Drafted by: Syed Ammad Sohail

Introduction

Kickstarter has emerged as a pivotal platform in the realm of crowdfunding, revolutionizing the way creative, technological, and entrepreneurial projects secure funding. Unlike traditional funding methods, Kickstarter provides an avenue for creators to present their ideas directly to the public, gaining financial support based on the collective interest and belief in their project's potential.

Objective: Utilizing Data for Strategic Decision-Making

The project's essence is to apply data science for analyzing Kickstarter projects, considering aspects like financial goals, categories, and timelines. It aims to develop a predictive model to determine project success or failure, aiding creators in planning and helping Kickstarter support projects requiring additional focus. Additionally, it includes cluster analysis to categorize projects into distinct groups, uncovering patterns and trends that showcase the platform's project diversity.

Classification Task

In the enhanced Exploratory Data Analysis (EDA), interactive tools like pygwalker and ydata_profiling were used for dynamic analysis, offering more depth than the static charts in Appendix I. We removed predictors identified as irrelevant after project launch (detailed in Appendix II). The analysis focused on projects classified as 'successful' or 'failed', using TF-IDF Vectorization to understand the significance of words in project names and to impute missing values in the 'category' column. This approach left only about 150 missing entries, which were subsequently dropped (Appendix IV). To capture the core target distribution, the dataset categorized countries as US, GB, CA, and others, and narrowed down project categories to the top 8 (Appendix III). Timeframes were divided into 'Beginning' (first ten days), 'Middle' (next ten days), and 'End' (remaining days), with 24-hour periods segmented into four 6-hour

intervals. Weekdays were classified as Monday to Friday and Weekend, while months were grouped into Quarters 1 through 4. We addressed skewness in the data through log transformation and identified outliers using an Isolation Forest algorithm with a contamination setting of 0.05 and 100 estimators. New features included the conversion of project goals to USD and the combination of 'create_to_launch_days' with 'launch_to_deadline_days' into 'create_to_deadline_days'. For the Random Forest and Gradient Boosting models, categorical columns involving text were converted to dummy variables through one-hot encoding, while ordinal data such as dates and months were treated as categorical (Appendix VII). For other algorithms, all categorical columns were converted to dummies and were standardized using Min-Max scaling (Appendix VIII). The year of project launch was retained as an integer. Feature selection aimed to eliminate multicollinearity, as shown in the correlation heatmap in Appendix V, and to evaluate feature importance using the Random Forest Algorithm (Appendix VI). Five algorithms were employed using random_state = 42, each with unique strengths: Random Forest for complex data, Gradient Boosting for imbalanced datasets, KNN for simplicity in small datasets, ANN for complex patterns, and Logistic Regression for binary classification efficiency. After hyperparameter tuning and testing all models, the Gradient Boosting Classifier proved most accurate, with an accuracy of nearly 75.07%. considering it uses boosted trees which are built sequentially, focusing more on removing classification inaccuracies in previous trees. This model can significantly enhance Kickstarter's platform by offering data-driven insights to help creators optimize their projects for success, thus attracting more backers and fostering a vibrant community of creators and supporters. Models with their respective accuracies and hyperparameters are highlighted in Appendix IX. The top 5 features can also be seen in Appendix VI.

Clustering:

After incorporating some features such as *state, backers_count, spotlight, staff_pick*, *create_to_launch,days* and *launch_to_deadline_days* in the preprocessed dataset, which was used in Random Forest and Gradient Boosting models, KPrototype was selected for its capacity to handle both categorical and numerical features. For clustering algorithms like KMeans, Hierarchical Clustering, DBSCAN, and Autoencoders, the dataset, employed in ANN, KNN, and Logistic Regression, was used, incorporating the added features. These were chosen for their abilities in forming cohesive clusters (Hierarchical Clustering), managing outliers (DBSCAN), dimensionality reduction (Autoencoders), and efficient data segmentation (KMeans). Although KMeans had a higher silhouette score (Appendix XII), KPrototype was favored for its more insightful interpretations of cluster centroids and detailed cluster analysis. I selected 4 clusters based on elbow plots in Appendix X and XI.

