

# PREDICTIVE MODELING ANALYSIS BOOSTS KICKSTARTER PROJECT SUCCESS FORECASTING TO 75% ACCURACY

## THE CLIENT

The project involved analysing the Kickstarter platform to predict the success of new projects and find insights for creators. The goal was to classify whether a project would be successful and cluster similar projects for better targeting.

## THE CHALLENGES

- Developing a model to predict project success based only on pre-launch data.
- Handling a dataset with mixed categorical and numerical variables.
- Grouping projects to identify common characteristics between successful campaigns.

## THE APPROACH

### Phase 1: Feature Engineering for Classification

- Variable Selection: Dropped variables that represented post-launch information or were not predictive at launch, such as:
  - *Pledged Amount & Backers Count*: post-launch success metrics.
  - *Staff Pick & Spotlight*: Assigned based on initial performance and known after launch.
  - *Time-Specific Variables* (e.g., *deadline\_day*, *created\_at\_hr*): Dropped due to low predictive power and high model complexity.
- Kept variables indicative of pre-launch planning and project characteristics:
  - *Goal (converted to USD)*, *Country*, *Category*, *Create to Launch Days*, *Name Length*, *Blurb Length*.

### Phase 2: Classification Model Development

- Model Selection:
  - Chose Random Forest for its high accuracy (75%) and robustness in managing large and complex datasets, showing good performance in identifying key drivers behind successful campaigns.
  - Pre-processed data by converting categorical variables and removing missing values.
  - Converted the dependent variable into binary (1 = successful, 0 = failed) to simplify prediction.

### Phase 3: Clustering Analysis

- Model Selection:
  - Used K-Prototypes for clustering since the dataset contained both numerical and categorical variables, ensuring that all relevant features were considered without losing interpretability.
  - Employed Huang's Initialization for cluster centroids to account for data variability.
  - Formed 2 clusters to improve insight and visibility into project characteristics.

## THE RESULTS

- Classification Accuracy:
  - The Random Forest model achieved a 75% accuracy rate, successfully predicting project success based on pre-launch data.
- Clustering Insights:
  - Well-prepared projects with shorter campaign lengths, realistic goals, and effective use of Kickstarter features (e.g., spotlight) were more likely to succeed.
  - Projects launched in the summer, particularly in the hardware category, demonstrated higher success rates, indicating a beneficial seasonality effect.

### Cluster Characteristics    Key Insights

Cluster 0	Projects with modest funding goals, longer campaigns, fewer backers, which launched fast with clear names and detailed descriptions had lower success
Cluster 1	Projects with more backers, higher funding goals, shorter campaigns, launched after intensive preparation with longer project names and clear descriptions (usually in hardware category) had better success