McGill | DESAUTELS

INSY-662 Data Mining

Predictive Modeling and Clustering Insights for Kickstarter Success

Abdulrahman AROWORAMIMO

Professor
Prof. Elizabeth Han

November, 2023

Contents

List of Figures

List of Tables

1 Introduction

Kickstarter, a renowned crowdfunding platform, operates on a unique premise where backers support projects of interest through financial pledges. The platform employs an "all or nothing" model, where each project sets a financial goal, and its outcome is categorized as either failed or successful based on whether the goal is achieved. The high stakes associated with this model underpin the importance of a predictive model that can accurately forecast a project's fate. Such a tool would be invaluable for project creators, allowing them to assess the suitability of Kickstarter for their endeavor before committing, ultimately saving time and resources. Moreover, delving into the diverse attributes of past projects can provide creators with insights to strategically position their initiatives for success. Recognizing this potential, this project aims to develop a classification model capable of predicting a project's success or failure. Additionally, the project seeks to employ clustering techniques on historical data to uncover inherent patterns and trends among past Kickstarter projects, offering creators a deeper understanding to enhance their project planning and execution strategies.

2 Predictive Modeling

2.1 Data Pre-processing

All variables were examined to determine predictors and identify columns for removal. First, identifier columns, including `id` and `name`, were dropped. `Currency` was excluded as it doesn't provide any relevant information beyond what the `country` variable offers. The `goal` variable was transformed into a new variable using the `conversion_rate` column to ensure comparability across all observations. Subsequently, both `goal` and `conversion_rate` columns were dropped. `Name_len` and `blurb_len` were also removed since their cleaned versions exist in the dataset and will be utilized instead.

Additionally, all date-time variables at a yearly granularity were removed. This decision stems from the consideration that variables at a yearly granularity are not likely to provide informative insights for future predictions. Future projects cannot adopt past years, and applying the model to upcoming projects is the objective of this project. While time series analysis could provide insights, it falls beyond the scope of this project. This filtering ensures that the model focuses on features more likely to contribute to predictive accuracy for future Kickstarter projects.

Furthermore, considering the project's aim to develop a model predicting a project's fate at its launch, columns containing information available only after the project's fate has been determined were excluded. These variables include `spotlight`,`staff_pick`, `pledged`, `backers_count`, `state_changed_at`, `state_changed_at_weekday`, `usd_pledged`, `static_usd_rate`, `state_changed_at_month`, `launch_to_state_change_days`, `state_changed_at_day`, `state_changed_at_yr`, and `state_changed_at_hr`. Also, `created_at`, `launched_at`, and `deadline` columns were dropped because their more granular forms are in the dataset and will be used instead. After removing the columns that were not useful given the goal of the project, null values were dropped from the dataset. Finally, all observations that have states other than `failed` and `successful` were dropped. Techniques such as scaling and PCA were not used because the algorithm used is not sensitive to feature scale and thrives with high dimensional datasets.

2.2 Modeling Process

Gradient boosted trees algorithm was selected for its good performance on high-dimensional datasets. Random forest was also considered, but the performance of gradient boosting was slightly better. To find the optimal mix of hyperparameters, randomized search CV was used to determine the combination of `max_depth`, `n_estimators`, `learning_rate`, and `subsample` that maximizes accuracy. The optimal mix

was found to be 0.1 learning_rate, 5 max_depth, 157 n_estimators, and 1 subsample. After hyperparameter tuning, threshold moving was applied to further enhance accuracy, given the slight imbalance in the dataset. The optimal threshold was determined using stratified kfold; this method was chosen because a stratified Kfold is more adept at handling imbalance than kfold. The average best threshold was found to be about 48% with an average accuracy of 78%. After the optimal threshold was determined, the optimal hyperparameters and threshold were applied to the entire dataset, and the accuracy score was 86% as depicted in appendix A.1.

