INSY 662 – Individual Project

Rodrigo Castro

Predictive Analysis and Cluster Identification in Kickstarter Projects

December 2023

1. Introduction

This project aims to develop a predictive model to determine the success or failure of Kickstarter projects at the time of their launch. Utilizing a comprehensive dataset from Kickstarter, the study employs various machine learning techniques, including Random Forest, Logistic Regression, Gradient Boosting, KNN, and Neural Networks, to analyze and predict project outcomes. Additionally, the project incorporates K-means clustering to identify distinct groups within the Kickstarter dataset, thereby uncovering patterns and trends that characterize successful crowdfunding campaigns.

2. Data pre-processing

The Kickstarter's predictors chosen for the classification model —category, name length, blurb length, days from creation to launch, days from launch to deadline, and the USD goal amount— are all available at the moment a project is launched. These features were carefully selected based on their significance in influencing project outcomes without exhibiting collinearity, as detailed in the correlation matrix from Appendix I. For this reason, no predictors were taken out during the modeling.

For the pre-processing steps, the 'goal_usd' column was created by converting all project goals into USD using the exchange rate, ensuring currency consistency. Only projects identified as "successful" or "failed" were included, for that reason the target variable "state" was binarized. On the other hand, missing rows and duplicates were removed, and the 'category' variable was dummified. For the clustering phase, the 'country' variable was binarized to distinguish U.S. projects, which represent more than 50% of the dataset, and categories with a frequency lower than 500 were grouped. Finally, for both models the dataset was cleansed of outliers using an Isolation Forest to focus on the most representative samples.

3. Classification Modeling

In the classification phase, the dataset was split into training and test sets, with numerical features standardized for model compatibility. To look for the appropriate algorithm, various models were trained to predict Kickstarter project outcomes, the result are summarized in Appendix II. The Accuracy across models suggest that while the models are relatively good at identifying successful projects, there is still room for improvement, especially in precision and recall. For that reason, the highest performance model, Gradient Boosting, underwent hyperparameter tuning to obtain the most reliable model for our prediction.

Model	Accuracy	Precision	Recall	F1 Score
RandomForest	0.7317	0.6232	0.4969	0.5529
LogisticRegression	0.7194	0.6106	0.4405	0.5118
GradientBoosting	0.7471	0.6649	0.4890	0.5636
KNN	0.6808	0.5257	0.4484	0.4840
NeuralNetwork	0.7333	0.6251	0.5023	0.5570
GradientBoosting (Tuned)	0.7484	0.6599	0.5086	0.5745

4. Cluster Analysis

The clustering process began with the standardization of the 'goal_usd' feature, to ensure equitable cluster analysis. Utilizing the elbow and silhouette methods, as seen in Appendix III and IV respectively, the optimal numbers of cluster that correctly segment the Kickstarter data was seven. Therefore, each project was assigned a cluster label for the analysis of distinctive characteristics. For instance, the success rate and geographic distribution varied across clusters, indicating different success probabilities and regional concentrations. To better interpret the clusters, Principal Component Analysis (PCA) was applied to reduce the data to two principal components for a visual representation, as seen in Appendix V. Following is a brief description of each cluster.

- **Cluster 1:** Low success rate of around 9%, with a moderate presence of U.S.-based projects and lower financial goals, and primarily includes Web projects.
- **Cluster 2:** High success rate of around 42%, has a strong U.S. project presence, moderate financial goals, and is exclusively focused on Hardware projects.
- **Cluster 3:** Low success rate of around 10%, has a strong U.S. project presence, stands out with very high financial goals and includes projects related to Apps and Gadgets.
- **Cluster 4:** Moderate success rate of around 42%, with a high presence of U.S.-based projects and a broad range of financial goals, with a notable presence of Plays projects.
- **Cluster 5:** Best-performing group, highest success rate of nearly 50%, with a high proportion of U.S.-based projects, lower financial goals, and a variety of project types, predominantly categorized as Others.
- **Cluster 6:** Smallest group, lowest success rate of around 3.6%, contains projects with highest financial goals, indicating high-cost projects, all from the U.S.
- **Cluster 7:** High success rate of around 17%, significant number of U.S.-based projects, and lower financial goals, predominantly comprising Software projects.