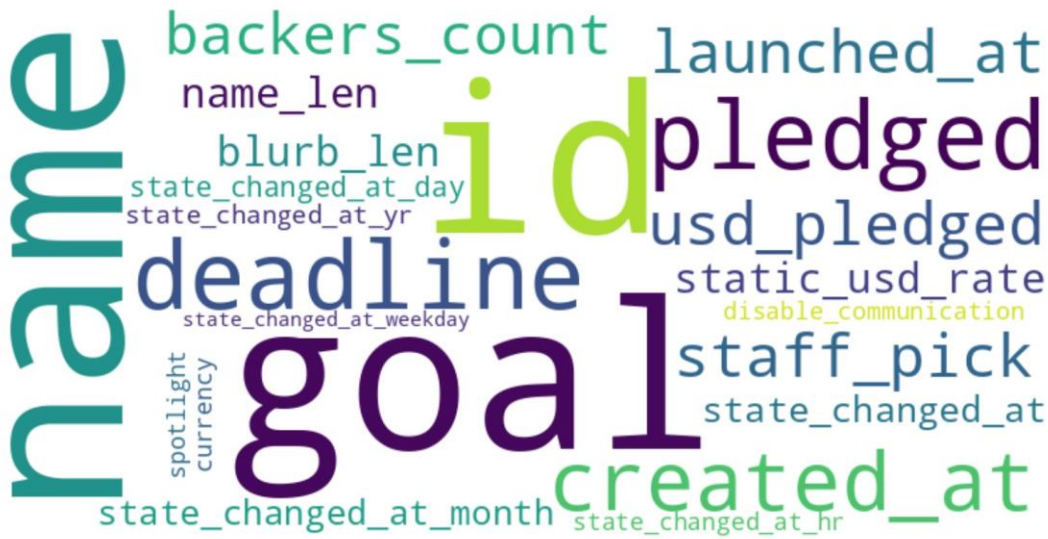
Recommendations based on insights (Insights Summary in Appendix XIII)

Successful Kickstarter projects, particularly those in Cluster 3, underscore the importance of detailed and comprehensive project descriptions. Creators should clearly articulate their project's goals, uniqueness, and the potential impact on backers. Projects hastily launched, like those in Cluster 2, often underperform, stressing the importance of extensive market research, a strong narrative, and strategic marketing. Active audience engagement is critical; creators must regularly communicate with backers, respond to feedback, and use social media for community building. Aligning with in-demand categories, such as successful hardware projects in Cluster 3, can also increase a project's appeal. Kickstarter should focus on increasing diversity and visibility of projects, particularly those not naturally attracting high engagement. This could involve revising algorithms and editorial practices and providing creators with resources like webinars and market research tools. Enhancing features for community engagement and offering creators data insights from successful projects, such as case studies and trend analyses, are also crucial for informed project development.