2.3 Model Interpretation

Utilizing gradient boosting for feature importance analysis, key predictors such as `goal_converted`, `create_to_launch_days`, `category_web`, `category_software`, `name_len_clean`, `created_at_day`, `launched_at_hr`, `deadline_hr`, `launch_to_deadline_days`, and `blurb_len_clean` were identified as the most influential features, as illustrated in appendix A.2.

However, while feature importance provides insight into the importance of each feature, it doesn't inherently reveal the relationship between these features and the target variable. To delve into this aspect, SHAP (SHapley Additive exPlanations) was employed. The SHAP summary showcases how each of the top features impacts the project outcome, providing a more nuanced understanding of their contributions (see appendix A.3). From the plots, the following insights were discerned, zeroing in on the top five features:

- The amount of the project's goal in USD is negatively correlated with the probability of a project's success, i.e., projects with higher goals are more likely to fail, and vice versa.
- Projects that have a low number of days between creation and launch are more likely to fail.
- Projects that belong to the web and software categories are more likely to fail.
- Projects with longer name lengths are more likely to succeed.

Note that these insights are based on this dataset and the model's predictions. They may or may not be representative of the actual situation.

3 K-Prototypes Clustering

K-Prototypes was chosen considering the presence of categorical variables in the dataset. The optimal number of clusters, determined using the elbow method, was found to be five (refer to appendix A.4). Due to the high dimensionality of the dataset, a subset of features was selected for clustering to optimize performance. Seven features were chosen based on their potential for valuable insights. While clustering with all features is ideal, it is not computationally efficient, justifying the decision to cluster a subset. The selected features include `goal_converted`, `staff_pick`, `create_to_launch_days`, `category`, `name_len_clean`, `blurb_len_clean`, and `launch_to_deadline_days`.

Before clustering, the numeric features, excluding the binary feature, were standardized to ensure that each feature contributes equally. However, to facilitate a more intuitive interpretation of the resulting clusters, these features were reverse-scaled. This process restores the original scale of the numeric features, providing a clearer understanding of the typical values associated with each cluster centroid (see appendix A.5). The interpretation of the cluster centroids is then based on the familiar, unscaled values, enhancing the practical relevance of the results. To gain deeper insights into each cluster, every observation in the original dataframe was labeled with its corresponding cluster, enabling a comprehensive analysis that extends beyond the information provided by cluster centroids (see appendix A.6).

3.1 Cluster Analysis

The resulting five clusters exhibit distinct properties, although some overlaps in their general characteristics are observed. Before delving into each cluster, some overarching insights were discerned. At a high level, it is evident that staff picks are both highly unlikely and uncommon. The likelihood decreases even further for high-goal projects. Projects in the hardware category are more likely to receive staff picks as long as they have reasonable goals and timelines, a crucial factor given the positive impact staff picks have on project success.

- **Cluster 0:** Predominantly comprised of hardware projects with a moderate goal, this cluster exhibits more staff picks than other groups, except for cluster 2, which is also dominated by hardware projects. The projects in this cluster have the highest duration between project creation and launch, along with a moderate timeframe between project launch and the funding deadline. Their name and description lengths are moderate compared to other clusters. About 0.03

- **Cluster 1:** Dominated by web projects with a low occurrence of staff picks, this cluster has the second-lowest goals, the shortest duration between creation and launch (except for cluster 4 where the duration is markedly shorter), and the lowest name length. They have the shortest duration between project launch and the funding deadline, indicating a relatively quick turnaround. About 39% of projects fall into this category, of which 28% are successful.

- **Cluster 2:** This cluster comprises projects with the highest occurrence of staff picks. Similar to cluster 0, they are predominantly hardware projects but stand out with a significantly lower number of days from project creation to launch. Additionally, they have slightly higher description and name lengths. Interestingly, they exhibit a lower number of days between launch and the deadline. Their goal is the lowest across all the clusters and significantly lower than cluster 0. The differences exhibited between this cluster and cluster 0 may be the reason this group of projects has higher staff picks. About 40% of projects fall into this category, of which 41% are successful, making it the most popular and successful group.