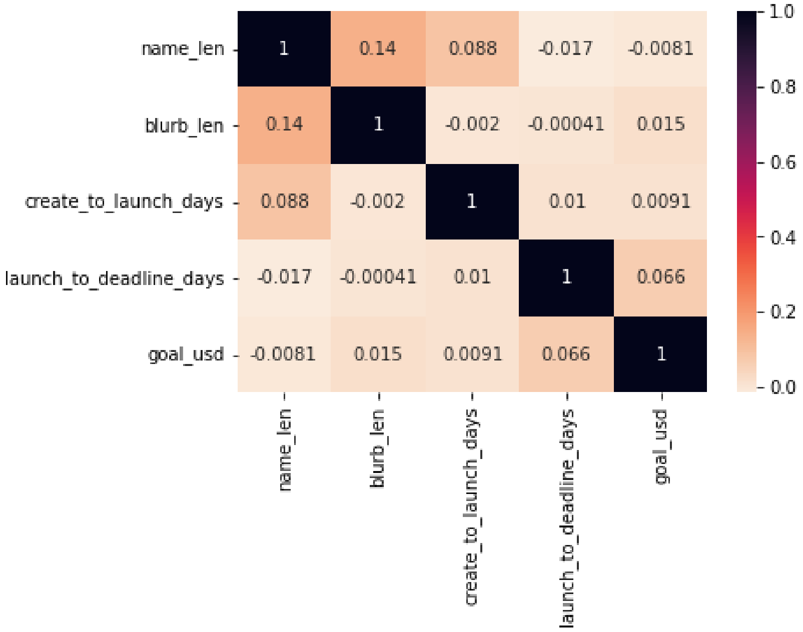
5. Business Insights

In a business context, the classification and clustering models derived from Kickstarter data offer valuable insights for both project creators and backers. The classification model predicts the likelihood of a project's success at launch, enabling creators to optimize key aspects like funding goals and marketing strategies, while backers can use it to assess investment risks. This aids in making informed decisions, potentially increasing the success rate of projects and the return on investment.

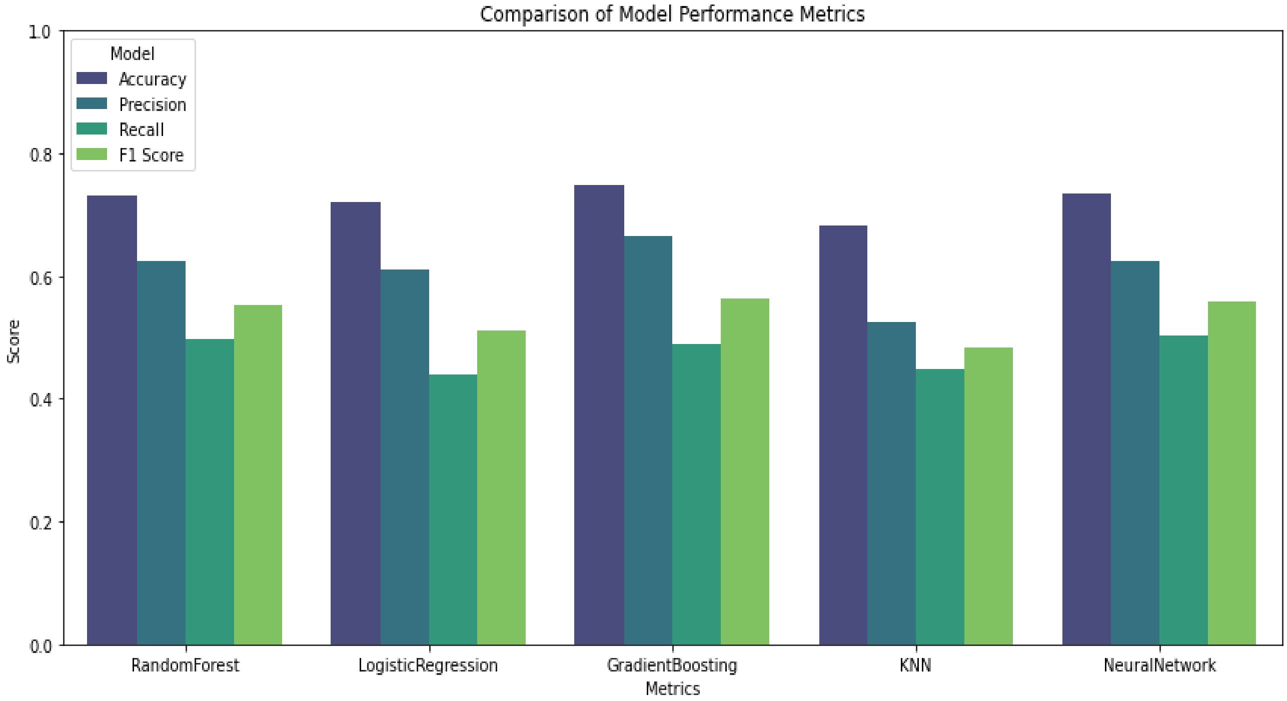
The clustering model further enhances understanding by identifying patterns within the Kickstarter ecosystem. It segments projects into groups with shared characteristics, revealing successful niches and riskier ventures. For Kickstarter as a platform, these insights are crucial for targeted support and marketing efforts, helping to improve overall project success rates and user satisfaction. This dual-model approach provides a comprehensive toolkit for strategic planning and decision-making in the crowdfunding domain.