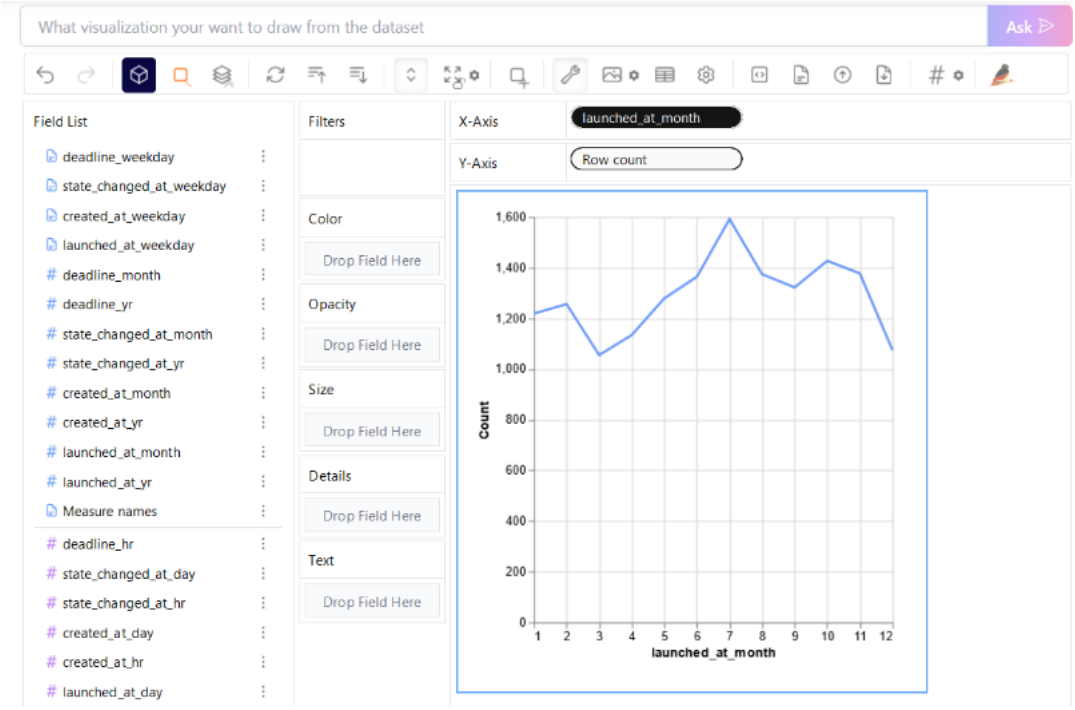
Drafted by: Syed Ammad Sohail

Appendix

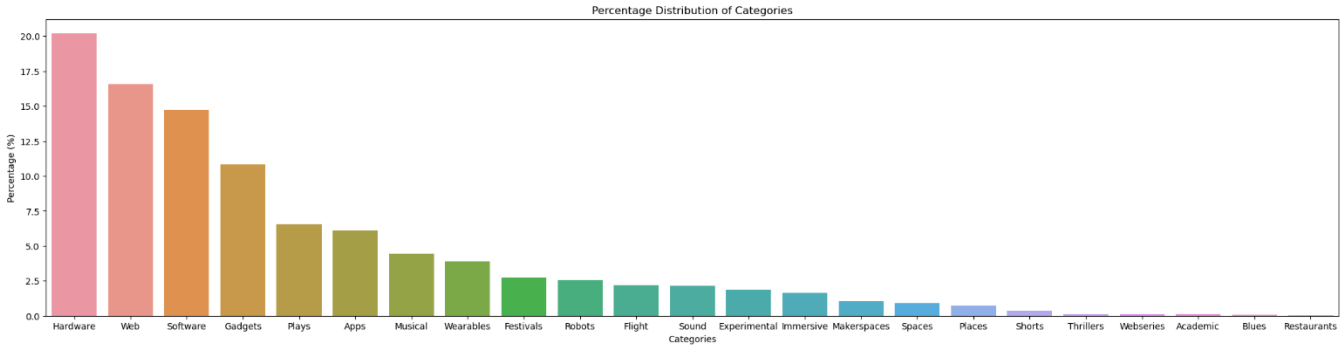
I. Invalid Predictors

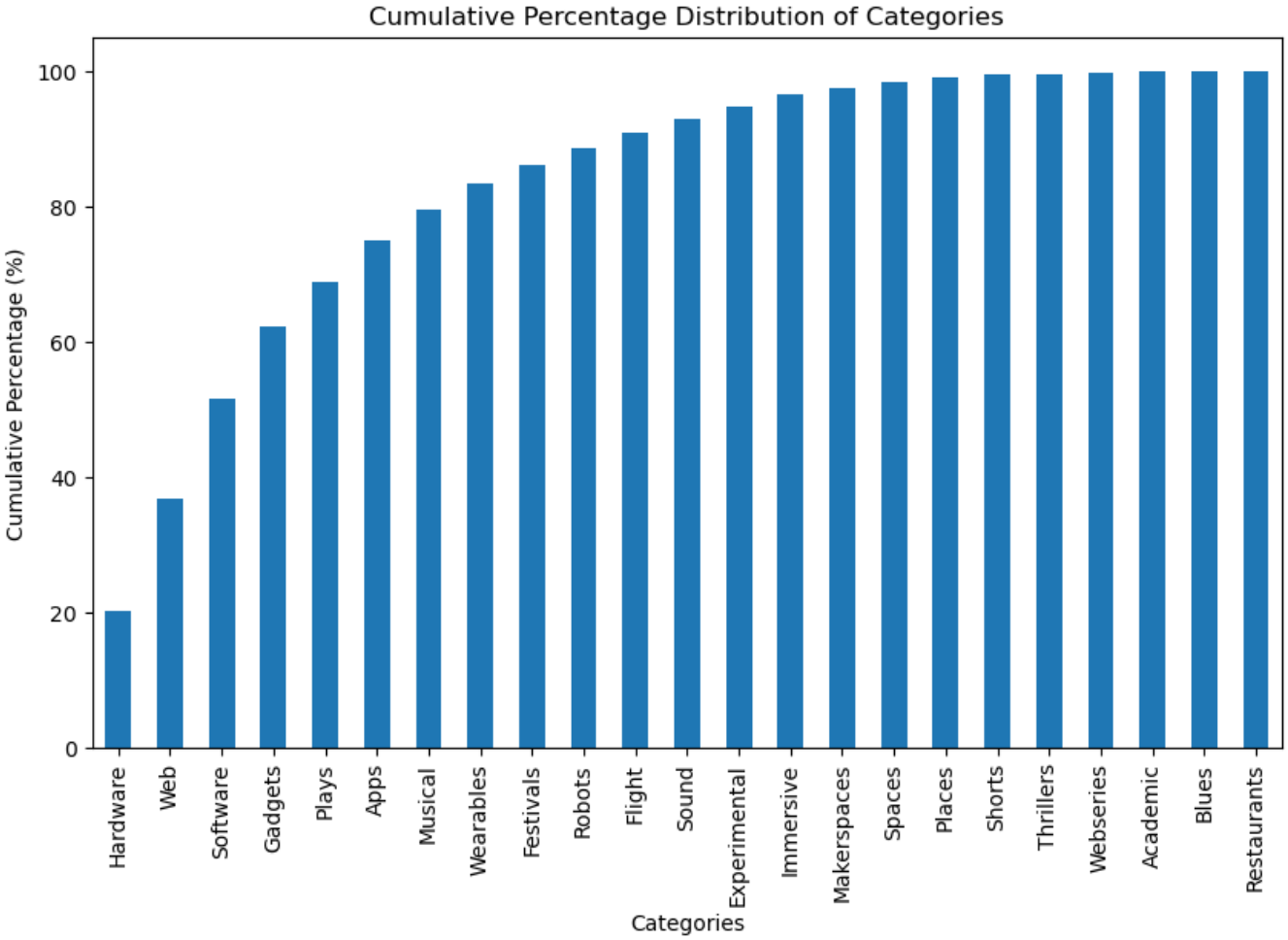


II. Interactive EDA using Pygwalker

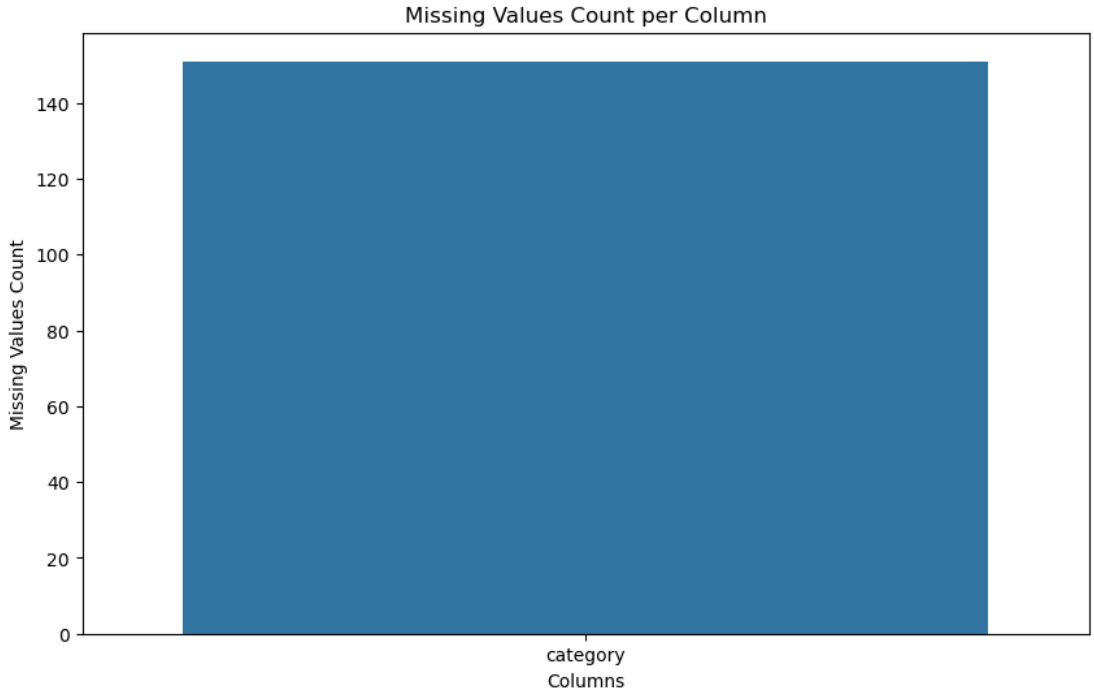


III. Grouping Top 8 categories

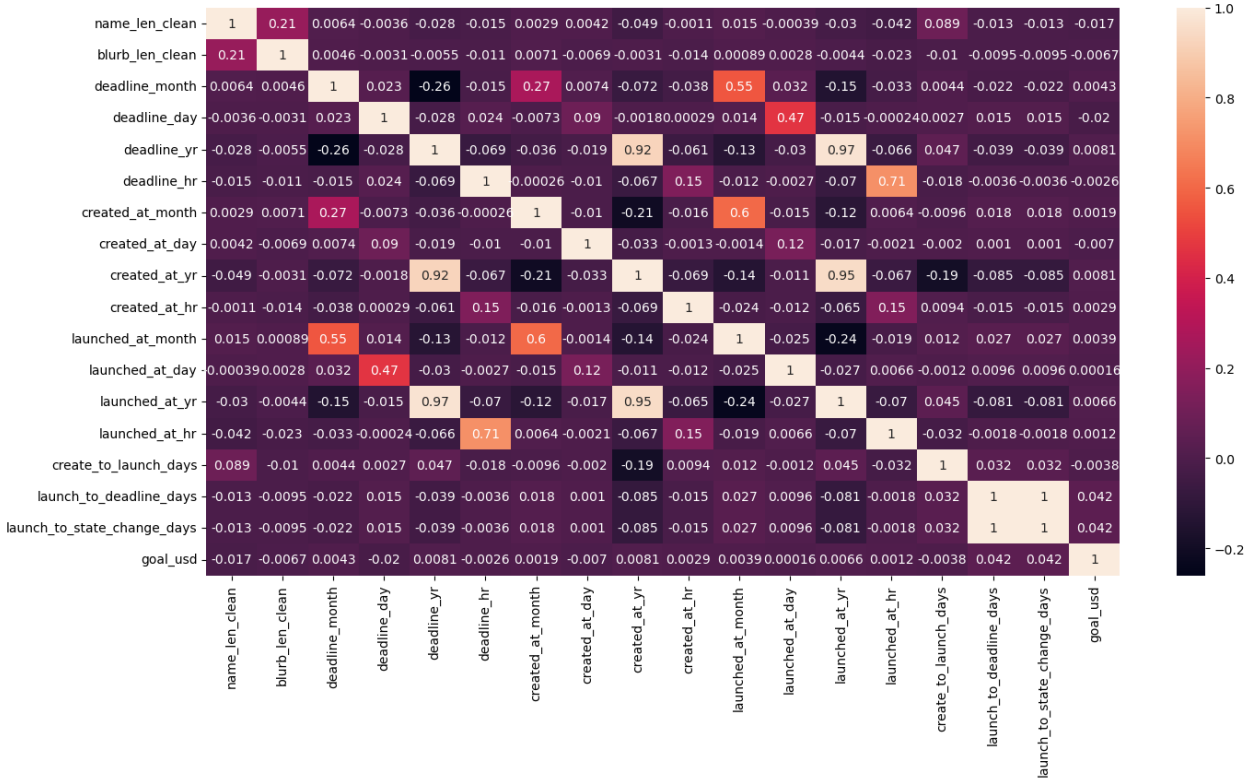




IV. Missing Values Count



V. Correlation Heatmap



VI. Random Forest Feature Importance

log_goal	0.111348
log_create_to_deadline_days	0.067487
launched_at_hr	0.054792
created_at_day	0.053509
name_len_clean	0.052431