- **Cluster 3:** Similar to cluster 1, this cluster is dominated by web projects but with a much higher goal. Surprisingly, the higher goal does not significantly affect the occurrence of staff picks, resulting in a marginally lower staff pick rate compared to cluster 1. Projects in this cluster also have a significantly higher number of days between project launch and the funding deadline, the highest across all clusters, and slightly higher duration between creation and launch than cluster 1. About 16% of projects are in this group with a 21% success rate.

- **Cluster 4:** Characterized by the highest goals and a complete absence of staff picks, this cluster is also predominantly composed of hardware projects. Projects in this cluster have the lowest name and description lengths and the shortest duration between project creation and launch. About 0.0003% are in this group with a 0% success rate. It is important to note that only four projects fall in this group, making them outliers. However, they still demonstrate how a high goal can negatively impact a project's success.

In summary, creators aiming for staff picks should note that hardware projects stand out as a favorable choice for staff picks; however, a high goal deters staff. They should also be mindful of the project duration from creation to launch, as a well-thought-out timeline appears to contribute to staff pick selections. Evidently, projects that match the characteristics of projects in cluster 2 have been the most successful thus far as depicted in appendix A.7.

A Appendix

A.1 Final Model Performance with Threshold of 0.48

Table 1: Final Model Performance with Threshold of 0.478

	Precision	Recall	F1-Score	Support
Failed	0.88	0.92	0.90	8218
Successful	0.81	0.74	0.78	3963
Accuracy			0.86	12181
Macro Avg	0.85	0.83	0.84	12181
Weighted Avg	0.86	0.86	0.86	12181

