Introduction

Kickstarter is a founding crowdfunding platform offering a diverse array of categories with art, music, film, technology, design, food, publishing, and more. Through the platform, creators present their projects, detailing their visions, funding goals, and proposed rewards for backers. With its "all-or-nothing" funding model, Kickstarter provides a unique environment where projects must meet their financial targets within specified timelines to secure funding. Once a project reaches its goal, Kickstarter facilitates the collection of pledged funds, enabling creators to realize their visions and deliver the result to backers. This dynamic website empowered creators to engage with a global community and bring their creative ideas to reality.

Therefore, we were provided with the Kickstarter dataset to first develop a classification model to predict if a new project will be successful or not when the owner submits it. After that, the goal is to develop a clustering model to group projects together.

Classification Model

Variables to drop

The first step in building the classification model involves removing variables considered post-launch information and those that do not contribute to predicting whether a project will succeed or fail.

1. **pledged & usd_pledged**: Excluded because these represent the amount of money pledged by backers after the project has launched.
2. **backers_count**: Excluded as it reflects the number of individuals who have supported the project financially, which is a result of accumulating post-launch.
3. **spotlight**: Excluded because it indicates whether a project was featured in the Kickstarter spotlight, a status known after a project's success, not available at launch.
4. **staff_pick**: Excluded because it's a status that could be assigned after evaluating the project's initial performance, thus not predetermined at launch.
5. **state_changed_at,state_changed_at_month,state_changed_at_day,state_changed_at_yr,state_changed_at_hr,state_changed_at_weekday**: Excluded as they pertain to the timing of the project's status changes, such as from active to successful or failed, which occurs after the project has been launched and is not predictive of initial success.
6. **disable_communication**: Excluded because it suggests whether communication with the project owner was disabled, a status that could change during or after the project lifecycle and does not inherently predict success at launch.
7. **currency**: Excluded considering that financial analysis will be conducted in a standardized currency (USD), making the original currency less relevant, especially since the country of origin provides a similar geographical context.
8. **name & blurb**: Excluded from direct use in predictive modeling without text analysis capabilities because their raw textual content requires preprocessing to be effectively used as features, which is beyond this project's scope.
9. **launch_to_state_change_days**: Excluded as it involves the duration from launch to when the project's status changes, incorporating post-launch activity data, thus unavailable at launch.
10. **id**: id is a label used for identifying and tracking projects within the dataset. It holds no information about the characteristics of the project that could influence its success or failure and can add unnecessary complexity.
11. **deadline,deadline_weekday, deadline_day, deadline_yr, and deadline_hr**: Excluded as converting weekdays into categorical variables increases the model's

complexity by adding multiple binary columns. For the other variables, they are too precise and are likely insignificant

12. **created_at,created_at_day,created_at_yr,created_at_hr,created_at_weekday**: Excluded because the precise time and date of project creation, while informative about the project's timeline, likely offer less direct predictive value for success than variables related to the project's content, financial goals, or market context.
13. **deadline_month,created_at_month**: Dropped because highly correlated variables
14. **launched_at_month**: dropped because converted into 'launched_at_season'

Variables to keep

1. **goa**l: Kept because it represents the financial target set by the project creator before the project launches, directly influencing and reflecting the project's ambition and planning. The goal variable of projects or campaigns is converted from their original currency to USD (U.S. dollars), resulting in the `usd_goal` variable. Consequently, **static_usd_rate and goal are dropped,** leaving **`usd_goal`** as the only variable representing the financial target in a standardized currency format.
2. **country**: Kept because it provides geographical context that could influence a project's success, reflecting market size, crowdfunding culture, and potential reach.
3. **category**: Kept because the project's category could significantly affect its appeal to backers, with some categories attracting more interest or support based on market trends and audience preferences.
4. **create_to_launch_days**: Kept as it indicates the amount of time spent preparing the project before it goes live, potentially reflecting the planning and quality of the project proposal.
5. **name_len_clean**: Kept as they could indirectly measure the effort put into making the project title concise and appealing, influencing first impressions.
6. **blurb_len_clean**: Kept for reasons similar to name_len and name_len_clean.
7. **created_at_yr**: Kept because the success of projects might be influenced by factors such as economic conditions, platform popularity, and competition, all of which can vary significantly from year to year. Including it in the model can help control for these temporal effects, leading to more accurate predictions.
8. **create_to_launch_days**: Keeping this variable allows for insights into the preparation period before a project goes live, which could be crucial in understanding factors influencing project success or failure. For example, shorter or longer periods of preparation may be associated with different outcomes like higher funding targets or more marketing activities.
9. **launch_to_deadline_days**: Keeping this variable allows analysis of the project's fundraising trajectory and dynamics over time. It shows how soon a project has to attract supporters, whether there are any ups and downs in funding traction, and if projects that reach their goal quickly differ from projects that take a longer time.

