Method involves applying the K-Means algorithm for k values ranging from 2 to 9 and plotting the sum of squared distances of samples to their closest cluster center (inertia). The Silhouette Method calculates silhouette scores for different k values (2 to 9) to assess cluster quality, where a higher score indicates better-defined clusters. This approach provides insights into the trade-off between cluster compactness and separation.

Clustering Insights:

I chose to create 5 clusters based on the Elbow and Silhouette scores. Employing Kmeans and Agglomerative clustering algorithms, I observed that Kmeans outperformed Agglomerative. Kmeans exhibited more balanced data distribution among clusters, with Silhouette scores around 0.27, indicating good cohesion. In contrast, Agglomerative clustering showed uneven data point distribution, with two clusters having a score of 0 and three clusters exceeding 0.60. Additionally, I assessed the distribution of the 'usd_amount' and ‘launch_to_dealine_days’ columns(Figure 3) and found out that the projects with high USD amounts were in Cluster 4. In contrast, the projects with high launch_to_dealine_days were in Cluster 3. Additionally, projects with high create_to_dealine_days’ were distributed among Clusters 0,1 and 2(Figure 4).

Appendix:

Figure 1:

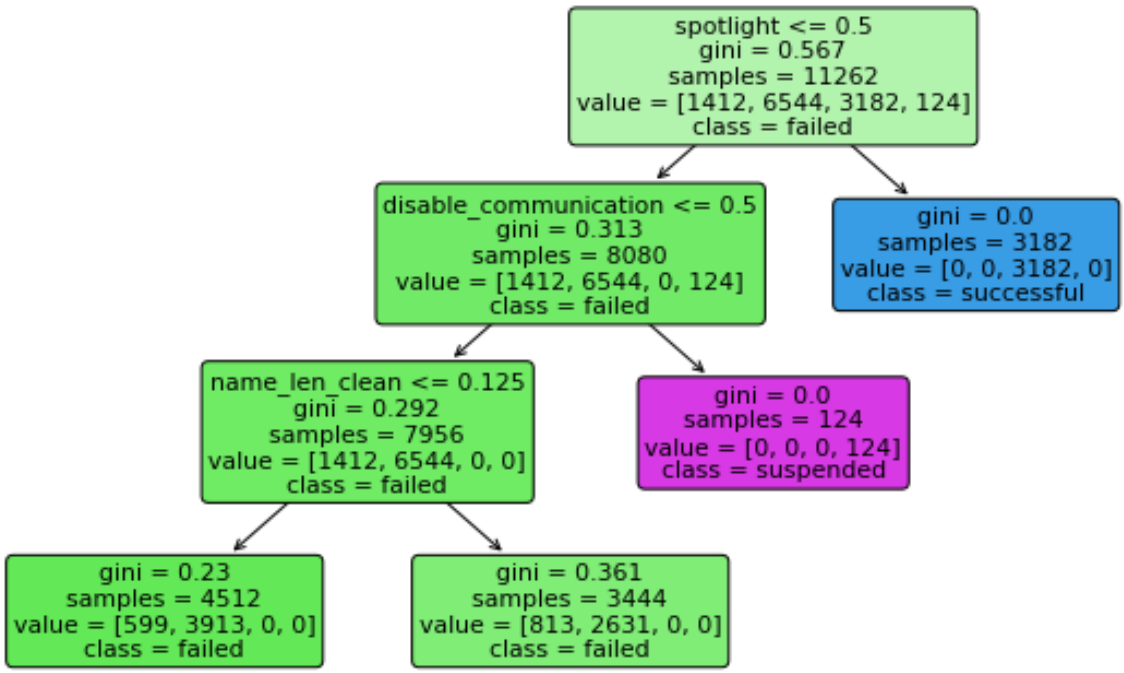

Figure 1

Figure 2:

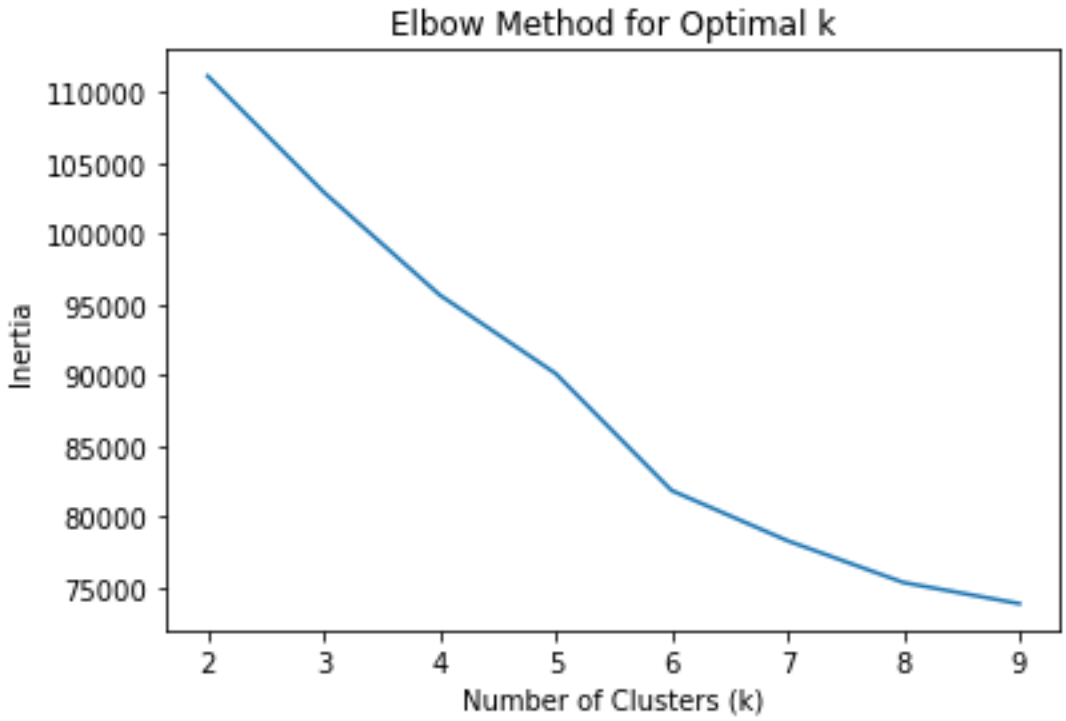

Figure 2

Figure 3:

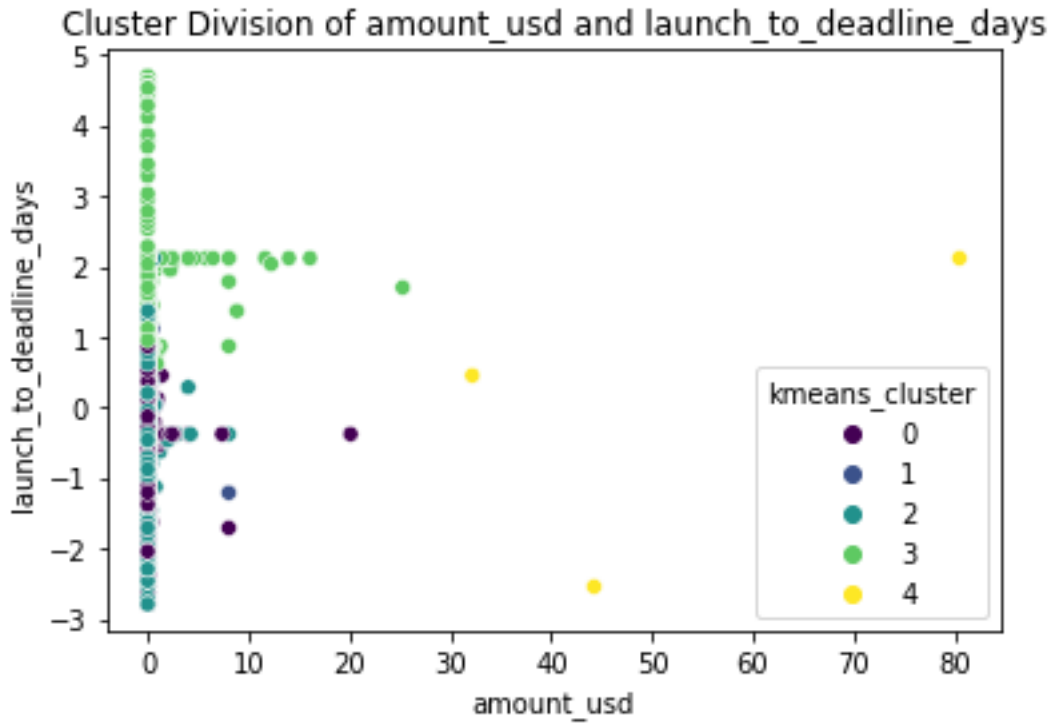

Figure 3

Figure 4:

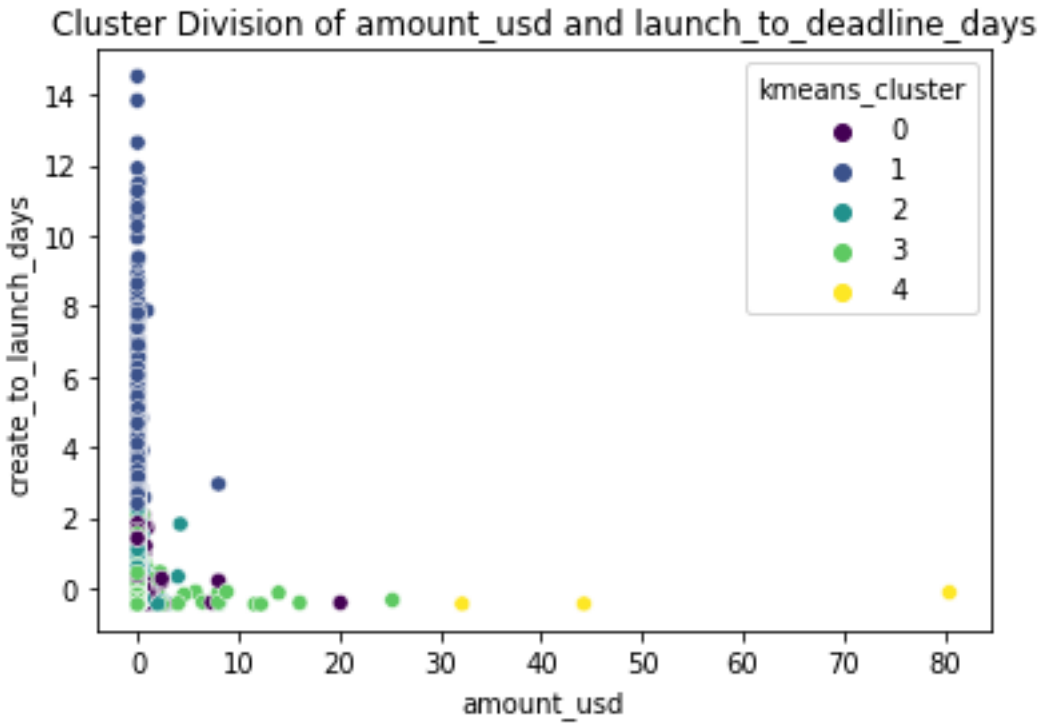

Figure 4