A.2 Feature Importance

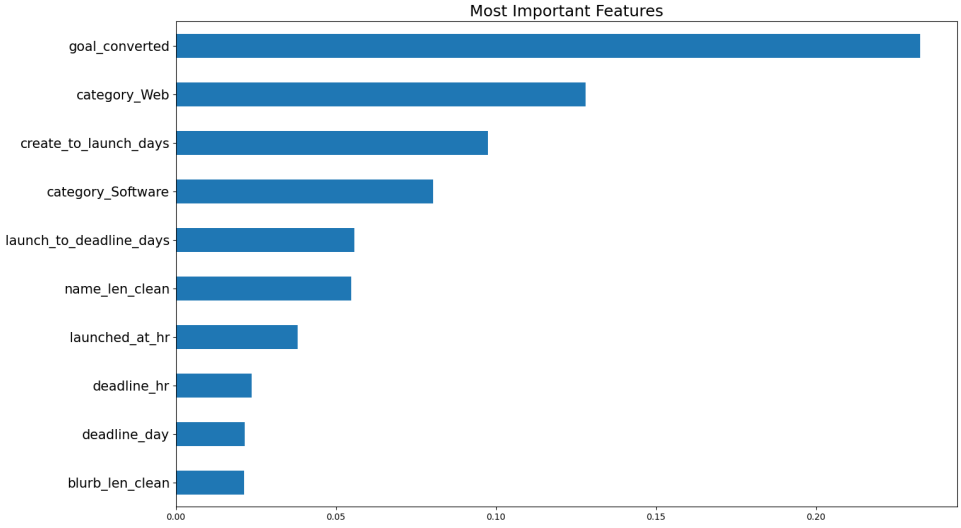

Figure 1: Feature Importance

A.3 SHAP Summary Plot

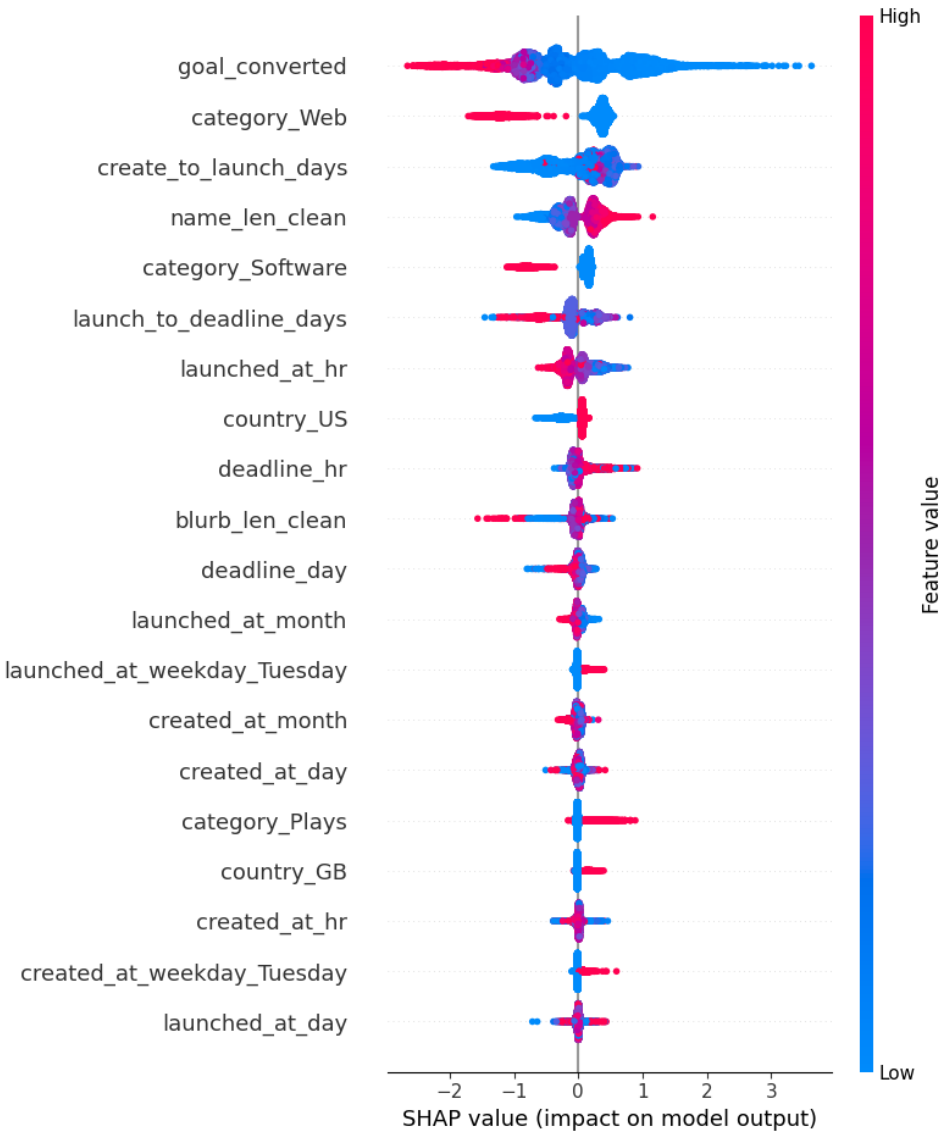

Figure 2: SHAP Summary Plot

A.4 Elbow Plot

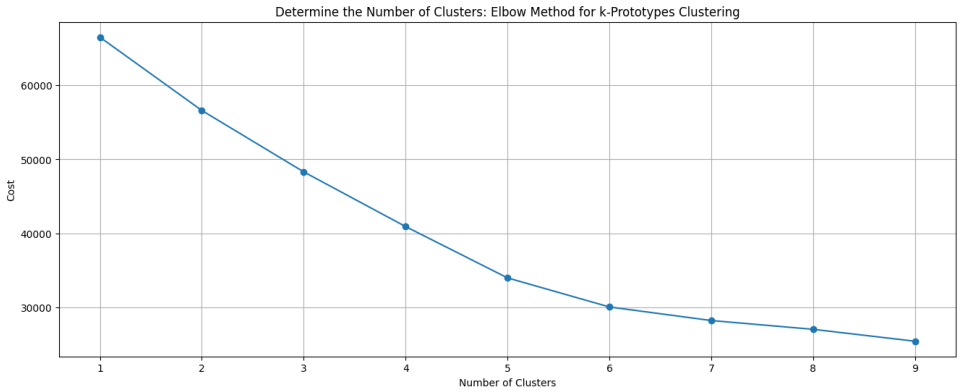

Figure 3: Determine the Number of Clusters: Elbow Method for k-Prototypes Clustering