launched_at_day	0.052269
deadline_day	0.052095
deadline_hr	0.051559
created_at_hr	0.050740
log_blurb_len_clean	0.045676
created_at_month	0.037603
category_Web	0.036967
launched_at_month	0.035048
deadline_month	0.034379
log_launched_at_yr	0.032612
category_Software	0.016596
category_Plays	0.012567
category_Other	0.011309
launched_at_weekday_Tuesday	0.009935
category_Hardware	0.009514
country_grouped_US	0.009319
created_at_weekday_Monday	0.009038
created_at_weekday_Tuesday	0.008978
launched_at_weekday_Wednesday	0.008854
launched_at_weekday_Monday	0.008749
deadline_weekday_Thursday	0.008723
created_at_weekday_Wednesday	0.008327
category_Musical	0.008282
deadline_weekday_Saturday	0.008209
created_at_weekday_Thursday	0.008066
deadline_weekday_Sunday	0.008031
deadline_weekday_Wednesday	0.007961
launched_at_weekday_Thursday	0.007577
deadline_weekday_Monday	0.007105
country_grouped_Other	0.007047
country_grouped_GB	0.006867
created_at_weekday_Saturday	0.006839
created_at_weekday_Sunday	0.006678
deadline_weekday_Tuesday	0.006438
category_Gadgets	0.006126
launched_at_weekday_Sunday	0.005083
launched_at_weekday_Saturday	0.004829
category_Wearables	0.004440