6. Appendix

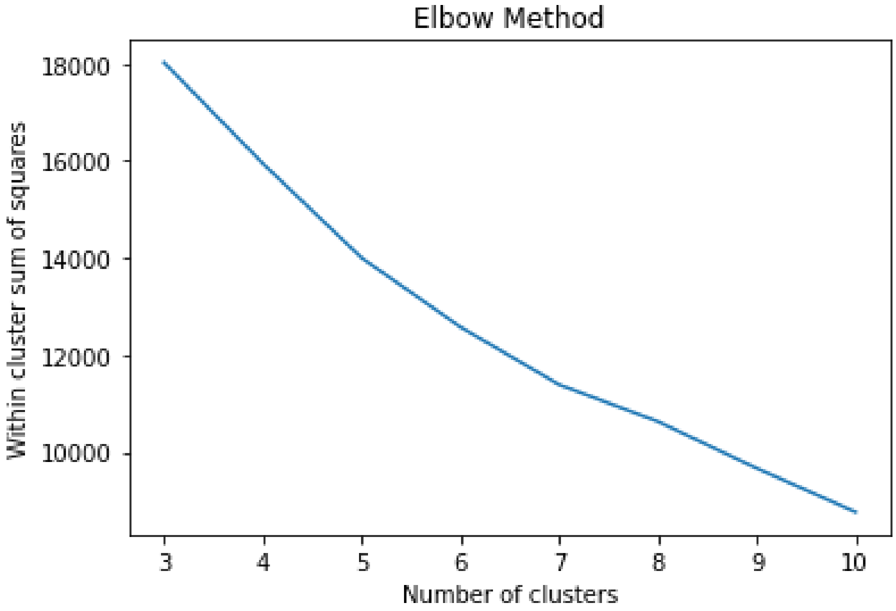
I. Correlation Matrix - Numerical Variables (Classification Model)



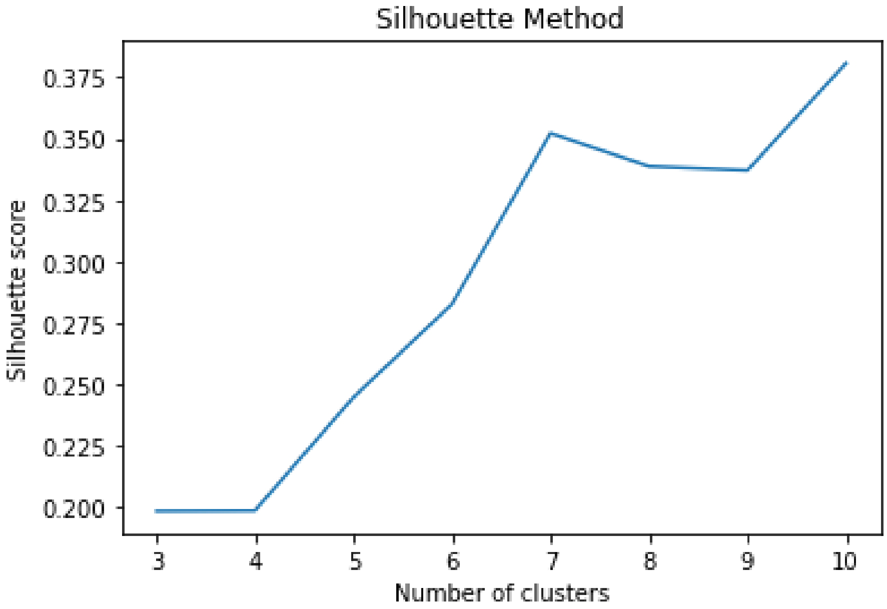
II. Classification Model Performance Comparison