A.5 Cluster Centroids

Table 2: Cluster Centroids

Cluster	Goal Converted	Create to Launch	Name Length	Launch to Deadline	Blurb Length	Staff Pick	Category
0	74296.44	475.01	5.71	35.70	12.84	0.13	Hardware
1	50125.67	25.52	3.32	29.32	11.54	0.09	Web
2	45335.85	36.96	6.99	30.52	14.60	0.15	Hardware
3	143768.74	30.54	4.64	56.21	12.94	0.07	Web
4	56606002.60	12.00	4.50	39.75	10.50	0.00	Hardware

A.6 Cluster Proportions and Project Success Rates

Table 3: Cluster Proportions and Project Success Rates

Cluster	Proportion	Project Success Rate
0	0.034890	Failed: 69.88% Successful: 30.12%
1	0.393400	Failed: 71.87% Successful: 28.13%
2	0.403661	Failed: 58.17% Successful: 41.83%
3	0.167720	Failed: 78.95% Successful: 21.05%
4	0.000328	Failed: 100%

A.7 Visual:Cluster Proportions and Project Success Rates

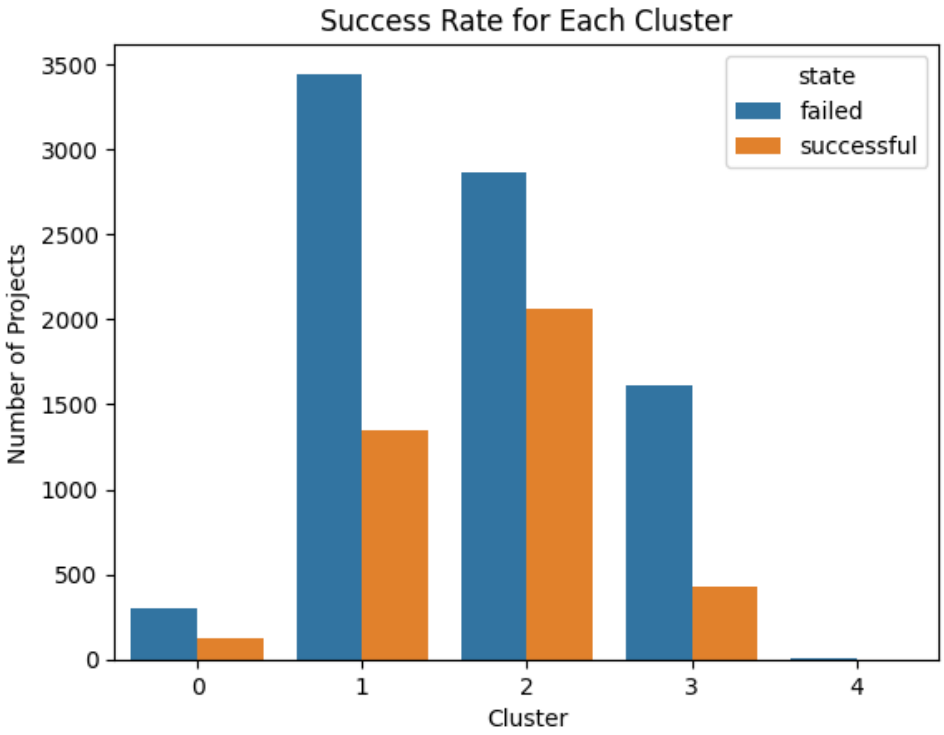

Figure 4: Cluster Proportions and Project Success Rates