**Model 1: Classification to predict state is “successful” or “failure”.**

The preprocessing of data began with the conversion of project goals into a uniform currency (USD), a crucial step to ensure comparability across projects from different countries. This standardization is particularly important in a global platform like Kickstarter, where projects come from various international backgrounds. Following this, a careful curation of features was undertaken. Null values are observed in Categories, so we replace null with ‘*None*’ to take them into calculations. Irrelevant or non-analyzable columns, such as *'name'* and detailed time variables like *‘state\_changed\_at’*, were removed. More importantly, variables that would not be available at the project's launch, such as the amount pledged, were excluded to align with the project's premise of making predictions at the time of launch.

Further refinement of the dataset involved focusing on projects that were either 'successful' or 'failed', discarding other states to maintain clarity in the target variable. This decision aligns with the business objective of discerning clear outcomes for new projects. The treatment of skewed variables in the dataset marks a significant departure from standard practices for Random Forest and Gradient Boosting models. Despite these models' typical insensitivity to scale, scaling down features like *'launch\_to\_deadline\_days', 'create\_to\_launch\_days',* and *'goal\_usd'* proved beneficial. This step likely improved model accuracy by reducing the variance bias caused by extreme skewness, leading to more balanced decision-making in the model.

Categorical encoding and splitting the dataset into training and testing sets were standard yet essential steps, ensuring that the model could learn from diverse scenarios and validate its predictions effectively. The use of an isolation forest in a grid search to identify and remove outliers further refined the training set, enhancing the model’s ability to generalize. Additionally, the application of Synthetic Minority Over-sampling Technique (SMOTE) addressed the issue of imbalance in the response variable, a critical step to prevent bias towards the majority class.

In building the model, various algorithms were tested, including logistic regression, artificial neural networks, Random Forest, and Gradient Boosting. The choice of Gradient Boosting as the final model was based on its superior performance, particularly in handling the dataset's nuances. Hyperparameter tuning through grid search, guided by the f1\_weighted scoring metric, further optimized the model. This choice of metric, favoring a weighted average of the F1 score over simple accuracy, was important in preventing overfitting and ensuring a balanced consideration of both precision and recall.

The achieved accuracy rates of 0.771 and 0.763 on the full and sample datasets, respectively, are commendable, given the challenge of predicting project success at launch with limited information. This performance not only demonstrates the model's technical efficacy but also its practical value in a business context. For Kickstarter, such a model offers a strategic tool to assess project viability, guide project creators, and streamline platform operations.

In conclusion, this project represents a harmonious blend of machine learning expertise and business acumen. By meticulously addressing each step of the model development process, from data preprocessing to model optimization, a robust and reliable tool was crafted. This tool not only achieves high accuracy in its predictions but also provides actionable insights, enhancing decision-making processes for both Kickstarter and its user community.

**Model 2: Unsupervised K-prototype Clustering**

In the multifaceted domain of machine learning, the clustering of Kickstarter projects based on mixed data types offers a nuanced exploration into the potential factors that drive project success or failure. The K-prototype clustering, adept at handling mixed data types, reveals patterns that could be invaluable for Kickstarter's strategic positioning and support for project creators.

The preprocessing of this dataset involved several steps to refine and prepare the data for clustering. The null categories were filled with 'none', which allows for the inclusion of projects without a defined category and prevents the exclusion of potentially insightful data. The dimensional reduction for time-related features was performed by creating a more generalized *'year\_season'* variable. This step aids in reducing the complexity of the data while still retaining the temporal information that could influence project outcomes.

A particularly interesting creation was a new variable derived from the ratio of pledge to goal amount. Due to its high skewness, this variable was categorized into whether the goal was fulfilled and to what extent it was overfulfilled. This categorization provides a clear and immediate indication of the project's funding success relative to its goal, a likely significant factor in the overall success of the project. Concentrating on the US and GB markets, which accounts for 83% of the countries, was a strategic choice to distill the analysis to where patterns are most pronounced, affecting the majority of projects. Through exploratory data analysis, only 18 variables of high importance were retained for clustering. This step ensured that the model was not overburdened with noise and could focus on the most predictive features.

While isolation forests are typically used for anomaly detection, their performance was suboptimal even after hyperparameter tuning through a grid search. The skewness and diversity of the data posed challenges, potentially leading to swamping and masking effects where normal instances might be incorrectly identified as anomalies and vice versa. This underscores the need for robust preprocessing and the careful selection of model types for different datasets.

The manual checking and removal of outliers from highly skewed data is an example of hands-on data curation. This step, although labor-intensive, ensures that extreme values do not distort the clustering process, thus enhancing the quality of the final clusters.