dtype: float64

VII. Final Features used in Gradient Boosting and Random Forest Classifier

0	name_len_clean	12619 non-null	float64
1	deadline_month	12619 non-null	category
2	deadline_day	12619 non-null	category
3	deadline_hr	12619 non-null	category
4	created_at_month	12619 non-null	category
5	created_at_day	12619 non-null	category
6	created_at_hr	12619 non-null	category
7	launched_at_month	12619 non-null	category
8	launched_at_day	12619 non-null	category
9	launched_at_hr	12619 non-null	category
10	launched_at_weekday_Monday	12619 non-null	uint8
11	launched_at_weekday_Saturday	12619 non-null	uint8
12	launched_at_weekday_Sunday	12619 non-null	uint8
13	launched_at_weekday_Thursday	12619 non-null	uint8
14	launched_at_weekday_Tuesday	12619 non-null	uint8
15	launched_at_weekday_Wednesday	12619 non-null	uint8
16	created_at_weekday_Monday	12619 non-null	uint8
17	created_at_weekday_Saturday	12619 non-null	uint8
18	created_at_weekday_Sunday	12619 non-null	uint8
19	created_at_weekday_Thursday	12619 non-null	uint8
20	created_at_weekday_Tuesday	12619 non-null	uint8
21	created_at_weekday_Wednesday	12619 non-null	uint8
22	deadline_weekday_Monday	12619 non-null	uint8
23	deadline_weekday_Saturday	12619 non-null	uint8
24	deadline_weekday_Sunday	12619 non-null	uint8
25	deadline_weekday_Thursday	12619 non-null	uint8
26	deadline_weekday_Tuesday	12619 non-null	uint8
27	deadline_weekday_Wednesday	12619 non-null	uint8
28	category_Gadgets	12619 non-null	uint8
29	category_Hardware	12619 non-null	uint8
30	category_Musical	12619 non-null	uint8
31	category_Other	12619 non-null	uint8
32	category_Plays	12619 non-null	uint8
33	category_Software	12619 non-null	uint8
34	category_Wearables	12619 non-null	uint8
35	category_Web	12619 non-null	uint8
36	log_goal	12619 non-null	float64
37	log_blurb_len_clean	12619 non-null	float64
38	log_create_to_deadline_days	12619 non-null	float64
39	log_launched_at_yr	12619 non-null	float64

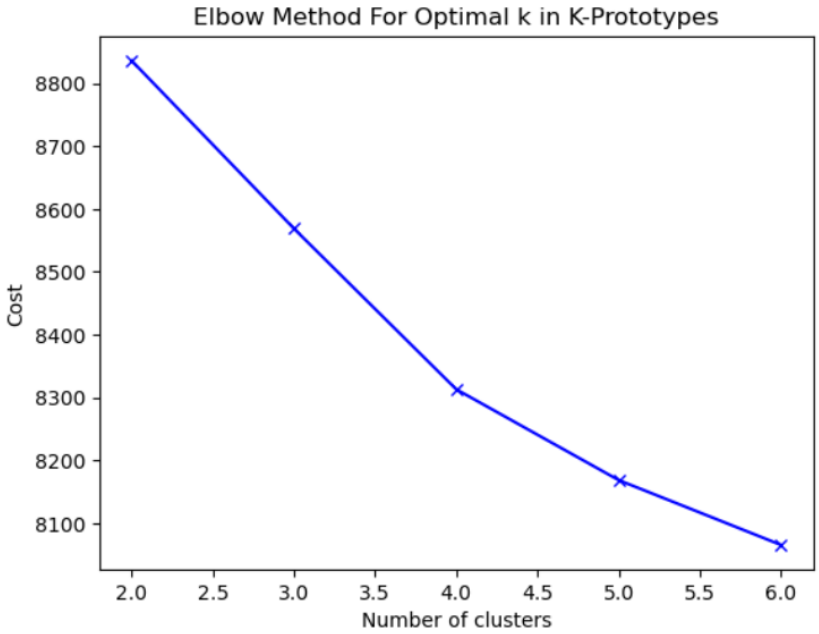
VIII. Final Features used in ANN, KNN and Logistic Regression

