Ethan Pirso
260863065
December 4, 2023

INSY 662 Individual Project Report

Navigating the dynamic world of crowdfunding, we set out to create models that predict Kickstarter campaign outcomes and decode the clustering of projects. We employed a Random Forest Classifier for its ability to discern successful campaigns with precision, and we complemented it with a DBSCAN clustering model to map out the landscape of Kickstarter projects. These models together provide a dual lens: one that forecasts success with targeted accuracy, and another that groups campaigns into strategic insights.

Classification Model

In the quest to predict the success of Kickstarter campaigns, we developed a classification model that prioritized recall over accuracy. This focus on recall is especially pertinent in the business context of crowdfunding, where the cost of missing a potentially successful campaign (a false negative) could mean forgoing a lucrative investment opportunity. Conversely, identifying successful campaigns (true positives) is of greater value to both investors seeking promising ventures and creators aiming to gauge their project's viability.

In tailoring our predictive model, we excluded post-launch predictors to ensure that our insights remain actionable from the outset of a campaign. Then, we removed outliers identified by an Isolation Forest. We engineered features such as 'goal_usd_rate', providing a standardized

financial target irrespective of currency fluctuations. After this, we streamlined our predictors, removing those with undue correlation and high variance inflation factors.

The best performing model, Random Forest Classifier, was rigorously tuned to enhance its predictive power (Figure 1). By carefully adjusting hyperparameters, we cultivated a model that balanced complexity with generalization. This fine-tuning led to impressive performance metrics, notably a recall rate of 0.78 on the training dataset, which slightly improved to 0.79 on the grading sample. Such consistency is indicative of the model's reliability and its capacity to generalize well to unseen data.

From a business standpoint, the insights drawn from the model are multi-fold. For one, platform operators like Kickstarter can deploy the model to flag campaigns with high success potential, offering them additional support or featured spots to maximize visibility. For creators, understanding the model's feature importances (Figure 2) can guide them in designing their campaigns to align with success patterns. For investors and backers, the model acts as a decision-support tool, indicating where to direct their attention and funds.

Clustering Model

For clustering, we implemented an autoencoder followed by DBSCAN. The autoencoder, trained once and saved for consistency, ensured that the encoded data fed into the clustering algorithm reflected the inherent patterns without random reinitialization in each run. DBSCAN was chosen over K-Means after a silhouette score comparison, with DBSCAN achieving a score of 0.54, signifying well-defined clusters and good evidence of the fit to the reality. Through the PCA visualization (Figures 3, 4), the DBSCAN clustering algorithm successfully identified 5 distinct

segments within the Kickstarter campaign data. The clusters formed were not only well-demarcated but also showed specific characteristics.

Upon analyzing the categorical and numerical features (Figures 5, 6), we uncovered distinctive patterns within the clusters. Cluster 1, for instance, stands out with a high concentration of web and app-related campaigns, which signals a cluster of digitally-focused initiatives. This inclination towards technology is further underscored by the numerical analysis showing this cluster's relatively moderate financial targets, suggesting a blend of ambitious yet achievable projects with realistic funding expectations.

Meanwhile, Cluster 3 is discernible by its strong affiliation with technology and design, coupled with elevated average goals and pledges. This pattern hints at a cluster composed of high-stake, innovative campaigns that are likely to be capital-intensive, reflecting a segment where backers are drawn to cutting-edge ideas with transformative potential.

In contrast, Cluster 4's landscape is dominated by creative arts, including theater, music, and festivals. These campaigns typically showcase lower financial thresholds yet exhibit a higher frequency of successful outcomes. This suggests a cluster marked by projects with attainable aspirations and possibly a robust network of community support, resonating well with their audience and securing the necessary funding with greater ease.

In summary, these insights not only inform Kickstarter on the diversity of campaigns hosted on their platform but also enable them to tailor their support and marketing efforts. Moreover, our model can guide campaign creators in understanding the competitive landscape and setting realistic goals based on cluster-specific benchmarks.

Appendix

Model	Accuracy	Recall	Precision	ROC-AUC
RandomForestClassifier	0.735786	0.777886	0.553236	0.817506
GradientBoostingClassifier	0.781088	0.567515	0.675991	0.831865
Ensemble Model	0.773183	0.493151	0.688525	0.828971
LogisticRegression	0.758589	0.443249	0.668142	0.801892
SVC	0.75342	0.385519	0.682842	0.767001

Figure 1: Table containing performance metrics for all the models trained. Models are ordered by recall, with Random Forest Classifier having the highest.