Upon model selection, the K-prototype was identified as the most suitable for this mixed data type over K-means, which struggles with non-numeric data. The use of the MinMax scaler standardized the data effectively to a range of 0 to 1, making it amenable to clustering and interpretation.

The elbow method (*Image1 in appendix*), a classic means of determining the optimal number of clusters, pinpointed four as the optimal number for this dataset. This quantitatively derived insight was validated qualitatively; the clusters displayed clear boundaries when visualized through PCA (*Image2 in appendix*), indicating a successful clustering with meaningful separation. The analysis of each cluster's mode—a statistical measure that represents the most frequent occurrence of a variable—further enhanced the understanding of the distinct characteristics of each cluster:

**Cluster 0's** mode revealed failed web projects with no backers or pledges, highlighting the challenge in this category to garner support and suggesting a need for better engagement strategies.

**Cluster 1's** mode showed that even a small number of backers could lead to the success of 'Plays' projects, especially when spotlighted, underscoring the importance of community engagement and platform visibility.

**Cluster 2's** mode, representing successful hardware projects, had a higher number of backers and pledges with a lower goal, indicating that in the hardware category, conservative funding targets combined with solid backing are keys to success.

**Cluster 3's** mode, by contrast, depicted failed hardware projects with no backing, suggesting that without sufficient community support and visibility, even the most innovative projects can falter.

It was observed that successful projects typically had some level of backing and were spotlighted, suggesting that visibility and early community engagement are vital. Failed projects, on the other hand, often had no backers or pledges, pointing to a potential lack of engagement or market fit. The clustering process, therefore, has not only categorized projects based on historical outcomes but also has illuminated the path forward. For instance, the importance of setting realistic funding goals is underscored by the success of projects with modest goals. This insight suggests that Kickstarter could play a more proactive role in guiding creators to set achievable targets.

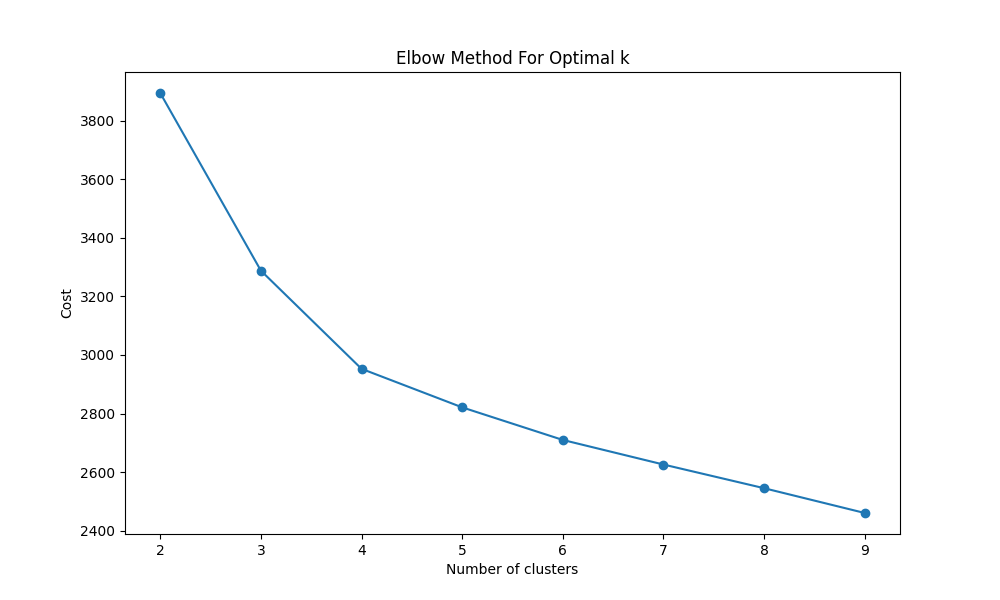
Moreover, the differentiation between categories like 'Plays' and 'Hardware' suggests that Kickstarter could offer category-specific advice and tailored support, thereby enhancing the likelihood of a project's success. The temporal aspect, while subtle, also hints at potential seasonal trends which could inform the timing of project launches.

This clustering and the subsequent insights offer Kickstarter actionable recommendations. Lower funding goals, especially in the Hardware category, appear more attainable. Engaging backers early and understanding category-specific success factors could be crucial. Moreover, the quest for a project to be spotlighted might be a worthy investment given its association with success.

In essence, this process has mapped the features that correlate with project success, equipping Kickstarter with a deeper comprehension of its platform's dynamics and offering a strategic pathway to bolster the success of its community of creators.

**Appendix:**

**Image1: Elbow method:**

****

**Image2: PCA Clustering Visualization**

**A chart of multiple colored dots

Description automatically generated with medium confidenceWhy isolation forests failed:**

https://thingsolver.com/blog/friday-talks-the-dark-horse-of-isolation-forest/