DESAUTELS Faculty of Management
Faculté de gestion

Kickstarter Data Analysis

INSY 662: Data Mining and Visualization

Individual Project Report

Submitted by:

Yash Sethi (261208170)

Fall 2024

Business Usecase

Kickstarter is a platform that allows creators to bring their ideas to life through crowdfunding. Now, imagine if Kickstarter could provide backers with success or failure predictions for projects before they pledge their support. By knowing which projects have a higher probability of success, backers might be more inclined to contribute larger amounts to a greater number of projects. Our goal is to analyze the characteristics of these projects and predict their likelihood of success or failure.

Task 1: Classification Model

Data Preparation and Feature Engineering The dataset contains fields that outline various characteristics of the projects. The 'state' field, which indicates the success or failure of a project, was used as the target variable for training the classification model. The initial step involved preparing the data to be input into various models:

- Rows where the 'state' field was not 'successful' or 'failed' were removed.
- Columns were dropped from the dataset for several reasons. These included those determined after the project's outcome (e.g., *backers_count*, *pledged*, *spotlight*, etc.), those insignificant for predicting success (e.g., *project_name*, *USD_rate*), and datetime columns that were not useful for prediction. Additionally, *disable_communication* was removed as it contained only a single value (*False*) across all rows.
- For feature engineering, several new features were created. The significant features are i) ***months_diff_bw_create_launch*:** The time taken from the creation of the project to its launch. ii) ***months_diff_bw_launch_deadline*:** The time difference between the project's launch and its funding deadline. iii) ***goal_usd*:** The standardize goal amount for the projects in USD for fair comparison.
- Dummy variables were created for categorical fields and integrated into the main dataset.
- Additionally, anomalies were removed from the dataset using the Isolation Forest algorithm(contamination factor (C) = 0.05) to prevent negative impact on model performance.

Next, exploratory data analysis (EDA) was conducted to evaluate predictors' individual predictive strength. (Please refer to the Jupyter Notebook for the EDA results.)

Model Development Ran various iterations of *Logistic Regression*, *Decision Tree Model*, *Random Forest Model* and *Gradient Boosting Model*. Further, performed feature selection for each model by retraining only the important features(Gini coeff. gr than 0.005). Subsequently, hyperparameter tuning was performed for each model using Grid Search technique to optimize their performance. Finally, the models' performances were evaluated and summarized using key metrics to do comparative analysis.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.743	0.819	0.698	0.754
Decision Tree Model	0.716	0.733	0.780	0.756
Random Forest Model	0.707	0.711	0.808	0.756
Gradient Boosting Model	0.778	0.786	0.827	0.806

Table 1: Comparison of Regression Models Based on Performance Metrics

Clearly, the **Gradient Boosting Model** achieved the highest performance, including an accuracy of 78%, and an F1-score of 80.6%, making it the most effective model for predicting Kickstarter project outcomes. Other models showed competitive performance in certain individual metrics; however, their overall performance fell short compared to Gradient Boosting.

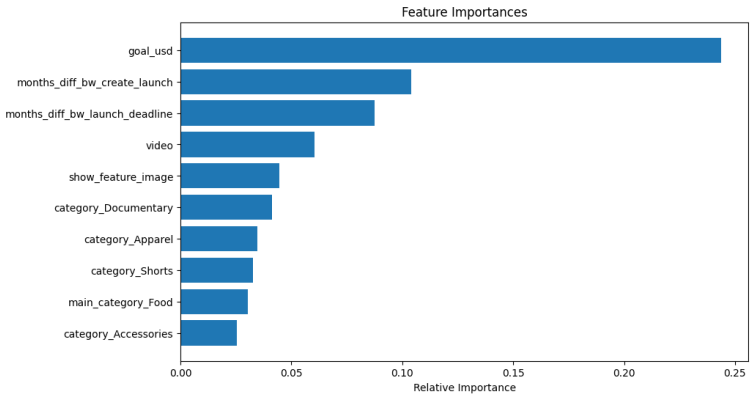
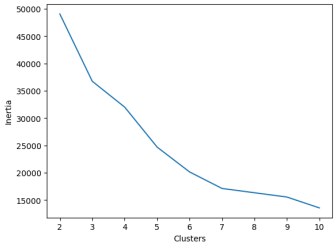


Figure 1: Feature Importance for the Gradient Boosting Model

Business Interpretation The analysis highlights key factors driving project success, such as lower *goal_usd*, the inclusion of videos and images in descriptions, and recent launch dates, which show higher success rates. The developed model can serve as a predictive tool, allowing creators to input project details and receive insights on success likelihood, helping them refine strategies. Kickstarter can also use these insights to recommend improvements to creators.

Task 2: Clustering Model:

The preprocessing steps for the unsupervised model mirrored those of previous models, including the Isolation Forest algorithm with a contamination factor of 0.05. Features associated with successful projects were selected based on feature importance from Random Forest model against *usd_pldeged*, identifying the top 4 numerical predictors identified were *usd_pledged*, *backers_count*, *goal_usd*,and *name_len_clean*. *state* was included as well. The optimal number of clusters, determined using the elbow method, was 5.



Cluster	Average Silhouette Score
Cluster 1	0.411
Cluster 2	0.311
Cluster 3	0.204
Cluster 4	0.391
Cluster 5	0.432

Figure 2: Comparison of Elbow Method and Silhouette Scores for Clustering Analysis

Cluster Insights

i) **Cluster 1 and Cluster 4** represents failed projects. The main difference between them lies in their goals with Cluster 1 having an average goal of **15,900 USD**, and Cluster 4 has a significantly higher average goal of **250,217 USD** which shows that cluster. ii) **Cluster 3, vs. Cluster 2, and Cluster 5:** These belong to successful projects. However, Cluster 3 stands out with significantly higher values for the number of backers, USD pledged, and USD goal compared to the other clusters. iii) **Cluster 2 vs. Cluster 5:** Although Cluster 2 has a higher number of backers and a larger amount pledged compared to Cluster 5, the most distinguishing factor is the average clean name length of the projects, which is higher for Cluster 2.

Potential Business Impacts: Projects with lower goals are more likely to succeed compared to projects with high goals. Creators should set realistic funding targets. Longer project names correlate with higher success compared to shorter names, emphasizing the importance of engaging descriptions. Kickstarter can use these insights to recommend goal adjustments and optimize visibility for high-potential campaigns based on clustering patterns.