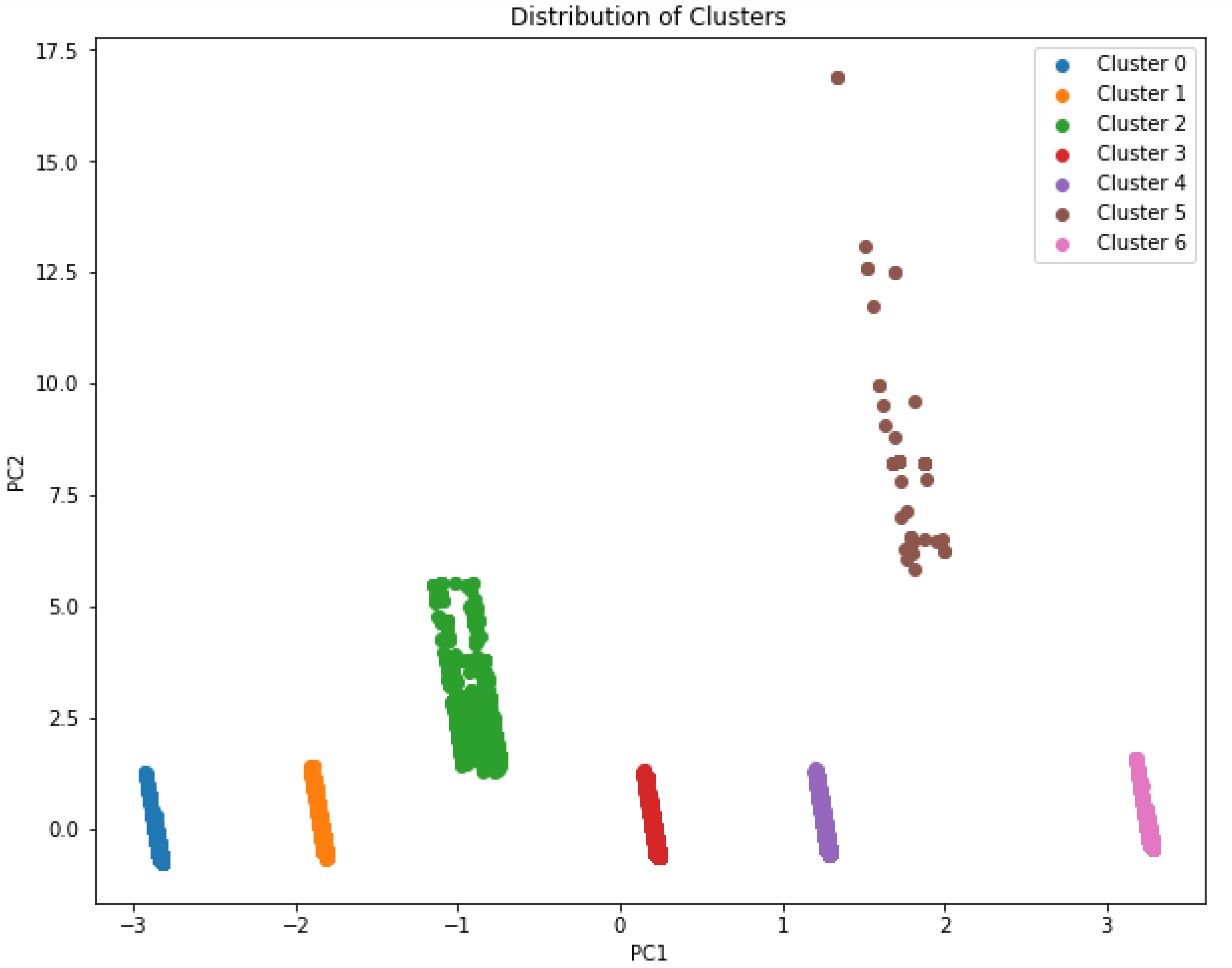
III. Within Cluster Sum of Squared



IV. Silhouette Score for each Cluster



V. Distribution of Clusters applying Feature Reduction


Distribution of Clusters
PC2
PC1
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6