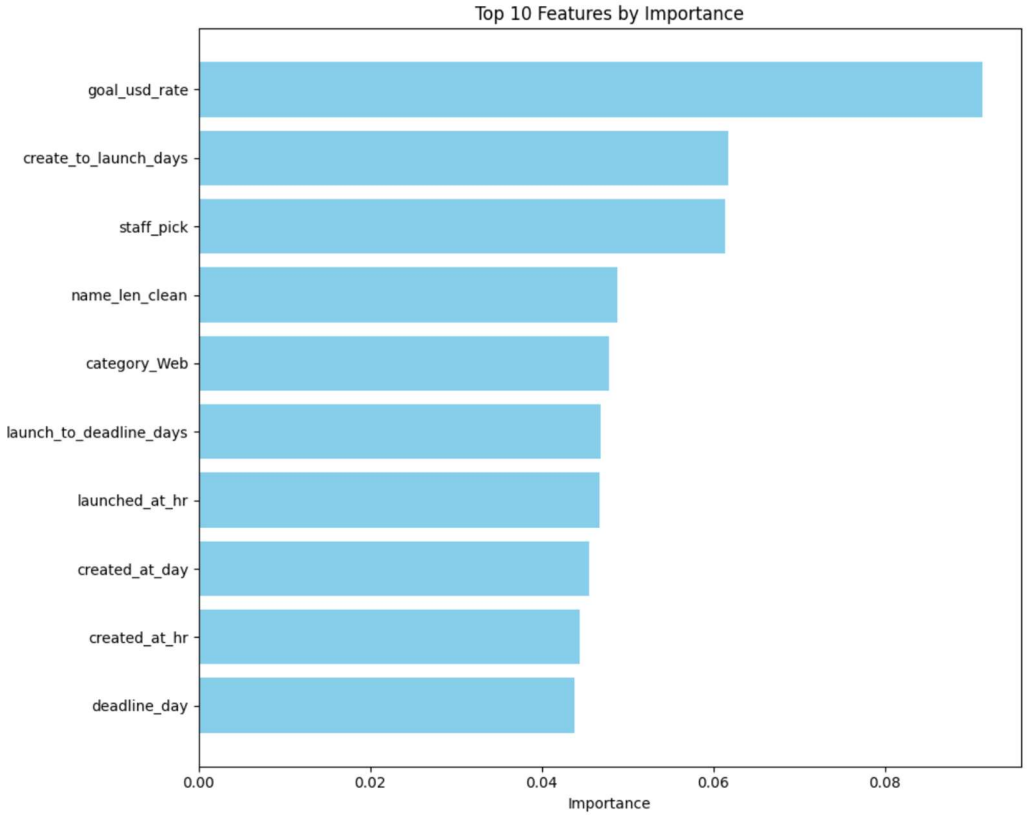

Figure 2: Plot of the top ten features ordered by importance.