Final Classification Model

The Random Forest model was selected as the best for the project due to its high accuracy (75%) and effectiveness in managing large datasets. This model, an ensemble method using multiple decision trees, excels in accuracy and handling complex data without overfitting. It was prepared by removing irrelevant features, excluding missing values for efficiency, and converting the dependent variable into a binary format (successful=1, failed=0). These steps ensured a streamlined analysis and significant statistical results.

Clustering Model

Pre Processing

The factors retained are essential and parsimonious, relating to pre- and post-launch, unlike the original model which only considers pre-launch changes Such as the previous method, the monetary targets were standardized to the USD for consistency and lanched_at_month converted into launched_at_season. Minimum-maximum scaling was used for its interpretive utility rather than z-score standardization, as it facilitates the understanding of the importance of each factor.

Final Clustering Model

The K-Prototypes model is ideal for our dataset containing both categorical and numerical variables. It uses means for numerical features and modes for categorical ones to define cluster centroids, enhancing robustness and interpretability compared to methods like k-means or DBSCAN, which either require one-hot encoding, so having much less interpretability, or clustering only numerical features, which would result in a loss of insights. We chose Huang's initialization over Cao's for its ability to account for data variability better by using k-modes for centroids. Two clusters were formed to improve interpretability and visibility. The model's effectiveness is confirmed by lower total costs in the second iteration of clustering, indicating optimal grouping. The model's gamma of 0.042 suggests a strong influence of categorical variables in cluster differentiation, with both types of features showing good cohesion within clusters.

Cluster 0 projects, which are mostly US-based and within the Web category, tend to have fewer backers and lower funding, with concise names but detailed descriptions. They launch quickly but run longer campaigns with modest financial goals and have a lower success rate, often lacking staff picks and spotlight features, typically launching in the Fall.

Cluster 1 projects, predominantly from the US and within the Hardware category, draw numerous backers and significant funding. These projects feature longer names with more concise descriptions, suggesting a different approach to capturing potential backers. Creators prepare thoroughly before launching, resulting in shorter campaigns and faster transitions to completion. Despite their success, they set surprisingly low funding goals, potentially affected by scaling or outliers. These successful projects typically didn't receive staff picks but effectively used the spotlight feature and were often launched in the Summer.

Kickstarter clustering analysis shows that well-prepared projects with shorter campaign lengths are more successful and attract greater funding compared to those that launch quickly and have longer campaigns, which often fail and secure less funding. Successful strategies for Kickstarter include launching in the summer, focusing on hardware projects, setting realistic funding goals, accepting tight deadlines, and providing detailed descriptions. These approaches benefit from the uplifting summer mood, growth in the hardware sector, achievable funding targets, the motivation created by deadlines, and clarity for investors. Kickstarter earns by taking a 5% cut from funded projects, which supports its operations. Data suggests that increasing project spotlight, particularly during low engagement periods, could increase success rates and site profitability.

	backers_count	usd_pledged	name_len_clean	blurb_len_clean	create_to_launch_days	launch_to_deadline_days	launch_to_state_chang
0	0.0002989211351053383 3	0.0005512047999035474	0.25430599814450777	0.4041417282968449	0.02808662354455996	0.394837732331174	0.3948391889
1	0.003918632000887881	0.007638382260185966	0.4195543090777106	0.43497655960073756	0.03463756708366722	0.33551748541865517	0.33551748541

launch_to_state_change_days	usd_goal	state	country	staff_pick	category	spotlight	launched_at_season	Cluster
0.3948391889529767	0.0010765922721579914	0.0	failed	US	False	Web	False	Fall
0.33551748541865517	0.0004252171829852 33	1.0	successful	US	False	Hardware	True	Summer