0	name_len_clean	12619 non-null	float64	27	log_blurb_len_clean	12619 non-null	float64
1	launched_at_weekday_Monday	12619 non-null	float64	28	log_create_to_deadline_days	12619 non-null	float64
2	launched_at_weekday_Saturday	12619 non-null	float64	29	log_launched_at_yr	12619 non-null	float64
3	launched_at_weekday_Sunday	12619 non-null	float64	30	launched_at_day_group_Middle	12619 non-null	float64
4	launched_at_weekday_Thursday	12619 non-null	float64	31	launched_at_day_group_Beginning	12619 non-null	float64
5	launched_at_weekday_Tuesday	12619 non-null	float64	32	created_at_day_group_Middle	12619 non-null	float64
6	launched_at_weekday_Wednesday	12619 non-null	float64	33	created_at_day_group_Beginning	12619 non-null	float64
7	created_at_weekday_Monday	12619 non-null	float64	34	deadline_day_group_Middle	12619 non-null	float64
8	created_at_weekday_Saturday	12619 non-null	float64	35	deadline_day_group_Beginning	12619 non-null	float64
9	created_at_weekday_Sunday	12619 non-null	float64	36	launched_at_month_quarter_Q1	12619 non-null	float64
10	created_at_weekday_Thursday	12619 non-null	float64	37	launched_at_month_quarter_Q3	12619 non-null	float64
11	created_at_weekday_Tuesday	12619 non-null	float64	38	launched_at_month_quarter_Q2	12619 non-null	float64
12	created_at_weekday_Wednesday	12619 non-null	float64	39	created_at_month_quarter_Q1	12619 non-null	float64
13	deadline_weekday_Monday	12619 non-null	float64	40	created_at_month_quarter_Q3	12619 non-null	float64
14	deadline_weekday_Saturday	12619 non-null	float64	41	created_at_month_quarter_Q2	12619 non-null	float64
15	deadline_weekday_Sunday	12619 non-null	float64	42	deadline_month_quarter_Q1	12619 non-null	float64
16	deadline_weekday_Thursday	12619 non-null	float64	43	deadline_month_quarter_Q2	12619 non-null	float64
17	deadline_weekday_Tuesday	12619 non-null	float64	44	deadline_month_quarter_Q3	12619 non-null	float64
18	deadline_weekday_Wednesday	12619 non-null	float64	45	launched_at_hour_group_Night	12619 non-null	float64
19	category_Gadgets	12619 non-null	float64	46	launched_at_hour_group_Afternoon/Evening	12619 non-null	float64
20	category_Hardware	12619 non-null	float64	47	created_at_hour_group_Morning	12619 non-null	float64
21	category_Musical	12619 non-null	float64	48	created_at_hour_group_Afternoon/Evening	12619 non-null	float64
22	category_Plays	12619 non-null	float64	49	created_at_hour_group_Night	12619 non-null	float64
23	category_Software	12619 non-null	float64	50	deadline_hour_group_Morning	12619 non-null	float64
24	category_Wearables	12619 non-null	float64	51	deadline_hour_group_Afternoon/Evening	12619 non-null	float64
25	category_Web	12619 non-null	float64	52	deadline_hour_group_Night	12619 non-null	float64
26	log_goal	12619 non-null	float64				