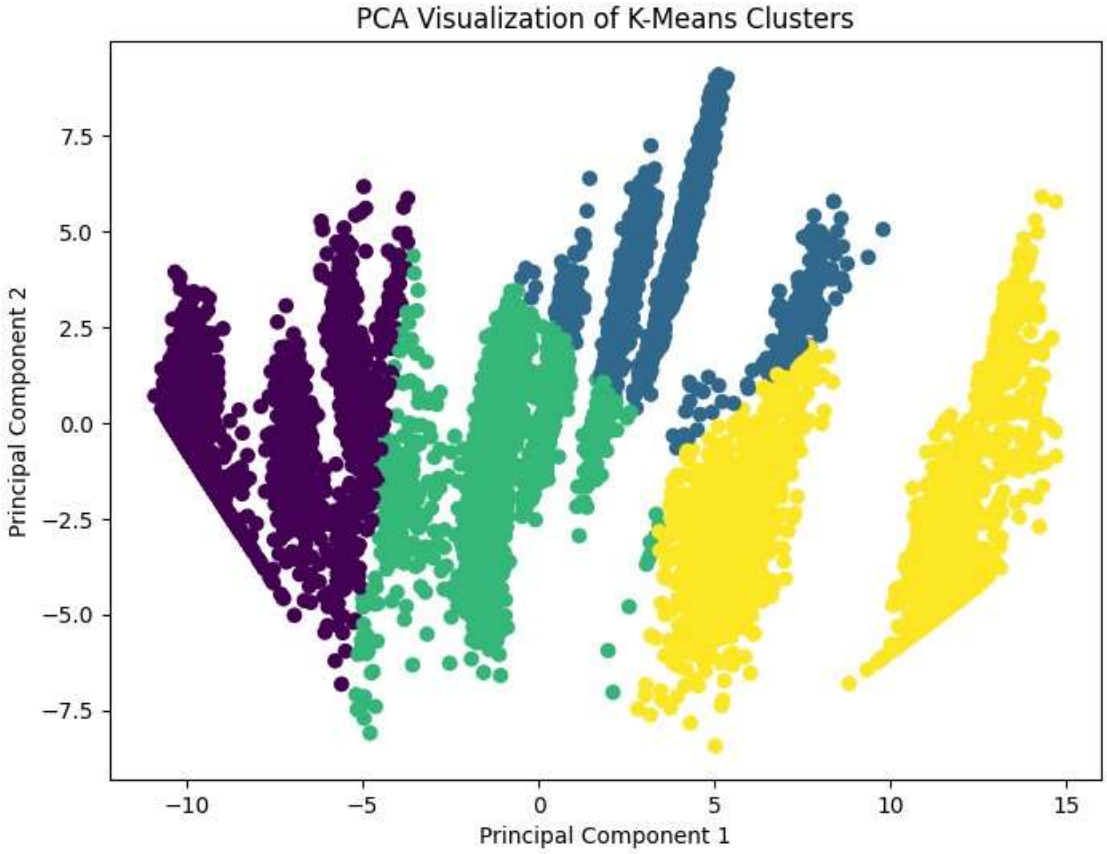

Figure 3: PCA visualization of K-Means clusters. PCA was performed on encoded data, silhouette score = 0.53.

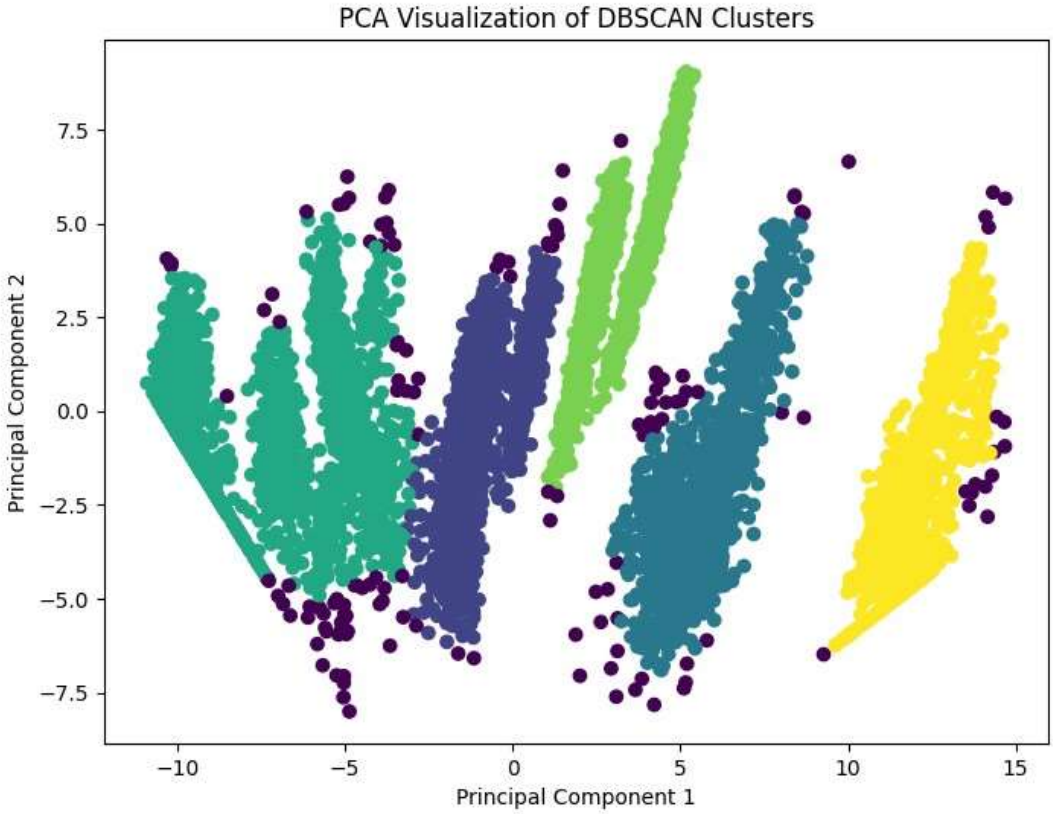

Figure 4: PCA visualization of DBSCAN clusters. PCA was performed on encoded data, silhouette score = 0.54.