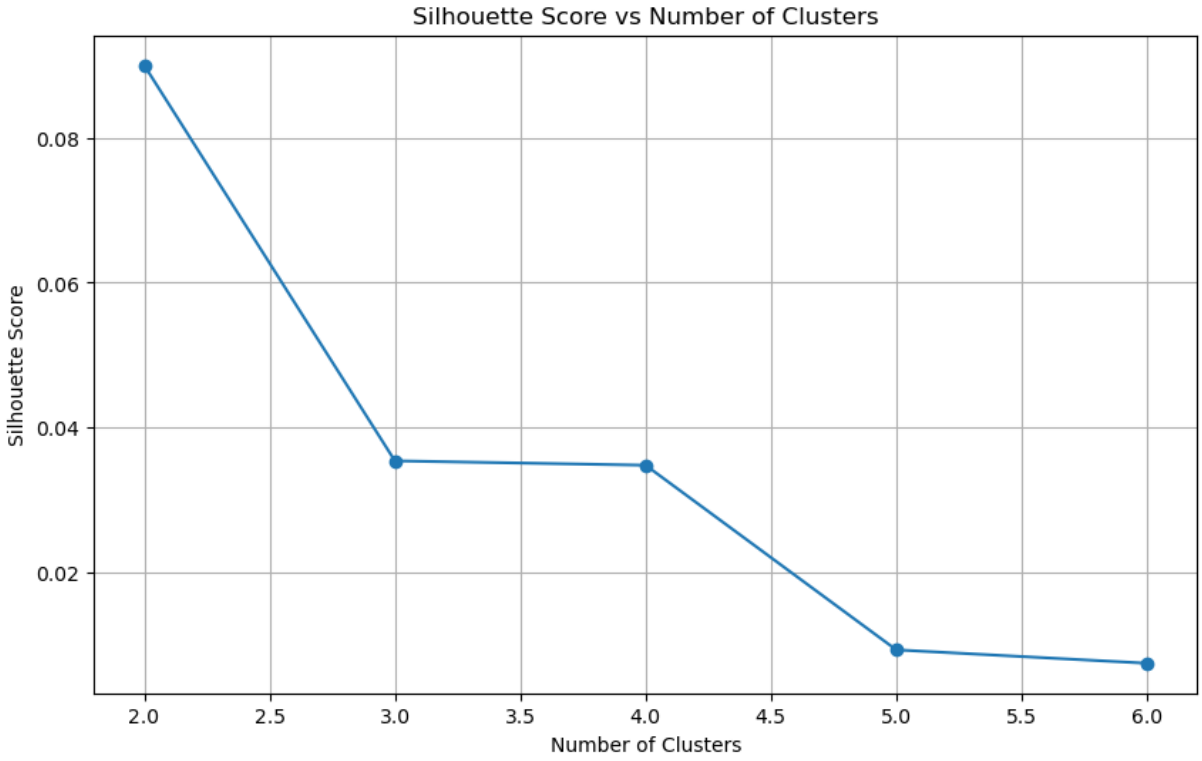
IX. Model Summary

Name	Accuracy/ %	Hyperparameters used
Random Forest Classifier	74.12	max_depth=20, min_samples_leaf=2, min_samples_split=3, n_estimators= 300, random_state=42, warm_start=True
Gradient Boosting Classifier	75.07	learning_rate=0.10, max_depth=3, n_estimators=205, random_state=42, min_samples_split=20, min_samples_leaf = 2
Artificial Neural Network (MLP Classifier)	72.53	alpha=0.0001, hidden_layer_sizes=(32, 16), max_iter=1000, activation='logistic', solver='adam', random_state=42
Logistic Regression	72.53	random_state=42, C=100, penalty='l1', solver='liblinear'
K-nearest Neighbour	67.91	n_neighbors=30

X. Elbow Method for KPrototype



XI. Silhouette Score v/s Number of Clusters for KPrototype



XII. Comparison of Silhouette Scores (number of clusters = 4, random_state=50)

Clustering Algorithm	Silhouette Scores	Hyperparameters (if any)
KPrototype	0.034780	init='Cao', n_init=5
KMeans	0.103195	None
Hierarchical Clustering (complete linkage)	0.034958	metric = "euclidean"
DBSCAN	-0.079918	eps=2.5, min_samples=30
Autoencoders	0.0789349	encoding_dim = 32, activation='relu'

XIII. Clusters Summary

Cluster	Description	Key Characteristics	Engagement and Success	Project Timelines	Geographic and Currency Details	Staff Engagement and Visibility
Cluster 0: The Modest Cluster	Modest-sized projects with standard detail in project presentations.	Average project name and blurb lengths. Launch and deadlines typically mid-week	Low backers count and goal amounts Predominantly failed projects	Short create-to-launch periods Moderate launch-to-deadline and launch-to-state-change durations	Mostly US-based projects, using USD	Low frequency of being a staff pick or featured in the spotlight
Cluster 1: The Niche Web Projects Cluster	Niche web projects with focused presentation	Shorter project names and blurbs Launches and deadlines mid-week	Slightly higher backers count than Cluster 0, yet low overall Majority of projects fail	Similar to Cluster 0 but with slightly longer launch-to-deadline days	US-based projects, using USD	- Rarely a staff pick or in the spotlight
Cluster 2: The Low Engagement Cluster	Projects with the lowest level of engagement.	Shortest project names and blurbs Launches end of the week, leading to Sunday deadlines	Lowest backers count among all clusters Predominantly failed projects	Shortest create-to-launch periods	Mainly US-based projects, using USD	Seldom a staff pick or spotlighted
Cluster 3: The Successful Hardware Cluster	Successful hardware projects with detailed presentation.	Longer project names and blurbs Launch and deadlines evenly spread across the week	Significantly higher backers' count High rate of project success	Moderate creation-to-launch and launch-to-deadline periods	US-based projects, predominantly	