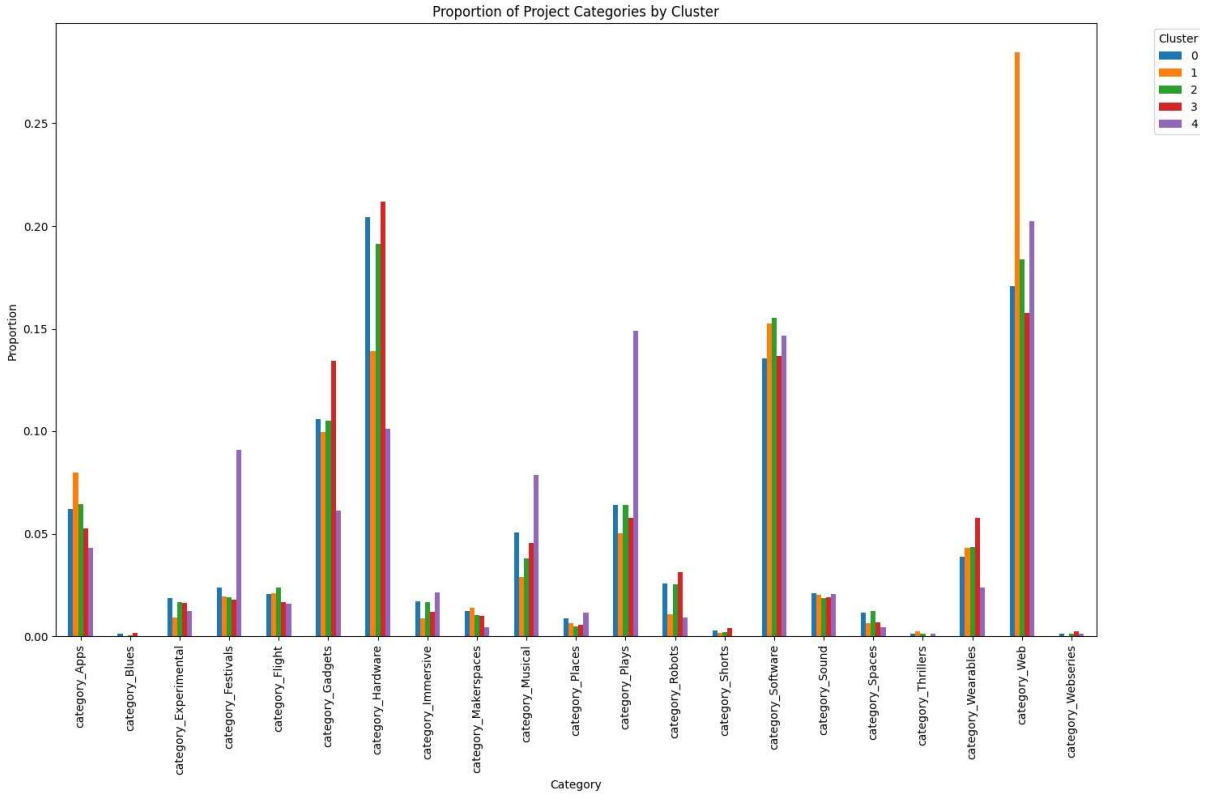

Figure 5: Feature analysis of project ‘category’ by DBSCAN cluster.

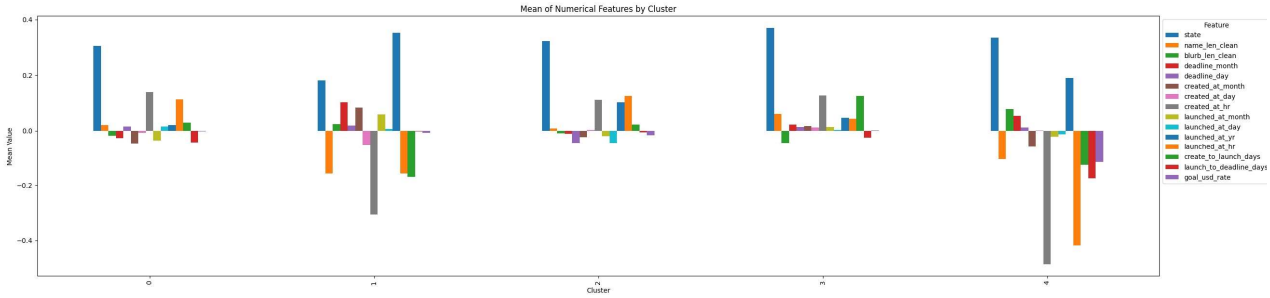

Figure 6: Feature analysis of numerical features by DBSCAN cluster.